Computational Representations for Graphical User Interfaces

Yue Jiang
yue.jiang@aalto.fi
Aalto University
Finland

ABSTRACT

Graphical User Interfaces (GUIs) have been widely used in daily life. To enhance GUI design and interaction experience on GUIs, it is important to understand GUIs and understand how individuals interact with them. Consequently, my thesis focuses on applying computational approaches to improve our understanding of GUIs and user interactions. First, I introduce novel GUI representations to capture the visual, spatial, and semantic factors of GUIs and improve the performance of downstream GUI tasks. Second, I simulate how users visually engage with GUIs to understand user interactions to help inform the design of GUI representations. Third, based on the understanding of GUIs and interactions on GUIs, I develop language language representations aimed at assisting users in understanding and more effectively interacting with GUIs.

ACM Reference Format:
Yue Jiang. 2024. Computational Representations for Graphical User Interfaces. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3613905.3638191

1 INTRODUCTION

Graphical User Interfaces (GUIs) have played an important role in enhancing digital interactions. Understanding GUIs and how users interact with them is essential for improving GUI design and user experience on GUIs [18–20]. In my thesis, I employ computational approaches to understand GUIs and user interactions with them.

Different graphical user interface (GUI) representations can capture various aspects of GUI design. Deep learning approaches heavily depend on how data are represented, and different representations can capture the different factors of GUI design. The choice of data representation for GUIs can affect the capabilities of downstream tasks within graphical user interfaces. The effectiveness of deep learning-based methods often depends on the choice of data representation. Present deep learning-based GUI methods often require the design of preprocessing pipelines and data transformations, which can be time-consuming and may struggle to extract discriminative information from the data. A well-designed representation can leverage prior knowledge to address this limitation.

Many downstream tasks on GUIs require the use of suitable representation. Recent computational approaches improved the GUI design workflows by giving suggestions on GUIs [14, 25, 27, 38, 44], adapted GUIs to different devices and user preferences, or user actions [10, 11, 13, 15, 28, 32, 33, 42, 43], reverse engineered GUIs to understand UIs and improve accessibility [9, 22, 35, 45, 51]. They use various GUI representations to solve individual downstream tasks but still lack a unified representation for various tasks.

Existing GUI representations have the limitation that they cannot capture all the visual, spatial, and semantic factors of GUIs. Some representations focus on textural content [29, 30], while others pay attention to visual appearance [2, 8, 37]. They only emphasize certain aspects of GUI properties while neglecting others. I aim to address it by bridging the gap between textual content, visual appearance, and layout-based constraints to enhance the interpretability and performance of downstream GUI tasks. The second aspect of my research focuses on understanding and simulating user interactions with GUIs to gain insights into how users visually engage with these interfaces. I aim to use eye tracking prediction to inform the design of representations by understanding how GUIs are perceived. Previous work often focused on eye tracking proxies, such as webcam or mouse, or concentrated on specific design types like mobile GUIs. Instead, I focus on simulating real eye tracking using the data collected from eye trackers and covering various GUI categories. Finally, building on our understanding of GUIs and user interactions, I developed a language representation with the goal of assisting users in better understanding and interacting with GUIs, serving as a UI-focused instruction-following visual agent.

Thus, the main research objectives of my thesis are

- To develop efficient GUI representations that capture visual, spatial, and semantic structures of GUIs for downstream tasks.
- To explore the simulation of human behaviors, such as eye tracking, for gaining insights into user interactions with GUIs, reducing the dependency on human analysis and user studies.
- To create a language agent that interactively assists users in understanding and interacting with GUIs.

2 RESULTS AND CONTRIBUTIONS

2.1 GUI Representation

Present-day graphical user interfaces (GUIs) consist of various arrangements of text, images, and interactive components. The challenge of effectively conveying a GUI's visual, spatial, and semantic structure in computational design remains unsolved. Current GUI representations have limitations as they tend to focus on specific aspects of GUI properties while overlooking others. Some previous methods focused on capturing textual content within GUIs [29, 30]. However, these approaches neglected the visual appearance and

