HOIGaze: Gaze Estimation During Hand-Object Interactions in Extended Reality Exploiting Eye-Hand-Head Coordination

ZHIMING HU*, University of Stuttgart, Germany
DANIEL HAEUFLE, University of Tuebingen, Germany
SYN SCHMITT, University of Stuttgart, Germany
ANDREAS BULLING, University of Stuttgart, Germany

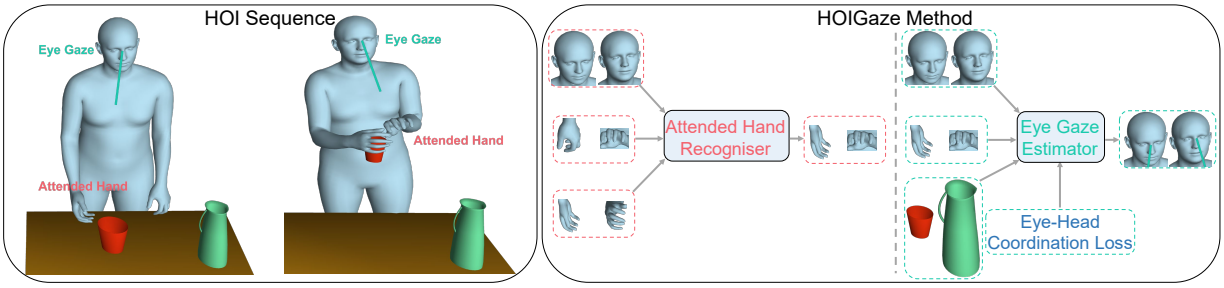

Fig. 1. HOIGaze is a novel method for estimating eye gaze during hand-object interactions in extended reality. The left figure shows an example sequence of daily HOI activity. HOIGaze uses a novel hierarchical framework that first recognises attended hand from head orientations, left and right hand gestures and then uses a gaze estimator that is trained with an eye-head coordination loss to estimate eye gaze from head orientations, attended hand, and scene objects.

We present HOIGaze – a novel learning-based approach for gaze estimation during hand-object interactions (HOI) in extended reality (XR). HOIGaze addresses the challenging HOI setting by building on one key insight: The eye, hand, and head movements are closely coordinated during HOIs and this coordination can be exploited to identify samples that are most useful for gaze estimator training – as such, effectively denoising the training data. This denoising approach is in stark contrast to previous gaze estimation methods that treated all training samples as equal. Specifically, we propose: 1) a novel *hierarchical framework* that first recognises the hand currently visually attended to and then estimates gaze direction based on the attended hand; 2) a new *gaze estimator* that uses cross-modal Transformers to fuse head and hand-object features extracted using a convolutional neural network and a spatio-temporal graph convolutional network; and 3) a novel *eye-head coordination loss* that upgrades training samples belonging to the coordinated eye-head movements. We evaluate HOIGaze on the HOT3D and Aria digital twin (ADT) datasets and show that it significantly outperforms state-of-the-art methods, achieving an average improvement of 15.6% on HOT3D and 6.0% on ADT in mean angular error. To demonstrate the potential of our method, we further report significant performance improvements for the sample downstream task of eye-based activity recognition on ADT. Taken together, our results underline the significant information content available in eye-hand-head coordination and, as such, open up an exciting new direction for learning-based gaze estimation.

CCS Concepts: • **Human-centred computing** → **Interaction techniques**; • **Computing methodologies** → **Neural networks**.

Additional Key Words and Phrases: Gaze estimation, eye-hand-head coordination, hand-object interaction, deep learning, extended reality

ACM Reference Format:
Zhiming Hu, Daniel Haeufle, Syn Schmitt, and Andreas Bulling. 2025. HOIGaze: Gaze Estimation During Hand-Object Interactions in Extended Reality Exploiting Eye-Hand-Head Coordination. *ACM Trans. Graph.* 1, 1 (April 2025), 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

With the growing popularity of extended reality (XR), analysing and understanding human behaviour in XR environments has become a popular research topic. Human eye gaze estimation in particular has significant relevance for a number of XR applications including 1) gaze-based interaction that employs eye movements to select or interact with 3D objects [Sidenmark and Gellersen 2019b]; 2) gaze-contingent rendering that maintains high rendering quality in gaze central region while reducing the fidelity in peripheral region to improve rendering efficiency [Patney et al. 2016]; 3) gaze-based intention estimation that uses eye gaze features to predict interaction intentions [Belardinelli et al. 2022; Sun et al. 2018]; or 4) eye-based activity recognition that recognises user activities based on their eye movements [Hu et al. 2022].

Estimating human eye gaze in XR environments is challenging because human gaze behaviour is influenced by both bottom-up scene content and various top-down factors, e.g. tasks that a user desires to finish [Hu et al. 2021, 2022]. Prior works on gaze estimation and analysis in XR typically focused on free-viewing conditions in which

*Corresponding author

Authors' addresses: Zhiming Hu, University of Stuttgart, Germany; Daniel Haeufle, University of Tuebingen, Germany; Syn Schmitt, University of Stuttgart, Germany; Andreas Bulling, University of Stuttgart, Germany.

ACM 0730-0301/2025/4-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

no specific task is assigned to users [Hu et al. 2019; Sitzmann et al. 2018] or non-interactive scenarios where users cannot naturally interact with the environment [Hadnett-Hunter et al. 2019; Hu et al. 2022]. However, free-viewing and non-interactive tasks have limited relevance for practical XR applications, in which users typically desire to naturally interact with the environment to perform a particular task. Gaze estimation in the more practically relevant but also significantly more challenging scenarios that involve hand-object interactions (HOIs) has been largely neglected so far.

