ST-DIFFEYE: Diffusion-based Continuous Gaze Generation via Joint Scanpath-Trajectory Modeling

Brian Nlong Zhao **Ozgur Kara**[†] **Junho Kim** **James M. Rehg**

University of Illinois Urbana-Champaign

Project Page Code Model

Abstract

We study the problem of human gaze modeling, which aims to generate the gaze patterns a viewer produces while observing a visual stimulus. Gaze is primarily captured through two modalities: continuous eye-tracking trajectories, which describe fine-grained motion dynamics, and discrete scanpaths, which describe high-level fixation structure. Because gaze varies substantially across viewers and trials, we treat this variability as a defining property rather than noise and model gaze as a stochastic generative process. Existing generative gaze models supervise on only one of these two representations in isolation. We hypothesize that trajectories and scanpaths describe gaze at complementary scales and are jointly informative during training, and test this hypothesis through ST-DIFFEYE, which jointly denoises trajectories and scanpaths by representing them as input and output channels within a diffusion framework, requiring no architectural changes beyond channel expansion. At sampling time, the denoised outputs are post-processed into final scanpaths. We further introduce a principled evaluation framework based on the Continuous Ranked Probability Score (CRPS), which jointly assesses the accuracy and diversity of generated gaze using existing distance or similarity measures. Experiments on task-driven visual search, covering both target-present and target-absent scenarios, and on free-viewing benchmarks demonstrate state-of-the-art performance. These results, along with detailed ablations, confirm the benefit of joint modeling and the value of distribution-aware evaluation in capturing the intrinsic variability of human gaze.

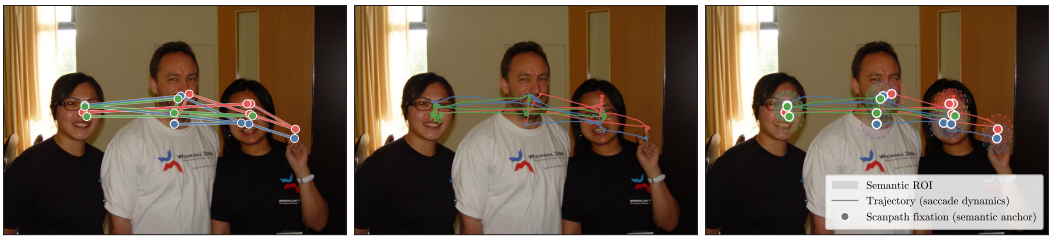
1 Introduction

Human visual attention can be measured via eye tracking, which records a viewer's gaze position over time to produce a continuous trajectory of high-frequency (x, y) coordinates [1, 2]. In the standard analysis pipeline, the continuous eye tracking trajectories are converted into a discrete representation called a scanpath, by identifying the fixations and the saccades that connect them [3, 4]. The scanpath representation is used in a broad set of application domains, ranging from cognitive psychology and advertising to virtual reality [5, 6].

With the emergence of generative AI, it is natural to ask whether realistic gaze behavior can be synthesized by a generative model that implicitly captures the dynamics of human attention. This is a challenging task because gaze emerges from a complex interaction between low-level image features, top-down goals, and individual traits, leading to high variability across subjects and even repeated viewings [7, 8]. Prior synthesis works have focused on two task conditions: 1) *free-viewing*, in which the observer explores the stimulus image without any explicit goal; and 2) *visual search*, in which the observer actively scans the stimulus to locate a designated target object category. Prior works have

[†]Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(**a**) Scanpaths (**b**) Trajectories (**c**) Trajectories + Scanpaths

Figure 1: **Complementary representations of human gaze.** Prior work models gaze using either scanpaths for fixation-level structure (a) or continuous trajectories for fine-grained motion dynamics (b). Our approach jointly leverages both (c), aligning fixation anchors with the underlying trajectory along a shared time axis.

adopted the scanpath representation for three reasons. First, the scanpath provides a low-dimensional output representation whose length is bounded by the expected number of fixations. Second, a variety of widely-used metrics exist for comparing real and synthesized scanpaths. Third, scanpaths are the predominant data modality in public eye tracking datasets. As a consequence, scanpath generation has emerged as the canonical task in gaze synthesis, and an on-going line of work has built increasingly expressive scanpath generators, from recurrent and reinforcement-learning formulations to recent diffusion-based models, in both free-viewing and visual search regimes [9–11].

However, recent works [12, 13] have departed from this convention by training generative models directly on raw eye-tracking trajectories. In particular, Kara et al. [12] demonstrated that training a trajectory generator to produce continuous gaze trajectory outputs, and then postprocessing the generated trajectories into scanpaths, yielded better performance in comparison to direct scanpath generation. They argued that continuous trajectories preserve information, for example saccade kinematics and gaze dynamics during fixation, which is useful for generative modeling and absent from the scanpath representation.

This paper advances the idea that continuous gaze trajectories and their associated scanpaths provide *complementary* sources of information for gaze synthesis: Scanpaths provide a coarse structural summary of *where* attention dwells, while eye tracking trajectories provide a fine-grained signal for *how* the eye moves between fixations (Figure 1). This decomposition is reminiscent of prior work on multi-scale representations in computer vision [14, 15]. We hypothesize that a generative model trained jointly on scanpaths and trajectories can exploit their complementary cues and produce samples that are both structurally coherent at the fixation level and dynamically variable at the trajectory level.

To test this hypothesis, we represent each scanpath as a continuous fixation index sequence of the same temporal length as its corresponding trajectory, and append it as additional channels to the trajectory data so that the diffusion model jointly denoises both modalities from a single augmented input. Scanpaths thus actively participate in the denoising process rather than serving only as auxiliary supervision, while the generated trajectories are post-processed into final scanpaths at sampling time. This formulation requires no architectural changes beyond a channel expansion in the data input and output, and induces a shared latent geometry in which the fixation structure of scanpaths and the continuous dynamics of trajectories are aligned along a common time axis. We show that the resulting joint representation improves both scanpath- and trajectory-level performance across free-viewing and task-driven benchmarks.

In addition, we observe that the standard sequence-level metrics compare generated scanpaths against human references using mean or best-match scores [13, 12], which reward point-estimate accuracy but do not capture distributional fidelity. KL-divergence-based protocols [10, 9] compare distributions of metric values, but provide only an indirect assessment that depends on the chosen metric. As a result, a model that repeatedly generates a small set of high-quality scanpaths may appear competitive with one that captures a diverse, human-like gaze distribution. To address this gap, we introduce a family of evaluation metrics built on the Continuous Ranked Probability Score (CRPS) [16, 17], which jointly assesses distributional accuracy and diversity using existing distance- or score-based sequence metrics. For negative-type distances, the resulting score retains the theoretical propriety guarantee; for other scanpath measures, we use it as a practical distribution-aware evaluation protocol. We evaluate ST-DIFFEYE on COCO-FreeView [18], MIT1003 [19], and the task-driven COCO-Search18 [20]

benchmarks, and report consistent gains over both trajectory-only and scanpath-only baselines. Our contributions are summarized as follows:

- A joint scanpath-trajectory modeling approach that couples both modalities through joint denoising via channel-level concatenation, exploiting their complementary information without additional architectural complexity.
- A CRPS-based metric family that incorporates existing sequence distance or similarity measures to jointly evaluate the distributional accuracy and diversity of generative gaze models, with formal propriety guaranteed for negative-type distances.
- State-of-the-art performance on free-viewing and task-driven benchmarks, with ablations confirming the contribution of joint training and the informativeness of the proposed metrics in distinguishing models that capture human gaze variability from those that do not.

2 Related Works

Scanpath Generation. Saliency prediction has a long history in visual perception research [21–23, 1, 24], and following Itti's computational model [25] became an active area of study in computer vision [26–36]. Saliency maps, however, capture only where attention is likely to land, not the temporal order in which a viewer visits those locations, motivating a shift to scanpath generation: given an image and an optional target category for task-driven search, the model predicts an ordered sequence of fixations. Early approaches sampled fixations from hand-crafted [25, 22] or deep [33, 34, 32] saliency maps, while later work modeled the fixation sequence directly. Recurrent and reinforcement-learning models [37, 38, 18] predict scanpaths autoregressively, and DeepGaze III [11] couples a saliency backbone with a fixation density model. Transformer-based models [39–41] extend this to longer-range dependencies and unified free-viewing/visual-search formulations, while individualized variants [42, 43, 10] condition on observer or task identity. Most of these models produce a single scanpath per stimulus and approximate average viewer behavior, leaving inter-subject variability unmodeled. Diffusion-based scanpath generation has only recently emerged: ScanDL [44] models reading scanpaths over text, and ScanDiff [9] extends this to natural images with textual conditioning for task-driven search. These models capture stochasticity at the fixation level but still supervise on scanpaths and discard the underlying continuous gaze signal.

Trajectory Generation. A second line of work supervises generative models on the raw eye-tracking trajectory, retaining the timing and geometry of saccades and the high-frequency dynamics during fixations. DiffEye [12] is, to our knowledge, the first diffusion-based model trained directly on continuous trajectories from natural images, and reports gains on scanpath-level metrics after converting the generated trajectory into a fixation sequence. Related efforts include DiffEyeSyn [45], which synthesizes user-specific high-frequency eye movements, and an autoregressive diffusion model for gaze on 360° panoramic content [13]. These methods establish that continuous gaze can be generated, but treat the trajectory as the sole training signal and do not exploit the discrete fixation structure that scanpath-based models make explicit. Our work bridges these two lines by training a single diffusion model jointly on trajectories and scanpaths, so that fine-grained dynamics and fixation-level structure inform a shared latent representation.

3 Continuous Gaze Generation via Joint Trajectory-Scanpath Modeling

3.1 Problem Formulation

Human visual attention can be represented as a stochastic spatio-temporal process conditioned on a visual stimulus and, optionally, a viewing task. To capture this process, we define a dataset $\mathcal{D}$ comprising N visual stimuli. We use the subscript $i \in \{1, \dots, N\}$ to index the specific image $I_i \in \mathbb{R}^{H \times W \times 3}$, and the superscript $j \in \{1, \dots, J_i\}$ to index the J_i distinct human subjects who viewed it. The complete dataset is

$$\mathcal{D} = \left\{ \left(I_i, c_i, \{R_i^j, S_i^j\}_{j=1}^{J_i} \right) \right\}_{i=1}^{N}, \quad R_i^j = \{(x_t^{\mathrm{tr}}, y_t^{\mathrm{tr}})\}_{t=1}^{T_i^j}, \quad S_i^j = \{(x_m^{\mathrm{sp}}, y_m^{\mathrm{sp}}, \Delta t_m^{\mathrm{sp}})\}_{m=1}^{M_i^j} \quad (1)$$

where c_i is the optional task-related conditioning variable (such as a target category in visual search), T_i^j is the number of recorded time steps in the raw trajectory R_i^j, and M_i^j is the number of fixations

