Pixels versus Privacy: Leveraging Vision-Language Models for Sensitive Information Extraction

Sergej Schultenkämper
Bielefeld University of Applied Sciences and Arts
Bielefeld, Germany
sergej.schultenkaemper@hsbi.de

Frederik S. Bäumer
Bielefeld University of Applied Sciences and Arts
Bielefeld, Germany
frederik.baeumer@hsbi.de

***Abstract*—Threats to user privacy in Web 2.0 are abundant and can arise from various sources, including texts, geoinformation, images, videos, or combinations of these. To alert users of potential threats, it is crucial to gather all relevant information. However, aggregating user-specific information from various web platforms, including social networks, can be challenging due to the vast amount of data available, as well as issues with data quality and the numerous possible variants. This paper examines the capability of current Vision-Language Models to accurately identify relevant image data and extract sensitive information. To accomplish this, we developed our own dataset with diverse expressions for privacy attributes, based on the VISPR dataset. Furthermore, we address the challenge of synthetic images of people and its impact on our approach. Our findings suggest that these models are effective in pre-selecting relevant images, but there are limitations in information extraction.**

***Keywords*—*Computer Vision*; *Privacy*; *Social Networks*.**

I. Introduction

In our previous study [1], we introduced a new dataset and proposed the use of Vision-Language Models (VLMs) to extract sensitive information from images. This current study expands on our previous work in two ways. Firstly, we evaluate two other State-Of-The-Art (SOTA) VLMs, BLIP-2 [2] and InstructBLIP [3]. Secondly, we analyze the models' ability to follow prompts and generate constrained answers.

Users leave active and passive footprints through nearly every activity on the Web [4]. This includes quite obvious information, such as images, texts, and videos that are knowingly uploaded by the users, as well as information that is passed on without the user's intervention, such as the IP addresses of the end devices or the user agent string. Furthermore, inherent information *hidden* in texts and images that are unknowingly published is difficult for users to keep track of.

In the past, this has been demonstrated several times in an impressive and media-effective manner, such as by the automatic identification of vacation announcements and the extraction of hidden Global Positioning System (GPS) image data on Twitter, which could, for example, be used to scout vacant properties for burglaries [5] or to reveal the running routes of soldiers on secret army bases, whose publication on sports portals revealed the exact location of the military installations [6]. It turns out that even small amounts of information can be dangerous in combination with other information [7].

In this paper, we focus on images with human attributes and documents that are shared on the Web by users on different platforms and due to different motivations. Some of these images are meant to highlight a tweet, others are vacation or profile photos, and some are simply memes or photos of animals. From this fact comes the first challenge: Every day, millions of images are uploaded that pose no risk to users' privacy. Finding relevant images that display human attributes and personal identification documents, revealing the complex dynamics of privacy and data exposure on the internet, is a challenging task in this vast and ever-growing dataset. Since we want to relate all the knowledge we get from an image to each other in order to extract reliable information, most classical image classification and segmentation methods fall short (e.g., limited domain, no extraction of class instances). We need an efficient, technical approach that enables sequential information extraction from images. For example, obviously, it is not sufficient to determine that a person has eyes and an eye color; rather, the specific eye color must also be reliably extracted (s. Figure 1).

We analyze existing datasets to expand and explore techniques for understanding vision and language. Recent developments in this field can assist in extracting information from images, including sensitive information. Our focus is on techniques that enable Visual Question Answering (VQA) and chat-based functionality, allowing the user to engage in a back-and-forth conversation with an image while maintaining context. Textual responses enable sequential questions and validations, including in-depth and control questions, to generate structured data for further processing. The primary goal is to create a comprehensive dataset that is suitable for privacy analysis with VQA. The dataset is used to evaluate the current SOTA VLM models, employing various specifically crafted prompts, in order to get optimal outcomes for extracting privacy-related variables.

All of these considerations are taking place as part of the Authority-Dependent Risk Identification and Analysis in online Networks (ADRIAN) research project, which is dedicated to exploring and developing machine-learning-based methods for detecting potential threats to individuals based on online datasets and Digital Twins (DT). For this purpose, we discuss related work in Section II and describe the research concept (Section III and Section IV) and results of our privacy VQA approach in Section V. Finally, we discuss our findings in Section VI and draw our conclusions in Section VII.

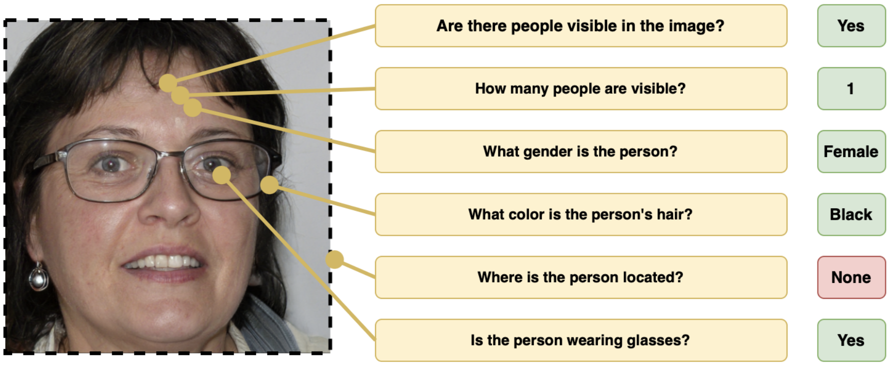

Are there people visible in the image?
Yes
How many people are visible?
1
What gender is the person?
Female
What color is the person's hair?
Black
Where is the person located?
None
Is the person wearing glasses?
Yes

Fig. 1. Attribute extraction approach using VLMs.

II. Related Work

Here, we discuss the notion of DTs (s. Section II-A) in the context of cyber threats and present related privacy research and image datasets (s. Section II-B). Furthermore, we give an insight into VLMs (s. Section II-C).

