Examining Human Perception of Generative Content Replacement in Image Privacy Protection

Anran Xu
Interactive Intelligent Systems Laboratory, The University of Tokyo
Tokyo, Japan
anran@iis-lab.org

Shitao Fang
Interactive Intelligent Systems Laboratory, The University of Tokyo
Tokyo, Japan
fst@iis-lab.org

Huan Yang
Microsoft Research
Beijing, China
hyang@fastmail.com

Simo Hosio
Center for Ubiquitous Computing, University of Oulu
Oulu, Finland
Tokyo College, The University of Tokyo
Tokyo, Japan
simo.hosio@oulu.fi

Koji Yatani
Interactive Intelligent Systems Laboratory, The University of Tokyo
Tokyo, Japan
koji@iis-lab.org

ABSTRACT

The richness of the information in photos can often threaten privacy, thus image editing methods are often employed for privacy protection. Existing image privacy protection techniques, like blurring, often struggle to maintain the balance between robust privacy protection and preserving image usability. To address this, we introduce a generative content replacement (*GCR*) method in image privacy protection, which seamlessly substitutes privacy-threatening contents with similar and realistic substitutes, using state-of-the-art generative techniques. Compared with four prevalent image protection methods, *GCR* consistently exhibited low detectability, making the detection of edits remarkably challenging. *GCR* also performed reasonably well in hindering the identification of specific content and managed to sustain the image's narrative and visual harmony. This research serves as a pilot study and encourages further innovation on *GCR* and the development of tools that enable human-in-the-loop image privacy protection using approaches similar to *GCR*.

CCS CONCEPTS

• **Security and privacy** → **Privacy protections**; **Usability in security and privacy**.

KEYWORDS

image privacy, usable security, generative artificial intelligence, diffusion models

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

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642103

ACM Reference Format:
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 (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3613904.3642103

1 INTRODUCTION

Images form the foundation of technology-supported communication, including online social networks [11], daily-life recording [13], and information exchanges at workplaces [68]. Meanwhile, privacy issues stemming from the pervasive image capturing and sharing culture online are highlighted and studied in the field of human-computer interaction (HCI) [41] and machine learning (ML) [46, 53, 80]. Image obfuscation is thus critical as one of the direct approaches to privacy protection. Obfuscation techniques help block potentially malicious detection of privacy-threatening visual information [41]. They also present a unique challenge: as obfuscation intensifies, both the usefulness of images and the willingness to share content diminish [29, 30, 44, 67]. To achieve a better balance of protection and utility of images, existing research has investigated various alternative obfuscation methods, including style transformations [24], sticker overlays [40], avatars [44], and cartoon substitutes [31]. However, as these methods introduce artificial visual components, resultant images may cause unnaturalness and may diminish viewers' experience.

Generative artificial intelligence (AI) can generate realistic images with various prompts [20, 52, 59] and enable various image editing applications [4]. We envision that generative AI can offer novel image obfuscation methods with unique value in image privacy protection, through seamless replacement of visual content in images. Computer vision research has already explored the technical development of such approaches for human faces [17, 62] and vehicle plates [76]. A deep understanding of how obfuscation by such technology could impact user perception and experience would benefit future interactive system and interface designs using generative AI for image privacy protection.

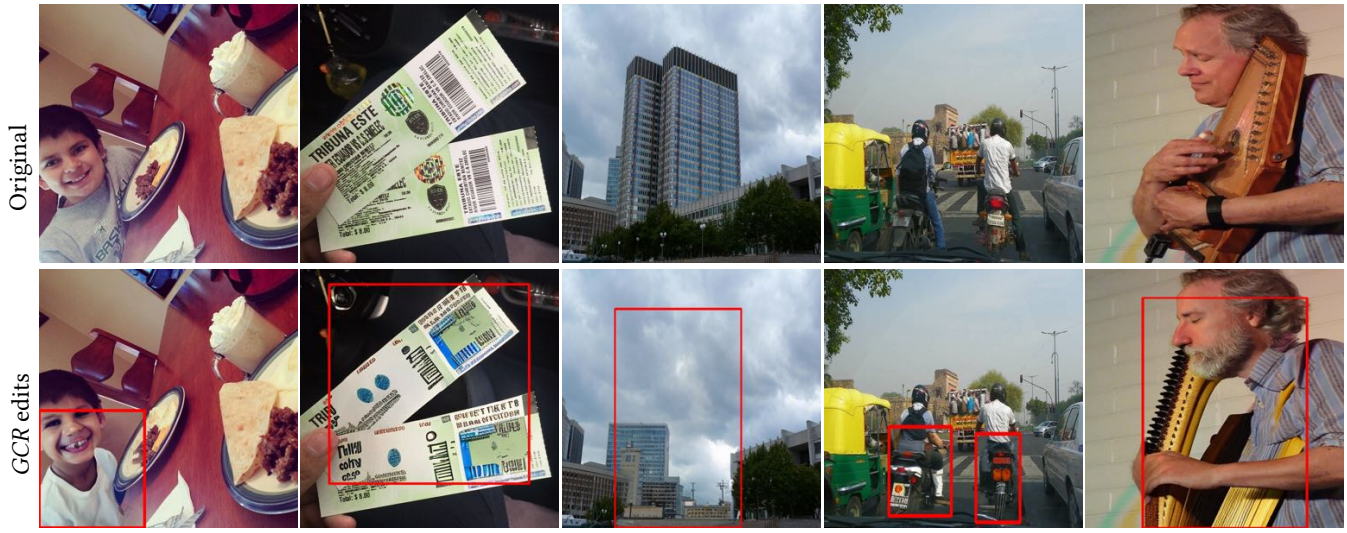

Figure 1: Examples of images processed by our proposed method of generative content replacement (*GCR*). In the provided images, the highlighted contents were annotated as privacy-threatening in an image privacy dataset called DIPA [73]. *GCR* removed these segments and introduced similar substitutes to replace them. This strategy ensures protection against revealing identifiable information but allows for the overall narrative to be communicated through shared images.

In this paper, we explore the feasibility of image-generative AI in obfuscating multiple different types of privacy-threatening content. This includes an examination of user perceptions and experiences against other commonly employed methods. Our method, *Generative Content Replacement* (*GCR*), leverages state-of-the-art image captioning [39] and generative diffusion AI [57] (Figure 1). GCR is not constrained by the types of objects to be obfuscated, providing a capability to enable image obfuscation in realistic, practical settings.

Specifically, this work offers the following three contributions to the field of HCI and machine learning:

- Development of *GCR*, using a state-of-the-art image captioning method [39] and generative diffusion AI [57].
- An experiment using *generative content replacement (GCR)* with 270 images covering 23 privacy-threatening categories against four common protection methods, to encompass a plethora of realistic image editing scenarios for privacy protection.
- Data analysis results that validate the advantages of *GCR* over existing common image protection methods.

We found that our participants were not able to detect edits by *GCR* for protection in 60% of the images tested, confirming high visual integration of *GCR*. The results also show that *GCR* maintained the story of the original images and demonstrated higher visual harmony in the edited images than *blurring*, *colorfilling*, and *removal*. These findings can inspire researchers to design interfaces incorporating *GCR* for image privacy protection and encourage machine learning (ML) experts to develop algorithms tailored for *GCR*.

2 RELATED WORK

Image privacy protection is a topic of significant interest in both the HCI and ML communities. Recent literature reviews have consolidated various methods of image privacy protection concerning human vision [1, 46], computer vision [28, 45], and both domains [80]. Here, we present related work on image protection methods and associated perceptual experiments from an HCI perspective. In addition, Section 2.3 briefly introduces generative models in visual content generation, explaining our proposition that *GCR* could be a novel method for image privacy protection, effective against both human vision and computer vision.

2.1 Image Protection Methods

Our primary focus is investigating the usability of *GCR* through human-centered approaches. Therefore, techniques tailored to thwarting malicious computer-vision detection, such as image perturbation [10], are not part of our discussion. Similarly, non-edit image protection approaches, like access control [49], are excluded due to our specific emphasis on direct image data protection.

Commonly employed filtering methods such as mosaics have been leveraged to prevent privacy breaches by photo owners [69], co-owners [64], and bystanders [7, 21, 55]. However, heavy implementation of these methods often results in a significant decline in both usability and sharing inclination of images [67], thus limiting their applicability. Alternatives that employ transformation effects such as aging [51] and style transformation [30] were proposed to offer better visual aesthetics for edited images. However, advancements in adversarial methods like upsampling [2] and reverse generation [81] can potentially revert transformed images to their original state, weakening the purpose of privacy protection.

A more robust approach involves the direct removal of privacy-threatening contents in images, termed *image inpainting* or *removal* [8]. While this eliminates the appearance of privacy-threatening content, it might lead to potential misinterpretations and unnatural feelings because of information loss [53]. Researchers have attempted to address this by introducing non-privacy-threatening substitutes like stickers [40], avatars [44], and cartoon replacements [31]. Although they fulfill their intended purpose, their practicality remains constrained. For instance, methods that depend on pre-existing material libraries for sticker or cartoon replacements might become ineffective if suitable matches are not found. Furthermore, these substitutions often exhibit a visual disconnect from the rest of the image due to their distinct styles. While some people might appreciate stylistic alterations in images, such modifications might be less preferred when individuals seek to share photos retaining their original aesthetic. We, therefore, argue that an image content replacement method using realistic substitutes can be more effective in ubiquitous scenarios of image privacy protection.

