Imago Obscura: An Image Privacy AI Co-pilot to Enable Identification and Mitigation of Risks

Kyzyl Monteiro*
kyzyl@cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Yuchen Wu*
wuyuchen21@mails.tsinghua.edu.cn
Tsinghua University
Beijing, China

Sauvik Das
sauvik@cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Figure 1: Imago Obscura: A privacy-focused image AI co-pilot that enables users to: 1) articulate their image sharing intent and privacy concerns; 2) become aware of multiple contextually pertinent image privacy risks; and 3) apply recommended obfuscation techniques for the risks they choose to address, enabling informed decision-making about image sharing.

ABSTRACT

Users often struggle to navigate the privacy / publicity boundary in sharing images online: they may lack awareness of image privacy risks or the ability to apply effective mitigation strategies. To address this challenge, we introduce and evaluate Imago Obscura, an intent-aware AI-powered image-editing copilot that enables users to identify and mitigate privacy risks in images they intend to share. Driven by design requirements from a formative user study with 7 image-editing experts, Imago Obscura enables users to articulate their image-sharing intent and privacy concerns. The system uses these inputs to surface contextually pertinent privacy risks, and then recommends and facilitates application of a suite of obfuscation techniques found to be effective in prior literature — e.g., inpainting, blurring, and generative content replacement. We evaluated Imago Obscura with 15 end-users in a lab study and found that it improved users' awareness of image privacy risks and their ability to address them, enabling more informed sharing decisions.

CCS CONCEPTS

• **Human-centered computing**; • **Security and privacy → Usability in security and privacy**;

KEYWORDS

usable privacy, human-AI teaming, generative AI, intent-aware tool

ACM Reference Format:
Kyzyl Monteiro, Yuchen Wu, and Sauvik Das. 2025. Imago Obscura: An Image Privacy AI Co-pilot to Enable Identification and Mitigation of Risks. In *The 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25), September 28-October 1, 2025, Busan, Republic of Korea.* ACM, New York, NY, USA, 26 pages. https://doi.org/10.1145/3746059.3747633

1 INTRODUCTION

One of the greatest usable privacy challenges of the modern social internet is helping users navigate what Palen and Dourish, in 2003, identified as the 'privacy/publicity' boundary: i.e., people's desire to share personal information with others without exposing themselves to undue risks [50]. This problem manifests acutely in the context of image sharing — people collectively share 14 billion images daily [6], for reasons ranging from sharing and documenting moments in their personal lives to collaborating and communicating in their professional lives [7, 60]. But sharing images also comes with risks: many personal images can reveal a wide range of potentially sensitive information: e.g., who one knows, where one goes, and what one likes to do [2, 13, 51]. These risks are not abstract: prior work has shown that violations of image privacy and security can lead to a spectrum of harms from interpersonal threats, including personal embarrassment, job loss, identity theft, stalking, and harassment [46, 53, 61, 63].

Despite these risks, a large body of prior art suggests that many users have trouble understanding and mitigating the privacy risks associated with sharing personal images online [22, 35, 41, 45, 48,

*Both authors contributed equally to this work.

UIST '25, September 28-October 1, 2025, Busan, Republic of Korea

ACM ISBN 979-8-4007-2037-6/2025/09
https://doi.org/10.1145/3746059.3747633

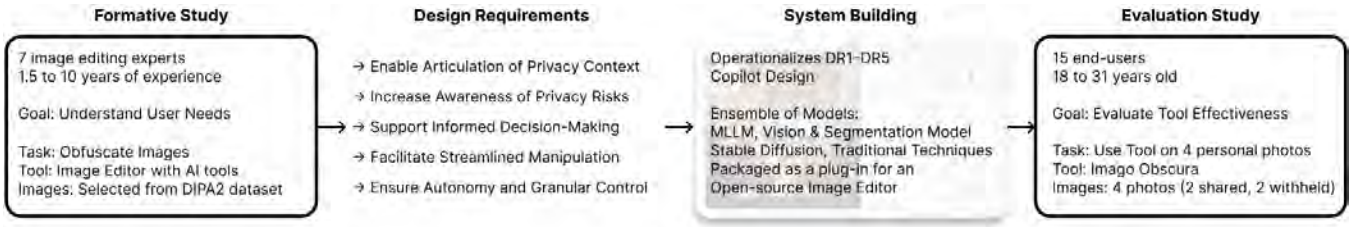


Figure 2: Overview of our methodology. We conducted a formative study to derive design requirements, built a tool based on those requirements, and evaluated it with end-users using their personal photos.

63]. Some users are ***unaware*** of the risks they expose themselves to while sharing an image [22, 35, 48, 63]; others may find they have ***limited ability*** to mitigate these risks because existing approaches to obfuscate or redact sensitive information from images require significant technical expertise [41, 45]. As a result, today, users generally employ crude and restrictive strategies when navigating the privacy/publicity boundary while sharing images online — i.e., they ignore image privacy risks altogether, self-censor themselves, or utilize insufficient and error-prone audience selection controls to try and limit who can see their images [14, 39, 58, 74].

How can we make it easier for users to effectively identify and mitigate interpersonal privacy risks[1] in images they want to share online? To answer that question, we introduce Imago Obscura—an image privacy copilot that leverages generative AI technologies to help end-users identify image privacy risks, mitigate those risks, and make more informed decisions about what images to share online. We employed a three-phased, human-centered design process to design and evaluate Imago Obscura.

First, we conducted **a formative study** with seven image-editing experts to understand how they approach making privacy-preserving image edits. From this study, we distilled five design requirements to prioritize when developing Imago Obscura. For example, we learned that while there is a general need to raise users' awareness of image privacy risks, it is imperative to do this in a manner that is customized to users' *lived concerns and contexts of use*.

Second, we **designed and developed Imago Obscura** based on the design requirements we derived from our formative study. Imago Obscura enables users to directly articulate their privacy concerns and sharing intent through natural language, and uses this information to identify pertinent risks in users' images. It then recommends appropriate obfuscation techniques while informing users of the implications of those techniques, and automatically applies obfuscation strategies users choose to implement.

Finally, we **evaluated Imago Obscura** through a lab user study with 15 participants. In short, participants found that Imago Obscura helped them mitigate the privacy concerns they really cared about, surfaced relevant risks that would have otherwise gone unnoticed, and aided them in making an informed decision about which risks to accept and which to mitigate based on their sharing intent. We also found some opportunities for improvement. For example, there is a need for guardrails to prevent malicious use — as by lowering the barrier to applying obfuscation techniques to their images, Imago Obscura also facilitates the creation of inauthentic or misleading images. On balance, however, we found that Imago Obscura helps users make more informed decisions balancing their desires for privacy and publicity when sharing images online.

In summary, this paper contributes:

(1) Five core design requirements for an image privacy copilot to help users make informed decisions about how to balance privacy concerns with sharing intent when obfuscating images they hope to share online.
(2) The design and implementation of a system, *Imago Obscura*, which demonstrates a novel orchestration of AI techniques to create an intent-aware, image privacy co-pilot that enables users to:
   (a) receive personalized image privacy support tailored to threats they express in natural language;
   (b) apply, compare, and contrast diverse semantic obfuscations to make informed decisions on risk mitigation.
(3) Insights from a summative user study in which we learned users valued a) the ability to articulate concerns as it helped them address risks they deemed most important; b) the comprehensive presentation of risks and recommended obfuscations, which supported informed decision-making; and c) the ability to apply recommended image obfuscations easily with precise control and high agency.

2 RELATED WORK

2.1 Identifying and Taxonomizing Privacy-Sensitive Content

Studies and Taxonomies on Sensitive Content. Categorizing sensitive content in images is well-studied: e.g., prior work highlights various sensitive content elements such as faces, objects, backgrounds, and phone screens [2, 4, 23]. A more recent meta-analysis of prior literature and an analysis of photos collected from participants has culminated in a taxonomy of 28 categories of sensitive content in images [42, 43].

Datasets on Sensitive Content. Prior work in computer vision has contributed a number of labeled datasets for sensitive content detection. Some datasets label images as simply private or public [59, 75], while others label more granular categories of sensitive content like nudity, violence, and drinking [49, 71–73, 76]. Recent work by Xu et al. adds detailed reasoning for image privacy labels, in addition to object-level annotations of sensitive content [67, 68] — this granular annotation is particularly important for content-level image privacy protection.

[1]Throughout this work we adopt an interpersonal threat model: the adversary is a human viewer who can identify and infer sensitive details from image content, rather than a large-scale automated classifier. A full description, including excluded institutional and purely algorithmic threats are mentioned in the Appendix A.1.

In short, prior art provides a comprehensive categorization of image privacy risks. We build on this prior art: Imago Obscura uses these taxonomies of sensitive content in order to surface content-level privacy risks to users in an easily accessible manner that promotes informed decision-making.

2.2 Image Obfuscation Techniques and Their Effectiveness

Image obfuscation techniques have evolved from traditional methods like blurring and pixelation [10, 33, 36] to advanced approaches including inpainting, avatar replacement, and generative content replacement [21, 30, 40, 66]. Various studies have examined the effectiveness of these methods [19, 20, 44]. With the advent of image generation models, recent studies have proposed generative content replacement as an obfuscation technique and have found them to be effective [30, 66].

While prior work has explored the effectiveness of different obfuscation techniques, it also calls for the need for a human-in-the-loop image privacy protection tool that helps users make informed decisions [44, 66]. Imago Obscura answers this call by raising user awareness of the various obfuscation techniques and their effectiveness as it relates to users' specific image sharing goals. Existing tools, ranging from advanced editors to simplified redaction apps [1, 11, 16], either require significant expertise or focus narrowly on specific elements like faces or text, and in both cases, offer little support for user intent or contextual awareness, making them difficult for end-users to use effectively. We adopt a user-centered design approach to develop an AI copilot, bridging the gap between advanced obfuscation techniques and end-user application.

2.3 Automated Systems for Image Privacy Identification and Protection

Researchers have made significant strides in developing automated tools for identifying and protecting sensitive content in images. Early efforts focused on specific elements, such as Hasan et al.'s tool to distinguish bystanders from subjects in photos [18] and Korayem et al.'s work on detecting screens [32]. Expanding on this body of work, researchers also began integrating automatic obfuscation techniques. For instance, Ilya et al.'s Face/Off system recognizes and blurs faces for which the owner lacks permission [24], while Frome et al. created a system for Google Street View that automatically detects and blurs faces and license plates [15]. More recently, automated obfuscation techniques have also been extended to video and device-based privacy contexts to replace sensitive content in video streams [21, 25].

More recent work has explored user-centered approaches. Li et al. proposed design considerations for an image obfuscation tool based on a Wizard of Oz study [41]. Additionally, Vishwamitra et al. introduced AutoPri, a novel system enabling automatic and user-specific content-based photo privacy control [62]. However, all automated obfuscation systems frame privacy as a classification task: identifying sensitive content and applying traditional obfuscation (e.g., blurring, masking) in a context-agnostic way without consideration of user sharing intent. They offer limited user control and rarely support understanding why elements are risky or how risks relate to sharing intent.

Our work both extends and challenges this line of work by re-imagining privacy protection not just as a matter of automation, but as a process of helping users balance the privacy–publicity trade-off. We argue that users need to be active participants in this pipeline, as the boundary between privacy and publicity depends on contextual knowledge and intent. Our approach, realized through a formative study, moves beyond automation toward a copilot model, where the AI and user collaboratively manage image privacy.

3 FORMATIVE STUDY

We build on prior research demonstrating the potential of AI for image privacy [45, 66], by exploring how AI-powered features can simplify the process of identifying and mitigating image privacy risks, without requiring specialized expertise. To inform the design of our intelligent image obfuscation tool, we conducted a formative study with image editing experts. This study aimed to uncover user needs beyond basic usability barriers, focusing instead on their workflows, decision-making processes, the techniques they employ, and the challenges they encounter.

3.1 Participants

We recruited seven image manipulation experts (E1-E7) with 18 months to 10 years of experience in image editing, with some being self-taught (E3, E5) and others having taken design and digital art courses (E1, E2, E4, E6, E7).

3.2 Study Procedure

The study consisted of two major components (1) an image obfuscation task, and (2) a post-task semi-structured interview. The task was screen recorded and the interview was audio recorded. After obtaining informed consent, participants completed a brief demographic questionnaire and received an introduction to the concept of "image obfuscation" and the study's goals.

For the image obfuscation task, participants were asked to select images from a subset of the DIPA datasets [67, 68], which is a recent dataset that has records of images, annotated sensitive elements, the associated risks, and the sensitivity of the element [67, 68]. The subset included over 115 images that were high quality and tagged as having sensitive content that had a user-reported score of more than 5 out of 7 to ensure that the images were ones that participants would have concerns about. Participants were asked to envision a relevant sharing intent and a privacy concern for each image they chose. Participants were then provided with Krita [34], an open-source image editor which we additionally equipped with AI tools such as object segmentation, bounding box segmentation, text-based generation replacement, reference image-based generation, and avatar replacement. Participants were instructed to use these tools to obfuscate their chosen images for privacy preservation. To contextualize their work, we provided participants with privacy knowledge materials, including: 1) A list of potential sensitive content in images 2) Potential threats associated with image sharing 3) Examples of obfuscated images (before and after) from previous work [44, 66], and a few examples the authors created.

Finally, for the post-task semi-structured interview, we asked participants questions about the images they obfuscated, their workflow and thought process, the rationale behind their choices, their

task experience, and suggestions and aspirations they have for an image obfuscation tool.

3.3 Analysis

To elicit user requirements for a system, we conducted a two-phased analysis. First, we compared and contrasted the risks participants identified in the images they chose, to the risks that were pre-identified in the DIPA dataset [67, 68]. Doing so allowed us to codify the elements and risks identified by our participants, which they identified beyond what was tagged in the original dataset, and which risks they seemed to miss. Second, we began with open coding of the interview transcripts, followed by a thematic analysis to identify patterns across participants [5]. Two researchers coded the transcripts independently, and then later came together to resolve any conflicts in coding. We also analyzed the screen recordings to understand how the users executed the tasks.

3.4 Findings

Our analysis revealed several findings, some of which confirmed previous knowledge, while others provided novel insights. Echoing prior literature [66], participants (E2, E4, E6, E7) identified both advantages and drawbacks of using AI-powered image editing techniques for privacy. For example, participants appreciated that AI techniques allow for natural outputs and faster editing. However, our study revealed that users also appreciated the determinism and reliability of non-AI editing techniques. The stochastic nature of AI outputs occasionally frustrated participants:"it generated weird people .. it looks like witch-craft .. makes it nonsensical" (E2). Also akin to prior work [41, 62], we found that users followed a fairly standard workflow consisting of: Identifying sensitive content, selecting the sensitive content, and applying an obfuscation technique (E1-7). However, we also identified previously undocumented pain points with image editing for privacy. These challenges sharpened our understanding of the goal of obfuscation—to find a user-acceptable balance between reaping the social advantages of sharing personal information and mitigating privacy risks. Through our analysis of user needs, we also realized an underlying theme: while participants appreciated automation, they frequently also wanted control to express their preferences. This finding further pointed to collaborative, co-pilot style system rather than one that fully automates the process of identifying and mitigating privacy risks. We next discuss these pain points and design opportunities as they relate to three different phases of the standard workflow we identified.

3.4.1 Identifying sensitive content — Pain points and design opportunities.

Sharing intent and lived privacy concern impact what people want to obfuscate. In a few cases, identifying sensitive content was simple for our participants. Indeed, we found that participants consistently associated certain objects in images with privacy risks: e.g., license plates and laptop screens (E1, E2, E4, E6). "I obscure the number plate...I think that's obvious"(E2). But there were other situations where this identification process was more nuanced. We found that when different participants chose the same image to obfuscate, they often chose *different* sensitive objects within the image to obfuscate.

When exploring why, we found that participants were heavily influenced by their envisioned sharing intent and privacy concern (E1-7): "maybe your family and friends aren't comfortable with the fact that you're gay" (E1); "maybe blonde hair is identifiable in this country" (E4). This finding — that different people have different intentions and concerns and thus take different approaches to mitigating privacy risks on the same image — led us to realize there is **a need for a flexible, content-aware system that can take into account users' sharing intent and privacy concerns**.

Participants needed content-relevant reminders of privacy risk. During the study, some participants referred to the provided list of privacy risks to identify sensitive content in images. Several found it helpful when identifying threats (E1, E3, E6). "it gave me a good idea of what different options I could do and what elements I should look for". However, others engaged with it at a surface level, describing it as too large or detailed (E1, E2, E6, E7). They noted that while the list was helpful, it wasn't convenient. Notably, some participants also said the list surfaced risks they had forgotten or hadn't previously considered (E1, E3, E4). "... I did not know .. that even appearance and self-presentation could be privacy concerns" (E3). Taken together, these findings suggest that **users need to be made aware of the broad spectrum of content-specific privacy risks**.

