Is this it? Benchmarking Scanpath Metrics for Information Display

Yao Wang
University of Stuttgart
Stuttgart, Germany
yao.wang@vis.uni-stuttgart.de

Junichi Nagasawa
University of Stuttgart
Stuttgart, Germany
jnagasawa@acm.org

Danqing Shi
University of Cambridge
Cambridge, United Kingdom
ds2206@cam.ac.uk

Chuhan Jiao
University of Stuttgart
Stuttgart, Germany
chuhan.jiao@vis.uni-stuttgart.de

Yue Jiang*
University of Utah
Salt Lake City, United States
yue.jiang@utah.edu

Andreas Bulling
University of Stuttgart
Stuttgart, Germany
andreas.bulling@vis.uni-stuttgart.de

Abstract

Scanpath prediction is a fundamental task in human visual attention research, aiming to simulate user viewing behaviour for given stimuli. While scanpath prediction methods have matured in natural scenes, recent research has expanded to information displays, such as graphical user interfaces and data visualisations. However, there is currently no consensus on which scanpath metrics to use for evaluation, raising concerns regarding the validity and comparability of the proposed methods. This paper benchmarks ten commonly used scanpath metrics across the MASSVIS and UEyes datasets by comparing model predictions with empirical gaze data. We evaluate these metrics with subjective expert ratings of scanpath similarity. Our analysis reveals that vector-based and region-based metrics align more closely with expert ratings than pixel-based and recurrence-based metrics. Based on these findings, we provide best practices for evaluating visual scanpaths in information displays, emphasising the urgent need for appropriate metrics to ensure the validity of future research.

CCS Concepts

• **Human-centered computing** → **Graphical user interfaces**; **Empirical studies in visualization**.

Keywords

Scanpath metrics, scanpath prediction, graphical user interface, data visualisation

ACM Reference Format:
Yao Wang, Junichi Nagasawa, Danqing Shi, Chuhan Jiao, Yue Jiang, and Andreas Bulling. 2026. Is this it? Benchmarking Scanpath Metrics for Information Display. In *2026 Symposium on Eye Tracking Research and Applications (ETRA '26), June 01–04, 2026, Marrakesh, Morocco.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3797246.3805691

*Corresponding author

ETRA '26, Marrakesh, Morocco

1 Introduction

Fixations and saccades are the most common eye movement events. Fixations occur when gaze velocity falls below a threshold for a sufficient duration (usually longer than 50 ms), whereas saccades occur when gaze rapidly shifts to another position. The combination of these fixations and saccades creates what is known as a visual scanpath [Noton and Stark 1971]. This spatial-temporal data on eye movements provides a valuable source of information for understanding individual viewing behaviour across several conditions in information displays (e.g. graphical user interfaces and data visualisations), such as free-browsing [Eraslan et al. 2016], visual search [Putkonen et al. 2025], and chart reading [Polatsek et al. 2018; Wang et al. 2024b].

Since collecting human gaze data is time-wise and budget-wise costly, researchers have focused on developing scanpath prediction methods that simulate human-like eye movements [Itti et al. 1998]. The modelling of human visual scanpaths originated with heuristic approaches based on bottom-up saliency [Brockmann and Geisel 2000] and cognitive biases, such as inhibition of return [Itti et al. 1998]. These works served as a foundation for more sophisticated models, such as feature-based machine learning [Judd et al. 2009], Hidden Markov Models [Coutrot et al. 2018], recurrent neural networks [Sun et al. 2019], and reinforcement learning [Yang et al. 2020]. With the availability of large-scale eye tracking datasets [Chen et al. 2021; Jiang et al. 2015] and modern architectures such as transformers, current state-of-the-art methods have achieved promising performance, including DeepGaze III [Kümmerer et al. 2022] and ScanDiff [Cartella et al. 2025]. While these methods matured in natural scenes, research has only a few pioneering works that have tailored scanpath prediction methods to information displays, specifically GUIs [Jiang et al. 2024; Jokinen et al. 2020; Shi et al. 2024] and data visualizations [Shi et al. 2025; Wang et al. 2024a].

Scanpath metrics are key for evaluating whether these prediction methods reproduce human-like viewing behaviour. Unlike saliency metrics that compare aggregated attention distributions [Judd et al. 2012], scanpath metrics typically operate on a pairwise basis – comparing a generated scanpath against a human ground truth. Consequently, evaluation results are derived by averaging these pairwise scores. The scanpath metric collapses two sets of fixation sequences into a single scalar, often obscuring underlying variance and potentially masking model failures. Additionally, previous work

has shown that standard scanpath metrics can contradict expert ratings when judging which predictions most closely match human viewing behaviour [Jiao et al. 2026; Wang et al. 2024a].

