Visual Explanation for Advertising Creative Workflow

Shoko Sawada
shoko@tkl.iis.u-tokyo.ac.jp
The University of Tokyo
Tokyo, Japan

Tomoyuki Suzuki
suzuki_tomoyuki@cyberagent.co.jp
CyberAgent
Tokyo, Japan

Kota Yamaguchi
CyberAgent
Tokyo, Japan

Masashi Toyoda
Institue of Industrial Science, The University of Tokyo
Tokyo, Japan

ABSTRACT

Explainable AI (XAI) attempts to produce interpretable results from highly complex AI systems, but its form and effectiveness vary depending on the application domain. In this paper, we explore how XAI techniques can help graphic designers work on advertising materials. A creative domain such as graphic design is often characterized by a weak connection between the individual work and the business goal; e.g., a small change in the design of a banner can result in a huge difference in the audience's reaction. We develop an XAI system for designers that provides visual feedback explaining which component of the design is likely to affect the business metric. Our user study shows that with our system, designers complete the project in fewer iterations and in less time to achieve the desired quality of work compared to naive score-based feedback. These findings highlight the benefits of leveraging XAI in creative domains.

CCS CONCEPTS

• **Human-centered computing** → **Visualization application domains**; **Empirical studies in visualization**; • **Computing methodologies** → *Graphics systems and interfaces*; • **Applied computing** → Arts and humanities.

KEYWORDS

XAI, Visual explanation, Graphic design

ACM Reference Format:
Shoko Sawada, Tomoyuki Suzuki, Kota Yamaguchi, and Masashi Toyoda. 2024. Visual Explanation for Advertising Creative Workflow. 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, 8 pages. https://doi.org/10.1145/3613905.3651123

1 INTRODUCTION

With the rise of the Internet, the online advertising market has experienced rapid growth. Machine learning models have played a pivotal role in this growth, facilitating tasks such as targeting audience [26], generating marketing copies [13], or bidding optimization [21, 31]. One of the most common and important applications in online advertising is the prediction of performance metrics such as click-through rate (CTR) or conversion rate (CVR) for creative work [4, 11, 18, 24, 30]. Advertising agencies use such machine learning models for quality assessment in the creative workflow, where designers receive instant feedback from the performance prediction models on the quality of banner designs before the banners are delivered to advertising platforms, such as social media. While this AI-assisted workflow helps designers of all skill levels produce high-quality work, naive machine-learning applications only provide numerical or categorical scores without any reasons or factors. It can often be difficult for designers to interpret and force designers to repeat the trial-and-error process to identify the key scoring factor.

We believe XAI techniques can address this interpretation problem in the advertising creative workflow. Previous studies have highlighted how explanations support humans in collaborative tasks with AI, spanning domains such as text and tabular data classification [2, 27], childcare [33], and healthcare [1]. The particular challenge in the creative workflow is its inherently creative nature; there is no definitive answer, and multiple graphic designs are acceptable. In addition, in online advertising, the business outcome depends on what campaign the project is concerned with, where and when the ad is being served, and which audience segment is being targeted, which is difficult for designers to account for. While XAI has been studied for creative tasks, such as UX design [14], games [32], or arts [3], we find that our advertising creative workflow presents a unique task setup to explore the use of the XAI approach.

This paper presents a case study of the XAI application in the advertising creative workflow. We developed a visual feedback system that provides a detailed assessment of the banner design quality, as shown in Figure 1. Our system accepts banner designs in Adobe Photoshop format and presents a quality score and visual feedback of the banners. During the development, we extensively reviewed online advertising strategies and interviewed professional graphic designers to identify the requirements for our system. We evaluated our system through a user study where we asked designers to create a banner with a quality score higher than existing banners with our system. We recorded each step and the duration of each designer's iteration until they reached the goal. The results show that our feedback system successfully supports designers in