3.4.2 Selecting Sensitive Content — Pain points and design opportunities.

Despite all participants having more than one year of experience with image editing, they struggled with the process of selecting sensitive content in an image after they had identified that content as sensitive. For example, they faced difficulties in precisely selecting an object or ensuring that the correct layer was selected when manipulating the image (E1, E4-7)."it was a little confusing for me with all the layer stuff, but it wasn't super confusing, especially since I had prior experience" (E1). Participants preferred to use the object selection tool (an AI-powered object segmentation tool) to precisely select an area of interest (E1, E2, E4). This observation further highlighted the **need to reduce the complexity in directly selecting and manipulating sensitive content in images**.

3.4.3 Obfuscating Images — Pain points and design opportunities.

Participants select obfuscation techniques based on prior familiarity, not effectiveness at mitigating privacy risks. Some participants implicitly considered the pros and cons of different obfuscation techniques. For example, one participant chose image generation over blurring as it produced "believable fake additions" (E4). However, many participants selected an obfuscation technique based on convenience and prior familiarity, rather than based on the appropriateness of a technique vis-a-vis a specific privacy risk and/or sharing intent (E1, E2, E4-6). For instance, participants often selected pixelation or blurring, even though these techniques were less effective than other options. When asked why they chose to use the same obfuscation techniques and not consider others, E6 rationalized: "I just forgot how to do it." These observations highlight the **need for a system that makes users aware of and understand the pros and cons of broad spectrum of obfuscation techniques**.

Figure 3: Imago Obscura addresses "self disclosure risks". (1) Identifies that the numbered candle can reveal personal information. (2) Recommends removing the candle from the image. (3) Precisely selects the sensitive area, the candle, and applies inpainting.

The most appropriate obfuscation technique varied based on sharing intent and participant preference. Our analysis showed that personal preference and sharing intent were the most important factors for participants when choosing obfuscation techniques. For instance, E2 chose to blur a bystander in an image for a research paper because they wanted the obfuscation to be apparent to the viewer. Others experimented with different obfuscation techniques and ultimately selected one that best matched their preference and comfort level. We noted that participants exhibited varying levels of comfort with sharing obfuscated images—E2 was reluctant to share a heavily obfuscated image online, while E7 was more open, even after heavily manipulating an image by adding obvious fictional elements, like a snow castle in the background. Participants also aimed to more precisely control the output of AI-powered obfuscation techniques by using prompts and reference images. E3, for example, replaced the face of a bystander in an image with one of a specific public figure by using a reference image of that public figure: "I replaced the face .. with the photo of a model... models are like very much public" (E3). Overall, these findings further highlight that a purely automated approach to obfuscation is unlikely to be widely accepted: **users want granular control when applying and fine-tuning obfuscations**.

4 IMAGO OBSCURA

We distilled five design requirements from our formative study. Guided by these requirements, Imago Obscura enables users to make informed decisions about if and how to obfuscate personal images by guiding users through a structured workflow. In this section, we describe the five design requirements and provide examples of how they were operationalized in Imago Obscura. Figure 1 shows the entire user workflow in sequence.

4.1 DR1: Enabling expressive articulation of privacy concern and sharing intent

An image privacy copilot should adapt to users' directly expressed privacy concerns and sharing intent.

Natural language expression. Imago Obscura allows users to directly articulate their privacy concerns and sharing intent through natural language. To scaffold this articulation, we ask users two questions inspired by a study conducted in previous work [41] — i.e., "What's your main purpose of sharing this image?" and "Do you have any privacy concerns about this image?". For example, a user can express that they want to announce the birth of their child, but do not want to show their face (Appendix Figure 13).

Visual annotation of areas of concern. Natural language is powerful, but sometimes it is faster and quicker for users to directly select and manipulate sensitive content in images. For example, when there are multiple people in the photo, it may be easier to directly click on a bystander's face than to type out a description of the bystander (Appendix Figure 14). Therefore, we gave users the option to directly select concerning content in an image.

4.2 DR2: Increasing awareness of content-level privacy risks

An image privacy copilot should proactively surface potentially risky content beyond a user's immediate privacy concern. Identifying and addressing privacy risks associated with image sharing is essential for informed decision-making. Imago Obscura identifies five content-related categories of image privacy risk, drawn from an analysis of prior art that includes taxonomies of sensitive content [42, 43, 68] and image privacy risk at large [2, 13, 51, 68]. Below, we describe each of the five image privacy risks Imago Obscura identifies, and provide illustrative examples of each.

Self-Disclosure Risk. Self-disclosure risk involves the unintended revelation of personal details that may compromise an individual's privacy. This risk occurs when subtle cues in an image reveal information such as personal habits, health conditions, private life events, etc, or allow outsiders to infer details about someone's interests, affiliations, or habits: e.g., a photo of a bookshelf might suggest certain intellectual or political leanings; a gym bag could hint at fitness routines; medication bottles can disclose health issues; specific types of food may imply dietary restrictions or choices. In Figure 3, a mother wanting to share a photo of her child cutting a birthday cake may unwittingly reveal the child's age through a numbered candle. Imago Obscura identifies this detail and flags it as a self-disclosure risk.

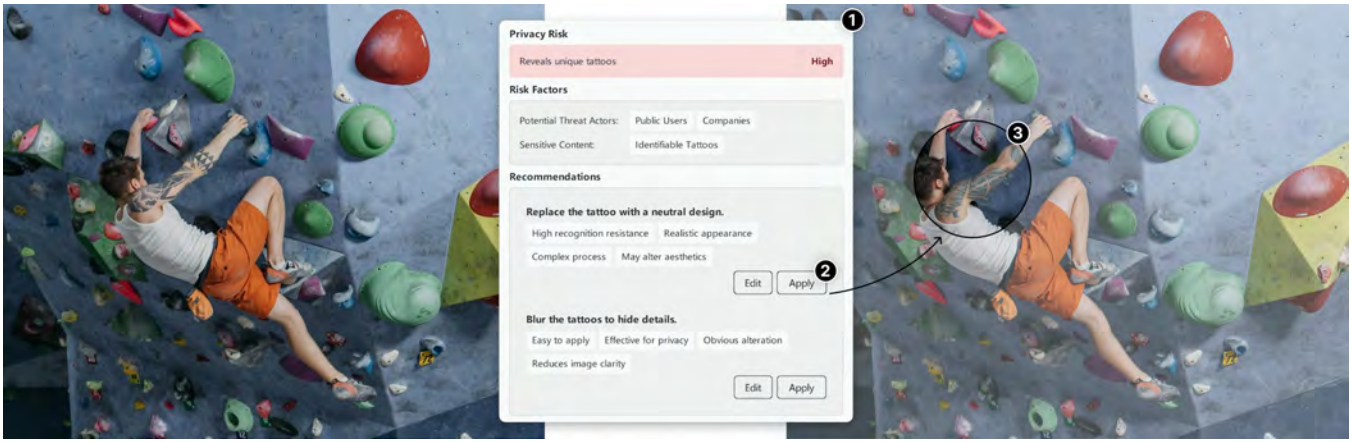


Figure 4: Imago Obscura addresses "identity exposure risk". (1) Identifies that the tattoo can reveal the person's identity. (2) Recommends to replace the tattoo with a new one. (3) Precisely selects the sensitive area, the tattoo, and applies generative content replacement.

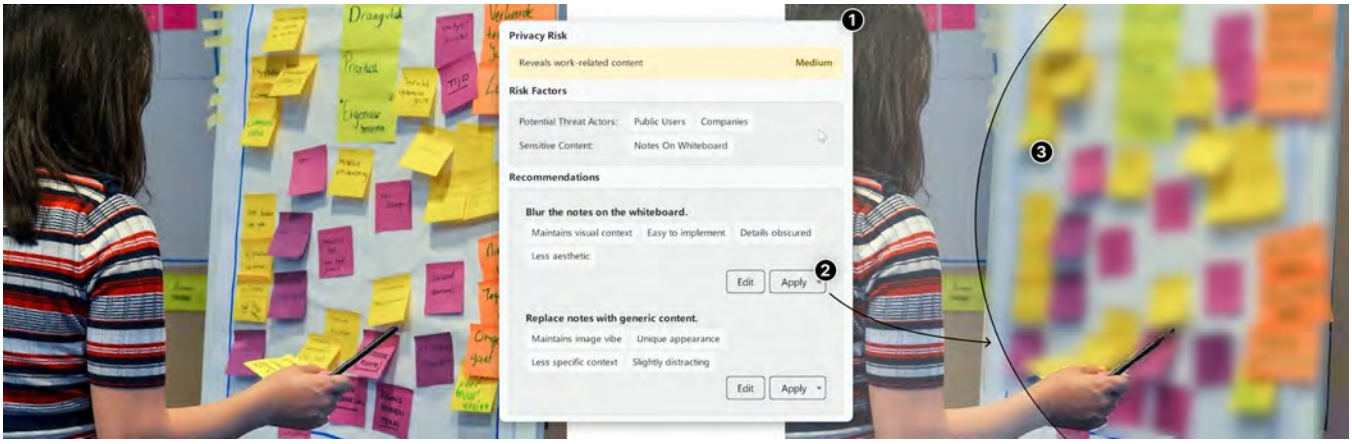


Figure 5: Imago Obscura addresses "confidential information leakage risk". (1) Identifies that the notes on the board can reveal confidential information. (2) Recommends to blur the notes on the board. (3) Precisely selects the sensitive area, the board, and applies blur.

Identity Exposure Risk. Identity exposure risk refers to the potential for an individual's identity to be uncovered through visible personally identifiable information (PII) or distinguishing information such as facial features, ID cards, or unique body marks. As shown in Figure 4, a climbing gym photo intended for the website's hero image includes a person whose face is obscured, but a visible tattoo could still reveal their identity.

Confidential Information Leakage Risk. Confidential information leakage risks occur when secret or proprietary information is visible in the background of an image. This can happen, for example, in business environments where whiteboards or computer screens are captured or documents are visible on a desk. For instance, in Figure 5, a researcher planning to share a photo from a collaborative workshop captures notes on a whiteboard. Imago Obscura flags these notes as having the potential for confidential information leakage and suggests techniques to mitigate this risk.

Location Exposure Risk. Location exposure risk involves identifiable location information being revealed, compromising the user's physical privacy and safety. These risks can arise when recognizable landmarks, or specific weather patterns are visible. For instance, in Figure 6, a person wants to share their home office setup, but the view from the window shows recognizable buildings, potentially disclosing the location of their home. Imago Obscura identifies and alerts the user to potential location exposure risks.

Bystander Risk. Bystander risks arise when individuals in the background of an image are unintentionally captured. This can occur in crowded public places or events, where bystanders may not be aware that they are being photographed. Examples include street scenes, public gatherings, or casual photos taken in parks. In Figure 7, a marathon runner shares a picture of themselves running, but a bystander's face is visible in the background. Imago Obscura flags the bystander, suggesting techniques to obscure their identity.

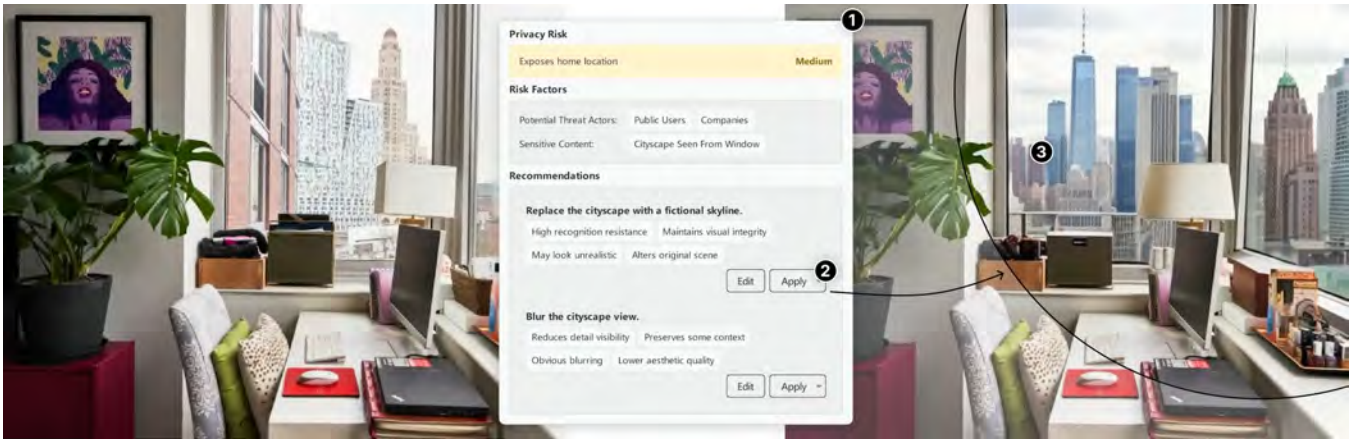


Figure 6: Imago Obscura addresses "location exposure risk". (1) Identifies that the window view can reveal the location. (2) Recommends to replace the window view. (3) Precisely selects the sensitive area, the window, and applies generative content replacement.

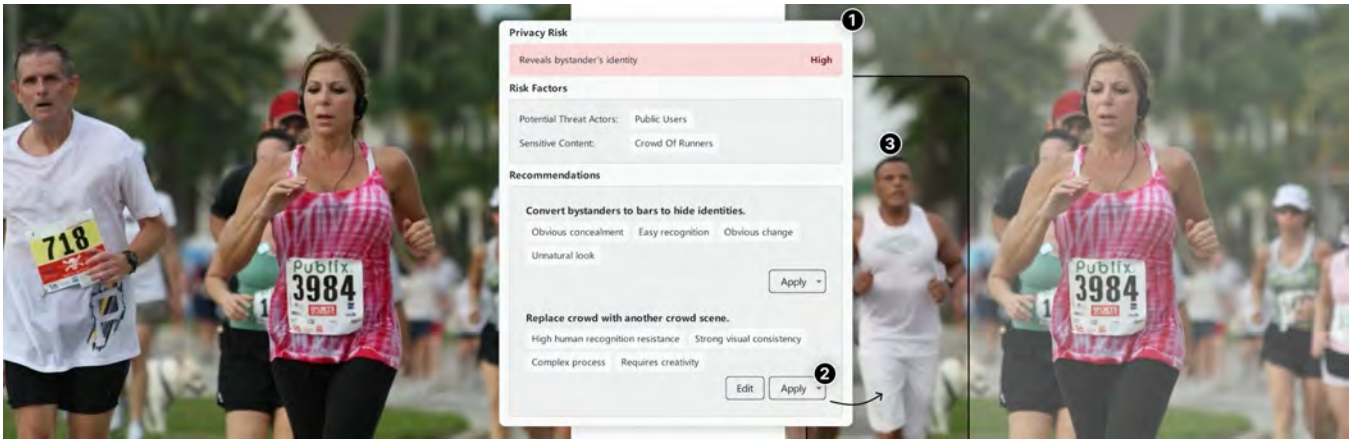


Figure 7: Imago Obscura addresses "bystander privacy risk". (1) Identifies that the bystanders' privacy might be at risk. (2) Recommends to generate a new running crowd scene. (3) Precisely selects the sensitive area, the bystander, and applies generative content replacement.

4.3 DR3: Promote informed decision-making

An image privacy copilot should provide users with explanations of privacy risks it identifies, and promote obfuscation strategies that minimally interfere with sharing intent. To promote informed decision-making, we provide users with detailed explanations of both the risks identified and obfuscation techniques recommended to address those risks.

Presenting Risks, Sensitive Content, Threat Actors, and Severity. While surfacing pertinent risks is crucial, we realized the importance of presenting these risks to users in a manner that is easy to understand and act upon. To ensure user comprehension, we present identified risks in natural language, phrasing them to account for the user's specific concerns and/or sharing intent. For example in Figure 6, Imago Obscura presents a location exposure risk with the label "Exposes your location" and provides an explanation of what content may result in the risk: "Cityscape seen from window". Furthermore, we present the sensitive content from which the risk arises and potential threat actors who might be able to exploit the risk. Recognizing that some sensitive elements can reveal more information than others and that risks vary in severity, we also classify and present the severity of each risk to the user as High, Medium, or Low.

