AttentionLeak: What Does Human Attention Reveal About Information Visualisation?

Malte Sönnichsen(✉), Mayar Elfares, Yao Wang, Ralf Küsters, Alina Roitberg, and Andreas Bulling

University of Stuttgart, Stuttgart, Germany
malte.soennichsen@ki.uni-stuttgart.de

Abstract. In scenarios where direct access to displayed content, such as secured web pages or confidential documents, is restricted, eye-tracking data can serve as a side channel for information inference. Represented as human attention maps, eye tracking data is widely used in research, for example, to quantify how users explore visual information. In this work, we specifically focus on visual question-answering (VQA) scenarios to demonstrate, for the first time, that a rich amount of information can be leaked solely from human attention maps. Hence, we assume that an adversary only has access to the gaze attention maps and aims to derive a range of attributes about the image (e.g. the chart type), the question (e.g. question type), and the answer (e.g. the accuracy-based complexity). This information leakage could be the first step towards potentially more complex insights about human perception and cognition. Our experiments demonstrate that deriving attributes is feasible, and simultaneously predicting multiple attributes improves the success rate for attributes that are difficult to infer. This paper highlights potential threats, encouraging the community to address these concerns and develop appropriate privacy-preserving solutions.

Keywords: VQA · Side-Channel Attack · Multitask Learning · Transformer · Privacy

1 Introduction

Attention maps – 2D maps that encode human gaze data – have become an indispensable tool in eye-tracking research, particularly in understanding how users engage with information visualisations within documents [3,17,26]. By identifying and highlighting areas of visualisation that attract attention, gaze attention maps help researchers and designers determine what aspects of visual content are most salient to the human visual system [38,47]. Particularly in visual question-answering scenarios, gaze attention maps bridge the gap between perception and

M. Sönnichsen and M. Elfares—Both authors contributed equally to this research.

X.-C. Yin et al. (Eds.): ICDAR 2025, LNCS 16026, pp. 77–95, 2026.
https://doi.org/10.1007/978-3-032-04627-7_5

cognition by showing how visual attention is distributed in response to a question, thus revealing insights into how individuals prioritize visual information, manage cognitive load, and integrate perceptual features to infer answers [43,50].

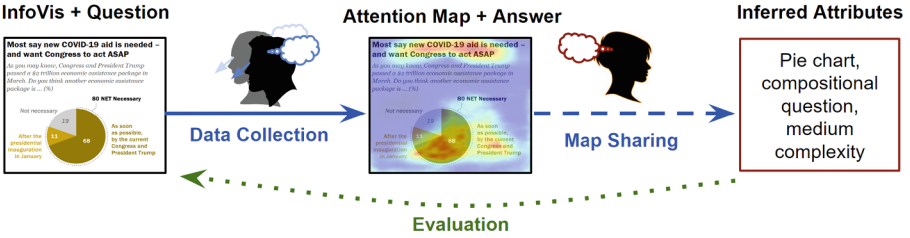

Fig. 1. Human attention maps can unintentionally reveal sensitive information in visual question-answering (VQA) scenarios where users' gaze patterns are collected and aggregated, and then attention maps are publicly shared. We demonstrate, for the first time, that adversaries with access only to gaze attention maps can infer various attributes related to the chart, the question, and the answer.

Even when the content or stimuli presented to a user, such as a secured webpage or protected document, cannot be directly accessed by an attacker, eye-tracking data can serve as a side-channel to infer what the user is viewing [46,48]. This, therefore, raises significant privacy concerns:

First, the visualisation image can be privacy-sensitive in several scenarios, particularly when they involve personal, confidential, or sensitive data [55,57]. For example (c.f. Fig. 1), a dashboard visualising patient health records, such as medical histories, diagnoses, or medication usage, can reveal sensitive health information, or a graph visualising income distribution or spending patterns in a community could expose individuals' financial data, including salaries, debts, or investment information. The visualisation image can also reveal information about the users' visual perception, which forms the foundation of further cognitive processing since human attention is not only naturally drawn to visually salient parts of an image but can also be modulated by the task at hand, where the question guides the focus to specific parts of the image that are relevant for answering [43,50].

Second, the question asked to the users is subsequently crucial and potentially privacy-sensitive since it can reveal information about the visualisations, such as the topic or some visual references [31]. Additionally, it acts as a cognitive guide, shaping where a person looks (or should look) in the image [36]. This is an example of top-down attention, where higher cognitive functions (the question) influence visual perception [53]. In addition, the question helps filter out unnecessary information by narrowing the search space within the image [16]. The cognitive system uses the question to determine which visual features and regions are worth attending to. Human cognition integrates multiple aspects of the visual scene (e.g., colours, shapes, sizes, and spatial relationships) to extract

meaningful information that answers the question. This integration relies on working memory and executive function, which keep track of relevant visual elements while the question is being processed cognitively [43]. Therefore, information about the question can leak information about the user's perceptual and cognitive processes.

Third, the answer can also raise privacy concerns since it might reveal information about the users' prior knowledge, memory recall, and semantic understanding [32,43]. Moreover, the complexity of the question can increase cognitive load, affecting both the processing time and accuracy of the answer [50].

Overall, VQA serves as an effective domain for handling multiple modalities, particularly images (visual representations) and text (questions and answers), relying on various human cognitive and perceptual factors.

In this paper, we aim to shed light on the amount of information that could be inferred from gaze attention maps solely, beyond the simple gaze location estimation [46,48]. More technically, we present AttentionLeak – an adversarial attack that only leverages the gaze attention maps to derive privacy-sensitive information in the form of a range of attributes about the image (e.g. the chart type), the question (e.g. question type), and the answer (e.g. the answer complexity). This could act as the first step towards potentially more complex insights about human perception and cognition[1]. In summary, our paper makes the following contributions:

- Our work is the first to demonstrate the potential of gaining information from gaze attention maps alone.
- We show that we can glean insights about the image, question, and answer solely from the gaze attention maps in VQA scenarios.
- Through extensive experiments[2], we demonstrate that an adversary, even with limited resources in terms of data and computing power, can successfully perform the attack.
- We further recommend the most suitable model architectures, optimizers, and augmentation techniques for our attack according to the characteristics of attention maps.