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

ACM ISBN 979-8-4007-0331-7/24/05
https://doi.org/10.1145/3613905.3651123

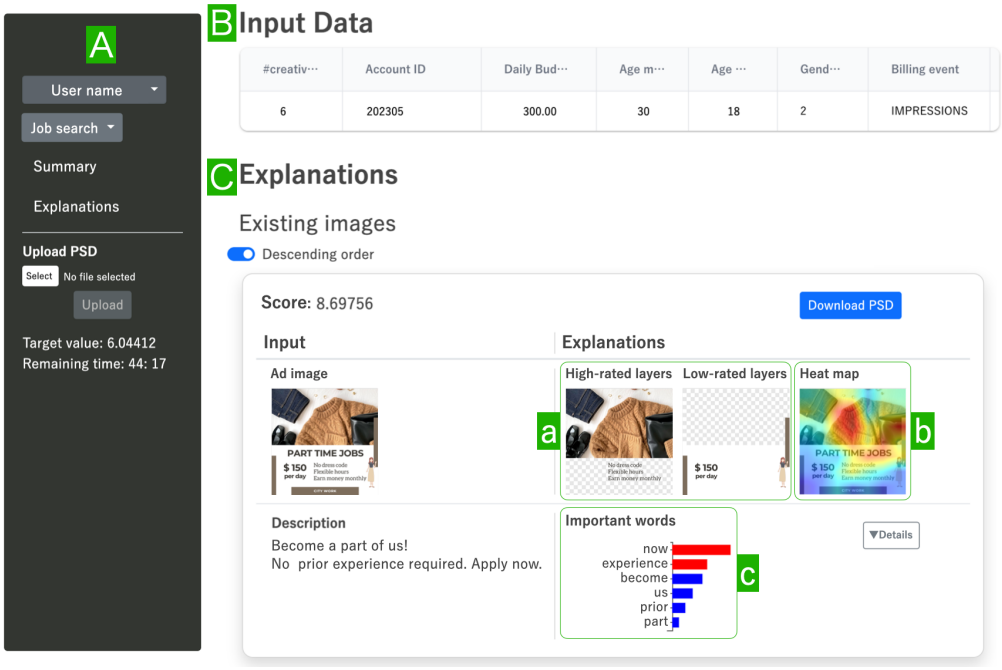

Figure 1: An overview of our visual feedback system for designers. For a campaign project, a designer uploads the current banner design and gets the quality score and an explanation of what element contributed to the score. (A) The operation panel equips designers with essential functions to select a campaign project and upload banner images. (B) The configuration table displays campaign details, including target audiences. (C) The explanation view presents the banners and their visual feedback, such as layer-based explanations, heat maps, and word-based explanations. A single campaign project can contain multiple banners, but we show only one banner here for simplicity.

reaching the goal with fewer steps and less time and in identifying design culprits.

In summary, our study provides the following contributions: the XAI application in the advertising domain, the user study to identify the desired feedback approach, and the empirical insights into how XAI techniques work in the practical creative workflow.

2 BACKGROUND

In our advertising agency, graphic designers create a campaign banner with score-based feedback from a regression model that predicts business metrics such as CTR or CVR, which we call a quality score. The model takes a banner image, associated textual description, and targeting configurations such as audience demographics or budget and predicts the quality score. Figure 2 illustrates how the model provides feedback to designers. The model is a neural network learned from historical business data and has been in production since 2020. It has three encoders, each tailored to a specific modality: a Convolutional Neural Network (CNN) for banner images, a Long Short-term Memory (LSTM) for associated text, and a Multi-Layer Perceptron (MLP) for target configuration. The training data consists of 100,000 ads actually delivered between June 2017 and August 2023. Each ad has a quality score based on its actual performance on the business metric. Our score-based feedback system helps designers at all levels reach the required quality of work before the actual delivery on the advertising platform. In this paper, we specifically focus on the *refinement* workflow where designers improve the existing banner designs during the campaign. This is a common scenario in campaigns where the audience gets bored by repeatedly seeing the same banner.

While score-based feedback effectively assesses the quality standard, designers sometimes struggle to identify influential design components and how to modify them to improve the score. Several studies have tried to interpret quality prediction models. Xia et al. visualized self-attention to identify the principal modality of inputs but did not focus on influential parts of respective banner images [29]. Others utilized Grad-CAM [25] to visualize crucial regions in banner images and discovered that text information notably affected the model's predictions [19]. However, these studies mainly aid model developers in understanding what the model has learned, unlike our study, which focuses on assisting designers in enhancing advertising effectiveness.

3 RESEARCH QUESTIONS

The following research questions emerged in the context of incorporating XAI techniques into the task of improving online advertisements.

Q1. Is an XAI approach effective for advertising creative workflow? Existing literature [2, 16] indicates that explanations could potentially mislead users by enhancing the human's trust. Although Feng et al. [6] showcase that experts tend to navigate explanations more

Figure 2: AI-assisted advertising creative workflow. The feedback model takes a banner image, associated text, and targeting configuration and predicts a quality score.

adeptly than non-expert users in a question-answering context, there are limited studies addressing creative domains. We wonder if XAI techniques can accelerate the refinement task where the goal is to improve the existing banner design to outperform the original quality score.

Q2. Does an XAI approach harm creativity? Another question that naturally arises in applying an XAI approach to creative workflow is the negative effect on creativity. A pertinent concern in the realm of AI-enabled workflow is the potential homogenization of resulting designs. If designers uniformly act on explanations, the resulting outputs may lack distinctiveness and diversity. We would like to answer if this concern is true.