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

ACM ISBN 979-8-4007-0331-7/24/05.
https://doi.org/10.1145/3613905.3638191

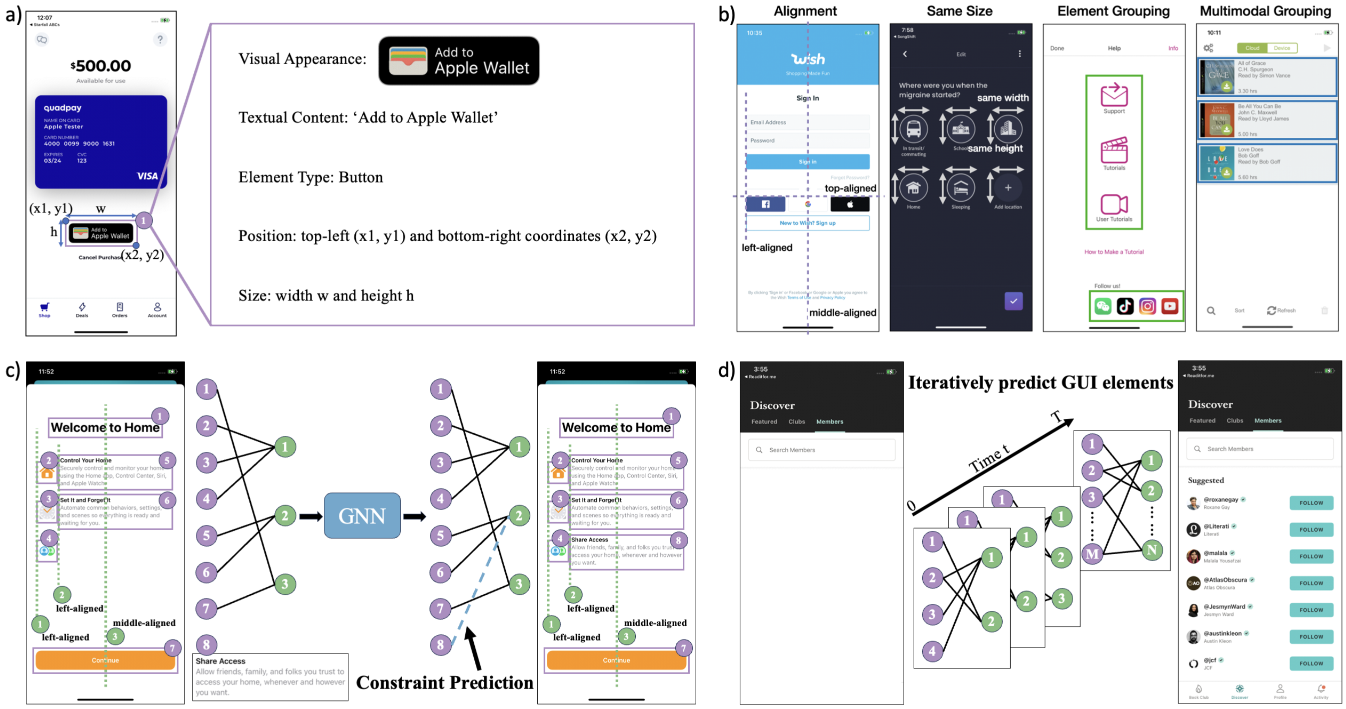

Figure 1: Our graph-based GUI representation connects between GUI element properties (a) and constraints (b) to capture the visual–spatial–semantic structure of a GUI. It is a bipartite graph containing element nodes (colored purple) representing the GUI elements' properties and constraint nodes (colored green) that can be fed into graph neural networks (GNNs) (c). We show that it can help GUI design by iteractively suggesting element sizes and positions (d).

types of GUI elements. On the other hand, some methods prioritized visual appearance and the types of GUI elements [2, 8, 37] but often neglected textual content. Consequently, they may treat GUIs with similar structural and visual features differently due to variations in textual content. Some attempts to bridge this gap between textual content and visual appearance [31, 34] require substantial task-specific datasets and manual data labeling. In contrast to neural networks, constraint-based approaches have also been explored, allowing explicit rules for element and layout constraints [15, 22, 23, 41]. While constraint-based approaches can provide explicit representations of GUIs and enhance the interpretability of the representation, they still often require the labor-intensive process of manually constructing constraints.

