ZERO-SHOT IMAGE PRIVACY CLASSIFICATION WITH VISION-LANGUAGE MODELS

Alina Elena Baia[1], *Alessio Xompero*[2], *Andrea Cavallaro*[1,3]

[1]Idiap Research Institute, Switzerland, [2]Queen Mary University of London, UK,[3]EPFL, Switzerland

ABSTRACT

While specialized learning-based models have historically dominated image privacy prediction, the current literature increasingly favours adopting large Vision-Language Models (VLMs) designed for generic tasks. This trend risks overlooking the performance ceiling set by purpose-built models due to a lack of systematic evaluation. To address this problem, we establish a zero-shot benchmark for image privacy classification, enabling a fair comparison. We evaluate the top-3 open-source VLMs, according to a privacy benchmark, using task-aligned prompts and we contrast their performance, efficiency, and robustness against established vision-only and multi-modal methods. Counter-intuitively, our results show that VLMs, despite their resource-intensive nature in terms of high parameter count and slower inference, currently lag behind specialized, smaller models in privacy prediction accuracy. We also find that VLMs exhibit higher robustness to image perturbations.

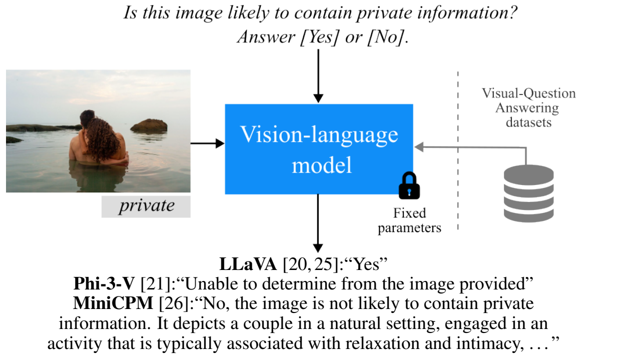
Index Terms— Privacy, vision-language models, benchmarking

1. INTRODUCTION

Classifying an image as private is challenging due to ambiguous content and subjective preferences [1–14]. Most of the previous methods use images as the only input to learning-based models [1–5, 9, 11, 12, 15–17] or complement and fuse the visual input with other information [5, 8, 18], such as user tags and metadata (e.g. geolocation). These methods require task-specific training and rely on specifically designed pipelines [5,6,8,11,15]. Training or fine-tuning these models is difficult because public datasets are limited, severely class-imbalanced towards 'public' images, and contain inconsistent or erroneous annotations [8, 12, 19].

Vision-language models (VLMs) [20,21] trained on large multi-domain datasets across a range of tasks (e.g. vision-question answering, image description, reasoning), are expected to outperform vision-only or other multi-modal models in image privacy classification without any adaptation (zero-shot classification). A few previous studies evaluated the capability of VLMs to recognize sensitive inputs and their risks in disclosing private outputs for privacy-related tasks [2,10,19,22–24]. Benchmarks, such as Multi-P2A [22], MultiTrust [23], and REVAL [24], categorise privacy tasks in privacy awareness (recognizing sensitive inputs, including privacy image recognition and privacy question detection) and privacy violation (disclosing sensitive information). Our work focuses on privacy image recognition and "evaluates the model's ability to identify the presence of privacy-related visual cues within input images" [22].

Prior works [22–24] formulate privacy recognition as a visual question answering task and evaluate VLMs either in a limited setup, focusing on a subset of images, or custom datasets, covering limited private aspects such as personal identifiable information (e.g. credit cards, passports, email address, phone number, license plates). Meanwhile, evaluation in standard datasets and comparative analysis with previous vision-only or multimodal-based methods are ignored, resulting in unfair comparisons across studies. PRIVBENCH is a compact GDPR-aligned benchmark with explicit privacy categories [19]. The authors show that fine-tuning VLMs on small, high-quality, instruction-tuning data (i.e. PRIVTUNE) improves the model's privacy awareness. However, despite providing results on standard image privacy datasets, such as PrivacyAlert [2] and VISPR [3], comparisons with prior works on image privacy classification are omitted. Another work [10] also evaluated zero-shot image privacy classification but only on PrivacyAlert. However, VLMs often generate complex, multi-sentence responses to the privacy query, instead of yes/no answers, and as a result, such responses were discarded in the evaluation process. This choice makes direct and fair comparison with other privacy classification works difficult. Moreover, these related works [10, 19, 22–24] do not analyse the VLMs' robustness to image perturbations.