To fill this gap, we present *HOIGaze* – the first gaze estimation method specifically geared to hand-object interactions in XR. Our key insight is that eye, hand, and head movements are strongly coordinated during HOIs and this coordination can be used to effectively denoise the training samples to improve gaze estimation performance. Specifically, we propose a novel *hierarchical framework* that first recognises the *attended hand* – the hand that is closest to eye gaze – by comparing the angular distance between gaze direction and a vector pointing from the eye to both hand centres. In a second step the method then estimates eye gaze based on the attended hand (see Figure 1). We further present a new *gaze estimator* that combines a convolutional neural network (CNN) to extract head features with a spatio-temporal graph convolutional network (GCN) to extract features from the attended hand gestures and scene object positions. The estimation further uses two cross-modal Transformers to fuse the head and hand-object features to estimate eye gaze. Finally, we introduce a novel *eye-head coordination loss* that upgrades training samples belonging to coordinated eye-head movements to improve the generalisation ability of the gaze estimator. We extensively evaluate our method on the HOT3D dataset [Banerjee et al. 2024] for HOIs as well as on the Aria digital twin (ADT) dataset [Pan et al. 2023] that contains a mixture of free-viewing, non-interactive, and HOI scenarios. Experimental results show that HOIGaze outperforms the state of the art by a large margin, achieving an average improvement of 15.6% on HOT3D and 6.0% on ADT in mean angular error. Complementing these evaluations, we also evaluate the effectiveness of our method for the sample downstream task of eye-based activity recognition on ADT and demonstrate that using our method results in significant performance improvements. [1]

The specific contributions of our work are three-fold:

- We propose HOIGaze – a novel method for estimating eye gaze during HOIs in extended reality that exploits the close coordination between eye, hand, and head movements. It combines a novel hierarchical framework, a new gaze estimator that uses cross-modal Transformers to fuse the head and hand-object features extracted using a CNN and a spatio-temporal GCN, and a novel eye-head coordination loss.
- We report extensive experiments on two public datasets for both HOI and mixed settings and demonstrate significant performance improvements over several state-of-the-art methods.
- We show the effectiveness of our method for the sample downstream task of eye-based activity recognition, also showing significant performance improvements.

[1]The full source code and trained models will be released upon acceptance.

2 RELATED WORK

2.1 Eye Gaze Estimation

Human eye gaze estimation or visual attention prediction has been a popular topic in the area of vision research for decades. Typical gaze estimation methods can be classified into bottom-up methods that focus on low-level visual scene content [Itti et al. 1998; Wang et al. 2023] or top-down approaches that take high-level context into consideration [Koulieris et al. 2016; Wang et al. 2024]. For example, Itti et al. extracted multiscale colour, intensity, and orientation features from 2D images to predict saliency map (density map of gaze distribution) [Itti et al. 1998]. Wang et al. predicted saliency map of information visualisations using both the visualisation content and the questions assigned to the viewers [Wang et al. 2024].

Recently, with the increasing use of extended reality, a lot of efforts have been devoted to analysing and estimating human eye gaze in XR environments. Some researchers focused on free-viewing settings where no specific task is assigned to the viewers. For example, Sitzmann et al. collected users' free-viewing eye gaze data on 360-degree images and adapted existing saliency predictors to predict saliency maps of 360-degree images [Sitzmann et al. 2018]. Hu et al. recorded users' eye movements during freely exploring static or dynamic virtual environments for developing eye gaze estimation models [Hu et al. 2020, 2019]. Other researchers devoted to analysing eye gaze in non-interactive scenarios where users cannot naturally interact with the environment. Specifically, Hadnett-Hunter et al. explored the effect of three tasks on visual attention in desktop monitor-based virtual environments [Hadnett-Hunter et al. 2019]. Hu et al. analysed the differences of eye gaze patterns under four different visual tasks during viewing 360-degree videos [Hu et al. 2022]. However, free-viewing or non-interactive settings have limited relevance for practical XR applications. In stark contrast, in this work we investigate gaze estimation in the more challenging but also more practically relevant hand-object interaction scenarios.

2.2 Eye-Hand-Head Coordination

Analysing and understanding the coordination of human eye, hand, and head movements is a significant topic in the areas of cognitive science and human-centred computing. Stahl [Stahl 1999] and Sidenmark et al. [Sidenmark and Gellersen 2019a] both found that eye gaze is coordinated with head movements during the gaze shift process. Hu et al. revealed that human eye movements in virtual environments have strong correlations with their head movements in both free-viewing [Hu et al. 2020, 2019] and task-oriented scenarios [Hu et al. 2021, 2022]. Kothari et al. observed coordinated patterns of human eye and head movements in real-world daily activities [Kothari et al. 2020]. Hu et al. learned generalisable joint representations of hand trajectories and head orientations in extended reality [Hu et al. 2024d]. Belardinelli et al. observed the coordinated patterns of human eye gaze and hand trajectories during daily pick and place activities in virtual environments [Belardinelli et al. 2022]. Hu et al. revealed the correlation between eye gaze direction and wrist movements in various daily activities [Hu et al. 2024b]. In stark contrast with prior works, we are the first to exploit eye-hand-head coordination to effectively denoise training samples to improve gaze estimation performance.

2.3 Hand-Object Interaction

Hand-object interaction is an important interaction paradigm in people's daily life and has been studied by many researchers. Damen et al. employed egocentric images in various HOI scenarios to recognise or anticipate user activities [Damen et al. 2022] while Zhang et al. used egocentric images with per-pixel segmentation labels of hands and objects for activity recognition [Zhang et al. 2022b]. Shi et al. took temporal inter-dependencies between HOI actions into consideration to generate procedure actions in instructional videos [Shi et al. 2025]. Liu et al. explored action recognition, motion forecasting, and cooperative grasp synthesis during bimanual hand-object manipulation process [Liu et al. 2024]. Zhan et al. investigated hand mesh reconstruction, task-aware motion fulfillment, and complex task completion during complex HOI activities [Zhan et al. 2024]. Despite plenty of research on HOIs, eye gaze estimation during HOI activities has been neglected so far. To fill this gap, in this work we explore gaze estimation under HOI scenarios.

3 METHOD

3.1 Method Design

Problem Formulation. We define gaze estimation during hand-object interactions in extended reality as the task of generating a sequence of eye gaze directions $G = \{g_i\}_{i=1}^{T} \in R^{3\times T}$, where g_i is a 3D unit vector and T is the sequence length, from hand gestures, scene objects, and head movements. To enhance hand gestures with more context, we represent hand gestures using the 3D positions of all the hand joints as well as the head and wrist positions. Specifically, the left and right hand gestures are represented as $LH = \{he_i, lw_i, rw_i, lh_i\}_{i=1}^{T} \in R^{3\times(N+3)\times T}$ and $RH = \{he_i, lw_i, rw_i, rh_i\}_{i=1}^{T} \in R^{3\times(N+3)\times T}$ respectively, where he_i, lw_i, and rw_i are the 3D positions of the head, left and right wrists, $lh_i \in R^{3\times N}$ and $rh_i \in R^{3\times N}$ refer to the 3D positions of the left and right hands and N is the number of hand joints. We denote scene objects using the 3D positions of the object centres $O = \{o_i^1, o_i^2, ..., o_i^J\}_{i=1}^{T} \in R^{3\times J\times T}$, where J is the number of objects. We use the head forward directions to represent head orientations $H = \{h_i\}_{i=1}^{T} \in R^{3\times T}$, where h_i is a 3D unit vector.

