Modeling Human Gaze Behavior with Diffusion Models for Unified Scanpath Prediction

[1]Giuseppe Cartella, [1]Vittorio Cuculo, [2]Alessandro D'Amelio,
[1]Marcella Cornia, [2]Giuseppe Boccignone, [1]Rita Cucchiara
[1]University of Modena and Reggio Emilia, Italy [2]University of Milan, Italy
[1]{name.surname}@unimore.it, [2]{name.surname}@unimi.it

Abstract

Predicting human gaze scanpaths is crucial for understanding visual attention, with applications in human-computer interaction, autonomous systems, and cognitive robotics. While deep learning models have advanced scanpath prediction, most existing approaches generate averaged behaviors, failing to capture the variability of human visual exploration. In this work, we present `ScanDiff`*, a novel architecture that combines diffusion models with Vision Transformers to generate diverse and realistic scanpaths. Our method explicitly models scanpath variability by leveraging the stochastic nature of diffusion models, producing a wide range of plausible gaze trajectories. Additionally, we introduce textual conditioning to enable task-driven scanpath generation, allowing the model to adapt to different visual search objectives. Experiments on benchmark datasets show that* `ScanDiff` *surpasses state-of-the-art methods in both free-viewing and task-driven scenarios, producing more diverse and accurate scanpaths. These results highlight its ability to better capture the complexity of human visual behavior, pushing forward gaze prediction research. Source code and models are publicly available at* https://aimagelab.github.io/ScanDiff.

1. Introduction

Understanding and predicting human visual attention remains a central problem in computer vision [10, 11, 15, 41, 48], with broad relevance to fields such as human-computer interaction [35], autonomous driving [52], and cognitive robotics [54]. Visual attention deployment is a dynamic and selective mechanism that allows humans to efficiently process the vast amount of information in complex visual stimuli. A critical aspect of computational modeling in this domain involves the prediction of human gaze scanpaths – the sequences of fixations and saccades that represent the dynamic process of visual exploration.

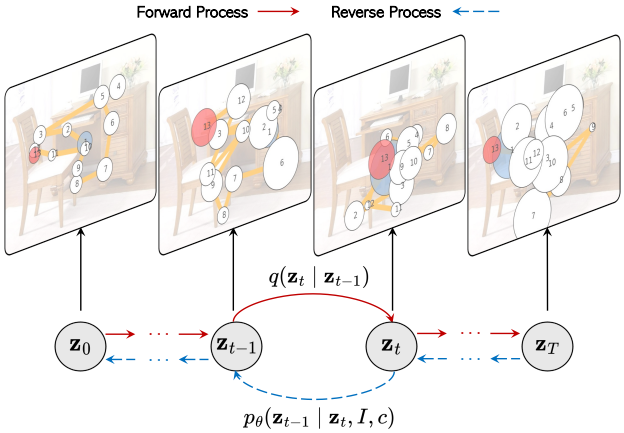

Figure 1. The diffusion process of `ScanDiff` that generates realistic scanpaths through learned transitions conditioned on image I and viewing task c.

Models based on deep convolutional [40] and recurrent architectures [17], as well as more recent Transformer-based methods [48, 69], have significantly improved the ability to predict eye movements. These models are effective in both free-viewing scenarios, where observers explore without an explicit task, and in visual search, where exploration follows predefined goals. However, most of these approaches generate scanpaths that reflect an averaged behavior, failing to capture the rich variability observed in individual visual exploration [42, 56, 63]. As noted in [9], the decision of where to look next at any given moment is neither entirely deterministic nor completely random. Modeling the effects of randomness allows us to efficiently address the influence of complex factors, such as oculomotor biases, traits, and motor response variability, at both the individual and group levels. Indeed, variability – and the resulting stochasticity of gaze allocation – goes beyond merely revealing individual idiosyncrasies, which are significant in clinical and psychological studies. It also enables the observer to remain responsive to new signals and promotes a flexible shift of attention. This flexibility, in turn, facilitates efficient learning and exploration of the en-

vironment, an essential capability for autonomous systems in computer vision and robotics [4, 58].

Recent advances in generative modeling, particularly diffusion probabilistic models [26, 32], offer a promising alternative by using stochastic sampling to learn and generate diverse sequential outputs. Early applications to scanpath prediction have shown promise in generating human-like gaze behaviors that capture the inherent variability of individual scanpaths on 360° images [36, 60] and text [8]. When combined with the sequence modeling strengths of Transformer-based architectures [27, 59], these approaches move beyond deterministic predictions to simulate a broader range of plausible scanpaths aligned with cognitive theories of attention [73].

Building on these insights, we propose `ScanDiff`, a unified architecture that integrates diffusion models with Vision Transformers [27, 51] to generate diverse and realistic gaze scanpaths. Unlike existing approaches, `ScanDiff` explicitly models scanpath variability by leveraging the stochastic nature of diffusion models, enabling the generation of diverse yet plausible gaze trajectories. Furthermore, our method incorporates textual conditioning and a length prediction module, allowing the model to flexibly adapt to diverse visual search objectives within a unified framework.

Through extensive experiments on COCO-FreeView [68], MIT1003 [37], and COCO-Search18 [16, 65], we demonstrate that `ScanDiff` sets a new state of the art in scanpath prediction across both free-viewing and task-driven scenarios. Additionally, we present a novel analysis of the variability of predicted scanpaths, highlighting that our approach generates highly diverse eye trajectories, better capturing human gaze behaviors than competitors. This is achieved by leveraging existing scanpath prediction metrics and incorporating a new measure that penalizes excessive similarity among generated scanpaths.

In summary, our key contributions are as follows:

- A novel diffusion-based architecture that models the inherent stochasticity of human gaze, enabling the generation of diverse and realistic gaze trajectories.
- A unified framework that integrates textual conditioning and a length prediction module, allowing the model to adapt to both free-viewing and task-driven scenarios.
- A comprehensive evaluation that includes a novel analysis of scanpath variability, demonstrating that `ScanDiff` outperforms existing methods in capturing the diversity of human gaze behavior, along with achieving state-of-the-art results in traditional scanpath prediction metrics.

2. Related Work

Scanpath Prediction. The study of visual attention in computer vision has seen significant progress since the seminal works in [1, 3, 5, 33]. In particular, research on modeling scanpaths – *i.e.*, the sequence of gaze fixations and subsequent shifts (saccades) – has surged in recent years, with applications expanding across multiple domains [10].

A considerable amount of research has been dedicated to predicting scanpaths under free-viewing conditions, where the observer has no predefined task [2, 7, 14, 17, 23, 33, 41]. Yet, echoing the foundational work in this field, some studies have shifted focus toward goal-driven attention modeling, in which an observer purposively engages in a specific task, such as locating an object within a scene [13, 23, 48, 66] or searching for targets not present in the image [67, 69]. Other works aimed at simulating human-like attention in visual question answering [13] and image captioning tasks [30].

Recent approaches have explored predicting attention dynamically as a person views an image while hearing a referring expression specifying the target object [46, 49]. Others have used vision-and-language models to jointly predict scanpaths and generate language-based explanations [15], or to predict subjective feedback like satisfaction and aesthetic quality alongside human attention patterns [43]. Notably, some efforts have focused on using diffusion models to generate scanpaths, though these have been limited to specific settings, such as reading [8] or viewing 360° images [36, 60]. To the best of our knowledge, we are the first to explore the potential of diffusion-based architectures for free-viewing and visual search tasks in natural scenes.

