Privacy in Image Datasets: A Case Study on Pregnancy Ultrasounds

Rawisara Lohanimit[1], Yankun Wu[2], Amelia Katirai[3], Yuta Nakashima[2], Noa Garcia[2]

[1]Massachusetts Institute of Technology
[2]The University of Osaka
[3]Tsukuba University
rloha@mit.edu, {yankun@is., n-yuta@, noagarcia@}ids.osaka-u.ac.jp, katirai.amelia.fu@u.tsukuba.ac.jp

Abstract

The rise of generative models has led to increased use of large-scale datasets collected from the internet, often with minimal or no data curation. This raises concerns about the inclusion of sensitive or private information. In this work, we explore the presence of pregnancy ultrasound images, which contain sensitive personal information and are often shared online. Through a systematic examination of LAION-400M dataset using CLIP embedding similarity, we retrieve images containing pregnancy ultrasound and detect thousands of entities of private information such as names and locations. Our findings reveal that multiple images have high-risk information that could enable re-identification or impersonation. We conclude with recommended practices for dataset curation, data privacy, and ethical use of public image datasets.

Introduction

The development of image generation models (Rombach et al. 2022; Saharia et al. 2022; Ramesh et al. 2022) has significantly increased the demand for publicly available visual data. Image datasets (Deng et al. 2009; Lin et al. 2014; Schuhmann et al. 2021) have become essential for training robust computer vision models across multiple tasks, including image generation, image recognition, object detection, and others. However, the process of compiling large-scale image datasets is heavily based on web scraping techniques, collecting images from a wide variety of online sources, such as social media platforms, websites, and public repositories. Although scraping images from the internet allows one to collect massive amounts of data, it typically results in uncurated collections that might not meet quality or ethical standards, causing several critical issues. For example, image datasets can contain harmful societal biases (Birhane, Prabhu, and Kahembwe 2021; Meister et al. 2023; Garcia et al. 2023), such as reinforcing stereotypes based on race, gender, or age. Furthermore, uncurated datasets can also include sensitive or problematic content like child sexual abuse material (Caetano et al. 2025; Thiel 2023), hate speeches (Birhane et al. 2024), or pornography (Birhane and Prabhu 2021), which can lead to misuse of creating harmful outputs, such as synthetic child exploitation content (Thiel 2023; Caetano et al. 2025). Another major concern is the potential exposure of individuals' personal information (Carlini et al. 2023), as some images may contain private or sensitive information that may be used without proper consent.

When computer vision models are trained on such data, there is a significant risk that the undesirable traits, such as biases, harmful content, or privacy violations, are perpetuated or even amplified, potentially placing affected individuals and groups at risk (Katirai et al. 2024). In the context of image generation, recent studies have highlighted that diffusion-based models can memorize training examples, producing exact or near-identical copies of images during the generation process (Carlini et al. 2023). This becomes particularly critical if the training dataset contains personal or sensitive data, as these generative models could potentially reproduce private information, allowing anyone to access it. To address these issues, the field of machine unlearning (Gandikota et al. 2023; Zhang et al. 2024; Lu et al. 2024) proposes algorithmic solutions to steer models away from generating unwanted concepts. However, there is evidence (Suriyakumar et al. 2024) showing that models are not able to fully forget a concept learned during training. To directly address the root causes of these issues, other efforts are focused on auditing datasets (Birhane and Prabhu 2021; Birhane, Prabhu, and Kahembwe 2021; Garcia et al. 2023; Thiel 2023; Birhane et al. 2024). For example, Birhane et al. (2024) audit LAION-400M (Schuhmann et al. 2021) and LAION-2B, two subsets of the full LAION-5B (Schuhmann et al. 2022), revealing an increasing prevalence of hate and toxic content as dataset sizes grow.

Inspired by previous works on dataset auditing, this paper aims to investigate privacy concerns in large-scale image datasets. As privacy in computer vision refers to a broad term that can encompass various types of sensitive information in visual data, ranging from recognizable facial features to medical information, we narrow our focus by investigating pregnancy ultrasound images. Pregnancy ultrasound images, or sonograms, are medical images produced using ultrasound technology to visualize a developing fetus in the womb. They are taken by using high-frequency sound waves to create real-time images of the fetus, placenta, and surrounding tissues with the aim of monitoring fetal growth, detecting any potential abnormalities or health concerns, and

determining important factors such as gestational age. Pregnancy ultrasounds can also provide visual information about the pregnant person's reproductive organs and the overall health of the pregnancy. While previous works have uncovered the existence of misogyny, pornography, malignant stereotypes (Birhane and Prabhu 2021; Birhane, Prabhu, and Kahembwe 2021; Birhane et al. 2024), medical images (Adams et al. 2023) and child sexual abuse material (CSAM) (Thiel 2023) in large uncurated image datasets, the choice of pregnancy ultrasounds as our case study is motivated by the unique characteristics of these images:

- Pregnancy ultrasounds often contain a significant amount of personal information, such as the name of the pregnant person, the hospital where the ultrasound was performed, the date and time of the procedure, the gestational age of the fetus, or the name of the sonographer (the person performing the ultrasound). This personal information can be highly sensitive and potentially identifiable (Leaver and Highfield 2018).
- Pregnancy ultrasounds are frequently shared online or on social media platforms as part of pregnancy announcements to family and friends, increasing the likelihood to be collected in image datasets used to train computer vision models and posing a substantial risk to privacy.
- Pregnancy ultrasounds can reveal potential malformations or health issues in the fetus and the pregnant person, such as developmental anomalies or genetic conditions, which are highly sensitive medical information. If such images are included in a dataset and leaked, they could expose private details about the health and future well-being of both the pregnant person and the fetus.
- The intimate nature of pregnancy ultrasounds makes them highly personal, and their inclusion in training datasets without consent can lead to unintended harm or emotional distress for the individuals involved, especially if these images are used in ways that the individual did not anticipate or approve.
- Ultrasound images are often taken in medical settings where the pregnant person is vulnerable. The improper use or public exposure of such images could cause distress, erode trust in healthcare systems, and further contribute to the broader societal issues on data privacy.

Overall, pregnancy ultrasound images have the potential to contain highly sensitive private information, while at the same time, they are frequently shared and celebrated online, increasing the likelihood to be included in image datasets. Thus, our work seeks to answer the following two key research questions (RQs):

1. Are pregnancy ultrasound images present in image datasets used to train computer vision models?
2. If so, do pregnancy ultrasound images contain private information that can be used to identify individuals?

To address RQ(1), we develop a method for detecting pregnancy ultrasound images based on large-scale image retrieval and classification techniques and apply it to the LAION-400M dataset (Schuhmann et al. 2021). Using this approach, we successfully identify 833 pregnancy ultrasound images, which we further analyze to respond to RQ(2). We run a personal information detection algorithm based on OCR text recognition and Named Entity Recognition on these identified images and uncover four types of personal information: name, location, date time, and phone numbers. While the number of detected images may seem small relative to the 400 million images originally in the dataset, it is important to emphasize that these are real individuals, with real pregnancies, and real lives at stake. The inclusion of such images in the dataset without consent is ethically problematic and must be addressed with great care. Thus, we conclude the paper with several recommendations for large-scale image collections, including the implementation of more robust data privacy and consent protocols, especially for images containing sensitive personal information.

Background

Private Information Definition

There are varied approaches and concepts related to private information and the question of how to define privacy and what should be considered to be private is often contentious, even among scholars in the area. Legal definitions of privacy vary depending on regional or cultural context, as evidenced, for example, by the overlapping but distinct definitions of personally identifiable information in the United States,[1] and personal or sensitive data in the European Union.[2] As online ultrasound sharing has been reported in the academic literature across multiple regional and cultural contexts (Zhu et al. 2019; Roberts et al. 2015), and working as an interdisciplinary and international team, we avoid a regionally-bound legal definition of privacy and instead use the broader term "private information" and an approach informed by recent academic debates on privacy.

