Evaluating the Impact of AI-Generated Visual Explanations on Decision-Making for Image Matching

Albatool Wazzan
Dept of Computer and Information Science
Temple University
Philadelphia, Pennsylvania
USA
tuk11806@temple.edu

Marcus Wright
Dept of Computer and Information Science
Temple University
Philadelphia, Pennsylvania
USA
marcus03wright@icloud.com

Stephen MacNeil
Dept of Computer and Information Science
Temple University
Philadelphia, Pennsylvania
USA
stephen.macneil@temple.edu

Richard Souvenir
Dept of Computer and Information Science
Temple University
Philadelphia, Pennsylvania
USA
souvenir@temple.edu

Abstract

Explanations have increasingly been incorporated into intelligent systems to offer insights into the underlying AI models. In this paper, we investigate the impact of AI-generated visual explanations on users' decision-making processes during an image matching task. Our work examines how these explanations affect correctness, timing, and confidence and explores the role of AI literacy in user behavior. We conducted a mixed-methods user study with 54 participants who were tasked to identify hotels from images using a specialized intelligent system. Participants were randomly assigned to use the system with or without visual explanation capabilities. Results showed that visual explanations did not affect the accuracy of the decision or the confidence of the user in image matching tasks. Participants with high-AI literacy outperformed those with lower literacy, but engaged less with explanations. Distinct matching strategies emerged between high-AI and low-AI participants, with high-AI participants systematically examining high-ranked images and using the explanation for verification purposes, while low-AI participants followed more exhaustive approaches.

CCS Concepts

• **Human-centered computing** → **User studies**; **Empirical studies in HCI**; *Empirical studies in visualization*; • **Computing methodologies** → *Matching*.

Keywords

XAI, visual explanations, post hoc explanations, decision-making, image matching, empirical studies, mixed-methods evaluation

ACM Reference Format:
Albatool Wazzan, Marcus Wright, Stephen MacNeil, and Richard Souvenir. 2025. Evaluating the Impact of AI-Generated Visual Explanations on Decision-Making for Image Matching. In *30th International Conference on Intelligent User Interfaces (IUI '25), March 24–27, 2025, Cagliari, Italy.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3708359.3712121

IUI '25, Cagliari, Italy

ACM ISBN 979-8-4007-1306-4/25/03
https://doi.org/10.1145/3708359.3712121

1 Introduction

Investigative image analysis involves decision-making workflows that increasingly rely on the use of intelligent systems. For applications such as geolocation or scene matching in image forensics and journalism, it is often necessary to examine or compare subtle details between multiple images [28, 50, 51]. Despite the growing availability of AI-powered tools, many users remain cautious due to the black-box nature of these applications [57]. To address these concerns, researchers in the field of explainable AI (XAI) have developed various explanation methods that reveal the inner workings of deep neural networks. By helping users understand the reasoning behind the results returned by the model, these methods aim to increase trust in these systems, promote transparency, and allow users to identify potential biases or errors in the model reasoning.

XAI methods are generally categorized into interpretable-by-design models [9, 12, 13, 20, 38], which offer built-in transparency but may be limited in complexity, and post hoc explanations [25, 46, 48] that seek to clarify the predictions of black box models. Post hoc explanations, particularly visual ones, have gained attention for their effectiveness in tasks involving visual data. Unlike textual explanations [34, 42], which provide narrative descriptions of model reasoning, visual explanations offer a direct and interpretable link between input and output.

Figure 1 outlines the decision-making process for an image matching task using an intelligent system with XAI features. Visual post hoc explanations have been extensively evaluated for algorithmic understanding, particularly in tasks such as image classification [18, 31, 47]. However, less work has been done to evaluate their impact on user decision-making processes when using intelligent systems. To evaluate the effectiveness of AI-generated visual explanations for decision-making, we conducted a mixed-methods user study with 54 participants. Participants were randomly assigned to use a system with or without visual explanation capabilities. Through both quantitative and qualitative evaluations, we aimed to address the following research questions:

RQ1 How do AI-generated visual explanations impact user decision-making in image matching?
RQ2 How does AI literacy affect how users engage with visual explanations in image matching?

Our findings revealed that visual explanations had no impact on task performance or user confidence. Users with higher AI literacy (high-AI) performed better on the task, but relied on the explanations less often than users with lower AI literacy (low-AI).

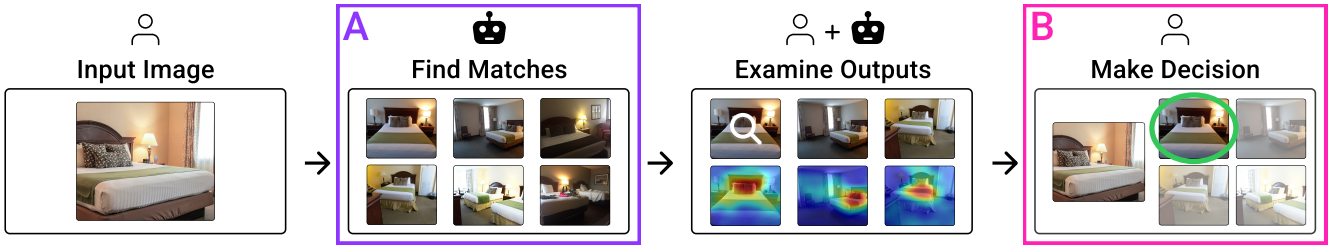


Figure 1: Human-AI workflow for modern image matching tasks with interactive intelligent systems. The user can examine the output with or without visual explanations. Previous work on evaluating visual explanations typically focused on (A) understanding the AI model. This paper evaluates how visual explanations impact (B) user decision-making.

Moreover, high-AI users with access to the explainability feature reported significantly lower confidence in their performance than those without. High-AI participants adopted systematic strategies, prioritizing high-ranked images and using explanations for verification. In contrast, low-AI participants employed more exhaustive strategies, frequently exploring lower-ranked images and relying more heavily on explanations. These results align with previous findings on the limited utility of visual explanations for algorithm understanding, while further expanding the scope to encompass their impact on user decision-making.

2 Related Work

Humans often rely on mental shortcuts or heuristics when interacting with intelligent systems [10]. To help users make informed decisions, the design of these systems frequently incorporates elements that enhance understanding and interpretation. For example, visualization components such as word clouds, bar charts, and Venn diagrams have been integrated into recommendation systems [54] and article search assistants [53]. In the domain of clinical decision support, text-based systems that explain diagnosis recommendations by highlighting contributing input factors, such as symptoms or laboratory results, have been shown to increase user trust and reliance [11, 40]. Similarly, for visual analysis tasks, techniques such as image augmentation [61] or image segmentation [59] have been used to facilitate the user decision-making process.

While these techniques aim to facilitate user decision-making, much of the research on human-AI interaction in visual analysis tasks has primarily focused on how explanations affect the user's understanding of the AI model itself. However, as AI systems become increasingly integrated into decision-making workflows, it is equally important to understand how these explanations impact the user's decision-making process. In this section, we review related methods for evaluating visual explanations and explore the role of AI literacy in the adoption of explainable AI (XAI).

2.1 Using Visual Explanations to Improve Model Interpretability

The proliferation of visual explanation methods was followed by efforts that aimed to evaluate their efficacy. Early work in this area focused on quantitative measures, such as faithfulness and completeness, which provide information on the alignment between the model and its explanation [26, 27, 41, 43, 52]. However, these methods do not capture how interpretable or helpful these explanations are from a user's perspective [21]. Human-centered evaluations, often conducted through psychophysical user studies, assess how well participants understand the model, which is often quantified through metrics such as accuracy, decision speed, and confidence [7]. In an image classification task, AlQaraawi [2] found that while LRP-generated saliency maps helped users accurately identify key features, they were less effective in predicting outcomes for new images. Similarly, Kim [31] demonstrated that explanations, such as GradCAM and BagNet, increased user confidence, but did not consistently help users distinguish between correct and incorrect model predictions. Nguyen [39] examined attribution maps in both generic and fine-grained image classification tasks, showing that the explanations not only failed to improve user performance, but worsened it for more complex, fine-grained tasks. For an image-based age prediction task, Chu [17] found that explanations did not have a significant effect on user accuracy or trust in model predictions, while Shen [47] found that saliency maps significantly reduced user prediction accuracy by 10%, suggesting that visual explanations may introduce confusion in evaluating predictions. Our work draws inspiration from these studies, but distinguishes itself by evaluating the impact of explanations on decision-making for a real-world task.