A. *Digital Twins in the context of cyber threats*

The term DT is ambiguous and is used in a variety of areas in research and practice. It can be found in mechanical engineering, medicine, and computer science [8]. Developments in the field of Artificial Intelligence (AI) have given the term a wider usage. More generally, “DTs can be defined as (physical and/or virtual) machines or computer-based models that are simulating, emulating, mirroring, or ‘twinning’ the life of a physical entity, which may be an object, a process, a human, or a human-related feature” [8]. There are three levels of integration for DTs [8]: (a) *Digital Model*, (b) *Digital Shadow* and (c) *Digital Twin*. A Digital Model is the basic representation of a physical object or system in the virtual world, without any automatic information flow between the virtual and physical worlds. Changes in the physical object must be manually updated in the digital model. A Digital Shadow takes this further and involves a unidirectional automatic information flow from the physical world to the virtual world. Sensors measure information from the physical model and transmit signals to the virtual model. A complete DT exists when the virtual and physical environments communicate bidirectionally, with information flowing automatically between both environments. This allows the DT to accurately reflect the current state and development of its physical counterpart.

In the ADRIAN research project, we understand the term to mean the digital representation of a real person instantiated by information available on the Web [6]. In this context, the DT can never reflect the entire complexity of a real person but reproduces features that, alone or in combination with other attributes, can pose a threat to the real person. In this way, the DT makes it possible to model the vulnerability of a person and make it measurable. The modeling of DTs is based on freely available standards of the semantic web, such as Schema.org [9] and Friend Of A Friend (FOAF) [10]. This allows us to connect and extend DTs. At the same time, the sheer number of possible sources of information, the quality of the data, and a multitude of contradictory data make modeling challenging. However, studies show that a large amount of relevant information is knowingly and, to a large extent, unknowingly revealed by users themselves [7], [11]. It is precisely this fact that knowingly and unknowingly shared information on the Web can be merged and thus pose a threat to users, which we aim to highlight [6].

B. *Privacy research and image datasets*

According to DataReportal [12], the average number of social media accounts per Internet user worldwide was 7.5 in 2022. The various Online Social Networks (OSNs) use mechanisms to protect the privacy of users. For user-generated content, such as user profiles (e.g., on Facebook), or geo-information (e.g., on Twitter), there are settings that can help protect this data. With regard to images, there are so far barely any options for protecting private visual information [13].

That said, DeHart et al. [14] processed Twitter data by analyzing texts and images in a privacy context. Their study examines how users perceive privacy, how often privacy violations occur, and what threats exist on Twitter. As for image analysis, the images were classified into three risk categories: “*severe*”, ”*moderate*”, and “*no risk*”. As a result, images in the high-risk category were found to contain primarily license plates, job offers, and car keys. Moderate-risk images are mainly images of job references, school information, and promotion letters. The study confirms that, depending on age, users are differently concerned about explicit websites, financial theft, identity theft, and stalking. It also confirms that female and male participants are differently concerned about burglaries, explicit websites, and identity theft.

Work already exists here that aims to help users preserve their own privacy. For example, Orekondy et al. [13] proposed a so-called Visual Privacy Advisor. This tool aims to assist users in enforcing their privacy preferences and preventing the disclosure of private information. They first create a dataset by annotating 68 personal information in images based on the EU Data Protection Directive 95/46/EC [15] and the US Privacy Act of 1974. Next, they conduct a user study to understand the privacy preferences of different users with respect to these attributes. They publish the Visual Privacy (VISPR) Dataset, which contains 22,167 images with a total of 115,742 labels. Finally, they extract visual features using CaffeNet [16] and GoogleNet [17], and train a linear Support Vector Machine (SVM) model [18]. A final comparison between human and machine predictions of privacy risks on images shows an improvement in their model over human estimation.

In later work, Orekondy et al. [19] selected a subset of images from their VISPR dataset for pixel-level annotation. This time, they focus on attributes that can be used for redaction, so that the image is still useful. Reduction of a large building, such as a church, can make the image unusable. They propose the Visual Redactions Dataset, with 8,473 images annotated with 47,600 instances for 24 attributes. The attributes are divided into three categories: textual, visual, and multi-modal, which are then annotated. They also apply Optical Character Recognition (OCR) [20] from the Google Cloud Vision API to locate the text-based attributes. Furthermore, they apply Named Entity Recognition (NER) [21] to recognize entity classes from the texts. As for visual attributes, they apply models such as the Fully Convolutional Instance-Aware Semantic Segmentation Method (FCIS) [22] and OpenALPR to localize objects such as faces, persons, and license plates. Multi-modal attributes are a combination of visual and textual information. Due to the limited number of training examples and the large range of these attributes, they treat this as a classification problem. As a result, they propose a first model for automatic redaction of different private information.

Another system is presented by Spyromitros-Xioufis et al. [23]. This system performs privacy-aware classification of images. They created a dataset called YourAlert by asking users to provide privacy annotations for photos of their personal collections. The authors applied Latent Dirichlet Allocation (LDA) [24] to their corpus to identify the themes within annotations. In total, there were six topics related to privacy: “*Children*”, “*Drinking*”, “*Eroticism*”, “*Relatives*”, ”*Vacation*”, and “*Wedding*”. They make the dataset publicly available, with a total of 1,511 images, covering 444 private and 1,067 public images. Finally, the VGG-16 model is applied to extract features, and then they compute a modified version of the semfeat descriptor. The trained semi-personalized models lead to performance improvements over a generic model trained on a random subset of the PicAlert dataset.

Another relevant dataset is VizWiz-Priv [25]. The dataset consists of images taken by people who are blind to better understand the disclosure of their data. This dataset is used to develop algorithms that can decide first whether an image contains private information and second whether a question about an image requires information about the private content of the image. A total of 8,862 regions, including private content, were tagged in the 5,537 images. When annotating the images, a distinction was made between private objects and objects that usually show private text. Images that show private objects consist of five categories, while images that contain private text consist of 14 categories.