4 CHOICE OF XAI APPROACHES

To determine the requirements of our system, we listed candidates of XAI approaches and had an interview session with five designers; a detailed account of the candidate approaches and the interview can be found in the appendix (Section A). We selected the following two XAI approaches for implementation based on the invaluable feedback.

Activation maps: Activation maps [25] highlight the regions of an image that are important for a neural network's prediction to help users discern which parts of the input image most influence the classification outcome. Other studies[19] employ it for model developers.

Feature-based approach: Banner images typically consist of multiple layered components: a background, a product image, a marketing copy, and other decorative elements, as illustrated in Figure 2. Feature-based explanation methods can identify influential features in the model output, such as LIME [22] and SHAP [17]. We apply these methods by treating each layer as a feature so that the XAI system can present which layers impact the quality score most.

5 PROTOTYPE SYSTEM

As Figure 1 shows, our system lets designers explore visual explanations within a particular campaign project during the refinement task. The system comprises three components: (A) the operation panel, (B) the configuration table, and (C) the explanation view. In the operation panel, designers select assigned projects from the dropdown lists and upload refined banners in Adobe Photoshop Data (PSD) format. The configuration table shows the metadata of a selected project so that designers can refer to details about the target audiences or other information about the project.

The explanation view serves as a space to display all banners within the selected project in a row format. Each banner and its visual feedback reside within an individual panel. Designers can scroll through the view and browse all panels. By default, these panels are organized based on the quality scores in descending order, as designers usually refine banners based on banners with higher scores. Each explanation panel shows the predicted quality score at the top and a button to download the PSD file. Designers can download existing banners when they create a new banner. The second row of each panel presents a banner and its visual feedback side by side. Our system generates both feature-based explanations and activation maps.

For the activation map, we incorporate Grad-CAM [25], which was originally developed for classification models. We adopt Grad-CAM for our regression model and render the explanations as heat maps as shown in Figure 1(b). The color map is *jet*, where red regions indicate increased value contributions and blue areas indicate decreased value contributions. On the feature-based front, we adopted LIME [22] and designed high-rated and low-rated layers (Figure 1(a)). LIME calculates the importance of layers, and we selected the layers with the top N and bottom N levels of importance as the high- and low-rated layers, respectively. High-rated layers positively affect the quality score, and low-rated layers are the opposite. We stack and present them in a single image, respectively. In this paper, we set n to six to balance informativeness and conciseness. The system also has two supplementary presentations of the layer-based explanation. Hovering over the high-rated or low-rated layers displays further details about each layer and its importance of both high-rated and low-rated layers. When calculating important layers, LIME generates perturbation images by randomly removing several layers from the original image and then inputs them into the model to obtain their quality score. An auxiliary on-demand view displays the perturbation data to facilitate designers' understanding of the layers influencing the quality score.

Table 1: Participant profiles. We excluded Designer 6 from the study as the designer misunderstood the setting.

Designers	Designer 1	Designer 2	Designer 3	Designer 4	Designer 5	Designer 6 (excluded)
Gender	female	female	female	female	female	male
Years of experience	3 years	2 years	3 years	3 years	1.5 years	26 years

To help designers develop appealing marketing copies embedded within a banner, we also utilize LIME to identify influential words for the quality score. Here, we present text associated with the banner (See Figure 2) instead of the copy inside the image. These texts usually span one to two sentences and offer more content than in-banner copies. We show the associated text and LIME's output in the third row of each panel. The bar chart shown in Figure 1(c) visually presents the importance of each word, with words that enhance the quality score depicted in red, and those that diminish the score shown in blue.

6 USER STUDY

To evaluate the practical impact of our XAI system within the refinement task, we organized a user study. It replicated a typical AI-enabled creative workflow.

6.1 Refinement task

Setting. We asked designers for the task of refining existing banners. Refinement is considered successful if the quality score of a new banner exceeds the target score, which is determined by the highest quality score among the existing banners in each campaign project. Primary assets such as product photos and marketing copies are typically designated by clients in real-world scenarios. We mimic this situation by reusing the main photo from those in existing banners within the same project. We asked designers to work assuming there is no specific restriction regarding the copies and other elements. All designers use Adobe Photoshop for the refinements. Designers can upload refined banners to our Web-based feedback system at their convenience without limitation.

We set a maximum work time of 45 minutes for each project, which is sufficient for most cases in our environment. Our system actively displays a countdown timer in the control panel (Figure 1 A) and notifies designers either when the quality score exceeds the target or when time runs out. At this point, we mark the task as completed. We provided 5 business days to complete this task.