2.2 Effects of AI Literacy on Model Interpretability

Research on explainable AI (XAI) explanations has increasingly recognized the critical role that AI literacy plays in how explanations are interpreted [36]. Users bring varying mental models, cognitive abilities, and domain expertise, all of which influence how they understand AI explanations [5, 24, 32]. For image classification tasks, studies showed mixed results regarding the effectiveness of attention maps for users with varying levels of expertise. For example, Zimmermann [63] found that standard activation maps did not significantly improve the mental models of novice users for convolutional neural network (CNN) processes compared to simpler alternatives. However, Shitole [48] demonstrated that carefully designed attention maps significantly improved causal understanding of the model for an occluded image classification task compared to standard attention map baselines. Ehsan [23] found that users with a high level of AI literacy preferred more technical and detailed explanations of the behavior of AI agents, while users with

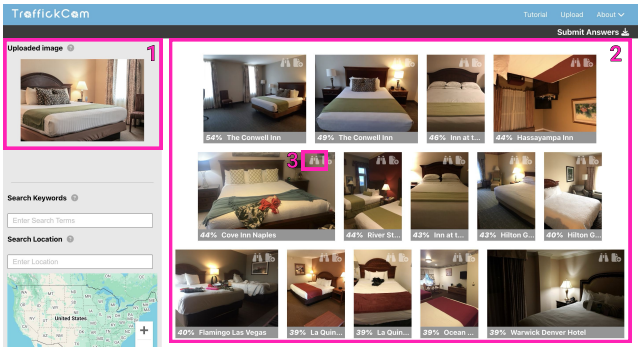


Figure 2: TraffickCam Interface. The user submits (1) an input image and the system uses a trained neural network to return (2) the most similar images from a large database. The pairwise visual similarity map between the retrieved image and input can be displayed by toggling the (3) visualization button on each retrieved image.

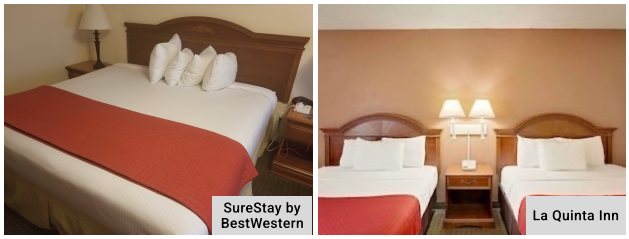
a low level of AI literacy preferred simpler and clearer narratives. A similar phenomenon is seen in how experienced users better calibrate their trust in AI compared to novice users [60]. In recommendation systems, Kühl [33] demonstrated that AI literacy not only shapes preferences for explanations, but also affects the user's willingness to comply with system recommendations; experienced users developed stronger mental models and required detailed explanations to trust and follow the recommendations. While prior work has explored how AI literacy affects users' understanding of AI models, our work shifts the focus to its potential impact on users' decision-making processes.

3 Methods

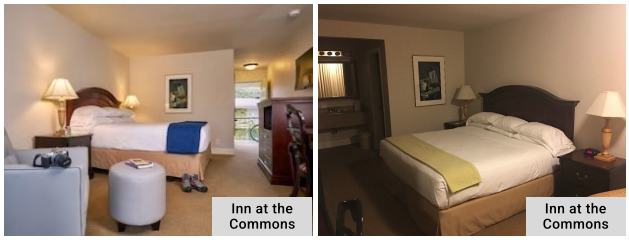
We conducted a mixed-methods user study with 54 participants, aimed at evaluating the impact of visual explanations on users' decision-making when interacting with an intelligent system for an image matching task. This section details the experimental platform, task design, the participant recruitment process and measures used for the evaluation.

3.1 Experimental Platform

TraffickCam [51] is a specialized reverse image search engine designed to help combat human trafficking by helping analysts identify hotel rooms from images, such as online advertisements. Figure 2 shows the TraffickCam interface. A user submits an input image, and the system uses a trained neural network to return the output images from a large database ranked by computed similarity. The system implements infinite scroll, where additional images automatically and continuously load as users scroll down the page. Like many retrieval systems, the top results may not always be the best matches. This issue is exacerbated in hotel room recognition, as visually similar images may not be from the same hotel, and visually dissimilar images may be a match, as shown in Figure 3. So, the user is left to complete the hotel identification task by carefully examining the results.



(a) Visually similar, but non-matching



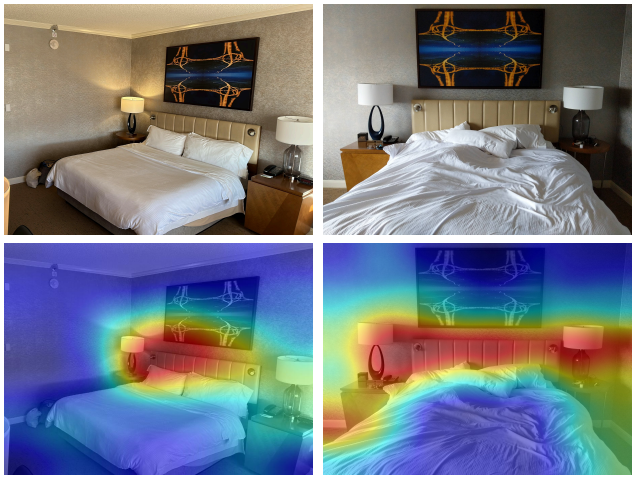
(b) Visually dissimilar, but matching

Figure 3: Challenges with hotel recognition from images. (a) Images representing different hotels that are visually similar. Close inspection of the objects (e.g., lamp, headboard) shows clear differences. (b) Images from the same hotel that are visually dissimilar, however, the objects (e.g., lamp, headboard, desk chair) within the scene match.

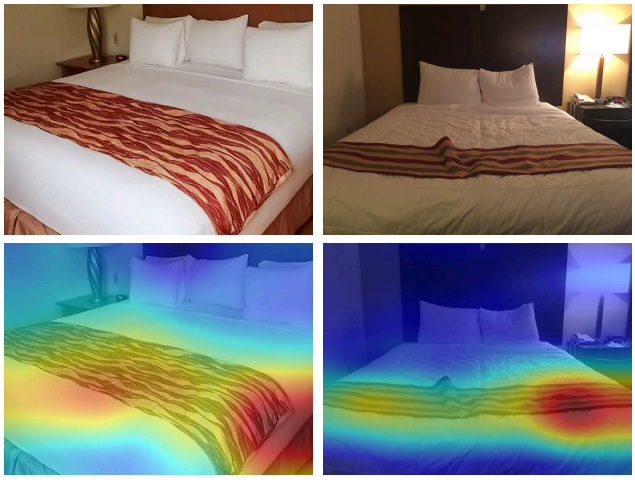
For the study, we modified the application to conditionally enable access to a visual explanation algorithm that produces a heatmap designed to highlight the regions of a pair of images that contributed the most to the pairwise similarity [6]. Users with the feature enabled could toggle a button on the retrieved images to be presented with the visual explanation of the selected image compared to the input. Figure 4 shows two examples of these visual explanations with pairs of images. In Figure 4a the heatmap highlights the visual similarity of the image regions that contain the headboard and lamps. Moreover, additional examination indicates that the artwork in both images is similar. Although the AI system did not rely on the similarity of the artwork to match the images, this evidence in the non-matched region could further confirm the match for the user. In Figure 4b the heatmap highlights the bed scarves as the regions of the image that contributed the most to the similarity score. However, on closer inspection, the striped patterns on the bed scarves are different. In addition, outside the highlighted region, the shape and style of the lamps also differ. Based on these discrepancies, users might choose to ignore the superficial visual similarity identified by the AI system and conclude that the pair of images do not match.

3.2 Study Design

We followed a factorial study design, with a between-subjects factor, where participants were randomly assigned to either the *Baseline* condition, where they used the TraffickCam system without the explanation feature, or the *Explanation-Enabled* condition, where the system included the explanation feature as an option. Participants completed the same four tasks, classified into easy and hard levels of difficulty, serving as a repeated within-subjects factor.



(a) Same hotel



(b) Different hotel

Figure 4: Pairwise visual similarity. For each pair of images, the heatmap highlights the regions that contributed the most to pairwise similarity (red = high, blue = low).

