VisRecall++: Analysing and Predicting Visualisation Recallability from Gaze Behaviour

YAO WANG*, University of Stuttgart, Germany
YUE JIANG, Aalto University, Finland
ZHIMING HU, University of Stuttgart, Germany
CONSTANTIN RUHDORFER, University of Stuttgart, Germany
MIHAI BÂCE†, KU Leuven, Belgium
ANDREAS BULLING, University of Stuttgart, Germany

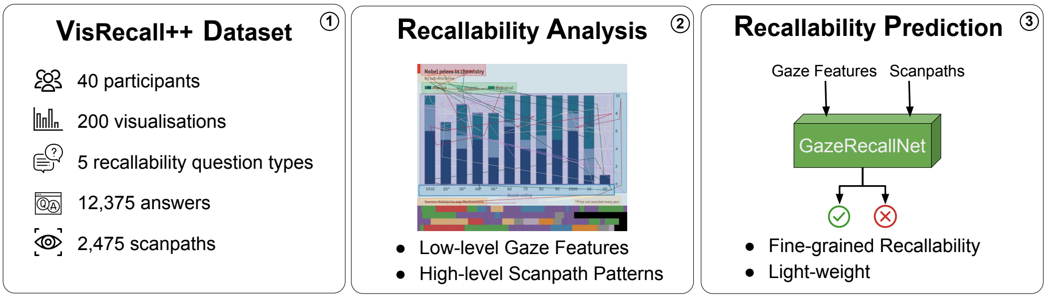

Fig. 1. ① We propose *VisRecall++*, a novel recallability dataset that contains gaze data from 40 participants on 200 information visualisations and five recallability question types. ② Our analyses on VisRecall++ show that low-level gaze features (saccade amplitude, the number of fixations, and fixation duration) significantly differ between high and low recallability groups. Moreover, we observe significant differences in high-level scanpath patterns, such as correct-answer scanpaths having significantly higher stationary entropy than wrong-answer scanpaths in every question type, and considerable variability in AOI transitions. ③ Inspired by our findings, we propose *GazeRecallNet*, a light-weight method to predict fine-grained recallability from three low-level gaze features and string-encoded scanpaths.

Question answering has recently been proposed as a promising means to assess the recallability of information visualisations. However, prior works are yet to study the link between visually encoding a visualisation in memory and recall performance. To fill this gap, we propose *VisRecall++* – a novel 40-participant recallability dataset that contains gaze data on 200 visualisations and 1,000 questions, including identifying the title and retrieving values. We measured recallability by asking participants questions after they observed the

*Corresponding author
†A significant part of this work was conducted while at the University of Stuttgart

Authors' addresses: Yao Wang, University of Stuttgart, Stuttgart, Germany, yao.wang@vis.uni-stuttgart.de; Yue Jiang, Aalto University, Espoo, Finland, yue.jiang@aalto.fi; Zhiming Hu, University of Stuttgart, Stuttgart, Germany, zhiming.hu@vis.uni-stuttgart.de; Constantin Ruhdorfer, University of Stuttgart, Stuttgart, Germany, constantin.ruhdorfer@vis.uni-stuttgart.de; Mihai Bâce, KU Leuven, Leuven, Belgium, mihai.bace@kuleuven.be; Andreas Bulling, University of Stuttgart, Stuttgart, Germany, andreas.bulling@vis.uni-stuttgart.de.

2573-0142/2024/5-ART239 $15.00
https://doi.org/10.1145/3655613

visualisation for 10 seconds. Our analyses reveal several insights, such as saccade amplitude, number of fixations, and fixation duration significantly differ between high and low recallability groups. Finally, we propose *GazeRecallNet* – a novel computational method to predict recallability from gaze behaviour that outperforms the state-of-the-art model RecallNet and three other baselines on this task. Taken together, our results shed light on assessing recallability from gaze behaviour and inform future work on recallability-based visualisation optimisation.

CCS Concepts: • **Human-centered computing → Information visualization**; **HCI theory, concepts and models**.

Additional Key Words and Phrases: Information visualisation, eye-tracking study, gaze behaviour, recallability, deep learning

ACM Reference Format:
Yao Wang, Yue Jiang, Zhiming Hu, Constantin Ruhdorfer, Mihai Bâce, and Andreas Bulling. 2024. VisRecall++: Analysing and Predicting Visualisation Recallability from Gaze Behaviour. *Proc. ACM Hum.-Comput. Interact.* 8, ETRA, Article 239 (May 2024), 18 pages. https://doi.org/10.1145/3655613

1 INTRODUCTION

Effective information visualisations convey information clearly to their target users [Fekete et al. 2008; Perer and Shneiderman 2008]. While this high-level goal is easy to formulate and clear, how to design visualisations that achieve this goal remains an open challenge [Bateman et al. 2010; Inbar et al. 2007]. Visualisation designers commonly rely on well-established guidelines [Tufte 1985] that recommend designing information visualisations with specific characteristics, such as a low visual density [Borkin et al. 2013] or a high data-ink ratio [Tufte et al. 1990]. However, all of these approaches focus on characteristics of the visualisation – they do not explicitly capture the users' perception when looking at a visualisation. For users, a key property that designers typically want to maximise is information recall, i.e. the challenge of making sure that users understand and remember key information (the "take home message") of a visualisation.

Despite its importance, few works have studied the recallability of information visualisations. Borkin et al. [2015] have used a qualitative score assigned to self-reported user descriptions by visualisation experts to quantify recallability. This approach is cumbersome and only provides a single score representing overall recallability. Wang et al. [2022a] have introduced a question-answering paradigm to assess both fine-grained and overall recallability by measuring the accuracy of answering five different types of questions about a visualisation. The five types of questions include identifying the theme, finding extreme values, filtering data, retrieving values, and understanding the structure or trend. While their paradigm and dataset allowed, for the first time, to understand how different visualisation characteristics impact users' recall performance, they did not analyse individual differences between users. It remains unclear why certain participants performed better in the recallability task than others.

We fill this gap by studying the link between visual encoding of a visualisation, captured using eye gaze data, and its impact on recall performance. In line with the recallability study conducted by Wang et al. [2022a], we first present *VisRecall++*, a novel dataset of gaze data collected from 40 participants to assess their recall performance on 200 visualisations and five question types. Complementing gaze, we provide rich semantic annotations of the visual elements of the visualisations, such as titles, axes, and labels. Using this dataset, we analysed the gaze behaviour on information visualisations during the encoding stage, i.e., when viewing visualisations for 10 seconds each and trying to memorise as much as possible without knowing the question in advance. We found that saccade amplitude, the number of fixations, and fixation duration significantly differ between high and low recallability groups. By analysing which visual elements attracted users' visual attention the most, we found that the stationary entropy [Krejtz et al. 2015] of scanpaths preceding a correct answer was significantly higher than those preceding a wrong one. These individual differences suggest that eye movements are directly linked to recall performance. We further propose *GazeRecallNet*, the first computational method to predict the ability of users to answer recallability questions on information visualisations only from their scanpaths and different low-level gaze features, such as saccade amplitudes, fixation duration, and the number of fixations. We show that GazeRecallNet outperforms the state-of-the-art models such as RecallNet, in terms of both overall recallability and fine-grained recallability.

As illustrated in Figure 1, the contributions of our work are three-fold:

(1) We introduce VisRecall++ , a novel recallability dataset that contains eye gaze data from 40 participants on 200 different information visualisations and five question types.
(2) We provide in-depth analyses on VisRecall++ that show how low and high-level gaze behaviour characteristics correlate with recall performance.
(3) We propose GazeRecallNet , the first computational method to predict recallability scores on information visualisations only from gaze behaviour.

2 RELATED WORK

Our work is related to previous works on 1) recallability of information visualisations, 2) gaze-based image analysis, and 3) gaze-based cognitive state estimation.

