Prompting Future Driven Diffusion Model for Hand Motion Prediction

Bowen Tang[1], Kaihao Zhang[2✉], Wenhan Luo[3✉], Wei Liu[4], and Hongdong Li[1]

[1] Australian National University
[2] Harbin Institute of Technology, Shenzhen
[3] The Hong Kong University of Science and Technology
[4] Tencent

Abstract. Hand motion prediction from both first- and third-person perspectives is vital for enhancing user experience in AR/VR and ensuring safe remote robotic arm control. Previous works typically focus on predicting hand motion trajectories or human body motion, with direct hand motion prediction remaining largely unexplored - despite the additional challenges posed by compact skeleton size. To address this, we propose a prompt-based Future Driven Diffusion Model (PromptFDDM) for predicting hand motion with guidance and prompts. Specifically, we develop a Spatial-Temporal Extractor Network (STEN) to predict hand motion with guidance, a Ground Truth Extractor Network (GTEN), and a Reference Data Generator Network (RDGN), which extract ground truth and substitute future data with generated reference data, respectively, to guide STEN. Additionally, interactive prompts generated from observed motions further enhance model performance. Experimental results on the FPHA and HO3D datasets demonstrate that the proposed PromptFDDM achieves state-of-the-art performance in both first- and third-person perspectives.

Keywords: Hand motion prediction · Generative models · Diffusion model

1 Introduction

Motion prediction underpins numerous computer vision applications, including autonomous driving [50] and human-robot interactions [33, 34]. Anticipating near-future human motions enables a comprehensive understanding of human intentions and informs automated decision-making. As the primary interface with the world, predicted hand motions enhance interactions but remain relatively unexplored. Compared to human body skeletons, hand skeletons are more compact and change more rapidly, making them challenging to predict.

In the existing literature, 3D hand motion prediction is largely overlooked in computer vision. Motion-related methods predict human motions [5, 11, 44–46, 70], which change in position, orientation, or shape during an action. They consider several physical body constraints and human behaviors for diverse motion prediction during training. However, they focus only on the human body

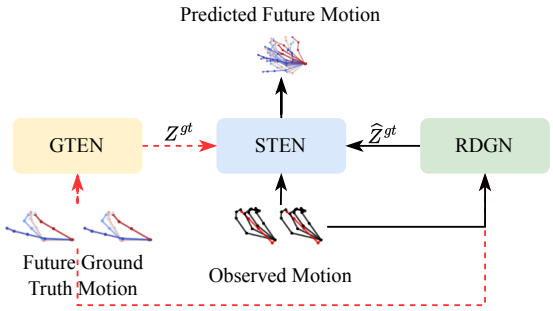


Fig. 1: PromptFDDM includes three networks: Spatial-Temporal Extraction Network (STEN), Ground Truth Extraction Network (GTEN), and Reference Data Generator Network (RDGN). The red dashed lines indicate the training-only pipeline. Black lines are used during both training and testing. GTEN and STEN extract latent features from 3D coordinates of 21 hand joints, from future ground truth and observed motion respectively. RDGN generates reference data to substitute for ground truth in the reference stage. STEN predicts future 3D coordinates of 21 hand joints using latent features from observed motion and differing guidance. It is guided by extracted ground truth during training and uses generated reference data during testing.

level, and the specific hand motion is neglected. Hand-related methods mainly forecast the hand trajectory, predicting its spatial movement as a whole but not the individual finger motions [4, 42]. In certain scenarios, only particular hand motions should be considered. For instance, in Augmented Reality (AR) and Virtual Reality (VR), predicted hand motion can revolutionize user experiences in gaming [4, 17, 41, 42]. Besides, in human-computer interaction and gesture-based control systems, predicted hand motion supports natural, contact-free interactions [20, 43, 60].

Our focus in this paper is to learn models of hand motion from observed data. More specifically, we are interested in hand motion prediction during daily activities, where we forecast the diverse future 3D motions of the human hand given its past motion. Hand motion prediction presents several challenges. Firstly, acquiring accurate 3D hand motion annotations at scale is labor-intensive and costly, often requiring wearable markers or multi-camera systems for hand motion capture in controlled environments [18, 23]. Secondly, hand motions are subtle and localized, with a limited range of motion. Previous works considering the full hand as a single joint ignore finger motions and hand articulations. Furthermore, hand motion occurs in a sophisticated, compact range compared to the larger motion range of the entire human skeleton. Thirdly, hand motion is naturally diverse and changes rapidly. Human uncertainties lead to unpredictable intentions, with multiple potential movements possible. Predicting as many reasonably is practically necessary. Additionally, hand movements change more quickly than full-body movements, challenging models to extract useful information. These factors make our task more challenging.

In this paper, we address these challenges by developing a novel prompt-based Future Driven Diffusion Model (PromptFDDM). Leveraging the powerful distribution mapping capabilities of Diffusion Models (DM), PromptFDDM offers an effective solution for future motion prediction. When trained with access to ground truth, the model inherently possesses the capacity to achieve heightened predictive accuracy. Specifically, we first stack ST-GCN-SE blocks [67] to construct a spatial-temporal extractor network (STEN) for compact skeleton prediction.

To incorporate ground truth, we train the proposed PromptFDDM in two stages shown in Fig. 1: (1) In the first training stage, a ground truth extractor network (GTEN) is designed to extract a compact latent feature map Z^{gt} from future motions to guide STEN in the prediction process. With guidance from the future, STEN naturally attains more reliable predictions. (2) In fact, such future guidance cannot be obtained in advance. Therefore, we train a reference data generator network (RDGN) to directly generate similar features $\hat{Z}^{gt}$ extracted by GTEN, using only the observed motions. Due to the compact nature of the latent feature map, RDGN can generate it accurately. With this generated guidance, STEN, lacking real future information, makes more reliable predictions at inference time than relying solely on itself. In addition to the above scheme and architectural novelties, we also introduce interactive prompt learning, enabling the model to dynamically adapt to and learn various motion behaviors. Future motions are usually highly related to past motions; therefore, the DM can utilize observed motions to estimate a more robust future latent features map.

The core contributions of PromptFDDM are highlighted as follows:

- We propose PromptFDDM, an approach using the strong mapping abilities of DM to estimate future information, guiding hand motion prediction.
- We utilize observed motion to generate interactive prompts, a set of tunable parameters that aid the denoising process of the RDGN. Introducing additional historical information prompts improves result reliability.
- We are the first to explore the hand motion prediction problem. Furthermore, we conduct comprehensive benchmarking of recent methods on the FPHA and HO3D datasets for the proposed task. Extensive experimental results demonstrate that our approach outperforms state-of-the-art baseline methods, achieving superior performance.