Evaluation protocol. To address *Q1*, we monitored the quality scores of refined banners and the designers' operational times, enabling us to evaluate the refinement efficacy in terms of both the number of iterations and the duration of the process. The number of iterations was measured by the count of iterative edits during a single project, and the work duration was the time from the first material download to the final banner upload.

To address *Q2*, we calculated the diversity among the refined banners. Given a group of refined banners for a certain project, we measure the mean distance between all pairs of banners in that group as the diversity score. As image distance, we used the L1 distance of image features and used ResNet-50 [10] pre-trained by CLIP [20] for image feature extraction.

We compare our prototype system with the plain score-based baseline. The baseline displays the quality scores of existing banners and runs inference when designers upload a new banner.

Dataset and participants. For our study, we selected six campaign projects to be delivered to a specific social media platform. Each campaign project has six to eleven banners. We invited six professional designers. Table 1 shows participant profiles. We select participants based on their recent usage frequency of the score-feedback system. We asked each participant to work on three campaign projects using our XAI system and the remaining three with the baseline system. In total, we obtained 36 refinement trials from all the participants. Prior to the task, the participants attended a 45-minute orientation session about the refinement task and our XAI system. Upon completing the task, we requested the participants to provide feedback via a questionnaire and an interview. Our study is approved by the REB review process.

6.2 Results and analysis

6.2.1 Efficacy evaluation. Figure 3 (a) plots the cumulative distribution of completion over the number of iterations, and Figure 3 (b) plots the cumulative distribution of completion over the work duration. Note that we exclude 6 cases worked by Designer 6 (D6) from these plots because subsequent interviews revealed that D6 mistakenly assumed that they were restricted from modifying any given images and copies. The remaining 30 cases achieved the target score within the time limit.

From Figure 3 (a), we can observe that participants completed tasks in fewer iterations using our system. Notably, in 9 out of 15 cases, participants reached target scores at the first iteration using our visual feedback system, compared to just 2 cases with the baseline system. As for the cumulative work duration (Figure 3 (b)), despite the additional overhead that the visual feedback system needs to generate explanations, the overall trend leaned toward reduced work duration when compared to the baseline system. In our setup, the baseline took 5 to 10 seconds per upload, and our system took up to 30 seconds to present all the visual feedback.

6.2.2 Diversity evaluation. We summarize the diversity scores in Table 2. We performed a Welch's t-test between the two systems' diversity scores and found no significant differences between them (p-value = 0.7092). We conjecture that using XAI does not necessarily harm the creativity of the human designers from this experiment.

6.2.3 User satisfaction. We asked all the participants to answer 5 questions after the refinement task. We collected responses on a 5-point Likert scale, which included the options: Agree, Somewhat Agree, Neutral, Somewhat Disagree, and Disagree. Table 3 summarizes the results of the questionnaire. The XAI system received significantly higher ratings in the questions about whether

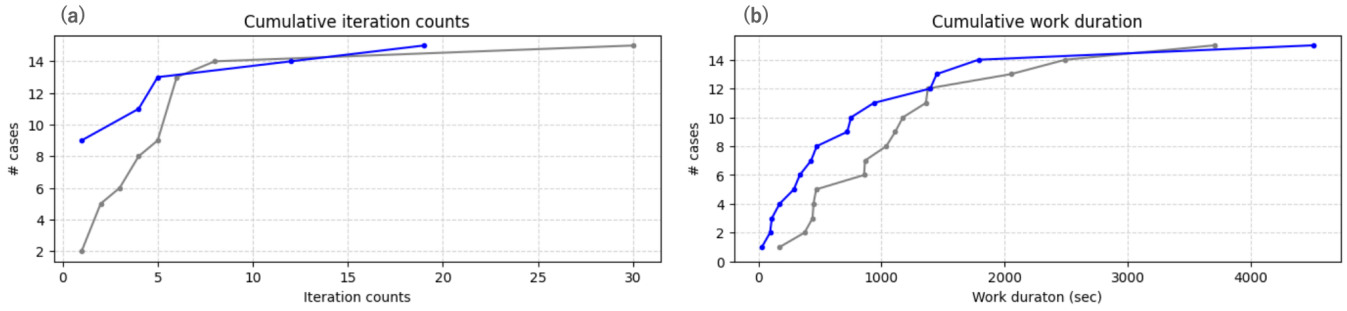

Figure 3: The cumulative distributions of the number of iterations (a) and work duration (b) of each trial case. The blue lines represent the distributions from our visual feedback system, and the gray lines represent those from the baseline system.

Table 2: Diversity score summary. The higher the value, the more diverse the result.