To bridge the gap in GUI representation, we proposed Graph4GUI [24], a bipartite graph-based approach that integrates the properties of GUI elements with their layout-based constraints (Figure 1). This bipartite graph structure expresses each GUI element and its connections to others through element and constraint nodes. Our approach, Graph4GUI, offers a solution that considers not only textual content, visual appearance, and GUI element types but also the constraints and interrelationships between GUI elements. The incorporation of graph neural networks (GNNs) facilitates the learning of domain-specific representations from this graph-structured data, as it aggregates information from neighboring nodes, facilitating the exchange of knowledge between element and constraint nodes. This approach recognizes that individual elements interact with the overall layout, such as icons often grouped and aligned within toolbars. The utilization of GNNs enables us to capture both the GUI layout's structure and the unique properties of GUI elements. To assess the effectiveness of this graph-based representation, we applied it to an autocompletion task, enhancing efficiency by iteratively predicting the locations of unplaced GUI elements and providing confidence levels to assist designers in prioritizing element placement. Compared to prior approaches, our graph-based representation can better understand GUI structures, enhancing the model's interpretability and explainability.

2.2 Simulating User Interactions for GUI Representation

The most distinctive aspect of GUIs compared to natural images is that GUIs are often interactive. Thus, it is important to understand how users interact with GUIs, such as how users look at GUIs. Thus, to better inform the design of GUI representation, we also need to understand and simulate how users interact with GUIs. Eye tracking data can be used to generate heatmaps and fixation maps, which offer visual representations of how users visually engage with and navigate through GUIs, determining where users focus their visual attention when interacting with a GUI. This helps GUI designers understand which elements of the GUI draw the most attention and which ones are often overlooked. Designers can then optimize the placement of important features or content to align with user attention.

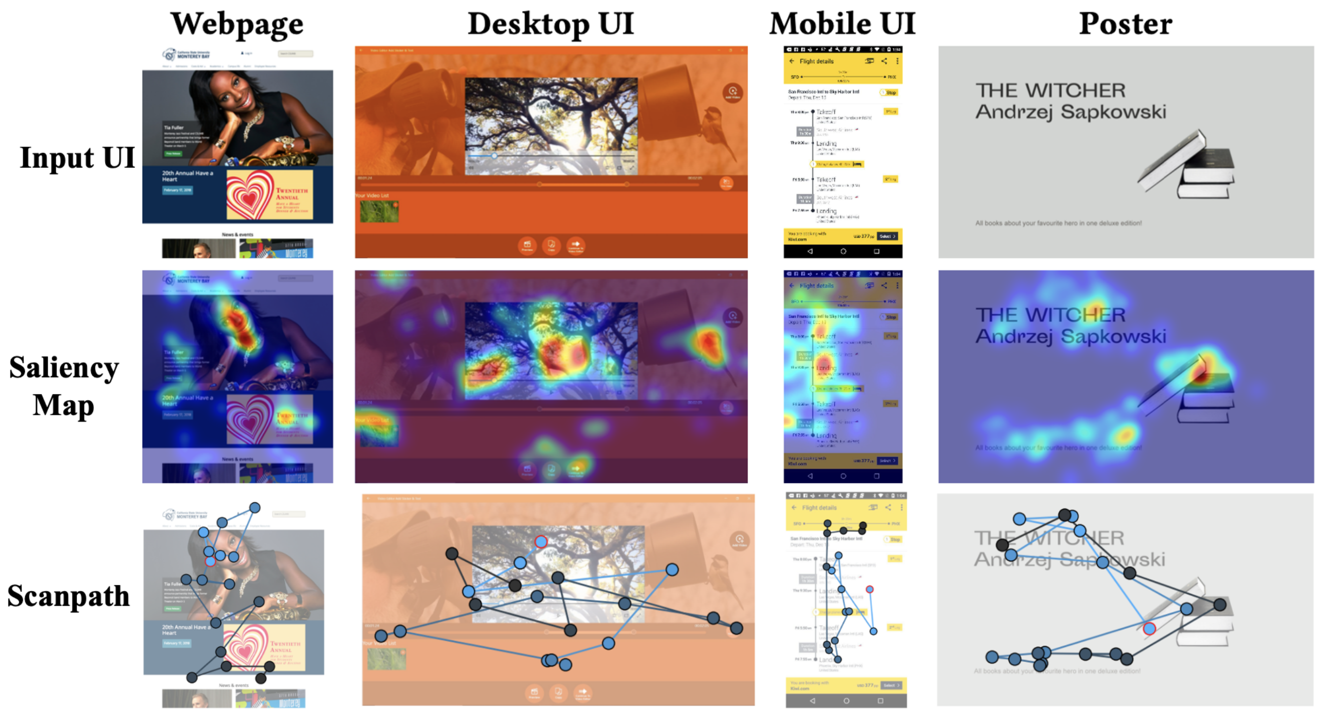

Figure 2: We developed models predicting saliency maps and scanpaths on GUIs.