Presenting Image Obfuscation Techniques and Their Attributes. Our formative study revealed that participants often chose image obfuscation techniques arbitrarily, partly due to forgetting about the gamut of available options. Therefore, beyond risk identification, Imago Obscura presents obfuscation techniques for each identified risk: converting the recall problem into one of recognition. Our tool enables the application of a broad range of obfuscation techniques curated from existing literature [30, 44, 66]. The list of curated obfuscation techniques is in Appendix A.3, and also visualized in Figure 8. The formative study also showed that participants often overlooked the effect of obfuscation techniques on the final image. Accordingly, we highlight each technique's unique properties and

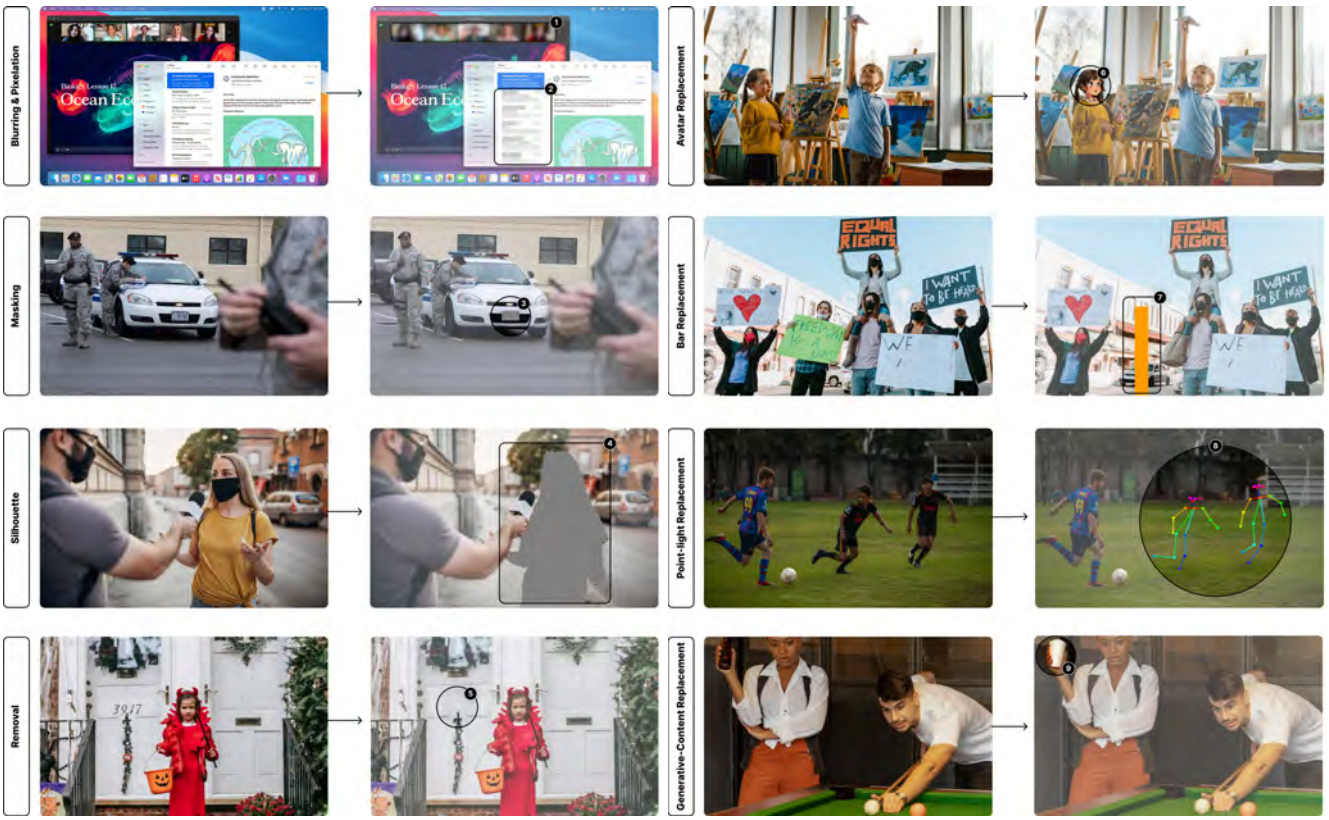


Figure 8: Demonstration of the diverse image obfuscation techniques enabled by Imago Obscura. Each pair shows the original (left) and obfuscated (right) image: (1-2) *Blurring* and *pixelation* of screen content to protect confidential information (3) *Masking* of license plate to preserve vehicle anonymity (4) *Silhouette masking* to anonymize a whistle-blower in a news article (5) *Removal* of house number to conceal the specific location (6) *Avatar replacement* to protect identity of a child's friend (7) *Bar replacement* to obscure a fellow participant (8) *Point-light representation* shows body pose while preserving anonymity; (9) *Generative replacement* of an alcohol bottle avoids promoting alcohol.

how it affects the image. Informed by prior art, we consider various attributes to define the effectiveness of each obfuscation technique: effectiveness against recognition, detectability, visual harmony, narrative coherence, realism, and vulnerability [19, 44, 66]. These attributes are presented in Appendix Table 1.

4.4 DR4: Facilitate easy and effective application of obfuscation techniques

An image privacy copilot should present users with a streamlined process to reduce the complexities of precisely selecting the sensitive content and applying effective obfuscation techniques. To simplify the mitigation privacy risks that users want to address, we made both selection of risky content pertinent to those risks and application of obfuscation techniques accessible through one-click interactions.

Precise selection of risky content. After a user chooses which risk they would like to address, Imago Obscura enables the selection of the sensitive content pertinent to the risk automatically with a one-click action. The system then precisely selects the sensitive content and awaits confirmation from the user.

Easy application of obfuscation techniques. On confirmation that the selection aligns with the user's intention, Imago Obscura automatically applies the chosen obfuscation technique, also with a one-click interaction. Figure 8 illustrates the different obfuscation techniques that are possible through Imago Obscura — these include traditional obfuscation techniques like blurring, pixelation, and masking, as well as AI-powered techniques like removal/inpainting, point light replacement, and generative content replacement.

4.5 DR5: Ensure autonomy and granular control

An image privacy copilot should remain just that — a copilot. It should afford users ultimate authority to make decisions about what and how to obfuscate an image. To ensure autonomy and granular control, Imago Obscura affords users choice at every step of the risk identification and mitigation workflow.

Choice over which risks to mitigate, and how to mitigate them. While we present various risks pertinent to the image and the user's sharing intent, users retain full control over which privacy risks they want to address. For instance, if a user is sharing a photo of themselves in front of the Eiffel Tower, they can make an intentional

choice to forgo location exposure risks. For each sensitive content segment linked to a risk, the system recommends at least two obfuscation techniques from which the user can choose.

Manual refinement of automatic selection. While the system automatically selects sensitive objects for users to obfuscate to address a specific risk, they are also afforded the option to refine their selection of objects in the image prior to applying the recommended obfuscation technique.

Ad hoc use of obfuscation techniques. Beyond the recommended obfuscation techniques, Imago Obscura also enables users to apply obfuscation techniques in an ad hoc manner. It does so by presenting a toolbar with an AI-powered precise selection option, two traditional image transformations (blurring and masking), and two AI-powered image generation-based obfuscation techniques (generative content replacement and avatar replacement).

Granular control over obfuscation techniques. Imago Obscura provides users with granular control over obfuscation techniques through the use of intensity control sliders, text prompts, and reference image upload options. For example, users can increase the blur on confidential information, replace a bystander's face with that of a reference photo, or create a fictional background.

5 IMPLEMENTATION

We implemented Imago Obscura as a plugin for the open-source graphics editor Krita [34]. Imago Obscura utilizes an integrated ensemble of four AI models to enable the workflow and design space we described. Specifically, it leverages a multimodal large language model, GPT-4o [3, 52], to identify privacy risks in images based on user's expressed privacy concerns and the taxonomy of privacy risks we described above; a vision model, Florence 2 [65], to automatically annotate images with bounding boxes and labels for objects found in the image; a segmentation model, SAM [31], to get precise selections of sensitive content in the image, helping us to associate privacy concerns with specific regions of the image; and, a text-to-image generation model, stable diffusion [28, 54], to automatically apply AI-powered image obfuscation techniques if the user so chooses. A full overview of this process can be seen in Figure 9.

Users select an image, and can articulate their privacy concerns and sharing intent of that image through natural language and visual annotation. Once the user presses a button to analyze privacy risks, we feed the image, users' concerns and sharing intent, and the taxonomy of image privacy risks we synthesized from prior literature through our ensemble of multimodal AI models. Using this input, Imago Obscura identifies sensitive content in the image that users may consider obfuscating. These risks are presented to users in the form of explanations of why that content may be risky. Finally, users can choose to act on any of the identified risks. For each risk, the system presents a subset of relevant obfuscation techniques from the nine techniques we found in prior literature (Appendix A.3). Users can easily apply these techniques through simple click-based interactions. The cumulative effect of this workflow is that users get highly customized and personalized assistance with identifying and mitigating pertinent privacy risks in images they hope to share online.

5.1 Pre-scan Process

Prior to engaging the Multi-modal Large Language Model (MLLM), the image undergoes a preliminary scan using the Florence vision model [65], chosen for its robust object detection and classification capabilities. Florence provides labeled bounding boxes for all detected objects, which are then supplied to the MLLM. This preliminary analysis serves as a form of visual prompting, employing the "set-of-mark prompting" technique that has been shown to improve MLLM's visual reasoning capabilities [70]. The annotated image functions as a visual guide, informing subsequent steps in the privacy risk analysis and obfuscation process, enabling the MLLM to focus on targeted areas within the image.

5.2 Prompting Techniques to Identify Risks

The tool begins by developing an understanding of the image, integrating the user's sharing intent and privacy concerns. This process is guided through a series of structured prompts, using chain-of-thought prompting [64] to instruct the Multimodal Large Language Model (MLLM) to systematically analyze the image content, user context, and potential privacy risks (see Appendix A.7.1 for the prompt).

The system initiates the analysis by instructing the MLLM to examine the image and user's sharing intent and privacy concerns highlighted through text or visual annotations. User-provided visual annotations — areas marked in green — are provided as visual prompts and guided to be interpreted as direct indicators of privacy concerns. The MLLM is specifically instructed to prioritize these user-indicated regions to ensure that the user's specific privacy concerns are addressed first. Once these concerns are identified, the MLLM follows a two-step process.

(1) *Identify Sensitive Elements*: The model is guided to scan the image to identify potentially sensitive content by referring to a curated list of potential sensitive elements. This curated list, derived from prior literature, encompassed detailed elements in categories such as identity and personal information, nudity, and social contexts.
(2) *Assess Privacy Risks*: The MLLM is further guided to evaluate the identified sensitive elements to determine potential privacy risks, while referencing a curated list of privacy risks (Section 4.2). These risks are assigned a severity level (High, Medium, Low) and potential threat actors (e.g., public users, companies, acquaintances) based on context and are instructed to be presented in natural language.

5.3 Prompting Technique to Recommend Obfuscation

Following the identification of privacy risks, the system instructs the MLLM to identify and recommend appropriate image obfuscation techniques (see Appendix A.7.2 for the prompt). The model is prompted to choose from a list of obfuscation techniques (Appendix A.3). The MLLM is then instructed to generate up to two recommendations per sensitive element, using user-friendly, non-technical language to describe the obfuscation methods. These recommendations are tailored based on the image, the user's provided concern, and attributes identified in previous literature (Appendix Table 1).

Figure 9: Step-by-step outputs of each model in the Imago Obscura pipeline. 1) The vision model detects and labels objects with bounding boxes. 2) The MLLM identifies sensitive content and recommends obfuscation strategies (shown as a JSON object). 3) The vision model re-localizes the sensitive elements identified by the MLLM. 4) The segmentation model refines the selected region with precision. 5) The image generator replaces the selected region using the chosen obfuscation method.

Finally, we receive a JSON object that includes the risks, their severity, relevant threat actors, sensitive elements, and recommended obfuscation techniques with their attributes.

5.4 Locating and Selecting Sensitive Elements

To apply obfuscations precisely, the tool first aggregates all identified sensitive elements from the JSON object output by the MLLM. It then uses the Florence vision model to locate each sensitive element, generating bounding boxes to provide approximate locations. For more precision, these bounding boxes are passed to the Segment Anything Model (SAM) [31], which generates detailed contours. Finally, the detailed contours are used by the graphic editor's selection tool to enable precise content selection.

5.5 Applying Obfuscations

Imago Obscura applies traditional image obfuscations, such as blurring and pixelation, through the integrated image editing tools of Krita. For the AI-driven techniques, the system combines AI methods with these traditional transformations. Techniques like removal/inpainting, avatar replacement, and generative content replacement leverage a stable diffusion model to replace sensitive content with generated elements. For bar and point-light replacements, the sensitive element is first removed via inpainting, followed by the application of the respective replacement.

6 EVALUATION

To evaluate Imago Obscura, we conducted an in-person lab user study with 15 participants. The goal of our user study was three-fold. First, drawing on the Security and Privacy Acceptance Framework (SPAF), which outlines three key barriers that inhibit end-user adoption of new and expert-recommended security and privacy tools [9], we wanted to assess Imago Obscura's impact on users' awareness of, motivation to address, and their ability to mitigate pertinent privacy risks in images.Next, we aimed to assess how well Imago Obscura fulfilled the five design requirements we distilled from our formative study (DR1–5).

Finally, we wanted to understand to what extent users found Imago Obscura useful and usable. We considered but ultimately excluded comparisons to existing tools like Photoshop or prior privacy-specific systems as these tools do not surface privacy risks or support intent-aware mitigation strategies. Instead, we opted for a within-subjects design using participants' own images, allowing us to assess how the tool supports real privacy goals and decision-making in context.

To these ends, we used pre-task and post-task surveys, a final survey including the System Usability Scale (SUS), and a semi-structured exit interview. This protocol was revised based on insights we gained from an initial set of pilot studies we conducted with a separate set of 12 participants.

Note that we also conducted a technical evaluation of the risk identification component of our model pipeline to ensure that its outputs were robust and accurate — we share the details of that evaluation in the Appendix A.6. In short, with GPT-4o [3], our approach achieved an accuracy of ~70% for sensitive object identification, ~83% for risk category classification and ~73% for severity assessment on the DIPA2 dataset [68].

We consider this evaluation peripheral because, to some degree, Imago Obscura is model agnostic — if more accurate models become available in the future, Imago Obscura will be able to take advantage of them. Moreover, since Imago Obscura is a copilot and not a full automation tool, we consider this performance good enough to support users in making informed decisions.

6.1 Participants

We recruited 15 end-users (P1-P15) who had previously shared personal images online. Participants ranged in age from 18 to 31 years old (5 male and 10 female). All participants had an academic background including undergraduate students, PhD candidates, and research assistants. Six participants reported previous experience with image obfuscation techniques, using tools such as Adobe Photoshop, Background Remover app, Adobe Firefly, Canva's blurring tool, and built-in smartphone editing features.

6.2 Study Procedure

Our study lasted approximately one hour. Participants were asked to bring four personal images each to the study: two they previously shared online (shared images) and two they wanted to share but had withheld due to privacy concerns (withheld images). Participants were first briefed on the goal of Imago Obscura and the study and were then shown a video walkthrough of the tool before beginning the tasks All study procedures were approved by our institution's Institutional Review Board (IRB). Participants provided informed consent and were explicitly informed that their uploaded images would be processed using third-party AI services. They were given the option to opt out or substitute alternative images if desired.

Task: Participants were then asked to use the tool on each of the four images they brought. They were required to load each image into Imago Obscura, express their privacy concerns and/or sharing intent with that image, and have a look at the privacy risks surfaced; they were *not* required to make changes to the images.

Measurements: Participants were asked to fill in a pre-task and post-task questionnaire for each image. These questionnaires focused on measuring the tool's effectiveness against our design requirements through Likert scale questions (see Appendix A.5.1 for the full set of questions). Generally, these scales comprised of attitudinal questions such as "I feel that the tool understood my privacy concerns and sharing intent" where participants had to rate agreement from a scale of ranging from 1 (strongly disagree) to 5 (strongly agree). Following best practices in questionnaire design, some of our questions were reverse-coded.

After completing the tasks with all four images, participants were asked to fill in a final questionnaire, which aimed to measure how Imago Obscura addressed the SPAF barriers. The final survey also had a section with the System Usability Scale (SUS) questionnaire to evaluate overall usability.