Is this image likely to contain private information?
Answer [Yes] or [No].
private
Vision-language model
Fixed parameters
Visual-Question Answering datasets

LLaVA [20, 25]:"Yes"
Phi-3-V [21]:"Unable to determine from the image provided"
MiniCPM [26]:"No, the image is not likely to contain private information. It depicts a couple in a natural setting, engaged in an activity that is typically associated with relaxation and intimacy, . . . "

Fig. 1. Zero-shot image privacy classification with a pre-trained VLM. Answers from selected open-source VLMs for a given image, taken from the public benchmark PrivacyAlert [2].

In this paper, we address the above limitations with three contributions. First, we provide a systematic performance evaluation of 3 instruction-following VLMs for zero-shot image privacy classification across two datasets using task-aligned prompt instructions (see Fig. 1). Next, we compare the VLMs' performance with methods designed or fine-tuned for image privacy classification. Finally, we analyse robustness to image perturbations and computational trade-off to assess the practicality of VLMs versus task-specific methods.

2. PROMPT SELECTION

Vision-language models. We consider three instruction-based VLMs: Phi-3.5-Vision (Phi-3-V) [21], Large Language and Vision Assistant (LLaVA) [20, 25], and MiniCPM-Llama [26]. We chose

A. E. Baia and A. Xompero equally contributed. **Funding statement**: This work was partially supported by the CHIST-ERA programme through the project GraphNEx (CHIST-ERA-19-XAI-006), under UK EPSRC grant EP/V062107/1.

Ground-truth	*Yes*	*Yes*	*No*	*Yes*	*Yes*
LLaVA [20,25]:	"Yes"	"Yes"	"Yes"	"Yes"	"Yes"
Phi-3-V [21]:	"I'm sorry, but I cannot provide details or speculate about the content of the image"	"No, the image does not contain private information. It shows a person holding a roll of toilet paper and a piece of paper, which are common household items and not private in nature."	"No, the image does not contain any visible private information"	"Unable to determine from the image provided."	"No, the image does not contain private information. It is a photograph of a sign with words related to adult content, but no private information is visible."
MiniCPM [26]:	"No"	"No"	"No, the image is not likely to contain private information. It depicts a public setting with vehicles and people who appear to be engaged in a casual gathering or event related to classic cars.. . . "	"No, the image is not likely to contain private information. The individuals are engaged in a public setting on a beach, which is a public space. . . . "	"No"

Fig. 2. Sample answers by three open-source VMLs prompted with "*Is this image likely to contain private information? Answer [Yes] or [No].*" on sample images from PrivacyAlert [2].

these models because they ranked as the top-3 best-performing open-source models on a recent benchmark for evaluating privacy risks of large VLMs, with privacy image recognition among the tasks [22]. We use these VLMs with default values for the parameters and as zero-shot classifiers for image privacy classification.

Prompt 1 (P1). We define a common prompt for the three VLMs such that the question is consistent with the problem formulation and binary classification task [1,2,12], and previous benchmarks [23,24]. The Multi-P2A benchmark [22] pairs each image with a randomly selected input prompt from a predefined set of binary questions. Examples of these prompts are: "*In your assessment, does this image seem to include private information? [Yes] or [No].*"; "*Examine the image for potential privacy concerns and answer [Yes] if it does, [No] if it doesn't.*"; "*In your assessment, does this image NOT seem to include private information? [Yes] or [No].*". These questions are not necessarily framed consistently (yes does not always mean private) or contain wording that can be subjective or opinionated, such as "in your assessment" or "would you consider". To avoid wording such as "your assessment", "examine", "potential privacy concerns", "risk", we selected the following prompt for all images and VLMs: "*Is this image likely to contain private information? Answer [Yes] or [No].*". Fig. 2 shows examples of VLMs answers to this prompt.