To understand this, we collected and analyzed a novel eye-tracking dataset, UEyes [16, 17], using a high-fidelity in-lab eye tracker. This dataset includes multi-duration saliency maps and scanpaths generated from the eye movements of 62 participants who interacted with a diverse set of 1,980 user interfaces. Compared to previous eye-tracking datasets, which were often limited in size[3, 50] and often concentrated on specific design types like mobile GUIs [26], our UEyes dataset comprises high-quality eye-tracking data covering various GUI categories, including webpages, mobile interfaces, desktop applications, and posters. We further analyzed saliency-related patterns across the GUI types and created improved models for predicting saliency maps or scanpaths, simulating how users perceive GUIs. These predictive models predict saliency maps and scanpaths, respectively (Figure 2). They offer insights for GUI designers by estimating where users are likely to direct their gaze within a GUI and enable designers to update their GUIs to emphasize important areas. With visual saliency models, designers can improve their designs by making decisions based on how users are likely to perceive their GUIs [4]. Designers can use these visual saliency models to make decisions based on how users are likely to perceive their GUIs. Additionally, the predictive models for scanpaths, which capture the sequence and temporal dynamics of fixations, assist designers in understanding and modifying the visual flow within their designs, encouraging users to engage with GUI elements in the desired sequence.

Furthermore, individuals exhibit diverse gaze patterns influenced by factors such as prior exposure and learning approaches. Thus, we further proposed a novel deep Reinforcement Learning (RL) model with a Transformer architecture, which advances the predictive modeling of personalized scanpaths in GUIs. It predicts both spatial and temporal scanpath characteristics, capturing users' viewing behaviors and long-range dependencies. It predicts a sequence of fixation points for a given GUI image, outperforming previous models. Additionally, it generates personalized scanpaths from a few user-specific scanpath samples, reflecting individual preferences. Unlike methods relying on mouse movements or manual annotations, our model replicates actual eye tracker-recorded scanpaths.

2.3 Language Representation for GUIs

Recent advancements in large language models (LLMs) and vision-language models (VLMs) have opened up new opportunities for computational understanding and interactions with GUIs [47–49]. As natural language is one of the main mediums for human communication and interaction, we can construct language representation to enhance the user experience of understanding and interacting with GUIs. Using text descriptions of GUIs alone with LLMs leaves out the rich visual information of the GUI. Although LLMs have demonstrated remarkable abilities to comprehend task instructions in natural language in many domains [6, 39, 40, 46, 52], fusing visual with textual information is important to understanding GUIs as it mirrors how many humans engage with the world. One approach that has sought to bridge this gap when applied to natural images is Vision-Language Models (VLMs), which accept *multi-modal* inputs of both images and text, typically output only text, and allow for general-purpose question answering, visual reasoning, scene descriptions, and conversations with image inputs [1, 5, 7, 36].