2 Related Work

In this section, we introduce the key related work on how an adversary can derive private information through inference attacks, what this information can reveal about the users and the corresponding information visualisation, and why privacy-preserving techniques are, therefore, crucial.

[1] Note that even if the data is aggregated (the common practice in gaze attention maps [50]), patterns could still re-identify individuals [23,34], leaking further information. However, this remains out of the scope of this paper.

[2] The implementation code will be publicly available upon acceptance.

2.1 How Can an Adversary Infer Private Information?

Inference attacks are privacy-violating strategies that aim to extract sensitive or private information from seemingly benign or aggregated data. They have become an increasingly important area of research as the reliance on data sharing and machine learning systems grows. The pioneering work of Denning et al. [11] on statistical database security showed that adversaries could aggregate query responses to piece together sensitive information, even when the data is aggregated, such as in the case of attention maps. Later works [23,34] demonstrated that databases could be de-anonymised through cross-referencing with auxiliary information (i.e. shadow models) and individuals can be re-identified. This concern further grows in our case since eye data includes identifiers [8] and quasi-identifiers (e.g. gender [39]).

Inference attacks include: (i) membership inference attacks, introduced by Shokri et al. (2017) [42], that allow adversaries to determine whether a particular data point was included in a model's training set, (ii) model inversion attacks, demonstrated by Fredrikson et al. (2015) [15], that enables attackers to reconstruct sensitive input data from model outputs, and (iii) attribute inference attacks where adversaries leverage access to trained models to infer missing dataset attributes. In this paper, we focus on attribute inference attacks since they exploit indirect information, which users might not even realize could reveal such insights.

Melis et al. (2019) [33] demonstrated the feasibility of such attacks in the context of collaborative learning, where participants may unknowingly expose sensitive information about their data to other participants. Recently, the work by Zhang et al. (2018) [55] revealed that attention maps generated by deep learning models, commonly used for interpretation, can unintentionally reveal sensitive patterns in the data, allowing attackers to infer underlying private information even without access to the original dataset. In contrast, we focus on human attention maps generated through gaze data.

2.2 What Does the Inferred Information Reveal About the Users and the Corresponding Information Visualisation?

Numerous studies in eye tracking research and cognitive science have revealed that human eye movements can provide insights into a user's mental state [6,7], and this has inspired a growing number of research in eye-based user modelling [21,37,47]. Previous works have also estimated participants' levels of text comprehension [1], intention [27,40,56], mind-wandering tendencies [22,54], and recallability [49] 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 [2,5,20,21].

More specifically, in information visualisations, several eye-tracking datasets have been collected by researchers to understand human visual attention for bottom-up [3,41] as well as top-down attention [17,26,38,50]. In this paper, we

focus on the most recent dataset by Wang et al. [50] that used the BubbleView technique [24] to collect SalChartQA, a large-scale question-driven dataset comprising 6,000 attention maps under analytical questions.

2.3 How Does Awareness of Attack Feasibility Aid in Mitigation Efforts?

Inspired by the famous side-channel attack on Apple passwords through gaze reflection [48], privacy-preserving eye tracking [4,9,28,29] started to attract attention but remains under-investigated [18]. In particular, privacy threats are not yet well-understood, and the community remains unaware of the potential risks. Nonetheless, it remains necessary to share eye tracking data in order to cover the large variety in eye data and infer insights or train ML models [13,14]. Hence, the associated privacy risks, the possible misuse of data copies, and the potential for personal information leakage, e.g. inference attacks, increase.

Hence, building on the above-mentioned works, our work investigates the feasibility of attribute inference attacks on information visualisations.

3 Attack Methodology

In this section, we present our threat model and assumption about the adversary's capabilities. Then, we introduce the dataset used and the related challenges. Finally, we present AttentionLeak, our attribute inference attack on attention maps.

3.1 Threat Model

Adversary's Goal. The main goal of the adversary is to infer attributes about the inputs (i.e. the visualisation images) given the gaze attention maps. In general, the attack can be generalised to *any* visual attention scenario. However, in this paper, we focus on data visualisation since it incorporates information about different fields. We further focus on visual question-and-answer (VQA) scenarios since they are information-rich and can reveal insights about the user perception (i.e. information about the charts) as well as the user cognition (i.e. the corresponding questions and answers).

Adversary's Knowledge. We assume that the adversary has access to the public/leaked/inferred (c.f. Sect. 2) attention maps and does not have access to the private stimuli (i.e. visualisation images, questions, and answers). Nonetheless, the adversary can create shadow datasets by collecting publicly available stimuli with her selected attributes and mimicking the gaze attention maps.

Adversary's Strategy. The adversary starts by compiling the shadow dataset from public knowledge. She then trains a model that takes the shadow gaze attention maps as input and outputs the corresponding attribute(s). The predicted attribute(s) is compared against the shadow ground-truth attribute(s). In this paper, we investigate two main types of models: single- and multi-class classification models with different architectures. Once the model is trained, the adversary uses the model in inference mode to reveal information about the victim dataset (i.e. the dataset to be attacked/targeted).

3.2 The Information Visualisation Dataset

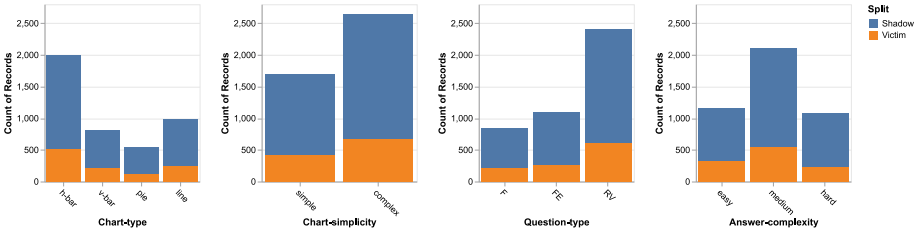

Fig. 2. Class distribution of the four tasks. Chart-type: horizontal bar chart (h-bar), vertical bar chart (v-bar), pie chart, and line chart. Chart-simplicity: simple and complex. Question-type: filtering (F), finding extremum (FE), and retrieving values (RV). Answer-complexity: easy, medium, and hard.