Many key privacy scholars argue for an anti-reductionist, pluralistic understanding of privacy which recognizes that privacy is interlinked with too many aspects, including specific circumstances or contexts, to be able to be reduced to a single definition (Trepte and Masur 2023). Moreover, as Solove (2023) has argued in relation to sensitive data, "the borderlines of many categories are so blurry that they are useless," and that, ultimately, "data is what data does".

Acknowledging these debates and varied concepts, we utilize Nissenbaum's theory of contextual integrity here, (Nissenbaum 2004; Malkin 2022), which highlights the importance of norms and expectations for how information will be used within particular contexts, and the resulting privacy violations that occur when these expectations are breached and data is utilized outside of these expected ways. Thus, *data which may not be too sensitive to share in one context, may become so when that data is removed from its original context and utilized for another purpose, as in the case of data scraped for training.* This is especially relevant in the area of pregnancy ultrasounds, where an ultrasound image

[1] https://www.dol.gov/general/ppii

[2] https://commission.europa.eu/law/law-topic/data-protection/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data/what-personal-data-considered-sensitive_en

may be shared to mark a particular social and medical milestone on an online platform where the poster expects that friends and family or other interested people may be able to view it, but a breach of contextual integrity occurs when this data is extracted from this intended context and used as a part of training datasets.

It is also noteworthy that unintended access or use to health-related data such as ultrasounds can create risks related to insurance or identity theft, and are frequently highly sensitive.[3][4] Pregnancy ultrasounds in particular can reveal sensitive information about parental or fetal health and healthcare access, and so usage of the images beyond their original intended scope can be highly problematic.[5]

Privacy Information in Images and in Datasets

With the growing concerns about the presence of private information in images, researchers have developed privacy detection methods to determine whether an image contains private information (Tonge and Caragea 2020; Tran et al. 2016; Tonge and Caragea 2016). Typical approaches extract features using a trained model and use them to train a classifier that predicts the presence of private information (Xu et al. 2024; Zhao et al. 2022). A more complex task involves identifying specific areas within an image that contain private information (Gurari et al. 2019).

The focus of image privacy often revolves around bioidentifiers, such as facial attributes (Zhang et al. 2014; Kumawat and Nagahara 2022; Carlini et al. 2023). For example, Zhang et al. (2014) proposed an anonymous camera that uses optical masking to obscure captured faces, thereby improving privacy during image collection. A recent study (Samson et al. 2024) finetuned a visual language model to improve the private information detection ability using a dataset constructed from LAION-5B. While this improved detection capabilities, it also raised concerns about the potential presence of private information in large-scale image datasets and the risks they pose.

Despite these advancements, the issue of private information presence within image datasets remains insufficiently explored compared to other types of datasets. For example, Li, Yang, and Lu (2024) evaluated the privacy and re-identification risks of open government data in China. In addition, millions of instances of private data, such as email addresses, phone numbers, IP addresses, credit card numbers, bank account numbers, and names, have been found across large text corpora (Subramani et al. 2023; Elazar et al. 2024; Jahan et al. 2023) such as the Colossal Clean Crawled Corpus (Raffel et al. 2020) and the Pile (Gao et al. 2020). In contrast, despite progress in studying other issues in image datasets, such as toxic content (Birhane et al. 2024), copyright concerns (Ma et al. 2024; Moayeri et al. 2024), metadata anonymization (Rempe et al. 2024; Jahan et al. 2023), and demographic bias (Garcia et al. 2023; Meister et al. 2023), privacy risks have not been widely evaluated. This gap creates potential risks of private information leakage (Hu et al. 2024; Zhang and Li 2022). Unlike medical datasets, which often implement privacy-preserving measures to remove sensitive data before publication (Zhang and Li 2022), such practices are rarely applied to image datasets from other domains (Deng et al. 2009; Lin et al. 2014), leaving them vulnerable to the exposure of private information. Co-current work by (Hong et al. 2025) investigates general privacy concerns in a public image dataset, DataComp CommonPool (Gadre et al. 2023), with a primary focus on the legal implications and the effect of data filtering. In contrast, our work studies pregnancy ultrasound images, analyzing subtypes and co-occurring private information that heighten re-identification risks.

The Pregnancy Ultrasound Image

Next, we briefly review prior research on the significance of ultrasound images, finding that their significance is not only medical, but also social, and is shaped by culture (Roberts et al. 2015).

First, pregnancy ultrasounds carry medical significance as they are perceived to be a part of appropriate antenatal care, and as necessary for a successful pregnancy (Øyen and Aune 2016). This is in part due to their use in diagnosing congenital conditions. However, there are also concerns over the role of ultrasounds in the medicalization of pregnancy, through which the pregnant person increasingly comes to be viewed as a patient, and pregnancy as a dangerous state, which must be controlled and monitored through technology. In addition, ultrasounds carry social significance as one method through which both parental and fetal identities are developed. Though there is a belief that ultrasounds contribute to parental bonding with a fetus (Skelton et al. 2024), this is contested in recent research. It is also noteworthy that the cultural weight of ultrasound images has been critically examined by feminist scholars, who argue that ultrasound images and their role in identity construction constrain the right of the pregnant person to choose (Lie et al. 2019).

The social significance is related to cultural pressures, including pressures to save the images "for posterity," and to share them with others, which increasingly occurs through online platforms (Harpel 2018). As such, the ultrasound has been described as part of an "online birth" taking place prior to the "physical birth" (Johnson 2014) as it becomes part of a digital footprint which extends across the life-course. Such images are primarily shared with family or friends, as one study of 117 pregnant women conducted through Facebook found that 77.6 percent shared only with friends, and 78.4 with family (Harpel 2018). Similarly, a study of Chinese expectant mothers found that pregnancy was perceived to be a time of vulnerability (Zhu et al. 2019) and thus news regarding it was shared primarily within closed groups of family or friends. However, they may also be shared more broadly, as Leaver and Highfield (2018) report the presence of ultrasounds publicly shared through Instagram . Further, their study identified privacy concerns including the presence of what the authors termed personally identifiable metadata.

[3]https://www.propublica.org/article/millions-of-americans-medical-images-and-data-are-available-on-the-internet

[4]https://techcrunch.com/2020/01/10/medical-images-exposed-pacs/

[5]https://www.curtin.edu.au/news/think-before-you-post-the-impact-of-sharing-photos-of-your-child-online/

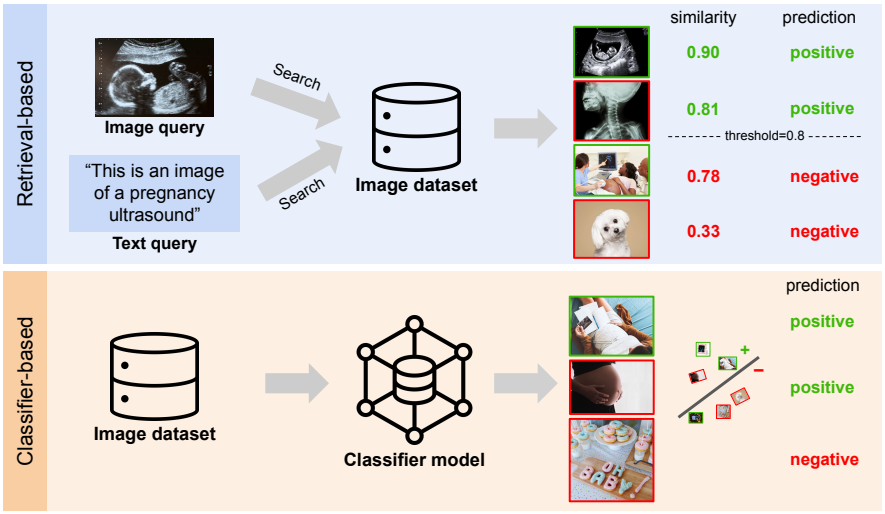

Figure 1: Pregnancy ultrasound image detection. For retrieval-based detection, we use image and text as queries to find images that have high similarity with the query. For classifier-based detection, we use the output of the classifier as the prediction.

