# TransFusion: A Practical and Effective Transformer-based Diffusion Model for 3D Human Motion Prediction

Sibo Tian[1], Minghui Zheng[1,*], and Xiao Liang[2,*]

***Abstract*—Predicting human motion plays a crucial role in ensuring a safe and effective human-robot close collaboration in intelligent remanufacturing systems of the future. Existing works can be categorized into two groups: those focusing on accuracy, predicting a single future motion, and those generating diverse predictions based on observations. The former group fails to address the uncertainty and multi-modal nature of human motion, while the latter group often produces motion sequences that deviate too far from the ground truth or become unrealistic within historical contexts. To tackle these issues, we propose TransFusion, an innovative and practical diffusion-based model for 3D human motion prediction which can generate samples that are more likely to happen while maintaining a certain level of diversity. Our model leverages Transformer as the backbone with long skip connections between shallow and deep layers. Additionally, we employ the discrete cosine transform to model motion sequences in the frequency space, thereby improving performance. In contrast to prior diffusion-based models that utilize extra modules like cross-attention and adaptive layer normalization to condition the prediction on past observed motion, we treat all inputs, including conditions, as tokens to create a more practical and effective model compared to existing approaches. Extensive experimental studies are conducted on benchmark datasets to validate the effectiveness of our human motion prediction model. The project page is available at https://github.com/sibotian96/TransFusion.**

***Index Terms*—Human Motion Prediction (HMP), Deep Learning, Diffusion Models, Human-Robot Collaboration (HRC)**

## I. Introduction

HUMAN-ROBOT collaboration (HRC) in the recycling of end-of-life electronic products has gained significant attention in recent years [1]–[4]. Unlike traditional remanufacturing, where industrial robots and humans perform separate tasks in isolation for safety purposes, HRC allows for synergistic utilization of both human and robot agents during collaborative disassembly. Humans excel at handling uncertainty, while robots are efficient at performing repetitive, and labor-intensive tasks. When humans and robots work in close proximity, it is crucial for robots to understand their collaborator's behavior to ensure safety and improve collaboration efficiency. Accurate prediction enables robots to anticipate how to assist humans and avoid potential collisions. Therefore, modeling human behavior and predicting human motion are essential for achieving safe and seamless HRC. Many works [4]–[18] have explored these areas.

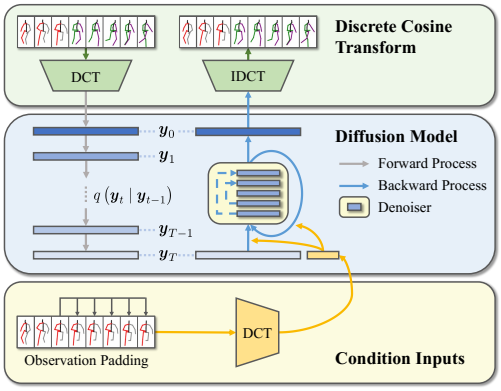


Fig. 1. An overview of the proposed HMP method. The proposed HMP method consists of a diffusion process and a reverse process. In the diffusion process, the motion sequence is transformed to the frequency domain using DCT. The noise is then progressively incorporated into the data over $T$ diffusion steps, resulting in white noise representation. In the reverse process, the observation is padded to match the length of the motion sequence. After applying DCT, the observed motion inputs guide the denoiser in recovering the data from the pure noise representation. Finally, IDCT is applied to reconstruct the motion sequence from its frequency components.

Previous state-of-the-art human motion prediction (HMP) works aimed to regress a single future sequence of human skeletal data based on observations; however, considering the inherent uncertainty and multi-modality of human motion, it is crucial to predict the distribution of potential human motions rather than relying on a single deterministic output, especially for safety-critical applications like HRC, as unforeseen human motion may cause serious collisions. Recent research on stochastic HMP has primarily focused on deep generative models. Previous works have utilized generative adversarial networks (GANs) [5] and variational autoencoders (VAEs) [6]–[13] to generate multiple future motions based on a short observation. These works typically incorporate multiple loss terms to ensure both the quality and diversity of generated samples. However, diversity-promoting techniques, such as diversity loss and diverse sampling, may lead to early deviations from the ground truth or sudden stagnation,

Manuscript received December 26, 2023; revised March 25, 2024; accepted April 28, 2024. This paper was recommended for publication by Editor Angelika Peer upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by the USA National Science Foundation under Grant No. 2026533/2422826 and 2132923/2422640. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

[1] Sibo Tian and Minghui Zheng are with the J. Mike Walker '66 Department of Mechanical Engineering, Texas A&M University, College Station, TX 77843, USA. Emails: sibotian, mhzheng@tamu.edu.

[2] Xiao Liang is with the Zachry Department of Civil and Environmental Engineering, Texas A&M University, College Station, TX 77843, USA. Email: xliang@tamu.edu.

[*] Corresponding Authors.

Digital Object Identifier (DOI):

resulting in unrealistic and implausible predictions. Overly diverse predictions can hinder downstream applications, such as motion planning for collaborative robot manipulators, as the excessive variety and out-of-context predictions may cover most of the shared working space when all predictions are considered valid. This leads to an overly conservative planned robot trajectory or even a failure to find a possible solution, contradicting the purpose of incorporating HMP in HRC.

Recently, a new deep generative model called denoising diffusion probabilistic model (DDPM) [20] has shown significant progress in generative tasks. DDPMs often learn the data distribution more accurately and produce samples of higher quality compared to VAEs and GANs. In this work, we propose a new diffusion model that incorporates a simple yet powerful transformer-based denoising neural network, aiming to predict several authentic within-context motions that are likely to happen. Specifically, we treat HMP as an inpainting problem, which facilitates consistency in generated motion and historical information, inspired by [18]. Unlike current state-of-the-art works that require additional modules to handle diffusion steps and observation, such as cross-attention in [16] or adaptive normalization in [15], [18], we treat all the inputs, including diffusion steps and observation, as tokens for the transformer. Additionally, we do not include any post-process blocks or blocks that need to be trained separately, such as the motion refinement block in [15], [17], and the autoencoder as well as behavior latent space in [16]. This enables us to train our model in an end-to-end fashion. Such innovations reduce the complexity of the network and make our model more practical than previous works. We utilize long skip connections to fuse information from shallow and deep layers for improved training. Unlike prior works that directly add two branches together, the long skip connection in our model is achieved by first concatenating two branches and passing them through a linear projection layer to match the dimensions. Furthermore, by treating the diffusion steps and historical motion as tokens, we aim for the self-attention layer in the transformer to more effectively learn token dependencies. We employ squeeze and excitation (SE) blocks [21] to the token dimensions in the transformer model. The SE mechanism enables dynamic recalibration of all tokens, including motion inputs and condition tokens, before they are passed to the self-attention layer. It assigns higher weights to tokens carrying more important information, resulting in more flexible and adaptive self-attention mechanisms and, consequently, better prediction performance. In addition, instead of representing human motion in the time domain, we adopt the discrete cosine transform (DCT) and learn the model in the frequency domain. DCT helps reduce the dimension of motion sequences while preserving important details by eliminating high-frequency components, which are primarily noise. This approach is beneficial for predicting continuous motions as it extracts time properties from sequential data. Overall, the main contributions of this work can be summarized as follows:

- We provide a detailed review of all diffusion-based HMP works known to us, which can serve as inspiration for future research in this area.
- We propose a novel, practical, and effective diffusion-based model called TransFusion for HMP. Unlike prior works, TransFusion does not require additional modules to handle diffusion steps and observation, nor does it need to process and refine the inputs and outputs of the diffusion model. It can be trained end-to-end and achieves state-of-the-art accuracy on three benchmark datasets.
- We conduct comprehensive experiments and ablation studies to validate the performance of TransFusion.

