GazeMoDiff: Gaze-guided Diffusion Model for Stochastic Human Motion Prediction

Haodong Yan[†1], Zhiming Hu[‡§2], Syn Schmitt[2,3] and Andreas Bulling[2,3]

[1]The Hong Kong University of Science and Technology (Guangzhou), China
[2]University of Stuttgart, Germany
[3]Center for Bionic Intelligence Tuebingen Stuttgart (BITS), Germany

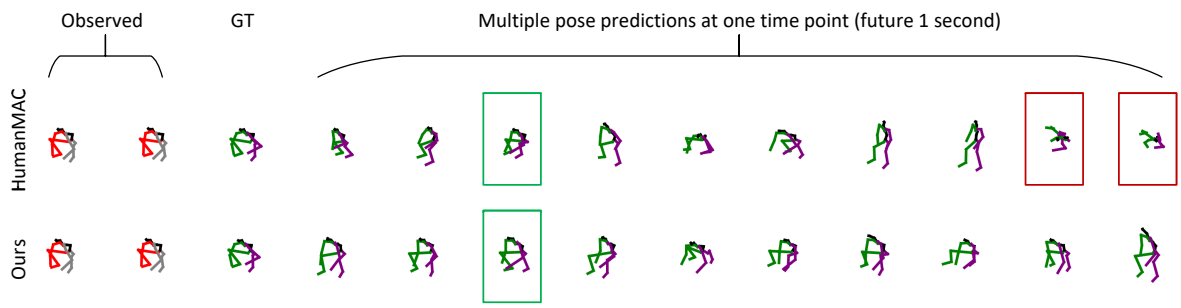

Figure 1: *Ground truth (GT) human pose at future one second and multiple pose predictions generated by different methods on the GIMO dataset [ZYM*22] with the lowest error prediction boxed in green and implausible cases marked in red. Our method can generate multiple reasonable motions while other methods produce some implausible predictions.*

Abstract

Human motion prediction is important for many virtual and augmented reality (VR/AR) applications such as collision avoidance and realistic avatar generation. Existing methods have synthesised body motion only from observed past motion, despite the fact that human eye gaze is known to correlate strongly with body movements and is readily available in recent VR/AR headsets. We present GazeMoDiff – a novel gaze-guided denoising diffusion model to generate stochastic human motions. Our method first uses a gaze encoder and a motion encoder to extract the gaze and motion features respectively, then employs a graph attention network to fuse these features, and finally injects the gaze-motion features into a noise prediction network via a cross-attention mechanism to progressively generate multiple reasonable human motions in the future. Extensive experiments on the MoGaze and GIMO datasets demonstrate that our method outperforms the state-of-the-art methods by a large margin in terms of multi-modal final displacement error (17.3% on MoGaze and 13.3% on GIMO). We further conducted a human study (N=21) and validated that the motions generated by our method were perceived as both more precise and more realistic than those of prior methods. Taken together, these results reveal the significant information content available in eye gaze for stochastic human motion prediction as well as the effectiveness of our method in exploiting this information.

CCS Concepts

• ***Computing methodologies*** → *Machine learning; Machine learning approaches; Neural networks;* • ***Human-centred computing*** → *Human computer interaction; Virtual reality; Interaction paradigms;*

† Equal Contribution
‡ Equal Contribution
§ Corresponding author, zhiming.hu@vis.uni-stuttgart.de

© 2024 The Authors.
Proceedings published by Eurographics - The European Association for Computer Graphics.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.2312/pg.20241315

delivered by EUROGRAPHICS DIGITAL LIBRARY www.eg.org diglib.eg.org

1. Introduction

Generating realistic human body movements is a key research challenge in the area of virtual and augmented reality (VR/AR) and is the basis for safe, smooth, and immersive human-environment [DJPZ*21, SPW*18] and human-human interactions [KALL17, Hol12]. Human motion prediction (HMP) enables a number of exciting applications such as redirected walking to create the illusion of unlimited virtual interaction spaces [AGBS16, SPW*18] or to steer users away from physical boundaries, such as walls, and thus avoid collisions [GIWW21, ZYM*22]. Human motion prediction can also provide users with a low-latency experience by preparing VR content in advance based on the predicted future human poses [HZBD19, CWL*22] and it has been used to generate human-like motions for virtual agents to enhance the interaction experience [BRB*21, DKP*23].

Previous work on HMP has typically generated human motions in a deterministic way, i.e. by producing only a single prediction at a time [MBR17, BBKK17, CGM*20, GMK*19]. Recently, in light of the fact that human motion is stochastic by nature [CZL*23, BEP22], researchers have turned to stochastic human motion prediction, i.e. generating multiple reasonable human motions at a time [CZL*23, YK20, BEP22, BKL18]. Stochastic human motion prediction suits the needs of many VR applications. For example, to minimise the collision risk in a virtual environment, it is necessary to consider multiple possible future trajectories to warn the users. To produce realistic virtual agents, it is also beneficial to synthesise multiple reasonable human motions that users can peruse and select from according to their personal preferences. However, existing stochastic HMP methods typically generate human motions using only past observed motions and neglect other modalities, in particular human eye gaze. With rapid advances in eye tracking technology, human eye gaze information has become readily available in many VR/AR head mounted displays (HMDs), such as HTC Vive Pro Eye, Varjo VR-3, and Vision Pro, and has demonstrated its potential for gaze-based interaction [SG19b] and gaze-contingent rendering [HZL*19, HLZ*20] in VR/AR. In addition, a large body of work in the cognitive sciences and human-centred computing has shown that human body movements are closely coordinated with human gaze behaviour [HBLW21a, HBLW21b, SG19a, Fre08]. Despite this close coordination, information on eye gaze has not been used for stochastic human motion prediction so far.

To address this limitation we propose *GazeMoDiff* – the first **Gaze**-guided human **Mo**tion **Diff**usion model to generate multiple reasonable human motions. Our method first uses a 1D convolutional neural network (1D CNN) and a graph attention network (GAT) to extract features from eye gaze and body motion respectively, then employs a spatio-temporal graph attention network to fuse these features, and finally injects the gaze-motion features into a noise prediction network via a cross-attention mechanism to generate multiple reasonable human future motions through a progressive denoising process. We extensively evaluate our method on the MoGaze dataset [KBM*20] for real-world settings as well as on the GIMO dataset [ZYM*22] for AR settings. Experimental results demonstrate that our method significantly outperforms the state-of-the-art methods that only use past body poses, achieving an improvement of 16.7% on MoGaze and 10.8% on GIMO in multi-modal average displacement error and 17.3% on MoGaze and 13.3% on GIMO in multi-modal final displacement error. To evaluate our method comprehensively, we further conducted a user study and the responses from 21 users validated that the motions generated by our method are perceived as both more precise and more realistic than predictions of prior methods. The full source code and trained models are available at zhiminghu.net/yan24_gazemodiff.

The main contributions of our work are three-fold:

- We propose a novel gaze-guided diffusion model for stochastic human motion prediction that uses a spatio-temporal graph attention network to fuse the gaze and motion features and then injects these features into a noise prediction network via a cross-attention mechanism to generate multiple reasonable human motions in the future.
- We conduct extensive experiments on two public datasets for both real-world and AR settings and demonstrate significant performance improvements over state-of-the-art methods.
- We report a user study that shows the motions generated by our method are perceived as more precise and more realistic than those from prior methods.

2. Related Work

2.1. Human Motion Prediction

Human motion prediction is a fundamental research topic in virtual and augmented reality. Early studies commonly considered HMP as a deterministic sequence prediction task, tackling it with recurrent neural network (RNN) [MBR17, GSAH17, LZX*17, FLFM15], graph neural network (GNN) [LCPW21, LCZ*21, HYH*24] and Transformers [AKCH21, MGVO21]. Noticing that human body motions are inherently stochastic, recent works began to predict human motions in a stochastic way using generative models such as variational autoencoder (VAE) [ASP*21, MLS21], generative adversarial network (GAN) [KGB19, BKL18, JZJ*20], and Flow networks [YK20]. These methods can generate diverse human motions through a diversity-aware loss or sampling strategy [MLS21, YK20], but their predictions are not physically plausible. Inspired by the recent success of denoising diffusion models in the field of image generation [HJA20, RDN*22, NDR*21, ZA23], researchers have adopted diffusion models to the task of stochastic human motion prediction and have achieved more realistic predictions than traditional generative approaches [BEP22, CZL*23]. However, existing methods only generate future motion predictions based on observed past motion, neglecting the fact that human eye gaze is known to correlate strongly with body movements and is readily available in recent VR/AR headsets. To fill this gap, in this work we propose the first gaze-guided diffusion model for stochastic human motion prediction.