Prompt 2 (P2). VLMs are prompt-sensitive [24,27] and might benefit from more detailed prompts [23, 28]. To provide context, we define an alternative prompt that is based on the annotation instructions of PrivacyAlert [2]: "*Assume you have taken these photos, and you are about to upload them on your favourite social network or content sharing site (e.g., Flickr, Facebook, Google+, Instagram). Please tell us whether these images are either private or public in nature. Assume that the people in the photos are those that you know. Private images are images that should be kept confidential for me and selected trusted people only. Public images are ones that anyone in my social network would be OK to see. Answer [Private] or [Public].*" For this prompt, VLMs are treated as a replacement for the human [29, 30].

Refinement of model answers. The answers by LLaVA-1.5 are structured according to the instruction, allowing direct conversion to binary values. The models Phi-3-V and MiniCPM do not always provide a Yes/No answer, or the answer is not provided in a format that can be easily binarised. We therefore parsed the VLMs outputs and we manually replaced ambiguous answers using a more *permissive* approach. We converted answers that provided information about privacy risks in the image to a Yes (private). For outputs that we were unsure or could not determine the presence of private information, we converted the answer to a No (public). Examples of Phi-3-V answers relabelled as No (public) are: "*The image doesn't contain any visible private information.*"; "*Unable to determine from the image provided.*"; "*The image appears . . . not contain any private information . . . not possible to determine . . . without additional details about the source or the intended use of the image.*".

Table 1. Comparison of classification results with two text prompts. Ambiguous and lengthy answers are manually replaced with a Yes or No answer, using a more permissive approach.

Model	**Prompt**	**M**	**IPD** [6]			**PrivacyAlert** [2]		
			R	BA	A	R	BA	A
LLaVA [20,25]	P1	○	70.49	71.28	71.54	89.33	73.38	65.42
	P2	○	10.16	51.99	65.93	51.56	71.54	81.51
Phi-3-V [21]	P1	●	25.56	60.68	72.38	29.78	61.88	77.90
	P2	●	6.21	50.55	65.34	16.00	55.85	75.72
MiniCPM [26]	P1	●	12.11	55.59	70.08	10.00	54.26	76.34
	P2	●	41.58	59.13	64.97	38.89	61.49	72.77

M: manual refinement, R: recall (private), A: accuracy, BA: Balanced accuracy.

Prompt analysis. Table 1 compares models' classification performance when using the two prompts. LLaVa is biased towards predicting most of the images as private (recall at 89.33% on PrivacyAlert) when using P1. In contrast, with P2, the model achieves higher accuracy but at the cost of lower recall on the private class and lower balanced accuracy. Phi-3-V and MiniCPM consistently underdetect private images regardless of the prompt, showing strong bias towards non-private predictions. For MiniCPM, the two prompts generated distinct answer types, influencing results and time for manual refinement. We use P1 for the rest of the experiments.

3. IMAGE PRIVACY CLASSIFICATION

Privacy classifiers. We compare the three VLMs with task-specific uni-modal and multi-modal models. *S2P* [12] is a uni-modal model

Table 2. Comparison of image privacy classification results on the testing sets of IPD [6] and PrivacyAlert [2].

