Privacy-Aware Visual Language Models

Anonymous authors
Paper under double-blind review

Abstract

As Visual Language Models (VLMs) become increasingly embedded in everyday applications. Ensuring they can recognise and appropriately handle privacy-sensitive content is thus essential to protect users. To this end, we conduct a comprehensive evaluation of ten state-of-the-art VLMs and identify limitations in their understanding of visual privacy. However, existing privacy-related datasets often suffer from label inconsistencies, limiting their reliability. To address this, we introduce two compact, high-quality benchmarks, PrivBench and PrivBench-H, that focus on commonly recognised visual privacy categories aligned with the General Data Protection Regulation (GDPR). Additionally, we present PrivTune, an instruction-tuning dataset specifically curated to improve privacy sensitivity. We obtain a Privacy VLM by fine-tuning an off-the-shelf VLM on only 100 samples from PrivTune, which leads to substantial gains on all benchmarks, surpassing even GPT-4, while maintaining strong performance on other tasks. Our findings show that privacy-awareness in VLMs can be substantially improved with minimal data and careful dataset design, setting the stage for safer, more privacy-aligned AI systems.

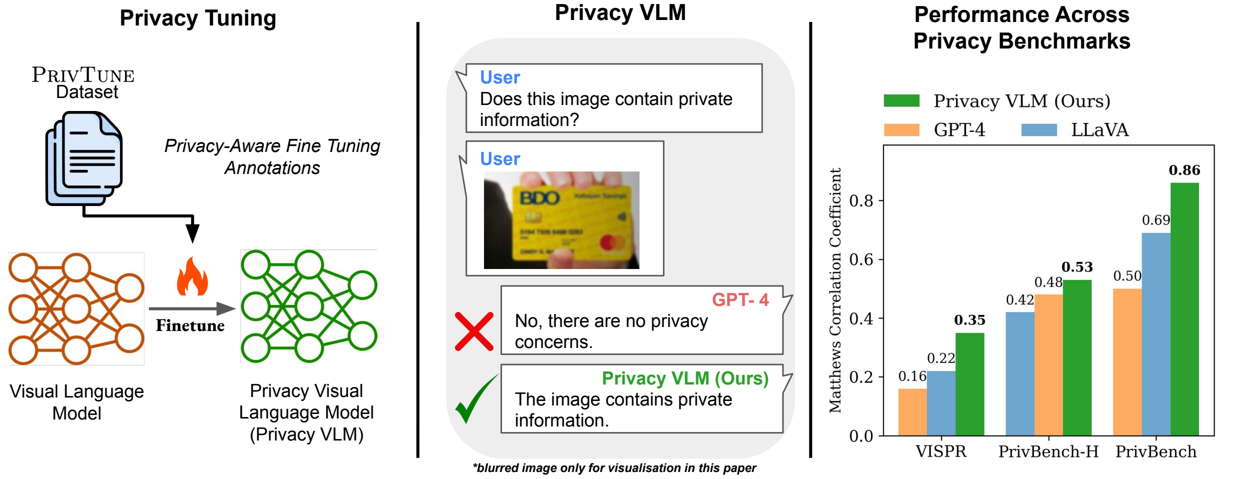
1 Introduction

Rapid advancements in Large Language Models (LLMs) have led to the development and widespread adoption of a new generation of Visual Language Models (VLMs) (Alayrac et al., 2022; Li et al., 2022; Liu et al., 2024b; Li et al., 2025; Liu et al., 2024a; Bavishi et al., 2024; Team et al., 2023; Achiam et al., 2023) that can process both image and text data. These models enable virtual assistants that assist with automated image reasoning tasks in the real world. However, with the increasing deployment of VLMs, the volume of data shared with these interactive agents is expected to grow significantly, raising questions about how to keep these interactions safe.

To this end, key regulatory frameworks like the European Union's General Data Protection Regulation (GDPR) (GDPR, 2016) and the proposed EU AI Act (European Commission, 2021) highlight the critical importance of privacy protection in AI. As VLMs integrate into everyday technologies, from smartphones to social media, compliance with these regulations becomes essential for responsibly handling sensitive information. In turn, a new family of 'privacy-aware' VLMs can serve as safety tools to make users aware of their data's sensitivity and prevent the inclusion of sensitive data, especially for minors or unaware users, or be used to help clean datasets before release.

While numerous benchmark datasets (Hudson & Manning, 2019; Hartvigsen et al., 2022; Zhao et al., 2018; Lin et al., 2021; Goyal et al., 2017; Li et al., 2023; Tömekçe et al., 2024) assess VLMs and LLMs for quality, bias, truthfulness, and toxicity, the essential domain of privacy awareness in visual contexts remains largely unexplored. To address this gap, we evaluate privacy awareness across existing datasets on ten state-of-the-art VLMs. Human evaluations reveal significant label noise within several datasets. Consequently, we introduce two high-quality benchmarks, PrivBench and PrivBench-H. These benchmarks focus on commonly recognised private categories aligned with the GDPR.

Our evaluation of several state-of-the-art VLMs generally reveals limitations in accurately identifying privacy-sensitive images. Motivated by this insight, we introduce the PrivTune dataset, which contains privacy conversations labelled into 8 categories. This dataset is explicitly designed to enhance the privacy awareness



Does this image contain private information?

No, there are no privacy concerns.

The image contains private information.