These issues are then further amplified when such images are scraped and used for purposes unanticipated by the original sharer.

Methodology

To address our research questions, RQ(1) and RQ(2), we develop a methodology that aims to: 1) detect pregnancy ultrasound images within large collections of images, and 2) identify private information in pregnant ultrasound images.

Pregnancy Ultrasound Image Detection

Given a dataset of images $\mathcal{D} = \{I_1, I_2, \cdots, I_N\}$ where I_j represents the j-th image in a dataset of size N, the goal is to identify the subset $\mathcal{U} \subset \mathcal{D}$ such that images in $\mathcal{U}$ contain pregnancy ultrasound images. To achieve this, we follow two approaches: a retrieval-based approach and a classifier-based approach. An overview is shown in Figure 1.

In both approaches, images in $\mathcal{D}$ are mapped to a semantic space with a pre-trained CLIP image encoder. The retrieval-based approach uses a query, which can be either a text description or an example ultrasound image, to compute the similarity between images in $\mathcal{D}$ and the queries. An image in $\mathcal{D}$ is identified as an ultrasound image when the similarity is sufficiently large. In contrast, the classifier-based approach uses a classifier trained on a training dataset containing both pregnancy images as positive samples and other images as negative samples.

Retrieval-Based Approach This approach uses a set $\mathcal{Q}_\text{t} = \{Q_\text{t}\}$ of text description queries Q_t or a set $\mathcal{Q}_\text{i} = \{Q_\text{i}\}$ of image queries Q_i that exemplify ultrasound images. We use pre-trained CLIP text and image encoders[6] to map the queries and images in $\mathcal{D}$ into the semantic space. Let CLIP_image and CLIP_text denote the CLIP image and text encoders, respectively. We compute image features $x_j = \text{CLIP}_\text{image}(I_j)$ for $I_j \in \mathcal{D}$, as well as text features $q_\text{t} = \text{CLIP}_\text{text}(Q_\text{t})$ for $Q_\text{t} \in \mathcal{Q}_\text{t}$ and image features $q_\text{i} = \text{CLIP}_\text{image}(Q_\text{i})$ for $Q_\text{i} \in \mathcal{Q}_\text{i}$. Some example text queries $\mathcal{Q}_\text{t}$ are: *Ultrasound imaging for prenatal screening*, *Pregnancy announcement photoshoot*, and *Prenatal ultrasound for comprehensive fetal anomaly screening*, which cover different types of possible pregnancy ultrasound images.

We use $|\mathcal{Q}_\text{i}| = 22$ images and $|\mathcal{Q}_\text{t}| = 100$ sentences as queries. We compute cosine similarity between the representations of queries and images in $\mathcal{D}$. For each image in $\mathcal{D}$, we identify the highest similarity among all similarities computed with the queries. If the highest similarity exceeds the threshold τ, the image is retrieved as a pregnancy ultrasound image.

Classifier-Based Approach This approach classifies whether I_j is a pregnancy ultrasound image using CLIP image features x_j. We compare three different methods: random forest (RF), support vector machine (SVM), and neural network (NN). We use scikit-learn (Pedregosa et al. 2011) implementation of RF and SVM classifiers, while the NN model is implemented with PyTorch (Paszke et al. 2019). The parameters of the RF and SVM models are optimized using grid search. The NN model consists of four layers with dimensionalities $1,024$, 256, 32, and 1, respectively, and ReLU activations (Agarap 2018). The binary cross entropy loss function is used with Adam optimizer (Kingma 2014), a learning rate of 0.0001, and a batch size of $1,024$.

The PIU Dataset To train the parameters of the classifiers, we collect the PIU (Pregnancy Image Ultrasound) dataset. PIU is a custom-made dataset with both positive and negative examples of pregnancy ultrasound images collected by using the easy-image-scraping library (Naumann et al. 2022). Due to the emotional significance of pregnancy ultrasound images for expectant parents, these images appear in various forms when shared online. Common presenta-

[6]https://github.com/rom1504/clip-retrieval

(a) pregnancy ultrasound images

(b) non-pregnancy ultrasound images

Figure 2: Examples of (a) positive images and (b) negative images in the PIU dataset. Faces and private information redacted for privacy.

tions include plain digital ultrasound images, photos of ultrasounds with decorative backgrounds, ultrasounds incorporated into event invitations such as baby showers, and family photos featuring members holding ultrasound images. To account for all this variability and get a sufficient representation of ultrasound images for training our classifiers, we query images using keywords such as *pregnancy ultrasounds*, *baby shower with ultrasounds*, and more. For negative examples, on top of collecting any non-related instances such as *animals*, *locations*, or *household objects*, we especially look for images under the concepts of *non-pregnancy ultrasound*, which tend to be the hardest cases to classify.

We split a preliminary collection of images into train, validation, and test splits. Using an active learning approach, we iteratively refine the training set while keeping the validation and test sets fixed. In each iteration, we train a model, evaluate it on the validation dataset, and identify misclassified or challenging images. Then, we collect similar images to add to the training set. To prevent data leakage, we remove images with CLIP similarity above 0.95 to those in the validation or test sets both before initiating active learning and during each iteration, followed by manual inspection. The model is retrained with the updated train split. We repeat this process until the validation performance stabilizes, allowing the model to adapt to diverse data in large image datasets without overfitting. The final dataset contains 9,900 images, evenly balanced between positive and negative, comprising 3,960 training, 990 validation, and 990 test images. Figure 2 shows example images.

Private Information Identification

After detecting pregnancy ultrasound images in the original dataset $\mathcal{D}$, the next step is to search for private information within them. We develop an approach (see Figure 3) that detects and reads text in images in three steps: image preprocessing, text recognition, and private information extraction.

Image Preprocessing Since pregnancy ultrasound images can appear in various formats and visual presentations, and due to the uncurated nature of large image datasets, text in the images may not always be clearly readable by algorithms. Challenges such as handwritten text, image tilt, and low resolution can hinder text detection. To address these issues, we apply image preprocessing. First, we upscale low-resolution images (i.e. one of the dimensions is less than 200 pixels) by a factor of four using the Real-ESRGAN super-resolution model (Wang et al.). Additionally, we augment each image by rotating it by 5 to 90 degrees, both clockwise and counterclockwise. These image preprocessing steps improve text detectability in subsequent stages.

Text Recognition Next, we employ an open-source optical character recognition (OCR) model called Tesseract (Kay 2007) to read text in images. Tesseract OCR is an LSTM-based model (Hochreiter and Schmidhuber 1997) that recognizes line and character patterns in images. However, despite the image preprocessing step, the initial OCR outputs are not always accurate. To refine the quality of the recognized text, we employ the LLaVa-Next model (Liu et al. 2023b,a), a general-purpose visual and language understanding model. LLaVa-Next can correct OCR errors by conditioning on the image with the prompt: *Correct the following OCR extracted text: [recognized text].*

Private Information Extraction From the recognized text in the previous step, we utilize Presidio (Mendels et al. 2018) to identify private information. Presidio is a software development kit capable of identifying and anonymizing private entities in text, such as credit card numbers, names, locations, and more. It leverages natural language processing techniques, including regular expressions for pattern matching, Named Entity Recognition to detect entities, and rule-based logic and checksum with relevant context. The software has both predefined and customizable recognizers to detect sensitive entities. Following Subramani et al. (2023), we consider various types of private information and narrow them down to four categories — Name, Location, Date Time, and Phone Number — based on their frequent occurrence in the images.[7]

Private Information in LAION-400M

Following the methodology described above, we examine the LAION-400M dataset (Schuhmann et al. 2021) in search for private information, specifically private information in

[7]Detailed definitions of types and recognizers can be found in the appendix.