2.1 Recallability of Information Visualisations

Recallability of information visualisations has recently become popular in the areas of cognitive science and visualisation [Borkin et al. 2015; Kim et al. 2012, 2017; Kong et al. 2019; Wang et al. 2022a]. Previous cognitive science literature usually measures image (visualisation) memorability by how recognisable they are [Borkin et al. 2013], that is, the tendency of visualisations for people to remember or forget. However, only a few works studied the recallability of visualisations [Bainbridge 2019], which is usually measured by quantifying how much information an observer remembers from a visualisation [Rust and Mehrpour 2020; Wang et al. 2022a] and is not necessarily related to recognisability [Borkin et al. 2015; Wang et al. 2022a]. Keskin et al. [2023] analysed users' attention during visual encoding to investigate how the attention in the encoding stage is linked to the cued-recall performance on 2D web maps. Borkin et al. [2015] proposed quantifying recallability by asking visualisation experts to assign a qualitative score to self-reported free-text descriptions from the observers. However, this approach is cumbersome and only provides an ordinal score representing overall recallability while hiding the contribution of individual visualisation characteristics. More recently, Wang et al. [2022a] introduced a question-answering paradigm to assess the recallability of information visualisations. Their approach measures recallability as the accuracy of answering questions about visualisations. They also proposed a computational method, RecallNet, to predict recallability from visual properties of visualisations. While promising, their work neglected to study the encoding stage and its importance for recallability, i.e. how observers look at visual elements of visualisations and how this process of visual inspection links to recallability. We fill this gap by introducing a new recallability dataset that offers gaze data. This provides an opportunity to analyse and understand how visualisations are visually encoded, if and how their properties influence recallability, and to predict recallability from gaze behaviour.

2.2 Gaze-based Image Analysis

Eye tracking technology has received increasing attention from computer vision and cognitive science researchers and has become a powerful tool for image analysis and understanding. Pioneering works have studied how eye fixations are linked to memory for pictures [Christianson et al. 1991; Loftus 1972]. More recently, gaze stationary entropy [Krejtz et al. 2015] quantifies the randomness and complexity of a person's eye movements while observing artworks. Scanpaths capture the spatiotemporal attention in an image and have been widely used to analyse images [Jiang et al. 2015], videos [Hu et al. 2021], webpages [Drusch et al. 2014], mobile user interfaces [Jokinen et al. 2020], as well as 3D virtual environments [Hu 2020]. Gazealytics [Chen et al. 2023b] is an eye-tracking analytics tool that unifies spatiotemporal exploration of fixations and scanpaths for various analytical tasks. Scanpath scarf plots summarise scanpath dynamics between AOIs [Blascheck et al. 2014]. A body of work has used the visual toolkit for exploratory scanpath and comparative gaze metrics analysis [Chen et al. 2023a], interactive data annotations with AOIs and data analysis [Cai et al. 2022; Pozdniakov et al. 2023]. In the area of information visualisation, gaze-based AOI analysis has been used to understand how people explore visualisations or assess the quality of visualisations [Borkin et al. 2015; Burch et al. 2017; Polatsek et al. 2018; Wang et al. 2023, 2022b]. However, despite the potential of the human eye gaze for analysing visualisations, little attention has been paid to specifically analysing the link between eye gaze and recall performance of information visualisations. We fill this gap by recording human gaze data in the context of recalling visualisations, allowing us to link eye gaze, visualisation elements, and recallability.

2.3 Gaze-based Cognitive State Estimation

Numerous studies in eye tracking research and cognitive science have revealed that human eye movements can provide insights into human cognitive behaviour [Bulling and Roggen 2011; Bulling and Zander 2014], and this has inspired a growing number of researches in gaze-based cognitive state estimation [Hu et al. 2021; Pfleging et al. 2016; Wang et al. 2019]. Specifically, Pfleging et al. [2016] proposed to estimate users' cognitive load by measuring users' pupil diameters under various controlled lighting conditions. Sattar et al. [2020] predicted user search intents using human gaze fixations, while Lethaus et al. [Lethaus et al. 2013] inferred driver intent using eye gaze features. Strohm et al. [2021] introduced a method to reconstruct mental images from eye movements visually. David et al. [2019] predicted artificial visual field losses from eye gaze features using Hidden Markov Models and recurrent neural networks. Previous works have also estimated participants' levels of text comprehension [Ahn et al. 2020] and mind-wandering tendencies [Huang et al. 2019; Zermiani et al. 2022] from their eye movements. In addition, an increasing number of researchers have studied the correlations between human eye movements and tasks and proposed many successful gaze-based task recognition methods [Boisvert and Bruce 2016; Braunagel et al. 2017; Hild et al. 2018; Hu et al. 2021]. Complementing these prior works, we focus on the problem of predicting recallability from human eye movements.

3 VisRecall++ Dataset

To investigate the link between participants' gaze behaviours and their recall of content from information visualisations, we propose the VisRecall++ – a novel dataset that contains eye gaze data from 40 participants on 200 information visualisations for five recallability question types. Our dataset and code are publicly available at https://doi.org/10.18419/darus-3138.

3.1 Data Collection

Stimuli. We used the 200 information visualisations from the VisRecall dataset [Wang et al. 2022a] as stimuli aligning with the prior work. The selection covers a variety of frequently used information visualisations, including 56 bar plots, 45 line plots, 27 scatter plots, 22 pie plots, 25 tables, and 25 complex visualisations (e.g. box charts and isotype charts). Figure 2 shows a sample

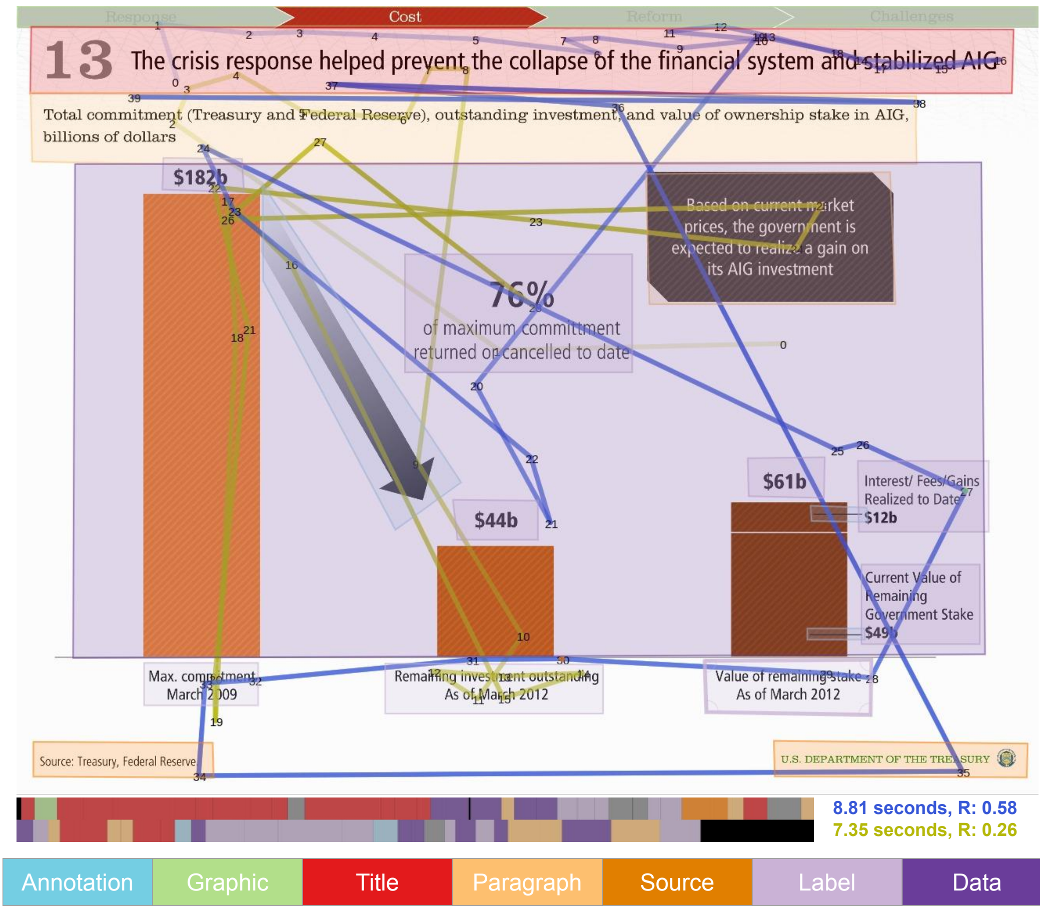

Fig. 2. Sample annotated AOIs and scanpaths from two participants overlaid on the same visualisation from VisRecall++ . The participant with the blue scanpath had a recallability score of 0.58 while the participant with the yellow scanpath had a recallability score of 0.26. For recallability, higher is better and a score of 1.0 indicates correctly answering all questions.

visualisation of VisRecall++ in a sample web application. We used all 1,000 recallability questions in five question types from VisRecall [Wang et al. 2022a] to collect gaze data.