Exit interview: After participants used Imago Obscura on all four of their images, we conducted a final semi-structured interview. The questions we asked participants were informed by their responses to the questionnaires they filled out for each image and centered around understanding whether the individual design requirements were met. For example, we asked questions like: "In the survey you indicated that the tool helped/did not help you identify privacy risks you hadn't considered before. Could you share more about what led you to this conclusion?". Participants were encouraged to refer to the four images they tested and give examples while answering these questions. We ended with a brief demographic questionnaire. This mixed-methods approach allowed us to evaluate Imago Obscura both quantitatively and qualitatively.

Analysis: We employed a mixed-methods approach to analyze our data. The interview responses were thematically analyzed by two researchers individually who later came together to resolve any conflicts [5]. Our approach combined both deductive and inductive coding. We began with a deductive coding frame to assess whether Imago Obscura adhered to our five design requirements (DR1–5) and addressed the SPAF barriers (awareness, motivation, ability) [9]. In parallel, we remained open to emergent themes that reflected participants' unanticipated concerns, reactions, or values. These inductive insights revealed additional opportunities and limitations that were not captured by the original design requirements.

To complement our qualitative analysis, we also analyzed the post- versus pre-task questionnaires participants filled out for each image. First, outside of the SUS scale items, all reverse-coded items were re-coded before analysis to ensure a consistent interpretation of scale direction (i.e., with a 1 indicating a negative impression, and a 5 indicating a positive impression). We then calculated descriptive statistics (means and standard deviations) for metrics related to each design requirement and SPAF barrier.

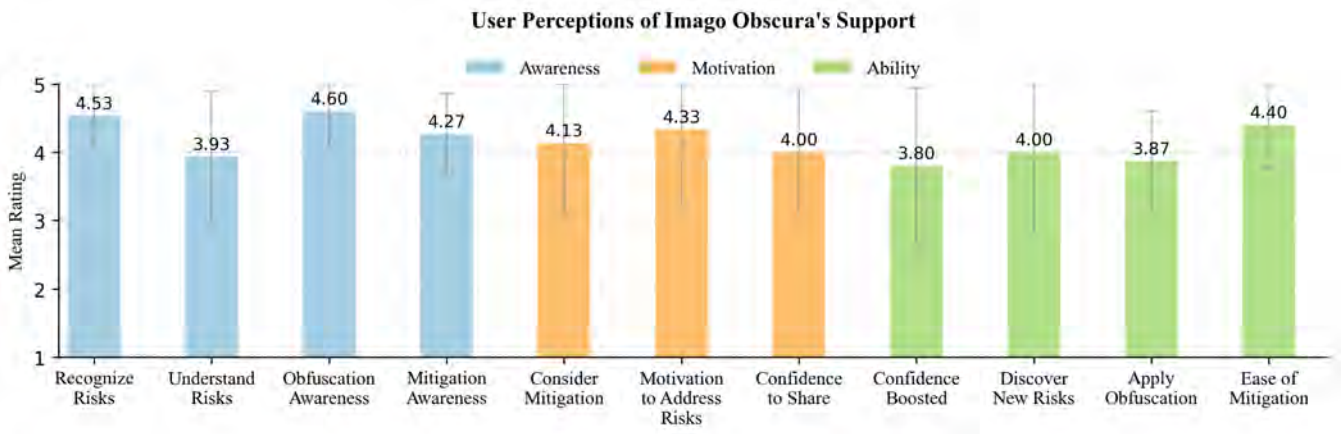
We next fit two random-intercept ordinal logistic regressions to model how use of Imago Obscura affected participants' belief that an image: (i) captured what they wanted to express by sharing it ("expression capture"), and (ii) contained concerning privacy risks ("perceived privacy risk"). The primary predictor variable in these models was whether the questionnaire was filled out before or after using Imago Obscura to modify an image (pre-task vs. post-task, with pre-task being the reference level). Each model was run for three groups of images: all images (4 per participant), previously shared images (2 per participant), and previously withheld images (2 per participant). We accounted for repeated measures with a random-intercept term for participant ID. We verified the proportional odds assumption using graphical diagnostics and found no violations. As noted in Figure 12, the reported p-values are derived from these regression models. For statistically significant coefficients ($p < .05$), we also report odds ratios (OR), calculated as e^{β}, to aid interpretation (Table 3).

6.3 Findings

Participants expressed a variety of sharing intents for the images they brought in — from documenting personal experiences, celebrating achievements, highlighting casual moments, and showcasing scenic or humorous content with friends and followers. The privacy concerns they expressed about their *withheld* images — i.e., images they *wanted* to share but did not for privacy reasons — included: violating the privacy of others pictured without consent; unintentionally disclosing sensitive personal spaces or geographic locations; and, sharing content that could be misunderstood or pose reputational risks.

6.3.1 *How does Imago Obscura impact users' awareness of, motivation to, and ability to address image privacy risks?*

Awareness Barrier: Participants strongly indicated that Imago Obscura enhanced their awareness of privacy risks in images (M=4.53, SD=0.64) and improved their understanding of potential privacy risks (M=3.93, SD=0.96). P13 noted during the interview, "It didn't occur to me that somebody might be able to identify the building, and that could be a privacy risk in a certain type of photo". P12 mentioned "I think definitely I would use something like [Imago Obscura], in the future, if I'm taking a photo, in my house where you could see more of a layout, especially being a small woman, I think more about personal safety and stuff". Participants also reported becoming more aware of different obfuscation techniques available to address image privacy risks (M=4.60, SD=0.50). The tool successfully enhanced participants' awareness of how to address privacy risks in images (M=4.26, SD=0.59). P8 explained, "it gave different suggestions for how you can replace it, like replacing



Note: Based on final questionnaire likert ratings (1 = strongly disagree to 5 = strongly agree); reverse-coded items were re-coded for consistency.

Figure 10: Participants rated Imago Obscura highly across all three SPAF barriers—awareness, motivation, and ability—suggesting the tool effectively supports users in adopting pro-image privacy behaviors.

the people with statues that look similar. It's not something that I would have thought of".

Motivation Barrier: Our results suggest that Imago Obscura positively influenced participants' motivation to address image privacy risks. Participants reported that using the tool made them more likely to consider mitigating privacy risks in their images (M=4.13, SD=1.06). Participants also felt motivated to take steps to address privacy risks in the images they share online (M=4.33, SD=1.11). Furthermore, participants expressed increased confidence in their ability to share images while mitigating key privacy risks (M=4.00, SD=0.92). As P3 remarked: "I started looking to make sure what could be a risk. And I think I didn't do that before, when I was even posting. I'm usually a very careful person, but it definitely helped me become, like, a little bit more aware of that."

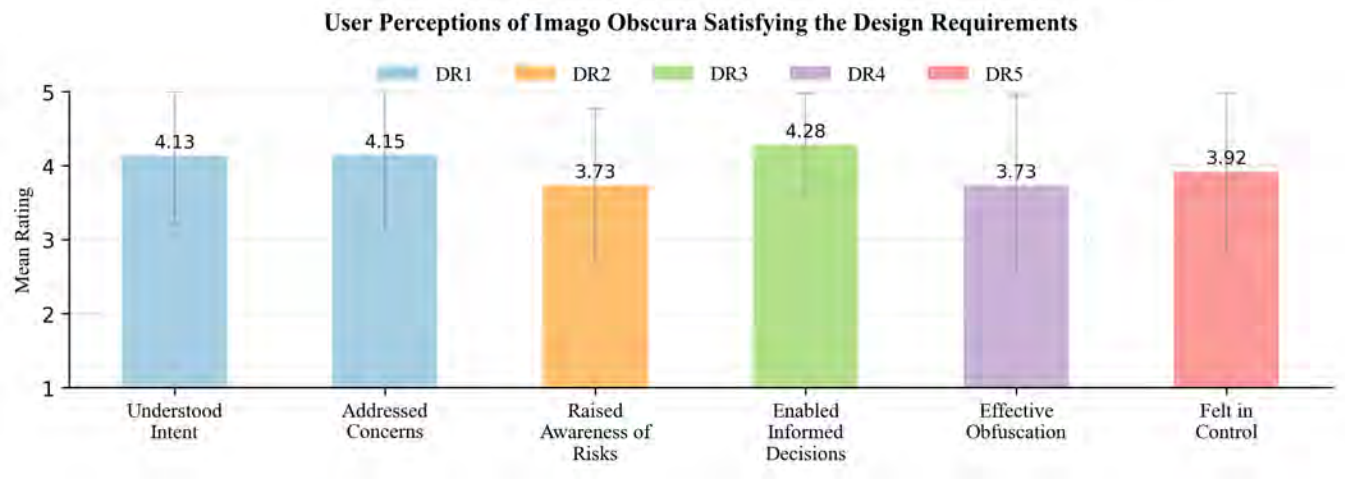
Ability Barrier: Imago Obscura successfully supported participants in overcoming the ability barrier by making it easier for them to address privacy risks in their images (M=4.40, SD=0.63). Participants felt that the tool helped them identify risks in their images that they hadn't considered before (M=4.00, SD=1.19), although there was more variation in responses to this question compared to others. P8 explained, "It did highlight almost all the privacy concerns I had, probably even uncovering some concerns which I did not think of, including geo tagging and so on." Participants also reported feeling confident about sharing their images online after using the tool (M=3.80, SD=1.14) and also felt supported in effectively applying techniques to mitigate pertinent privacy risks (M=3.8, SD=0.74). P12 stated: "I think it made me feel more comfortable posting the photos that I didn't post, right? Being able to, like, pick and choose what I wanted out and cover what necessarily shouldn't be online."

6.3.2 *How well does Imago Obscura adhere to design requirements?*

DR1: Understands and Accounts for User-Articulated Privacy Concerns. Participants felt that Imago Obscura understood their privacy concerns and sharing intent (M=4.13, SD=0.92) and effectively addressed their concerns (M=4.15, SD=1.05). For example, P1 stated: "[...] I had stated my privacy concerns, and the tool was able to pick up on that and also identify objects or people in the image that I had not considered obscuring before [...]" More generally, users appreciated how the system recognized a broad spectrum of concerns, from identifying individuals in backgrounds to detecting revealing location information.

DR2: Expands Awareness of Content-level Privacy Risks. Many participants reported becoming more aware of privacy risks they hadn't previously considered (M=3.73, SD=1.04): "I never really thought about specifics, like flags of places that I was visiting [...] People can look up the town and they could figure out everything." (P10) The tool particularly heightened awareness around location leakage, the presence of bystanders, and background elements that could compromise privacy. "...[I] actually have people in the background... I know one of them is specifically very obsessed with privacy issues. I have shared this photo before without blurring, but I think if I want to share it again now I will blur." (P15) Accordingly, participants reported Imago Obscura having an educational effect: "I think it's giving a pretty good taxonomy of risks with their rating on each picture... by running these four pictures, I kind of learned different ways of framings of some potential privacy risk and how they are ranked in the pictures" (P14). Similarly, P3 was surprised when the tool identified a logo that could reveal their location, noting it was "something that I completely brushed over." Beyond physical locations, participants recognized subtle contextual identifiers in images: from visible phone screens (P7) to religious buildings (P8) and even flags (P9), which could reveal personal information. By highlighting overlooked risks, the tool broadened users' privacy awareness beyond their initial concerns.

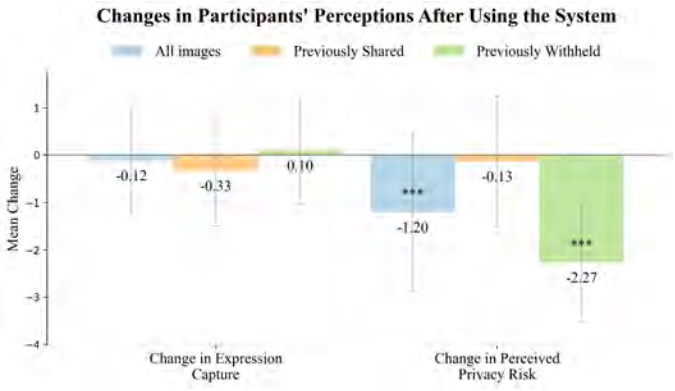
DR3: Empowers Informed Decision-Making. Participants reported feeling empowered to make informed decisions about addressing privacy risks in their images (M=4.28, SD=0.69). In particular, they appreciated that the tool explained both the identified risks and the corresponding obfuscation strategies, enabling them to weigh trade-offs. When asked what helped them decide which



Note: Based on Post-tool use questionnaire likert ratings (1 = strongly disagree to 5 = strongly agree); reverse-coded items were re-coded for consistency.

Figure 11: Participants reported that Imago Obscura satisfied all five design requirements (DR1–DR5).

mitigation strategy to employ, a participant highlighted how "when I hover the mouse... it showed you, like, a box" and offered an "explanation that what it does", referring to how the tool visually highlighted areas of concern and clearly described the suggested mitigation strategy along with its relevant attributes. Participants also appreciated the variety of obfuscation options provided, and that they could easily apply and remove the obfuscations to chose which they technique they preferred more.



Significance: $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Note: p-values come from a random-intercepts ordinal regression.

Figure 12: Using Imago Obscura significantly reduces the perceived privacy risk in an image.

Our quantitative results align with the finding that Imago Obscura enabled more informed choices (Fig 12, Appendix Table 3). Across all images, participants' perceived privacy risks decreased significantly after using the tool (M=−1.2, SD=1.69; $\beta = -1.66, e^{\beta} = 0.19, p < .001$ - *statistically significant*). This effect was most pronounced for *withheld* images where risk perceptions dropped by more than two points on average (M=-2.26, SD=1.26 $\beta = -4.43$, $e^{\beta} = 0.012$, $p < 0.001$), with an odds ratio of 0.012, meaning participants were over 80 times less likely to report higher risk—an extremely strong effect. In contrast, for shared images, the average reduction in perceived privacy risk was not significant (M=-0.13, SD=1.382 $\beta = -0.56$, $p = 0.285$)— largely because participants did not harbor strong privacy concerns for these images in the first place.

Moreover, we found no significant change in participants' belief that using Imago Obscura to address privacy risks changed how well that image captured their sharing intent (M=−0.12, SD=1.15; $\beta = -0.24$, $p = 0.518$). While absence of evidence cannot be considered evidence of absence, our findings at the very least suggest that *if* the modifications introduced by Imago Obscura are negatively impacting sharing intent, the effect is quite small.

In sum, we can surmise that Imago Obscura helped participants greatly reduce perceived privacy risk without compromising sharing intent — especially for images that participants *wanted* to share but withheld for privacy reasons.

We identified four different scenarios that showcased how Imago Obscura impacted users' decision-making on whether and how to mitigate image privacy risks:

(1) Users with no awareness of privacy risks in an image became aware of potential risks and took steps to mitigate them.
(2) Users aware but unconcerned about certain risks, upon receiving more information and mitigation options, made more confident decisions after weighing sharing intent against concerns.
(3) Users aware of specific privacy risks in an image found that the tool effectively identified and helped mitigate them.
(4) Users uncertain about the validity of their concerns gained clarity when the tool highlighted risk severity (i.e., low/medium/high), often increasing confidence in their decision to share or not share.

DR4: Facilitates application of obfuscation techniques. Participants rated the tool positively for its effectiveness at applying image obfuscation techniques (M=3.73, SD=1.21). The seamless integration of risk identification and mitigation capabilities received positive feedback, highlighting the tool's success in meeting DR4 by making privacy protection techniques accessible and effective.

"I think it was very intuitive, like you type and then you press generate, and then it gives you all the options, right? I think it's very clear" (P5). Participants found that the tool made it easier to implement privacy protections that would otherwise require specialized skills: "it's like, more user friendly than Photoshop because of the AI part of it, and specifically that it specified things that you can change or you shouldn't change." (P4) Participants also mentioned that Imago Obscura had a low learning curve, so that even users without technical backgrounds could apply sophisticated obfuscation techniques: "The learning curve was very low, so it's pretty easy to learn." (P1)

DR5: Ensures autonomy and granular control. Imago Obscura successfully provided users with a sense of autonomy and control during the risk mitigation process (M=3.91, SD=1.06). Participants reported feeling in control while applying risk mitigation techniques, appreciating the ability to selectively address specific risks according to their preferences.

The system's approach of providing options rather than making unilateral changes was particularly valued: "I felt in control, because I could, like, discard the changes or type like a more specific prompt if I wanted to, and still have, like, the autonomy to choose from, like the generated images [...] So I was able to go back in and select only like specific amount and generate based on that." (P3) This agency allowed users to carefully consider the privacy-publicity tradeoff for each image, weighing what elements were important to preserve sharing intent while mitigating potential privacy risks.