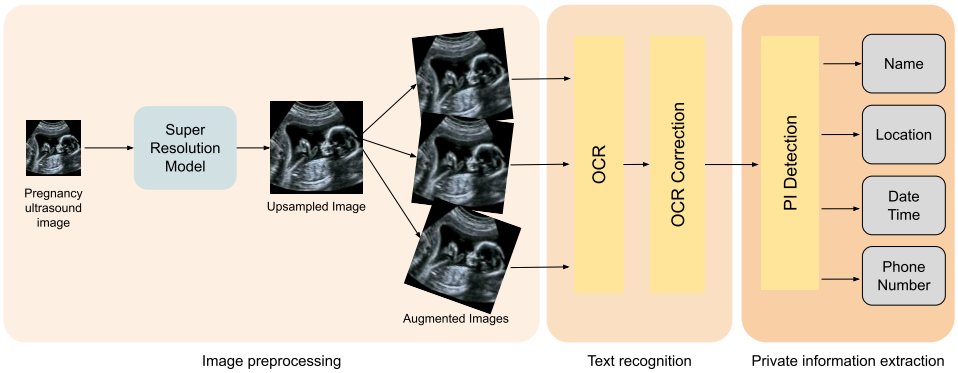

Figure 3: Private information identification. Detected pregnancy ultrasound images are 1) preprocessed with super-resolution and rotation for horizontal text alignment, 2) processed for text recognition and correction, and 3) passed to a private information detection system to extract Name, Location, Date Time, and Phone Number entities.

pregnancy ultrasound images. LAION-400M is a dataset with 400 million image-text pairs in English extracted from the Common Crawl web data dump between 2014 and 2021 and filtered with CLIP (Radford et al. 2021) to remove non-matching semantic text-image data. We select this dataset due to its public availability and its widespread adoption in training vision-language models, such as the popular Stable Diffusion (Rombach et al. 2022). We use the precomputed CLIP image embeddings as image representations. Our findings are presented below.

Pregnancy Ultrasound Images in LAION-400M

Search Method Evaluation First, we evaluate the detection performance of the two approaches, i.e. the retrieval-based detection and the classifier-based detection, on the PIU dataset. For the retrieval-based approach, we select an optimal threshold on the PIU validation set, and then apply the chosen threshold to compute results on the test set. When images are used as queries (i.e., image-to-image retrieval), the threshold is set at $\tau = 0.7$, while for text queries (i.e., text-to-image retrieval), $\tau = 0.3$.

The results are presented in Table 1. With retrieval-based detection, despite achieving high similarity scores, the accuracy in detecting pregnancy ultrasound images is relatively low, particularly when using images as references. This could be attributed to the high-level semantics in the CLIP embedding space. For example, images of babies or pregnant individuals have a high similarity score with pregnancy ultrasound images in the CLIP embedding space, leading to the lower accuracy of the retrieval of *pregnancy ultrasound images* specifically and reducing retrieval accuracy. This indicates that more specialized models are needed to accurately identify pregnancy ultrasound images. In contrast, the classifier-based detection, which trains a dedicated classifier on top of the CLIP embeddings using PIU training images, obtains much better performance. From the three proposed classifiers, the SVM stands out for its high accuracy and low false positive rate. This is particularly important given the large volume of data; even a small false positive rate can lead to a significant number of irrelevant results that require manual inspection when applied to a large dataset. The SVM's false negative rate is also comparable to the best performing classifier, which is crucial for ensuring that target images are not missed. Therefore, for the rest of the paper, we present results using the SVM classifier.

Detection approach	Method	Acc	FP rate	FN rate
Retrieval-based	Image	77.17	10.55	34.75
	Text	83.03	8.52	25.05
Classifier-based	RF	94.64	5.05	5.65
	SVM	**97.27**	**1.41**	4.04
	NN	96.67	3.23	**3.43**

Table 1: Pregnancy ultrasound image detection performance on the PIU test set.

Detected Pregnancy Ultrasound Images in LAION-400M Using the trained SVM model on the whole LAION-400M dataset, we identify $1,364$ pregnancy ultrasound images. To remove duplicates, we use a copy detection method SSCD (Pizzi et al. 2022) with a similarity threshold of 0.92, and keep one image per duplicate set. This process reduces the image count to $1,045$. We subsequently apply manual inspection, further refining it to 833 unique images, highlighting the significant presence of duplicate data within the dataset, as previously noted by (Touvron et al. 2023).

False Positive Rate To validate the efficiency of our model on the target dataset, we inspect potential false negative images. As finding actual false negatives among 400 million unlabeled images is highly challenging, we conducted a manual inspection of images near the classifier's decision boundary. Specifically, we applied the SVM classifier to a partition of the LAION-400M dataset (i.e. 1 million images) and manually reviewed images whose negative prediction scores fell within 2 standard deviations (S.D.) of the decision boundary. We selected this threshold because only

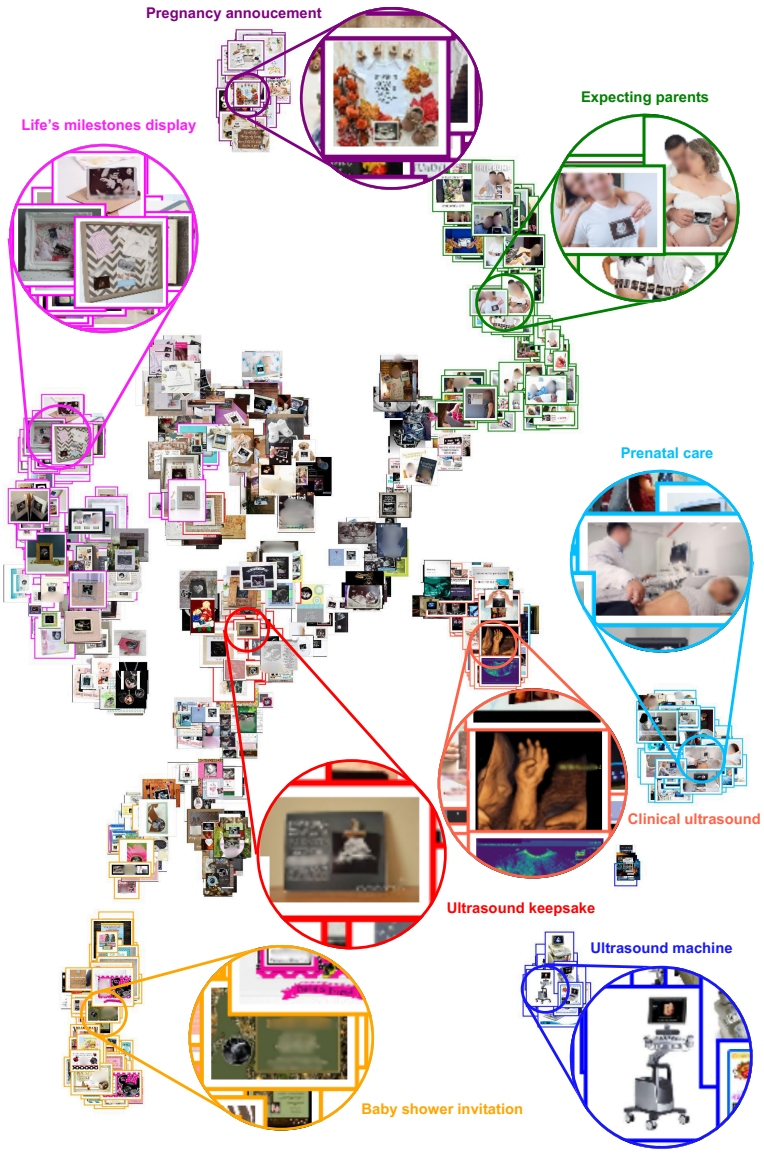

Figure 4: t-SNE visualization of the pregnancy ultrasound images found in LAION-400M. Colors represent each of the cluster themes (names shown next to each cluster) found with HDBSCAN. Faces and private information redacted for privacy.