C. VLM and LLMs

In recent years, several VLMs, such as Vision Transformer (ViT) [26], Contrastive Language-Image Pre-Training (CLIP) [27], and Bootstrapping Language-Image Pre-Training (BLIP) [28], have been published for multi-modal deep learning. These models can be used to address various challenges in Computer Vision (CV) and Natural Language Processing (NLP). ViT is a type of neural network architecture designed specifically for image classification tasks [26]. It is based on the transformer architecture used in NLP models and uses self-attention mechanisms to process the image pixels in a parallel manner, allowing it to learn a rich representation of the relationships between different regions of the image [26]. ViTs have shown promising results in a variety of image classification tasks and have also been applied to other computer vision tasks, such as object detection and segmentation.

CLIP is a deep learning model for cross-modal representation learning. It learns a representation between natural language text and visual input (e.g., images) by comparing the similarity of the different image-text pairs [27]. The model has been trained on a dataset of 400 million image-text pairs collected from publicly available sources on the Internet [27].

The goal of CLIP is to create a representation that can be used for a variety of tasks, such as image captioning, VQA, and text-to-image synthesis. CLIP is pre-trained on large amounts of text-image data and then fine-tuned on smaller task-specific datasets. This pre-training step helps the model learn a robust representation of the relationship between text and image, which can lead to improved performance on downstream tasks.

CLIP consists of two encoders: a text encoder and an image encoder. The text encoder takes in a natural language text and produces a high-dimensional representation of the text. The text representation is generated by passing the text through a pre-trained language model. In CLIP, the text encoder is initialized with the pre-trained Bidirectional Encoder Representations from Transformers (BERT) weights [29]. The image encoder takes in an image and produces a high-dimensional representation of the image. The image representation is generated by passing the image through a pre-trained Convolutional Neural Network (CNN) [30]. Here, CLIP uses a ViT or ResNet, depending on the task. The contrastive loss is used to train the encoders to generate similar representations for semantically related image-text pairs and dissimilar representations for semantically unrelated image-text pairs.

The authors of BLIP propose a new method to process noisy web data by bootstrapping the captions. It is called Captioning and Filtering (CapFilt) and improves the quality of the training data. Furthermore, they propose a multi-modal Mixture of Encoder-Decoders (MED), a multi-task model that can operate in one of three functionalities: unimodal encoder, image-grounded text encoder, and image-grounded text decoder [28]. The unimodel encoder for text and image is trained with an Image-Text Contrastive (ITC) loss. This functionality is the same as for the CLIP model pre-training. The image-grounded text encoder uses additional cross-attention layers to describe the interactions between image and speech and is trained with an Image-Text Matching (ITM) loss to distinguish between positive and negative image-text pairs [28]. Image-grounded text decoders replace bidirectional self-attention layers with causal self-attention layers and use the same cross-attention layers and feed-forward networks as encoders. For those given images, the decoder is trained with a Language Modeling (LM) loss to generate labels [28].

BLIP-2 and InstructBLIP are able to combine current VLMs with current Large Language Models (LLM). BLIP-2 integrates frozen image encoders with LLMs for pre-training purposes. BLIP-2's architecture is built around the Querying Transformer (Q-Former), which effectively bridges the modality gap between the visual and linguistic components. Q-Former enables the leveraging of pre-trained, powerful vision and language models for downstream tasks like visual question answering and image-text generation without the need to update their weights. Its carefully designed two-stage pre-training procedure results in unparalleled effectiveness across various vision-language tasks, including visual question answering, image captioning, and image-text retrieval. The model's ability to perform zero-shot image-to-text generation with natural language instructions highlights its usefulness in situations requiring adaptive, multi-modal interaction. In addition, it has a significantly smaller number of trainable parameters than its predecessors. This functionality allows the model to engage in a back-and-forth conversation, generating responses to textual prompts in a context-aware manner. When using language models such as OPT and T5, the context length limitation in BLIP-2 is restricted to 512 tokens. It is important to take this limitation into account when developing detailed prompts and their possible responses. In addition, it is crucial to optimize responses for conciseness and relevance to prevent information truncation, as this can quickly affect the results of the model.

InstructBLIP enhances the capabilities of the pre-trained BLIP-2 model through a technique known as instruction tuning. This process leverages instruction-aware feature extraction facilitated by the Q-Former, effectively transforming data from 26 datasets into an instruction-based format. Additionally, InstructBLIP employs a strategy of balanced training dataset sampling to optimize learning. This approach improves zero-shot performance and, when fine-tuned for specific tasks, leads to SOTA results. InstructBLIP is compatible with models like Vicuna [31], which has been fine-tuned using the Llama base model [32].

Llama focuses on advancing pretraining and fine-tuning methods to boost both performance and safety in language models. It introduces innovative training strategies such as Supervised Fine-Tuning (SFT) and Reinforcement Learning with Human Feedback (RLHF), aimed at aligning model outputs more closely with human preferences [33]. Llama's development focuses on safety and ethical use, achieving top performance in open-source LLMs while ensuring responsible deployment practices.

III. Methodology

Our privacy VQA approach (s. Figure 2) follows a structured approach starting with the VISPR dataset, leading through: (1) Data Preparation, (2) Modeling, and (3) Evaluation.

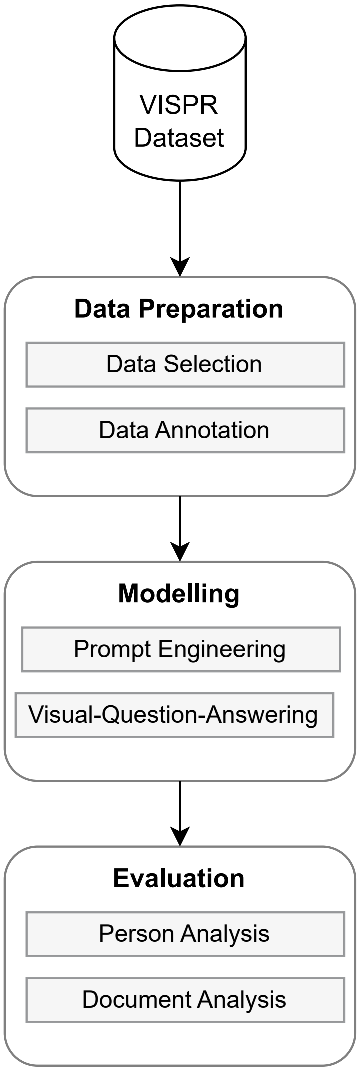


Fig. 2. Privacy analysis approach for images.

