

A Appendix / supplemental material

A.1 Classifier Free Guidance

Given a noised eye movement trajectory $R^{(t_{\text{diff}})}$ at diffusion timestep t_{diff} , we perform two forward passes through the noise prediction network ϵ_θ : one conditioned on the visual stimulus I , yielding $\epsilon_\theta(R^{(t_{\text{diff}})}, t_{\text{diff}}, I)$, and another using an unconditional input I_{uc} , defined as a zero matrix, yielding $\epsilon_\theta(R^{(t_{\text{diff}})}, t_{\text{diff}}, I_{\text{uc}})$. The final guided noise prediction $\hat{\epsilon}_\theta$ is computed as:

$$\hat{\epsilon}_\theta = (1 - c) \cdot \epsilon_\theta(R^{(t_{\text{diff}})}, t_{\text{diff}}, I_{\text{uc}}) + c \cdot \epsilon_\theta(R^{(t_{\text{diff}})}, t_{\text{diff}}, I), \quad (3)$$

where c is the classifier-free guidance scale, which we set to 4 during inference. To enable this, we simulate the unconditional setting during training by randomly replacing the conditioning input I with a zero matrix in 10% of the training samples. This encourages the model to learn both conditional and unconditional denoising behavior, supporting effective guidance at inference time.

A.2 Evaluation Algorithm

Below is the algorithm we used to compute the *mean* and *best* scores for each of the scanpath and continuous trajectory metrics.

Algorithm 1 Evaluation of Scanpath and Continuous Trajectory Generation Metrics

Require: Test set of images \mathcal{I} ; ground truth scanpaths $\mathcal{G}_i = \{g_1, \dots, g_N\}$; generated scanpaths $\mathcal{S}_i = \{s_1, \dots, s_M\}$; evaluation metric $d(\cdot, \cdot)$
Ensure: Overall *best* and *mean* scores across test images
1: Initialize `best_scores` $\leftarrow []$ and `mean_scores` $\leftarrow []$
2: **for** each image $i \in \mathcal{I}$ **do**
3: Initialize `image_best` $\leftarrow []$ and `image_mean` $\leftarrow []$
4: **for** each $g \in \mathcal{G}_i$ **do**
5: Compute distances $\{d(g, s) \mid s \in \mathcal{S}_i\}$
6: `best_g` $\leftarrow \min_{s \in \mathcal{S}_i} d(g, s)$
7: `mean_g` $\leftarrow \frac{1}{M} \sum_{s \in \mathcal{S}_i} d(g, s)$
8: Append `best_g` to `image_best`
9: Append `mean_g` to `image_mean`
10: **end for**
11: Append $\frac{1}{N} \sum \text{image_best}$ to `best_scores`
12: Append $\frac{1}{N} \sum \text{image_mean}$ to `mean_scores`
13: **end for**
14: **return** Overall Best = $\frac{1}{|\mathcal{I}|} \sum \text{best_scores}$, Overall Mean = $\frac{1}{|\mathcal{I}|} \sum \text{mean_scores}$

A.3 Additional Scanpath Distributions and Evaluations

Please see additional examples of scanpaths generated by DiffEye for the MIT1003 dataset in Fig. 6.

In addition to the trajectory-based metrics, we evaluate our model using Sequence Score (SS) and Semantic Sequence Score (Sem SS). The results, presented in Table 4, show that our model, DiffEye, achieves state-of-the-art performance on the MIT1003 dataset, outperforming all baselines in both SS and Sem SS. Furthermore, our model generalizes effectively by remaining highly competitive on the unseen OSIE dataset. This robust performance is particularly notable given that DiffEye was trained on significantly less data than competing models. For instance, while DeepGazeIII was trained on approximately 600,000 scanpaths, our method achieved these results with only 8,900 trajectories, demonstrating our approach’s ability to produce high-quality scanpaths by leveraging rich, raw trajectory information.

A.4 Additional Continuous Eye Movement Trajectory Distributions

Please see additional examples of continuous eye movement trajectories generated by DiffEye for the MIT1003 dataset in Fig. 7.

A.5 Analysis of Saliency Prediction

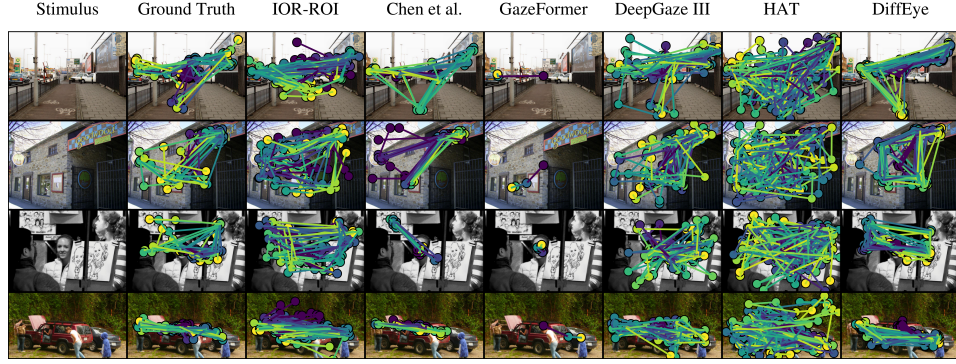


Figure 6: **Additional scanpath generation results.** Scanpaths generated by DiffEye and baseline models are shown alongside ground truth annotations across four different scenes. Each row represents a unique stimulus, and each column shows the generated scanpaths for each method.

