SalChartQA: Question-driven Saliency on Information Visualisations

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

Weitian Wang
University of Stuttgart
Stuttgart, Germany
st176544@stud.uni-stuttgart.de

Abdullah Abdelhafez*
German University in Cairo
Cairo, Egypt
abdullahabdelhafez01@gmail.com

Mayar Elfares
University of Stuttgart
Stuttgart, Germany
mayar.elfares@vis.uni-stuttgart.de

Zhiming Hu†
University of Stuttgart
Stuttgart, Germany
zhiming.hu@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

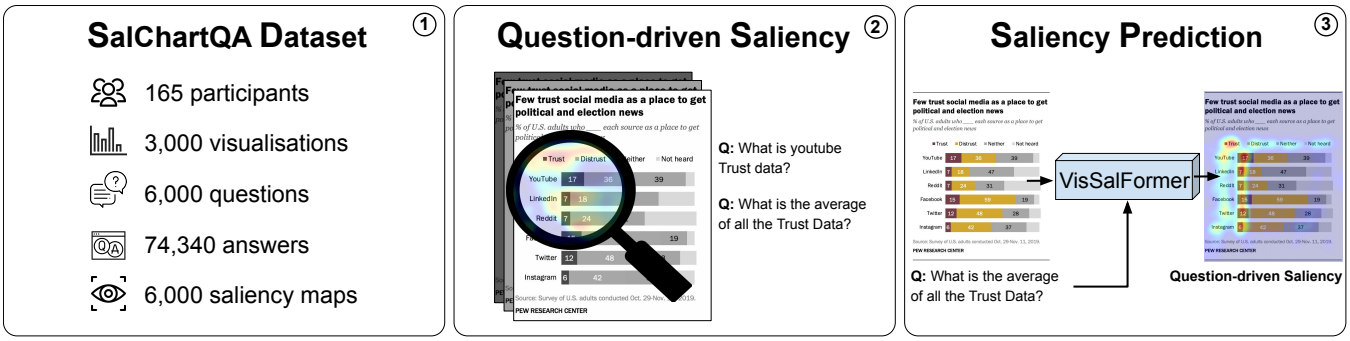


Figure 1: This work makes three distinct contributions: ① We propose *SalChartQA* – a novel large-scale dataset consisting of more than 74,000 answers to 6,000 questions on 3,000 information visualisations and corresponding crowd-sourced human attention data (N = 165). ② Analyses on SalChartQA demonstrate the strong impact of the question on visual saliency. Informed by these findings, we propose ③*VisSalFormer*: a Transformer-based model to predict question-driven saliency maps on information visualisations.

ABSTRACT

Understanding the link between visual attention and users' information needs when visually exploring information visualisations is under-explored due to a lack of large and diverse datasets to facilitate these analyses. To fill this gap we introduce SalChartQA – a novel crowd-sourced dataset that uses the BubbleView interface to track user attention and a question-answering (QA) paradigm to induce different information needs in users. SalChartQA contains 74,340 answers to 6,000 questions on 3,000 visualisations. Informed by our analyses demonstrating the close correlation between information needs and visual saliency, we propose the first computational method to predict question-driven saliency on visualisations. Our method outperforms state-of-the-art saliency models for several metrics, such as the correlation coefficient and the Kullback-Leibler divergence. These results show the importance of information needs for shaping attentive behaviour and pave the way for new applications, such as task-driven optimisation of visualisations or explainable AI in chart question-answering.

CCS CONCEPTS

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

KEYWORDS

Information visualisation, eye-tracking study, gaze bahaviour, visual saliency, deep learning

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

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM Reference Format:
Yao Wang, Weitian Wang, Abdullah Abdelhafez, Mayar Elfares, Zhiming Hu, Mihai Bâce, and Andreas Bulling. 2024. SalChartQA: Question-driven Saliency on Information Visualisations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3613904.3642942

1 INTRODUCTION

Analysis and computational modelling of human visual attention have emerged as key research topics in information visualisation research. The analysis of where users look at – measured, for example, using eye tracking [6] or mouse clicks as a proxy to gaze [23, 33] – provides rich information on how users perceive, explore, and process visualisations [48, 53]. This information is highly valuable for designers as it allows them to optimise the design of information visualisations for better clarity [37], memorability [6], and intelligibility [24]. Computational modelling of visual attention, also known as saliency prediction [30], is equally powerful as it enables many intelligent visual analytics applications, such as video compression [29], semantic segmentation [32], interactive UI design [21], and forecasting fixations in virtual reality (VR) [27, 28].

Users' attentive behaviour is, in part, driven by characteristics of the information visualisations themselves, such as the appearance of different visualisation elements (colour, size, font size, and weight, etc.) and their visual arrangements [6, 49, 65]. Complementing these bottom-up factors, where and in which order humans look at is also influenced by the task they are performing [53, 55], their goals and intentions [16, 70], or – in case of information visualisations – their information needs. For example, Figure 1-② shows how trustful social media is for political and election news. It is reasonable to expect that a user interested in average values will look at the visualisation differently than one interested in particular social media outlets. While bottom-up factors and their impact on users' attention have been widely studied in information visualisation research, the top-down influence of users' information needs remains largely unexplored. Similarly, there is a lack of computational models to predict saliency on information visualisations for different information needs. One reason for this is that studying different information needs and their influence on human visual attention is challenging: There is a wide range of needs, potential interaction effects between concurrent needs, as well as individual differences across users in how visual attention is deployed, specifically for various information needs while viewing visualisations. These challenges also apply to saliency prediction for which training of robust learning-based methods requires large amounts of ground-truth annotated data. This work addresses these challenges in two ways, allowing us to study the dependency between information needs and attention and advance saliency prediction methods.

First, we propose to use a question-answering (QA) paradigm to "induce" different information needs in users in a controlled manner. Users are tasked to answer questions about different information visualisations while their attention is being tracked. By varying the type of visualisation and question, we can isolate information needs and study the link between information needs and visual attention in a controlled manner. As such, the questions can be considered operational actions for different information needs. QA has attracted significant research interests in the computer vision and natural language processing communities in recent years [2, 60] but has only recently started to be explored in information visualisation research, e.g., to study visual attention behaviour [23, 53], chart question answering [47, 57], and information recall [66].

Second, we introduce SalChartQA – a large-scale saliency dataset collected using the QA paradigm. Given the difficulties in scaling up laboratory eye tracking studies, to collect the dataset, we opt for an online study on Amazon Mechanical Turk (AMT) using the popular BubbleView interface to track participants' visual attention [33]. This approach allows us to obtain data from a larger number of people and, as such, to cover individual differences in attention deployment more comprehensively. It also allows us to scale up the number of stimuli and answers in SalChartQA by an order of magnitude compared to previous datasets [6, 23, 37, 53] (also see Table 1 for a comparison). The size of SalChartQA has two notable advantages. On the first hand, it ensures the statistical power of our analysis. On the other hand, it finally allows us to propose a new learning-based method for question-driven saliency prediction on information visualisations (VisSalFormer) that achieves significant improvements over several baseline methods geared to predict question-driven saliency or importance maps for natural images and information visualisations.

In summary, our work makes the following three contributions:

(1) We introduce SalChartQA – a large-scale saliency dataset collected online from 165 participants who provided more than 74,000 answers to 6,000 questions posed for 3,000 information visualisations.
(2) We provide in-depth analyses on SalChartQA , including the correlation of questions and mouse clicks, and how questions influence human visual attention.
(3) We propose VisSalFormer – the first computational method to predict question-driven saliency on information visualisations that significantly outperforms several state-of-the-art methods.

2 RELATED WORK

2.1 Computational Modelling of Visual Attention

In the research area of human vision, computational modelling of visual attention is a key topic and has been well-studied in the past few decades. Existing works typically focused on modelling visual attention on natural images [15, 22, 30, 45]. For example, Itti et al. [30] proposed one of the earliest saliency prediction models that combine multiscale image features, including colour, intensity, and orientation, to predict the saliency maps of natural images. In addition, Cheng et al. [13] predicted saliency maps for natural images using the global contrast features of the images. The Multi-Duration Saliency Excited Model (MD-SEM) [22] was proposed to capture attention across multiple viewing durations, providing insights into the temporal dynamics of visual attention. Ramanishka et al. [54] proposed the caption-guided visual saliency, while Liu et al. [42] and Lou et al. [45] introduced the vision transformer [20] to saliency prediction, which all demonstrate the link between natural language and image saliency.

