SituationalPriv: A Context-Aware Framework for Privacy Detection and Protection in Vision-Language Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

With the widespread adoption of vision-language models (VLMs), users increasingly transmit large amounts of visual information, making context-aware privacy protection essential. Existing benchmarks for privacy detection are limited: some degrade image quality by blurring sensitive regions, others narrowly target predefined categories, and most overlook the contextual nature of privacy. As a result, current static evaluations fail to capture VLMs' real-world privacy recognition capabilities.

To address this, we introduce **SituationalPriv**, a benchmark for evaluating context-aware privacy understanding. It contains 440 high-quality, privacy-relevant images from the DIPA2 dataset, each paired with two distinct usage contexts that assign different privacy attributes to the same content. This design realistically simulates privacy-sensitive scenarios, enabling more comprehensive evaluation.

We further propose a **training-free framework** that leverages pretrained VLMs and large language models (LLMs) to improve context-aware privacy detection. Unlike prior fine-tuning approaches limited to fixed domains, our method demonstrates strong generalization across open-domain datasets.

1 Introduction

People today routinely share vast numbers of images through online applications, many of which contain sensitive personal information. Users often expose such data unintentionally, making it crucial for VLMs to detect potential privacy leaks in context and alert users. However, existing benchmarks for privacy detection face significant limitations: some degrade image quality by blurring sensitive regions [2], while others restrict evaluation to predefined categories and single scenarios [3, 10]. More importantly, they overlook the contextual nature of privacy [12], where identical content may or may not constitute a privacy risk depending on situational factors (e.g., a license plate shared with an insurance company vs. on social media). As a result, static evaluations fail to reflect VLMs' ability to handle real-world privacy-sensitive use cases.

To address these gaps, we introduce **SituationalPriv**, a benchmark designed for context-aware privacy detection. It includes 440 high-quality, privacy-relevant images curated from the DIPA2 dataset [14], each expanded into two distinct scenarios where the same object can be privacy-sensitive in one context but not in the other. This pairwise setup enables realistic and comprehensive evaluation of VLMs' ability to interpret contextual privacy.

We further propose a **training-free framework** that combines pretrained VLMs and LLMs. By using LLMs to guide object-level queries and reason over contextual information, our method significantly improves context-aware privacy detection without fine-tuning, ensuring generalization across domains and unseen privacy categories.

Our contributions are threefold:

- We present **SituationalPriv**, the first benchmark to systematically evaluate VLMs' context-aware privacy understanding with paired sensitive vs. non-sensitive scenarios.
- We design a **training-free LLM–VLM framework** that leverages pretrained knowledge for privacy detection without domain-specific fine-tuning.
- We demonstrate significant performance improvements on **SituationalPriv**, showing that our approach generalizes beyond in-domain data and fixed privacy categories, making it suitable for real-world applications.

2 Related Work

Vision-Language Models have demonstrated strong performance across various tasks [5, 7, 9]. As VLMs are increasingly integrated into everyday applications, concerns regarding privacy and security have also gained significant attention [2, 10]. Several new benchmarks have been proposed to assess VLMs' ability to recognize and protect privacy-sensitive information [2, 13]. However, these benchmarks often struggle to accurately reflect model performance in real-world applications.

For example, the VizWiz dataset [2, 4] is designed to evaluate a model's ability to detect privacy-sensitive objects in images; however, it employs mosaic obfuscation on those regions—which, while preserving privacy, also degrades image quality and obscures critical visual features. PRIVBENCH [10] evaluates VLMs' ability to identify sensitive objects in images, such as passports, license plates, debit cards, and faces. Similarly, Caldarella et al. [3] developed a face dataset to test VLMs' facial recognition performance under varying input conditions. In another study, Tömekçe et al. [13] introduced a dataset to assess how well VLMs can infer personal attributes from image inputs.

While these benchmarks provide valuable insights, they often overlook the context-dependent nature of privacy [12]. Privacy is a highly situational attribute. For example, sharing an image with a visible license plate might be appropriate when communicating with the DMV or an auto insurance company, but sharing the same image on public social media could lead to privacy breaches. The existing benchmarks predominantly focus on static detection methods that measure VLMs' ability to identify specific object categories without considering the contextual nuances of privacy. This limitation can result in scenarios where a model performs well on a static benchmark but fails to handle real-world tasks effectively, posing potential privacy risks.

