

Grid Labeling: Crowdsourcing Task-Specific Importance from Visualizations

Minsuk Chang¹, Yao Wang², Huichen Will Wang³, Andreas Bulling², and Cindy Xiong Bearfield¹

¹Georgia Institute of Technology, USA

²University of Stuttgart, Germany

³University of Washington, USA

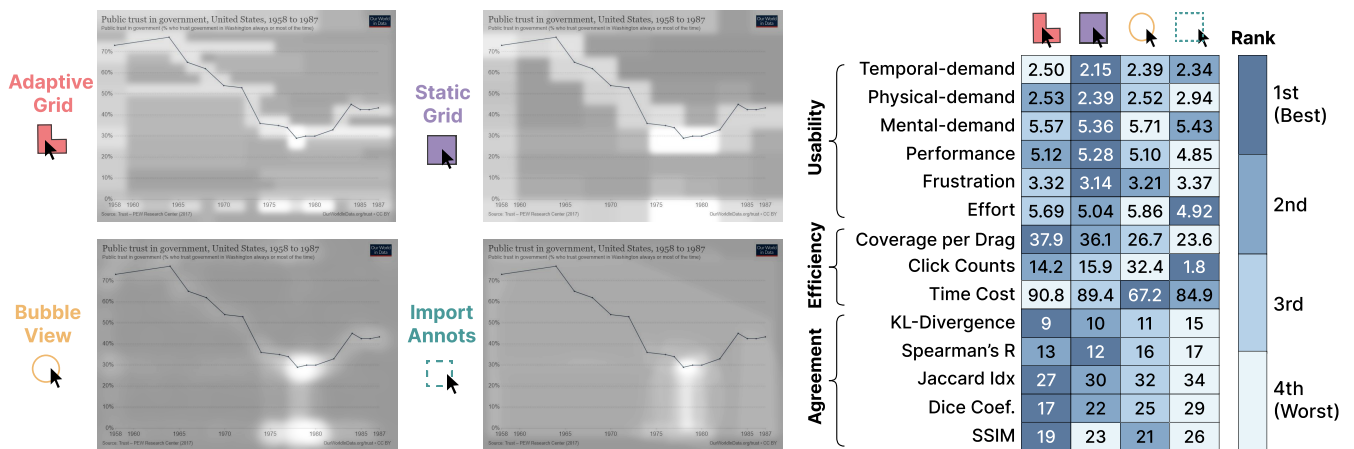


Figure 1: (left) Example importance annotations for the task “Between which two years was public trust in government the lowest?” obtained using three tools: Adaptive/Static Grid (Ours), BubbleView [KBB*17], and ImportAnnots [OAH14]. (right) Average quantity and ranking for each annotation method for all metrics used in our experiment. A darker color represents a higher ranking in each row.

Abstract

Knowing where people look in visualizations is key to effective design. Yet, existing research primarily focuses on free-viewing-based saliency models—although visual attention is inherently task-dependent. Collecting task-relevant importance data remains a resource-intensive challenge. To address this, we introduce Grid Labeling – a novel annotation method for collecting task-specific importance data to enhance saliency prediction models. Grid Labeling dynamically segments visualizations into Adaptive Grids, enabling efficient, low-effort annotation while adapting to visualization structure. We conducted a human-subject study comparing Grid Labeling with existing annotation methods, ImportAnnots, and BubbleView across multiple metrics. Results show that Grid Labeling produces the least noisy data and the highest inter-participant agreement with fewer participants while requiring less physical (e.g., clicks/mouse movements) and cognitive effort. An interactive demo and the accompanying dataset are available at <https://github.com/jangsus1/Grid-Labeling>.