Design of HOIGaze. We observe that during hand-object interactions human visual attention is usually attracted to one hand at a specific time. For example, in a scenario where a user first picks up a cup on the table using their right hand and then picks up a jug with their left hand, the user would pay their attention to the right hand first and then turn to the left hand (see Figure 1). The attended hand is highly correlated with eye gaze while the unattended hand has little correlation. Therefore, knowing which hand is the attended one has significant potential for estimating eye gaze. Based on this observation, we propose a novel hierarchical framework that combines an attended hand recogniser and an eye gaze estimator (see Figure 2 for an overview of our method). The attended hand recogniser uses a convolutional neural network and two spatio-temporal graph convolutional networks to extract features from head orientations, left and right hand gestures respectively, and then concatenates these features to recognise attended hand via a convolutional neural network. The gaze estimator first uses a convolutional neural network to extract head features and a spatio-temporal graph convolutional network to extract features from the attended hand gestures and scene object positions, then employs two cross-modal Transformers to fuse the head and hand-object features, and finally uses a convolutional neural network to estimate eye gaze from the fused features.

3.2 Attended Hand Recogniser

Head Orientation Feature Extraction. Considering the good performance of 1D CNN for processing head movement data [Hu et al. 2021, 2022, 2020], we used three 1D CNN layers, each with 32 channels and a kernel size of three, to extract features from head orientations $H \in R^{3\times T}$. The first two CNN layers were followed by a layer normalisation (LN) and a Tanh activation while the third CNN layer was followed by a Tanh activation. After the three CNN layers, we obtained the head orientation features $f_{he} \in R^{32\times T}$.

Hand Gesture Feature Extraction. In light of the superior performance of graph convolutional networks for processing body and hand pose data [Hu et al. 2024c; Tang et al. 2024], we used two spatio-temporal GCNs to extract features from the left and right hand gestures respectively. Specifically, we modelled the hand gesture data ($LH \in R^{3\times(N+3)\times T}$ or $RH \in R^{3\times(N+3)\times T}$) as fully connected spatial and temporal graphs with their adjacency matrices measuring the weights between each pair of nodes. The spatial graph consists of $N + 3$ joints representing the hand joints, head, and wrists, respectively, while the temporal graph contains T nodes corresponding to hand gesture at different time steps. We first mapped the original hand gesture data into a latent feature space using a spatial-temporal graph convolutional network (ST-GCN). The ST-GCN first multiplied the data with a temporal adjacency matrix $A^T \in R^{T\times T}$ to perform temporal convolution, then used a feature matrix $W \in R^{3\times 8}$ to map the original node features (3 dimensions) into latent space (8 dimensions), and finally multiplied the data with a spatial adjacency matrix $A^S \in R^{(N+3)\times(N+3)}$ to perform spatial convolution. We copied the output of the ST-GCN along the temporal dimension ($R^{8\times(N+3)\times T} \rightarrow R^{8\times(N+3)\times 2T}$) to enhance the features [Ma et al. 2022]. We further used a residual GCN module that contains two GCN blocks to process the enhanced data. Each GCN block consists of an ST-GCN, a layer normalisation, a Tanh activation, and a dropout layer with a dropout rate of 0.3 to avoid overfitting. The feature matrix of the ST-GCN used in the GCN block was set to $W \in R^{8\times 8}$, ensuring that the input and output of the GCN block had the same size. A residual connection was applied for each GCN block to improve the network flow. We finally cut the output of the residual GCN module in half along the temporal dimension to obtain the hand gesture features ($f_{lh} \in R^{8\times(N+3)\times T}$ and $f_{rh} \in R^{8\times(N+3)\times T}$).

Attended Hand Recognition. To recognise attended hand, we first aggregated the hand gesture features along the spatial dimension ($R^{8\times(N+3)\times T} \rightarrow R^{8(N+3)\times T}$). We then concatenated the head orientation, left and right hand gesture features along the spatial dimension and obtained $f \in R^{(16(N+3)+32)\times T}$. We finally applied two CNN layers, each with a kernel size of three, to process the concatenated features. The first CNN layer had 64 channels and was

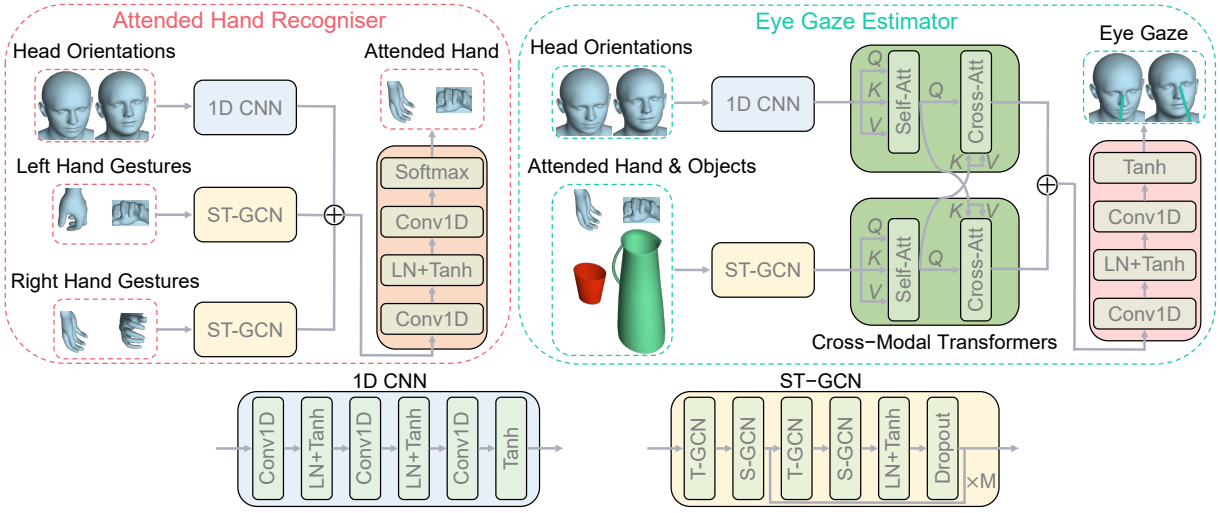

Fig. 2. HOIGaze combines an attended hand recogniser and an eye gaze estimator. The attended hand recogniser uses a 1D CNN and two ST-GCNs to extract features from head orientations, left and right hand gestures, respectively, to recognise the attended hand. The gaze estimator uses an ST-GCN to extract features from the attended hand and scene objects, and then fuses the head and hand-object features using cross-modal Transformers to estimate eye gaze.