## II. Related Work

### A. Diffusion Models

Diffusion models [20], considered a new and promising addition to deep generative model family, have gained attention for their ability to generate high-quality samples through a simple training procedure. These models consist of two key processes: the forward process and the backward process. The forward process introduces noise gradually to the original data. After $T$ steps, the data will turn into pure noise. The transition of each step is represented as:

$$q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}\right)=\mathcal{N}\left(\boldsymbol{x}_t \mid \sqrt{\alpha_t} \boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}\right) \tag{2}$$

where $\boldsymbol{x}_t$ is the perturbed data at diffusion step $t$, $\beta_t$ is the pre-defined noise schedule, and $\alpha_t = 1 - \beta_t$. Note that $\boldsymbol{x}_t$ for an arbitrary step $t$ can be sampled directly using $\boldsymbol{x}_0$ in a closed form with the notation $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$:

$$q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_0\right)=\mathcal{N}\left(\boldsymbol{x}_t \mid \sqrt{\bar{\alpha}_t} \boldsymbol{x}_0,\left(1-\bar{\alpha}_t\right) \boldsymbol{I}\right). \tag{3}$$

Regarding the backward process, a natural approach is to reverse the steps applied in the forward process, aiming to restore the clean data from pure noise. A denoising process is proposed to approximate the true backward transition $q\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right)$ by learning a Gaussian model:

$$p_\theta\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right)=\mathcal{N}\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{\mu}_\theta\left(\boldsymbol{x}_t, t\right), \boldsymbol{\Sigma}_\theta\left(\mathbf{x}_t, t\right)\right). \tag{4}$$

Instead of predicting $\boldsymbol{x}_{t-1}$ from $\boldsymbol{x}_t$ directly at each step $t$, DDPM proposes that predicting the injected noise generates better results, and the mean $\boldsymbol{\mu}_\theta\left(\boldsymbol{x}_t, t\right)$ can be represented as:

$$\boldsymbol{\mu}_\theta\left(\boldsymbol{x}_t, t\right)=\frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_t, t\right)\right). \tag{5}$$

$\boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_t, t\right)$ is the noise-predicting neural network with a simple loss function:

$$\mathcal{L}(\theta)=\mathbb{E}_{t, \boldsymbol{x}_0, \epsilon}\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t} \boldsymbol{x}_0+\sqrt{1-\bar{\alpha}_t} \boldsymbol{\epsilon}, t\right)\right\|_2^2. \tag{6}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. As for the covariance $\boldsymbol{\Sigma}_\theta\left(\boldsymbol{x}_t, t\right)$, DDPM sets it as time-dependent constants for simplicity, i.e., $\boldsymbol{\Sigma}_\theta\left(\boldsymbol{x}_t, t\right)=\sigma_t^2 \boldsymbol{I}$ where $\sigma_t^2=\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \beta_t$. Finally, $\boldsymbol{x}_{t-1}$ can be sampled from $p_\theta\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right)$ as below:

$$\boldsymbol{x}_{t-1}=\frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_t, t\right)\right)+\sigma_t \boldsymbol{z} \tag{7}$$

where $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ is the Gaussian noise.

While DDPM excels at generating high-quality samples without complex adversarial training, it does have a drawback in terms of inference time. Some techniques, such as the denoising diffusion implicit model (DDIM) [22], are proposed to accelerate the sampling process while maintaining the generation of high-quality samples.

### B. Diffusion-based Human Motion Prediction

Although current VAE-based methods like STARS [13] achieve state-of-the-art prediction accuracy, they often suffer from generating excessively diverse and out-of-context predictions because of the diversity-promoting techniques they employ. The emergence of diffusion models has provided a new promising direction for predicting human motion while considering uncertainty. For example, one study [19] proposed using spatial and temporal Transformers arranged in series or in parallel as the motion denoiser. The observed past motion sequence and perturbed future motion sequence are concatenated together and passed into the noise-predicting network. The diffusion step $t$ is injected by first projecting it to the vector space and then adding it directly to the input sequence. While the model does not achieve state-of-the-art performance, it demonstrated that diffusion models can strike a balance between diversity and accuracy, generating motion predictions that are contextually appropriate.

Two works, MotionDiff [15] and TCD [17], also utilized the spatial-temporal Transformer in the diffusion model. MotionDiff [15] employs an encoder-decoder structure, where the past motion sequence is first processed by the encoder network and then concatenated with the diffusion time encoding of step $t$. This encoded information serves as a condition for motion generation and is used multiple times within a single denoising step through a module that combines linear transformations with gating and bias addition. For the decoder part, MotionDiff first utilizes the spatial-temporal Transformer [23] to extract information from the noised future motion sequence. The hidden vector, along with the condition, is then fed into another Transformer block to predict the noise. After obtaining the outputs from the diffusion model, MotionDiff uses a GCN to refine the results, making the approach more intricate and preventing it from being trained in an end-to-end manner. On the other hand, TCD [17] adopts a similar approach to condition the denoiser on the observations and diffusion step $t$, much like as [19] does. However, TCD takes a different perspective on the prediction task by breaking it into two parts: short-term and long-term predictions. Instead of generating the entire prediction sequence at once, the short-term diffusion block aims to predict the first few frames of the future motion sequence based on the observation. The long-term diffusion block then generates the remaining frames using both the observation and the outputs of the short-term diffusion block as the new condition. Another contribution of TCD is its ability to handle imperfect observation by introducing noise to missing elements in the past motion sequence. The authors trained various models for different data-missing situations to validate the effectiveness of their approach.

In [18], an end-to-end diffusion model called HumanMAC was proposed, which solves the prediction problem from the perspective of masked completion. Specifically, the model is trained to generate the entire motion sequence, encompassing both the observed and future motion, starting from random noise. During the inference stage, the future motion is treated as a missing part within the complete sequence, as only the observation part is available. In each denoising step, the noisy known region is sampled from the observation, and the inpainted region is sampled from the output of the previous iteration. These two samples are combined through a mask operation before being passed to the denoiser. Moreover, HumanMAC represents the motion sequence in the frequency space using DCT, thereby reducing the computational cost by discarding high-frequency components. Additionally, adaptive normalization modules are introduced after the self-attention layer and the feed-forward network in the Transformer to guide the prediction using historical information and diffusion steps.