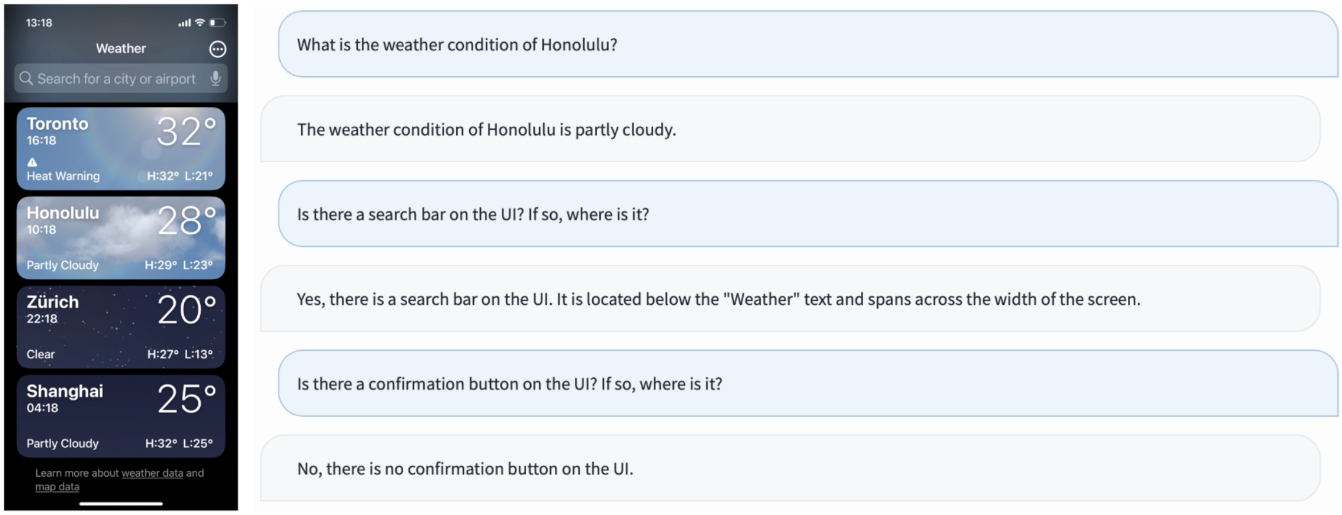

Figure 3: We created a language agent for GUIs, which can perform many UI-related tasks, including conversations, detailed descriptions, listing available actions, predicting UI action outcomes, selecting UI elements, and goal-based planning.

However, the performance of these models on GUI tasks falls short compared to natural images because of the lack of GUI examples in their training data. As shown in Figure 3, we generated paired text-image data to train a VLM model, ILuvUI [21], which is a GUI-focused instruction-following visual agent. It can perform many GUI-related tasks, including conversations, detailed descriptions, listing available actions, predicting GUI action outcomes, selecting GUI elements, and goal-based planning.

3 FUTURE DIRECTIONS

3.1 GUI Representation

I plan to focus on extending the graph-based GUI representation to apply to more diverse GUI-related tasks. I have demonstrated the effectiveness of our graph-based GUI representation in the context of GUI autocompletion; my future work will explore its applicability in more diverse domains, such as evaluation and improvement of accessibility, as well as the detection of hierarchy and grouping within GUIs. For accessibility, I aim to represent accessibility needs as constraints [12] and use such constraints to train and predict layout constraints, potentially leading to improvements in accessibility. In addition, I also plan to employ the graph-based GUI representation to detect hierarchy and grouping within GUIs with high accuracy. I intend to make our method and model open-source and hope to open up future opportunities for GUI-related research.

3.2 Simulating User Interactions for GUI Representation

Our current model is only designed for free-viewing eye tracking within GUIs. My future research can extend to scope to solve task-based eye tracking problems, such as visual search tasks on GUIs. This extension will help us better understand human behaviors and interactions on GUIs, especially predict how users perform tasks while navigating through GUIs. Future work can also explore gaze patterns on GUIs with dynamic elements on GUI transitions and other types of user interactions on GUIs, such as text entry.

3.3 Language Representation for GUIs

I plan to improve both the quality of the language dataset on GUIs and the vision-language model to create a better language agent for GUIs. I will analyze our current dataset and create synthetic data based on the weaknesses, such as reducing hallucination and including more types of tasks for GUIs. On the other hand, I plan to improve the performance by building better vision-language models that can accept high-resolution GUIs. In addition, I also plan to support machine-interpretable output, such as JSON. For example, applications like UI navigation and software testing can benefit from machine-interpretable output.

4 DISSERTATION STATUS AND LONG TERM GOALS

I am currently a PhD student at Aalto University and the Finnish AI Center (FCAI) in Finland under the supervision of Professor Antti Oulasvirta (primary) and Professor Vikas Garg. I aim to have my Ph.D. defense in the spring of 2025. I have never attended any other doctoral consortium. My long-term goal is to get a tenure-track faculty job or be a research scientist in a relevant industrial group.

5 BENEFITS AND CONTRIBUTION STATEMENT

I hope to gain valuable feedback for my dissertation direction from the mentors and peers at the CHI2024 Doctoral Consortium. I think it is an opportunity to engage in discussion with peers to refine my research directions. In addition, I will be honored to attend the CHI2024 Doctoral Consortium to connect with other Ph.D. students working on different topics of HCI to expand my network within the academic community and potentially explore future collaborations. I will actively engage in discussions, provide feedback to other attendees, and communicate with them. Together with other peers, I will help create a supportive community and keep the connection after the Doctoral Consortium.