This paper argues that the key challenge for current scanpath metrics lies in the fundamental difference between natural scenes and information displays. Although a pioneering study highlighted disagreement between several scanpath metrics and expert ratings [Wang et al. 2024a] in data visualisations, the validity of scanpath metrics for information displays has not been systematically examined. All scanpath metrics were designed for natural scenes, but previous research has directly used them to evaluate scanpath methods in information displays [Jiang et al. 2023] and even to guide design decisions [Emami et al. 2024]. User interfaces and data visualisations are text-rich, structured, hierarchical, and contain significant white spaces [Matzen et al. 2017]. Therefore, human viewing behaviour in information displays differs fundamentally from that in natural scenes. The information displays induce distinct viewing patterns that metrics optimised for natural scenes fail to capture, leaving a huge question mark about whether scanpath metrics remain reliable. To systematically examine the validity of scanpath metrics in the context of information displays, this paper makes the following contributions:

- We benchmark ten commonly used scanpath metrics across four state-of-the-art prediction models. We provide AOI-based gaze statistics as an alternative comprehensive perspective for scanpath evaluation.
- We conduct a user study with eight experts to assess scanpath similarity, revealing that widely used pixel-based and recurrence-based metrics do not align with expert ratings.
- We provide actionable recommendations and best practices for evaluating scanpath prediction methods on information displays, ensuring the validity of future research.

2 Scanpath Metrics Benchmarking

2.1 Datasets, Metrics, Methods, and Evaluation Protocol

Datasets. We used 108 stimuli from the UEyes dataset [Jiang et al. 2023] and 65 stimuli from the MASSVIS dataset [Borkin et al. 2015, 2013] for our benchmark. All stimuli were from the test sets, ensuring that the predictive models had no prior exposure to them during training. The selection covers a diverse range of scenes, consisting of 27 webpages, 27 desktop UIs, 27 mobile UIs, and 27 posters from UEyes, along with 28 infographics and 38 charts from MASSVIS.

Metrics. We benchmarked ten commonly used scanpath metrics categorised into a) pixel-based, b) vector-based, c) region-based, and d) recurrence-based groups:

a.1 *Dynamic Time Warping (DTW) [Müller 2007]* minimises the distance between two temporal sequences by non-linearly warping the time dimension. Lower values indicate higher similarity.

a.2 *Scaled Time-Delay Embedding (sTDE) [Takens 2006; Wang et al. 2011]* compares scanpaths by embedding them into a phase space that preserves temporal dynamics. The score is normalised between 0 (worst) and 1 (best).

b.1 *Eyenalysis [Mathot et al. 2012]* aligns scanpaths as continuous trajectories in the temporal domain to minimise spatial distance. The resulting similarity score is normalised between 0 (worst) and 1 (best).

b.2 *MultiMatch (MM) [Jarodzka et al. 2010]* computes similarity across five dimensions: shape, length, position, direction, and duration. Each dimension is normalised between 0 (worst) and 1 (best).

c.1 *Levenshtein Distance (LEV) [Levenshtein 1965]* measures the minimum number of single-character edits required to transform one string-encoded scanpath into another. The value is normalised by sequence length, where 0 represents a perfect match.

c.2 *ScanMatch [Cristino et al. 2010]* uses a substitution matrix to spatially weight costs, then penalises mismatches by the Needleman–Wunsch [Needleman and Wunsch 1970] algorithm. The score is normalised between 0 (worst) and 1 (best).

c.3 *Sequence Score [Yang et al. 2020]* aligns AOI-based strings with the Needleman-Wunsch algorithm. The score is normalised between 0 (worst) and 1 (best).

d.1 *Cross-Recurrence (REC) [Anderson et al. 2013]* measures the percentage of fixations in the predicted scanpath that spatially overlap with the ground truth, regardless of temporal order.

d.2 *Weighted Determinism (DET) [Anderson et al. 2013]* quantifies the percentage of recurrence points forming diagonal lines, indicating matching temporal sub-sequences.

d.3 *Center of Recurrence Mass (CORM) [Anderson et al. 2013]* calculates the distance of recurrence points from the main diagonal to assess temporal delay. Lower values indicate better temporal alignment.

Furthermore, we report the distribution of several foundational gaze statistics [Goldberg and Helfman 2010]. These include the total number of fixations, fixation ratios on specific elements (titles, data marks, and axes), the number of AOI transitions, and revisit counts for those elements. AOI annotations for MASSVIS were derived from Wang et al. [2024a], and UEyes is omitted due to the lack of AOI annotations.

Baseline scanpath prediction methods. We choose four scanpath prediction methods as baselines for benchmarking scanpath metrics. UMSS [Wang et al. 2024a] and EyeFormer [Jiang et al. 2024] are the state-of-the-art methods on the MASSVIS dataset and the UEyes dataset, respectively. Two strong baselines trained on natural images, DeepGaze III [Kümmerer et al. 2022] and ScanDiff [Cartella et al. 2025], are also included.

Evaluation protocol. For each visual stimulus and prediction method, we generated a set of predicted scanpaths equal in number to the ground truth scanpaths. The metrics are presented using *mean* and *best* protocols. For the *mean* protocol, all pairwise combinations of predicted and ground truth scanpaths are computed, and the average is reported. For the *best* protocol, each predicted scanpath was matched with the single ground truth scanpath that yielded the highest similarity score (or lowest cost), averaging these optimal values [Wang et al. 2024a]. For MultiMatch, the *best* match

is defined as the pair that maximised the mean value across all five dimensions.

