HAGI: Head-Assisted Gaze Imputation for Mobile Eye Trackers

Chuhan Jiao
University of Stuttgart
Stuttgart, Germany
chuhan.jiao@vis.uni-stuttgart.de

Zhiming Hu*
University of Stuttgart
Stuttgart, Germany
zhiming.hu@vis.uni-stuttgart.de

Andreas Bulling
University of Stuttgart
Stuttgart, Germany
andreas.bulling@vis.uni-stuttgart.de

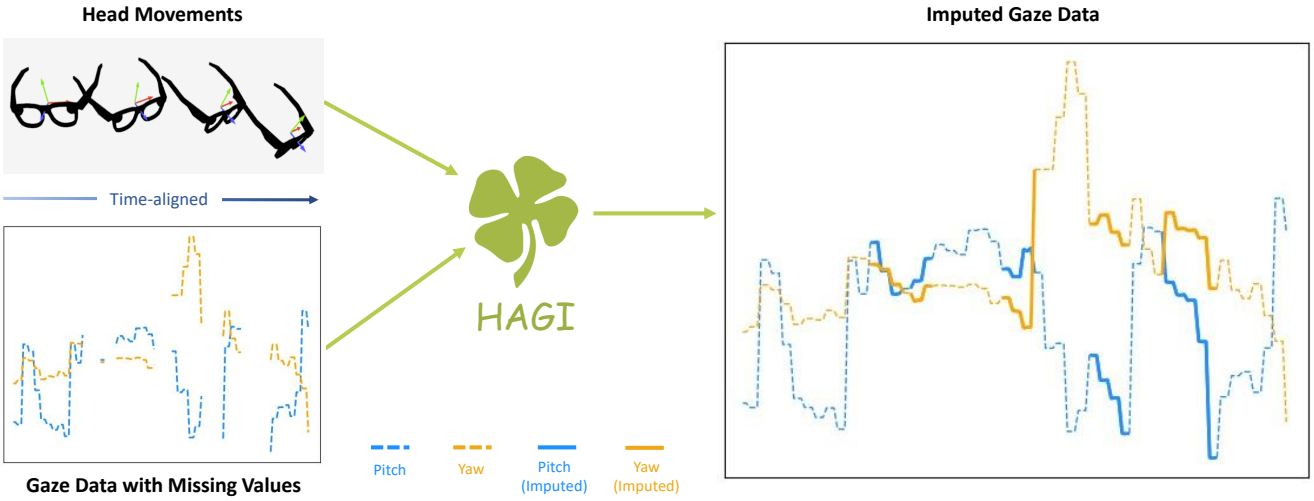

Figure 1: Missing data is inevitable in mobile eye trackers. HAGI is a novel multi-modal approach for gaze data imputation that exploits the close coordination between eye and head movements. Input to our method is gaze data with missing values and time-aligned head movements captured using sensors readily available in mobile eye trackers (Left). HAGI uses a novel head-conditional diffusion model to impute the missing gaze data (Right) with lower mean angular error and is more realistic than previous methods.

ABSTRACT

Mobile eye tracking plays a vital role in capturing human visual attention across both real-world and extended reality (XR) environments, making it an essential tool for applications ranging from behavioural research to human-computer interaction. However, missing values due to blinks, pupil detection errors, or illumination changes pose significant challenges for further gaze data analysis. To address this challenge, we introduce HAGI – a multi-modal diffusion-based approach for gaze data imputation that, for the first time, uses the integrated head orientation sensors to exploit the inherent correlation between head and eye movements. Our method includes a head-movement feature extraction module alongside a novel hybrid feature fusion mechanism that effectively integrates gaze and head motion features at multiple levels. Additionally, we introduce a tailored loss function to enhance gaze imputation accuracy further. Extensive evaluations on the large-scale Nymeria, Ego-Exo4D, and HOT3D datasets demonstrate that HAGI consistently outperforms conventional interpolation methods and deep learning-based time-series imputation baselines, reducing mean angular error by up to 22%. Furthermore, statistical analyses confirm that HAGI produces gaze velocity distributions that more closely match actual human gaze behaviour than baselines, ensuring more realistic gaze imputations. Our method paves the way for more complete and accurate eye gaze recordings in real-world settings and has significant potential for enhancing gaze-based analysis and interaction across various application domains.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Machine learning**; • **Human-centered computing**;

1 INTRODUCTION

Mobile eye tracking has become an essential tool for studying human behaviour [28, 80], attention [59, 69], cognition [42, 51], and decision-making processes [42, 88], and has emerged as an attractive modality for interaction in real-world environments [45, 52, 53, 84]. Also, latest commercial head-mounted extended reality (XR) devices, such as the Apple Vision Pro [4], Meta Quest Pro [56], Microsoft HoloLens 2 [58], and Project Aria Glasses [15] are equipped with integrated eye tracking functionality.

A fundamental challenge in working with gaze data recorded using mobile eye trackers is the prevalence of missing values arising from blinks, pupil detection failures, occlusions, or illumination changes [9, 63, 65]. Missing values can significantly degrade data quality and, in the worst case, render gaze recordings unusable for further analysis and downstream applications [20]. Prior research

*Corresponding author

has largely adopted two strategies to address this issue: discarding missing values entirely [37, 38, 60] or imputing them using interpolation methods, such as linear [5, 10, 36, 55] or nearest neighbour interpolation [97]. While discarding data ensures data integrity, it results in discontinuities, making gaze data unsuitable for applications requiring temporal completeness and consistency, such as for training machine learning models [27, 28, 30]. Data imputation preserves continuity but fails to accurately reconstruct naturalistic gaze trajectories or match the velocity profile of real human eye movements [20, 55]. As such, none of these existing approaches is fully satisfying and there remains a critical need for advanced gaze imputation techniques that can handle missing gaze data in a robust, continuity-preserving, and biologically plausible manner.

We introduce HAGI – a novel multi-modal approach that leverages the close coordination between eye and head movements (also known as eye-head coordination) for gaze imputation. Our method exploits the fact that along with gaze data, the latest mobile eye trackers and XR headsets are also readily equipped with sensors for head tracking, such as inertial sensors or vision-based approaches for self-localisation and mapping. Unlike prior methods that treat gaze as a standalone signal, HAGI is the first to integrate head movement information to infer missing gaze values. Specifically, we propose (1) a head movement feature extraction module that encodes both head rotation and translation, (2) a novel hybrid feature fusion mechanism combing early fusion and skip fusion that effectively combines head and gaze signals at multiple levels, and (3) a loss function designed to improve gaze imputation accuracy. We evaluate HAGI on three large-scale gaze datasets that cover different everyday indoor and outdoor activities: Nymeria [54], Ego-Exo4D [19], and HOT3D [7]. Results of these evaluations show that our approach significantly outperforms both interpolation-based methods and deep learning-based time-series imputation baselines. Our method reduces mean angular error (MAE) by up to 22%, and yields gaze velocity distributions that more closely resemble real human gaze movements than other methods. Our main contributions are as follows:

(1) We introduce HAGI, the first multi-modal diffusion-based approach for gaze imputation that exploits the coordination between eye and head movements.
(2) We propose a novel head-movement feature extraction module and a hybrid feature fusion mechanism, and a tailored loss function that integrates head and gaze data for robust gaze imputation.
(3) We conduct extensive evaluations on three large-scale egocentric gaze datasets, demonstrating that HAGI outperforms existing methods—achieving up to a 22% reduction in mean angular error—and generates gaze velocity distributions that closely resemble real human gaze dynamics, ensuring more naturalistic and biologically plausible imputed gaze trajectories.

By leveraging eye-head coordination, HAGI enables more reliable gaze-based analysis and interaction in mobile eye-tracking applications. Our approach enhances the quality of gaze data in real-world scenarios, making it particularly beneficial for applications in XR, behavioural research, and gaze-based human-computer interaction.

2 RELATED WORK

2.1 Gaze Imputation

The task of gaze imputation, which involves reconstructing missing values within recorded gaze data, remains an underexplored area of research. The predominant approach to handle missing values in most prior studies has been to exclude instances containing missing gaze data entirely [1, 3, 6, 8, 23, 87]. Among the limited works that do address missing data, classical interpolation techniques have been the primary method of choice. Notably, Huang and Bulling [36] applied linear interpolation to fill gaps in gaze data lasting less than 50 milliseconds, while Mannaru et al. [55] similarly employed linear interpolation for recovering missing gaze data in time-domain analyses. Alternative interpolation methods, such as nearest neighbours interpolation [97], unsupervised Expectation-Maximization algorithm [47], have also been explored in this context. However, as emphasised by Grootjen et al. [20], interpolation methods often struggle to accurately replicate the velocity distribution of real gaze data. On the other hand, machine learning techniques have also been applied to gaze super-resolution - a special type of gaze imputation. Jiao et al. [38] introduced SUPREYES, an implicit neural representation learning method for gaze super-resolution. However, it is important to note that SUPREYES primarily operates on low-resolution gaze data without any missing values, focusing on resolution enhancement rather than imputing missing data at arbitrary positions. In stark contrast to previous methods, HAGI can impute missing gaze data at any arbitrary location while preserving the natural dynamics of human gaze behaviour by closely mimicking the gaze velocity distribution.

2.2 Eye-head Coordination

Eye-head coordination refers to the coordinated movements between the eyes and the head and has been extensively investigated in the areas of cognitive science and human-centered computing. Specifically, Stahl studied the process of gaze shift and found that the amplitude of head movement is proportional to the amplitude of gaze shift [79]. Fang et al. investigated the gaze fixation process and revealed that eye-head coordination plays a significant role in visual cognitive processing [16]. Hu et al. analysed eye-head coordination in immersive virtual environments and discovered that human eye gaze positions are strongly correlated with head rotation velocities [29, 30, 34]. Sidenmark et al. focused on the process of gaze shift in virtual reality and identified the coordination of eye, head, and body movements [74]. Emery et al. studied the process of performing various tasks, e.g. reading, drawing, shooting, and object manipulation, in virtual environments and identified general eye, hand, and head coordination patterns [14]. Recently, inspired by the strong link between eye and head movements, researchers started to use both eye and head information in many applications and have achieved great success [18, 31, 33, 44, 46, 76, 90]. For example, Gandrud et al. predicted users' locomotion directions in virtual reality using their gaze directions and head orientations [18]. Sidenmark et al. [75] and Kytö et al. [46] employed users' eye and head movements in virtual reality to improve the accuracy of target selection. Kothari et al. classified eye gaze events, i.e. fixations, pursuits, and saccades, from the magnitudes of eye and head movements [44]. Hu et al. predicted eye fixations in the future using

historical gaze positions and head rotation velocities [27] and proposed to recognise the task a user is performing from user's eye and head movements [28].