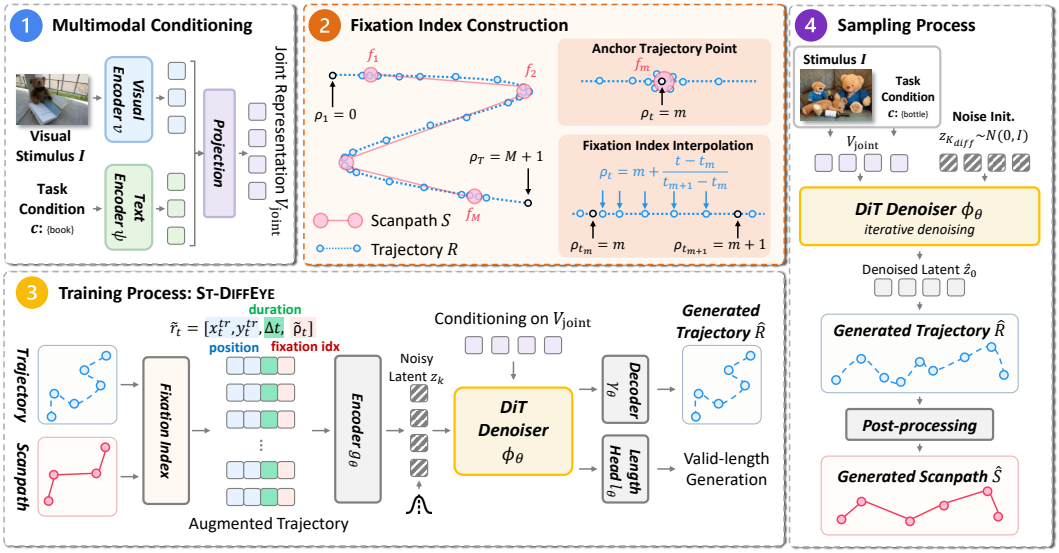


Figure 2: **Overview of ST-DIFFEYE.** (1) The visual stimulus I and task condition c are encoded by v and ψ and fused into a joint representation V_{joint}. (2) Each trajectory point is assigned a continuous fixation index $\tilde{\rho}_t$ by anchoring fixations to their nearest trajectory points and linearly interpolating between anchors. (3) The augmented trajectory tokens are denoised by a DiT ϕ_θ conditioned on V_{joint}, then decoded into a trajectory with a predicted valid length. (4) At sampling time, the generated trajectory $\hat{R}$ is post-processed via fixation extraction to produce the final scanpath $\hat{S}$.

in the scanpath S_i^j, with $(x_m^{\text{sp}}, y_m^{\text{sp}})$ the spatial centroid and Δt_m^{sp} the duration of the m-th fixation. The superscripts tr and sp distinguish trajectory and scanpath coordinates throughout.

Our primary goal is to learn a generative model of human gaze behavior. To train such models, previous approaches typically rely on either continuous trajectories R or discrete scanpaths S in isolation. In contrast, our framework utilizes both R and S jointly during training to capture their complementary information. Ultimately, we obtain a learned generative distribution $p_\theta(R \mid I, c)$ that, conditioned on a new visual stimulus I and an optional task variable c, produces a continuous trajectory R, which is then postprocessed to yield the corresponding discrete scanpath S.

3.2 ST-DIFFEYE

Overview. Motivated by the capacity of DiTs [46], and unlike the U-Net backbone of prior gaze models [12], we follow ScanDiff [9] in adopting a DiT and adapt it to operate directly on continuous trajectory data. To guide the generative process, we extract contextual features from the visual stimulus via a visual backbone $v(\cdot)$ and task-specific conditions through a text encoder $\psi(\cdot)$, fusing them into a joint multimodal representation $V_{\text{joint}} \in \mathbb{R}^{hw \times d}$. Rather than treating continuous trajectories and discrete scanpaths in isolation, we bridge them by augmenting each trajectory token with an additional channel that encodes the higher-level scanpath structure through a continuous *fixation index*. The augmented sequence is projected into a latent token space $z_0 = g_\theta(\tilde{R}) \in \mathbb{R}^{T \times d}$ and corrupted through a forward diffusion process. The denoising network ϕ_θ models spatial and temporal dependencies of the gaze sequence and fuses the noisy gaze tokens z_k with the multimodal condition V_{joint} using self/cross-attention. Finally, to account for the natural variation in human gaze duration, the predicted clean latent $\hat{z}_0 = \phi_\theta(z_k, V_{\text{joint}})$ is processed by a decoder γ_θ alongside a length-prediction head l_θ that estimates token validity, enabling variable-length generation. An overview is shown in Figure 2.

Fixation Index. While continuous trajectories provide rich spatio-temporal dynamics, scanpaths encode higher-level semantic structure through discrete fixations. We exploit the natural spatial alignment between these two modalities to inject scanpath information into our trajectory modeling pipeline. Intuitively, trajectories can be viewed as a high-resolution signal, whereas scanpaths represent a compressed, low-resolution abstraction that preserves the most salient information. In

particular, fixations correspond to semantically meaningful regions, while the intermediate saccades are more stochastic and less structured. Our goal is therefore to regularize these intermediate trajectory segments by aligning them with fixation-level structure.

To this end, we introduce a simple yet effective representation, termed the *fixation index*, which assigns each trajectory point a continuous value indicating its relative position within the scanpath. Dropping the per-image and per-subject indices (i, j) for clarity, we work with a single trajectory $R = \{(x_t^{\mathrm{tr}}, y_t^{\mathrm{tr}})\}_{t=1}^{T}$ and its corresponding scanpath $S = \{(x_m^{\mathrm{sp}}, y_m^{\mathrm{sp}}, \Delta t_m^{\mathrm{sp}})\}_{m=1}^{M}$ as defined above; the durations Δt_m^{sp} are not used in this construction. We first identify a set of *anchor trajectory points* in the trajectory by aligning each fixation to its nearest trajectory point:

$$t_m = \arg\min_t \left\| (x_t^{\mathrm{tr}}, y_t^{\mathrm{tr}}) - (x_m^{\mathrm{sp}}, y_m^{\mathrm{sp}}) \right\|_2, \quad m = 1, \ldots, M, \tag{2}$$

while enforcing temporal consistency $1 \le t_1 < t_2 < \cdots < t_M \le T$. We additionally define boundary anchors $t_0 = 1$ and $t_{M+1} = T$. We then construct a fixation index sequence $\rho = \{\rho_t\}_{t=1}^{T}$, where each anchor point is assigned its corresponding index:

$$\rho_{t_m} = m, \quad m = 1, \ldots, M, \qquad \rho_{t_0} = 0, \quad \rho_{t_{M+1}} = M + 1. \tag{3}$$

For trajectory points between two consecutive anchors t_m and t_{m+1}, we assign values by linear *fixation index interpolation*:

$$\rho_t = m + \frac{t - t_m}{t_{m+1} - t_m}, \quad t \in [t_m, t_{m+1}], \quad m = 0, \ldots, M. \tag{4}$$

This yields a continuous-valued index that encodes the progression of each trajectory point within the fixation sequence. Values close to m indicate proximity to the m-th fixation, while intermediate values such as $m + 0.5$ correspond to mid-saccade transitions. Finally, we normalize the index as $\tilde{\rho}_t = \rho_t/(M + 1)$ and include the per-sample time step $\Delta t = 1/\text{sample rate}$ to encode the temporal resolution of the trajectory. The final augmented trajectory token is

$$\tilde{r}_t = [x_t^{\mathrm{tr}},\, y_t^{\mathrm{tr}},\, \Delta t,\, \tilde{\rho}_t]. \tag{5}$$

This augmentation provides each trajectory point with explicit global context about its role in the overall gaze process, enabling the model to better structure stochastic saccadic motion while preserving the semantic organization induced by fixations.

Training. Given the augmented trajectory sequence $\tilde{R} = \{\tilde{r}_t\}_{t=1}^{T}$ with $\tilde{r}_t = [x_t^{\mathrm{tr}}, y_t^{\mathrm{tr}}, \Delta t, \tilde{\rho}_t]$, we first project the sequence into a latent representation $z_0 = g_\theta(\tilde{R})$. Following the Denoising Diffusion Probabilistic Model (DDPM) framework [47], we define a forward diffusion process that gradually adds Gaussian noise to z_0 over K steps according to a predefined variance schedule $\beta_1, \ldots, \beta_K$, given by $q(z_k \mid z_{k-1}) = \mathcal{N}(z_k;\, \sqrt{1 - \beta_k}\, z_{k-1},\, \beta_k I)$. This formulation allows us to sample the noisy latent z_k at an arbitrary step k directly from z_0 in closed form as $q(z_k \mid z_0) = \mathcal{N}(z_k;\, \sqrt{\bar{\alpha}_k}\, z_0,\, (1 - \bar{\alpha}_k)I)$, where $\alpha_k = 1 - \beta_k$ and $\bar{\alpha}_k = \prod_{i=1}^{k} \alpha_i$. The denoising network ϕ_θ is trained to reverse this process and recover z_0, conditioned on the diffusion step k and the multimodal stimulus representation V_{joint}. Following ScanDiff [9], we adopt the same training objective (see Appendix for implementation details), which minimizes a weighted sum of four components:

$$\mathcal{L} = \mathcal{L}_{\mathrm{diff}} + \lambda_{\mathrm{recon}}\mathcal{L}_{\mathrm{recon}} + \lambda_{\mathrm{val}}\mathcal{L}_{\mathrm{val}} + \lambda_{\mathrm{reg}}\mathcal{L}_{\mathrm{reg}}. \tag{6}$$

The diffusion loss $\mathcal{L}_{\mathrm{diff}}$ supervises the latent denoising process. The reconstruction loss $\mathcal{L}_{\mathrm{recon}}$ ensures that the decoded output $\hat{R} = \gamma_\theta(\hat{z}_0)$ accurately maps the predicted clean latent back to the augmented trajectory space. The validity loss $\mathcal{L}_{\mathrm{val}}$ trains the length-prediction head to distinguish valid trajectory tokens from padding elements. The terminal regularization loss $\mathcal{L}_{\mathrm{reg}}$ encourages the distribution of the final diffusion state z_K to match an isotropic Gaussian prior.

Sampling. Given a visual stimulus and task condition, we first sample $z_K \sim \mathcal{N}(0, I)$ and iteratively apply the learned reverse diffusion process conditioned on V_{joint} to obtain the predicted clean latent $\hat{z}_0$. The decoder γ_θ then maps $\hat{z}_0$ back to the augmented trajectory representation $\hat{R} = \{\hat{r}_t\}_{t=1}^{T}$, where each predicted token is $\hat{r}_t = [\hat{x}_t^{\mathrm{tr}}, \hat{y}_t^{\mathrm{tr}}, \Delta t, \hat{\rho}_t]$. Since the fixation index is introduced only as an auxiliary structural signal for trajectory generation, we discard the predicted index channel $\hat{\rho}_t$ after decoding and retain the continuous trajectory $\{(\hat{x}_t^{\mathrm{tr}}, \hat{y}_t^{\mathrm{tr}})\}_{t=1}^{T}$ together with the fixed step size Δt. Finally, we convert the generated trajectory into a discrete scanpath using the fixation extraction algorithm provided by [19]. By sampling different initial noise sequences z_K for the same stimulus, the model produces multiple plausible trajectories and scanpaths that reflect the stochastic variability of human gaze. More implementation details are shown in Appendix Sections C and D.

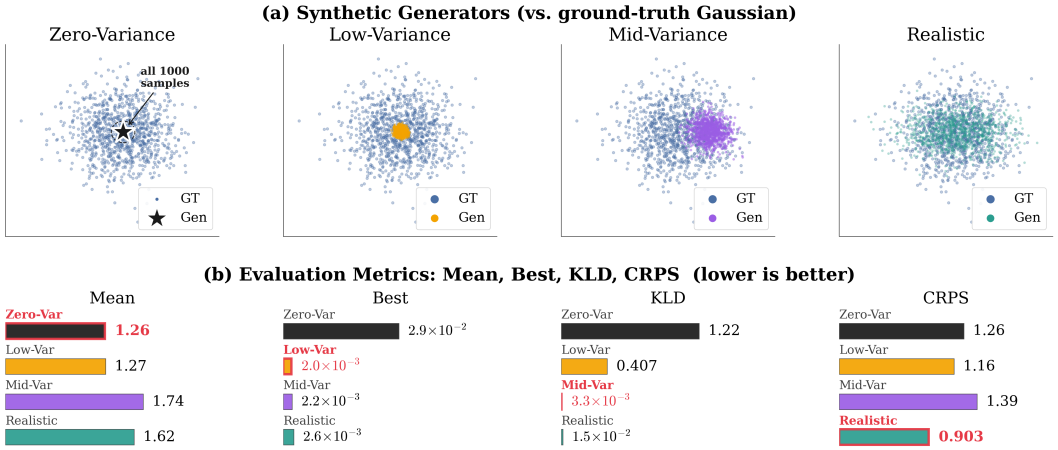


Figure 3: **Failure modes of mean, best, and KLD evaluation protocols.** (a) Four generators evaluated against the same ground-truth Gaussian (blue): zero-variance, low-variance, mid-variance, and realistic. (b) Scores under each protocol (lower is better). Each non-CRPS protocol is won by a different degenerate generator (mean: zero-variance; best: low-variance; KLD: mid-variance), while only CRPS correctly identifies the realistic generator.