2.2 Results

Table 1 and Table 2 present benchmarking results on the MASSVIS and UEyes datasets, comparing the human gold standard (computed using a leave-one-out policy) with four computational models using both *mean* and *best* aggregation methods. Our analysis reveals a misalignment between scanpath metrics and the actual model performance in information displays.

Pixel-based metrics. For MASSVIS, DTW and sTDE fail to meaningfully distinguish human behaviour from model-generated predictions. They incorrectly assign the highest scores to ScanDiff and EyeFormer, which generate significantly fewer fixations compared to the human gold standard. For UEyes, UMSS ranked last for DTW and sTDE due to the large number of fixations (22.5). This aligns with prior findings that such metrics exhibit a strong bias towards shorter scanpaths [Wang et al. 2024a].

Vector-based metrics. MultiMatch and Eyenalysis are highly discriminative, consistently ranking the human baseline above model predictions for both datasets. For UEyes, the human has the highest Eyenalysis score under *mean* protocol, and all other MultiMatch and Eyenalysis scores prefer EyeFormer.

Region-based metrics. The *mean* LEV scores are indistinguishable between humans and models in information displays. Furthermore, its *best* scores incorrectly favour unfinetuned predictive models (EyeFormer for MASSVIS, and DeepGaze III for UEyes). In contrast, ScanMatch and Sequence Score reliably separate human behaviour from model predictions.

Recurrence-based metrics. Recurrence-based metrics largely fail in information displays. While the *mean* of REC correctly ranks human spatial overlap highest, all other measures break down. DET always favours human scanpaths, whereas CORM assigns exceptionally high scores to DeepGaze III and EyeFormer.

Gaze statistics. While scanpath models struggle with fixation counts, they are all within one standard deviation of human behaviour for Fixation-on-axis, Fixation-on-title ratios, and Title, Mark, and Axis Revisits. For AOI Transitions, only UMSS (12.4) performs close to human behaviour (14.1).

3 User Study

While scanpath metrics are widely used for natural scenes, the validity remains unknown for information displays. Given the evidence that human experts can reliably interpret gaze patterns [Jiao et al. 2026; Wang et al. 2024a], we use expert ratings as a ground truth for what constitutes human-like viewing behaviour. We recruited eight researchers with extensive experience, each having conducted at least three eye tracking studies. Each expert evaluated 20 images (10 MASSVIS, 10 UEyes). For each image, three human scanpaths were randomly picked as the reference scanpath. The tasks were designed as forced-choice ranking questions. For each question, participants were presented with one human scanpath as a reference, alongside another unlabelled human scanpath shuffled with four model predictions (UMSS, DeepGaze III, ScanDiff, EyeFormer). All scanpaths were shown in looped animation and overlaid on the image. Participants were asked to carefully examine and rank the 5 scanpaths relative to the reference scanpath, from most similar (1) to least similar (5), based on their professional judgment. To mitigate layout and order biases, candidate presentation was randomised using a 5×5 Latin square. The study took approximately 20 minutes to complete.

Results. Expert ratings are reported in Table 1, Table 2, and illustrated in Figure 1. A Friedman Test revealed significant differences in the overall ratings among the five methods for both the MASSVIS ($\chi^2(4, N = 30) = 34.03, p < 0.001$) and UEyes ($\chi^2(4, N = 30) = 24.43, p < 0.001$) datasets. For MASSVIS, the human ground truth scanpaths achieved the lowest mean rating (1.54). Post-hoc Wilcoxon Signed-rank Tests with a Bonferroni correction confirmed that the human scanpaths were rated significantly better than all other prediction methods ($p < 0.05$). For UEyes, EyeFormer achieved the lowest mean rating (2.21), closely followed by the human scanpaths (2.34). Post-hoc comparisons indicated no significant difference between EyeFormer and the human ($p = 0.77$), though both performed significantly better than the remaining methods ($p < 0.05$).

4 Discussion

4.1 Influence of Visual Stimuli

From natural scenes to information displays. Gaze behaviour during task-free viewing differs fundamentally between natural scenes and structured information displays. Unlike natural scenes, information displays – such as GUIs and data visualisations – impose strong structural and semantic constraints that drive distinct, goal-directed visual search patterns. Our results demonstrate that pixel-based metrics (DTW, sTDE), recurrence-based metrics (REC, DET, CORM), and the standard Levenshtein Distance (LEV) fail to capture these domain-specific constraints and frequently contradict expert human ratings. Specifically, pixel- and recurrence-based metrics lack semantic awareness and thus ignore the functional hierarchies inherent to information displays. Furthermore, the foundational assumptions of recurrence [Anderson et al. 2013] do not translate to the non-linear, hierarchical scanning behaviours typical of information displays. Similarly, the failure of LEV highlights that basic string-alignment metrics can be easily misled by scanpaths that visit the correct semantic elements (e.g. data marks) but in a spatially incoherent order. Conversely, the ScanMatch and Sequence Score are reliable metrics. They incorporate spatial weighting or element annotations, which are essential for navigating grid-like stimuli such as charts or mobile UIs. Vector-based metrics, including MultiMatch and Eyenalysis, also demonstrate high reliability. By evaluating the holistic geometric trajectory rather than isolated fixation coordinates, they remain robust against the structural domain gaps between natural scenes and information displays.