Despite the fact that both eye and head motions are beneficial for many applications, they have not been studied together for eye gaze imputation yet. And existing gaze prediction methods [27, 34] cannot be directly applied to gaze imputation, since they are only able to predict future gaze, but imputation requires a bi-directional method (e.g., filling in the previous gaze samples according to the observation). To the best of our knowledge, we are the first to demonstrate that eye-head coordination can be successfully transferred to the task of gaze imputation and lead to significant performance gains.

2.3 Time-Series Imputation

Since gaze data can be represented as a time series, in theory, existing time-series imputation methods can be directly applied to gaze data imputation. Deep learning-based methods have been shown to outperform statistical approaches in prior time-series imputation studies. Existing methods explore various neural network architectures, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Transformers, and Multilayer Perceptrons (MLPs). For example, TimesNet [89] transforms the input time series into the frequency domain and processes it using a CNN model; BRITS [11] employs a bidirectional RNN to capture temporal patterns in time series data; iTransformer [49], Informer [98], and Crossformer [96] utilise different attention mechanisms within Transformers to better model time-series dynamics; DLinear [92] applies MLPs for computationally efficient time-series imputation. Generative methods such as Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and diffusion models have also been explored for time-series imputation, including US-GAN [57], GP-VAE [17], and CSDI [82].

While these methods offer powerful general-purpose solutions, they are typically designed for single-modal signals and do not account for the unique characteristics of gaze data or the availability of complementary head movement signals in mobile eye tracking. In contrast, HAGI is specifically tailored to the gaze imputation task and is, to our knowledge, the first to explicitly incorporate head movement information as an auxiliary modality. Our approach is the first method that is specifically geared to the gaze imputation task and that leverages head movement information to enhance gaze imputation performance.

3 BACKGROUND

3.1 Gaze Data Imputation

Let $\mathbf{X} = \{x_1, x_2, ..., x_L\} \in \mathbb{R}^{L\times K}$ be a sequence of gaze directions, where L represents the length of the gaze sequence, determined by the sampling rate of the eye tracker and the duration of data collection. In our case, we set $K = 2$, as each gaze direction at any given time step is represented by (pitch, yaw). Additionally, we define $\mathbf{M} = \{m_1, m_2, ..., m_L\} \in \{0, 1\}^{L\times 1}$ to denote an observation mask, where $m_l = 1$ indicates that the eye tracker produces a valid output for x_l, while $m_l = 0$ denotes that x_l is invalid. Gaze data imputation is the task of estimating the gaze directions for invalid values within $\mathbf{X}$ by leveraging the valid gaze observations in $\mathbf{X}$.

3.2 Conditional Score-based Diffusion Model for Time-Series Imputation (CSDI)

Denoising Diffusion Probabilistic Models (DDPMs) [25] are a class of generative models that have achieved state-of-the-art performance across a range of domains, including image generation [25], audio synthesis [43], and more recently, eye movement synthesis [39, 40]. These models learn to generate realistic data by reversing a gradual noising process, enabling fine-grained control over the generative trajectory.

Tashiro et al. [82] extended the diffusion framework to time-series imputation by proposing Conditional Score-based Diffusion for Imputation (CSDI). CSDI models the conditional distribution of missing values given the observed portions of the time series. It introduces a conditional training scheme in which a masking function stochastically selects observed and unobserved regions of the input during training, allowing the model to learn flexible imputation strategies across various missing patterns.

Since gaze data is also time-series, CSDI can be adapted for gaze data imputation. CSDI is trained in a self-supervised manner. During training, given an input time series $\mathbf{x}_0$, CSDI randomly generates an observation mask that separates $\mathbf{x}_0$ into the observed part $\mathbf{x}_0^{co}$ and the target part requiring imputation $\mathbf{x}_0^{ta}$. As with the original DDPM, CSDI consists of two processes: The forward process is a Markov chain that progressively adds noise to $\mathbf{x}_0^{ta}$, to transform $\mathbf{x}_0^{ta}$ into random noise following a Gaussian distribution. The forward process is defined as follows:

$$q\left(\mathbf{x}_t^{ta} \mid \mathbf{x}_0^{ta}\right) = \mathcal{N}\left(\mathbf{x}_t^{ta}; \sqrt{\alpha_t}\mathbf{x}_0^{ta}, (1-\alpha_t)\, I\right) \tag{1}$$

where t denotes the time step, and α_t is a constant determined by a predefined noise schedule. More specifically,

$$\mathbf{x}_t^{ta} = \sqrt{\alpha_t}\mathbf{x}_0^{ta} + (1-\alpha_t)\epsilon \tag{2}$$

where $\epsilon \sim \mathcal{N}(0, I)$ is random Gaussian noise.

The reverse process aims to start from pure Gaussian noise, similar to $\mathbf{x}_t^{ta}$, and iteratively denoise it to reconstruct a sample resembling the original data distribution. Since, in imputation, we additionally have conditional information from the observed sequence $\mathbf{x}_0^{co}$, the reverse process is defined as follows:

$$p_\theta\left(\mathbf{x}_{t-1}^{ta} \mid \mathbf{x}_t^{ta}, \mathbf{x}_0^{co}\right) = \mathcal{N}\left(\mathbf{x}_{t-1}^{ta}; \mu_\theta(\mathbf{x}_t^{ta}, t \mid \mathbf{x}_0^{co}), \sigma_\theta(\mathbf{x}_t^{ta}, t \mid \mathbf{x}_0^{co})I\right) \tag{3}$$

where θ represents the trainable parameters of the neural network,

$$\mu_\theta(\mathbf{x}_t^{ta}, t \mid \mathbf{x}_0^{co}) = \frac{1}{\alpha_t}\left(\mathbf{x}_t^{ta} - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}}\epsilon_\theta\left(\mathbf{x}_t^{ta}, t \mid \mathbf{x}_0^{co}\right)\right), \tag{4}$$

and $\sigma_\theta(\mathbf{x}_t^{ta}, t \mid \mathbf{x}_0^{co})I$ is a constant defined by the noise schedule. The term $\epsilon_\theta\left(\mathbf{x}_t^{ta}, t \mid \mathbf{x}_0^{co}\right)$ in Equation 4 is a trainable denoising deep learning model. The training objective is to minimise the difference between the prediction and the actual added noise in Equation 2:

$$\mathcal{L}_{noise} = \left\|\epsilon_t - \epsilon_\theta\left(\mathbf{x}_t^{ta}, t \mid \mathbf{x}_0^{co}\right)\right\|_2 \tag{5}$$

4 HEAD-ASSISTED CONDITIONAL DIFFUSION MODEL FOR GAZE IMPUTATION

Previous studies have demonstrated that human eye movements are strongly correlated with head movements, a phenomenon known as eye-head coordination [14, 28, 34, 74, 75, 79]. Most commercially

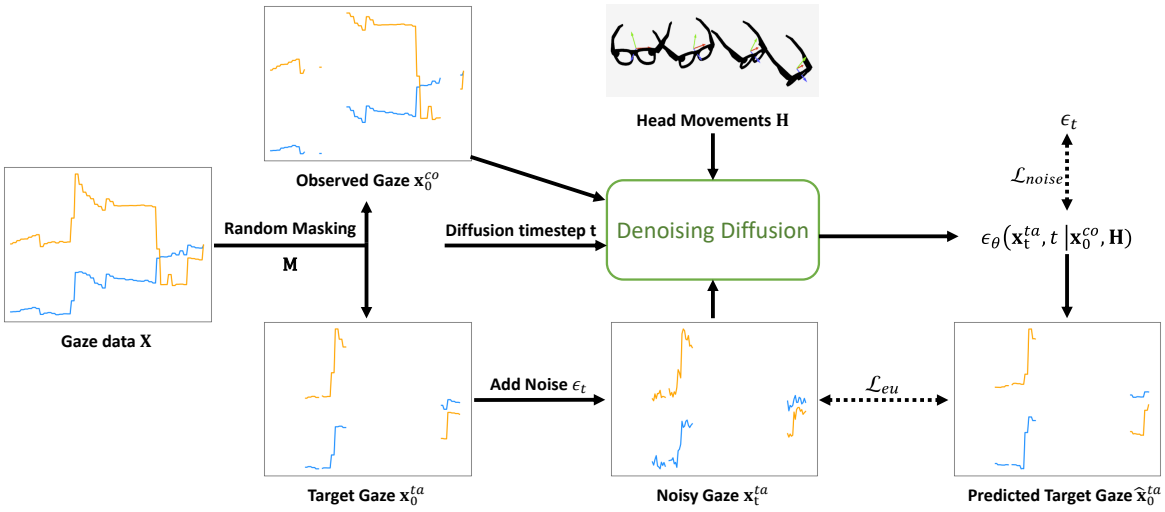

Figure 2: The training pipeline of HAGI. The detailed architecture of the denoising diffusion (Green box) is shown in Figure 3.

available head-mounted devices equipped with eye-tracking capabilities, such as the Apple Vision Pro [4], Meta Quest Pro [56], Microsoft HoloLens 2 [58], or Project Aria Glasses [15], are readily fitted with sensors for tracking users' head movements. Unlike gaze data, which inevitably contains missing values [22, 65], head movements are recorded continuously without any missing data, provided the sensor remains operational. The key idea of HAGI is to exploit the close coordination between head and eye movements to impute missing values in gaze data. At its core, HAGI uses a novel head-conditioned diffusion model with a tailored loss function for gaze imputation (see Section 4.2). Additionally, we introduce a dedicated head feature extractor that both encodes head rotation and translation, and a hybrid fusion mechanism to effectively merge the head and gaze features at different levels (see Section 4.3), enhancing the accuracy and realism of gaze imputation.

