DiffGaze: A Diffusion Model for Modelling Fine-grained Human Gaze Behaviour on 360° Images

CHUHAN JIAO, YAO WANG, and GUANHUA ZHANG, University of Stuttgart, Germany
MIHAI BÂCE[*], KU Leuven, Belgium
ZHIMING HU[*], The Hong Kong University of Science and Technology (Guangzhou), China
ANDREAS BULLING, University of Stuttgart, Germany

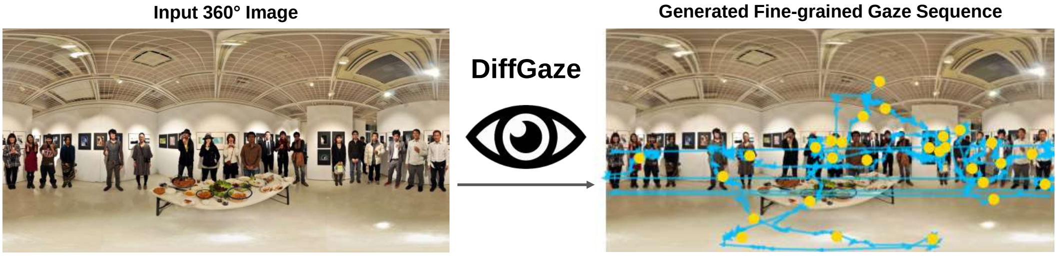

Fig. 1. In stark contrast to *scanpath prediction*, which only focuses on predicting sequences of fixations (yellow), *fine-grained gaze sequence generation* is the much more challenging task of faithfully and holistically modelling human gaze behaviour (blue) that resemble the gaze data recorded by eye trackers (normally 30 Hz - 1000 Hz). We present *DiffGaze* — the first method for fine-grained gaze sequence generation on 360° images. DiffGaze aims to model eye events with very different velocity distributions, e.g. fixations and saccades, formulating the fine-grained gaze sequence generation task as a conditional diffusion process conditioned on the image features.

Modelling human gaze behaviour on 360° images is important for various human-computer interaction applications. However, existing methods are limited to predicting discrete fixation sequences or aggregated saliency maps, thereby neglecting fine-grained gaze behaviour such as saccadic eye movements that can be captured by commercial eye-trackers. We introduce a more challenging task—*fine-grained gaze sequence generation*. This task aims to generate eye-tracker-like gaze data for given stimuli. We propose *DiffGaze*, a diffusion-based method for generating realistic and diverse fine-grained human gaze sequences conditioned on 360° images. We evaluate DiffGaze on two 360° image benchmarks for fine-grained gaze sequence generation as well as two downstream tasks, scanpath prediction and saliency prediction. Our evaluations show that DiffGaze outperforms the fine-grained gaze generation baselines in all tasks on both benchmarks. We also report a 21-participant survey study showing that our method generates gaze sequences that are indistinguishable from real human sequences. Taken together, our evaluations not only demonstrate the effectiveness of DiffGaze but

[*]Work done while at the University of Stuttgart.

Authors' addresses: Chuhan Jiao; Yao Wang; Guanhua Zhang, University of Stuttgart, Stuttgart, Germany, firstname.lastname@vis.uni-stuttgart.de; Mihai Bâce, KU Leuven, Leuven, Belgium, mihai.bace@kuleuven.be; Zhiming Hu, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China, zhiminghu@hkust-gz.edu.cn; Andreas Bulling, University of Stuttgart, Stuttgart, Germany, firstname.lastname@vis.uni-stuttgart.de.

ACM 0004-5411/2025/8-ART111
https://doi.org/XXXXXXX.XXXXXXX

also point towards a new generation of methods that faithfully model the rich spatial and temporal nature of natural human gaze behaviour.

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

Additional Key Words and Phrases: Gaze Behaviour Modelling, Scanpath Prediction, Eye Tracking, Eye Movement Synthesis

ACM Reference Format:
Chuhan Jiao, Yao Wang, Guanhua Zhang, Mihai Bâce, Zhiming Hu, and Andreas Bulling. 2025. DiffGaze: A Diffusion Model for Modelling Fine-grained Human Gaze Behaviour on 360° Images. *J. ACM* 37, 4, Article 111 (August 2025), 23 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

With recent advances in camera technology, capturing high-resolution 360° images enables a new generation of immersive experiences in virtual reality (VR). This potential has led to rising consumer interest in adopting this new technology and growing research efforts in understanding how humans perceive and explore these 3D virtual environments [44, 49, 50]. To understand and enhance VR experiences, computational user modelling – a core research area within interactive intelligent system research – is key. User modelling involves learning computational representations of user behaviours, enabling predictions, personalisation, and adaptation of interactive intelligent systems to suit human needs and capabilities better. While traditional human-computer interaction has long focused on user modelling tasks in desktop or mobile contexts (e.g. typing modelling [47], mouse movement modelling [67], gaze modelling [9]), immersive environments present new challenges and opportunities for modelling user behaviour in richer and more complex ways.

In VR, visual attention is a crucial aspect of user behaviour that provides insights into cognitive and perceptual processes during interactions. Eye tracking technology, which captures gaze data to reveal where a person is looking, is widely recognised as a valuable tool for studying attention and decision-making in physical and virtual spaces. While eye tracking has become more widely available and affordable, [27, 56] – and is integrated into an ever-increasing number of VR headsets – collecting gaze data, particularly at scale, remains tedious and time-consuming and is often not feasible at all. This has led to a strong interest in HCI and adjacent fields in computational models that can predict human gaze behaviour without specialised eye tracking hardware. Such models of visual attention are essential not only for understanding perception in VR but also for developing applications that adapt to users' visual behaviour.

Within HCI, computational models of gaze have been used to support various applications, including adaptive user interfaces [21], predicting users' cognitive load [52]. Despite these advances, prior works on computational modelling of visual attention on 360° images have focused on saliency [7, 50] or scanpath prediction [1, 37, 53]. Despite significant advances, both tasks still only tackle a simplified problem. Saliency prediction yields one or multiple fixation density maps, which do not capture the temporal nature of human gaze behaviour. Scanpath prediction only yields sequences of gaze fixations, neglecting those gaze samples between fixations and raw gaze samples that form the fixations. As a result, neither of these tasks – nor any existing method developed in the past to tackle them – allow to faithfully model the rich spatial and temporal nature of natural human gaze behaviour on 360° images that are akin to the gaze data captured by commercial eye trackers operating at a typical sampling rate of at least 30 Hz.

To address these limitations, we introduce a challenging task – *fine-grained gaze sequence generation*, which aims to generate temporally dense gaze data that resembles the continuous output of commercial eye trackers, usually with a sampling rate from 30 Hz to 1,000 Hz. Unlike traditional scanpath prediction that models only discrete fixation locations, fine-grained sequences capture the continuous trajectory of the eye, including the raw gaze samples that form and occur between fixations. The key challenge of this task is to model eye movements with various velocity distributions.

Inspired by recent success in using diffusion models to model complex distributions across different domains [28, 45, 54, 68], we propose *DiffGaze* – the first generative method to model fine-grained human gaze behaviour on 360° images. DiffGaze is based on a score-based denoising diffusion model [54] that is conditioned on the features extracted from 360° images. Taking 30 Hz as an example, we evaluate our method on fine-grained gaze sequence generation on two datasets, Sitzmann [50] and Salient360! [42, 43]. Since fixations are encoded in fine-grained gaze sequence and can be detected using eye event detection algorithms, such as Identification by Velocity-Thresholding [46], we further evaluate the performance of DiffGaze on two downstream tasks, scanpath prediction and saliency prediction. The results show that our method achieves state-of-the-art performance on fine-grained gaze sequence generation and scanpath prediction and outperforms other fine-grained gaze sequence generation baselines in saliency prediction. We also show that our model reproduces natural eye movement characteristics in humans, such as the mean number of saccades, mean saccade velocity, mean number of fixations, or mean fixation duration. Finally, through an online user study with 21 participants, we show that the fine-grained gaze sequences generated by DiffGaze are practically indistinguishable from real human gaze behaviour. Taken together, DiffGaze represents a novel approach in computational user modelling within HCI, demonstrating how generative models can faithfully capture the complex, fine-grained aspects of human gaze in VR. As such, our method and the fine-grained gaze sequence generation task not only have the potential to unify long-standing, yet so far largely separated, efforts on saliency and scanpath prediction. Our method also promises significant improvements in a range of applications that can directly build on it, such as high-frequency animation of virtual humans, synthesising eye tracking datasets in VR or assisting designers to improve the design of VR environments.
The specific contributions of our work are three-fold:

(1) We propose DiffGaze – the first generative method to model fine-grained human gaze sequences on 360° images.
(2) Through extensive evaluations, we show that our method achieves state-of-the-art performance on fine-grained gaze sequence generation and scanpath prediction.
(3) Through an online survey study with 21 participants, we show that the gaze sequences generated by DiffGaze are practically indistinguishable from real human gaze behaviour.

2 RELATED WORK

Our work on fine-grained gaze sequence generation is related to previous work on 1) saliency modelling, 2) scanpath prediction on 360° images, and 3) realistic eye movement generation.

2.1 Saliency Modelling

Saliency modelling aims at predicting the spatial distribution of human gaze fixations, also known as saliency map, on an image and has been extensively studied in the past few decades. Traditional saliency prediction methods usually focus on predicting the saliency map of 2D images and can be classified into bottom-up and top-down approaches. Bottom-up methods predict saliency maps using low-level image features such as intensity, colour, and orientation [11, 20]. Top-down approaches use high-level features such as specific tasks and context information to predict saliency maps [5, 63]. Recently, with the development of virtual reality, human visual saliency on 360° images has become an essential research topic in computer vision. Specifically, Sitzmann et al. proposed to combine existing 2D saliency predictors with a central bias to generate saliency maps for 360° images [50] while Chen et al. proposed a local-global bifurcated deep network for saliency prediction on 360° content [7]. However, saliency prediction methods can only model the spatial distribution of human visual attention and ignore the temporal dynamics. To solve this limitation, researchers focused on scanpath prediction that models the temporal sequence of gaze fixations.