The question types are:

- Identify the title or theme (**T-question**): T-questions require participants to identify the title or the general theme of the corresponding visualisation and are used to test participants' ability to recall the general story of visualisations [Borkin et al. 2015]. Examples: *What is the theme of the visualisation? What is the title of the visualisation?*
- Find extreme values (**FE-question**): FE-questions ask participants to find certain extreme values in the visualisation and are used to measure participants' low-level recall ability of the stimuli [Polatsek et al. 2018; Schulz et al. 2013]. Examples: *Which particle is the latest discovered? Which area had the lowest level of urbanisation in 1950?*
- Filter data elements (**F-question**): F-questions request participants to filter data elements based on some specific criteria and are used to calculate participants' ability to recall multiple

elements in the stimuli [Kim and Heer 2018; Polatsek et al. 2018]. Examples: *Which particle is Bosons? What is the source of the data?*

- Retrieve values (**RV-question**): RV-questions require participants to retrieve the value for a specific visual element and are utilised to evaluate participants' recall of detailed information in the visualisations [Kim and Heer 2018; Polatsek et al. 2018]. Examples: *What percentage of Indians are expected to live in urban areas by 2045? What is the maximum percentage of aid allocated?*
- Understand the structure or trend (**U-question**): U-questions ask participants to understand the structure or the trend of the visualisation [Schulz et al. 2013] and are used to quantify participants' high-level recall ability of the visualisation [Chaudhry et al. 2020; Methani et al. 2020]. Examples: *What decreases as time goes by? What does the purple curve represent?*

VisRecall++ includes 196 T-questions, 302 FE-questions, 276 F-questions, 125 RV-questions, and 101 U-questions. Each visualisation has five associated questions, containing at least two different question types. Each question has four possible answer options, and only one option is correct.

Apparatus. Gaze data for VisRecall++ was collected using an EyeLink 1000 Plus eye tracker running at 2,000 Hz in binocular mode, providing an accuracy of 0.5° under proper calibration [Ehinger et al. 2019]. Information visualisations were presented on a 24.5" monitor with a resolution of 1920×1080 pixels at 90 cm from the participant using a high-performance desktop computer. Visualisations were shown in the screen centre, covering a visual angle of around $21.1° \times 14.8°$. Participants used a desk-mounted chin rest to minimise the influence of head movements on gaze data quality. We used the JavaScript-based web application provided by the authors of VisRecall[1]. The web application runs in a browser in full-screen mode and is embedded in the WebLink recording software provided by the manufacturer[2].

Participants. We recruited 43 participants from the local university[3]. Three participants quit the experiment due to a self-reported lack of visual literacy. The final participants are 15 females and 25 males. No participants are colour-blind. All participants reported normal or corrected-to-normal vision and were aged between 19 and 32 years ($\mu = 24.7$, $\sigma = 2.2$), with an English level of C1 or better. They were compensated for their participation for $15 per hour and could stop without adverse consequences. All personal information was fully pseudonymised.

Experimental Design. The 200 visualisations were randomly divided into 10 trials based on the original split of VisRecall [Wang et al. 2022a], where each trial contains 20 visualisations. Each trial takes approximately 20 to 25 minutes to complete. Each participant randomly took 2 to 6 trials. We ensured that each visualisation was observed by at least 12 participants ($\mu = 15.8$, $\sigma = 1.08$). We limited participants to a maximum of three trials in one day.

Procedure. Prior to the study, we provided participants with a comprehensive explanation of the various question types. We used the question-answering paradigm proposed in [Wang et al. 2022a] to quantify participants' recallability of information visualisations, following the two-by-two setting [Wang et al. 2022a] for demonstrating visualisations to reduce the effect of working memory [Owen et al. 2005]. Specifically, in the encoding stage, we showed two information visualisations sequentially to the participants, each for 10 seconds. The observation duration aligns with prior work [Borkin et al. 2013; Wang et al. 2022a]. We then sequentially presented the questions regarding the two visualisations in the recall stage, i.e. five questions for the first

[1] https://doi.org/10.18419/darus-2826
[2] https://www.sr-research.com/weblink/
[3] The university ethics committee approved our study prior to data collection.

visualisation and then five questions for the second visualisation. The questions were displayed sequentially with the blurred visualisation next to the question in the recall stage. After answering one question, participants had to click a button to proceed, and they could not return to previous questions. During the experiments, participants' eye gaze data and their answers to the questions were recorded for further analysis.

Table 1. The number and percentage of correct and incorrect answers for each question type (overall, T-, FE-, F-, RV-, and U-questions).

Groups	Overall	T	FE	F	RV	U
Correct	5,574 (45%)	1,593	1,644	1,273	494	569
Incorrect	6,801 (55%)	847	2,146	2,097	1,014	678
Total	12,375	2,440 (20%)	3,790 (31%)	3,370 (27%)	1,508 (12%)	1,247 (10%)

3.2 Data Processing

Area of Interest (AOI) Annotation. To analyse the correlations between recallability scores and the visual elements inspected by the participants, we recruited three scientific researchers with more than three years of experience in information visualisation. They first independently annotated the areas of interest (AOIs) for all 200 visualisations and were then asked to discuss and reach a consensus on which areas represent AOIs. We divided all elements containing texts [Borkin et al. 2015] into Labels, Titles, Paragraphs, and Sources. We used a single area in one visualisation to annotate Data, such as including all bars in a bar graph or all points in a scatterplot. The hierarchical order of bounding boxes from high to low is defined as Annotations (annotated visual elements), Axes (axis location, including tick marks and numeric values), Graphical Elements (non-data-related visual elements), Legends (data visual encoding explanations), Objects (human recognisable objects), Titles, Paragraphs, Sources, Labels, Data. In total, we collected 996 Labels, 338 Data, 283 Axes, 199 Titles, 180 Paragraphs, 156 Sources, 105 Legends, 104 Annotations, 92 Graphical Elements, and 36 Objects [Borkin et al. 2015]. See Figure 2 and Figure 4 for the annotated AOIs overlaid on sample visualisations from VisRecall++ , and supplementary material for more examples.

Gaze Data Processing. We detected eye fixations using the Identification by Dispersion-Threshold (IDT) algorithm [Salvucci and Goldberg 2000] in the EyeLink software with velocity and acceleration thresholds of $30°/s$ and $8000°/s^2$, respectively. Since each visualisation was shown for 10 seconds, we identified those scanpaths with a total fixation time (duration) shorter than 2 seconds as outliers and discarded those scanpaths. We further calculated the Hit-Any-AOI Rate (HAAR) [Wang et al. 2022b] to check the quality of the scanpaths, and removed the scanpaths whose HAAR was less than 0.5. This resulted in a mean HAAR for our dataset of 0.827, which indicates a good quality of gaze data [Wang et al. 2022b]. See supplementary material for further details on the gaze data processing procedure.