3.3 CRPS-based Evaluation

Generative gaze models are commonly evaluated under two point-estimate protocols [12, 13]: a *mean* protocol averaging a similarity metric over all generated–GT pairs, and a *best* protocol reporting only the closest generated sample per GT. Both have well-known failure modes, illustrated on synthetic 2D Gaussian generators in Figure 3: the mean protocol rewards mode collapse (zero-variance generator), while the best protocol rewards low-variance generators that produce at least one close sample per GT but otherwise ignore the multi-modality of human gaze (low-variance generator). ScanDiff [9] and TPP-Gaze [10] instead adopt a Kullback-Leibler divergence (KLD) protocol over score histograms, but the same histogram can arise from very different sample distributions (mid-variance generator).

We therefore propose a distribution-aware protocol based on the Continuous Ranked Probability Score (CRPS) [16], inspired by [17], which evaluates the predictive distribution directly by jointly measuring accuracy and diversity. Given ground-truth samples $\mathbf{x} = \{x^{(i)}\}_{i=1}^{N}$ and generated samples $\hat{\mathbf{x}} = \{\hat{x}^{(j)}\}_{j=1}^{J}$,

$$\mathrm{CRPS}(\hat{\mathbf{x}}, \mathbf{x}, d) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{J}\sum_{j=1}^{J} d\left(\hat{x}^{(j)}, x^{(i)}\right) - \frac{1}{2J^2}\sum_{j=1}^{J}\sum_{k=1}^{J} d\left(\hat{x}^{(j)}, \hat{x}^{(k)}\right)\right]. \tag{7}$$

The first term pulls generated samples toward the ground truth, while the second penalizes collapse through their pairwise distances. For negative-type distances, such as Euclidean distance, this score is proper and is minimized in expectation by the true distribution. For other scanpath measures such as DTW, ScanMatch, MultiMatch, and Sequence Score, we use it as a practical distribution-aware metric without claiming formal propriety. Thus, the protocol jointly evaluates accuracy and diversity, resolving the illustrated failure modes of mean, best, and KLD, with only the realistic generator winning.

4 Experiments

4.1 Evaluation Setup

Datasets and Preprocessing. We evaluate our framework on two canonical gaze generation tasks: task-driven visual search and free-viewing. Each sample comprises an image stimulus, a discrete scanpath, and its corresponding continuous gaze trajectory, with visual search samples additionally including a target object description. For visual search, we use **COCO-Search18** [20] (6,202 images, 18 target categories, 10 subjects, ∼300K fixations). For free-viewing, we use **COCO-FreeView** [18]

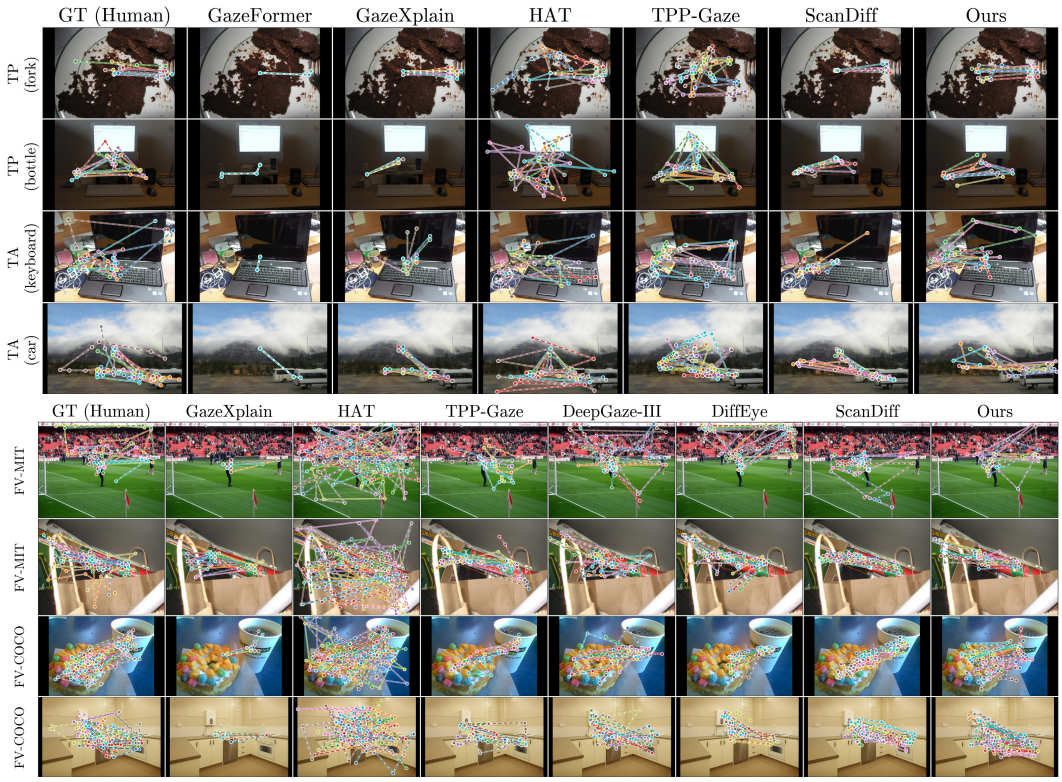


Figure 4: **Qualitative comparison with baselines.** Generated scanpaths on COCO-Search18 (top, target-present and target-absent rows) and the free-viewing benchmarks COCO-FreeView and MIT1003 (bottom).

(the same COCO images under a free-viewing protocol, >800K fixations) and **MIT1003** [19] (1,003 images, 15 subjects). Because only discrete scanpaths are publicly available for the COCO datasets, we processed the original EDF files to extract the raw continuous trajectories. To align these newly extracted trajectories with the public scanpaths, we trim the leading and trailing trajectory segments that fall outside a 20-pixel radius of the initial and final fixations. All trajectories are then downsampled to 240 Hz and padded or truncated to a maximum of 1500 points for COCO-FreeView data and 750 for others, while scanpaths are bounded to a maximum of 16 fixations. Image preprocessing, text feature extraction, and train/test splits strictly follow ScanDiff [9].

Metrics. We primarily evaluate models using the CRPS protocol from Section 3.3, instantiated with two complementary metric families. Following DiffEye [12], we use *distance-based* metrics that quantify geometric and temporal misalignment: Levenshtein Distance (LD), Discrete Fréchet Distance (DFD), Dynamic Time Warping (DTW), and Time Delay Embedding (TDE). Following ScanDiff [9], we use *similarity-based* metrics: MultiMatch (MM), ScanMatch (SM), Sequence Score (SS), and Semantic Sequence Score (SSS). SSS is reported only for visual search since it relies on target-category supervision. For SM, SS, and SSS, we additionally report duration-aware variants (†) that incorporate fixation durations. KLD, Mean, Best-protocol results are included in the appendix.

Baselines. We compare against representative gaze models grouped by modeling strategy. The first group covers non-diffusion scanpath predictors: GazeFormer [39], a task-conditioned transformer producing a single sequence per stimulus; HAT [40], a transformer unifying top-down and bottom-up scanpath prediction; and DeepGaze III [11], a saliency-based probabilistic model for free-viewing. The second adds extra conditioning or stochasticity: GazeXplain [43] uses language-based supervision, and TPP-Gaze [10] models fixation arrivals with neural temporal point processes. The third comprises diffusion-based generative models: ScanDiff [9] generates scanpaths with a task-conditioned diffusion transformer, and DiffEye [12] generates continuous trajectories and converts them into scanpaths for free-viewing. ScanDiff and DiffEye are most directly comparable to our

Table 1: **CRPS metrics on free-viewing and visual-search datasets.** Datasets: COCO-FV (COCO-FreeView), MIT-FV (MIT1003), COCO-TP (COCO-Search18, Target Present), COCO-TA (COCO-Search18, Target Absent). ScanMatch (SM), Sequence Score (SS), and Semantic Sequence Score (SSS); † marks the duration-aware variant. **Bold**: best; underlined: second best within each dataset and comparison block. Dashes indicate methods that do not predict durations, or metrics that do not apply (SSS is undefined for free-viewing).

Dataset	Comparison	Method	LD↓	DFD↓	DTW↓	TDE↓	MM↑	SM↑	SM† ↑	SS↑	SS† ↑	SSS↑	SSS† ↑
COCO-FV	*Scanpath*	GazeXplain [43]	95.512	168.087	1418.642	0.059	0.323	−0.232	−0.231	0.099	−0.220	–	–
		HAT [40]	50.825	106.551	944.485	0.052	0.413	0.099	–	−0.170	–	–	–
		TPP-Gaze [10]	64.219	109.774	1076.181	**0.035**	0.385	0.055	0.022	0.163	0.019	–	–
		DeepGaze III [11]	66.587	104.779	998.289	0.040	**0.414**	0.033	–	−0.235	–	–	–
		DiffEye [12]	56.368	121.355	1116.594	0.051	0.370	0.025	−0.018	0.197	−0.029	–	–
		ScanDiff [9]	**41.856**	104.744	910.433	0.049	0.387	**0.102**	0.089	**0.274**	0.075	–	–
		ST-DIFFEYE	47.108	**103.009**	**851.792**	0.048	0.377	0.089	**0.108**	0.206	**0.088**	–	–
	Trajectory	DiffEye [12]	56.349	121.493	1120.990	0.051	0.289	0.025	−0.196	**0.187**	−0.147	–	–
		ST-DIFFEYE	**53.487**	**105.121**	**957.552**	**0.044**	**0.299**	**0.028**	**−0.097**	0.102	**−0.090**	–	–
MIT-FV	*Scanpath*	GazeXplain [43]	32.979	115.568	640.062	0.071	0.350	0.009	0.049	0.285	0.036	–	–
		HAT [40]	34.782	119.965	667.106	0.084	0.399	0.080	–	−0.153	–	–	–
		TPP-Gaze [10]	26.109	103.796	502.235	**0.052**	0.378	0.112	0.094	0.255	0.089	–	–
		DeepGaze III [11]	26.787	102.286	465.898	0.058	**0.419**	**0.130**	–	−0.107	–	–	–
		DiffEye [12]	27.530	111.506	516.821	0.062	0.377	0.112	0.125	0.328	0.121	–	–
		ScanDiff [9]	25.189	102.216	496.327	0.058	0.379	0.111	0.118	**0.342**	0.128	–	–
		ST-DIFFEYE	**24.961**	**100.903**	**456.985**	0.058	0.380	0.129	**0.145**	0.321	**0.146**	–	–
	Trajectory	DiffEye [12]	**26.351**	111.252	496.530	0.061	**0.378**	**0.123**	0.129	**0.333**	0.124	–	–
		ST-DIFFEYE	26.589	**108.085**	**486.137**	**0.060**	0.373	0.114	**0.139**	0.287	**0.133**	–	–
COCO-TP	*Scanpath*	GazeFormer [39]	15.115	181.157	284.039	0.099	0.327	0.044	0.008	**0.540**	0.052	0.001	−0.045
		GazeXplain [43]	13.808	184.248	283.833	0.094	0.334	0.066	0.066	0.535	0.102	0.039	0.037
		HAT [40]	10.784	122.199	227.450	0.080	**0.398**	0.201	–	0.197	–	0.185	–
		TPP-Gaze [10]	25.299	179.339	476.733	0.121	0.348	−0.014	−0.049	0.180	−0.088	−0.036	−0.095
		ScanDiff [9]	9.653	124.455	219.855	0.076	0.371	0.165	0.104	0.521	0.148	0.164	0.089
		ST-DIFFEYE	**8.716**	**110.038**	**196.631**	**0.074**	0.381	**0.206**	**0.146**	0.504	**0.180**	**0.211**	**0.134**
COCO-TA	*Scanpath*	GazeFormer [39]	32.439	188.240	622.325	0.100	0.308	−0.161	−0.171	0.293	−0.170	−0.141	−0.160
		GazeXplain [43]	28.302	152.507	514.880	0.074	0.350	−0.009	0.001	**0.347**	0.010	0.002	0.009
		HAT [40]	24.371	117.266	439.956	**0.064**	**0.399**	0.095	–	−0.051	–	0.100	–
		TPP-Gaze [10]	27.822	125.431	545.535	0.085	0.370	0.035	0.007	0.186	−0.001	0.038	−0.006
		ScanDiff [9]	21.252	119.689	426.350	0.069	0.378	0.086	0.077	0.305	0.090	0.109	0.092
		ST-DIFFEYE	**21.128**	**115.683**	**410.444**	**0.064**	0.381	**0.110**	**0.095**	0.288	**0.109**	**0.130**	**0.109**

setting, capturing stochasticity at the fixation and trajectory levels respectively, while we model both within a single diffusion process.