*blurred image only for visualisation in this paper

Figure 1: **Privacy-tuning Overview and Benchmark Results.** From left to right: (i) our privacy-tuning pipeline, (ii) a qualitative example from the tuned model, (iii) Matthews Correlation Coefficient (MCC↑) comparison of our Privacy VLM with state-of-the-art VLMs on PrivBench and VISPR.

of VLMs and comprises high-quality human annotations. Example images from PrivTune are illustrated in Figure 2. We employ this dataset for privacy-tuning, *i.e.*, fine-tuning VLMs to improve their understanding and management of visual privacy concerns (see Figure 1). We demonstrate substantial improvements in the model's ability to identify and address privacy-related content, generalising robustly across all privacy benchmarks. Moreover, we show that privacy-tuning is remarkably efficient: fine-tuning on just 100 images from our PrivTune is sufficient to achieve an 85% F1 score on PrivBench. Figure 1 illustrates the process of privacy-tuning, showcases qualitative outcomes from our privacy-tuned models, and provides quantitative comparisons of privacy perception between our model and other state-of-the-art VLMs on different privacy benchmarks. Our Privacy VLM, obtained from privacy-tuning an off-the-shelf VLM, consistently outperforms leading state-of-the-art VLMs on privacy image datasets, including prominent models such as LLaVA (Liu et al., 2024a), CogVLM (Wang et al., 2024), and GPT-4 (Achiam et al., 2023), while minimally impacting performance on other conventional benchmarks.

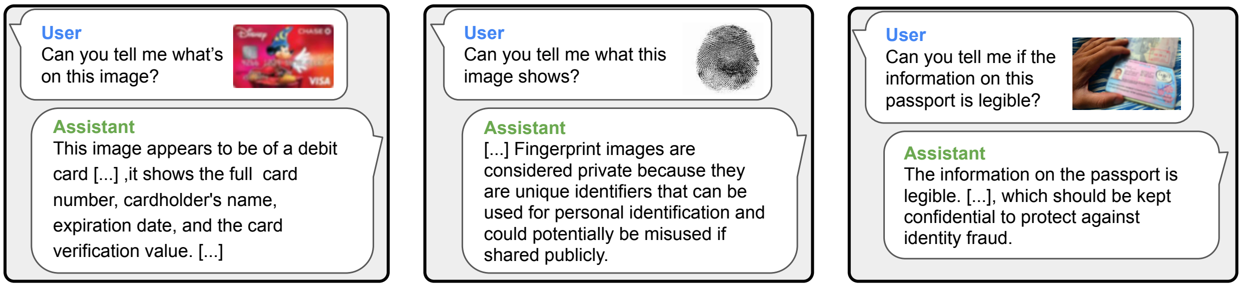
Privacy varies across cultures and contexts and is an ever-evolving concept, making it particularly interesting to study whether privacy awareness can extend beyond a fixed set of categories. Our PrivBench and PrivTune focus on a carefully chosen subset of commonly accepted private classes, enabling us to test on broader benchmarks which adopt a wider privacy spectrum. For example, we train a VLM on only license plates and faces and obtain a high F1 on credit cards. Moreover, applying our privacy-tuning pipeline to a million-scale computer vision corpus demonstrates its practical value, automatically flagging sensitive content across diverse real-world contexts without explicit exposure to those environments. These experiments collectively demonstrate the ability of privacy-aware VLMs to adapt to unseen private attributes.

Our work makes three key contributions toward Privacy-Aware Visual Language Models:

- We introduce two human-curated high-quality benchmarks, PrivBench and PrivBench-H, enabling assessment of privacy-awareness in VLMs.

- Through comprehensive evaluations, we reveal critical shortcomings in current VLMs' capacity to accurately recognise privacy-sensitive visual content.

- We introduce PrivTune and demonstrate that privacy-tuning VLMs using this dataset significantly enhances their privacy awareness without compromising their performance on standard tasks.



User
Can you tell me what's on this image?

Assistant
This image appears to be of a debit card [...] ,it shows the full card number, cardholder's name, expiration date, and the card verification value. [...]

User
Can you tell me what this image shows?

Assistant
[...] Fingerprint images are considered private because they are unique identifiers that can be used for personal identification and could potentially be misused if shared publicly.

User
Can you tell me if the information on this passport is legible?

Assistant
The information on the passport is legible. [...], which should be kept confidential to protect against identity fraud.

Figure 2: **Examples from the PrivTune dataset**: This figure shows sample privacy-aware dialogues, each paired with human ground-truth labels and GPT-4-generated conversations. Images are blurred for visualisation.

2 Related Work

Sensitive Attribute Inference Beyond memorisation and data leakage (Neel & Chang, 2023; Carlini et al., 2022; Brown et al., 2021; Tirumala et al., 2022), recent research have highlighted LLMs' capability to infer sensitive attributes such as age, gender, and location during inference (Staab et al., 2024). Subsequent work extended the scope to VLMs and showed that they can infer private attributes from visual content (Tömekçe et al., 2024). However, their work focused on extracting locations and other private attributes from social media imagery, whereas we aim to measure whether models have an understanding of private categories in images.