2.2 Perception of Image Protection Methods

HCI research has deeply investigated how people perceive image protection techniques. One pivotal finding was the negative correlation between the sharing intentions of images and the intensity of their protection, through applying mosaics of varying intensities on the same set of images [67].

In subsequent studies, researchers further explored the visual perception of images edited using diverse methods. Hasan et al. conducted a thorough examination of five prevalent image protection strategies: *Blurring*, *Silhouette*, *Pixel*, *Masking*, and *Edging* [29]. By using 100 images across 20 unique scenarios, they introduced four metrics to assess the trade-off between privacy and utility for each method, providing a comprehensive understanding of how each technique performed in various scenarios. Building on this, Hasan et al. studied if style transformation outside of the edited areas could increase the aesthetic of images, though their findings didn't indicate a significant boost in viewers' satisfaction [30].

Li et al.'s comparative analysis examined the influence on the identification of 14 specific individuals using eight different protection methods [44]. Their results emphasized *inpainting* and *avatar* as techniques that best balanced between avoiding privacy leakages and maintaining viewer perception.

Zhao et al. recently characterized the dimensions of human vision adversary protection into two primary categories: imperceptibility and perceptibility [80]. Imperceptibility, in particular, is a critical metric in image protection and evaluates the degree to which the edits are seamlessly integrated into images, serving as an indicator of the "naturalness" of a protection approach. Following these foundational studies, our research adopts their methodologies, selecting representative image protection techniques for comparison with *GCR* in a two-stage experiment for evaluating *GCR*'s feasibility in various aspects of human perception.

2.3 Generative Models of Images

In this section, we introduce mainstream generative models, including the generative adversarial network (GAN) and diffusion models, and then introduce their applications and argue their potential in privacy protection.

GAN was one of the most common generative model algorithms [26]. In general, GANs are finetuned by training two neural networks concurrently. The generator network, *G*, produces synthetic content, while the discriminator network, *D*, evaluates the authenticity of that content against ground-truth data. The generator's goal is to continually improve its content generation until the discriminator can no longer reliably discern between the synthetic and real data. In privacy protection, researchers leveraged the potential of GANs to replace identifiable details with generated content. Recent advancements have exploited the capability to modify the latent space within GANs, leading to nuanced alterations in the generated outputs, therefore preventing privacy-threatening details from being identified. Examples include subtle changes in facial features [62], adding makeup effects [32], and the transformation of identifiable components such as vehicle plates in privacy-threatening images [76].

Diffusion models, on the other hand, can create visual content through a structured denoising procedure. Unlike traditional image generation, which starts from scratch, diffusion models initiate with a noisy version of the target image and iteratively refine it to achieve the desired output [57, 58]. These models are typically initialized based on prompts, which can range from textual descriptions to visual cues such as image contours [77]. Recent advancements in interactive applications of diffusion models enable people to specify the desired content to replace marked areas in target images, ensuring a coherent blend with the rest of the image [52, 79].

Images generated by diffusion models have been demonstrated to be highly realistic, closely resembling photographs taken in the real world [50]. Given the capabilities of diffusion models, we posit that these interactive generative techniques hold great promise for boosting privacy protection in ubiquitous scenarios. When detecting privacy-threatening content, users could seamlessly replace it with realistic generative content that retains the image's narrative coherence, and original style as well as blocks any potential malicious observation or detection. This work investigated how *GCR* works in ubiquitous scenarios of image privacy protection, scrutinizing the effectiveness of state-of-the-art diffusion model-generated visual content in protecting image privacy from a human-centered perspective.

3 STUDY DESIGN

3.1 Research Questions

To frame our research, we first define Generative Content Replacement (*GCR*) as "an image privacy protection method employing generated visual substitutes to replace content in images that may compromise privacy". To examine the user perception and experience of *GCR*, we frame our investigation around three research questions:

- **RQ1.** To what extent can viewers detect the edits with *GCR* (detectability)?
- **RQ2.** To what extent can viewers identify original content from images edited with *GCR* (vulnerability)?

- **RQ3.** How well do images processed by *GCR* perform using common evaluation metrics derived from related image protection research?

One of the main objectives of our study is human vision adversary protection, as discussed by Zhao et al. [80], who identify the ability of humans to detect image manipulation as a key evaluation metric across modern image protection methods. In our study, we evaluate this as *detectability*, and, additionally, evaluate the related cognitive effort of detecting the edit. We also study the *vulnerability* of image protection methods, referring to the degree to which people are able to identify original contents covered or replaced with image protection methods. Prior to gauging vulnerability, the participants were informed that the image being observed was edited. This ensures that vulnerability measurements are not biased towards just the participants who correctly recognized an image being edited.

Finally, we include the following evaluation metrics in the study to emphasize human factors of using *GCR*:

- *Narrative coherence*, inspired by work by Hasan et al. [29, 30], which emphasizes the importance of retaining sufficient information from the original image. Similarly, in this work, we define the preservation of information sufficiency as *narrative coherence* that measures how well the original story is preserved in edited images. As such, it is a natural choice in assessing image privacy protection methods in contexts where people e.g. wish to share images online to share an experience but hide specific aspects from the images.
- *Visual harmony*, is likewise inspired by Hasan et al. [29] who discuss how disruptive changes to the visual or aesthetic appeal of the image deteriorate the user experience of image protection methods.
- *User satisfaction*, discussed in this context by Li et al. [44] and Hasan et al. [29], and directly indicative of people's willingness to share photos online [18].

3.2 Generative Content Replacement

Figure 2 illustrates the workflow of our *GCR* method implemented for this study. It leverages textual prompts to produce visual substitutes that match the type of the original content. When there exists visual content within an image that threatens privacy, *GCR* utilizes BLIP-2 [39], a state-of-the-art image captioning model, to automatically produce two textual descriptions: T_C for the specific privacy-threatening content and T_I for the entire image. Merging the two descriptions yields a cohesive prompt, framed as "T_C within the context of T_I". Subsequently, the image stripped of its privacy-threatening content is combined with this prompt and processed using the stable diffusion v2.1 model [57]. This model performs a 50-iteration denoising cycle with DPM-Solver++ [48] as the sampling solver, eventually producing an image that integrates generative substitutes seamlessly into the removed area. We directly used the bounding box annotations available in an image privacy dataset called DIPA [74] (detailed in Section 3.4) to inform our system of where the edit should be placed. While further improving this *GCR* method may be possible (e.g., employing models and algorithms specifically designed for this purpose or exploring more appropriate prompts), our present work focuses on understanding user perceptions of *GCR* implemented with the latest machine learning models available as of summer in 2023. The rightmost column in Figure 3 shows example images with *GCR*.

3.3 Reference Image Privacy Protection Methods

While there exist numerous editing methods for image privacy protection, previous research did not show strong differences in human perception among blurring and pixelation [29], as well as among various style transformations [30]. To simplify our experimental designs, we selected four representative methods, each occupying a distinct space in terms of detectability and vulnerability, summarised in Table 1. The four columns besides the leftmost in Figure 3, show example images with these approaches.

- *Blurring*. Our *blurring* method entails copying pixels from an image downsampled by a factor of 30x to the original, thereby emphasizing its role in privacy protection. As the *blurring* approach essentially presents a reduced-resolution version of the original content, we posited that it possesses **high** detectability and **high** vulnerability.
- *Cartooning*. We decided to leverage a GAN-based model to perform direct cartoonization for realizing arbitrary privacy-threatening content protection [71]. It serves to obfuscate details, such as eyes in a face or text in a document, to prevent potential privacy leakages. As cartoonization is not as striking as *blurring* or *colorfilling* and maintains some information about the original content (e.g., shapes, sizes, and colors), we consider this cartooning transformation an image privacy protection approach with **low** detectability and **high** vulnerability.
- *Colorfilling*. We adopted a color-filling approach where each position within the edited areas is assigned a random pixel color. Although the term *encryption* has been used in prior work to describe similar methods [55], our implementation does not involve actual information encryption, but rather randomly changes the color value of each pixel. Given its pronounced visual effect, which entirely obfuscates the information in the edited regions, we consider that the *colorfilling* is an approach that exhibits **high** detectability and **low** vulnerability.
- *Removal*. While diffusion models can remove visual content [57], they may introduce unrelated foreground elements rather than seamlessly blending the stripped parts of images with their adjacent backgrounds. Therefore, we leveraged the best pretrained LAMA image inpainting model [66] to obliterate privacy-threatening content while extending adjacent portions of the image stably. We consider that the *removal* technique is a method with **low** detectability and **low** vulnerability in both object-level and category-level identification.

3.4 Image Selection and Processing

Our objective was to deliver a comprehensive examination of *GCR* with reference methods. We chose DIPA [74], an image privacy dataset comprising 5,897 content-level annotations across 1,304 images highlighting privacy-threatening content, as our data source.