4 DATA ANALYSIS

Since several studies have demonstrated that human eye movements can provide insights into human cognitive behaviour [Bulling and Roggen 2011; Bulling and Zander 2014], we analysed the link between human gaze behaviour and recallability of information visualisations. Specifically, we analysed the characteristics of three low-level gaze features (saccade amplitudes [Baloh et al. 1975], number of fixations [Loftus and Mackworth 1978], and fixation duration [Baloh et al. 1975; Loftus and Mackworth 1978]) and scanpath patterns on AOIs.

4.1 Recallability Group Separation

To analyse how observers visually process the recallability questions, we first split the participants' gaze data according to the question type and correctness of the answer. As each answer corresponds to a scanpath, we created subsets of scanpaths preceding correct answers and wrong answers from participants, denoted as correct-answer scanpaths and wrong-answer scanpaths. Since each trial presented different images and questions, we further divided participants into a high and a low recallability group based on the mean recallability of each trial. The separation of high recallability and low recallability groups is for understanding how observers' gaze features are correlated with recallability in subsequent sections. See supplementary material for additional statistics on these two participant groups.

4.2 Dataset Statistics

VisRecall++ contains 2,475 valid scanpaths with 12,375 answers (each scanpath corresponds to exactly 5 answers) from 40 participants. The scanpaths in VisRecall++ have a mean recording duration of 7.17 seconds ($\sigma = 1.67$ s). The scanpaths have a mean length of 32.37 fixations ($\sigma = 8.29$ fixations) with a mean fixation duration of 222 milliseconds ($\sigma = 132$ ms), and a mean saccade amplitude of 3.47° ($\sigma = 3.63°$). Each question, offering four options, presents a baseline accuracy of 25% by random chance alone. As shown in Table 1, Despite an overall correctness rate of 45.0%, 20% surpassing random chance, most questions were answered incorrectly. Questions involving general or extreme information, such as theme identification (T-question), were answered correctly at a rate of 65.3%, contrasting with lower rates for detailed tasks like data element filtering (F-question) at 37.8%. This trend underscores the ease of perceiving general or extreme information compared to more intricate details, as indicated by varying correctness rates across question types.

4.3 Low-level Gaze Features

To understand how eye gaze events (fixations and saccades) are linked with recallability, we analysed three low-level gaze features:

Number of fixations. We first compared the number of fixations in scanpaths between the high and low recallability groups. The mean number of fixations in the high recallability group was 33.09 ($\sigma = 8.34$) and 31.45 ($\sigma = 8.15$) in the low recallability group. This difference was statistically significant in a Student's T-test as $t(2{,}570) = 5.014$, $p < 0.001$.

Fixation Duration. Fixations in the correct-answer scanpaths had a mean duration of 219.51ms ($\sigma = 129.60$ ms) and the low recallability group has a mean fixation duration of 226.29 ms ($\sigma = 134.19$ ms). Statistical significance was found as $t(83{,}262) = 7.352$, $p < 0.001$.

Saccade Amplitude. Finally, we calculated the saccade amplitudes for each group as the Euclidean distances between subsequent fixations in degrees of visual angle. The mean saccade amplitude in the high recallability group was 3.53° ($\sigma = 3.69°$), and 3.40° ($\sigma = 3.56°$) in the low recallability group. Statistical significance was found as $t(80{,}690) = 4.853$, $p < 0.001$.

4.4 Scanpath Patterns

Visual elements (AOIs)-based visual analysis is a widely used approach in information visualisation research to analyse scanpath patterns [Borkin et al. 2015; Polatsek et al. 2018; Wang et al. 2022b]. We performed two analyses to understand the semantics of scanpaths, i.e. how the viewing behaviour on AOIs is linked to recallability. Stationary entropy [Krejtz et al. 2015] focuses on attention distribution among AOIs while scanpath scarf plot [Stellmach et al. 2010] illustrates qualitative gaze transitions.

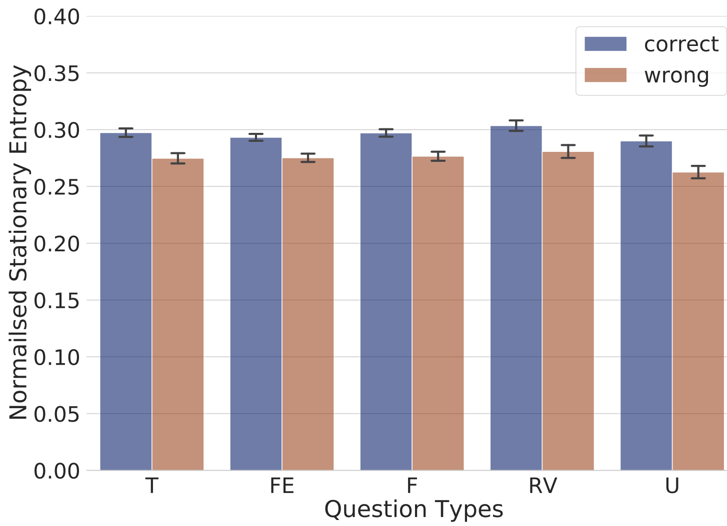

Fig. 3. The normalised mean gaze stationary entropy [Krejtz et al. 2015] of correct-answer and wrong-answer scanpaths in every recallability question type (T-, FE-, F-, RV-, and U-questions). Error bars indicate the standard error. The stationary entropy of correct-answer scanpaths in all question types is significantly lower than wrong-answer scanpaths.

Stationary Entropy. Gaze stationary entropy [Krejtz et al. 2015] is a metric to quantify how equally attention is distributed among AOIs. The normalised stationary entropy ranges from 0 to 1, and a higher value means that the subject distributes their visual attention more equally among AOIs. We analysed the gaze stationary entropy of the correct-answer and wrong-answer scanpaths. The stationary entropy of correct-answer scanpaths in all question types is significantly lower than wrong-answer scanpaths (see Figure 3): $t(2{,}492) = 3.866$, $p < 0.001$ for T-questions, $t(3{,}851) = 3.813$, $p < 0.001$ for FE-questions, $t(3{,}443) = 4.000$, $p < 0.001$ for F-questions, $t(1{,}536) = 3.138$, $p < 0.001$ for RV-questions, $t(1{,}313) = 3.779$, $p < 0.001$ for U-questions, respectively.

Scanpath Scarf Plot. After assigning a unique colour to each type of AOI, the scanpaths can be visualised as scarf plots. The lengths represent the sum of fixation durations, and colour changes represent attention shifts between AOIs. Figure 4 showcases two examples from VisRecall++ , each divided into groups of high and low recallability. The visualisations include fixation contours in the form of Bell Curves, scanpath scarf plots, and tables displaying the percentage fixation duration. The scanpaths with high recallability have a shorter percentage dwell time on Data (D), and a longer percentage dwell time on Axes (X) and Legends (L) for all two examples. Additionally, the high recallability group usually has a longer scarf, indicating a longer total fixation time, which agrees with our finding in subsection 4.3. Moreover, the high recallability group generally has a lower percentage dwell time on Data (D), and a longer percentage dwell time on Axes (X) and Legends (L). This visualisation was created using Gazealytics[4] [Chen et al. 2023b]. See supplementary material for more examples.