2.2. Correlations between Eye Gaze and Body Motion

Intuitively, a period of human eye gaze information seems related to human intention [KN18], which can drive future motion trends. Extensive works in cognitive science and human-centred computing have also demonstrated this strong correlation between human eye gaze and subsequent body motion. Some researchers have revealed that in many everyday activities, such as free viewing or

© 2024 The Authors.
Proceedings published by Eurographics - The European Association for Computer Graphics.

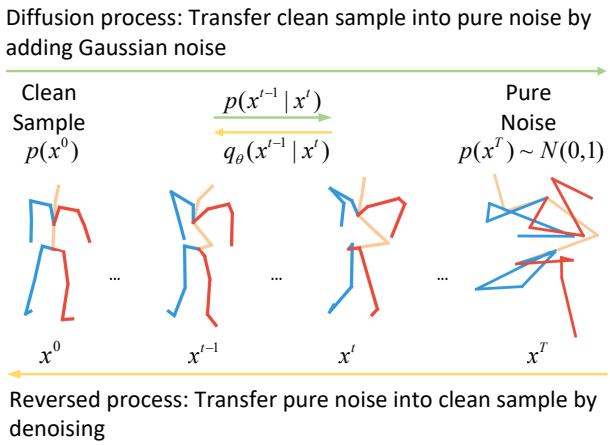

Figure 2: *Diffusion process and reversed process in DDPM.*

object searching, human head movements are closely associated with eye movements [ZS82, HZL*19, KYK*20, HXSB24]. Emery et al. further explored the coordination of human eye, hand, and head movements in virtual environments and leveraged this coordination to improve the performance of gaze estimation [EZW*21]. Sidenmark et al. identified the coordinated movements between eye, head, and torso during gaze shifts in virtual reality [SMG*20, SG19a]. Despite the strong link between human eye gaze and body motion, prior work has only explored to use eye gaze to filter information from the full 3D scene environment and then employed the scene features for deterministic human motion prediction [ZYM*22]. We are the first to directly leverage eye gaze information (without requiring information about the full 3D environment) to further boost the performance of stochastic human motion prediction.

2.3. Denoising Diffusion Models

Denoising diffusion models, or more precisely, denoising diffusion probabilistic models (DDPM) [HJA20, SME20, DN21, RBL*22] are a group of the most ingenious generative models. They aim to model reversing a Markov chain of the diffusion process illustrated in Figure 2. During training, noisy samples are obtained by incrementally adding noise to raw samples. The DDPM model then progressively reverses the diffusion process by predicting noise and denoising the samples. The loss is computed as the difference between the predicted noise and the Gaussian noise added during the diffusion process. In the inference stage, given a well-trained DDPM, it can generate realistic motions from a Gaussian distribution via the reversed diffusion process.

Owing to their power of generating realistic and high-quality samples, diffusion models have been applied in image/video generation [RBL*22, HSG*22], anomaly detection [WLSW22], objection detection [CSSL22], 3D reconstruction [XWC*22], time series forecasting [RSSV21] and imputation [TSSE21]. In a similar context, Tevet et al. [TRG*22] proposed a text-driven human motion synthesis method with diffusion models. Recently, Barquero et al. [BEP22] utilised a latent diffusion model to sample the diverse behaviour code to predict stochastic motions. Chen et al. [CZL*23] presented an end-to-end motion prediction framework based on diffusion models without complicated loss constraints and training processes. However, existing methods usually produce unrealistic predictions, probably because they only focus on past human poses and neglect important information from other modalities. To address this limitation, our method injects gaze-motion features into the denoising step through a cross-modal attention mechanism to generate more reasonable motion predictions.

3. GazeMoDiff Model

Our diffusion model for gaze-guided stochastic human motion prediction takes past body poses and eye gaze information as input. We represent human pose $p \in \mathbb{R}^{j\times 3}$ using the 3D positions of all human joints, where j denotes the number of human joints. Eye gaze $g \in \mathbb{R}^{1\times 3}$ is defined as a unit direction vector. Given H frames of observed sequence $\mathbf{x} = [(g_1, p_1), (g_2, p_2), ..., (g_H, p_H)]$, our goal is to predict human motions in the future F frames. Considering the stochastic nature of human motion, we generate k different future motion trajectories to provide multiple reasonable predictions $\mathbf{P} = \{\mathbf{p}_0, \mathbf{p}_1, ..., \mathbf{p}_k\}$, where $\mathbf{p}_i = [p_{H+1}, p_{H+2}, ..., p_{H+F}]$. An overview of our model is shown in Figure 3. Our model consists of three modules: a gaze-motion feature extraction module that uses a gaze encoder and a motion encoder to extract gaze and motion features respectively, a gaze-motion feature fusion module that fuses the gaze-motion features using a spatio-temporal graph attention network, as well as a diffusion-based motion generation module that employs a cross-attention mechanism to inject the gaze-motion features into a noise prediction network to generate multiple reasonable human future motions through progressive denoising.

3.1. Gaze-motion Feature Extraction

3.1.1. Gaze Encoder

We first pad raw observed gaze data $G_{raw} \in \mathbb{R}^{3\times 1\times H}$ to full-length gaze data $G \in \mathbb{R}^{3\times 1\times (H+F)}$ following prior work [CZL*23]. We then employed an approximate discrete cosine transform (DCT) [MLS21] that selected the first L components to process the full-length gaze data $G \in \mathbb{R}^{3\times 1\times (H+F)}$. By leveraging this transformation, the time dimension of eye gaze signals was reduced from $H+F$ to L, which both improved the smoothness of generation and reduced computational complexity [MLSL19]. We further employed a 1D convolutional neural network as an encoder to extract features from the transformed eye gaze sequence $G_{dct} \in \mathbb{R}^{3\times 1\times L}$. Specifically, we used four CNN layers with the kernel size of three. The channel size of the first three layers is 32 and each layer is followed by a layer normalisation (LN) and a Tanh activation function. After the three CNN layers, we employed an extra CNN layer with three channels and a Tanh activation function to obtain the final eye gaze feature $f_{gaze} \in \mathbb{R}^{3\times 1\times L}$.

3.1.2. Motion Encoder

Recent works have demonstrated that a graph structure is particularly effective for learning inherent correlations between human body joints [LCPW21, LCZ*21, MNL*22]. Inspired by these works, we treated human motions as a fully-connected spatio-temporal graph $P \in \mathbb{R}^{3\times j\times (H+F)}$, where later F frames of poses were padded with

© 2024 The Authors.
Proceedings published by Eurographics - The European Association for Computer Graphics.