Our work aims to bridge this gap by introducing a novel approach that emphasizes the context of privacy. We selected 440 high-quality images from the DIPA2 dataset and generated both privacy-sensitive and privacy-insensitive contexts for each image. This design simulates the application of VLMs in real-world scenarios, offering a more robust evaluation of their privacy awareness.

In addition to dataset innovations, there has been growing interest in leveraging Large Language Models due to their superior language understanding and reasoning capabilities compared to VLMs. Previous research has explored using LLMs as agents or guides to assist VLMs in tasks like visual question answering [1, 6, 8], enhancing overall task performance. However, the potential of LLMs to guide VLMs specifically for privacy detection and protection remains largely unexplored. Our work contributes to this emerging field by proposing a framework where LLMs act as privacy-aware guides, dynamically interpreting context and instructing VLMs to enhance privacy protection in practical applications.

3 SituationalPriv

In this section, we introduce the situationalPriv framework, which includes an automated data construction pipeline and a context-aware evaluation of privacy detection in VLMs.

3.1 Privacy leakage in SituationalPriv

Following Shao et al. [11], we focus on the risk of unintentional personal information leakage that arises when data senders are unaware that sharing certain data in a given context may inadvertently expose sensitive personal information. For example, consider an image `Img` containing m objects:

LLM-guidance framework							
Model Type	VLM	Q-LLM	D-LLM	Recall	Precision	F1	Accuracy
Hybrid	LLaVA-1.5 13B	GPT-4o	GPT4o	77.73	60.21	67.86	63.18
Hybrid	LLaVA-1.5 13B	GPT-4o	GPT-4o mini	26.59	58.21	36.51	53.75
Hybrid	LLaVA-1.5 13B	GPT-4o mini	GPT-4o	59.77	44.65	51.12	42.84
Hybrid	LLaVA-1.5 13B	GPT-4o mini	GPT-4o mini	25.23	37.12	30.04	41.25
Hybrid	LLaVA-1.6 7B	GPT-4o	GPT-4o	48.86	75.97	59.47	66.70
Hybrid	LLaVA-1.6 7B	GPT-4o mini	GPT-4o mini	10.23	70.31	17.86	52.95
Hybrid	GPT-4o	GPT-4o	GPT-4o	34.55	81.28	48.48	63.30
Baselines							
Model Type	VLM	LLM	–	Recall	Precision	F1	Accuracy
LLM-only	–	GPT-4o	–	62.87	68.49	65.56	67.01
VLM-only	LLaVA-1.5 13B	–	–	13.18	93.55	23.11	56.14

Table 1: Results of different models on the SituationalPriv benchmark. The LLM-guidance framework with LLaVA-1.5 13B achieves the best performance, with Recall substantially surpassing all other models, highlighting its strong sensitivity to privacy-related information.

$o_1, o_2, \ldots, o_m$, where each object o_i may convey potential personal information I_i. In a particular context C, the data sender may intend to share `Img` to communicate the object o_i and its associated information I_i with the recipient. However, sharing `Img` in such context C may also expose another object o_j and its corresponding sensitive information I_j, which the sender did not intend to disclose. We define such unintended exposure of I_j as a *privacy leakage*, to which should be paid special attention in real-world application.

3.2 SituationalPriv data Construction

Context-Aware Privacy Seed We selected 440 images from the DIPA2 dataset that were manually annotated as highly informative and containing personal information. Inspired by Shao et al. [11], we define a context-aware privacy seed as a tuple (IMG_i, PO_i), where IMG_i denotes the transmitted image and PO_i represents a privacy-sensitive object within the image.

Constructing Privacy Scenario Each context-aware privacy seed is further expanded into a 5-tuple scenario: $(IMG_i, PO_i, DataSender_i, DataRecipient_i,$ $Context_i)$, where $DataSender_i$ refers to the individual sharing the image, $DataRecipient_i$ refers to the data receiver, and $Context_i$ describes the context or situational setting of the image transmission.