[4]https://github.com/gazealytics/gazealytics-master

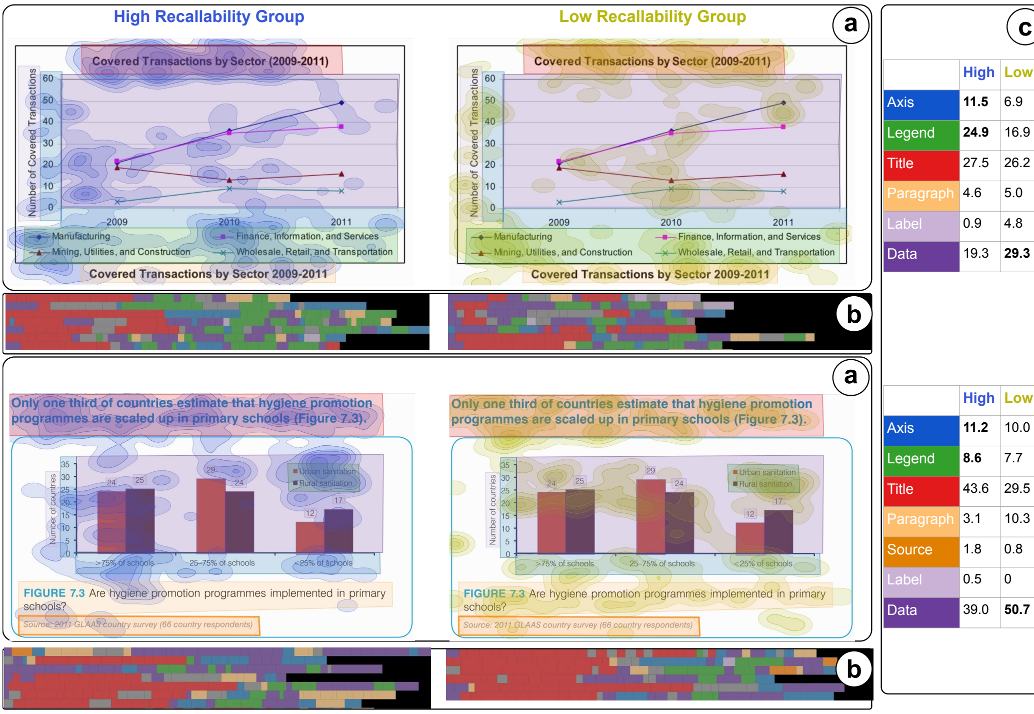

Fig. 4. Two examples from VisRecall++ , each divided into groups of high and low recallability group. The visualisations included ⓐ fixation contours (Bell Curve), ⓑ scanpath scarf plots, and ⓒ tables displaying percentage fixation duration.

5 GazeRecallNet

The findings in Section 4 demonstrate a link between human gaze behaviour and recallability of information visualisations, raising the question of whether recallability of information visualisation can be predicted from gaze behaviour. To this end, we propose GazeRecallNet – a predictive model designed for fine-grained recallability prediction on five question types. Figure 5 shows an overview of our GazeRecallNet model. Given the three low-level eye gaze features of the scanpath (the number of fixations, saccade amplitudes, and fixation durations) and the string-encoded scanpaths, i.e., the string representing the sequence of AOI annotations as described in subsection 3.2, the model predicts whether a given scanpath can lead to a correct answer to recallability questions.

5.1 Gaze Encoding

We encode gaze features and string-encoded scanpaths into embedding vectors and concatenate them to form a single gaze embedding vector. It is then fed into a network to predict the accuracy of responses to the recallability questions.

Gaze Features. The lengthiest scanpath in our dataset includes less than 80 fixations. Thus, we encode employ a trainable parametric matrix of size 80 $\times$ 64 to encode the number of fixations. The matrix maps the number of fixations to a 64-dimensional vector. To represent the sequence of saccade amplitudes and fixation durations, we use gated recurrent unit networks (GRUs) [Cho et al. 2014] to encode them, resulting in 64-dimensional vectors as their embeddings respectively. We

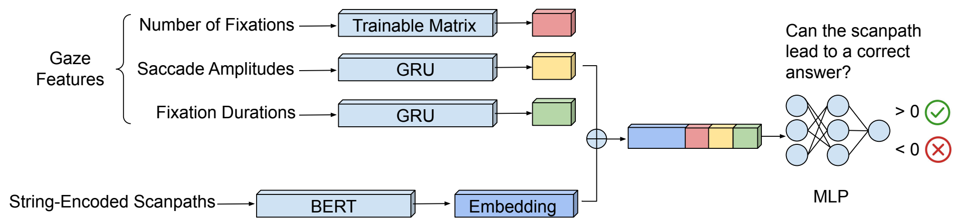

Fig. 5. Overview of GazeRecallNet . Three low-level gaze features (the number of fixations, saccade amplitudes, and fixation durations) and string-encoded scanpaths are encoded in parallel. All the gaze feature embeddings are concatenated to train a classifier to predict whether an observer can correctly answer a recallability question.

chose GRUs since they are empirically superior to process sequential and temporal data, which is a common choice in encoding gaze features [Palmero et al. 2018; Park et al. 2020].

String-encoded Scanpaths. Our analysis in Figure 3 demonstrates that the specific scanpath patterns on visual elements during the encoding stage correlated with the recallability scores in the recall stage. Therefore, we assigned each fixation a character to represent the AOI it landed on, resulting in string-encoded scanpaths [Bulling and Roggen 2011; Wang et al. 2022b]. For instance, "D" denotes data, "L" represents labels, and "T" signifies titles. Consecutive fixations on the same type of AOIs are counted only once. The number of fixations in the resulting scanpaths ranges from 2 to 25. To represent these string-encoded scanpaths, we use the pre-trained bidirectional encoder representations from transformers (BERT) [Devlin et al. 2019] to generate a 768-dimensional embedding vector. BERT has been widely applied in language understanding tasks ranging from textual classification [Sun et al. 2019] to reading comprehension [Xu et al. 2019] and can generate embeddings for scanpath strings of an arbitrary length.

5.2 Recallability Prediction

We concatenated all the generated gaze features and scanpath embedding to form the gaze embedding, which is fed into a three-layer perceptron (MLP) model for training a classifier for predicting the accuracy of responses to recallability questions. MLPs are known for their simplicity and effectiveness in regression tasks, particularly when handling gaze data [Jiao et al. 2023; Zhang et al. 2017]. We apply the binary cross entropy loss (BCE Loss) during training to classify recallability questions as answerable or unanswerable — where a positive output means the answer was predicted to be correct, otherwise wrong.

6 EXPERIMENT RESULTS

We conducted a series of experiments to compare the performance of GazeRecallNet with recallability prediction methods on VisRecall++ . Different ablated versions of the GazeRecallNet were also evaluated.

6.1 Training Details

As described in subsection 4.4, we discarded all fixations that did not land on any AOIs and removed repeating characters in the string-encoded scanpaths. To train our GazeRecallNet to predict fine-grained recallability scores for a certain question type, we only used those scanpaths that have proceeded at least one question for that question type. There are 2440, 3790, 3370, 1508, and 1247 scanpaths for T-, FE-, F-, RV-, and U-question, respectively (see Table 1). We did a five-fold cross-validation for a training and testing set of VisRecall++ in every question type across visualisations. In each fold, we did cross-participant separation; that is, all data from a single participant are in either training or testing set. GazeRecallNet was trained for 50 epochs with the Adam [Kingma and Ba 2015] optimiser with a learning rate of $1E-4$ [Fosco et al. 2020]. All experiments were conducted on a single NVIDIA GeForce RTX 2060 Super GPU with 8GB VRAM.

6.2 Baseline Methods and Evaluation Metrics

Baseline Methods. The only method that predicts visualisation recallability is the RecallNet [Wang et al. 2022a]. Besides, we created three simple but effective baselines: CoordLSTM, a *Mean*, and a *Random* predictor. The descriptions of all baselines are as follows:

- RecallNet [Wang et al. 2022a] aims at predicting recall accuracy using a visualisation as input and predicts one recallability score independent of the user.
- We designed CoordLSTM as a one-layer Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber 1997] model with 16 hidden neurons that predicts the recallability score with scanpath coordinates as input. We opted to include it here given the recent success of LSTMs for a range of sequential modelling tasks, such as scanpath prediction [Chen et al. 2021] or encoding video frames [Xu et al. 2018].
- The *Mean* predictor calculates the mean recallability score of every question type in the training set, then uses this value as the probability of predicting the answer to be correct.
- The *Random* predictor predicts randomly whether an observer will answer a question correctly or incorrectly.