Unlike the aforementioned works, BeLFusion [16] takes a different approach by interpreting diversity from a behavioral perspective rather than focusing on skeleton joint dispersion. The diffusion model utilized in BeLFusion is based on the U-net with cross-attention [24], which allows the model to sample behavior codes in the latent space. These behavior codes are then transferred to the ongoing motion through a behavior coupler, resulting in more realistic predictions. However, BeLFusion requires multiple training stages and complex adversarial training to disentangle behavior from pose and motion, which makes their model difficult to implement.

## III. Methodology

### A. Problem Definition and Notations

We note the complete sequence of human motion as $\boldsymbol{x} = \left[\boldsymbol{q}^{(t-H)}, \dots, \boldsymbol{q}^{(t-2)}, \boldsymbol{q}^{(t-1)}, \boldsymbol{q}^{(t)}, \boldsymbol{q}^{(t+1)}, \dots, \boldsymbol{q}^{(t+F-1)}\right] \in \boldsymbol{R}^{(H+F)\times 3J}$, where $\boldsymbol{q}^{(t)} \in \boldsymbol{R}^{3J}$ is the Cartesian coordinates of human skeleton at the frame $t$, and $J$ is the number of human joints. The first $H$ frames of $\boldsymbol{x}$ correspond to the observation, denoted as $\boldsymbol{x}^O$, and the following $F$ frames represent the future motion $\boldsymbol{x}^P$ to be predicted. Given the observed human motion $\boldsymbol{x}^O$, the objective of HMP consists in predicting the future motion sequence $\boldsymbol{x}^P$. We use $\boldsymbol{y}$ to represent the frequency components after the DCT operation.

### B. Transformer-based Diffusion Model (TransFusion)

We propose a direct adaptation of the diffusion model to the HMP problem. As shown in Fig. 1, in the forward process, the motion sequence $\boldsymbol{x}$ is first projected to the frequency domain via the DCT operation, which is commonly used in HMP [10], [12], [18] due to its ability to encode the temporal nature of human motion. Thus, $\boldsymbol{y} = \mathrm{DCT}(\boldsymbol{x}) = \boldsymbol{D}x$, where $\boldsymbol{D} \in \boldsymbol{R}^{(H+F)\times(H+F)}$ is the DCT basis. Since the DCT operation is an orthogonal transform, we can always recover the original motion sequence from frequency coefficients $\boldsymbol{y}$ by applying the inverse discrete cosine transform (IDCT), i.e., $\boldsymbol{x} = \mathrm{IDCT}(\boldsymbol{y}) = \boldsymbol{D}^\top \boldsymbol{y}$. What's more, considering that the most important and relevant information of human motion is concentrated in the lower frequency coefficients, and the higher frequency terms are mainly related to the noise, we simply keep the first $L$ rows of DCT basis and ignore the remaining $H + F - L$ rows to reduce the dimensionality of the data while processing the DCT and IDCT. Then, the noisy DCT coefficients $\boldsymbol{y}_t$ at any diffusion step $t$ can be sampled by the reparameterization trick:

$$\boldsymbol{y}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{y}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \tag{8}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and $\boldsymbol{y}_0$ equals to $\mathrm{DCT}(\boldsymbol{x})$.

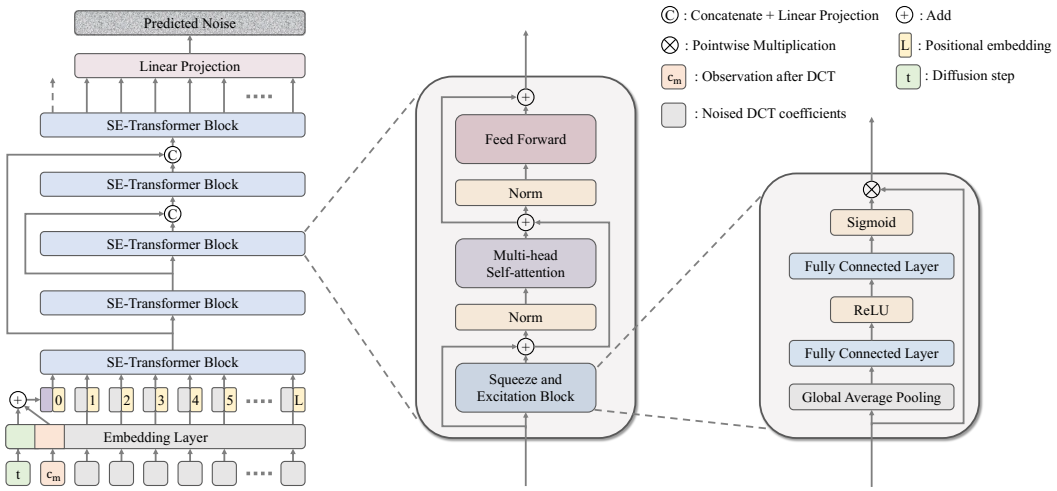


Fig. 2. Architecture of the noise prediction network. The noise prediction network consists of several SE-Transformer blocks. At each diffusion step $t$, and with the inclusion of historical information, the tokens are embedded and combined. Along with positional embeddings, these tokens are then processed through the SE-Transformer blocks. The final outputs from the last layer of SE-Transformer blocks are passed to a linear projection layer, which yields the predicted noise for the given input.

Regarding noise prediction in the backward process for human motion sequences, we propose an alternative approach to the U-net used in the original DDPM [20]. Our approach involves a Transformer-based network denoted as $\boldsymbol{\epsilon}_\theta$, and its architecture is given in Fig. 2. Specifically, in order to enable predictions conditioned on the observation, we leverage the conditional DDPM to guide the sample generation. The observation sequence $\boldsymbol{x}^O$ is first padded to match the length of the complete motion sequence, with the last frame of $\boldsymbol{x}^O$ extended accordingly. Subsequently, the padded sequence is processed through the DCT operator to obtain compact historical information $\boldsymbol{c}_m$. Different from prior works that employed extra modules like cross-attention and adaptive normalization to inject the observation and diffusion step, we calculate the condition by directly adding the encoded diffusion step $t$ and historical data $\boldsymbol{c}_m$ together. This condition, along with all the noisy DCT coefficients, are treated as tokens and fed into the denoiser network which comprises several Transformer layers with long skip connections between shallow and deep layers. The skip connection is achieved by concatenating two tensors and passing them through a linear projection layer. Each Transformer layer is composed of an SE block [21], a self-attention module, and a feed-forward network. The SE block functions as an attention mechanism, but with significantly fewer parameters compared to the self-attention module. It consists of only two fully connected layers with a single pointwise multiplication. The SE block adaptive rescales each token by modeling inter-dependencies among different tokens. It optimizes the Transformer encoder's learning process, and enhances network performance.

