1. Supplementary Material

1.1. Discrete cosine transformation

First, the full gaze-motion sequence $X \in \mathbb{R}^{3\times(j+1)\times(T+F)}$ is encoded to transformed sequence $Y \in \mathbb{R}^{3\times(j+1)\times L}$ temporally by discrete cosine transform (DCT) [ANR74] as follows:

$$Y = \mathrm{DCT}(X) = DX \tag{1}$$

where $D \in \mathbb{R}^{(T+F)\times(T+F)}$ is the DCT matrix. The reason we applied DCT first is that extensive work [MLSL19, CZL*23] has demonstrated such a transformer can extract both current and periodic temporal features, which leads to smoother predictions and a more compact representation.

Due to the orthogonality of DCT, the original gaze-motion sequence can be recovered from the transformed sequence:

$$X = \mathrm{iDCT}(X) = D^{\top}Y \tag{2}$$

To reduce the computational complexity and weaken the effects of high-frequency noise, we select the first L row of D, $D_L \in \mathbb{R}^{L\times(T+F)}$ to replace D. In a subsequent paper, we will replace D with D_L for simplicity.

1.2. Details spatial graph attention network

Given features $H' = [h'_1, h'_2, ..., h'_L] \in \mathbb{R}^{f\times j\times L}$ extracted by the temporal GAT, the spatial GAT aggregated features across spatial dimensions as follows:

$$h''_i = \mathrm{LeakyReLU}\left(\frac{1}{\bar{N}_{head}}\sum_{n=1}^{\bar{N}_{head}}\sum_{k=1}^{j}\bar{\alpha}^n_{ik}\bar{h}'_k\right), \tag{3}$$

where $h''_i \in \mathbb{R}^{f'\times L}$ is the output feature of node i and $\bar{N}_{head}$ denotes the number of heads for attention. We also applied an average function to fuse different output features from each head in spatial GAT. For each head, the attention matrix $\bar{\alpha}^n_{ik}$ represents interactions between each timestamp, calculated as follows:

$$\bar{\alpha}^n_{ik} = \frac{\exp\left(\mathrm{LeakyReLU}\left(\bar{\mathbf{a}}^n\left[\bar{h}'_i \oplus \bar{h}'_k\right]\right)\right)}{\sum_{l=1}^{j}\exp\left(\mathrm{LeakyReLU}\left(\bar{\mathbf{a}}^n\left[\bar{h}'_i \oplus \bar{h}'_l\right]\right)\right)}, \tag{4}$$

where $\bar{\mathbf{a}}^n$ is a parameter vector $\in \mathbb{R}^{2f'L\times 1}$

1.3. Details of self-attention block

Given noised sequence Y'^t at timestep t, We then apply an efficient self-attention block [ZCP*22] to further model temporal correlations between each frame:

$$\mathbf{Y} = \mathrm{Dropout}\left(\mathrm{softmax}(\mathbf{Q})\,\mathrm{softmax}\left(\mathbf{K}^{\top}\right)\right)\mathrm{LN}(\mathbf{V}) + Y'^t, \tag{5}$$

where LN is layer normalisation, $\mathbf{Q} \in \mathbb{R}^{L\times d}$, $\mathbf{K} \in \mathbb{R}^{L\times d}$, and $\mathbf{V} \in \mathbb{R}^{L\times d}$ are calculated using the original self-attention mechanism [VSP*17]:

$$\mathbf{Q} = \mathbf{W}_q Y'^t, \mathbf{K} = \mathbf{W}_k Y'^t, \mathbf{V}\mathbf{W}_v Y'^t, \tag{6}$$

where $\mathbf{W}_q$, $\mathbf{W}_k$ and $\mathbf{W}_v$ are learnable parameter matrices.

1.4. Details of step hint module

In this paper, we employed feature-wise linear modulation to inject the fused embedding $\mathbf{e}$ into the output $\mathbf{Y}$ of the self-attention block:

$$w = \phi_w(\psi(\mathbf{e}_t)),\ b = \phi_b(\psi(\mathbf{e}_t)),\ \mathbf{Y}' = \mathrm{SiLU}(w\cdot\mathbf{Y} + b), \tag{7}$$

where $(\cdot)$ denotes element-wise multiplication, ϕ_w and ϕ_b are linear projections, and ψ is a single layer MLP with SiLU activation function. This modulation allowed the step hint embedding to influence the self-attention features. We also applied this block after each cross-attention and MLP block, enabling the timestep embedding to provide hints throughout the network.