We target the SalChartQA dataset [50]. The dataset consists of 3,000 charts, with 2 questions per chart, resulting in 6,000 question-driven attention maps. The mean participant number is 13.1 (with a minimum of 10 participants) per question. For each chart, participants' attention maps are aggregated with a Gaussian filter with a 1-degree visual angle. We split the dataset into two subsets (shadow and target dataset) with an 8:2 ratio. The shadow dataset is further split into train and validation sets with a 3:1 ratio. We ensure that each chart and participant occurs only in one subset to prevent information leakage from the same charts or participants into the different subsets. Therefore, we modelled the participant, question, image, and relationship as a bipartite graph. We then calculated the connected components in this graph, where each node represents either a participant or a chart, and the edge represents the question. In other words, a participant node and a chart node are connected if the participant answers a question related to that chart. Then, we ensured that an entire component only occurs either in the shadow dataset or in the target dataset.

Targetted Attributes. In our attack construction, the adversary targets four key attributes of SalChartQA to infer information about the chart, question, and answer. Note that the attribute distribution is highly imbalanced in the dataset, as shown in Fig. 2:

- **Chart-type**: the dataset consists of three commonly used chart types: bar, line, and pie charts with an approximate ratio of 4:1:1, respectively.
- **Chart-simplicity**: the charts are categorised into *simple* and *complex* according to the visual complexity of the image, e.g. the number of columns or the existence of stacked or grouped bars. The simple-to-complex ratio is 8:3 for bar charts and 6:4 for line charts. All pie charts are classified as *simple*.
- **Question-type**: Each chart image includes two questions of the following types: (i) compositional questions that contain mathematical/logical operations such as sum, difference or average, (ii) visual questions that refer to the image visual attributes such as colour, length/height of graphical marks, and (iii) data retrieval questions. Namely, we use the top-3 most occurring questions: filtering (F), finding extremum (FE) and retrieving values (RV).
- **Answer-complexity**: To illustrate that an adversary can infer additional, unintended attributes beyond those initially available in the dataset, we compute the *answer-complexity* attribute. *Answer-complexity* is derived from the number of clicks performed by users to answer a question. This metric, therefore, outlines the complexity of answering a question. Hence, we calculate the 0.25- and 0.75-quantile of the number of clicks per image type and assign the saliency maps with the number of clicks below the 0.25- quantile to the complexity *easy* between the quantiles to *medium* and the ones over the 0.75-quantile to *hard*.

Attention Maps. The SalChartQA dataset includes three types of attention maps: aggregation of all answers, correct answers only, and incorrect answers only. For our experiments, we mostly focus on the aggregation of correct answers because it helps to separate genuine visual cues from biases that might arise due to question phrasing, misinterpretation, or other cognitive and linguistic factors. By analyzing only the attention maps from correct answers, we can identify unbiased, task-relevant attention patterns that contribute to accurate responses to improve the reliability and fairness of our findings across diverse users and contexts.

3.3 The AttentionLeak Attack

Using the shadow dataset, the adversary maps each attribute inference attack to an image single- or multi-classification task, taking the attention map as input and predicting the respective attribute class(es). The adversary can further employ different types of model architectures depending on the access to resources (e.g. GPUs and datasets). We demonstrate the feasibility of the attack through three different model types: (i) a convolutional neural network (CNN), e.g. Resnet101 [19], (ii) a vision transformer (ViT), e.g. ViT-b/16 [12], and (iii) a foundation model, e.g. Dino v2 [35]:

Convolutional Neural Network (CNN). ResNet-101 [19] is a deep convolutional neural network (CNN) that embeds strong inductive biases about visual data

directly into its architecture. Through its convolutional operations, it processes images in a way that inherently accounts for the locality and hierarchical nature of visual information – nearby pixels are more likely to be related than distant ones and visual features build up from simple to complex. This architectural bias, combined with its local receptive fields, makes ResNet-101 naturally data-efficient for image processing tasks. The network consists of 101 layers organized into residual blocks, where each block can learn additional features while preserving already learned information through skip connections. This design allows for very deep networks while maintaining stable training dynamics, making ResNet-101 particularly effective even with moderate-sized datasets.

Vision Transformer (ViT). Vision Transformer (ViT) [12] represents a departure from traditional computer vision architectures by having minimal built-in assumptions about image structure. Unlike CNNs, ViT treats images as sequences of patches and relies on self-attention mechanisms to learn relationships between these patches from scratch. The "B/16" variant processes images by dividing them into 16×16 pixel patches. These patches are linearly embedded and combined with position embeddings before being fed into a series of transformer encoder blocks [45]. While this architecture is extremely flexible and can theoretically learn any kind of spatial relationship in the data, this flexibility comes at a cost of data efficiency. The model must learn these visual relationships from the data itself, rather than having them built into its architecture. This explains why ViT models often underperform CNNs like ResNet101 on smaller datasets, where the benefits of their flexibility cannot overcome the advantage of CNNs' built-in inductive biases.

Probed Foundation Model. Nowadays, it is becoming more and more common to work with embeddings of large-scale pre-trained foundation models and apply classical classification algorithms on top of these embeddings. We calculate embeddings of all attention maps using the image feature extraction pipeline from Huggingface [52] with Dino v2 [35]. These embeddings are then classified for the respective attributes using Logistic Regression or Random Forest.

Single- and Multi-task Training. The adversary inputs the attention maps to the attack model after rescaling them to a maximum height/width of 224, without changing the aspect ratio (no normalisation is computed). The model is trained using a cross-entropy loss. In the single-task training, the adversary trains the attack model on one of the four tasks and outputs one class.

$$\sum_{i=1}^{\#\text{tasks}} \left(\frac{1}{\sigma_i^2} \mathcal{L}_i + \log \sigma_i \right),$$

where $\mathcal{L}_i$ is the respective cross entropy loss of task i and $\log \sigma_i$ is learned. This method is effective, simple to implement, and keeps the number of hyper-parameters low.

4 Evaluation

In this section, we show, through extensive experiments, (i) if it is possible to infer private information solely from the attention maps, (ii) the effect of multi- and single-task attacks, (iii) the shadow model variations that an adversary can use (e.g. architectures and hyperparameters), (iv) the quantitative and qualitative information that an adversary can gain.

4.1 Implementation Details

We use ResNet-101 and ViT-B/16 models from the pytorch-image-models library [51], pretrained on the ImageNet dataset [10]. We trained all models on a single NVIDIA A100-40 GPU.