38 images were within 1 S.D., making it insufficient for a meaningful analysis. Within the 2 S.D. range, we identified 1,872 images, none of which were pregnancy ultrasound images.

Data Visualization and Clustering We further analyze the different types of detected pregnancy ultrasound images using data visualization tools. First, we transform the CLIP image representation of each image from the original 512-dimensional space into a 5-dimensional vector using UMAP (McInnes, Healy, and Melville 2020). We then apply HDBSCAN (McInnes et al. 2017) for clustering, with a minimum cluster size of 20 images. This approach reveals 8 distinct clusters. In contrast, when HDBSCAN is applied directly to the original CLIP image representation without UMAP dimensionality reduction, no clusters are detected, highlighting the necessity of reducing the dimensionality to reduce information from the 512-dimensional space, which may otherwise introduce distracting noise.

Next, we visualize the 833 pregnancy ultrasound image space in two dimensions by reducing the image features to 2D using the t-SNE algorithm (van der Maaten and Hinton 2008). Figure 4 shows the spatial arrangement of the clusters, revealing varying degrees of semantic relatedness, with some outliers such as images of ultrasound machines and pregnancy announcement photos. Despite efforts to eliminate duplicates, some similar images remain due to resolution differences, which caused the image representations to be less semantically aligned.

ID	Theme	Most frequent words	Num. images
0	Ultrasound machine	ultrasound, doppler, color, scanner, machine	33
1	Prenatal care	ultrasound, pregnant, woman, stock, doctor	55
2	Expecting parents	pregnant, pregnancy, stock, baby, ultrasound	120
3	Clinical ultrasound	ultrasound, fetal, fetus, pregnancy, abortion	33
4	Pregnancy announcement	announcement, pregnancy, baby, social, media	40
5	Baby shower invitation	baby, shower, invitations, invitation, photo	67
6	Ultrasound keepsake	frame, photo, baby, picture, sonogram	27
7	Life's milestones display	baby, frame, photo, box, shower	82
-1	Not in any cluster	-	376
All	-	-	833

Table 2: Cluster analysis of pregnancy ultrasound images detected in the LAION-400M dataset. Each image is assigned to a cluster based on HDBSCAN results. To characterize the themes of each cluster, we extract the top 5 most frequent words from the captions of the images within each cluster and use them to assign theme names.

Pregnancy Ultrasound Images Themes When analyzing each of the 8 detected clusters in detail, we observe unique themes across them. To effectively label these themes, we compute word frequency counts from the captions associated with the images in each cluster, as the captions are semantically aligned with the images they describe. The top 5 most frequent words for each cluster are used to manually set cluster names that capture the predominant themes in the images. Additionally, clusters are color-coded in the t-SNE representation map shown in Figure 4, providing an intuitive visualization of the types of pregnancy ultrasound images found in the LAION-400M dataset. A detailed breakdown, including the themes, the top 5 most common words, and the number of images per cluster, is provided in Table 2.

While our initial expectation of the dataset's ultrasound images leaned towards the stereotypical medical film format, black and white with technical details like date, time, and hospital information, our analysis reveals a broader variety. Although these ideal information-rich images are still present in the dataset, they are a minority. Upon visual and semantic examination, none of the clusters distinctly represents this stereotypical format. The most prevalent category is *expecting parents* (ID 3), which accounts for 14.4% of the images, depicting women or couples holding ultrasound images. This aligns with real-world scenarios where expectant parents share their pregnancy journey online, using these images for announcements and to foster social engagement (Harpel 2018; Roberts et al. 2015). Furthermore, *life's milestones display* (ID 7) represents 9.8% of the images, emphasizing their sentimental value as keepsakes. Another notable category is *baby shower invitation* (ID 5) accounting for 8% of the images. While the actual ultrasound images are smaller, these images contain significant private information relevant to event details such as names, addresses, phone numbers, and dates. This presents a potential risk regarding data privacy as the inclusion of such detailed private information in publicly shared or inadequately secured images can lead to unauthorized access and misuse of this data.

The majority of images did not belong to any cluster, potentially due to their unique characteristics or insufficient numbers to form a separate cluster. These images include various items such as pendants, decorations, and book covers. Although some of these images bear a superficial resemblance to existing clusters, their distinct visual features prevent them from being grouped with other more cohesive clusters. Moreover, another observation is the low density of these images in Figure 4, suggesting sparse representation in the feature space.

	Num. instances	
Type	All images	Unique images
Name	387	228
Location	238	120
Date Time	513	299
Phone Number	55	30
Total	1,193	677

Table 3: Summary of private information instances extracted from pregnancy ultrasound images in LAION-400M.

Type	Precision	Recall	F1 score
Name	0.62	0.35	0.45
Location	0.50	0.67	0.57
Date Time	0.70	0.60	0.65
Phone Number	0.58	0.70	0.63

Table 4: Private information detection method evaluation as precision, recall, and F1 score.

Private Information in Pregnancy Ultrasound

Types of Private Information With the detected pregnancy ultrasound images in the LAION-400M dataset, we apply the methodology described previously to analyze the content and search for private information. Running Presidio on the 833 detected images, we find a total of 677 instances of private information. Table 3 reports results both with and without duplicates (i.e., *all images* and *unique images*, respectively) to show how the raw, unmodified data behaves. A large number of the detected instances are in the Date Time category, which is logical considering that ultrasound images typically include timestamps to document the exact time of the medical examination. Other prevalent private information categories detected in the pregnancy ultrasound images are Name and Location.

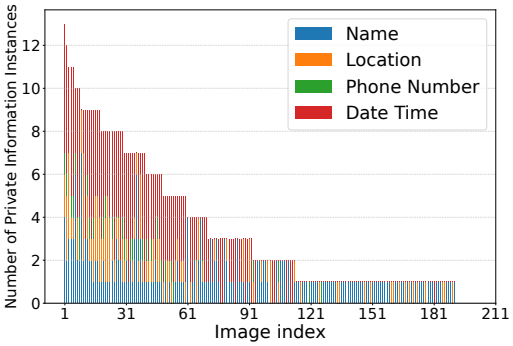

Figure 5: Number of private information detected in each pregnancy ultrasound within LAION-400M.

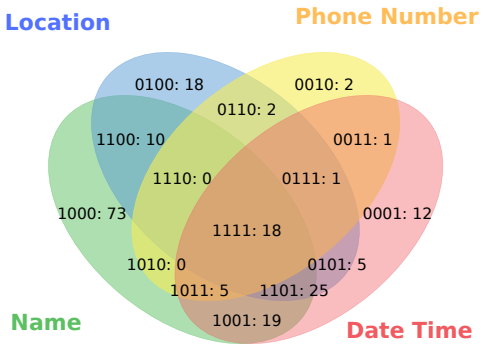

Figure 6: Overlaps of private information types in the LAION-400M dataset. The four-digit binary code represents Name, Location, Phone Number, and Date Time, where "1" indicates presence and "0" indicates absence.