Visual Privacy Datasets Several image privacy datasets, such as Biv-Priv (Sharma et al., 2023), PrivacyAlert (Zhao et al., 2022), *PicAlert* (Zerr et al., 2012), VISPR (Orekondy et al., 2017), and VizWiz-Priv (Gurari et al., 2019), have been developed to support classifiers targeting privacy-sensitive content. VizWiz-Priv employs blurring to protect privacy, thus limiting its effectiveness in evaluating a model's detailed privacy comprehension. Our analysis revealed significant labelling noise within PrivacyAlert and Biv-Priv, a problem we empirically document. Additionally, Biv-Priv uses staged props distributed among only 26 individuals, constraining diversity and realism. Datasets from autonomous driving, namely *PP4AV* (Trinh et al., 2023) and *ADD* (Wu et al., 2023), specialise in detecting and anonymising faces and license plates in street scenes, thus lacking a comprehensive taxonomy of general-purpose privacy attributes. VISPR resembles our dataset, labelling private attributes to predict user-specific privacy risks. Unlike VISPR, where images may provide only partial identifying cues (e.g., a hand displaying skin tone), our dataset aims that each image is explicitly traceable to an individual, offering a robust testbed for identity-level privacy detection. Also, our datasets do not include debatable classes such as ethnic clothing, landmarks, or car ownership; instead, we utilise commonly accepted private classes.

High-Quality Evaluation Datasets Our benchmarks align with a tradition of developing compact, high-quality evaluation datasets designed for tracking progress. Notable examples include reannotations of ImageNet (Deng et al., 2009) in ImageNetV2 (Recht et al., 2019), and CIFAR-10 (Krizhevsky et al., 2009) via CIFAR-10.1 (Recht et al., 2018) and CIFAR-10H (Peterson et al., 2019). Research has shown that evaluating LLMs using smaller, carefully annotated datasets (even as few as 100 samples) can provide reliable insights (Polo et al., 2024). Our proposed benchmarks similarly leverage high-quality annotations to measure noise-free and precise insights.

Safety in LLMs Prior research has identified multiple safety challenges in LLMs, such as truthfulness, jailbreaking, hallucinations, and biases (Zou et al., 2023; Yong et al., 2023; Yuan et al., 2023; Gallegos et al., 2024; Huang et al., 2025). Correspondingly, several benchmarks were introduced to systematically address these concerns (Zhao et al., 2018; Nangia et al., 2020; Lin et al., 2021; Askell et al., 2021; Hartvigsen et al., 2022; Gehman et al., 2020). In contrast to earlier work, our study focuses explicitly on whether models appropriately recognise and manage privacy-sensitive content, thereby addressing a critical gap in current

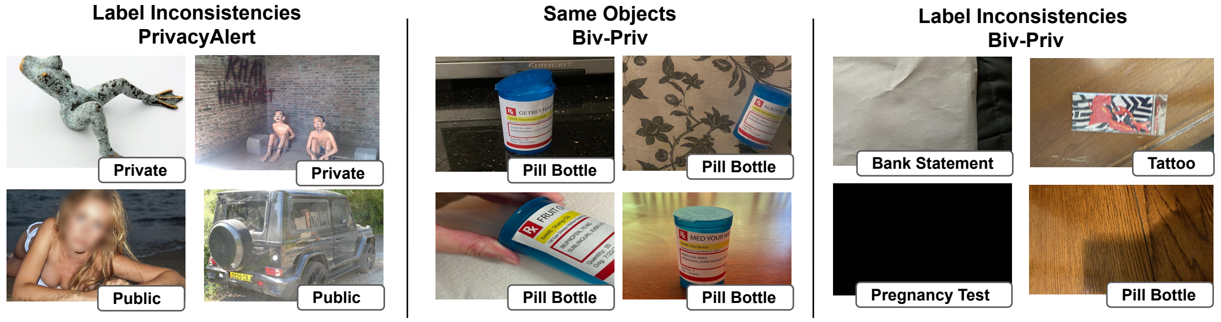


Figure 3: **Common Labelling Errors and Limited Diversity in Biv-Priv and PrivAlert Datasets**. Left: PrivAlert mislabels images containing people (blurring: ours) as non-private, while labeling dolls and paintings as private. Center: Repeated objects (17 of 56 images) within the 'pill bottle' class of Biv-Priv, illustrating limited diversity. Right: Biv-Priv labeling errors including black screens, empty sheets, and object-free images incorrectly labeled as private. Labels assigned by datasets appear at the bottom-right of each image.

safety evaluations. Our training dataset uniquely targets the alignment of VLMs to recognize and respect visual privacy.

3 Quality and Consistency of Privacy Datasets

In this section, we assess three commonly used image privacy datasets with qualitative analysis. Furthermore, we describe the human evaluation that quantitatively measures the quality of the labels.

Biv-Priv Dataset Within this dataset (Sharma et al., 2023), we identified significant labeling inconsistencies. Among false negatives, we discovered 60 images containing empty white papers incorrectly labeled as private documents such as doctor's prescriptions, medical records, or bank statements (see Appendix H). Additionally, we found 28 images depicting completely black screens across multiple classes. Combining only these inconsistencies already sums up to 8.8% of the private images. Furthermore, we observed that many images contain the exact same objects, questioning the diversity of the dataset. We also observed other types of issues in the dataset, such as images featuring fake removable sleeve tattoos, blurry images, and incorrectly labelled public images (see Figure 3).

PrivAlert For the PrivAlert dataset (Zhao et al., 2022), we noted numerous images containing people labeled as non-private, despite the dataset explicitly defining people as private. Using the DETR object detector (Carion et al., 2020), we identified 1,707 individuals present in 540 out of 1,254 images labeled as public. Additionally, we encountered inconsistencies such as statues and paintings of people labelled as private. Examples of these inconsistencies are shown in Figure 3, with more images in Appendix G.