Recently, given the importance of information visualisations for communicating information effectively, researchers started to

model human visual attention on visualisations, which is fundamentally different from that on natural images [49, 56, 65] due to the highly structured nature of the visualisations. Specifically, Matzen et al. [49] proposed a saliency prediction model for data visualisations that combines the output of a typical saliency model designed for natural images and the output of a text recogniser to generate saliency maps that are specialised for data visualisations. Shin et al. [56] focused on human visual attention on chart images and presented a convolutional neural network-based encoder-decoder architecture to predict saliency maps for data visualisations. Sood et al. [61] leveraged a cognitive model of visual attention as an inductive bias to train saliency models on both natural scenes and information visualisations. Wang et al. [65] proposed a unified model of saliency and scanpaths (i.e. sequences of eye fixations) for information visualisations that first predicts multi-duration element-level saliency maps and then samples scanpaths from the saliency maps. However, existing visual attention models on information visualisations have only one "averaged" saliency map as ground truth, which means they can only predict visual saliency under a "free-viewing setting", i.e. users could freely explore a visualisation. However, in practice, users often have specific information needs they want to address. For example, what is the average value for a certain time series, or what is the minimum or maximum? Such information needs will likely lead to different attentive behaviour and, because of this, we explored the more challenging but practically relevant question-driven settings [47, 53], in which users are assigned a specific question while exploring the visualisations.

2.2 Task-driven Visual Attention

In recent years, human visual attention in task-driven settings has been a long-lasting, challenging topic in cognitive science and computer vision [51]. Borji et al. [5] and Koulieris et al. [35] both focused on human visual attention in video games. Borji et al. predicted saliency maps using players' input such as 2D mouse positions and joystick buttons. Koulieris et al. used game state variables to predict users' gaze positions in video games. Zheng et al. [71] predicted visual saliency on webpages under different tasks such as information browsing and form filling using an end-to-end learning framework. In graphical user interfaces, Xu et al. [69] proposed a spatio-temporal model to predict visual attention under a text editing task, while Jokinen et al. [31] modelled visual search with a cognitive model. Bâce et al. [3] studied users' visual attention during everyday mobile device interaction in different applications and usage contexts. Hu et al. [26, 27] analysed visual attention in task-driven VR environments and used task-related features to predict users' gaze positions. Even though task-driven visual attention has been well-studied in other scenarios, limited work exists on information visualisations. In this work, we fill the gap by conducting comprehensive analysis and experiments to study and predict task- (question-) driven human visual attention on information visualisations.

2.3 Visual Attention Datasets on Information Visualisations

Researchers have collected many datasets using eye tracking technology to study human visual attention on information visualisations since eye gaze provides rich information about visual search and decision-making [9, 36]. Some researchers investigated visual attention on information visualisations under free-viewing settings. For example, Borkin et al. [6] labelled a dataset of 393 visualisations and analysed the eye movements of 33 participants and thousands of participant-generated text descriptions of the visualisations. Shin et al. [56] collected a large-scale dataset that contains 10,960 visualisations and 12,504 user responses using a webcam-based eye-tracking approach. Other researchers studied task-driven visual attention on information visualisations. Specifically, Gomez et al. [23] collected a dataset that covers 20 information visualisations to analyse human visual attention under visual analysis tasks. Lallé et al. [37] collected a dataset containing 40 interactive visualisations to predict occurrences of confusion during the interaction using eye tracking and mouse data. Polatsek et al. [53] performed an eye-tracking study using 30 charts to analyse human visual attention when solving three low-level analytical questions. Lastly, Kim et al. [33] demonstrated how mouse-contingent data approximates eye fixations on information visualisations under a free-viewing and descriptive task. However, existing task-driven eye tracking datasets on information visualisations are limited in their sizes (usually less than 100 stimuli) [23, 33, 37, 53], and there exists no large-scale task- (question-) driven visual attention dataset. In this work, we fill this gap by collecting a large-scale question-driven visual attention dataset comprising thousands of information visualisations.

3 SalChartQA DATASET

To better understand and predict question-driven human attention on information visualisations we propose the SalChartQA – a novel large-scale dataset consisting of more than 74,000 answers to 6,000 questions on 3,000 information visualisations together with crowd-sourced human attention data obtained from 165 users (see Figure 2). Visualisations and questions in our dataset are sourced from the ChartQA dataset [47]. Table 1 shows a comparison between SalChartQA and other visualisation saliency datasets that contain questions. Existing datasets that were collected using an eye tracker [6, 23, 37, 53] are limited in size to tens or hundreds of stimuli. BubbleView [33] is a demonstrated approach to collect human visual attention from mouse clicks, used previously in other datasets [8, 11]. We facilitate the large-scale human attention data collection using the BubbleView [33] interface to track participants' visual attention. Our dataset and code are publicly available at https://doi.org/10.18419/darus-3884.

3.1 Collection of Visualisations and Questions

ChartQA-H dataset [47] is a human-annotated Chart Question Answering dataset, covering the three most commonly used visualisation types: bar plots, line plots, and pie charts. We randomly selected 3,000 out of 4,800 visualisations from the ChartQA-H dataset. Our selection contains 1,958 bar plots (1,417 horizontal, 541 vertical), 672 line plots, and 370 pie charts. There were 1,197 *simple charts* in our selection whose data table had exactly two columns [47]

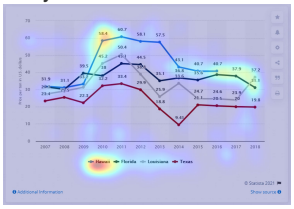
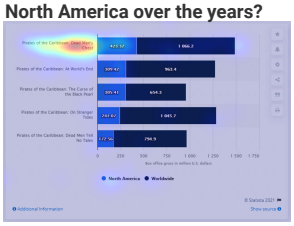
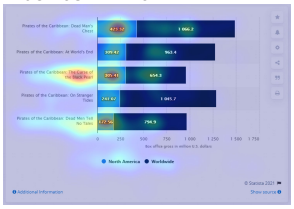
	Q: What was the lower number of the red line?	Q: What state had the biggest one year increase?	Q: Which pirates of Caribbean collected highest revenue in North America over the years?	Q: In which movie the difference between the revenue collection between North America and world wide was minimum?
				
Question Type	Finding Extremum	Filtering	Filtering	Comparison
Question length	34	38	73	94
Is Answer Numeric?	Yes	No	No	No
Question Accuracy	56.25 %	52.94 %	86.67 %	56.25 %

Figure 2: Sample saliency maps from SalChartQA including the corresponding questions at the top and four question characteristics.

and 1,803 more *complex charts*. Every visualisation had two corresponding human-annotated reasoning questions. There were 3,996 numerical-answer questions and 2,004 non-numerical-answer questions. We noticed 239 out of 6,000 answers (3.98%) were incorrectly labelled in ChartQA. We corrected these cases and made them available as part of SalChartQA . Since ChartQA-H provides no question category annotations, we manually checked 600 questions in our selection for five well-established question categories: comparison, computing derived value, data retrieval, finding extremum, and filtering questions. We created a keyword list for each question category, such as "how many" and "median" for computing derived value questions. Then, we used these keyword lists to classify the 6,000 questions into five categories (see supplementary material for specification of keyword lists):

(1) 1,841 comparison (CP) questions [1]. Comparison questions are judgments about data properties, especially to compare two data points. Example questions: "Which country has the lesser protected areas over the years, Lithuania or Saudi Arabia?", and "How many times is France bigger than Congo?".
(2) 1,496 computing derived value (CV) questions [55]. Participants were asked to compute a derived value of some given data points for these questions. Example questions: "What is the average of 2019 and 2020 blue bar?", and "What is the difference of value in the yield of South America and India?".
(3) 1,920 retrieving value (RV) questions [47, 53, 55]. For these questions, participants were asked to identify the values of attributes for given data points. Example questions: "How many people died from Brown coal production?", and "How many countries have less than 60% of people that are confident in Obama?".
(4) 390 find extremum (FE) questions [53, 55, 66]. These questions require finding data points that have an extreme data attribute value. Example questions: "What is the highest channel distribution by number of dispensed prescriptions in the U.S. in 2018?", and "In which year the white members were maximum?".
(5) 285 filtering (F) questions [53, 55, 66]. Participants were asked to find data points satisfying several given concrete conditions on data attribute values for these questions. Example questions: "In which year the Mexican government's campaign against drugs traffickers is making 47 percent progress?", and "Which country has its value 5.97%?".