For evaluation, we use both micro- and macro-average accuracy to better assess the performance on majority as well as minority classes:

$$\text{Micro-Averaging Accuracy (Acc)} = \frac{\sum_{i=1}^{C}(\text{TP}_i + \text{TN}_i)}{\sum_{i=1}^{C}(\text{TP}_i + \text{TN}_i + \text{FP}_i + \text{FN}_i)} \quad (1)$$

$$\text{Macro-Averaging Accuracy (Macro Acc)} = \frac{1}{C}\sum_{i=1}^{C}\frac{\text{TP}_i + \text{TN}_i}{\text{TP}_i + \text{TN}_i + \text{FP}_i + \text{FN}_i} \quad (2)$$

where C denotes the number of classes.

We use two baselines for comparison: (i) random guessing and (ii) predicting the majority class. In the case of random guessing, the micro- and macro-accuracy is $\frac{1}{C}$. For predicting the majority class, the macro-accuracy is $\frac{1}{C}$ and the micro-accuracy depends on how imbalanced the label distribution is. The more imbalanced, the higher the micro-accuracy.

We conducted extensive hyperparameter optimization across five distinct classification tasks: multi-task, chart-type, chart-simplicity, question-type, and answer-complexity classification. For each task, we optimized for maximum macro accuracy on the validation set. The search space was consistent across all tasks and included three optimizer variants (Adam [25], AdamW [30], and SGD with Nesterov momentum [44]), with learning rates ranging from $1e-4$ to $5e-1$. We also explored the impact of data augmentation and sampling strategies (random vs. balanced) on model performance.

4.2 Is It Possible to Infer Chart-, Question-, and Answer-Related Attributes from Attention Maps Alone?

Our main goal is to evaluate the feasibility of inferring chart-, answer- and question-related attributes solely from gaze attention maps. Table 1 compares various models, including random and majority baselines, CNNs, ViTs, and probed foundation models, in predicting attributes such as *chart-type*, *chart-simplicity*, *question-tpe*, and *answer-complexity*. All models substantially outperform random chance and majority baselines, with fine-tuned neural architectures achieving notably higher accuracy. Fully fine-tuned models such as ResNet-101 and ViT-B/16 demonstrated strong performance, particularly in detecting

chart-type and *chart-simplicity*, achieving a macro accuracy of up to 89.59% and 88.32%, respectively. These results indicate that human attention maps contain a significant amount of information that can reveal private details about the chart (e.g. chart-type and chart-simplicity), the question (e.g. question-type), and the answer (e.g. answer-complexity), underscoring potential privacy risks of attention maps in gaze-based applications and side channel attacks.

Table 1. Different neural architectures used for inferring attributes from attention maps alone. All models outperform the baselines, highlighting the privacy risks.

Model	Fully Fine-tuned	Chart-type		Chart-simplicity		Question-type		Answer-complexity	
		Macro Acc	Acc	Macro Acc	Acc	Macro Acc	Acc	Macro Acc	Acc
Baselines									
Majority	-	25.00	47.61	50.00	60.20	33.33	55.96	33.33	50.63
Random	-	25.00	25.00	50.00	50.00	33.33	33.33	33.33	33.33
CNNs									
ResNet-101	✓	**89.59**	**89.42**	**88.32**	**89.17**	**50.79**	52.54	**64.52**	59.78
ViTs									
ViT-B/16	✓	87.69	89.08	86.48	86.73	47.70	56.51	63.88	**61.13**
Probed Foundation Models									
Dino v2 + Random Forest	×	69.67	75.65	78.88	80.19	39.70	**58.63**	51.36	58.02
Dino v2 + Logistic Regression	×	80.30	81.52	80.54	81.78	45.30	53.74	53.04	54.07

Interestingly, despite the advancements of vision transformers in computer vision, our findings show that CNNs outperform ViTs mixed ViT vs. CNN result. While ViTs B/16 has a roughly double parameter count, ResNet-101 outperforms the ViT model on three out of the four attributes tested—namely, *chart-type*, *chart-simplicity*, and *question-type*. Conversely, ViT achieves (an insignificant) higher accuracy on *answer-complexity*. One possible explanation is the nature of the pre-training data. Both ViTs and CNNs, typically undergo pre-training on large image datasets that may not align well with the unique characteristics of gaze attention maps. As a result, features from these large models may not transfer as effectively to gaze data, potentially leading to overfitting. While ResNet-101 is also pre-trained on image data, it has fewer parameters, potentially leading to less overfitting when faced with this unfamiliar data. In addition, gaze attention maps typically have strong local spatial correlations (e.g., concentrated fixations or heatmaps), which CNNs, like ResNet-101, are designed to exploit through hierarchical feature extraction. ResNet uses convolutional filters to capture local textures and structures efficiently, whereas ViTs rely on global self-attention, which may struggle with localized patterns. Moreover, ResNet-101 has strong inductive biases (translation invariance, local receptive fields), making it better suited for structured, spatially dependent data like attention maps. ViTs rely on self-attention mechanisms without built-in spatial biases, requiring large-scale training to learn such relationships effectively. Similarly, the foundation models (Dino v2, also ViT-based) are pre-trained on very large datasets in a self-supervised manner, and instead of being fully fine-tuned, the models remain frozen with only a classifier applied to their extracted features.

Table 2. Multi-Task (MT) vs Single-Task (ST) training regime for inferring attributes from attention maps. Performance deltas show that MT training particularly improves question-type classification, which was the primary objective for model selection on the validation set, while showing varying effects on other attributes.