CCS Concepts

• Human-centered computing → Visualization techniques; Empirical studies in visualization;

1. Introduction

Where do people look in visualizations under tasks? Understanding salient parts of visualizations is crucial for designing compelling visualizations that optimally support analytic tasks [BKO*17]. How-

ever, modeling saliency is challenging as where people visually focus is inherently task-dependent. For example, during free viewing, participants may primarily engage in bottom-up processes driven by visually salient elements, such as a bright red patch [KW79]. However, when given an analytic task, participants are more likely

to engage in top-down processing, directing their attention toward task-relevant visualization regions. For example, in a “find the extreme” task, they may focus on the tip of the highest bar [GL13]. In this context, task-dependent saliency is strongly related to importance, which involves actively filtering areas with sufficient information for task-solving. Existing work has defined the image’s “importance” as regions where individual annotators believe as important [OAH14]. We supplement this definition by taking a task-dependent approach to propose a new alternate definition for “task-specific importance” as “the minimum area in a visualization required for a user to complete a task successfully.”

However, existing mouse-tracking-based saliency collection methods rely on free-viewing [NMF*20, OAH14, KBB*17], as they are designed to encourage users to explore and describe an image. However, because visualizations are often used for analytic tasks [AES05], we posit that an effective model for predicting where people look should consider task relevance when identifying regions of importance. To address this limitation, we contribute *Grid Labeling*, a toolkit to enhance existing saliency prediction models with task-specific annotation. A key advantage of our tool is that it is more resource-efficient than traditional methods, such as eye-tracking or mouse-tracking [BKO*17, OAH14, KBB*17, NMF*20, WWA*24].

Grid Labeling segments visualizations into Adaptive Grids that dynamically adjust based on existing graphical elements, making it easily adapted to various visualization sizes and designs. This approach enables participants to identify critical areas simply by clicking relevant grids, eliminating the need for cumbersome mouse interactions, such as free-form drawing to annotate regions [OAH14] or clicking the same regions multiple times [KBB*17]. Moreover, Grid Labeling streamlines data collection, reducing the number of participants required to converge to a stable importance map. In a human-subject experiment, we demonstrate that, compared to the two popular approaches, ImportAnnots [OAH14] and BubbleView [KBB*17], Grid Labeling produces less noisy data with higher levels of agreement between participant responses. Additionally, participants reported lower perceived effort when using our method. We also explore the key distinction between saliency and importance, contributing to differences in annotation duration.

The specific contributions of our work are three-fold:

- We introduce the Grid Labeling method for capturing “task-specific importance” in visualizations, which enhances task-specific saliency modeling.
- Through a human-subject study, we illustrate the importance of considering task-specific “areas of importance” in visualizations.
- We quantitatively demonstrate that our Grid Labeling method outperforms traditional crowd-sourcing methods for collecting task-specific importance data in visualizations. Participants in our study could identify important areas with less effort and with higher inter-participant agreement.

2. Related Work

Researchers have been leveraging eye tracking methodologies from human perception research to model how people perceive im-

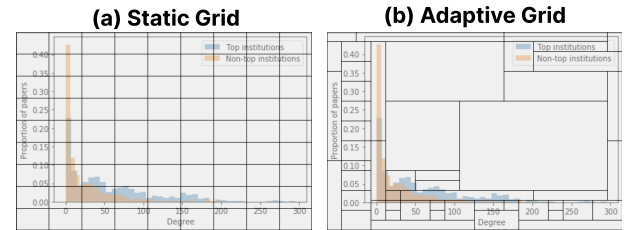


Figure 2: Comparison of Static and Adaptive Grid segmentations applied to a histogram from CharXiv [WXH*24]. Static Grid splits the visualization into equal-sized rectangles, while Adaptive Grid dynamically produces patches that fit the visualization layout.

ages [KNEM15, CPS16]. These models help assess the appearance and salience of visual representations, enabling eye movement tracking to understand the perceptual and cognitive mechanisms of scene perception [IKN98] and object detection [BCJL15]. The existing saliency models perform well in naturalistic scenes; however, there are unique perception rules and cognitive biases in the artificial world of data visualization [CAFG12], and, thus, these models do not accurately predict where people would look in visualizations. Visualization researchers have been building visual saliency models geared to visualizations [MHD*18, BRB*16]. However, these models rely on handcrafted features, making it difficult to generalize to complex visualizations. Additionally, these models cannot incorporate textual information to generate task-specific saliency maps since the prediction is solely based on visual inputs.