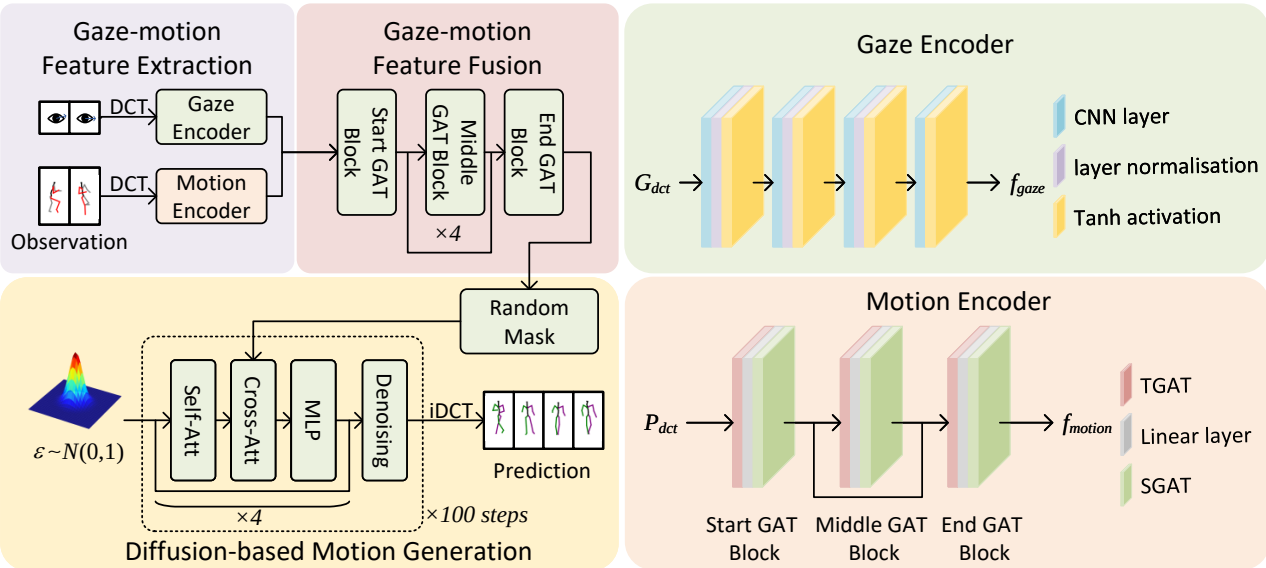

Figure 3: *Overview of the proposed method GazeMoDiff. GazeMoDiff first uses a gaze encoder and a motion encoder to extract the gaze and motion features respectively, then employs a spatio-temporal graph attention network to fuse these features, and finally injects the gaze-motion features into a noise prediction network via a cross-attention mechanism to generate multiple reasonable human future motions through a progressive denoising process.*

the last observed frame. Similarly, we employed DCT to obtain the transformed spatio-temporal graph $P_{dct} \in \mathbb{R}^{3\times j\times L}$ for smoother generation and lower computational complexity. We designed a GAT-based motion encoder to model the spatio-temporal correlations in the motion sequence for extracting motion features $f_{motion} \in \mathbb{R}^{3\times j\times L}$. The motion encoder consisted of three GAT blocks i.e. a start GAT block, a middle GAT block, and an end GAT block.

Start GAT Block. Given the transformed spatio-temporal graph $P_{dct} \in \mathbb{R}^{3\times j\times L}$, we first enhance the features using a multi-head self-attention mechanism via the temporal GAT layerand obtained $H' = [h'_1, h'_2, ..., h'_L] \in \mathbb{R}^{3\times j\times L}$ in the following way:

$$h'_i = \text{LeakyReLU}\left(\frac{1}{N_{head}}\sum_{n=1}^{N_{head}}\sum_{k=1}^{L}\alpha^n_{ik}h_k\right), \tag{1}$$

where $h'_i \in \mathbb{R}^{3\times j}$ is the output feature of node i, h_k is the input feature of node k, and $N_{head} = 8$ denotes the number of heads for attention. We fused different output features from each head by averaging them. For each head, the attention matrix α^n_{ik} represents interactions between each timestamp, calculated as follows:

$$\alpha^n_{ik} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}^n\left[h_i \oplus h_k\right]\right)\right)}{\sum_{l=1}^{T}\exp\left(\text{LeakyReLU}\left(\mathbf{a}^n\left[h_i \oplus h_l\right]\right)\right)}, \tag{2}$$

where $\mathbf{a}^n$ is a parameter vector $\in \mathbb{R}^{6\cdot j\times 1}$ and $\oplus$ denotes concatenation operation.

A linear layer was then used to transfer feature dimension from $H' \in \mathbb{R}^{3\times j\times L}$ to $\bar{H}' \in \mathbb{R}^{16\times j\times L}$.We further proposed a spatial graph attention network layer to extract features between different joints and obtain the output of the start GAT block $H_{sta} \in \mathbb{R}^{16\times j\times L}$. Specifically, the spatial GAT layer was similar to the temporal GAT layer , differing primarily in how attention coefficients were computed. Rather than extracting features in the temporal dimension as in the temporal GAT, the spatial GAT extracts features across the spatial dimension.

Middle GAT Block. We designed a middle GAT block to further extract the body pose features. Taking the output of the Start GAT block $H_{sta} \in \mathbb{R}^{16\times j\times L}$ as input, we first duplicated the temporal features for enhancement [MNL*22] $\left(\mathbb{R}^{16\times j\times L} \rightarrow \mathbb{R}^{16\times j\times 2L}\right)$. Then we used a temporal GAT layer, a linear layer, and a spatial GAT layer to further learn correlations between joints and extract features. Note that the input and output dimensions of the linear layer in the middle GAT block are both 16. Then a layer normalisation, a Tanh activation function, and a dropout layer with 0.3 dropout rate was used after the spatial GAT layer to avoid overfitting. A skip connection was added to improve the network flow. To keep the temporal dimension of the original input to this block the same, the middle GAT block's output f_{mid} was divided in half in the temporal dimension $\left(\mathbb{R}^{16\times j\times 2L} \rightarrow \mathbb{R}^{16\times j\times L}\right)$.

End GAT Block. The end GAT block was used to obtain the final motion features with the same feature dimension in gaze features f_{gaze} for further fusion. This block consists of a temporal GAT, a linear layer (reducing feature dimension from 16 to 3), a spatial GAT, a layer normalisation (LN), a Tanh activation function, and a dropout layer with 0.3 dropout rate. Given $f_{mid} \in \mathbb{R}^{16\times j\times L}$, the end GAT block outputs $f_{motion} \in \mathbb{R}^{3\times j\times L}$.

© 2024 The Authors.
Proceedings published by Eurographics - The European Association for Computer Graphics.

3.2. Gaze-motion Feature Fusion

To fuse gaze and motion features, we treated gaze and motion features as "virtual joints" and combined gaze and poses together as a fully connected graph $f_{in} \in \mathbb{R}^{3\times(j+1)\times L}$. Then a similar spatio-temporal GAT was employed to fuse gaze and motion features.

The gaze-motion fusion module was composed of a start GAT block, four middle GAT blocks, and an end GAT block. The structures of those blocks are the same as the blocks in the motion encoder. The gaze-motion features $f_{fus} \in \mathbb{R}^{3\times(j+1)\times L}$ were calculated as:

$$f_{fus} = GazePoseFuse\left(f_{in}\right). \tag{3}$$

3.3. Diffusion-based Motion Generation

We formulated the motion generation task as an inverse diffusion process, which contains iterative noise prediction and denoising steps.

Noise prediction. Inspired by text-driven noise prediction network [ZCP*22], we designed a noise prediction network ϵ_θ that employed the cross attention mechanism to guide the denoising process using the gaze-motion features. This network is developed to predict the noise at the current time step given the gaze-motion features f_{fus} and noisy motion sequence $Y^t \in \mathbb{R}^{3\times(j+1)\times L}$ from the previous time step. As illustrated in Figure 3, ϵ_θ consists of n stacked self-attention block, cross-attention block and multi-layer perception (MLP) block with skip connections.Additionally, to trade off diversity and realism of predictions, we train our model using classifier-free guidance [HS22]. Specifically, the model predicted noises from both conditional and unconditional prior by randomly setting $c = None$ for 10% possibility and $c = f_{fus}$ for 90% possibility. The predicted noise at time step t is formulated as follows:

$$\bar{\epsilon} = \epsilon_\theta\left(Y^t, c, t\right). \tag{4}$$

To better model the temporal correlations within the noisy gaze-motion sequence $Y^t \in \mathbb{R}^{3\times(j+1)\times L}$, we first employed a linear transformation to project the sequence into a higher-dimension latent space $Y'^t \in \mathbb{R}^{3\times(j+1)\times L'}$. We then applied an efficient self-attention block [ZCP*22] to further model temporal correlations between each frame, $\mathbf{Y} = \text{self-att}\left(Y'^t\right) + Y'^t$

We then applied a step hint module to inform about how many steps of noise have been added thus far. We first obtained the timestep embedding $\mathbf{e}_t$ using position embedding [VSP*17]. The gaze-motion historical features c were also fused with a learned linear projection of the timestep embedding, $\mathbf{e} = \mathbf{e}_t + \mathbf{W}'c$. This fused embedding $\mathbf{e}$ was then injected into the output of the self-attention block, $\mathbf{Y}' = \text{step-hint}\left(\mathbf{Y}, \mathbf{e}\right)$.