Diffusion Models for Sequence Modeling. Recently, diffusion models have emerged as one of the most successful probabilistic generative architectures across various fields, particularly in computer vision [22]. They have also gained popularity as a non-autoregressive alternative for modeling sequences [64], demonstrating success in generating various types of sequences, including continuous time-series [19, 20, 45], text [28, 44], and audio [38]. Notably, these models have recently been adopted for generating spatio-temporal data, such as GPS trajectories [61, 74, 75] and human motion data [18, 57, 72], including eye movement patterns [8, 36, 60]. In contrast to these methods, we focus on both standard free-viewing and goal-oriented settings, introducing a novel approach that can predict scanpaths of variable lengths, thereby enabling greater variability and more realistic gaze behavior.

3. Proposed Method

Scanpath generation aims to predict the spatial and temporal dynamics of human eye movements in response to a given visual stimulus. The generation can be performed under the free-viewing task or the goal-directed task, as in the case of object visual search [48, 65, 67], visual question answering [13] or incremental object referral [49]. We propose `ScanDiff`, a novel scanpath prediction architecture based on diffusion models to generate realistic and diverse gaze

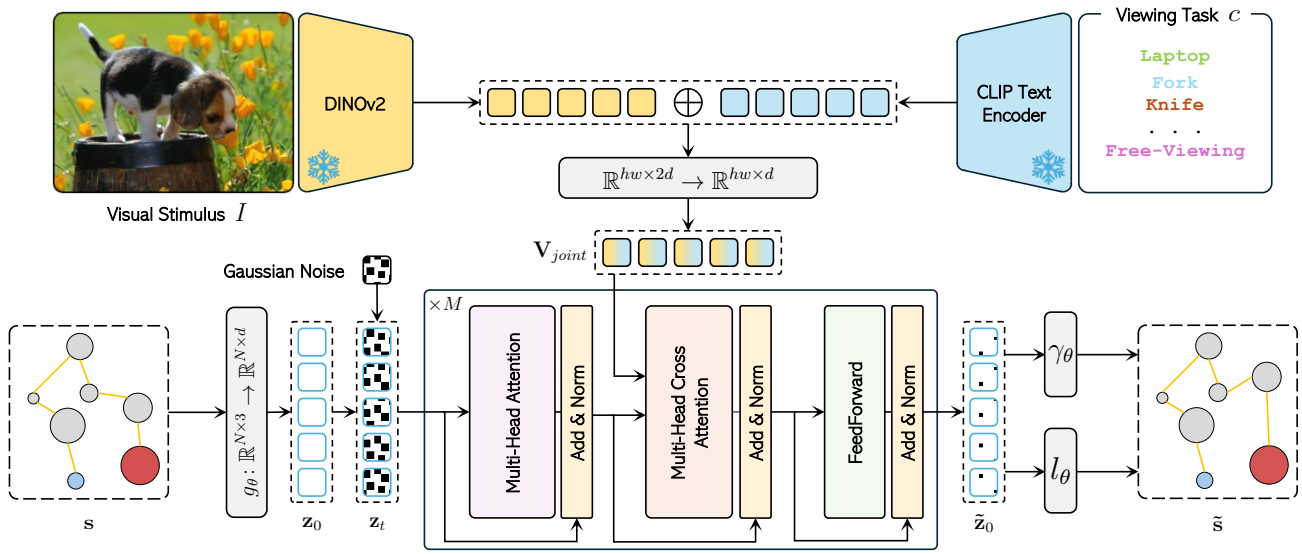

Figure 2. Overview of ScanDiff. Given a stimulus I and a viewing task c, a scanpath $\tilde{s}$ is generated through a diffusion process.

patterns (see Fig. 2). Its multimodal nature enables the unified prediction of various types of visual attention, seamlessly adapting to different viewing tasks and stimuli.

3.1. Preliminaries

Diffusion models are a class of generative models able to model the ground-truth distribution of a given dataset by reversing a diffusion process that gradually adds noise to the input data. They consist of a forward and a backward process. Given a sample $\mathbf{x}_0$ drawn from a real-world data distribution $\mathbf{x}_0 \sim q(\mathbf{x})$, the forward process gradually corrupts the input data by adding Gaussian noise for a number of timesteps T according to a variance schedule $\beta_1, \ldots, \beta_T$. This produces, at each timestep t, a latent variable $\mathbf{x}_t$ with distribution $q(\mathbf{x}_t \mid \mathbf{x}_{t-1})$, defined as:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \boldsymbol{\mathcal{I}}\right), \quad (1)$$

with $\boldsymbol{\mathcal{I}}$ being the identity matrix. In the reverse process, the final goal is to recover $\mathbf{x}_0$ by denoising $\mathbf{x}_T$. This process is defined by a Markov chain parameterized by θ:

$$p_\theta(\mathbf{x}_{0:T}) := p_\theta(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t). \quad (2)$$

In particular, each transition $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}\big(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)\big)$ is parameterized by a function ϕ_θ, where μ_θ and Σ_θ represent the predicted mean and variance of the true posterior distribution, respectively.

Problem Definition. Given an image or stimulus $I \in \mathbb{R}^{H\times W\times 3}$ and a viewing task c, the final objective is to predict a human-like scanpath represented as an ordered sequence of N fixations $\mathbf{s} = \{f_1, f_2, \ldots, f_N\}$. Each fixation f_i consists of a pair (r_i, m_i), where $r_i = (x_i, y_i) \in \mathbb{R}^2$ is the 2D spatial fixation location, while $m_i \in \mathbb{R}^+$ is the fixation duration. In this work, we propose a non-autoregressive approach to generate scanpath trajectories by learning a diffusion model ϕ_θ. Starting from a noisy sample drawn from a Gaussian distribution, the learned model iteratively refines it to produce the final scanpath trajectory.

3.2. ScanDiff Model

3.2.1. Forward Process: Scanpath Embedding

Let $\mathbf{s} \in \mathbb{R}^{N\times 3}$ be a sequence of N ground-truth fixations. To enable a structured latent space that better captures temporal and spatial dependencies of the scanpath, we learn a linear projection $g_\theta : \mathbb{R}^{N\times 3} \rightarrow \mathbb{R}^{N\times d}$ to map the scanpath $\mathbf{s}$ into an augmented embedding space, thus obtaining the initial uncorrupted latent variable $\mathbf{z}_0 = g_\theta(\mathbf{s}) \in \mathbb{R}^{N\times d}$. During the forward process (see Fig. 1), we gradually corrupt the whole embedded sequence $\mathbf{z}_0$ by adding Gaussian noise over T timesteps, following a predefined variance schedule. At each timestep t, the noisy latent representation $\mathbf{z}_t$ is obtained through a Markovian diffusion process, as defined in Eq. 1. By the final timestep, the representation $\mathbf{z}_T$ approaches an isotropic Gaussian distribution, effectively removing any trace of the original scanpath structure.