Human Evaluation To quantitatively assess dataset quality, we randomly sampled 50 images (25 private, 25 public) from each dataset for human evaluation (details in Appendix I). Five reviewers judged images based on the original privacy class definitions provided by the dataset creators. We measured binary accuracy by comparing the dataset labels with reviewers' majority selections and calculated inter-rater agreement using Fleiss' kappa (Fleiss & Cohen, 1973). The results, presented in Figure 4, demonstrate that labels in our PrivBench dataset exhibit greater consistency with its privacy definition.

The Visual Privacy Dataset (VISPR). Our human evaluation confirmed its high overall quality. However, we argue that some classes included in VISPR (Orekondy et al., 2017), such as hair color, are debatable in their privacy status. Figure 5 provides examples that, although technically containing private attributes, are insufficient to uniquely identify a person without context. Additionally, our analysis with DETR revealed that VISPR is highly skewed towards images containing people, which makes up 74.6 % of its private class.

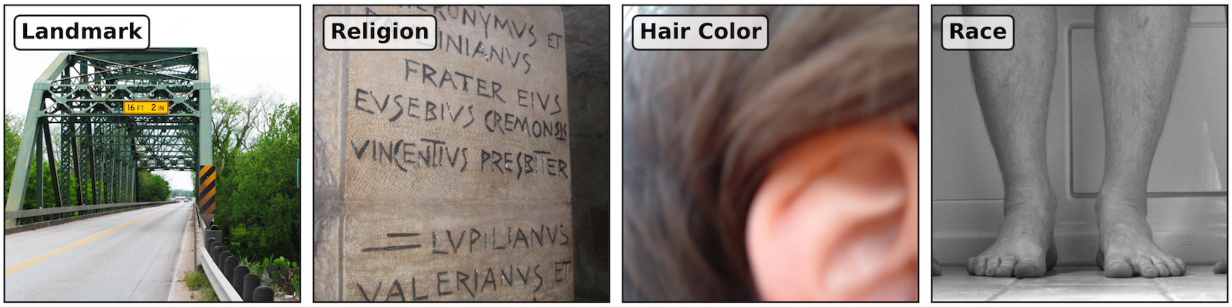


Figure 5: **Samples from the VISPR dataset.** Examples of privacy attributes (e.g., hair colour) that are insufficient on their own to identify individuals; class labels are shown in the top-left corner.

Based on these findings and observed qualitative issues, we conclude PrivAlert and Biv-Priv are unsuitable as benchmarks due to excessive label noise. However, for completeness, we include detailed scores for these datasets in the Appendix.

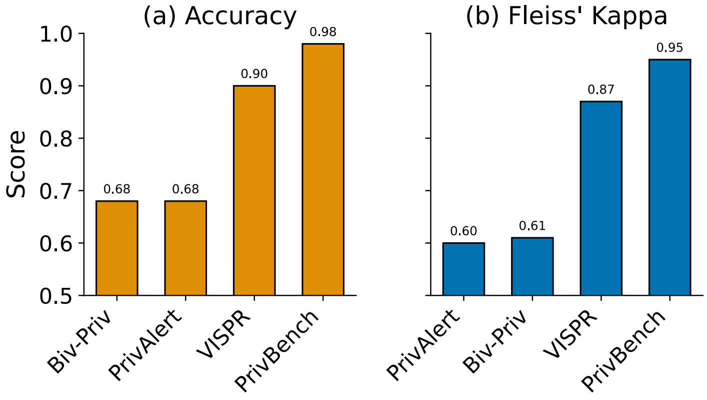


Figure 4: **Human Evaluation on Privacy Datasets**. We report the accuracy and inter-rater agreement (Fleiss' Kappa) for PrivAlert, Biv-Priv, VISPR and our PrivBench.

4 Methodology

Privacy Datasets We introduce three datasets, PrivBench, PrivBench-H(ard), and PrivTune , each containing 160 private and 160 public images (Table 1). Each dataset comprises unique public images, whereas PrivBench and PrivBench-H share the 'private' category images. The private set includes explicit private items (e.g., passports, debit cards), while public images contain no private content, such as landscapes and food pictures. In Figure 2, some private samples are shown.

All datasets are a subset of the Re-LAION-5B, a cleaned version of the original LAION-5B dataset (LAION, 2023; Schuhmann et al., 2022). To ensure quality, we first applied keyword-based caption filtering (e.g., "selfie", "person" or "face" for faces). Subsequently, images were manually selected according to strict guidelines, accepting only clearly private images (e.g., excluding closed passports without visible personal data). Detailed guidelines are provided in Appendix B. Figure 4 demonstrates via human evaluation that our dataset achieves higher accuracy and inter-rater agreement compared to existing privacy datasets.

Per GDPR Article 4 (GDPR, 2016), personal data encompasses any information relating to identifiable individuals. Consequently, all classes listed in Table 1 qualify as private under GDPR. Appendix B offers a detailed justification for classifying each category as private.

PrivTune **Training Dataset:** To effectively privacy-tune a VLM, we collect privacy-aware fine-tuning annotations consisting of multi-turn dialogues between a simulated user and a visual assistant.

We utilised GPT-4 (Achiam et al., 2023) for generating these dialogues, providing explicit instructions to simulate dialogues where the assistant responds to user inquiries and discusses potential privacy concerns. Generation was conditioned on class names and privacy labels, formulated as $p(d|i, l, c)$, where d is the dialogue, i represents instructions, l indicates the binary privacy label, and c specifies the class. This means GPT-4 is not used to classify the privacy of the image itself. An example annotation was included to guide model responses. Due to policy constraints restricting GPT-4's handling of nudity samples, we utilized ShareGPT (Chen et al., 2024) for these cases.