To incorporate the historical gaze-motion features f_{fus} and leverage their impact on noise prediction, we utilised cross-attention blocks. These blocks enable a deeper exploration of how the historical features influence the noise predictions at different denoising steps. In addition, computing the contributions of different attention heads in parallel better integrates information from both modalities (gaze and motion). Potentially, the noisy motions can also provide feedback to update c, leading to a collaborative learning process. In the cross-attention block, $\mathbf{Q}_c$ and $\mathbf{V}_c$ were calculated by gaze-motion historical features c while $\mathbf{K}_c$ was calculated using the output of self-attention blocks $\mathbf{Y}'$. The output of the cross-attention block was calculated as follows:

$$\mathbf{Y}_c = \text{Dropout}\left(\text{softmax}\left(\mathbf{Q}_c\right)\text{softmax}\left(\mathbf{K}_c^\top\right)\right)\text{LN}\left(\mathbf{V}_c\right) + \mathbf{Y}'. \tag{5}$$

$$\mathbf{Q}_c = \mathbf{W}'_q c, \mathbf{K}_c = \mathbf{W}'_k \mathbf{Y}', \mathbf{V}_c = \mathbf{W}'_v c. \tag{6}$$

Finally, we employed a two-layer MLP with GELU activation function and a dropout layer with 0.2 dropout rate to further extract features. A linear layer was added at the end to align with the noise dimension.

In our implementation, the noise prediction network contained four self-attention blocks, four cross-attention blocks, and four MLP blocks. Each attention operation used eight attention heads. The timestep embedding dimension and the latent space dimension were both set to 512.

Denoising. In a vanilla approach, the denoised sequence was generated directly from the Gaussian noise input. However, due to the accumulation of prediction noise errors, the observation information in the padded sequence is far from the truth in the latter steps. In each denoising step, the observed sequence was also available to guide the generation in the original space. Thus, we employed an ingenious prediction mask mechanism [CZL*23] to obtain denoised motions $\bar{\mathbf{p}} \in \mathbb{R}^{3\times j\times F}$ progressively. The detailed denoising process is illustrated in the supplementary material.

3.4. Training

In the training stage, the observation and prediction sequences are both available. Thus, we trained the model on the full-motion sequence $X_{full} \in \mathbb{R}^{3\times(j+1)\times(H+F)}$. First, we also transferred X_{full} into the DCT space $Y_{full} \in \mathbb{R}^{3\times(j+1)\times L}$. We then added noise to Y_{full} to generate a noisy sequence Y^t_{full}. Then we predicted the noise through Equation 4.

We then optimised all parameters in our pipeline by minimising the l_2 loss between the predicted noise $\epsilon_\theta\left(Y^t, c, t\right)$ and the true noise ϵ:

$$\mathscr{L} = \mathbb{E}_{\epsilon,t}\left[\left\|\epsilon - \epsilon_\theta\left(Y^t_{full}, c, t\right)\right\|^2\right]. \tag{7}$$

The detailed training and inference procedure is illustrated in the supplementary material.

4. Experiments

4.1. Datasets

Only a limited number of datasets contain synchronised recordings of both eye gaze and full-body human motion. We evaluated our method on two such public datasets, i.e. the MoGaze [KBM*20] dataset for real-world setting and the GIMO [ZYM*22] dataset for AR setting.

MoGaze. The MoGaze dataset provides motion capture and eye-tracking data recorded simultaneously from 6 participants performing *pick* and *place* actions in an indoor environment. It contains

© 2024 The Authors.
Proceedings published by Eurographics - The European Association for Computer Graphics.

over 3 hours of body movement and eye gaze recordings captured at 30 Hz. The pose of the human body is expressed using the 3D coordinates of 21 body joints while the eye gaze is represented as a direction vector. Following common settings in stochastic human motion prediction [CZL*23, BEP22, YK20], we used recordings from $p1$, $p2$, $p4$, $p5$, and $p6$ for training and employed the data from $p7$ for testing.

GIMO. The GIMO dataset contains body motion and eye gaze data captured from 11 participants in various indoor AR environments. The action categories include *sitting* or *laying on objects*, *touching*, *holding*, *reaching to objects*, *opening*, *pushing*, *transferring*, *throwing*, *swapping objects*, etc. We represented the body pose in the GIMO dataset using the 3D coordinates of 23 body joints and denoted the eye gaze using a direction vector. In the evaluation process, we followed the official dataset splits provided in GIMO [ZYM*22]. The training set comprises motion and eye gaze recordings from 12 scenes, while the test set contains data from 14 scenes, including 12 known environments and 2 new, unseen environments. This evaluation protocol allows the assessment of generalisation capabilities to new scenes.

4.2. Evaluation Metrics

To evaluate the performance of our model, we employed four commonly used metrics following prior works on stochastic HMP [YK20, SRM*23, CZL*23]:

- Average displacement error (ADE): ADE measures the average l_2 distance between the ground truth and the predicted motions over the whole future sequence. The lower the ADE, the more precise the prediction.
- Final displacement error (FDE): FDE calculates the l_2 distance between the ground truth and the predicted future motions at the final timestep. The lower the FDE, the more precise the prediction at the final timestep.
- Multi-modal average displacement error (MMADE): MMADE is designed to handle the multi-modal nature of the predictions in the task of stochastic HMP. It takes into account the fact that there can be multiple reasonable ground truth sequences for a given input. The multiple ground truth is obtained from the future motions of similar observed sequences (Sequences with l_2 distance below a given threshold are treated as similar sequences). MMADE is the mean value of the ADE calculated using the predictions and the multiple ground truth. The lower the MMADE, the better the ability to generate multiple reasonable predictions.
- Multi-modal final displacement error (MMFDE): MMFDE is calculated in a similar way as MMADE. The difference is that MMADE calculates the average l_2 distance over all future timesteps while MMFDE measures the l_2 distance only at the final timestep. The lower the MMFDE, the better the ability to generate multiple reasonable predictions at the final timestep.

4.3. Baselines

We compared our method with the following state-of-the-art methods in stochastic human motion prediction:

- *DLow* [YK20]: *DLow* is a latent flow-based model to generate multiple future motions via diversity-promoting sampling and loss.
- *CVAE* [YK20]: *CVAE* is a conditional variational autoencoder utilised in DLow [YK20] as a pre-trained generative model which can also forecast stochastic future motions.
- *BeLFusion* [BEP22]: *BeLFusion* is a latent diffusion-based model to predict reasonable motions based on disentangling the behavioural representation from past motions.
- *HumanMAC* [CZL*23]: *HumanMac* is a diffusion-based stochastic HMP model with a DCT completion fashion in the inference.

4.4. Implementation Details

We set the observation and prediction time windows to 0.5 seconds (15 frames) and 2 seconds (60 frames) respectively for both the MoGaze and GIMO datasets, following common practice in stochastic human motion prediction [CZL*23, BEP22, YK20]. For a fair comparison, we retrained all the baseline methods from scratch using their default training parameters. We trained all the modules in our method in an end-to-end manner for 300 epochs using Adam optimiser [KB14] with an initial learning rate of 0.0003 and a batch size of 32. The learning rate was decayed at 75, 150, 225, 275 epochs, respectively, with a multiplicative factor of 0.9. We applied a standard diffusion process as proposed by [SME20] that degraded 1500 steps in the training and sampled 100 steps in the inference. We selected the Cosine noise scheduler [ND21] in the diffusion following HumanMAC [CZL*23]. The multi-modal ground truth threshold was set to 0.4 following previous work [BEP22]. All experiments were conducted in an Nvidia TITAN X GPU with 12GB memory using the PyTorch 1.7.1 framework.

4.5. Quantitative Results

The quantitative results on the MoGaze [KBM*20] and GIMO [ZYM*22] datasets are shown in Table 1. Overall, our model outperforms the state-of-the-art methods on all metrics.