Participants were eligible for the study if they were 18 years of age or older and fluent in English. Prior to starting, they were informed of the task details, estimated duration, compensation, and right to withdraw at any time. The participants interacted with the web-based application using a keyboard and mouse. They were initially presented with a page that described the purpose of the experiment and the TraffickCam application before completing a pre-study survey, which included demographics (i.e., age, gender identity, academic background, and education level) and a question to assess AI literacy. The pre-survey also included the General Attitudes toward Artificial Intelligence Scale (GAAIS) [45], which measured their attitudes toward AI and a Subjective Technical Competence Scale (STC) [4], which assessed their self-reported technical proficiency and comfort with using technology.

Figure 5: The two *easy* (top) and two *hard* (bottom) input images used in the experiment. The level of difficulty was based on a combination of the level of difficulty in reporting a correct match for the AI system and human users in a pilot study.

After completing the pre-study survey, participants engaged in a training session that mirrored the main task of determining the hotel from which the input image was captured. For a provided input image, the participants reviewed the results returned by the system, similar to Figure 2. Depending on the condition, they could also view pairwise visualization heatmaps to assist in their assessment. To indicate a potential match, users toggled a selection icon next to the chosen image. Once participants completed their selections, they were shown the images they identified as matches, along with the corresponding hotel names. At this point, they had the option to reorder the hotel list to reflect their confidence in identifying the correct match to the input image.

The main experiment involved four tasks, each corresponding to the input images shown in Figure 5. A set of candidate images were selected on the basis of the level of difficulty in reporting a correct match for the AI system. We determined the rank and total number of correct matches in the returned results. The task was considered easier when the output included more correct matches ranked higher due to their greater similarity to the input image. In contrast, harder tasks had fewer and/or lower ranked correct matches, based on similarity scores computed by the AI model. Those images were then used in pilot studies to collect feedback from human participants. Tasks were perceived as easy when there was high intraclass similarity, that is, correctly matched images shared distinctive features with the input image, and hard when incorrect images closely resembled the input. Together, the AI and human criteria informed the final selection of two *easy* and two *hard* images for the experiment.

The tasks were presented to the participants in randomized order, and there was no time limit to complete each task. After submitting a hotel prediction for each task, participants rated their confidence on a scale from 1 (not confident) to 5 (very confident) before moving

on to the next task. Once the image matching portion was completed, participants completed a post-study survey that included four open-response questions about their strategies, challenges, and experiences. An additional question focused on their understanding of how the system generated the image results which was used to assess their mental models of the system. After completion of the experiment, the participants were compensated with a $5 coffee shop gift card.

3.3 Participants

To determine the appropriate sample size, we conducted a priori power analysis to detect a medium effect size (Cohen's d = 0.5) at a significance level of 0.05 with 85% power. This analysis indicated that 50 participants would be sufficient to detect meaningful differences between the experimental conditions, which is consistent with other studies of this type [14]. We recruited 54 participants from a college campus. The study, approved by the Institutional Review Board, was conducted over three weeks by two members of the research team. The mean age of the participants was 26 (*SD* = 6.28). The participants came from diverse academic backgrounds, including computer science, physics, chemistry, marketing, neuroscience, and engineering. The sample reflects a random selection process; no demographic criteria were applied and no explicit effort was made to balance any demographic attribute in the sample.

Of the 54 participants, we excluded a total of 9 participants from the analysis. Two participants, one from each condition, did not submit results for the majority of tasks. Seven participants from the Explanation-Enabled condition did not use the explanation feature at all, and were excluded based on the per-protocol analysis. Consequently, our analysis includes data from 26 participants in the Baseline condition and 19 in the Explanation-Enabled condition.

To ensure there was no sampling bias, we analyzed the GAAIS and STC responses from the pre-survey and found no significant differences in the participants' scores across the experimental conditions for either attitudes toward AI or self-reported technical proficiency, which suggests no sampling bias for these traits.

3.4 Mixed-Methods Analysis

We employed a mixed-methods approach that combines an analysis of quantitative measures from the experiment with rich qualitative insights collected through surveys. To investigate RQ1, we analyzed the effects of condition (Baseline and Explanation-Enabled) across the three dependent variables of task completion time, performance, and participant's confidence. For RQ2, we stratified participants into two groups: *low-AI* and *high-AI* to understand whether AI literacy had an effect on these same dependent variables. We also investigated participants' interaction patterns to understand their strategies and experiences.

3.4.1 Identifying Low and High AI Participants. To understand the AI literacy of participants in this study, we stratified our participants into two groups: *high-AI* and *low-AI* based on their responses to the open-ended pre-survey question. Previous studies have categorized AI expertise through self-reported measures, such as general AI knowledge [32] or professional background [23]; however, the reliability of self-reported data can vary. Given the more focused nature of our task, we instead asked participants to explain how they believe reverse image search functions by posing the question: "In a few sentences, describe how you think reverse image search works."

Two members of the research team independently categorized the responses based on their perception of whether the response indicated the participant was highly knowledgeable or not knowledgeable about reverse image search. The inter-rater reliability, as measured using Maxwell's RE coefficient [37] was 0.843, indicating a high level of agreement between the evaluators. Subsequently, the evaluators met to resolve the discrepancies through a consensus discussion.

3.4.2 Dependent Variables. We assessed three key dependent variables. *Performance*, measured by the mean reciprocal rank (MRR) [19] of the correct hotel among the ranked list submitted by the participant. That is, the reciprocal rank is maximized (1.0) when the correct match is ranked first and minimized (0.0) when the correct match is not included in the participant-provided list. *Task completion time* calculated as the duration taken by participants to complete each task, starting when the task was presented and ending when the ranked list was submitted. *Confidence ratings* were collected after each task, where participants rated their confidence in their selections on a scale from 1 (not confident) to 5 (very confident).

3.4.3 Quantitative Analysis. To evaluate RQ1, we used a linear mixed-effects model to account for the between-subjects categorical independent variable Condition (Baseline and Explanation-Enabled), and the repeated measure, Task Difficulty, allowing us to model the fixed effect of the experimental condition while accounting for random effects due to individual differences. To evaluate RQ2, we used the non-parametric Mann-Whitney U test to examine the impact of our categorical independent variables Condition (Baseline and Explanation-Enabled) and AI Literacy (high-AI and low-AI) on our continuous dependent variables (e.g., MRR and timing).

To identify strategies participants employed during the image matching task, we encoded the sequence of logged actions for each user and performed an analysis using the Levenshtein distance, which measures the minimum number of operations required to transform one sequence into another. The operations considered were insertions and deletions, each assigned a cost of 1, and substitutions, which were assigned a cost of 2. The total edit distance for each sequence pair was calculated as the sum of these costs, providing a straightforward measure of the dissimilarity between sequences of potentially different lengths. We applied t-SNE [56] to the pairwise dissimilarity matrix to embed the sequences into a two-dimensional space, followed by k-means clustering to identify groupings within the data.

3.4.4 Qualitative Analysis. In the post-survey, we collected open responses from the participants to questions about their task strategies, the challenges they faced, and their general experience during the tasks. These responses were analyzed using reflexive thematic analysis [8]. First, two members of the research team independently open coded the responses. Next, the two researchers met to discuss their codes and resolve any discrepancies. The coding process was iterative and involved multiple rounds of review and refinement.

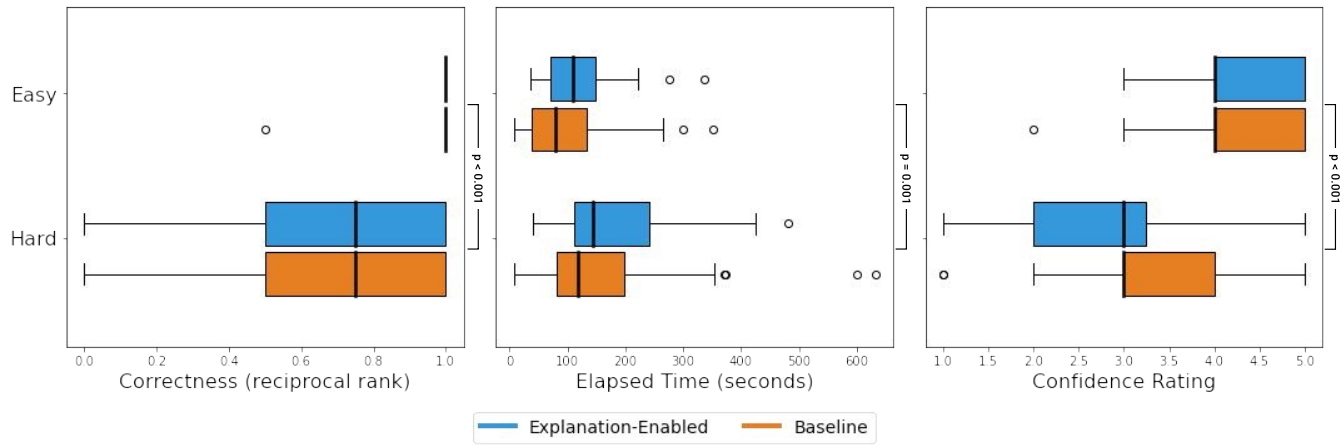


Figure 6: The plots show the (left) correctness, (middle) timing, and (right) confidence between the Explanation-Enabled and Baseline conditions in easy and hard tasks.