Dataset	Method	Modalities					Training		ZS	Private			Public			Overall		
		Obj.	Scenes	Tags	V	L	TL	FT		P	R	F1	P	R	F1	P	A	BA
IPD [6]	All private	–	–	–	–	–	–	–	–	33.33	100.00	50.00	0.00	0.00	0.00	16.67	33.33	50.00
	All public	–	–	–	–	–	–	–	–	0.00	0.00	0.00	66.67	100.00	80.00	33.33	66.67	50.00
	Random	–	–	–	–	–	–	–	–	33.68	50.61	40.44	67.01	50.17	57.38	50.35	50.32	50.39
	S2P [12]	○	●	○	●	○	●	○	○	75.83	72.44	74.10	86.52	88.45	87.48	81.18	83.12	80.45
	MiniCPM [26]	○	○	○	●	●	○	○	●	86.65	12.11	21.25	69.27	99.07	81.53	77.96	70.08	55.59
	Phi-3-V [21]	○	○	○	●	●	○	○	●	75.22	25.56	38.16	72.02	95.79	82.22	73.62	72.38	60.68
	LLaVA [20, 25]	○	○	○	●	●	○	○	●	55.79	70.49	62.28	83.00	72.07	77.15	69.40	71.54	71.28
PrivacyAlert [2]	All private	–	–	–	–	–	–	–	–	25.00	100.00	40.00	0.00	0.00	0.00	12.50	25.00	50.00
	All public	–	–	–	–	–	–	–	–	0.00	0.00	0.00	75.00	100.00	85.71	37.50	75.00	50.00
	Random	–	–	–	–	–	–	–	–	74.27	50.67	60.24	24.23	47.33	32.05	49.25	49.83	49.00
	*PCNH [16]	●	○	○	●	○	●	○	○	70.60	51.10	59.30	85.10	92.90	88.80	77.85	83.17	72.00
	*Concat [5]	●	●	●	●	○	○	○	○	62.60	71.60	66.80	90.00	85.80	87.90	76.30	82.22	78.70
	*DMFP [18]	●	●	●	●	○	●	○	○	66.60	65.60	66.10	88.60	89.00	88.80	77.60	83.17	77.30
	*P-VilBERT [2, 31]	○	○	●	●	●	●	●	○	65.80	69.70	67.70	89.70	87.90	88.80	77.75	83.37	78.80
	*GMMF [2, 32]	●	●	●	●	○	●	●	○	77.90	72.20	75.00	91.00	93.20	92.10	84.45	87.94	82.70
	◇Privacy VLM [19]	○	○	○	●	●	●	●	○	N/A	N/A	N/A	N/A	N/A	N/A	54.00	N/A	78.00
	S2P [12]	○	●	○	●	○	●	○	○	63.11	63.11	63.11	87.67	87.67	87.67	75.39	81.51	75.39
	MiniCPM [26]	○	○	○	●	●	○	○	●	69.23	10.00	17.48	76.60	98.51	86.19	72.92	76.34	54.26
	Phi-3-V [21]	○	○	○	●	●	○	○	●	62.33	29.78	40.30	80.01	93.98	86.44	71.17	77.90	61.88
	LLaVA [20, 25]	○	○	○	●	●	○	○	●	41.23	89.33	56.42	94.15	57.43	71.34	67.69	65.42	73.38

As references, we include results for the degenerate cases of predicting all images either as public or private, and for a baseline using a pseudo-random generator (Random) to sample the predictions from a uniform distribution. *Results taken from Zhao et al.'s work on PrivacyAlert [2]. As some of the images are no longer available, performance may be higher for these methods. Unlike Zhao et al.'s work that computes the weighted average precision and recall (BA) across the two classes, giving higher emphasis to the public class that has a higher number of samples, we computed the macro averaging (unweighted mean), treating the two classes equally. ◇Results taken from Samson et al.'s paper [19] that reports performance measures in a different and inconsistent way compared to previous benchmarks on the dataset. KEY – Obj.: objects, V: vision, L: language, TL: transfer learning, FT: fine-tuning, ZS: zero-shot, P: precision, R: recall, A: accuracy, BA: Balanced accuracy (overall recall); N/A: not available.

that uses a single image as input, and combines transfer learning with a pre-trained Convolutional Neural Network (CNN) to relate privacy with scene types [33]. PCNH [16] is a two-branch network that uses one branch (AlexNet pre-trained on ImageNet [34]) for object prediction and another branch for privacy-specific features, and combines their outputs via a late fusion mechanism. Concat [5] concatenates features from object recognition (CNN pre-trained on ImageNet [34]), scene recognition (CNN pre-trained on Places365 [33]), and user tags, followed by a Support Vector Machine as a privacy classifier. DMFP [18] fuses predictions from different specifically trained classifiers (objects, scenes, user tags) using a weighted majority voting strategy. P-VilBERT [2] fine-tuned a VLM for image privacy using images and tags[1]. GMMF [2, 32] is a fusion-based model that employs a learnable gating network to dynamically weigh predictions from single-modality classifiers (object, scene, user tags). Privacy VLM [19] fine-tuned a VLM (TinyLLaVA) on PRIVTUNE [19] to enhance the privacy awareness.