Figure 2 illustrates representative samples with partial fine-tuning annotations. Appendix A details the PrivTune dataset, including prompts and metrics related to collected dialogues.

PrivBench Set-up: For the PrivBench benchmark, we employ private images listed in Table 1. The public class includes straightforward examples like landscapes, empty streets, and food images to evaluate basic privacy comprehension. Public samples are shown in Appendix E.

PrivBench-H Set-up: The PrivBench-Hard benchmark employs the same private images as the standard PrivBench, introducing complexity through challenging negatives such as fake debit cards, blurred faces, dolls resembling humans, simulated scenes, and non-private documents or objects (e.g., brochures, toy cars without plates). These selections intentionally resemble private classes to increase classification difficulty. As the metric we use for evaluating privacy is sensitive to false positives, the samples in PrivBench-H significantly raise the complexity as shown in the results. Public samples are shown in Appendix E.

Access to Privacy Datasets Due to the sensitive nature of the images in our dataset, it is available for research purposes upon request. We ask researchers to delete the dataset after use to ensure the privacy of the individuals represented. Further reflections on the ethics of the dataset are described in the Discussion.

Table 1: **Class Taxonomy in privacy datasets.** Taxonomy of private and public categories used in PrivBench, PrivBench-H, and PrivTune, with corresponding GDPR articles.

Classes	Description	GDPR Article
Debit Card 💳	Debit cards, credit cards.	§4
Face 👴	Portraits, facial images, personal identification.	§4, §9
Fingerprint 🫆	Fingerprints, biometric identifiers, close-up images.	§4, §9
License Plate 47-35	Vehicle license plates, cars, motorcycles.	§4
Nudity ▨	Nude images, explicit content, sensitive material.	§4
Passport 🪪	Passports, visas, personal identification documents.	§4
Private Chat 💬	Emails, personal messages, digital conversations.	§4
Tattoo Tattoo	Tattoos, body art, personal identifiers.	§4
Public	General content, non-sensitive information, landscapes, food.	-

5 Results

Measuring the Understanding of Privacy We denote VLMs as $f(x, p)$ where x is the image input and p is the text prompt containing instructions to analyse the image and provide a privacy score indicating whether the image is private or non-private. We frame the task mathematically as: $f(x, p) \rightarrow s$, where $s \in \{0, 1\}$ represents a binary privacy score assigned by the model, indicating whether the image contains any private information. We also experimented with scores ranging from 1 to 5 to capture gradations in privacy levels. However, we found that all models consistently provided only two options, failing to show variance in their responses.

We instructed the VLMs to analyse the image for any personally identifiable information and to provide a "Yes" or "No" response indicating whether it contained private information. We accepted all answers containing "Yes" or "No"; all other responses were rejected and classified as mistakes.

To address prompt sensitivity, we utilise PrivTune to measure performance on variations of the prompt to see whether the results differ. Generally speaking, we see stable results with some exceptions, particularly for models performing below average. We utilise the best performing prompt on PrivTune per model to evaluate on PrivBench. There is no overlap between the train and evaluation datasets.

Evaluation To evaluate VLMs' understanding of privacy, we assessed several state-of-the-art models on VISPR (Orekondy et al., 2017) as well as on our proposed benchmarks: PrivBench and PrivBench-H. We primarily report the Matthews Correlation Coefficient (MCC), as it provides a robust and balanced evaluation even under significant class imbalance(Chicco & Jurman, 2020; Matthews, 1975):

$$\text{MCC (Matthews Correlation Coefficient)} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP}+\text{FP})\,(\text{TP}+\text{FN})\,(\text{TN}+\text{FP})\,(\text{TN}+\text{FN})}}$$

Additional metrics are detailed in Appendix. For consistency, we set the decoding temperature to zero (greedy decoding) across all experiments.

Privacy-tuning Furthermore, we establish *Privacy VLM* by privacy-tuning a TinyLLaVA (Zhou et al., 2024) model using the fine-tuning annotations from our PrivTune , testing is done on PrivBench. The model was fine-tuned using LoRA (Hu et al., 2022) with 20 training epochs. This fine-tuning strategy aims to enhance model sensitivity to privacy without significantly compromising performance for other tasks by keeping the training time as short as possible. The hyperparameters of these experiments can be found in the Appendix D.

Performance of VLMs on Visual Privacy. As detailed in Table 2, privacy-tuning significantly boosts model performance. The Privacy VLM consistently outperforms other VLMs on all benchmarks. Among models tested in a zero-shot context, MoeLLaVA (Lin et al., 2024) performs best on PrivBench, maintaining stable performance across private classes (see Table 3). CoAgent achieves the highest scores without any privacy-tuning on VISPR, which is rich in images with people, aligning with its strong performance on face-related classes in PrivBench. GPT-4 rejects all nudity samples, and therefore these scores are not utilized for the overall score.

Table 2: **Performance across Privacy Benchmarks.** Results Matthews Correlation Coefficient (MCC↑) demonstrate that our Privacy VLM consistently achieves superior performance across all benchmarks.