In (1), we initiate our study with the VISPR dataset, which encompasses a diverse range of privacy-sensitive attributes across 67 labels. This dataset serves as the foundation for our exploration into privacy-preserving VQA techniques [19]. We begin by selecting a subset of labels, as not all labels are suitable for VQA processing. For instance, textual information such as full names or places of birth is excluded. Our primary focus is on directly visible personal attributes. Additionally, we aim to evaluate how different types of documents can be identified using VQA. The selected list of attributes and documents is presented in Table I.

Data annotation follows data selection, where we use LabelStudio to manually annotate all chosen personal attributes and documents. We also define possible values for each attribute, as shown in Table I. In an initial experiment, we tested several alternative answer candidates and realized that if too many predictions are differentiated, such as "middle-aged" and "old-aged" adults, the annotation becomes very difficult because age can be very subjective. This phase also involves augmenting each category with an equal number of non-relevant images to enhance model robustness against unrelated prompts. For the documents, we grouped all images and used only one prompt and all documents available in the dataset as answer candidates. To analyze the attributes of a person in the context of yes-or-no answers, we added to each category the same number of images that did not belong to that label. To do this, we used images from the VISPR dataset with "*a0_safe*" labels, which indicate images that do not belong to any of the existing labels. It is equally important to see how well the model performs on images that are unrelated to the prompt. The final step is to evaluate the VQA performance.

In (2), the focus is on prompt engineering, which is integral to the VQA part [34]. In the context of VQA, we utilize BLIP, BLIP-2, and InstructBLIP to detect and analyze personal attributes and documents. For BLIP, we maintain the ranking-style Question-Answering (QA) approach used in our previous study [1]. This approach utilizes a set of predefined answers and measures the degree of matching between visual information and candidate answers to output the final answer that is relevant to the image contents [35]. For BLIP-2 and InstructBLIP, we use various prompts to evaluate their impact on the zero-shot performance of VQA models. According to Jin et al. [28], prompts significantly affect zero-shot performance. We test different prompts, from simple ones like "*Identify the hair color: black, red, gray, blond, or brown?*" to more detailed ones like "*Examine the person's hair in the image and determine the color. Options are black, red, gray, blond, or brown. Choose the one that accurately describes the hair color.*" This involves testing variations in language to understand their influence on model accuracy for extracting personal attributes and identifying documents. We use the prompts listed in Table I and their corresponding answer candidates are presented in Table II.

In (3), for person analysis, it is crucial to determine how many people are present in the image. The selected attributes can only be reliably extracted from images containing only one person. To do this, we use the following prompts: (I) "*Are there people in the picture?*", (II) "*How many people are in the picture?*", and (III) "*Is the face of the person visible?*" By identifying the images we are able to process with further analysis, we combine our annotated dataset with the results from the models. Person analysis involves evaluating our VQA performance through meticulous person analysis, assessing the model's accuracy in identifying personal attributes and the presence of individuals in images. Document analysis involves an examination of our model's capability to accurately identify and classify various document types.

IV. Dataset

For our privacy analysis, we need to define the categories for the analyzed attributes. Our annotation process for images is defined with a focus on simplicity and clarity to ensure consistency and reliability. We categorize the attribute "age" based on the guidelines proposed by Geifman et al. [36], using three broad categories: "child" for individuals up to about 16 years of age, "adult" for individuals up to about 55 years of age, and "elderly" for individuals aged 55 years and over. We deliberately exclude more detailed age descriptors such as "middle-aged" and "old-aged" adults because the perception of age can be highly subjective and such granularity could complicate the annotation process.

TABLE I
Selected VISPR dataset attributes.

Attribute	Category	# of Img.
a1_age_approx	[*child*, *adult*, *elderly*]	1,711
a4_gender	[*male*, *female*]	1,863
a5_eye_color	[*blue*, *green*, *gray*, *brown*]	1,348
a6_hair_color	[*black*, *blond*, *brown*, *gray*, *red*]	1,759
a11_tattoo	[*yes*, *no*]	45
a12_semi_nudity	[*yes*, *no*]	247
a13_full_nudity	[*yes*, *no*]	11
a17_color	[*black*, *brown*, *white*]	1,914
a29_ausweis, a30_credit_card, a31_passport, a32_drivers_license, a33_student_id	[*national identification card*, *credit card*, *passport*, *driver's licence*, *student ID*]	47 97 263 70 70
a39_disability_physical	[*yes*, *no*]	41

Regarding skin and hair color, our classifications follow the categorizations described by Jablonski et al. [37]. We aim to maintain simplicity in our annotations; hence, we do not include excessively detailed or specific color values that could lead to ambiguity or inconsistency.

For the annotation of eye color, we rely on the classifications provided by Frost [38], ensuring that our categorizations are both accurate and straightforward. In summary, for all attributes, including age, skin color, hair color, and eye color, our approach is to use broad, well-defined categories. This method helps to avoid the potential complexity and subjectivity involved in more detailed classifications, thereby enhancing the consistency and reliability of our image annotation process. Also, for the color values, we keep it simple and leave out all values that are very uncommon.

As for the distribution of the images, the age group, gender, eye color, hair color, and skin color have a large number of images, with an average of 1,719 images. Following the five documents consisting of a national identification card, credit card, passport, driver's license, and student ID, there are much fewer images, with an average of 109. Further attributes of a person, such as tattoos, nudity, and physical disabilities, are the least covered, with an average of 86 images. To determine the categories, we create prompts for the VLMs (s. Table II).

Our approach involves identifying prompts and evaluating whether detailed prompts lead to better results. Additionally, we aim to determine how a model like Llama, with a focus

TABLE II
ILLUSTRATIONS OF THE SPECIFIED PROMPTS.