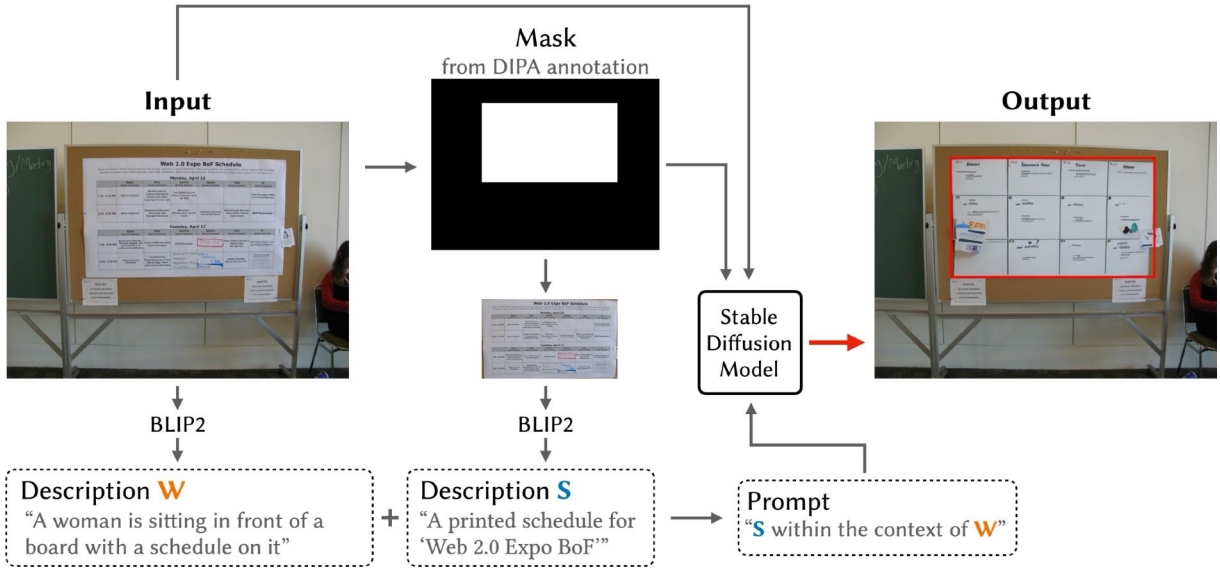

Figure 2: The proposed workflow of generative content replacement (*GCR*) method. Utilizing an image input, our *GCR* method employs the BLIP-2 model [39] to formulate two distinct prompts: prompt T_I that encapsulates the entire image description, and prompt T_C which focuses on the specific privacy-threatening content, as identified in DIPA [74]. Following this, *GCR* incorporates the original image, the mask obtained from DIPA, and a prompt articulated as "T_C within the context of T_I", into the stable diffusion v2.1 model [57]. This produces an output (here, highlighted by a red bounding box) where the privacy-infringing content has been seamlessly substituted with similar generated counterparts.

Table 1: Comparison of the four image privacy protection methods in terms of their detectability and vulnerability. A lower level of detectability suggests that human vision (HV) is less likely to detect any edits, while higher vulnerability means that upon being informed of edits, individuals can more easily identify the original content from edits.

	High Vulnerability	**Low Vulnerability**
High Detectability	*Blurring*	*Colorfilling*
Low Detectability	*Cartooning*	*Removal*

DIPA provides annotations of multiple objects and categories of content in each image if they exist. In our study, however, we used the DIPA images to obfuscate only one category of privacy-threatening content (e.g., person, place identifier) per image. Thus, the privacy-threatening content in each image in the study could consist of multiple objects, each encapsulated within a bounding box, but all within the same category. We then classified the data based on three parameter spaces to adequately represent and diversify the types of privacy-threatening content.

- *Relative Size.* This represents the proportion of the content (as defined by its bounding box) to the image's overall size. For annotated content containing multiple bounding boxes, we summed the sizes altogether. We filtered out content that was either too small (<0.005) or too large (>0.7) relative to the entire image size. The remaining images were divided into three categories: small (lowest 30%), medium (middle 40%), and large (highest 30%) based on their relative sizes.
- *Relative Position.* This parameter space represents where the privacy-threatening content is located in the image. We computed the distance between the center of the bounding box and the image's center, divided by the image's diagonal. For multi-object contents, we took an average number of all results of relative positions. The processed data was then divided into three segments: near, mid-range, and far from the image center with the same 30%-40%-30% split as the relative sizes.
- *Aspect Ratio.* This parameter space represents the ratio of an object's width to its height. We calculated an average number of all aspect ratio values for multiple objects in one image. We also devised three categories of vertical, balanced, and horizontal, and the same 30%-40%-30% split was applied.

Our selection included 270 unique images so that they were uniformly distributed across the three parameters discussed above. While satisfying the above conditions, we ensured that the images we selected, to the extent feasible, evenly encompassed the 23 privacy-threatening categories identified in DIPA rather than maintained the same occurrence distribution of these categories. Our data balancing, however, did not consider the factor of higher-level

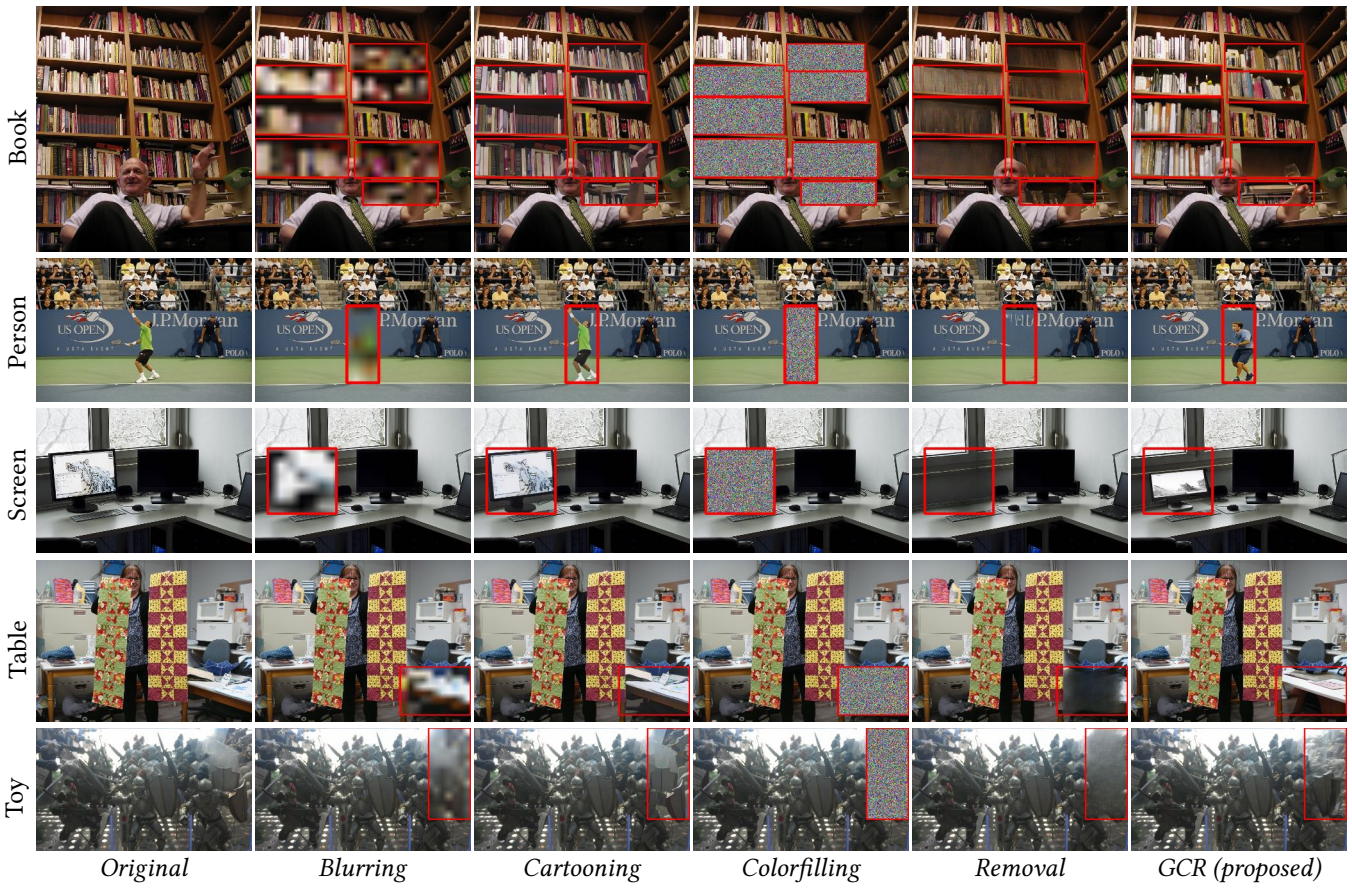

Figure 3: Sample images from our selection. Each image was processed by *GCR* as well as the other four reference methods. The edited image regions are highlighted with red bounding boxes.

descriptions of each category (Section 4.2) because annotations in some DIPA images exhibited high divergence.

Figure 4 and Table 2 detail the distributions of bounding boxes and distributions of privacy-threatening content categories, respectively. Each image was processed with *GCR* and the four reference methods explained in Section 3.3, resulting in five images with obfuscation for each original image (1,350 processed images in total). While this selection might not mirror the exact occurrence of privacy-threatening content in real-world photos, it was designed to ensure sample balance and consider different obfuscation scenarios.

4 USER STUDY

We conducted a two-stage experiment to benchmark *GCR* and investigate the research questions presented earlier. The procedure was approved by the institutional review board at the first author's institution.

4.1 Participant Recruitment

We recruited participants through Prolific [56], a reliable crowdsourcing platform that proved to offer high-quality responses from participants [23]. The inclusion criteria of participants were native-level English proficiency; primary usage of English in daily life; being 18 or older; and a willingness to view photographs.