For each seed, we construct two 5-tuple scenarios: one non-privacy-sensitive and one privacy-sensitive. In the non-privacy-sensitive case, the data sender intentionally shares the privacy-sensitive object PO_i within IMG_i under the given context, and thus no privacy leakage occurs. In contrast, in the privacy-sensitive case, the data sender want to share other information while unintentionally reveals PO_i to the data recipient under the given context, despite not intending to do so. This unintended disclosure constitutes a privacy leakage.

4 Context-Aware Privacy Detection

4.1 Evaluation setup

Our evaluation procedure involves providing the model with an input image to be transmitted, along with the scenario, which includes the *data sender*, *data recipient*, and the *context*. The model is then requested to determine whether sharing such an image in the given scenario would result in a privacy leakage.

Baseline We assess a VLM on its ability to perform context-aware privacy detection. Specifically, for each image, we evaluate the model's performance in both a *privacy-sensitive* and a *non-privacy-sensitive* scenario. This pairwise evaluation setup enables us to analyze the model's ability to understand the context-dependent attribute of privacy information, and to examine whether it can

accurately distinguish privacy risks under different contextual conditions—or whether it tends to exhibit over-sensitivity or under-sensitivity toward privacy sensitive content.

LLM-guidance framework We propose an LLM-guided framework to enhance the context-aware privacy detection capabilities of VLMs. An overview of the framework is shown in Figure 1. We begin by using the Recognize Anything Model[15] to identify all objects within the input image and generate a corresponding tag list. The tag list and the context are then jointly input into a LLM, which is prompted to select potentially privacy-sensitive objects from the tag list based on the given context and generates a set of questions $Q = [q_1, q_2, \ldots, q_m]$ related to selected objects to be asked to the VLM.

The VLM answers each question based on the input image, producing a response set $A = [a_1, a_2, \ldots, a_m]$. This answer set, along with the context, is then passed to the LLM with the context, which is prompted to assess whether a privacy leakage occurs in the given scenario, based on the responses and the contextual information.

Our framework leverages the LLM both to guide the VLM to focus on the context-relevant objects in the image and to reason about whether privacy leakage happens in the given context. By combining VLM's visual grounding with LLM's contextual reasoning, our approach enhances dynamic, context-aware privacy detection in multimodal settings.

Metric We use Recall and F1 for evaluation. Recall measures the proportion of privacy-sensitive images correctly identified, indicating whether a model is "under-sensitive" to privacy detection. Precision, the counterpart of Recall, reflects whether a model is "over-sensitive," i.e., how many predicted privacy-sensitive cases are truly sensitive. In practice, missing privacy leaks (low Recall) is more harmful than raising occasional false alarms (lower Precision). Therefore, we report both metrics and use F1, the harmonic mean of Precision and Recall, to capture overall performance.

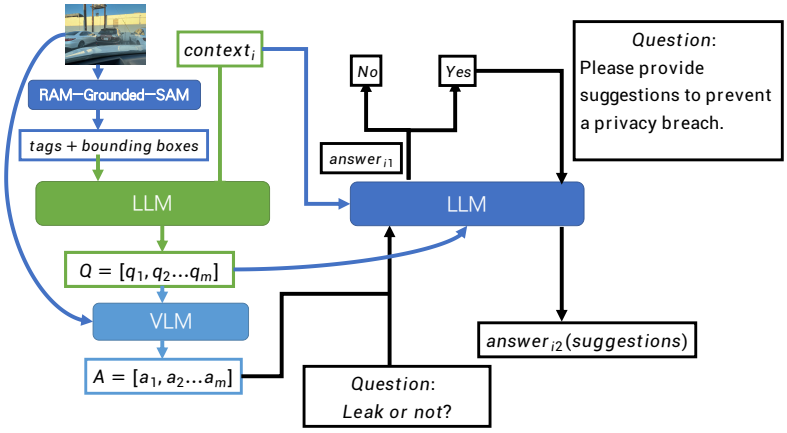

Figure 1: LLM-guidance privacy detection and protection framework.

4.2 Results

In Table 1, we report the performance of proposed LLM-guidance framework and two baselines (LLM-only and VLM-only) on primary metrics. Within the LLM-guidance framework, Q-LLM is responsible for jointly identifying potentially privacy-sensitive objects from the provided tag list and context, generating questions for the VLM. D-LLM assesses whether privacy leakage occurs based on the VLM's answer set and context. We employ widely adopted and size-varying VLMs (Llava-1.5 13B and Llava-1.6 7B) and LLMs (GPT-4o and GPT-4o mini).