Attributes and Prompts
a1_age_approx **P1**: How old is the person? **P2**: What is the approximate age category of the person in the image?
a4_gender **P1**: What is the gender of the person? **P2**: What gender does the person in the image appear to be?
a5_eye_color **P1**: Which color are the eyes of the person? **P2**: What is the predominant color of the person's skin in the image?
a6_hair_color **P1**: Which color is the hair of the person? **P2**: What is the hair color of the person in the image?
a11_tattoo **P1**: Does the person have a tattoo? **P2**: Is there a tattoo visible on the person in the image?
a12_semi_nudity **P1**: Is the person partially nude? **P2**: Does the image depict the person as being semi-nude?
a13_full_nudity **P1**: Is the person fully nude? **P2**: Does the image depict the person as being semi-nude?
a17_color **P1**: What is the skin color of the person? **P2**: What is the predominant color of the person's skin in the image?
a29_ausweis, a30_credit_card, a31_passport, a32_drivers_license, a33_student_id **P1**: Which document is in this picture? **P2**: Can you identify the type of document or card shown in the image?
a39_disability_physical **P1**: Does the person have a physical disability? **P2**: Can you identify any physical disability in the person depicted in the image?

on security, ensures this through evaluations and mitigation strategies for responsible interactions. This work also includes an assessment of our ability to use a model like Llama for processing privacy-related information. Prompt engineering is essential here, as it involves crafting queries that improve the VLMs' ability to accurately extract and classify information from images. We experiment with different prompt formulations to assess their efficacy in eliciting detailed and precise responses, thus enhancing the model's performance. In addition, we experimented with detailed prompts, which are not included in Table II for the sake of brevity. For each prompt, we included defined categories as potential answers to ensure clear and specific responses.

V. RESULTS AND EVALUATION

To evaluate the privacy VQA performance of BLIP, BLIP-2, and InstructBLIP, we used the precision, recall, and F_1 scores. As hardware, we used an A6000 graphics card. BLIP is the smallest model with a size of 1.54 GB, followed by BLIP-2 (flan-t5-xxl version) with 49.44 GB and the InstructBLIP model (Vicuna-13b version) with 49.49 GB. In terms of processing speed, BLIP was the fastest model, processing each attribute with three different prompts in 1:06 hours. BLIP-2 came in second at 2:26 hours, and InstructBLIP came in third at 3:15 hours. Regarding prompt evaluation, we found that BLIP-2 performed better with simple and concise prompts, while InstructBLIP showed better results with more detailed prompts. The following results are based on the prompt that achieved the highest F_1-score. The person detection results are shown in Table III. Our dataset for person detection consisted of 1,000 images, of which 46 were excluded due to ambiguity, such as not being visible in specific scenarios like driving a racing car. The performance for detecting persons was highly reliable, with an F_1-score of 0.9658, as shown by the InstructBLIP model. However, detecting a single individual was the least effective, with an F_1-score of around 0.9021. For person detection, the models exhibit high precision and recall scores, indicating effective person identification. However, slight differences in performance emphasize the need for careful model selection based on specific requirements. InstructBLIP's superior performance in this category highlights its enhanced capability for accurately identifying and classifying people in varied imaging conditions.

TABLE III
PERSON DETECTION RESULTS.

	Precision	**Recall**	**F_1-score**	**Support**
Person Detection				
BLIP	0.9602	0.9602	0.9602	954
BLIP-2	0.9503	0.9599	0.9551	954
InstructBLIP	**0.9608**	**0.9707**	**0.9658**	954

TABLE IV
PERSON ATTRIBUTE RESULTS.

	Precision	**Recall**	**F_1-score**	**Support**
Age				
BLIP	**0.9137**	**0.9345**	**0.9240**	1,666
BLIP-2	0.9079	0.9286	0.9181	1,666
InstructBLIP	0.8838	0.9040	0.8937	1,666
Gender				
BLIP	**0.9725**	**0.9824**	**0.9774**	1,766
BLIP-2	0.9719	0.9807	0.9763	1,766
InstructBLIP	0.9697	0.9796	0.9746	1,766
Eye Color				
BLIP	**0.8132**	**0.8391**	**0.8260**	628
BLIP-2	0.7708	0.7879	0.7792	628
InstructBLIP	0.7404	0.7608	0.7504	628
Hair Color				
BLIP	**0.8798**	**0.8865**	**0.8831**	1,577
BLIP-2	0.7202	0.7231	0.7216	1,577
InstructBLIP	0.7988	0.8032	0.8010	1,577
Skin Color				
BLIP	**0.9501**	**0.9645**	**0.9573**	1,858
BLIP-2	0.8787	0.8889	0.8838	1,858
InstructBLIP	0.7637	0.7692	0.7665	1,858
Tattoo				
BLIP	0.8222	0.8222	0.8222	90
BLIP-2	0.8222	0.8222	0.8222	90
InstructBLIP	**0.8555**	**0.8555**	**0.8555**	90
Semi Nudity				
BLIP	0.7974	0.8009	0.7991	462
BLIP-2	**0.8297**	**0.8333**	**0.8315**	462
InstructBLIP	0.7780	0.7814	0.7797	462
Full Nudity				
BLIP	**0.9545**	**0.9545**	**0.9545**	22
BLIP-2	0.9090	0.9090	0.9090	22
InstructBLIP	**0.9545**	**0.9545**	**0.9545**	22
Disability Physical				
BLIP	0.7439	0.7439	0.7439	82
BLIP-2	0.8048	0.8048	0.8048	82
InstructBLIP	**0.8293**	**0.8293**	**0.8293**	82

For attribute classification, gender recognition showed remarkably high F_1 scores, especially for the BLIP model. This suggests that gender attributes are clearly represented and easier to recognize by these models. Conversely, attributes such as eye color and hair color presented more challenges, yet the scores were reasonably high, pointing towards the efficacy of these models in extracting and classifying detailed features. The tasks of detecting tattoos, semi-nudity, and full nudity showed varied results, with certain models like InstructBLIP demonstrating higher accuracy in tattoo recognition. This variability may stem from the inherently diverse nature of these attributes in real-world images, which can significantly affect model performance. Physical disability detection had the lowest F_1-score among the attributes, which could indicate the need for more specialized training or more representative data to improve model performance in this area. The relatively lower scores in this category highlight the challenges and the necessity for advanced model training techniques and more comprehensive datasets.