2.2 Scanpath Prediction on 360° Images

Scanpath prediction aims to generate a discrete sequence of eye gaze fixations from a given image and has been extensively explored in vision research [2, 21, 35, 60]. In the rapidly evolving virtual reality domain, predicting scanpaths on 360° images has emerged as a pivotal research topic. Researchers have explored directly adapting scanpath prediction methods from 2D images, such as DeepGaze III [30] and CLE [3], to 360° images. Additionally, methods exclusive to 360° images including SaltiNet [1], ScanGAN360 [37], ScanDMM [53], and Pathformer3D [41] have been proposed. The key idea of these works can be divided into two categories.

(1) Sample fixation locations from predicted saliency information. CLE [3] samples fixation locations from a saliency map through a three-stage processing, which relies on a centre-bias model, a context/layout model, and an object-based model, respectively; DeepGaze III [30] takes previous fixation history to predict the next fixation on saliency feature map; SaltiNet [1] samples fixations from predicted saliency volumes, i.e. a temporal-aware representation of saliency information, of a given image. These methods did not consider fixation durations, but only sampled the durations from the duration distribution [1].

(2) End-to-end approaches, i.e. directly generate fixation locations. Variant lengths in ground truth human scanpaths pose an issue for training end-to-end models like generative adversarial networks (GANs), which require a fixed output length. To tackle this problem, Martin et al. simplified the task by predicting 1 Hz gaze data over a fixed duration (generating one gaze location per second) and proposed ScanGAN360 [37], a GAN network combined with a dynamic time warping-based loss function. Following the same setting, Sui et al. presented ScanDMM – a deep Markov model to generate 1 Hz gaze data from 360° images [53]. The assumption behind these methods is that all 1 Hz gaze samples are fixations lasting for 1 second. The generated fixation numbers are fixed given a duration, which is unrealistic compared with human scanpaths with variant lengths. In addition, the 1 Hz gaze samples are not guaranteed to belong to fixations they might belong to other types of eye movements.

As summarized in Table 1, previous gaze modeling methods primarily concentrate on generating discrete gaze fixation sequences, i.e. scanpaths, neglecting the fine-grained nature of human eye movements. In contrast, our work aims to predict fine-grained gaze sequences from 360° images that resemble the data recorded using eye trackers rather than only generating discrete gaze fixations, thereby capturing a more realistic representation of human visual attention. The fixations and saccades can also be detected in our generated fine-grained gaze sequence by applying eye movement detection algorithms like Identification by Velocity-Thresholding [46]. Given a fixed duration, the number of fixations is variable in our method, and the duration of each fixation can be calculated by multiplying the number of samples forming the fixation by the sampling frequency.

2.3 Eye Movement Generation

Generating realistic eye movements has emerged as an important research topic in human-computer interaction given its relevance for several applications, such as virtual character animation or human attention analysis. In early work, Lee et al. proposed to use an empirical model of saccades and a statistical model of eye-tracking data to generate eye movements [34] while Lan et al. presented a novel set of psychology-inspired generative models to synthesise eye movements in reading, verbal

Table 1. Comparison of abilities of gaze behaviour modelling methods for 360° images. DiffGaze is the first method to generate all kinds of gaze behaviours including raw gaze samples at 30 Hz, fixations, saccadic eye movements, fixation duration, and variant number of fixations (Fixation Count). EM: eye movements, ?: unsure eye movement type.

Method	Sampling Frequency	Saccadic EM	Fixation	Fixation Duration	Fixation Count
End-to-end [37, 53]	1 Hz	?	?	1 second	fixed
Saliency-based [1, 3, 30]	×	×	✓	sample from distribution	variant
Fine-grained EM (DiffGaze)	30 Hz	✓	✓	detect from raw gaze	variant

communication, and scene perception [31]. Considering the strong correlation between eye and head movements, Sitzmann et al. [50] and Rai et al. [42] proposed to use head orientation as a proxy to eye gaze. Hu et al. proposed to generate realistic eye movements in free-viewing situations from users' head movements and scene content [18, 19] and to predict eye movements using task-related information in task-oriented situations [16, 17]. Xu et al. proposed to predict eye movements in graphical user interface from users' mouse and keyboard inputs while Koulieris et al. used game state variables in video games to predict users' gaze positions [29]. However, existing methods usually rely on other input modalities, such as head movements, mouse and keyboard inputs, to generate eye movements. In stark contrast, we generate realistic and fine-grained gaze sequences only using the information available in 360° images.

3 METHOD

An overview of our method is shown in Figure 2. Human gaze behaviour consists of a mix of rapid eye movements and predominantly slower ones, resulting in a complex and diverse velocity distribution. To better capture these dynamics, we leverage diffusion models, which have been shown to effectively model complex distributions across domains [15, 24, 25, 28, 68]. We formulate fine-grained gaze sequence generation on 360° images as a conditional diffusion task. That is, DiffGaze is trained to denoise human gaze data corrupted by Gaussian noise, conditioned on the spherical convolution of a 360° image. During diffusion sampling, DiffGaze generates continuous gaze data from random noise and extracted image features.

3.1 Problem Definition

Eye tracking data are usually denoted as a sequence of two-dimensional locations (i, j). Projected on 360° images, these locations are represented by their latitude and longitude (ϕ, λ), where $\phi \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ and $\lambda \in [-\pi, \pi]$. However, this representation introduces discontinuities at the image borders and leads to periodic values, causing problems during training. For example, λ and $\lambda + 2\pi$ represent the same meridian, including the two image boarders ($\lambda = \pm 180°$). Thus, we instead project the image onto a unit sphere and represent each gaze point as a three-dimensional vector (x, y, z) following [37]:

$$x = cos(\phi)cos(\lambda), y = cos(\phi)sin(\lambda), z = sin(\phi) \tag{1}$$

Given a 360° image, our goal is to generate a *fine-grained gaze sequence*—that is, a high-frequency, temporally dense sequence of gaze points that approximates the raw output of a commercial eye tracker. Formally, we denote a fine-grained gaze sequence as $X \in \mathbb{R}^{3 \times L}$, where L is the number of gaze samples and each column corresponds to a gaze vector $(x, y, z)^\top$ at a particular timestamp. These sequences are sampled at a fixed rate and capture both slow and fast eye movements, different from scanpaths that only contain sparse fixation locations. The generated sequence is projected

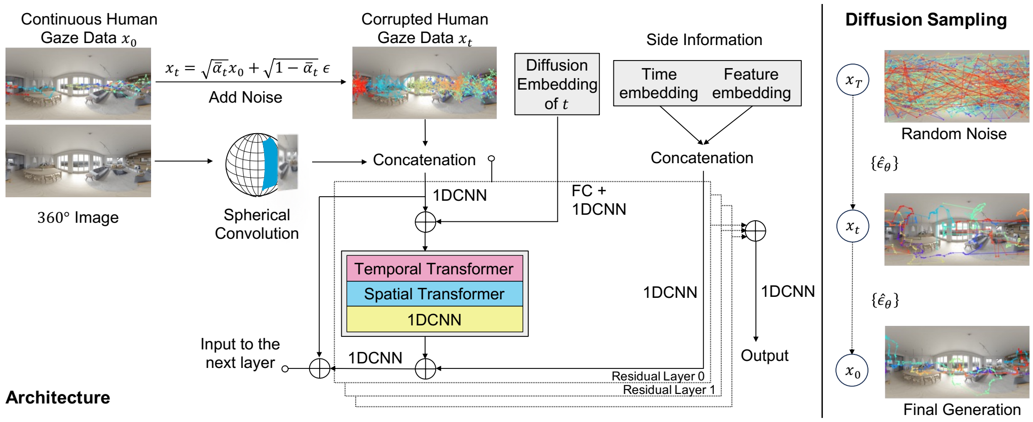

Fig. 2. Overview of our proposed DiffGaze method. We cast continuous gaze sequence generation as a conditional diffusion task. Our model is trained to recover the original gaze trajectory from the corrupted, noisy data. The condition to guide this diffusion process includes a spherical convolution for the 360° image, and side information (time and feature embedding). We apply two Transformers to learn both the temporal and spatial attention. Please refer to the text for details about the architecture, diffusion process and the loss function. 1DCNN refers to the one-dimensional convolutional neural network and FC refers to the fully-connected layer.

back to the latitude-longitude representation (ϕ, λ) for evaluation and visualisation by

$$\phi = arctan2(z, \sqrt{x^2 + y^2}), \lambda = arctan2(y, x) \tag{2}$$

3.2 Diffusion Model

At its core, DiffGaze uses a denoising diffusion model that has recently shown impressive results in generative image tasks [54]. Diffusion models are probabilistic models, consisting of a forward and a reverse process. The aim is to learn a distribution $p_\theta(X)$ to approximate the data distribution $q(X)$. The **forward process** gradually adds Gaussian noise to the original data X_0 for T timesteps. At each step $t \in [1, T]$, the sampling of X_t can be obtained via a Markov chain:

$$q(X_t \mid X_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}X_{t-1}; \beta_t \mathbf{I}) \tag{3}$$

where $\beta_t \in [0, 1]$ is a scalar representing the noise level. Using $\alpha_t = \prod_{i=1}^{t}(1-\beta_i)$, the noise distribution at any intermediate timestamp can be obtained by

$$q(X_t \mid X_0) = \mathcal{N}(X_t; \sqrt{\alpha_t}X_0, (1-\alpha_t)\mathbf{I}) \tag{4}$$