3.2 Crowd-sourcing Study Set-up & Participants

We used the widely adopted BubbleView approach [33] for collecting human attention data. Participants used mouse clicks as a proxy to eye fixations to deblur small "bubble" regions. There were no hold and drag operations, and once the user clicked another location, the previous region was blurred again. In our study, the circle surrounding each mouse click had a radius of 30 pixels (~1.5 visual angle) [33]. We deployed our study on Amazon Mechanical Turk (AMT), a crowd-sourcing platform to collect human attention data and question answers on all 6,000 questions, splitting them randomly and evenly into 300 human intelligence tasks (HITs). The HIT layout is illustrated in Figure 3, where every question was shown alongside a blurred visualisation, and a text input field. Each visualisation was shown in its original size and blurred by a 40-pixel Gaussian filter to make any text labels unreadable. Participants were instructed to click on as many areas of interest on the visualisation as required to provide an answer to the question. There was no time limit to clicking and providing an answer. Once the answer was found, participants were asked to type the answer in the text field and proceed to the next question. Every HIT contained 20 visualisation-question pairs, and crowd workers were not allowed to participate in HITs that contained the same visualisations. We implemented the procedures as a web application and integrated it into the BubbleView interface [33]. We incorporated attention checks in the form of mathematics tasks that randomly appeared

Table 1: A comparison between our dataset and other visualisation saliency datasets that contain questions. Our dataset is the first large-scale question-driven saliency dataset of information visualisations. Vis: number of visualisations, Q: number of questions. F: filtering, FE: finding extremum, RV: retrieving value, CP: comparison, CV: computing derived value.

Datasets	Vis * Q	Participants	Answers	Collection	Questions
Gomez et al. [23]	20 * 1	100 + 18	2,000 + 360	Cursor + Eye Tracker	Visual Analysis
Borkin et al. [6]	393 * 1	33	6,563	Eye Tracker	Description
Lallé et al. [37]	40 * 1	136	5,440	Eye Tracker	Confusion
Polatsek et al. [53]	30 * 3	47	1,211	Eye Tracker	F, FE, RV
SalChartQA (Ours)	3,000 * 2	165	74,360	Mouse Click	F, FE, RV, CP, CV

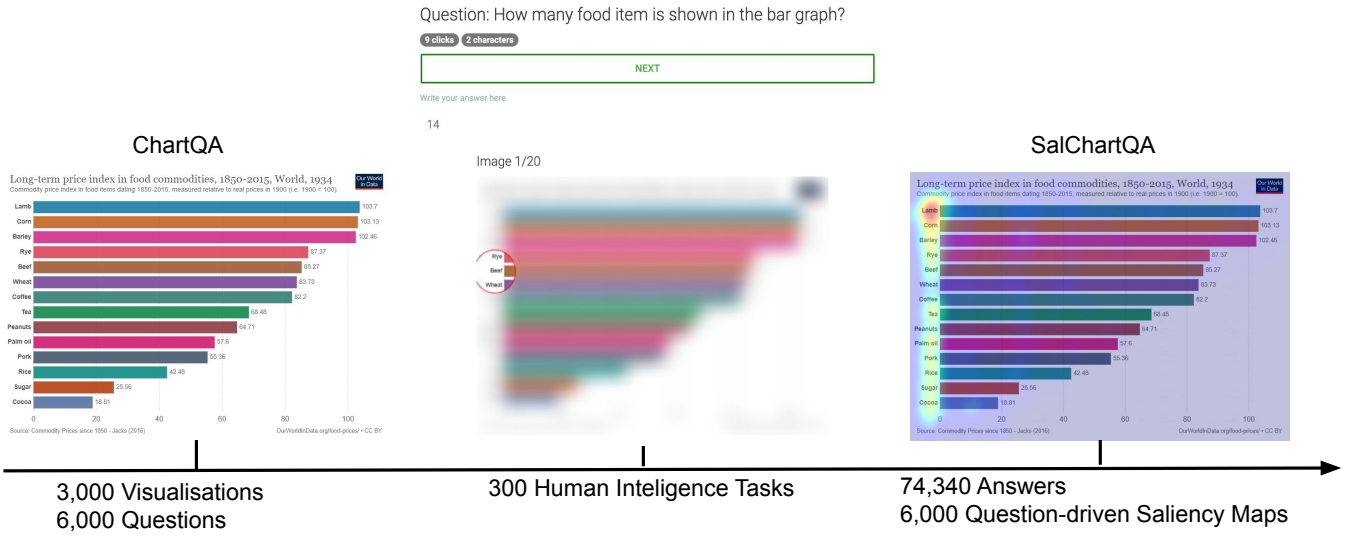


Figure 3: Our online study was based on 3,000 visualisations and 6,000 questions from the ChartQA dataset [47] (left). Implementing the BubbleView method [33], each visualisation was blurred and participants were asked to answer the question given at the top by clicking on different parts of the visualisation, thereby unblurring small parts of the image with each click (middle). 300 human intelligence tasks yielded over 74,000 answers and 6,000 question-driven saliency maps (right).

once within the HIT to ensure data quality [25]. If they failed the attention check, the study was immediately terminated. Workers were paid $ 1.00 for completing each HIT. All results remained pseudonymised to the experimenters.

3.3 Responses

In total, we received 5,152 HIT responses, of which 424 did not pass the attention check. We noticed that some answers were entirely unrelated to questions in low-accuracy responses, such as answering "No" for the question "What is the average of all the Trust Data?". To further improve data quality, we filtered the HIT responses with low accuracy by a 1.5 Inter Quartile Range (IQR) [64], i.e., we discarded 1,010 HITs with a question-answering accuracy lower than 35.34%. See supplementary material for further analysis of the discarded HIT responses. 3,718 HITs remained after filtering, with a mean accuracy of 80.67% (σ = 16.22%), corresponding to 74,360 answers to 6,000 questions. The mean number of answers to one question is 13.10, with a variance of 6.39 answers. The minimum number of answers to one question is 10, which accounts for up to 97 – 98% of the optimal performance for covering an average of 89 – 90% of eye fixations [33]. For all 74,360 answers, the mean number of clicks is 19.20 (σ = 22.09), and the mean duration between the first and the last clicks is 17.56 s (σ = 44.78 s).

4 ANALYSIS

To study the influence of information needs on attention, operationalised through the question-answering paradigm, we first performed a series of analyses on the collected data.

4.1 Number of Mouse Clicks

The number of clicks performed in BubbleView was shown to serve as a good measure of visual importance and attention [33]. In Figure 4 (top), we examined whether questions impact the *click count* by grouping answers based on *question length* and *whether the answer is numerical*. The results revealed a positive correlation between the *number of clicks* and *question length*, as evidenced by Spearman's rank correlation coefficient (r = 0.244, p < 0.001). The *number of clicks* did not conform to a normal distribution,

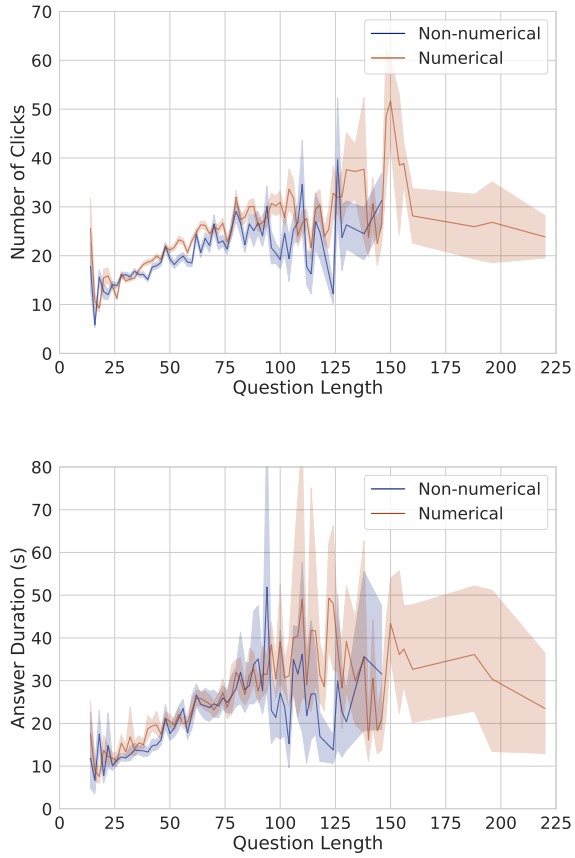


Figure 4: Relationship between the number of clicks and question length (top) and the answer duration and question length (bottom) for numerical and non-numerical answers. Question length means the number of characters in a question. Both the number of clicks and answer duration positively correlate to question length.

as indicated by the Kolmogorow-Smirnov normality test ($D = 0.999$, $p < 0.001$). Consequently, we employed a Welch-Satterthwaite T-test to assess the differences between numerical and non-numerical questions. The result revealed a statistically significant difference, with non-numerical questions receiving significantly fewer clicks than numerical questions (t (53,030) = 13.072, $p < 0.001$). In Figure 4 (bottom), we similarly examined the relationship between *answer duration* (between the first and the last mouse click) and *question length*. The *answer duration* positively correlated with *question length* as shown by the Spearman's rank correlation coefficient ($r = 0.303$, $p < 0.001$). Moreover, we observed a larger variance in the *number of clicks* and *answer duration* when questions are longer than ~ 100 characters. This can be attributed to only 171 questions having more than 100 characters.