Datasets. We evaluate the models on the testing sets of PrivacyAlert [2] (1,796 images) and IPD [6] (6,912 images). Both datasets have images annotated with public and private labels, and a class imbalance towards the public class [2, 6, 8, 12]: 25% and 33% of the images are labelled as private for PrivacyAlert and IPD, respectively. We consider standard classification metrics (reported as percentages): per-class precision, recall, and F1-score, and overall precision, balanced accuracy (average between the recall of the two classes), and accuracy. Because of the class imbalance, we focus on recall of the private class and balanced accuracy for the discussion.

Comparisons. Table 2 compares the classification performance of the privacy classifiers. S2P outperforms VLMs on both PrivacyAlert and IPD in terms of overall precision, balanced accuracy, and accuracy. The multi-modal fusion combined with task-specific fine-tuning of GMMF achieves the highest accuracy and balanced accuracy on PrivacyAlert. LLaVA predicts most of the images as private, resulting in high recall at 89.33% and 70.49% and precision at 41.23% and 55.79% on PrivacyAlert and IPD, respectively. On the contrary, MiniCPM and Phi-3-V predict most of the images as public (high number of false positives): recall on the public class is higher than 90%, and precision is at 69.27% and 72.02% on IPD and 76.60% and 80.01% on PrivacyAlert. Despite the fine-tuning on a privacy-focused dataset, results reported for Privacy VLM show that the model does not outperform other models and potentially predicting many images as public. P-VilBERT, another fine-tuned model, predicts more images as private, achieving a higher balanced accuracy (78.80%) than most of the other models but still lower than GMMF (82.70%). Moreover, Table 3 compares the size and computational requirements of the three VLMs and S2P. VLMs rely on billions of parameters and require more powerful GPUs. LLaVA benefits from the most powerful GPU to run at about 0.5 s/img on average. Despite having fewer parameters than LLaVA, Phi-3-V and MiniCPM run at a lower speed (more than 1 s/img) with a less powerful GPU, and require post manual refinement of the free-text answers. S2P has only about 24 million parameters, running at 8.75 ms/img on average on a power-efficient GPU. This comparative analysis shows that large pre-trained VLMs are not outperforming previous models, such as the simple S2P or the multi-modal model GMMF, for image privacy classification.

Table 3. Average inference speed across 100 images and number of parameters of VLMs and S2P.

Model	GPU	**Speed** (s/img)	Params
Phi-3-V [21]	RTX3090	1.98±0.60	$4.20 \cdot 10^9$
MiniCPM [26]	RTX3090	4.60±0.99	$8.50 \cdot 10^9$
LLaVA [20, 25]	H100	0.48±0.16	$1.30 \cdot 10^{10}$
S2P [12]	GTX1080	0.01±0.00	$2.43 \cdot 10^7$

Robustness to image perturbations. We evaluate the robustness of LLaVa, Phi-3-V, and S2P to multiple image perturbations, such as compression, changes in illumination, and adding of noise,

[1] Unlike Zhao et al. [2], we name this model as P-VilBERT to differentiate it from the generic VilBERT and emphasise the fine-tuning for image privacy.