Hence, X_t is

$$X_t = \sqrt{\alpha_t}X_0 + (1-\alpha_t)\epsilon \tag{5}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. In the **reverse process**, a generative model θ denoises X_t to recover X_0 by the transition:

$$p_\theta(X_{t-1} \mid X_t) = \mathcal{N}(X_{t-1}; \mu_\theta(X_t, t), \sigma_\theta(X_t, t)\mathbf{I}) \tag{6}$$

$\sigma_\theta(X_t, t)$ is fixed to a constant β_t for easier optimisation. $\mu_\theta(X_t, t)$ can be decomposed into the linear combination of X_t and a denoising function $\epsilon_\theta(X_t, t)$, which is learnt via the following optimisation [15]:

$$\underset{\theta}{min}\mathcal{L}(\theta) = \underset{\theta}{min}\mathbb{E}_{X_0\sim q(X_0),\epsilon\sim\mathcal{N}(0,\mathbf{I}),t}\,||\epsilon - \epsilon_\theta(X_t, t)||_2^2 \tag{7}$$

3.3 Conditioning Mechanism for 360° Images

Since objects in the image, along with their semantics and location, have a significant influence on human gaze behaviour, it is essential to consider image features when predicting gaze patterns. Therefore, we use the 360° image itself as a condition to the diffusion process. More specifically, given that gaze points are projected to a spherical space (Equation 1), we propose a spherical embedding of 360° image as the condition in the reverse diffusion process. Following [37], we encode the information of the image into $c \in R^{m \times 1}$ using the sphere convolutional neural network (S-CNN) [12] that can handle space-varying distortions resulting from the equirectangular projection. In S-CNN, we apply a CoordConv layer [36] to extract the spatial embedding. CoordConv adds extra coordinate channels to make convolutions more robust to coordinate transformations, with few parameters and efficient computation. After introducing the condition c to the reverse process, the goal is to approximate the learnt distribution $p_\theta(X \mid c)$ to the genuine distribution $q(X \mid c)$. Correspondingly, we extend the reverse process in Equation 6 as

$$p_\theta(X_{t-1} \mid X_t, c) = \mathcal{N}(X_{t-1}; \mu_\theta(X_t, t \mid c), \sigma_\theta(X_t, t \mid c)\mathbf{I}) \tag{8}$$

with the training objective

$$\underset{\theta}{min}\mathcal{L}(\theta) = \underset{\theta}{min}\mathbb{E}_{X_0 \sim q(X_0), \epsilon \sim \mathcal{N}(0,\mathbf{I}), t} \, \|\epsilon - \epsilon_\theta(X_t, t \mid c))\|_2^2 \tag{9}$$

3.4 DiffGaze Architecture

An overview of DiffGaze architecture are shown in Figure 2. DiffGaze takes the noisy gaze data $X_t \in R^{L \times 3}$, the conditioning 360° image, and the time step t as inputs to predict the added noise ϵ at the time step for the denoising reverse process.

The key to generating fine-grained gaze data is learning the relationship between image features and gaze locations. To achieve this, DiffGaze leverages a commonly used feature extractor for 360° images, S-CNN [12], to extract spatial features. To ensure fair comparison with prior work [37, 53], we retain all layers of the S-CNN except the final one, which is modified to produce output features with the same shape as X_t. We then apply early fusion by concatenating the extracted image features with X_t along a new dimension, resulting in a tensor of shape $(2, L, 3)$. Before further processing, this fused tensor is passed through a 1D convolutional layer (kernel size 1) to increase the number of channels, yielding a tensor of shape $(C, L, 3)$, where C denotes the number of channels.

Following commonly used diffusion model designs [15, 28, 54, 68], DiffGaze contains N residual layers. Prior work has shown that Transformer-based architectures are more effective than convolutional layers at capturing complex temporal and spatial patterns in time-series data [54]. As human gaze trajectories are time-series in nature, we adopt a similar architecture to CSDI [54], equipping each residual layer with two separate one-layer Transformer encoders: a *Temporal Transformer*, which models dependencies along the time axis (L), and a *Spatial Transformer*, which operates across the three gaze dimensions (x, y, z). In contrast to CSDI, which considers only self-attention within the time-series, our design additionally incorporates image features alongside gaze data. This enables the attention mechanism to learn temporal dependencies as well as relationships between visual content and gaze patterns across both time and space.

At the beginning of each residual layer, we add a diffusion embedding of the current time step t to the fused input tensor (shape: $(C, L, 3)$) via element-wise addition. This tensor is then processed by the two Transformers. The output is subsequently upsampled to $(2C, L, 3)$ using a 1D convolutional layer with kernel size one.

We incorporate two types of side information, following prior work [28, 54]. First, we encode the timestamp sequence $S = \{S_1, ..., S_L\}$ using a 128-dimensional positional encoding [59, 69]:

$$S_{embedding}(S_l) = (sin(S_l/\tau^{0/64}), ..., sin(S_l/\tau^{63/64}), \\ cos(S_l/\tau^{0/64}), ..., cos(S_l/\tau^{63/64})) \tag{10}$$

where $l \in [1, L]$, $\tau = 10,000$. Second, we use a 16-dimensional categorical embedding for the three gaze features (x, y, z), as in [54]. We concatenate these two embeddings and pass them through a 1DCNN layer (kernel size one), producing a side information tensor of shape $(2C, L, 3)$. This side information is fused with the main feature tensor via element-wise addition. The result is passed through a gated activation unit [28, 57, 58], followed by two 1DCNN layers (kernel size one). The output of this block serves both as input to the next residual layer and as a residual connection. Finally, the output of the last residual layer, with shape $(C, L, 3)$, is projected to the predicted noise tensor of shape $(L, 3)$ using a final 1DCNN layer.

4 EXPERIMENTS

4.1 Datasets

Most existing scanpath and saliency datasets for 360° do not offer real gaze data captured using an eye tracker to train and evaluate attention models. Two noteworthy exceptions are Sitzmann [50] and Salient360! [42, 43]. The Sitzmann dataset contains 22 images with 1,980 gaze trajectories from 169 participants, recorded at a sampling rate of 120 Hz for 30 seconds per image. Salient360! contains binocular gaze data of between 40 - 42 observers looking at 85 images. Gaze data was recorded at a sampling rate of 60 Hz for 25 seconds per image. Following previous works on scanpath prediction [37, 53], we used the same 19 images from the Sitzmann dataset for training and the remaining three for testing. We used the entire Salient360! dataset as a test set for cross-dataset evaluation.

In eye tracking, the number of gaze samples should (roughly) be the sampling frequency multiplied by the recording duration. We discarded recordings with fewer gaze samples than the minimum number of gaze samples. For Sitzmann, we set the minimum number to 3,481, corresponding to 29 seconds of viewing time. For Salient360!, we set the minimum number to 1,441, corresponding to 24 seconds of viewing time. We opted to generate fine-grained gaze sequences at a sampling frequency of 30 Hz as commonly offered by commercial mobile eye trackers. To this end, we first downsampled the data from both datasets to 30 Hz. To fix the output size for the diffusion model, the downsampled data was truncated to 871 samples for Sitzmann and 721 samples for Salient360!.

4.2 Experimental Setup

Implementation Details. We used the downsampled Sitzmann dataset to train our model. Similar to [53], we resized all images to a resolution of (128, 256) before training. We adapted the diffusion process parameters from CSDI [54] which includes four residual layers, $C = 64$ residual channels, and eight attention heads for each Transformer, and set the noise level range from 0.0001 to 0.5. We also used a quadratic schedule to update the noise level at each diffusion step, following [39, 51, 54] to enhance the quality of the generated samples. We set the total diffusion step $T = 200$, the batch size as 16, and the learning rate to 0.001. We used the Adam optimiser to train our model for 500 epochs, and reduce the learning rate by a factor of 0.1 at epoch 375 and epoch 450. All the experiments were run on a single NVIDIA V100 32GB GPU.

Baselines. To the best of our knowledge, DiffGaze is the first method for generating 30 Hz fine-grained gaze sequences on 360° images. The closest methods to ours are ScanGAN360 [37] and ScanDMM [53] but they were designed for generating 1 Hz gaze sequences to simulate the human

Table 2. Quantitative evaluation of fine-grained gaze sequence generation methods on Sitzmann and Salient360! dataset in terms of Levenshtein distance (LEV), Dynamic Time Warping (DTW), mean absolute error (MAE), and root mean squared error (RMSE). Best results are shown in **bold**.