From 2D displays to 3D environments. Immersive environments introduce more challenges for the validity of scanpath metrics. First, the resolution of immersive environments is usually 8192×4096 pixels or bigger, and the viewing duration easily exceeds 30 seconds. This extended image space and duration (number of fixations) degrade scanpath similarity scores significantly [Jiao et al. 2026].

Table 1: Scanpath evaluation on the MASSVIS dataset in terms of scanpath metrics (a.1-d.3) and gaze statistics (bottom 8 rows). The best results for each metric are shown in bold, and those that contradict expert ratings are coloured in red. Gaze statistics within 1 standard deviation of humans are marked in light green.

Group	Metrics		Human†	UMSS	DeepGaze III	ScanDiff	EyeFormer
	Expert Rating ↓		**1.54** (0.32)	2.88 (0.50)	3.93 (0.44)	4.00 (0.41)	2.65 (0.29)
a.1	**DTW ↓**	*mean*	9,518 (2,901)	9,209 (2,421)	9,301 (2,721)	8,964 (2,738)	**7,716** (2,294)
		best	7,177 (2,138)	6,447 (1,575)	5,426 (1,666)	5,254 (1,569)	**4,608** (1,235)
a.2	**sTDE ↑**	*mean*	**.928** (.018)	.914 (.018)	.914 (.023)	.921 (.025)	.925 (.017)
		best	.940 (.015)	.927 (.013)	.933 (.016)	**.941** (.015)	.939 (.012)
a.3	**Eyenalysis ↓**	*mean*	**67.2** (27.9)	90.4 (28.5)	138.9 (42.4)	130.3 (44.4)	102.3 (28.9)
		best	**49.2** (17.1)	67.7 (16.5)	98.3 (27.4)	92.7 (28.3)	72.0 (16.8)
b.1	**MM-Shape ↑**	*mean*	**.957** (.010)	.934 (.007)	.945 (.008)	.953 (.015)	.952 (.009)
		best	**.964** (.013)	.939 (.015)	.952 (.014)	.960 (.012)	.957 (.012)
b.1	**MM-Direction ↑**	*mean*	**.784** (.010)	.703 (.044)	.659 (.034)	.662 (.105)	.714 (.047)
		best	**.847** (.065)	.764 (.073)	.753 (.081)	.751 (.090)	.790 (0.071)
b.1	**MM-Length ↑**	*mean*	**.952** (.012)	.926 (.024)	.941 (.023)	.942 (.025)	.948 (.019)
		best	**.959** (.016)	.933 (.021)	.948 (.019)	.952 (.018)	.954 (.016)
b.1	**MM-Position ↑**	*mean*	**.796** (.032)	.751 (.071)	.742 (.072)	.757 (.088)	.787 (.060)
		best	**.867** (.046)	.799 (.051)	.790 (.056)	.820 (.060)	.833 (.045)
b.2	**MM-Duration ↑**	*mean*	**.710** (.024)	.618 (.089)	-	.707 (.099)	.512 (.075)
		best	.749 (.044)	.687 (.068)	-	**.756** (.078)	.593 (.079)
c.1	**LEV ↓**	*mean*	60.9 (8.6)	**60.6** (10.6)	64.5 (12.5)	63.3 (12.3)	60.8 (11.5)
		best	52.7 (8.1)	42.8 (7.9)	40.7 (12.6)	39.8 (12.1)	**39.1** (10.3)
c.2	**ScanMatch ↑**	*mean*	**.516** (.118)	.363 (.087)	.220 (.052)	.239 (.077)	.301 (.059)
		best	**.639** (.010)	.471 (.087)	.318 (.073)	.341 (.104)	.416 (.091)
c.3	**Sequence Score ↑**	*mean*	**.522** (.117)	.388 (.015)	.238 (.071)	.247 (.085)	.315 (.084)
		best	**.540** (.209)	.499 (.112)	.322 (.078)	.330 (.010)	.416 (.088)
d.1	**REC ↑**	*mean*	**3.26** (2.36)	1.99 (1.68)	2.10 (2.51)	2.67 (4.18)	1.70 (1.74)
		best	5.54 (3.14)	4.48 (2.36)	6.55 (3.19)	**9.28** (7.41)	4.55 (2.16)
d.2	**DET ↑**	*mean*	**23.7** (19.9)	3.96 (12.1)	2.05 (11.6)	4.03 (15.2)	4.17 (15.8)
		best	**46.4** (18.3)	30.4 (25.6)	22.5 (34.0)	32.3 (36.0)	34.3 (37.3)
d.3	**CORM ↓**	*mean*	28.1 (13.8)	30.2 (20.2)	19.8 (25.7)	**15.2** (21.1)	26.6 (28.9)
		best	14.7 (10.5)	4.69 (7.84)	**0.12** (1.39)	0.15 (1.52)	**0.12** (1.39)
	Number of Fixations		37.6 (6.7)	22.6 (3.7)	11.0 (0.0)	11.4 (2.9)	15 (0.0)
	Fixation-on-title Ratio (%)		10.0 (15.3)	6.6 (12.3)	6.0 (13.8)	8.1 (16.4)	5.3 (10.9)
	Fixation-on-mark Ratio (%)		27.3 (20.8)	39.7 (25.6)	52.1 (30.2)	37.8 (25.9)	46.2 (27.3)
	Fixation-on-axis Ratio (%)		3.2 (9.9)	3.6 (9.7)	3.7 (12.4)	2.7 (8.7)	3.3 (10.0)
	AOI Transitions		14.1 (5.5)	12.4 (4.7)	5.2 (2.4)	5.2 (2.4)	6.2 (2.5)
	Title Revisit		1.8 (2.3)	1.1 (1.7)	0.4 (0.9)	0.5 (0.8)	0.6 (1.0)
	Mark Revisit		4.0 (2.3)	3.9 (1.9)	2.0 (1.1)	1.8 (1.0)	2.2 (1.2)
	Axis Revisit		0.43 (1.2)	0.57 (1.4)	0.23 (0.7)	0.20 (0.6)	0.25 (0.7)