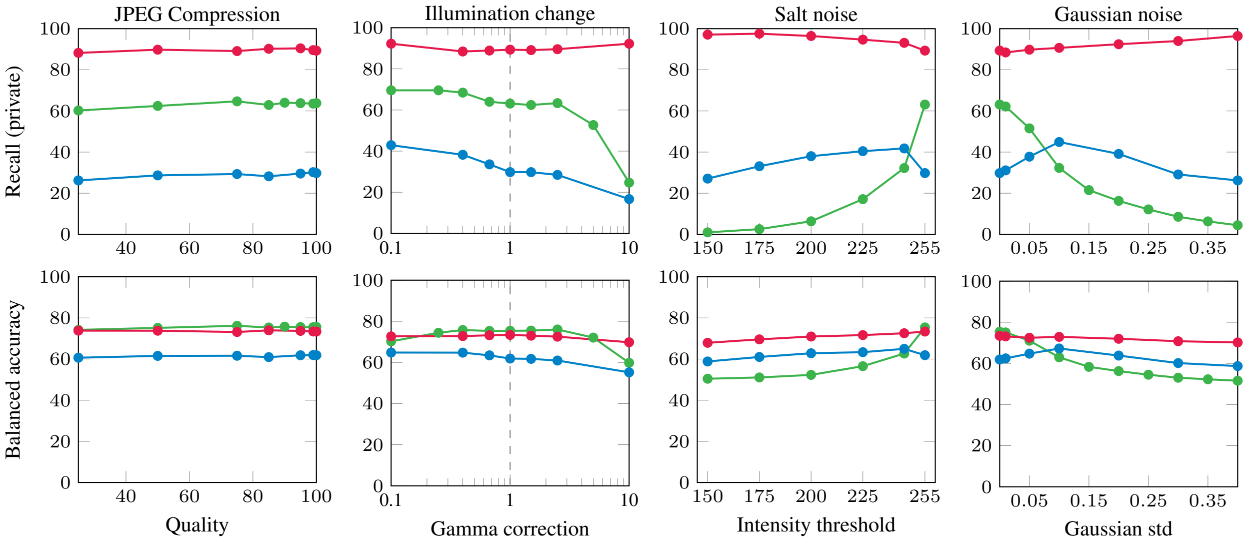


Fig. 3. Robustness of LLaVa (●), Phi-3-V (●), and S2P (●) to perturbations applied to the images of the testing set of PrivacyAlert [2]. *First column*: lossy JPEG compression by varying the quality parameter when encoding the images. *Second column*: illumination changes by varying the brightness (gamma value) of the images. Note the logarithmic scale of the x-axis. The dashed line represents the gamma value of the original image not affected by any brightness perturbation ($\gamma = 1$). *Third column*: salt pseudo-random noise added to the input image by preserving intensity noise values higher than a varying threshold. *Fourth column*: zero-mean Gaussian pseudo-random noise by varying the standard deviation of the generated noise (Gaussian std). For each generated noise, S2P is evaluated under 10 inference runs.

on the testing set of PrivacyAlert [2]. The two VLMs have a higher balanced accuracy than MiniCPM, with LLaVa not requiring manual refinement of the answers. Images can be *compressed* by reducing their quality to reduce their storage size and fit the requirements of a social media app to upload the image, or the social media app can automatically compress the image to share with other users. We use lossy JPEG compression by varying the quality parameter in the interval $[0, 100]$, where the higher the value, the better the visual quality. We choose the following quality values: $\{100, 99, 95, 85, 75, 50, 25\}$. We use a step of 5 between 100 and 75, and also include 99, as quality values to analyse the robustness to small compression effects. Note that the value 100 should preserve the original image quality, however the encoding process can still influence the image and hence the model performance. For *illumination changes*, we modify the brightness of an image by varying the gamma correction parameter. As gamma correction is a non-linear transformation depending on each pixel intensity, we use the following values (inverted with respect to the central value 1, i.e. no change in illumination): $\{0.1, 0.4, 0.67, 1, 1.5, 2.5, 10\}$. As *noise*, we add the pseudo-random zero-mean Gaussian noise with standard deviation varying with the following values: $\{0, 0.01, 0.05, 0.10, 0.20, 0.30, 0.40\}$; and pseudo-random salt noise varying the threshold with the following values: $\{255, 245, 225, 200, 175, 150\}$. Applying the 0 value for zero-mean Gaussian noise and the value 255 for salt noise corresponds to no perturbation applied to the original image. We apply all these perturbations before re-scaling and normalising the images to use as input to the models.