1.5. Detailed training process

The detailed training procedure is shown in algorithm 1.

Algorithm 1: Training procedure of proposed method

Input: observed gaze-motion sequence X_{obs}, noising steps T, the initialized gaze-motion fusion network *GazePoseFuse*, the initialized gaze encoder *GazeEncoder*, the initialized gaze encoder *PoseEncoder*, the initialized noise prediction network ϵ_θ, full gaze-motion sequence X_{full}, max iterations I_{max}

Output: trained *GazePoseFuse* and ϵ_θ

$Y = \mathrm{DCT}(\mathrm{Pad}(X))$;
$H'' = GazePoseFuse(Y)$;
$Y_{full} = \mathrm{DCT}(X_{full})$;
for $t = 1, 2, ..., I_{max}$ **do**
  $Y^0_{full} = Y_{full}$;
  $t = \mathrm{Uniform}(\{1, 2, \cdots, T\})$;
  $\epsilon \sim N(0, I)$;
  $\theta = \theta - \nabla_\theta \left\|\epsilon - \epsilon_\theta\left(Y_{full}, H'', t\right)\right\|^2$
return *GazePoseFuse* and ϵ_θ.

1.6. Detailed Completion denoising process

We employed an ingenious prediction mask mechanism [CZL*23] to denoise progressively:

$$Y^{t-1} = \mathrm{DCT}\left(\mathbf{M}\odot\mathrm{iDCT}\left(Y^{t-1}_{orig}\right) + (\mathbf{1}-\mathbf{M})\odot\mathrm{iDCT}\left(Y^{t-1}_{pred}\right)\right), \tag{8}$$

where $\mathbf{M} = [\underbrace{1,1,\ldots,1}_{H},\underbrace{0,0,\ldots 0}_{F}]^{\top} \in \mathrm{R}^{(H+F)\times 1}$ is a mask vector indicating which frames are observed. Y^{t-1}_{orig} is obtained by adding $t-1$ iterations of Gaussian noise, and Y^{t-1}_{pred} is obtained by denoising the output Y^t from the previous iteration:

$$Y^{t-1}_{orig} = \sqrt{\bar{\beta}_{t-1}}Y + \sqrt{1-\bar{\beta}_{t-1}}\epsilon,\ \bar{\beta}_t = \prod_{i=1}^{t}\beta_i, \beta_i \in [0,1]. \tag{9}$$

$$Y^{t-1}_{pred} = \frac{1}{\sqrt{\beta_t}}\left(Y^t - \frac{1-\beta_t}{\sqrt{1-\bar{\beta}_t}}\epsilon_\theta\left(Y^t, H'', t\right)\right) + (1-\beta)\epsilon, \tag{10}$$

where t denotes the t-th noise iteration, β_i controls noise level, and $\epsilon \sim N(0,I)$. At the start, Y^t is sampled from Gaussian noise.

Through this iteration process, we can obtain the full generated sequence $\bar{\mathbf{p}} = \mathrm{iDCT}(Y^0)$. The last F frames were predicted motions.

The detailed denoising procedure is shown in algorithm 2.

Algorithm 2: Inference procedure of proposed method

Input: observed gaze-motion sequence X, noising steps T, the trained gaze-motion fusion network $GazePoseFuse$, the trained noise prediction network ϵ_θ, noising steps T

Output: future motions $\mathbf{p}$

$Y^T \sim N(0,I)$;

$Y = \mathrm{DCT}\left(\mathrm{Pad}\left(X_{full}\right)\right)$;

$H'' = GazePoseFuse\left(Y_{full}\right)$;

for $t \in T, T-1, ..., 1$ **do**

$\epsilon \sim N(0,I)$ if $t > 1$, else $\mathbf{z} = 0$;

$Y^{t-1}_{orig} = \sqrt{\bar{\beta}_t} Y + \sqrt{1-\bar{\beta}_t}\epsilon$;