Private Information Identification Method Evaluation Relying solely on these counts provides limited insight, as they lack context regarding the accuracy and reliability of the method, as well as the potential errors that may arise across different components or modules. To address this, we conduct a manual evaluation on a randomly selected subset of 200 pregnancy ultrasound images detected in the LAION-400M dataset. Two of the authors of the paper served as annotators to examine each of the 200 images, classifying the instances of private information within these images into four categories following the same definition as Presidio. Each annotator reviewed the same set of 200 images, allowing for cross-checking and discussion to resolve uncertainties. Eventually, the annotations from each annotator are consolidated to create a unified ground truth.

We report precision, recall, and F1 score in Table 4 by comparing our private information identification method's detected strings to the ground truth from annotators, allowing for a margin of error to account for variations in string matching. Specifically, we consider two strings matched if their Levenshtein distance is less than 2 or the similarity is greater than 70%. The Levenshtein distance measures how many single-character edits are needed to transform one word into another, which indicates the distance between two sequences. The similarity is calculated by taking twice the length of the longest common subsequence between two strings and dividing by the total number of characters in both strings. This flexibility was essential given the challenges of OCR on images with varying quality, which often resulted in incomplete or inaccurate text extraction.

The results in Table 4 show that our method performs well in detecting Phone Number and Date Time information, as these primarily consist of numerical values and formatted strings, making them less susceptible to errors introduced by OCR. Conversely, recognizing Location and Name is more challenging. Manual inspection reveals that the framework exhibits difficulties in accurately extracting text associated with addresses. This is likely attributed to poor image resolution or distorted text rendering within the images. Similarly, off-the-shelf name recognition models displayed inconsistencies. For instance, the model could recognize *Chole* but failed to identify *Jessica*, despite both being common English names. This implies that the model failed to recognize a significant number of actual Name instances. The low F1 scores suggest that the 228 detected Name instances from Table 3 may not accurately represent the true number of Name instances in the dataset, with potentially more instances that have not been detected. To address these inaccuracies, instead of using a predefined recognizer, a more accurate and customized name recognizer may be necessary.

Linked Instances of Private Information As discussed above, we detect various types of private information in the pregnancy ultrasound images in LAION-400M, often linking multiple details to specific individuals or events, elevating the potential risk of identity exposure. Some combinations of identifiers can be used to uniquely identify people, especially the mix of gender, birth date, and postal code, which can identify 87% of individuals in the United States (Sweeney 2000). Our findings from Table 2 highlight that one of the largest clusters represents *baby shower invitation cards*, which typically contain comprehensive details like names, locations, dates, and times. This is problematic because it aggregates and displays private data in a publicly accessible format, increasing the potential for misuse or misappropriation of private information and making individuals susceptible to privacy breaches and identity theft.

We analyze the linked instances of detected private information in Figures 5 and 6. Figure 5 shows the distribution of the number of private information instances detected per pregnancy ultrasound image. While the majority of the images have minimal or no private information, 86 images contain more than one type of sensitive information, with some images containing up to 13 distinct instances of private information. More specifically, Figure 6 shows the intersections between the different types of private information per image. The results show that 22.9% (191 images) contain at least one type of private information. Of these, 10.3% (86 images) have more than one type of private information, and a smaller subset, 2.1% (18 images), include all four types of private information. The most common overlap, occurring in 3% (25 images), involves the combination of Name, Location, and Date Time information.

Recommendations

Our work focuses on detecting private information in pregnancy ultrasound images within the LAION-400M dataset, representing only a fraction of the potential private data it may contain. There are other image types that could have private information, which we did not explore in this project. However, the focus on pregnancy ultrasound images is justified by their significant real-world privacy implications and the sensitivity of reproductive health data. Contrary to the view that this focus is narrow, we emphasize that half the population may undergo such ultrasounds. When including the fetus in these images, the privacy concern extends to virtually everyone at that stage of life, underscoring its broad relevance. Below, we outline several recommendations based on our findings.

Detection and De-Identification of Private Information When creating a new dataset or utilizing a new dataset, it is crucial to make sure that individuals' privacy is preserved because the data will be used to train a model that might be vulnerable to membership attack, leaking the training data and sensitive information contained within the data (Carlini et al. 2021, 2023). To address this similar issue, Yu et al. (2021) applied neural network to detect private information and generative adversarial models to de-identify sensitive information such as facial features and car license plates on images collected from multimedia recording devices. This approach not only protects individuals' privacy, but also maintains the utility of the images. In similar manner, dataset creators or users should employ various methods, such as Presidio (Mendels et al. 2018) or neural networks, to detect and de-identify private information. Developing systematic automated tool that can perform these tasks is essential to establish privacy-preserving dataset.

Practice Consent Consent is the best tool for limiting excessive collection of personal data and respecting individual autonomy (Froomkin 2019). Informed consent is necessary to protect individuals and uphold their rights. Without it, people often worry their information might be sold or analyzed by AI for marketing purposes, leading to feelings of rights violations (Andreotta, Kirkham, and Rizzi 2021). The internet offers a vast pool of data, but collecting data from it may violate people's rights. We should strive to collect datasets that respect individual rights by ensuring people are well-informed and offering them the option to opt-out, rather than assuming consent by default, especially in machine learning where data is used in diverse applications. Obtaining consent from both data providers and owners fosters ethical and transparent practices within the field and more effectively aligns their goals and expectations.

Privacy-Preserving Training Given our findings on private information in image datasets that may be memorized by models (Ju et al. 2025), we recommend robust privacy-preserving methods during training in addition to preprocessing datasets. Differential Privacy (DP) protects individual data by adding noise to query results or representations while preserving statistical utility (Dwork 2006). It has been applied across different domains, such as facial recognition (Chamikara et al. 2020), image generation (Yu et al. 2021), and medical images classification (Wu et al. 2019). However, our discovery of 531 duplicate images (38.9% of retrieved data) challenges DP's core assumption that adjacent datasets differ by only one record. Duplicates allow an individual's data to persist even after removal, undermining DP guarantees and highlighting the critical role of careful data collection in privacy-preserving systems.

Beyond DP, other methods include a distributed selective SGD framework, where multiple models collaboratively learn from their datasets without directly sharing them (Shokri and Shmatikov 2015). Moreover, the concept of knowledge transfer through a teacher-student model, where the teacher model is trained on disjoint data and the student model learns from the teacher's noisy aggregated responses, illustrated innovative strategies to protect privacy during the training phase (Papernot et al. 2016, 2018; Liu et al. 2020). By integrating these privacy-preserving techniques, we can improve the confidentiality of the data while maintaining the integrity and utility of the training process.

Further Research Applying Nissenbaum's theory of contextual integrity, we bring attention to the issues that arise when data that is shared in one context and with a certain intention is then scraped and used without consent in a domain and for purposes other than those originally intended. At the same time, the contextual nature of what should be considered private makes it difficult to create clear guarantees that privacy will be respected. In the case of pregnancy ultrasounds, their dual medical and social functions make them especially sensitive data, and their out-of-context use and the privacy issues which result from it are of particular concern. There is an urgent need for attention to issues at the intersection of reproductive health, privacy, and data management. We aim to draw needed attention to these issues and to stimulate social and academic debate to clarify the relevant norms and build consensus around the appropriate handling of such data.

Conclusion

We explored the presence of private information within the LAION-400M dataset, focusing specifically on the retrieval of pregnancy ultrasound images. We employed both retrieval-based and classifier-based approaches to identify relevant images within the dataset. The images we found are not only typical clinical ultrasound images but also included baby shower invitations, pregnancy announcements, images of ultrasound machines, and ultrasound photos within frames. These findings underscore the diverse contexts in which ultrasound images are shared, reflecting real-world uses. We found presence of private information in those images, with many instances where multiple types of personal data co-occur within a single image, which increases the risk of identity exposure.

We believe that this work represents a significant step toward privacy in image datasets, shedding light on the prevalence of embedded private information and highlighting the importance of open-source data auditing for a safe and responsible machine learning community.