followed by a layer normalisation and a Tanh activation function while the second CNN layer had two channels and was followed by a Softmax activation to generate the probabilities of the left and right hands being the attended hand.

3.3 Eye Gaze Estimator

Head Orientation Feature Extraction. We employed the same CNN module as used in the attended hand recogniser (subsection 3.2) to obtain head orientation features $f_{he} \in R^{32\times T}$.

Hand-Object Feature Extraction. Considering that human visual attention is more likely to be attracted by the scene objects that are close to the attended hand, we first calculated the average distance between object centre and all the joints of the attended hand for every scene object and then added the nearest scene object to the representation of the attended hand $AH = \{he_i, lw_i, rw_i, ah_i, o_i^{ah}\}_{i=1}^{T} \in R^{3\times(N+4)\times T}$, where ah_i refers to the attended hand joints and o_i^{ah} denotes the 3D positions of the nearest scene object. We further used an ST-GCN and a residual GCN module that contains four GCN blocks to extract features from the hand-object data. The ST-GCN and GCN block had the same architecture as used in the attended hand recogniser (subsection 3.2) except that the spatial graph had $N+4$ nodes rather than $N+3$. We finally aggregated the hand-object features along the spatial dimension and obtained $f_{ah} \in R^{8(N+4)\times T}$.

Head-Hand-Object Feature Fusion. In light of the good performance of cross-modal transformers for fusing different features [Yan et al. 2024; Zhang et al. 2022a], we used cross-modal transformers to fuse the head and hand-object features by modelling correlations between different time steps. To this end, we first applied a self-attention block to enhance the head and hand-object features respectively. Given input features $X \in R^{T\times n}$ where T refers to sequence length and n denotes the feature dimension, the self-attention block first calculated query feature vectors $Q \in R^{T\times n}$, key feature vectors $K \in R^{T\times n}$, and value feature vectors $V \in R^{T\times n}$ using $Q = W_q X$, $K = W_k X$, and $V = W_v X$, where W_q, W_k, and W_v are the linear projections to generate Q, K, and V. The self-attention block then enhanced the input features X using

$$Y = X + softmax(\frac{Q \otimes K^T}{\sqrt{n}}) \otimes V, \tag{1}$$

where $Y \in R^{T\times n}$ is the enhanced features and $\otimes$ refers to matrix multiplication. After the self-attention block, we further used a cross-attention block to fuse the head and hand-object features. Specifically, we used one modality to calculate Q and the other modality to compute K and V and then applied Equation 1 to fuse the two modalities. After the cross-modal transformers, we obtained the enhanced head features $f_{he}^{'} \in R^{32\times T}$ and hand-object features $f_{ah}^{'} \in R^{8(N+4)\times T}$.

Eye Gaze Estimation. To estimate eye gaze, we first concatenated the head and hand-object features along the spatial dimension and obtained $f \in R^{(8(N+4)+32)\times T}$. We then used two 1D CNN layers, each with a kernel size of three, to process the concatenated features. The first CNN layer had 64 channels and was followed by a layer normalisation and a Tanh activation while the second CNN layer used three channels and a Tanh activation to generate eye gaze. We finally normalised the output to unit vectors to represent eye gaze directions $\hat{G} = \{\hat{g}_i\}_{i=1}^{T} \in R^{3\times T}$.

3.4 Loss Function

We first trained the attended hand recogniser and then used the recognised attended hand to train the gaze estimator. Specifically, we trained the recogniser using the cross entropy loss given its good performance for classification task. To train the gaze estimator, we proposed a novel eye-head coordination loss that increases the weights of the training samples belonging to eye-head coordinated

movements:

$$\ell_i = \begin{cases} f_{eh} * (g_i - \hat{g}_i)^2, & \text{if } g_i \cdot h_i > Cos_{eh} \\ (g_i - \hat{g}_i)^2, & \text{otherwise} \end{cases} \quad (2)$$

where $g_i \cdot h_i$ calculates the cosine similarity between eye gaze direction and head orientation, Cos_{eh} is the threshold for eye-head cosine similarity and is set to 0.8, and f_{eh} is the weighting factor and is set to 4.0. The insight behind this loss function is that the coordinated eye-head movements are much more pervasive than the movements with little eye-head correlation [Hu et al. 2024b; Nakashima et al. 2015; Sidenmark and Gellersen 2019a; Sitzmann et al. 2018]. Therefore, increasing the weights of coordinated eye-head training samples can improve the generalisation ability of the gaze estimator.

4 EXPERIMENTS AND RESULTS

4.1 Datasets

To evaluate our method's generalisation capability for different scenarios, we tested our method on the HOT3D dataset [Banerjee et al. 2024] for HOI setting as well as on the ADT dataset [Pan et al. 2023] for mixed setting that contains a mixture of free-viewing, non-interactive, and HOI scenarios.

HOT3D Dataset. The HOT3D dataset contains human eye gaze, head pose, wrist pose, hand gestures, and 3D scene objects recorded at 30 Hz during various HOI activities in three different environments including *living room*, *kitchen*, and *office*. The original test set of HOT3D is not publicly available, so we only used HOT3D's original training set that contains 136 recordings from nine subjects and each recording lasts for around two minutes. To evaluate our method's generalisation capability for different users and environments, we respectively performed a cross-user evaluation and a cross-scene evaluation. For cross-user evaluation, we split the data into three user sets, i.e. {P1, P2, P3}, {P9, P10, P11}, and {P12, P14, P15}, trained on two sets and tested on the remaining one, and repeated this procedure three times by testing for a different user set. For cross-scene evaluation, we trained on two environments and tested on the remaining one, and repeated this procedure three times by testing for a different scene.