4.1 Problem Definition and Data Preparation

We extend the definition of gaze data imputation outlined in Section 3.1. For each gaze sequence $\mathbf{X} = \{x_1, x_2, ..., x_L\}$, we have a corresponding sequence of time-aligned head movements $\mathbf{H} = \{h_1, h_2, ..., h_L\}$. The task is to predict the gaze directions for missing values within $\mathbf{X}$ by leveraging both the head movements $\mathbf{H}$ and the existing gaze observations in $\mathbf{X}$. Since head movements correspond to the movements of head-mounted devices, we obtain $\mathbf{H}$ by processing the SLAM poses of the head-mounted device. Let us denote $T^l_{\text{world, tracker}} \in SE(3)$ as the pose of the mobile eye tracker in the world coordinate system at time step l. The head movement at time step l is represented as the relative transformation matrix between the SLAM poses at two consecutive time steps:

$$h_l = \Delta T^{l,l+1}_{\text{tracker}} = (T^l_{\text{world, tracker}})^{-1} T^{l+1}_{\text{world, tracker}} \tag{6}$$

Additionally, for each gaze sample x_l, represented as (pitch, yaw), we normalise the values to the range $[0, 1]$ using a sine transformation before further processing.

4.2 HAGI Diffusion Process

At the heart of HAGI is a conditional diffusion model that performs gaze imputation by leveraging the strong correlation between eye and head movements. While diffusion models have shown promise in general time-series imputation tasks (Section 3.2), they have not been adapted for multi-modal signals exhibiting biomechanical coordination such as gaze and head motion.

To this end, we design a head-conditioned diffusion framework that integrates head movement signals throughout the denoising process. Let $\mathbf{x}_0$ be the input gaze sequence, with $\mathbf{x}_0^{co}$ and $\mathbf{x}_0^{ta}$ representing the observed and target parts, respectively, separated by a randomly generated binary observation mask. The forward process of HAGI remains identical to that of CSDI [82]. Given that $\mathbf{H}$ represents the head movements corresponding to $\mathbf{x}_0$, the reverse process of HAGI extends Equation 3 as follows:

$$p_\theta\left(\mathbf{x}_{t-1}^{ta} \mid \mathbf{x}_t^{ta}, \mathbf{x}_0^{co}, \mathbf{H}\right) = \mathcal{N}\left(\mathbf{x}_{t-1}^{ta}; \mu_\theta\left(\mathbf{x}_t^{ta}, t \mid \mathbf{x}_0^{co}, \mathbf{H}\right), \sigma_\theta\left(\mathbf{x}_t^{ta}, t \mid \mathbf{x}_0^{co}, \mathbf{H}\right) I\right) \tag{7}$$

As with standard denoising diffusion models [25, 35, 43, 73, 82, 93], one of the training objectives of HAGI is to accurately predict the added noise at diffusion time step t:

$$\mathcal{L}_{noise} = \left\| \epsilon_t - \epsilon_\theta\left(\mathbf{x}_t^{ta}, t \mid \mathbf{x}_0^{co}, \mathbf{H}\right) \right\|_2 \tag{8}$$

Since we transform gaze data into sine space for normalisation, the model does not understand how gaze moves in Euclidean space. Therefore, we introduce an additional loss function, $\mathcal{L}_{eu}$, that minimises the mean squared error between the imputed gaze $\hat{\mathbf{x}}_0^{ta}$ at diffusion time step t and the ground truth gaze target $\mathbf{x}_0^{ta}$ in Euclidean space. More specifically,

$$\hat{\mathbf{x}}_0^{ta} = \frac{\mathbf{x}_t^{ta} - \sqrt{1-\alpha_t}\theta\left(\mathbf{x}_t^{ta}, t \mid \mathbf{x}_0^{co}, \mathbf{H}\right)}{\sqrt{\alpha_t}} \tag{9}$$

$$\mathcal{L}_{eu} = \left\| \arcsin(\mathbf{x}_0^{ta}) - \arcsin(\hat{\mathbf{x}}_0^{ta}) \right\|_2 \tag{10}$$

Algorithm 1 HAGI Training Procedure

Input: gaze data $\mathbf{X}$, corresponding head movements $\mathbf{H}$, total diffusion step T
repeat
  $\mathbf{x}_0 = \sin(\mathbf{X})$
  $\mathbf{M} \sim$ Random Mask Generator
  $\mathbf{x}_0^{co} = \mathbf{M} \odot \mathbf{x}_0$
  $\mathbf{x}_0^{ta} = (1 - \mathbf{M}) \odot \mathbf{x}_0$
  $t \sim \text{Uniform}(\{1, ..., T\})$
  $\epsilon \sim \mathcal{N}(0, I)$
  $\mathbf{x}_t^{ta} = \sqrt{\alpha_t}\mathbf{x}_0^{ta} + (1 - \alpha_t)\epsilon$
  $\mathcal{L}_{noise} = \left\| \epsilon_t - \epsilon_\theta \left(\mathbf{x}_t^{ta}, t \mid \mathbf{x}_0^{co}, \mathbf{H} \right) \right\|_2$
  $\hat{\mathbf{x}_0^{ta}} = \frac{\mathbf{x}_t^{ta} - \sqrt{1-\alpha_t}\theta(\mathbf{x}_t^{ta}, t | \mathbf{x}_0^{co}, \mathbf{H})}{\sqrt{\alpha_t}}$
  $\mathcal{L}_{eu} = \left\| \arcsin(\mathbf{x}_0^{ta}) - \arcsin(\hat{\mathbf{x}_0^{ta}}) \right\|_2$
  Take gradient descent step on
  $\nabla_\theta(\mathcal{L}_{noise} + \lambda \mathcal{L}_{eu})$
until converged

The final training objective of HAGI is given by:

$$\mathcal{L} = \mathcal{L}_{noise} + \lambda \mathcal{L}_{eu} \tag{11}$$

Figure 2 shows the HAGI training pipeline. Algorithm 1 and Algorithm 2 summarise HAGI's training and sampling procedures.

Algorithm 2 HAGI Sampling Procedure

Input: gaze data $\mathbf{X}$, corresponding head movements $\mathbf{H}$, observation mask $\mathbf{M}$
$\mathbf{x}_0^{co} = \mathbf{M} \odot \sin(\mathbf{X})$
Sample $\mathbf{x}_T^{ta} \sim \mathcal{N}(0, I)$
for $t = T, T-1, \ldots, 1$ **do**
  $\mu_\theta(\mathbf{x}_t^{ta}, t \mid \mathbf{x}_0^{co}, \mathbf{H}) = \frac{1}{\alpha_t} \left(\mathbf{x}_t^{ta} - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}} \epsilon_\theta \left(\mathbf{x}_t^{ta}, t \mid \mathbf{x}_0^{co}, \mathbf{H} \right) \right)$
  Sample $\mathbf{x}_{t-1}^{ta} \sim p_\theta \left(\mathbf{x}_{t-1}^{ta} \mid \mathbf{x}_t^{ta}, \mathbf{x}_0^{co}, \mathbf{H} \right)$ using Equation 7
end for
Output: $\arcsin(\mathbf{x}_0^{ta})$

4.3 HAGI Architecture

Figure 3 shows an overview of the HAGI architecture which is represented as the green denoising diffusion box in Figure 2. The primary challenge in designing the model architecture is to effectively integrate information from head movements $\mathbf{H}$ with the observed gaze sequence $\mathbf{x}_0^{co}$. To address this, we propose a head feature extractor that encodes head rotation and translation (green box in Figure 3) and a novel hybrid fusion mechanism, incorporating both early and skip fusion, to merge head and gaze features. This design enables HAGI to better capture the correlation between head and gaze information.

As outlined in Sections 3.1 and 4.1, the input gaze data is represented as a sequence of (pitch, yaw) with shape $(L, 2)$, where L denotes the sequence length. The input gaze data is then divided into the observed gaze $\mathbf{x}_0^{co} \in \mathbb{R}^{L\times 2}$ and the noisy target part $\mathbf{x}_t^{ta} \in \mathbb{R}^{L\times 2}$ using a randomly generated mask and noise addition, as described in Algorithm 1. Following prior work [40, 82], we concatenate the gaze-related inputs $\mathbf{x}_0^{co}$ and $\mathbf{x}_t^{ta}$ along a new dimension. The concatenated input, now of shape $(2, L, 2)$, is subsequently passed through a 1D convolutional layer with a kernel size of one to upsample it into an intermediate gaze feature of shape $(C, L, 2)$, where C denotes the number of channels.

Head feature extractor. The input head movements $\mathbf{H} \in \mathbb{R}^{L\times 4\times 3}$ consist of a sequence of transformations $h_l = [R, \mathbf{t}] \in \mathbb{R}^{4\times 3}$, where $R \in \mathbb{R}^{3\times 3}$ denotes head rotation and $\mathbf{t} \in \mathbb{R}^{1\times 3}$ refers to head translation. Following common practice in processing transformation matrices [48, 91], we first flatten each transformation into a 12-dimensional vector and apply Fourier encoding, a frequency-based positional encoding, to the flattened $\mathbf{H} \in \mathbb{R}^{L\times 12}$. The encoded features are then passed through a fully connected layer with two output features, followed by a 1D convolutional layer with a kernel size of one, reshaping the head features to match the intermediate gaze feature shape $(C, L, 2)$.

Hybrid Fusion Mechanism. Similar to popular diffusion model architectures [40, 43, 72, 82, 93], HAGI consists of N residual layers with $2C$ residual channels. We use **early fusion** by concatenating the gaze and head features along the channel dimension. The fused feature of shape $(2C, L, 2)$ is the input to the first residual layer. Each residual layer consists of a temporal Transformer that applies self-attention along the time dimension L and a spatial Transformer that applies self-attention along the feature dimension $K = 2$. This dual-transformer design has been widely adopted in time-series imputation tasks [62, 82]. However, unlike prior works focusing on single-modal time-series data, our approach captures spatial and temporal correlations in-between and across head and gaze.