3.2.2. Conditional Denoising Process

Scanpath prediction involves the generation of an ordered sequence of N fixations conditioned on a given stimulus I and a viewing task c. Therefore, referring to Eq. 2, the conditioned denoising process can be rewritten as:

$$p_\theta(\mathbf{z}_{0:T} \mid I, c) := p(\mathbf{z}_T) \prod_{t=1}^{T} p_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t, I, c). \quad (3)$$

Our model ϕ_θ is based on an encoder-only Transformer [59], modified to incorporate an additional multi-

head cross-attention layer between the self-attention and feed-forward layers [53]. This enables the model to effectively condition on the image I and the viewing task c.

Image Encoding. We process each stimulus $I \in \mathbb{R}^{H \times W \times 3}$ using a Transformer-based visual backbone $v(\cdot)$ which outputs a dense feature map $v(I) \in \mathbb{R}^{h \times w \times d_v}$. Here, h and w denote the number of patches along the height and width of the image, respectively, while d_v refers to the dimensionality of the visual embedding space.

Task Encoding. Our model features a unified architecture that seamlessly adapts to different tasks without requiring any architectural modifications. This flexibility is achieved by using a text encoder to represent the viewing task c. Specifically, for the free-viewing task, we represent c as an empty string. In contrast, for visual search, c corresponds to the textual label of the target object to look for in the image (*e.g.* "laptop"). To extract task representations, we employ a pre-trained text encoder $\psi(\cdot)$ which maps c to a feature vector $\psi(c) \in \mathbb{R}^{d_t}$ in the textual embedding space, where d_t denotes the dimensionality of the textual features.

Multimodal Conditioning. To condition the denoising process on both the image I and task c, we project the visual features $v(I)$ and the textual features $\psi(c)$ in a joint multimodal embedding space. Following previous works [48], we first map the visual and textual features into a common d-dimensional space using two independent linear transformations. The textual features are then repeated hw times and concatenated with the visual features along the channel dimension, resulting in a feature map of size $hw \times 2d$. Finally, this feature map is linearly projected into a d-dimensional feature space to obtain the final multimodal embedding $\mathbf{V}_{joint} \in \mathbb{R}^{hw \times d}$, which effectively combines visual and task semantic information. This multimodal conditioning enables a unified model to adapt to various tasks, from free-viewing to visual search. The resulting multimodal embedding $\mathbf{V}_{joint}$ is then passed through the cross-attention layer of the Transformer. In parallel, $\mathbf{z}_t$ is augmented with a learnable positional encoding and a sinusoidal diffusion timestep embedding. For the sake of simplicity, in what follows, we refer to $\mathbf{z}_t$ as the combination of the noisy scanpath embedding, the positional, and the diffusion timestep embeddings.

Unlike previous approaches [36, 60] that directly concatenate the noisy gaze sequence with the image embedding, we combine $\mathbf{z}_t$ and the visual-semantic features $\mathbf{V}_{joint}$ only in the cross-attention layer. This design choice allows the model to dynamically modulate the interaction between gaze dynamics and visual-semantic information, rather than enforcing a rigid concatenation.

Scanpath Reconstruction and Length Prediction. The Transformer encoder output is defined as $\tilde{\mathbf{z}}_0 = \phi_\theta(\mathbf{z}_t, \mathbf{V}_{joint}) \in \mathbb{R}^{N \times d}$. The spatial coordinates and the duration of the fixations are reconstructed starting from the predicted sample $\tilde{\mathbf{z}}_0$. Specifically, a feed-forward network γ_θ with three linear layers followed by a ReLU activation function is adopted to decode $\tilde{\mathbf{z}}_0$ to an approximation of the original scanpath $\tilde{\mathbf{s}} = \{\tilde{f}_1, \tilde{f}_2, \ldots, \tilde{f}_N\}$, $\tilde{f}_i \in \mathbb{R}^3$. The visual response to a given stimulus and the corresponding scanpath length can vary across subjects. Existing works that leverage diffusion models for scanpath generation [36, 60] typically produce fixed-length scanpaths. In contrast and crucially, we take a different approach: we introduce a length prediction module in the model architecture, which allows for greater flexibility. In particular, this module predicts the probability $\tilde{u}_i$ of each token in the reconstructed sample $\tilde{\mathbf{z}}_0$ to be valid through a linear function $l_\theta : \mathbb{R}^{N \times d} \to \mathbb{R}^N$. The final predicted length is equal to the number of consecutive valid tokens.

3.3. Training and Inference

The training objective $\mathcal{L}$ is defined as the combination of four different components:

$$\mathcal{L} = \mathcal{L}_{\text{VLB}} + \mathcal{L}_{rec} + \mathcal{L}_{val} + \mathcal{L}_{T}. \tag{4}$$

The first component $\mathcal{L}_{\text{VLB}}$ aims to minimize the difference between the uncorrupted sample $\mathbf{z}_0$ and the model prediction. Formally, it is defined as:

$$\mathcal{L}_{\text{VLB}} = \sum_{t=1}^{T} \|\mathbf{z}_0 - \phi_\theta(\mathbf{z}_t, \mathbf{V}_{joint})\|^2. \tag{5}$$

This simplification can be derived from the variational lower bound [8]. To reduce the noise in the optimization of $\mathcal{L}_{\text{VLB}}$ we adopt importance sampling [50].

The second term $\mathcal{L}_{rec}$ measures the scanpath reconstruction error and is defined as the L_1 loss between the ground-truth scanpath $\mathbf{s}$ and the reconstructed one $\tilde{\mathbf{s}}$:

$$\begin{aligned} \mathcal{L}_{rec} &= \|\mathbf{s} - \tilde{\mathbf{s}}\| \\ &= \frac{1}{N} \sum_{i=1}^{N} \left(|x_i - \tilde{x}_i| + |y_i - \tilde{y}_i| + |m_i - \tilde{m}_i| \right). \end{aligned} \tag{6}$$

Here, $\tilde{x}_i$ and $\tilde{y}_i$ denote the spatial coordinates of the reconstructed fixation $\tilde{f}_i$, and $\tilde{m}_i$ is the relative fixation duration. During training, the ground-truth scanpaths are padded or truncated to a maximum length of L. The $L - N$ padding fixations are masked out during the computation of $\mathcal{L}_{rec}$.

The term $\mathcal{L}_{val}$ represents binary cross-entropy loss for predicting the validity of each reconstructed fixation:

$$\mathcal{L}_{val} = \frac{1}{L} \sum_{i=0}^{L} \text{BCE}(u_i, \tilde{u}_i). \tag{7}$$

At the final diffusion step, the mean prediction should converge to zero under the assumption that the noise prior

follows a standard Gaussian distribution. To stabilize the diffusion process, we define the loss $\mathcal{L}_T = \|\mu(\mathbf{z}_T)\|^2$ which penalizes any residual bias in the mean prediction at the final timestep ensuring it ends in a clean isotropic Gaussian. This also regularizes training by enforcing theoretical constraints otherwise weakened by finite-step approximations.

Sampling a Scanpath. At inference time, our model generates a scanpath in response to a visual stimulus I and a viewing task c (*e.g.* free-viewing or visual-search). We sample $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\mathcal{I}}) \in \mathbb{R}^{N \times d}$ and the model ϕ_θ iteratively denoises $\mathbf{z}_T$ to $\mathbf{z}_0$. At each sampling step t the multimodal features $\mathbf{V}_{joint}$ are fused with $\mathbf{z}_t$ in the cross-attention layer to condition the reverse process. After denoising $\mathbf{z}_T$ into $\mathbf{z}_0$, this is fed through two independent inverse embedding layers to obtain the predicted scanpath and its relative length N. To account for the variability in the observational patterns across different subjects, we generate multiple scanpath trajectories by sampling distinct noisy samples $\mathbf{z}_T$.