Many participants appreciated the object selection feature because it gave them granular control. 'P10 stated: "I was able to very clearly specify, like, I want to blur out this part of the image... being able to just click in... when I did click specific parts would usually get what I wanted, so being able to have more fine tuned control was a good feeling." This sentiment was echoed by others, with P3 explaining that precise selection helped them feel in control when iteratively refining obfuscations: "...it kind of altered my face. So I was able to go back in and select only like specific amount and generate based on that." Several participants (P6, P14, P15) requested even more advanced selection tools like those found in professional editing software, with P6 suggesting "if there was something like a lasso tool or something that could be more flexible as compared to just, you know, explicit object selection in the image... that will be pretty useful."

6.3.3 ***How usable is Imago Obscura?*** Overall, participants reported a positive usability experience, particularly in terms of ease of use and learnability. The usability evaluation revealed an average estimated SUS score of 70.1, indicating a good overall level of system usability. It is worth noting, however, that we accidentally omitted the system inconsistency question typically present in the SUS scale —– effectively reducing the maximum possible score to 90 from 100 (since we assume the most pessimistic case where all participants answered strongly agree to the question "I thought there was too much inconsistency in this system."). Thus the score we report above should not be compared against the standard SUS benchmarks. Instead, we focused more specifically on the individual items of the SUS. Participants demonstrated particularly high ratings for the system's ease of use (M = 4.2) and low perceived need for technical support (M = 4.47), suggesting a user-friendly interface. The system's learnability was also perceived positively, with users indicating they would quickly learn to use the system (M = 4.0). Conversely, the lowest-scoring areas included the perceived frequency of use (M = 3.47) and system integration (M = 3.87), which may warrant further investigation to enhance overall user engagement and system cohesiveness.

6.3.4 ***What other values, concerns, or reactions did participants express?***

Increases Social Context and Consent Considerations. Participants became more aware of location risks and bystander privacy after using Imago Obscura. P1, for example, noticed how the tool "was able to pick up that there were cashiers and workers in the photos... that I had not considered obscuring." This recognition extended to interpersonal dynamics, as P7 discovered privacy concerns in captured interactions like a subject's hands touching another subjects shoulder that they hadn't previously considered problematic. Another notable finding was that Imago Obscura caused participants to reflect on consent in image sharing. P12 stated: "I don't like posting other people when I don't have their consent to do it," while P2 appreciated how the tool recognized that "I need their permission to share it and their face." P14's experience exemplified this reflection on bystander consent, noting that even after previously sharing an image, they reconsidered an image as they remembered a subject from the photo was particular about privacy.

Image Privacy Can Come at the "Cost" of Authenticity and Self-Presentation. Users face a fundamental tension between protecting privacy and maintaining authentic self-expression. Many participants discussed how image obfuscations could undermine the authenticity and communicative intent of their images. This tension was explicitly described by P14 who noted: "I feel like it's really hard to balance... removing privacy concerns while maintaining the authenticity of the picture... that's really a trade-off from my perspective." Many participants rejected certain modifications that appeared artificial, with P12 explaining: "I thought the avatar one... felt a lot more cheesy to me... if I'm editing a photo to get rid of something for a privacy concern, I don't necessarily want people to know that I had a privacy concern." This tension influenced sharing decisions, as P2 explained: "I will not still use it... because I feel that the image itself will not have that spirit or will not share what that moment was." Some participants expressed a willingness to make minor privacy-enhancing edits but drew the line at modifications that fundamentally altered the image's meaning or appearance. P3 articulated this threshold clearly: "If there are smaller things that really need to be blurred, I would use it for very small things, but for bigger things... I wouldn't use it... because it completely alters the entire setting that I'm in. And if it does that, then there's no point of posting the picture in the first place."

Improving the Quality of Generative Replacements Can Reduce the Authenticity "Cost" of Privacy. While privacy obfuscations sometimes came with a perceived authenticity "cost", participants found the quality of generative replacements helped them navigate this trade-off. Generative content replacement techniques were accepted only to the extent that the replaced content was realistic. P3 more broadly captured this sentiment: "I think it makes it seem fake, which I don't want it to appear faked or masked,

because then people know that, and like, they will probably be more curious to look into the image, right?" However, participants also indicated that high-quality generative replacements could mitigate this trade-off: "after replacing the roof, it does not make the picture look bad... if it's not making the picture look bad or look weird, I don't see any trade off." (P14) The quality of generative content replacement algorithms, thus, plays a critical role in determining whether users would adopt privacy-enhancing modifications. As these algorithms improve over time, we might expect more broad acceptance of privacy-preserving obfuscations.

7 DISCUSSION

To summarize, we engaged in a user-centered design process to build an image privacy AI-copilot — Imago Obscura — that enables users to identify and mitigate privacy risks in images that they seek to share online.

Our findings suggest that Imago Obscura appeared to strike an appropriate balance between raising awareness of privacy risks and affording users agency to address those risks in a manner that did not compromise their sharing intentions. Users felt more confident in their decision-making, whether or not they chose to mitigate risks. We nevertheless discovered a number of considerations, limitations, and opportunities for future work that we discuss in more detail here.

7.1 Human-AI collaborative systems can overcome the limits of purely automated systems

Reducing the role of the "human-in-the-loop" through automation has long been a top-level objective of the security community [8]. The usable security community, in contrast, has pushed back against this narrative by outlining how automation has its limits [12] — for example, all automated systems have failure conditions, and purely automated security systems can make handling failure cases even more difficult for users. Generative AI technologies — for all the risks they bring to privacy and security [37] — provide an interesting new opportunity to create human-AI collaborative systems that overcome the limitations of purely automated approaches in helping users make security and privacy decisions without overwhelming them with choice.

For example, unlike traditional image-generation or obfuscation tools, our approach with Imago Obscura focuses on enabling users to navigate the nuanced decisions involved in managing the privacy/publicity boundary of content in an image being shared online, rather than merely automating the obfuscation process.

Indeed, our findings indicate that it is critical to raise users' awareness of image privacy risks, expose available mitigation strategies, and make clear the implications of each choice. While prior research has emphasized automating privacy protection, our work demonstrates that keeping users in the loop allows for a deeper and more meaningful engagement with their privacy decisions. The copilot design strikes a balance between automation and manual control, enhancing user agency through informed decision-making and simplifying action.

7.2 Generative AI privacy-copilots should be scaffolded with theory-informed prompting

A key risk of incorporating generative AI into user-facing products is that large language models can hallucinate and be inaccurate [26, 47]. While techniques — like retrieval augmented generation [27, 38] — have been proposed to reduce the likelihood of hallucinations in some task contexts, prior work has argued that "hallucination" may be an inevitable outcome of the stochasticity of large language models [69]. This context begs the question: if generative AI systems hallucinate, can they be "trusted" to help people make privacy and security decisions?

The approach we took in the development of Imago Obscura was to use a scaffolded prompting process that grounded outputs in existing literature and theory on image privacy, usable security, and privacy. For example, we did not simply prompt GPT-4o to identify privacy risks in an input image — we first identified and segmented the image, we provided user-specified concerns, and we provided a taxonomy of image privacy risks distilled from prior literature. This process constrained the output space of our model pipeline to reduce the likelihood of catastrophic or ungrounded hallucination, thereby increasing accuracy and robustness.

Indeed, our evaluation revealed that users valued the system's ability to highlight privacy risks specific to their concerns and sharing intent while also informing them of risks that they had not considered but could see the value in identifying. This approach, theory-informed prompting, can also help with quickly translating empirical findings into actionable design implications — one can imagine, for example, improving the outputs of Imago Obscura by having it distill new empirical findings from the usable privacy literature on image privacy risks as these findings emerge.

7.3 Opportunities for improvement

In the process of creating and evaluating Imago Obscura, we identified opportunities for improvement that merit further consideration.

7.3.1 Fostering agency, preventing over-reliance. Our study revealed broader conceptual implications that warrant careful consideration. Participants exhibited contrasting reactions after using our tool, highlighting potentially unintended consequences. While some indicated they would likely share fewer images due to heightened risk awareness, others expressed increased confidence in sharing, believing all risks were adequately addressed. These opposing reactions point to a possible over-reliance on the tool: users who are more sensitive to privacy risks may use it to confirm their fears and self-censor; users who are less sensitive to privacy risks may feel like the tool definitively covers all their bases. While improving user confidence is generally positive, it's crucial to ensure users understand that residual risks may still exist. This is especially important given the current 70% detection accuracy of our pipeline, which could lead to false confidence, particularly among users with low privacy awareness. This observation underscores the delicate balance required between informing users of potential risks and inadvertently overemphasizing them.

To better align user expectations with system capabilities, future iterations should explore trust calibration strategies such as

displaying uncertainty cues alongside flagged risks, integrating onboarding that sets expectations, and including disclaimers to guard against over-trust. These strategies could help preserve user agency while empowering users with knowledge and encouraging critical thinking about image sharing and privacy risks.

7.3.2 Guardrails to prevent malicious use. While it is widely acknowledged that generative AI technology can be used maliciously [17] — including in ways that exploit user privacy [29, 37] — our work seeks to shift the balance of power back to the end-users by leveraging the capabilities of these technologies to enable users to preserve their privacy more effectively. We also recognize that the generative technology in Imago Obscura could be misused, for example, to deceive viewers or spread misinformation. For instance, one participant wanted to replace their younger sibling with another person, while another chose to replace fast food on their plate with healthier alternatives. These examples may seem innocuous, but they raise concerns around consent and content integrity — especially when manipulated content could mislead others or be taken out of context, leading to unintended consequences.

To help mitigate such risks, future versions of the tool could incorporate provenance markers (e.g., embedded metadata, lightweight annotations, cryptographic watermarking) to signal when and how an image was edited. These markers can improve transparency and integrity, especially in public sharing contexts.

Our user study revealed that many participants were mindful of these ethical concerns and found creative ways to balance privacy protection with ethical responsibility. For example, the participant who initially wanted to replace their sibling with another person later opted to replace them with a pet. They felt this substitution was more plausible and less deceptive, making it a good compromise. This finding further motivates approaches proposed in prior work, which explored replacing sensitive content with similar, non-sensitive alternatives to preserve context [30, 66].

In sum, there remains a need for future research to explore and evaluate both nudges and guardrails to ensure consensual uses of reference images, and to appropriately watermark generated content so that it is clear when an image has been altered.

7.4 Limitations

It is important to note that our evaluation, while designed to be representative of real-world scenarios by using users' own images, was conducted in a simulated laboratory setting with a limited number of participants who are not representative of all people who share images online. As such, the results should be interpreted with appropriate caution. Despite this limitation, we believe the insights gained remain valid and offer interesting perspectives on user interaction with privacy-enhancing tools.

Another limitation lies in our deviation from a fully iterative human-centered design cycle with a consistent user population. Rather than refining the system through repeated cycles with end-users, we began with expert image editors to inform design and then evaluated the system with lay users. While this progression reflects our intent to transfer expert strategies to everyday users, positioning the system as a privacy copilot akin to an expert assistant, it may limit insights into how user needs evolve across iterative refinements.

To support this expert-driven phase, we used the DIPA dataset to provide a consistent and diverse annotated images, enabling us to surface generalizable strategies across experts. While this limited early exploration of user-specific concerns, it was addressed in the summative study, where end-users engaged with their own images.

We note that for the purposes of this work, our threat model excluded the institutional privacy risks that emerge from using third-party AI model providers. While our participants did not directly express privacy concerns regarding the use of third-party models, we recognize that using these models introduces risks such as metadata leakage and cloud-based exposure, and may exclude users who do not trust external service providers. There are a number of locally deployable vision-language models that can be integrated to help address these concerns, but we leave it to future work to explore trade-offs between use of local models and the quality / acceptability of the obfuscations they generate.

Finally, we used ordinal logistic regression models for hypothesis testing, treating Likert responses as ordinal outcomes and modeling random intercepts to account for repeated measures. As ordinal rating data has known interpretive limitations, these findings should accordingly be considered with appropriate caution.

8 CONCLUSION

We present Imago Obscura, an AI-powered image privacy copilot that helps users make more informed decisions when navigating the privacy/publicity boundary [50] in online image sharing. We distilled five concrete design requirements for our tool following a formative study with seven image manipulation experts. Based on these requirements, Imago Obscura enables users to articulate their image-sharing intent and privacy concerns, surfaces contextually relevant privacy risks, and recommends appropriate obfuscation techniques to address privacy concerns while minimally compromising sharing intent. Our implementation integrates a pipeline of generative AI models into an image editing tool, scaffolded by a theory-grounded prompting approach that leverages prior literature on image privacy and usable security. Through a summative evaluation with 15 participants who tested Imago Obscura on their own photos, we found that the system enhanced participants' awareness of, motivation to address, and their ability to mitigate relevant privacy risks in images they wanted to share online. As a result, it improved participants' confidence that they understood and could address pertinent privacy risks in their images, and thus their ability to make informed decisions about whether or not to share an image with or without obfuscation. More generally, our findings demonstrate how the interactive capabilities of modern generative AI technologies can help strike an effective balance between the benefits of automation and manual control for technologies that aim to simplify end-user privacy decision-making.

ACKNOWLEDGMENTS

This work was generously supported, in part, by NSF SaTC Grant #231629. We thank the anonymous reviewers for their valuable feedback. We are also grateful to Isadora Krsek for her suggestions on improving the figures, Hank Lee for his insights on the analysis, and Yuxuan Li, Pradyumna Shome, Minjung Park, and Pranav Khadpe for their thoughtful comments on earlier drafts.

REFERENCES

[1] Adobe Inc. 2025. Adobe Photoshop. https://www.adobe.com/products/photoshop.html.

[2] Shane Ahern, Dean Eckles, Nathaniel S Good, Simon King, Mor Naaman, and Rahul Nair. 2007. Over-exposed? Privacy patterns and considerations in online and mobile photo sharing. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 357–366.

[3] Open AI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[4] Tuomas Aura, Thomas A Kuhn, and Michael Roe. 2006. Scanning electronic documents for personally identifiable information. In *Proceedings of the 5th ACM workshop on Privacy in electronic society*. 41–50.

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

[6] Matic Broz. 2024. How many pictures are there (2024): Statistics, trends, and forecasts. https://photutorial.com/photos-statistics/ Accessed: 2024-09-02.

[7] Daniel Castro, Steven Hickson, Vinay Bettadapura, Edison Thomaz, Gregory Abowd, Henrik Christensen, and Irfan Essa. 2015. Predicting daily activities from egocentric images using deep learning. In *proceedings of the 2015 ACM International symposium on Wearable Computers*. 75–82.

[8] Lorrie F Cranor. 2008. A framework for reasoning about the human in the loop. *Conference on Usability, Psychology, and Security* 1 (2008).

[9] Sauvik Das, Cori Faklaris, Jason I Hong, Laura A Dabbish, et al. 2022. The security & privacy acceptance framework (spaf). *Foundations and Trends® in Privacy and Security* 5, 1-2 (2022), 1–143.

[10] Jelle Demanet, Kristof Dhont, Lies Notebaert, Sven Pattyn, and André Vandierendonck. 2007. Pixelating familiar people in the media: Should masking be taken at face value? *Psychologica belgica* 47, 4 (2007), 261–276.

[11] Early Moon, LLC. 2025. DAMA - Auto Redact Privacy. https://apps.apple.com/us/app/dama-auto-redact-privacy/id1534690075.

[12] W Keith Edwards, Erika Shehan Poole, and Jennifer Stoll. 2008. Security automation considered harmful?. In *Proceedings of the 2007 Workshop on New Security Paradigms*. 33–42.

[13] Michael Fire, Roy Goldschmidt, and Yuval Elovici. 2014. Online social networks: threats and solutions. *IEEE Communications Surveys & Tutorials* 16, 4 (2014), 2019–2036.

[14] Ricard L Fogues, Jose M Such, Agustin Espinosa, and Ana Garcia-Fornes. 2017. Exploring the viability of tie strength and tags in access controls for photo sharing. In *Proceedings of the Symposium on Applied Computing*. 1082–1085.

[15] Andrea Frome, German Cheung, Ahmad Abdulkader, Marco Zennaro, Bo Wu, Alessandro Bissacco, Hartwig Adam, Hartmut Neven, and Luc Vincent. 2009. Large-scale privacy protection in google street view. In *2009 IEEE 12th international conference on computer vision*. IEEE, 2373–2380.

[16] Guardian Project. 2025. ObscuraCam: The Privacy Camera. https://guardianproject.info/apps/obscuracam/.

[17] Maanak Gupta, CharanKumar Akiri, Kshitiz Aryal, Eli Parker, and Lopamudra Praharaj. 2023. From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy. *IEEE Access* (2023).