4.2 Pre-experimental Stage

At the beginning of the experiment, we explained the overview of the tasks and expected workload to the participants. The participants could first preview sample images from DIPA [73] to help them decide whether to proceed with viewing similar images during the study. As a safety measure, we informed participants that they were allowed to withdraw from the study at any time and for any reason. After their consent to participate, we asked them how carefully they would consider the potential of leaking five different types of privacy information (*personal information*, *location of shooting*, *individual preferences/pastimes*, *social circle*, and *others' private/confidential information*) derived from higher-level descriptions of different privacy-threatening types in DIPA (e.g., *vehicle plate* may indicate *personal information* and *location of shooting*) [73]. In other words, the five types assessed in the **PreQ** encompass the 23 distinct categories outlined in Table 2. Through

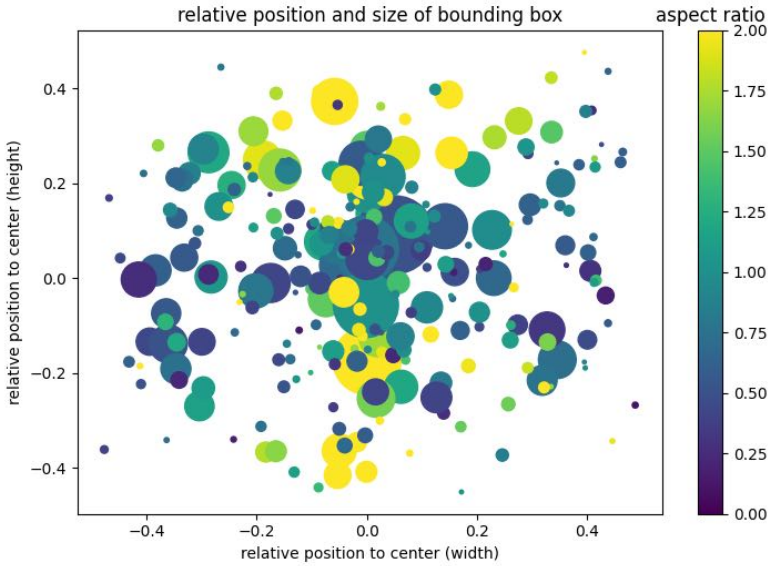

Figure 4: The distribution of bounding boxes on 270 privacy-threatening content selected from DIPA [74] in our study. Each circle symbolizes visual content to be processed in one image and is positioned relative to the image's center. The circle's diameter reflects the relative size of the bounding box within its respective image, though it is presented in a scaled manner to prevent visual clutter. The color variation indicates the aspect ratio of the bounding boxes, derived from their width-to-height ratio.

this approach, we gauged the level of participants' concerns regarding privacy-threatening information while keeping the task workload low.

- **PreQ.** How carefully do you consider the potential of leaking <each of the five privacy information types, one at a time> when you are sharing photos?
  Response: [1: not at all carefully] – 2 – 3 – [4: neutral] – 5 – 6 – [7: extremely carefully]

4.3 Stage I - Assessing Detectability

After the pre-experimental questionnaire, participants were asked to perform tasks associated with edit detectability of given images to answer **RQ1**. 10 randomly-selected images from our 1,350 edited images were shown to each participant. If they identified edits, we further required them to locate the edited region by clicking in the given images, and rate the difficulty of detecting the corresponding edit. We note that participants were not informed whether the given images contained any edits, to avoid potentially biasing their responses. The specific questions were listed as follows:

- **Q1-1: Edit Detectability.** Do you believe the photo has been digitally edited (i.e., has some visual information been changed from the original version taken in real life)?
  Response: yes / no
  If the response to *Q1-1* was *yes*, participants were presented with the subsequent questions.
- **Q1-2.** Please click on the part of the photo where you believe the editing has been applied.
- **Q1-3: Perceived Difficulty of Edit Detection.** How difficult or easy is it to detect that the photo has been edited? Rate on a scale of 1-7.
  Response: [1: extremely easy]– 2 – 3 – [4: neutral] – 5 – 6 – [7: extremely difficult]

In one randomly chosen image of the 10 images processed by each participant, we also added an additional question item as an integrity check that instructed the participant to choose the option *strongly agree*:

- **Integrity Check.** For this question, please select the option "strongly agree".
  Response: [1: strongly disagree] – [2: disagree] – [3: neutral] – [4: agree] – [5: strongly agree]

In the subsequent analysis, we ignored data from participants who failed the integrity check.

4.4 Stage II - Assessing Vulnerability And other Perception Evaluation

After participants finished all the tasks in the first stage, we disclosed that all these images had been edited and asked them to respond to additional questions to answer **RQ2** and **RQ3**. We juxtaposed each of the edited images assigned in Stage I and its original image with the location of the edit highlighted in a red bounding box. A toggle button next to each image set allowed participants to view only the region where the edit was employed. We did not reveal the name of the obfuscation method to participants at this stage to avoid potential biases (e.g., *GCR* may create an impression

Table 2: The distribution of selected privacy-threatening content in our research according to 23 privacy-threatening categories identified by DIPA [74], also presented by three parameter spaces we defined in Section 3.4. All divisions were 30%-40%-30% split. We selected 10 images for each combination of factors: *relative size*, *relative position*, and *aspect ratio*. This resulted in a total of 270 images, with each split interval of every parameter space comprising 90 images, as depicted in the last row.

Category	Relative Size			Relative Position			Aspect Ratio			All
	small	*medium*	*large*	*near*	*mid-range*	*far*	*vertical*	*balanced*	*horizontal*	
Accessory	7	6	1	5	5	4	4	4	6	14
Book	5	4	6	5	6	4	4	5	6	15
Cigarette	2	0	0	1	1	0	0	1	1	2
Clothing	5	4	5	4	4	6	5	4	5	14
Cosmetics	0	1	4	2	2	1	3	1	1	5
Electronic Device	2	2	1	1	2	2	2	1	2	5
Finger	2	1	4	1	3	3	1	5	1	7
Food	2	5	2	2	4	3	3	2	4	9
Home Interior	5	4	5	4	5	5	6	4	4	14
Identity	3	5	5	5	4	4	1	8	4	13
Machine	4	2	2	2	5	1	3	2	3	8
Musical Instrument	3	6	5	5	6	3	6	5	3	14
Person	6	4	5	5	4	6	8	5	2	15
Pet	3	7	5	7	3	5	5	6	4	15
Photo	2	2	2	4	1	1	0	5	1	6
Place Identifier	6	4	5	5	4	6	8	4	3	15
Printed Material	5	4	5	3	5	6	4	5	5	14
Scenery	4	5	5	6	3	5	6	2	6	14
Screen	5	4	5	5	5	4	3	7	4	14
Table	5	4	4	4	5	4	1	3	9	13
Toy	4	6	4	5	4	5	5	4	5	14
Vehicle Plate	5	5	5	3	6	6	5	3	7	15
Others	5	5	5	6	3	6	7	4	4	15
Sum of The All	**90**	**90**	**90**	**90**	**90**	**90**	**90**	**90**	**90**	**270**

that it is a novel approach). We asked the following questions for each image set:

- **Q2-1: Edit vulnerability.** To what extent do you agree that you can accurately identify the particular original object, by only giving its corresponding edit?
- **Q2-2: Perceived Confidence in Maintaining Narrative Coherence.** To what extent do you agree that you can recognize the object in the edited image as (a/an) *category name of the content*, despite not being the same one as the original one?
- **Q2-3: Perceived Visual Harmony with Original Images.** To what extent do you agree that the edited content visually blends with the rest of the image, maintaining the visual harmony as well as the original object did?
  Response for Q2-1, Q2-2 and Q2-3: [1: strongly disagree] – 2 – 3 – [4: neutral] – 5 – 6 – [7: strongly agree]
- **Q2-4: Perceived Overall Satisfaction on Edited Images.** To what extent do you find the photo with the edit satisfying?
  Response: 1: [extremely unsatisfying] – 2 – 3 – [4: neutral] – 5 – 6 – [7: extremely satisfying]
- **Q2-5.** Please click on the corresponding part of the original image that matches the object edited in the given edited image.

The *category name* in **Q2-2** was derived from DIPA's annotations according to 23 identified categories of privacy-threatening content (Table 2). For **Q2-3**, while Hasan et al. used the phrase *"This photo looks visually appealing"* to understand the aesthetic impact of edits [29, 30], we believe that aesthetic perceptions can be deeply subjective. Therefore, we chose "visual harmony" as a more objective measurement that focuses on the preservation of the image's original style. To ensure that participants understood where exactly the edit occurred in each of the given images, we included **Q2-5**. In addition, we incorporated a similar integrity

check in this stage for another randomly chosen image as in stage I.

5 RESULTS

5.1 Participants

We recruited a total of 135 participants. 14 participants did not pass our integrity check, leaving us with 121 valid participants. The demographic distribution based on gender, age, student status, and employment status is detailed in Table 3. The "Others" employment status includes participants who were "not in paid work", "start a new job within the next month", and cases where the data had expired in Prolific. For their ethnicity, 92 (76.0%) participants were white, 15 (12.4%) were Black, 7 (5.8%) were Asian, 6 (5.0%) were mixed and 1 claimed as others (0.8%). They came from 9 different countries: United Kingdom (84 participants), United States (17), South Africa (7), Nigeria (4), Ireland (3), Canada (2), Zimbabwe (2), India (1), and New Zealand (1). The average time taken by participants from starting the task to its completion was 21.6 minutes. Every participant was compensated £3 upon task completion, regardless of their performance in the integrity check.

5.2 Edit Detectability Evaluation

Table 4 presents our experimental results for *Edit Detectability* and *Perceived Difficulty of Edit Detection* of all valid responses. In addition to reviewing answers for the sanity check questions, we analyzed if participants accurately identified the edited regions by examining the responses for **Q1-2** and **Q2-5**. We detected 48 occurrences in Stage I and 14 in Stage II where participants' clicks were outside the bounding boxes in DIPA [73] even with a 10% error tolerance. We excluded responses for images associated with these inaccurate clicks as we regarded that participants were not able to identify the locations of edits precisely or misunderstood. This removal resulted in 1,148 sets of valid responses.