† Scanpaths are not compared with themselves

Second, gaze in head-mounted devices relies on combined eye and head movements. Combined with the low sampling rates (30–60 Hz) of integrated mobile trackers, this composite motion increases gaze estimation error, leading to high uncertainty in AOI-based metrics [Wang et al. 2022]. Finally, gaze depth creates spatial ambiguity, particularly under visual occlusion, where objects overlap from the viewer's perspective [Koch et al. 2024]. To address this, future metrics should incorporate depth information to ensure robust evaluation in 3D environments [Chen et al. 2024].

Table 2: Scanpath evaluation on the UEyes Dataset. The best results for each metric are shown in bold, and those contradicted with expert ratings are marked in red.

Group	Metrics		Human†	UMSS	DeepGaze III	ScanDiff	EyeFormer
	Expert Rating ↓		2.34 (0.40)	3.27 (0.61)	3.71 (0.51)	3.47 (0.71)	**2.21** (0.55)
a.1	**DTW ↓**	*mean*	4,713 (1,932)	6,963 (2,606)	4,933 (2,026)	5,044 (2,122)	**4,229** (1,471)
		best	3,772 (1,768)	5,519 (2,376)	3,487 (1,616)	3,496 (1,699)	**3,070** (1,253)
a.2	**sTDE ↑**	*mean*	.899 (.106)	.894 (.076)	.895 (.078)	.900 (.079)	**.912** (.076)
		best	.919 (.076)	.917 (.015)	.923 (.017)	.930 (.018)	**.936** (.012)
a.3	**Eyenalysis ↓**	*mean*	**110.2** (54.4)	124.9 (56.5)	153.8 (71.6)	146.1 (67.6)	114.8 (44.5)
		best	87.7 (46.0)	90.3 (39.4)	108.9 (49.0)	101.3 (46.6)	**84.2** (33.0)
b.1	**MM-Shape ↑**	*mean*	.933 (.021)	.916 (.019)	.928 (.020)	.939 (.019)	**.941** (.016)
		best	.935 (.079)	.923 (.037)	.939 (.016)	.947 (.015)	**.951** (.013)
b.1	**MM-Direction ↑**	*mean*	.738 (.094)	.712 (.083)	.705 (.093)	.694 (.106)	**.748** (.087)
		best	.798 (.091)	.777 (.065)	.791 (.057)	.777 (.073)	**.819** (0.053)
b.1	**MM-Length ↑**	*mean*	.930 (.027)	.908 (.029)	.922 (.031)	.928 (.030)	**.939** (.023)
		best	.933 (.080)	.915 (.042)	.934 (.024)	.935 (.024)	**.947** (.017)
b.1	**MM-Position ↑**	*mean*	.805 (.076)	.773 (.068)	.770 (.076)	.782 (.088)	**.819** (.055)
		best	.843 (.088)	.806 (.060)	.815 (.056)	.829 (.058)	**.853** (.041)
b.1	**MM-Duration ↑**	*mean*	.696 (.095)	.577 (.085)	-	.533 (.113)	**.737** (.075)
		best	.745 (.093)	.633 (.073)	-	.587 (.107)	**.782** (.051)
c.1	**LEV ↓**	*mean*	25.8 (5.7)	35.5 (6.4)	**25.3** (5.4)	26.3 (4.9)	25.5 (4.1)
		best	22.5 (7.0)	31.3 (8.9)	**19.0** (5.4)	20.6 (5.9)	20.8 (5.2)
c.2	**ScanMatch ↑**	*mean*	.490 (.149)	.547 (.116)	.453 (.146)	.467 (.170)	**.501** (.150)
		best	.593 (.131)	.431 (.134)	.564 (.135)	.596 (.156)	**.605** (.134)
d.1	**REC ↑**	*mean*	**2.87** (3.36)	1.72 (2.36)	1.59 (2.41)	1.87 (2.99)	1.90 (2.86)
		best	4.92 (4.66)	4.08 (4.00)	4.34 (3.61)	**5.31** (4.91)	5.16 (5.63)
d.2	**DET ↑**	*mean*	**6.36** (17.2)	2.31 (11.0)	1.31 (9.19)	2.64 (12.67)	2.51 (12.3)
		best	**16.8** (26.7)	12.2 (24.8)	8.16 (22.3)	15.2 (28.6)	14.9 (28.2)
d.3	**CORM ↓**	*mean*	27.0 (21.4)	25.7 (22.7)	18.9 (25.3)	**15.8** (20.7)	23.8 (23.6)
		best	12.0 (15.3)	3.37 (7.97)	**0.77** (4.16)	1.27 (5.07)	2.21 (6.30)
	Number of Fixations		15.5 (3.8)	22.5 (3.8)	11.0 (0.0)	13.0 (3.0)	15 (0.0)