Furthermore, in the inference stage, as we already have the historical motion, inspired by [18], we propose the integration of noisy observation guidance at the beginning of each denoising step. This guidance involves several sequential operations. Firstly, we project the denoised DCT coefficients obtained from the previous denoising step and the noisy frequency coefficients acquired from the observation into the temporal domain using IDCT. Subsequently, these components are mixed together via a mask operation, where we define the mask as $\boldsymbol{M} = [1, 1, \ldots, 1, 0, 0, \ldots 0]^\top$. The mask $\boldsymbol{M}$ consists of $H$ elements set to one, representing the noisy observation, and all other elements are set to zero, representing the denoised motion. To distinguish between samples from the last denoising step and the observation, we use $\boldsymbol{y}_t^D$ to denote the denoised samples and $\boldsymbol{y}_t^O$ to denote the observed samples. With these notations, the process of the noisy observation guidance can be summarized as follows:

$$\begin{aligned}\boldsymbol{y}_t = \mathrm{DCT}\big[\boldsymbol{M} \odot \mathrm{IDCT}\left(\boldsymbol{y}_t^O\right) + \\ (\mathbf{1} - \boldsymbol{M}) \odot \mathrm{IDCT}\left(\boldsymbol{y}_t^D\right)\big]\,.\end{aligned} \tag{9}$$

We present the workflow for model training and inference, as outlined in Algorithm 1 and 2.

## IV. Experiments

### A. Experimental Setup

**Datasets.** We evaluate the performance of TransFusion on three benchmark datasets: Human3.6M [26], HumanEva-I [27], and Amass [28]. Human3.6M is a commonly-used benchmark dataset for HMP, comprising 3.6 million frames of human poses recorded at 50 Hz. The dataset consists of 15 daily actions, such as walking and smoking, performed by 11 actors. To ensure a fair comparison with other works, we adopt the widely-used setting proposed by [6]. Specifically, we represent the human pose using 17 joints. The model is trained on 5 subjects (S1, S5, S6, S7, and S8) and tested on 2 subjects (S9 and S11). We utilize 25 frames (0.5 seconds) as the observation to forecast the following 100 frames (2 seconds). Compared to Human3.6m, HumanEva-I is a relatively smaller dataset and exhibits less variation in motion and is recorded at a higher frequency of 60Hz. In line with prior works, we represent the human skeleton using 15 joints and follow the official train/test split provided in the original dataset. The prediction horizon is set to 60 frames (1 second) given an observation of 15 frames (0.25 seconds). AMASS is a large-scale dataset that currently combines 24 extremely varied datasets, all with a standardized joint configuration. It contains

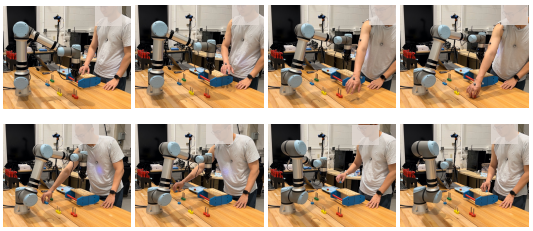

Fig. 3. The HRC reaching motion dataset scenario, depicting human-robot collaboration in collecting screws on the table. The top row illustrates the human reaching out to pick up a screw, while the bottom row showcases the human bringing a screw back. Motion data is recorded using markers attached to the human body.

9 million frames when downsampled to 60 Hz. We follow the same training and evaluation pipeline proposed in [16], and use 30 frames (0.5 seconds) to predict 120 frames (2 seconds).

We also conduct performance evaluations on a HRC reaching motion dataset [4], where the human and robot collaborate in collecting screws from various locations on a table, as depicted in Fig. 3. Mocap system is employed to record the human motion at a frequency of 50 Hz. The human agent is represented as a 5-joint skeleton, consisting of the xiphoid process, the incisura jugularis, the shoulder, the elbow, and the wrist. For training purposes, we utilize 400 data samples, and 63 data samples are reserved for testing. The prediction period is set to 60 frames (equivalent to 1.2 seconds), given an observation window of 15 frames (equivalent to 0.3 seconds).

**Algorithm 1** Training procedure of TransFusion

**Input:** complete motion sequence $\boldsymbol{x}$, diffusion steps $T$, denoiser $\epsilon_\theta$, maximum training epoch $E_{max}$.

**for** $i = 0, 1, \cdots, E_{max}$ **do**
  $\boldsymbol{x}_0 \sim p(\boldsymbol{x}), \quad \boldsymbol{y}_0 = \mathrm{DCT}(\boldsymbol{x}_0)$
  $\boldsymbol{c}_m = \mathrm{DCT}(\mathrm{Padding}(\boldsymbol{x}_0^O))$
  $t \sim \mathrm{Uniform}(\{1, 2, \cdots, T\}), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
  $\boldsymbol{\theta} = \boldsymbol{\theta} - \nabla_{\boldsymbol{\theta}} \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}} \left( \sqrt{\bar{\alpha}_t} \boldsymbol{y}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}, \boldsymbol{c}_m, t \right) \right\|^2$
**end for**

**Output:** trained denoiser network $\epsilon_\theta$

**Algorithm 2** Inference procedure of TransFusion

**Input:** observed motion sequence $\boldsymbol{x}^O$, diffusion steps $T$, the mask of the observation $\boldsymbol{M}$, trained denoiser $\epsilon_\theta$.

$\boldsymbol{y}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
$\boldsymbol{y} = \boldsymbol{c}_m = \mathrm{DCT}(\mathrm{Padding}(\boldsymbol{x}^O))$
**for** $t = T, T-1, \cdots, 1$ **do**
  $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ if $t > 1$, else $\boldsymbol{z} = \mathbf{0}$
  $\boldsymbol{y}_{t-1}^O = \sqrt{\bar{\alpha}_{t-1}} \boldsymbol{y} + \sqrt{1 - \bar{\alpha}_{t-1}} \boldsymbol{z}$
  $\boldsymbol{y}_{t-1}^D = \frac{1}{\sqrt{\alpha_t}} \left( \boldsymbol{y}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_{\boldsymbol{\theta}} (\boldsymbol{y}_t, \boldsymbol{c}_m, t) \right) + \sigma_t \boldsymbol{z}$
  $\boldsymbol{x}_{t-1}^O = \mathrm{IDCT}\left(\boldsymbol{y}_{t-1}^O\right), \quad \boldsymbol{x}_{t-1}^D = \mathrm{IDCT}\left(\boldsymbol{y}_{t-1}^D\right)$
  $\boldsymbol{y}_{t-1} = \mathrm{DCT}\left(\boldsymbol{M} \odot \boldsymbol{x}_{t-1}^O + (\mathbf{1} - \boldsymbol{M}) \odot \boldsymbol{x}_{t-1}^D\right)$
**end for**
$\boldsymbol{x} = \mathrm{IDCT}(\boldsymbol{y}_0)$

**Output:** complete motion sequence $\boldsymbol{x}$

**Evaluation metrics.** We adopt the commonly-used pipeline proposed in [6] and use five metrics to evaluate our model's performance: (1) Average Pairwise Distance (APD): This metric computes the average L2 distance between all pairs of motion samples, serving as a measure of diversity within the predicted future motions. (2) Average Displacement Error (ADE): ADE calculates the smallest average L2 distance over all time steps between the ground truth and predicted samples, evaluating the prediction accuracy. (3) Final Displacement Error (FDE): FDE measures the smallest L2 distance in the last time frame between the prediction results and ground truth, also evaluating the prediction accuracy. (4) Multi-Modal ADE (MMADE): This metric is the multi-modal version of ADE, assessing the model's ability to capture the multi-modality nature of human motion. Multi-modal ground truth future motions are obtained by grouping similar observations. (5) Multi-Moddal FDE (MMFDE): Similar to MMADE, MMFDE is the multi-modal version of FDE.