With deep learning, gaze data became the ground truth for saliency models [SCHE23, WBB24], increasing prediction performance and enabling task-specific saliency [WWA*24]. These models require large datasets, but collecting accurate gaze data is expensive and cumbersome. To address this, researchers proposed gaze proxies. For example, WebGaze [PSL*16] offers low-cost webcam-based data collection for online studies, though it struggles with quality due to low resolution and poor calibration. Therefore, mouse cursor-based annotation tools [JHDZ15, KBB*17, OAH14] were proposed to improve data quality. Among these methods, BubbleView [KBB*17] was the most widely used tool for capturing visual saliency and importance [BKO*17, WWA*24]. However, BubbleView is primarily designed for exploring images and gathering information, which differs slightly from the goal of capturing perceived importance. As a result, while BubbleView is well-suited for measuring visual saliency, it may not be the best tool for capturing task-specific importance [NMF*20]. Built upon these prior approaches’ limitations, our Grid Labeling aims to collect responses that cover all essential areas of the visualization with minimum noise, leading to more efficient data collection.

3. Grid Labeling

Existing saliency and importance annotation methods use circular [KBB*17] or freeform [OAH14] shapes, which do not preserve the structure of visual elements, particularly when creating saliency maps by Gaussian kernels [WBB24]. Inspired by Google’s reCaptcha [Goo25], we propose **Grid Labeling**, a patch-based annotation approach that addresses this limitation. With Grid Labeling, users annotate specific areas by clicking on image patches

(Figure 2). This binary interaction—clicking or not clicking—enforces discrete annotations, facilitating faster response aggregation by promoting higher consensus.

3.1. Static Grid

As a baseline, we first designed the Static Grid by dividing the visualization’s height and width into N equal sections, resulting in N^2 patches. We leave N as a hyperparameter, which was set to 8 in our experiment. This made the patch size approximately equivalent to the recommended circle size in BubbleView [KBB*17].

3.2. Adaptive Grid

To further reduce annotation time and effort, we introduce an Adaptive Grid that groups smaller *tiles* into larger *blocks* aligned with the visualization’s layout.

Step 1: Split Regions. We divide the visualization into three regions: text, edge, and background. We filter out the text area with PaddlePaddle OCR, followed by the Canny edge detection algorithm to extract graphical elements. The remaining tiles not identified as text or edges are labeled as background.

Step 2: Defining the Tile Space. Let’s assume we are filling in the visualization with small tiles with the size of t (e.g., 32px), which is the minimum patch size. Then the image would be covered with a grid of tiles with dimensions $M \times N$ ($M = \lfloor \text{Height}/t \rfloor$, $N = \lfloor \text{Width}/t \rfloor$), forming $G \in \{0, 1\}^{M \times N}$. We individually build the binary grid for each region (text, edge, or background), which will be covered with larger blocks in Step 3. Each entry $G_{i,j}$ is set to 1 if tile (i, j) belongs to the selected region and 0 otherwise.

Step 3: Optimizing Block Arrangement. We then assign larger rectangular blocks that can minimally cover the entire grid. Define binary decision variables $B_{i,j}^{w,h} \in \{0, 1\}$, where $B_{i,j}^{w,h} = 1$ indicates that a rectangular block of size $w \times h$ tiles is placed with its top-left corner at tile (i, j) . Our objective is to minimize the total number of blocks: $\min \sum_{i,j,w,h} B_{i,j}^{w,h}$, while the entire region must be covered once without overlap between the blocks. This merges coverage and non-overlap requirements into a single constraint:

$$\sum_{\substack{(i',j',w,h) \\ i' \leq i < i'+h, j' \leq j < j'+w}} B_{i',j'}^{w,h} = G_{i,j}, \quad \forall (i, j).$$

We solve this optimization problem using Integer Linear Programming (ILP) with OR-Tools’ constraint programming solver. By enforcing the exact coverage of each tile, we get patches that cover the visualization while respecting the background contents.