† Scanpaths are not compared with themselves

4.2 Influence of Evaluation Procedure

Expert rating. Finetuned scanpath prediction models specific to each dataset (UMSS for MASSVIS and EyeFormer for UEyes) consistently ranked in the top two positions alongside the human ground truth. This provides strong evidence that *1) experts are able to distinguish human-like viewing behaviours*, and *2) domain-specific finetuning significantly improves scanpath models.* The stark contrast between several metric scores and expert ratings exposes a critical vulnerability in current evaluation pipelines. Experts self-reported that their primary criteria are reading behaviour, the overall shape of the scanpath, and accurate fixations on critical semantic areas such as chart titles.

Human data scale. The statistical power of scanpath evaluation depends heavily on the scale of human data. However, the number of participants per stimulus is significantly lower than in natural scene benchmarks. For instance, the CAT2000 [Judd et al. 2012] benchmark has 120 participants per stimulus, whereas the MASSVIS and UEyes datasets, on average, have only 16.7 and 10.8 participants per stimulus, respectively. Given that human viewing behaviour is shaped by individual cognitive load and prior experience, small participant samples may lead to misinterpreting cross-subject variance as model failure.

Evaluation pipeline. The lack of a standardised evaluation procedure undermines the reliability of several scanpath metrics. For instance, handling fixations that fall outside stimulus boundaries remains a challenge; while excluding them or assigning them to the nearest AOI are common fixes, both methods introduce artificial bias into the resulting similarity scores. Meanwhile, several metrics are highly sensitive to thresholds. For example, REC requires a threshold of 1.9 visual degrees [Anderson et al. 2015], which

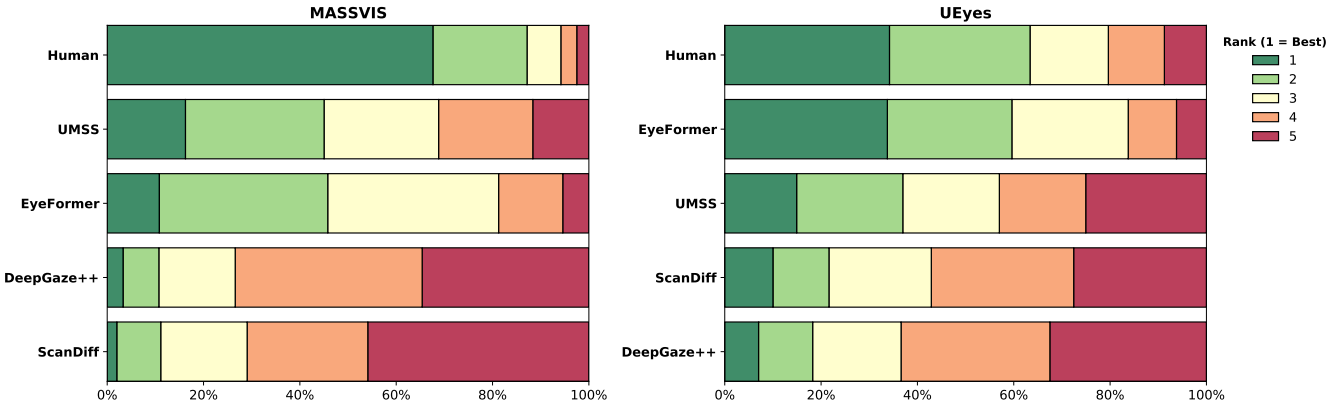

Figure 1: Aggregated expert ratings for the MASSVIS (left) and UEyes datasets (right).

cannot be accurately calculated if a dataset fails to report screen distance or physical stimulus size. Similarly, semantic-based metrics such as Sequence Score are further constrained by inconsistent AOI definitions. The inherent sensitivity of these metrics to AOI boundaries and foveal tolerance settings complicates cross-dataset comparisons and limits the generalisability of findings [Wang et al. 2022]. A robust solution would be to develop unified, automated AOI segmentation models specifically for information displays.