Attribute	Model	Macro Acc. MT	Macro Acc. ST	Δ Macro Acc.	Acc. MT	Acc. ST	Δ Acc.
Chart-type	ResNet-101	87.80	89.59	−1.79	89.01	89.42	−0.41
Chart-type	ViT-B/16	80.76	87.69	−6.93	83.38	89.08	−5.70
Chart-simplicity	ResNet-101	89.97	88.32	+1.65	91.04	89.17	+1.87
Chart-simplicity	ViT-B/16	86.24	86.48	−0.24	87.17	86.73	+0.44
Question-type	ResNet-101	53.18	50.79	+2.39	60.48	52.54	+7.94
Question-type	ViT-B/16	52.45	47.70	+4.75	54.85	56.51	−1.66
Answer-complexity	ResNet-101	59.84	64.52	−4.68	62.51	59.78	+2.73
Answer-complexity	ViT-B/16	60.40	63.88	−3.48	62.33	61.13	+1.20

Looking at different attributes, distinguishing *chart-type* and *chart-simplicity* appears to be relatively simple, with ResNet-101 achieving 89.59% and 88.32% macro accuracy. In contrast, inferring *question-type* and *answer-complexity* are more challenging, with the best macro accuracy score of 50.79% and 64.52. This is not surprising since *chart-type* and *chart-simplicity* are visually-driven attributes while identifying cognitively-driven attributes, such as the type of question and the answer that users attempt to solve, might require very subtle eye gaze cues.

Overall, these findings suggest that gaze data alone can effectively reveal chart-, question-, and answer-related attributes, posing privacy concerns. We show that even if an adversary has limited access to resources (e.g. GPUs or shadow data), she can effectively infer attributes through lower-parameter architectures (e.g. CNNs) as this task and modality are relatively data-scarce.

4.3 Single vs. Multi-task Models for Inferring Multiple Attributes from Attention Maps

In addition to specific targetted attributes, an adversary might be interested in inferring multiple sensitive attributes simultaneously. We, therefore, further investigate the effectiveness of Multi-Task (MT) versus Single-Task (ST) training for inferring multiple chart-, question-, and answer-related attributes from attention maps.

As shown in Table 2 and Fig. 3, ST training performs best in visually-driven attributes such as *chart-type* and *answer-complexity* while MT performs better for the other attributes. These findings are due to the fact that ST excels in tasks requiring in-depth modelling of nuanced features or patterns by allocating all resources to capturing these intricate details.

Furthermore, our analysis of the confusion matrices, shown in Fig. 4, reveals biases toward over-represented classes, particularly for the *question-type* attribute (*FE* class) and the *answer-complexity* attribute (*medium* class), with a notable amount of false positives. This is mostly because, in attention-based models (e.g., Vision Transformers, attention layers in CNNs), these classes

receive higher attention weights, making their activation maps more pronounced and consistent, exhibiting highly predictable attention distributions. Nevertheless, these results highlight that, despite some biases, the models can accurately infer sensitive information from attention maps, often revealing details about the chart, question, and answer.

Multi-task (MT) training demonstrates notable improvements in *question-type* classification, the most challenging attribute. While *question-type* accuracy remains lower compared to other attributes, both architectures show substantial gains under MT training, with macro accuracy improvements of +2.39% and +4.75% for ResNet-101 and ViT-B/16 respectively, indicating that the shared representations learned through MT training effectively address our primary objective of enhancing *question-type* inference.

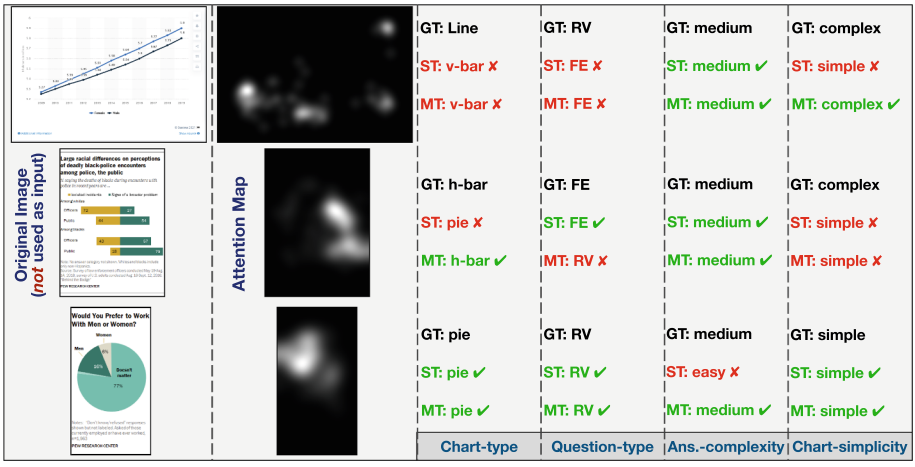

Fig. 3. Qualitative results for Multi-Task (MT) vs Single-Task (ST) training.

4.4 Optimizer Effect on the Attack Success

We demonstrate the effect of the different optimization techniques on the attack success by pairing the ViT and CNN models with Adam, AdamW, and Stochastic Gradient Descent (SGD) optimizers for all four attributes. Figure 5 shows that SGD optimization performs particularly well with the ViT model. Since the ViT model relies on self-attention mechanisms that capture long-range dependencies and global contextual information in images, ViT models are therefore, sensitive to weight updates that align well with global features. SGD, with its steady convergence, often works well for ViTs, as it avoids the risk of overfitting and allows for stable learning of these global patterns across layers.

For CNN-based models, the results are more mixed, with some tasks benefiting from AdamW and others from SGD or Adam. This is due to the fact

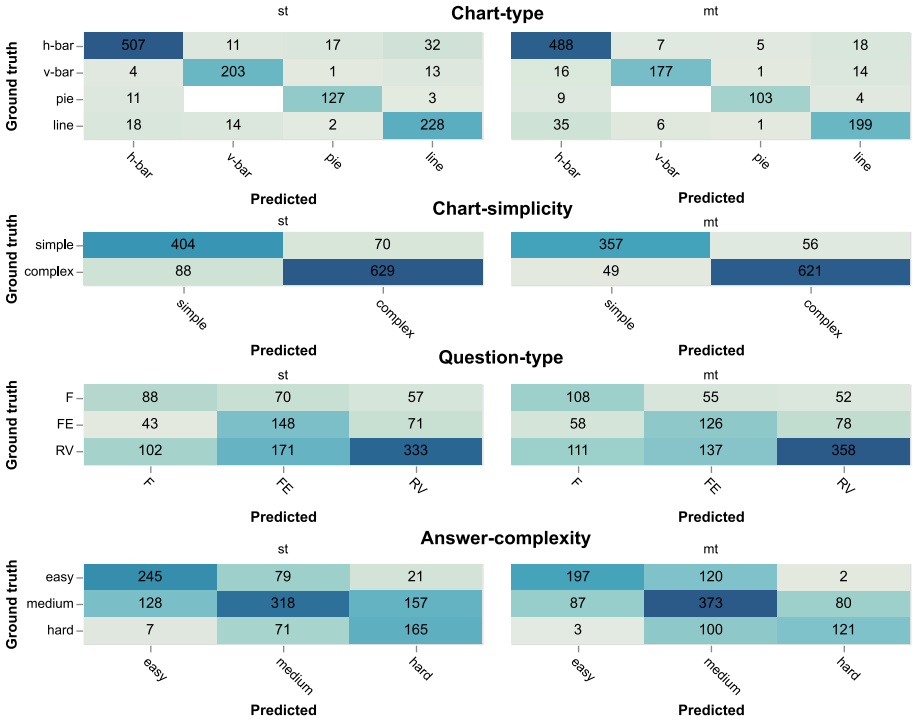

Fig. 4. Confusion matrices results on the SalChartQA dataset.