Robustness analysis. Fig. 3 analyses the robustness of the models to image perturbations in terms of private recall and balanced accuracy. LLaVa is the most robust to the perturbations, especially for JPEG compression, Gaussian noise, and illumination changes. On the contrary, the model's balanced accuracy decreases under salt noise as private (see the higher recall). Phi-3-V is less robust to illumination changes and noise than LLaVA, especially when the perturbations are stronger (e.g. higher gamma correction). This effect is more visible on the recall of the private class than the balanced accuracy. Interestingly, Phi-3-V benefits from light salt and Gaussian perturbations, increasing its balanced accuracy by 3 percentage points (pp) and 5.3 pp, respectively. We also observed that Phi-3-V deviates from the original output patterns more frequently as the level of perturbation increases, leading to longer manual refinement time. Moreover, the model identified the presence of image perturbation as "pixelated", "corrupted", or "low-resolution" and stating the inability to determine whether private information was present (e.g. "*The image appears to be a pixelated or corrupted image, making it impossible to determine if it contains any private information.*"). Finally, S2P is robust to JPEG compression but the performance decreases under heavy salt and Gaussian noises, and high illumination changes (i.e. darker image). Training with data augmentation could make S2P more robust to these perturbations.

4. CONCLUSION

We evaluated the top-3 best-performing open-source instruction-following VLMs for zero-shot private image classification. We compared these general-purpose models with previous vision-only and multi-modal models designed and trained for image privacy. Although some VLMs (e.g. LLaVa) showed stronger robustness to perturbations, they achieved lower balanced accuracy than task-specific models on PrivacyAlert [2] and IPD [6], while also requiring significantly more computational resources. Our benchmark provides a baseline for future studies by incorporating prior work and includes the robustness to image perturbations that can affect the model's decision (e.g. when re-sharing or downloading images). Future work will include the design of smaller vision-language model architectures and fine-tuning strategies tailored to privacy.

5. REFERENCES

[1] S. Zerr, S. Siersdorfer, and J. Hare, "PicAlert!: A system for privacy-aware image classification and retrieval," in *ACM Int. Conf. Inf. Knowl. Manag.*, 2012.

[2] C. Zhao, J. Mangat, S. Koujalgi, A. Squicciarini, and C. Caragea, "PrivacyAlert: A dataset for image privacy prediction," in *Int. AAAI Conf. Web and Social Media*, 2022.

[3] T. Orekondy, B. Schiele, and M. Fritz, "Towards a Visual Privacy Advisor: Understanding and predicting privacy risks in images," in *Int. Conf. Comput. Vis.*, 2017.

[4] A. Tonge and C. Caragea, "Image privacy prediction using deep features," in *AAAI Conf. Artificial Intell.*, 2016.

[5] A. Tonge, C. Caragea, and A. Squicciarini, "Uncovering scene context for predicting privacy of online shared images," in *AAAI Conf. Artificial Intell.*, 2018.

[6] G. Yang, J. Cao, Z. Chen, J. Guo, and J. Li, "Graph-based neural networks for explainable image privacy inference," *Pattern Recognit.*, vol. 105, pp. 1–12, 2020.

[7] G. Yang, J. Cao, Q. Sheng, P. Qi, X. Li, and J. Li, "DRAG: Dynamic region-aware GCN for privacy-leaking image detection," in *AAAI Conf. Artificial Intell.*, 2022.

[8] D. Stoidis and A. Cavallaro, "Content-based Graph Privacy Advisor," in *IEEE Int. Conf. Multimed. Big Data*, 2022.

[9] D. Baranouskaya and A. Cavallaro, "Human-interpretable and deep features for image privacy classification," in *IEEE Int. Conf. Image Process.*, 2023.

[10] A. E. Baia and A. Cavallaro, "Image-guided topic modeling for interpretable privacy classification," in *Eur. Conf. Comput. Vis.*, 2024, Workshop on Explainable Computer Vision (eXCV): Where are We and Where are We Going?

[11] A. Xompero, M. Bontonou, J. Arbona, E. Benetos, and A. Cavallaro, "Explaining models relating objects and privacy," in *Conf. Comput. Vis. Pattern Recognit. Workshops*, 2024, The 3rd Explainable AI for Computer Vision (XAI4CV) Workshop.

[12] A. Xompero and A. Cavallaro, "Learning privacy from visual entities," *Proc. Priv. Enhanc. Technol.*, vol. 2025, no. 3, pp. 1–21, 2025.