Model	**LLM**	**Vision Encoder**	PrivBench	PrivBench-H	**VISPR**
Otter	MPT-7B	CLIP ViT-L/14	−0.87	−0.85	−0.83
Fuyu	Persimmon-8B	–	−0.02	−0.04	−0.03
InstructBLIP	Vicuna-7B	Q-Former	0.19	0.08	0.16
GPT-4	–	–	0.50	0.48	0.16
ShareGPT	Vicuna-7B	CLIP	0.52	0.42	0.17
CogVLM	Vicuna-7B	EVA2-CLIP-E	0.59	0.31	0.19
CoAgent	Vicuna-7B	EVA2-CLIP-L	0.62	0.25	0.27
LLaVA	Vicuna-7B	CLIP ViT-L/14	0.69	0.42	0.22
MoE-LLaVA	Phi-2-2.7B	CLIP ViT-L/14	0.72	0.40	0.16
TinyLLaVA	Phi-2-2.7B	SigLIP	0.56	0.42	0.18
Privacy VLM (Ours)	Phi-2-2.7B	SigLIP	0.86	0.53	0.35

Detailed in Table 3, we show performance with adding the class definitions (passport, face, etc.) to the prompt, thereby changing the problem into detection. For some models, we see similar performance as to the standard task, which indicates that the models lack vision capability to solve the task. For other models, we observe improved performance, such as GPT-4, TinyLLaVa and MoeLLaVA. This suggests that while these models can detect these objects, they do not inherently consider them private themselves. Interestingly, GPT-4 (Achiam et al., 2023) does not classify fingerprints, faces, and tattoos as private. However, when asked to define privacy in images with only a text prompt, it explicitly mentions these classes (see Appendix C). This suggests a misalignment between the image and text spaces: GPT-4 defines these objects as private in text and can detect them with vision, yet it does not conclude that images containing them are private. This is a potential safety risk that should be further studied.

Table 3: **Class scores on PrivBench:** This table compares Matthews Correlation Coefficient (MCC↑) scores of our Privacy VLM and other VLMs on all classes in PrivBench. Since GPT-4 rejects all nudity samples, these do not contribute to its overall score. The last displays results for the case when private class names are added to the input prompt.

Model	All	💳	👦	🫆	47-35	▨	🪪	💬	Tattoo	Class
Otter	-0.87	-0.63	-0.63	-0.63	-0.63	-0.63	-0.63	-0.63	-0.63	-0.95
Fuyu	-0.02	-0.16	0.12	-0.02	0.12	0.07	-0.16	-0.11	0.07	0.30
BLIP	0.19	0.37	-0.06	0.11	-0.13	-0.06	0.43	0.18	0.11	0.04
GPT-4	0.50	0.94	0.00	0.57	0.72	-	1.00	0.43	0.30	0.95
ShareGPT	0.52	0.79	0.21	0.82	0.00	0.94	1.00	0.37	0.00	0.54
CogVLM	0.59	0.97	0.23	0.61	0.23	0.48	0.97	0.76	0.72	0.56
CoAgent	0.62	0.63	0.13	0.63	0.29	0.33	0.63	0.60	0.50	0.56
LLaVA	0.69	1.00	0.37	1.00	0.30	1.00	1.00	0.65	0.69	0.59
MoELLaVA	0.72	0.90	0.32	0.90	0.27	0.90	0.90	0.87	0.71	0.82
TinyLLaVA	0.56	0.72	0.21	0.88	0.21	0.82	1.00	0.61	0.48	0.85
Privacy VLM (Ours)	0.86	0.88	0.72	0.88	0.65	0.88	0.88	0.88	0.72	0.94

Table 4: **Generalization when leaving out one private class during training:** We omit one class at training time and report MCC(↑) and F1(↑) for that left-out class on PrivBench.

	Performance on left-out class							
	💳	👦	🫆	47-35	▨	🪪	💬	Tattoo
MCC	0.90	0.89	0.85	0.79	0.85	0.89	0.84	0.88
F1	1.00	0.82	0.99	0.18	0.98	1.00	0.99	0.78

Generalization Privacy is a broad concept that poses challenges for models to generalise beyond the categories they were trained on. We evaluated this by omitting one class at a time from the PrivTune training data and assessing the model's performance on these classes during testing. For instance, we excluded credit cards during training to evaluate Privacy VLM's ability to recognise the sensitivity of credit card data, where it obtains a 0.9 MCC score. We trained the models using the same configuration as before. Table 4 shows that Privacy VLM effectively generalises to new categories, although its generalisation was less optimal when license plates and tattoos were excluded.

Amount of Training Data Initial experiments suggest that not much data is required to effectively privacy-tune a model, prompting us to investigate the minimal amount of training data needed. Therefore, we conducted multiple experiments using varying amounts of training data. The results, depicted in Figure 6, show that using approximately 100 samples of the PrivTune dataset is sufficient to privacy-tune a model, which translates to less than 10 images per class to achieve at least a 0.75 MCC (or 85% F1 score, see Appendix E) score on the PrivBench benchmark.

Table 5: **Performance Difference After Privacy Tuning:** Absolute percentage change in Privacy VLM performance on various benchmarks before applying PrivTune (TinyLLaVA).

Metric	PrivBench	PrivBench-H	VQAv2 (Goyal et al., 2017)	POPE (Li et al., 2023)	ScienceQA (Lu et al., 2022)
Original	55.6	42.2	81.5	87.7	69.7
Priv-Tuned	86.4 $_{+30.8}$	52.8$_{+10.6}$	79.9$_{-1.6}$	86.4$_{-1.3}$	69.1$_{-0.6}$

