ICCV HONOLULU HAWAII
OCT 19-23, 2025

Project Website

Modeling Human Gaze Behavior with Diffusion Models for Unified Scanpath Prediction

Giuseppe Cartella[1], Vittorio Cuculo[1], Alessandro D'Amelio[2], Marcella Cornia[1], Giuseppe Boccignone[2], Rita Cucchiara[1]

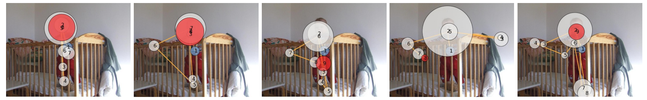
[1]{name.surname}@unimore.it [2]{name.surname}@unimi.it

UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA — AImage Lab

UNIVERSITÀ DEGLI STUDI DI MILANO — PHUSE Lab

1. The Variability of Human Visual Exploration

Human Visual Exploration is inherently variable, even for the same image/task.



- The decision of where to look next at any given moment is neither entirely deterministic nor completely random [1].
- The stochasticity of gaze allocation enables the observer to remain responsive to new signals and promotes a **flexible shift of attention**.

The Failure of SOTA Models

Most existing approaches generate **averaged behaviors**, failing to capture the variability of human visual exploration.

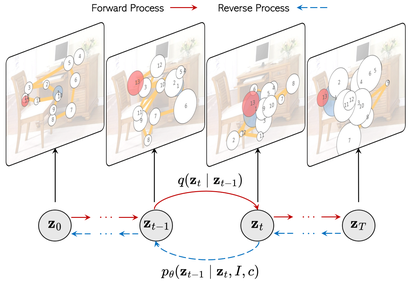
However, achieving good performance in terms of MultiMatch (MM), ScanMatch (SM) or Sequence Score (SS) **DOES NOT** mean that the model is able to consider the inter-subject variability.

2. Diffusion Models are a Natural Fit

Our Idea

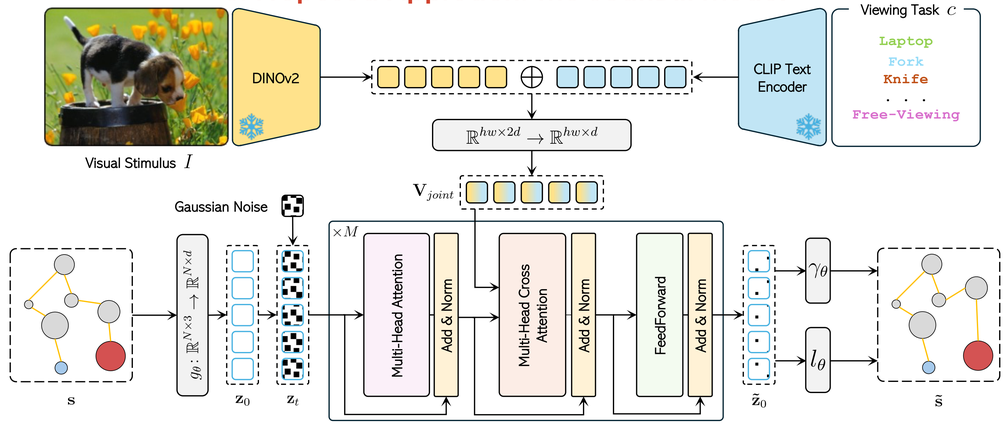
Leverage **stochastic sampling** to model **uncertainty** in gaze allocation.

Main Contributions



- A **diffusion-based architecture** for scanpath prediction that reflects the inter-group variability.
- A **novel analysis** on the diversity of the generated scanpaths.
- Introduction of the **Diversity-Aware Sequence Score**.

3. Proposed Approach: The ScanDiff Model



1. The **scanpath** is embedded into the **initial uncorrupted** latent variable z_0 .
2. **Corrupt** the whole embedded sequence z_0 by adding **Gaussian noise** over T timesteps.
3. The **stimulus** is encoded through a **Transformer-based visual backbone**.
4. A text encoder embeds the **viewing task**, enabling a unified architecture.
5. Visual and textual features are projected in a **joint multimodal embedding space**.
6. A modified Transformer encoder processes the noisy embedded scanpath sequence z_t and the multimodal features serve as **conditioning** for the **denoising process**.
7. A three layers MLP γ_θ **reconstructs** the original scanpath and a **length prediction module** l_θ predicts the **scanpath length**.

The **training objective** is defined as the combination of four different components: $\mathcal{L} = \mathcal{L}_{\text{VLB}} + \mathcal{L}_{rec} + \mathcal{L}_{val} + \mathcal{L}_T$