Table 4: **Quantitative comparison using Sequence Score (SS) and Semantic Sequence Score (Sem SS) on the MIT1003 and OSIE datasets.** We report the mean performance for each method. **Bold** values indicate the best scores, while underlined values denote the second-best scores.

Model	MIT1003		OSIE	
	SS (\uparrow)	Sem SS (\uparrow)	SS (\uparrow)	Sem SS (\uparrow)
DiffEye (Ours)	0.4782	0.6611	0.4371	0.5837
HAT	0.4079	0.5794	0.4002	0.5791
Gazeformer	0.3531	0.4522	0.2713	0.3602
DeepGazeIII	0.4440	<u>0.6604</u>	0.4623	0.6459
ROI	<u>0.4506</u>	0.6603	<u>0.4404</u>	<u>0.6110</u>
Chen et al.	<u>0.4237</u>	0.6397	0.4333	0.5711

Saliency Prediction To evaluate the spatial realism of our generated eye-tracking sequences, we compared our method against several models trained specifically for the saliency prediction task. The baselines include: the SUM model [50], which integrates the Mamba architecture with a U-Net to output saliency maps across diverse image types; TranSalNet [49], which leverages transformers for saliency prediction; and DeepGaze I [47] and DeepGaze IIE [48], which are based on pretrained convolutional neural networks.

Table 5: Saliency prediction comparison. Bold is best and underline is second best.

Method	AUC-Judd \uparrow	AUC-Borji \uparrow	NSS \uparrow	SIM \uparrow	CC \uparrow	KL \downarrow
DeepGaze I	0.883	0.766	2.306	0.484	0.580	<u>0.980</u>
DeepGaze IIE	<u>0.923</u>	0.830	<u>3.321</u>	0.618	0.794	0.552
TranSalNet	0.896	0.874	2.443	0.508	0.658	6.369
SUM	0.931	<u>0.8458</u>	3.611	0.727	0.878	1.438
DiffEye	0.832	0.737	1.991	0.447	0.527	1.991

For each image in the test set, we generated 15 eye-tracking trajectories using our method. From each trajectory, we extracted fixation points using a script provided by [18], and used these to create individual fixation maps. These maps were aggregated and convolved with a Gaussian kernel to produce a single saliency map per image. For the baseline methods, which directly output saliency maps, we passed the same test images through each model. We then compared all predicted saliency maps, including ours, to the ground truth saliency maps provided in the dataset using six standard metrics: AUC-Judd, AUC-Borji, Normalized Scanpath Saliency (NSS), Similarity (SIM), Pearson’s Correlation Coefficient (CC), and Kullback–Leibler Divergence (KL). Please refer to [69] for a comprehensive detailing of the saliency metrics used. Fig. 8 shows additional examples of saliency maps generated by DiffEye and the baselines for the MIT1003 dataset and Table 5 reports the quantitative results.