4.2 Saliency Map Generation and Metrics

We generated the saliency maps by blurring the BubbleView click locations with a Gaussian kernel with a sigma of 19px (~1 visual angle [10, 33]). We used the Shannon entropy (SE) and saliency coverage (SC) to characterise the resulting saliency maps.

- **Shannon entropy** is widely used in analysing saliency maps [40, 46]. It characterises how much a saliency map differs from a uniform distribution (all pixels with the same value) [7] by a positive value ranging from 0 to infinity. The lower the Shannon entropy is, the closer the saliency map to a uniform distribution.
- **Saliency coverage** quantifies how many pixels are activated given a certain threshold on a grayscale map [12]. It indicates how big the areas participants have explored. SC ranges between 0 and 1 (the higher the saliency coverage is, the larger the areas observers explore). We used the mean threshold value of Otsu's thresholding algorithm [52] of all saliency maps (0.238) as the global threshold [38] to generate binary maps. Saliency coverage was then calculated as the percentage of activated pixels in binary maps.

For comparing the similarity between two saliency maps, we used Pearson's correlation coefficient (CC) and normalised scanpath saliency (NSS) to evaluate the similarities of saliency maps:

- **Pearson's correlation coefficient (CC)** is the covariance of two maps divided by the product of their standard deviations. The CC score is -1 when the two maps are complementary and 1 when they are identical.
- **Normalized scanpath saliency (NSS)** measures the saliency values at fixation locations (clicks) along a participant's scanpath. A positive NSS indicates map correspondence, chance at NSS = 0, and a negative NSS indicates anti-correspondence [10].

4.3 Saliency Similarity within Participants

We analysed participants' visual behaviour variances to ensure that participants who answered questions correctly had consistent visual behaviour according to their information needs. We randomly and evenly split all participants who answered a question correctly into two groups. We generated a saliency map for every group and computed the similarity between the two maps. 1,612 questions were evaluated, with a minimum of four participants who answered correctly and four wrongly. The mean NSS between correct-answer saliency maps was 2.182 ($\sigma = 0.949$), which was significantly higher than between wrong-answer saliency maps ($\mu = 1.491$, $\sigma = 0.905$), paired-sample T-test, t (1,611) = 27.104, $p < 0.001$. The mean CC between split correct-answer saliency maps was 0.760 ($\sigma = 0.145$), which was significantly higher than between wrong-answer saliency maps ($\mu = 0.549$, $\sigma = 0.232$), paired-sample T-test, t (1,611) = 34.514, $p < 0.001$). Therefore, the high and stable NSS and CC of correct-answer saliency maps indicated consistent participant viewing behaviour.

4.4 Visualisation and Saliency

We investigated the question of whether visualisation characteristics influence question-driven saliency. When comparing mean Shannon entropies between *simple charts* and *complex charts*, we found that *simple charts* had a significantly lower entropy of 3.818

(σ = 1.347) compared to *complex charts* (μ = 4.845, σ = 1.231), Welch-Satterthwaite T-test, t (4,803) = 29.919, p < 0.001). We used the mean number of areas of interest (AOIs) to measure the complexity of each visualisation following prior work [4],. Based on annotations provided by ChartQA, we found visualisations had a mean of 26.05 AOIs (σ = 19.12). Using correlation analysis between the number of AOI and the Shannon Entropy for every visualisation we found a significant positive correlation, as supported by Spearman's rank correlation coefficient ($r = 0.063, p < 0.001$). Then, we analysed whether the visualisation types influence the Shannon entropy (SE) and saliency coverage (SC). Pie charts, line plots, vertical bars, and horizontal bars have a mean SC of 15.2%, 13.1%, 11.4%, and 7.6%, respectively. The observed differences in SC across visualisation types were statistically significant according to the Kruskal-Wallis Test (H = 273.578, p < 0.001). Subsequent post-hoc Dunn's Test further confirmed the significance of differences between all pairs of visualisation types (p < 0.001). Line plots, pie charts, horizontal bars, and vertical bars have a mean SE of 5.129, 4.797, 4.205, and 3.930, respectively. SE also significantly differed across visualisation categories, Kruskal-Wallis Test (H = 632.459, p < 0.001). A post-hoc Dunn's Test confirmed the significance between all visualisation types (p < 0.001). Therefore, we conclude that visualisation types strongly influence question-driven saliency. Pie charts and line plots have larger regions activated in saliency maps, and their saliency maps have higher Shannon entropy than bar plots.

4.5 Questions and Saliency

To better understand the interplay between visual saliency and characteristics of the questions, we conducted further analyses on whether question categories, question length, and question accuracy influence visual saliency.

Question categories. FE, CP, CV, RV, and F questions have a mean SE of 4.575, 4.574, 4.533, 4.455, and 4.278, respectively. SE significantly differs across question categories according to the Kruskal-Wallis Test (H = 19.379, p < 0.001). A post-hoc Dunn's Test confirmed the significance between CP and RV questions (p = 0.024), CP and F questions (p = 0.002), CV and F questions (p = 0.024), and FE and F questions (p = 0.027). For saliency coverage, CP questions have the highest SC of 12.9%, followed by CV (12.4%), RV (11.4%), FE (9.4%), and F questions (8.4%). SC significantly differs across question categories according to the Kruskal-Wallis Test (H = 165.202, p < 0.001). A post-hoc Dunn's Test confirmed the significance between all question types (p < 0.001). The findings indicate that participants explore the largest area to answer comparison questions, while they explore the smallest regions for filtering questions.

Question length. Our analyses showed a significant positive correlation between SE and question length, supported by Spearman's rank correlation coefficient (r = 0.192, p < 0.001). Similarly, a positive correlation between SC and question length was observed, as evidenced by Spearman's rank correlation coefficient (r = 0.195, p < 0.001). This suggests increasing question length is linked to higher Shannon entropy and saliency coverage.

Question accuracy. This analysis applied to 4,326 questions that exhibited a combination of at least one correct and one wrong answer. Figure 5 illustrates example saliency maps from an easy question (accuracy > 85%) and a hard question (accuracy < 25%) overlaid on the corresponding visualisations from SalChartQA. As can be seen from the figure, the SE negatively correlated with question accuracy, Spearman's rank correlation coefficient (r = -0.172, p < 0.001). Similarly, a negative correlation was observed between SC and question accuracy, as evidenced by Spearman's rank correlation coefficient (r = -0.226, p < 0.001). We conclude that the more areas participants explore (higher SC), the less likely they will answer correctly.

5 VisSalFormer

Our analyses revealed that human visual attention is strongly influenced by users' information needs when looking at information visualisations. Informed by these findings, we then focused on developing a computational method to predict saliency on information visualisations taking the information needs into account. To this end, we propose VisSalFormer – a Transformer-based saliency model that is specifically geared to information visualisations and that takes both visualisations and questions as input. As shown in Figure 6, VisSalFormer consists of two main branches: In the upper branch, a pre-trained Swin Transformer [43] encodes the visualisation into vision features. In the lower branch, a pre-trained bidirectional encoder representation from Transformers (BERT) [19] encodes the question string into text features. In the cross-modality feature fusion module, vision and text features are effectively fused together into combined latent feature maps. Finally, a CNN-based decoder converts latent features into saliency maps. We explain the individual components of the model in detail in the following subsections.

5.1 Visualisation Embeddings

We leveraged the capabilities of the Swin Transformer for encoding visualisations. The Swin Transformer excels in producing robust hierarchical features, which has demonstrated its effectiveness in serving as a general backbone for various tasks such as image classification, object detection, semantic segmentation [43], and image restoration [39]. The hierarchical features from the Swin Transformer aligned with the intrinsic complexity and element-rich nature of visualisations [65]. We fed a visualisation into a Swin Transformer, which is processed to 224×224. Subsequently, we directed the sequence of hidden states from the final layer of the Swin Transformer through two linear layers with the size of 49×768, each followed by a rectified linear unit (ReLU) activation function. The linear layers and ReLU aimed to map the extracted visual features into a unified fused feature space.