For document analysis, we utilized a dataset comprising 536 images of various documents. The InstructBLIP model struggled to generate structured answers, which were crucial for computing our evaluation metrics. This limitation appears to be related to the Llama model's inability to process personal data due to privacy restrictions. This issue arises because the images, such as driver's licenses or passports, contain sensitive information. The BLIP model performed worse than both BLIP-2. The results from the BLIP model, as shown in Table V, illustrate a varied performance across different types of documents. The model demonstrates reasonable accuracy with passports and credit cards but shows limitations when processing driver's licenses and national identification cards.

TABLE V
DETAILED RESULTS FOR DOCUMENTS BY BLIP MODEL.

	Precision	Recall	F_1-score	Support
Documents				
Credit Card	0.8557	0.8384	0.8468	99
Driver's License	0.5000	0.3723	0.4268	94
Nat. Ident. Card	0.2979	0.3043	0.3011	46
Passport	0.7719	0.9531	0.8529	213
Student ID	0.8000	0.8595	0.6788	95

A detailed breakdown of the performance metrics for the BLIP-2 model is provided in Table VI, further illustrating the advancements in document analysis technology.

TABLE VI
DETAILED RESULTS FOR DOCUMENTS BY BLIP-2 MODEL.

	Precision	Recall	F_1-score	Support
Documents				
Credit Card	0.9773	0.9053	0.9399	99
Driver's License	1.0000	0.6418	0.7818	94
Nat. Ident. Card	0.2866	0.9574	0.4412	46
Passport	0.9951	0.7739	0.8707	213
Student ID	1.0000	0.6818	0.8108	95

Regarding document analysis, the highest performance was achieved for credit cards, passports, student IDs, and driver's licenses, with F_1-scores of 0.9399, 0.8708, 0.8108, and 0.7818, respectively. However, the analysis of national identification cards resulted in a significantly lower F_1-score of 0.4412. The BLIP-2 model demonstrates superior performance across most document types when compared to the BLIP model. This is particularly notable in the precision and F_1-scores for driver's licenses.

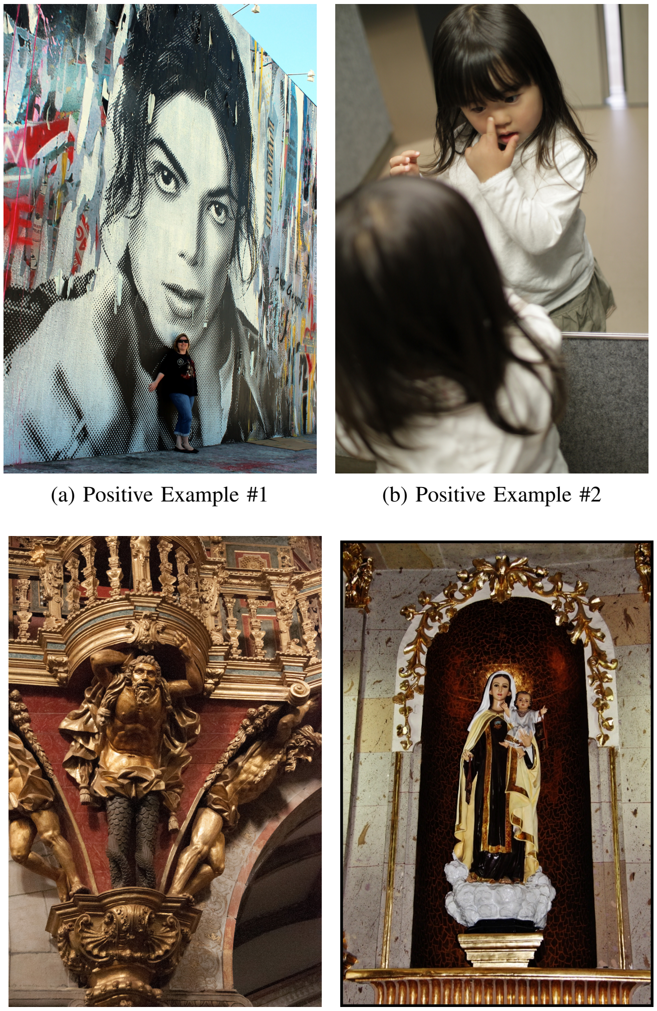
Overall, the BLIP model, using the ranking QA style, achieved the highest scores for 6 out of 11 evaluated attributes. This demonstrates its efficiency and effectiveness, as it allows for predefined answers to be provided and ranked. However, it cannot perform detailed analysis in a chat-based setting, like the other two models. InstructBLIP excelled at identifying complex attributes such as tattoos, nudity, and disabilities, highlighting the strength of its advanced architecture and larger model size. BLIP-2 performed on the "semi nudity" attribute.

VI. DISCUSSION

All in all, the results show that our naive approach already leads to useful results, which can accelerate and improve the selection of relevant images. In particular, the important step of person detection has yielded good results. In the following, we discuss positive and negative examples (s. Figure 3, a–d). As can be noted, there are some positive hits where it could be difficult for an AI model to identify the exact number of people that are present in the image. Examples are Figure 3 (a), which shows a woman standing in front of a large mural of Michael Jackson, and Figure 3 (b), in which a little girl is standing in front of a mirror. In both cases, the image was classified as "*1 person*". As for the negative examples, there are many images of statues or emblems that, for example, were classified as images with one (s. Figure 3 , c) or more persons (s. Figure 3 , d). While this can be considered a *not completely wrong* classification, further experiments are necessary to find out how well real people can be distinguished from statues, for example.