6 ACKNOWLEDGMENT

I would like to thank my supervisor, Professor Antti Oulasvirta, and all the collaborators who provided invaluable guidance and support throughout my Ph.D. Additionally, I am thankful for the enjoyable internship experience at Apple, where I was mentored by Jeffrey Nichols, Eldon Schoop, and Amanda Swearngin. I also extend my appreciation for the long-term support from Wolfgang Stuerzlinger and Christof Lutteroth. This work was supported by Aalto University's Department of Information and Communications Engineering, the Research Council of Finland (flagship program: Finnish Center for Artificial Intelligence, FCAI, grants 328400, 345604, 341763; Human Automata, grant 328813; Subjective Functions, grant 357578), and the Meta PhD Fellowship.

REFERENCES

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. arXiv:2204.14198 [cs.CV]

[2] Gary Ang and Ee-Peng Lim. 2022. Learning and Understanding User Interface Semantics from Heterogeneous Networks with Multimodal and Positional Attributes. *ACM Trans. Interact. Intell. Syst.* (dec 2022). https://doi.org/10.1145/3578522 Just Accepted.

[3] Michelle A Borkin, Zoya Bylinskii, Nam Wook Kim, Constance May Bainbridge, Chelsea S Yeh, Daniel Borkin, Hanspeter Pfister, and Aude Oliva. 2015. Beyond memorability: Visualization recognition and recall. *IEEE transactions on visualization and computer graphics* 22, 1 (2015).

[4] 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 *Proc. UIST*.

[5] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023. PaLI: A Jointly-Scaled Multilingual Language-Image Model. arXiv:2209.06794 [cs.CV]

[6] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)* (2023).

[7] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500 [cs.CV]

[8] Biplab Deka, Zifeng Huang, and Ranjitha Kumar. 2016. ERICA: Interaction Mining Mobile Apps. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 767–776. https://doi.org/10.1145/2984511.2984581

[9] Morgan Dixon and James Fogarty. 2010. Prefab: Implementing Advanced Behaviors Using Pixel-Based Reverse Engineering of Interface Structure. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 1525–1534. https://doi.org/10.1145/1753326.1753554

[10] Krzysztof Gajos and Daniel S. Weld. 2004. SUPPLE: Automatically Generating User Interfaces. In *Proceedings of the 9th International Conference on Intelligent User Interfaces* (Funchal, Madeira, Portugal) *(IUI '04)*. Association for Computing Machinery, New York, NY, USA, 93–100. https://doi.org/10.1145/964442.964461

[11] Krzysztof Z. Gajos, Daniel S. Weld, and Jacob O. Wobbrock. 2010. Automatically Generating Personalized User Interfaces With Supple, In Proceedings of the 9th International Conference on Intelligent User Interfaces. *Artif. Intell* 174, 12-13, 910–950. https://doi.org/10.1016/j.artint.2010.05.005

[12] Krzysztof Z. Gajos, Jacob O. Wobbrock, and Daniel S. Weld. 2008. Improving the Performance of Motor-Impaired Users With Automatically-Generated, Ability-Based Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. ACM, 1257–1266. https://doi.org/10.1145/1357054.1357250

[13] Lena Hegemann, Yue Jiang, Joon Gi Shin, Yi-Chi Liao, Markku Laine, and Antti Oulasvirta. 2023. Computational Assistance for User Interface Design: Smarter Generation and Evaluation of Design Ideas. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–5.

[14] Forrest Huang, John F. Canny, and Jeffrey Nichols. 2019. Swire: Sketch-Based User Interface Retrieval. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3290605.3300334

[15] Yue Jiang, Ruofei Du, Christof Lutteroth, and Wolfgang Stuerzlinger. 2019. ORC Layout: Adaptive GUI Layout with OR-Constraints. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300643

[16] Yue Jiang, Luis A. Leiva, Hamed Rezazadegan Tavakoli, Paul R. B. Houssel, Julia Kylmälä, and Antti Oulasvirta. 2023. UEyes: Understanding Visual Saliency across User Interface Types. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 285, 21 pages. https://doi.org/10.1145/3544548.3581096

[17] Yue Jiang, Luis A Leiva, Hamed Rezazadegan Tavakoli, Paul RB Houssel, Julia Kylmälä, and Antti Oulasvirta. 2023. UEyes: An Eye-Tracking Dataset across User Interface Types. In *Workshop Paper at the 2023 CHI Conference on Human Factors in Computing Systems*.