4 Results

We present our findings, organized to address the two research questions, evaluating the impact of (RQ1) visual explanations on decision-making and (RQ2) AI literacy on engaging with visual explanations in image matching.

4.1 RQ1: Impact of Visual Explanations on Decision-Making in Image Matching

We evaluated correctness, decision time, and user confidence for the four tasks, which were aggregated by the two difficulty levels. Figure 6 shows the (left) correctness, measured by MRR, (middle) timing, and (right) confidence. In the following, we describe each result in detail.

4.1.1 Visual Explanations and Performance . Performance was measured using MRR, which accounts for the rank, if present, of the correct hotel in the list submitted by the participant. For Easy tasks, participants in the Baseline condition ($M = 0.98$, $SD = 0.10$) performed similar to participants in the Explanation-Enabled condition ($M = 1.00$, $SD = 0.00$). For Hard tasks, performance was worse for both groups. Participants in the Explanation-Enabled condition ($M = 0.69$, $SD = 0.35$) slightly outperformed those in the Baseline condition ($M = 0.63$, $SD = 0.39$). Figure 6 (left) shows the distribution of performance scores for both easy and hard tasks. The Easy tasks were quite manageable for all participants, as all but one predicted the correct hotel as their first choice. For Hard tasks, there was a wider range of performance, but still similar whether explanations were available or not. Table 1 shows the results of the linear mixed-effects model analysis on the correctness metric. There was a significant difference in performance for the users in both conditions for the Hard tasks ($p < 0.001$). There was no significant effect for the Explanation-Enabled condition on performance.

4.1.2 Visual Explanations and Timing. We analyzed the average time participants spent on the task computed from the time the output images are displayed to when the participant submitted their decision. For Easy tasks, participants took an average of $M = 98.00$, $SD = 75.30$ seconds in the Baseline condition and $M = 118.68$, $SD = 67.23$ seconds in the Explanation-Enabled condition. For Hard tasks, in the Baseline condition, participants spent an average of $M = 161.72$, $SD = 132.72$ seconds, while, in the Explanation-Enabled condition, an average of $M = 180.03$, $SD = 105.83$ seconds. Figure 6 (middle) shows the distribution of time spent on the tasks. Analysis indicated a significant difference in the time spent between easy and hard tasks ($\beta = 61.35$, $p = 0.001$). There was no significant difference between the experimental conditions ($\beta = -20.69$, $p = 0.421$) nor a significant interaction effect between condition and task difficulty ($\beta = 4.38$, $p = 0.85$).

Table 1: Analysis of correctness, as measured by mean reciprocal rank.

	Estimate	Std. Error	z-value	p-value
Intercept	0.981	0.048	20.24	<0.001
Condition (Explanation-Enabled)	0.019	0.076	0.254	0.800
Difficulty (Hard)	-0.351	0.060	-5.88	<0.001
Condition*Difficulty	0.041	0.093	0.437	0.662

4.1.3 Visual Explanations and Confidence. After each image matching task, participants rated their confidence in their selections on a scale from 1 (not confident) to 5 (very confident). For Easy tasks, confidence ratings were generally high in both conditions, with participants in the Baseline condition reporting a mean confidence of $M = 4.29$, $SD = 0.70$ and those in the Explanation-Enabled condition reporting $M = 4.36$, $SD = 0.68$. For Hard tasks, participants in the Baseline condition reported a mean confidence of $M = 3.28$, $SD = 1.01$, while those in the Explanation-Enabled condition reported ($M = 3.03$, $SD = 0.94$). Figure 6 (right) shows the distribution of confidence ratings. Analysis showed a significantly different mean confidence rating for Hard tasks ($\beta = -1.015$, $p < 0.001$). There was no significant difference between experimental conditions ($\beta = 0.073$, $p = 0.72$) nor a significant interaction effect between condition and task difficulty ($\beta = -0.318$, $p = 0.18$).

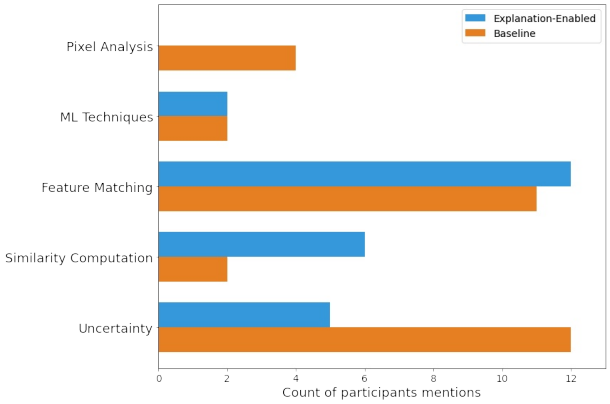


Figure 7: Comparison of participants' understanding of reverse image search between the Explanation-Enabled (blue) and Baseline (orange) conditions. The figure illustrates differences across five codes.

4.1.4 Visual Explanations and Users' Mental Models. To evaluate whether the availability of visual explanations influenced participants' mental models and understanding of reverse image search, we examined their responses to a post-study survey question regarding their understanding of image matching. The reflexive thematic analysis identified five topics that encapsulate the range of participants' conceptual understanding:

- **Pixel Analysis**: Describing the process of analyzing pixel values and their relationships to surrounding pixels for image matching.
- **ML Techniques**: Referring to various artificial intelligence methods used in reverse image search, including machine learning algorithms, contrastive learning, and neural networks.
- **Feature Matching**: Identifying and comparing distinct features within images (e.g., shapes, patterns, or objects) to detect similarities.
- **Similarity Computation**: Describing the process of calculating similarity scores or percentages between images to determine matches.
- **Uncertainty**: Expressing confusion or providing vague or incomplete descriptions about how reverse image search systems work.

Two researchers coded the responses, and disagreements were mediated through discussion until consensus was reached. Figure 7 illustrates the distribution of each code for both conditions. In the Explanation-Enabled condition, fewer participants expressed uncertainty about the image matching process than in the Baseline condition, suggesting that they were able to form clearer mental models after interacting with the system, which may have helped bridge gaps in their initial understanding. For many of the themes, the relative change between conditions was similar. Mentions of matching features or objects were high for both conditions, suggesting task-driven learning gains irrespective of the availability of explanations.

One difference pattern between conditions was references to computing similarity. Participants in the Explanation-Enabled condition made such references at a much higher rate. For example, participants provided responses such as: "It estimates the similarity based on features and expresses it as a percentage" and "the system returns images in descending order of priority based on the similarity score." However, for the Baseline condition, fewer participants mentioned the similarity computation. These results suggest that the visual explanations helped participants internalize the computational process underlying the image matching algorithm.

4.2 RQ2: Impact of AI Literacy on Interpreting and Engaging with Visual Explanations in Image Matching

To assess how AI literacy affects users' interpretation and engagement with visual explanations, we categorized participants into two groups: *high-AI* and *low-AI* based on their responses to the open-ended pre-survey question. Following this categorization, 21 participants were rated as having high-AI literacy (11 in the Baseline and 10 in the Explanation-Enabled condition) and 24 participants were rated as low-AI literacy (15 in the Baseline and 9 in the Explanation-Enabled condition). Given the universally high performance, the Easy tasks were excluded from the analysis in this section.

4.2.1 AI Literacy and Decision-Making in Image Matching. We compared the performance of the high-AI and low-AI groups on correctness, timing, and confidence for the Hard tasks.

Figure 8 (left) shows the distribution of performance scores for high-AI and low-AI participants. High-AI participants performed similarly in both conditions, with a median MRR of 1.0 in both the Explanation-Enabled and Baseline conditions. Analysis using the Mann-Whitney U test showed no significant differences between conditions ($U = 230.50$, $p = 0.74$). Low-AI participants performed better in the Explanation-Enabled condition ($Mdn = 0.75$) than those in the Baseline condition ($Mdn = 0.37$), though the difference was not statistically significant ($U = 231.50$, $p = 0.84$). Overall, high-AI participants significantly outperformed low-AI participants for both conditions ($U = 645.0$, $p = 0.002$), suggesting that AI literacy played a key role in task performance.