2 Related Works

Human Motion Prediction. Early in the research, traditional works [36,38,46, 47, 55, 57, 63, 75] predict a single future motion based on past poses. RNN-based models like [16, 21, 28, 47, 62] improve temporal dependency modeling. CNN-based models [7, 22, 38] predict whole sequences without accumulation error. As a derivative of CNNs, GCNs have inherent advantages to represent human skeletons as graphs and capture spatial dependencies [2,12,13,32,39,46]. ST-GCN [67]

proposes a spatial-temporal skeleton joints graph to better maintain temporally consecutive poses. Besides GCN-based models, transformer-based models [1, 8, 44, 48] adapt attention mechanism to model pairwise spatial-temporal dependencies. These methods struggle with long-term prediction. Considering that human motion is highly subjective and uncertain, similar observed motion sequences can lead to diverse future motions.

Addressing this has been the focus of stochastic motion prediction methods. Existing ones are mainly based on the deep generative models [3, 6, 9, 35, 45, 61, 65, 66, 68, 70, 71] such as variational autoencoders (VAEs) [31], generative adversarial networks (GANs) [19], normalizing flows (NFs) [53] and denoising diffusion probabilistic models (DDPMs) [25]. However, these works model the whole human skeleton and neglect the specific hand motion prediction. Here, we propose an ST-GCN-SE block to extract information from a compact hand skeleton.

Diffusion Models. Denoising Diffusion Probabilistic Models (DDPMs) are inspired by the laws of thermodynamics. These models attempt to learn the inverse process of gradually adding noise to the data distribution until it becomes random noise. Due to its dynamic principles, DDPMs generate diverse and high-quality results. Researchers in [25, 58] further advance practices for image generation applications. TCD [56] formulates prediction as a denoising problem, directly forecasting the observation and prediction motion sequence from noise. However, they suffer from high computational complexity of diffusion models. An end-to-end diffusion model, HumanMAC [11], is proposed for direct prediction of future motion from masked noise. BeLFusion [5] aims to learn behavior codes from observed motions and sample these in a latent space for diversity. However, its training requires multiple stages to disentangle behavior codes and heavily relies on pre-trained motion encoders and decoders, making the model challenging to implement. In this paper, our PromptFDDM performs DM on generating only compact reference data, providing accurate future guidance for STEN in the generation process.

Prompt Learning. This approach is initially employed in the NLP domain [37, 40]. It is motivated by pre-trained language models, such as BERT [14] and GPT [52]. The basic idea is that pre-trained models provide knowledge useful for downstream tasks. With some adaptations, the prompt idea is introduced to V-L [76, 77] and vision-only [29, 64] models. The concept of classifier-guided diffusion is first introduced by [15] and later adapted by [49] to enable conditioning based on CLIP textual representations. MDM [59], MotionDiffuse [73], and T2M-GPT [72] introduce prompts in the field of text-driven motion synthesis. They employ text prompts from CLIP to generate diverse motions. Here, we generate prompts from a specific history motion sequence and they are highly related to generating its future motion.

3 Preliminaries: Discrete Cosine Transform

In the motion prediction literature, human skeletons are often represented as a sequence of 3D joint coordinates. This encourages valid pose generation but does not guarantee smooth, natural results. To ensure temporal continuity, we adopt a trajectory representation using the Discrete Cosine Transform (DCT), as proposed in [46]. Discarding high frequencies provides a more compact representation capturing motion sequence smoothness, especially in 3D coordinates.

Given a $H + P$ frames motion sequence $\mathbf{m} \in \mathbb{R}^{(H+P)\times 3J}$, with each row of $\mathbf{m}$ representing the skeleton of hand joints at each frame. We project the motion sequences of $\mathbf{m}$ into the frequency domain via Discrete Cosine Transform (DCT) as

$$\widetilde{\mathbf{m}} = \mathrm{DCT}\left(\mathbf{m}\right) = \mathbf{D}\mathbf{m}, \tag{1}$$

where $\mathbf{D} \in \mathbb{R}^{M\times(H+P)}$ represents the pre-defined DCT basis as described in [45]; $\widetilde{\mathbf{m}} \in \mathbb{R}^{M\times 3J}$ with each row of $\widetilde{\mathbf{m}}$ representing the first $M < H + P$ DCT coefficients for the trajectory.

As the DCT operation is an orthogonal transform, we can consistently recover the original motion sequence from the DCT domain through an inverse DCT operation as

$$\mathbf{m} = \mathrm{IDCT}\left(\widetilde{\mathbf{m}}\right) = \mathbf{D}^{\top}\widetilde{\mathbf{m}}. \tag{2}$$

4 Methodology

Problem Formulation. We denote the specific length H of the observed hand motion sequence T_H as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_H] \in \mathbb{R}^{H\times 3J}$, where $\mathbf{x}_h \in \mathbb{R}^{3J}$ represents the Cartesian coordinates of hand skeletons at the h-th frame, and J is the number of hand joints. Given the observed motion sequence, the objective of the hand motion prediction task is to predict the subsequent motion sequence T_P, which is represented as $\mathbf{Y} = [\mathbf{x}_{H+1}, \mathbf{x}_{H+2}, ..., \mathbf{x}_{H+P}]_K \in \mathbb{R}^{P\times 3J}$, where K represents the number of possible diversities of hand motions for the next P frames. Our objectives are: (1) making one of the K predictions as close to the ground truth as possible, and (2) maximizing the diversity among the K sequences.

Fig. 2 shows the framework of our approach. We tackle the problem of hand motion prediction through the following key designs: (1) We first utilize *ground truth (Sec. 4.1)* as guidance for motion prediction. (2) We design a *diffusion model (Sec. 4.2)* to generate reference data substituting for ground truth. (3) We adopt observed motion sequences to *generate prompts (Sec. 4.3)* to improve performance.