Evaluation Metrics. We compute the accuracy of recallability questions by $\frac{TP+TN}{TP+TN+FP+FN}$, where TP and TN represent the right predictions of correct and wrong answers, respectively, while FP and FN represent wrong predictions of correct and wrong answers, respectively. The overall accuracy was computed as the average accuracy of all question types, weighted by the distribution of the number of test samples in each question type.

6.3 Model Evaluation

Table 2 shows the overall and fine-grained recallability accuracy of GazeRecallNet and four baselines: RecallNet [Wang et al. 2022a], CoordLSTM, a *Mean*, and a *Random* predictor. GazeRecallNet outperformed all four baselines for recallability prediction on every question type. Our model reached 63.0% prediction accuracy in terms of the overall recallability prediction compared to the baselines (46.1% for RecallNet, 60.8% for CoordLSTM, 53.2% for the *Mean*, 50.5% for the *Random* predictor), and reached state-of-the-art performance for every fine-grained recallability prediction task (T-, FE, F, RV-, and U-questions). Moreover, GazeRecallNet has only 236,641 trainable parameters, compared to 25,592,362 trainable parameters for RecallNet.

6.4 Ablation Study

We further carried out an ablation study to investigate how each branch in GazeRecallNet contributes to overall and fine-grained recallability (see Table 3). We first evaluated the model by removing the string-encoded scanpaths (the second row) and all the low-level gaze features, i.e. the number of fixations, fixation durations, and saccade amplitudes (the third row). Even when AOI information is unavailable (w/o scanpaths), our method still achieves close-to-top performance in terms of overall accuracy (61.7% vs 63.0%). To evaluate the importance of each gaze feature, we removed each feature from the training process (the last three rows). Removing any gaze feature reduced

Table 2. Accuracy of fine-grained recallability on VisRecall++ under five-fold cross-validation evaluation across visualisations, reported in mean and standard deviation (%). The best results of each recallability question type are shown in **bold**.

Methods	Overall	T	FE	F	RV	U
GazeRecallNet (ours)	**63.0 (2.0)**	**66.4 (5.0)**	**58.7 (2.2)**	**62.3 (2.3)**	**67.3 (2.0)**	**58.8 (3.1)**
RecallNet	46.1 (1.9)	64.6 (6.2)	43.8 (3.2)	38.8 (1.5)	32.8 (2.1)	51.9 (1.3)
CoordLSTM	60.8 (1.9)	65.8 (5.3)	54.7 (3.1)	59.7 (1.6)	64.9 (3.7)	54.3 (4.8)
Mean	53.2 (1.0)	55.2 (1.5)	51.5 (1.7)	52.7 (2.1)	57.1 (2.9)	51.8 (3.1)
Random	50.5 (0.8)	50.9 (1.4)	50.4 (1.7)	49.5 (1.3)	51.8 (2.2)	51.0 (3.8)

Table 3. GazeRecallNet ablation study, reported in mean and standard deviation of recallability accuracy (%). The three gaze features are denoted as NF: number of fixations, FD: fixation duration, and SA: saccade amplitudes.

Methods	Overall	T	FE	F	RV	U
Full Model	**63.0 (2.0)**	**66.4 (5.0)**	**58.7 (2.2)**	**62.3 (2.3)**	**67.3 (2.0)**	**58.8 (3.1)**
w/o scanpaths	61.7 (2.4)	65.8 (5.3)	54.9 (2.6)	62.2 (2.3)	67.2 (2.1)	55.1 (2.7)
w/o NF, FD, SA	62.0 (2.2)	65.8 (5.3)	56.3 (3.1)	62.2 (2.3)	67.2 (2.1)	55.0 (3.5)
w/o NF	62.3 (2.1)	65.8 (5.3)	57.4 (2.5)	62.2 (2.3)	67.2 (2.1)	56.2 (3.4)
w/o FD	61.9 (2.0)	64.7 (3.8)	56.8 (2.3)	62.1 (2.4)	67.2 (2.1)	57.1 (2.0)
w/o SA	62.0 (2.4)	65.8 (5.3)	57.1 (2.5)	62.0 (2.0)	67.2 (2.1)	56.2 (2.6)

the overall recallability prediction accuracy for the number of fixations to 62.3%, fixation durations to 61.9%, and saccade amplitudes to 62.0%. Results demonstrate that all gaze features contribute to the full model.

In a nutshell, this section demonstrates the superiority of GazeRecallNet over four baseline methods in predicting recallability scores across various question types on VisRecall++ . The results highlight the robustness and effectiveness of GazeRecallNet in predicting fine-grained recallability.

7 DISCUSSION

Understanding the link between visually encoding a visualisation and the ability to recall details from memory afterward is essential and lays the foundation not only for understanding human behaviours, i.e. whether certain viewing behaviour is an "optimal" strategy for remembering better, but also for potentially optimising visualisations for increased recallability. Toward this goal, our work proposed several original contributions.

7.1 Recallability Dataset with Gaze Data

Given that there was no suitable dataset to study the link between recallability and gaze features in the encoding stage, we proposed VisRecall++ – a novel dataset that contains 2,475 scanpaths from 40 participants in 5 recallability question types. Using VisRecall++ , we identified several findings that link gaze features in the encoding stage of visualisation in memory to correct or incorrect recall afterwards. As noted in subsection 4.3, there were statistically significant differences between the high and low recallability groups regarding the three low-level gaze features: number of fixations, fixation duration, and saccade amplitude. When only using these three gaze features as input to the model, the overall accuracy of our method only decreases from 63.0% to 61.7% (*w/o scanpaths*, see Table 3). This finding underlines the strong link between these low-level gaze features and recallability. When analysing high-level scanpath patterns, we also found a significant difference in stationary entropy between the different question types (see Figure 3). This suggests that the way users explored the visualisations significantly differed between high and low recallability groups: The scarf plots shown in Figure 4 qualitatively illustrate that the high recallability group distributed their visual attention more equally among AOIs. In contrast, the low recallability group focused more on Data and Titles. This indicates that some specific gaze patterns may be beneficial for recall performance.

Taken together, these differences point towards differences in encoding "strategy", i.e. how humans encode information in memory, and may lead to applications that teach users how to improve their encoding ability and, consequently, recallability. Our VisRecall++ enables future work to link top-down recallability with bottom-up visual saliency of the information visualisations. Given the detailed AOI annotations and the corresponding gaze data that VisRecall++ provides, future work could investigate whether and how saliency contributes to visual encoding abilities.

7.2 Predicting Recallability from Gaze

Building on our analyses (subsection 4.3, subsection 4.4) that demonstrated a strong link between gaze features and recall performance, we proposed GazeRecallNet – a computational method for gaze-based recallability prediction, that is, the task of predicting whether a question will be answered correctly or not only from gaze features and scanpaths. While earlier work [Wang et al. 2022a] has relied only on *image features*, thus ignoring differences in user behaviour, our GazeRecallNet leverages gaze features (scanpath length, saccade amplitude, and fixation duration) and scanpaths that encode the semantic meaning of different visualisation elements. Our experiments showed that our method achieved state-of-the-art performance on fine-grained recallability prediction (see Table 2). RecallNet [Wang et al. 2022a] was en par with our model only for predicting T-question recallability. For all other question types, performance was below even the naïve baselines. We also compared GazeRecallNet with a model that instead used the scanpath coordinates as input (CoordLSTM). Results showed that our approach was still the best-performing one, highlighting the importance of AOI and semantic scanpaths rather than absolute fixation locations.

Furthermore, our ablation study demonstrated that removing any component from our combination of gaze features reduces the prediction accuracy in both overall and fine-grained recallability. This underlines the importance of combining scanpaths and low-level gaze features to achieve high performance in recallability tasks: While *string-encoded scanpaths capture content-based semantics*, focusing on transitions across AOI types, *low-level gaze features capture individual details of eye movement behaviour*. Finally, the previous RecallNet [Wang et al. 2022a] used a pre-trained image encoder for classification. In stark contrast, GazeRecallNet does not require encoding of image features, thus resulting in a light-weight model with only 135,233 trainable parameters vs 25,592,362 for RecallNet.