Figure 8 (middle) shows the distribution of time spent by the high-AI and low-AI participants. Participants in the high-AI group averaged $Mdn = 122.76$ seconds in the Explanation-Enabled condition and $Mdn = 118.79$ seconds in the Baseline condition. Participants in the low-AI group spent $Mdn = 211.51$ seconds in the Explanation-Enabled condition compared to $Mdn = 119.05$ seconds in the Baseline condition. Neither difference was significant, $U = 208.0$, $p = 0.97$ and $U = 169.0$, $p = 0.14$, respectively. Overall, there was no significant difference in the time spent between the two AI literacy groups ($U = 1032.0$, $p = 0.35$).

Figure 8 (right) shows the distribution of confidence ratings for high-AI and low-AI participants. In line with their better overall performance, high-AI participants expressed higher confidence on average than low-AI participants, with a difference that approached significance ($U = 717.5$, $p = 0.06$). Within the high-AI group,

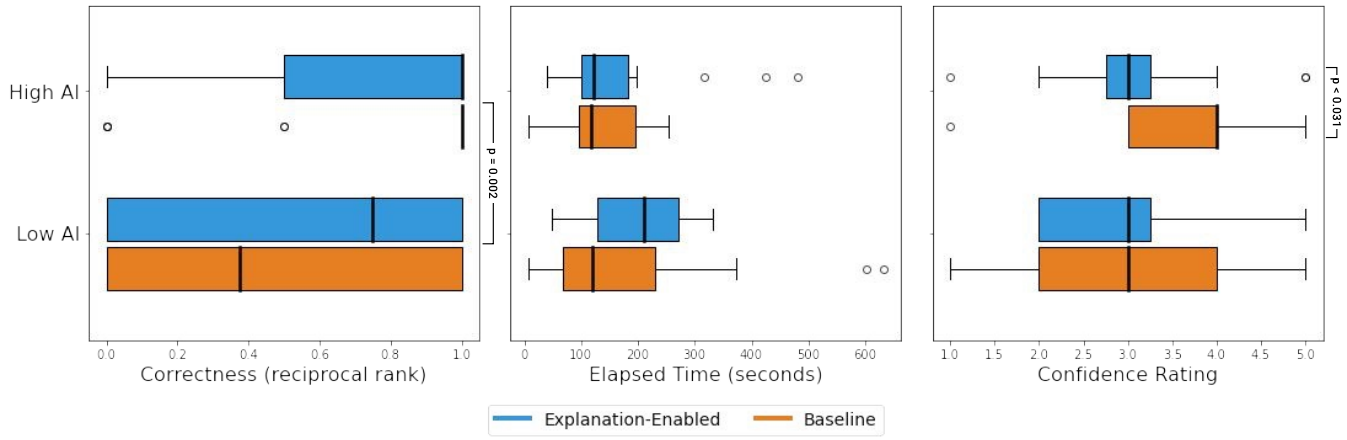


Figure 8: The plots show the (left) correctness, (middle) timing, and (right) confidence between Explanation-Enabled and Baseline conditions between users with high-AI, and low-AI literacy.

the confidence reported was lower for those in the Explanation-Enabled condition ($Mdn = 3.0$) compared to the Baseline condition ($Mdn = 4$); the analysis indicated that this difference was statistically significant ($U = 288.5$, $p = 0.03$), suggesting that access to the explanation feature decreased user confidence even though there was no corresponding decrease in performance. Within the low-AI group, the confidence ratings were similar in both conditions, with a median confidence of $Mdn = 3.00$ in both conditions, suggesting that visual explanations had little or no impact on the confidence of participants with lower AI literacy.

4.2.2 AI Literacy and Interaction Patterns in Image Matching. Table 2 describes the actions logged during the experiment and includes the average frequency of each action for each task aggregated by condition and the participant's AI literacy. We observed some differences. High-AI participants inspected approximately 44.97% more images per session than low-AI participants. High-AI participants used the pairwise visualization feature an average of 4.20 times per session, while the average for low-AI participants was nearly double at 7.87.

Participants had the option to select (and deselect) matches from the (primary) output returned by the system or the additional (secondary) images from the same hotel that were displayed upon inspection of a primary image. In general, users selected similar numbers of images. On average, users in all conditions performed 9.48 total primary and secondary selections. While high-AI users performed a similar number of primary selections in each condition, low-AI users averaged 4.6 primary selections in the Baseline condition and nearly double ($M = 8.12$) in the Explanation-Enabled condition. For secondary selections, all subgroups averaged around 4, with the exception of high-AI users in the Explanation-Enabled condition, who averaged 2.7. For the participants in the Explanation-Enabled condition, high-AI participants were more decisive in their selections; they made fewer selections and deselections than their low-AI counterparts.

In addition to the actions of the participants, we examined the ranks of images inspected throughout the tasks. High-AI participants inspected higher-ranked images, with a median rank of Mdn (IQR) = 5 (2,8). While low-AI participants, inspected a wider range of images, with a median rank of Mdn (IQR) = 8 (2,21). The results of Mann-Whitney U test showed a significant difference between the two groups ($p = 0.002$). This behavior suggests that the high-AI and low-AI participants employed different exploration strategies.

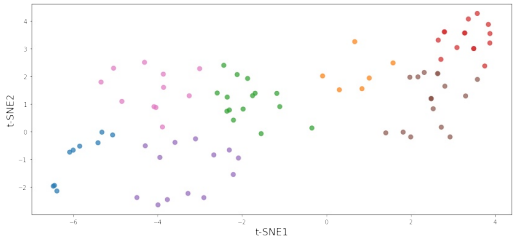
4.2.3 AI Literacy and Image Matching Strategies. To further analyze the matching strategies employed, we encoded participants' sequences of actions, then analyzed them using the Levenshtein distance. We applied t-SNE for dimensionality reduction, followed by k-means clustering to identify groupings. We determined the optimal number of clusters to be $k = 7$ using the elbow method. The resultant embedding and clustering of the interaction sequences are shown in Figure 9a.

Examining the resultant clusters revealed different patterns of interaction. Figure 9b shows a *representative sequence* of each cluster. For cluster 1, the sequences contained repeated instances of INS-SL1 or INS-VIS-SL1, suggesting that the users systematically used the features of the application to examine images before moving on to the next image. The sequences in cluster 2 consisted mostly of selection actions SL1 with some use of visual explanations (VIS), indicating that users selected primary output images as matches with minimal inspection. Cluster 3 contained sequences with multiple instances of INS-VIS, but unlike cluster 1, they were not immediately followed by a selection. Cluster 4 sequences included mainly selection (SL1), suggesting quick decision-making with minimal interactions. Sequences in cluster 5 tended to start with multiple INS actions and include both SL1 and SL2 actions. These sequences demonstrated an exhaustive search strategy in which participants inspected multiple primary and secondary images. For Cluster 6, the sequences included many SL2 actions without many SL1 actions, indicating that users often selected matching images from those not initially returned by the AI system. There was no readily discernible pattern to the sequences in Cluster 7. The cluster includes a variety of sequences, including long ones with repeated actions on the same image.

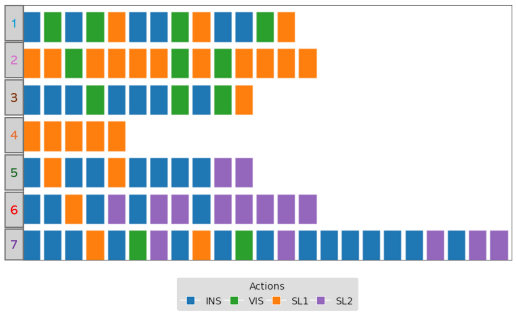
Figure 10 shows the distribution of each sequence pattern grouped by the participant's level of AI literacy. The analysis is based on 41

Action	Description	Mean Usage		
		AI Lit.	Baseline	Explanation
Inspect (INS)	Examined enlarged image and browsed additional images from the same hotel	High	17.09	17.60
		Low	11.33	13.12
Visualize (VIS)	Examined pairwise visual explanation heatmaps for an output image and the input image.	High	N/A	4.20
		Low	N/A	7.87
Select Primary (SL1)	Selected an image returned by the system as a match.	High	5.80	6.50
		Low	4.60	8.12
Select Secondary (SL2)	Selected one of the additional images as a match.	High	3.90	2.70
		Low	4.40	4.12
Deselect (DSL)	Deselected a previously selected image.	High	0.45	0.10
		Low	0.26	0.87

Table 2: Summary of logged actions, indicating the average usage by participants throughout the experiment. N/A values indicate the absence of the Visualize action in the Baseline condition.