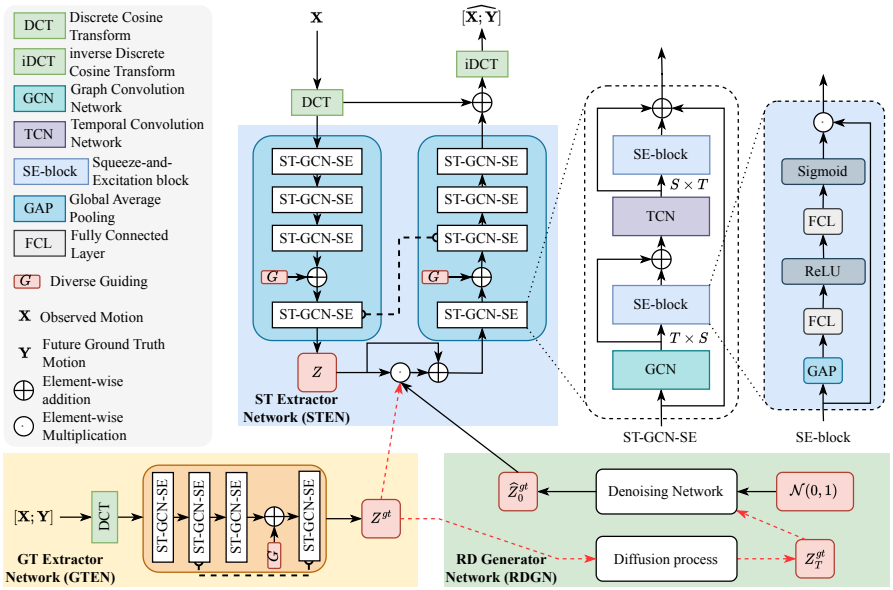


Fig. 2: Overview of the PromptFDDM approach. PromptFDDM consists of STEN, GTEN, and RDGN. STEN leverages information from observed motion and is guided by the other two networks to perform prediction in two stages: (1) In the first stage, we only train STEN with GTEN. GTEN extracts the latent feature map Z^{gt} from the ground truth to guide STEN in the prediction process. (2) In the second stage, we only train RDGN to generate the value $\widehat{Z}^{gt}$ that approximates Z^{gt} as closely as possible for STEN to use. In the inference stage, we only use STEN and RDGN to conduct the prediction. In the denoising network, to achieve more reliable results, we introduce interactive prompts that utilize observed information to assist the denoising process. Notably, we do not input the future ground truth into DM at the inference stage. The red dashed lines are only used in the training stage. In the testing stage, we only use the reverse process of DM.

4.1 Learning from Ground Truth

In the first stage, STEN is guided by ground truth extracted from GTEN for prediction. The structure of GTEN is illustrated by the yellow box in Fig. 2. It mainly consists of ST-GCN-SE blocks, as depicted in the Fig. 2, based on ST-GCN [67] with squeeze-and-excitation (SE) blocks [27]. Diverse guidance modules G are introduced to assist STEN and GTEN in calibrating learning input motion behaviors. GTEN is designed to extract the compact latent feature map from Concat($\mathbf{X}, \mathbf{Y}$), denoted as $Z^{gt} \in \mathbb{R}^{n \times d}$. After that, STEN utilizes the information to predict future motions. The structure of STEN is shown in Fig. 2 blue box, containing the encoder $\mathcal{E}$ and decoder $\mathcal{D}$. Both are similar to GTEN.

Within each ST-GCN-SE block, we utilize the SE-block to aggregate local features from the spatial dimension following GCN and the temporal dimension following TCN, respectively. It serves as an attention mechanism but with

significantly fewer parameters compared to the self-attention module. Adaptive scaling is applied to individual channels to model dependencies among different channels, thus optimizing the learning process of ST-GCN and improving network performance, especially in this compact hand skeleton information scenario. Following the method in [45, 46], we replicate the final frame of $\mathbf{X}$ for P times to construct $\widehat{\mathbf{X}}$ before computing the DCT coefficients, which translates to estimating a residual vector in frequency space. The STEN encoder $\mathcal{E}$ solely extracts the latent feature map $Z = \mathcal{E}(\widehat{\mathbf{X}})$. GTEN extracts the latent feature map Z^{gt} from both observed and future motions Concat $(\mathbf{X}, \mathbf{Y})$. The fusion block combines the two latent feature maps ($\bar{Z} = \text{fusion}\,(Z, Z^{gt})$). Subsequently, the STEN decoder $\mathcal{D}$ reconstructs the motions from the fused latent feature maps $\bar{Z}$. Since the hand structure is inherently similar to the human skeleton, we follow similar training strategies as the human motion prediction tasks to train STEN with GTEN. And the parameters of the trained STEN will be fixed for the second stage usage.

4.2 Guiding STEN by Generated Reference Data

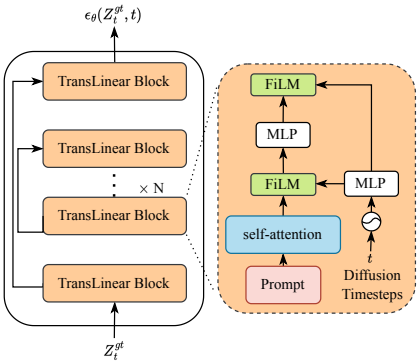


Fig. 3: The architecture of the noise prediction network TransLinear, which takes the latent feature map at diffusion step t as input. TransLinear is composed of N blocks with skip connections. Each block contains an interactive prompt block that utilizes extra observed motion information for better denoising performance.

In the second stage, conditioned on the latent features Z extracted by trained STEN from historical motion $\mathbf{X}$, RDGN generates reference data to substitute for the ground truth, guiding the trained and fixed STEN to perform prediction. We leverage the powerful data estimation capabilities of DM to generate $\widehat{Z}^{gt}$. First, we employ the pretrained GTEN to extract Z^{gt} from the concatenation of $\mathbf{X}$ and $\mathbf{Y}$. Then, in the DM training stage, we continuously add noise to transform the input Z^{gt} into Gaussian noise $Z_T^{gt} \sim \mathcal{N}(0, 1)$ over T time steps. Each step is modeled as a Markov noise process using the following equation

$$q\left(Z_t^{gt} \mid Z_{t-1}^{gt}\right) = \mathcal{N}\left(Z_t^{gt}; \sqrt{\alpha_t} Z_{t-1}^{gt}, \beta_t \mathbf{I}\right), \tag{3}$$

where the β_t represents the pre-defined scale factor, and $\alpha_t = 1 - \beta_t$. Here, $\mathcal{N}$ denotes the Gaussian distribution. As one of the properties of the forward process, it allows for the sampling of Z_t^{gt} at any time step t, which can be represented in the following form: by using $\bar{\alpha}_t = \Pi_{i=1}^{t} \alpha_i$, we have

$$q\left(Z_t^{gt} \mid Z_0^{gt}\right) = \mathcal{N}\left(Z_t^{gt}; \sqrt{\bar{\alpha}_t} Z_0^{gt}, (1 - \bar{\alpha}_t)\, \mathbf{I}\right). \tag{4}$$

In the inference stage, DM methods sample Gaussian random noise Z_T^{gt} and gradually denoise it until it reaches a high-quality output, denoted as Z_0^{gt}:

$$q\left(Z_{t-1}^{gt} \mid Z_t^{gt}, Z_0^{gt}\right) = \mathcal{N}\left(Z_{t-1}^{gt}; \tilde{\mu}_t\left(Z_t^{gt}, Z_0^{gt}\right), \tilde{\beta}_t \mathbf{I}\right), \tag{5}$$

where the mean $\tilde{\mu}_t\left(Z_t^{gt}, Z_0^{gt}\right) = \frac{1}{\sqrt{\alpha_t}}\left(Z_t^{gt} - \epsilon \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\right)$ and variance $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$. The symbol ϵ represents the noise in Z_t^{gt} and is the only source of uncertainty in the reverse process.