4. Experimental Results

4.1. Experimental Setup

Datasets. We evaluate our model on both the free-viewing and visual search tasks. For free-viewing experiments, we adopt COCO-FreeView [68] and MIT1003 [37], which comprise 6,202 and 1,003 images, respectively. To train our model, we combine images from both datasets, using 70% for training, 15% for validation, and 15% for testing.To assess the performance on the visual search task, we employ COCO-Search18 [16, 65], which features eye gaze behavior from 10 people while searching for the presence of a specific object (among 18 diverse categories) in the scene. Images are divided into target-present and target-absent splits, with 3,101 items each. In this setting, we employ the training, validation, and test sets used in previous works [15, 48].

Evaluation Metrics. The similarity between generated and human scanpaths is measured through the MultiMatch (MM) [25, 34], ScanMatch (SM) [21], Sequence Score (SS) [65], and Semantic Sequence Score (SemSS) [67] metrics. In particular, we adopt the same evaluation protocol proposed in [23] where the distribution of human vs. generated metrics is compared against the distribution of the human consistency metrics, using the Kullback-Leibler divergence. Beyond similarity, capturing the diversity of generated scanpaths is crucial to prevent the model from collapsing into a deterministic solution, thereby preserving the natural variability observed in human eye movements. To this end, we adopt two additional metrics: the Individual Scanpath Recall [69], which we rename as Recall Sequence Score (RSS), and a newly introduced metric that favors the diversity of generated scanpaths termed as Diversity-aware Sequence Score (DSS).

In particular, the RSS measures the extent to which the generated scanpaths cover the variability of human scanpaths for a given stimulus. For a human scanpath in the dataset, it is considered covered if its SS with at least one generated scanpath surpasses a predefined threshold. The RSS is then computed as the ratio of covered human scanpaths to the total number of human scanpaths. This metric evaluates whether the model can replicate the range of individual behaviors observed in humans.

The novel DSS we propose extends the standard sequence similarity measures by incorporating a term that penalizes excessive similarity among the generated scanpaths when humans do not reflect such behavior. Given a set of generated scanpaths $\mathbf{s}_g$ and corresponding human scanpaths $\mathbf{s}_h$ for a specific visual stimulus, DSS is computed as

$$\mathrm{DSS}(\mathbf{s}_g, \mathbf{s}_h) = \frac{\mathrm{SS}(\mathbf{s}_g, \mathbf{s}_h)}{1 + |\mathrm{SS}(\mathbf{s}_g, \mathbf{s}_g) - \mathrm{SS}(\mathbf{s}_h, \mathbf{s}_h)|} \tag{8}$$

where SS is the average sequence score calculated over the possible combinations of different scanpaths. The denominator penalizes models that produce overly uniform predictions, encouraging outputs that not only match human behavior but also reflect its natural variability.

Implementation and Training Details. We employ the DINOv2 ViT-B/14 model [51] as the pre-trained visual backbone, considering its rich semantic understanding of the visual scene. In particular, we use the DINOv2 variant with registers [24]. To align with its training resolution, we resize all images to a resolution of 518×518, resulting in a feature map $v(I)$ of 37×37 patches, with an embedding dimension of $d_v = 768$. For textual encoding, we utilize the pre-trained CLIP ViT-B/32 model [55], which projects the viewing task c into a feature space of dimension $d_t = 512$. The modified architecture of the Transformer encoder consists of $M = 6$ layers, each with 8 attention heads and a hidden dimension $d = 512$. The model is trained with the AdamW optimizer, a batch size of 128, a learning rate set to 1×10^{-4}, a weight decay of 1×10^{-2}, and a number of diffusion steps $T = 1000$. In addition, a squared-root noise schedule is adopted. Following previous works [13, 15], the maximum scanpath length is set to 16. During training, the spatial coordinates of each fixation are scaled in the range $[0, 1]$, and fixation durations are retained in seconds.

4.2. Comparison with the State of the Art

We evaluate `ScanDiff` by comparing it with existing scanpath prediction models in both free-viewing and visual search tasks. Our evaluation includes a diverse range of approaches and architectures, covering both traditional model-based methods (*e.g.* Itti-Koch [33], CLE [7], and G-Eymol [70]) and deep learning-based models (*e.g.* PathGAN [2], IOR-ROI-LSTM [17], DeepGazeIII [41], ChenLSTM [13], Gazeformer [48], HAT [69], ChenLSTM-ISP [14], GazeXplain [15], and TPP-Gaze [23]).

	COCO-FreeView										MIT1003									
	MM ↓						SM ↓		SS ↓		MM ↓						SM ↓		SS ↓	
	Sh	Len	Dir	Pos	Dur	Avg	w/ Dur	w/o Dur	w/ Dur	w/o Dur	Sh	Len	Dir	Pos	Dur	Avg	w/ Dur	w/o Dur	w/ Dur	w/o Dur
Itti-Koch [33]	0.504	0.507	0.237	1.325	-	0.643	-	4.317	-	1.624	1.040	0.702	0.353	2.900	-	1.249	-	3.233	-	6.639
CLE (Itti) [7, 33]	0.052	0.317	0.427	1.966	-	0.691	-	3.576	-	1.747	0.061	0.124	0.414	1.515	-	0.529	-	3.454	-	5.397
CLE (DG) [7, 40]	0.037	0.180	0.323	1.823	-	0.591	-	3.657	-	1.750	0.099	0.038	0.458	1.066	-	0.415	-	2.566	-	6.234
PathGAN [2]	0.070	0.406	1.009	0.073	0.031	0.318	1.210	1.383	0.718	1.012	0.063	0.234	1.603	0.514	0.165	0.516	2.255	1.069	1.087	1.190
G-Eymol [70]	0.583	0.741	1.296	0.550	0.676	0.769	9.350	8.990	8.622	4.127	0.870	0.523	0.444	0.431	0.187	0.491	9.942	2.513	9.799	3.771
IOR-ROI-LSTM [17]	1.107	0.442	0.013	0.444	0.028	0.407	1.540	1.520	0.546	1.005	0.677	0.446	0.021	1.099	0.051	0.459	0.985	0.875	0.437	5.302
DeepGazeIII [41]	0.037	0.016	0.019	0.028	-	0.025	-	0.368	-	0.393	-	-	-	-	-	-	-	-	-	-
ChenLSTM [13]	0.034	0.128	0.105	0.045	0.189	0.100	0.574	0.373	0.344	0.442	0.028	0.073	0.149	0.107	0.110	0.094	0.168	0.161	0.192	0.316
HAT [69]	1.099	0.434	0.042	0.444	-	0.505	-	1.025	-	0.331	1.196	0.522	0.381	2.386	-	1.121	-	2.112	-	1.305
ChenLSTM-ISP [14]	0.038	0.173	0.166	0.077	0.188	0.128	0.683	0.576	0.377	0.579	0.034	0.124	0.175	0.114	0.095	0.108	0.264	0.214	0.267	0.240
GazeXplain [15]	0.151	0.195	0.874	0.164	0.382	0.353	3.915	3.423	2.278	5.616	0.018	0.065	0.079	0.058	0.188	0.082	0.035	0.094	0.072	1.419
DeepGazeIII [41]	**0.009**	**0.017**	0.059	0.038	-	**0.031**	-	0.348	-	0.417	**0.025**	0.020	0.210	0.074	-	0.082	-	0.210	-	3.878
ChenLSTM [13]	0.715	0.411	0.056	0.129	0.092	0.280	0.116	0.110	0.022	0.093	0.251	0.153	0.181	0.136	0.059	0.156	0.373	0.251	0.284	0.236
GazeXplain [15]	0.346	0.226	0.032	**0.033**	0.068	0.141	0.049	0.038	0.017	**0.007**	0.060	0.046	0.065	0.025	0.044	0.048	0.158	**0.051**	0.128	**0.043**
TPP-Gaze [23]	0.063	**0.017**	0.061	0.038	**0.010**	0.038	0.125	0.226	0.033	0.130	0.039	0.036	0.139	0.068	**0.027**	0.062	0.244	0.257	0.144	0.280
ScanDiff (Ours)	0.131	0.048	**0.021**	0.037	0.151	0.078	**0.015**	**0.027**	**0.013**	0.038	0.050	**0.015**	**0.042**	**0.019**	0.072	**0.040**	**0.041**	0.065	**0.026**	0.047