Results on MoGaze. As shown in Table 1, our model achieves the lowest ADE, FDE, MMADE and MMFDE compared to prior methods, substantially surpassing the best state-of-the-art HumanMAC [CZL*23]. For average displacement error, our method achieves an improvement of 12.8% (0.638 *vs*. 0.732) over HumanMAC [CZL*23]. On final displacement error, our method obtains a 13.7% performance gain (0.939 *vs*. 1.089). On multi-modal accuracy metrics, our method improves the MMADE performance by 16.7% (0.649 *vs*. 0.779) and the MMFDE performance by 17.3% (0.946 *vs*. 1.144). The greater gains on the multi-modal metrics reveal that our model is more capable of generating multiple reasonable predictions aligned with the stochasticity of human nature. We further conducted a Wilcoxon signed-rank test to compare our method and the state-of-the-art and the results demonstrate that the differences between our method and the state-of-the-art method are statistically significant ($p < 0.01$). In addition, we also presented the performance of our method without using eye gaze in Table 1. Compared to our method, the ablated version of not using eye gaze presents lower accuracy in terms of ADE, FDE, MMADE and MMFDE, demonstrating the effectiveness of eye gaze information for stochastic human motion prediction. It is also notable that our method without using eye gaze

© 2024 The Authors.
Proceedings published by Eurographics - The European Association for Computer Graphics.

Table 1: *Comparison of our method with the state-of-the-art methods on both MoGaze [KBM*20] and GIMO [ZYM*22] (unit: meters). The best results are in bold while the second best are underlined.*

	Results on MoGaze [KBM*20]				Results on GIMO [ZYM*22]			
	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓
CVAE [YK20]	1.070	1.644	1.091	1.667	1.292	2.059	1.294	2.054
DLow [YK20]	0.807	1.261	0.840	1.274	1.084	1.688	1.090	1.687
BeLFusion [BEP22]	0.899	1.306	0.908	1.313	0.840	1.220	0.845	1.221
HumanMAC [CZL*23]	0.732	1.089	0.779	1.144	0.815	1.121	0.821	1.118
$Ours_{w/o\ gaze}$	0.664	0.982	0.696	0.996	0.812	1.120	0.818	1.116
$Ours_{head}$	0.669	0.971	0.678	0.974	0.734	0.979	0.737	0.975
Ours	**0.638**	**0.939**	**0.649**	**0.946**	**0.729**	**0.974**	**0.732**	**0.969**

still outperforms the state-of-the-art methods, achieving an improvement of 9.8% on MMADE and 12.9% on MMFDE. These results underscore the superiority of our model architecture. Considering that eye gaze is not always available, we further present a novel variant that replaces gaze with head direction. We can see from Table 1 that although this variant is not as good as our full model, it outperforms the variant without gaze. These results demonstrate that head direction can be a reliable alternative when gaze is not available.

Results on GIMO. As shown in Table 1, our method outperforms the state-of-the-art methods in all the metrics. Specifically, compared to the state-of-the-art method HumanMAC, our method achieves an improvement of 10.5% (0.815 to 0.729) in ADE as well as an improvement of 13.1% (1.121 to 0.974) in FDE. On multi-modal accuracy metrics, our method obtains an improvement of 10.8% in MMADE and 13.3% in MMFDE, respectively. The results from a Wilcoxon signed-rank test validate that the differences between our method and the state-of-the-art method are statistically significant ($p < 0.01$). We also find that our method significantly outperforms the ablated version of not using eye gaze, revealing the importance of eye gaze for generating stochastic human motions. The results of $Ours_{head}$ on GIMO also validate that eye gaze performs better than head orientation in terms of human motion prediction and revealed that head orientation can be used as a proxy to gaze when gaze is not available.

4.6. Visualisation Analysis

We visualised the multiple predicted poses at the time point of the future one second. We compared our method with the state-of-the-art method HumanMAC [CZL*23]. For each method, we randomly generated 10 predictions for comparison.

Visualisation results on the MoGaze dataset. We illustrated a representative visualisation from the MoGaze dataset [ZYM*22] in Figure 4. The observed motion sequence involves a turn to the right, and the ground truth pose shows that at the future one second, this person continues the turn by about 100 degrees further to the right. We can observe that predictions of HumanMAC [CZL*23] generally continue walking in the same forward direction as the observed last frame, failing to anticipate the full turning trajectory. The best prediction generated by HumanMAC [CZL*23] (labelled in green) also differs from the ground truth. In contrast, our method can recognise that the person intends to keep turning right. As a result, the best prediction from our method is precisely facing the true direction of the ground truth and other predictions generally align with the ground truth. In addition, we can see that the poses generated by our method are physically plausible without any angle distortions or strange limb lengths.

Visualisation results on the GIMO dataset. We also illustrated a typical visualisation from the GIMO dataset [ZYM*22]. As depicted in Figure 1, predictions from our method are more similar to the ground truth. Furthermore, all predictions from our method are generally realistic. In contrast, HumanMAC yielded predictions that include some implausible cases (boxed in red). These results demonstrate that our method can generate more reasonable motion predictions than the state-of-the-art.

Visualisation results of not using eye gaze. To further evaluate the effectiveness of eye gaze for stochastic human motion prediction, we showed a visualisation example to compare our method with the ablated version of not using eye gaze. As shown in Figure 5, the observed motions showed a person turning to the right and then standing still. Our method was able to predict that the person would continue going right in the future while the ablated version of not using eye gaze failed to recognise this intention and stayed in place. These results demonstrate that eye gaze signals provide information on human intention and can help improve the performance of stochastic human motion prediction methods.

More visualisation results are provided in the supplementary material.

4.7. Adding Eye Gaze to Prior Methods

To further evaluate our method's effectiveness for extracting and fusing gaze and motion features, we compared our methods with prior methods' variants that add eye gaze as input. The results on MoGaze and GIMO are shown in Table 2, where Hum_{gaze} and BeL_{gaze} denotes the variants of HumanMAC and BeLFusion that add eye gaze as an extra input joint. Overall, our approach outperforms both BeLFusion and HumanMAC which incorporate eye gaze information in terms of all metrics, demonstrating our method's superiority in exploiting eye gaze information. Additionally, compared to the performance of original HumanMAC and BeLFusion shown in Table 1, all the metrics improve on the MoGaze dataset while getting worse slightly on the GIMO dataset. It demonstrates that simply adding eye gaze as a new joint to an existing method does not always have a positive effect. Instead, the internal features of the eye gaze and its synergistic correlations with other human joint nodes need to be considered. Our approach outperforms the state-of-the-art methods significantly by introducing novel feature

© 2024 The Authors.
Proceedings published by Eurographics - The European Association for Computer Graphics.

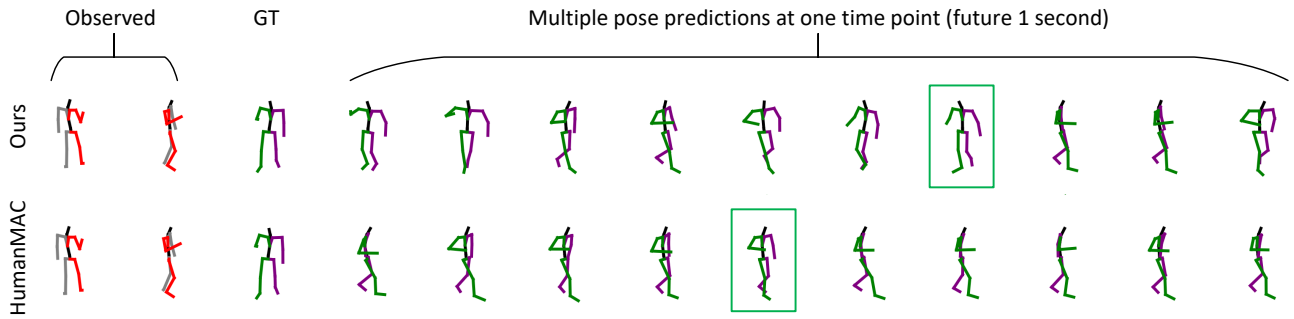

Figure 4: *Ground truth (GT) human pose at future one second and multiple pose predictions generated by different methods on the MoGaze dataset [KBM*20] with the best prediction (lowest l_2 distance to GT) boxed in green. Our method can generate motions that are closer to the ground truth than the state-of-the-art method HumanMAC [CZL*23].*