For the noise prediction, we introduce the denoising network ϵ_θ (see Fig. 3), which is a transformer-based U-net [54]. We utilize a transformer-based denoising model with long skip connections [11] on the motion latent $Z^{gt} \in \mathbb{R}^{n \times d}$. We use $\{Z_t^{gt}\}_{t=0}^T$ to denote the noisy sequence, and $Z_{t-1}^{gt} = \epsilon_\theta\left(Z_t^{gt}, t\right)$ for t-step denoising. We also focus on unconditional generation with a simple objective [25]

$$\mathcal{L}_{LDM} := E_{\epsilon,t}\left[\|\epsilon - \epsilon_\theta\left(Z_t^{gt}, t\right)\|_2^2\right], \tag{6}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, and $Z_0^{gt} = \text{GTEN}\left(\text{Concat}\left(\mathbf{X}, \mathbf{Y}\right)\right)$. During training $\boldsymbol{\epsilon_\theta}$, the GTEN is frozen to compress the motion sequence into Z_0^{gt}. The samples of the diffusion forward process are from the latent distribution $p\left(Z_0^{gt}\right)$. During the diffusion reverse stage, ϵ_θ first predicts $\widehat{Z}_0^{gt}$ with T iterative denoising steps, then the STEN decoder $\mathcal{D}$ reconstructs the latent $\bar{Z} = \text{fusion}\left(Z, \widehat{Z}_0^{gt}\right)$ into the motion sequence.

4.3 Prompts for Generating Reference Data

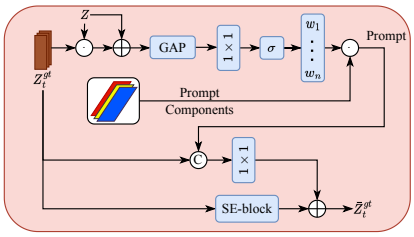


Fig. 4: Illustration of the interactive prompt block. The future motions are inherently related to the observed motions. Fusing extracted history Z and future Z_t^{gt} latent feature maps makes the generated prompts inherit extra history information. Interactive prompts with input features can enrich the content inside the $\bar{Z}_t^{gt}$.

In NLP [26, 37, 40] and vision tasks [29], prompting-based techniques are explored for the parameter-efficient fine-tuning of large frozen models trained on a source task (S) for application to a target task (T). The effective performance of prompting-based techniques is attributed to their ability to encode task-specific contextual information within prompt components efficiently. Our interactive prompt blocks are intended to generate historical condition prompts, added at every layer in RDGN to generate better reference data $\widehat{Z}_0^{gt}$.

Prompt Components form a set of learnable parameters interacting with the information from the input features. The most efficient method for feature-prompt interaction is through the learned prompt

components to calibrate features, see Fig. 4. Given that future motions are inherently correlated with past motions, we introduce fusing the latent feature map Z encoded from observed motions with Z_t^{gt} ($\bar{Z} = \text{fusion}\left(Z, Z_t^{gt}\right)$) to generate prompt in the denoising stage.

To generate the interactive prompts from input features $\bar{Z}$, we first apply global average pooling (GAP) across the spatial dimension to generate the feature $\mathbf{v} \in \mathbb{R}^M$. Then, we obtain a more compact vector through compression and use the softmax operation to calculate the prompt weight $w \in \mathbb{R}^N$. Finally, we utilize these weights to guide the parameters in the generated prompt components through a dot product operation to generate the prompts $\mathbf{P}$.

To enable interaction between prompt weights and input features, we begin by concatenating Z_t^{gt} with the prompts $\mathbf{P}$ in the spatial channel. Next, a 1D convolution is used to fuse them and exploit the degradation information encoded in the prompts to transform the input features.

5 Experiments

5.1 Datasets

First-Person Hand Action (FPHA). The FPHA dataset [18] collects first-person RGB-D videos capturing diverse hand-object interactions. It includes ground-truth 3D hand pose, 6D object pose, and hand joint locations from magnetically-tracked mocap sensors. Object pose is annotated for 4 objects in a subset of videos, also via magnetic sensors. Our model observes 15 past frames (0.5s) and predicts 60 future frames (2.0s).

HO-3D. The HO-3D dataset [23] captures hand-object interactions from a third-person perspective. It contains $77,558$ frames with 3D hand joint positions and 3D object bounding boxes. In our study, we only use the 3D hand position annotations. The dataset [23] is divided into a training set of $66,034$ frames and an evaluation set of $11,524$ frames. In the evaluation set, only wrist coordinates are labeled for the hands, and full hand annotations are unavailable. Therefore, we utilize only the original training set and follow the same action split as in [10, 24, 69]. Additionally, we select one video sequence per object for the test set manually. Our model observes 20 past frames and predicts 80 future frames.

5.2 Implementation

Evaluation metrics. Consistent with the evaluation protocols in [70], we employ five metrics to assess the diversity and accuracy of our model. (1) **Average Pairwise Distance (APD)**: computing the L2 distance among all pairs of predicted motion poses at each time step. (2) **Average Displacement Error (ADE)**: the smallest average L2 distance between the ground truth and the predicted motions for the entire sequence. (3) **Final Displacement Error (FDE)**: the smallest average L2 distance between the ground truth and the predicted motions for the last frame of the sequence. (4) **Multi-Modal-ADE**

Table 1: Quantitative results on the FPHA and HO3D datasets. **Bold** numbers indicate the best results. The lower is better for all metrics except for APD. The symbol '-' indicates that certain methods are not reported in this setting.