that CNNs focus on local features through convolutional layers, which capture spatial hierarchies by learning increasingly complex feature maps from layer to layer. The features are often localized, meaning that different parts of the model capture different aspects of the data. This architectural difference creates diverse learning dynamics across layers, leading to mixed results with different optimizers depending on the task.

4.5 Data Augmentation Effect on the Attack Success

Depending on the adversary's access to shadow data, she might exploit data augmentation techniques to compensate for limited data access. Hence, we investigate the effect of training with and without data augmentation on the attack success. For non-augmented input, the longer edge of the image is rescaled to 224 pixels, the shorter edge is padded equally on both sides to 224 pixels. This ensures that the entire attention map, including borders, is always present. This procedure is done in both training on the shadow dataset and attacking the victim dataset. With augmentation, we first randomly flip the image over the horizontal axis. Then, we randomly crop out a patch with the scale of 90% to 110% of the original image and an aspect ratio between 0.9 and 1.1. Scaling a

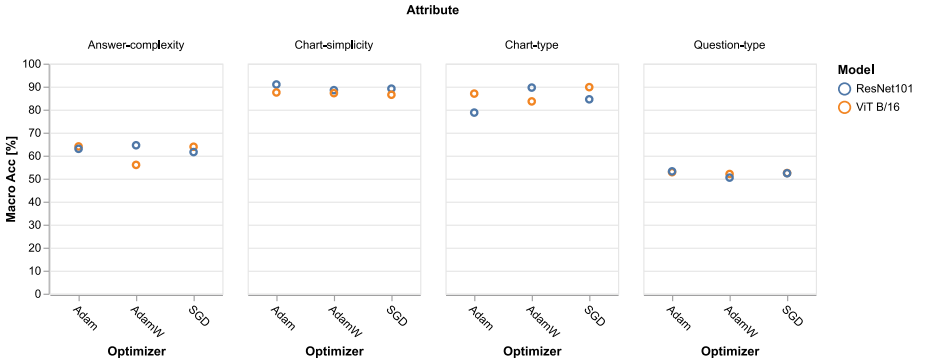

Fig. 5. The effect of the optimiser choice on the attack success. Each dot is the respective best-performing attack model measured by macro accuracy.

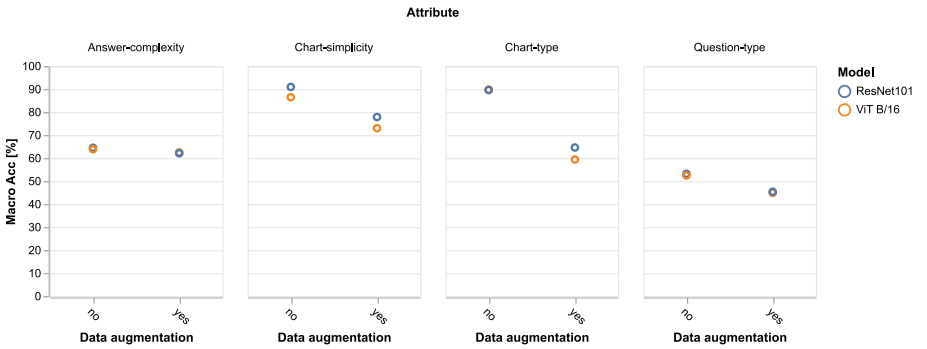

Fig. 6. The effect of data augmentation on the attack success. Even minimal and carefully chosen data augmentation techniques (e.g. horizontal flipping, scaling and changes in aspect ratio) led to decreased classification performance, suggesting that preserving the exact visual characteristics of the attention maps is crucial for a successful attack.

plot and the attention map or changing the aspect ratio, i.e., stretching or compressing the image, does not change any information present in both images.

Despite the relatively simple image transformations, we observe a consistent performance drop across all attributes when using augmentation (Fig. 6), showcasing that typical augmentations used in visual recognition might not be suitable for attention maps. This effect is particularly pronounced for *chart-type*, where accuracy is ∼30% higher without augmentation. For *question-type*, accuracy improves by around ∼5% without augmentation. These findings suggest that standard transformations like shifting and cropping, which are usually effective in computer vision, disrupt the spatial and relational integrity crucial for gaze data and special data augmentation methods need to be developed for this modality. In other words, an adversary requires a real-world shadow dataset or a specially-designed augmentation technique for attention maps.

5 Discussion

We demonstrate that gaze-based attention maps alone can effectively reveal sensitive information about the underlying chart, question, and answer in information visualisations, raising potential privacy concerns. Despite no direct access to the chart, our results indicate that attention maps can expose both visual content (e.g. chart-type) and aspects of user intent (e.g. answer-complexity).

Our results show that an adversary with limited compute power can still be successful through (i) lower-parameter models like CNNs (e.g., ResNet-101) may outperform larger models like ViTs, presumably due to the data scarcity of the gaze modality, or (ii) training only one model since multi-task (MT) training improves accuracy for cognitively-driven attributes, like *question-type*. Nonetheless, we show that standard augmentation techniques used in visual recognition (e.g., cropping, flipping) are not suitable for attention maps and, therefore, do not solve the limited access to data issue.

Limitations and Future Work. We mainly focused on demonstrating the feasibility of attribute inference attacks on information visualisations. Nonetheless, further attributes could be inferred about (i) the information visualisation, such as the linguistics of the question and answer (e.g. number of characters and the visual references), and the topics (e.g. politics, economy, health, and society), as well as (ii) the users in the data such as their cognitive states and attention models.