Table 1. Performance comparison of different models on the COCO-FreeView [68] and MIT1003 [37] datasets. Models trained using identical settings and datasets to ScanDiff are highlighted in **gray**. Among these, the highest performance for each metric is marked in **bold**. Underlined values denote the top overall performance across all models and metrics.

Free-Viewing Results. Table 1 presents a comprehensive evaluation across the considered free-viewing datasets. For a fair comparison, the most recent models were re-trained using identical settings and datasets to ScanDiff. These results are reported in gray color at the bottom of the table.

For the COCO-FreeView dataset, ScanDiff demonstrates competitive performance across multiple metrics. Our approach achieves the best results in the MM-direction metric among models trained with identical settings. This indicates the superior ability of our model to predict saccade directions that match human scanpaths. Furthermore, our model shows strong performance in the MM-position metric and achieves the best overall SM and SS metrics when considering fixation duration, demonstrating effective modeling of temporal dynamics. The MIT1003 dataset results further validate the effectiveness of ScanDiff. Our model achieves the highest MM average score among all the competitors, demonstrating its overall strong performance. The superior results in SM and SS with duration is confirmed also on this dataset, outperforming all competing methods. This strong performance on the duration-aware metrics highlights the ability of our model to effectively represent the temporal aspects of visual attention. The unified diffusion-based architecture allows capturing both the spatial patterns of eye movements and their temporal characteristics, enabling more realistic scanpath generation.

Visual Search Results. The results summarized in Table 2 demonstrate the superior performance of our proposed model on the COCO-Search18 dataset across multiple evaluation metrics and search scenarios. We evaluate performance in both target-present and target-absent conditions, which represent fundamentally different search behaviors in human visual attention. In the target-present scenario, ScanDiff achieves state-of-the-art performance across all metrics. Most notably, our model exhibits a significant improvement in MM distributions, with an average KL divergence of 0.048, which represents a 71.3% reduction compared to the second best model (*i.e.*, GazeXplain at 0.167). For SemSS, ScanDiff attains KL divergence values of 0.072 and 0.078 with and without duration information, respectively, demonstrating consistent performance across temporal aspects of gaze behavior. The target-absent condition presents a particularly challenging scenario, as human attention patterns become more exploratory when the target object cannot be found. Even in this setting, ScanDiff outperforms all baseline methods by substantial margins.

It is worth noting that while some competing methods like GazeXplain perform reasonably well in specific metrics, they lack the consistent performance across all evaluation dimensions that ScanDiff demonstrates. This consistency across metrics and conditions indicates that our model better captures the underlying mechanisms of human visual search behavior in both goal-directed (target-present) and exploratory (target-absent) scenarios.

Qualitative Results. Fig. 3 shows some qualitative results comparing ScanDiff with other competitors in both free-viewing and visual search settings. These results confirm the effectiveness of our model also from a qualitative point-of-view, highlighting its ability in generating human-like eye movement trajectories across diverse scenarios.

4.3. Ablation Studies

To provide insights into the design choices of our model, we conduct a series of ablation studies examining the impact of different components on scanpath prediction performance. Tables 3 and 4 summarize these results across both free-viewing and visual search tasks.

Effect of Textual and Visual Backbones. We first investi-

	Target-Present							Target-Absent						
	MM ↓	SM ↓		SS ↓		SemSS ↓		MM ↓	SM ↓		SS ↓		SemSS ↓	
	Avg	w/ Dur	w/o Dur	w/ Dur	w/o Dur	w/ Dur	w/o Dur	Avg	w/ Dur	w/o Dur	w/ Dur	w/o Dur	w/ Dur	w/o Dur
PathGAN [2]	0.513	1.891	2.451	0.808	0.939	0.313	0.468	0.125	0.792	0.357	0.498	0.944	0.212	0.465
ChenLSTM [13]	0.197	0.011	0.236	0.040	0.262	0.084	0.162	0.075	0.010	0.012	0.036	0.775	0.044	0.677
Gazeformer [48]	0.281	0.027	0.340	0.119	0.268	0.131	0.208	0.089	0.061	0.075	0.102	0.245	0.085	0.139
HAT [69]	0.118	-	0.263	-	0.148	-	3.837	0.052	-	0.063	-	0.097	-	3.472
ChenLSTM-ISP [14]	0.174	0.013	0.306	0.015	0.257	0.043	0.097	0.082	0.028	0.146	0.063	0.727	0.052	0.561
GazeXplain [15]	0.166	0.023	0.237	0.070	0.232	0.140	0.188	0.046	0.062	0.048	0.038	0.191	0.043	0.206
TPP-Gaze [23]	0.524	1.618	3.218	0.579	1.590	0.554	1.147	0.098	0.511	0.529	0.242	0.325	0.093	0.164
Gazeformer [48]	0.251	0.045	0.508	0.048	0.349	0.095	0.262	0.526	1.184	1.688	0.319	0.520	0.671	1.043
GazeXplain [15]	0.167	**0.010**	0.238	0.050	0.217	0.092	0.224	0.037	0.030	0.026	0.028	0.146	0.038	0.143
TPP-Gaze [23]	0.507	2.317	3.995	0.893	2.381	0.736	1.605	0.135	0.775	0.887	0.427	0.537	0.231	0.300
ScanDiff (Ours)	**0.048**	0.037	**0.079**	**0.019**	**0.074**	**0.072**	**0.078**	**0.020**	**0.005**	**0.008**	**0.008**	**0.031**	**0.007**	**0.024**

Table 2. Performance comparison of different models on the COCO-Search18 dataset [16, 65] for both target-present and target-absent settings. Models trained using identical settings and training splits to ScanDiff are highlighted in **gray**. Among these, the highest performance for each metric is marked in **bold**. Underlined values denote the top overall performance across all models and metrics.