4.2 Analysis and Discussion

Main results. We report CRPS-based results across all four datasets in Table 1. **Visual search.** On COCO-Search18, ST-DIFFEYE achieves the strongest overall performance in both target-present and target-absent regimes, obtaining the best score on 9 of 11 metrics in COCO-TP and 8 of 11 in COCO-TA. The largest improvements over ScanDiff, the closest diffusion-based scanpath baseline, are on the distance-based metrics (DFD, DTW, TDE) and the duration-aware variants (SM†, SS†, SSS†), suggesting that trajectory information provides complementary fine-grained temporal cues beyond discrete fixation sequences. **Free-viewing.** On COCO-FreeView and MIT1003, ST-DIFFEYE again leads the scanpath comparison block, ranking first on 4 of 9 and 5 of 9 metrics respectively, with consistent gains on DFD, DTW, and the duration-aware similarity metrics. The improvements are less pronounced than in visual search, which we attribute to the greater inter-subject variability of unconstrained free-viewing; in contrast, task constraints in visual search reduce gaze ambiguity and provide stronger shared structure across observers. Nevertheless, ST-DIFFEYE exceeds ScanDiff on every duration-aware variant across both free-viewing benchmarks, indicating that joint trajectory modeling captures temporal information that scanpath-only supervision discards. While DeepGaze III and HAT remain competitive on MM and non-duration-aware SM, our consistent gains on trajectory-sensitive and duration-aware metrics show that joint modeling better captures the distributional properties of free-viewing gaze. **Trajectory comparison.** Beyond scanpath-level evaluation, the trajectory comparison block shows that ST-DIFFEYE also improves over DiffEye, the only continuous-trajectory baseline, on COCO-FreeView (best on 7 of 9 metrics) and on MIT1003 (best on 5 of 9 metrics). This indicates that joint scanpath supervision benefits not only the extracted scanpaths but

Table 2: **Ablation studies on COCO-Search18 (Target Present).** We isolate four design choices: (a) input modality, (b) fixation-index interpolation, (c) concatenation stage, and (d) fixation-index encoding. The shaded row is our full model, combining the default of each group (trajectory + scanpath input, uniform interpolation, raw input channel concatenation, raw scalar encoding). **Bold** marks the best in each column.

Group	Variant	LD↓	DFD↓	DTW↓	TDE↓	MM↑	SM↑	$SM^{\dagger}$ ↑	SS↑	$SS^{\dagger}$ ↑	SSS↑	$SSS^{\dagger}$ ↑
(a) Input modality	Scanpath only	9.653	124.455	219.855	0.076	0.371	0.165	0.104	**0.521**	0.148	0.164	0.089
	Trajectory only	9.180	121.767	237.234	0.090	0.372	0.164	0.112	0.445	0.160	0.178	0.110
(b) Interpolation	Distance-based	9.063	120.182	219.703	0.085	0.371	0.176	0.113	0.483	0.160	0.179	0.098
(c) Concat. stage	Latent-space	8.984	120.450	219.644	0.086	0.376	0.168	0.114	0.481	0.156	0.177	0.110
	Embedding-space	16.949	323.933	838.854	0.272	0.075	−0.302	−0.281	−0.204	−0.135	−0.288	−0.112
(d) Index encoding	Learnable	8.941	115.029	214.956	0.083	0.372	0.182	0.125	0.495	0.169	0.186	0.108
	Sinusoidal	9.282	119.815	222.527	0.083	0.372	0.199	0.105	0.519	0.139	0.207	0.093
ST-DIFFEYE		**8.716**	**110.038**	**196.631**	**0.074**	**0.381**	**0.206**	**0.146**	0.504	**0.180**	**0.211**	**0.134**

also the underlying continuous trajectories. Overall, jointly modeling scanpaths and trajectories yields consistent gains across both task regimes and evaluation granularities, with larger improvements in the more constrained visual-search setting and competitive gains under the higher variability of free-viewing. The strongest improvements occur on trajectory-sensitive and duration-aware metrics, supporting the complementary role of trajectory information. More results under the mean, best, and KLD evaluation protocols are provided in Appendix Section A.

Metric interpretation. The reported metrics capture complementary aspects of gaze generation quality. LD measures fixation ordering; DFD and DTW measure geometric similarity under different alignment strategies; TDE captures gaze dynamics; MM evaluates multiple scanpath properties, including shape, direction, length, and position; and SM and SS measure fixation-level and semantic sequence similarity, respectively. Their duration-aware variants additionally account for fixation duration. Thus, strong performance on one metric does not necessarily imply strong performance on another; for example, scanpaths may visit similar regions in the same order while differing in geometry or temporal dynamics. We therefore report a comprehensive suite of established metrics and evaluate each under the proposed CRPS protocol, which provides a common distribution-aware framework while preserving the complementary information captured by each metric.

Ablation Study. We conduct ablation studies on the COCO-Search18 target-present split to validate our design choices, as shown in Table 2. **Input modality.** The full model outperforms scanpath-only and trajectory-only variants on most metrics, confirming that fixation-level structure and continuous gaze dynamics provide complementary information. **Fixation index construction.** Uniform interpolation performs better than the distance-based alternative, which assigns each trajectory point a fixation index using an inverse-distance-weighted average of nearby fixation indices, suggesting that preserving temporal progression between fixations is more effective than relying only on spatial proximity. **Concatenation stage.** Injecting the fixation index as a raw input channel gives the best results, while latent-space concatenation is weaker and embedding-space conditioning substantially degrades performance. **Encoding type.** Using the raw scalar index generally outperforms using the index to retrieve learnable and sinusoidal embeddings, indicating that the continuous fixation index already provides sufficient alignment information. Overall, these results show that our simple joint input design with raw-channel fixation index conditioning is effective and robust.

Qualitative Evaluation. We visualize generated scanpaths in Figure 4. On COCO-Search18, our model produces accurate and task-consistent scanpaths in the target-present setting: the generated fixations reliably move toward the queried object and land within the target region, while still following the overall spatial pattern of human scanpaths. In the target-absent setting and free-viewing tasks, where gaze behavior is less goal-directed and more exploratory, our samples generally traverse image regions consistent with the ground-truth scanpath distribution. In contrast, several baselines either produce deterministic or low-diversity outputs, such as GazeFormer, GazeXplain, and ScanDiff, or fail to match the visual structure of human gaze patterns, as observed for HAT, TPP-Gaze, and DiffEye. These qualitative results suggest that our joint trajectory-scanpath modeling

better captures both task relevance and human-like variability. More visualizations and statistical analysis of generation results are shown in Appendix Sections E and F.

5 Conclusion

We present ST-DIFFEYE, a joint trajectory-scanpath diffusion framework for stochastic human gaze modeling. By aligning scanpaths with continuous trajectories through fixation-index channels, our method combines fixation-level structure with fine-grained gaze dynamics using only minimal input-output modifications. We also introduce CRPS-based metrics for evaluating both accuracy and diversity. Experiments on free-viewing and visual search benchmarks show that joint modeling outperforms single-representation baselines and achieves state-of-the-art performance, highlighting the complementarity of scanpaths and trajectories for generative gaze modeling. Limitations, broader societal impacts, and the licenses of assets used in this work are discussed in Appendix Sections G to I.

6 Acknowledgment

We acknowledge the support of the UIUC Investment for Growth Program, the Health Care Engineering Systems Center within the Coordinated Science Lab at UIUC, and the National Institutes of Health (NIH) under award P41EB028242.

References

[1] Alfred L Yarbus. *Eye movements and vision*. Springer, 2013.

[2] Constantin A Rothkopf, Dana H Ballard, and Mary M Hayhoe. Task and context determine where you look. *Journal of vision*, 7(14):16–16, 2007.

[3] David Noton and Lawrence Stark. Scanpaths in eye movements during pattern perception. *Science*, 171(3968):308–311, Jan 1971.

[4] David Noton and Lawrence Stark. Scanpaths in saccadic eye movements while viewing and recognizing patterns. *Vision Research*, 11(9):929–IN8, 1971.

[5] Adrienne Moore, Madeline Wozniak, Andrew Yousef, Cindy Carter Barnes, Debra Cha, Eric Courchesne, and Karen Pierce. The geometric preference subtype in asd: identifying a consistent, early-emerging phenomenon through eye tracking. *Molecular Autism*, 9(1):19, Mar 2018.

[6] Isayas Berhe Adhanom, Paul MacNeilage, and Eelke Folmer. Eye tracking in virtual reality: a broad review of applications and challenges. *Virtual Real.*, 27(2):1481–1505, January 2023.

[7] Antonio Torralba, Aude Oliva, Monica S. Castelhano, and John M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, 113(4):766–786, Oct 2006.

[8] Benjamin de Haas, Alexios L. Iakovidis, D. Samuel Schwarzkopf, and Karl R. Gegenfurtner. Individual differences in visual salience vary along semantic dimensions. *Proceedings of the National Academy of Sciences*, 116(24):11687–11692, 2019.

[9] Giuseppe Cartella, Vittorio Cuculo, Alessandro D'Amelio, Marcella Cornia, Giuseppe Boccignone, and Rita Cucchiara. Modeling human gaze behavior with diffusion models for unified scanpath prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16206–16216, 2025.

[10] 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 *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 8775–8784, 2025.

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

[12] Ozgur Kara, Harris Nisar, and James Matthew Rehg. Diffeye: Diffusion-based continuous eye-tracking data generation conditioned on natural images. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[13] Chuhan Jiao, Yao Wang, Guanhua Zhang, Mihai Bâce, Zhiming Hu, and Andreas Bulling. Diffgaze: A diffusion model for modelling fine-grained human gaze behaviour on 360 images. *ACM Transactions on Interactive Intelligent Systems*, 16(1):1–23, 2026.

[14] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[15] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

[16] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[17] Abbas Mammadov, So Takao, Bohan Chen, Ricardo Baptista, Morteza Mardani, Yee Whye Teh, and Julius Berner. Variational flow maps: Make some noise for one-step conditional generation. *arXiv preprint arXiv:2603.07276*, 2026.

[18] Xianyu Chen, Ming Jiang, and Qi Zhao. Predicting human scanpaths in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10876–10885, 2021.