5.2.1 Q1-1: Edit Detectability. The columns under Q1-1 and NP in Table 4 show the number of responses and the ratio of how frequently participants responded negatively (i.e., they did not find any edits), respectively. *GCR* exhibited the lowest edit detectability rate. The results revealed that, without explicit pointers, 60% of the edits processed through *GCR* escaped recognition by participants. We conducted a logistic regression analysis to determine whether participants could detect edits (0 = yes, 1 = no). We summed the responses to the five questions of **PreQ** per participant as their *privacy concern level*, inspired by previous studies [3, 54]. Our independent variables included *protection methods* (i.e., *GCR* and four *reference methods*), *relative size*, *relative position*, *aspect ratio*, and *privacy concern level*. *Protection methods* was treated as a categorical variable, and *GCR* was set as the reference group. Other variables were treated as continuous variables. We weighted the data according to the proportion of *yes* and *no* responses to avoid biases toward the majority class.

Table 5 details the coefficients, p-values of coefficients, odds ratio, and 95% confidence interval of odds ratio (95% CI) of each independent variable resulting from the logistic regression model. We observe that all reference methods except *removal* were statistically significant negative predictors, indicating that it was more likely to detect the edits by *blurring*, *cartooning*, and *colorfilling*. When *relative size* of edit areas became larger, participants were inclined to detect more edits. In addition, if the edit was increasingly far from the image center, our participants tended to detect fewer edits.

5.2.2 Q1-3: Perceived Difficulty of Edit Detection. *GCR* obtained the highest score of 3.33 in perceived difficulties of detecting edits. We conducted a Kruskal-Wallis test first, a method for analyzing non-normally distributed ordinal data [34], to investigate whether different *protection methods* influenced the perceived difficulty of identifying edits. The test confirmed a significant difference in the methods ($\chi^2(4) = 166.45$, $p < 0.001$, $\eta^2 = 0.23$). We then employed the Mann-Whitney U test, a robust non-parametric analysis for ordinal data between two independent groups [82]. This allowed us to closely examine the pairwise mean differences between *GCR* and each reference method, highlighting the distinctiveness of *GCR*. Bonferroni correction was further implemented to account for the multiple comparisons and adjust the p-values accordingly. We also used Cliff's Delta (δ) as a measure of effect size to quantify the magnitude of differences between *GCR* and other *protection methods*. Additionally, we reported the confidence intervals (95% CI) for the median difference in Mann-Whitney U tests. The Mann-Whitney U tests with each reference method revealed that *GCR* edits were more challenging to spot than those made by other methods (between *GCR* and *blurring*: (U = 15,150.0, corrected $p < 0.001$, $\delta = 0.58$, 95% CI = [1.63, 2.37]); between *GCR* and *cartooning*: (U = 7,735.5, corrected $p < 0.01$, $\delta = 0.31$, 95% CI = [0.57, 1.43]); between *GCR* and *colorfilling*: (U = 15,436.0, corrected $p < 0.001$, $\delta = 0.65$, 95% CI = [1.63, 2.37]); between *GCR* and *removal*: (U = 6,530.5, corrected $p < 0.01$, $\delta = 0.29$, 95% CI = [0.50, 1.50])). These analyses confirmed that *GCR* was the most difficult method to recognize in our study.

We constructed a linear regression model to examine how other independent variables, including *relative size*, *relative position*, *aspect ratio*, and *privacy concern level*, influence the responses of this question. Our linear regression model showed that only *relative size* significantly predicted the perceived difficulty of edit detection (b = -1.499, $t(731)$ = -4.490, $p < 0.001$), indicating larger edits were more easily to detect. We acknowledge that the model doesn't offer strong predictive power for the response but presents it due to its statistical significance (adjusted $R^2 = 0.025$, $F(4, 731) = 5.806$, $p < 0.001$).

5.3 Vulnerabilities Evaluation

In evaluating participants' ability to identify original content based solely on the edited image, the *colorfilling* method was the most effective in obscuring identification, with an average rating of 2.82. The average rating for *GCR* was 3.79.

Through a Kruskal-Wallis test, we confirmed that *protection methods* affected the perceived vulnerability ($\chi^2(4) = 176.19$, $p < 0.001$, $\eta^2 = 0.15$). More specifically, Mann-Whitney U tests with Bonferroni corrections revealed significant differences between *GCR* and reference methods except *blurring* (between *GCR* and *cartooning*: (U = 15,469.5, corrected $p < 0.001$, δ = -0.40, 95% CI = [-2.39, -1.61]); between *GCR* and *colorfilling*: (U = 33,131.0, corrected $p < 0.001$, $\delta = 0.26$, 95% CI = [2.58, 3.43]); between *GCR* and *removal*: (U = 29,989.0, corrected $p < 0.01$, $\delta = 0.20$, 95% CI = [1.58, 2.42])). The perceived vulnerability of *GCR* was not significantly different

Table 3: Demographic information of our participants.The "Others" in employment status includes "not in paid work", "start a new job within the next month", and "data expired".

	Age					Student or Not			Employment				All
	18–24	25–34	35–44	45–54	55–	Yes	No	Data Expired	Full-Time	Part-Time	Unemployment	Others	
Male	7	22	33	12	6	17	58	5	51	6	10	13	80
Female	7	14	6	9	5	6	33	2	22	8	2	9	41
All	**14**	**36**	**39**	**21**	**11**	**23**	**91**	**7**	**73**	**14**	**12**	**22**	**121**

Table 4: Distribution of answers collected in Stage I. Detailed references to the question descriptions are in Section 4. The term *NP* refers to the percentage of participants who failed to recognize edits, assessing detectability. Bold font indicates the maximum values of *NP* and question Q1-3, indicating the perceived difficulty of edit detection.

	Q1-1		NP	Q1-3	
	yes	no		mean	std
blurring	206	24	0.1	1.51	0.95
cartooning	119	115	0.49	2.35	1.31
colorfilling	201	34	0.14	1.30	0.77
removal	107	118	0.52	2.45	1.76
GCR (proposed)	89	135	**0.6**	**3.33**	1.80

Table 5: Results from the logistic regression model of edit detectability (0 = yes, 1 = no). We treated *relative size, relative position, aspect ratio,* and *privacy concern level* as continuous variables and treated *protection methods* (i.e., *GCR* and four reference methods) as categorical variables. We set *GCR* as the reference group for *protection methods*. Significant factors are in bold. The Akaike Information Criterion (AIC) was 1214.43.

	Estimated Coefficients	P-Values	Odds Ratios (OR)	95% CI for OR
Intercept	0.450	0.130	1.584	[0.874, 2.869]
***protection methods* (blurring)**	−2.724	< 0.001	0.066	[0.039, 0.110]
***protection methods* (cartooning)**	−0.493	0.014	0.611	[0.413, 0.904]
***protection methods* (colorfilling)**	−2.366	< 0.001	0.094	[0.059, 0.150]
protection methods (removal)	−0.380	0.076	0.698	[0.470, 1.038]
relative size	−4.069	< 0.001	0.017	[0.005, 0.054]
relative position	1.493	0.006	4.451	[1.550, 12.782]
aspect ratio	−0.088	0.136	0.915	[0.815, 1.028]
privacy concern level	0.009	0.400	1.009	[0.989, 1.029]

from that of *blurring* (U = 28211.0, corrected p = 1.0, δ = 0.10, 95% CI = [0.59, 1.41]).

We performed linear regression to analyze the influence of each parameter space on perceived confidence in retrieving obfuscated objects. However, none of *relative size, relative position, aspect ratio,* and *privacy concern level* was a significant factor, and the overall model could not predict this response well (adjusted R^2 = 0.0008, $F(4, 1140) = 0.75$, p = 0.56). This result suggests that these four factors would not offer a clear influence on vulnerability.

5.4 Perception Evaluation

5.4.1 Q2-2: Perceived Confidence in Maintaining Narrative Coherence. In terms of enabling human perception to grasp the overall narrative of images by retaining category-level information, *GCR* obtained a compatible score of 4.34, which was only surpassed by *cartooning* with a score of 5.38.

The Kruskal-Wallis test demonstrated that the choice of *protection methods* significantly influenced the perceived confidence in maintaining narrative coherence of the edit parts for original images ($\chi^2(4) = 212.72$, $p < 0.001$, $\eta^2 = 0.19$). Upon conducting Mann-Whitney U tests with Bonferroni corrections, we found that the responses to each reference method were significantly different from those to *GCR* (between *GCR* and *blurring*: (U = 32,644.0, corrected $p < 0.001$, δ = 0.27, 95% CI = [2.58, 3.42]); between *GCR* and *cartooning*: (U = 18,744.5, corrected $p < 0.001$, δ = -0.28, 95% CI = [-1.39,-0.61]); between *GCR* and *colorfilling*: (U = 36,908.5, corrected

Table 6: Distribution of answers collected in Stage II. Detailed references to the question descriptions are in Section 4. Bold font indicates the maximum or minimum values based on each question's context. Specifically, lower values are preferable for Q2-1 (marked as *L* in the table), indicating decreased vulnerability. Higher scores for Q2-2, Q2-3, and Q2-4 (marked as *H* in the table) are supposed to better preserve *narrative coherence*, *visual harmony*, and *user satisfaction*, respectively.