Method	FPHA [18]					HO3D [23]				
	APD↑	ADE↓	FDE↓	MMADE↓	MMFDE↓	APD↑	ADE↓	FDE↓	MMADE↓	MMFDE↓
cVAE [70]	5.168	0.671	0.754	0.672	0.751	2.729	0.198	0.261	0.221	0.264
DLOW [70]	9.616	0.668	0.723	0.669	0.723	5.744	0.222	0.292	0.243	0.292
MOJO [74]	7.695	0.532	0.607	0.535	0.608	**6.335**	0.180	0.255	0.214	0.260
GSPS [45]	9.738	0.494	0.654	0.495	0.650	2.808	0.187	0.225	0.204	0.227
HumanMAC [11]	2.831	0.684	0.781	0.686	0.780	1.821	**0.155**	0.217	0.209	0.236
BeLFusion [5]	2.822	0.696	0.869	0.699	0.870	-	-	-	-	-
STEN alone	18.627	0.473	0.602	0.474	0.600	4.090	0.168	0.215	0.195	0.221
Ours	**19.798**	**0.413**	**0.535**	**0.416**	**0.534**	4.541	0.163	**0.203**	**0.190**	**0.210**

(MMADE): the multi-modal version of ADE, grouped by similar observed past motion status. (5) **Multi-Modal-FDE (MMFDE)**: similarly, the multi-modal version of FDE.

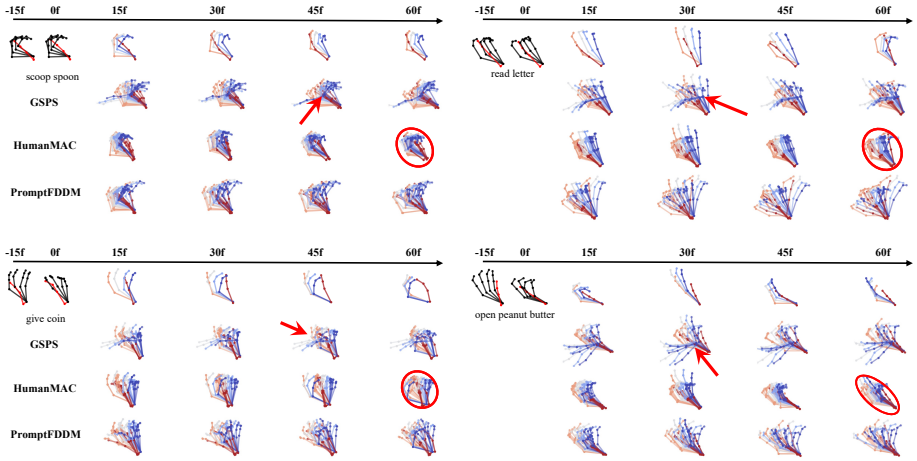
Baselines. To comprehensively evaluate our method, we compare it with several motion prediction models, including cVAE [70] and DLow [70], MOJO [74], GSPS [45], HumanMAC [11], and BeLFusion [5]. All comparative methods are trained on the two hand datasets from scratch. In this task, all models are configured to match the hand skeleton data structure. In qualitative comparison, the employed competitors are GSPS and HumanMAC.

Implementation details. We use similar loss functions with previous human motion prediction works [45,70]. The dimensionality of latent feature maps Z and Z^{gt} is 16 for all methods. The number of diverse samples K for each prediction is 10. For comparisons, the encoder and decoder of our STEN and GTEN models both have 4 layers of ST-GCN-SE block. For FPHA and HO3D, the model is trained with a batch size of 16 for 500 epochs using 1000 training examples per epoch. Adam optimizer [30] with base learning rate 1e-3, is used, decayed after 100 epochs.

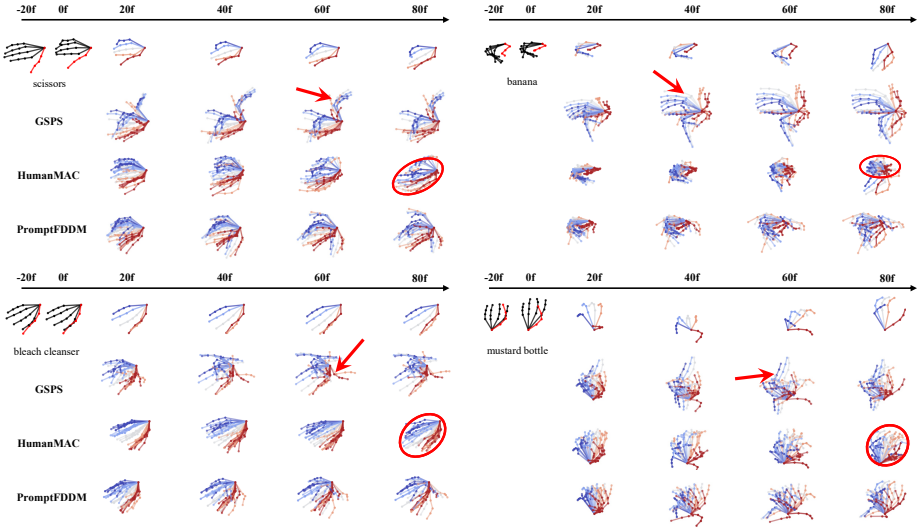
For the denoising model, we train ϵ_θ on both datasets with a 1000-step DDPM [25] and sample with a 100-step DDIM [58]. The Linear scheduler is exploited for variance scheduling in our model. The noise prediction network contains 4-layer and 12-layer TransLinear blocks for FPHA and HO3D respectively.

5.3 Comparison to state-of-the-art approaches

Quantitative results. We compare our method with counterparts in Tab. 1 on FPHA and HO3D. As observed, for FPHA, our method achieves state-of-the-art across all metrics, demonstrating validity. Previous methods exhibit relatively low diversity (APD) and accuracy (ADE, FDE, MMADE, MMFDE). For HO3D, our method provides relatively reliable predictions compared to MOJO, with the highest diversity. Additionally, our method provides relatively higher diversity

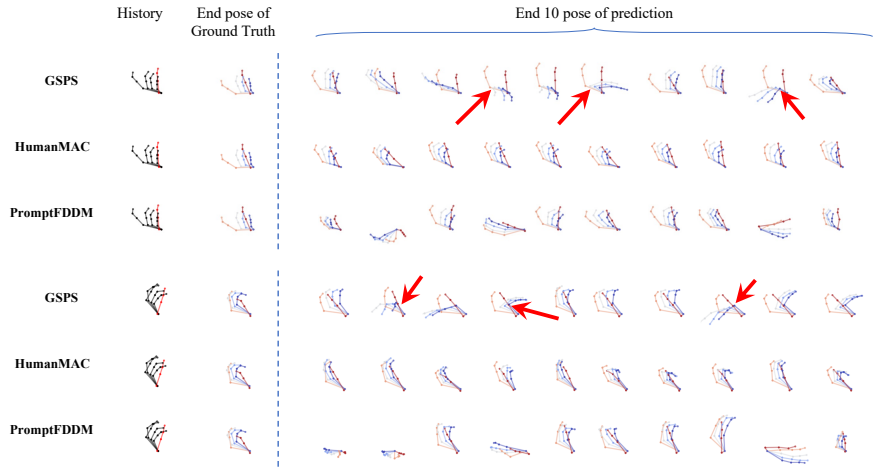


(a) Diversity motion sequences prediction visualization on FPHA.