(a) 2D t-SNE embedding



(b) Representative sequences (examples) from each cluster

Figure 9: Sequence analysis. (a) 2D t-SNE embedding of the participant sequence data with points colored based on k-means clustering ($k = 7$). (b) Representative sequences (examples) from each cluster. Each row shows the sequence of actions, indicated by color.

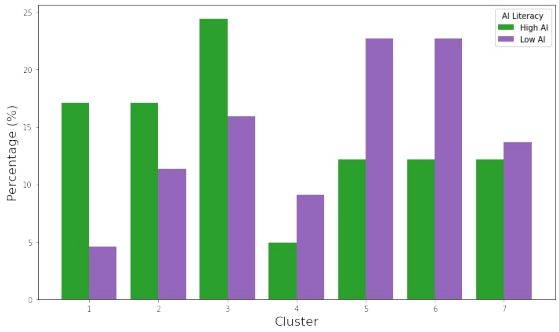


Figure 10: Distribution of high-AI and low-AI Literacy participants across the seven clusters.

interaction sequences from high-AI participants and 44 interaction sequences from low-AI participants. We noted some differences in the interaction patterns between these groups.

Cluster 1 sequences demonstrated a depth-first strategy in which the user thoroughly evaluated an output image and decided on whether to select it before proceeding to the next; this strategy was primarily employed by high-AI participants. The distribution of sequences in cluster 2 was nearly balanced between high-AI and low-AI participants; these users, at times, inspected or examined the visualizations, but mainly focused on selecting matches. Cluster 3 sequences suggested a breadth-first search approach in which the images were thoroughly examined, but, unlike cluster 1, selection occurred only after several images had been examined. Clusters 4, 5, and 6 sequences were dominated by low-AI participants. Cluster 4 sequences suggested quick decision-making; these participants adopted an economic search approach without using the features of the system for further inspection. Clusters 5 and 6 sequences were nearly twice as common among low-AI participants. In Cluster 5, participants examined and selected images from both primary and secondary results, while in Cluster 6, they reviewed both but mainly chose from the secondary results. These participants followed an exhaustive search approach, exploring not only the main system output but also secondary alternatives. For Cluster 7, the sequence pattern was not discernable; there was a near-even split between high-AI participants and low-AI participants.

The self-described strategies of high-AI participants during the post-study survey further contextualized these findings. Among the 10 high-AI participants in the Explanation-Enabled condition, five mentioned using the visual explanation as a key step of their process. For three of these participants, the heatmaps served as a verification tool, helping them confirm their initial assessments before making final decisions. For example, P52 described a two-step process, "I would examine the image by looking first at things like the headboard or lamps if they are unique then use the heatmap to confirm my thoughts." Similarly, P10 explained, "I was mostly checking whether the beds and walls looked alike. If I was skeptical,

the heatmap would help me." Another participant (P12) further elaborated:

> "I would first review the initial image used for comparison and then compare it to the top row of results, which had the highest confidence level. If there were clear similarities between the images (e.g., pillows, bed frame, lamps, rugs, curtains, etc.) I would make sure they were actually similar. In the case that nothing stood out as the same, I would just select the closest possible image. To confirm things were actually the same, I would then use the heatmap visualization to confirm my thoughts."

P8's approach differed slightly from the others, as they used the heatmap more selectively, turning to it only when something did not seem right: "I look for distinct features, like the view through the windows, the carpet, and the color of the walls. Then I would use the heatmap when I couldn't figure out why an image was ranked high." P50 relied on the heatmap from the very beginning, stating, "I initially followed the given heatmap and similarity, and then started noticing the details in the image." Despite their heavy reliance on the heatmap throughout the task, none of the low-AI participants mentioned it to describe their approach. Instead, these participants described comparing objects, colors, and patterns. For instance, P28 said, "I focused on similarities in terms of room orientation, the carpet pattern/color, the designs of bed lamps, and a few more things."

5 Discussion

The results suggest two key findings: 1) visual explanations have little or no impact on decision-making for this image matching task, and 2) AI literacy plays an important role, with high-AI users performing better and applying different strategies than those with lower AI literacy. In this section, we discuss these results and their broader implications.

5.1 Utility of Visual Explanations

In response to **RQ1**, in general, visual explanations did not seem to improve user decision-making for this image matching task. When the tasks were harder and cognitively demanding, the performance was significantly lower compared to easy tasks. The heatmaps, intended to provide clarity by highlighting regions of similarity, may have been too coarse, as users in previous studies have noted [32], which may have led to increased cognitive burden. Using the dual-process theory [15] as a lens, we interpret the observed behavior as users potentially shifting from quick, instinctive System 1 thinking to slower, more deliberate System 2 thinking, as they tried to reconcile their own interpretations and decisions with the heatmap's highlighted areas. Instead of facilitating faster, more confident decisions, the explanations required users to engage in critical thinking, as they needed to interpret the heatmap while simultaneously analyzing image details, which may have increased the effort required to complete the task. This aligns with prior research, which suggests that explanations may not always improve decision outcomes [62], particularly when users must exert more cognitive effort to reconcile the explanations with their own domain knowledge.

Although the visual explanations did not lead to more accurate decisions, participants in the Explanation-Enabled condition demonstrated a better conceptual understanding of reverse image search. Multiple participants included more detailed descriptions for AI-guided image matching, including descriptions of how similarity is computed. This suggests that while the heatmaps may not have improved immediate task performance, they did foster an understanding of the underlying AI processes, aligning with findings from prior research [2, 48]. However, this better conceptual understanding was not enough to overcome the intrinsic complexity of the task, which still required careful analysis.

5.2 Role of AI Literacy

Our findings regarding **RQ2** revealed that AI literacy played an important role in this image matching task, influencing participants' performance and decision-making strategies. High-AI participants consistently outperformed their low-AI counterparts. However, among high-AI participants, those with access to explanations reported lower confidence. This paradox may be explained by the expertise reversal effect [30], where experienced users, having developed their own decision-making strategies, found the explanations to introduce cognitive interference, causing them to second-guess their judgments. Despite this reduced confidence, their ability to navigate the task effectively stemmed from their established mental models, allowing them to engage with the visual explanations efficiently without relying on them heavily. This is consistent with prior research, which suggests that familiarity with AI systems improves the user's ability to use them effectively [23, 29, 32].

Low-AI participants struggled to achieve similar levels of performance and confidence, which was reflected in their interaction patterns. Low-AI participants heavily relied on the explanations much more than high-AI participants. Due to their lack of foundational knowledge, they likely found it difficult to grasp the new information conveyed, requiring more cognitive resources to process this unfamiliar content [44].

The differences in interaction patterns and strategies between high-AI and low-AI participants underscore these broader findings. High-AI participants tended to inspect more higher-ranked images and use the visual explanation feature for verification purposes, demonstrating that visual explanations can help users calibrate trust in AI by confirming their own assessments [62]. These participants employed systematic strategies, such as depth-first search, suggesting that they had well-developed mental models of the system's functionality and felt confident using its features to refine their decisions. In addition, they employed strategies that involved offloading parts of the task by using the system's features, such as the heatmap, to examine multiple images and focus their mental resources on key decision points later in the task [58]. This strategic use of the system reflects their ability to balance exploration with efficient decision-making.

Low-AI participants showed different patterns of behavior. They tended to use the explanation more frequently and inspect fewer images per session. Rather than using the explanations to support their decisions, they may have relied on the heatmaps as a crutch for decision-making. This behavior aligns with the theory of satisficing, where users opt for an adequate solution rather than the optimal one

when faced with high cognitive demands [49]. Additionally, these participants engaged in exhaustive search strategies [3], exploring low-ranked images and various system features before reaching a decision. Their exhaustive approach appeared to compensate for a lack of confidence in the system's outputs and uncertainty about how to best navigate the task.

These findings highlight a key issue with current explanation designs: they are often designed with the intent and understanding of the expert designers, rather than the needs of novice users [22]. This one-size-fits-all approach can make it difficult for less familiar users to effectively leverage explanations in decision-making tasks. Future efforts should focus on designing explanations that prioritize the user's perspective, particularly novice users, ensuring that they receive just enough information without feeling overwhelmed [1].

5.3 Limitations

Several limitations were identified in the experimental design and implementation. These limitations may have influenced the outcomes and should be considered when interpreting the results.