Ethical Statement

We recognize the ethical implications inherent in our research. We are committed to upholding the privacy of individuals, particularly in cases where data may be collected inadvertently from the internet without explicit consent. Our project involves the creation of a dataset that includes pregnancy ultrasound images, which contain highly sensitive personal information. To protect these data, we will take the following measures: the dataset will not be shared publicly, and all images and corresponding trained models will be securely deleted upon the completion of the project. We take these steps to ensure the highest respect for individual privacy and data integrity.

Adverse Impact Statement

While the main goal of our work is to identify and highlight the presence of private information in publicly available image datasets, we are aware of the dual-use risk. Specifically, malicious actors could adapt the proposed methodology to extract private information from other datasets and use it to harm or threaten individuals through identity theft, harassment, targeted misinformation, or other violations of privacy (King and Meinhardt 2024). These risks point to a serious issue: many large datasets contain personal and identifiable information, often collected through data scraping without proper protection or consent (Birhane, Prabhu, and Kahembwe 2021).

We hope our work draws attention to these concerns and encourages the research community to approach this space with caution and to develop ethical frameworks, auditing mechanisms, and clear guidelines for dataset collection, release, and downstream use. Future work might also explore technical safeguards to help mitigate potential harms from existing datasets.

Acknowledgements

This work was supported by JSPS KAKENHI No. JP23H00497 and JP22K12091, JST CREST Grant No. JPMJCR20D3, JST FOREST Grant No. JPMJFR216O, and The University of Osaka IDS Co-Creation Project Na22990007.

References

Adams, L. C.; Busch, F.; Truhn, D.; Makowski, M. R.; Aerts, H. J.; and Bressem, K. K. 2023. What does DALL-E 2 know about radiology? *Journal of Medical Internet Research*, 25: e43110.

Agarap, A. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

Andreotta, A. J.; Kirkham, N.; and Rizzi, M. 2021. Ai, Big Data, and the future of consent. *AI & SOCIETY*, 37(4): 1715–1728.

Birhane, A.; Han, S.; Boddeti, V.; Luccioni, S.; et al. 2024. Into the laion's den: Investigating hate in multimodal datasets. *NeurIPS*.

Birhane, A.; and Prabhu, V. U. 2021. Large image datasets: A pyrrhic win for computer vision? In *WACV*.

Birhane, A.; Prabhu, V. U.; and Kahembwe, E. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*.

Caetano, C.; Santos, G. O. d.; Petrucci, C.; Barros, A.; Laranjeira, C.; Ribeiro, L. S.; de Mendonça, J. F.; Santos, J. A. d.; and Avila, S. 2025. Neglected Risks: The Disturbing Reality of Children's Images in Datasets and the Urgent Call for Accountability. *arXiv preprint arXiv:2504.14446*.

Carlini, N.; Hayes, J.; Nasr, M.; Jagielski, M.; Sehwag, V.; Tramèr, F.; Balle, B.; Ippolito, D.; and Wallace, E. 2023. Extracting training data from diffusion models. In *USENIX Conference on Security Symposium*.

Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, U.; et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, 2633–2650.

Chamikara, M.; Bertok, P.; Khalil, I.; Liu, D.; and Camtepe, S. 2020. Privacy Preserving Face Recognition Utilizing Differential Privacy. *Computers & Security*, 97: 101951.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Dwork, C. 2006. Differential Privacy. In *International Colloquium on Automata, Languages and Programming*.

Elazar, Y.; Bhagia, A.; Magnusson, I. H.; Ravichander, A.; Schwenk, D.; Suhr, A.; Walsh, E. P.; Groeneveld, D.; Soldaini, L.; Singh, S.; et al. 2024. What's In My Big Data? In *ICLR*.

Froomkin, A. M. 2019. Big data: Destroyer of Informed Consent.

Gadre, S. Y.; Ilharco, G.; Fang, A.; Hayase, J.; Smyrnis, G.; Nguyen, T.; Marten, R.; Wortsman, M.; Ghosh, D.; Zhang, J.; et al. 2023. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36: 27092–27112.

Gandikota, R.; Materzynska, J.; Fiotto-Kaufman, J.; and Bau, D. 2023. Erasing concepts from diffusion models. In *ICCV*.

Gao, L.; Biderman, S.; Black, S.; Golding, L.; Hoppe, T.; Foster, C.; Phang, J.; He, H.; Thite, A.; Nabeshima, N.; Presser, S.; and Leahy, C. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. arXiv:2101.00027.

Garcia, N.; Hirota, Y.; Wu, Y.; and Nakashima, Y. 2023. Uncurated image-text datasets: Shedding light on demographic bias. In *CVPR*.

Gurari, D.; Li, Q.; Lin, C.; Zhao, Y.; Guo, A.; Stangl, A.; and Bigham, J. P. 2019. Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 939–948.

Harpel, T. 2018. Pregnant Women Sharing Pregnancy-Related Information on Facebook: Web-Based Survey Study. *Journal of Medical Internet Research*, 20(3).

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Comput.*, 9(8): 1735–1780.

Hong, R.; Hutson, J.; Agnew, W.; Huda, I.; Kohno, T.; and Morgenstern, J. 2025. A Common Pool of Privacy Problems: Legal and Technical Lessons from a Large-Scale Web-Scraped Machine Learning Dataset. *arXiv preprint arXiv:2506.17185*.

Hu, X.; Liu, D.; Li, H.; Huang, X.; and Shao, J. 2024. VLS-Bench: Unveiling Visual Leakage in Multimodal Safety. *arXiv preprint arXiv:2411.19939*.

Jahan, S.; Ge, Y.-F.; Kabir, E.; and Wang, H. 2023. Analysis and Protection of Public Medical Dataset: From Privacy Perspective. In *International Conference on Health Information Science*. Springer.

Johnson, S. A. 2014. "Maternal devices", social media and the self-management of pregnancy, mothering and child health. *Societies*, 4(2): 330–350.

Ju, T.; Hua, Y.; Fei, H.; Shao, Z.; Zheng, Y.; Zhao, H.; Lee, M.-L.; Hsu, W.; Zhang, Z.; and Liu, G. 2025. Watch Out Your Album! On the Inadvertent Privacy Memorization in Multi-Modal Large Language Models. *arXiv preprint arXiv:2503.01208*.

Katirai, A.; Garcia, N.; Ide, K.; Nakashima, Y.; and Kishimoto, A. 2024. Situating the social issues of image generation models in the model life cycle: a sociotechnical approach. *AI and Ethics*, 1–18.

Kay, A. 2007. Tesseract: an open-source optical character recognition engine. *Linux J.*, 2007(159): 2.

King, J.; and Meinhardt, C. 2024. Rethinking Privacy in the AI Era: Policy Provocations for a Data-Centric World. *Stanford University Human-Centered Artificial Intelligence*.

Kingma, D. P. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kumawat, S.; and Nagahara, H. 2022. Privacy-preserving action recognition via motion difference quantization. In *ECCV*.

Leaver, T.; and Highfield, T. 2018. Visualising the ends of identity: pre-birth and post-death on Instagram. *Information, Communication & Society*, 21(1): 30–45.

Li, Y.; Yang, R.; and Lu, Y. 2024. A privacy risk identification framework of open government data: A mixed-method study in China. *Government Information Quarterly*, 41(1): 101916.

Lie, M.; Graham, R.; Robson, S. C.; and Griffiths, P. D. 2019. "He looks gorgeous"–iu MR images and the transforming of foetal and parental identities. *Sociology of Health & Illness*, 41(2): 360–377.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Liu, C.; Zhu, Y.; Chaudhuri, K.; and Wang, Y. 2020. Revisiting Model-Agnostic Private Learning: Faster Rates and Active Learning. *CoRR*, abs/2011.03186.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023a. Improved Baselines with Visual Instruction Tuning.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual Instruction Tuning. In *NeurIPS*.