4. User Study

We investigate how crowdworkers annotate important areas in visualizations with different annotation methods. We first demonstrate that annotations differ when users are instructed to annotate task-specific vs. task-agnostic areas, motivating the need for more task-specific annotation data to be collected. We then compare four methods participants can use to identify minimal task-relevant areas in a visualization: BubbleView, ImportAnnots, Static Grid, and

Adaptive Grid. We evaluate these methods based on four metrics: task completion time, the number of participants required for convergence, annotation effort (e.g., number of clicks), and usability.

4.1. Participants and Design

We conducted a power analysis based on pilot results. Considering the smallest effect size across all comparison metrics (Cohen’s $f = 0.2715$ for cognitive load), a target sample of 152 participants would yield 80% power to detect an overall difference between annotation methodologies at an α level of 0.05. We recruited participants from Prolific ($M_{age} = 38.5$, $\sigma = 11.9$, 52 females) and compensated them \$12 per hour. We curated a set of 18 visualizations from ChartQA [MLT*22] and CharXiv [WXH*24], covering a diverse range of chart types (e.g., bar, stacked bar, pie, line, Choropleth map, heatmap, histogram, scatterplot, and contour plots). Each participant was then shown these selected charts in a random order. In a between-subject set-up, participants were randomly assigned to annotate them via one of four tools: BubbleView, ImportAnnots, and Static/Adaptive Grid.

4.2. Procedure

The study was conducted as a between-subject experiment. After consenting to the experiment, participants were instructed on how to use the assigned annotation tool (BubbleView, ImportAnnots, Static Grid, or Adaptive Grid). They first solved an example task with a simple bar chart with one of the following prompts: *annotate the important area and describe key points*, *annotate the area minimally required for you to identify the highest value in the chart* (results see Section 4.3). Then, they labeled 18 visualizations using the same tool. For ImportAnnots, participants were instructed as *annotate important areas related to answering the question*. For BubbleView, they were just asked to answer the question following the prior work’s design [WWA*24]. For Grid Labeling (Adaptive/Static), they were instructed to annotate “Task-Specific Importance.” In the end, they reported the tool’s usability using NASA-TLX [HS88], completed an assessment of their visualization literacy [PO23], and provided demographic information.

4.3. Results: Importance vs. Free-Viewing

We demonstrate the participants’ annotation behavior appeared significantly different when they were instructed to annotate components of the visualization they found important during free-viewing, compared to when they were instructed to annotate the importance area in response to a specific task, as shown in Figure 3. During task-agnostic annotation, the importance area is more evenly distributed across the visualization with a slight emphasis on the top of the visualization and the title text (aligned with existing work such as patterns identified by [BKO*17]). In contrast, the annotations cluster around the area with the smallest bar at the bottom of the visualization in response to the find minimum task. This further validates our case that existing models trained on free-viewing annotation and eye-tracking data might not be the most predictive for visualization saliency, considering salient regions can vary with user intent and tasks.

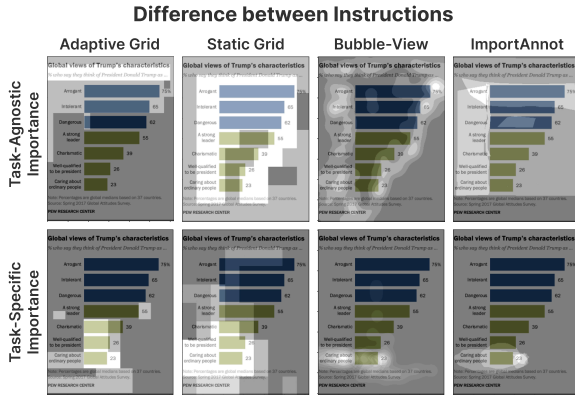


Figure 3: Averaged participants' annotations for four tools when applying different instructions. The task-specific instruction was "What is the important area regarding this question: What is the minimum value of the bar?".