Since the head feature is concatenated with the gaze feature and fed into the first residual layer, information from head movements may degrade as it propagates through multiple residual layers due to operations such as nonlinear activation functions. To mitigate this, inspired by the skip connections in ResNet [24], we introduce a *skip fusion* mechanism that integrates the head features extracted by our head feature extractor into each residual layer. Specifically, we concatenate the head feature with other commonly used side information in time-series analysis, including time embedding [39, 82, 85, 99] which is obtained by applying standard positional encoding along the time-axis L, feature embedding [39, 82] that is a categorical feature embedding for the feature-axis (pitch, yaw), and the observation mask on the channel dimension [2, 82].

A 1D convolutional layer in each residual layer maps this combined information to a shape of $(4C, L, 2)$, while another 1D convolutional layer maps the Transformer output to the same shape. These two representations are fused via element-wise addition. Following [40, 43, 82], the fused tensor is then passed through a gated activation unit and two 1D convolutional layers to produce the input for the next residual layer and the skip connections. With shape $(2C, L, 2)$, the final residual layer output is projected by two 1D convolutional layers into the predicted noise, maintaining the same shape $(L, 2)$ as the input gaze data. All convolutional layers in HAGI are used for up- and down-sampling tensors at the channel dimension; we set the kernel size to one for all these layers.

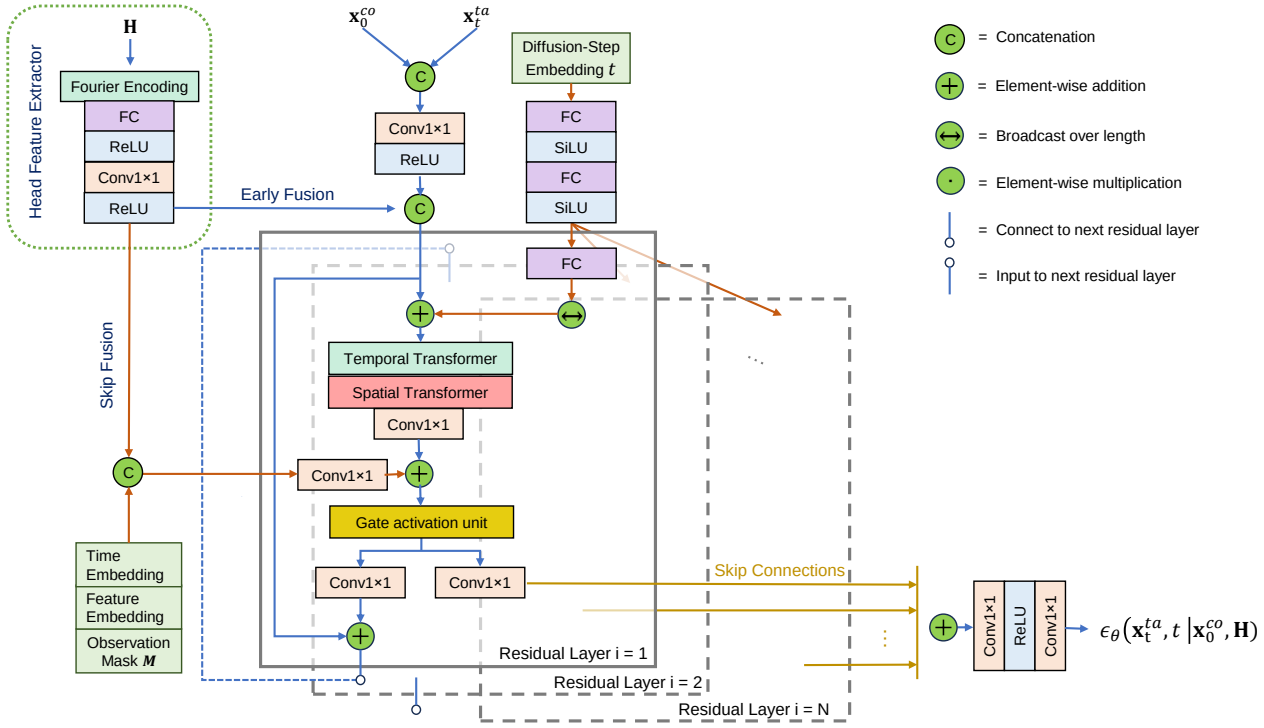

Figure 3: Overview of HAGI architecture. The model takes the observed gaze sequence $\mathbf{x}_0^{co}$ and head movement sequence H as input. A head feature extractor (green box) encodes H into a latent representation. Gaze and head features are then fused using a novel hybrid fusion mechanism that combines early fusion (concatenation before residual blocks) and skip fusion (via residual connections). The fused representation is processed by a denoising diffusion model consisting of residual blocks with spatial and temporal Transformers. This design enables HAGI to capture eye-head coordination for accurate and realistic gaze imputation.

We follow prior work [39, 40, 43, 82] to employ positional encoding to transform the diffusion step t into a 128-dimensional embedding. This embedding is passed through three fully connected layers and summed with the fused head and gaze tensor before being processed by the Transformers in each residual layer.

5 EXPERIMENTS

5.1 Datasets

To assess the performance of HAGI across diverse real-world scenarios, we evaluated our method on three publicly available, large-scale datasets that include mobile eye-tracking and head movement recordings captured during everyday activities:

Nymeria [54] is the world's largest in-the-wild human motion dataset, featuring over 300 hours of multimodal recordings from 264 participants engaged in everyday activities across 50 distinct indoor and outdoor environments. The dataset includes 30 Hz gaze data, head motion, wrist motion, body motion, and natural language descriptions. Due to its large scale and coverage of varied real-world settings, we used Nymeria for training and evaluation. Specifically, we selected recordings with well-calibrated personalised gaze data, SLAM poses from the eye tracker. We randomly partitioned these recordings into training (80%, 593 recordings, 145.8 hours), validation (5%, 38 recordings, 8.4 hours), and test (15%, 111 recordings, 28.7 hours) sets.

Ego-Exo4D [19] is a large-scale multimodal dataset that provides synchronised egocentric and exocentric recordings of skilled human activities, covering activities such as cooking, football, music, dance, basketball, bicycle repair, and rock climbing. The egocentric recordings contain 30 Hz gaze data and SLAM-derived head poses from head-mounted devices. To assess the generalisability of HAGI across diverse indoor and outdoor activities, we used all 72 egocentric recordings (4.6 hours) with well-calibrated personalised gaze data for cross-dataset evaluation.

HOT3D [7] is an egocentric multimodal dataset for studying hand-object interactions in indoor environments. Similar to the other two datasets, it includes gaze data and head motion recordings. Since handling tools and interacting with objects are fundamental aspects of everyday activities, we incorporated HOT3D for cross-dataset evaluation. Our evaluation used all 111 (3.48 hours), containing well-calibrated personalised gaze data and SLAM poses.

5.2 Evaluation Settings

Baselines. We compared HAGI against several baseline approaches. These include widely used interpolation methods for gaze imputation as well as state-of-the-art deep learning-based time-series imputation techniques.

- **Head direction**: Given the strong correlation between head and eye movements, head direction was widely used as a proxy for gaze direction in prior work [30, 32, 61, 78].
- **Linear interpolation** and **Nearest interpolation**: These two interpolation techniques are the most commonly used methods for handling missing values in eye-tracking research [22, 41].
- **iTransformer** [49]: A Transformer-based model designed for multiple time-series tasks, including time-series imputation.
- **DLinear** [92]: A multilayer perceptron (MLP)-based method that models time-series trends using a moving average kernel and a seasonal component.
- **TimesNet** [89]: Converts 1D time-series data into 2D tensors with Fast Fourier Transformation and processes them using a CNN-based architecture. It was designed for multi-time-series tasks.
- **BRITS** [11]: A bidirectional RNN-based approach for time-series imputation.
- **CSDI** [82]: A diffusion-based generative model for time-series imputation, demonstrating superior performance over GAN-based [57] and VAE-based [17] methods in various imputation tasks.

Input duration. We chose five seconds as the duration for all inputs, same as prior gaze-based deep learning models [40, 50]. We clipped the gaze recordings in the Nymeria dataset into non-overlapped five-second segments according to the time stamps in the atomic motion descriptions. For the Ego-Exo4D and HOT3D datasets, we ignored the data within the very first and last second in each recording to ensure the data quality and clipped the recording into non-overlapping five-second segments. Following prior work [50], all segments with more than 5% invalid gaze samples were discarded. To ensure the model only learns to reconstruct real gaze movements, all the remaining invalid gaze samples were excluded from the loss computation during training.

Data loss ratio. As in general time-series imputation, we masked a certain proportion of valid gaze data for evaluating imputation performance. The selected data loss ratios were informed by blink duration statistics and missing data ratios reported in prior research: Blinks occur approximately 20 times per minute, with each blink lasting between 150–450 milliseconds [64, 81]. Consequently, up to 10% missing data can be attributed solely to blinks [65]. Additionally, prior studies have reported missing data ratios ranging from 20% to 60% [26, 66, 71]. Moreover, an empirical research shows that consecutive missing values last 1,325 milliseconds on average with the standard deviation of 4,076 milliseconds, and the majority of the missing segments are shorter than 1 second [21]. Based on these findings, we evaluated HAGI under missing data conditions of 10%, 30%, and 50%. Furthermore, to assess HAGI's robustness in extreme scenarios, we also tested it with 90% missing data. Since the longest blink duration (450 ms) corresponds to approximately 10% of our five-second input window, we primarily evaluated HAGI on long blinks using a 10% missing ratio. Specifically, for each five-second gaze trajectory in our test sets, we randomly masked a continuous 10% segment of valid data. For 30%, 50%, and 90% missing ratios, we simulated real-world data loss by ensuring that each masked segment lasted at least 150 milliseconds, corresponding to the shortest blink duration. This strategy ensured that the missing data patterns realistically reflected natural gaze data loss.