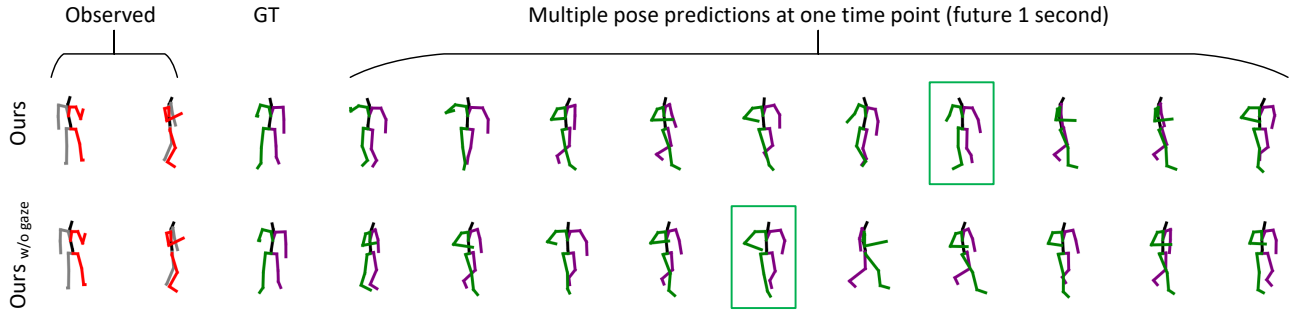

Figure 5: *Ground truth (GT) human pose at future one second and multiple pose predictions generated by different methods on the MoGaze dataset [KBM*20] with the best prediction (lowest l_2 distance to GT) boxed in green. Our method using eye gaze can generate motions that are closer to the ground truth than the ablated version of not using eye gaze.*

extraction and fusion modules for gazes and motions and using the fused features to guide future motion generation.

Table 2: *Comparison of our method with baselines that add eye gaze as input on both MoGaze [KBM*20] and GIMO [ZYM*22] (unit: meters). The best results are in bold while the second best are underlined.*

Dataset	Method	ADE	FDE	MMADE	MMFDE
MoGaze	Hum_{gaze}	0.654	0.965	0.679	0.980
	BeL_{gaze}	0.654	1.018	0.659	1.012
	Ours	**0.638**	**0.939**	**0.649**	**0.946**
GIMO	Hum_{gaze}	0.850	1.135	0.852	1.137
	BeL_{gaze}	0.858	1.126	0.861	1.221
	Ours	**0.729**	**0.974**	**0.732**	**0.969**

4.8. Ablation Study

We conducted an ablation study to comprehensively analyse the individual contributions of various modules within our framework. We thoroughly assessed the performance of our model by comparing it to variants that exclude specific modules:

- $Ours_{w/o\ gat}$: Without spatio-temporal graph attention network.
- $Ours_{w/o\ CA}$: Without cross attention in the noise predicting network.
- $Ours_{w/o\ GE}$: Without gaze encoder.
- $Ours_{w/o\ ME}$: Without motion encoder.

As we can see from Table 3, our model outperforms all the ablated versions. It reveals that each module within our framework significantly contributes to the generation of stochastic human future motions.

Table 3: *Ablation study on MoGaze [KBM*20]. The best results are in bold.*

	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓
$Ours_{w/o\ gat}$	0.655	0.990	0.663	0.993
$Ours_{w/o\ CA}$	0.697	1.001	0.709	1.000
$Ours_{w/o\ GE}$	0.655	0.946	0.665	**0.946**
$Ours_{w/o\ ME}$	0.642	0.976	0.654	0.985
Ours	**0.638**	**0.939**	**0.649**	**0.946**

© 2024 The Authors.
Proceedings published by Eurographics - The European Association for Computer Graphics.

5. User Study

To perform a comprehensive evaluation of the perceived precision and realism of the generated samples, we conducted a user study to assess model performance based on human perception and intuition. Participants were tasked with random samples predicted by different motion forecasting models. By collecting subjective rankings directly from individuals, the user study provided an intuitive approach to evaluate the aspects of the predictions that are challenging to assess objectively, such as naturalness, continuity, and the overall plausibility of the motions.

5.1. Assessment Details

We randomly selected 4 sequences from the MoGaze dataset and 12 sequences from the GIMO dataset. The difference in the number of selections between the two datasets can be attributed to the difference in the variety of actions included. While the MoGaze dataset consists of only two actions, the GIMO dataset contains a larger number of different actions. We compared our method with the best baseline HumanMAC. The compared conditions were randomly provided to the users in order to avoid any potential bias. Each method generated 10 predictions from random noise. We set the sample size to 10 for a trade-off between the predictions' diversity and users' burden. The participants in our study were then instructed to rank these predictions based on the following two key aspects:

- **Realism**: If these poses are physically plausible. You can check if there are any angle distortions, too short/long limbs, implausible poses, or any sudden or unreasonable changes during the whole motion.
- **Precision**: If these motions align with the ground truth. You can measure the similarity between each motion and the ground truth.

The questionnaire utilised in our study was created using the *Jotform* platform The interface of our user study can be found in the supplementary material. A total of 21 individuals (11 males and 10 females, aged between 21 and 39 years, Mean= 26.3, SD=4.4) were recruited to take part in our user study via university mailing lists and social networks. This user study is approved by the university.

5.2. Evaluation

As shown in Table 4, our method outperforms HumanMAC in terms of both precision and realism. 74.6% of the total responses considered predictions of our method as more precise while only 25.4% regarded HumanMAC's results as more precise. From the perspective of realism, 67.7% of the total responses considered the motions generated by our method as more realistic than HumanMAC while only 33.3% of the responses considered HumanMAC were able to generate more realistic predictions. To further validate our findings, we conducted a paired Wilcoxon signed-rank test, which revealed a statistically significant difference ($p < 0.01$) between our method and HumanMAC in terms of both precision and realism.

Table 4: *Results of the user study. Participants were required to measure the predictions from different methods according to precision and realism.*

	More precise	More realistic
HumanMAC	25.4%	32.3%
Ours	74.6%	67.7%

6. Discussion

Our study marks a notable advancement in the emerging field of incorporating eye gaze cues for stochastic human motion prediction in VR/AR research. We have successfully demonstrated the effectiveness of our proposed method in generating multiple reasonable human motion predictions.

Performance of our method. Our method outperforms the state-of-the-art methods in terms of all metrics on both the MoGaze and GIMO datasets (See Table 1). In addition, the visualisation results on both datasets showed that our method can generate multiple reasonable predictions with high precision while the state-of-the-art may produce erroneous or implausible predictions (Figure 1 and Figure 4). The results from a user study further confirmed that the predictions generated by our method were more realistic and more precise than that from prior method (Table 4). These results demonstrate that our work have made a significant improvement in stochastic human motion prediction in terms of improving both precision and realism.

The effectiveness of eye gaze for stochastic HMP. To the best of our knowledge, our method is the first to introduce eye gaze in the stochastic HMP task to obtain a significant performance improvement. Compared to the full model, the ablated version of not using eye gaze performed significantly worse on both datasets (Table 1). It demonstrated that introducing the historical eye gaze as input can benefit the generation of more precise future motions. The visualisation example further demonstrated the effectiveness of incorporating eye gaze. As shown in Figure 5, the observed motion showed a person turned to the right and then standing. The ground truth pose at future one second showed the person turning right again from a standstill. Our full model with eye gaze can predict this intention of turning right while the ablated version of not using eye gaze failed to recognise it and generated predictions that were all staying in place. We argue that the eye gaze implies the intention of future body movements. Therefore, our method can recognise more complex and precise intentions and make reasonable and precise predictions.

The approach to introducing gaze into stochastic HMP. Our method showed a significant improvement in the stochastic HMP task by adding eye gaze as input. A natural question is whether adding eye gaze as additional input to an existing method has an effect. In subsection 4.7 we explored existing diffusion-based methods with eye gaze as an additional joint input. The accuracy declined slightly on the GIMO dataset (Table 2). Compared to the MoGaze dataset, the GIMO dataset contains more complex actions and motions, and the correlation between eye gaze and future action intent is much harder to mine. Therefore, simply adding eye gaze as an additional joint makes it difficult to model this complex correlation. Our method extracted the gaze and motion features via two encoders

© 2024 The Authors.
Proceedings published by Eurographics - The European Association for Computer Graphics.

and fused them via GAT. Experimental results demonstrated the effectiveness of our gaze encoder, motion encoder, as well as the GAT (Table 3).