Comprehensive Performance Evaluation Based on the harmonic mean of precision and recall (F1 score), our evaluation shows that the LLM-guidance framework significantly outperforms both the VLM-only baseline and the LLM-only baseline using identical models, clearly demonstrating its effectiveness in enhancing privacy leakage detection. Moreover, introducing advanced VLMs and Q-LLMs further improves the framework's overall F1 score, indicating that better question quality and more accurate visual analyses of potentially privacy-sensitive objects positively contribute to detection performance. Notably, when GPT-4o is employed as the D-LLM, all combinations consistently yield F1 scores above 50, substantially surpassing combinations involving lower-performing D-LLMs. This

observation highlights the critical importance of effectively aggregating and analyzing contextual information and question-answer interactions concerning potential privacy-sensitive objects within the LLM-guidance framework.

Analysis of Detection Sensitivity and Precision Trade-offs We analyze Recall to assess the models' ability to detect actual privacy leakage cases. The VLM-only baseline shows very low Recall (13.18), missing most violations, while the LLM-guidance framework achieves much higher Recall, confirming its stronger sensitivity. Consistent with F1 trends, more advanced VLMs and Q-LLMs further improve detection, and the D-LLM plays a key role in boosting overall capability.

For Precision, the VLM-only baseline performs well but mainly captures only the most explicit cases, explaining its high Precision but low Recall. The LLM-guidance framework, though slightly more prone to false positives, maintains Precision above 30 across all settings, effectively controlling false alarms. Stronger Q-LLM and D-LLM components further enhance reliability. Notably, the combination of LLaVA-1.5 13B and GPT-4o achieves the best balance, with a Recall of 77.73 and Precision of 60.21, demonstrating the framework's ability to capture subtle privacy risks without sacrificing accuracy.

5 Conclusion

We presented **SituationalPriv**, a dataset with scenario-specific contexts to systematically evaluate VLMs' ability in context-aware privacy detection. We further introduced a **training-free LLM-guidance framework** that leverages LLMs to enhance VLMs' performance in detecting privacy-sensitive content. Experiments demonstrate clear improvements over baseline methods, showing that our approach provides a robust solution for real-world privacy protection tasks.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[2] J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342, 2010.

[3] S. Caldarella, M. Mancini, E. Ricci, and R. Aljundi. The phantom menace: Unmasking privacy leakages in vision-language models, 2024. URL `https://arxiv.org/abs/2408.01228`.

[4] D. Gurari, Q. Li, C. Lin, Y. Zhao, A. Guo, A. Stangl, and J. 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*, pages 939–948, 2019.

[5] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[6] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[7] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975, 2022.

[8] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[10] L. Samson, N. Barazani, S. Ghebreab, and Y. M. Asano. Little data, big impact: Privacy-aware visual language models via minimal tuning, 2025. URL https://arxiv.org/abs/2405.17423.

[11] Y. Shao, T. Li, W. Shi, Y. Liu, and D. Yang. Privacylens: Evaluating privacy norm awareness of language models in action. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 89373–89407. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/a2a7e58309d5190082390ff10ff3b2b8-Paper-Datasets_and_Benchmarks_Track.pdf.

[12] Y. Shao, T. Li, W. Shi, Y. Liu, and D. Yang. Privacylens: Evaluating privacy norm awareness of language models in action. *Advances in Neural Information Processing Systems*, 37:89373–89407, 2024.

[13] B. Tömekçe, M. Vero, R. Staab, and M. Vechev. Private attribute inference from images with vision-language models. *arXiv preprint arXiv:2404.10618*, 2024.

[14] A. Xu, Z. Zhou, K. Miyazaki, R. Yoshikawa, S. Hosio, and K. Yatani. Dipa2: An image dataset with cross-cultural privacy perception annotations. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(4):1–30, 2024.

[15] Y. Zhang, X. Huang, J. Ma, Z. Li, Z. Luo, Y. Xie, Y. Qin, T. Luo, Y. Li, S. Liu, Y. Guo, and L. Zhang. Recognize anything: A strong image tagging model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1724–1732, June 2024.