Dataset	**Method**	**LEV ↓**		**DTW ↓**		**MAE ↓**		**RMSE ↓**	
		mean	*best*	*mean*	*best*	*mean*	*best*	*mean*	*best*
Sitzmann [50]	Human	1,167†	992†	1,708,925†	944,914†	1,587†	1,123†	2,418†	1,813†
	ScanGAN360 [37]	1,406	1,327	2,084,017	1,709,852	1,795	1,422	2,532	2,051
	ScanDMM [53]	1,293	1,161	2,139,311	1,467,909	1,710	1,279	2,519	1,966
	DiffGaze (Ours)	**1,272**	**1,148**	**1,785,109**	**1,163,822**	**1,623**	**1,195**	**2,365**	**1,810**
Salient360! [42, 43]	Human	1,060†	928†	360,523†	215,671†	402†	285†	587†	434†
	ScanGAN360 [37]	1,163	1,084	421,229	337,640	435	344	602	478
	ScanDMM [53]	1,092	**951**	441,725	285,542	421	301	611	456
	DiffGaze (Ours)	**1,079**	957	**380,857**	**245,206**	**413**	**298**	**591**	**439**

† Gaze sequences are not compared with themselves

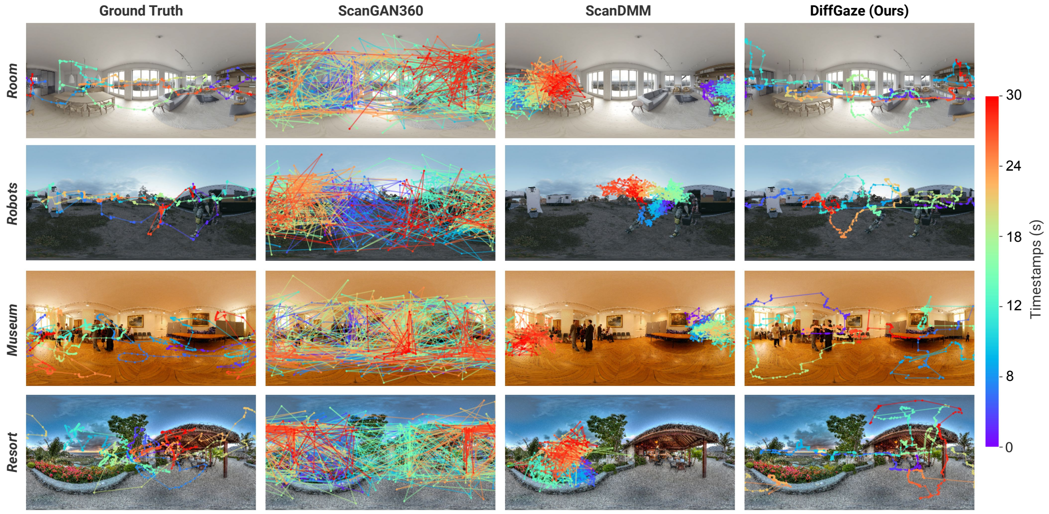

Fig. 3. Qualitative comparison of fine-grained gaze data generation models in four scenes. From left to right: gaze samples from a human observer, generated 30 Hz eye movement sequences from the ScanGAN360 method, ScanDMM, and our proposed model. From top to bottom: the Room and the Robots from the Sitzmann dataset, the Museum and Resort from the Salient360! dataset.

scanpaths. We adapted these methods by changing their output resolutions and by retraining them using our settings. We used the same hyperparameters as given in the original papers in retraining.

4.3 Fine-grained Gaze Sequence Generation

We conducted an experiment to assess the quality of the gaze sequences generated by our method and the baselines at 30 Hz, compared to the human ground truth. All methods are generative models that produce diverse gaze sequences, therefore we generated 100 sequences per method for each image to reduce the sampling bias.

Evaluation Metrics. We evaluated the generated gaze sequences using four time-series metrics commonly used in scanpath prediction and gaze data super-resolution: Levenshtein distance (LEV), dynamic time warping (DTW), mean absolute error (MAE), and root mean square error (RMSE) [23, 37, 53]. These metrics assess the temporal alignment, spatial distance, and overall deviation between two gaze sequences, with lower values indicating better performance. All gaze locations were converted to image space for metric calculation. For the Sitzmann dataset, metrics were calculated on the original image size due to uniform resolution. For the Salient360! dataset, metrics were calculated on the provided saliency maps' size (1024, 2048) due to varying image sizes.

Given that each image has multiple human ground truth gaze sequences and multiple generated gaze sequences, we report two numbers for each metric: *best* and *mean*, following prior work [60]. These numbers capture two important aspects of the quality of the generated sequences. The best number reflects how realistic the generated gaze sequence is, i.e., if there exists a human gaze sequence that is close to the generated one. We computed the best number as follows: For each generated sequence, we computed each metric to all ground truth sequences and recorded the best score for each metric. We then averaged the best scores over all generated sequences for each image and reported the result. In contrast to *best*, the *mean* number reflects how representative the generated gaze sequence is, i.e., how well it captures the average behaviour of all human gaze sequences. We computed the mean number as follows: For each generated sequence, we computed each metric with respect to all ground truth sequences and recorded the average score for each metric. Then we averaged the means over all generated sequences and reported the result.

To compare our results with human performance, we also computed a human baseline [62] for each metric. For each image, we selected one ground truth sequence and computed the metrics with the remaining ground truth sequences. We recorded the best score for each metric and averaged them over all ground truth sequences for each image. We then calculated the mean over all images for the human baseline.

Quantitative Results. Table 2 shows the results on both datasets. DiffGaze demonstrates superior performance on the Sitzmann dataset across all metrics and outperforms other baselines on three out of four metrics on Salient360!. This shows our method's effectiveness in generating fine-grained gaze sequences that closely resemble human behaviour.

Qualitative Results. Figure 3 presents examples of the generated gaze sequences. The gaze samples produced by ScanGAN360 and ScanDMM did not resemble the human eye behaviour of exploring 360° images. In particular, ScanGAN360 created many rapid eye movements across the images, unlike the human ground truth. The samples produced by ScanDMM concentrate on a few large regions on the images. This suggests that these two methods do not capture the relation between fixation and saccade in human eye movements. On the other hand, DiffGaze shows similar behaviours to the human ground truth. Additional qualitative results can be found in the supplementary material.

We also observed that visual inspection of human gaze sequences reveals significant variation in gaze patterns, even when viewing the same image. This variability results in poor performance of human agreement on existing scanpath metrics and time-series metrics (see Table 2). Therefore, these metrics' comparison with human agreement may not fully reflect the model's performance.

Survey Study. To gain a more nuanced understanding of model performance, we designed a survey, following prior work in scanpath prediction [60], that involved users in rating visualisations of different fine-grained gaze trajectories. The user study was structured into three stages:

- *Training*: we randomly picked three 360° images from both dataset. For each image, we showed 10 randomly chosen human gaze sequences to familiarise participants with the visual appearance of real human gaze data.
- *Rating*: Participants were then asked to rate the realism of gaze sequences on a scale of 1 to 10, where 1 meant highly unrealistic, and 10 highly realistic. We randomly selected 20 images

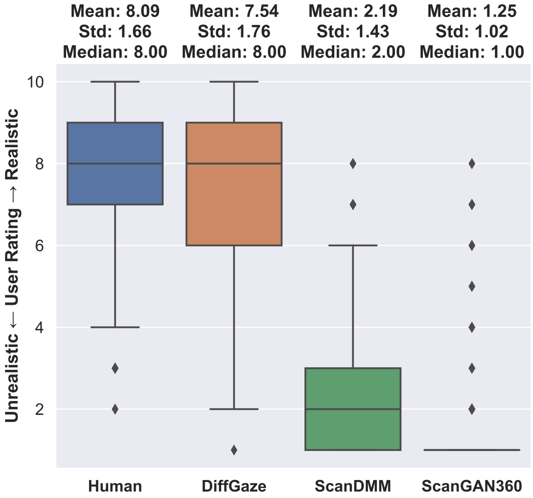

Fig. 4. User ratings of the realism of gaze sequences generated by Human, DiffGaze, ScanDMM, and ScanGAN360. 1: most unrealistic, 10: most realistic.

from both datasets. For each image, we randomly showed one real human gaze sequence, one sequence from our method, ScanDMM, and ScanGAN360, respectively. To minimise any potential positional bias, the presentation order of the four visualisations was randomised.

- *Validation*: We randomly selected 10 images from both datasets. For each image, we randomly picked four real human gaze sequences and asked participants to rate them in the same way as before. We then computed the average rating for each participant. We discarded all participants with an average rating lower than five.

The user study is approved by the ethics committee at the local university. We recruited 21 participants (nine females and 12 males) from different universities for the study, 15 of them were familiar with gaze data or had participated in eye tracking studies. The study was done online. No personal information of participants was collected. Figure 4 shows the average user ratings for realism achieved by the three different methods. The average rating of DiffGaze (7.54) is significantly higher than those of ScanDMM (t (63) = 2.19, p<0.001) and ScanGAN360 (t (93) = 1.25, p<0.001). Although the average of DiffGaze is slightly lower than the humans (7.54 vs 8.09), the median (8) is the same. This indicates that in many cases the generated fine-grained gaze sequences of our method are indistinguishable from real human sequences.

Analysis of Eye Movements. We further evaluated DiffGaze by comparing the fixation and saccade statistics of the generated gaze sequences with human ground truth. We used a velocity-based saccade detection algorithm [14] to segment the gaze sequences in spherical coordinates into fixations and saccades, with a velocity threshold $\lambda = 2$.

Table 3 shows the results of this analysis. In terms of saccade statistics, DiffGaze and ScanGAN360 yield similar mean saccade velocities as human observers in both datasets. However, ScanDMM and ScanGAN360 generate more saccades than human observers, generating more frequent attention shifts. DiffGaze achieves a comparable number of saccades as human observers, suggesting more natural gaze behaviour. Moreover, ScanDMM and ScanGAN360 generate more fixations than human observers, with shorter mean fixation durations. DiffGaze produces more realistic fixation statistics in the Salient360! dataset, with a similar number and duration of fixations as human observers. On the Sitzmann dataset, DiffGaze generates slightly fewer fixations with slightly longer duration than

Table 3. Eye movement statistics on Sitzmann and Salient360! dataset. Numbers closest to human statistic are marked in **bold**.