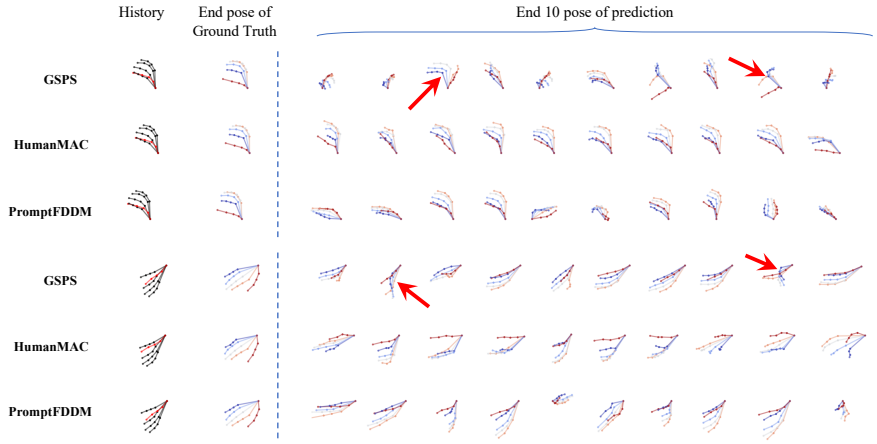


(b) Diversity motion sequences prediction visualization on HO3D.

Fig. 5: Qualitative results on motion sequence prediction visualization. The first line of each sample is the ground truth. The negative and positive frame numbers on the time axis indicate the observed motions and future motions at certain time steps, respectively. The red-black skeletons and red-to-blue skeletons denote the observed and predicted motions, respectively. We focus on the valid poses and range of motion for overlapping predicted motions. GSPS normally generates some invalid poses, indicated by red arrows. HumanMAC generates a limited range of motion, indicated by red circles.



(a) Comparison on end poses visualization on FPHA.



(b) Comparison on end poses visualization on HO3D.

Fig. 6: Results of end poses visualization. Each row denotes the 10 predicted end poses from three models. The red-black skeletons and red-to-blue skeletons denote the observed and predicted motions, respectively. GSPS generates some invalid poses. HumanMAC generates very similar 10 motions, indicating a lack of diversity and resulting in low APD results.

Table 2: Experimental results of the ablation study on a different number of the first L rows of DCT.

L	FPHA			HO3D		
	APD↑	ADE↓	FDE↓	APD↑	ADE↓	FDE↓
5	15.193	0.417	0.560	1.532	0.176	0.232
10	**20.634**	0.462	0.610	3.002	**0.162**	0.204
20	19.798	**0.413**	**0.535**	**4.541**	0.163	**0.203**

than HumanMAC, the most accurate ADE. Our HO3D predictions strike an excellent balance between diversity and accuracy. As mentioned in Sec. 5.1, FPHA is collected from a first-person perspective and HO-3D from a third-person perspective. In the egocentric scenario, the world coordinates continually change with sensor movement, so FPHA exhibits higher diversity and lower accuracy than HO-3D.

Qualitative comparison. To visually evaluate the diversity and accuracy of results, we follow the approach of previous works in human motion prediction by using overlapping predicted motions and end poses for visualization, see Fig. 5. We compare our method with GSPS [45] and HumanMAC [11]. Since hand motions are localized with limited range, for better visualization, we show only the five least overlapping predicted motions per frame. In Fig. 6, we show the end poses of all predicted results separately.

For future predictions, GSPS consistently generates unreasonable failure cases like finger joint fractures (annotated by red arrows) in Fig. 5 and Fig. 6. In contrast, HumanMAC produces more reasonable cases but with limited diversity. Its range of finger movements is insufficient to cover general daily activities (annotated by red circles). The results in Fig. 5 and Fig. 6 demonstrate that our method generates more physically valid and diverse results in daily activities.

5.4 Ablation Study

We conduct comprehensive ablation studies, including (1) the value of L in DCT/iDCT; (2) the design of the noise-denoising network; (3) the effectiveness of the SE-block; and (4) the impact of interactive prompts. We provide detailed explanations for each aspect. In all tables, our final choices for our method are highlighted with gray shading.

Value of L. As discussed in Sec. 4.1, we approximate the DCT and iDCT operations by selecting the first L rows of $\mathbf{D}$ as $\mathbf{D}_L$ to improve computing efficiency. In Tab. 2, we present the optimal choice for L, 20 for both datasets.

Design of the noise prediction network. RDGN's LDM adopts a U-Net architecture from [11]. More precisely, U-Net comprises a stack of TransLinear blocks, and within each transformer encoder, two FiLM-like conditioning modules [51] are integrated. Tab. 3 shows reasonable layer choices considering average rank across three metrics; the best values are 4 layers for FPHA and 12 layers for HO3D.

Table 3: Experimental results of the ablation study on a different number of layers in the noise prediction network.

Layer	FPHA			HO3D		
	APD↑	ADE↓	FDE↓	APD↑	ADE↓	FDE↓
2	19.645	0.419	0.563	4.530	0.165	0.207
4	19.798	0.413	**0.535**	4.500	**0.163**	0.205
8	19.742	**0.406**	0.541	**4.558**	0.164	0.204
12	**19.825**	0.422	0.544	4.541	**0.163**	**0.203**

Table 4: Experimental results of the ablation study on different schedulers in the diffusion model.

Scheduler	FPHA			HO3D		
	APD↑	ADE↓	FDE↓	APD↑	ADE↓	FDE↓
Linear	**19.798**	**0.413**	**0.535**	**4.541**	**0.163**	**0.203**
Sqrt	18.794	0.452	0.568	4.441	0.164	0.210
Cosine	18.940	0.448	0.585	4.517	0.165	0.209

Table 5: Experimental results of the ablation study with respect to the SE-block in the model.