[19] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*, pages 2106–2113. IEEE, 2009.

[20] 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):8776, 2021.

[21] Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, January 1980.

[22] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology*, 4(4):219–227, 1985.

[23] John M Findlay and Iain D Gilchrist. Visual attention: The active vision perspective. In *Vision and attention*, pages 83–103. Springer, 2001.

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

[25] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998.

[26] David J Berg, Susan E Boehnke, Robert A Marino, Douglas P Munoz, and Laurent Itti. Free viewing of dynamic stimuli by humans and monkeys. *Journal of vision*, 9(5):19–19, 2009.

[27] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):185–207, 2012.

[28] Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille. The secrets of salient object segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 280–287, 2014.

[29] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1072–1080, 2015.

[30] Saumya Jetley, Naila Murray, and Eleonora Vig. End-to-end saliency mapping via probability distribution prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5753–5761, 2016.

[31] Srinivas SS Kruthiventi, Kumar Ayush, and R Venkatesh Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, 26(9):4446–4456, 2017.

[32] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing*, 27(10):5142–5154, 2018.

[33] M Kümmerer, L Theis, and M Bethge. Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet. In *International Conference on Learning Representations (ICLR 2015)*, pages 1–12, San Diego, CA, USA, 2014. International Conference on Learning Representations (ICLR 2015).

[34] Akis Linardos, Matthias Kümmerer, Ori Press, and Matthias Bethge. Deepgaze iie: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 12919–12928, 2021.

[35] Jianxun Lou, Hanhe Lin, David Marshall, Dietmar Saupe, and Hantao Liu. Transalnet: Towards perceptually relevant visual saliency prediction. *Neurocomputing*, 494:455–467, July 2022.

[36] Alireza Hosseini, Amirhossein Kazerouni, Saeed Akhavan, Michael Brudno, and Babak Taati. Sum: Saliency unification through mamba for visual attention modeling. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, page 1597–1607. IEEE, February 2025.

[37] Marc Assens, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. Pathgan: Visual scanpath prediction with generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[38] Wanjie Sun, Zhenzhong Chen, and Feng Wu. Visual scanpath prediction using ior-roi recurrent mixture density network. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):2101–2118, 2019.

[39] 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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 1441–1450. IEEE, June 2023.

[40] 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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 1683–1693. IEEE, June 2024.

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

[42] Xianyu Chen, Ming Jiang, and Qi Zhao. Beyond average: Individualized visual scanpath prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25420–25431, 2024.

[43] Xianyu Chen, Ming Jiang, and Qi Zhao. Gazexplain: Learning to predict natural language explanations of visual scanpaths. In *European Conference on Computer Vision*, pages 314–333. Springer, 2024.

[44] Lena Bolliger, David Reich, Patrick Haller, Deborah Jakobi, Paul Prasse, and Lena Jäger. Scandl: A diffusion model for generating synthetic scanpaths on texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15513–15538, 2023.

[45] Chuhan Jiao, Guanhua Zhang, Yeonjoo Cho, Zhiming Hu, and Andreas Bulling. Diffeyesyn: User-specific subtle eye movement synthesis using diffusion models. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2026.

[46] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

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

[48] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Featured Certification.

[49] 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 *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Appendix

Contents

A Additional Evaluation Protocols **15**

A.1 Mean Protocol 15

A.2 Best-of-N Protocol 16

A.3 KLD Protocol 17

B Statistical Significance **18**

C Background on Model Architecture, Training, and Sampling **19**

D Implementation Details **20**

D.1 Training Details 20

D.2 Dataset-Specific Implementation Details 21

E Statistical Analysis of Generated Scanpaths **22**

F Additional Qualitative Results **24**

F.1 Additional Scanpath Results 24

F.2 Additional Trajectory Results 24

G Limitations and Future Work **26**

H Broader Impacts **27**

I Asset Licenses **28**

A Additional Evaluation Protocols

We report additional results under the mean, best-of-N, and KLD evaluation protocols.

A.1 Mean Protocol

Table 3: **Mean metrics across free-viewing and visual-search datasets.** Datasets: COCO-FV (COCO-FreeView), MIT-FV (MIT1003), COCO-TP (COCO-Search18, Target Present), COCO-TA (COCO-Search18, Target Absent). ScanMatch (SM), Sequence Score (SS), and Semantic Sequence Score (SSS); † marks the duration-aware variant. **Bold**: best; underlined: second best within each dataset. Dashes indicate methods that do not predict durations, or metrics that do not apply (SSS is undefined for free-viewing).

Dataset	Method	LD↓	DFD↓	DTW↓	TDE↓	MM↑	SM↑	SM† ↑	SS↑	SS† ↑	SSS↑	SSS† ↑
COCO-FV	GazeXplain [43]	98.724	208.090	1476.637	0.081	0.785	0.163	0.141	0.155	0.134	–	–
	HAT [40]	76.518	204.100	1636.314	0.101	0.671	0.304	–	0.289	–	–	–
	TPP-Gaze [10]	84.113	185.634	1454.071	0.080	0.792	0.243	0.218	0.213	0.186	–	–
	DeepGaze III [11]	81.550	188.692	**1359.499**	0.085	0.677	0.279	–	0.266	–	–	–
	DiffEye [12]	78.838	181.144	1406.843	0.078	0.803	0.309	0.283	0.285	0.253	–	–
	ScanDiff [9]	78.396	**171.134**	1480.892	**0.076**	**0.831**	**0.356**	**0.346**	**0.324**	**0.299**	–	–
	ST-DIFFEYE	**72.386**	186.934	1434.554	0.092	0.778	0.301	0.320	0.298	0.276	–	–
MIT-FV	GazeXplain [43]	45.582	183.752	859.060	0.098	0.805	0.347	0.375	0.335	0.359	–	–
	HAT [40]	60.118	221.411	1374.953	0.136	0.659	0.285	–	0.305	–	–	–
	TPP-Gaze [10]	43.426	184.633	824.114	0.101	0.777	0.306	0.294	0.305	0.280	–	–
	DeepGaze III [11]	**41.517**	183.371	801.176	0.101	0.684	0.387	–	0.393	–	–	–
	DiffEye [12]	47.670	177.656	791.000	**0.092**	**0.813**	**0.402**	**0.435**	**0.411**	**0.430**	–	–
	ScanDiff [9]	46.085	**176.326**	812.946	0.093	0.812	0.366	0.387	0.374	0.378	–	–
	ST-DIFFEYE	41.641	187.029	**744.834**	0.105	0.773	0.324	0.365	0.350	0.358	–	–
COCO-TP	GazeFormer [39]	**15.114**	**181.152**	**284.021**	**0.099**	**0.827**	**0.545**	**0.508**	**0.614**	**0.552**	**0.501**	**0.455**
	GazeXplain [43]	15.926	200.329	307.095	0.104	0.807	0.508	0.476	0.591	0.532	0.474	0.439
	HAT [40]	18.652	224.163	433.123	0.144	0.651	0.467	–	0.565	–	0.460	–
	TPP-Gaze [10]	44.976	254.692	851.326	0.166	0.755	0.173	0.145	0.230	0.190	0.199	0.157
	ScanDiff [9]	17.325	192.211	332.698	0.117	0.804	0.485	0.413	0.583	0.479	0.467	0.375
	ST-DIFFEYE	15.382	200.597	359.013	0.137	0.788	0.472	0.357	0.610	0.441	0.476	0.334
COCO-TA	GazeFormer [39]	32.439	**188.240**	622.325	0.100	0.808	0.339	0.329	0.345	0.330	0.359	0.340
	GazeXplain [43]	**31.986**	189.654	**572.228**	**0.094**	**0.813**	**0.381**	**0.371**	**0.397**	**0.382**	**0.377**	**0.365**
	HAT [40]	32.378	200.521	630.468	0.121	0.651	0.340	–	0.360	–	0.363	–
	TPP-Gaze [10]	47.572	200.067	916.626	0.129	0.777	0.224	0.203	0.236	0.209	0.362	0.344
	ScanDiff [9]	35.257	200.334	654.090	0.115	0.807	0.344	0.332	0.361	0.344	0.273	0.243
	ST-DIFFEYE	32.894	202.098	632.080	0.116	0.788	0.340	0.320	0.363	0.322	0.360	0.330

Following DiffEye [12], the mean protocol computes a distance or similarity score between every pair of ground-truth and generated scanpaths for each input stimulus, and averages the scores across all pairs. Results are reported in Table 3.

A.2 Best-of-N Protocol

Table 4: **Best-of-N metrics across free-viewing and visual-search datasets.** Datasets: COCO-FV (COCO-FreeView), MIT-FV (MIT1003), COCO-TP (COCO-Search18, Target Present), COCO-TA (COCO-Search18, Target Absent). ScanMatch (SM), Sequence Score (SS), and Semantic Sequence Score (SSS); † marks the duration-aware variant. **Bold**: best; underlined: second best within each dataset. Dashes indicate methods that do not predict durations, or metrics that do not apply (SSS is undefined for free-viewing).

Dataset	Method	LD↓	DFD↓	DTW↓	TDE↓	MM↑	SM↑	SM† ↑	SS↑	SS† ↑	SSS↑	SSS† ↑
COCO-FV	GazeXplain [43]	97.256	180.455	1312.892	0.063	0.838	0.190	0.169	0.192	0.173	–	–
	HAT [40]	71.187	147.122	1233.285	0.080	0.722	0.389	–	0.398	–	–	–
	TPP-Gaze [10]	74.643	134.642	1065.083	0.054	0.878	0.353	0.302	0.347	0.305	–	–
	DeepGaze III [11]	78.371	136.125	**1014.752**	0.061	0.729	0.354	–	0.369	–	–	–
	DiffEye [12]	71.682	142.949	1132.701	0.063	0.875	0.387	0.343	0.384	0.339	–	–
	ScanDiff [9]	**70.118**	**125.174**	1111.774	**0.060**	**0.892**	**0.445**	**0.434**	**0.445**	**0.418**	–	–
	ST-DIFFEYE	71.237	133.596	1050.138	0.071	0.865	0.391	0.426	0.400	0.404	–	–
MIT-FV	GazeXplain [43]	42.563	134.709	655.104	0.080	0.879	0.422	0.457	0.439	0.470	–	–
	HAT [40]	49.778	158.156	995.366	0.107	0.719	0.370	–	0.425	–	–	–
	TPP-Gaze [10]	34.951	117.262	535.849	**0.062**	0.888	0.458	0.436	0.490	0.467	–	–
	DeepGaze III [11]	38.182	117.225	542.954	0.072	0.744	0.489	–	0.530	–	–	–
	DiffEye [12]	36.852	119.422	515.773	0.067	**0.897**	**0.526**	**0.543**	**0.558**	**0.565**	–	–
	ScanDiff [9]	35.347	**113.097**	523.003	0.064	**0.897**	0.491	0.505	0.532	0.539	–	–
	ST-DIFFEYE	**34.812**	114.222	**485.741**	0.065	0.884	0.471	0.504	0.520	0.542	–	–
COCO-TP	GazeFormer [39]	15.115	181.157	284.039	0.099	0.827	0.544	0.508	0.614	0.552	0.501	0.456
	GazeXplain [43]	14.731	179.029	266.416	0.092	0.863	0.574	0.560	0.632	0.612	0.544	0.541
	HAT [40]	**11.086**	67.599	138.744	0.058	0.756	**0.712**	–	0.825	–	0.719	–
	TPP-Gaze [10]	23.307	164.028	481.844	0.121	0.874	0.305	0.238	0.410	0.312	0.366	0.263
	ScanDiff [9]	11.299	69.390	137.772	0.057	0.923	0.689	0.584	0.804	0.666	0.685	0.576
	ST-DIFFEYE	11.104	**60.648**	**128.501**	**0.055**	**0.925**	0.710	**0.595**	**0.837**	**0.684**	**0.740**	**0.629**
COCO-TA	GazeFormer [39]	32.439	188.240	622.325	0.100	0.808	0.339	0.329	0.345	0.330	0.359	0.340
	GazeXplain [43]	30.281	146.649	464.407	0.074	0.885	0.461	0.452	0.485	0.482	0.469	0.471
	HAT [40]	27.281	111.527	395.514	0.069	0.738	0.491	–	0.541	–	0.547	–
	TPP-Gaze [10]	31.723	130.847	573.821	0.091	0.888	0.359	0.311	0.407	0.350	0.444	0.376
	ScanDiff [9]	**26.136**	**103.716**	385.415	**0.066**	**0.911**	**0.512**	0.481	0.546	0.533	0.559	0.531
	ST-DIFFEYE	26.738	105.025	**376.639**	0.068	0.907	0.511	**0.484**	**0.557**	**0.541**	**0.568**	**0.542**

Following DiffEye [12], the best-of-N protocol selects, for each ground-truth scanpath, the closest match among all generated scanpaths for the same stimulus, using the minimum score for distance metrics and the maximum score for similarity metrics. We then average these best scores across all ground-truth scanpaths. Results are reported in Table 4.