Diversity reduction due to eye gaze guidance. As shown in Table 1, our model reported obvious improvements in terms of all accuracy metrics. Intuitively, the diversity of predictions will decrease. To further explore whether and to what extent the diversity of our predictions were reduced, we used the average pairwise distance (APD), i.e. the average L2 distance between every pair of motion predictions, to measure the diversity. We observed that when ablating eye gaze from our model, the APD value increases from 15.572 to 15.799 on MoGaze and from 15.519 to 20.199 on GIMO. These results validate that while adding eye gaze can improve accuracy, it inevitably decreases the diversity of the motion predictions.

Limitations. First, only two public datasets contain both the eye gaze and body motions, limiting our evaluation's generalisability. In addition, existing datasets contain limited action categories, which make predictions that tend to be limited to a few number of actions and can not generalise to unseen actions. . Finally, our approach uses observed motions and eye gaze as input. However, the input data may be missing due to tracking errors in practical applications, which may limit the application of our method in real-world scenarios.

Future work. Our proposed approach, GazeMoDiff, has the potential to be extended to incorporate additional modalities beyond eye gaze. It serves as a multi-modal human motion prediction framework, where various guidance information can be utilised to enhance the generation of reasonable motions. By incorporating modalities such as facial expressions, hand gestures, and physiological signals, the GazeMoDiff framework can be expanded to capture a more comprehensive range of cues for motion prediction. In addition, we are also looking forward to integrating our method with real VR/AR applications. Finally, as a diffusion-based prediction model, our method needs 100 DDIM sampling steps to denoise and obtain predictions, which limits to implementation into real-time scenarios. We will try some fast sampling methods [LZB*22a, LZB*22b] to improve generation speed.

7. Conclusion

In this work, we were the first to explore the effectiveness of eye gaze on generating multiple reasonable human motions in the future. We proposed a novel gaze-guided diffusion model that fuses the gaze and motion features using a spatio-temporal graph attention network and then injects these features into a noise prediction network via a cross-attention mechanism to generate multiple reasonable human future motions. Extensive experiments demonstrated that our method outperforms the state-of-the-art methods by a large margin and a user study validated that our method can generate motions that are more precise and more realistic than prior methods. As such, our work makes an important step towards generating more realistic human motions for virtual agents and guides future work on cross-modal human behaviour generation.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2075 – 390740016.

References

[AGBS16] AZMANDIAN M., GRECHKIN T., BOLAS M., SUMA E.: Automated path prediction for redirected walking using navigation meshes. In *2016 IEEE Symposium on 3D User Interfaces (3DUI)* (2016), IEEE, pp. 63–66. 2

[AKCH21] AKSAN E., KAUFMANN M., CAO P., HILLIGES O.: A spatio-temporal transformer for 3d human motion prediction. In *2021 International Conference on 3D Vision (3DV)* (2021), IEEE, pp. 565–574. 2

[ASP*21] ALIAKBARIAN S., SALEH F., PETERSSON L., GOULD S., SALZMANN M.: Contextually plausible and diverse 3d human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 11333–11342. 2

[BBKK17] BUTEPAGE J., BLACK M. J., KRAGIC D., KJELLSTROM H.: Deep representation learning for human motion prediction and classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 6158–6166. 2

[BEP22] BARQUERO G., ESCALERA S., PALMERO C.: Belfusion: Latent diffusion for behavior-driven human motion prediction. *arXiv preprint arXiv:2211.14304* (2022). 2, 3, 6, 7

[BKL18] BARSOUM E., KENDER J., LIU Z.: Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (2018), pp. 1418–1427. 2

[BRB*21] BHATTACHARYA U., REWKOWSKI N., BANERJEE A., GUHAN P., BERA A., MANOCHA D.: Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE virtual reality and 3D user interfaces (VR)* (2021), IEEE, pp. 1–10. 2

[CGM*20] CAO Z., GAO H., MANGALAM K., CAI Q.-Z., VO M., MALIK J.: Long-term human motion prediction with scene context. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16* (2020), Springer, pp. 387–404. 2

[CSSL22] CHEN S., SUN P., SONG Y., LUO P.: Diffusiondet: Diffusion model for object detection. *arXiv preprint arXiv:2211.09788* (2022). 3

[CWL*22] CHEN H., WEI L., LIU H., SHI B., ZHANG G., ZHA H.: Mount: Learning 6dof motion prediction based on uncertainty estimation for delayed ar rendering. *IEEE Transactions on Visualization and Computer Graphics* (2022), 1–12. doi:10.1109/TVCG.2022.3228807. 2

[CZL*23] CHEN L.-H., ZHANG J., LI Y., PANG Y., XIA X., LIU T.: Humanmac: Masked motion completion for human motion prediction. *arXiv preprint arXiv:2302.03665* (2023). 2, 3, 5, 6, 7, 8

[DJPZ*21] DAVID-JOHN B., PEACOCK C., ZHANG T., MURDISON T. S., BENKO H., JONKER T. R.: Towards gaze-based prediction of the intent to interact in virtual reality. In *ACM Symposium on Eye Tracking Research and Applications* (2021), pp. 1–7. 2

[DKP*23] DU Y., KIPS R., PUMAROLA A., STARKE S., THABET A., SANAKOYEU A.: Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 481–490. 2

[DN21] DHARIWAL P., NICHOL A.: Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems 34* (2021), 8780–8794. 3

© 2024 The Authors.
Proceedings published by Eurographics - The European Association for Computer Graphics.

[EZW*21] EMERY K. J., ZANNOLI M., WARREN J., XIAO L., TALATHI S. S.: Openneeds: A dataset of gaze, head, hand, and scene signals during exploration in open-ended vr environments. In *ACM Symposium on Eye Tracking Research and Applications* (2021), pp. 1–7. 3

[FLFM15] FRAGKIADAKI K., LEVINE S., FELSEN P., MALIK J.: Recurrent network models for human dynamics. In *Proceedings of the IEEE international conference on computer vision* (2015), pp. 4346–4354. 2

[Fre08] FREEDMAN E. G.: Coordination of the eyes and head during visual orienting. *Experimental brain research 190* (2008), 369–387. 2

[GIWW21] GAMAGE N. M., ISHTAWEERA D., WEIGEL M., WITHANA A.: So predictable! continuous 3d hand trajectory prediction in virtual reality. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (2021), pp. 332–343. 2

[GMK*19] GOPALAKRISHNAN A., MALI A., KIFER D., GILES L., ORORBIA A. G.: A neural temporal model for human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 12116–12125. 2

[GSAH17] GHOSH P., SONG J., AKSAN E., HILLIGES O.: Learning human motion models for long-term predictions. In *2017 International Conference on 3D Vision (3DV)* (2017), IEEE, pp. 458–466. 2

[HBLW21a] HU Z., BULLING A., LI S., WANG G.: Ehtask: Recognizing user tasks from eye and head movements in immersive virtual reality. *IEEE Transactions on Visualization and Computer Graphics* (2021). 2

[HBLW21b] HU Z., BULLING A., LI S., WANG G.: Fixationnet: Forecasting eye fixations in task-oriented virtual environments. *IEEE Transactions on Visualization and Computer Graphics 27*, 5 (2021), 2681–2690. 2

[HJA20] HO J., JAIN A., ABBEEL P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems 33* (2020), 6840–6851. 2, 3

[HLZ*20] HU Z., LI S., ZHANG C., YI K., WANG G., MANOCHA D.: Dgaze: Cnn-based gaze prediction in dynamic scenes. *IEEE Transactions on Visualization and Computer Graphics 26*, 5 (2020), 1902–1911. 2

[Hol12] HOLM J. E.: *Collision prediction and prevention in a simultaneous multi-user immersive virtual environment*. PhD thesis, Miami University, 2012. 2

[HS22] HO J., SALIMANS T.: Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022). 5

[HSG*22] HO J., SALIMANS T., GRITSENKO A., CHAN W., NOROUZI M., FLEET D. J.: Video diffusion models. *arXiv preprint arXiv:2204.03458* (2022). 3