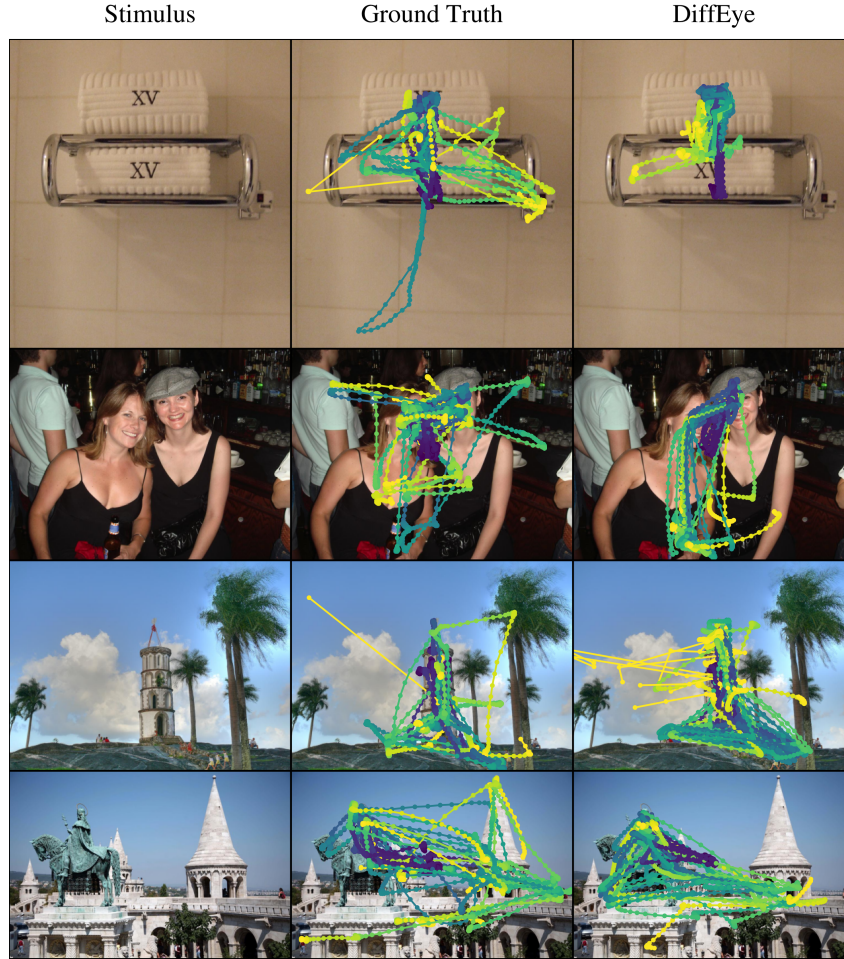


Figure 7: **Additional qualitative results of continuous eye movement trajectory generation.** Additional eye movement trajectories generated by DiffEye alongside ground truth annotations across four different scenes.



Figure 8: **Additional qualitative results for saliency prediction.** Saliency maps generated by DiffEye and baseline models are shown alongside ground truth maps for four different scenes. Each row corresponds to a different stimulus image, with columns displaying the stimulus, ground truth saliency map, and predictions.

A.6 Analysis of Statistical Properties of Scanpath Generation

To evaluate the plausibility of the generated scanpaths, we compare their statistical properties against those of ground truth human eye movements. Figure 9 shows the distributions of three key metrics: saccade amplitude, saccade direction, and inter-saccade angle for both the MIT1003 and OSIE datasets.

Our model’s performance (blue line) demonstrates a strong alignment with the ground truth distributions (black dashed line) across all three metrics. In the saccade amplitude distribution, our model successfully captures the peak at lower pixel values. For saccade direction, our model accurately reflects the horizontal bias present in human vision (peaks at 0 and ± 180 degrees). Finally, in the inter-saccade angle distribution, our model correctly shows a strong tendency for forward movements (peak near 0 degrees) and return saccades (smaller peak near ± 180 degrees). These results are consistent across both datasets, confirming that our approach generates statistically more realistic scanpaths than the compared methods.

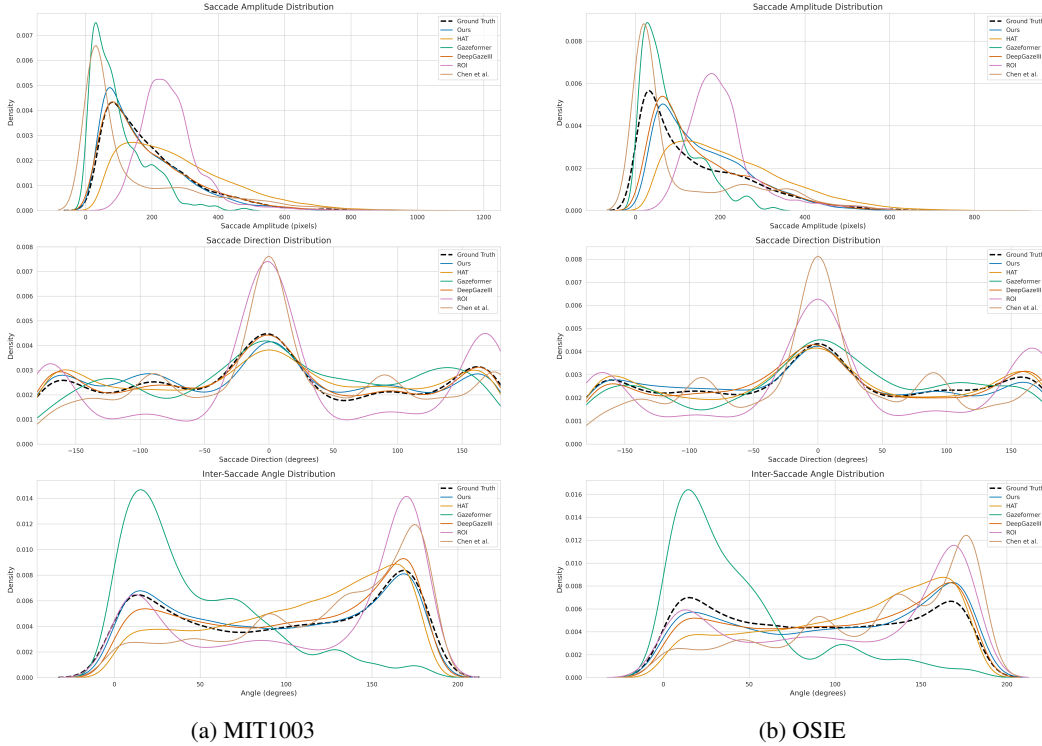


Figure 9: Comparison of statistical properties for generated scanpaths on the (a) MIT1003 and (b) OSIE datasets. The distributions for saccade amplitude, saccade direction, and inter-saccade angle of our model (blue) are compared against ground truth human scanpaths (black, dashed) and several other methods. Our model consistently provides the closest fit to the ground truth distributions.