	Backbones		COCO-FreeView			COCO-Search18			
	Textual	Visual	MM ↓	SM ↓	SS ↓	MM ↓	SM ↓	SS ↓	SemSS ↓
Effect of Textual Backbone									
	RoBERTa	DINOv2	0.110	0.181	0.111	0.076	0.143	0.072	0.132
ScanDiff	CLIP	DINOv2	**0.078**	**0.015**	**0.013**	**0.048**	**0.037**	**0.019**	**0.072**
Effect of Visual Backbone									
	CLIP	RN50	**0.049**	0.019	0.029	0.070	0.052	0.031	0.084
	CLIP	CLIP	0.058	0.199	0.112	0.090	0.079	0.033	**0.069**
ScanDiff	CLIP	DINOv2	0.078	**0.015**	**0.013**	**0.048**	**0.037**	**0.019**	0.072

Table 3. Performance comparison of different textual and visual backbones on COCO-FreeView [66] and COCO-Search18 [16] (TP) datasets. Best results are highlighted in **bold**.

gate the influence of textual and visual backbones on model performance. As shown in Table 3, replacing the CLIP text encoder with RoBERTa [47] leads to a degradation in performance across all metrics on both datasets. This highlights the importance of vision-language pre-training for scanpath prediction, as CLIP's joint embedding space better captures the semantic relationships between textual queries and visual features that guide human attention.

For the visual backbone comparison, we test our model with ResNet-50-FPN [29], CLIP visual encoder (always using the ViT-B version), and DINOv2. The results indicate that while ResNet-50 achieves the best MM average score on COCO-FreeView, DINOv2 consistently outperforms other visual backbones across most metrics, particularly on the more challenging COCO-Search18 dataset. Interestingly, the CLIP visual encoder performs best on the SemSS metric, suggesting its strength in capturing semantic relationships between fixations and image regions, likely due to its vision-language pre-training.

Effect of $\mathcal{L}_T$ Loss Function. We evaluate the contribution of the diffusion prior alignment loss $\mathcal{L}_T$. As shown in Table 4, including $\mathcal{L}_T$ improves overall performance, but the benefits are more pronounced on the COCO-Search18 dataset, where it improves MM, SM and SemSS. This suggests that the convergence loss is particularly valuable for modeling the sequential nature of fixations in goal-directed

			COCO-FreeView			COCO-Search18			
	$\mathcal{L}_T$	T	MM ↓	SM ↓	SS ↓	MM ↓	SM ↓	SS ↓	SemSS ↓
Effect of $\mathcal{L}_T$ Loss Function									
	✗	1000	0.088	**0.011**	**0.009**	0.058	0.040	**0.018**	0.076
ScanDiff	✓	1000	**0.078**	0.015	0.013	**0.048**	**0.037**	0.019	**0.072**
Varying Diffusion Timesteps									
	✓	200	**0.043**	0.072	0.064	0.069	0.040	0.023	0.085
	✓	500	0.049	0.145	0.085	**0.046**	0.050	0.030	0.098
	✓	1500	0.045	0,131	0.122	0.057	0.086	0.061	0.130
ScanDiff	✓	1000	0.078	**0.015**	**0.013**	0.048	**0.037**	**0.019**	**0.072**

Table 4. Ablation study on the effect of alignment loss ($\mathcal{L}_T$) and diffusion timesteps (T) on COCO-FreeView [66] and COCO-Search18 [16] (TP) datasets. Best results are highlighted in **bold**.

visual search tasks, where the temporal order of fixations is more structured compared to free-viewing scenarios.

Validating Diffusion Timesteps. Finally, we investigate the impact of diffusion timesteps (T) on model performance by varying T from 200 to 1500. As shown in Table 4, $T = 1000$ achieves the best overall balance across metrics and datasets. While smaller timesteps (*e.g.*, $T = 200$) can achieve better MM scores on COCO-FreeView, they underperform on the other spatial metrics. Similarly, $T = 500$ achieves a slightly better MM score on COCO-Search18, but at the cost of poorer performance on other metrics. Increasing T beyond 1000 (*i.e.*, $T = 1500$) leads to deteriorated performance across most metrics, likely due to overfitting to noise patterns. These results highlight the critical role of properly calibrating the diffusion process: sufficient timesteps are needed to learn complex distributions, but excessive noise can degrade the ability of the model to capture meaningful patterns in scanpath data.

4.4. Scanpath Variability Analysis

Human visual exploration is inherently variable. Individuals perceive the same stimulus in different manners depending on factors such as attention, context, and cognitive processes [6, 31]. Capturing such variability is essential for developing models that accurately reflect the diverse range of

Figure 3. Comparison of simulated and human scanpaths across different datasets for both free-viewing and visual search tasks. From top to bottom: results on COCO-FreeView [66], MIT1003 [37], COCO-Search18 TP [16], and COCO-Search18 TA [67] datasets.

	COCO-FV		MIT1003		TP		TA	
	DSS ↑	RSS ↑	DSS ↑	RSS ↑	DSS ↑	RSS ↑	DSS ↑	RSS ↑
IOR-ROI-LSTM [17]	0.185	0.428	0.264	0.579	-	-	-	-
ChenLSTM [13]	0.174	0.420	0.257	0.534	0.386	0.635	0.247	0.591
Gazeformer [48]	-	-	-	-	0.377	0.578	0.206	0.417
HAT [69]	-	0.645	-	0.615	-	0.861	-	0.748
ChenLSTM-ISP [14]	0.190	0.501	0.264	0.619	0.423	0.735	0.268	0.670
GazeXplain [15]	0.099	0.032	0.302	0.674	0.406	0.689	0.283	0.716
TPP-Gaze [23]	0.271	0.732	0.313	0.758	0.284	0.516	0.221	0.663
ScanDiff (Ours)	**0.277**	**0.736**	**0.354**	**0.815**	**0.425**	**0.747**	**0.312**	**0.800**

Table 5. Analysis of scanpath variability on free-viewing (COCO-FreeView [66] and MIT1003 [37]) and task oriented datasets (COCO-Search18 target-present [16] and target-absent [67]).

human traits. However, existing scanpath prediction models tend to align closely with the statistical mean of human gaze behavior. While this approach may improve performance on traditional evaluation metrics, it fails to reflect the natural variability in human visual attention. Commonly used metrics such as MM, SM, and SS tend to reward predictions that closely match an aggregated ground truth, thus favoring models that generate a single representative scanpath [39]. This is clear in several works [13, 15, 48] where scanpath models surpass human consistency. Indeed, the average similarity between ground-truth scanpaths can be smaller than the average similarity between generated scanpaths if these well reflect an average behavior.

Building upon these considerations, we present a first attempt to quantitatively assess the ability of a model to generate diverse, yet human-like, gaze trajectories. Specifically, for this study, we adopt the DSS and RSS metrics introduced in Sec. 4.1. Table 5 reports the results on both free-viewing and visual search. Notably, ScanDiff achieves the best overall performance on all settings and datasets, highlighting its effectiveness in predicting accurate eye movement trajectories well aligned with the human scanpath variability. Interestingly, the performance gap between ScanDiff and state-of-the-art methods becomes even more evident in the visual search task and further supported by qualitative results in the Supplementary Material. Goal-oriented scanpaths tend to be more deterministic [12, 71], particularly in the target-present setting, and are generally shorter than those in free-viewing scenarios. Nevertheless, our model effectively captures even the more subtle variability present in human gaze behavior.