[18] Rakibul Hasan, David Crandall, Mario Fritz, and Apu Kapadia. 2020. Automatically detecting bystanders in photos to reduce privacy risks. In *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 318–335.

[19] Rakibul Hasan, Eman Hassan, Yifang Li, Kelly Caine, David J Crandall, Roberto Hoyle, and Apu Kapadia. 2018. Viewer experience of obscuring scene elements in photos to enhance privacy. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[20] Rakibul Hasan, Yifang Li, Eman Hassan, Kelly Caine, David J Crandall, Roberto Hoyle, and Apu Kapadia. 2019. Can privacy be satisfying? On improving viewer satisfaction for privacy-enhanced photos using aesthetic transforms. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[21] Rakibul Hasan, Patrick Shaffer, David Crandall, Eman T Apu Kapadia, et al. 2017. Cartooning for enhanced privacy in lifelogging and streaming videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 29–38.

[22] Benjamin Henne and Matthew Smith. 2013. Awareness about photos on the web and how privacy-privacy-tradeoffs could help. In *Financial Cryptography and Data Security: FC 2013 Workshops, USEC and WAHC 2013, Okinawa, Japan, April 1, 2013, Revised Selected Papers 17*. Springer, 131–148.

[23] Roberto Hoyle, Robert Templeman, Denise Anthony, David Crandall, and Apu Kapadia. 2015. Sensitive lifelogs: A privacy analysis of photos from wearable cameras. In *Proceedings of the 33rd Annual ACM conference on human factors in computing systems*. 1645–1648.

[24] Panagiotis Ilia, Iasonas Polakis, Elias Athanasopoulos, Federico Maggi, and Sotiris Ioannidis. 2015. Face/off: Preventing privacy leakage from photos in social networks. In *Proceedings of the 22nd ACM SIGSAC Conference on computer and communications security*. 781–792.

[25] Yasha Iravantchi, Thomas Krolikowski, William Wang, Kang G Shin, and Alanson Sample. 2024. PrivacyLens: On-Device PII Removal from RGB Images using Thermally-Enhanced Sensing. *Proceedings on Privacy Enhancing Technologies 2024)* 1 (2024), 20.

[26] Adam Tauman Kalai and Santosh S Vempala. 2024. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*. 160–171.

[27] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906* (2020).

[28] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020. Training generative adversarial networks with limited data. *Advances in neural information processing systems* 33 (2020), 12104–12114.

[29] Patrick Gage Kelley, Celestina Cornejo, Lisa Hayes, Ellie Shuo Jin, Aaron Sedley, Kurt Thomas, Yongwei Yang, and Allison Woodruff. 2023. " There will be less privacy, of course": How and why people in 10 countries expect {AI} will affect privacy in the future. In *Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023)*. 579–603.

[30] Mohamed Khamis, Habiba Farzand, Marija Mumm, and Karola Marky. 2022. DeepFakes for privacy: Investigating the effectiveness of state-of-the-art privacy-enhancing face obfuscation methods. In *Proceedings of the 2022 International Conference on Advanced Visual Interfaces*. 1–5.

[31] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4015–4026.

[32] Mohammed Korayem, Robert Templeman, Dennis Chen, David Crandall, and Apu Kapadia. 2016. Enhancing lifelogging privacy by detecting screens. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 4309–4314.

[33] Pavel Korshunov, Andrea Melle, Jean-Luc Dugelay, and Touradj Ebrahimi. 2013. Framework for objective evaluation of privacy filters. In *Applications of Digital Image Processing XXXVI*, Vol. 8856. SPIE, 265–276.

[34] Open-source Krita Foundation. 2025. Krita: Free and open-source digital painting application. https://krita.org/en/.

[35] Damera Vijay Kumar, P Satya Shekar Varma, and Shyam Sunder Pabboju. 2013. Security issues in social networking. *International Journal of Computer Science and Network Security (IJCSNS)* 13, 6 (2013), 120.

[36] Karen Lander, Vicki Bruce, and Harry Hill. 2001. Evaluating the effectiveness of pixelation and blurring on masking the identity of familiar faces. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 15, 1 (2001), 101–116.

[37] Hao-Ping Lee, Yu-Ju Yang, Thomas Serban Von Davier, Jodi Forlizzi, and Sauvik Das. 2024. Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–19.

[38] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[39] Fenghua Li, Zhe Sun, Ang Li, Ben Niu, Hui Li, and Guohong Cao. 2019. Hideme: Privacy-preserving photo sharing on social networks. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 154–162.

[40] Wenjie Li, Rongrong Ni, and Yao Zhao. 2017. JPEG photo privacy-preserving algorithm based on sparse representation and data hiding. In *Image and Graphics: 9th International Conference, ICIG 2017, Shanghai, China, September 13-15, 2017, Revised Selected Papers, Part III 9*. Springer, 575–586.

[41] Yifang Li and Kelly Caine. 2022. Obfuscation remedies harms arising from content flagging of photos. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–25.

[42] Yifang Li, Wyatt Troutman, Bart P Knijnenburg, and Kelly Caine. 2018. Human perceptions of sensitive content in photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 1590–1596.

[43] Yifang Li, Nishant Vishwamitra, Hongxin Hu, and Kelly Caine. 2020. Towards a taxonomy of content sensitivity and sharing preferences for photos. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[44] Yifang Li, Nishant Vishwamitra, Bart P Knijnenburg, Hongxin Hu, and Kelly Caine. 2017. Effectiveness and users' experience of obfuscation as a privacy-enhancing technology for sharing photos. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–24.

[45] Chi Liu, Tianqing Zhu, Jun Zhang, and Wanlei Zhou. 2022. Privacy intelligence: A survey on image privacy in online social networks. *Comput. Surveys* 55, 8 (2022), 1–35.

[46] Virginia Mantouvalou. 2019. 'I lost my job over a Facebook post: Was that fair?'Discipline and dismissal for social media activity. *International Journal of Comparative Labour Law and Industrial Relations* 35, 1 (2019).

[47] Gary Marcus. 2020. The next decade in AI: four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177* (2020).

[48] Phillip Nyoni and Mthulisi Velempini. 2018. Privacy and user awareness on Facebook. *South African Journal of Science* 114, 5-6 (2018), 1–5.

[49] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. 2017. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *Proceedings of the IEEE international conference on computer vision*. 3686–3695.

[50] Leysia Palen and Paul Dourish. 2003. Unpacking" privacy" for a networked world. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 129–136.

[51] Constantinos Patsakis, Athanasios Zigomitros, Achilleas Papageorgiou, and Agusti Solanas. 2015. Privacy and security for multimedia content shared on OSNs: issues and countermeasures. *Comput. J.* 58, 4 (2015), 518–535.

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[53] Yasmeen Rashidi, Tousif Ahmed, Felicia Patel, Emily Fath, Apu Kapadia, Christena Nippert-Eng, and Norman Makoto Su. 2018. " You don't want to be the next meme": College Students' Workarounds to Manage Privacy in the Era of Pervasive Photography. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*. 143–157.

[54] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[55] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. 2023. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *32nd USENIX Security Symposium (USENIX Security 23)*. 2187–2204.

[56] Shawn Shan, Wenxin Ding, Josephine Passananti, Haitao Zheng, and Ben Y Zhao. 2023. Prompt-specific poisoning attacks on text-to-image generative models. *arXiv preprint arXiv:2310.13828* (2023).

[57] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. 2020. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th USENIX security symposium (USENIX Security 20)*. 1589–1604.

[58] Manya Sleeper, Rebecca Balebako, Sauvik Das, Amber Lynn McConahy, Jason Wiese, and Lorrie Faith Cranor. 2013. The post that wasn't: exploring self-censorship on facebook. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 793–802.

[59] Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, Adrian Popescu, and Yiannis Kompatsiaris. 2016. Personalized privacy-aware image classification. In *Proceedings of the 2016 ACM on international conference on multimedia retrieval*. 71–78.

[60] Nancy Van House, Marc Davis, Morgan Ames, Megan Finn, and Vijay Viswanathan. 2005. The uses of personal networked digital imaging: an empirical study of cameraphone photos and sharing. In *CHI'05 extended abstracts on Human factors in computing systems*. 1853–1856.

[61] Nishant Vishwamitra, Hongxin Hu, Feng Luo, and Long Cheng. 2021. Towards understanding and detecting cyberbullying in real-world images. In *2020 19th IEEE international conference on machine learning and applications (ICMLA)*.

[62] Nishant Vishwamitra, Yifang Li, Hongxin Hu, Kelly Caine, Long Cheng, Ziming Zhao, and Gail-Joon Ahn. 2022. Towards automated content-based photo privacy control in user-centered social networks. In *Proceedings of the Twelfth ACM Conference on Data and Application Security and Privacy*. 65–76.

[63] Yang Wang, Gregory Norcie, Saranga Komanduri, Alessandro Acquisti, Pedro Giovanni Leon, and Lorrie Faith Cranor. 2011. " I regretted the minute I pressed share" a qualitative study of regrets on Facebook. In *Proceedings of the seventh symposium on usable privacy and security*. 1–16.

[64] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[65] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. 2023. Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks. arXiv:2311.06242 [cs.CV] https://arxiv.org/abs/2311.06242

[66] Anran Xu, Shitao Fang, Huan Yang, Simo Hosio, and Koji Yatani. 2024. Examining Human Perception of Generative Content Replacement in Image Privacy Protection. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–16.

[67] Anran Xu, Zhongyi Zhou, Kakeru Miyazaki, Ryo Yoshikawa, Simo Hosio, and Koji Yatani. 2023. DIPA: An Image Dataset with Cross-cultural Privacy Concern Annotations. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*. 259–266.

[68] Anran Xu, Zhongyi Zhou, Kakeru Miyazaki, Ryo Yoshikawa, Simo Hosio, and Koji Yatani. 2024. DIPA2: An Image Dataset with Cross-cultural Privacy Perception Annotations. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 4 (2024), 1–30.

[69] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817* (2024).

[70] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441* (2023).

[71] Jun Yu, Zhenzhong Kuang, Zhou Yu, Dan Lin, and Jianping Fan. 2017. Privacy setting recommendation for image sharing. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 726–730.

[72] Jun Yu, Zhenzhong Kuang, Baopeng Zhang, Wei Zhang, Dan Lin, and Jianping Fan. 2018. Leveraging content sensitiveness and user trustworthiness to recommend fine-grained privacy settings for social image sharing. *IEEE transactions on information forensics and security* 13, 5 (2018), 1317–1332.

[73] Jun Yu, Baopeng Zhang, Zhengzhong Kuang, Dan Lin, and Jianping Fan. 2016. iPrivacy: image privacy protection by identifying sensitive objects via deep multi-task learning. *IEEE Transactions on Information Forensics and Security* 12, 5 (2016), 1005–1016.

[74] Sergej Zerr, Stefan Siersdorfer, and Jonathon Hare. 2012. Picalert! a system for privacy-aware image classification and retrieval. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. 2710–2712.

[75] Sergej Zerr, Stefan Siersdorfer, Jonathon Hare, and Elena Demidova. 2012. Privacy-aware image classification and search. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. 35–44.

[76] Chenye Zhao, Jasmine Mangat, Sujay Koujalgi, Anna Squicciarini, and Cornelia Caragea. 2022. Privacyalert: A dataset for image privacy prediction. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 1352–1361.

A APPENDIX

A.1 Threat Model

Our adversary seeks to learn or infer sensitive personal information about the subject/owner of a photo, based on the content in the photo, that is peripheral to the sharing intent of the photo. To that end, we focus on *observable privacy* and *inferential privacy* as outlined by Liu et al. [45], covering both visible and inferable content-based risks. We exclude threats from automated adversaries without an analyst-in-the-loop and do not consider *contextual privacy* [45], which relates to external data like captions or metadata. Consequently, we do not implement defenses against algorithmic adversaries (e.g., adversarial perturbations [55–57]). Use case scenarios aligned with this model are provided in Appendix A.2.

A.2 Use case scenarios

To further contextualize how we envision end-users might use Imago Obscura, consider the following two common scenarios.

A.2.1 ***Scenario 1: Inadvertent Sharing of Sensitive Information.*** Users often share images online without careful consideration of pertinent risks, which can lead to the inadvertent disclosure of sensitive information [22, 35, 48, 63]. In turn, these accidental disclosures can lead to, for example, embarrassment, regret, harassment, and job loss [46, 53, 61, 63].

Example: Alice, a 25-year-old marketing executive, takes a selfie at her desk to share her excitement about a new project. She posts it on her public Instagram story without noticing that her computer screen in the background displays confidential client information. A competitor sees the post, leading to a breach of client confidentiality. Alice receives a formal warning and nearly loses her job.

A.2.2 ***Scenario 2: Privacy Concerns Inhibiting Image Sharing.*** Users may want to share images online for informational and/or emotional support, but hesitate due to privacy concerns. As a result, the user might either self-censor entirely and miss out on accessing support they seek, or they attempt to find a workaround that is difficult to implement, ineffective at addressing the privacy concern, or diminishes their sharing intent [41].

Example: Bob, a 40-year-old father, wants to share photos from a recent family vacation on Instagram so he can keep his extended family updated. However, he's concerned about his children's privacy and the potential for their images to be misused online. He considers several options: 1) not sharing the photos at all, missing out on connecting with friends and family, 2) spending hours manually editing each photo to blur his children's faces, which is time-consuming and diminishes the quality of the images. and 3) sharing only scenery photos without people, which fails to capture the family moments he wanted to share. Ultimately, Bob feels frustrated by the lack of an easy solution that balances his desire to share with his need for privacy.

A.3 Curated list of obfuscation techniques

Our tool enables the application of nine obfuscation techniques curated from existing literature [30, 44, 66].

Masking. Replaces sensitive content with a solid box for complete obfuscation.

Silhouette Masking. Replaces content with shapes, preserving context without revealing identity.

Blurring. Softens details while retaining visual context, offering moderate privacy.

Pixelation. Enlarges pixels to obscure details while keeping recognizable forms.

Bar Replacement. Covers sensitive content with a thin bar, highlighting presence but hiding specifics.

Point-Light Replacement. Uses dots to represent movement, preserving dynamics without revealing identities.

Removal. Eliminates sensitive content entirely, filling in the background seamlessly.

Avatar Replacement. Replaces individuals with avatars, maintaining social cues while protecting identity.

Generative Content Replacement. Replaces sensitive elements with realistic alternatives, ensuring coherence.

A.4 Formative Study Material

A.4.1 Task Material. To help users in this process, we explained potential privacy threats and provide a list of sensitive content that may exist in images, which they could consider as they edit.

(1) Threats
  (a) Interpersonal Threats
    - Threats within the social circle or specific people with whom they were connected on online social networks
    - Threats outside the social circle or strangers with whom they were not directly connected
  (b) Institutional Threats
    - Companies, including employees within companies
(2) Sensitive Content
  (a) Personal Identification
    - Faces and identities of individuals (including photo owner, family, friend, bystander)
    - Personal documents (ID cards, passports, licenses)
    - Contact information (addresses, phone numbers)
    - Vehicle plates and identifying markers
  (b) Nudity and Sexuality
    - Full or partial nudity
    - Sexual content or suggestive poses
    - Revealing or immodest clothing
  (c) Privacy and Personal Space
    - Home interiors and private areas (bedrooms, bathrooms)
    - Personal belongings and assets
    - Screens displaying private information
    - Unorganized or messy living spaces
  (d) Sensitive Personal Information
    - Medical conditions and treatments
    - Financial information (bank accounts, credit cards)
    - Legal documents and sensitive printed materials
    - Educational records
  (e) Behavioral and Social Content
    - Alcohol consumption and party scenes
    - Smoking and drug use

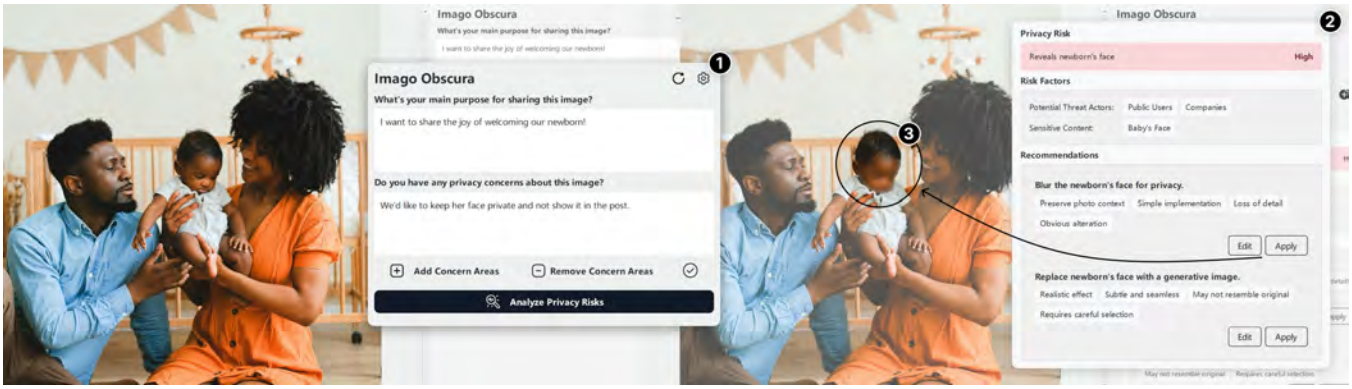


Figure 13: Imago Obscura enables users to express their sharing intent and their privacy concerns in natural language, subsequently identifying pertinent risks and recommending obfuscation techniques.