	Project 1	Project 2	Project 3	Project 4	Project 5	Project 6	Average
Score-based feedback	26.35	23.59	18.55	32.79	32.29	43.40	29.66
Our visual feedback	27.60	30.51	43.68	29.00	25.05	31.35	31.22

designers understood where and how to improve, as shown in the first two rows. All participants gave positive responses to our visual feedback. In contrast, all designers rated the baseline system as Neutral or below in the question of whether they understood where to improve, which confirmed the designers' dissatisfaction with the current score-based feedback. Several designers mentioned in the interview that with the baseline method, they typically begin the refinement by making trivial changes repeatedly to identify areas impacting the score. With our visual explanation, participants praised the ability to show where refinements would influence the scores immediately, and they were able to increase the scores by modifying or replacing low-rated layers.

We could not identify a statistically significant difference between the two systems regarding overall satisfaction, as shown in the last three rows in Table 3. However, everyone responded that they want to use our visual feedback system in the future, which indicates their great satisfaction with our system.

6.2.4 Usage analysis. We discuss how designers utilized our visual feedback system. We share their preference, feedback, and the strategies they applied to illustrate the system's practical application in the refinement process.

Preference for XAI approaches. We asked the designers about their preferences and satisfaction with the layer-based and heat map explanations. While designers utilized the layer-based explanations more often, the heat maps received higher satisfaction ratings. Two designers found that though they prefer layer-based explanations, sometimes the layer-based method suggested trivial and unhelpful layers, such as a part of decorations, and the heat map was particularly useful in those cases. The other designer responded that it was still better to see multiple types of visual feedback even when they show conflicting results. Based on the designers' feedback, while the layer-based explanations highlighted layers that clearly should be refined, the heat maps seemed to suggest areas that they would not have thought to change without the visual feedback. We speculated that gaining such clear insights might contribute to the high satisfaction with the heat map in the questionnaire.

The most effective or desirable method seems to depend on the input banner, suggesting that the two XAI techniques we have chosen help designers in complementary ways.

Refinement strategy. To understand how designers employ visual feedback, we conducted interviews following their completion of the questionnaire. Designer 3 (D3) struggled to achieve the target score in one project. D3 found that a copy text was in the low-rated layers and began changing the words in the copy. The score eventually exceeded the target after D3 removed the entire copy. Despite taking longer than other projects, D3 expressed satisfaction with the visual feedback system, as it eliminated the need for an inefficient trial-and-error process to identify areas to modify.

Designer 4 (D4), who completed two projects with the first uploads using our system, said that D4 relied exclusively on word-based explanations for all refinements. This was unexpected, as we considered the word importance as an adjunct to image explanations. Typically, when improving copies, designers handpick words that are known to be effective, such as *sale, discount*, and *best*, and display them in large. D4 noticed that those words were not always highlighted as the important words, and D4 reduced the font size of those words and enlarged the important words instead, which led to immediate score improvements. The ability to produce effective copies from word-based explanations beyond commonly known words is a significant discovery.

7 LESSONS LEARNED

In this study, we developed a visual feedback system to help designers improve advertising banners. The system employs two explanation methods and provides various forms of visual feedback. The user study results indicate that designers more efficiently refined banners using our visual feedback system than conventional score-based feedback and expressed considerable satisfaction with their

Table 3: The results of the questionnaire. We collect responses from six designers using a 5-point Likert scale, where 5 indicates agreement, and 1 indicates disagreement. The average scores were then subjected to a t-test analysis.

Question	Average (Score-based feedback)	Average (Visual feedback)	p-value
Did you understand which layer to edit to increase the predicted effectiveness value?	2.17	4.5	0.00525
Did you understand how to edit to increase the predicted effectiveness value?	2.67	4.67	0.000573
Did you find the tool effective in improving ad images?	4.0	4.67	0.102
Did you find the tool helpful in improving ad images efficiently?	3.67	4.83	0.0583
Would you consider using the tool in your future work?	3.5	5	0.0599

experience. It was also conjectured that the use of visual feedback did not necessarily have a negative effect on design diversity.

While a previous study [33] revealed that users may not use all the features of a versatile XAI visualization system, the designers enthusiastically embraced various types of feedback in our system and reasoned with them, thereby exploring the refinement ideas. Notably, even when designers faced with conflicts between distinct explanations, they showed a keen ability to select proper explanations by utilizing their expertise in design. From the above, we believe that in the AI-assisted creative workflow, it is important to design a system that relies on and amplifies designers' reasoning ability, rather than just aiming to provide limited types of accurate feedback from the system.

As described in Section 6.2.4, some designers utilized visual feedback in the way we did not expect. This underscores the importance of iteratively refining the prototype system based on designer feedback and user evaluation.