Dataset	Method	Mean saccade number	Mean saccade velocity (°/s)	Mean fixation number	Mean fixation duration (s)
Sitzmann [50]	Human	47.73 ± 9.86	224.23 ± 24.66	41.11 ± 7.29	0.69 ± 0.13
	ScanDMM [53]	103.28 ± 32.85	192.12 ± 42.52	60.89 ± 16.45	0.46 ± 0.16
	ScanGAN360 [37]	154.93 ± 14.46	217.24 ± 6.06	91.22 ± 9.82	0.26 ± 0.03
	DiffGaze (Ours)	**41.50 ± 9.11**	**220.97 ± 28.05**	**32.80 ± 6.91**	**0.88 ± 0.22**
Salient360! [42, 43]	Human	44.05 ± 11.70	209.01 ± 24.73	30.94 ± 7.25	0.79 ± 0.30
	ScanDMM [53]	91.34 ± 29.46	193.54 ± 46.33	54.68 ± 15.45	0.44 ± 0.15
	ScanGAN360 [37]	144.32 ± 23.41	216.13 ± 7.62	84.80 ± 13.25	0.29 ± 0.06
	DiffGaze (Ours)	**35.94 ± 8.35**	**219.76 ± 29.85**	**28.58 ± 6.42**	**0.84 ± 0.21**

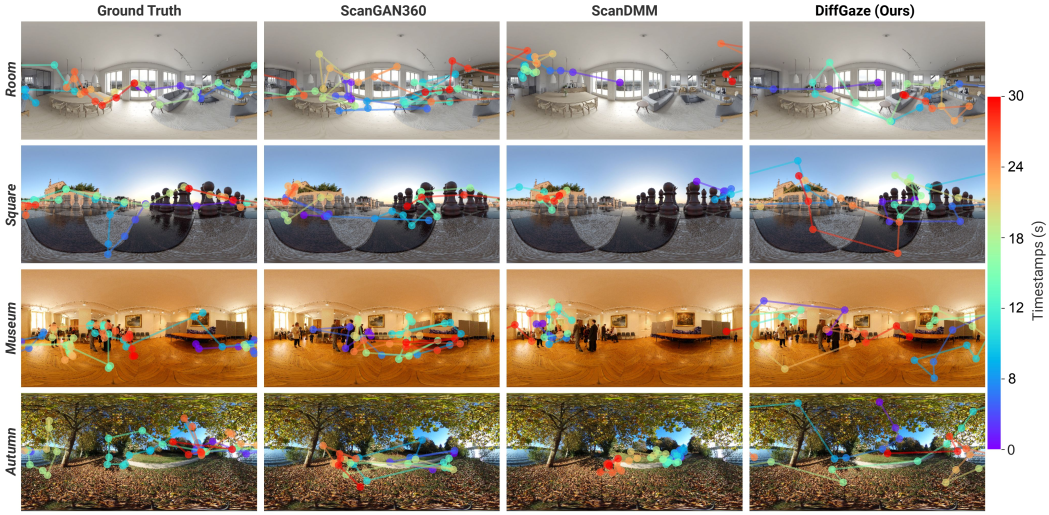

Fig. 5. Qualitative comparison to scanpath prediction models in four scenes. From left to right: scanpaths obtained by a human observer, generated 30 Hz scanpaths obtained by ScanGAN360, ScanDMM, and the proposed model. From top to bottom: the Room and the Square from Sitzmann dataset, the Museum and Autumn from Salient360! dataset.

human observers, but is still closer to the ground truth than the baselines. Taken together, these results show that DiffGaze can generate gaze sequences with characteristics that are highly similar to real human gaze behaviour.

4.4 Scanpath Prediction

Given that fixations and their durations are readily encoded in fine-grained gaze sequences, to further evaluate the quality of fixations within the generated fine-grained gaze sequences, we also evaluated DiffGaze on scanpath prediction. We used the scanpaths obtained from our gaze sequence extraction procedure (Section 4.3) as ground truth on the Sitzmann dataset, given it does not provide them originally. We discarded fixations shorter than 150 ms, following [50] and

Table 4. Quantitative evaluation of scanpath prediction methods on Sitzmann and Salient360! dataset in terms of Levenshtein distance (LEV), Dynamic Time Warping (DTW), and Recurrence (REC). Best results of each dataset are shown in **bold**.

Dataset	Method	LEV ↓		DTW ↓		REC ↑	
		mean	*best*	*mean*	*best*	*mean*	*best*
Sitzmann [50]	Human	55.82†	44.89†	73,751†	43,655†	0.108†	0.541†
	Human (1 Hz)	43.64	35.51	35,712	24,088	0.459	1.393
	ScanGAN360 (1 Hz) [37]	**54.29**	41.92	70,764	46,247	0.044	0.361
	ScanDMM (1 Hz) [53]	**54.29**	41.25	71,250	45,452	0.049	0.375
	DiffGaze (1 Hz, Ours)	54.88	41.45	71,861	45,466	0.046	0.378
	Saltinet [1]	64.73	54.35	89,722	66,610	0.022	0.249
	DeepGaze III [30]	59.48	51.94	78,083	57,813	0.071	0.407
	CLE [3]	63.01	53.63	83,846	46,037	**0.091**	**1.313**
	ScanGAN360 [37]	60.35	50.73	74,644	50,635	0.037	0.302
	ScanDMM [53]	57.22	42.80	88,513	52,559	0.049	0.475
	DiffGaze (Ours)	55.84	**39.40**	**69,381**	**44,065**	0.043	0.439
Salient360! [42, 43]	Human	99.40 †	76.44†	28,657†	15,998†	0.092†	0.479†
	ScanGAN360 (1 Hz) [37]	94.81	47.53	24,782	14,223	0.040	0.381
	ScanDMM (1 Hz) [53]	96.32	42.73	**24,347**	12,765	0.040	0.464
	Saltinet [1]	105.88	55.45	44,626	25,623	0.009	0.177
	DeepGaze III [30]	100.87	83.69	33,390	25,573	0.041	0.255
	CLE [3]	100.11	79.06	32,535	16,020	**0.088**	**1.046**
	ScanGAN360 [37]	**94.62**	53.79	25,281	15,261	0.036	0.324
	ScanDMM [53]	96.71	45.26	32,674	14,550	0.045	0.552
	DiffGaze (Ours)	98.85	**42.20**	25,238	**12,641**	0.038	0.535

† Scanpaths are not compared with themselves

MIT1003 [26]. The remaining fixations formed the scanpaths for our analyses. For the Salient360! dataset, we used the provided scanpaths as ground truth.

Baselines. In addition to the two baselines trained on fine-grained gaze sequences, we also compared our method with the state-of-the-art scanpath prediction methods, Saltinet [1], DeepGaze III [30], CLE [3], and ScanGAN360 [37] and ScanDMM [53] which trained on 1 Hz gaze sequences. For Saltinet, DeepGaze III, ScanGAN360 (1 Hz), and ScanDMM (1 Hz), we used the pre-trained weights provided by the authors. Since CLE generates scanpath on given saliency maps, we applied BMS360 [33] to obtain the saliency maps. For CLE, Saltinet, and DeepGaze III, the number of generated fixations has to be pre-defined. Therefore, we set the number to the mean number of fixations of each dataset for these methods. Moreover, to link with 1 Hz gaze sequence prediction, we trained a 1 Hz DiffGaze model using the processed 1 Hz Sitzmann dataset provided by ScanDMM authors [53] with the same setting as described in 4.2. Moreover, we compared the 1 Hz human ground truth of the Sitzmann dataset with the ground truth scanpaths extracted from human gaze sequences. Note that ScanDMM authors compared their predicted scanpaths with head movements instead of the ground truth scanpaths[1] of the Salient360! dataset [53], we compared all predicted scanpaths with ground truth scanpaths in both datasets.

[1] https://github.com/xiangjieSui/ScanDMM/issues/1

Evaluation Metrics. Following [53], we used Levenshtein distance (LEV), dynamic time warping (DTW), and recurrence measure (REC) as evaluation metrics. REC measures the proportion of gaze points that are close to each other in two scanpaths. We opted for two degrees of visual angle as the threshold following [43]. All metrics and the human baseline [62] were computed as described in Section 4.3.

Quantitative Results. Table 4 shows 1 Hz human gaze sequences exhibit superior metrics compared to scanpaths extracted from fine-grained human gaze sequences. This suggests that the scanpath prediction task is oversimplified by downsampling the human gaze sequence to 1 Hz. Consequently, on the Sitzmann dataset, the models trained with data have higher upper-bound performance than the actual human scanpath. Therefore, 1 Hz models have overall good performance. For a fair comparison, we excluded the 1 Hz models from other evaluations. However, it is worth noting that, even if generating fine-grained human gaze data is a more difficult task compared with 1 Hz data generation, the scanpaths extracted from DiffGaze's generated fine-grained gaze trajectories still outperform those 1 Hz models on best DTW, best REC, and best LEV. And it rivals those 1 Hz models in other metrics.

DiffGaze also shows promising results compared to scanpath prediction models and other baselines on both datasets. On the Sitzmann dataset, which has only three test images, DiffGaze achieves the best results and reaches human performance on LEV and DTW. On Salient360!, which has 85 test images, DiffGaze outperforms all the baselines and matches human performance on LEV and DTW, showing its strong generalisation ability in cross-dataset evaluation. Interestingly, DiffGaze has slightly higher LEV and DTW scores than human agreement, which suggests that DiffGaze produces more consistent scanpaths than human observers. As for REC, CLE has the best performance on both datasets, because REC only measures the spatial similarity of fixations and ignores the temporal aspect. CLE generates scanpaths from saliency maps, which gives it an advantage in capturing salient regions. However, DiffGaze still shows comparable performance to models without saliency prior, demonstrating its ability to generate spatially accurate fixations.

Qualitative Results. Figure 5 shows example scanpaths obtained from generated fine-grained gaze sequences of DiffGaze, ScanGAN360, and ScanDMM. It can be observed that ScanDMM tends to produce fixations that are concentrated within a limited area, thereby failing to replicate the exploratory nature of human gaze. Conversely, the scanpaths generated by our method exhibit greater resemblance to the human ground truth in terms of the number, location, and sequence of fixations. For example, in the *Museum* scene, our method successfully emulates the shift in human attention between the crowd on the left and the painting on the right. While the scanpaths produced by ScanGAN360 display a pattern akin to ours, Table 3 reveals that the average duration of fixations for ScanGAN360 is significantly reduced compared to human fixations. In contrast, the distribution of fixation durations for DiffGaze aligns closely with that of humans. Overall, the scanpaths derived from the gaze sequences generated by DiffGaze demonstrate a higher degree of similarity to human scanpaths compared to the other two baseline methods.