5. Conclusion

In this paper, we introduced ScanDiff, a novel diffusion-based architecture for scanpath prediction that significantly advances the state-of-the-art by modeling the inherent stochastic nature of human visual attention. Experimental results on multiple benchmark datasets demonstrate that ScanDiff not only achieves state-of-the-art performance in traditional scanpath prediction metrics but also generates more diverse scanpaths that better capture the variability inherent in human visual exploration. This diversity is crucial for applications requiring realistic simulation of human visual behavior, such as human-computer interaction, autonomous systems, and cognitive robotics. These results highlight the importance of modeling stochasticity in visual attention deployment, suggesting that future research in gaze prediction should consider the probabilistic nature of human gaze beyond deterministic approaches.

Acknowledgments

We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support. This work was supported by the PNRR project "Italian Strengthening of Esfri RI Resilience (ITSERR)" funded by the European Union - NextGenerationEU (CUP B53C22001770006).

References

[1] J Aloimonos, Isaac Weiss, and Amit Bandyopadhyay. Active Vision. In *ICCV*, 1987. 2

[2] Marc Assens, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. Pathgan: Visual scanpath prediction with generative adversarial networks. In *ECCV Workshops*, 2018. 2, 5, 6, 7, 1

[3] Ruzena Bajcsy and Mario Campos. Active and Exploratory Perception. *CVGIP: Image Understanding*, 56(1):31–40, 1992. 2

[4] Ruzena Bajcsy, Yiannis Aloimonos, and John K Tsotsos. Revisiting active perception. *Autonomous Robots*, 42:177–196, 2018. 2

[5] Dana H Ballard. Animate Vision. *Artificial Intelligence*, 48 (1):57–86, 1991. 2

[6] James W Bisley. The neural basis of visual attention. *The Journal of physiology*, 589(1):49–57, 2011. 7

[7] Giuseppe Boccignone and Mario Ferraro. Modelling gaze shift as a constrained random walk. *Physica A: Statistical Mechanics and its Applications*, 331(1-2):207–218, 2004. 2, 5, 6, 1

[8] Lena S Bolliger, David R Reich, Patrick Haller, Deborah N Jakobi, Paul Prasse, and Lena A Jäger. ScanDL: A diffusion model for generating synthetic scanpaths on texts. In *EMNLP*, 2023. 2, 4

[9] Roxanne L Canosa. Real-world vision: Selective perception and task. *ACM Transactions on Applied Perception*, 6(2): 1–34, 2009. 1

[10] Giuseppe Cartella, Marcella Cornia, Vittorio Cuculo, Alessandro D'Amelio, Dario Zanca, Giuseppe Boccignone, and Rita Cucchiara. Trends, Applications, and Challenges in Human Attention Modelling. In *IJCAI*, 2024. 1, 2

[11] Giuseppe Cartella, Vittorio Cuculo, Marcella Cornia, and Rita Cucchiara. Unveiling the Truth: Exploring Human Gaze Patterns in Fake Images. *IEEE Signal Processing Letters*, 31: 820–824, 2024. 1

[12] Monica S Castelhano, Michael L Mack, and John M Henderson. Viewing task influences eye movement control during active scene perception. *Journal of vision*, 9(3):6–6, 2009. 8

[13] Xianyu Chen, Ming Jiang, and Qi Zhao. Predicting Human Scanpaths in Visual Question Answering. In *CVPR*, 2021. 2, 5, 6, 7, 8, 1, 3, 4, 9

[14] Xianyu Chen, Ming Jiang, and Qi Zhao. Beyond average: Individualized visual scanpath prediction. In *CVPR*, 2024. 2, 5, 6, 7, 8, 1

[15] Xianyu Chen, Ming Jiang, and Qi Zhao. GazeXplain: Learning to Predict Natural Language Explanations of Visual Scanpaths. In *ECCV*, 2024. 1, 2, 5, 6, 7, 8, 3, 4, 9

[16] Yupei Chen, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Minh Hoai, and Gregory Zelinsky. COCO-Search18 fixation dataset for predicting goal-directed attention control. *Scientific Reports*, 11(1):1–11, 2021. 2, 5, 7, 8, 1

[17] Zhenzhong Chen and Wanjie Sun. Scanpath Prediction for Visual Attention using IOR-ROI LSTM. In *IJCAI*, 2018. 1, 2, 5, 6, 8, 3, 4

[18] Chen Chu, Hengcai Zhang, Peixiao Wang, and Feng Lu. Simulating human mobility with a trajectory generation framework based on diffusion model. *Int. J. of Geographical Information Science*, 38(5):847–878, 2024. 2

[19] Andrea Coletta, Sriram Gopalakrishnan, Daniel Borrajo, and Svitlana Vyetrenko. On the constrained time-series generation problem. In *NeurIPS*, 2023. 2

[20] Jonathan Crabbé, Nicolas Huynh, Jan Stanczuk, and Mihaela Van Der Schaar. Time series diffusion in the frequency domain. In *ICML*, 2024. 2

[21] Filipe Cristino, Sebastiaan Mathôt, Jan Theeuwes, and Iain D Gilchrist. ScanMatch: A novel method for comparing fixation sequences. *Behavior Research Methods*, 42(3): 692–700, 2010. 5

[22] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Trans. PAMI*, 45(9):10850–10869, 2023. 2

[23] Alessandro D'Amelio, Giuseppe Cartella, Vittorio Cuculo, Manuele Lucchi, Marcella Cornia, Rita Cucchiara, and Giuseppe Boccignone. TPP-Gaze: Modelling Gaze Dynamics in Space and Time with Neural Temporal Point Processes. In *WACV*, 2025. 2, 5, 6, 7, 8, 1, 3, 4, 9

[24] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, 2024. 5

[25] Richard Dewhurst, Marcus Nyström, Halszka Jarodzka, Tom Foulsham, Roger Johansson, and Kenneth Holmqvist. It depends on how you look at it: Scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach. *Behavior Research Methods*, 44(4):1079–1100, 2012. 5

[26] Prafulla Dhariwal and Alexander Nichol. Diffusion Models Beat GANs on Image Synthesis. In *NeurIPS*, 2021. 2

[27] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021. 2

[28] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. DiffuSeq: Sequence to Sequence Text Generation with Diffusion Models. In *ICLR*, 2022. 2

[29] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 7

[30] Sen He, Hamed R Tavakoli, Ali Borji, and Nicolas Pugeault. Human attention in image captioning: Dataset and analysis. In *ICCV*, 2019. 2

[31] John M Henderson. Human gaze control during real-world scene perception. *Trends in cognitive sciences*, 7(11):498–504, 2003. 7

[32] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *NeurIPS*, 2020. 2

[33] L. Itti, C. Koch, and E. Niebur. A Model of Saliency-based Visual Attention for Rapid Scene Analysis. *IEEE Trans. PAMI*, 20:1254–1259, 1998. 2, 5, 6, 1

[34] Halszka Jarodzka, Kenneth Holmqvist, and Marcus Nyström. A Vector-based, Multidimensional Scanpath Similarity Measure. In *ETRA*, 2010. 5

[35] Yue Jiang, Luis A Leiva, Hamed Rezazadegan Tavakoli, Paul RB Houssel, Julia Kylmälä, and Antti Oulasvirta. UEyes: Understanding Visual Saliency across User Interface Types. In *ACM CHI*, 2023. 1