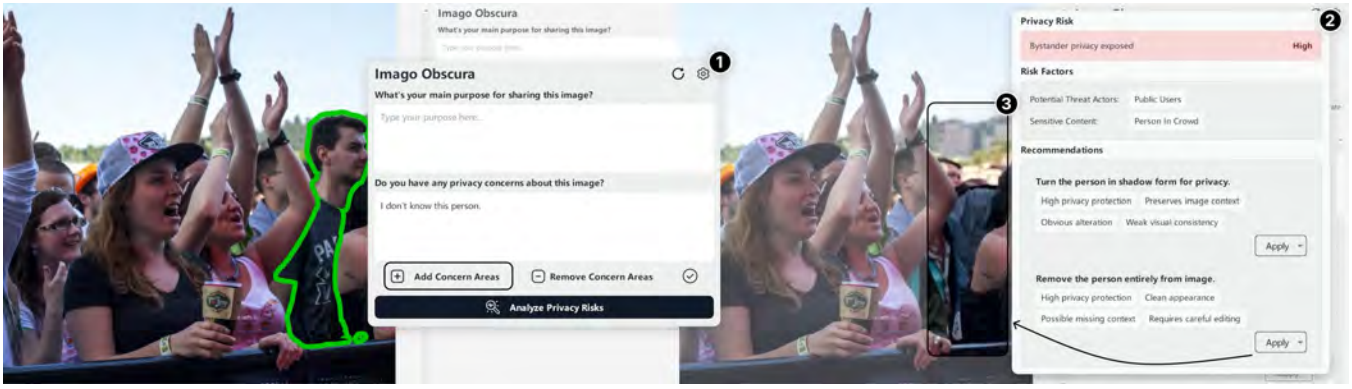


Figure 14: Imago Obscura enables the user to directly select areas of concerns which the tool will automatically precisely select and highlight in green, subsequently identifying pertinent risks and recommending obfuscation techniques.

Manipulation Technique	Effectiveness (Human Recognition)	Detectability	Visual Harmony	Narrative Coherence	Realism	Vulnerability
Masking/Colorfilling	High	Obvious	Weak	Low	Unnatural	Low
Silhouette Masking	High	Obvious	Weak	Medium	Unnatural	Medium
Blurring	Low	Obvious	Weak	High	Unnatural	High
Pixelating	Low	Obvious	Weak	Medium	Unnatural	High
Bar Replacement	High	Obvious	Weak	Medium	Unnatural	Low
Point Light Replacement	High	Obvious	Weak	Medium	Unnatural	Low
Cartoon Replacement	High	Obvious	Strong	High	Unnatural	Medium
Inpainting/Removal	High	Subtle	Strong	Low	Realistic	Low
Generative Content Replacement	High	Subtle	Strong	High	Realistic	Low

Table 1: Effectiveness Attributes of Image Obfuscation Techniques from Literature

- Inappropriate or illegal activities
- Unprofessional behavior at work

(f) Appearance and Self-Presentation
- Unflattering images or angles
- Embarrassing expressions or poses
- Grooming and sleep-related content

(g) Location and Environmental Identifiers
- Specific locations or landmarks
- Event attendance (revealing time and place)
- Workplace or school environments

(h) Relationships and Personal Moments
- Intimate or affectionate interactions
- Family gatherings or private events
- LGBTQ+ related content

(i) Political, Religious, and Controversial Content
- Political affiliations or activities
- Religious symbols or practices
- Controversial texts or memes

(j) Potential Safety Concerns
- Weapons (real or fake)
- Dangerous situations involving children or pets
- Accident scenes

(k) Digital Privacy
- Screenshots of private conversations
- Social media content without permission
- Unauthorized photos of others

(l) Miscellaneous Sensitive Content
- Food and dietary habits
- Pets and their behavior
- Personal interests and hobbies
- Low-quality or old photos

A.4.2 Post-task Interview Questions. Note: This component employed a semi-structured interview approach, with a pre-defined set of questions serving as a guide for the interviewer. The question bank covered various topics. However, not all questions were necessarily asked during each interview. The interviewer selected the most relevant questions based on the participant's responses and the available time, in order to maintain a focused and efficient interview process. The interview was rephrased and conducted in a conversational manner to ensure participants felt comfortable throughout the process.

(1) Participant Background and Experience
  (a) Could you briefly describe your background and experience with photo editing or graphic editing tools?
  (b) Have you ever used image editing specifically for obfuscation? If so, can you describe your experience?

(2) Task Workflow and Thought Process
  (a) Could you walk us through your thought process while completing the task?
  (b) Can you walk us through how you approached obfuscating one of the images?

(3) Techniques and Rationale Behind Choices
  (a) I noticed you used different techniques (e.g., blurring, pixelation, removal, generative content replacement). Can you explain why you chose this (refer to a technique) technique?
  (b) Could you describe the techniques you used in the images? Why did you select these techniques?
  (c) Can you explain why you chose one technique over another for a specific element of the image (e.g., why you used blurring instead of removal)?

(4) Obfuscation Tool Design and User Experience
  (a) What features or design elements would make the tool easier for you to use?
  (b) What changes or improvements would make the tool easier to use for non-experts?

(5) Experience with Task Materials and Suggestions
  (a) What role, if any, did the examples and sensitive content lists play in your decision-making during the task?
  (b) Did the list of sensitive content and threats influence your thinking during the task? If so, how?

A.5 Evaluation Study

A.5.1 Survey Questions.

(1) Pre Task
  (a) Image ID
  (b) This image captures what I'm trying to share or express online.
  (c) There are privacy risks in this image that would make me hesitate to share it online.
  (d) I feel comfortable sharing this image online.
  (e) Why and with whom would you like to share this image online?
  (f) Could you describe what privacy concerns you have with this image, if any?

(2) Post Task
  (a) Image ID
  (b) I feel comfortable sharing the original image online.
  (c) There are privacy risks in this modified image that make me hesitate to share it online.
  (d) This modified image captures what I'm trying to share or express online.
  (e) I feel uncomfortable sharing this modified image online for reasons other than privacy.
  (f) I feel that the tool understood my privacy concerns and sharing intent.
  (g) I feel that the tool failed at addressing my concerns.
  (h) I already knew about all of the privacy risks the tool showed me.
  (i) I was able to make an informed decision about if and how to address privacy risks in my image.
  (j) The tool failed to effectively apply image obfuscation techniques.
  (k) I felt in control while applying risk mitigation techniques.

(3) SUS (1 Question was accidentally omitted)
  (a) I think that I would like to use this system frequently.
  (b) I found the system unnecessarily complex.
  (c) I thought the system was easy to use.
  (d) I think that I would need the support of a technical person to be able to use this system.
  (e) I found the various functions in this system were well integrated.
  (f) I would imagine that most people would learn to use this system very quickly.
  (g) I found the system very cumbersome to use.
  (h) I felt very confident using the system.
  (i) I needed to learn a lot of things before I could get going with this system.

(4) Final Questions
  (a) I feel that the tool helped me recognize privacy risks in the images I share online.
  (b) The tool did not increase my understanding of potential privacy risks in my images.
  (c) I feel more aware of the different obfuscation techniques available to address image privacy risks.
  (d) The tool did not enhance my awareness of how to address privacy risks in images.

(e) Using this tool made me more likely to consider mitigating privacy risks in my images.
(f) I feel unmotivated to take steps to address privacy risks in images I share online.
(g) Using this tool, I feel confident that I can share images while mitigating key privacy risks.
(h) The tool made me feel more confident about sharing this image online.
(i) The tool did not help me identify risks in my images that I hadn't considered before.
(j) I feel that the tool supported me in effectively applying techniques to mitigate privacy risks.
(k) The tool did not make it easier for me to address privacy risks in my images.

A.5.2 Semi-structured Interview Questions.

(1) General Experience with the Tool
  (a) How would you describe your overall experience using Imago Obscura on the four images you brought to the study?
  (b) Were there any features that stood out to you? Why?
  (c) Were there any challenges you encountered while using the tool? If so, can you describe them?
  (d) Can you recall the last time you attempted to obfuscate an image you shared online? How would you compare using Imago Obscura to this previous experience?
(2) Design Requirements
  (a) Looking at the survey responses for the four images, it seems you felt that the tool did/did not understand your privacy concerns and sharing intent. Could you elaborate on why you felt this way? Were there specific moments or features that influenced your experience? [DR1]
  (b) You indicated that the tool helped/did not help you identify privacy risks you hadn't considered before. Could you share more about what led you to this conclusion? Were there specific risks that stood out or were overlooked? [DR2]
  (c) In the survey, you mentioned that the tool did/did not help you make informed decisions about addressing privacy risks. Could you explain why you feel this way? Were there aspects of the tool that supported or hindered your decision-making process? [DR3]
  (d) Your responses suggest that applying the image obfuscation techniques was easy/difficult. Can you describe your experience with this process? Were there specific parts that you found straightforward or challenging? [DR4]
  (e) You noted that you did/did not feel in control while using the tool. Could you explain what contributed to this feeling? Were there features or interactions that enhanced or diminished your sense of control? [DR5]
(3) Other
  (a) I see you feel more/less comfortable sharing this image, after using the tool. Can you explain what led to this change in comfort? (ask for 2 images, if possible of opposing results)
  (b) You indicated that the new image did/did not capture what you were trying to share or express online for image [Image ID]. Could you elaborate on how the modifications affected your ability to communicate your intent?
  (c) You mentioned that the tool made you feel more/less confident about sharing the image online. Can you explain why?

A.5.3 Demographic Survey.

(1) What is your age? [Number]
(2) What is your gender? [Options: Male, Female, Non-binary, Prefer not to say, Other (please specify)]
(3) How often do you post or send images to others? [Scale: 1 - Almost every day, 2 - A few times a week, 3 - Once a month, 4 - Rarely, 5 - Never]
(4) Have you used any image obfuscation techniques before? If yes, what forms or tools have you used? [Short answer]

A.6 Technical Evaluation

We focused our technical evaluation on the risk identification component because it informs the users subsequent actions. We qualitative assess the other components of our tool.

To assess the performance of Imago Obscura's risk identification component, we conducted an evaluation using the DIPA2 dataset [68]. The DIPA2 dataset was released in 2024, and provides object-level annotations of sensitive elements and their corresponding privacy risk category. The granularity and recency of this dataset makes it an ideal baseline for our evaluation.

A.6.1 Dataset and Methodology. We evaluated our model's performance on three attributes of the dataset which were relevant to Imago Obscura:

(1) **Object sensitivity**: Identifying whether an object in the image may be a privacy risk (binary classification)
(2) **Risk category** Assessing the category of risk (multi-class classification, 0-5 categories, as defined by DIPA2 —personal information, location of shooting, individual preferences/ pastimes, social circle, others' private/ confidential information or Other)
(3) **Severity** Determining the severity of the risk (High / Medium / Low). DIPA2 usees a 1–7 Likert scale for severity. However, for our tool, we adopted a more user-friendly representation by prompting the MLLM to predict High, Medium, or Low. Accordingly, we reduced DIPA2's baseline to a 1–3 scale to compare it with the output from our pipeline for analysis.

A.6.2 Results. Table 2 presents the performance of the model pipeline we use in Imago Obscura on these three tasks:

Task	Accuracy (%)	Precision (%)	Recall (%)
Object sensitivity(binary)	69.65	63.02	53.22
Risk category(multi-class)	82.93	16.48	57.05
Severity(High/Med/Low)	72.86	-	-

Table 2: Imago Obscura's Risk Identification Component Performance

Measure	Image Type	Mean Change (SD)	β	OR	SE	z	p	Sig.
Change in Expression Capture	All images	-0.116 (1.151)	-0.235		0.364	-0.646	0.518	
	Previously Shared	-0.333 (1.154)	-0.427		0.556	-0.769	0.442	
	Previously Withheld	0.1 (1.124)	0.181		0.499	0.364	0.716	
Change in Perceived Privacy Risk	All images	-1.200 (1.695)	-1.665	0.19	0.364	-4.566	<.001	***
	Previously Shared	-0.133 (1.382)	-0.563		0.526	-1.07	0.285	
	Previously Withheld	-2.266 (1.257)	-4.428	0.012	0.823	-5.379	<.001	***

Significance: $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Note: Significance and effect direction are derived from cumulative link mixed models (random-intercept ordinal logistic regression), accounting for repeated measures and participant-level variation.