ADT Dataset. The ADT dataset collects human eye gaze, head pose, wrist pose and 3D scene objects at 30 Hz during various indoor activities including *work*, *room decoration*, and *meal preparation*, in which free-viewing, non-interactive, and HOI scenarios are mixed together. The dataset contains 34 sequences and each sequence lasts for around two minutes. For evaluation on ADT, we followed prior works [Hu et al. 2024b,c] to use 24 sequences for training and the remaining 10 sequences for testing. The ADT dataset does not record dynamic hand gestures but provides a static hand gesture, in which the 3D positions of the fingers are determined only by the wrist pose while the relative finger pose doesn't change over time. For experiments on ADT, we used the static hand gesture to train and test our method.

4.2 Evaluation Settings

Evaluation Metric. As is common in gaze estimation [Hu et al. 2021, 2020, 2024b], we used the mean angular error between the estimated gaze directions and the ground truth as the metric to evaluate different methods.

Baselines. We compared our method with the following state-of-the-art gaze estimation methods designed for XR environments:

- *Head Direction*: *Head Direction* is frequently used as a proxy for eye gaze in XR due to the strong correlation between eye and head movements [Hu et al. 2020, 2019; Sitzmann et al. 2018].
- *DGaze* [Hu et al. 2020]: *DGaze* estimates eye gaze from the scene content and head movements via convolutional neural networks.
- *FixationNet* [Hu et al. 2021]: *FixationNet* combines prior knowledge of gaze distribution with head and scene features extracted by convolutional neural networks to estimate eye gaze.
- *Pose2gaze* [Hu et al. 2024b]: *Pose2gaze* estimates eye gaze from body movements using graph convolutional networks.

Time Horizon. We used 15 frames (corresponding to 500 ms) of hand-head-object data as input to estimate the corresponding gaze directions $G_{t:t+14} = \{g_t, g_{t+1}, ..., g_{t+14}\}$ following common evaluation settings for gaze estimation in XR [Hu et al. 2020, 2024b].

Implementation Details. We trained the baseline methods from scratch using their default parameters. We trained our attended hand recogniser using the AdamW optimiser with an initial learning rate of 0.005 and a weight decay coefficient of 0.05. We decayed the learning rate by 0.95 every epoch and trained the recogniser for 60 epochs using a batch size of 32. We trained our gaze estimator using the Adam optimiser with an initial learning rate of 0.005 that was decayed by 0.95 every epoch. We used a batch size of 32 to train the gaze estimator for a total of 80 epochs. We implemented our method with the PyTorch framework using a Linux machine with one NVIDIA V100 GPU.

4.3 Gaze Estimation Results

Results on HOT3D (Cross-User). Table 1-*HOT3D (Cross-User)* summarises the performances of different methods on the HOT3D dataset for cross-user evaluation. We can see that our method consistently outperforms the state of the art in terms of both average performance and performances at different user sets. Specifically, our method achieves an average improvement of 15.6% (9.37° *vs.* 11.10°) in mean angular error, an improvement of 13.7% (9.23° *vs.* 10.69°) on {P1, P2, P3}, 14.6% (9.16° *vs.* 10.73°) on {P9, P10, P11}, and 17.9% (9.69° *vs.* 11.80°) on {P12, P14, P15}. We further performed a paired Wilcoxon signed-rank test to compare the mean angular error of our method with that of the state-of-the-art methods and the results validated that the differences between our method and prior methods are statistically significant ($p < 0.01$). The above results demonstrate that our method has a strong generalisation capability for different users. Figure 3 shows the visualisation of the gaze estimation results from our method and the state-of-the-art method *Pose2Gaze* [Hu et al. 2024b]. We can see that our method achieves better performance than the state-of-the-art method at different scenarios and different activities. More visualisation results are provided in the supplementary video.

Results on HOT3D (Cross-Scene). Table 1-*HOT3D (Cross-Scene)* shows the mean angular errors of different methods on HOT3D for

Table 1. Mean angular errors of different methods on the HOT3D and ADT datasets. Best results are in bold.

	HOT3D (Cross-User)				HOT3D (Cross-Scene)				ADT			
	{P1, P2, P3}	*{P9, P10, P11}*	*{P12, P14, P15}*	Average	*Room*	*Kitchen*	*Office*	Average	*Work*	*Decoration*	*Meal*	Average
Head Direction	23.24°	28.00°	17.85°	23.20°	23.69°	22.83°	23.16°	23.20°	22.88°	18.44°	25.23°	22.25°
DGaze [Hu et al. 2020]	12.17°	15.08°	14.87°	14.29°	13.37°	12.98°	11.39°	12.81°	8.84°	10.53°	10.77°	9.92°
FixationNet [Hu et al. 2021]	11.90°	14.60°	14.78°	14.00°	12.78°	12.84°	11.34°	12.53°	8.82°	10.50°	10.83°	9.92°
Pose2Gaze [Hu et al. 2024b]	10.69°	10.73°	11.80°	11.10°	9.79°	9.73°	9.96°	9.80°	8.25°	9.71°	10.43°	9.34°
Ours	**9.23°**	**9.16°**	**9.69°**	**9.37°**	**8.55°**	**8.69°**	**8.69°**	**8.64°**	**7.81°**	**9.46°**	**9.41°**	**8.78°**
Ours w/o attended hand	9.89°	11.24°	10.57°	10.67°	9.71°	9.32°	9.16°	9.43°	8.26°	9.97°	9.87°	9.25°
Ours w/o Transformers	9.60°	10.07°	10.24°	10.02°	8.87°	8.85°	9.17°	8.92°	8.03°	9.74°	9.96°	9.12°
Ours w/o eye-head coord. loss	9.83°	9.48°	9.70°	9.64°	8.79°	8.71°	8.84°	8.76°	7.87°	9.49°	9.71°	8.90°
Ours w/ GT attended hand	**8.68°**	**8.82°**	**9.35°**	**8.98°**	**8.41°**	**8.25°**	**8.40°**	**8.34°**	**7.59°**	**9.18°**	**9.28°**	**8.57°**

cross-scene evaluation. It can be seen that our method consistently outperforms other methods in both average performance and performances at different environments. Specifically, our method achieves an average improvement of 11.8% (8.64° *vs.* 9.80°), an improvement of 12.7% (8.55° *vs.* 9.79°) at *living room*, 10.7% (8.69° *vs.* 9.73°) at *kitchen*, and 12.8% (8.69° *vs.* 9.96°) at *office*. A paired Wilcoxon signed-rank test was conducted and the results indicated that the differences between our method and the state of the art are statistically significant ($p < 0.01$). These results show that our method has a strong generalisation capability for different XR environments.