[18] Yue Jiang, Yuwen Lu, Clara Kliman-Silver, Christof Lutteroth, Toby Jia-Jun Li, Jeffrey Nichols, and Wolfgang Stuerzlinger. 2024. Computational Methodologies for Understanding, Automating, and Evaluating User Interfaces. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*.

[19] Yue Jiang, Yuwen Lu, Christof Lutteroth, Toby Jia-Jun Li, Jeffrey Nichols, and Wolfgang Stuerzlinger. 2023. The Future of Computational Approaches for Understanding and Adapting User Interfaces. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 367, 5 pages. https://doi.org/10.1145/3544549.3573805

[20] Yue Jiang, Yuwen Lu, Jeffrey Nichols, Wolfgang Stuerzlinger, Chun Yu, Christof Lutteroth, Yang Li, Ranjitha Kumar, and Toby Jia-Jun Li. 2022. Computational Approaches for Understanding, Generating, and Adapting User Interfaces. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 74, 6 pages. https://doi.org/10.1145/3491101.3504030

[21] Yue Jiang, Eldon Schoop, Amanda Swearngin, and Jeffrey Nichols. 2023. ILuvUI: Instruction-tuned LangUage-Vision modeling of UIs from Machine Conversations. arXiv:2310.04869 [cs.HC]

[22] Yue Jiang, Wolfgang Stuerzlinger, and Christof Lutteroth. 2021. ReverseORC: Reverse Engineering of Resizable User Interface Layouts with OR-Constraints. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 316, 18 pages. https://doi.org/10.1145/3411764.3445043

[23] Yue Jiang, Wolfgang Stuerzlinger, Matthias Zwicker, and Christof Lutteroth. 2020. ORCSolver: An Efficient Solver for Adaptive GUI Layout with OR-Constraints. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376610

[24] Yue Jiang, Changkong Zhou, Garg Vikas, and Antti Oulasvirta. 2024. Graph4GUI: Graph Neural Networks for Representing Graphical User Interfaces. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3613904.3642822

[25] Hsin-Ying Lee, Lu Jiang, Irfan Essa, Phuong B. Le, Haifeng Gong, Ming-Hsuan Yang, and Weilong Yang. 2020. Neural Design Network: Graphic Layout Generation with Constraints. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III* (Glasgow, United Kingdom). Springer-Verlag, Berlin, Heidelberg, 491–506. https://doi.org/10.1007/978-3-030-58580-8_29

[26] Luis A Leiva, Yunfei Xue, Avya Bansal, Hamed R Tavakoli, Tuðçe Köroðlu, Jingzhou Du, Niraj R Dayama, and Antti Oulasvirta. 2020. Understanding visual saliency in mobile user interfaces. In *22nd International conference on human-computer interaction with mobile devices and services*. 1–12.

[27] Jianan Li, Jimei Yang, Jianming Zhang, Chang Liu, Christina Wang, and Tingfa Xu. 2021. Attribute-Conditioned Layout GAN for Automatic Graphic Design. *IEEE Transactions on Visualization and Computer Graphics* 27, 10 (oct 2021), 4039–4048. https://doi.org/10.1109/TVCG.2020.2999335

[28] Toby Jia-Jun Li, Amos Azaria, and Brad A. Myers. 2017. SUGILITE: Creating Multimodal Smartphone Automation by Demonstration. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6038–6049. https://doi.org/10.1145/3025453.3025483

[29] Toby Jia-Jun Li, Amos Azaria, and Brad A. Myers. 2017. SUGILITE: Creating Multimodal Smartphone Automation by Demonstration. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6038–6049. https://doi.org/10.1145/3025453.3025483

[30] Toby Jia-Jun Li, Jingya Chen, Haijun Xia, Tom M. Mitchell, and Brad A. Myers. 2020. Multi-Modal Repairs of Conversational Breakdowns in Task-Oriented Dialogs. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 1094–1107. https://doi.org/10.1145/3379337.3415820

[31] Toby Jia-Jun Li, Lindsay Popowski, Tom Mitchell, and Brad A Myers. 2021. Screen2Vec: Semantic Embedding of GUI Screens and GUI Components. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 578, 15 pages. https://doi.org/10.1145/3411764.3445049

[32] Toby Jia-Jun Li and Oriana Riva. 2018. KITE: Building conversational bots from mobile apps. In *Proceedings of the 16th ACM International Conference on Mobile Systems, Applications, and Services (MobiSys 2018)*. ACM.