SE-block	FPHA			HO3D		
	APD↑	ADE↓	FDE↓	APD↑	ADE↓	FDE↓
w/o SE	16.008	0.422	**0.523**	3.685	0.196	0.249
w SE	**19.798**	**0.413**	0.535	**4.541**	**0.163**	**0.203**

Table 6: Experimental results of the ablation study on interactive prompts usage and its length.

Prompt len	FPHA			HO3D		
	APD↑	ADE↓	FDE↓	APD↑	ADE↓	FDE↓
w/o	18.884	0.423	0.562	3.851	**0.161**	0.206
5	**19.847**	0.420	0.539	**4.541**	0.163	0.203
10	19.798	**0.413**	**0.535**	4.513	0.162	**0.200**

We conduct an investigation to assess the impact of various predefined diffusion variance schedulers, including Linear, Cosine, and Sqrt. As shown in Tab. 4, the Linear scheduler is the most optimal choice for both datasets.

Effectiveness of the SE-block. The SE-block is a plug-in module within ST-GCN, which serves as the basic unit of our model. We analyze its ability to extract spatial-temporal information. Tab. 5 shows that this module demonstrates highly efficient capability for comprehensively capturing both temporal and skeletal information.

Impact of the interactive prompts. As the interactive prompts are generated based on observed motions, we analyze the impact of generated prompts in terms of their size. Tab. 6 shows the optimal choices are size 10 for FPHA and size 5 for HO3D.

6 Conclusion

In this paper, we propose to predict hand motion in 3D physical space from both first- and third-person viewpoints. We introduce a novel prompt-based Future Driven Diffusion Model (PromptFDDM) to fill the gap in the motion prediction field. The future-driven prompting method ensures more accurate and diverse outcomes. Quantitative and qualitative experiments demonstrate the effectiveness of our approach, with superior performance on the hand datasets FPHA and HO3D. The various action behaviors learned during training contribute to predicting future motions based on a single observed motion, yielding reliable and diverse results. There are some limitations in the trade-off between diversity and accuracy. We expect PromptFDDM can serve as a solid baseline and provide a new perspective for modeling 3D hand motion prediction. Future research efforts will focus on improving the accuracy of third-person viewpoints and consider the motion involved in hand-object interactions.

Acknowledgment

This work is funded in part by the National Natural Science Foundation of China (Grant No. 62372480), in part by ARC-Discovery (DP 220100800), in part by CCF-Tencent Rhino-Bird Open Research Fund (No. CCF-Tencent RAGR20230118).

References

1. Aksan, E., Kaufmann, M., Cao, P., Hilliges, O.: A spatio-temporal transformer for 3d human motion prediction. In: 3DV (2021)
2. Aksan, E., Kaufmann, M., Hilliges, O.: Structured prediction helps 3d human motion modelling. In: ICCV (2019)
3. Aliakbarian, M.S., Saleh, F.S., Salzmann, M., Petersson, L., Gould, S.: A stochastic conditioning scheme for diverse human motion prediction. In: CVPR (2020)
4. Bao, W., Chen, L., Zeng, L., Li, Z., Xu, Y., Yuan, J., Kong, Y.: Uncertainty-aware state space transformer for egocentric 3d hand trajectory forecasting. In: CVPR (2023)
5. Barquero, G., Escalera, S., Palmero, C.: Belfusion: Latent diffusion for behavior-driven human motion prediction. In: ICCV (2023)
6. Barsoum, E., Kender, J.R., Liu, Z.: HP-GAN: probabilistic 3d human motion prediction via GAN. In: CVPRW (2018)
7. Bütepage, J., Black, M.J., Kragic, D., Kjellström, H.: Deep representation learning for human motion prediction and classification. In: CVPR (2017)
8. Cai, Y., Huang, L., Wang, Y., Cham, T., Cai, J., Yuan, J., Liu, J., Yang, X., Zhu, Y., Shen, X., Liu, D., Liu, J., Magnenat-Thalmann, N.: Learning progressive joint propagation for human motion prediction. In: ECCV (2020)
9. Cai, Y., Wang, Y., Zhu, Y., Cham, T., Cai, J., Yuan, J., Liu, J., Zheng, C., Yan, S., Ding, H., Shen, X., Liu, D., Magnenat-Thalmann, N.: A unified 3d human motion synthesis model via conditional variational auto-encoder*. In: ICCV (2021)
10. Cao, Z., Radosavovic, I., Kanazawa, A., Malik, J.: Reconstructing hand-object interactions in the wild. In: ICCV (2021)
11. Chen, L., Zhang, J., Li, Y., Pang, Y., Xia, X., Liu, T.: Humanmac: Masked motion completion for human motion prediction. In: ICCV (2023)
12. Chen, M., Wei, Z., Huang, Z., Ding, B., Li, Y.: Simple and deep graph convolutional networks. In: ICML (2020)
13. Dang, L., Nie, Y., Long, C., Zhang, Q., Li, G.: MSR-GCN: multi-scale residual graph convolution networks for human motion prediction. In: ICCV (2021)
14. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL HLT (2019)
15. Dhariwal, P., Nichol, A.Q.: Diffusion models beat gans on image synthesis. In: NeurIPS (2021)
16. Fragkiadaki, K., Levine, S., Felsen, P., Malik, J.: Recurrent network models for human dynamics. In: ICCV (2015)
17. Gamage, N.M., Ishtaweera, D., Weigel, M., Withana, A.: So predictable! continuous 3d hand trajectory prediction in virtual reality. In: ACM Int. Conf. User. Inter. Soft. Tech (2021)
18. Garcia-Hernando, G., Yuan, S., Baek, S., Kim, T.: First-person hand action benchmark with RGB-D videos and 3d hand pose annotations. In: CVPR (2018)

19. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: NeurIPS (2014)
20. Gourob, J.H., Raxit, S., Hasan, A.: A robotic hand: controlled with vision based hand gesture recognition system. In: International Conference on Automation, Control and Mechatronics for Industry (ACMI) (2021)
21. Gui, L., Wang, Y., Liang, X., Moura, J.M.F.: Adversarial geometry-aware human motion prediction. In: ECCV (2018)
22. Guo, X., Choi, J.: Human motion prediction via learning local structure representations and temporal dependencies. In: AAAI (2019)
23. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: Honnotate: A method for 3d annotation of hand and object poses. In: CVPR (2020)
24. Hasson, Y., Tekin, B., Bogo, F., Laptev, I., Pollefeys, M., Schmid, C.: Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In: CVPR (2020)
25. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
26. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for NLP. In: ICML (2019)
27. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR (2018)
28. Jain, A., Zamir, A.R., Savarese, S., Saxena, A.: Structural-rnn: Deep learning on spatio-temporal graphs. In: CVPR (2016)
29. Jia, M., Tang, L., Chen, B., Cardie, C., Belongie, S.J., Hariharan, B., Lim, S.: Visual prompt tuning. In: ECCV (2022)
30. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
31. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: ICLR (2014)
32. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: ICLR (2017)
33. Koppula, H.S., Saxena, A.: Anticipating human activities for reactive robotic response. In: IROS (2013)
34. Koppula, H.S., Saxena, A.: Anticipating human activities using object affordances for reactive robotic response. IEEE TPAMI (2016)
35. Kundu, J.N., Gor, M., Babu, R.V.: Bihmp-gan: Bidirectional 3d human motion prediction GAN. In: AAAI (2019)
36. Lehrmann, A.M., Gehler, P.V., Nowozin, S.: Efficient nonlinear markov models for human motion. In: CVPR (2014)
37. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. In: EMNLP (2021)
38. Li, C., Zhang, Z., Lee, W.S., Lee, G.H.: Convolutional sequence to sequence model for human dynamics. In: CVPR (2018)
39. Li, M., Chen, S., Zhao, Y., Zhang, Y., Wang, Y., Tian, Q.: Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In: CVPR (2020)
40. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. In: ACL IJCNLP (2021)
41. Li, Y., Cao, Z., Liang, A., Liang, B., Chen, L., Zhao, H., Feng, C.: Egocentric prediction of action target in 3d. In: CVPR (2022)
42. Liu, S., Tripathi, S., Majumdar, S., Wang, X.: Joint hand motion and interaction hotspots prediction from egocentric videos. In: CVPR (2022)

43. Mangukiya, Y., Purohit, B., George, K.: Electromyography (emg) sensor controlled assistive orthotic robotic arm for forearm movement. In: IEEE Sensors Applications Symposium (SAS) (2017)
44. Mao, W., Liu, M., Salzmann, M.: History repeats itself: Human motion prediction via motion attention. In: ECCV (2020)
45. Mao, W., Liu, M., Salzmann, M.: Generating smooth pose sequences for diverse human motion prediction. In: ICCV (2021)
46. Mao, W., Liu, M., Salzmann, M., Li, H.: Learning trajectory dependencies for human motion prediction. In: ICCV (2019)
47. Martinez, J., Black, M.J., Romero, J.: On human motion prediction using recurrent neural networks. In: CVPR (2017)
48. Martínez-González, Á., Villamizar, M., Odobez, J.: Pose transformers (POTR): human motion prediction with non-autoregressive transformers. In: ICCVW (2021)
49. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In: ICML (2022)
50. Paden, B., Cáp, M., Yong, S.Z., Yershov, D.S., Frazzoli, E.: A survey of motion planning and control techniques for self-driving urban vehicles. IEEE Trans. Intell. Veh. (2016)
51. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.C.: Film: Visual reasoning with a general conditioning layer. In: AAAI (2018)
52. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog (2019)
53. Rezende, D.J., Mohamed, S.: Variational inference with normalizing flows. In: ICML (2015)
54. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI) (2015)
55. Ruiz, A.H., Gall, J., Moreno, F.: Human motion prediction via spatio-temporal inpainting. In: ICCV (2019)
56. Saadatnejad, S., Rasekh, A., Mofayezi, M., Medghalchi, Y., Rajabzadeh, S., Mordan, T., Alahi, A.: A generic diffusion-based approach for 3d human pose prediction in the wild. In: ICLR (2023)
57. Sofianos, T., Sampieri, A., Franco, L., Galasso, F.: Space-time-separable graph convolutional network for pose forecasting. In: ICCV (2021)
58. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
59. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. In: ICLR (2023)
60. von Tiesenhausen, J., Artan, U., Marshall, J.A., Li, Q.: Hand gesture-based control of a front-end loader. In: IEEE Canadian Conference on Electrical and Computer Engineering (CCECE) (2020)
61. Walker, J., Marino, K., Gupta, A., Hebert, M.: The pose knows: Video forecasting by generating pose futures. In: ICCV (2017)
62. Wang, B., Adeli, E., Chiu, H., Huang, D., Niebles, J.C.: Imitation learning for human pose prediction. In: ICCV (2019)
63. Wang, J.M., Fleet, D.J., Hertzmann, A.: Gaussian process dynamical models for human motion. IEEE TPAMI (2008)
64. Wang, Z., Zhang, Z., Lee, C., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J.G., Pfister, T.: Learning to prompt for continual learning. In: CVPR (2022)
65. Wei, D., Sun, H., Li, B., Lu, J., Li, W., Sun, X., Hu, S.: Human joint kinematics diffusion-refinement for stochastic motion prediction. In: AAAI (2023)

66. Xu, S., Wang, Y., Gui, L.: Diverse human motion prediction guided by multi-level spatial-temporal anchors. In: ECCV (2022)
67. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: AAAI (2018)
68. Yan, X., Rastogi, A., Villegas, R., Sunkavalli, K., Shechtman, E., Hadap, S., Yumer, E., Lee, H.: MT-VAE: learning motion transformations to generate multimodal human dynamics. In: ECCV (2018)
69. Ye, Y., Gupta, A., Tulsiani, S.: What's in your hands? 3d reconstruction of generic objects in hands. In: CVPR (2022)
70. Yuan, Y., Kitani, K.: Dlow: Diversifying latent flows for diverse human motion prediction. In: ECCV (2020)
71. Zand, M., Etemad, A., Greenspan, M.A.: Flow-based spatio-temporal structured prediction of motion dynamics. IEEE TPAMI (2023)
72. Zhang, J., Zhang, Y., Cun, X., Zhang, Y., Zhao, H., Lu, H., Shen, X., Ying, S.: Generating human motion from textual descriptions with discrete representations. In: CVPR (2023)
73. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. IEEE TPAMI (2024)
74. Zhang, Y., Black, M.J., Tang, S.: We are more than our joints: Predicting how 3d bodies move. In: CVPR (2021)
75. Zhong, C., Hu, L., Zhang, Z., Ye, Y., Xia, S.: Spatio-temporal gating-adjacency GCN for human motion prediction. In: CVPR (2022)
76. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: CVPR (2022)
77. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. IJCV (2022)