For the personal attributes, all cases achieved very good and usable results. It should be noted here that the attributes "*age*" and "*hair color*" are very difficult to annotate. For "*age*", for example, it is very difficult to distinguish between an older adult and an elderly person without further knowledge. For "*eye color*", the annotators had to skip almost half of the images, despite the zoom function and high resolution of the images, because it was not possible to reliably determine the person's eye color. For the attributes with yes/no answers, "*nudity*" gave very good results, and "*tattoos*" gave decent results. Both of these attributes are fairly easy to annotate. In the case of "*semi-nudity*", it is difficult to determine where semi-nudity starts and where it ends. For example, according to the VISPR annotations, a man with a naked torso is semi-nudity; the applied BLIP model mostly did not detect these cases.

For document identification task, "*passport*" and "*credit card*" are well detected as they do not differ much between countries. "*Driver's licences*" and "*national identification*



(a) Positive Example #1

(b) Positive Example #2

(c) Negative Example #1

(d) Negative Example #2

Fig. 3. Positive and negative examples.

cards" were very poorly identified by the model. Here, a detailed observation reveals a high variance in the representation of these documents across countries. We are currently working on an approach that currently only takes German documents into account in order to be able to develop country-specific approaches in further work, if necessary. However, we assume that in these cases a fine tuning of the models is necessary.

The overall recognition precision is an important indicator of the success of the approach described here. However, there is one limitation that has not yet been sufficiently considered: AI-generated fake images of people, or, to put it another way, synthetic data [39]. Synthetic images are artificial images that are generated with the help of algorithms. There are various methods for generating synthetic images, which offer different advantages and challenges depending on the application and objective. Some common methods are GANs, diffusion models, VAEs, and neural style transfer [40]. These synthetic images must be detected [41] and ignored before the approach presented in this paper is applied, as this would significantly corrupt the resulting DTs. Next to that, the recognition of artificially generated privacy-relevant images is of great importance to ensure security on social media and to detect and prevent criminal activities that arise from these new technologies more quickly. We are giving high priority to the topic of synthetic data in further research work, as the influence of synthetic images on the quality of DTs is immense.

VII. Conclusion

This study presents valuable findings in the field of VLMs, demonstrating the efficiency of BLIP-based models in capturing and extracting predefined privacy attributes from images using a newly created dataset for privacy analysis. The evaluation shows that the model can extract attributes from images with high accuracy, achieving high micro-average values for person detection, attribute classification, and document analysis. Additionally, this study aimed to investigate whether an exemplar-based method for visual question answering (VQA) can assist in pre-selecting relevant images from a given dataset and extracting specific human attributes. This could be a crucial pre-processing step in our research project ADRIAN, which seeks to extract pertinent attributes for various OSN users and initiate a DT.

IntructBLIP and BLIP-2 were able to identify complex identifications of nudity, tattoos, and physical disability. We were able to show that the BLIP-based models in their original form, i.e., without further fine tuning, can already demonstrate a very good detection rate for the number of people in an image and also shine in the recognition of human attributes. However, in terms of documents, the model is only suitable for identifying specific documents, such as credit cards, and fails to detect country-specific types of documents.

For the future, there are already new models to analyze, such as CogVLM [42]. The promise of models like CogVLM lies in the integration of visual and linguistic data. While traditional VLMs often struggle with the challenge of deeply fusing these two types of data, CogVLM represents a promising advance. It demonstrates how deep fusion of these data can be achieved without compromising the performance of a pre-trained large-scale language model. By being able to identify and label objects in images and accurately extract their coordinates, CogVLM opens up new perspectives for processing and analyzing visual-linguistic information. This could not only improve accuracy and efficiency in existing use cases but also open the door to new applications in artificial intelligence.

Building on these initial findings, we plan to develop richer datasets to enhance the analysis of privacy vulnerabilities. For instance, the VISPR dataset can detect images containing sensitive elements, such as signatures, personal phone numbers, identifiable landmarks, or street signs. After detection, the next step involves extracting this sensitive information, which is crucial for assessing privacy risks. Accurately identifying specific data points, such as residential addresses, workplace locations, and direct contact phone numbers, allows for the assessment of potential privacy threats. This detailed information is particularly valuable in understanding the scope and scale of social engineering attacks. The goal is to leverage this enriched data to develop advanced predictive models that can foresee and neutralize such threats before they materialize. By taking this approach, we aim to proactively protect individuals from privacy breaches and reduce the risks associated with unauthorized data exploitation. This proactive approach is crucial in the digital age, where data privacy and security are of the utmost importance.

ACKNOWLEDGMENT

This research is funded by dtec.bw – Digitalization and Technology Research Center of the Bundeswehr. dtec.bw is funded by the European Union – NextGenerationEU.

REFERENCES

[1] S. Schultenkämper and F. S. Bäumer, “Looking for a needle in a haystack: How can vision-language understanding help to identify privacy-threatening images on the web,” in *Proceedings of the 18th International Conference on Internet and Web Applications and Services, ICIW 2023*. IARIA, 2023, pp. 1–6.

[2] J. Li, D. Li, S. Savarese, and S. Hoi, “BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models,” in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML’23. Journal of Machine Learning Research, 2023.

[3] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi, “InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[4] C. Iordanou, G. Smaragdakis, I. Poese, and N. Laoutaris, “Tracing Cross Border Web Tracking,” in *Proceedings of the Internet Measurement Conference 2018*. New York, NY, USA: Association for Computing Machinery, 2018, pp. 329–342.

[5] M. B. Flinn, C. J. Teodorski, and K. L. Paullet, “Raising Awareness: An examination of embedded GPS data in images posted to the social networking site twitter,” *Issues in Information Systems*, vol. 11, no. 1, pp. 432–438, 2010.

[6] F. S. Bäumer, S. Denisov, Y. Su Lee, and M. Geierhos, “Towards Authority-Dependent Risk Identification and Analysis in Online Networks,” in *Proceedings of the IST-190 Research Symposium (RSY) on AI, ML and BD for Hybrid Military Operations (AI4HMO)*, A. Halimi and E. Ayday, Eds., 10 2021.