As a budding effort to apply XAI to creative workflows, this study demonstrated its effectiveness and provided the empirical evidence to stimulate future research.

8 LIMITATIONS AND FUTURE WORK

Our study has several limitations, while it provides insights on using XAI techniques in banner design tasks for online advertising. Firstly, this study was conducted with a specific advertising system within a single agency, and the scale of the user study remained small in terms of the number of designers and the variety of campaign projects. User studies with diverse advertising systems, more designers, and larger data sets are needed for further analysis.

Another limitation is that we organize the layers of the banners in a specific format before the study. While PSD files can have arbitrary layers, we group layers into hierarchies such as copies, decorations, the main product, or the background to prevent our feature-based feedback from unintentionally suggesting meaningless minor layers to the designers. We would like to see how the results change when different formats are used in the future. The results with the single model may have some limitations because the impact of XAI techniques on users may vary depending on the target models. We plan to investigate the sensitivity of XAI techniques with several variations of the model.

We haven't conducted a distinct examination of the impact of each explanation method. Future endeavors should consider these limitations and potential improvements to our approach. Our system supports identifying where to edit but cannot fully assist designers in suggesting refinement plans. Presenting alternative options for low-rated layers may be more effective, sourced from stock materials or AI-generated images.

REFERENCES

[1] Dario Antweiler and Georg Fuchs. 2022. Visualizing Rule-based Classifiers for Clinical Risk Prognosis. In *2022 IEEE Visualization and Visual Analytics (VIS)*. IEEE, Sankt Augustin Germany, 55–59. https://doi.org/10.1109/VIS54862.2022.00020

[2] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, USA, Article 81, 16 pages. https://doi.org/10.1145/3411764.3445717

[3] Nick Bryan-Kinns, Corey Ford, Alan Chamberlain, Steven David Benford, Helen Kennedy, Zijin Li, Wu Qiong, Gus G. Xia, and Jeba Rezwana. 2023. Explainable AI for the Arts: XAIxArts. In *Proceedings of the 15th Conference on Creativity and Cognition (C&C '23)*. ACM, New York, NY, USA, 1–7. https://doi.org/10.1145/3591196.3593517

[4] Junxuan Chen, Baigui Sun, Hao Li, Hongtao Lu, and Xian-Sheng Hua. 2016. Deep CTR Prediction in Display Advertising. In *Proceedings of the 24th ACM International Conference on Multimedia (MM '16)*. ACM, New York, NY, USA, 811–820. https://doi.org/10.1145/2964284.2964325

[5] Furui Cheng, Yao Ming, and Huamin Qu. 2020. DECE: Decision Explorer with Counterfactual Explanations for Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (Aug. 2020), 1438–1447. arXiv:2008.08353 [cs, stat]

[6] Shi Feng and Jordan Boyd-Graber. 2019. What Can AI Do for Me? Evaluating Machine Learning Interpretations in Cooperative Play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. ACM, New York, NY, USA, 229–239. https://doi.org/10.1145/3301275.3302265

[7] Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. 2021. AdViCE: Aggregated Visual Counterfactual Explanations for Machine Learning Model Validation. In *2021 IEEE Visualization Conference (VIS)*. IEEE, New York, NY, USA, 31–35. https://doi.org/10.1109/VIS49827.2021.9623271

[8] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual Visual Explanations. arXiv:1904.07451 [cs, stat]

[9] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. 2018. Local Rule-Based Explanations of Black Box Decision Systems. arXiv:1805.10820 [cs]

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, Redmond WA USA, 770–778. https://doi.org/10.1109/CVPR.2016.90

[11] Xinran He, Junfeng Pan, Ou Jin, Tianbing Xu, Bo Liu, Tao Xu, Yanxin Shi, Antoine Atallah, Ralf Herbrich, Stuart Bowers, and Joaquin Quiñonero Candela. 2014. Practical Lessons from Predicting Clicks on Ads at Facebook. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising* (New York, NY, USA) *(ADKDD'14)*. ACM, New York, NY, USA, 1–9. https://doi.org/10.1145/2648584.2648589

[12] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating Visual Explanations. In *Computer Vision – ECCV 2016 (Lecture Notes in Computer Science)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 3–19. https://doi.org/10.1007/978-3-319-46493-0_1

[13] Hidetaka Kamigaito, Peinan Zhang, Hiroya Takamura, and Manabu Okumura. 2021. An Empirical Study of Generating Texts for Search Engine Advertising. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*. Association for Computational Linguistics, Tokyo Japan, 255–262. https://doi.org/10.18653/v1/2021.naacl-industry.32