	Q2-1 (*L*)		**Q2-2** (*H*)		**Q2-3** (*H*)		**Q2-4** (*H*)	
	mean	std	mean	std	mean	std	mean	std
blurring	3.38	2.19	3.23	2.16	2.85	1.81	2.69	1.7
cartooning	5.38	1.73	**5.38**	1.91	4.72	1.89	4.12	1.78
colorfilling	**2.82**	2.3	2.66	2.22	1.81	1.54	2.0	1.74
removal	3.04	2.22	2.99	2.16	4.29	2.13	3.94	2.12
GCR (proposed)	3.79	2.23	4.34	2.23	**5.34**	1.73	**4.98**	1.75

$p < 0.001$, $\delta = 0.41$, 95% CI = [3.56, 4.44]); between *GCR* and *removal*: (U = 33,412.5, corrected $p < 0.001$, $\delta = 0.33$, 95% CI = [2.58, 3.43])).

Our linear regression model indicated that *relative size* significantly predicted the response of perceived confidence in maintaining narrative coherence ($b = 1.196$, $t(1140) = 2.542$, $p < 0.05$). The overall model, however, predicted the responses of **Q2-2** very weakly (adjusted $R^2 = 0.004$, $F(4, 1140) = 2.43$, $p < 0.05$).

5.4.2 Q2-3: Perceived Visual Harmony with Original Images. *GCR* achieved the highest score (5.34) in seamlessly integrating its generated content with the original image style, ensuring that visual harmony remains while protecting privacy.

Differences in *protection methods* could significantly influence the perception of visual harmony, as our Krustal-Wallis test showed ($\chi^2(4) = 394.22$, $p < 0.001$, $\eta^2 = 0.34$). The Mann-Whitney U tests with Bonferroni corrections revealed that each reference method was significantly different with *GCR* (between *GCR* and *blurring*: (U = 42,540.0, corrected $p < 0.001$, $\delta = 0.66$, 95% CI = [3.60, 4.40]); between *GCR* and *cartooning*: (U = 30,756.0, corrected $p < 0.01$, $\delta = 0.19$, 95% CI = [0.66, 1.34]); between *GCR* and *colorfilling*: (U = 47,737.0, corrected $p < 0.001$, $\delta = 0.82$, 95% CI = [4.56, 5.44]); between *GCR* and *removal*: (U = 32,084.0, corrected $p < 0.001$, $\delta = 0.28$, 95% CI = [0.63, 1.37])). Given it had the highest score of 5.34, we verified that *GCR* significantly obtained more agreement in preserving the visual harmony of the original images compared to reference methods.

Our linear regression model demonstrated that *privacy concern level* significantly predicted the response of perceived visual harmony with original images ($b = 0.024$, $t(1140) = 2.507$, $p < 0.05$), indicating people who considered about privacy more would believe image protection methods could maintain visual harmony better. The overall model also showed a weak prediction capability (adjusted $R^2 = 0.04$, $F(4, 1140) = 13.27$, $p < 0.001$).

5.4.3 Q2-4: Perceived Overall Satisfaction on Edited Images. Images edited with *GCR* achieved significantly higher overall satisfaction compared to those edited with the four reference methods. This was confirmed by a Krustal-Wallis test ($\chi^2(4) = 302.18$, $p < 0.001$, $\eta^2 = 0.26$), showing at least one *protection method* was significantly different to others. Further, Mann-Whitney U tests with Bonferroni corrections confirmed significant differences when comparing *GCR* with each reference method individually (between *GCR* and *blurring*: (U = 41,833.0, corrected $p < 0.001$, $\delta = 0.63$, 95% CI = [2.62, 3.38]); between *GCR* and *cartooning*: (U = 33,240.5, corrected $p < 0.001$, $\delta = 0.29$, 95% CI = [0.67, 1.33]); between *GCR* and *colorfilling*: (U = 45,236.0, corrected $p < 0.001$, $\delta = 0.73$, 95% CI = [3.58, 4.42]); between *GCR* and *removal*: (U = 32,016.0, corrected $p < 0.001$, $\delta = 0.28$, 95% CI = [0.63, 1.37])).

Our linear regression model showed that *relative size* ($b = -2.960$, $t(1140) = -7.228$, $p < 0.001$) and *privacy concern level* ($b = 0.026$, $t(1140) = 2.983$, $p < 0.01$) could be a negative influencer and a positive influencer to the overall satisfaction of edited images, respectively. In addition, the overall model could weakly predict these responses (adjusted $R^2 = 0.054$, $F(4, 1140) = 17.44$, $p < 0.001$).

6 DISCUSSION

6.1 Balancing Privacy Protection, Detectability, and Narrative Coherence

Our study affirmed that approximately 60% of the edits made by *GCR* were undetectable to participants, with the remaining edits necessitating a higher degree of scrutiny to be detected. In assessing the vulnerability of *GCR*, we observed that participants' confidence in identifying specific objects within *GCR*'s edits paralleled the confidence levels observed with edits made through *blurring*, as determined by Mann-Whitney U tests (Section 5.3). Although we classified *blurring* as a method with high vulnerability, we utilized a high intensity in our implementation, reducing the original content to one-thirtieth of its original length (a 900-fold downsampling in size). This protection level guarantees that *GCR* effectively obscures many types of content deemed to be privacy-threatening, such as humans, documents, and screens, as highlighted in prior studies on intensified *blurring* effects [29].

The low edit detectability of *GCR* offers a promising avenue for individuals who wish to keep image alterations subtle, an aspect largely overlooked in existing research. If alterations are unnoticed, it may reduce the likelihood of malicious entities trying to reverse-engineer the edits using advanced image generation techniques, such as image upsampling [2]. Furthermore, the risk of information leakage through interpersonal connections could potentially be reduced [19]. For instance, an individual might share a photo from a gathering, opting to blur certain faces to preserve their friends' privacy. However, if someone detects these edits and scrutinizes interpersonal relationships on social networking service (SNS) platforms, they might decipher the identities of the blurred

faces. If they are unable to detect the edits, they might simply move on, believing that someone else appears in the photos. Therefore, we advocate for a heightened focus on "detectability" when developing future image privacy protection strategies.

The high levels of narrative coherence, visual harmony, and overall satisfaction observed with images edited by *GCR* might bolster its use in various image protection scenarios. For instance, in the context of photo sharing on SNS platforms, where the perceived willingness to share photos has been found to be negatively related to the degree of information that is downsampled or obfuscated [67], utilizing *GCR* might appeal to individuals who have previously refrained from employing privacy protection due to concerns over altering the original visual appearance of images. Furthermore, the widespread use of SNS platforms has significantly increased the prevalence of owning and sharing images that contain others in social groups [25, 42]. *GCR* could potentially alleviate privacy conflicts between those willing to share photos and those preferring not. For those inclined to share photos, intensive protection methods proposed by other stakeholders might be ignored [65]. Implementing *GCR*, which maintains the overall visuals of photos, might facilitate a more acceptable balance among individuals from various privacy concern groups. Subsequent research might explore the applicability of *GCR* in diverse sharing settings and conduct qualitative studies to comprehend user preferences regarding the utilization of *GCR*.

6.2 Prospective Applications of *GCR*

GCR shows promise as a tool for image privacy protection, especially when images are intended to convey information, but without revealing personal or identifiable details. Similar to our previous discussions on image sharing on SNS platforms, individuals may wish to share snapshots capturing unique moments involving bystanders. Using *GCR* allows them to preserve the essence of the captured moment without infringing on the privacy of those unwittingly included in the picture. *GCR* can also provide benefits in image-sharing circumstances where contents in the background may potentially expose privacy-threatening information while other information sharing is necessary, such as in workspaces and participatory sensing. In such instances, certain content could be discreetly replaced to ensure the usability of the resulting image and prevent malicious reverting as illustrated in Section 6.1.

Recent HCI research has explored utilizing real substitutes to replace sensitive objects during image data collection [63]. We believe that employing *GCR* could make this process more convenient and expansive in the future, if its integrated diffusion model can generate samples without causing data poisoning that degrades the model training process [6]. When tasks require capturing visual content from the surrounding environment, it can replace privacy-threatening content with substitutes to maintain the usability of image data without needing specific instructions or physical items for participants. This may not only elevate participants' comfort but also augment the volume and diversity of data they're willing to contribute. In the future, it would be intriguing to see researchers deploy *GCR* in crowdsourced data collections and subsequently assess the integrity and quality of the sanitized images.

GCR also holds promise for applications in specific communities. For example, visually impaired individuals who rely on remote assistance often need to capture their surroundings to share with helpers [36]. By integrating *GCR* into these systems, privacy-threatening content can be replaced, thus alleviating concerns from bystanders about their privacy being compromised. This application of *GCR* could enhance the independence and confidence of visually impaired users in their interactions with the environment, while simultaneously respecting the privacy of those around them. *GCR* may also facilitate studies on indigenous peoples, allowing researchers to build trust by respecting private lives and the secret of their unique cultures while recording intangible cultural property [35]. We expect to leverage interdisciplinary collaborations to explore the benefits of deploying *GCR* across various communities.

6.3 Ethical Considerations

Ethical implications of image manipulation techniques are timely in HCI. In our approach, the substitutes generated by *GCR* originate from random noise, which makes it improbable for them to resemble real-world objects. This sets our method apart from harmful impersonation techniques used in Deepfake technologies, as discussed in [72]. However, being powered by advanced generative diffusion models [57], *GCR* highlights the potential responsible AI issues of diffusion algorithms [12, 14], and meets challenges similar to those in AI-driven face synthesis [9]. The implicit prompts within *GCR* could theoretically overlook certain details of the original content (Figure 2), leading the diffusion models to unintentional bias into the outputs. For instance, *GCR* could change ethnic features when replacing human-related content (e.g., altering skin colors). Whether this is an undesired feature in the specific context of obfuscating an image can be of course debated. Nevertheless, future development of *GCR* could focus on enhancing its interactivity and customization capabilities, affording more control to users over the generation process and producing outputs that are both satisfying and responsible.