Results on ADT. The gaze estimation performances of different methods on the ADT dataset are presented at Table 1-*ADT*. We can see from the table that our method consistently outperforms prior methods at different activities including *work*, *room decoration*, and *meal preparation*, and achieves an average improvement of 6.0% (8.78° *vs.* 9.34°). A paired Wilcoxon signed-rank test validated that our improvement is statistically significant ($p < 0.01$). The above results demonstrate that our method maintains superior performance than prior methods for mixed setting that contains a mixture of free-viewing, non-interactive, and HOI scenarios.

4.4 Ablation Study

Attended Hand. To test the effectiveness of the recognised attended hand, we re-trained our gaze estimator using both the left and right hands rather than the attended hand. We can see from Table 1 that our method significantly outperforms the ablated version of not using attended hand (paired Wilcoxon signed-rank test, $p < 0.01$). In addition, we also re-trained our gaze estimator using the ground truth attended hand and the results in Table 1 further validate the usefulness of the attended hand. Furthermore, we analysed the error cases when the recognised attended hand is wrong and found that our method can achieve superior or comparable performance with the state of the art in these cases, demonstrating the robustness of our method (see supplementary material for details).

Cross-Modal Transformer. We re-trained our gaze estimator without using the cross-modal Transformers and the results in Table 1 demonstrate that cross-modal Transformers help improve the performance significantly (paired Wilcoxon signed-rank test, $p < 0.01$). We also validated that both the self-attention and cross-attention blocks used in the cross-modal Transformers contribute to the overall performance (see supplementary material for details).

Table 2. Mean angular errors of our method's different ablated versions on the HOT3D and ADT datasets. Best results are in bold.

	HOT3D-User	HOT3D-Scene	ADT
w/o recogniser GCN	9.75°	8.78°	8.99°
w/o estimator GCN	9.87°	9.13°	9.20°
w/o head orientations	10.53°	9.29°	9.42°
w/o head-wrist positions	12.82°	11.47°	9.57°
w/o hand gestures	9.63°	8.74°	-
w/o scene objects	10.34°	9.19°	9.03°
Ours	**9.37°**	**8.64°**	**8.78°**

Eye-Head Coordination Loss. We replaced the eye-head coordination loss with a mean squared error (MSE) loss to re-train our method. The results in Table 1 verify that the eye-head coordination loss can significantly improve our method's gaze estimation performance (paired Wilcoxon signed-rank test, $p < 0.01$).

GCN in Attended Hand Recogniser and Gaze Estimator. We respectively removed the residual GCNs used in our attended hand recogniser and gaze estimator to re-train our method. We can see from the results in Table 2 that using residual GCNs achieves significantly better performances than not using them (paired Wilcoxon signed-rank test, $p < 0.01$). We also changed the number of GCN layers and validated that using two residual GCNs in the attended hand recogniser and four residual GCNs in the gaze estimator achieves the best performance (see supplementary material for details).

Input Modality. We respectively tested different ablated versions of our method that did not contain head orientations, head-wrist positions, dynamic hand gestures (i.e., replace dynamic hand gestures with static ones), and scene objects. We can see from Table 2 that our method consistently outperforms the ablated versions and the results are statistically significant (paired Wilcoxon signed-rank test, $p < 0.01$), thus underlining the effectiveness of each input modality used in our method. We also changed the number of scene objects and validated that using the nearest scene object achieves the best performance (see supplementary material for details).

5 EYE-BASED ACTIVITY RECOGNITION

Activity recognition is important for many XR scenarios such as low-latency predictive interfaces [David-John et al. 2021; Keshava

Table 3. Eye-based activity recognition accuracies of different methods on ADT. Best results are in bold while the second best are underlined.

GT	Ours	*Pose2Gaze*	*FixationNet*	*DGaze*	*Head Direction*	*Chance*
72.9%	<u>71.8%</u>	68.7%	66.6%	66.0%	47.1%	33.3%

et al. 2020], adaptive virtual environment design [Hadnett-Hunter et al. 2019], or human-aware intelligent systems [Vortmann and Putze 2020]. It is well-known that human eye gaze can be directly used to recognise user activities [Bulling et al. 2010; Coutrot et al. 2018; Hu et al. 2022]. Therefore, eye-based activity recognition is a particularly relevant sample downstream task to further evaluate the quality of the estimated eye gaze.

Dataset. We tested on the ADT dataset given that it provides activity labels for the recorded sequences. We used the same training and test sets as described in subsection 4.1.

Activity Recognition Method. We used *EHTask* [Hu et al. 2022] – the state-of-the-art eye-based activity recogniser to evaluate the estimated eye gaze. *EHTask* employs a 1D CNN and a bidirectional gated recurrent unit (GRU) to extract eye gaze features and then uses fully-connected layers to recognise activities from the eye features.

Procedure. We trained *EHTask* using its default parameters to recognise three activities, i.e. *work*, *room decoration*, and *meal preparation*, from the ground truth eye gaze. At test time, we used the eye gaze generated from different methods as input to *EHTask* to evaluate their effectiveness on activity recognition.

Results. Table 3 shows the activity recognition accuracies of using the ground truth eye gaze and the eye gaze generated from different methods on the ADT dataset. We can see that our method achieves better recognition performance than prior methods (71.8% vs. 68.7%) and the result is statistically significant (paired Wilcoxon signed-rank test, $p < 0.01$). We also find that the activity recognition performance of using our estimated eye gaze is comparable with that of using ground truth eye gaze (71.8% vs. 72.9%). The above results demonstrate the effectiveness of using our method to estimate eye gaze for XR-related downstream tasks such as activity recognition.