5.3.1 Participant Sample. The application used in the experiment was designed for a specialized image matching task and provided a convenient, real-world platform to evaluate decision-making with visual explanations. Although some of the users, recruited from a university setting, may have been familiar with the TraffickCam project or the researchers on the team, none had previously used any version of the application nor were expert image analysts. The fact that all participants performed universally well on the easier tasks suggests that this type of analysis may be approachable for the lay user. We followed the human-grounded evaluation framework [21], where lay participants serve as proxies to observe general behavioral patterns. The limited experience with the system may not fully capture the learning gains accumulated over time. Explanations that initially seemed unhelpful or confusing might become more effective as users grow more familiar with the interface and task. Future studies could explore longitudinal evaluations of non-expert users or they could involve expert users directly.

5.3.2 Visual Explanation Method. This study employed a CAM-based method to highlight paired image similarity in a Transformer-based feature-encoding setting. This approach was selected for its ability to provide detailed insights into paired image embeddings, making it particularly suited to the hotel-matching task. However, alternative methods may emphasize different aspects of model behavior. Techniques like BagNet [9], RISE [41], and prototype-based approaches such as ProtoPNet [16] focus on aspects such as semantic regions and conceptual prototypes. Future research could explore these alternatives to assess their impact on user decision-making and evaluate their effectiveness in real-world scenarios.

5.3.3 Task Difficulty. TraffickCam indexes a vast database of millions of images, sourced from travel websites and user-uploaded photos, so the outputs returned by the system may include blurry, low-quality, or otherwise nonstandard images of hotel rooms. Several participants reported difficulties analyzing images due to poor quality, mentioning issues like lens glare or inadequate lighting that obscured key details. They also struggled with images taken from different angles, which made it harder to visualize accurate matches.

Many participants found the high visual similarity between hotel room images challenging. In some cases, images appeared almost identical, differing only in minor details such as carpets or wall patterns. Conversely, some participants noted that visually dissimilar images contained similar objects, leading to confusion when trying to determine the correct match. Others found it difficult to distinguish between images dominated by generic objects, such as TV stands, white bed comforters, or neutral desks. These objects offered few clues for identifying the correct match, as they are commonly found in many hotel rooms. These types of comments came from participants in both conditions and irrespective of AI literacy. These challenges are inherent in this real-world matching task.

Some comments were specific to the interface as participants suggested new features for the application, most related to the limited functionality with the secondary images. Some noted the inability to enlarge secondary images while others wanted to apply the visual explanations to secondary images. These comments were not specific to a particular group as these issues had the potential to affect all participants equally.

6 Conclusion and Future Work

Much of the existing work in explainable AI (XAI) focuses on how explanations enhance algorithmic understanding. Our study shifted focus to evaluate their impact on decision-making in a real-world investigative image matching task. We found that the visual explanations had no impact on decision accuracy or user confidence. AI literacy was a key factor in this task, with participants who had higher levels of AI literacy consistently outperforming those with lower literacy. High-AI participants used efficient strategies, such as focusing on high-ranked images and using explanations to verify their judgments. In contrast, low-AI participants adopted less efficient and exhaustive search strategies.

Our findings highlight the complexity of integrating visual explanations into decision-making workflows, particularly in cognitively demanding tasks. While explanations can improve users' understanding of AI systems, their utility in guiding decisions remains limited. These insights are important for decision-making, particularly in high stakes scenarios such as medical diagnosis and financial analysis where much of the decision-making still relies on human judgment.

Users with higher AI literacy demonstrated a more strategic use of visual explanations. With that in mind, future work should focus on techniques that adjust the depth of information presented based on users' AI literacy, such as adaptive [55] and selective explanation techniques [35], integrating them into decision-making workflows. By progressively introducing details as needed, cognitive load can be minimized for less experienced users, ultimately enhancing decision-making for a broader range of users.

Acknowledgments

Research was sponsored by the DEVCOM Analysis Center and was accomplished under Cooperative Agreement Number W911NF-22-2-0001. The views and conclusions contained in this document are

those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the DEVCOM Analysis Center or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] Ashraf Abdul, Christian Von Der Weth, Mohan Kankanhalli, and Brian Y Lim. 2020. COGAM: measuring and moderating cognitive load in machine learning model explanations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. ssociation for Computing Machinery, Honolulu HI USA, 1–14.

[2] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th international conference on intelligent user interfaces*. ACM IUI, Cagliari, Italy, 275–285.

[3] Anne Aula, Päivi Majaranta, and Kari-Jouko Räihä. 2005. Eye-tracking reveals the personal styles for search result evaluation. In *Human-Computer Interaction-INTERACT 2005: IFIP TC13 International Conference, September 12-16, 2005. Proceedings 10*. Springer, Springer, Rome, Italy, 1058–1061.

[4] G Beier. 1999. Locus of control when interacting with technology (Kontrollüberzeugungen im Umgang mit Technik). *Report Psychologie* 24, 9 (1999), 684–693.

[5] Maalvika Bhat and Duri Long. 2024. Designing Interactive Explainable AI Tools for Algorithmic Literacy and Transparency. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 939–957. https://doi.org/10.1145/3643834.3660722

[6] Samuel Black, Abby Stylianou, Robert Pless, and Richard Souvenir. 2022. Visualizing paired image similarity in transformer networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. IEEE/CVF, Waikoloa, HI, USA, 3164–3173.

[7] Judy Borowski, Roland Simon Zimmermann, Judith Schepers, Robert Geirhos, Thomas SA Wallis, Matthias Bethge, and Wieland Brendel. 2021. Exemplary Natural Images Explain CNN Activations Better than State-of-the-Art Feature Visualization. In *International Conference on Learning Representations*. ICLR, Austria, 9 pages.

[8] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[9] Wieland Brendel and Matthias Bethge. 2019. Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. arXiv:1904.00760 [cs.CV] https://arxiv.org/abs/1904.00760

[10] Nadia Burkart and Marco F Huber. 2021. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research* 70 (2021), 245–317.

[11] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In *2015 International Conference on Healthcare Informatics*. IEEE, Dallas, TX, USA, 160–169. https://doi.org/10.1109/ICHI.2015.26

[12] Moritz Böhle, Mario Fritz, and Bernt Schiele. 2021. Convolutional Dynamic Alignment Networks for Interpretable Classifications. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE/CVF, Nashville, Tennessee, 10024–10033. https://doi.org/10.1109/CVPR46437.2021.00990

[13] Moritz Böhle, Mario Fritz, and Bernt Schiele. 2022. B-cos Networks: Alignment is All We Need for Interpretability. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE/CVF, New Orleans, LA, 10319–10328. https://doi.org/10.1109/CVPR52688.2022.01008

[14] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 981–992. https://doi.org/10.1145/2858036.2858498

[15] Shelly Chaiken. 1999. Dual-process theories in social psychology. *Guilford Press google schola* 2 (1999), 206–214.

[16] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. 2019. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems* 32 (2019), 12 pages.

[17] Eric Chu, Deb Roy, and Jacob Andreas. 2020. Are Visual Explanations Useful? A Case Study in Model-in-the-Loop Prediction. arXiv:2007.12248 [cs.LG] https://arxiv.org/abs/2007.12248

[18] Julien Colin, Thomas Fel, Remi Cadene, and Thomas Serre. 2023. What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods. arXiv:2112.04417 [cs.CV] https://arxiv.org/abs/2112.04417

[19] Nick Craswell. 2009. *Mean Reciprocal Rank*. Springer US, Boston, MA, 1703–1703. https://doi.org/10.1007/978-0-387-39940-9_488

[20] Jon Donnelly, Alina Jade Barnett, and Chaofan Chen. 2022. Deformable protopnet: An interpretable image classifier using deformable prototypes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. IEEE/CVF, New Orleans, LA, 10265–10275.

[21] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv: Machine Learning* none (2017), 9 pages. arXiv:1702.08608 https://api.semanticscholar.org/CorpusID:11319376

[22] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding Explainability: Towards Social Transparency in AI systems. 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 82, 19 pages. https://doi.org/10.1145/3411764.3445188

[23] Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I-Hsiang Lee, Michael Muller, and Mark O Riedl. 2024. The Who in XAI: How AI Background Shapes Perceptions of AI Explanations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24, Vol. 104)*. ACM, Hawaii, 1–32. https://doi.org/10.1145/3613904.3642474

[24] Upol Ehsan and Mark O Riedl. 2020. Human-centered explainable ai: Towards a reflective sociotechnical approach. In *HCI International 2020-Late Breaking Papers: Multimodality and Intelligence: 22nd HCI International Conference, HCII 2020, July 19–24, 2020, Proceedings 22*. Springer, Springer, Copenhagen, Denmark, 449–466.