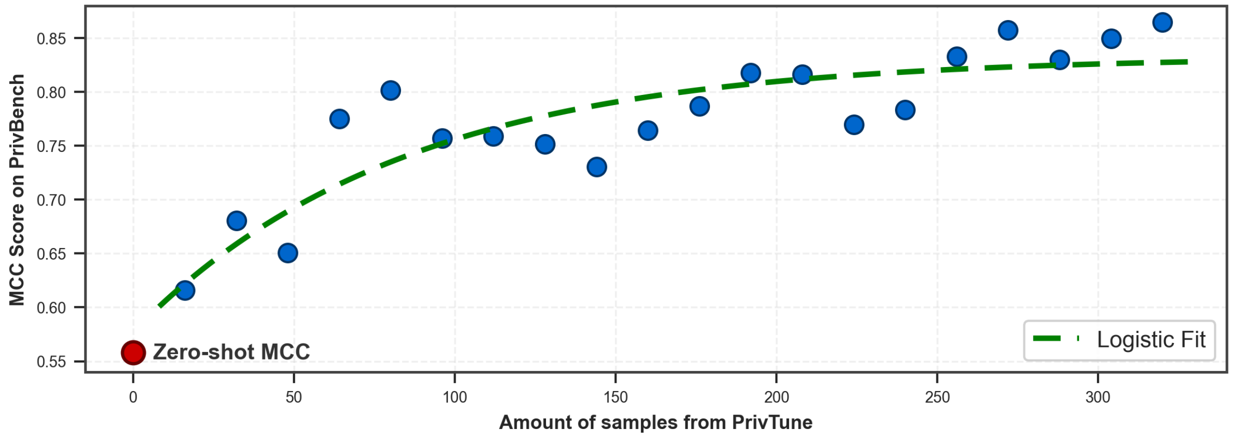


Figure 6: **Little data required for privacy-tuning.** This figure shows how the amount of training data affects the performance of privacy-tuning. The experiments demonstrate that as few as 100 samples from PrivTune are sufficient to achieve a 0.75 MCC score on PrivBench.

Impact on other VLM tasks To assess the cost of privacy tuning, we measured changes on standard benchmarks. Table 5 displays these results, revealing a slight decrease in performance on other tasks due to privacy tuning. However, this minor decrease is offset by a substantial improvement in the model's understanding of privacy.

Table 6: **Cross-validated Performance on PrivBench:** We perform 100 runs on random 50% splits on PrivBench. We report the mean and standard deviation for the MCC.

MCC (%)	Tiny LLaVA	CoAgent	Moe-LLaVA	Privacy VLM (Ours)
Mean	55.2	59.0	72.5	86.1
Std.	0.24	0.40	0.68	0.58

Size of Benchmarks Recognising the modest size of our datasets, we conducted cross-validation experiments on PrivBench. Consistent with our human evaluation and generalisation results on VISPR, these experiments demonstrate stable model performance, with MCC scores varying within just 1% standard deviation when using 100 random 50% splits (Table 6).

Privacy in different Languages Privacy perception varies culturally, prompting us to evaluate multilingual privacy recognition capabilities. Using GPT-4 (Achiam et al., 2023) and ShareGPT (Chen et al., 2024), we tested model performance across languages, by translating input and output, to diverse privacy cultures: German, English, Russian, and Chinese. Table 7 indicates improved performance in German compared to English, potentially reflecting the greater societal focus on privacy in Germany (Stevens Institute of Technology, 2023). However, further research is needed to conclusively attribute differences to data biases.

Table 7: PrivBench **in Different Languages:** This table presents the MCC scores when prompting VLMs in different languages.

	🇩🇪	🇬🇧	🇨🇳	🇷🇺
ShareGPT	0.75	0.60	0.00	0.00
GPT-4	0.53	0.50	0.64	0.60

5.1 Use Case: Privacy Analysis of Datasets

Building on our privacy-tuned model's strong performance and generalisation capabilities, we apply it to analyse large collections of images for privacy concerns. We use the Places365 dataset (Zhou et al., 2017), running our privacy-tuned model on a random sample of 100,000 images across various location categories.

Table 8: **Privacy Rate Analysis Across Locations.** This table presents our privacy-tuned model's assessment of 100,000 images from Places365 (Zhou et al., 2017). It lists the top 15 locations with the highest and lowest privacy rates, highlighting where the model detects significant private information, particularly in areas populated by cars and people, as well as sensitive locations like military bases and medical facilities.

(a) Lowest-15 (least private→less private).

atrium, hotel outdoor, sky, windmill, tower, courthouses, synagogue outdoor, viaduct, canal urban, library outdoor, shopping mall indoor, fishpond, islet, moat water

(b) Highest-15 (most private→less private).

car interior, nursing home, army base, operating room, aeroplane, cabin cockpit, dressing room, pub indoor, server room, beauty salon, berth, martial arts gym, physics laboratory, hospital

Table 9: **In-depth Analysis of Privacy Assessments**. This table presents GPT4 detailed analysis of the privacy explanations made by our Privacy VLM for 3 of the top 15 privacy-rated location types.

Class Name	**Analysis of Privacy VLM's Privacy Explanations By GPT4-V**
Army Base	The general trend for classifying the location as private is due to the presence of individuals in military uniforms, which could reveal their personal identities, affiliations, and sensitive operations related to national security. The presence of identifiable features, such as faces and uniforms, suggests a need for confidentiality to protect the privacy and safety of the individuals depicted.
Dressing Room	The general trend for classifying the dressing room location as private is centered around the presence of personally identifiable information, particularly individuals' faces, which could be used to recognise or track them. Additionally, the setting of a dressing room is inherently private due to the personal activities, such as dressing or grooming, that occur there.
Operating Room	The general trend for classifying the operating room as a private location is due to the presence of sensitive medical procedures, personal health details, and identifiable features of patients and medical professionals that are not meant for public disclosure to protect patient privacy.