While these metrics provide valuable insights, we argue that relying solely on them may not be sufficient. Prior works often prioritize increasing the APD to enhance diversity, sometimes leading to unrealistic and implausible predictions. Moreover, the accuracy evaluation of previous works is based on the best-of-many strategy, where only the closest sample to the ground truth is considered, potentially overlooking predictions that deviate significantly from reality. This approach might not be suitable for real-world tasks, such as HRC.

In our study, we aim to generate not just one good sample close to the ground truth but as many good predictions as possible while maintaining a certain degree of diversity. To achieve this, we introduce additional strategies: worst-of-many and median-of-many, along with their corresponding metrics: (6) ADE-W, (7) FDE-W, (8) MMADE-W, (9) MMFDE-W for worst-of-many evaluation, and similarly, (10) ADE-M, (11) FDE-M, (12) MMADE-M, (13) MMFDE-M for median-of-many evaluation. By incorporating these strategies, we offer a more comprehensive evaluation of our model's performance, considering both accuracy and diversity in the predictions.

**Baselines.** To assess the effectiveness of our model, we conduct a comparative evaluation with several state-of-the-art works, which include DeLiGAN [5], DLow [6], DivSamp [7], BoM [8], DSF [9], MOJO [10], MultiObj [11], GSPS [12], STARS [13], Motron [14], MotionDiff [15], BeLFusion [16], and HumanMAC [18]. Notably, we exclude TCD [17] as it adopts a two-stage prediction strategy, which can be readily combined with other prediction models. Therefore, considering TCD as a baseline would not be fair.

**Implementation details.** We train TransFusion with 1,000 diffusion steps and the cosine variance schedule [25]. During training, we sample 50,000 data from the training set in each epoch for all benchmark datasets. We train the model for 1,500 epochs on Human3.6M with a batch size of 64, 100 epochs on HumanEva-I, and 3000 epochs on AMASS with the same batch size. We also disregard historical observation with a probability of 0.2 during training to regularize the model. The learning rate is initialized to $3 \times 10^{-4}$ and is decayed by a ratio of 0.8 every 100 epochs for Human3.6M and HumanEva-I, and every 200 epochs for AMASS. For the noise prediction network, we use 9-layer SE-Transformer blocks for Human3.6M, 5-layer SE-Transformer blocks for HumanEva-I, and 13-layer SE-Transformer blocks for AMASS. Furthermore, we use the first 20 rows of DCT coefficients for all datasets. Following common practice, we set the dimension of the hidden state to 512 for all benchmark datasets. To expedite the inference process, we leverage a 100-step DDIM [22] and generate 50 predictions for each single observation. All experiments are conducted using PyTorch and a single NVIDIA A100 GPU.

TABLE I
QUANTITATIVE RESULTS WITH BEST-OF-MANY STRATEGY ON HUMAN3.6M AND HUMANEVA-I

| Method | Human3.6M | | | | | HumanEva-I | | | | | AMASS | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | APD | ADE | FDE | MMADE | MMFDE | APD | ADE | FDE | MMADE | MMFDE | APD | ADE | FDE | MMADE | MMFDE |
| DeLiGAN [5] | 6.509 | 0.483 | 0.534 | 0.520 | 0.545 | 2.177 | 0.306 | 0.322 | 0.385 | 0.371 | - | - | - | - | - |
| DLow [6] | 11.741 | 0.425 | 0.518 | 0.495 | 0.531 | 4.855 | 0.251 | 0.268 | 0.362 | 0.339 | 13.170 | 0.590 | 0.612 | 0.618 | 0.617 |
| DivSamp [7] | 15.310 | 0.370 | 0.485 | 0.475 | 0.516 | **6.109** | 0.220 | 0.234 | 0.342 | **0.316** | **24.724** | 0.564 | 0.647 | 0.623 | 0.667 |
| BoM [8] | 6.265 | 0.448 | 0.533 | 0.514 | 0.544 | 2.846 | 0.271 | 0.279 | 0.373 | 0.351 | - | - | - | - | - |
| DSF [9] | 9.330 | 0.493 | 0.592 | 0.550 | 0.599 | 4.538 | 0.273 | 0.290 | 0.364 | 0.340 | - | - | - | - | - |
| MOJO [10] | 12.579 | 0.412 | 0.514 | 0.497 | 0.538 | 4.181 | 0.234 | 0.244 | 0.369 | 0.347 | - | - | - | - | - |
| MultiObj [11] | 14.240 | 0.414 | 0.516 | - | - | 5.786 | 0.228 | 0.236 | - | - | - | - | - | - | - |
| GSPS [12] | 14.757 | 0.389 | 0.496 | 0.476 | 0.525 | 5.825 | 0.233 | 0.244 | 0.343 | 0.331 | 12.465 | 0.563 | 0.613 | 0.609 | 0.633 |
| STARS [13] | **15.884** | **0.358** | **0.445** | **0.442** | **0.471** | 6.031 | 0.217 | 0.241 | **0.328** | 0.321 | - | - | - | - | - |
| Motron [14] | 7.168 | 0.375 | 0.488 | - | - | - | - | - | - | - | - | - | - | - | - |
| MotionDiff [15] | 15.353 | 0.411 | 0.509 | 0.508 | 0.536 | 5.931 | 0.232 | 0.236 | 0.352 | 0.320 | - | - | - | - | - |
| BeLFusion [16] | 7.602 | 0.372 | 0.474 | 0.473 | 0.507 | - | - | - | - | - | 9.376 | 0.513 | 0.560 | **0.569** | **0.591** |
| HumanMAC [18] | 6.301 | 0.369 | 0.480 | 0.509 | 0.545 | 6.554** | 0.209 | **0.223** | 0.342 | 0.335 | 9.321 | 0.511 | **0.554** | 0.593 | **0.591** |
| TransFusion | 5.975 | **0.358** | 0.468 | 0.506 | 0.539 | 1.031 | **0.204** | 0.234 | 0.408 | 0.427 | 8.853 | **0.508** | 0.568 | 0.589 | 0.606 |

* Bolded numbers indicate the best results, and numbers with underline represent the second best results. For all accuracy metrics, lower values are preferred. It is important to note that APD measures the difference among the 50 prediction results, and a larger APD does not necessarily indicate better performance. The symbol '-' indicates that certain results are not reported in the baselines, and '**' denotes that the result reported in the baseline comes from dropping the observation during inference with a probability of 50%. This will increase APD but will generate some unrealistic predictions.