4.3 (Properly) Using Scanpath Metrics in Information Displays

Best practice. Based on our benchmark, the most reliable metrics with high discriminative power are *MultiMatch*, *ScanMatch*, and *Sequence Score*. For gaze statistics, the AOI Transition is a good indicator of scanpath similarity, as it distinguishes the finetuned UMSS from other unfinetuned methods (see Table 1). Only UMSS lies within 1 standard deviation of the human distribution. For the evaluation protocol, we strongly advise researchers evaluating information displays to aggregate scores using both the *best* and *mean* protocol. The *mean* protocol evaluates the alignment of the overall population distribution, rewarding models that capture the full diversity of human strategies. Meanwhile, the *best* protocol measures individual plausibility. Since human gaze in information displays exhibits high inter-observer variance, a generated scanpath might match a valid strategy, though having low similarity scores with most human scanpaths. Additionally, researchers must be aware of the limitations of metrics. For instance, Eyenalysis is insensitive to temporal order because the Mannan distance is permutation-invariant. Therefore, it functions more as a spatial saliency metric than a sequential measurement.

Beyond pairwise metrics: alternatives for evaluation. Given the limitations of existing pairwise metrics, one direction for future research is to compare population-level dynamics, such as Aggregated Scanpath Flow [Peysakhovich and Hurter 2018], or Multi-Duration Saliency [Fosco et al. 2020]. Alternatively, metrics could assess downstream utility, evaluating generated scanpaths based on their effectiveness in downstream tasks such as gaze-based UI personalisation or user authentication [Khamis et al. 2016]. If a predicted scanpath can successfully drive a personalised interface or provide a unique biometric signature, it may possess a functional fidelity that current metrics fail to measure.

5 Conclusion

As visual scanpath research expands to information displays like GUIs and data visualisations, ensuring the validity of scanpath metrics has become imperative. We benchmarked ten common scanpath metrics across the MASSVIS and UEyes datasets, comparing four scanpath prediction models with human ground truth in a user study with expert rating of scanpath similarity. Our analysis reveals significant vulnerabilities in pixel-based and recurrence-based metrics, which generate inconsistent and unreliable outputs in information displays. Conversely, vector-based metrics and region-based metrics are more reliable. Based on these findings, we recommend best practices and highlight the critical need to apply appropriate scanpath metrics to information displays to ensure the validity of future research.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161 (Project A07). D.S. acknowledges support from Turing AI Fellowship: Advancing Modern Data-Driven Robust AI (grant no. EP/W002965/1). C.J. is funded by the European Union's Horizon Europe research and innovation funding program under grant agreement No. 101072410.

References

Nicola C Anderson, Fraser Anderson, Alan Kingstone, and Walter F Bischof. 2015. A comparison of scanpath comparison methods. *Behavior research methods* 47, 4 (2015), 1377–1392.

Nicola C Anderson, Walter F Bischof, Kaitlin EW Laidlaw, Evan F Risko, and Alan Kingstone. 2013. Recurrence quantification analysis of eye movements. *Behavior research methods* 45, 3 (2013), 842–856.

Michelle A Borkin, Zoya Bylinskii, Nam Wook Kim, Constance May Bainbridge, Chelsea S Yeh, Daniel Borkin, Hanspeter Pfister, and Aude Oliva. 2015. Beyond memorability: Visualization recognition and recall. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2015), 519–528.

Michelle A Borkin, Azalea A Vo, Zoya Bylinskii, Phillip Isola, Shashank Sunkavalli, Aude Oliva, and Hanspeter Pfister. 2013. What Makes a Visualization Memorable? *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2306–2315.

Dirk Brockmann and Theo Geisel. 2000. The ecology of gaze shifts. *Neurocomputing* 32 (2000), 643–650.

Giuseppe Cartella, Vittorio Cuculo, Alessandro D'Amelio, Marcella Cornia, Giuseppe Boccignone, and Rita Cucchiara. 2025. Modeling Human Gaze Behavior with Diffusion Models for Unified Scanpath Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 16206–16216.

Xiaolin Chen, Lihua Lu, and Hui Wei. 2024. Identifying Fixations and Saccades in Virtual Reality. In *Proceedings of the 2024 International Conference on Virtual Reality Technology*. 24–31.

Yupei Chen, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Minh Hoai, and Gregory Zelinsky. 2021. Coco-search18 fixation dataset for predicting goal-directed attention control. *Scientific reports* 11, 1 (2021), 8776.

Antoine Coutrot, Janet H Hsiao, and Antoni B Chan. 2018. Scanpath modeling and classification with hidden Markov models. *Behavior Research Methods* 50, 1 (2018), 362–379.

Filipe Cristino, Sebastiaan Mathôt, Jan Theeuwes, and Iain D Gilchrist. 2010. ScanMatch: A novel method for comparing fixation sequences. *Behavior research methods* 42, 3 (2010), 692–700.

Parvin Emami, Yue Jiang, Zixin Guo, and Luis A Leiva. 2024. Impact of Design Decisions in Scanpath Modeling. *Proceedings of the ACM on Human-Computer Interaction* 8, ETRA (2024), 1–16.

Sukru Eraslan, Yeliz Yesilada, and Simon Harper. 2016. Scanpath trend analysis on web pages: Clustering eye tracking scanpaths. *ACM Transactions on the Web (TWEB)* 10, 4 (2016), 1–35.

Camilo Fosco, Anelise Newman, Pat Sukhum, Yun Bin Zhang, Nanxuan Zhao, Aude Oliva, and Zoya Bylinskii. 2020. How Much Time Do You Have? Modeling Multi-Duration Saliency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4473–4482.