We prompt Private VLM to classify the image for privacy with a short explanation for its decision. To quantify the model's interpretation of privacy, we calculated a "private image rate" for each location type, the ratio of images classified as private to the total number of images for that location: Privacy Rate $= N_{\text{private}}/N_{\text{total}}$

We find that the model effectively classifies images as private for place categories that typically have a high human presence, such as a cockpit or dressing room. Additionally, the model effectively generalises to inherently sensitive categories like military bases and medical facilities, even though these were not present in the PrivTune training dataset. To further understand why the privacy-tuned model classified certain images as private, we used GPT-4 to perform an automated analysis of our model's reasoning for its scoring. In Table 9, we provide examples for three location types. This revealed that the model is aware of people and license plates, as well as locations and situations. We provide more detailed results in Appendix J.

6 Discussion

Our results show that privacy-tuning improves privacy understanding while causing a slight performance degradation on standard benchmarks. We believe that integrating privacy-tuning into the regular fine-tuning phase of a VLM would be even more effective, although limited computational resources prevented us from testing.

We are aware that our datasets contain sensitive images, such as individuals' passports and debit cards. To protect individual privacy, we have implemented ethical safeguards. Researchers must request access to the datasets through a form where they specify the purpose of their use, agree to use the data responsibly and commit to deleting it after use. Moreover, we only accept requests for research purposes. We also emphasize that this data is already publicly available, as our datasets are subsets of Re-LAION-5B.

Finally, techniques such as in-context learning and chain-of-thought hold promise for enhancing VLMs' understanding of privacy. Although our experiments with these methods did not yield immediate improvements, we believe they could boost performance. However, since not all users are familiar with advanced prompting strategies, we argue that VLMs should be inherently privacy-aware by design to ensure safe deployment.

7 Conclusion

We investigate the ability of VLMs to handle privacy-sensitive information. Our results reveal that existing models, including state-of-the-art systems like GPT-4, fall short in recognizing visual privacy risks. This gap is compounded by inconsistencies in popular privacy datasets. To address this, we introduce two benchmarks, PrivBench and PrivBench-H, and an effective fine-tuning dataset, PrivTune. Our experiments demonstrate that tuning on as few as 100 examples significantly enhances privacy recognition across benchmarks, with minimal cost to overall performance. These findings underscore the feasibility of aligning VLMs with privacy expectations through compact, well-curated datasets, even in low-data regimes. This approach significantly boosts the models' sensitivity to privacy without compromising performance on other benchmarks, suggesting a robust strategy towards VLMs that can safely handle any sensitive real-world data.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Fuyu-8b: A multimodal architecture for ai agents, 2024.

Gavin Brown, Mark Bun, Vitaly Feldman, Adam Smith, and Kunal Talwar. When is memorization of irrelevant training data necessary for high-accuracy learning? In *Proceedings of the 53rd annual ACM SIGACT symposium on theory of computing*, pp. 123–132, 2021.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2022.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pp. 370–387. Springer, 2024.

Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21:1–13, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

European Commission. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206, 2021. COM/2021/206 final.

Joseph L Fleiss and Jacob Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619, 1973.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 2024.

General Data Protection Regulation GDPR. General data protection regulation. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC*, 2016.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P Bigham. 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*, pp. 939–948, 2019.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2): 1–55, 2025.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

LAION. Relaion-5b. https://laion.ai/blog/relaion-5b/, 2023. Accessed: 2024-04-27.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Joshua Adrian Cahyono, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *CoRR*, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024b.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel Bowman. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1953–1967, 2020.

Seth Neel and Peter Chang. Privacy issues in large language models: A survey. *arXiv preprint arXiv:2312.06717*, 2023.

Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9617–9626, 2019.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 34303–34326, 2024.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Tanusree Sharma, Abigale Stangl, Lotus Zhang, Yu-Yun Tseng, Inan Xu, Leah Findlater, Danna Gurari, and Yang Wang. Disability-first design and creation of a dataset showing private visual information collected with people who are blind. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2023.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. In *International Conference on Learning Representations 2024*, 2024.

Stevens Institute of Technology. Countries ranked by internet privacy, 2023. URL `https://online.stevens.edu/info/countries-ranked-by-internet-privacy/`. Accessed: 2024-05-19.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35:38274–38290, 2022.

Batuhan Tömekçe, Mark Vero, Robin Staab, and Martin Vechev. Private attribute inference from images with vision-language models. *Advances in Neural Information Processing Systems*, 37:103619–103651, 2024.

Linh Trinh, Phuong Pham, Hoang Trinh, Nguyen Bach, Dung Nguyen, Giang Nguyen, and Huy Nguyen. Pp4av: A benchmarking dataset for privacy-preserving autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1206–1215, 2023.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024.

Zizhang Wu, Xinyuan Chen, Hongyang Wei, Fan Song, and Tianhao Xu. Add: An automatic desensitization fisheye dataset for autonomous driving. *Engineering Applications of Artificial Intelligence*, 126:106766, 2023.

Zheng-Xin Yong, Cristina Menghini, and Stephen H Bach. Low-resource languages jailbreak gpt-4. *arXiv preprint arXiv:2310.02446*, 2023.

Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Pinjia He, Shuming Shi, and Zhaopeng Tu. Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. In *The Twelfth International Conference on Learning Representations*, 2023.

Sergej Zerr, Stefan Siersdorfer, and Jonathon S. Hare. Picalert: A system for privacy-aware image classification and retrieval. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 2710–2712, 2012.

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

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*, 2018.

Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*, 2024.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.