Evaluation metrics. We used two metrics for evaluation:

- Mean Angular Error (MAE): MAE is the most commonly used evaluation metric in previous gaze estimation research [30, 32, 34, 94, 95]. It measures the angular difference (in degrees) between the predicted and the ground truth gaze vectors. Specifically, we first converted the gaze direction from the unit spherical coordinate system (pitch, yaw) to a 3D Cartesian vector (x, y, z) and then computed the MAE as follows:

$$\text{MAE} = \frac{1}{J}\sum_{j=1}^{J} \arccos\left(\frac{g_j \cdot \hat{g_j}}{|g_j||\hat{g_j}|}\right) \tag{12}$$

  where J is the total number of missing frames, and g_j and $\hat{g_j}$ represent the ground truth gaze vector and the predicted gaze vector, respectively.
- Jensen–Shannon divergence (JS): While MAE evaluates the accuracy of gaze direction reconstruction, it does not assess whether the imputed gaze movements exhibit realistic human gaze dynamics. To address this, we incorporated JS divergence, following prior work on eye movement synthesis [40, 67, 68]. JS divergence measures the similarity between the velocity distribution of imputed and real human gaze movements. Let P and Q be the distributions of the predicted and ground truth gaze velocities at missing frames, respectively. JS divergence is defined as:

$$\text{JS}(P||Q) = \frac{1}{2}\text{KL}\left(P\middle\|\frac{1}{2}(P+Q)\right) + \frac{1}{2}\text{KL}\left(Q\middle\|\frac{1}{2}(P+Q)\right), \tag{13}$$

  where KL denotes the Kullback–Leibler (KL) divergence. JS divergence ranges between zero and one, with lower values indicating better performance.

A good gaze imputation method should achieve lower JS on the basis of lower MAE.

Implementation details. All datasets used in our experiments provide gaze data at 30 Hz, and we set the input duration to five seconds, resulting in a total sequence length of $L = 150$. For HAGI, we set the residual layers to $N = 4$, with $C = 64$ channels and eight attention heads per Transformer. The diffusion process consisted of $T = 50$ steps, using a cosine noise schedule with a minimum noise level of $1 - \alpha_1 = 10^{-4}$ and a maximum noise level of $1 - \alpha_T = 0.5$. We set $\lambda = 1.5$ for $\mathcal{L}_{eu}$. The training was conducted for 500 epochs with a batch size of 256, using the Adam optimizer with an initial learning rate of 10^{-3}, which decayed to 10^{-4} at epoch 375 and further to 10^{-5} at epoch 450.

Since CSDI [82] can be considered an ablated version of HAGI without head movement information, we ensured a fair comparison by training CSDI with the same hyperparameters as HAGI using its official implementation. For other deep learning-based time-series imputation baselines, we leveraged PyPOTS [12], a popular Python toolbox that includes various time-series imputation methods [13]. We modified only the input shape while keeping

Nymeria [54]	Data loss ratio			
	10% (Long Blinks)	30%	50%	90%
Head direction	23.32	23.43	23.44	23.44
Linear	4.96	6.88	9.68	11.54
Nearest	5.29	6.52	8.34	12.61
iTransformer [49]	8.75	11.10	16.57	24.05
DLinear [92]	12.12	12.25	12.97	14.01
TimesNet [89]	19.53	18.59	20.32	22.91
BRITS [11]	10.25	11.98	14.06	17.58
CSDI [82]	4.72	5.90	7.44	10.54
HAGI (Ours)	**3.67**	**4.55**	**5.77**	**8.53**

Table 1: Mean angular error (MAE) of gaze imputation across different methods and data loss ratios on the Nymeria [54] test set. The best results are marked in bold, and the second-best are underlined.

Nymeria [54]	Data loss ratio			
	10%	30%	50%	90%
Head direction	0.139	0.137	0.139	0.146
Linear	0.129	0.078	0.089	0.150
Nearest	0.081	0.073	0.103	0.135
CSDI [82]	0.044	0.042	0.037	0.030
HAGI (Ours)	**0.042**	**0.040**	**0.035**	**0.017**

Table 2: Jensen–Shannon divergence (JS) of gaze imputation across different methods and data loss ratios on the Nymeria [54] test set. The best results are marked in bold, and the second-best are underlined.

the optimal hyperparameters provided for the PhysioNet2012 [77] dataset and trained all models with the same number of epochs and batch size. All deep learning-based methods were trained on the Nymeria training set, and we selected the best-performing model on the validation set for the final evaluation. For classical interpolation methods, we used the built-in functions from the SciPy library [86]. As a baseline head direction proxy, we filled in missing frames with $(\text{pitch}, \text{yaw}) = (0, 0)$. Since HAGI and CSDI are generative models and do not produce deterministic outputs, we followed [2, 82] and used the median of 100 generated samples for evaluation.

5.3 Gaze Imputation Results

Quantitative results. Table 1 shows the mean angular error (MAE) of gaze imputation across different methods and missing ratios for a within-dataset evaluation on the Nymeria test set. As can be seen from the table, HAGI outperforms all baselines across all missing ratios, achieving improvements of 22% on long blinks (3.67° vs. 4.72°), 23% on 30% missing data (4.55° vs. 5.90°), 22% on 50% missing data (5.77° vs. 7.44°), and 19% on 90% missing data (8.53° vs. 10.54°) compared with the second-best method. In contrast, except for CSDI, time-series imputation methods did not achieve performance comparable to traditional interpolation methods. This suggests that these time-series imputation methods cannot be directly adapted to gaze data without modification. Consequently, we excluded these methods from further evaluations (see Appendix A if interested).

Table 2 shows the Jensen–Shannon divergence (JS) of gaze imputation across different methods and data loss ratios, also on the Nymeria test set. The velocity distribution of HAGI-imputed gaze achieved the lowest JS across all experimental settings, indicating its superior ability to preserve natural gaze dynamics. In contrast, traditional interpolation methods perform substantially worse in replicating human-like eye movements than diffusion-based approaches.

We then conducted a cross-dataset evaluation on the Ego-Exo4D and HOT3D datasets to assess whether HAGI can generalise to diverse everyday settings. The results are shown in Table 3. Similar to the results on the Nymeria dataset, despite HAGI not being trained or fine-tuned on these datasets, it achieved the lowest MAE across all methods and data loss ratios, with a minimum improvement of 11%, a maximum improvement of 21%, and an average improvement of 18% over the second-best method. Furthermore, HAGI also attained the lowest JS among all methods, suggesting that its imputed gaze velocity distribution is the closest to the real human gaze velocity distribution among the baselines. Moreover, it is worth noting that in the cross-dataset evaluation, the absolute MAE across different missing ratios did not decrease compared with the within-dataset results on the Nymeria dataset (see Table 1). This suggests that HAGI learns robust correlations between eye and head movements that generalise well across datasets featuring diverse everyday activities and environments, demonstrating strong generalisation performance and reflecting the benefits of training on a larger and more diverse dataset.

Qualitative results. We present four sample gaze imputation results from the cross-dataset evaluation for four methods with relatively low MAE in Figure 4. As shown in the figure, HAGI benefits from incorporating head movement information, resulting in imputed gaze trajectories that follow a similar trend and are spatially closer to the ground truth human eye movements (see the results of 30% and 50% missing ratios in Figure 4). In the scenario with 90% missing values, HAGI imputed gaze samples are more closely aligned with the ground truth. This suggests that leveraging head movements provides valuable contextual information for gaze imputation, enabling HAGI to generate more naturalistic and human-like gaze trajectories compared to baseline methods, aligning with results of MAE in Table 3. In contrast, the gaze trajectories imputed by traditional interpolation methods are visually dissimilar to real human eye movements. This visual assessment aligns with the JS values reported in Table 3. Although CSDI achieved a JS score comparable to HAGI in Table 3, its imputed gaze data exhibits substantial spatial deviation from real human eye movements.

5.4 Ablation Study

Head rotation and translation. The evaluation results so far show that incorporating head information enables HAGI to achieve superior performance over single-modal baseline approaches. However, as discussed in Section 4.3, the input head movements comprise

	Mean angular error (MAE)								Jensen–Shannon divergence (JS)							
Dataset	Ego-Exo4D				HOT3D				Ego-Exo4D				HOT3D			
Data loss ratio	10%	30%	50%	90%	10%	30%	50%	90%	10%	30%	50%	90%	10%	30%	50%	90%
Head direction	25.82	25.76	25.88	25.82	23.63	23.81	23.70	23.73	0.126	0.125	0.124	0.131	0.148	0.148	0.147	0.145
Linear	3.92	5.54	7.90	9.29	3.87	5.64	7.80	8.98	0.094	0.085	0.073	0.135	0.277	0.102	0.096	0.148
Nearest	4.18	5.18	6.68	10.13	4.14	5.23	6.61	9.86	0.062	0.066	0.081	0.121	0.080	0.081	0.100	0.135
CSDI [82]	3.78	4.78	6.07	8.91	3.67	4.69	5.85	8.28	0.042	**0.041**	0.033	0.026	0.052	0.051	0.042	0.029
HAGI (Ours)	**3.06**	**3.80**	**4.86**	**7.37**	**3.01**	**3.91**	**4.99**	**7.34**	**0.040**	**0.041**	**0.033**	**0.024**	**0.049**	**0.050**	**0.040**	**0.021**

Table 3: Cross-dataset evaluation results: Mean angular error (MAE) and Jensen–Shannon divergence (JS) of gaze imputation across different methods and data loss ratios on the Ego-Exo4D [19] and HOT3D [7] datasets. The best results are marked in bold, and the second-best are underlined.

Figure 4: Four examples of gaze imputation results at different missing ratios (10%, 30%, 50%, 90%) using different methods in the cross-dataset evaluation. The bottom row shows the visualisations of ground truth human eye movements.

rotation and translation. It remains unclear how head rotation and head translation separately contribute to gaze imputation performance. To gain further insight into this question, we trained two ablated versions of HAGI: one that receives only head rotation matrices as input and another that receives only head translation vectors.