GCR aims to preserve the original category of the content being replaced, therefore minimizing potential misinterpretations of the edited photos. However, using textual prompts to regulate content generation, it is still possible for *GCR* to generate visual objects that do not belong to the original categories. Future research could focus on evaluating the effectiveness of AI transparency strategies (e.g., content authenticity detection [22] and 'warning flag' [38]) to prevent misunderstanding during the perception of *GCR*'s edits in the context of image privacy protection.

Previous studies have also expressed concerns about human memory distortion stemming from existing image protection methods [33]. Given its ability to alter details in real photos, *GCR* might similarly influence people's recollections of actual events, as suggested by prior psychological research [61, 70]. Therefore, the potential for *GCR* to cause memory distortion should be thoroughly examined alongside existing image privacy protection methods. On the other hand, *GCR* could potentially aid individuals in avoiding specific phobias by replacing real objects in photos with substitutes that do not evoke stress. HCI researchers can deploy *GCR* in various contexts, exploring both its ethical challenges and potential benefits.

Finally, it has been argued that humans should have the right to represent themselves in the world in a transparent way. There exists the question of whether people have the right to take photos of people and share them in an obfuscated form, even if the altercations were performed in the name of privacy protection [16]. Further research should delve into how photo owners and bystanders perceive the disclosure of such edits and how viewers react to AI-altered images, exploring e.g. various AI transparency methods in this context.

7 LIMITATIONS

The distribution of the images used in our data collection might cause biases in participants' perceptions due to uneven exposure to different obfuscation methods. Moreover, we did not conduct a power analysis before the study to confirm the required sample size and each of the images received ratings from only one participant in our data collection. This was to obtain the ratings for a broad range of images. Collecting multiple responses for each edited image could reveal individual variances in how participants perceive edits made by *GCR*. Future research should expand data collection to deepen the understanding of user perception of *GCR* as well as existing image protection methods. This would also help to establish a further understanding of *GCR*'s effectiveness and potential weakness.

In our vulnerability evaluation, we presented the original images alongside their edited versions, asking participants to share their confidence level in identifying particular visual content. Ideally, *GCR* should exhibit a low vulnerability level, close to methods like *removal* or *colorfilling*, given that the generated content has no relation to the original content. The question used to derive this metric was Q2-1 ("To what extent do you agree that you can accurately identify the particular original object, by only giving its corresponding edit?"). *GCR* may produce generated similar objects to the original objects. Our participants thus might have considered that the edited images would convey some information about the original objects.

Future studies should contemplate leveraging techniques such as perceptual area detection [78] and image matching algorithms [60] to further substantiate the effectiveness of *GCR*, or engaging expert participants to provide a more precise analysis of its vulnerability.

In our analysis of the results, we employed linear regression to determine whether the selected parameter spaces (i.e., *relative size*, *relative position*, *aspect ratio*, and *privacy concern level*) had significant impacts on participant responses. Despite these parameter spaces exhibiting significance in various perceptual metrics, e.g., the influence of *privacy concern level* on *perceived overall satisfaction*, the R^2 values in these models remained generally very low. This suggests that these independent variables did not account for a substantial portion of the variation in the responses. We hope that future research will be able to identify more influential parameter spaces and provide a more accurate assessment of *GCR*'s usability.

The subjective nature of privacy protection means that no single approach can universally cater to the diverse needs of all demographic groups. Our participant selection, despite including nine countries, lacked significant cultural diversity. Approximately 83% of participants were recruited from either the U.K. or the U.S., predominantly representing Western culture. Previous research has underscored the cultural influence on privacy preferences [5, 15, 47], which, in turn, might impact satisfaction levels with image editing as our results showed (Section 5.4.3). Therefore, subsequent studies should factor in cultural backgrounds to discern if perceptions of *GCR* differ across cultures.

The images from DIPA [73] were specifically chosen from two large-scale datasets, OpenImages [37] and LVIS [27], and annotated from a perspective of detecting any potential privacy threats to mitigate annotators' privacy concerns about the leakage of their own photos. This raises questions about whether the perceived image editing areas in our study can represent the desired editing areas in real-life image privacy protection practices, even though we filtered out a diverse set of 270 visual content and corresponding images. For example, the rating results might be biased since participants only encountered random people and objects from the DIPA dataset, rather than familiar things from their own lives. Future studies should focus more on human-centered taxonomies of privacy-threatening content from the perspective of online users (e.g., work by Li et al. [43]), and conduct in-the-wild experiments to analyze how individuals engage with *GCR* in actual image-capturing and -sharing situations.

The current version of *GCR* sometimes faces challenges when processing complex scenarios, particularly those involving humans or textual documents (Figure 5). By incorporating specialized models, e.g., those tailored for altering human appearances [62] or creating visual text [75], we can potentially bolster *GCR*'s abilities. An advanced implementation of *GCR* would offer a better showcase of the cutting-edge capabilities of generative models to participants, possibly leading to even more promising experimental outcomes. However, as this paper focused on validating the utility of *GCR* from a human-centered perspective, we leave exploring more advanced *GCR* algorithms to future work.

8 CONCLUSION

In this paper, we introduced a novel method for image privacy protection, named generative content replacement (*GCR*), and verified its effectiveness using human-centered approaches. This innovative method extracts high-level information about privacy-threatening content within images and subsequently generates authentic content to replace the originals, effectively impeding access to privacy-threatening details. We show that *GCR* excels in concealing edit traces (ensuring low detectability) and obfuscates particular privacy-threatening content to be identified as effectively as a 30x downsampling *blurring* method. Furthermore, our participants believed edits processed by *GCR* maintained *narrative coherence* and *visual harmony* of original images, and expressed high *overall satisfaction* with these edits. As image capturing and sharing become increasingly pervasive in our daily lives, we believe that *GCR* will inspire researchers on future innovations in image protection methodologies. Moreover, we expect human-centered investigations for a better understanding of the future establishment of supportive interfaces to enable the general usage of *GCR*.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of the Microsoft Research D-core 2023 program for this study. Their contribution has been

Figure 5: Examples of edited images where the generated content was not as natural as objects in the original ones. The left image displays an unnatural hand, the center one shows a face without eyes and nose, and the right one presents unrecognizable characters in the substitute. In intricate generation tasks, such as replacing parts of the human body, *GCR* sometimes struggled to produce realistic substitutes, pinpointing a clear avenue for future development work.

invaluable in facilitating the research and findings presented in this paper. We sincerely thank Zefan Sramek and Zhongyi Zhou for their precise and insightful suggestions, which greatly enhanced the quality of our research.

REFERENCES

[1] Jemal H Abawajy, Mohd Izuan Hafez Ninggal, and Tutut Herawan. 2016. Privacy preserving social network data publication. *IEEE communications surveys & tutorials* 18, 3 (2016), 1974–1997.

[2] Shady Abu-Hussein and Raja Giryes. 2023. UDPM: Upsampling Diffusion Probabilistic Models. *arXiv:2305.16269* (2023).

[3] Mark S Ackerman, Lorrie Faith Cranor, and Joseph Reagle. 1999. Privacy in e-commerce: examining user scenarios and privacy preferences. In *Proceedings of the ACM Conference on Electronic Commerce*. 1–8.

[4] Adobe. 2023. https://www.adobe.com/products/photoshop/ai.html. Accessed: 2023-9-14.

[5] Mahdi Nasrullah Al-Ameen, Tanjina Tamanna, Swapnil Nandy, MA Manazir Ahsan, Priyank Chandra, and Syed Ishtiaque Ahmed. 2020. We don't give a second thought before providing our information: understanding users' perceptions of information collection by apps in Urban Bangladesh. In *Proceedings of the ACM SIGCAS Conference on Computing and Sustainable Societies*. 32–43.

[6] Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G Baraniuk. 2023. Self-consuming generative models go mad. *arXiv preprint arXiv:2307.01850* (2023).

[7] Rawan Alharbi, Mariam Tolba, Lucia C Petito, Josiah Hester, and Nabil Alshurafa. 2019. To mask or not to mask? balancing privacy with visual confirmation utility in activity-oriented wearable cameras. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 3, 3 (2019), 1–29.

[8] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. 2000. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. 417–424.

[9] Fadi Boutros, Vitomir Struc, Julian Fierrez, and Naser Damer. 2023. Synthetic data for face recognition: Current state and future prospects. *Image and Vision Computing* (2023), 104688.

[10] Kieran Browne, Ben Swift, and Terhi Nurmikko-Fuller. 2020. Camera adversaria. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–9.

[11] ByMatic Broz. 2023. Number of photos (2023): Statistics, facts, & predictions. https://photutorial.com/photos-statistics/. Accessed: 2023-9-14.

[12] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. 2023. Extracting training data from diffusion models. In *32nd USENIX Security Symposium*. 5253–5270.

[13] 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*. Association for Computing Machinery, New York, NY, USA, 75–82. https://doi.org/10.1145/2802083.2808398

[14] Chen Chen, Jie Fu, and Lingjuan Lyu. 2023. A pathway towards responsible ai generated content. *arXiv preprint arXiv:2303.01325* (2023).