Joseph H Goldberg and Jonathan I Helfman. 2010. Comparing information graphics: a critical look at eye tracking. In *Proceedings of the 3rd BELIV'10 Workshop: BEyond time and errors: novel evaLuation methods for Information Visualization*. 71–78.

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

Halszka Jarodzka, Kenneth Holmqvist, and Marcus Nyström. 2010. A vector-based, multidimensional scanpath similarity measure. In *Proceedings of the 2010 symposium on eye-tracking research & applications*. 211–218.

Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. 2015. Salicon: Saliency in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1072–1080.

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

Yue Jiang, Luis A Leiva, Hamed Rezazadegan Tavakoli, Paul RB Houssel, Julia Kylmälä, and Antti Oulasvirta. 2023. Ueyes: Understanding visual saliency across user interface types. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–21.

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

Jussi PP Jokinen, Zhenxin Wang, Sayan Sarcar, Antti Oulasvirta, and Xiangshi Ren. 2020. Adaptive feature guidance: Modelling visual search with graphical layouts. *International Journal of Human-Computer Studies* 136 (2020), 102376.

Tilke Judd, Frédo Durand, and Antonio Torralba. 2012. A benchmark of computational models of saliency to predict human fixations. (2012).

Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. 2009. Learning to predict where humans look. In *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 2106–2113.

Mohamed Khamis, Florian Alt, Mariam Hassib, Emanuel von Zezschwitz, Regina Hasholzner, and Andreas Bulling. 2016. Gazetouchpass: Multimodal authentication using gaze and touch on mobile devices. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 2156–2164.

Maurice Koch, Nelusa Pathmanathan, Daniel Weiskopf, and Kuno Kurzhals. 2024. How Deep Is Your Gaze? Leveraging Distance in Image-Based Gaze Analysis. In *Proceedings of the 2024 Symposium on Eye Tracking Research and Applications*. 1–7.

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

V Levenshtein. 1965. Leveinshtein distance. *Levenshtein_distance [accessedon 02/22/10]* (1965).

S Mathot, F Cristino, ID Gilchrist, and J Theeuwes. 2012. Eyenalysis: A similarity measure for eye movement patterns. *Journal of Eye Movement Research* 5 (2012), 1–15.

Laura E Matzen, Michael J Haass, Kristin M Divis, Zhiyuan Wang, and Andrew T Wilson. 2017. Data visualization saliency model: A tool for evaluating abstract data visualizations. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2017), 563–573.

Meinard Müller. 2007. Dynamic time warping. *Information retrieval for music and motion* (2007), 69–84.

Saul B Needleman and Christian D Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology* 48, 3 (1970), 443–453.

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

Vsevolod Peysakhovich and Christophe Hurter. 2018. Scanpath visualization and comparison using visual aggregation techniques. *Journal of Eye Movement Research* 10, 5 (2018), 10–16910.

Patrik Polatsek, Manuela Waldner, Ivan Viola, Peter Kapec, and Wanda Benesova. 2018. Exploring visual attention and saliency modeling for task-based visual analysis. *Computers & Graphics* 72 (2018), 26–38.

Aini Putkonen, Yue Jiang, Jingchun Zeng, Olli Tammilehto, Jussi PP Jokinen, and Antti Oulasvirta. 2025. Understanding visual search in graphical user interfaces. *International Journal of Human-Computer Studies* 199 (2025), 103483.

Danqing Shi, Yao Wang, Yunpeng Bai, Andreas Bulling, and Antti Oulasvirta. 2025. Chartist: Task-driven Eye Movement Control for Chart Reading. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–14.

Danqing Shi, Yujun Zhu, Jussi PP Jokinen, Aditya Acharya, Aini Putkonen, Shumin Zhai, and Antti Oulasvirta. 2024. CRTypist: Simulating touchscreen typing behavior via computational rationality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–17.

Wanjie Sun, Zhenzhong Chen, and Feng Wu. 2019. Visual Scanpath Prediction using IOR-ROI Recurrent Mixture Density Network. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 6 (2019), 2101–2118.

Floris Takens. 2006. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Warwick 1980: proceedings of a symposium held at the University of Warwick 1979/80*. Springer, 366–381.

Wei Wang, Cheng Chen, Yizhou Wang, Tingting Jiang, Fang Fang, and Yuan Yao. 2011. Simulating human saccadic scanpaths on natural images. In *CVPR 2011*. IEEE, 441–448.

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

Yao Wang, Yue Jiang, Zhiming Hu, Constantin Ruhdorfer, Mihai Bâce, and Andreas Bulling. 2024b. VisRecall++: Analysing and predicting visualisation recallability from gaze behaviour. *Proceedings of the ACM on Human–Computer Interaction* 8, ETRA, Article 339 (2024), 18 pages.

Yao Wang, Maurice Koch, Mihai Bâce, Daniel Weiskopf, and Andreas Bulling. 2022. Impact of Gaze Uncertainty on AOIs in Information Visualisations. In *2022 Symposium on Eye Tracking Research and Applications*. 1–6.

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