Table 3: Changes in Participants' Perceptions Before and After Using the System

A.7 MLLM Prompts

A.7.1 Image Privacy Risk Identification Prompt. -

```
[Background]: You are an AI assistant with expertise in privacy
      and social media, tasked with protecting the user's privacy
      when sharing photos online, by identifying potential risks in
       a specific image and communicating them concisely and in non
      -technical language to the user.

[Goal]: Analyze the provided image and associated information to:

1. Understand the context of the image
   * Examine the photo [image]
   * Consider the user's purpose for sharing, if provided [text]
   * Address user's privacy concerns, if any [text, image with
       green annotations]
2. Identify potential sensitive content
   * Refer to the Sensitive Content list [text list]
   * Analyze all objects in the photo [text, annotated images,
       object list]
3. Determine privacy risks based on steps 1 & 2
   * Refer to common privacy risks in photo sharing [text list]
   * Identify user's concern specific privacy risks, if any [text]
4. For each risk, categorize its severity and specify potential
     threat actors

Your analysis will help you identify and communicate potential
     privacy risks to the user in a clear and actionable manner.

[MATERIALS] To achieve your goal, you have access to:

1. Primary Image [Original Image]
   * The image the user wants to share
2. User-Provided Context (optional)
   * Sharing intent in the user's words [User Input]
   * Privacy concerns expressed by the user
     * Textual description in users words [User Input]
     * Annotated image with concerns marked in green by the user [
         User Concern Region]
3. Image Analysis [Pre-Scan Data]
   * Visually annotated photo with red boxes marking all objects
   * JSON dictionary of object annotations, including position,
       length, and width of bounding boxes
4. Reference Materials
   * Curated list of Potential Sensitive Elements
   * Curated list of Potential Risks in sharing images online

Remember to prioritize user-provided privacy concerns when
     identifying risks and sensitive content.

[TASKS] Please follow these tasks to analyze the image and provide
      necessary privacy risk assessments:

1. Understand the Image Context:
   1. analyze the image and the users sharing intent
   2. Describe elements within green-bordered areas as user
       concerns (if present)
   3. analyze all user concern (if provided)
   4. Focus on specific elements, not general categories (e.g., "
       license plate" instead of "car")
   5. use concise phrases for each element
2. Identify Sensitive Elements
   1. Reference the curated list of potential sensitive elements
   2. Scan the entire image for sensitive elements
   3. Scan the annotated image for sensitive elements
   4. Scan the objects identified in the dictionary for potential
       sensitive elements
   5. Include user-highlighted concerns as sensitive elements
   6. Consider context-specific sensitive elements not in the
       curated list
   7. When conducting analysis, first examine each object
       individually and assess it for sensitivity, and then
       analyze the relationships between objects in the image to
       identify potential sensitive information inferred in the
       image.
   8. Combine similar elements to avoid duplicates. For example, "
       person 1", "person 2", and "person 3" can be combined as "
       person"
3. Determine Privacy Risks
   1. Identify potential privacy risks for each sensitive element
   2. Refer to the curated list of potential privacy risks to
       identify risks present in the image that the user might
       have forgotten to consider
   3. Combine the same risks which have different sensitive
       elements
   4. Use clear, non-technical phrases (max 5 words per risk)
       Example: "Reveals personal information" instead of "Self
       Disclosure"
4. Assess Each Privacy Risk
   1. Categorize severity: High, Medium, or Low. If the risk
       contains elements marked by the user, prioritize those
       risks as high severity.
   2. Specify potential threat actors (e.g., Public Users,
       Companies, Family/Friends)
   3. List associated sensitive elements using concise phrases
   4. Consider user intent and privacy concern: Ensure that the
       severity prediction accounts for the user's mentioned
       intent and specific privacy concerns.
5. Ensure Comprehensive Coverage
   1. All risks should be identified
   2. Every sensitive element should have at least one associated
       privacy risk
   3. All user concerns must be addressed in at least one privacy
       risk
6. Review and Refine
   1. Verify all tasks are completed thoroughly
   2. Ensure clarity and consistency in assessments

CURATED LIST OF POTENTIAL SENSITIVE ELEMENTS

1. Identity and Personal Information
   1. Person: Faces and identities of individuals (including photo
       owner, family members, children, friends, bystanders)
   2. Identity: Personal documents (e.g., ID cards, passports,
       licenses), contact information (e.g., home address, phone
       numbers)
   3. Place Identifier: Locations (e.g., home, workplace), scenery
       , or vacation spots that may be private
   4. Vehicle Plate: Vehicle license plates and identifying
       markers
2. Nudity and Sexual Content
   1. Full or partial nudity or semi-nudity
   2. Sexual content, suggestive poses, or erotic imagery
   3. Revealing, immodest, or inappropriate clothing (e.g.,
       swimsuits, underwear)
3. Other People and Social Contexts
   1. Person: Photos featuring others (e.g., family, friends,
       coworkers, bystanders)
   2. Group events and social gatherings (e.g., parties, weddings)
   3. Interactions with significant others or personal moments
       with others
4. Embarrassing or Unorganized Environments
   1. Table: Messy, unorganized, or cluttered home spaces (e.g.,
       kitchen, living room, bathroom)
```

```
   2. Unflattering grooming or sleeping shots
   3. Low-quality or outdated photos that do not reflect the
       current state
5. Violence and Criminal Activity
   1. Weapon: Scenes depicting violence or harm (e.g., battlefield
       , firearms)
   2. Criminal behavior or unlawful activities (e.g., drugs,
       vandalism, theft)
   3. Dangerous objects (e.g., weapons, guns)
6.  Medical and Health Conditions
   1. Visible injuries, medical conditions, or medical treatments
   2. Unflattering depictions of physical health (e.g., acne,
       wounds, bad teeth)
   3. Photos taken during medical procedures or showing medical
       equipment
7.  Alcohol, Drugs, and Partying
   1. Cigarettes: Images showing drinking, smoking, or substance
       use
   2. Social gatherings involving alcohol, drugs, or related
       paraphernalia
   3. Partying or celebratory events with potentially
       controversial behaviors
8.  Appearance, Grooming, and Physical Attributes
   1. Cosmetics: Unflattering body features or grooming (e.g.,
       messy hair, weight issues)
   2. Clothing: Tattoos, piercings, or unusual fashion choices
       that may be controversial
   3. Finger: Poses or expressions that reflect poorly on personal
        character
9. Religious and Cultural Sensitivity
   1. Religious symbols, clothing, or practices that might be
       sensitive
   2. Cultural references or behaviors that could be
       misinterpreted or offensive
   3. LGBTQ+ content that may be sensitive in certain contexts
10.  Sensitive and Private Information
    1. Screen: Screens displaying sensitive or personal
        information (e.g., emails, documents, monitor screens)
    2. Printed Materials: Handwritten or printed details revealing
        personal or professional data
    3. Unique or personal belongings that reveal too much about
        the owner
11.  Illegal, Unlawful, or Copyrighted Content
    1. Printed Materials: Images associated with illegal
        activities (e.g., drug use, piracy)
    2. Content that might suggest unlawful behavior (e.g.,
        trespassing, theft, vandalism)
    3. Book: Copyrighted materials or unauthorized content (e.g.,
        photos of artwork, copyrighted documents)
12.  Politically and Socially Offensive Content
    1. Printed Materials: Politically sensitive or controversial
        subjects (e.g., North Korean leader, racism memes)
    2. Vulgar gestures, symbols, or language (e.g., middle finger,
        offensive memes)
    3. Racism, hate speech, or other socially offensive materials
13.  Personal Assets and Belongings
    1. High-Value Assets: Cars, jewelry, antiques, art, and other
        valuable personal belongings
    2. Pet: Photos of personal pets or animals that the individual
        owns
    3. Electronic Devices: Personal electronics (e.g., laptops,
        phones)
    4. Musical Instrument: Musical instruments and other personal
        items that might be sensitive to the owner
14.  Factors Affecting Public Image and Reputation
    1. Photo: Unflattering or embarrassing shots that may harm
        public perception (e.g., unflattering facial expressions,
        bad hair days)
    2. Machine: Activities or settings that can be misinterpreted
        negatively (e.g., unorganized home, awkward social
        situations)
    3. Old, poor-quality, or technically flawed photos that do not
        reflect current image
15.  Food, Lifestyle, and Leisure
    1. Food: Unhealthy or unappealing food (e.g., junk food, fast
        food)
    2. Lifestyle: Overindulgence or gluttony in food or drink,
        smoking, cigars
    3. Toy: Personal items such as toys that might reflect a
        certain lifestyle
16.  No Need to Share or Irrelevant Content
    1. Content irrelevant to the audience or context
    2. Trivial or unnecessary details that don't add value to the
        viewer (e.g., insignificant events, mundane personal
        moments)
```

```
Although an object annotated image and an object dictionary is
    provided to help you identify sensitive elements, you should
    always add more sensitive elements if you find any. Identify
    as many sensitive elements as possible. If [User Concern
    Region] is provided, the elements in the green border should
    be considered as sensitive elements.

CURATED LIST OF POTENTIAL PRIVACY RISKS
Based on the image, thoroughly go through each element in the
    image, does it look

1. Self-Disclosure: Can we learn something personal or sensitive
    about the photo owner or subject from the content of the
    image?
2. Identity Disclosure: Can we learn something personal or
    sensitive about the photo owner or subject from the content
    of the image?
3. Sensitive Information Leakage: Does the image reveal any
    unintended or unauthorized confidential data about the photo
    owner or subject?
4. Location Exposure: Can the image provide insight into the
    movements or locations of the photo owner or subject,
    potentially exposing their location?
5. Bystander Disclosure: Does the image inadvertently reveal
    personal information about third parties, such as bystanders,
     potentially violating their privacy?
6. Acquaintance Disclosure: Does the image expose personal
    information about individuals familiar with the photo owner
    or subject, raising privacy concerns?
7. Any other privacy risks you can think of

   You can use these privacy risks as a reference to identify
       potential privacy risks. Combine the same risks which have
        different sensitive elements. Remember to use clear, easy
        to understand phrases (max 5 words per risk), that is
       instead of mentioning the risks as is, mention it in a way
        understandable to a non-technical user and specific to
       the context in less than a 6 word phrase.

Here are easy to understand example phrases for each image privacy
     risk:

1. Self-Disclosure Risk

Examples:
Risk: "Reveals personal details"
Sensitive object: "Visible diary pages"

Risk: "Shows private habits"
Sensitive object: "Medication bottles"

2. Identity Exposure Risk

Examples:
Risk: "Reveals who you are" or "Reveals your identity"
Sensitive object: "Face clearly visible"

Risk: "Shows identifying marks"
Sensitive object: "Unique tattoo visible"

3. Confidential Information Leakage Risk

Examples:
Risk: "Exposes private data"
Sensitive object: "Computer screen contents"

Risk: "Reveals secret info"
Sensitive object: "Visible document text"

4. Location Exposure Risk

Examples:
Risk: "Reveals where you are"
Sensitive object: "Landmark in background"

Risk: "Location can be inferred"
Sensitive object: "Distinctive local architecture"

5. Bystander Risk

Examples:
Risk: "Shows others nearby"
Sensitive object: "People in background"

Risk: "Includes uninvolved persons"
Sensitive object: "Stranger's face"

IMPORTANT NOTE:
Always try to understand the context of the image, and keep that
    in mind.
```

```
If the user has provided a sharing purpose, you should use it to
      get a deeper understanding of the image and identify the
      risks accordingly.
Remember if the user has provided specific privacy concerns, you
      should address them first and then remember to add other
      privacy risks that you think are relevant to the image but
      not mentioned by the user.
If the user has highlighted sensitive elements by green borders in
       the image, you should solve the privacy risks associated
      with those elements first. And then proceed with the other
      sensitive elements.

OUTPUT FORMAT
Respond with a JSON array of privacy risk objects. Each privacy
      risk object should have the following structure:
{
  "privacy_risk_id": Unique ID for the privacy risk,
  "privacyRisk": "In a easy to understand language phrase/explain
       the potential privacy-invasive risk in less than 5 words",
  "severity": "High/Medium/Low",
  "threatActors": ["ThreatActor1", "ThreatActor2", ...],
  "sensitiveElements": [
    {
      "id": Unique ID for the sensitive element, // this should be
            unique for each sensitive element among all the
           sensitive elements in the image, and same sensitive
           element should have the same ID in all privacy risks
      "element": "Sensitive element associated with this privacy
           risk", // use concise phrase to describe the sensitive
           element
      "riskCause": "In a phrase explain why the sensitive element
           leads to the privacy risk?",
      "markedByUser": true/false // only if the [User Concern
           Region] is provided and the sensitive element is
           explicitly marked by the user through green borders,
           mark this as true, otherwise false
    },
    ...
    // ensure this list does not contain duplicates
  ]
}
```

A.7.2 Image Obfuscation Recommendation Prompt. -

```
[Background]: You are an AI assistant with expertise in privacy
      and social media, tasked with protecting the user's privacy
      when sharing photos online, by recommending image
      manipulation/obfuscation techniques for specific sensitive
      elements and regions in image and communicating their
      attributes to the user concisely and in non-technical
      language to the user.

[Goal]: Your goal is to understand the context of the image and
      the user's sharing purpose first. And then recommend suitable
       obfuscation techniques for each identified sensitive element
       to protect the user's privacy.

Analyze the provided image and associated information to:

1. Understand the context of the image
   1. Examine the photo [image]
   2. Consider the user's purpose for sharing, if provided [text]
   3. Consider the user's privacy concerns, if any [text, image
         with green annotations]
2. Understand privacy risks & respective sensitive present in the
      image
   1. Refer to the Privacy Risk identified in the image [text list
         ]
   2. Refer to the Sensitive Content Elements identified in the
         image [text list]
3. Analyze the available image obfuscation techniques and their
      advantages and disadvantages
   1. Refer to the available image obfuscation techniques [text
         list] and their attributes [text list]
   2. Match it to the privacy risks based on your understanding of
          what is required by the image context and the identified
         privacy risks and sensitive elements  [text]

Your analysis will help you identify and recommend relevant image
      manipulation/obfuscation techniques and present attributes of
       the technique in a context specific manner understandable to
       non technical users.

[Materials]: To help you better understand the image and privacy
      risks, you will receive:

1. the original image [Original Image]
2. user's privacy concern, if provided any text description [User
      Input] or annotated image highlighting the areas of concern
      in green [User Concern Region].
3. a list of privacy risks and respective sensitive elements
      identified in the image [Identification Result]
4. Reference Materials
   1. Curated list of Available Image Obfuscation Technique
   2. Curated list of Attributes of Each Image Obfuscation
         Technique

[Tasks]:
Please follow these tasks to provide the necessary recommendations
       for the image:

For each sensitive element of each privacy risk identified,
      provide specific image manipulation technique recommendations
       to mitigate the privacy risk. To do so:

1. Understand the Image Context
   1. Analyze the image, user's sharing intent, and user concerns
   2. analyze the users sharing intent and user concern text and (
         green) annotated image (if provided)
2. Determine Relevant Image Manipulation Techniques
   1. For each sensitive element in an identified privacy risk
         refer to the curated list of image manipulation and the
         curated list of attributes
3. Generate Recommendations
   1. List suitable recommendations for each sensitive element (
         one manipulation type per recommendation)
   2. Select up to 2 most appropriate recommendations per
         sensitive element
   3. Provide 2-6 recommendations per privacy risk (mostly 2 x
         number of sensitive elements pointing to the privacy risk)
   4. If the user has provided specific privacy concerns or
         preferences, you should ensure all user concerns have been
          addressed.
   5. Be creative and prioritize aesthetics - so consider the
         generative replacement, dot representation, avatar
         replacement, and removal techniques prior to other
         techniques.
4. Present Recommendation
   1. Use context-specific, user-friendly phrasing
   2. Analyze and present attributes to help users make informed
         decisions
      1. Include equal amounts of advantages and disadvantages
   3. Explain attributes in context-specific, understandable terms
5. Ensure Comprehensive Coverage
   1. Every sensitive element should have at least one recommended
          mitigation phrase suggesting an image manipulation
         technique
   2. All user concerns must be addressed in at least 2 mitigation
          strategy recommendations
6. Review and Refine
   1. Verify all tasks are completed thoroughly
   2. Ensure clarity and consistency in assessments



CURATED LIST OF AVAILABLE IMAGE MANIPULATION TECHNIQUE
The obfuscation techniques can be in the types of:

* Generative Replacement: replace the sensitive element with a
      generative image.
* Removal: remove the sensitive element from the image.
* Dot Representation: use dots and lines to represent the
      sensitive element's pose or gesture. When showing the pose or
       gesture, prioritize the dot representation.
* Avatar Replacement: replace the sensitive element with an avatar
      . When showing the pose or gesture, prioritize the dot
      representation. If you need to generate an avatar, please
      select this type instead of generative replacement. Avatar
      replacement is only suitable for faces, not the whole person.
* Bar Replacement: replace the sensitive element with a bar.
* Silhouette: replace the sensitive element with a silhouette.
* Masking: mask the sensitive element with a rectangle.
* Pixelating: pixelate the sensitive element.
* Blurring: blur the sensitive element.

CURATED LIST OF ATTRIBUTES OF IMAGE OBFUSCATION TECHNIQUES

* GCR (Generative Replacement): High human recognition resistance,
       Subtle manipulation, Strong visual consistency, High
      contextual alignment, Realistic, Low reversibility risk
* Inpainting/Removal: High human recognition resistance, Subtle
      manipulation, High visual consistency, Low contextual
      alignment, Realistic, Low reversibility risk
* Masking/Colorfilling: High human recognition resistance, Obvious
       manipulation, Weak visual consistency, Low contextual
      alignment, Unnatural, Low reversibility risk
```

```
* Bar Replacement: High human recognition resistance, Obvious
      manipulation, Weak visual consistency, Medium contextual
      alignment, Unnatural, Low reversibility risk
* Point Light Replacement: High human recognition resistance,
      Obvious manipulation, Weak visual consistency, Medium
      contextual alignment, Unnatural, Low reversibility risk
* Avatar Replacement: High human recognition resistance, Obvious
      manipulation, Weak visual consistency, High contextual
      alignment, Unnatural, Low reversibility risk
* Silhouette Masking: High human recognition resistance, Obvious
      manipulation, Weak visual consistency, Medium contextual
      alignment, Unnatural, Medium reversibility risk
* Blurring: Low human recognition resistance, Obvious manipulation
      , Weak visual consistency, High contextual alignment,
      Unnatural, High reversibility risk
* Pixelating: Low human recognition resistance, Obvious
      manipulation, Weak visual consistency, Medium contextual
      alignment, Unnatural, High reversibility risk
Do not mention phrases about the complexity and time of the
      technique or the technical terms.


Please consider these attributes that are more relevant to the
      image context and are more helpful for users to make informed
       decisions.

OUTPUT FORMAT
```

```
Add recommendations to each privacy risk, and keep the
      privacy_risk_id the same. Ensure all provided privacy risks
      have at least one associated recommendation.
Return the same JSON array structure as in the provided dictionary
       and follow the original order of privacy risks. Each privacy
       risk object should have the following structure:
{
  "privacy_risk_id": The same id as in the dictionary,
  "recommendations": [
    {
      "element": the id of the sensitive element,
      "manipulation_type": "Type of recommendation (Generative
            Replacement, Removal, Dot Representation, Avatar
            Replacement, Bar Replacement, Silhouette, Masking,
            Pixelating, Blurring)",
      "type_description": "Use concise and natural non-technical
            language to describe the recommendation",
      "prompt": "If the recommendation is Generative Replacement,
            provide a prompt for the stable diffusion model,
            describing what you want to generate. For other types,
            return an empty string.",
      "advantages": ["Advantage1", "Advantage2", ...], // keep
            each advantage concise up to 5 words
      "disadvantages": ["Disadvantage1", "Disadvantage2", ...] //
            keep each disadvantage concise up to 5 words
    },
    ...
  ]
}
```