4.5 Saliency Prediction

The gaze sequence and scanpath evaluation focused on the temporal perspective. To better evaluate the generated fine-grained gaze sequence spatially, we evaluated the different gaze sequence generation approaches, DiffGaze, ScanGAN360 [37], and ScanDMM [53] on the saliency prediction task. We used each method to generate 1,000 30 Hz gaze sequences and applied the eye event detection algorithm mentioned in Section 4.3 to obtain gaze fixations. Similar to subsection 4.4, we filtered out all the fixations with duration less than 150 ms. We then used the script provided by Sitzmann et al. [50] to convert the fixation maps into continuous saliency maps by applying a Gaussian filter with a standard deviation of 1° of visual angle.

Baselines. Our method was trained to generate fine-grained gaze data and was not optimised for saliency prediction. Saliency prediction is a downstream task to evaluate generated fine-grained gaze trajectories. ScanDMM [53] claimed state-of-the-art saliency performance because they used a very large Gaussian kernel size (19 pixels) on a small image size (128×256) to generate the saliency map. We followed the process outlined in the Sitzmann dataset [50], using 1° of visual angle (roughly corresponding to the foveal area) as the kernel size on the image size, which corresponds to 0.71 pixels on a (128×256) image. For a fair comparison, we did not compare to deep learning saliency prediction baselines. Instead, we report results using BMS360 [33] and GBVS360 [33], which are two bottom-up saliency models that perform well compared to the state-of-the-art methods in previous saliency prediction works [6, 38, 53].

Evaluation Metrics. We report performance using five common metrics: Area under the ROC Curve (AUC), Normalised Scanpath Saliency (NSS), Similarity (SIM), Pearson's Correlation Coefficient (CC), and Kullback-Leibler divergence (KL). All metrics were calculated on image size (4096, 8192) for Sitzmann and (1024, 2048) for Salient360!, respectively.

Quantitative Results. Table 5 shows the results for saliency prediction. Our method achieves superior performance over ScanDMM and ScanGAN360 on all evaluation metrics on the Sitzmann dataset. On Salient360!, our method outperforms the two baselines on four out of five metrics, except for KL. Moreover, DiffGaze ranks second in AUC on both datasets and improves the AUC of GBVS360.

Qualitative Results. Figure 6 shows sample saliency maps obtained from generated 30 Hz gaze sequences of DiffGaze, ScanGAN360, and ScanDMM. See supplementary material for more examples. For instance, in the *Gallery* scene, our method assigns high saliency to the sculpture and most of the paintings, while the other methods mainly focus on the sculpture and the rightmost painting. Similarly, in the *Robots* scene, our method correctly identifies the salient regions, such as the largest robot and the robot in front of the cabin, while the other methods either miss them or include irrelevant regions, such as the whole cabin. These results demonstrate that our method can better capture the spatial location of human attention than the baselines.

4.6 Ablation Study

We conducted an ablation study to understand how the spatial and temporal Transformers in DiffGaze contribute to gaze sequence generation. As DiffGaze is a generative model, we evaluate the overall distribution of generated gaze sequences rather than individual sequences. Following the approach in subsection 4.5, we compared the aggregated saliency maps produced by the full model and its ablated variants (see Table 6). The full version of DiffGaze outperforms its ablated variants on four out of five metrics on the Sitzmann dataset, and on three out of five metrics on the Salient360! dataset. This suggests that the dual-Transformer architecture allows the model to capture a more accurate distribution of gaze data.

Qualitative results are shown in Figure 7 and in the supplementary material. These results show that removing the temporal Transformer leads to poor modelling of gaze dynamics: the model produces fixations mostly at the centre of the image, lacking temporal variation. When only the temporal Transformer is used, the model generates more sparse fixations across the scene, but fails to focus on key salient regions, unlike real human gaze data. In contrast, the full DiffGaze with both spatial and temporal Transformers produces fixation distributions that more closely match the ground truth. This highlights the complementary roles of the two components: the temporal Transformer is essential for capturing the sequence and timing of gaze behaviour, while the spatial Transformer helps the model attend to visually salient regions in the scene.

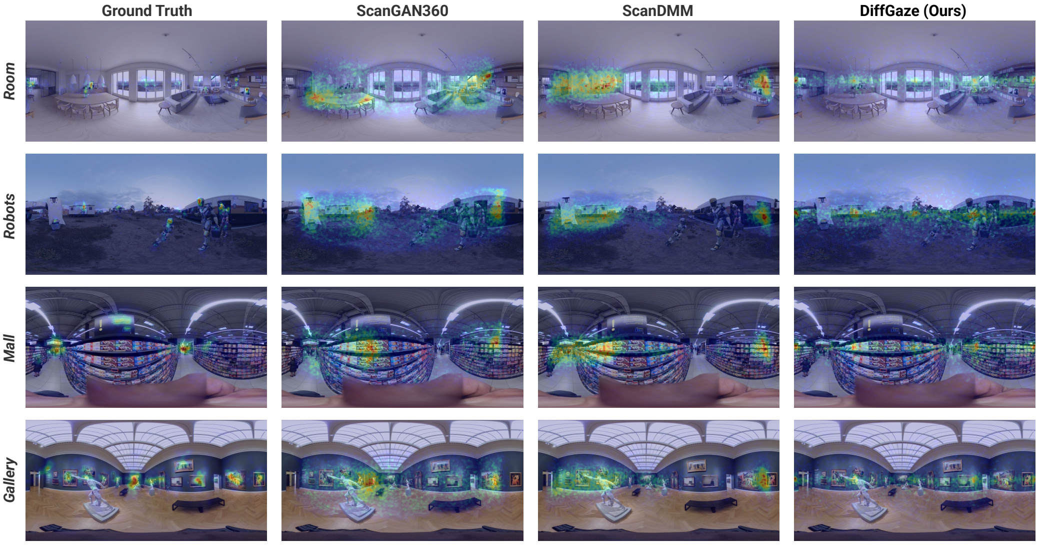

Fig. 6. Qualitative comparison to saliency prediction models in four scenes. From left to right: scanpaths obtained by a human observer, generated saliency maps obtained by ScanGAN360, ScanDMM, and the proposed model. From top to bottom: the Room and the Robots from Sitzmann dataset, the Mall and Gallery from Salient360! dataset.

Table 5. Evaluation of saliency methods on Sitzmann and Salient360! dataset. The best results of each dataset are shown in **bold**.

Dataset	Method	NSS ↑	CC ↑	AUC ↑	SIM ↑	KL ↓
Sitzmann [50]	BMS360 [33]	**1.175**	0.509	**0.883**	0.421	0.225
	GBVS360 [33]	1.144	**0.567**	0.597	**0.503**	**0.057**
	ScanDMM [53]	0.662	0.253	0.548	0.331	0.346
	ScanGAN360 [37]	0.810	0.272	0.419	0.348	0.343
	DiffGaze (Ours)	0.903	0.407	0.812	0.358	0.318
Salient360! [42, 43]	BMS360 [33]	**0.243**	**0.570**	**0.737**	**0.608**	2.146
	GBVS360 [33]	0.201	0.562	0.601	0.583	**1.609**
	ScanDMM [53]	0.149	0.368	0.454	0.476	2.468
	ScanGAN360 [37]	0.119	0.260	0.411	0.400	3.337
	DiffGaze (Ours)	0.176	0.487	0.638	0.559	2.861

4.7 Runtime Evaluation

To assess the efficiency of DiffGaze (17.9M parameters) in generating fine-grained human gaze sequences, we conducted a runtime evaluation comparing it against two baseline methods: ScanGAN360 (19.9M parameters) and ScanDMM (4.7M parameters). We measured the time required to generate a 30-second gaze sequence at 30 Hz on a single NVIDIA Tesla V100 GPU. To reduce potential bias, we repeated each method 10 times and reported the average runtime. On average,

Table 6. The saliency prediction results of the ablation study on DiffGaze components on Sitzmann and Salient360! dataset. The best results are shown in **bold**. TT: Temporal Transformer, ST: Spatial Transformer.

Dataset	Method	NSS ↑	CC ↑	AUC ↑	SIM ↑	KL ↓
Sitzmann [50]	DiffGaze w/o TT	0.487	0.290	0.446	0.266	**0.317**
	DiffGaze w/o ST	0.879	0.356	0.794	0.344	0.347
	DiffGaze (Ours)	**0.903**	**0.407**	**0.812**	**0.358**	0.318
Salient360! [42, 43]	DiffGaze w/o TT	0.080	0.275	0.433	0.251	**2.475**
	DiffGaze w/o ST	**0.185**	0.455	0.605	0.548	2.934
	DiffGaze (Ours)	0.176	**0.487**	**0.638**	**0.559**	2.861

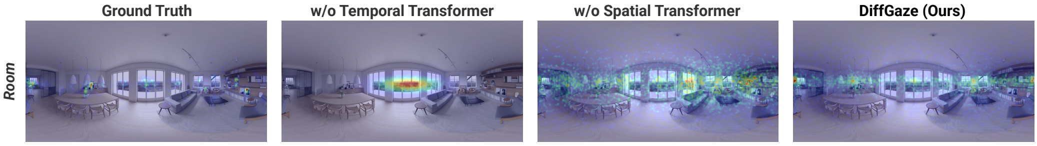

Fig. 7. Qualitative comparison between DiffGaze and its ablated versions in saliency prediction.

DiffGaze required 0.82 seconds to generate a full sequence, compared to 0.63 seconds for ScanGAN360 and 1.37 seconds for ScanDMM. The slower performance of ScanDMM can be attributed to its Markov chain-based approach, which generates each gaze point sequentially based on the previous observation state. In contrast, both ScanGAN360 and DiffGaze generate the entire sequence, resulting in faster inference times. These results suggest that DiffGaze balances generation speed and modelling fidelity, making it a practical solution for synthesising fine-grained gaze data in interactive systems.