TABLE II
QUANTITATIVE RESULTS WITH MEDIAN-OF-MANY AND WORST-OF-MANY STRATEGY ON HUMAN3.6M

| Model | Human3.6M (Median-of-many / Worst-of-many) | | | | AMASS (Median-of-many / Worst-of-many) | | | |
|---|---|---|---|---|---|---|---|---|
| | ADE-M/W | FDE-M/W | MMADE-M/W | MMFDE-M/W | ADE-M/W | FDE-M/W | MMADE-M/W | MMFDE-M/W |
| DLow [6] | 0.896 / 1.763 | 1.285 / 2.655 | 0.948 / 1.804 | 1.290 / 2.657 | 0.977 / 2.138 | 1.186 / 2.994 | 0.996 / 2.156 | 1.181 / 2.991 |
| DivSamp [7] | 0.924 / 2.497 | 1.344 / 3.263 | 1.001 / 2.530 | 1.359 / 3.267 | 1.958 / 3.269 | 1.718 / 4.479 | 1.970 / 3.272 | 1.715 / 4.478 |
| GSPS [12] | 1.014 / 2.458 | 1.375 / 2.964 | 1.066 / 2.480 | 1.383 / 2.964 | 1.089 / 1.799 | 1.364 / 2.642 | 1.099 / 1.808 | 1.358 / 2.638 |
| STARS [13] | 0.815 / 3.389 | 1.159 / 3.715 | 0.870 / 3.396 | 1.164 / 3.711 | - | - | - | - |
| BeLFusion [16] | 0.673 / 1.355 | 0.976 / 2.038 | 0.767 / 1.418 | 1.009 / 2.046 | 0.817 / 1.791 | 1.069 / 2.237 | 0.857 / 1.815 | **1.074** / 2.236 |
| HumanMAC [18] | 0.585 / 1.085 | 0.911 / 1.843 | 0.736 / 1.205 | 0.977 / 1.877 | - | - | - | - |
| TransFusion | **0.575 / 1.063** | **0.898 / 1.758** | **0.729 / 1.179** | **0.967 / 1.791** | **0.758 / 1.339** | **1.060 / 2.063** | **0.832 / 1.389** | 1.080 / **2.067** |

* Quantitative results of baselines are calculated from pretrained models. The symbol '-' indicates that certain results are not reported in the baselines.

Training on Human3.6M, HumanEva-I, and AMASS takes around 24, 1, and 49 hours respectively. Adam is used as the optimizer for all experiments.

TABLE III
COMPLEXITY COMPARISON

| Model | Human3.6M (Best-of-many) | | | | |
|---|---|---|---|---|---|
| | #Params | Avg. Inf. Time (Sec) | APD | ADE | FDE |
| HumanMAC-8 [18] | 28.40M | 1.266 | 6.301 | 0.369 | 0.480 |
| TransFusion-7 | **15.52M** | 0.899 | 6.537 | 0.363 | 0.471 |
| TransFusion-9 | 19.73M | 1.110 | 5.975 | **0.358** | **0.468** |
| TransFusion-9-DDIM10 | 19.73M | **0.123** | **6.941** | 0.362 | 0.471 |

* The number following the model name indicates the number of layers in the noise prediction network. If not specified, the model uses 100-step DDIM for sampling.

### *B. Comparison with the State-of-the-Arts*

We first follow the best-of-many strategy to compare TransFusion with existing works. The quantitative results are presented in Table I. For Human3.6M and HumanEva-I, TransFusion achieves the best result on ADE and the second-best result on FDE, and for AMASS, TransFusion outperforms all the baselines on ADE, and gets the second-best results on MMADE and MMFDE. These findings indicate that our model generates predictions that are closer to the ground truth. While the APD results from our model are not as favorable as some prior works due to not considering explicit diversity-prompting techniques, and opting instead to let DDPM learn the true data distribution from training set, we have already established that the higher APD does not necessarily imply better predictions, and sometimes the situation can be opposite. Excessive diversity can lead to predictions that significantly deviate from reality, resulting in unrealistic and overly conservative outcome. As a result, such methods may hinder efficiency in downstream tasks or even fail to meet the requirements of certain applications, such as motion planning for robot manipulators in HRC. Therefore, we prioritize a balanced approach and do not solely focus on increasing the APD in this work.

Since the commonly-used strategy only evaluates the quality of the closest prediction, we also assess the overall prediction quality by providing results following the median-of-many and worst-of-many strategies in Table II. As shown in the table, TransFusion surpasses all other baselines across all accuracy metrics, highlighting its superior overall prediction performance. We further illustrate the quality of the motion predicted by our model through visualization in Fig. 4. From the visualization, we can conclude that the predictions generated by our model closely match the ground truth. Furthermore, while maintaining semantic consistency with historical information, the predictions still exhibit a certain level of diversity.

We also observe that HumanMAC [18] achieves slightly inferior performance compared to ours. To delve deeper, we compare the number of parameters and average inference time of both models in Table III. Remarkably, TransFusion achieves better results than HumanMAC while utilizing only 54.6% of the parameters and reducing inference time by 29.0%. This efficiency gain is attributed to the fact that our model incorporates conditions without relying on any additional modules, unlike other existing works. The average inference time can be further reduced to 0.123 seconds when we use 10 sampling steps, without compromising sample quality.

Moreover, we adopt the best-of-50 strategy to evaluate our model on the HRC reaching motion dataset, and we retrain HumanMAC [18] on our dataset. The results demonstrate that our model has an APD equal to 0.727, an ADE equal to 0.035,