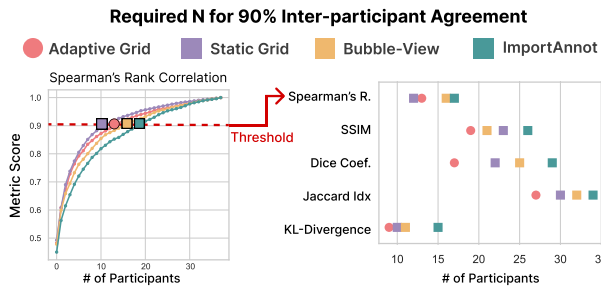


Figure 4: Required number of participants to reach 90% similarity compared to the aggregated importance. Five different similarity metrics were used to measure the convergence.

4.4. Results: Methods Comparison

In the user study, we compared four annotation methods using five metrics: usability, click counts, annotated area per mouse travel distance, annotation speed, and inter-participant agreement. The group means per metric are summarized in Figure 1.

Usability: Participant Task-Load Analysis To measure the usability of the annotation methods, we used the NASA-TLX scale, which has six categories (mental demand, physical demand, temporal demand, effort, performance, and frustration). The MANOVA suggested that the annotation methods statistically differ (Wilks' $\Lambda = 0.8626$, $F(18, 7362.88) = 21.94$, $p < 0.001$), while independent ANOVAs showed differences except for frustration. A post hoc Tukey test followed the five remaining categories. ImportAnnots showed higher physical demand than others ($p < 0.001$). Static Grid had lower temporal demand than Adaptive Grid and Bubble-View ($p \leq 0.023$) while also having higher performance than Bubble-View and ImportAnnots ($p \leq 0.028$). For effort, ImportAnnots and Static Grid were better than other tools ($p < 0.001$).

Interaction Efficiency: Click Count Comparison. The ANOVA ($F(32,608) = 248.995$) and post hoc Tukey test suggested that the significant effect is driven by the difference in click count between all pair-wise comparisons among methods except adaptive and static grid ($p = 0.42$). Participants labeled important areas with

adaptive and static grids using fewer clicks than Bubble-View. The grid-based methods required more clicks than ImportAnnots, but that is caused by the stroke tool, which allowed participants to annotate a large area while dragging the mouse with a single click.

Coverage Efficiency: Annotated Area per Mouse Travel Distance. We measured the coverage efficiency by dividing the annotated area size by the mouse travel distance during annotation. With ANOVA ($F=30.33$) and post hoc Tukey ($p < 0.001$), we observed a significant difference between Adaptive Grid and Static Grid, having higher coverage efficiency than Bubble-View and ImportAnnots.

Time Efficiency: Annotation Speed Across Methods. ANOVA ($F=11.79$) and the following Tukey test showed Bubble-View required less time per annotation compared to other tools ($p < 0.001$), demonstrating its efficiency in annotation speed. However, we argue that the underlying reason stems from the difference between saliency and importance, where capturing importance may naturally involve more intention during the annotation process [NMF*20].

Convergence Speed: Agreement Across Participants. We examined the number of participants needed to achieve 90% similarity with the aggregated mask using five similarity metrics: Spearman's Rank Correlation [Spe04], Structural Similarity Index [WBSS04], Dice Coefficient [Dic45], Jaccard Index [Jac12], and Kullback-Leibler (KL) Divergence [KL51], often used to measure difference between continuous 2d maps. We measured the convergence of 10 different randomized orders of responses for more generalized results with smoother graphs. As shown in Figure 4, the Adaptive Grid and Static Grid generally converge faster than the other tools across most metrics, while the Adaptive Grid was the best performing in all metrics except Spearman's R.

5. Discussion & Future Work

We contribute Grid Labeling, an annotation method for efficiently crowdsourcing task-dependent important areas of visualizations. Grid Labeling outperformed other approaches across all metrics, as shown in Figure 1. While ImportAnnots [OAH14] had the lowest click count and Bubble-View [KBB*17] required the least time, Adaptive Grid achieved the highest inter-participant agreement with the fewest participants across multiple metrics (e.g., SSIM, Dice, Jaccard, KL). Meanwhile, Static Grid demonstrated higher usability, as indicated by NASA-TLX [HS88]. These results highlight the potential of Grid Labeling in training task-specific saliency models, minimizing text overemphasis, and enhancing predictive accuracy. Considering the trade-offs, we recommend using an Adaptive Grid for maximizing convergence and a Static Grid to improve usability.