[36] Chuhan Jiao, Yao Wang, Guanhua Zhang, Mihai Bâce, Zhiming Hu, and Andreas Bulling. DiffGaze: A Diffusion Model for Continuous Gaze Sequence Generation on 360° Images. *arXiv preprint arXiv:2403.17477*, 2024. 2, 4, 1

[37] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *ICCV*, 2009. 2, 5, 6, 8

[38] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In *ICLR*, 2021. 2

[39] Matthias Kümmerer and Matthias Bethge. State-of-the-art in human scanpath prediction. *arXiv preprint arXiv:2102.12239*, 2021. 8

[40] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep Gaze I: Boosting saliency prediction with feature maps trained on ImageNet. *arXiv preprint arXiv:1411.1045*, 2014. 1, 6

[41] Matthias Kümmerer, Matthias Bethge, and Thomas SA Wallis. DeepGaze III: Modeling free-viewing human scanpaths with deep learning. *J. of Vision*, 22(5):7–7, 2022. 1, 2, 5, 6, 3

[42] Thomas Le Bras, Benoit Allibe, and Karine Doré-Mazars. The way we look at an image or a webpage can reveal personality traits. *Scientific Reports*, 14(1):15488, 2024. 1

[43] Peizhao Li, Junfeng He, Gang Li, Rachit Bhargava, Shaolei Shen, Nachiappan Valliappan, Youwei Liang, Hongxiang Gu, Venky Ramachandran, Yang Li, et al. UniAR: A Unified model for predicting human Attention and Responses on visual content. In *NeurIPS*, 2024. 2

[44] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-LM Improves Controllable Text Generation. In *NeurIPS*, 2022. 2

[45] Haksoo Lim, Minjung Kim, Sewon Park, and Noseong Park. Regular Time-series Generation using SGM. *arXiv preprint arXiv:2301.08518*, 2023. 2

[46] Xiaochuan Liu, Xin Cheng, Yuchong Sun, Xiaoxue Wu, Ruihua Song, Hao Sun, and Denghao Zhang. Eyear: Learning audio synchronized human gaze trajectory based on physics-informed dynamics. *arXiv preprint arXiv:2502.20858*, 2025. 2

[47] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019. 7

[48] Sounak Mondal, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Gregory Zelinsky, and Minh Hoai. Gazeformer: Scalable, Effective and Fast Prediction of Goal-Directed Human Attention. In *CVPR*, 2023. 1, 2, 4, 5, 7, 8, 6, 9

[49] Sounak Mondal, Seoyoung Ahn, Zhibo Yang, Niranjan Balasubramanian, Dimitris Samaras, Gregory Zelinsky, and Minh Hoai. Look Hear: Gaze Prediction for Speech-directed Human Attention. In *ECCV*, 2024. 2

[50] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 4

[51] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning Robust Visual Features without Supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 5

[52] Anwesan Pal, Sayan Mondal, and Henrik I Christensen. Looking at the Right Stuff - Guided Semantic-Gaze for Autonomous Driving. In *CVPR*, 2020. 1

[53] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 4

[54] Kun Qian, Zhuoyang Zhang, Wei Song, and Jianfeng Liao. Gvgnet: Gaze-directed visual grounding for learning underspecified object referring intention. *IEEE RA-L*, 8(9), 2023. 1

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 5

[56] Evan F Risko, Nicola C Anderson, Sophie Lanthier, and Alan Kingstone. Curious eyes: Individual differences in personality predict eye movement behavior in scene-viewing. *Cognition*, 122(1):86–90, 2012. 1

[57] Yiwen Song, Jingtao Ding, Jian Yuan, Qingmin Liao, and Yong Li. Controllable human trajectory generation using profile-guided latent diffusion. *ACM Trans. KDD*, 19(1):1–25, 2024. 2

[58] Yang Tang, Chaoqiang Zhao, Jianrui Wang, Chongzhen Zhang, Qiyu Sun, Wei Xing Zheng, Wenli Du, Feng Qian, and Jürgen Kurths. Perception and navigation in autonomous systems in the era of learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):9604–9624, 2022. 2

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2, 3

[60] Yujia Wang, Fang-Lue Zhang, and Neil A Dodgson. ScanTD: 360° Scanpath Prediction based on Time-Series Diffusion. In *ACM Multimedia*, 2024. 2, 4, 1

[61] Tonglong Wei, Youfang Lin, Shengnan Guo, Yan Lin, Yiheng Huang, Chenyang Xiang, Yuqing Bai, and Huaiyu Wan. Diff-RNTraj: A Structure-Aware Diffusion Model for Road Network-Constrained Trajectory Generation. *IEEE Trans. KDD*, 36(12):7940–7953, 2024. 2

[62] Juan Xu, Ming Jiang, Shuo Wang, Mohan S Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *J. of Vision*, 14(1):28–28, 2014. 1, 2

[63] Jumpei Yamashita, Yoshiaki Takimoto, Haruo Oishi, and Takatsune Kumada. How do personality traits modulate real-world gaze behavior? generated gaze data shows situation-dependent modulations. *Frontiers in Psychology*, 14:1144048, 2024. 1

[64] Yiyuan Yang, Ming Jin, Haomin Wen, Chaoli Zhang, Yuxuan Liang, Lintao Ma, Yi Wang, Chenghao Liu, Bin Yang, Zenglin Xu, et al. A survey on diffusion models for time series and spatio-temporal data. *arXiv preprint arXiv:2404.18886*, 2024. 2

[65] Zhibo Yang, Lihan Huang, Yupei Chen, Zijun Wei, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. Predicting Goal-Directed Human Attention Using Inverse Reinforcement Learning. In *CVPR*, 2020. 2, 5, 7

[66] Zhibo Yang, Lihan Huang, Yupei Chen, Zijun Wei, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. Predicting Goal-Directed Human Attention Using Inverse Reinforcement Learning. In *CVPR*, 2020. 2, 7, 8

[67] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Target-absent human attention. In *ECCV*, 2022. 2, 5, 8

[68] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Predicting Human Attention using Computational Attention. *arXiv preprint arXiv:2303.09383*, 2023. 2, 5, 6

[69] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Ruoyu Xue, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Unifying top-down and bottom-up scanpath prediction using transformers. In *CVPR*, 2024. 1, 2, 5, 6, 7, 8

[70] Dario Zanca, Stefano Melacci, and Marco Gori. Gravitational laws of focus of attention. *IEEE Trans. PAMI*, 42(12): 2983–2995, 2020. 5, 6, 1

[71] Gregory J Zelinsky. A theory of eye movements during target acquisition. *Psychological review*, 115(4):787, 2008. 8

[72] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Trans. PAMI*, 46(6):4115–4128, 2024. 2

[73] Minglu Zhao, Dehong Xu, and Tao Gao. From cognition to computation: A comparative review of human attention and transformer architectures. *arXiv preprint arXiv:2407.01548*, 2024. 2

[74] Yuanshao Zhu, Yongchao Ye, Shiyao Zhang, Xiangyu Zhao, and James Yu. DiffTraj: Generating GPS Trajectory with Diffusion Probabilistic Model. In *NeurIPS*, 2023. 2

[75] Yuanshao Zhu, James Jianqiao Yu, Xiangyu Zhao, Qidong Liu, Yongchao Ye, Wei Chen, Zijian Zhang, Xuetao Wei, and Yuxuan Liang. Controltraj: Controllable trajectory generation with topology-constrained diffusion model. In *KDD*, 2024. 2