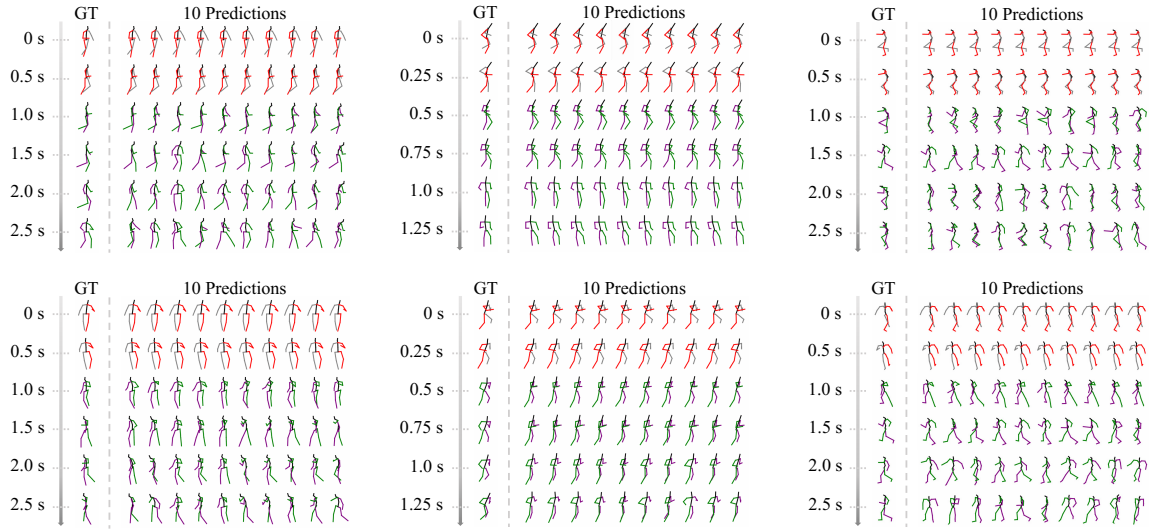


Fig. 4. Visualization results. The left column displays the results from Human3.6M, with each example corresponding to 'Walking' and 'Smoking'. The middle column presents the results from HumanEva-I, with the labels being 'Jogging' and 'Boxing'. The right column presents the results from AMASS. For each sample, the first column represents the ground truth, while the following 10 columns depict the prediction results. The observed context for each motion is represented by red and black skeletons, while the future motion is indicated by green and purple skeletons.

and a FDE equal to 0.047, while HumanMAC has an APD equal to 0.837, an ADE equal to 0.037, and a FDE equal to 0.050. These metrics reinforce the applicability of our model to HRC scenarios while accounting for uncertainty.

### C. Ablation Study

We conduct comprehensive ablation studies to explore different design choices of TransFusion, including the skip connections, the utilization of SE block and DCT, the number of rows ($L$) we used in DCT/IDCT, and the number of layers in the noise prediction network. The following evaluations are based on the best-of-many strategy.

We begin by evaluating the effectiveness of long skip connections in our network. As shown in Table IV, it is evident that the inclusion of skip connections enhances the model's performance, as both the concatenation and addition-based skip connections outperform the model without skip connections. Moreover, the model with concatenation-based skip connections demonstrates superior performance compared to the one with addition-based skip connections. We argue that the addition-based method of simply adding the branches together, without employing a linear projection, does not significantly benefit the model learning process. This is because in the addition-based method, the deeper layers already have a direct path from shallower layers due to the presence of residual connections in SE-Transformer blocks. As a result, the concatenation-based skip connections offer more advantages, leading to improved model performance.

TABLE IV
QUANTITATIVE RESULTS OF THE ABLATION STUDY ON THE SKIP CONNECTIONS

| Model | Human3.6M | | | HumanEva-I | | | AMASS | | |
|---|---|---|---|---|---|---|---|---|---|
| | APD | ADE | FDE | APD | ADE | FDE | APD | ADE | FDE |
| Concat + Proj | 5.975 | **0.358** | **0.468** | **1.031** | **0.204** | **0.234** | **8.853** | **0.508** | **0.568** |
| Add | 7.224 | 0.404 | 0.508 | 0.996 | 0.206 | 0.235 | 8.824 | 0.511 | 0.571 |
| w/o skip | **7.447** | 0.411 | 0.511 | 0.945 | 0.205 | 0.237 | 8.723 | 0.512 | 0.573 |

As previously mentioned, we use DCT to transfer the data from time domain to frequency domain, and add an SE block in the Transformer encoder to optimize learning. The results shown in Table V indicate that the utilization of DCT and the addition of the SE module indeed enhance the model's performance, as evidenced by improved ADE and FDE results.

TABLE V
QUANTITATIVE RESULTS OF THE ABLATION STUDY ON SE AND DCT

| Model | Human3.6M | | | HumanEva-I | | | AMASS | | |
|---|---|---|---|---|---|---|---|---|---|
| | APD | ADE | FDE | APD | ADE | FDE | APD | ADE | FDE |
| w SE & DCT | 5.975 | **0.358** | **0.468** | **1.031** | **0.204** | **0.234** | 8.853 | **0.508** | **0.568** |
| w/o SE | **6.071** | 0.361 | 0.472 | 1.004 | 0.210 | 0.237 | **9.018** | 0.509 | 0.571 |
| w/o DCT | 0.622 | 0.600 | 0.855 | 0.549 | 0.566 | 0.647 | 0.786 | 0.760 | 0.973 |

TABLE VI
QUANTITATIVE RESULTS OF THE ABLATION STUDY ON L

| DCT-L | Human3.6M | | | HumanEva-I | | | AMASS | | |
|---|---|---|---|---|---|---|---|---|---|
| | APD | ADE | FDE | APD | ADE | FDE | APD | ADE | FDE |
| 5 | 5.738 | 0.381 | 0.493 | 0.797 | 0.227 | 0.284 | 8.429 | 0.529 | 0.573 |
| 10 | 5.790 | 0.364 | 0.473 | 0.931 | **0.204** | 0.244 | 8.967 | 0.513 | **0.567** |
| 20 | 5.975 | **0.358** | **0.468** | 1.031 | **0.204** | 0.234 | 8.853 | **0.508** | 0.568 |
| 30 | **6.018** | 0.360 | 0.470 | 1.136 | **0.204** | 0.229 | 8.884 | 0.513 | 0.574 |
| Full | 5.798 | 0.365 | 0.472 | **1.589** | 0.209 | **0.216** | **9.027** | 0.517 | 0.575 |

Additionally, we investigate the design choice of DCT, and the impact of the dimensionality of the problem by using only the first $L$ rows of DCT basis. Smaller values of $L$ may result in the loss of important information, while larger values may add computational burden to the model, and include the irrelevant noise information during the training process. Therefore, we assess the influence of $L$ on the model's performance, and the results are provided in Table VI. For Human3.6M and AMASS, the best ADE and FDE metrics are achieved when $L = 20$, whereas for HumanEva-I, more than one model could achieve best ADE. We opt to set $L$ equal to 20 for all datasets, as it achieves good accuracy without imposing excessive computational burden on the model.

Table VII presents the results of experiments with different number of layers. For Human3.6M, we use 9 layers in our model as it yields the best performance in both ADE and FDE. For HumanEva-I, a 5-layer network shows the best FDE, while a 7-layer network performs best in terms of ADE.