The results are shown in Table 4. Since prior work [39, 40] and our previous findings demonstrated that diffusion-based approaches effectively model gaze velocity distributions, we report only the MAE in our results. Compared to CSDI, which does not use head information, HAGI trained with only head rotation and HAGI trained with only head translation achieved lower MAE across all settings and datasets. This suggests that both head rotation and translation are correlated with human eye gaze and are both important to achieve performance improvements. Furthermore, HAGI rotation consistently outperforms HAGI translation, suggesting that head rotation contributes more to gaze imputation performance than head translation. The full HAGI, trained with the complete head transformation matrices, outperforms both ablated versions, demonstrating its ability to leverage the full range of head motion for enhanced gaze imputation accuracy.

HAGI components. We finally conducted an ablation study to show the effectiveness of our different design choices for HAGI. Table 5 presents the MAE results for ablated versions of our method across different data loss ratios on the Nymeria, Ego-Exo4D, and HOT3D datasets. Early fusion is the most naive approach to integrating head information extracted from the head feature extractor. As can be seen from the table, it consistently surpasses CSDI, a model without the head feature extractor, across all evaluation settings, indicating the effectiveness of the proposed head feature extractor. However, HAGI with early fusion only consistently achieved the highest MAE among all its ablated versions across all settings. The proposed skip fusion mechanism reduces the MAE, achieving an additional 5% improvement on average compared with the version using only early fusion. Furthermore, with the proposed loss function $\mathcal{L}_{eu}$, the full HAGI consistently outperformed the version employing only the hybrid fusion mechanism across all settings, with an average improvement of 2%. These results underline the importance of the proposed components within HAGI for achieving the reported gaze imputation performance.

6 DISCUSSION

6.1 Performance

What all of our evaluations show is that HAGI not only achieves superior MAE reductions but also generates gaze trajectories that more faithfully resemble natural human eye movements, highlighting its robustness and effectiveness in real-world applications. Our method consistently outperforms baseline approaches both quantitatively (see Tables 1, 2, and 3) and qualitatively (see Figure 4) for all considered data loss ratios and datasets. In terms of mean angular error (MAE), HAGI reduces MAE by an average of 19.34% compared with the previous state-of-the-art method. Notably, Table 1 and Table 3 show that HAGI's MAE on 30% missing values is comparable to the MAEs of baseline methods on only 10% missing data, and its MAE on 50% missing values is similar to the MAEs of baseline methods on 30% data loss. This suggests that in real-world scenarios, HAGI can impute gaze data with an additional 20% missing values while maintaining existing methods' performance level. This finding is significant as it shows not only the effectiveness of our method on benchmarks but also the concrete benefits it provides for practical mobile eye-tracking applications.

Unlike MAE, which tends to increase with higher levels of missing data, the Jensen-Shannon (JS) divergence of HAGI *decreases* as the missing ratio increases. We compute gaze velocity distributions using the `numpy.histogram` function with `bins=100` across all settings. However, the resulting absolute JS values are not directly comparable across missing ratios or datasets. For example, at 90% missingness, the ground-truth velocity distribution is considerably broader than that at 10%, with the latter effectively forming a subset of the former. Applying the same number of bins to both leads to differing bin widths, which in turn affects the scale of JS divergence. This makes direct cross-ratio or cross-dataset comparisons of JS scores inappropriate. Nonetheless, within each dataset and missing ratio, comparisons across methods remain valid. HAGI consistently achieves the lowest JS divergence in all such settings, indicating that it generates gaze trajectories with the most plausible velocity distributions among all evaluated methods.

As shown in Table 1, existing time-series imputation models that are not based on diffusion perform worse than even standard interpolation techniques when applied to gaze data. This supports prior findings [39] that diffusion models are better suited to modelling gaze velocity distributions. Non-diffusion methods often overfit to slow eye movements, such as fixations, which dominate real-world gaze recordings. While augmenting non-diffusion baselines with head movement information is technically feasible, it is unlikely to yield significant improvements, as the underlying distribution of gaze velocities remains unchanged. We therefore focus on diffusion-based approaches and demonstrate that integrating head input within this framework yields further improvements in performance.

6.2 Head Rotation vs. Translation

To investigate the independent contributions of head rotation and head translation to gaze imputation, we conducted an ablation study. The results indicate that gaze imputation benefits from head rotation and translation; however, head rotation exhibits a stronger correlation with human gaze than head translation (Table 4). This finding aligns well with prior research on vestibular function testing. In particular, dynamic visual acuity (DVA) is primarily governed by the vestibulo-ocular reflex (VOR), which stabilises gaze during head movements [70]. Ramaioli et al. [70] demonstrated that DVA is consistently lower during head translations (tVOR) than during head rotations (rVOR), further supporting our conclusion that head rotation plays a more dominant role in eye-head coordination.

6.3 Generalisability and Application

As demonstrated in Section 5.3, HAGI's performance remained consistent in cross-dataset evaluations, showing no degradation in MAE compared to within-dataset evaluations. This suggests that HAGI can be directly applied in real-world scenarios without requiring retraining or fine-tuning, provided that the SLAM poses of the mobile eye tracker are available. Furthermore, our results indicate that even when using only head rotation or head translation, HAGI still outperforms existing approaches. This highlights its adaptability; for mobile eye trackers lacking a SLAM system, HAGI can utilise rotation data from IMUs to achieve robust gaze

	Head movements		Nymeria [54]				Ego-Exo4D [19]				HOT3D [7]			
	Rotation	Translation	10%	30%	50%	90%	10%	30%	50%	90%	10%	30%	50%	90%
CSDI [82]	✗	✗	4.72	5.90	7.44	10.54	3.78	4.78	6.07	8.91	3.67	4.69	5.85	8.28
HAGI (Ours)	✗	✓	4.66	5.75	7.29	10.47	3.72	4.63	5.91	8.79	3.60	4.57	5.74	8.19
	✓	✗	4.41	5.36	6.66	9.56	3.55	4.35	5.43	8.23	3.42	4.33	5.38	7.79
	✓	✓	**3.67**	**4.55**	**5.77**	**8.53**	**3.06**	**3.80**	**4.86**	**7.37**	**3.01**	**3.91**	**4.99**	**7.34**

Table 4: The results (mean angular error) of the ablation study on head rotation and translation across different data loss ratios in the Nymeria [54], Ego-Exo4D [19], and HOT3D [7] datasets. The best results are marked in bold, and the second-best are underlined.

	HAGI components			Nymeria [54]				Ego-Exo4D [19]				HOT3D [7]			
	Early Fusion	Skip Fusion	$\mathcal{L}_{eu}$	10%	30%	50%	90%	10%	30%	50%	90%	10%	30%	50%	90%
CSDI [82]	✗	✗	✗	4.72	5.90	7.44	10.54	3.78	4.78	6.07	8.91	3.67	4.69	5.85	8.28
HAGI (Ours)	✓	✗	✗	3.97	4.93	6.31	9.28	3.28	4.10	5.29	7.97	3.13	4.10	5.21	7.68
	✓	✓	✗	3.70	4.61	5.88	8.78	3.08	3.87	5.02	7.87	3.04	3.96	5.06	7.46
	✓	✓	✓	**3.67**	**4.55**	**5.77**	**8.53**	**3.06**	**3.80**	**4.86**	**7.37**	**3.01**	**3.91**	**4.99**	**7.34**

Table 5: The results (mean angular error) of the ablation study on different HAGI components across different data loss ratios in the Nymeria [54], Ego-Exo4D [19], and HOT3D [7] datasets. The best results are marked in bold, and the second-best are underlined.

imputation. While our task definition assumes perfect time alignment between head and eye movements, the real-world dataset used in our experiments includes a natural temporal offset of approximately 20–50 milliseconds. The strong performance of HAGI under these conditions suggests that it is resilient to small delays between head and gaze signals, further supporting its practicality in real-world use.

HAGI can be employed in two primary ways. First, it can be a post-processing method, particularly beneficial for preparing gaze data for machine learning models. Since machine learning approaches cannot handle missing values directly, HAGI can impute these values in a human-like manner, enhancing data utilisation efficiency. Second, a key advantage of HAGI is its ability to impute gaze data at arbitrary locations within a 5-second time window. This means that after an initial 5-second recording period, any newly encountered missing values can be imputed directly with HAGI. This capability makes HAGI promising for gaze-based interactive systems. However, the current inference speed—approximately 2.9 seconds per 5-second segment on an NVIDIA V100 32 GB GPU—remains a limitation for real-time deployment. We plan to optimise the model's efficiency to support real-time gaze interaction, such as gaze-based selection, foveated rendering, or attention-aware interfaces, in future work.

6.4 Limitations and Future Work

HAGI is designed for gaze imputation in mobile eye tracking scenarios, and all evaluations in this work were conducted on egocentric datasets involving everyday human activities. While this setting reflects realistic use cases for head-mounted eye trackers, advanced stationary eye trackers—such as the Tobii Eye Tracker 5 [83]—also support head tracking and may benefit from our approach. In future work, we aim to investigate the adaptability of HAGI to such desktop setups.

While our quantitative and distributional results demonstrate that HAGI produces realistic gaze trajectories at 30 Hz, we have not evaluated its performance on higher-frequency gaze data. As modern mobile eye trackers typically provide head-tracking sensors at higher sampling rates, head features can, in principle, support higher-frequency gaze imputation. However, this would require retraining the model, and current public datasets with high-frequency gaze recordings are limited. We aim to investigate this direction in future work.

Additionally, although our task assumes synchronised head and eye data, our experiments used datasets where head and gaze signals have an inherent delay of around 20–50 ms. This suggests that HAGI is robust to small misalignments. In future work, we intend to assess its tolerance to longer delays, which are common in practice due to imperfect sensor synchronisation.

7 CONCLUSION

In this work, we introduced HAGI– a novel multi-modal diffusion-based approach for gaze imputation that leverages eye-head coordination to enhance the reconstruction of missing gaze data. We evaluated HAGI on three large-scale egocentric gaze datasets –Nymeria, Ego-Exo4D, and HOT3D – and demonstrated that it significantly outperforms traditional interpolation methods and state-of-the-art deep learning baselines. Specifically, HAGI achieves up to 22% improvement in mean angular error and generates more realistic gaze velocity distributions that closely match human eye movements.