7.3 Limitations

The ability to accurately recall information from memory in our study may not only have been influenced by the visualisations or gaze patterns but also by other characteristics, such as participants' personal experience and working memory capacity [Unsworth et al. 2010]. To reduce such influences, our work specifically focused on *short-term* recallability, and each trial involved encoding two visualisations before assessing the users' recall using multiple-choice questions. This study design followed the prior work [Wang et al. 2022a] that showed that two visualisations in the encoding stage were appropriate for the question-answering (QA) paradigm. Moreover, the dataset is not classified according to its recallability before the experiment. Therefore, the difficulties across experimental trials varied and might be a confound to the visualisation recallability. The participants' English fluency is another confound. The study required participants to be proficient in English and, as such, there could have been differences between native vs non-native speakers. We addressed this by only recruiting participants who reported at least an English level of C1 in the Common European Framework of Reference for Languages. Still, more English native speakers in future studies would likely further reduce this influence. Lastly, our dataset has an uneven distribution of participants, including 15 females and 25 males. Future research could explore how gender bias, such as potential differences in memory retention between females and males, may impact the model's generalisability.

7.4 Privacy and Ethics Statement

The ethical approval of this study was obtained from the University's Ethics Committee. Data was collected with pseudonymisation of personal data and secure encryption of data storage systems. Access to the data is restricted to the research team and is used exclusively for this study. Plans for data sharing are designed to respect consent terms and privacy standards, ensuring any future use aligns with the same ethical approval.

8 CONCLUSION

In this work, we introduce VisRecall++ – a novel recallability dataset that contains gaze data from 40 participants on 200 visualisations and five question types. Our analyses show statistically significant differences between high and low recallability groups regarding low-level and high-level gaze features. Inspired by our findings, we then propose GazeRecallNet, a novel method to predict recallability from scanpaths and gaze features. Extensive experiments on VisRecall++ show that our method outperforms several baselines in overall and fine-grained recallability prediction. As such, our work sheds light on assessing recallability from gaze behaviour and informs future work on enhancing recallability through the optimisation of information visualisations.

ACKNOWLEDGMENTS

Y. Wang was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161. Y. Jiang was supported by the Research Council of Finland (Subjective Functions, grant 357578). Z. Hu was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2075 – 390740016. M. Bâce was funded by the Swiss National Science Foundation (SNSF) through a Postdoc.Mobility Fellowship (grant number 214434) while at the University of Stuttgart. C. Ruhdorfer and A. Bulling were funded by the European Research Council (ERC) under grant agreement 801708. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting C. Ruhdorfer.

REFERENCES

Seoyoung Ahn, Conor Kelton, Aruna Balasubramanian, and Greg Zelinsky. 2020. Towards Predicting Reading Comprehension From Gaze Behavior. In *Proceedings of the ACM International Symposium on Eye Tracking Research and Applications (ETRA)*. 1–5.

Wilma A Bainbridge. 2019. Memorability: How what we see influences what we remember. In *Psychology of learning and motivation*. Vol. 70. Elsevier, 1–27.

Robert W Baloh, Andrew W Sills, Warren E Kumley, and Vicente Honrubia. 1975. Quantitative measurement of saccade amplitude, duration, and velocity. *Neurology* 25, 11 (1975), 1065–1065.

Scott Bateman, Regan L Mandryk, Carl Gutwin, Aaron Genest, David McDine, and Christopher Brooks. 2010. Useful junk? The effects of visual embellishment on comprehension and memorability of charts. In *Proceedings of the 2010 CHI Conference on Human Factors in Computing Systems (CHI)*. 2573–2582.

Tanja Blascheck, Kuno Kurzhals, Michael Raschke, Michael Burch, Daniel Weiskopf, and Thomas Ertl. 2014. State-of-the-art of visualization for eye tracking data. In *EuroVis (STARs)*. 1–20.

Jonathan FG Boisvert and Neil DB Bruce. 2016. Predicting task from eye movements: On the importance of spatial distribution, dynamics, and image features. *Neurocomputing* 207 (2016), 653–668.

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 (TVCG)* 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 (TVCG)* 19, 12 (2013), 2306–2315.

Christian Braunagel, David Geisler, Wolfgang Rosenstiel, and Enkelejda Kasneci. 2017. Online Recognition of Driver-Activity Based on Visual Scanpath Classification. *IEEE Intelligent Transportation Systems Magazine (ITS)* 9 (2017), 23–36.

Andreas Bulling and Daniel Roggen. 2011. Recognition of visual memory recall processes using eye movement analysis. In *Proceedings of the 13th International Conference on Ubiquitous Computing (UbiComp)*. 455–464.

Andreas Bulling and Thorsten O Zander. 2014. Cognition-aware computing. *Proceedings of the International Conference on Pervasive Computing (Pervasive)* 13, 3 (2014), 80–83.

Michael Burch, Lewis Chuang, Brian Fisher, Albrecht Schmidt, and Daniel Weiskopf. 2017. *Eye tracking and visualization: Foundations, Techniques, and applications. ETVIS 2015.*

Minghao Cai, Bin Zheng, and Carrie Demmans Epp. 2022. Towards Supporting Adaptive Training of Injection Procedures: Detecting Differences in the Visual Attention of Nursing Students and Experts. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. 286–294.

Ritwick Chaudhry, Sumit Shekhar, Utkarsh Gupta, Pranav Maneriker, Prann Bansal, and Ajay Joshi. 2020. LEAF-QA: Locate, encode & attend for figure question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 3512–3521.

Kun-Ting Chen, Quynh Quang Ngo, Kuno Kurzhals, Kim Marriott, Tim Dwyer, Michael Sedlmair, and Daniel Weiskopf. 2023a. Reading Strategies for Graph Visualizations that Wrap Around in Torus Topology. In *Proceedings of the ACM International Symposium on Eye Tracking Research and Applications (ETRA)*. 1–7.

Kun-Ting Chen, Arnaud Prouzeau, Joshua Langmead, Ryan T Whitelock-Jones, Lee Lawrence, Tim Dwyer, Christophe Hurter, Daniel Weiskopf, and Sarah Goodwin. 2023b. Gazealytics: A Unified and Flexible Visual Toolkit for Exploratory and Comparative Gaze Analysis. In *Proceedings of the ACM International Symposium on Eye Tracking Research and Applications (ETRA)*. 1–7.

Xianyu Chen, Ming Jiang, and Qi Zhao. 2021. Predicting Human Scanpaths in Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10871–10880.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1–11.

Sven-Åke Christianson, Elizabeth F Loftus, Hunter Hoffman, and Geoffrey R Loftus. 1991. Eye fixations and memory for emotional events. *Journal of Experimental Psychology: Learning, memory, and cognition* 17, 4 (1991), 693.

Erwan Joël David, Pierre Lebranchu, Matthieu Perreira Da Silva, and Patrick Le Callet. 2019. Predicting artificial visual field losses: a gaze-based inference study. *Journal of Vision (JOV)* 19, 14 (2019), 22–22.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (NAACL)*. 4171–4186.

Gautier Drusch, JC Bastien, and Stéfane Paris. 2014. Analysing eye-tracking data: From scanpaths and heatmaps to the dynamic visualisation of areas of interest. *Advances in Science, Technology, Higher Education and Society in the Conceptual Age (STHESCA)* 20, 205 (2014), 25.

Benedikt V Ehinger, Katharina Groß, Inga Ibs, and Peter König. 2019. A new comprehensive eye-tracking test battery concurrently evaluating the Pupil Labs glasses and the EyeLink 1000. *PeerJ* 7 (2019), e7086.

Jean-Daniel Fekete, Jarke J Van Wijk, John T Stasko, and Chris North. 2008. The value of information visualization. *Information Visualization: Human-Centered Issues and Perspectives* (2008), 1–18.