Privacy and Ethics Statement. Demonstrating the feasibility and simplicity of these attacks is critical for raising awareness within the community about the potential privacy and ethical risks associated with human attention data. By showcasing the vulnerabilities, we underscore the need for robust safeguards to prevent the unintentional leakage of sensitive information, foster responsible development, and motivate the creation of privacy-preserving solutions. Without a clear understanding of the risks, researchers and developers may inadvertently overlook the ethical implications, leaving systems exposed to exploitation and users' data privacy at risk.

6 Conclusion

For the first time, we were able to demonstrate the feasibility of gaining information from gaze attention maps alone. We further show that an adversary, even with limited resources in terms of data and compute power, is able to retrieve the private information encoded and glean insights about the chart, question, and answer in information visualisations. Our work, therefore, highlights these potential threats to increase awareness and encourage the community to develop appropriate privacy-preserving solutions.

Acknowledgments. M. Elfares was funded by the Ministry of Science, Research and the Arts Baden-Württemberg in the Artificial Intelligence Software Academy (AISA).

M. Sönnichsen and A. Roitberg were funded by the Baden-Württemberg Stiftung (Elite Postdoc Program, PRIDRIVE project). Y. Wang was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161. R. Küsters was supported by the German Federal Ministry of Education and Research under Grant Agreement No. 16KIS1441 (CRYPTECS project) and the German Research Foundation under Grant No. 548713845. The authors gratefully acknowledge the computing time provided on the high-performance computer HoreKa by the National High-Performance Computing Center at KIT (NHR@KIT). HoreKa is partly funded by the German Research Foundation (DFG).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ahn, S., Kelton, C., Balasubramanian, A., Zelinsky, G.: Towards predicting reading comprehension from gaze behavior. In: Proceedings of the ACM International Symposium on Eye Tracking Research and Applications (ETRA), pp. 1–5 (2020)
2. Boisvert, J.F., Bruce, N.D.: Predicting task from eye movements: on the importance of spatial distribution, dynamics, and image features. Neurocomputing **207**, 653–668 (2016)
3. Borkin, M.A., et al.: Beyond memorability: visualization recognition and recall. IEEE Trans. Vis. Comput. Graph. (TVCG) **22**(1), 519–528 (2015)
4. Bozkir, E., Günlü, O., Fuhl, W., Schaefer, R.F., Kasneci, E.: Differential privacy for eye tracking with temporal correlations. PLoS ONE **16**(8), e0255979 (2021)
5. Braunagel, C., Geisler, D., Rosenstiel, W., Kasneci, E.: Online recognition of driver-activity based on visual scanpath classification. IEEE Intell. Transp. Syst. Mag. (ITS) **9**, 23–36 (2017)
6. Bulling, A., Roggen, D.: Recognition of visual memory recall processes using eye movement analysis. In: Proceedings of the 13th International Conference on Ubiquitous Computing (UbiComp), pp. 455–464 (2011)
7. Bulling, A., Zander, T.O.: Cognition-aware computing. In: Proceedings of the International Conference on Pervasive Computing (Pervasive), vol. 13, no. 3, pp. 80–83 (2014)
8. Cantoni, V., Galdi, C., Nappi, M., Porta, M., Riccio, D.: Gant: gaze analysis technique for human identification. Pattern Recogn. **48**(4), 1027–1038 (2015)
9. David-John, B., Butler, K., Jain, E.: Privacy-preserving datasets of eye-tracking samples with applications in XR. IEEE TVCG **29**(5), 2774–2784 (2023). https://doi.org/10.1109/TVCG.2023.3247048
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848
11. Denning, D.E.: Secure statistical databases with random sample queries. ACM Trans. Database Syst. (TODS) **5**(3), 291–315 (1980)
12. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale (2020)
13. Elfares, M., Hu, Z., Reisert, P., Bulling, A., Küsters, R.: Federated learning for appearance-based gaze estimation in the wild. NeurIPS-GMML (2022). https://doi.org/10.48550/arXiv.2211.07330

14. Elfares, M., Reisert, P., Tang, W., Hu, Z., Küsters, R., Bulling, A.: Privateyes: appearance-based gaze estimation using federated secure multi-party computation. In: ACM Symposium on Eye Tracking Research & Applications (ETRA) (2024). https://doi.org/10.48550/arXiv.2402.18970
15. Fredrikson, M., Jha, S., Ristenpart, T.: Model inversion attacks that exploit confidence information and basic countermeasures. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pp. 1322–1333 (2015)
16. Gao, L., Cao, L., Xu, X., Shao, J., Song, J.: Question-led object attention for visual question answering. Neurocomputing **391**, 227–233 (2020)
17. Gomez, S.R., Jianu, R., Cabeen, R., Guo, H., Laidlaw, D.H.: Fauxvea: crowdsourcing gaze location estimates for visualization analysis tasks. IEEE Trans. Visual Comput. Graphics **23**(2), 1042–1055 (2016)
18. Gressel, C., et al.: Privacy-aware eye tracking: challenges and future directions. IEEE Pervasive Comput. **22**(1), 95–102 (2023). https://doi.org/10.1109/MPRV.2022.3228660
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR abs/1512.03385 (2015). http://arxiv.org/abs/1512.03385
20. Hild, J., Voit, M., Kühnle, C., Beyerer, J.: 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), pp. 1–5 (2018)
21. Hu, Z., Bulling, A., Li, S., Wang, G.: EHTask: recognizing user tasks from eye and head movements in immersive virtual reality. IEEE Trans. Vis. Comput. Graph. (TVCG) **29**(4), 1992–2004 (2021)
22. Huang, M.X., Li, J., Ngai, G., Leong, H.V., Bulling, A.: Moment-to-moment detection of internal thought during video viewing from eye vergence behavior. In: Proceedings of ACM Multimedia (MM), pp. 1–9 (2019)
23. Kifer, D., Gehrke, J.: Injecting utility into anonymized datasets. In: Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data, pp. 217–228 (2006)
24. Kim, N.W., et al.: Bubbleview: an interface for crowdsourcing image importance maps and tracking visual attention. ACM Trans. Comput.-Hum. Interact. (TOCHI) **24**(5), 1–40 (2017)
25. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. CoRR abs/1412.6980 (2014). https://api.semanticscholar.org/CorpusID:6628106
26. Lallé, S., Conati, C., Carenini, G.: Predicting confusion in information visualization from eye tracking and interaction data. In: IJCAI, pp. 2529–2535 (2016)
27. Lethaus, F., Baumann, M.R., Köster, F., Lemmer, K.: A comparison of selected simple supervised learning algorithms to predict driver intent based on gaze data. Neurocomputing **121**, 108–130 (2013)
28. Li, J., Chowdhury, A.R., Fawaz, K., Kim, Y.: {Kalεido}:{Real-Time} privacy control for {Eye-Tracking} systems. In: 30th USENIX Security Symposium, pp. 1793–1810 (2021)
29. Liu, A., Xia, L., Duchowski, A., Bailey, R., Holmqvist, K., Jain, E.: Differential privacy for eye-tracking data. In: ACM Symposium on Eye Tracking Research & Applications, pp. 1–10 (2019)
30. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam. arXiv abs/1711.05101 (2017). https://api.semanticscholar.org/CorpusID:3312944