These results highlight the effectiveness of integrating head movement information in gaze imputation tasks and establish HAGI as a robust solution for improving gaze data quality with significant potential for real-world applications.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon Europe research and innovation funding programme under grant agreement No. 101072410.

Funded by
the European Union

REFERENCES

[1] Yasmeen Abdrabou, Yomna Abdelrahman, Mohamed Khamis, and Florian Alt. 2021. Think harder! Investigating the effect of password strength on cognitive load during password creation. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.

[2] Juan Lopez Alcaraz and Nils Strodthoff. 2022. Diffusion-based Time Series Imputation and Forecasting with Structured State Space Models. *Transactions on Machine Learning Research* (2022). https://openreview.net/forum?id=hHiIbk7ApW

[3] Tobias Appel, Natalia Sevcenko, Franz Wortha, Katerina Tsarava, Korbinian Moeller, Manuel Ninaus, Enkelejda Kasneci, and Peter Gerjets. 2019. Predicting cognitive load in an emergency simulation based on behavioral and physiological measures. In *2019 International Conference on Multimodal Interaction*. 154–163.

[4] Apple Inc. 2023. *Apple Vision Pro*. Mixed Reality Headset. Available at: https://www.apple.com/apple-vision-pro/ (Accessed on: 21 March 2025).

[5] Claudio Aracena, Sebastián Basterrech, Václav Snáel, and Juan Velásquez. 2015. Neural networks for emotion recognition based on eye tracking data. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 2632–2637.

[6] Mohammed Safayet Arefin, J Edward Swan II, Russell A Cohen Hoffing, and Steven M Thurman. 2022. Estimating perceptual depth changes with eye vergence and interpupillary distance using an eye tracker in virtual reality. In *2022 Symposium on Eye Tracking Research and Applications*. 1–7.

[7] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. 2024. HOT3D: Hand and Object Tracking in 3D from Egocentric Multi-View Videos. *arXiv preprint arXiv:2411.19167* (2024).

[8] Esubalew Bekele, Zhi Zheng, Amy Swanson, Julie Crittendon, Zachary Warren, and Nilanjan Sarkar. 2013. Understanding how adolescents with autism respond to facial expressions in virtual reality environments. *IEEE transactions on visualization and computer graphics* 19, 4 (2013), 711–720.

[9] Pieter Blignaut and Daniël Wium. 2014. Eye-tracking data quality as affected by ethnicity and experimental design. *Behavior research methods* 46 (2014), 67–80.

[10] Andreas Bulling, Daniel Roggen, and Gerhard Tröster. 2008. It's in your eyes: Towards context-awareness and mobile HCI using wearable EOG goggles. In *Proceedings of the 10th international conference on Ubiquitous computing*. 84–93.

[11] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. 2018. Brits: Bidirectional recurrent imputation for time series. *Advances in neural information processing systems* 31 (2018).

[12] Wenjie Du. 2023. PyPOTS: a Python toolbox for data mining on Partially-Observed Time Series. *arXiv preprint arXiv:2305.18811* (2023).

[13] Wenjie Du, Jun Wang, Linglong Qian, Yiyuan Yang, Fanxing Liu, Zepu Wang, Zina Ibrahim, Haoxin Liu, Zhiyuan Zhao, Yingjie Zhou, Wenjia Wang, Kaize Ding, Yuxuan Liang, B. Aditya Prakash, and Qingsong Wen. 2024. TSI-Bench: Benchmarking Time Series Imputation. *arXiv preprint arXiv:2406.12747* (2024).

[14] Kara J Emery, Marina Zannoli, James Warren, Lei Xiao, and Sachin S Talathi. 2021. OpenNEEDS: A dataset of gaze, head, hand, and scene signals during exploration in open-ended vr environments. In *Proceedings of the 2021 ACM Symposium on Eye Tracking Research and Applications*. 1–7.

[15] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. 2023. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561* (2023).

[16] Yu Fang, Ryoichi Nakashima, Kazumichi Matsumiya, Ichiro Kuriki, and Satoshi Shioiri. 2015. Eye-head coordination for visual cognitive processing. *PloS One* 10, 3 (2015), e0121035.

[17] Vincent Fortuin, Dmitry Baranchuk, Gunnar Rätsch, and Stephan Mandt. 2020. Gp-vae: Deep probabilistic time series imputation. In *International conference on artificial intelligence and statistics*. PMLR, 1651–1661.

[18] Jonathan Gandrud and Victoria Interrante. 2016. Predicting destination using head orientation and gaze direction during locomotion in vr. In *Proceedings of the 2016 ACM Symposium on Applied Perception*. 31–38.

[19] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. 2024. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19383–19400.

[20] Jesse W Grootjen, Henrike Weingärtner, and Sven Mayer. 2023. Highlighting the Challenges of Blinks in Eye Tracking for Interactive Systems. In *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*. 1–7.

[21] Jesse W Grootjen, Henrike Weingärtner, and Sven Mayer. 2024. Investigating the Effects of Eye-Tracking Interpolation Methods on Model Performance of LSTM. In *Proceedings of the 2024 Symposium on Eye Tracking Research and Applications*. 1–6.

[22] Jesse W Grootjen, Henrike Weingärtner, and Sven Mayer. 2024. Uncovering and Addressing Blink-Related Challenges in Using Eye Tracking for Interactive Systems. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–23.

[23] Nishan Gunawardena, Michael Matscheko, Bernhard Anzengruber, Alois Ferscha, Martin Schobesberger, Andreas Shamiyeh, Bettina Klugsberger, and Peter Solleder. 2019. Assessing surgeons' skill level in laparoscopic cholecystectomy using eye metrics. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*. 1–8.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

[26] Kenneth Holmqvist, Marcus Nyström, and Fiona Mulvey. 2012. Eye tracker data quality: What it is and how to measure it. In *Proceedings of the symposium on eye tracking research and applications*. 45–52.

[27] 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 (TVCG)* 27, 5 (2021), 2681–2690. https://doi.org/10.1109/TVCG.2021.3067779

[28] Zhiming Hu, Andreas Bulling, Sheng Li, and Guoping Wang. 2023. EHTask: Recognizing User Tasks From Eye and Head Movements in Immersive Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics* 29, 4 (2023), 1992–2004. https://doi.org/10.1109/TVCG.2021.3138902

[29] 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. In *Proceedings of the 2025 ACM Special Interest Group on Computer Graphics and Interactive Techniques*.

[30] 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.

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

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

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

[34] 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.

[35] Zhiming Hu, Guanhua Zhang, Zheming Yin, Daniel Haeufle, Syn Schmitt, and Andreas Bulling. 2025. HaHeAE: Learning Generalisable Joint Representations of Human Hand and Head Movements in Extended Reality. *IEEE Transactions on Visualization and Computer Graphics* (2025).

[36] Michael Xuelin Huang and Andreas Bulling. 2019. SacCalib: reducing calibration distortion for stationary eye trackers using saccadic eye movements. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*. 1–10.

[37] Ryo Ishii, Yukiko I Nakano, and Toyoaki Nishida. 2013. Gaze awareness in conversational agents: Estimating a user's conversational engagement from eye gaze. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 3, 2 (2013), 1–25.

[38] Chuhan Jiao, Zhiming Hu, Mihai Bâce, and Andreas Bulling. 2023. SUPREYES: SUPer Resolution for EYES Using Implicit Neural Representation Learning. In *Proc. ACM Symposium on User Interface Software and Technology (UIST)*. 1–13.