Camilo Luciano Fosco, Anelise Newman, Patr 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)*. IEEE, 4472–4481.

Jutta Hild, Michael Voit, Christian Kühnle, and Jürgen Beyerer. 2018. Predicting observer's task from eye movement patterns during motion image analysis. In *Proceedings of the ACM International Symposium on Eye Tracking Research and Applications (ETRA)*. 1–5.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

Zhiming Hu. 2020. Gaze analysis and prediction in virtual reality. In *Proceedings of the 2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VR)*. 543–544.

Zhiming Hu, Andreas Bulling, Sheng Li, and Guoping Wang. 2021. EHTask: Recognizing User Tasks from Eye and Head Movements in Immersive Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 29, 4 (2021), 1992–2004.

Michael Xuelin Huang, Jiajia Li, Grace Ngai, Hong Va Leong, and Andreas Bulling. 2019. Moment-to-Moment Detection of Internal Thought during Video Viewing from Eye Vergence Behavior. In *Proceedings of ACM Multimedia (MM)*. 1–9.

Ohad Inbar, Noam Tractinsky, and Joachim Meyer. 2007. Minimalism in information visualization: attitudes towards maximizing the data-ink ratio. In *Proceedings of the 14th European Conference on Cognitive Ergonomics: Invent! Explore! (ECCE)*. 185–188.

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

Chuhan Jiao, Zhiming Hu, Mihai Bâce, and Andreas Bulling. 2023. SUPREYES: SUPer Resolutin for EYES Using Implicit Neural Representation Learning. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–13.

Jussi P.P. 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.

Merve Keskin, Vassilios Krassanakis, and Arzu Çöltekin. 2023. Visual Attention and Recognition Differences Based on Expertise in a Map Reading and Memorability Study. *ISPRS International Journal of Geo-Information* 12, 1 (2023), 21.

Sung-Hee Kim, Zhihua Dong, Hanjun Xian, Benjavan Upatising, and Ji Soo Yi. 2012. Does an eye tracker tell the truth about visualizations?: Findings while investigating visualizations for decision making. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 18, 12 (2012), 2421–2430.

Younghoon Kim and Jeffrey Heer. 2018. Assessing effects of task and data distribution on the effectiveness of visual encodings. In *Computer Graphics Forum*, Vol. 37. Wiley Online Library, 157–167. Issue 3.

Yea-Seul Kim, Katharina Reinecke, and Jessica Hullman. 2017. Explaining the gap: Visualizing one's predictions improves recall and comprehension of data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI)*. 1375–1386.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. 1–11.

Ha-Kyung Kong, Zhicheng Liu, and Karrie Karahalios. 2019. Trust and recall of information across varying degrees of title-visualization misalignment. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI)*. 1–13.

Krzysztof Krejtz, Andrew Duchowski, Tomasz Szmidt, Izabela Krejtz, Fernando González Perilli, Ana Pires, Anna Vilaro, and Natalia Villalobos. 2015. Gaze transition entropy. *ACM Transactions on Applied Perception (TAP)* 13, 1 (2015), 1–20.

Firas Lethaus, Martin RK Baumann, Frank Köster, and Karsten Lemmer. 2013. A comparison of selected simple supervised learning algorithms to predict driver intent based on gaze data. *Neurocomputing* 121 (2013), 108–130.

Geoffrey R Loftus. 1972. Eye fixations and recognition memory for pictures. *Cognitive psychology* 3, 4 (1972), 525–551.

Geoffrey R Loftus and Norman H Mackworth. 1978. Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human perception and performance* 4, 4 (1978), 565.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WCACV)*. 1527–1536.

Adrian M Owen, Kathryn M McMillan, Angela R Laird, and Ed Bullmore. 2005. N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping* 25, 1 (2005), 46–59.

Cristina Palmero, Javier Selva, Mohammad Ali Bagheri, and Sergio Escalera. 2018. Recurrent cnn for 3d gaze estimation using appearance and shape cues. In *British Machine Vision Conference (BMVC)*. 1–13.

Seonwook Park, Emre Aksan, Xucong Zhang, and Otmar Hilliges. 2020. Towards end-to-end video-based eye-tracking. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*. Springer, 747–763.

Adam Perer and Ben Shneiderman. 2008. Integrating statistics and visualization: case studies of gaining clarity during exploratory data analysis. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 265–274.

Bastian Pfleging, Drea K Fekety, Albrecht Schmidt, and Andrew L Kun. 2016. A model relating pupil diameter to mental workload and lighting conditions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI)*. 5776–5788.

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.

Stanislav Pozdniakov, Roberto Martinez-Maldonado, Yi-Shan Tsai, Vanessa Echeverria, Namrata Srivastava, and Dragan Gasevic. 2023. How Do Teachers Use Dashboards Enhanced with Data Storytelling Elements According to their Data Visualisation Literacy Skills?. In *LAK23: 13th International Learning Analytics and Knowledge Conference*. 89–99.

Nicole C Rust and Vahid Mehrpour. 2020. Understanding image memorability. *Trends in Cognitive Sciences* 24, 7 (2020), 557–568.

Dario D Salvucci and Joseph H Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*. 71–78.

Hosnieh Sattar, Mario Fritz, and Andreas Bulling. 2020. Deep Gaze Pooling: Inferring and Visually Decoding Search Intents From Human Gaze Fixations. *Neurocomputing* 387 (2020), 369–382.

Hans-Jörg Schulz, Thomas Nocke, Magnus Heitzler, and Heidrun Schumann. 2013. A design space of visualization tasks. *IEEE Transactions on Visualization and Computer Graphics* 19 (2013), 2366–2375. Issue 12.

Sophie Stellmach, Lennart Nacke, and Raimund Dachselt. 2010. Advanced gaze visualizations for three-dimensional virtual environments. In *Proceedings of the 2010 symposium on eye-tracking research & Applications*. 109–112.

Florian Strohm, Ekta Sood, Sven Mayer, Philipp Müller, Mihai Bâce, and Andreas Bulling. 2021. Neural Photofit: Gaze-based Mental Image Reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 245–254.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification?. In *China National Conference on Chinese Computational Linguistics*. 194–206.

Edward R Tufte. 1985. The visual display of quantitative information. *The Journal for Healthcare Quality (JHQ)* 7, 3 (1985), 15.

Edward R Tufte, Nora Hillman Goeler, and Richard Benson. 1990. *Envisioning information*. Vol. 126.

Nash Unsworth, Gregory J Spillers, and Gene A Brewer. 2010. The contributions of primary and secondary memory to working memory capacity: an individual differences analysis of immediate free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 36, 1 (2010), 240.

Xi Wang, Andreas Ley, Sebastian Koch, David Lindlbauer, James Hays, Kenneth Holmqvist, and Marc Alexa. 2019. The mental image revealed by gaze tracking. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI)*. 1–12.

Yao Wang, Mihai Bâce, and Andreas Bulling. 2023. Scanpath Prediction on Information Visualisations. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* (2023), 1–15.

Yao Wang, Chuhan Jiao, Mihai Bâce, and Andreas Bulling. 2022a. VisRecall: Quantifying Information Visualisation Recallability via Question Answering. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 28, 12 (2022), 4995–5005.

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

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, Vol. 1. 2324–2335.

Yanyu Xu, Yanbing Dong, Junru Wu, Zhengzhong Sun, Zhiru Shi, Jingyi Yu, and Shenghua Gao. 2018. Gaze Prediction in Dynamic 360° Immersive Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5333–5342.

Francesca Zermiani, Andreas Bulling, and Maria Wirzberger. 2022. Mind Wandering Trait-level Tendencies During Lecture Viewing: A Pilot Study. In *Proceedings of the EduEye Workshop on Eye Tracking in Learning and Education (EduEye)*. 1–7.

Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2017. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence* 41, 1 (2017), 162–175.

Received November 2023; revised January 2024; accepted March 2024