[7] F. S. Bäumer, N. Grote, J. Kersting, and M. Geierhos, “Privacy Matters: Detecting Nocuous Patient Data Exposure in Online Physician Reviews,” in *International Conference on Information and Software Technologies*. Springer, 2017, pp. 77–89.

[8] B. R. Barricelli, E. Casiraghi, and D. Fogli, “A Survey on Digital Twin: Definitions, Characteristics, Applications, and Design Implications,” *IEEE Access*, vol. 7, pp. 167 653–167 671, 2019.

[9] R. V. Guha, D. Brickley, and S. Macbeth, “Schema.org: Evolution of Structured Data on the Web,” *Communications of the ACM*, vol. 59, no. 2, pp. 44–51, 01 2016.

[10] N. Pankong, S. Prakancharoen, and M. Buranarach, “A Combined Semantic Social Network Analysis Framework to Integrate Social Media Data,” in *Knowledge and Smart Technology (KST)*, 2012, pp. 37–42.

[11] F. S. Bäumer, J. Kersting, M. Orlikowski, and M. Geierhos, “Towards a Multi-Stage Approach to Detect Privacy Breaches in Physician Reviews.” in *SEMANTICS Posters&Demos*, 2018.

[12] Data Reportal, 01 2022, https://datareportal.com/reports/digital-2022-global-overview-report, retrieved 2024/03/11.

[13] T. Orekondy, B. Schiele, and M. Fritz, “Towards a Visual Privacy Advisor: Understanding and Predicting Privacy Risks in Images,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3706–3715.

[14] J. DeHart, M. Stell, and C. Grant, “Social Media and the Scourge of Visual Privacy,” *Information*, vol. 11, no. 2, 2020.

[15] E. Directive, “95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data,” *Official Journal of the EC, L 281*, vol. 38, pp. 31–50, 11 1995.

[16] Y. Jia *et al.*, “Caffe: Convolutional Architecture for Fast Feature Embedding,” in *Proceedings of the 22nd ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2014, pp. 675–678.

[17] C. Szegedy *et al.*, “Going Deeper with Convolutions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[18] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[19] T. Orekondy, M. Fritz, and B. Schiele, “Connecting Pixels to Privacy and Utility: Automatic Redaction of Private Information in Images,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8466–8475.

[20] S. N. Srihari, A. Shekhawat, and S. W. Lam, *Optical character recognition (OCR)*. GBR: John Wiley and Sons Ltd., 2003, p. 1326–1333.

[21] E. F. Tjong Kim Sang and F. De Meulder, “Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition,” in *Proc. of the Seventh Conf. on Natural Language Learning at HLT-NAACL 2003*, 2003, pp. 142–147.

[22] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, “Fully Convolutional Instance-aware Semantic Segmentation,” in *Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2359–2367.

[23] E. Spyromitros-Xioufis, S. Papadopoulos, A. Popescu, and Y. Kompatsiaris, “Personalized privacy-aware image classification,” in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. New York, United States: Association for Computing Machinery, Inc, 06 2016, pp. 71–78.

[24] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[25] D. Gurari *et al.*, “VizWiz-Priv: A Dataset for Recognizing the Presence and Purpose of Private Visual Information in Images Taken by Blind People,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 939–948.

[26] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in *International Conference on Learning Representations*, 2021.

[27] A. Radford *et al.*, “Learning Transferable Visual Models from Natural Language Supervision,” in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139. PMLR, 07 2021, pp. 8748–8763.

[28] W. Jin, Y. Cheng, Y. Shen, W. Chen, and X. Ren, “A Good Prompt Is Worth Millions of Parameters? Low-resource Prompt-based Learning for Vision-Language Models,” in *Proceedings of the 60th Annual Meeting of the ACL (Volume 1: Long Papers)*. Dublin, Ireland: ACL, 05 2022, pp. 2763–2775.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: ACL, 06 2019, pp. 4171–4186.

[30] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, pp. 436–44, 05 2015.

[31] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, “Judging llm-as-a-judge with mt-bench and chatbot arena,” 2023.

[32] H. Touvron *et al.*, “Llama 2: Open Foundation and Fine-Tuned Chat Models,” 2023.

[33] L. Ouyang *et al.*, “Training language models to follow instructions with human feedback,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 730–27 744, 2022.

[34] Y. Lan, X. Li, X. Liu, Y. Li, W. Qin, and W. Qian, “Improving Zero-shot Visual Question Answering via Large Language Models with Reasoning Question Prompts,” in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 4389–4400.

[35] Y. Qiao, Z. Yu, and J. Liu, “Rankvqa: Answer Re-Ranking For Visual Question Answering,” in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 2020, pp. 1–6.

[36] N. Geifman and E. Rubin, “Towards an age-phenome knowledge-base,” *BMC bioinformatics*, vol. 12, no. 1, pp. 1–9, 2011.

[37] N. G. Jablonski, “The Evolution of Human Skin and Skin Color,” *Annual Review of Anthropology*, vol. 33, no. 1, pp. 585–623, 2004.

[38] P. Frost, “European hair and eye color: A case of frequency-dependent sexual selection?” *Evolution and Human Behavior*, vol. 27, no. 2, pp. 85–103, 2006.

[39] L. Papa, L. Faiella, L. Corvitto, L. Maiano, and I. Amerini, “On the use of Stable Diffusion for creating realistic faces: from generation to detection,” in *2023 11th International Workshop on Biometrics and Forensics (IWBF)*, 2023, pp. 1–6.

[40] L. A. Gatys, A. S. Ecker, and M. Bethge, “Image Style Transfer Using Convolutional Neural Networks,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2414–2423.

[41] H. Agarwal, A. Singh, and R. D, “Deepfake Detection Using SVM,” in *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 2021, pp. 1245–1249.

[42] W. Wang *et al.*, “Cogvlm: Visual expert for pretrained language models,” *arXiv preprint arXiv:2311.03079*, 2024.