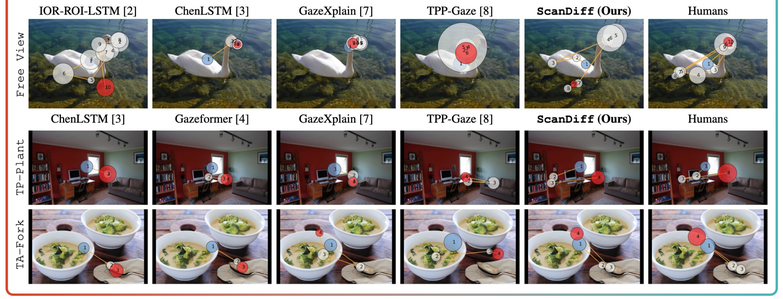
4. Comparison with the State-of-the-Art

	COCO-FreeView			MIT1003			COCO-TP				COCO-TA			
	MM ↓	SM ↓	SS ↓	MM ↓	SM ↓	SS ↓	MM ↓	SM ↓	SS ↓	SemSS ↓	MM ↓	SM ↓	SS ↓	SemSS ↓
ChenLSTM [3]	0.100	0.574	0.344	0.094	0.168	0.192	0.197	0.011	0.040	0.084	0.075	0.010	0.036	0.044
Gazeformer [4]	-	-	-	-	-	-	0.281	0.027	0.119	0.131	0.089	0.061	0.102	0.085
ChenLSTM-ISP [5]	0.128	0.683	0.377	0.108	0.264	0.267	0.174	0.013	0.015	0.043	0.082	0.028	0.063	0.052
HAT [6]	0.505	-	-	1.121	-	-	0.118	-	-	-	0.052	-	-	-
GazeXplain [7]	0.353	3.915	2.278	0.082	0.035	0.072	0.166	0.023	0.070	0.140	0.046	0.062	0.038	0.043
ChenLSTM [3]	0.280	0.116	0.022	0.156	0.373	0.284	-	-	-	-	-	-	-	-
Gazeformer [4]	-	-	-	-	-	-	0.251	0.045	0.048	0.095	0.526	1.184	0.319	0.671
GazeXplain [7]	0.141	0.049	0.017	0.048	0.158	0.128	0.167	**0.010**	0.050	0.092	0.037	0.030	0.028	0.038
TPP-Gaze [8]	**0.038**	0.125	0.033	0.062	0.244	0.144	0.507	2.317	0.893	0.736	0.135	0.775	0.427	0.231
ScanDiff (Ours)	0.078	**0.015**	**0.013**	**0.040**	**0.041**	**0.026**	**0.048**	0.037	**0.019**	**0.072**	**0.020**	**0.005**	**0.008**	**0.007**

We adopt the same evaluation protocol proposed in [7], and use the **KL divergence**.

ScanDiff achieves SOTA In **scanpath prediction** generating **diverse** and **plausible** gaze trajectories!

5. ScanDiff produces Human-Like Scanpaths



6. Scanpath Variability Analysis

Metrics such as MM, SM, and SS tend to reward models that generate a single representative scanpath **(averaged behavior)** [9].

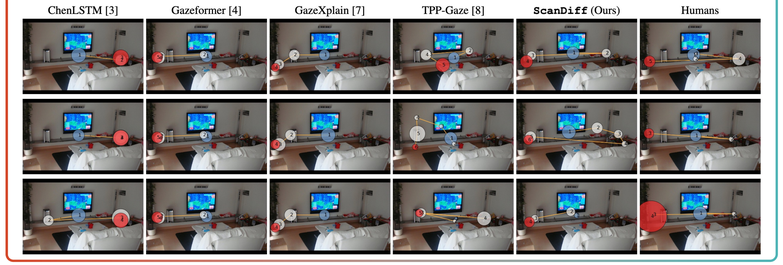
This problem stems from the way metrics are computed.

We propose the Diversity-Aware Sequence Score

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

A **penalization term** that penalizes excessive similarity among the generated scanpaths when humans **do not reflect** such behavior.

	COCO-FV		MIT1003		TP		TA	
	DSS ↑	RSS ↑	DSS ↑	RSS ↑	DSS ↑	RSS ↑	DSS ↑	RSS ↑
ChenLSTM [3]	0.174	0.420	0.257	0.534	0.386	0.635	0.247	0.591
Gazeformer [4]	-	-	-	-	0.377	0.578	0.206	0.417
ChenLSTM-ISP [5]	0.190	0.501	0.264	0.619	0.423	0.735	0.268	0.670
HAT [6]	-	0.645	-	0.615	-	0.861	-	0.748
GazeXplain [7]	0.099	0.032	0.302	0.674	0.406	0.689	0.283	0.716
TPP-Gaze [8]	0.271	0.732	0.313	0.758	0.284	0.516	0.221	0.663
ScanDiff (Ours)	**0.277**	**0.736**	**0.354**	**0.815**	**0.425**	**0.747**	**0.312**	**0.800**



References

[1] Canosa et al., "*Selective Perception and Task*", ACM Transactions on Applied Perception 2009
[2] Chen et., "Scanpath Prediction for Visual Attention using IOR-ROI LSTM", IJCAI 2018
[3] Chen aet al., "*Predicting Human Scanpaths in Visual Question Answering*", CVPR 2021
[4] Mondal et al, "*Gazeformer: Scalable, Effective and Fast Prediction of Goal-Directed Human Attention*", CVPR 2023
[5] Chen et al., "*Beyond Average: Individualized Visual Scanpath Prediction*", CVPR 2024
[6] Yang et al., "*Unifying Top-down and Bottom-up Scanpath Prediction Using Transformers*", CVPR 2024
[7] Chen et al., "*GazeXplain: Learning to Predict Natural Language Explanations of Visual Scanpaths*", ECCV 2024
[8] D'Amelio et al., "*TPP-Gaze: Modelling Gaze Dynamics in Space and Time with Neural Temporal Point Processes*", WACV 2025
[9] Kümmerer et al., "*State-of-the-art in human scanpath prediction*", arXiv 2021