6 DISCUSSION

Significance of Our Method. Our method outperforms state-of-the-art methods by an average improvement of 15.6% for HOI setting and 6.0% for mixed setting (Table 1), validating the overall superiority of our model architecture. In addition, our method achieves superior performances than prior methods for both cross-user and cross-scene evaluations (Table 1), demonstrating that our method has strong generalisation capabilities for different users and different XR environments. Furthermore, the results from the sample application of eye-based activity recognition confirm that our method can be more effective in real applications (Table 3).

Usability of Our Method. Our method exploits information about hand gestures to estimate eye gaze. Hand gesture information is readily available in many XR devices such as HTC Vive Focus 3 and Meta Quest 3 while eye gaze information is still not provided in such devices. Our method has significant potential to be integrated into such XR devices to enable numerous eye gaze-based applications including gaze-contingent rendering [Patney et al. 2016], gaze-based interaction [Duchowski 2018; Sidenmark and Gellersen 2019b], or virtual content design and optimisation [Alghofaili et al. 2019]. In addition, even if dynamic hand gestures are not available, our method using static hand gestures still outperforms prior methods by a large margin, achieving an average improvement of 13.2% (9.63° *vs.* 11.10°) on HOT3D (Cross-User), 10.8% (8.74° *vs.* 9.80°) on HOT3D (Cross-Scene), and 6.0% (8.78° *vs.* 9.34°) on ADT (Table 1 and Table 2). These results further demonstrate the usability of our method in real applications.

Eye-Hand-Head Coordination. Our key insight is that the eye, hand, and head movements are closely coordinated during HOIs and this coordination can be exploited to identify samples that are most useful for gaze estimator training – as such, effectively denoising the training data. Experimental results showed that using the attended hand rather than both hands and increasing the weights of coordinated eye-head training samples can significantly improve the gaze estimation performance (Table 1), validating the effectiveness of our insight. This insight is not only useful for developing future gaze estimation methods but can also guide future research on related topics such as human motion forecasting [Hu et al. 2024a,c; Yan et al. 2024], human motion synthesis [Sampieri et al. 2024], and hand motion prediction [Tang et al. 2024].

Limitations and Future Work. Despite these advances, we identified several limitations that we plan to address in future work. First, to the best of our knowledge, the HOT3D and ADT datasets are the only public datasets that provide eye gaze, head movements, hand gestures, and scene objects, thus unfortunately limiting the generalisability of our evaluation. In future work, we are looking forward to assessing our method for a broader range of activities and environments. In addition, our method is specifically designed for hand-object interactions and may not work well for other scenarios such as human-human interactions. It will be interesting to explore how to adapt our method to other scenarios. Furthermore, integrating our method into XR devices to enable eye gaze-based applications is an interesting avenue of future work. Finally, adding prior knowledge on human intention during hand-object interactions, e.g. the target position in a *place* activity, to our method may further boost the gaze estimation performance.

7 CONCLUSION

In this work, we explored the challenging task of estimating human eye gaze during hand-object interactions in extended reality. We proposed a learning-based method that features a novel hierarchical framework, a new gaze estimator that uses CNN, GCN, and cross-modal Transformers to extract features from head movements, hand gestures, and scene objects, and a novel eye-head coordination loss. Through extensive experiments on two public datasets, we showed that our method consistently outperforms several state-of-the-art methods by a large margin. We also validated the effectiveness of our method for the sample application of eye-based activity recognition. As such, our work reveals the significant information content

available in eye-hand-head coordination for gaze estimation during HOIs and informs future work on this promising research direction.

REFERENCES

Rawan Alghofaili, Michael S Solah, Haikun Huang, Yasuhito Sawahata, Marc Pomplun, and Lap-Fai Yu. 2019. Optimizing visual element placement via visual attention analysis. In *Proceedings of the 2019 IEEE Conference on Virtual Reality and 3D User Interfaces*. IEEE, 464–473.

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, et al. 2024. Introducing HOT3D: An Egocentric Dataset for 3D Hand and Object Tracking. *arXiv preprint arXiv:2406.09598* (2024).

Anna Belardinelli, Anirudh Reddy Kondapally, Dirk Ruiken, Daniel Tanneberg, and Tomoki Watabe. 2022. Intention estimation from gaze and motion features for human-robot shared-control object manipulation. In *Proceedings of the 2022 IEEE International Conference on Intelligent Robots and Systems*. IEEE, 9806–9813.

Andreas Bulling, Jamie A Ward, Hans Gellersen, and Gerhard Troster. 2010. Eye movement analysis for activity recognition using electrooculography. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 4 (2010), 741–753.

Antoine Coutrot, Janet H Hsiao, and Antoni B Chan. 2018. Scanpath modeling and classification with hidden Markov models. *Behavior Research Methods* 50, 1 (2018), 362–379.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2022. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision* (2022), 1–23.

Brendan David-John, Candace E Peacock, Ting Zhang, T Scott Murdison, Hrvoje Benko, and Tanya R Jonker. 2021. Towards gaze-based prediction of the intent to interact in virtual reality. In *Proceedings of the 2021 ACM Symposium on Eye Tracking Research and Applications*. 1–7.

Andrew T Duchowski. 2018. Gaze-based interaction: a 30 year retrospective. *Computers and Graphics* 73 (2018), 59–69.

Jacob Hadnett-Hunter, George Nicolaou, Eamonn O'Neill, and Michael Proulx. 2019. The effect of task on visual attention in interactive virtual environments. *ACM Transactions on Applied Perception* 16, 3 (2019), 1–17.

Zhiming Hu, Andreas Bulling, Sheng Li, and Guoping Wang. 2021. FixationNet: forecasting eye fixations in task-oriented virtual environments. *IEEE Transactions on Visualization and Computer Graphics* 27, 5 (2021), 2681–2690.

Zhiming Hu, Andreas Bulling, Sheng Li, and Guoping Wang. 2022. EHTask: recognizing user tasks from eye and head movements in immersive virtual reality. *IEEE Transactions on Visualization and Computer Graphics* (2022).

Zhiming Hu, Sheng Li, Congyi Zhang, Kangrui Yi, Guoping Wang, and Dinesh Manocha. 2020. DGaze: CNN-based gaze prediction in dynamic scenes. *IEEE Transactions on Visualization and Computer Graphics* 26, 5 (2020), 1902–1911.