A.3 KLD Protocol

Table 5: **KLD metrics across free-viewing and visual-search datasets.** Datasets: COCO-FV (COCO-FreeView), MIT-FV (MIT1003), COCO-TP (COCO-Search18, Target Present), COCO-TA (COCO-Search18, Target Absent). ScanMatch (SM), Sequence Score (SS), and Semantic Sequence Score (SSS); † marks the duration-aware variant. **Bold**: best; underlined: second best within each dataset. Dashes indicate methods that do not predict durations, or metrics that do not apply (SSS is undefined for free-viewing).

Dataset	Method	LD↓	DFD↓	DTW↓	TDE↓	MM↓	SM↓	SM† ↓	SS↓	SS† ↓	SSS↓	SSS† ↓
COCO-FV	GazeXplain [43]	0.630	0.116	0.026	0.322	0.369	3.412	4.907	5.640	2.407	–	–
	HAT [40]	0.115	0.191	0.149	1.002	4.218	0.559	–	0.344	–	–	–
	TPP-Gaze [10]	0.163	0.010	0.026	0.037	0.077	1.001	1.914	0.774	0.619	–	–
	DeepGaze III [11]	0.152	0.011	0.077	**0.036**	3.855	0.682	–	1.266	–	–	–
	DiffEye [12]	**0.098**	0.010	0.043	0.038	**0.022**	0.230	0.522	0.248	0.120	–	–
	ScanDiff [9]	0.100	0.036	**0.011**	0.045	0.070	**0.022**	**0.042**	**0.036**	**0.012**	–	–
	ST-DIFFEYE	0.314	**0.005**	0.053	0.062	0.237	0.676	0.103	0.382	0.102	–	–
MIT-FV	GazeXplain [43]	1.137	0.041	0.075	0.041	0.086	0.202	0.071	2.197	0.072	–	–
	HAT [40]	4.237	1.554	3.986	2.985	4.572	1.937	–	1.743	–	–	–
	TPP-Gaze [10]	**0.020**	0.015	0.027	**0.012**	0.074	0.342	0.575	0.289	0.428	–	–
	DeepGaze III [11]	0.741	0.031	0.113	0.075	3.721	0.106	–	1.035	–	–	–
	DiffEye [12]	0.080	**0.011**	0.018	0.021	**0.024**	**0.014**	**0.018**	0.071	**0.018**	–	–
	ScanDiff [9]	0.041	0.014	0.019	0.037	0.044	0.073	0.045	**0.055**	0.032	–	–
	ST-DIFFEYE	0.029	0.021	**0.013**	0.038	0.095	0.164	0.056	0.142	0.038	–	–
COCO-TP	GazeFormer [39]	0.169	**0.039**	0.097	0.164	0.111	**0.107**	0.063	0.109	0.104	0.072	0.127
	GazeXplain [43]	0.324	0.061	0.135	0.238	0.119	0.149	**0.018**	0.176	0.061	0.130	0.145
	HAT [40]	0.147	0.149	0.183	0.176	3.835	0.170	–	0.151	–	0.140	–
	TPP-Gaze [10]	2.893	2.265	3.736	3.209	0.715	5.206	5.886	3.170	3.137	2.229	1.751
	ScanDiff [9]	0.090	0.047	0.078	0.095	0.051	0.109	0.065	0.085	**0.022**	0.085	**0.076**
	ST-DIFFEYE	**0.028**	0.068	**0.066**	**0.081**	**0.042**	0.152	0.206	**0.045**	0.048	**0.054**	0.090
COCO-TA	GazeFormer [39]	0.118	0.085	0.069	0.109	0.093	0.038	0.030	0.235	0.101	0.152	0.086
	GazeXplain [43]	0.112	0.027	0.066	0.109	0.044	0.033	0.041	0.162	0.044	0.188	0.041
	HAT [40]	0.062	0.024	0.036	0.044	3.799	0.042	–	0.106	–	0.070	–
	TPP-Gaze [10]	0.665	0.182	0.476	0.266	0.129	0.752	0.917	0.540	0.628	0.281	0.227
	ScanDiff [9]	**0.022**	0.012	**0.014**	**0.007**	0.020	**0.012**	**0.008**	**0.031**	**0.008**	**0.020**	0.008
	ST-DIFFEYE	0.039	**0.007**	0.024	0.016	**0.018**	0.018	0.014	0.045	0.011	0.039	**0.006**

Following ScanDiff [9] and TPP-Gaze [10], the KLD protocol pools human-consistency scores from ground-truth versus ground-truth scanpath pairs and model-alignment scores from ground-truth versus generated pairs across the evaluation set, converts each pooled score set into a normalized histogram, and reports the KL divergence between the human and model score distributions. Results are reported in Table 5.

B Statistical Significance

Table 6: **Statistical significance on free-viewing datasets.** CRPS and KLD results on COCO-FreeView and MIT1003, reported with 95% bootstrap confidence intervals. For each test image, we sample 10 generated scanpaths from 100 generations and compare them against 10 ground-truth scanpaths, repeating the procedure 50 times. † denotes duration-aware variants. **Bold**: best within each dataset and evaluation protocol.

Protocol	Method	LD	DTW	DFD	TDE	MM	SM	SM†	SS	SS†
COCO-FreeView										
CRPS	DiffEye [12]	56.37±0.73	1117.1±19.6	121.37±1.91	0.051±0.001	0.369±0.002	0.025±0.004	-0.018±0.004	0.196±0.008	-0.029±0.005
	ScanDiff [9]	**46.50±0.30**	968.1±13.4	105.06±1.12	0.047±0.001	**0.388±0.001**	**0.102±0.002**	0.096±0.002	**0.273±0.007**	0.073±0.005
	ST-DIFFEYE	53.08±0.54	**911.9±11.1**	**102.51±1.01**	**0.047±0.001**	0.377±0.001	0.092±0.002	**0.123±0.002**	0.206±0.007	**0.088±0.004**
KLD	DiffEye [12]	0.101±0.009	0.046±0.009	0.013±0.004	**0.042±0.011**	**0.025±0.003**	0.240±0.033	0.534±0.058	0.250±0.021	0.125±0.018
	ScanDiff [9]	0.105±0.011	**0.014±0.004**	0.039±0.007	0.050±0.010	0.074±0.005	**0.027±0.008**	**0.047±0.011**	**0.040±0.007**	**0.016±0.005**
	ST-DIFFEYE	**0.084±0.010**	0.024±0.006	**0.013±0.004**	0.206±0.027	0.268±0.014	0.477±0.046	0.144±0.019	0.308±0.025	0.076±0.010
MIT1003										
CRPS	DiffEye [12]	27.49±0.50	516.2±21.9	111.53±4.44	0.062±0.003	**0.377±0.004**	0.112±0.009	0.125±0.011	**0.329±0.019**	0.122±0.015
	ScanDiff [9]	**26.03±0.59**	513.9±19.5	105.39±2.97	0.060±0.002	0.376±0.002	**0.115±0.006**	0.121±0.008	0.325±0.017	0.117±0.015
	ST-DIFFEYE	26.92±0.65	**472.6±17.2**	**104.35±2.79**	**0.055±0.002**	0.373±0.003	0.115±0.005	**0.139±0.006**	0.287±0.018	**0.134±0.013**
KLD	DiffEye [12]	0.087±0.031	**0.024±0.009**	**0.018±0.006**	**0.038±0.019**	**0.032±0.007**	**0.020±0.012**	**0.026±0.013**	0.078±0.023	**0.026±0.007**
	ScanDiff [9]	**0.047±0.023**	0.026±0.013	0.024±0.012	0.055±0.026	0.052±0.008	0.086±0.029	0.058±0.023	**0.066±0.030**	0.045±0.019
	ST-DIFFEYE	0.049±0.011	0.028±0.013	0.031±0.011	0.047±0.020	0.118±0.020	0.236±0.048	0.104±0.028	0.237±0.089	0.047±0.013

We report 95% bootstrap confidence intervals on COCO-FreeView and MIT1003. Specifically, for each test image, we generate 100 scanpaths, randomly sample 10 generated scanpaths to compare against the 10 ground-truth scanpaths, and repeat this procedure 50 times to estimate confidence intervals. The resulting intervals are reported in Table 6. Overall, they support the conclusions of the paper: under the proposed CRPS protocol, ST-DIFFEYE continues to outperform the baselines on the metrics reported in the paper while remaining competitive on the others, whereas under the KLD protocol the intervals similarly confirm the protocol-dependent trends discussed in the paper. These results indicate that the observed differences are not attributable to sampling noise.

C Background on Model Architecture, Training, and Sampling

Our model architecture primarily follows ScanDiff [9], a conditional diffusion model for scanpath generation. Given an image stimulus $I \in \mathbb{R}^{H\times W\times 3}$ and a viewing task c, the goal is to generate a scanpath $S = \{f_m\}_{m=1}^{M}$, where each fixation $f_m = (r_m, \Delta t_m^{\text{sp}})$ contains a 2D gaze location $r_m = (x_m^{\text{sp}}, y_m^{\text{sp}})$ and a fixation duration Δt_m^{sp}. Rather than diffusing directly in the raw scanpath space, ScanDiff first maps the scanpath into a latent token space through a learnable projection $g_\theta : \mathbb{R}^{M\times 3} \to \mathbb{R}^{M\times d}$, producing $z_0 = g_\theta(S)$. The forward process gradually corrupts z_0 into noisy latents z_k with a Gaussian diffusion process:

$$q(z_k \mid z_{k-1}) = \mathcal{N}\left(z_k; \sqrt{1-\beta_k} z_{k-1}, \beta_k I\right).$$

The conditional reverse process is then defined as

$$p_\theta(z_{0:K} \mid I, c) = p(z_K) \prod_{k=1}^{K} p_\theta(z_{k-1} \mid z_k, I, c).$$

The denoising network ϕ_θ is implemented as an encoder-only Transformer with an additional cross-attention layer inserted between self-attention and feed-forward layers. Self-attention models dependencies among scanpath tokens, while cross-attention conditions the denoising process on visual and task information. Specifically, ScanDiff extracts dense image features $v(I) \in \mathbb{R}^{h\times w\times d_v}$ using a DINOv2 [48] visual encoder and encodes the viewing task with a CLIP [49] text encoder $\psi(c) \in \mathbb{R}^{d_t}$. The visual and textual features are projected into a shared d-dimensional space, concatenated along the channel dimension, and linearly projected to obtain the joint multimodal feature $V_{\text{joint}} \in \mathbb{R}^{hw\times d}$. The noisy latent sequence z_k is augmented with positional and diffusion timestep embeddings, and the denoised latent prediction is obtained as $\hat{z}_0 = \phi_\theta(z_k, V_{\text{joint}})$. A feed-forward decoder γ_θ reconstructs the scanpath $\hat{S} = \gamma_\theta(\hat{z}_0)$, while a length prediction head $l_\theta : \mathbb{R}^{M\times d} \to \mathbb{R}^M$ predicts token validity for variable-length generation. ScanDiff is trained with a combination of four losses:

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{val}} + \mathcal{L}_{\text{reg}}.$$

The latent diffusion loss encourages the denoised prediction to recover the clean latent sequence:

$$\mathcal{L}_{\text{diff}} = \sum_{k=1}^{K} \|z_0 - \phi_\theta(z_k, V_{\text{joint}})\|_2^2.$$

The reconstruction loss supervises the decoded scanpath with an ℓ_1 loss over fixation coordinates and durations:

$$\mathcal{L}_{\text{recon}} = \frac{1}{M} \sum_{m=1}^{M} \left(|x_m^{\text{sp}} - \hat{x}_m^{\text{sp}}| + |y_m^{\text{sp}} - \hat{y}_m^{\text{sp}}| + |\Delta t_m^{\text{sp}} - \Delta \hat{t}_m^{\text{sp}}|\right),$$

where padded tokens are masked out. The validity loss $\mathcal{L}_{\text{val}}$ is a binary cross-entropy loss between the predicted token-validity probabilities and the ground-truth validity labels, and $\mathcal{L}_{\text{reg}} = \|\mu(z_K)\|_2^2$ regularizes the terminal diffusion state so that it matches a standard Gaussian prior. At inference time, ScanDiff samples an initial latent sequence $z_K \sim \mathcal{N}(0, I)$ and iteratively applies the learned reverse process conditioned on V_{joint} until obtaining $\hat{z}_0$. The decoded output $\hat{S} = \gamma_\theta(\hat{z}_0)$ gives the predicted fixation coordinates and durations, while the length head determines the valid prefix of the sequence. Multiple diverse scanpaths for the same image and task are produced by sampling different initial noise sequences z_K.

D Implementation Details

D.1 Training Details

We implement all models in PyTorch and train them using AdamW with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-2}. All models are trained for 300 epochs on a single NVIDIA GH200 GPU using internal clusters. The training time of each model is about 1 day. Unless otherwise specified, other training configurations and model architectures, such as length prediction and image and text conditioning, follow the official implementation of ScanDiff [9]. We also follow ScanDiff to train separate models for different datasets to account for their different task conditions and data distributions. Empirically, the best fixation-index injection point varies by dataset, with latent-space augmentation performing slightly better on MIT1003.

D.2 Dataset-Specific Implementation Details

Table 7: **Ablation study on concatenation location (MIT1003).** We compare appending the fixation index at the raw data-channel level versus the latent-channel level. Metrics follow the main paper CRPS protocol; † marks duration-aware variants.

Concat Location	LD↓	DFD↓	DTW↓	TDE↓	MM↑	SM↑	SM† ↑	SS↑	SS† ↑
Data channel	26.706	108.258	494.392	0.062	0.371	0.104	0.128	0.281	0.127
Latent space	**24.961**	**100.903**	**456.985**	**0.058**	**0.380**	**0.129**	**0.145**	**0.321**	**0.146**

Through empirical experiments, we found that injecting the fixation index in the latent space performs better on MIT1003. This may be due to the noisier trajectory annotations in MIT1003, where some raw trajectories move out of frame between fixations, making coordinate-level trajectory-scanpath alignment less reliable. Latent-space injection can partially reduce the effect of this misalignment. Specifically, instead of concatenating the fixation index as an additional raw trajectory channel, we first project the trajectory into the latent sequence of size $T \times 256$ and then concatenate the fixation index in the latent space, resulting in a $T \times 257$ latent representation. During sampling, we initialize Gaussian noise with the same latent dimensionality and discard the additional fixation-index dimension before passing the denoised latent sequence to the decoder. All other settings follow the experiments on the remaining datasets. We show the numbers in Table 7.

E Statistical Analysis of Generated Scanpaths

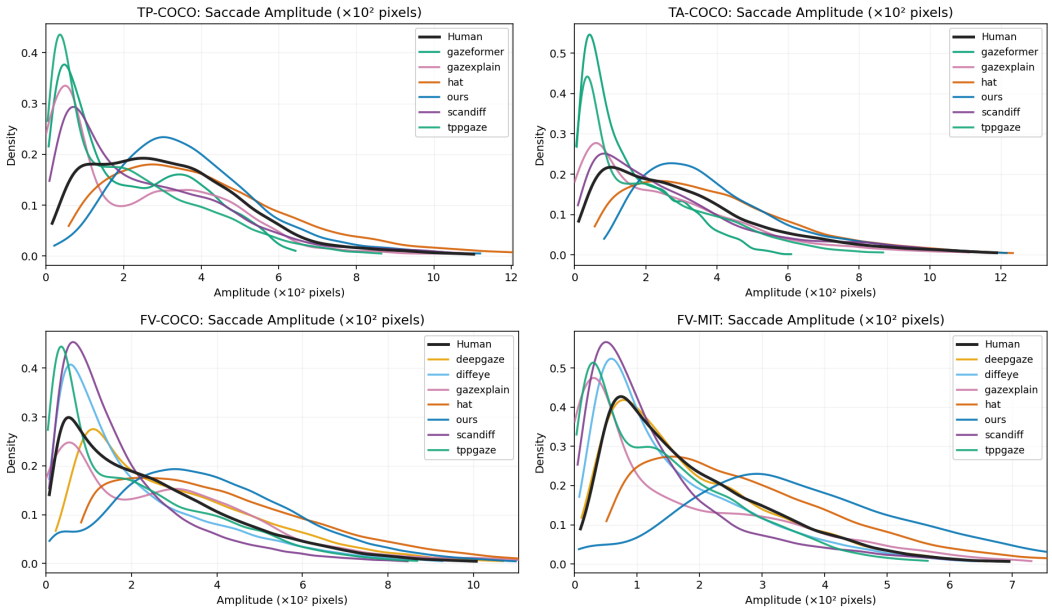


Figure 5: **Distributions of saccade amplitude for different datasets.** The horizontal axis represents the amplitude (length) of saccades in the unit of 10^2 pixels. The vertical axis represents the probability density.

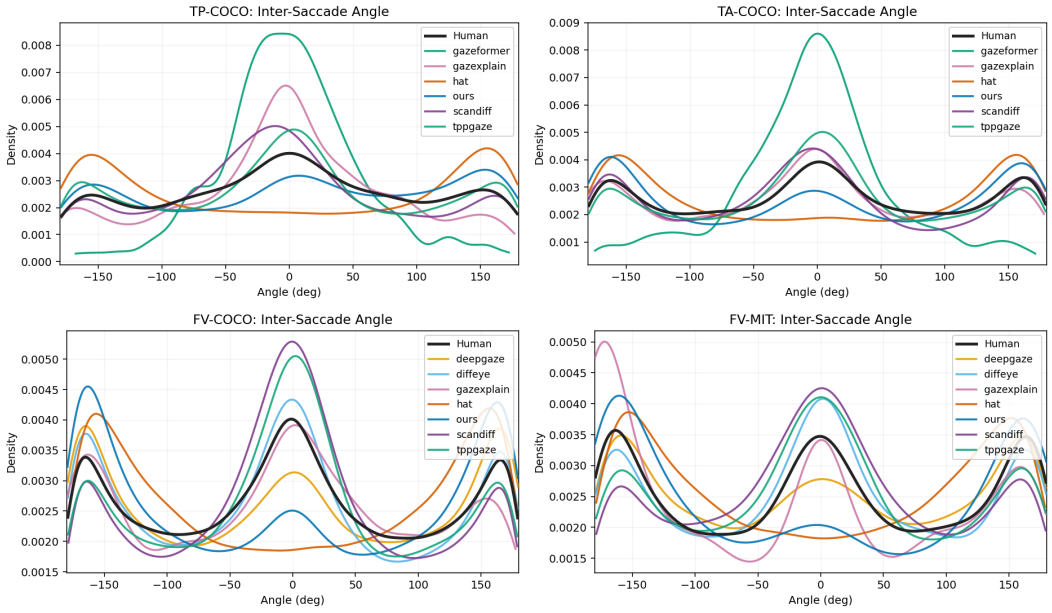


Figure 6: **Distributions of inter-saccade angle for different datasets.** The horizontal axis represents the degree of angles between consecutive saccades. The vertical axis represents the probability density.

We analyze the statistical distributions of real and generated scanpaths in terms of saccade amplitude, saccade direction, and inter-saccade angle. The corresponding plots are shown in Figures 5 to 7. The amplitude distributions indicate that human saccades typically follow a long-tailed pattern, with a peak around 100 pixels. The generated scanpaths generally follow a similar amplitude distribution.

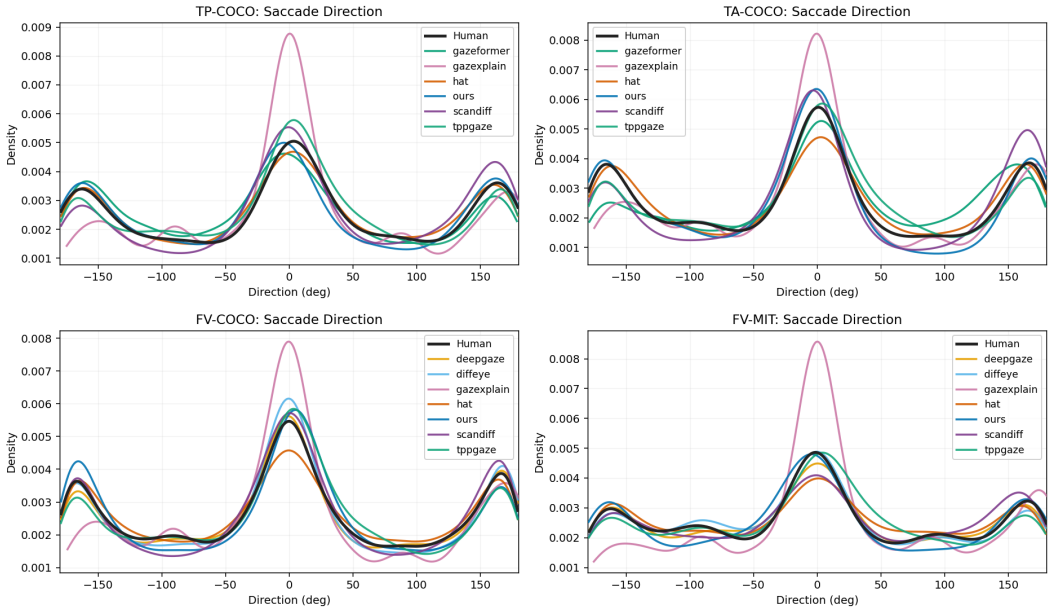


Figure 7: **Distributions of saccade direction for different datasets.** The horizontal axis represents the direction of the saccades in degrees. The vertical axis represents the probability density.

For inter-saccade angles, we observe peaks near 0 and 180 degrees. This suggests that gaze transitions often either continue in the previous direction or reverse to the opposite direction, indicating that consecutive eye movements are frequently colinear. For saccade direction, the overall distribution shapes are also similar between real and generated scanpaths. Both show a stronger horizontal bias than other directions. Overall, these results suggest that the generated scanpaths capture the key statistical properties of human eye movements.

F Additional Qualitative Results

F.1 Additional Scanpath Results

We provide additional qualitative comparisons of generated scanpaths against baselines on the free-viewing benchmarks (COCO-FreeView and MIT1003) in Figure 8 and on the visual-search benchmark (COCO-Search18, target-present and target-absent) in Figure 9. Across both settings, our model produces scanpaths that better align with the spatial structure and exploration patterns of the human ground truth, while baselines tend to either collapse to low-diversity outputs or drift away from the salient regions covered by human viewers.