Lu, S.; Wang, Z.; Li, L.; Liu, Y.; and Kong, A. W.-K. 2024. Mace: Mass concept erasure in diffusion models. In *CVPR*.

Ma, R.; Zhou, Q.; Xiao, B.; Jin, Y.; Zhou, D.; Li, X.; Singh, A.; Qu, Y.; Keutzer, K.; Xie, X.; et al. 2024. A Dataset and Benchmark for Copyright Protection from Text-to-Image Diffusion Models. *arXiv preprint arXiv:2403.12052*.

Malkin, N. 2022. Contextual integrity, explained: A more usable privacy definition. *IEEE Security & Privacy*, 21(1): 58–65.

McInnes, L.; Healy, J.; Astels, S.; et al. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*.

McInnes, L.; Healy, J.; and Melville, J. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426.

Meister, N.; Zhao, D.; Wang, A.; Ramaswamy, V. V.; Fong, R.; and Russakovsky, O. 2023. Gender artifacts in visual datasets. In *ICCV*.

Mendels, O.; Peled, C.; Vaisman Levy, N.; Hart, S.; Rosenthal, T.; Lahiani, L.; et al. 2018. Microsoft Presidio: Context aware, pluggable and customizable PII anonymization service for text and images.

Moayeri, M.; Basu, S.; Balasubramanian, S.; Kattakinda, P.; Chengini, A.; Brauneis, R.; and Feizi, S. 2024. Rethinking Artistic Copyright Infringements in the Era of Text-to-Image Generative Models. *arXiv preprint arXiv:2404.08030*.

Naumann, A.; Hertlein, F.; Zhou, B.; Dörr, L.; and Furmans, K. 2022. Scrape, Cut, Paste and Learn: Automated Dataset Generation Applied to Parcel Logistics. In *ICMLA*.

Nissenbaum, H. 2004. Privacy as contextual integrity. *Wash. L. Rev.*, 79: 119.

Øyen, L.; and Aune, I. 2016. Viewing the unborn child–pregnant women's expectations, attitudes and experiences regarding fetal ultrasound examination. *Sexual & Reproductive Healthcare*, 7: 8–13.

Papernot, N.; Abadi, M.; Erlingsson, U.; Goodfellow, I.; and Talwar, K. 2016. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*.

Papernot, N.; Song, S.; Mironov, I.; Raghunathan, A.; Talwar, K.; and Erlingsson, Ú. 2018. Scalable private learning with pate. *arXiv preprint arXiv:1802.08908*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32*, 8024–8035. Curran Associates, Inc.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *JMLR*.

Pizzi, E.; Roy, S. D.; Ravindra, S. N.; Goyal, P.; and Douze, M. 2022. A Self-Supervised Descriptor for Image Copy Detection. In *CVPR*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*. PMLR.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*.

Rempe, M.; Heine, L.; Seibold, C.; Hörst, F.; and Kleesiek, J. 2024. De-Identification of Medical Imaging Data: A Comprehensive Tool for Ensuring Patient Privacy. *arXiv preprint arXiv:2410.12402*.

Roberts, J.; Griffiths, F. E.; Verran, A.; and Ayre, C. 2015. Why do women seek ultrasound scans from commercial providers during pregnancy? *Sociology of health & illness*, 37 4: 594–609.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*.

Samson, L.; Barazani, N.; Ghebreab, S.; and Asano, Y. M. 2024. Privacy-aware visual language models. *arXiv preprint arXiv:2405.17423*.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*.

Schuhmann, C.; Kaczmarczyk, R.; Komatsuzaki, A.; Katta, A.; Vencu, R.; Beaumont, R.; Jitsev, J.; Coombes, T.; and Mullis, C. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. In *NeurIPS Workshop Datacentric AI*, FZJ-2022-00923. Jülich Supercomputing Center.

Shokri, R.; and Shmatikov, V. 2015. Privacy-Preserving Deep Learning. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, 1310–1321. New York, NY, USA: Association for Computing Machinery. ISBN 9781450338325.

Skelton, E.; Cromb, D.; Smith, A.; Harrison, G.; Rutherford, M.; Malamateniou, C.; and Ayers, S. 2024. The influence of antenatal imaging on prenatal bonding in uncomplicated pregnancies: a mixed methods analysis. *BMC Pregnancy and Childbirth*, 24(1): 265.

Solove, D. J. 2023. Data Is What Data Does: Regulating Based on Harm and Risk Instead of Sensitive Data. *Nw. UL Rev.*, 118: 1081.

Subramani, N.; Luccioni, S.; Dodge, J.; and Mitchell, M. 2023. Detecting personal information in training corpora: an analysis. In *TrustNLP 2023*.

Suriyakumar, V. M.; Alur, R.; Sekhari, A.; Raghavan, M.; and Wilson, A. C. 2024. Unstable Unlearning: The Hidden Risk of Concept Resurgence in Diffusion Models. *arXiv preprint arXiv:2410.08074*.

Sweeney, L. 2000. Simple Demographics Often Identify People Uniquely.

Thiel, D. 2023. Identifying and eliminating csam in generative ml training data and models. Technical report, Technical Report. Stanford University, Palo Alto, CA.

Tonge, A.; and Caragea, C. 2016. Image privacy prediction using deep features. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Tonge, A.; and Caragea, C. 2020. Image Privacy Prediction Using Deep Neural Networks. *ACM Trans. Web*, 14(2).

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tran, L.; Kong, D.; Jin, H.; and Liu, J. 2016. Privacy-cnh: A framework to detect photo privacy with convolutional neural network using hierarchical features. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Trepte, S.; and Masur, P. K. 2023. Definitions of Privacy. In *The Routledge Handbook of Privacy and Social Media*, 3–15. Routledge.

van der Maaten, L.; and Hinton, G. 2008. Visualizing Data using t-SNE. *JMLR*.

Wang, X.; Xie, L.; Dong, C.; and Shan, Y. ???? Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data. In *International Conference on Computer Vision Workshops (ICCVW)*.

Wu, B.; Zhao, S.; Sun, G.; Zhang, X.; Su, Z.; Zeng, C.; and Liu, Z. 2019. P3sgd: Patient privacy preserving sgd for regularizing deep cnns in pathological image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2099–2108.

Xu, A.; Fang, S.; Yang, H.; Hosio, S.; and Yatani, K. 2024. Examining human perception of generative content replacement in image privacy protection. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–16.

Yu, J.; Xue, H.; Liu, B.; Wang, Y.; Zhu, S.; and Ding, M. 2021. GAN-Based Differential Private Image Privacy Protection Framework for the Internet of Multimedia Things. *Sensors*, 21(1).

Zhang, G.; Wang, K.; Xu, X.; Wang, Z.; and Shi, H. 2024. Forget-Me-Not: Learning to Forget in Text-to-Image Diffusion Models. In *CVPR Workshops*.

Zhang, S.; and Li, X. 2022. Differential privacy medical data publishing method based on attribute correlation. *Scientific Reports*.

Zhang, Y.; Lu, Y.; Nagahara, H.; and Taniguchi, R.-i. 2014. Anonymous camera for privacy protection. In *ICPR*.

Zhao, C.; Mangat, J.; Koujalgi, S.; Squicciarini, A.; and Caragea, C. 2022. Privacyalert: A dataset for image privacy prediction. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 1352–1361.

Zhu, C.; Zeng, R.; Zhang, W.; Evans, R.; and He, R. 2019. Pregnancy-related information seeking and sharing in the social media era among expectant mothers: qualitative study. *Journal of medical Internet research*, 21(12): e13694.