TABLE VII
QUANTITATIVE RESULTS OF THE ABLATION STUDY ON #LAYERS

| | Human3.6M | | | HumanEva-I | | | AMASS | | |
|---|---|---|---|---|---|---|---|---|---|
| #Layers | APD | ADE | FDE | APD | ADE | FDE | APD | ADE | FDE |
| 3 | **8.172** | 0.416 | 0.533 | **1.288** | 0.210 | 0.238 | - | - | - |
| 5 | 6.494 | 0.374 | 0.485 | 1.031 | 0.204 | **0.234** | - | - | - |
| 7 | 6.537 | 0.363 | 0.471 | 0.923 | **0.203** | 0.236 | 8.777 | 0.530 | 0.588 |
| 9 | 5.975 | **0.358** | **0.468** | 0.872 | 0.206 | 0.239 | **9.044** | 0.523 | 0.580 |
| 11 | 5.722 | 0.360 | 0.472 | 0.846 | 0.205 | 0.243 | 8.841 | 0.513 | 0.573 |
| 13 | - | - | - | - | - | - | 8.853 | 0.508 | 0.568 |
| 15 | - | - | - | - | - | - | 8.771 | **0.505** | **0.567** |

Considering efficiency, we finally choose the 5-layer network for HumanEva-I. Similarly, we choose the 13-layer network for AMASS, as it outperforms the baselines and offers greater efficiency compared to the 15-layer network.

## V. CONCLUSIONS

This paper presents TransFusion, a practical and effective diffusion-based HMP method, leveraging the Transformer as the backbone with long skip connections between shallow and deep layers. We design the model in the frequency domain, utilizing the DCT operation. To condition the predictions on historical information, we treat the conditions as a token, avoiding the use of any additional module like cross-attention and adaptive normalization. The extensive experimental studies demonstrate that our model achieves state-of-the-art prediction results in terms of accuracy. In contrast to prior works often prioritize diversity and produce unrealistic future motions, TransFusion stands out by offering superior overall prediction quality while still maintaining a certain degree of diversity. The model's ability to strike a balance between accuracy and diversity makes it a promising solution for HMP tasks. A future research direction could involve proposing more efficient sampling methods, such as reducing denoising steps and parallelizing the denoising process [29], and integrating them with TransFusion to further reduce inference time without retraining the model.

## REFERENCES


[1] Lee, M. L., Behdad, S., Liang, X., & Zheng, M. (2022). Task allocation and planning for product disassembly with human–robot collaboration. Robotics and Computer-Integrated Manufacturing, 76, 102306.

[2] Sajedi, S., Liu, W., Eltouny, K., Behdad, S., Zheng, M., & Liang, X. (2022). Uncertainty-assisted image-processing for human-robot close collaboration. IEEE Robotics and Automation Letters, 7(2), 4236-4243.

[3] Zhang, X., Yi, D., Behdad, S., & Saxena, S. (2023). Unsupervised Human Activity Recognition Learning for Disassembly Tasks. IEEE Transactions on Industrial Informatics.

[4] Tian, S., Liang, X., & Zheng, M. (2023, May). An Optimization-Based Human Behavior Modeling and Prediction for Human-Robot Collaborative Disassembly. In 2023 American Control Conference (ACC) (pp. 3356-3361). IEEE.

[5] Gurumurthy, S., Kiran Sarvadevabhatla, R., & Venkatesh Babu, R. (2017). Deligan: Generative adversarial networks for diverse and limited data. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 166-174).

[6] Yuan, Y., & Kitani, K. (2020). Dlow: Diversifying latent flows for diverse human motion prediction. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16 (pp. 346-364). Springer International Publishing.

[7] Dang, L., Nie, Y., Long, C., Zhang, Q., & Li, G. (2022, October). Diverse Human Motion Prediction via Gumbel-Softmax Sampling from an Auxiliary Space. In Proceedings of the 30th ACM International Conference on Multimedia (pp. 5162-5171).

[8] Bhattacharyya, A., Schiele, B., & Fritz, M. (2018). Accurate and diverse sampling of sequences based on a "best of many" sample objective. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 8485-8493).

[9] Yuan, Y., & Kitani, K. (2019). Diverse trajectory forecasting with determinantal point processes. arXiv preprint arXiv:1907.04967.

[10] Zhang, Y., Black, M. J., & Tang, S. (2021). We are more than our joints: Predicting how 3d bodies move. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3372-3382).

[11] Ma, H., Li, J., Hosseini, R., Tomizuka, M., & Choi, C. (2022). Multi-objective diverse human motion prediction with knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8161-8171).

[12] Mao, W., Liu, M., & Salzmann, M. (2021). Generating smooth pose sequences for diverse human motion prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 13309-13318).

[13] Xu, S., Wang, Y. X., & Gui, L. Y. (2022, October). Diverse human motion prediction guided by multi-level spatial-temporal anchors. In European Conference on Computer Vision (pp. 251-269). Cham: Springer Nature Switzerland.

[14] Salzmann, T., Pavone, M., & Ryll, M. (2022). Motron: Multimodal probabilistic human motion forecasting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6457-6466).

[15] Wei, D., Sun, H., Li, B., Lu, J., Li, W., Sun, X., & Hu, S. (2023, June). Human joint kinematics diffusion-refinement for stochastic motion prediction. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 5, pp. 6110-6118).

[16] Barquero, G., Escalera, S., & Palmero, C. (2023). Belfusion: Latent diffusion for behavior-driven human motion prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2317-2327).

[17] Saadatnejad, S., Rasekh, A., Mofayezi, M., Medghalchi, Y., Rajabzadeh, S., Mordan, T., & Alahi, A. (2023, May). A generic diffusion-based approach for 3D human pose prediction in the wild. In 2023 IEEE International Conference on Robotics and Automation (ICRA) (pp. 8246-8253). IEEE.

[18] Chen, L. H., Zhang, J., Li, Y., Pang, Y., Xia, X., & Liu, T. (2023). Humanmac: Masked motion completion for human motion prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 9544-9555).

[19] Ahn, H., Mascaro, E. V., & Lee, D. (2023, May). Can we use diffusion probabilistic models for 3d motion prediction?. In 2023 IEEE International Conference on Robotics and Automation (ICRA) (pp. 9837-9843). IEEE.

[20] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33, 6840-6851.

[21] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141).

[22] Song, J., Meng, C., & Ermon, S. (2020). Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502.

[23] Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., & Ding, Z. (2021). 3d human pose estimation with spatial and temporal transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 11656-11665).

[24] Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems, 34, 8780-8794.

[25] Nichol, A. Q., & Dhariwal, P. (2021, July). Improved denoising diffusion probabilistic models. In International Conference on Machine Learning (pp. 8162-8171). PMLR.

[26] Ionescu, C., Papava, D., Olaru, V., & Sminchisescu, C. (2013). 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.

[27] Sigal, L., Balan, A. O., & Black, M. J. (2010). Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. International journal of computer vision, 87(1-2), 4.

[28] Mahmood, N., Ghorbani, N., Troje, N. F., Pons-Moll, G., & Black, M. J. (2019). AMASS: Archive of motion capture as surface shapes. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 5442-5451).

[29] Shih, A., Belkhale, S., Ermon, S., Sadigh, D., & Anari, N. (2024). Parallel sampling of diffusion models. Advances in Neural Information Processing Systems, 36.