[HXSB24] HU Z., XU J., SCHMITT S., BULLING A.: Pose2gaze: Eye-body coordination during daily activities for gaze prediction from full-body poses. *IEEE Transactions on Visualization and Computer Graphics* (2024). 3

[HYH*24] HU Z., YIN Z., HAEUFLE D., SCHMITT S., BULLING A.: Hoimotion: Forecasting human motion during human-object interactions using egocentric 3d object bounding boxes. *IEEE Transactions on Visualization and Computer Graphics* (2024). 2

[HZBD19] HOU X., ZHANG J., BUDAGAVI M., DEY S.: Head and body motion prediction to enable mobile vr experiences with low latency. In *2019 IEEE Global Communications Conference (GLOBECOM)* (2019), IEEE, pp. 1–7. 2

[HZL*19] HU Z., ZHANG C., LI S., WANG G., MANOCHA D.: Sgaze: A data-driven eye-head coordination model for realtime gaze prediction. *IEEE transactions on visualization and computer graphics 25*, 5 (2019), 2002–2010. 2, 3

[JZJ*20] JAIN D. K., ZAREAPOOR M., JAIN R., KATHURIA A., BACHHETY S.: Gan-poser: an improvised bidirectional gan model for human motion prediction. *Neural Computing and Applications 32*, 18 (2020), 14579–14591. 2

[KALL17] KIM Y., AN J., LEE M., LEE Y.: An activity-embedding approach for next-activity prediction in a multi-user smart space. In *2017 IEEE International Conference on Smart Computing (SMARTCOMP)* (2017), IEEE, pp. 1–6. 2

[KB14] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014). 6

[KBM*20] KRATZER P., BIHLMAIER S., MIDLAGAJNI N. B., PRAKASH R., TOUSSAINT M., MAINPRICE J.: Mogaze: A dataset of full-body motions that includes workspace geometry and eye-gaze. *IEEE Robotics and Automation Letters 6*, 2 (2020), 367–373. 2, 5, 6, 7, 8

[KGB19] KUNDU J. N., GOR M., BABU R. V.: Bihmp-gan: Bidirectional 3d human motion prediction gan. In *Proceedings of the AAAI conference on artificial intelligence* (2019), vol. 33, pp. 8553–8560. 2

[KN18] KOOCHAKI F., NAJAFIZADEH L.: Predicting intention through eye gaze patterns. In *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)* (2018), IEEE, pp. 1–4. 2

[KYK*20] KOTHARI R., YANG Z., KANAN C., BAILEY R., PELZ J. B., DIAZ G. J.: Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities. *Scientific reports 10*, 1 (2020), 2539. 3

[LCPW21] LI Q., CHALVATZAKI G., PETERS J., WANG Y.: Directed acyclic graph neural network for human motion prediction. In *2021 IEEE International Conference on Robotics and Automation (ICRA)* (2021), IEEE, pp. 3197–3204. 2, 3

[LCZ*21] LI M., CHEN S., ZHAO Y., ZHANG Y., WANG Y., TIAN Q.: Multiscale spatio-temporal graph neural networks for 3d skeleton-based motion prediction. *IEEE Transactions on Image Processing 30* (2021), 7760–7775. 2, 3

[LZB*22a] LU C., ZHOU Y., BAO F., CHEN J., LI C., ZHU J.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems 35* (2022), 5775–5787. 10

[LZB*22b] LU C., ZHOU Y., BAO F., CHEN J., LI C., ZHU J.: Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095* (2022). 10

[LZX*17] LI Z., ZHOU Y., XIAO S., HE C., HUANG Z., LI H.: Auto-conditioned recurrent networks for extended complex human motion synthesis. *arXiv preprint arXiv:1707.05363* (2017). 2

[MBR17] MARTINEZ J., BLACK M. J., ROMERO J.: On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 2891–2900. 2

[MGVO21] MARTÍNEZ-GONZÁLEZ A., VILLAMIZAR M., ODOBEZ J.-M.: Pose transformers (potr): Human motion prediction with non-autoregressive transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 2276–2284. 2

[MLS21] MAO W., LIU M., SALZMANN M.: Generating smooth pose sequences for diverse human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 13309–13318. 2, 3

[MLSL19] MAO W., LIU M., SALZMANN M., LI H.: Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 9489–9497. 3

[MNL*22] MA T., NIE Y., LONG C., ZHANG Q., LI G.: Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 6437–6446. 3, 4

[ND21] NICHOL A. Q., DHARIWAL P.: Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning* (2021), PMLR, pp. 8162–8171. 6

[NDR*21] NICHOL A., 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. *arXiv preprint arXiv:2112.10741* (2021). 2

© 2024 The Authors.
Proceedings published by Eurographics - The European Association for Computer Graphics.

[RBL*22] ROMBACH R., BLATTMANN A., LORENZ D., ESSER P., OMMER B.: High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 10684–10695. 3

[RDN*22] RAMESH A., DHARIWAL P., NICHOL A., CHU C., CHEN M.: Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125 1*, 2 (2022), 3. 2

[RSSV21] RASUL K., SEWARD C., SCHUSTER I., VOLLGRAF R.: Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International Conference on Machine Learning* (2021), PMLR, pp. 8857–8868. 3

[SG19a] SIDENMARK L., GELLERSEN H.: Eye, head and torso coordination during gaze shifts in virtual reality. *ACM Transactions on Computer-Human Interaction (TOCHI) 27*, 1 (2019), 1–40. 2, 3

[SG19b] SIDENMARK L., GELLERSEN H.: Eye&head: Synergetic eye and head movement for gaze pointing and selection. In *Proceedings of the 2019 ACM Symposium on User Interface Software and Technology* (2019), pp. 1161–1174. 2

[SME20] SONG J., MENG C., ERMON S.: Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020). 3, 6

[SMG*20] SIDENMARK L., MARDANBEGI D., GOMEZ A. R., CLARKE C., GELLERSEN H.: Bimodalgaze: Seamlessly refined pointing with gaze and filtered gestural head movement. In *ACM Symposium on Eye Tracking Research and Applications* (2020), pp. 1–9. 3

[SPW*18] SUN Q., PATNEY A., WEI L.-Y., SHAPIRA O., LU J., ASENTE P., ZHU S., MCGUIRE M., LUEBKE D., KAUFMAN A.: Towards virtual reality infinite walking: dynamic saccadic redirection. *ACM Transactions on Graphics (TOG) 37*, 4 (2018), 1–13. 2

[SRM*23] 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 *2023 IEEE International Conference on Robotics and Automation (ICRA)* (2023), IEEE, pp. 8246–8253. 6

[TRG*22] TEVET G., RAAB S., GORDON B., SHAFIR Y., COHEN-OR D., BERMANO A. H.: Human motion diffusion model. *arXiv preprint arXiv:2209.14916* (2022). 3

[TSSE21] TASHIRO Y., SONG J., SONG Y., ERMON S.: Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in Neural Information Processing Systems 34* (2021), 24804–24816. 3

[VSP*17] VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł., POLOSUKHIN I.: Attention is all you need. *Advances in neural information processing systems 30* (2017). 5

[WLSW22] WYATT J., LEACH A., SCHMON S. M., WILLCOCKS C. G.: Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 650–656. 3

[XWC*22] XU J., WANG X., CHENG W., CAO Y.-P., SHAN Y., QIE X., GAO S.: Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. *arXiv preprint arXiv:2212.14704* (2022). 3

[YK20] YUAN Y., KITANI K.: 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* (2020), Springer, pp. 346–364. 2, 6, 7

[ZA23] ZHANG L., AGRAWALA M.: Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543* (2023). 2

[ZCP*22] ZHANG M., CAI Z., PAN L., HONG F., GUO X., YANG L., LIU Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001* (2022). 5

[ZS82] ZANGEMEISTER W. H., STARK L.: Gaze latency: variable interactions of head and eye latency. *Experimental Neurology 75*, 2 (1982), 389–406. 3

[ZYM*22] ZHENG Y., YANG Y., MO K., LI J., YU T., LIU Y., LIU C. K., GUIBAS L. J.: Gimo: Gaze-informed human motion prediction in context. In *European Conference on Computer Vision* (2022), Springer, pp. 676–694. 1, 2, 3, 5, 6, 7, 8

© 2024 The Authors.
Proceedings published by Eurographics - The European Association for Computer Graphics.