$Y^{t-1}_{pred} = \frac{1}{\sqrt{\beta_t}}\left(Y^t - \frac{1-\beta_t}{\sqrt{1-\bar{\beta}_t}}\epsilon_\theta\left(Y^t, H'', t\right)\right) + (1-\beta)\epsilon$;

$Y^{t-1} = \mathrm{DCT}[\mathbf{M} \odot \mathrm{iDCT}\left(Y^{t-1}_{orig}\right) + (\mathbf{1}-\mathbf{M}) \odot \mathrm{iDCT}\left(Y^{t-1}_{pred}\right)]$;

$\mathbf{p} = \mathrm{iDCT}\left(Y^0\right)$;

$\bar{\mathbf{p}} = \mathbf{p}_{H:H+F}$;

return $\bar{\mathbf{p}}$

1.7. User Study Detail

Figure 1 shows an example of the instructions and definition of realism shown to the user before starting the test. In our user study, we randomly selected 18 samples for comparison. Figure 2 shows an example of our interface to enable participants to rank them. Note that in each question, the order of each method is randomly shuffled.

1.8. Further Visualisation Results

We illustrated more visualisation samples in Figure 3, Figure 4,Figure 5, Figure 6, and Figure 7

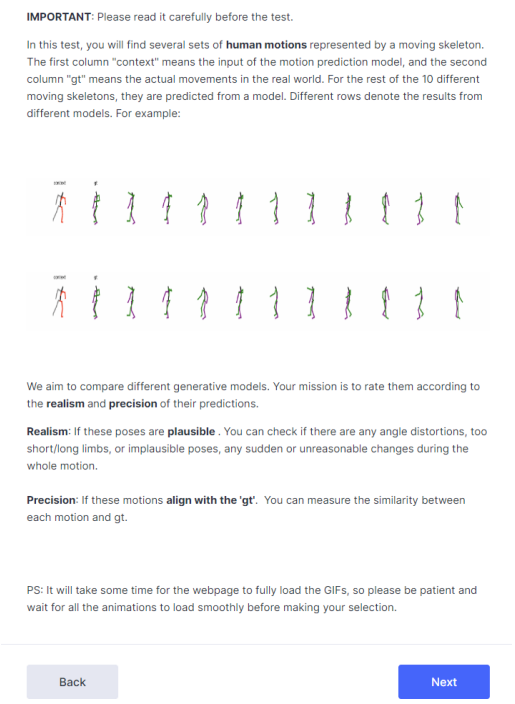

Figure 1: *The instructions and definition of realism of our questionnaire in user study*

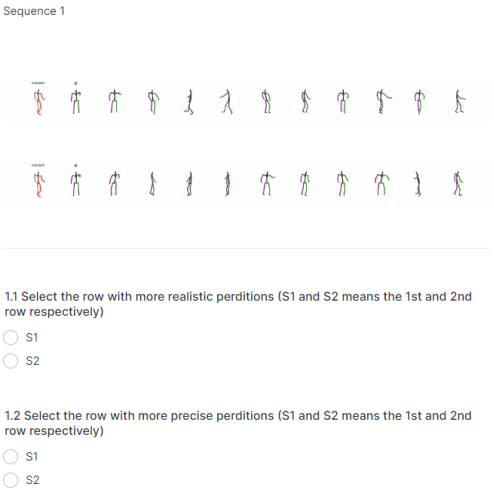

Figure 2: *The interface for ranking the predictions from different models according to realism and precision.*

References

[ANR74] AHMED N., NATARAJAN T., RAO K. R.: Discrete cosine transform. *IEEE transactions on Computers 100*, 1 (1974), 90–93. 1