5 DISCUSSION

5.1 Fine-grained Gaze Sequence *vs.* Discrete Scanpath in Real-world Applications

Previous works on scanpath prediction have focused on generating discrete sequences of fixations on 360° images. In contrast, we propose the first method to model and generate fine-grained gaze sequences that resemble the raw eye tracking data captured by real eye trackers. Scanpaths only contain gaze fixation information, while fine-grained gaze sequences include both fixation and saccade signals. Therefore, fine-grained gaze sequence generation can benefit not only the applications that rely on scanpath prediction, such as saliency prediction [37, 53] and image quality assessment [53], assisting data chart [61] and user interface design [21, 22].

For example, HCI and computer graphics researchers have long sought to generate fine-grained gaze sequences for digital human animation [55]. Existing methods rely on other inputs like head movements [18, 19] or speech [32]. Our method, DiffGaze, enables high frame rate animation of natural eye movements on a virtual human in a given scene. Compared to scanpath prediction approaches, it mimics more fine-grained human exploration of the virtual environment. Additionally, fine-grained gaze sequence generation holds promising potential for immersive media authoring and personalised content adaptation. Recent advances in 360° image synthesis, such as conditional scene generation for virtual interior design [48], could benefit from realistic gaze data to guide camera placement, highlight focal areas, or simulate user attention during design previews. Similarly, personalised visual content rendering methods—like the dual-level colour grading framework for 360° images [65]—could leverage individual gaze patterns to dynamically adjust tone and contrast in attention-relevant regions, enhancing both visual comfort and narrative emphasis. By incorporating high-frequency gaze data, systems can adapt to nuanced viewer preferences and perceptual cues, offering a pathway toward gaze-driven personalisation in immersive environments. Another potential application of DiffGaze is to synthesise large-scale eye tracking datasets on 360° images, which are difficult to collect due to the labour-intensive and privacy-compromising data collection process. Such datasets can help analyse human gaze behaviour in immersive environments and facilitate various gaze-based downstream tasks, such as eye event detection [66] or activity recognition [31] in VR. Moreover, contemporary commercial VR headsets often suffer from limited battery life. Recent research proposed a gaze-contingent system that reduces pixel luminance outside users' field of view to conserve power [13]. Our method provides fine-grained, realistic simulations of human gaze data for VR headsets without built-in eye trackers. Similarly, areas without predicted gaze samples can have reduced luminance to save power while maintaining users' perceptual fidelity.

5.2 Diffusion Model *vs.* Other Methods

We compared our diffusion-based model with two leading scanpath prediction methods—a GAN-based approach [37] and a Deep Markov Model (DMM) [53]—as strong baselines. As demonstrated in Figure 3 and the supplementary material, both the GAN-based and DMM-based methods cannot generate plausible fine-grained human gaze data. This is due to the significantly higher complexity of fine-grained gaze data than scanpaths that consist of only discrete fixation events. Scanpath prediction models only need to account for fixation locations. However, fine-grained gaze sequence generation models must account for various human eye movement events, including fixations and saccades. Moreover, generating a fixation involves predicting all raw gaze samples that form the fixation, not just a single location. This requires a model to generate numerous gaze samples with minimal location shifts to form a fixation, then perform several saccades to the next fixation after a human-like duration, and repeat this process.

Different individuals may exhibit different eye movements on the same image, which complicates the training of an effective discriminator in ScanGAN360. Furthermore, ScanGAN360 incorporates the spherical DTW with the loss function of the GAN generator, which directly penalises the difference between the generated samples and one ground truth. Given that one image can have multiple, vastly different ground truths, ScanGAN360 exhibits an unusual pattern in modelling fine-grained eye movement data. DMM models predict the likelihood of the next gaze location based on the current state information. As most gaze samples in fine-grained gaze sequences belong to fixations, these samples' pattern has a higher likelihood in the prediction. Since the displacements of these samples are rare, ScanDMM tends to generate large clusters as a result (see Figure 3).

Our diffusion-based approach overcomes these limitations by modelling the complex distribution of fine-grained gaze data through a forward and backward process. Compared with ScanGAN360, our method does not require a discriminator or a distance metric, which makes it easier to train and more robust to different ground truths. Moreover, unlike ScanDMM, our method simultaneously generates the whole gaze sequence, enabling it to capture human gaze behaviour's global spatial and temporal coherence. We demonstrate that our method can generate fine-grained gaze sequences that match the statistics and characteristics of human gaze data and are visually indistinguishable from real human gaze sequences.

5.3 Performance Metrics

Our work also sheds light on the challenge of evaluating the performance of attention models, as well as the shortcomings of widely used evaluation metrics and their discrepancies from human visual judgement. As discussed in subsection 4.3, human agreement performs poorly on selected time-series metrics due to the significant differences between real human gaze sequences. Additionally, we observed disagreements between the Levensthein distance and qualitative evaluation results. Specifically, ScanDMM performed the best in the best LEV on Salienct360! (see Table 2) but the generated gaze samples are visually implausible and received very low user ratings for realism in our user study. Interestingly, we also observed the disagreement between the qualitative results of generated gaze sequences and scanpath metrics. As shown in Table 4, the extracted scanpaths of ScanDMM and ScanGAN360 achieve good overall performance. Especially, for ScanGAN360, the generated 30 Hz gaze sequences are entirely visually unrealistic according to our user study result. This indicates that evaluating the scanpaths extracted from generated gaze sequences cannot reflect the method's performance on fine-grained gaze sequence generation. Statistics at the eye movement level (Table 3) and the user study results align with our visual intuitions. However, the spatial and temporal aspects are not shown in statistics. Besides, having a user study to evaluate a large scale of generated fine-grained trajectories is highly impractical. Therefore, we see an urgent need for designing effective gaze metrics on 360° images in future work.

Misalignment in the scale of evaluation metrics. Notably, a misalignment was identified between the two previous in 1 Hz scanpath evaluation in ScanDMM [53] and ScanGAN360 [37]. As the image resolution used for computing the metrics was not reported in either study, we computed metrics across various resolutions. Our findings indicate that while the metrics are sensitive to image resolution, alterations in image size do not impact the overall performance ranking. Although the scale of metrics reported in this paper differs from previous works, our method consistently outperforms baseline methods across all image resolutions.

5.4 Limitations and Future Work

The Sitzmann [50] and Salient360! [42, 43] datasets contain only 107 images, and we used 19 to train DiffGaze. The limited training data size may make the generalisability of our model unclear. However, these two are the only public 360° datasets providing raw gaze data. Once the generalisability of DiffGaze is confirmed, DiffGaze can then be modified to generate fine-grained gaze sequences for other tasks, e.g. visual search [10, 64], visual question answering [8], and natural images [4, 26]. However, among these, only the MIT1003 dataset [26] provides raw gaze data. More datasets with raw eye tracking data would enhance models like DiffGaze. We encourage the research community to make such data public.

In addition, since running time-series metrics on high-frequency gaze data is time-intensive, we only tested our method in generating 30 Hz gaze trajectories. Future work will explore our model's performance on higher sampling frequencies to see whether it can model other eye events. To detect certain types of eye events, such as microsaccades, prior work [40] suggests using the highest available sampling frequency, such as 1,000 Hz, to capture the eye tracking data. However, current 360° image eye tracking datasets typically have a maximum sampling frequency of 240 Hz. We anticipate the availability of more high-frequency VR eye tracking datasets to enhance fine-grained gaze sequence generation. In this work, we focused on bottom-up (stimulus-driven) modeling of fine-grained human gaze behaviour. For future work, we plan to explore top-down (intention-driven) gaze behaviour modeling to understand how users' intentions influence generated fine-grained gaze patterns.

6 CONCLUSION

This paper introduced DiffGaze, a conditional diffusion model for generating realistic and diverse fine-grained human gaze sequences in 360° environments. This method significantly advances the field by moving beyond the prediction of scanpaths to model more complex eye movements.

The effectiveness of DiffGaze was demonstrated through rigorous evaluation on two 360° image datasets across three different tasks. Not only did DiffGaze outperform previous methods in terms of gaze sequence generation, scanpath prediction and saliency prediction, but it also showed comparable performance with human baseline, underscoring its ability to simulate human-like gaze behaviour. These results highlight the potential of DiffGaze to facilitate further research in gaze behaviour analysis in immersive environments. By providing high-quality simulated eye-tracking data, DiffGaze opens up new possibilities for human-computer interaction and computer vision applications, paving the way for more intuitive and immersive user experiences.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon Europe research and innovation funding programme under grant agreement No. 101072410. Y. Wang was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161. The authors would like to thank Matthieu Perreira Da Silva for providing access to the raw gaze data of Salient360! dataset.

Funded by
the European Union

REFERENCES

[1] Marc Assens Reina, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. 2017. Saltinet: Scan-path prediction on 360 degree images using saliency volumes. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2331–2338.

[2] Wentao Bao and Zhenzhong Chen. 2020. Human scanpath prediction based on deep convolutional saccadic model. *Neurocomputing* 404 (2020), 154–164.

[3] Giuseppe Boccignone, Vittorio Cuculo, and Alessandro D'Amelio. 2020. How to look next? A data-driven approach for scanpath prediction. In *Formal Methods. FM 2019 International Workshops: Porto, Portugal, October 7–11, 2019, Revised Selected Papers, Part I 3*. Springer, 131–145.

[4] Ali Borji and Laurent Itti. 2015. Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581* (2015).

[5] Ali Borji, Dicky N Sihite, and Laurent Itti. 2012. Probabilistic learning of task-specific visual attention. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 470–477.

[6] Fang-Yi Chao, Lu Zhang, Wassim Hamidouche, and Olivier Deforges. 2018. Salgan360: Visual saliency prediction on 360 degree images with generative adversarial networks. In *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 01–04.