5.2 Question Embeddings

We employed the BERT model to encode questions, as it has gained widespread recognition for its efficacy in various language understanding tasks, spanning from textual classification [62] to reading comprehension [68]. We fed a question string into BERT to generate 768-neuron contextual word embeddings. To extract keyword features from the question, we initialised a random 10×768 vectors as the query and used the text features as the key and value for a cross-attention module. Then, the output of the cross attention was

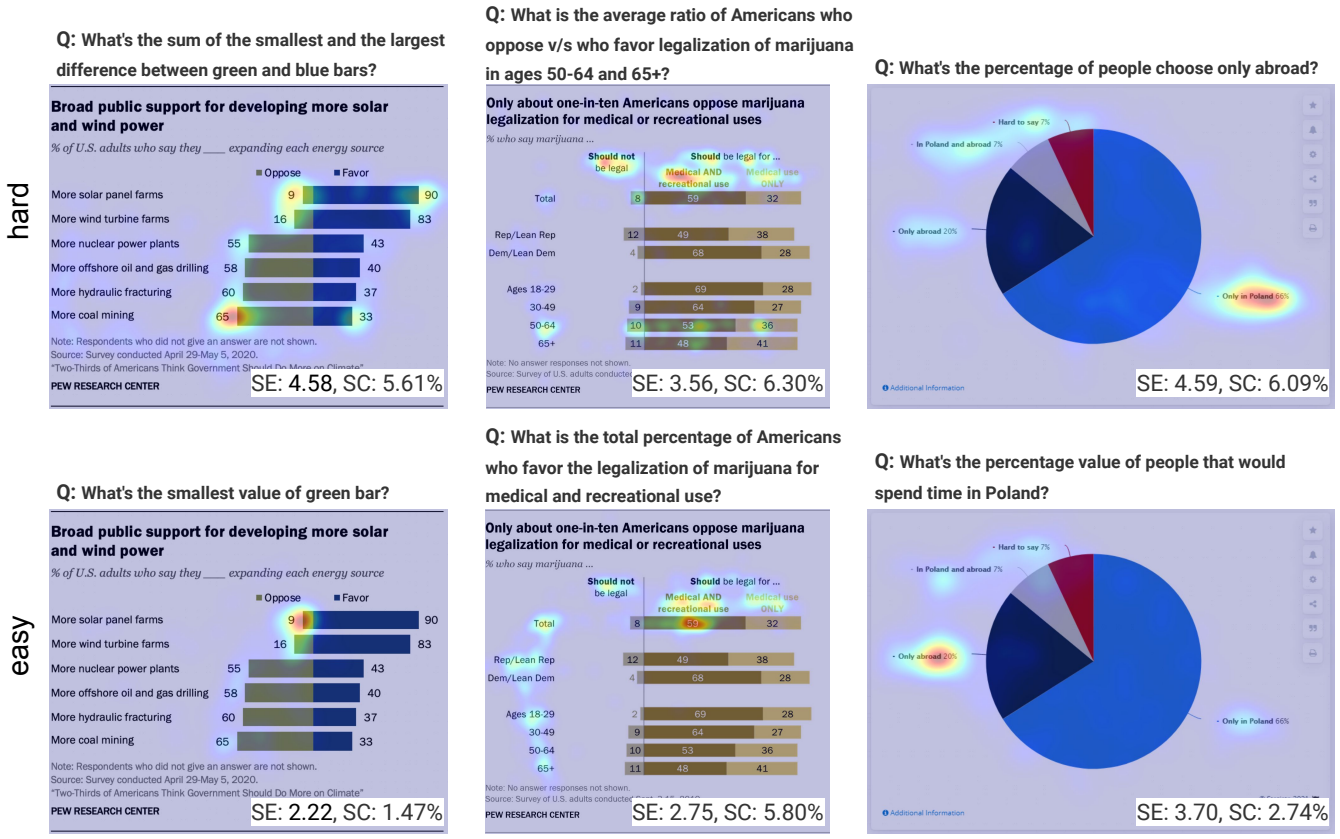


Figure 5: Sample saliency maps with computed Shannon entropy (SE) and saliency coverage (SC) under one easy question (accuracy > 85%) and another hard question (accuracy < 25%) overlaid on visualisations in SalChartQA. Decreasing question accuracy corresponds to higher Shannon entropy and saliency coverage.

fed to two linear layers with the size of 10×768, each followed by a ReLU activation function.

5.3 Cross-Modality Feature Fusion

We proposed a cross-modality feature fusion module to fuse visualisation and question embeddings effectively. We concatenated the output features of both vision and text features into a 59×768 vector. Then, we added the output of self-attention to itself and did a layer normalisation. The output was used as the key and value of another cross attention, and the vision features were used as the query. Then, we added the output of the cross attention to the vision features and did a layer normalisation. Finally, we reshaped the output vector to a $7 \times 7 \times 768$ latent feature maps for decoding into saliency maps.

5.4 Decoder

We implemented a CNN decoder to convert the latent features into saliency maps. The decoder contains 7 sequential blocks, each starting with a 3×3 2D convolutional layers. For blocks 1 to 8, a Batch Normalization (BN), a ReLU activation function, and a dropout layer with 0.1 probability are applied after the convolutional layer. After blocks 2 and 4, a 2-scale upsampling that adopts bilinear interpolation is applied to the feature map. After block 6, a 4-scale upsampling that adopts bilinear interpolation is applied to the feature map. For block 7, a BN and a Sigmoid activation function are applied after the convolutional layer to predict saliency maps. The output of the decoder is 128×128, and we resized it with bilinear interpolation back to the original image resolution. See supplementary materials for a visualisation of the decoder's architecture.

6 EXPERIMENTS

We conducted a series of experiments to compare the performance of VisSalFormer with state-of-the-art saliency prediction methods on SalChartQA . Different ablated versions of the VisSalFormer were also evaluated.

6.1 Baseline Methods

Five baseline methods, TranSalNet [45], MD-SEM [22], MD-EAM [65], DVS [49], and UMSI [21], encompass a variety of approaches designed to predict saliency or importance maps for visual stimuli (TranSalNet and MD-SEM for natural scenes, others for information visualisations).

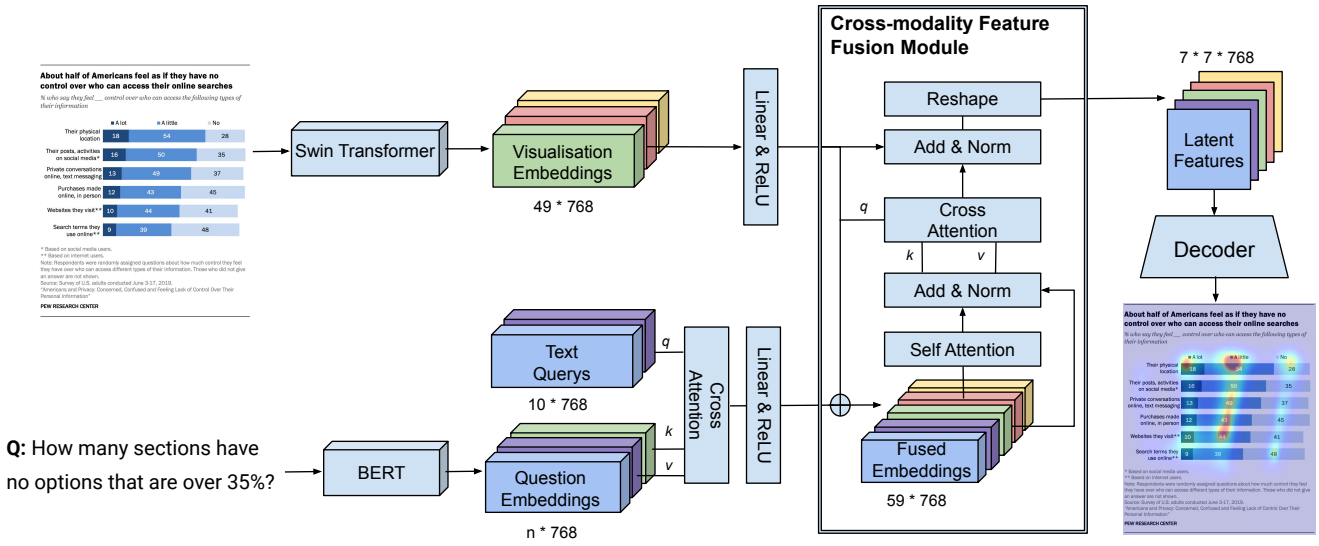


Figure 6: Overview of our proposed *VisSalFormer* model. Core to *VisSalFormer* is a cross-modality feature fusion module to handle multi-modal input consisting of a visualisation and corresponding question. Input to the fusion module is the visualisation and question embeddings obtained using pre-trained Swin and BERT Transformers. Finally, a CNN-based decoder uses the fused latent features to predict a saliency map.