31. Masry, A., Long, D., Tan, J.Q., Joty, S., Hoque, E.: ChartQA: a benchmark for question answering about charts with visual and logical reasoning. In: Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, pp. 2263–2279. Association for Computational Linguistics (2022)
32. Matzen, L.E., Haass, M.J., Divis, K.M., Wang, Z., Wilson, A.T.: Data visualization saliency model: a tool for evaluating abstract data visualizations. IEEE Trans. Vis. Comput. Graph. (TVCG) **24**(1), 563–573 (2017)
33. Melis, L., Song, C., De Cristofaro, E., Shmatikov, V.: Exploiting unintended feature leakage in collaborative learning. In: 2019 IEEE Symposium on Security and Privacy (SP), pp. 691–706. IEEE (2019)
34. Narayanan, A., Shmatikov, V.: Robust de-anonymization of large sparse datasets. In: 2008 IEEE Symposium on Security and Privacy (SP 2008), pp. 111–125. IEEE (2008)
35. Oquab, M., et al.: Dinov2: learning robust visual features without supervision (2023)
36. Patro, B., Namboodiri, V., et al.: Explanation vs attention: a two-player game to obtain attention for VQA. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 11848–11855 (2020)
37. Pfleging, B., Fekety, D.K., Schmidt, A., Kun, A.L.: 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), pp. 5776–5788 (2016)
38. Polatsek, P., Waldner, M., Viola, I., Kapec, P., Benesova, W.: Exploring visual attention and saliency modeling for task-based visual analysis. Comput. Graph. **72**, 26–38 (2018)
39. Sammaknejad, N., Pouretemad, H., Eslahchi, C., Salahirad, A., Alinejad, A.: Gender classification based on eye movements: a processing effect during passive face viewing. Adv. Cogn. Psychol. **13**(3), 232 (2017)
40. Sattar, H., Fritz, M., Bulling, A.: Deep gaze pooling: Inferring and visually decoding search intents from human gaze fixations. Neurocomputing **387**, 369–382 (2020)
41. Shin, S., Chung, S., Hong, S., Elmqvist, N.: A scanner deeply: predicting gaze heatmaps on visualizations using crowdsourced eye movement data. IEEE Trans. Vis. Comput. Graph. (TVCG) **29**(1), 396–406 (2022)
42. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 3–18. IEEE (2017)
43. Sood, E., Shi, L., Bortoletto, M., Wang, Y., Müller, P., Bulling, A.: Improving neural saliency prediction with a cognitive model of human visual attention. In: Proceedings the 45th Annual Meeting of the Cognitive Science Society (CogSci), pp. 3639–3646 (2023)
44. Sutskever, I., Martens, J., Dahl, G., Hinton, G.: On the importance of initialization and momentum in deep learning. In: Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28, ICML 2013, pp. III-1139–III-1147. JMLR.org (2013)
45. Vaswani, A.: Attention is all you need. In: Advances in Neural Information Processing Systems (2017)
46. Wang, H., Zhan, Z., Shan, H., Dai, S., Panoff, M., Wang, S.: Gazeploit: remote keystroke inference attack by gaze estimation from avatar views in VR/MR devices. In: Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, pp. 1731–1745 (2024)

47. Wang, X., et al.: The mental image revealed by gaze tracking. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI), pp. 1–12 (2019)
48. Wang, Y., Cai, W., Gu, T., Shao, W.: Your eyes reveal your secrets: an eye movement based password inference on smartphone. IEEE Trans. Mob. Comput. **19**(11), 2714–2730 (2019)
49. Wang, Y., Jiang, Y., Hu, Z., Ruhdorfer, C., Bâce, M., Bulling, A.: Visrecall++: analysing and predicting visualisation recallability from gaze behaviour. Proc. ACM Hum.-Comput. Interact. **8**(ETRA), 1–18 (2024)
50. Wang, Y., et al.: Salchartqa: question-driven saliency on information visualisations. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–14 (2024)
51. Wightman, R.: Pytorch image models (2019). https://github.com/rwightman/pytorch-image-models. https://doi.org/10.5281/zenodo.4414861
52. Wolf, T., et al.: Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45. Association for Computational Linguistics, Online (2020). https://www.aclweb.org/anthology/2020.emnlp-demos.6
53. Yu, D., Fu, J., Mei, T., Rui, Y.: Multi-level attention networks for visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4709–4717 (2017)
54. Zermiani, F., Bulling, A., Wirzberger, M.: 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), pp. 1–7 (2022)
55. Zhang, B., He, X., Shen, Y., Wang, T., Zhang, Y.: A plot is worth a thousand words: model information stealing attacks via scientific plots. In: 32nd USENIX Security Symposium (USENIX Security 2023), pp. 5289–5306 (2023)
56. Zhang, G., et al.: Predicting next actions and latent intents during text formatting. In: Proceedings of the CHI Workshop Computational Approaches for Understanding, Generating, and Adapting User Interfaces, pp. 1–6 (2022)
57. Zhang, S., Ma, D., Wang, Y.: Don't peek at my chart: privacy-preserving visualization for mobile devices. In: Computer Graphics Forum, vol. 42, pp. 137–148. Wiley Online Library (2023)