Since the present study did not explore why participants labeled certain grids as important and relied on an inter-participant agreement for quality control, future work could investigate the reasoning behind these selections to provide a more high-level, representational explanation, collecting a large-scale dataset and training a task-specific importance prediction model. Additionally, future work could refine Adaptive Grid generation using vision-based LLMs to enhance annotation usability by semantically filtering less important visualization grids through an automated pipeline.

Acknowledgements

This work was supported by the National Science Foundation under grants IIS-2237585 and IIS-2311575. Y. Wang was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161. A. Bulling was funded by the European Research Council (ERC) under grant agreement 801708.

References

- [AES05] AMAR R., EAGAN J., STASKO J.: Low-level components of analytic activity in information visualization. In *Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization (USA, 2005)*, INFOVIS '05, IEEE Computer Society, p. 15. doi:10.1109/INFOVIS.2005.24. 2
- [BCJL15] BORJI A., CHENG M.-M., JIANG H., LI J.: Salient object detection: A benchmark. *IEEE Transactions on Image Processing* 24, 12 (2015), 5706–5722. doi:10.1109/TIP.2015.2487833. 2
- [BKO*17] BYLINSKII Z., KIM N. W., O'DONOVAN P., ALSHEIKH S., MADAN S., PFISTER H., DURAND F., RUSSELL B., HERTZMANN A.: Learning visual importance for graphic designs and data visualizations. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology (New York, NY, USA, 2017)*, UIST '17, Association for Computing Machinery, p. 57–69. doi:10.1145/3126594.3126653. 1, 2, 3
- [BRB*16] BYLINSKII Z., RECASENS A., BORJI A., OLIVA A., TORRALBA A., DURAND F.: Where should saliency models look next? In *Computer Vision – ECCV 2016 (Cham, 2016)*, Leibe B., Matas J., Sebe N., Welling M., (Eds.), Springer International Publishing, pp. 809–824. 2
- [CAFG12] CORRELL M., ALBERS D., FRANCONERI S., GLEICHER M.: Comparing averages in time series data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (New York, NY, USA, 2012)*, CHI '12, Association for Computing Machinery, p. 1095–1104. doi:10.1145/2207676.2208556. 2
- [CPS16] CONKLIN K., PELLICER-SÁNCHEZ A.: Using eye-tracking in applied linguistics and second language research. *Second Language Research* 32, 3 (2016), 453–467. 2
- [Dic45] DICE L. R.: Measures of the amount of ecologic association between species. *Ecology* 26, 3 (1945), 297–302. doi:10.2307/1932409. 4
- [GL13] GILBERT C. D., LI W.: Top-down influences on visual processing. *Nature reviews neuroscience* 14, 5 (2013), 350–363. doi:10.1038/nrn3476. 2
- [Goo25] GOOGLE: reCAPTCHA: Easy on Humans, Hard on Bots. <https://developers.google.com/recaptcha>, 2025. Accessed: 2025-02-18. 2
- [HS88] HART S. G., STAVELAND L. E.: Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in Psychology* 52 (1988), 139–183. doi:10.1016/S0166-4115(08)62386-9. 3, 4
- [IKN98] ITTI L., KOCH C., NIEBUR E.: A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 11 (1998), 1254–1259. doi:10.1109/34.730558. 2
- [Jac12] JACCARD P.: The distribution of the flora in the alpine zone. *New Phytologist* 11, 2 (1912), 37–50. doi:10.1111/j.1469-8137.1912.tb05611.x. 4
- [JHDZ15] JIANG M., HUANG S., DUAN J., ZHAO Q.: Salicon: Saliency in context. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)*, pp. 1072–1080. doi:10.1109/CVPR.2015.7298710. 2