- **TranSalNet** [45] integrated transformer components to CNNs to capture the long-range contextual visual information. TranSalNet was trained with the Adam [34] optimiser for 10 epochs with a start learning rate of $1E-5$, which was decreased by a factor of 10 every three epochs. Models are trained with a batch size of 4 for 30 epochs with a stop patience of 5 epochs.
- **Multi-Duration Saliency Excited Model (MD-SEM)** [22] aimed to predict saliency maps for three viewing duration. We split mouse clicks in SalChartQA into 0–3 s, 3 – 5 s, and 5 – 10 s saliency maps for training. The MD-SEM model was trained from scratch with a batch size of 8. We set the loss weights to 3 for CCM, 10 for KL, -5 for CC and -1 for NSS during training. MD-SEM was trained with the Adam optimiser for 10 epochs with a start learning rate of $1E-4$, which was decreased by a factor of 10 every three epochs. Table 2 reports the 5 – 10 s branch that has the highest metrics.
- **Multi-Duration Element Attention Model (MD-EAM)** [65] used the MD-SEM architecture to train element fixation density maps in the field of visualisations. We generated element fixation density maps on SalChartQA (0 –3 s, 3 – 5 s, and 5 – 10 s) for training. The MD-EAM model was trained from scratch with a batch size of 8. We set the loss weights to 3 for CCM, 10 for KL, -5 for CC, and -1 for NSS during training. MD-EAM was trained with the Adam optimiser for 10 epochs with a start learning rate of $1E-4$, which was decreased by a factor of 10 every three epochs. Table 2 reports the 5 – 10 s branch that has the highest metrics.
- **Data Visualisation Saliency (DVS) Model** [49] integrated bottom-up saliency maps of the Itti-Koch [30] model with text-region maps. Since it is not a deep-learning-based approach, we used the official code[1] in our evaluation.
- **Unified Model of Saliency and Importance (UMSI)** [21] was designed to predict importance maps for various visual stimuli, including infographics, movie posters, mobile user interfaces, advertisements, and webpages. Since there were no importance map annotations on our SalChartQA , we directly used the official weights in our evaluation.

6.2 Training and Implementation Details

Dataset preparation. We randomly split the SalChartQA into training, validation, and test subsets with a ratio of 7:2:1. This partitioning resulted in 2,114 visualisations (4,228 questions) in the training set, 595 visualisations (1,190 questions) in the validation set, and 291 visualisations (582 questions) in the test set. VisSalFormer and all baseline methods were trained using the above partition of SalChartQA . Notably, since all baseline methods are limited to predicting a single saliency map for each visualisation, we merged two question-driven saliency maps from one visualisation as the ground truth map for training baseline methods.

Implementation details. VisSalFormer used the Swin-T[2] model and started training from the ImageNet-1K [18] pre-trained weights, and used pre-trained base-uncased weights for BERT[3]. The loss of VisSalFormer is a weighted combination of Kullback Leibler divergence (KL), Pearson's Correlation Coefficient (CC) and Normalized Scanpath Saliency (NSS). We set the loss weights to 10 for KL, -5 for CC, and -2 for NSS during training. VisSalFormer was trained with a batch size of 32 for 200 epochs with the AdamW [34, 44] optimiser with a learning rate of $6E-5$ and incorporated a weight decay of

[1]Code available at http://www.cs.sandia.gov/~atwilso/get_dvs.html

[2]https://huggingface.co/microsoft/swin-tiny-patch4-window7-224

[3]https://huggingface.co/bert-base-uncased

$1E-4$. All experiments were conducted on a single NVIDIA Tesla V100 GPU with 32 GB VRAM.

6.3 Quantitative Evaluation

Metrics. We used five popular metrics for evaluating performance: Normalized Scanpath Saliency (NSS), Pearson's Correlation Coefficient (CC), Kullback-Leibler divergence (KL), Similarity or histogram intersection (SIM), and Area Under the Curve (AUC). NSS and AUC are calculated on fixation maps, while CC, KL and SIM are on saliency maps.

Results. We compared the performance of our VisSalFormer to the five baseline methods, TranSalNet, MD-SEM, MD-EAM, DVS, and UMSI. Table 2 demonstrated that VisSalFormer dominated baselines in all five metrics. We conducted paired-sample T-Tests between VisSalFormer and the best baseline model in each metric. Significance was found in all five metrics with a significant level of 0.001. After being trained on SalChartQA from scratch, all baseline methods performed better than their official weights. The performance of all baseline methods without training on SalChartQA was inferior on SalChartQA, which suggested that the free-viewing visual saliency / importance is fundamentally different from question-driven saliency.

Ablation study. We further carried out three ablation studies to evaluate the effectiveness of our model. First, we replaced the Swin Transformer with two commonly used image encoders to see the influence of the vision features on question-driven saliency. We substituted the Swin Transformer with two commonly used image encoders, ViT [20] and Xception [14], to validate the effectiveness of the visualisation embeddings in VisSalFormer . The first two rows in Table 3 demonstrate that the Swin Transformer dominates the other two encoders in CC, KL, SIM, and AUC, but Xception has the highest NSS score. Second, we replaced the Bert with two newer large language models (LLMs), Llama[4] [63] and Bloom[5] [67]. Due to the limited computational resources at our disposal, we froze the weights of the Llama and Bloom models. Third, we removed the components in our model to analyse how each component contributes to the full model. We evaluated the contribution of question embeddings in three ablated versions. In these versions, we removed the entire question embedding branch from VisSalFormer , the cross-attention layer in the cross-modality fusion module, and the cross-attention and self-attention layers in the cross-modality fusion module. The fifth to seventh rows in Table 3 show that VisSalFormer outperforms all three ablated versions, highlighting all the components in our pipeline is essential.

6.4 Qualitative Evaluation

Figure 7 depicts six predictions from TranSalNet [45] and VisSalFormer. TranSalNet consistently generates identical saliency maps for one visualisation despite variations in the posed questions. The line plot accentuates this incongruity further, as TranSalNet exhibits an expansive region of false positives on the left side compared with the ground truth saliency maps. In contrast, the predictions from VisSalFormer demonstrate a remarkable adaptability across diverse questions, effectively encompassing most ground truth maps.

However, VisSalFormer still confronts three persistent challenges. Firstly, the pie chart predictions from VisSalFormer have similar saliency maps under distinct questions. This indicates a potential limitation in effectively biasing saliency maps towards question-specific regions by incorporating question embeddings. Secondly, examining fixation clusters in VisSalFormer unveils a higher incidence of false positives than the ground truth. This divergence becomes particularly evident in the ground truth map corresponding to the question "how much percentage showed in red color" where solely the red colour label is salient. In contrast, VisSalFormer activates labels for light blue and navy blue, unveiling a perceptible misalignment with the ground truth. Thirdly, we found a notable misalignment between VisSalFormer 's predictions and the ground truth when the question involves similar colours in charts, such as "light blue" and "navy blue". The model struggles to discern subtle distinctions between these closely related hues, leading to occasional activation of saliency for visually resembling colours. This introduces an open challenge in charts that have similar colour coding. Addressing these intricacies is imperative for refining the fidelity and robustness of VisSalFormer.

7 DISCUSSION

7.1 SalChartQA

In this work, we introduced the SalChartQA dataset, the first large-scale question-driven saliency dataset on information visualisations, both in terms of the number of participants and the number of visualisations/questions. The crowd-sourced dataset was collected on AMT with BubbleView clicks which has been proven to successfully approximate eye fixations on information visualisations [33]. There are two noteworthy advantages of BubbleView over eye trackers. Firstly, BubbleView plays a pivotal role in achieving data minimisation by selectively gathering information required for specific computations, thereby excluding extraneous elements such as face and eye images to protect user privacy online. Secondly, BubbleView proves to be a cost-effective alternative, presenting a more budget-friendly option (~ $1 per image) while ensuring on-the-fly applicability within browser environments without compromising data quality. Our large-scale data collection requirement (1) highly benefits training saliency prediction models (c.f. Section 6.3) and (2) inferring insights such as information needs with high statistical power. Although we mainly focus on question-driven saliency prediction, SalChartQA can further enable many applications, such as visualisation optimisation (redesign) [50, 58] and explainability in Chart Question Answering (CQA) [17].

7.2 VisSalFormer – A novel method for question-driven saliency prediction

As demonstrated in Section 6.1, all established baseline saliency methods devised for information visualisation have relied only on *image features*. Consequently, these methods can merely produce an "averaged" saliency map, devoid of any fine-grained sensitivity to specific questions or user queries. VisSalFormer , in stark contrast, leverages both embeddings from visualisation and question in the cross-modality feature fusion module, enabling VisSalFormer to

[4]https://huggingface.co/Enoch/llama-7b-hf
[5]https://huggingface.co/bigscience/bloom-3b

Table 2: Evaluation of saliency methods on SalChartQA. The best results are shown in bold, second best are underlined. Stars indicate statistical significance for the difference between VisSalFormer and the best baseline model (*: p < .001).