[25] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. 2019. Understanding Deep Networks via Extremal Perturbations and Smooth Masks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE/CVF, Seoul, Korea, 2950–2958. https://doi.org/10.1109/ICCV.2019.00304

[26] Ruth C. Fong and Andrea Vedaldi. 2017. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Venice, Italy, 1–10. https://doi.org/10.1109/iccv.2017.371

[27] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A Benchmark for Interpretability Methods in Deep Neural Networks. arXiv:1806.10758 [cs.LG] https://arxiv.org/abs/1806.10758

[28] Mohammad T. Islam, Scott Workman, Hui Wu, Nathan Jacobs, and Richard Souvenir. 2014. Exploring the geo-dependence of human face appearance. In *IEEE Winter Conference on Applications of Computer Vision*. IEEE, CO, USA, 1042–1049. https://doi.org/10.1109/WACV.2014.6835989

[29] Anniek Jansen, François Leborgne, Qiurui Wang, and Chao Zhang. 2024. Contextualizing the "Why": The Potential of Using Visual Map As a Novel XAI Method for Users with Low AI-literacy. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 87, 7 pages. https://doi.org/10.1145/3613905.3650812

[30] Slava Kalyuga. 2009. The expertise reversal effect. In *Managing cognitive load in adaptive multimedia learning*. IGI Global, Pennsylvania, USA, 58–80.

[31] Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, and Olga Russakovsky. 2022. HIVE: Evaluating the Human Interpretability of Visual Explanations. arXiv:2112.03184 [cs.CV] https://arxiv.org/abs/2112.03184

[32] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. 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 250, 17 pages. https://doi.org/10.1145/3544548.3581001

[33] Niklas Kühl, Christian Meske, Maximilian Nitsche, and Jodie Lobana. 2024. Investigating the Role of Explainability and AI Literacy in User Compliance. arXiv:2406.12660 [cs.AI] https://arxiv.org/abs/2406.12660

[34] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2019. An Evaluation of the Human-Interpretability of Explanation. arXiv:1902.00006 [cs.LG] https://arxiv.org/abs/1902.00006

[35] Vivian Lai, Yiming Zhang, Chacha Chen, Q. Vera Liao, and Chenhao Tan. 2023. Selective Explanations: Leveraging Human Input to Align Explainable AI. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 357 (Oct. 2023), 35 pages. https://doi.org/10.1145/3610206

[36] Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. 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, 1–16. https://doi.org/10.1145/3313831.3376727

[37] A. E. Maxwell. 1977. Coefficients of Agreement Between Observers and Their Interpretation. *The British Journal of Psychiatry* 130, 1 (1977), 79–83. https://doi.org/10.1192/bjp.130.1.79

[38] Meike Nauta, Ron Van Bree, and Christin Seifert. 2021. Neural prototype trees for interpretable fine-grained image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. IEEE/CVF, Nashville, Tennessee, 14933–14943.

[39] Giang Nguyen, Daeyoung Kim, and Anh Nguyen. 2022. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. arXiv:2105.14944 [cs.CV] https://arxiv.org/abs/2105.14944

[40] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. 2022. Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 568, 9 pages. https://doi.org/10.1145/3491102.3502104

[41] Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. RISE: Randomized Input Sampling for Explanation of Black-box Models. arXiv:1806.07421 [cs.CV] https://arxiv.org/abs/1806.07421

[42] Jean-Philippe Poli, Wassila Ouerdane, and Régis Pierrard. 2021. Generation of textual explanations in XAI: The case of semantic annotation. In *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, IEEE, Luxembourg City, Luxembourg, 1–6.

[43] Samuele Poppi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2021. Revisiting The Evaluation of Class Activation Mapping for Explainability: A Novel Metric and Experimental Analysis. arXiv:2104.10252 [cs.CV] https://arxiv.org/abs/2104.10252

[44] Travis R Ricks, Kandi Jo Turley-Ames, and Jennifer Wiley. 2007. Effects of working memory capacity on mental set due to domain knowledge. *Memory & cognition* 35 (2007), 1456–1462.

[45] Astrid Schepman and Paul Rodway. 2023. The General Attitudes towards Artificial Intelligence Scale (GAAIS): Confirmatory validation and associations with personality, corporate distrust, and general trust. *International Journal of Human–Computer Interaction* 39, 13 (2023), 2724–2741.

[46] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2019. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision* 128, 2 (Oct. 2019), 336–359. https://doi.org/10.1007/s11263-019-01228-7

[47] Hua Shen and Ting-Hao Kenneth Huang. 2020. How Useful Are the Machine-Generated Interpretations to General Users? A Human Evaluation on Guessing the Incorrectly Predicted Labels. arXiv:2008.11721 [cs.HC] https://arxiv.org/abs/2008.11721

[48] Vivswan Shitole, Fuxin Li, Minsuk Kahng, Prasad Tadepalli, and Alan Fern. 2021. One Explanation is Not Enough: Structured Attention Graphs for Image Classification. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., Virtual, 11352–11363. https://proceedings.neurips.cc/paper_files/paper/2021/file/5e751896e527c862bf67251a474b3819-Paper.pdf

[49] Herbert Alexander Simon. 1997. *Models of bounded rationality: Empirically grounded economic reason*. Vol. 3. MIT press, Cambridge, MA.

[50] Paras Nath Singh and Tara P. Gowdar. 2021. Reverse Image Search Improved by Deep Learning. In *2021 IEEE Mysore Sub Section International Conference (MysuruCon)*. IEEE, Mysuru, India, 596–600. https://doi.org/10.1109/MysuruCon52639.2021.9641572

[51] Abby Stylianou, Jessica Schreier, Richard Souvenir, and Robert Pless. 2017. TraffickCam: Crowdsourced and Computer Vision Based Approaches to Fighting Sex Trafficking. In *2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*. IEEE, Washington, DC, USA, 1–8. https://doi.org/10.1109/AIPR.2017.8457947

[52] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. arXiv:1703.01365 [cs.LG] https://arxiv.org/abs/1703.01365

[53] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 109–119. https://doi.org/10.1145/3397481.3450662

[54] Chun-Hua Tsai and Peter Brusilovsky. 2019. Evaluating Visual Explanations for Similarity-Based Recommendations: User Perception and Performance. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization* (Larnaca, Cyprus) *(UMAP '19)*. Association for Computing Machinery, New York, NY, USA, 22–30. https://doi.org/10.1145/3320435.3320465

[55] Tommaso Turchi, Alessio Malizia, Fabio Paternò, Simone Borsci, and Alan Chamberlain. 2024. Adaptive XAI: Towards Intelligent Interfaces for Tailored AI Explanations. In *Companion Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24 Companion)*. Association for Computing Machinery, New York, NY, USA, 119–121. https://doi.org/10.1145/3640544.3645253

[56] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008), 24 pages.

[57] Warren J Von Eschenbach. 2021. Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology* 34, 4 (2021), 1607–1622.

[58] Basil Wahn, Laura Schmitz, Frauke Nora Gerster, and Matthias Weiss. 2023. Offloading under cognitive load: Humans are willing to offload parts of an attentionally demanding task to an algorithm. *Plos one* 18, 5 (2023), e0286102.

[59] Bo Wang, Shuo Jin, Qingsen Yan, Haibo Xu, Chuan Luo, Lai Wei, Wei Zhao, Xuexue Hou, Wenshuo Ma, Zhengqing Xu, et al. 2021. AI-assisted CT imaging analysis for COVID-19 screening: Building and deploying a medical AI system. *Applied soft computing* 98 (2021), 106897.

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

[61] Albatool Wazzan, Imtiaz Ahmad, Stephen Macneil, and Richard Souvenir. 2024. Context or Clutter? Efficiently Matching Objects Across Scenes. In *Proceedings of the 2024 International Conference on Multimedia Retrieval* (Phuket, Thailand) *(ICMR '24)*. Association for Computing Machinery, New York, NY, USA, 404–413. https://doi.org/10.1145/3652583.3658090

[62] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. Association for Computing Machinery, Barcelona Spain, 295–305.

[63] Roland S Zimmermann, Judy Borowski, Robert Geirhos, Matthias Bethge, Thomas Wallis, and Wieland Brendel. 2021. How well do feature visualizations support causal understanding of CNN activations? *Advances in Neural Information Processing Systems* 34 (2021), 11730–11744.