[7] Dongwen Chen, Chunmei Qing, Xiangmin Xu, and Huansheng Zhu. 2020. Salbinet360: Saliency prediction on 360 images with local-global bifurcated deep network. In *Proceedings of the 2020 IEEE Conference on Virtual Reality and 3D User Interfaces*. IEEE, 92–100.

[8] Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. 2020. Air: Attention with reasoning capability. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 91–107.

[9] Xiuli Chen, Aditya Acharya, Antti Oulasvirta, and Andrew Howes. 2021. An adaptive model of gaze-based selection. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–11.

[10] Yupei Chen, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Minh Hoai, and Gregory Zelinsky. 2021. COCO-Search18 fixation dataset for predicting goal-directed attention control. *Scientific reports* 11, 1 (2021), 1–11.

[11] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. 2015. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 3 (2015), 569–582.

[12] Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. 2018. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *Proceedings of the European conference on computer vision (ECCV)*. 518–533.

[13] Budmonde Duinkharjav, Kenneth Chen, Abhishek Tyagi, Jiayi He, Yuhao Zhu, and Qi Sun. 2022. Color-Perception-Guided Display Power Reduction for Virtual Reality. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)* 41, 6 (2022), 144:1–144:16.

[14] Ralf Engbert, Lars OM Rothkegel, Daniel Backhaus, and Hans Arne Trukenbrod. 2016. Evaluation of velocity-based saccade detection in the SMI-ETG 2W system. *Technical report, Allgemeine und Biologische Psychologie, Uni-versität Potsdam, March* (2016).

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

[16] Zhiming Hu. 2021. Eye fixation forecasting in task-oriented virtual reality. In *Proceedings of the 2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops*. IEEE, 707–708.

[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.

[18] 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.

[19] 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.

[20] 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.

[21] Yue Jiang, Zixin Guo, Hamed Rezazadegan Tavakoli, Luis A Leiva, and Antti Oulasvirta. 2024. EyeFormer: predicting personalized scanpaths with transformer-guided reinforcement learning. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–15.

[22] Yue Jiang, Luis A Leiva, Hamed Rezazadegan Tavakoli, Paul RB Houssel, Julia Kylmälä, and Antti Oulasvirta. 2023. UEyes: Understanding visual saliency across user interface types. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.

[23] 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

[24] Chuhan Jiao, Zhiming Hu, and Andreas Bulling. 2025. HAGI: Head-Assisted Gaze Imputation for Mobile Eye Trackers. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–14.

[25] Chuhan Jiao, Guanhua Zhang, Yeonjoo Cho, Zhiming Hu, and Andreas Bulling. 2024. DiffEyeSyn: Diffusion-based User-specific Eye Movement Synthesis. *arXiv preprint arXiv:2409.01240* (2024).

[26] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. 2009. Learning to Predict Where Humans Look. In *IEEE International Conference on Computer Vision (ICCV)*.

[27] Moritz Kassner, William Patera, and Andreas Bulling. 2014. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Adj. Proc. ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. 1151–1160. https://doi.org/10.1145/2638728.2641695

[28] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2020. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761* (2020).

[29] 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.

[30] Matthias Kümmerer, Matthias Bethge, and Thomas SA Wallis. 2022. DeepGaze III: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision* 22, 5 (2022), 7–7.

[31] Guohao Lan, Tim Scargill, and Maria Gorlatova. 2022. Eyesyn: Psychology-inspired eye movement synthesis for gaze-based activity recognition. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 233–246.

[32] Binh H. Le, Xiaohan Ma, and Zhigang Deng. 2012. Live Speech Driven Head-and-Eye Motion Generators. *IEEE Transactions on Visualization and Computer Graphics* 18, 11 (2012), 1902–1914. https://doi.org/10.1109/TVCG.2012.74

[33] Pierre Lebreton and Alexander Raake. 2018. GBVS360, BMS360, ProSal: Extending existing saliency prediction models from 2D to omnidirectional images. *Signal Processing: Image Communication* 69 (2018), 69–78.

[34] Sooha Park Lee, Jeremy B Badler, and Norman I Badler. 2002. Eyes alive. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*. 637–644.

[35] Huiying Liu, Dong Xu, Qingming Huang, Wen Li, Min Xu, and Stephen Lin. 2013. Semantically-based human scanpath estimation with HMMs. In *Proceedings of the IEEE International Conference on Computer Vision*. 3232–3239.

[36] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. 2018. An intriguing failing of convolutional neural networks and the coordconv solution. *Advances in neural information processing systems* 31 (2018).

[37] Daniel Martin, Ana Serrano, Alexander W Bergman, Gordon Wetzstein, and Belen Masia. 2022. Scangan360: A generative model of realistic scanpaths for 360 images. *IEEE Transactions on Visualization and Computer Graphics* 28, 5

(2022), 2003–2013.

[38] Rafael Monroy, Sebastian Lutz, Tejo Chalasani, and Aljosa Smolic. 2018. Salnet360: Saliency maps for omni-directional images with cnn. *Signal Processing: Image Communication* 69 (2018), 26–34.

[39] Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*. PMLR, 8162–8171.

[40] Marcus Nyström, Diederick C Niehorster, Richard Andersson, and Ignace Hooge. 2021. The Tobii Pro Spectrum: A useful tool for studying microsaccades? *Behavior Research Methods* 53 (2021), 335–353.

[41] Rong Quan, Yantao Lai, Mengyu Qiu, and Dong Liang. 2024. Pathformer3D: A 3D Scanpath Transformer for 360 degree Images. In *European Conference on Computer Vision*. Springer, 73–90.

[42] Yashas Rai, Jesús Gutiérrez, and Patrick Le Callet. 2017. A dataset of head and eye movements for 360 degree images. In *Proceedings of the 8th ACM on Multimedia Systems Conference*. 205–210.

[43] Yashas Rai, Patrick Le Callet, and Philippe Guillotel. 2017. Which saliency weighting for omni directional image quality assessment?. In *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 1–6.

[44] Radiah Rivu, Ville Mäkelä, Mariam Hassib, Yomna Abdelrahman, and Florian Alt. 2021. Exploring How Saliency Affects Attention in Virtual Reality. In *IFIP Conference on Human-Computer Interaction*. Springer, 147–155.

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV]

[46] Dario D Salvucci and Joseph H Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*. 71–78.

[47] Danqing Shi, Yujun Zhu, Jussi PP Jokinen, Aditya Acharya, Aini Putkonen, Shumin Zhai, and Antti Oulasvirta. 2024. CRTypist: Simulating Touchscreen Typing Behavior via Computational Rationality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.

[48] Ka Chun Shum, Hong-Wing Pang, Binh-Son Hua, Duc Thanh Nguyen, and Sai-Kit Yeung. 2023. Conditional 360-degree image synthesis for immersive indoor scene decoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4478–4488.

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

[50] 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.

[51] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

[52] Ben Steichen, Giuseppe Carenini, and Cristina Conati. 2013. User-adaptive information visualization: using eye gaze data to infer visualization tasks and user cognitive abilities. In *Proceedings of the 2013 international conference on Intelligent user interfaces*. 317–328.

[53] Xiangjie Sui, Yuming Fang, Hanwei Zhu, Shiqi Wang, and Zhou Wang. 2023. ScanDMM: A Deep Markov Model of Scanpath Prediction for 360deg Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6989–6999.

[54] 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.

[55] Marcus Thiebaux, Brent Lance, and Stacy Marsella. 2009. Real-time expressive gaze animation for virtual humans. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*. 321–328.

[56] Marc Tonsen, Julian Steil, Yusuke Sugano, and Andreas Bulling. 2017. InvisibleEye: Mobile Eye Tracking Using Multiple Low-Resolution Cameras and Learning-Based Gaze Estimation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 1, 3 (2017), 1–21. https://doi.org/10.1145/3130971

[57] Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, et al. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* 12 (2016).

[58] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. 2016. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems* 29 (2016).

[59] 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).

[60] Yao Wang, Mihai Bâce, and Andreas Bulling. 2024. Scanpath Prediction on Information Visualisations. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 30, 7 (2024), 3902–3914.

[61] Yao Wang, Weitian Wang, Abdullah Abdelhafez, Mayar Elfares, Zhiming Hu, Mihai Bâce, and Andreas Bulling. 2024. SalChartQA: Question-driven Saliency on Information Visualisations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–14.

[62] Chen Xia, Junwei Han, Fei Qi, and Guangming Shi. 2019. Predicting human saccadic scanpaths based on iterative representation learning. *IEEE Transactions on Image Processing* 28, 7 (2019), 3502–3515.

[63] Pingmei Xu, Yusuke Sugano, and Andreas Bulling. 2016. Spatio-temporal modeling and prediction of visual attention in graphical user interfaces. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 3299–3310.

[64] Zhibo Yang, Lihan Huang, Yupei Chen, Zijun Wei, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. 2020. Predicting Goal-directed Human Attention Using Inverse Reinforcement Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 193–202.

[65] Lin-Ping Yuan, John J Dudley, Per Ola Kristensson, and Huamin Qu. 2025. Personalized Dual-Level Color Grading for 360-degree Images in Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics* (2025).

[66] Raimondas Zemblys, Diederick C Niehorster, and Kenneth Holmqvist. 2019. gazeNet: End-to-end eye-movement event detection with deep neural networks. *Behavior research methods* 51 (2019), 840–864.

[67] Guanhua Zhang, Zhiming Hu, Mihai Bâce, and Andreas Bulling. 2024. Mouse2Vec: Learning Reusable Semantic Representations of Mouse Behaviour. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.

[68] Guanhua Zhang, Zhiming Hu, and Andreas Bulling. 2024. DisMouse: Disentangling Information from Mouse Movement Data. In *Proc. ACM Symposium on User Interface Software and Technology (UIST)*. 1–13.

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