https://doi.org/10.1145/3586183.3606780
[39] Chuhan Jiao, Yao Wang, Guanhua Zhang, Mihai Bâce, Zhiming Hu, and Andreas Bulling. 2024. DiffGaze: A Diffusion Model for Continuous Gaze Sequence Generation on 360 {\deg} Images. *arXiv preprint arXiv:2403.17477* (2024).
[40] Chuhan Jiao, Guanhua Zhang, Zhiming Hu, and Andreas Bulling. 2024. DiffEyeSyn: Diffusion-based User-specific Eye Movement Synthesis. *arXiv preprint arXiv:2409.01240* (2024).
[41] Enkelejda Kasneci, Hong Gao, Suleyman Ozdel, Virmarie Maquiling, Enkeleda Thaqi, Carrie Lau, Yao Rong, Gjergji Kasneci, and Efe Bozkir. 2024. Introduction to Eye Tracking: A Hands-On Tutorial for Students and Practitioners. *arXiv preprint arXiv:2404.15435* (2024).
[42] Peter Kiefer, Ioannis Giannopoulos, Martin Raubal, and Andrew Duchowski. 2017. Eye tracking for spatial research: Cognition, computation, challenges. *Spatial Cognition & Computation* 17, 1-2 (2017), 1–19.
[43] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2021. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In *International Conference on Learning Representations*. https://openreview.net/forum?id=a-xFK8Ymz5J
[44] 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), 2539.
[45] Tiffany CK Kwok, Peter Kiefer, Victor R Schinazi, Benjamin Adams, and Martin Raubal. 2019. Gaze-guided narratives: Adapting audio guide content to gaze in virtual and real environments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.
[46] Mikko Kytö, Barrett Ens, Thammathip Piumsomboon, Gun A Lee, and Mark Billinghurst. 2018. Pinpointing: precise head-and eye-based target selection for augmented reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.
[47] Beibin Li, Erin Barney, Caitlin Hudac, Nicholas Nuechterlein, Pamela Ventola, Linda Shapiro, and Frederick Shic. 2020. Selection of eye-tracking stimuli for prediction by sparsely grouped input variables for neural networks: towards biomarker refinement for autism. In *ACM Symposium on Eye Tracking Research and Applications*. 1–8.
[48] Jiaman Li, Karen Liu, and Jiajun Wu. 2023. Ego-Body Pose Estimation via Ego-Head Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17142–17151.
[49] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. 2024. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=JePfAI8fah
[50] Dillon Lohr and Oleg V. Komogortsev. 2022. Eye Know You Too: Toward Viable End-to-End Eye Movement Biometrics for User Authentication. *IEEE Transactions on Information Forensics and Security* 17 (2022), 3151–3164. https://doi.org/10.1109/TIFS.2022.3201369
[51] Paula López, Ana María Bernardos, and José Ramón Casar. 2024. Eye-Tracking Analysis for Cognitive Load Estimation in Wearable Mixed Reality. In *Proceedings of the 2024 ACM Symposium on Spatial User Interaction*. 1–2.
[52] Mathias N Lystbæk, Thorbjørn Mikkelsen, Roland Krisztandl, Eric J Gonzalez, Mar Gonzalez-Franco, Hans Gellersen, and Ken Pfeuffer. 2024. Hands-on, Hands-off: Gaze-Assisted Bimanual 3D Interaction. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–12.
[53] Mathias N Lystbæk, Ken Pfeuffer, Jens Emil Sloth Grønbæk, and Hans Gellersen. 2022. Exploring gaze for assisting freehand selection-based text entry in ar. *Proceedings of the ACM on Human-Computer Interaction* 6, ETRA (2022), 1–16.
[54] Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, et al. 2024. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *European Conference on Computer Vision*. Springer, 445–465.
[55] Pujitha Mannaru, Balakumar Balasingam, Krishna Pattipati, Ciara Sibley, and Joseph T Coyne. 2017. Performance evaluation of the gazepoint GP3 eye tracking device based on pupil dilation. In *Augmented Cognition. Neurocognition and Machine Learning: 11th International Conference, AC 2017, Held as Part of HCI International 2017, Vancouver, BC, Canada, July 9-14, 2017, Proceedings, Part I 11*. Springer, 166–175.
[56] Meta Platforms Inc. 2022. *Meta Quest Pro*. Virtual Reality Headset. Available at: https://www.meta.com/quest/quest-pro/ (Accessed on: 21 March 2025).
[57] Xiaoye Miao, Yangyang Wu, Jun Wang, Yunjun Gao, Xudong Mao, and Jianwei Yin. 2021. Generative semi-supervised learning for multivariate time series imputation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 8983–8991.
[58] Microsoft Corporation. 2019. *Microsoft HoloLens 2*. Augmented Reality Headset. Available at: https://www.microsoft.com/hololens (Accessed on: 21 March 2025).
[59] Philipp Müller, Daniel Buschek, Michael Xuelin Huang, and Andreas Bulling. 2019. Reducing calibration drift in mobile eye trackers by exploiting mobile phone usage. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*. 1–9.
[60] Yukiko I Nakano and Ryo Ishii. 2010. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In *Proceedings of the 15th international conference on Intelligent user interfaces*. 139–148.
[61] 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.
[62] Tong Nie, Guoyang Qin, Wei Ma, Yuewen Mei, and Jian Sun. 2024. ImputeFormer: Low rankness-induced transformers for generalizable spatiotemporal imputation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2260–2271.
[63] Marcus Nyström, Richard Andersson, Kenneth Holmqvist, and Joost Van De Weijer. 2013. The influence of calibration method and eye physiology on eyetracking data quality. *Behavior research methods* 45 (2013), 272–288.
[64] Marcus Nyström, Richard Andersson, Diederick C Niehorster, Roy S Hessels, and Ignace TC Hooge. 2024. What is a blink? Classifying and characterizing blinks in eye openness signals. *Behavior research methods* 56, 4 (2024), 3280–3299.
[65] Marcus Nyström, Ignace TC Hooge, Roy S Hessels, Richard Andersson, Dan Witzner Hansen, Roger Johansson, and Diederick C Niehorster. 2025. The fundamentals of eye tracking part 3: How to choose an eye tracker. *Behavior Research Methods* 57, 2 (2025), 67.
[66] Kara Pernice and Jakob Nielsen. 2009. How to Conduct and Evaluate Usability Studies Using Eyetracking. *Nielsen Norman Group: Fremont, CA, USA* (2009).
[67] Paul Prasse, David Robert Reich, Silvia Makowski, Seoyoung Ahn, Tobias Scheffer, and Lena A Jäger. 2023. Sp-eyegan: Generating synthetic eye movement data with generative adversarial networks. In *Proceedings of the 2023 symposium on eye tracking research and applications*. 1–9.
[68] Paul Prasse, David R Reich, Silvia Makowski, Tobias Scheffer, and Lena A Jäger. 2024. Improving cognitive-state analysis from eye gaze with synthetic eye-movement data. *Computers & Graphics* 119 (2024), 103901.
[69] Zhehan Qu, Ryleigh Byrne, and Maria Gorlatova. 2024. "Looking" into Attention Patterns in Extended Reality: An Eye Tracking-Based Study. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 855–864.
[70] Cecilia Ramaioli, Luigi F Cuturi, Stefano Ramat, Nadine Lehnen, and Paul R MacNeilage. 2019. Vestibulo-ocular responses and dynamic visual acuity during horizontal rotation and translation. *Frontiers in neurology* 10 (2019), 321.
[71] Susan K Schnipke and Marc W Todd. 2000. Trials and tribulations of using an eye-tracking system. In *CHI'00 extended abstracts on Human factors in computing systems*. 273–274.
[72] Lei Shi and Andreas Bulling. 2025. CLAD: Constrained Latent Action Diffusion for Vision-Language Procedure Planning. *arXiv preprint arXiv:2503.06637* (2025).
[73] 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)*.
[74] Ludwig Sidenmark and Hans Gellersen. 2019. Eye, head and torso coordination during gaze shifts in virtual reality. *ACM Transactions on Computer-Human Interaction* 27, 1 (2019), 1–40.
[75] Ludwig Sidenmark and Hans Gellersen. 2019. 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*. 1161–1174.
[76] Ludwig Sidenmark, Diako Mardanbegi, Argenis Ramirez Gomez, Christopher Clarke, and Hans Gellersen. 2020. BimodalGaze: Seamlessly Refined Pointing with Gaze and Filtered Gestural Head Movement. In *Proceedings of the 2020 ACM Symposium on Eye Tracking Research and Applications*. 1–9.
[77] Ikaro Silva, George Moody, Daniel J Scott, Leo A Celi, and Roger G Mark. 2012. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. *Computing in cardiology* 39 (2012), 245.
[78] 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.
[79] John S Stahl. 1999. Amplitude of human head movements associated with horizontal saccades. *Experimental Brain Research* 126, 1 (1999), 41–54.
[80] Julian Steil and Andreas Bulling. 2015. Discovery of Everyday Human Activities From Long-term Visual Behaviour Using Topic Models. In *Proc. ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. 75–85. https://doi.org/10.1145/2750858.2807520
[81] John A Stern, Larry C Walrath, and Robert Goldstein. 1984. The endogenous eyeblink. *Psychophysiology* 21, 1 (1984), 22–33.
[82] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. 2021. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in Neural Information Processing Systems* 34 (2021), 24804–24816.
[83] Tobii AB. 2023. *Tobii Eye Tracker 5*. Eye tracker for gaming. Available at: http://gaming.tobii.com/product/eye-tracker-5/ (Accessed on: 21 March 2025).
[84] Rumeysa Turkmen, Zeynep Ecem Gelmez, Anil Ufuk Batmaz, Wolfgang Stuerzlinger, Paul Asente, Mine Sarac, Ken Pfeuffer, and Mayra Donaji Barrera Machuca. 2024. EyeGuide & EyeConGuide: Gaze-based Visual Guides to Improve 3D Sketching Systems. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–14.

	Ego-Exo4D [19]				HOT3D [7]			
	10%	30%	50%	90%	10%	30%	50%	90%
iTransformer [49]	7.89	11.04	17.55	26.36	6.81	9.83	15.80	24.46
DLinear [92]	12.12	12.14	12.88	14.00	9.83	10.00	10.62	11.52
TimesNet [89]	21.66	20.27	22.37	25.23	19.83	18.79	20.51	23.23
BRITS [11]	10.51	12.47	14.96	19.17	9.47	11.36	13.50	17.20

Table 6: Mean angular error (MAE) of iTransformer [49], Dlinear [92], TimesNet [89], BRITS [11] on the Ego-Exo4D [19] and HOT3D [7] datasets.

[85] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[86] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17 (2020), 261–272. https://doi.org/10.1038/s41592-019-0686-2

[87] Hana Vrzakova, Mary Jean Amon, McKenzie Rees, Myrthe Faber, and Sidney D'Mello. 2021. Looking for a deal? visual social attention during negotiations via mixed media videoconferencing. *Proceedings of the ACM on Human-computer Interaction* 4, CSCW3 (2021), 1–35.

[88] Katarzyna Wisiecka, Yuumi Konishi, Krzysztof Krejtz, Mahshid Zolfaghari, Birgit Kopainsky, Izabela Krejtz, Hideki Koike, and Morten Fjeld. 2023. Supporting complex decision-making: evidence from an eye tracking study on in-person and remote collaboration. *ACM Transactions on Computer-Human Interaction* 30, 5 (2023), 1–27.

[89] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. 2023. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=ju_Uqw384Oq

[90] 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*.

[91] Brent Yi, Vickie Ye, Maya Zheng, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. 2024. Estimating Body and Hand Motion in an Ego-sensed World. *arXiv preprint arXiv:2410.03665* (2024).

[92] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. 2023. Are Transformers Effective for Time Series Forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*.

[93] Guanhua Zhang, Zhiming Hu, and Andreas Bulling. 2024. DisMouse: Disentangling Information from Mouse Movement Data. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–13.

[94] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2015. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4511–4520.

[95] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2017. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence* 41, 1 (2017), 162–175.

[96] Yunhao Zhang and Junchi Yan. 2023. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The eleventh international conference on learning representations*.

[97] Zhaobo Zheng, Kumar Akash, Teruhisa Misu, Vidya Krishnamoorthy, Miaomiao Dong, Yuni Lee, and Gaojian Huang. 2022. Identification of adaptive driving style preference through implicit inputs in sae l2 vehicles. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. 468–475.

[98] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 11106–11115.

[99] Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. 2020. Transformer hawkes process. In *International conference on machine learning*. PMLR, 11692–11702.

A RESULT OF TIME-SERIES IMPUTATION BASELINES ON EGO-EXO4D AND HOT3D

We provide MAE of iTransformer [49], Dlinear [92], TimesNet [89], BRITS [11] on the Ego-Exo4D [19] and HOT3D [7] datasets in Table A.