[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). 1

[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. 4

[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. 1

[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). 1

[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). 1

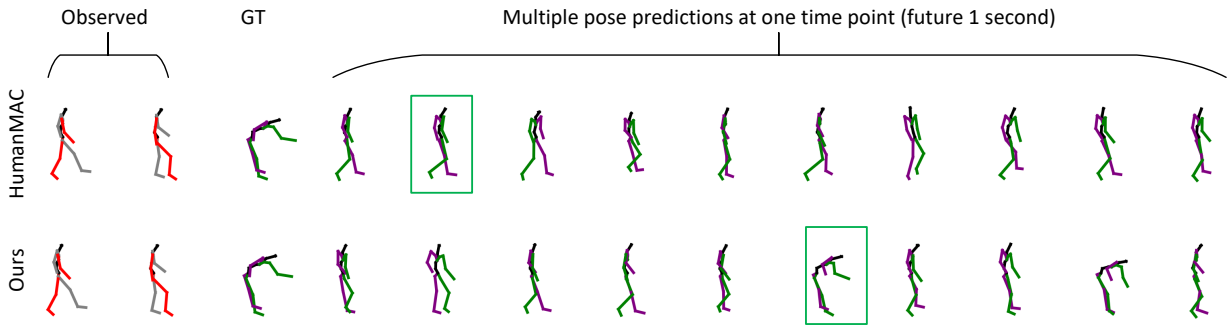

Figure 3: *Ground truth (GT) human pose at future one second and multiple pose predictions generated by different methods on the GIMO dataset [ZYM*22] with the best prediction (lowest l_2, distance to GT) boxed in green. The ground truth poses show a motion an action to bend down and take an object. Predictions from HumanMAC generally fail to forecast this intention, while our gaze-guided method is better able to predict future motion aligned better with this intention.*

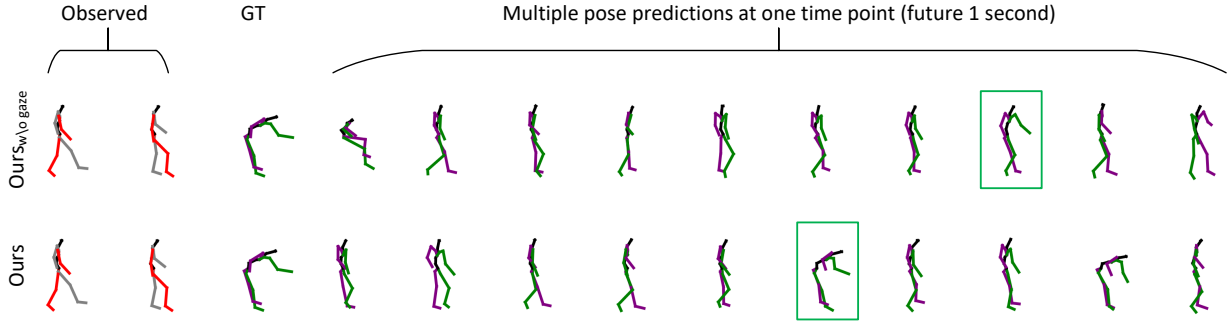

Figure 4: *Ground truth (GT) human pose at future one second and multiple pose predictions generated by different methods on the GIMO dataset [ZYM*22] with the best prediction (lowest l_2, distance to GT) boxed in green. The ground truth poses show a motion an action to bend down and take an object. Predictions from Ours$_{w/ogaze}$ generally fail to forecast this intention, while our full model is better able to predict future motion aligned better with this intention.*

[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. 3, 4

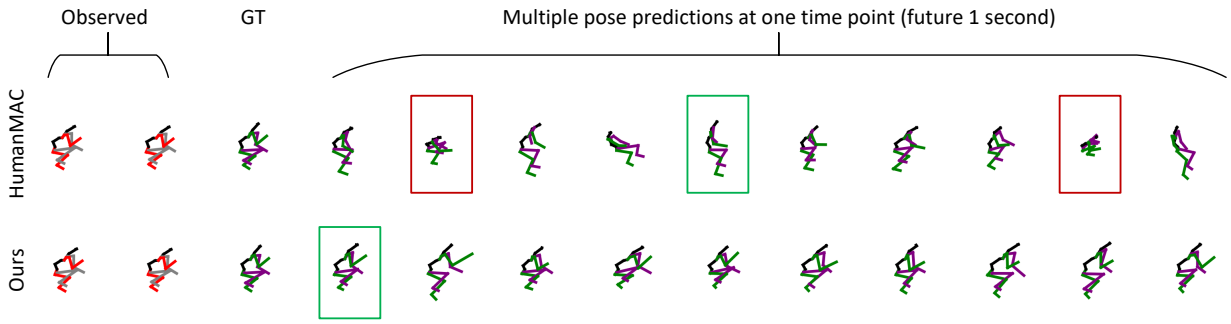

Figure 5: *Ground truth (GT) human pose at future one second and multiple pose predictions generated by different methods on the GIMO dataset [ZYM*22] with the best prediction (lowest l_2, distance to GT) boxed in green and implausible cases marked in red. Our method can generate multiple reasonable and precise predictions while some poses predicted by HumanMAC are physically implausible.*