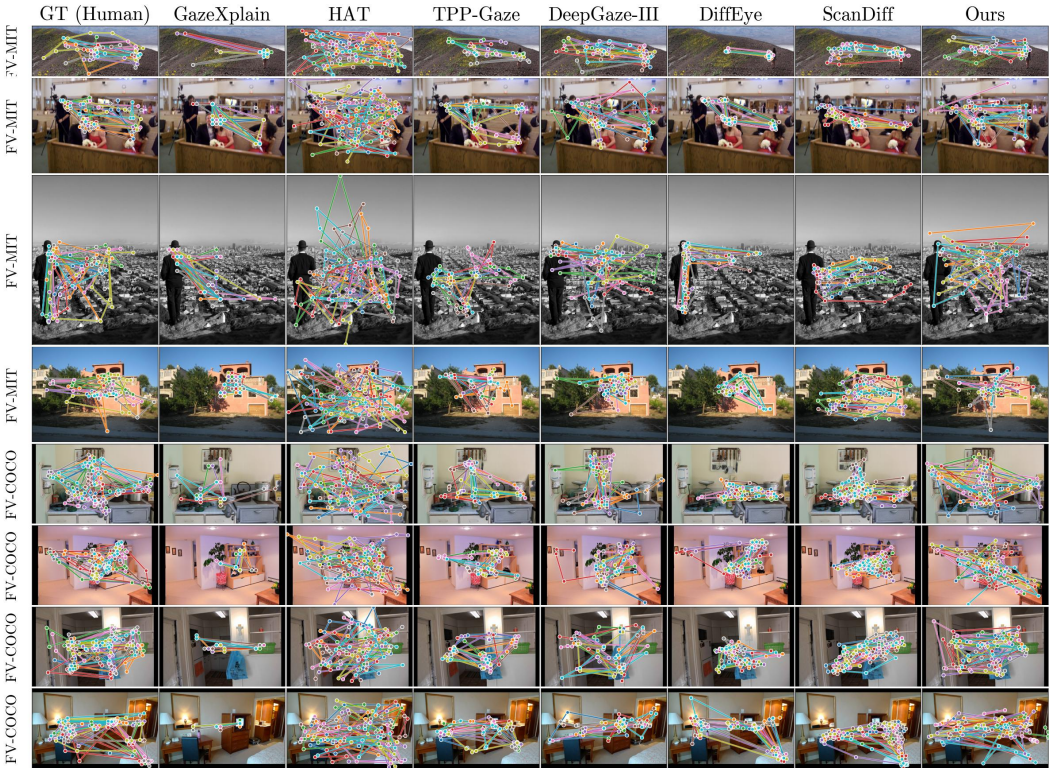


Figure 8: **Additional scanpath comparisons on free-viewing datasets.** Generated scanpaths from our method and baselines on COCO-FreeView and MIT1003.

F.2 Additional Trajectory Results

We further compare generated continuous trajectories against the only existing continuous-trajectory baseline, DiffEye [12], in Figure 10. Our trajectories more faithfully reproduce the spatial coverage and density patterns of the human ground truth, capturing both the fixation clusters and the saccadic transitions between them. In contrast, DiffEye tends to produce smoother, less structured paths that under-represent the discrete fixation organization characteristic of human gaze.



Figure 9: **Additional scanpath comparisons on visual-search datasets.** Generated scanpaths from our method and baselines on COCO-Search18 (target-present and target-absent).

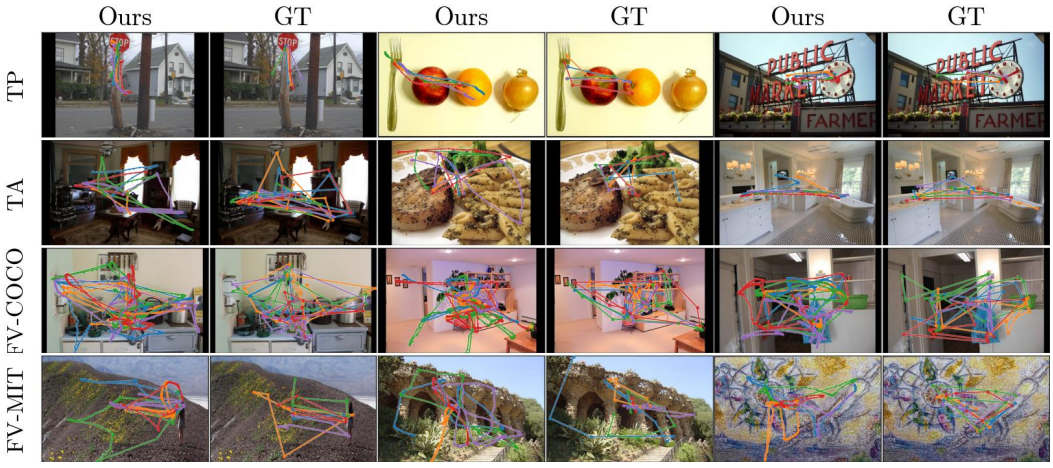


Figure 10: **Additional trajectory comparisons.** Generated continuous trajectories from our method and DiffEye [12] across free-viewing and visual-search datasets.

G Limitations and Future Work

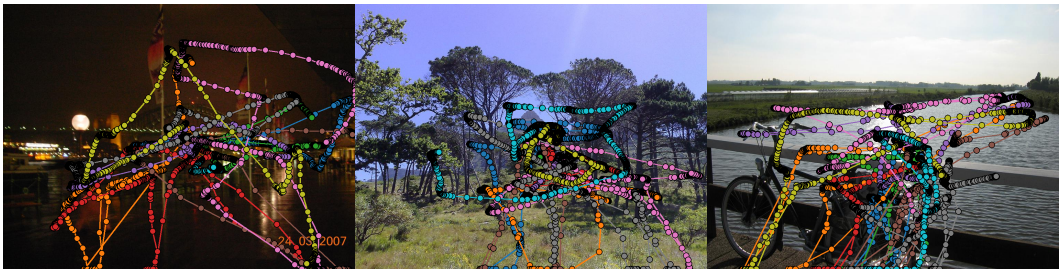


Figure 11: **Examples of failure cases caused by generated trajectories going out of frame.** Some generated trajectories tend to go outside the image frame, which can lead to inaccurate scanpaths after trajectory-to-scanpath conversion.

Table 8: Effect of the number of DDPM sampling steps on COCO-Search18 target-present (TP) and target-absent (TA). Best results are shown in bold.

Split	Model	LD↓	DFD↓	DTW↓	TDE↓	MM↑	SM↑	SM†↑	SS↑	SS†↑	SSS↑	SSS†↑
TP	500 Steps	15.006	378.970	927.431	0.319	0.019	-0.342	-0.062	-0.330	-0.050	-0.343	-0.030
	1000 Steps	**8.716**	**110.038**	**196.631**	**0.074**	**0.381**	**0.206**	**0.146**	**0.504**	**0.180**	**0.211**	**0.134**
TA	500 Steps	30.479	320.979	1310.391	0.273	0.139	-0.315	-0.083	-0.240	-0.071	-0.314	-0.022
	1000 Steps	**21.128**	**115.683**	**410.444**	**0.064**	**0.381**	**0.110**	**0.095**	**0.288**	**0.109**	**0.130**	**0.109**

While joint trajectory-scanpath modeling improves generation quality, it also introduces additional computational and storage costs. Raw trajectories are substantially longer than scanpaths, increasing memory usage, training time, and sampling cost. Moreover, our current implementation follows the baseline 1000-step DDPM sampling schedule, requiring approximately 40 seconds per stimulus to generate 10 trajectories in parallel on an RTX 4090 GPU. As shown in Table 8, reducing the sampling schedule to 500 steps noticeably degrades generation quality across both target-present and target-absent settings, suggesting that faster inference may require more efficient generative paradigms, such as flow matching, temporal compression, or hierarchical generation. The framework also requires paired trajectory and scanpath data; while scanpaths can typically be extracted automatically from raw eye-tracking recordings, storing high-frequency trajectories introduces additional storage overhead.

A second limitation concerns the noisy and dataset-dependent nature of trajectory data. Raw trajectories may contain blinks, missing samples, or out-of-frame gaze points, which can affect training and lead to unrealistic generations; examples are shown in Figure 11. Moreover, while raw-channel concatenation provides a robust default across most benchmarks, datasets with noisier trajectories or larger domain gaps, such as MIT1003, can benefit from latent-space injection. In practice, we recommend using raw-channel concatenation by default and selecting alternative injection strategies on a held-out validation set. Future work can further explore trajectory regularization, automatic detection of abnormal gaze patterns, and more efficient joint modeling strategies.

H Broader Impacts

Modeling human gaze has several positive societal applications. More accurate generative gaze models can support research in cognitive science, psychology, human-computer interaction, accessibility, virtual and augmented reality, and visual interface design. In clinical settings, gaze modeling may help study atypical visual attention patterns and support screening or diagnosis of developmental and mental health conditions, such as autism spectrum disorder or attention-related disorders, when used together with appropriate clinical expertise. Synthetic gaze generation may also reduce the cost of collecting large-scale eye-tracking data and enable safer evaluation of gaze-aware systems without exposing private user recordings. At the same time, gaze data is sensitive because it can reveal information about attention, intent, cognitive state, preference, and potentially health-related traits. Misuse of gaze modeling could enable invasive user profiling, manipulative interface or advertisement design, surveillance, or attention monitoring without informed consent. The technology may also be applied in high-stakes or ethically sensitive domains, including military training or assessment systems. These risks highlight the need for careful data governance, informed consent, privacy-preserving preprocessing, and restrictions on deployment in settings where gaze data could be used to infer sensitive personal attributes or manipulate user behavior. Our work focuses on improving scientific modeling and evaluation of gaze generation, and we encourage future uses to follow ethical review, transparency, and privacy safeguards.

I Asset Licenses

Our codebase is primarily developed based on ScanDiff [9] and DiffEye [12]. ScanDiff is publicly available on GitHub but does not specify an explicit license, whereas DiffEye is released under the MIT License. The datasets used in this work, including COCO-Search18 [20], COCO-FreeView [18], and MIT1003 [19], are stated to be distributed under the MIT License. However, the terms specified by the authors of COCO-Search18 and COCO-FreeView impose additional restrictions on the use and redistribution of the datasets. Specifically, they prohibit the reproduction, copying, sale, exchange, resale, or commercial exploitation of the images and derived data, as well as the further publication or distribution of any portion of the datasets. Therefore, we are unable to publicly release the raw trajectory data of COCO-Search18 and COCO-FreeView due to the datasets' usage and redistribution restrictions. For further information regarding access to these data, please contact the respective dataset authors. We acknowledge these licensing terms and usage restrictions and commit to complying with the applicable licenses and terms of use for all codebases and datasets used in this work.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly state in the abstract and introduction that our main contribution is about jointly modeling the trajectory and scanpath data and a new evaluation method for the task of eye gaze scanpath/trajectory generation.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitation in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe detailed model architecture, training, and sampling procedures.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The raw trajectory data for COCO-Search18 and COCO-FreeView is not yet public. We will discuss possible release strategies with the original dataset authors after redacting potentially sensitive personal information from the data upon acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: We include all details on data splits, training hyperparameters, optimizer settings, and other implementation choices in the implementation details.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: It would be too computationally expensive to report error bars.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We discuss the compute resources in the implementation details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We conform to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the border societal impacts in the Appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper has no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All code and datasets used are properly cited in the paper. We list the license of assets used in the Appendix.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method does not involve the use of LLMs. LLMs are used only for writing, editing, and formatting the paper.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.