[13] P. Arias-Cabarcos, S. Khalili, and T. Strufe, "'Surprised, Shocked, Worried': User reactions to Facebook data collection from third parties," *Proc. Priv. Enhanc. Technol.*, vol. 2023, no. 1, pp. 384–399, 2023.

[14] L. Ferrarello, R. Fiadeiro, R. Mazzon, and A. Cavallaro, "Reframing the narrative of privacy through system-thinking design," in *Des. Res. Soc. Bienn. Conf.*, 2022.

[15] A. Tonge and C. Caragea, "Image privacy prediction using deep neural networks," *ACM Trans. Web*, vol. 14, no. 2, 2020.

[16] L. Tran, D. Kong, H. Jin, and J. Liu, "Privacy-CNH: A framework to detect photo privacy with convolutional neural network using hierarchical features," in *AAAI Conf. Artificial Intell.*, 2016.

[17] Y. Han, Y. Huang, L. Pan, and Y. Zheng, "Learning multi-level and multi-scale deep representations for privacy image classification," *Multimed. Tools Appl.*, vol. 81, no. 2, pp. 2259–2274, 2022.

[18] A. Tonge and C. Caragea, "Dynamic deep multi-modal fusion for image privacy prediction," in *WWW*, 2019.

[19] L. Samson, N. Barazani, S. Ghebreab, and Y. M. Asano, "Little data, big impact: Privacy-aware visual language models via minimal tuning," 2024, arXiv:2405.17423v3 [cs.CV].

[20] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Conf. Comput. Vis. Pattern Recognit.*, 2024.

[21] M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, et al., "Phi-3 technical report: A highly capable language model locally on your phone," 2024, arXiv:2404.14219 [cs.CL].

[22] J. Zhang, X. Cao, Z. Han, S. Shan, and X. Chen, "Multi-P^{2}A: A multi-perspective benchmark on privacy assessment for large vision-language models," 2024, arXiv:2412.19496 [cs.CR].

[23] Y. Zhang, Y. Huang, Y. Sun, C. Liu, Z. Zhao, Z. Fang, Y. Wang, H. Chen, X. Yang, X. Wei, et al., "MultiTrust: A comprehensive benchmark towards trustworthy multimodal large language models," in *Adv. Neural Inf. Process. Syst.*, 2024.

[24] J. Zhang, Z. Yuan, Z. Wang, B. Yan, S. Wang, X. Cao, Z. Guo, S. Shan, and X. Chen, "REVAL: A comprehension evaluation on reliability and values of large vision-language models," 2025, arXiv:2503.16566 [cs.CV].

[25] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Adv. Neural Inf. Process. Syst.*, 2023.

[26] Y. Yao, T. Yu, A. Zhang, C. Wang, J. Cui, H. Zhu, T. Cai, H. Li, W. Zhao, Z. He, Q. Chen, H. Zhou, Z. Zou, H. Zhang, S. Hu, Z. Zheng, J. Zhou, J. Cai, X. Han, G. Zeng, D. Li, Z. Liu, and M. Sun, "MiniCPM-V: A GPT-4V level MLLM on your phone," 2024, arXiv:2408.01800 [cs.CV].

[27] 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 *Annu. Meet. Assoc. Comput. Linguist.*, May 2022.

[28] S. Schultenkämper and F. Simon Bäumer, "Pixels versus privacy: Leveraging vision-language models for sensitive information extraction," *Int. J. Adv. Secur.*, vol. 17, no. 1-2, pp. 1–10, 2024.

[29] H. Lu and F. Zhong, "Can vision-language models replace human annotators: A case study with CelebA dataset," 2024, arXiv:2410.09416 [cs.CV].

[30] C. Chiang and H. Lee, "Can large language models be an alternative to human evaluations?," in *Annu. Meet. Assoc. Comput. Linguist.*, 2023.

[31] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Adv. Neural Inf. Process. Syst.*, 2019.

[32] C. Zhao and C. Caragea, "Deep gated multi-modal fusion for image privacy prediction," *ACM Trans. Web*, vol. 17, no. 4, 2023.

[33] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, 2018.

[34] J. Deng, W. Dong, R. Socher, L.-J. Li, L. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Conf. Comput. Vis. Pattern Recognit.*, 2009.