- [KBB*17] KIM N. W., BYLINSKII Z., BORKIN M. A., GAJOS K. Z., OLIVA A., DURAND F., PFISTER H.: Bubbleview: an interface for crowdsourcing image importance maps and tracking visual attention. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 5 (2017), 36. doi:10.1145/3131275. 1, 2, 3, 4
- [KL51] KULLBACK S., LEIBLER R. A.: *On information and sufficiency*, vol. 22. Institute of Mathematical Statistics, 1951. doi:10.1214/aoms/1177729694. 4
- [KNEM15] KARTHIKEYAN S., NGO T., ECKSTEIN M., MANJUNATH B.: Eye tracking assisted extraction of attentionally important objects from videos. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)*, pp. 3241–3250. doi:10.1109/CVPR.2015.7298944. 2
- [KW79] KINCHLA R. A., WOLFE J. M.: The order of visual processing: “top-down,” “bottom-up,” or “middle-out”. *Perception & psychophysics* 25 (1979), 225–231. doi:https://doi.org/10.3758/BF03202991. 1
- [MHD*18] 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 Transactions on Visualization and Computer Graphics* 24, 1 (2018), 563–573. doi:10.1109/TVCG.2017.2743939. 2
- [MLT*22] MASRY A., LONG D. X., TAN J. Q., JOTY S., HOQUE E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244* (2022). 3
- [NMF*20] NEWMAN A., MCNAMARA B., FOSCO C., ZHANG Y. B., SUKHUM P., TANCIK M., KIM N. W., BYLINSKII Z.: Turkeyes: A web-based toolbox for crowdsourcing attention data. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (2020)*, CHI '20, p. 1–13. doi:10.1145/3313831.3376799. 2, 4
- [OAH14] O'DONOVAN P., AGARWALA A., HERTZMANN A.: Learning Layouts for Single-Page Graphic Designs. *IEEE Transactions on Visualization and Computer Graphics* 20, 8 (Aug. 2014), 1200–1213. doi:10.1109/TVCG.2014.48. 1, 2, 4
- [PO23] PANDEY S., OTTLEY A.: Mini-vlat: A short and effective measure of visualization literacy. In *Computer Graphics Forum* (2023), vol. 42, Wiley Online Library, pp. 1–11. doi:10.1111/cgf.14809. 3
- [PSL*16] PAPOUTSAKI A., SANGKLOY P., LASKEY J., DASKALOVA N., HUANG J., HAYS J.: Webgazer: scalable webcam eye tracking using user interactions. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (2016)*, IJCAI'16, p. 3839–3845. 2
- [SCHE23] SHIN S., CHUNG S., HONG S., ELMQVIST N.: A scanner deeply: Predicting gaze heatmaps on visualizations using crowdsourced eye movement data. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2023), 396–406. doi:10.1109/TVCG.2022.3209472. 2
- [Spe04] SPEARMAN C.: The proof and measurement of association between two things. *The American Journal of Psychology* 15, 1 (1904), 72–101. doi:10.2307/1412159. 4
- [WBB24] WANG Y., BÂCE M., BULLING A.: Scanpath prediction on information visualisations. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 30, 7 (2024), 3902–3914. 2
- [WBSS04] WANG Z., BOVIK A. C., SHEIKH H. R., SIMONCELLI E. P.: Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612. doi:10.1109/TIP.2003.819861. 4
- [WWA*24] WANG Y., WANG W., ABDELHAFEZ A., ELFARES M., HU Z., BÂCE M., BULLING A.: Salchartqa: Question-driven saliency on information visualisations. In *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems (CHI) (2024)*, pp. 1–14. doi:10.1145/3613904.3642942. 2, 3
- [WXH*24] WANG Z., XIA M., HE L., CHEN H., LIU Y., ZHU R., LIANG K., WU X., LIU H., MALLADI S., CHEVALIER A., ARORA S., CHEN D.: Charxiv: Charting gaps in realistic chart understanding in multimodal llms. 113569–113697. 2, 3