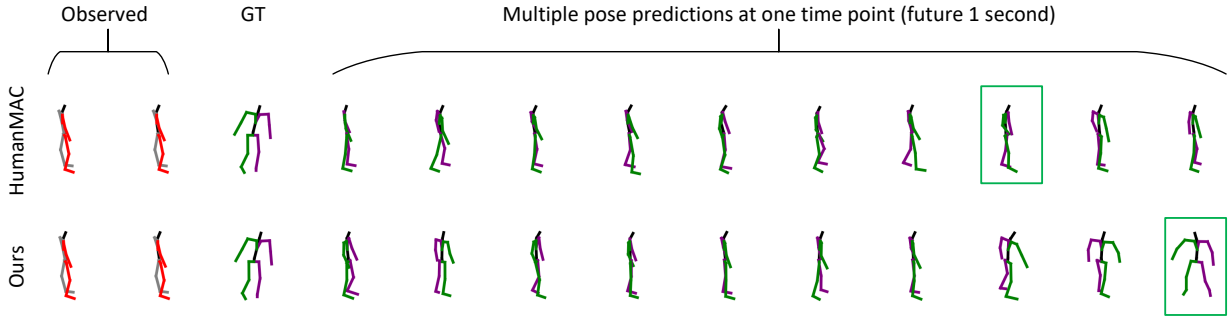

Figure 6: *Ground truth (GT) human pose at future one second and multiple pose predictions generated by different methods on the MoGaze dataset [KBM*20] with the best prediction (lowest l_2, distance to GT) boxed in green. The observed motion sequence is in place, and the groundtruth pose a second later is a sudden turn to the right at roughly 90 degrees. Predictions from HumanMAC are all in place still, while our gaze-guided method can recognise this possible intention and one of our generation accurately predicts the future of this pose.*

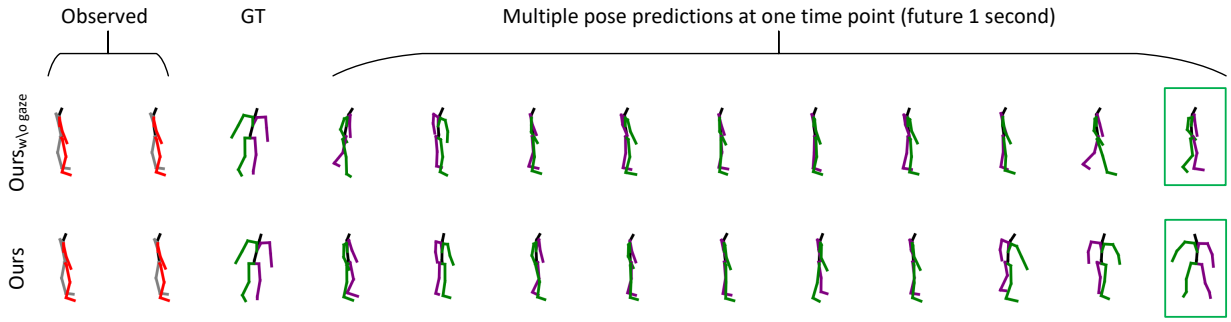

Figure 7: *Ground truth (GT) human pose at future one second and multiple pose predictions generated by different methods on the MoGaze dataset [KBM*20] with the best prediction (lowest l_2, distance to GT) boxed in green. The observed motion sequence is in place, and the groundtruth pose a second later is a sudden turn to the right at roughly 90 degrees. Predictions from $Ours_{w/ogaze}$ are all in place still, while our full model can recognise this possible intention and one of our generation accurately predicts the future of this pose.*