Method	SalChartQA Training	NSS ↑	CC ↑	KL ↓	SIM ↑	AUC ↑
UMSI [21]		0.538	0.196	1.196	0.401	0.687
DVS [49]		0.522	0.168	1.284	0.375	0.687
MD-EAM [65]		0.705	0.245	1.150	0.406	0.706
	✓	1.362	0.450	1.250	0.476	0.769
MD-SEM [22]		0.794	0.323	1.315	0.419	0.706
	✓	1.409	0.565	0.701	0.560	0.807
TranSalNet [45]		0.655	0.254	1.188	0.417	0.688
	✓	1.666	0.606	0.645	0.565	0.830
VisSalFormer (Ours)	✓	**1.782***	**0.674***	**0.532***	**0.615***	**0.839***

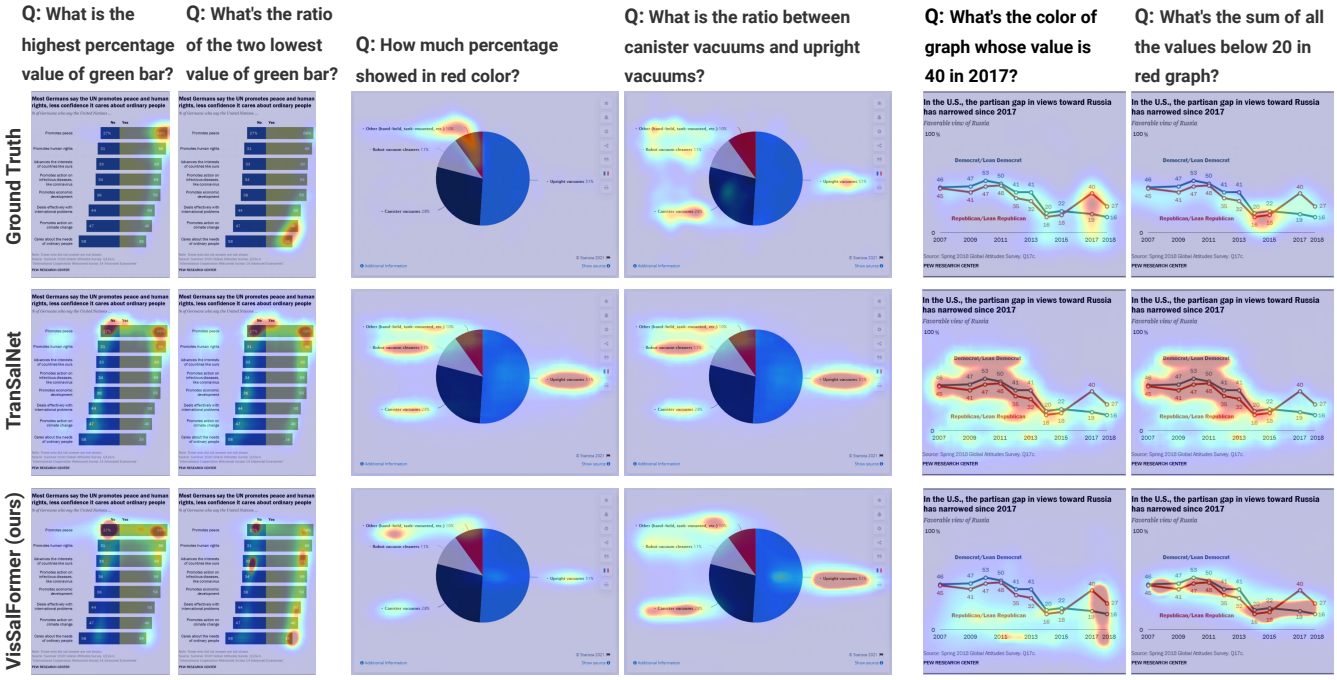


Figure 7: Sample visualisations and questions from the test set with human ground truth saliency maps (top row). Predictions from the strongest baseline method (cf. Table 2 TranSalNet, middle row) and our VisSalFormer (bottom row) are shown below. Our method predicts question-specific saliency maps that more closely resemble the human ground truth.

predict saliency maps intricately tied to the question, i.e. the user's information need. VisSalFormer takes an arbitrary question, then extracts the question embedding for a fused embedding with the visualisation. The question input of VisSalFormer goes beyond previous task-based saliency prediction models ([27, 71]), as previous works are usually limited to a predefined set of tasks or objects. Our experiments showed state-of-the-art performance of our method on SalChartQA, the first question-driven saliency dataset (see Table 2).

The ablation study of replacing Bert with Llama and Bloom yields two insights. Firstly, the question encoders, including Bert and the two large language models [63, 67], exhibit comparable performance in question-driven saliency prediction (see Table 3). Given that Bert has fewer parameters than other large language models (110 million [19] versus 3 billion or more), we conclude that a much bigger large language model than Bert is unnecessary as a question encoder for effective question-driven saliency prediction. Secondly, the last two rows in Table 3 underline the importance of updating the weights in question encoders. When comparing the updating of weights to freezing Bert's weights, the performance of updating weights in Bert surpasses freezing weights across all five saliency metrics. Thus, this paper not only presents a pioneering method but

Table 3: Ablation study for VisSalFormer . The best results are shown in bold, second best are underlined. QE: question encoder, FF: cross-modality feature fusion module. Stars indicate the statistical significance of the difference between the best and second best for each metric (*: p < .001).

Methods	Frozen QE	NSS ↑	CC ↑	KL ↓	SIM ↑	AUC ↑
Swin [43] → Xception [14]		**1.787**	0.643	0.567	0.604	0.831
Swin [43] → ViT [20]		1.718	0.646	0.578	0.588	0.832
Bert [19] → Llama [63]	✓	1.681	0.655	0.569	0.602	0.832
Bert [19] → Bloom [67]	✓	1.685	0.659	0.559	0.603	0.834
w/o question embedding		1.581	0.575	0.656	0.567	0.825
w/o cross-attention in FF		1.681	0.628	0.587	0.592	0.827
w/o cross-attention & self-attention in FF		1.689	0.632	0.590	0.597	0.827
VisSalFormer (ours)	✓	1.681	0.659	0.575	0.602	0.832
VisSalFormer (ours)		1.782	**0.674***	**0.532***	**0.615***	**0.839***

also lays the groundwork for future improvements and applications in question-driven saliency on information visualisation.

7.3 Limitations and Future Work

Our research demonstrated the importance of question-driven saliency and the link between the question or information need and the user's visual attention when visually exploring information visualisations. Our work refrained from studying text saliency since BubbleView was not validated for text reading. While our study has primarily centred on visualisation saliency, an intriguing prospect for future research is to delve into the domain of text saliency. However, the interplay between text and image saliency holds considerable promise [60]. Future research should investigate this dynamic relationship and develop models that can jointly assess the saliency of both textual and visual components.

Integrating question-driven saliency into CQA models for improved performance [41, 59] is one possible future work. By doing so, we anticipate the potential for substantial improvements in their performance and increased compatibility with eXplainable Artificial Intelligence (XAI) systems. Another avenue for exploration is using question accuracy as a metric for assessing visualisation quality. This approach could provide a quantifiable measure of how well a visualisation aligns with the questions it aims to address, ultimately contributing to more informed design choices. For example, Wang et al. [66] demonstrated the potential utility of using question accuracy as a metric to quantify recallability, which could help guide the creation of more memorable visualisations.

8 CONCLUSION

This work addresses a critical gap in understanding the influence of users' information needs on their visual attention when exploring information visualisations. To address the data scarcity issue of question-driven saliency on information visualisations, we proposed a novel large-scale saliency dataset that was collected online using the BubbleView interface and contains 6,000 question-driven saliency maps on 3,000 visualisations. Our analyses on SalChartQA demonstrated that information needs, operationalised through a QA paradigm, and visual saliency are tightly correlated. Using our dataset, we then proposed VisSalFormer – the first computational method for predicting question-driven saliency on information visualisations. Our method outperformed existing state-of-the-art saliency prediction models in the most common saliency metrics. By shedding first light on the interplay between information needs and visual attention, our work provides a new perspective on saliency prediction, a benchmark of question-driven saliency prediction, and informs the future development of new visual analytics applications that are able to take users' interests and needs into account.

ACKNOWLEDGMENTS

Y. Wang was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161. M. Elfares was funded by the Ministry of Science, Research and the Arts Baden-Württemberg in the Artificial Intelligence Software Academy (AISA). 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. A. Bulling was funded by the European Research Council (ERC) under grant agreement 801708.

REFERENCES

[1] Danielle Albers, Michael Correll, and Michael Gleicher. 2014. Task-driven evaluation of aggregation in time series visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 551–560.
[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2425–2433.
[3] Mihai Bâce, Sander Staal, and Andreas Bulling. 2020. Quantification of users' visual attention during everyday mobile device interactions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 1–14.
[4] Tanja Blascheck, Kuno Kurzhals, Michael Raschke, Stefan Strohmaier, Daniel Weiskopf, and Thomas Ertl. 2016. AOI hierarchies for visual exploration of fixation sequences. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*. ACM, 111–118.
[5] Ali Borji, Dicky N Sihite, and Laurent Itti. 2012. Probabilistic learning of task-specific visual attention. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 470–477.
[6] 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.

[7] Neil Bruce and John Tsotsos. 2005. Saliency based on information maximization. *Advances in Neural Information Processing Systems* 18 (2005), 1–8.