[14] Tiffany Knearem, Mohammed Khwaja, Yuling Gao, Frank Bentley, and Clara E Kliman-Silver. 2023. Exploring the Future of Design Tooling: The Role of Artificial Intelligence in Tools for User Experience Professionals. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. ACM, New York, NY, USA, Article 384, 6 pages. https://doi.org/10.1145/3544549.3573874

[15] Pang Wei Koh and Percy Liang. 2017. Understanding Black-box Predictions via Influence Functions. In *International Conference on Machine Learning*. PMLR, Stanford, CA, USA, 1885–1894. http://proceedings.mlr.press/v70/koh17a.html

[16] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Hum. Factors* 46, 1 (2004), 50–80. https://doi.org/10.1518/hfes.46.1.50.30392

[17] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., Seattle, WA, USA, 4765–4774. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[18] Junwei Pan, Jian Xu, Alfonso Lobos Ruiz, Wenliang Zhao, Shengjun Pan, Yu Sun, and Quan Lu. 2018. Field-Weighted Factorization Machines for Click-Through Rate Prediction in Display Advertising. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1349–1357. https://doi.org/10.1145/3178876.3186040

[19] Kyung-Wha Park, Jung-Woo Ha, JungHoon Lee, Sunyoung Kwon, Kyung-Min Kim, and Byoung-Tak Zhang. 2021. M2FN: Multi-step Modality Fusion for Advertisement Image Assessment. *Applied Soft Computing* 103 (May 2021), 107116. https://doi.org/10.1016/j.asoc.2021.107116

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, ACM, San Francisco, CA, USA, 8748–8763. http://proceedings.mlr.press/v139/radford21a.html

[21] Kan Ren, Weinan Zhang, Ke Chang, Yifei Rong, Yong Yu, and Jun Wang. 2018. Bidding Machine: Learning to Bid for Directly Optimizing Profits in Display Advertising. *IEEE Trans. Knowl. Data Eng.* 30, 4 (2018), 645–659. https://doi.org/10.1109/TKDE.2017.2775228

[22] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[23] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence*. AAAI Press, New York, NY, USA, 1527–1535. https://doi.org/10.1609/aaai.v32i1.11491

[24] Matthew Richardson, Ewa Dominowska, and Robert Ragno. 2007. Predicting Clicks: Estimating the Click-through Rate for New Ads. In *Proceedings of the 16th International Conference on World Wide Web* (Banff, Alberta, Canada) *(WWW '07)*. ACM, New York, NY, USA, 521–530. https://doi.org/10.1145/1242572.1242643

[25] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, Blacksburg, VA, USA, 618–626. https://doi.org/10.1109/ICCV.2017.74

[26] Jianqiang Shen, Sahin Cem Geyik, and Ali Dasdan. 2015. Effective Audience Extension in Online Advertising. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 2099–2108. https://doi.org/10.1145/2783258.2788603

[27] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces (IUI '21)*. ACM, New York, NY, USA, 318–328. https://doi.org/10.1145/3397481.3450650

[28] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2020. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (Jan. 2020), 56–65. https://doi.org/10.1109/TVCG.2019.2934619

[29] Bohui Xia, Xueting Wang, Toshihiko Yamasaki, Kiyoharu Aizawa, and Hiroyuki Seshime. 2019. Deep Neural Network-Based Click-Through Rate Prediction Using Multimodal Features of Online Banners. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*. IEEE, Tokyo, Japan, 162–170. https://doi.org/10.1109/BigMM.2019.00-29

[30] Shuangfei Zhai, Keng-hao Chang, Ruofei Zhang, and Zhongfei Mark Zhang. 2016. DeepIntent: Learning Attentions for Online Advertising with Recurrent Neural Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, New York, NY, USA, 1295–1304. https://doi.org/10.1145/2939672.2939759

[31] Weinan Zhang, Shuai Yuan, and Jun Wang. 2014. Optimal Real-Time Bidding for Display Advertising. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*. ACM, New York, NY, USA, 1077–1086. https://doi.org/10.1145/2623330.2623633

[32] Jichen Zhu, Antonios Liapis, Sebastian Risi, Rafael Bidarra, and G. Michael Youngblood. 2018. Explainable AI for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation. In *Proceedings of the IEEE Conference on Computational Intelligence and Games*. IEEE, Philadelphia, PA, USA, 1–8. https://doi.org/10.1109/CIG.2018.8490433

[33] Alexandra Zytek, Dongyu Liu, Rhema Vaithianathan, and Kalyan Veeramachaneni. 2022. Sibyl: Understanding and Addressing the Usability Challenges of Machine Learning In High-Stakes Decision Making. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (Jan. 2022), 1161–1171. https://doi.org/10.1109/TVCG.2021.3114864

A CHOICE OF XAI APPROACHES

In this section, we describe how we selected the XAI approaches for our system. Considering that our target users are more interested in understanding the rationale behind the prediction for a specific input rather than grasping the entire mechanics of the model, our focus was directed toward post-hoc and local explanations. While there are various XAI approaches, we initially picked six post-hoc local explanation methods.