Zhiming Hu, Syn Schmitt, Daniel Haeufle, and Andreas Bulling. 2024a. GazeMotion: Gaze-guided Human Motion Forecasting. In *Proceedings of the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Zhiming Hu, Jiahui Xu, Syn Schmitt, and Andreas Bulling. 2024b. Pose2Gaze: Eye-body Coordination during Daily Activities for Gaze Prediction from Full-body Poses. *IEEE Transactions on Visualization and Computer Graphics* (2024).

Zhiming Hu, Zheming Yin, Daniel Haeufle, Syn Schmitt, and Andreas Bulling. 2024c. HOIMotion: Forecasting Human Motion During Human-Object Interactions Using Egocentric 3D Object Bounding Boxes. *IEEE Transactions on Visualization and Computer Graphics* (2024).

Zhiming Hu, Congyi Zhang, Sheng Li, Guoping Wang, and Dinesh Manocha. 2019. SGaze: a data-driven eye-head coordination model for realtime gaze prediction. *IEEE Transactions on Visualization and Computer Graphics* 25, 5 (2019), 2002–2010.

Zhiming Hu, Guanhua Zhang, Zheming Yin, Daniel Haeufle, Syn Schmitt, and Andreas Bulling. 2024d. HaHeAE: Learning Generalisable Joint Representations of Human Hand and Head Movements in Extended Reality. *arXiv preprint arXiv:2410.16430* (2024).

Laurent Itti, Christof Koch, and Ernst Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 11 (1998), 1254–1259.

Ashima Keshava, Anete Aumeistere, Krzysztof Izdebski, and Peter Konig. 2020. Decoding task from oculomotor behavior in virtual reality. In *Proceedings of the 2020 ACM Symposium on Eye Tracking Research and Applications*. 1–5.

Rakshit Kothari, Zhizhuo Yang, Christopher Kanan, Reynold Bailey, Jeff B Pelz, and Gabriel J Diaz. 2020. Gaze-in-wild: a dataset for studying eye and head coordination in everyday activities. *Scientific Reports* 10, 1 (2020), 1–18.

George Alex Koulieris, George Drettakis, Douglas Cunningham, and Katerina Mania. 2016. Gaze prediction using machine learning for dynamic stereo manipulation in games. In *Proceedings of the 2016 IEEE Virtual Reality*. IEEE, 113–120.

Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. 2024. Taco: Benchmarking generalizable bimanual tool-action-object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21740–21751.

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

Ryoichi Nakashima, Yu Fang, Yasuhiro Hatori, Akinori Hiratani, Kazumichi Matsumiya, Ichiro Kuriki, and Satoshi Shioiri. 2015. Saliency-based gaze prediction based on head direction. *Vision Research* 117 (2015), 59–66.

Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. 2023. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 20133–20143.

Anjul Patney, Marco Salvi, Joohwan Kim, Anton Kaplanyan, Chris Wyman, Nir Benty, David Luebke, and Aaron Lefohn. 2016. Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions on Graphics* 35, 6, Article 179 (2016), 12 pages.

Alessio Sampieri, Alessio Palma, Indro Spinelli, and Fabio Galasso. 2024. Length-Aware Motion Synthesis via Latent Diffusion. In *European Conference on Computer Vision*. Springer, 107–124.

Lei Shi, Paul-Christian Bürkner, and Andreas Bulling. 2025. ActionDiffusion: An Action-aware Diffusion Model for Procedure Planning in Instructional Videos. In *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Ludwig Sidenmark and Hans Gellersen. 2019a. Eye, head and torso coordination during gaze shifts in virtual reality. *ACM Transactions on Computer-Human Interaction* 27, 1 (2019), 1–40.

Ludwig Sidenmark and Hans Gellersen. 2019b. Eye&head: Synergetic eye and head movement for gaze pointing and selection. In *Proceedings of the 2019 Annual ACM Symposium on User Interface Software and Technology*. 1161–1174.

Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. 2018. Saliency in VR: how do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics* 24, 4 (2018), 1633–1642.

John S Stahl. 1999. Amplitude of human head movements associated with horizontal saccades. *Experimental Brain Research* 126, 1 (1999), 41–54.

Qi Sun, Anjul Patney, Li-Yi Wei, Omer Shapira, Jingwan Lu, Paul Asente, Suwen Zhu, Morgan McGuire, David Luebke, and Arie Kaufman. 2018. Towards virtual reality infinite walking: dynamic saccadic redirection. *ACM Transactions on Graphics* 37, 4 (2018), 1–13.

Bowen Tang, Kaihao Zhang, Wenhan Luo, Wei Liu, and Hongdong Li. 2024. Prompting Future Driven Diffusion Model for Hand Motion Prediction. In *European Conference on Computer Vision*. Springer, 169–186.

Lisa-Marie Vortmann and Felix Putze. 2020. Attention-aware brain computer interface to avoid distractions in augmented reality. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.

Yao Wang, Mihai Bâce, and Andreas Bulling. 2023. Scanpath Prediction on Information Visualisations. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 30, 7 (2023), 3902–3914. https://doi.org/10.1109/TVCG.2023.3242293

Yao Wang, Weitian Wang, Abdullah Abdelhafez, Mayar Elfares, Zhiming Hu, Mihai Bâce, and Andreas Bulling. 2024. SalChartQA: Question-driven Saliency on Information Visualisations. In *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. 1–14. https://doi.org/10.1145/3613904.3642942

Haodong Yan, Zhiming Hu, Syn Schmitt, and Andreas Bulling. 2024. GazeMoDiff: Gaze-guided Diffusion Model for Stochastic Human Motion Prediction. In *Proceedings of the 2024 Pacific Conference on Computer Graphics and Applications*.

Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. 2024. OAKINK2: A Dataset of Bimanual Hands-Object Manipulation in Complex Task Completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 445–456.

Lingzhi Zhang, Shenghao Zhou, Simon Stent, and Jianbo Shi. 2022b. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *European Conference on Computer Vision*. Springer, 127–145.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2022a. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001* (2022).

Fig. 3. Visualisation of the gaze estimation results from our method and the state-of-the-art method *Pose2Gaze* [Hu et al. 2024b] on the HOT3D dataset. The green arrow represents the ground truth eye gaze, the red arrow denotes our method, the blue arrow refers to *Pose2Gaze*. Our method exhibits higher estimation accuracy than the state-of-the-art method at different scenarios and different activities.