[33] Yang Li, Jiacong He, Xin Zhou, Yuan Zhang, and Jason Baldridge. 2020. Mapping Natural Language Instructions to Mobile UI Action Sequences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL, Online, 8198–8210. https://doi.org/10.18653/v1/2020.acl-main.729

[34] Yang Li, Jiacong He, Xin Zhou, Yuan Zhang, and Jason Baldridge. 2020. Mapping Natural Language Instructions to Mobile UI Action Sequences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 8198–8210. https://doi.org/10.18653/v1/2020.acl-main.729

[35] Yang Li, Gang Li, Luheng He, Jingjie Zheng, Hong Li, and Zhiwei Guan. 2020. Widget Captioning: Generating Natural Language Description for Mobile User Interface Elements. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, Online, 5495–5510. https://doi.org/10.18653/v1/2020.emnlp-main.443

[36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485* (2023).

[37] Thomas F. Liu, Mark Craft, Jason Situ, Ersin Yumer, Radomir Mech, and Ranjitha Kumar. 2018. Learning Design Semantics for Mobile Apps. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (Berlin, Germany) *(UIST '18)*. Association for Computing Machinery, New York, NY, USA, 569–579. https://doi.org/10.1145/3242587.3242650

[38] Yuwen Lu, Chengzhi Zhang, Iris Zhang, and Toby Jia-Jun Li. 2022. Bridging the Gap between UX Practitioners' work practices and AI-enabled design support tools. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7.

[39] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155 [cs.CL]

[40] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277* (2023).

[41] Erica Sadun. 2013. *iOS Auto Layout Demystified*. Addison-Wesley Professional, Boston, US.

[42] Alborz Rezazadeh Sereshkeh, Gary Leung, Krish Perumal, Caleb Phillips, Minfan Zhang, Afsaneh Fazly, and Iqbal Mohomed. 2020. VASTA: A Vision and Language-Assisted Smartphone Task Automation System. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 22–32. https://doi.org/10.1145/3377325.3377515

[43] Amanda Swearngin, Amy J. Ko, and James Fogarty. 2017. Genie: Input Retargeting on the Web through Command Reverse Engineering. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 4703–4714. https://doi.org/10.1145/3025453.3025506

[44] Amanda Swearngin, Chenglong Wang, Alannah Oleson, James Fogarty, and Amy J. Ko. 2020. *Scout: Rapid Exploration of Interface Layout Alternatives through High-Level Design Constraints*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376593

[45] Maryam Taeb, Amanda Swearngin, Eldon School, Ruijia Cheng, Yue Jiang, and Jeffrey Nichols. 2023. AXNav: Replaying Accessibility Tests from Natural Language. *arXiv preprint arXiv:2310.02424* (2023).

[46] 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).

[47] Bryan Wang, Gang Li, and Yang Li. 2023. Enabling conversational interaction with mobile ui using large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[48] Hao Wen, Yuanchun Li, Guohong Liu, Shanhui Zhao, Tao Yu, Toby Jia-Jun Li, Shiqi Jiang, Yunhao Liu, Yaqin Zhang, and Yunxin Liu. 2023. Empowering LLM to use Smartphone for Intelligent Task Automation. *arXiv preprint arXiv:2308.15272* (2023).

[49] Hao Wen, Hongming Wang, Jiaxuan Liu, and Yuanchun Li. 2023. DroidBot-GPT: GPT-powered UI Automation for Android. *arXiv preprint arXiv:2304.07061* (2023).

[50] Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. 2015. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755* (2015).

[51] Xiaoyi Zhang, Lilian de Greef, Amanda Swearngin, Samuel White, Kyle Murray, Lisa Yu, Qi Shan, Jeffrey Nichols, Jason Wu, Chris Fleizach, et al. 2021. Screen recognition: Creating accessibility metadata for mobile applications from pixels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[52] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685* (2023).