[8] Alexandre Bruckert, Lucie Lévêque, Matthieu Perreira Da Silva, and Patrick Le Callet. 2023. A Dataset of Gaze and Mouse Patterns in the Context of Facial Expression Recognition. In *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences*. 157–164.

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

[10] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. 2018. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 41, 3 (2018), 740–757.

[11] Zoya Bylinskii, Nam Wook Kim, Peter O'Donovan, Sami Alsheikh, Spandan Madan, Hanspeter Pfister, Fredo Durand, Bryan Russell, and Aaron Hertzmann. 2017. Learning visual importance for graphic designs and data visualizations. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology (UIST)*. 57–69.

[12] Yeongnam Chae, Mitsuru Nakazawa, and Björn Stenger. 2018. Enhancing Product Images for Click-Through Rate Improvement. In *2018 25th IEEE International Conference on Image Processing (ICIP)*. 1428–1432.

[13] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. 2015. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 3 (2015), 569–582.

[14] François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 1251–1258.

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

[16] Carlos Gomez Cubero and Matthias Rehm. 2021. Intention Recognition in Human Robot Interaction Based on Eye Tracking. In *Proceedings of the International Conference on Human-Computer Interaction (INTERACT)*. Springer, 428–437.

[17] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding* 163 (2017), 90–100.

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 248–255.

[19] 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)*. ACL, 4171–4186.

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*. 1–22.

[21] Camilo Fosco, Vincent Casser, Amish Kumar Bedi, Peter O'Donovan, Aaron Hertzmann, and Zoya Bylinskii. 2020. Predicting Visual Importance Across Graphic Design Types. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology (UIST)*. ACM, 249–260.

[22] 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.

[23] Steven R Gomez, Radu Jianu, Ryan Cabeen, Hua Guo, and David H Laidlaw. 2016. Fauxvea: Crowdsourcing gaze location estimates for visualization analysis tasks. *IEEE Transactions on Visualization and Computer Graphics* 23, 2 (2016), 1042–1055.

[24] Vladimir Guchev, Paolo Buono, and Cristina Gena. 2018. Towards intelligible graph data visualization using circular layout. In *Proceedings of the 2018 International Conference on Advanced Visual Interfaces*. ACM, 1–3.

[25] David J Hauser and Norbert Schwarz. 2016. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior research methods* 48 (2016), 400–407.

[26] 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.

[27] Zhiming Hu, Andreas Bulling, Sheng Li, and Guoping Wang. 2021. FixationNet: Forecasting Eye Fixations in Task-Oriented Virtual Environments. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 27, 5 (2021), 2681–2690.

[28] Zhiming Hu, Sheng Li, Congyi Zhang, Kangrui Yi, Guoping Wang, and Dinesh Manocha. 2020. DGaze: Cnn-based gaze prediction in dynamic scenes. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 26, 5 (2020), 1902–1911.

[29] Laurent Itti. 2004. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE transactions on image processing* 13, 10 (2004), 1304–1318.

[30] 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 (TPAMI)* 20, 11 (1998), 1254–1259.

[31] 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.

[32] Chanho Jung and Changick Kim. 2011. A unified spectral-domain approach for saliency detection and its application to automatic object segmentation. *IEEE Transactions on Image Processing* 21, 3 (2011), 1272–1283.

[33] Nam Wook Kim, Zoya Bylinskii, Michelle A Borkin, Krzysztof Z Gajos, Aude Oliva, Fredo Durand, and Hanspeter Pfister. 2017. BubbleView: an interface for crowdsourcing image importance maps and tracking visual attention. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 5 (2017), 1–40.

[34] 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.

[35] George Alex Koulieris, George Drettakis, Douglas Cunningham, and Katerina Mania. 2016. Gaze prediction using machine learning for dynamic stereo manipulation in games. In *Proceedings of the 2016 IEEE Virtual Reality*. IEEE, 113–120.

[36] Kuno Kurzhals, Brian Fisher, Michael Burch, and Daniel Weiskopf. 2014. Evaluating visual analytics with eye tracking. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*. 61–69.

[37] Sébastien Lallé, Cristina Conati, and Giuseppe Carenini. 2016. Predicting Confusion in Information Visualization from Eye Tracking and Interaction Data.. In *IJCAI*. 2529–2535.

[38] Olivier Le Meur and Thierry Baccino. 2013. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods* 45, 1 (2013), 251–266.

[39] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*. IEEE, 1833–1844.

[40] Max Limper, Arjan Kuijper, and Dieter W Fellner. 2016. Mesh Saliency Analysis via Local Curvature Entropy.. In *Eurographics (Short Papers)*. 13–16.

[41] Yuetan Lin, Zhangyang Pang, Donghui Wang, and Yueting Zhuang. 2017. Task-driven visual saliency and attention-based visual question answering. *arXiv preprint arXiv:1702.06700* (2017), 1–8.

[42] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. 2021. Visual saliency transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*. IEEE, 4722–4732.

[43] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*. IEEE, 10012–10022.

[44] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*. 1–11.

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

[46] Xiaolong Ma, Xudong Xie, Kin-Man Lam, and Yisheng Zhong. 2015. Efficient saliency analysis based on wavelet transform and entropy theory. *Journal of Visual Communication and Image Representation* 30 (2015), 201–207.

[47] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). ACL, Dublin, Ireland, 2263–2279. https://doi.org/10.18653/v1/2022.findings-acl.177

[48] Laura E Matzen, Michael J Haass, Kristin M Divis, and Mallory C Stites. 2017. Patterns of attention: How data visualizations are read. In *International Conference on Augmented Cognition (AC)*. 176–191.

[49] 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 (TVCG)* 24, 1 (2017), 563–573.

[50] Luana Micallef, Gregorio Palmas, Antti Oulasvirta, and Tino Weinkauf. 2017. Towards Perceptual Optimization of the Visual Design of Scatterplots. *IEEE Transactions on Visualization and Computer Graphics* 23, 6 (2017), 1588–1599.

[51] Vidhya Navalpakkam and Laurent Itti. 2005. Modeling the influence of task on attention. *Vision Research* 45, 2 (2005), 205–231.

[52] Nobuyuki Otsu. 1979. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics* 9, 1 (1979), 62–66.

[53] 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.

[54] Vasili Ramanishka, Abir Das, Jianming Zhang, and Kate Saenko. 2017. Top-down visual saliency guided by captions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, 7206–7215.

[55] Bahador Saket, Alex Endert, and Çağatay Demiralp. 2018. Task-based effectiveness of basic visualizations. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 25, 7 (2018), 2505–2512.

[56] Sungbok Shin, Sunghyo Chung, Sanghyun Hong, and Niklas Elmqvist. 2022. A scanner deeply: Predicting gaze heatmaps on visualizations using crowdsourced eye movement data. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 29, 1 (2022), 396–406.

[57] Sicheng Song, Juntong Chen, Chenhui Li, and Changbo Wang. 2023. GVQA: Learning to Answer Questions about Graphs with Visualizations via Knowledge Base. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[58] Sicheng Song, Chenhui Li, Dong Li, Juntong Chen, and Changbo Wang. 2022. GraphDecoder: Recovering Diverse Network Graphs from Visualization Images via Attention-Aware Learning. *IEEE Transactions on Visualization and Computer Graphics* Early Access (2022), 1–17.

[59] Ekta Sood, Fabian Kögel, Philipp Müller, Dominike Thomas, Mihai Bâce, and Andreas Bulling. 2023. Multimodal integration of human-like attention in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2647–2657.

[60] Ekta Sood, Fabian Kögel, Florian Strohm, Prajit Dhar, and Andreas Bulling. 2021. VQA-MHUG: A gaze dataset to study multimodal neural attention in VQA. In *Proceedings of the ACL SIGNLL Conference on Computational Natural Language Learning (CoNLL)*. ACL, 27–43.

[61] Ekta Sood, Lei Shi, Matteo Bortoletto, Yao Wang, Philipp Müller, and Andreas Bulling. 2023. Improving Neural Saliency Prediction with a Cognitive Model of Human Visual Attention. In *Proc. the 45th Annual Meeting of the Cognitive Science Society (CogSci)*. 3639–3646.

[62] 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.

[63] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023), 1–16.

[64] Xiang Wan, Wenqian Wang, Jiming Liu, and Tiejun Tong. 2014. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC medical research methodology* 14 (2014), 1–13.

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

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

[67] BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100* (2022), 1–73.

[68] 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. ACL, 2324–2335.

[69] Pingmei Xu, Yusuke Sugano, and Andreas Bulling. 2016. Spatio-Temporal Modeling and Prediction of Visual Attention in Graphical User Interfaces. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 3299–3310.

[70] 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)*. IEEE, 193–202.

[71] Quanlong Zheng, Jianbo Jiao, Ying Cao, and Rynson WH Lau. 2018. Task-driven webpage saliency. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 287–302.