[15] Hichang Cho, Milagros Rivera-Sánchez, and Sun Sun Lim. 2009. A multinational study on online privacy: global concerns and local responses. *New media & society* 11, 3 (2009), 395–416.

[16] Julie E Cohen. 2008. Privacy, visibility, transparency, and exposure. *The University of Chicago Law Review* 75, 1 (2008), 181–201.

[17] Daniel Da Silva Costa, Pedro Nuno Moura, and Ana Cristina Bicharra Garcia. 2021. Improving human perception of GAN generated facial image synthesis by filtering the training set considering facial attributes. In *IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 100–106.

[18] Dianne Cyr, Milena Head, Hector Larios, and Bing Pan. 2009. Exploring human images in website design: a multi-method approach. *Management Information Systems Quarterly* (2009), 539–566.

[19] Ralf De Wolf. 2020. Contextualizing how teens manage personal and interpersonal privacy on social media. *New media & society* 22, 6 (2020), 1058–1075.

[20] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* 34 (2021), 8780–8794.

[21] Mariella Dimiccoli, Juan Marín, and Edison Thomaz. 2018. Mitigating bystander privacy concerns in egocentric activity recognition with deep learning and intentional image degradation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–18.

[22] Chengbo Dong, Xinru Chen, Ruohan Hu, Juan Cao, and Xirong Li. 2022. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 3 (2022), 3539–3553.

[23] Benjamin D Douglas, Patrick J Ewell, and Markus Brauer. 2023. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLOS ONE* 18, 3 (2023).

[24] Adám Erdélyi, Tibor Barát, Patrick Valet, Thomas Winkler, and Bernhard Rinner. 2014. Adaptive cartooning for privacy protection in camera networks. In *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance*. IEEE, 44–49.

[25] Ricard L. Fogues, Pradeep K. Murukannaiah, Jose M. Such, and Munindar P. Singh. 2017. Sharing Policies in Multiuser Privacy Scenarios: Incorporating Context, Preferences, and Arguments in Decision Making. *ACM Transactions on Computer-Human Interaction* 24, 1, Article 5 (2017), 29 pages.

[26] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. *arXiv:1406.2661* (2014).

[27] Agrim Gupta, Piotr Dollar, and Ross Girshick. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5356–5364.

[28] Md Rezwan Hasan, Richard Guest, and Farzin Deravi. 2023. Presentation-Level Privacy Protection Techniques for Automated Face Recognition-A Survey. *ACM Computing Surveys* (2023).

[29] 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 CHI Conference on Human Factors in Computing Systems*. 1–13.

[30] 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 CHI Conference on Human Factors in Computing Systems*. 1–13.

[31] 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.

[32] Shengshan Hu, Xiaogeng Liu, Yechao Zhang, Minghui Li, Leo Yu Zhang, Hai Jin, and Libing Wu. 2022. Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15014–15023.

[33] 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.

[34] Azmeri Khan and Glen D Rayner. 2003. Robustness to non-normality of common tests for the many-sample location problem. *Advances in Decision Sciences* 7, 4 (2003), 187–206.

[35] Chandran Kukathas. 2008. Cultural privacy. *The Monist* 91, 1 (2008), 68–80.

[36] Bineeth Kuriakose, Raju Shrestha, and Frode Eika Sandnes. 2022. Tools and technologies for blind and visually impaired navigation support: a review. *IETE Technical Review* 39, 1 (2022), 3–18.

[37] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020. The open images dataset v4. *International Journal of Computer Vision* 128, 7 (2020), 1956–1981.

[38] Andrew Lewis, Patrick Vu, Areeq Chowdhury, et al. 2022. Do content warnings help people spot a deepfake? Evidence from two experiments. (2022).

[39] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv:2301.12597* (2023).

[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*. Springer, 575–586.

[41] Yifang Li and Kelly Caine. 2022. Obfuscation Remedies Harms Arising from Content Flagging of Photos. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–25.

[42] Yao Li and Xinning Gui. 2022. Examining co-owners' privacy consideration in collaborative photo sharing. *Computer Supported Cooperative Work* 31, 1 (2022), 79–109.

[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 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 CHI Conference on Human Factors in Computing Systems* (2017), 1–24.

[45] Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. 2021. When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys* 54, 2 (2021), 1–36.

[46] 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.

[47] Paul Benjamin Lowry, Jinwei Cao, and Andrea Everard. 2011. Privacy concerns versus desire for interpersonal awareness in driving the use of self-disclosure technologies: The case of instant messaging in two cultures. *Journal of Management Information Systems* 27, 4 (2011), 163–200.

[48] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv:2211.01095* (2022).

[49] Michelle Madejski, Maritza Lupe Johnson, and Steven Michael Bellovin. 2011. The failure of online social network privacy settings. (2011).

[50] Juliette Mansour. 2023. AI-generated images can fool people. Here are tips to identify them. https://factcheck.afp.com/doc.afp.com.33BZ68V. Accessed: 2023-9-14.

[51] Reham Ebada Mohamed and Sonia Chiasson. 2018. Online privacy and aging of digital artifacts. In *Fourteenth Symposium on Usable Privacy and Security*. 177–195.

[52] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv:2112.10741* (2021).

[53] José Ramón Padilla-López, Alexandros Andre Chaaraoui, and Francisco Flórez-Revuelta. 2015. Visual privacy protection methods: A survey. *Expert Systems with Applications* 42, 9 (2015), 4177–4195.

[54] Sangkeun Park, Joohyun Kim, Rabeb Mizouni, and Uichin Lee. 2016. Motives and concerns of dashcam video sharing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 4758–4769.

[55] JithendraK Paruchuri, Sen-chingS Cheung, and MichaelW Hail. 2009. Video data hiding for managing privacy information in surveillance systems. *EURASIP Journal on Information Security* (2009), 1–18.

[56] Prolific. 2022. https://www.prolific.co/. Accessed: 2023-5-15.

[57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752

[58] 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.

[59] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22500–22510.

[60] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2020. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4938–4947.

[61] Daniel L Schacter. 1999. The seven sins of memory: Insights from psychology and cognitive neuroscience. *American psychologist* 54, 3 (1999), 182.

[62] Fahad Shamshad, Muzammal Naseer, and Karthik Nandakumar. 2023. CLIP2Protect: Protecting Facial Privacy using Text-Guided Makeup via Adversarial Latent Search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20595–20605.

[63] Tanusree Sharma, Abigale Stangl, Lotus Zhang, Yu-Yun Tseng, Inan Xu, Leah Findlater, Danna Gurari, and Yang Wang. 2023. Disability-First Design and Creation of A Dataset Showing Private Visual Information Collected With People Who Are Blind. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15.

[64] Jiayu Shu, Rui Zheng, and Pan Hui. 2018. Cardea: Context-aware visual privacy protection for photo taking and sharing. In *Proceedings of the 9th ACM Multimedia Systems Conference*. 304–315.

[65] Jose M Such, Joel Porter, Sören Preibusch, and Adam Joinson. 2017. Photo privacy conflicts in social media: A large-scale empirical study. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 3821–3832.

[66] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2149–2159.

[67] Yasuhiro Tanaka, Akihisa Kodate, Yu Ichifuji, and Noboru Sonehara. 2015. Relationship between willingness to share photos and preferred level of photo blurring for privacy protection. In *Proceedings of the ASE BigData & SocialInformatics*. 1–5.

[68] 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 *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1853–1856.

[69] Emanuel von Zezschwitz, Sigrid Ebbinghaus, Heinrich Hussmann, and Alexander De Luca. 2016. You Can't Watch This! Privacy-Respectful Photo Browsing on Smartphones. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 4320–4324.

[70] Kimberley A Wade, Maryanne Garry, J Don Read, and D Stephen Lindsay. 2002. A picture is worth a thousand lies: Using false photographs to create false childhood memories. *Psychonomic bulletin & review* 9, 3 (2002), 597–603.

[71] Xinrui Wang and Jinze Yu. 2020. Learning to cartoonize using white-box cartoon representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8090–8099.

[72] Mika Westerlund. 2019. The emergence of deepfake technology: A review. *Technology innovation management review* 9, 11 (2019).

[73] 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.

[74] 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.

[75] Yukang Yang, Dongnan Gui, Yuhui Yuan, Haisong Ding, Han Hu, and Kai Chen. 2023. GlyphControl: Glyph Conditional Control for Visual Text Generation. *arXiv:2305.18259* (2023).

[76] Jinao Yu, Hanyu Xue, Bo Liu, Yu Wang, Shibing Zhu, and Ming Ding. 2020. Gan-based differential private image privacy protection framework for the internet of multimedia things. *Sensors* 21, 1 (2020), 58.

[77] Lvmin Zhang and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. *arXiv:2302.05543* (2023).

[78] Lingzhi Zhang, Yuqian Zhou, Connelly Barnes, Sohrab Amirghodsi, Zhe Lin, Eli Shechtman, and Jianbo Shi. 2022. Perceptual artifacts localization for inpainting. In *European Conference on Computer Vision*. Springer, 146–164.

[79] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N Metaxas, and Jian Ren. 2023. Sine: Single image editing with text-to-image diffusion models. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 6027–6037.
[80] Ruoyu Zhao, Yushu Zhang, Tao Wang, Wenying Wen, Yong Xiang, and Xiaochun Cao. 2023. Visual Content Privacy Protection: A Survey. *arXiv:2303.16552* (2023).
[81] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.
[82] Ryszard Zieliński. 1990. Robustness of the one-sided Mann—Whitney—Wilcoxon test to dependency between samples. *Statistics & probability letters* 10, 4 (1990), 291–295.