A.1 Candidate approaches

Below, we list the initial selection of XAI approaches, from which we selected the activation maps and the feature-based approach.

Activation maps: Activation maps [25] highlight the regions of an image that are important for a neural network's prediction to help users discern which parts of the input image most influence the classification outcome.

Feature-based approach: Banner images typically consist of multiple layered components: a background, a product image, a marketing copy, and other decorative elements, as illustrated in Figure 2. Feature-driven explanation methods can identify influential features in the final quality score, like LIME [22] and SHAP [17]. We can apply these feature-based methods by treating each layer as a feature so that the XAI system can present which layer impacts the quality score when edited.

Rule-based approach: Rule-based methods [9, 23] translate models' internal logic into digestible if-then rules or decision trees. The rules specify essential features required for a particular classification prediction.

Counterfactual approach: The counterfactual method [8] revolves around the concept of "what if" scenarios in classification tasks by illustrating how altering a certain input would lead the model to make a different prediction.

Sample-based approach: Sample-based approach [15] elucidates the predictions by identifying specific training samples that greatly impact a given prediction, offering insight into foundational patterns the model detects.

Captioning approach: Captioning approach [12] leverages natural language to offer a concise and human-readable summary of the models' decision-making process.

A.2 Interview

To choose the appropriate methods, we interviewed five designers who generally use the score-based feedback system. We begin the session with an overview of how six techniques could be used to produce feedback, along with their respective advantages and disadvantages. We then asked the designers two questions about each explanation method: 1) Would you like to use an explanation in your AI design assistant? 2) Do you have any concerns about the explanation? If so, what are they?

The feedback revealed a strong preference for feature-based explanations and activation maps; all designers responded that they wanted to use these two explanation techniques. While activation maps were generally well received, some designers felt that they leaned more toward feature-based explanations, pointing out that highlighting pixels did not clearly identify the influential layers. One designer expressed concern that such explanations might unduly influence less experienced people. Interestingly, counterfactual explanations didn't resonate, with four out of five designers showing little interest. One designer commented on counterfactual explanations, "I prefer to rely on my own judgment on how to modify the images." Although end-user interfaces have actively adopted counterfactual explanations [5, 7, 28] for their user-friendly nature and actionable insights, they were not attractive to the designers in this task. Notably, although we hadn't specified our intention to integrate multiple explanations, three designers wanted to access both feature-based explanations and activation maps simultaneously.

B USER STUDY DETAILS

B.1 Task preparation

We calibrate the difficulty of the refinement task beforehand. Since our campaign projects may contain ads that have already undergone multiple refinements, their quality scores may be saturated, and it may be quite challenging to make further improvements. In this user study, we excluded the banners with quality scores in the top 1-sigma range in each campaign, and the target scores were set using the highest scores after exclusion.

As discussed in Section 8, our campaign project selection was deliberate. We intentionally left out projects where advertisers seemed to have strict guidelines for materials or layouts. This decision aligns with the design of our system, which encourages designers to explore explanations across different existing images to gain insights for improvement. We also avoided projects with many existing images, assuming that reviewing numerous explanations might overwhelm designers.

B.2 Interview feedback

This section depicts further details of the interview feedback of the user study (Section 6.2.4). D3, who faced the challenging project, completed the other two projects on the first iteration. The project that took D3 20 iterations proved to be quite challenging, given that the average number of iterations was higher than in other projects. In tackling this complex task, D3 found that a copy text was in the low-rated layers and began changing the words in the copy. The score eventually exceeded the target after D3 removed the copy entirely. Despite taking longer than other projects, D3 expressed satisfaction with the visual feedback system, as it eliminated the need for a trial-and-error process to identify areas for modification. while our system did not offer comprehensive guidance on improving it regarding this project, D3 leveraged high-rated layers to think up how to revise the banners in the remaining two projects completed on the first iteration. In one project, D3 enlarged a high-rated layer, and in another, D3 noticed that the background was in high-rated layers and then incorporated a main image from a different candidate banner with a similar background color to the refining banner.

D6, who worked under stricter assumptions than instructed and failed to achieve the target score in one project when using our system, still recognized the value of the visual feedback system. D6 considered that the low-rated layers and the head maps were helpful in readily identifying the areas needing refinements, more so than with the baseline system, and the visual feedback gave D6 novel ideas for refinements.