Sensitive Data Leakage Detection Based on Vision-Language Models

1st Ran Dubin
Department of Computer Science
Ariel University, Ariel, Israel
rand@ariel.ac.il

***Abstract*—The identification of sensitive information in text and images is a critical challenge in privacy and data security. This project aims to explore and evaluate existing methods for detecting sensitive content in multimodal contexts, particularly combining text and visual data. The research focuses on analyzing various detection approaches to understand their strengths, limitations, and practical applicability in real-world scenarios.**

A key aspect of this project is to address the challenges posed by current datasets, including their limitations and how these affect the performance of detection algorithms. To tackle these issues, the project will create a novel dataset that integrates sensitive text and image examples. Additionally, the study will evaluate state-of-the-art techniques and investigate the potential of Vision-Language Models (VLMs) for multimodal analysis. This work aims to enhance the understanding of sensitive information detection and contribute to the development of more robust tools for safeguarding privacy, while identifying key areas for future research.

I. INTRODUCTION

In today's digital age, the spread of personal and confidential data has created significant challenges in safeguarding sensitive information. Sensitive data, which includes personal identifiers, financial records, medical histories, and proprietary information, is constantly at risk of being exposed or leaked. Detecting and protecting sensitive information is crucial in both research and industry. As data leakage becomes increasingly common, developing accurate and efficient ways to identify sensitive information is essential.

From an industrial perspective, the threat of sensitive data leakage has broad implications for privacy, security, and regulatory compliance. Industries such as finance, healthcare, and social media are under increasing pressure to protect customer data from breaches. Current solutions in the industry are primarily driven by rule-based systems or standard machine learning models that analyze text data. However, these approaches often struggle to detect sensitive content in complex, multimodal formats-such as images combined with text-leading to gaps in security. Furthermore, industrial applications often face issues related to scalability, real-time detection, and ensuring compliance with strict privacy regulations like GDPR and HIPAA.

On the academic side, researchers have developed a variety of methodologies to address these challenges. From machine learning techniques to deep learning models, significant progress has been made in detecting sensitive information. Recent advances in Vision-Language Models (VLMs), which combine computer vision and natural language processing, promise better analysis of image and text data. However, academic research often faces constraints related to the availability of large, labeled datasets, particularly those that contain both sensitive text and images, which can limit the effectiveness of detection models.

This project seeks to bridge the gap between academic advancements and industrial needs by developing a comprehensive dataset that combines text and image data related to sensitive information, and by evaluating cutting-edge multimodal detection techniques. This work aims to enhance both academic understanding and practical industry tools for protecting privacy in the digital world.

II. CONTRIBUTION

This research advances the field of sensitive data detection by leveraging Vision-Language Models (VLMs) to analyze multimodal content, focusing on practical and scalable solutions for identifying sensitive information embedded in text and images. The main contributions of this work include:

1) **PII-Document Image Dataset & Schema:**
   - We built a synthetic but realistic dataset of document-like images covering 12 PII classes (e.g., CREDIT CARD NUMBER, NAME, ADDRESS, PHONE) plus a binary any-PII label.
   - We collected and filtered relevant datasets from online sources.
   - We constructed a training-ready data pipeline that pairs images with their corresponding structured text annotations. We normalized all annotations into a compact JSONL schema used consistently across training and evaluation. This format enables direct input into Vision-Language Models (VLMs), supporting efficient multimodal processing for sensitive data detection.
   - We created an unseen evaluation dataset with known per-class supports in order to test generalization.
2) **Lightweight Fine-Tuning of BLIP-2 with LoRA:**
   - We fine-tuned and evaluated BLIP-2 using LoRA adapters on our dataset.

- We saved the adapters as portable checkpoint-4000 and evaluated them without modifying the base weights.

3) **Comparative Lightweight Fine-Tuning of Qwen2.5-VL with LoRA:**
   - In addition to BLIP-2, We fine-tuned Qwen2.5-VL-7B-Instruct using LoRA adapters on our dataset, enabling gradient checkpointing and TF32 for efficiency.
   - We also saved the adapters as portable checkpoint-4000 and evaluated them without modifying the base weights.
   - This provides a cross-model comparison of fine-tuning behavior on identical data, highlighting strengths and weaknesses of different VLM backbones when adapted to sensitive-information detection.
4) **Empirical Evidence of Benefit on Unseen Data:**
   - On **Qwen2.5-VL-7B-Instruct** we achieved greatly improved precision while maintaining maximum recall.
   - We also observed strong per-class gains on high-impact categories, and near-perfect results on structured document types.
   - On **BLIP-2** the model achieved strong performance on the binary task of distinguishing sensitive from non-sensitive images. However, its performance on per-class classification was substantially weaker, with poor results across individual PII categories.

By addressing both academic and applied aspects of sensitive data detection in multimodal systems, this project contributes a novel dataset, a fine-tuned VLM pipeline, and an analysis framework that collectively push forward the capabilities of privacy-aware AI systems.

III. RELATED WORK

The detection of sensitive data leakage in Vision-Language Models (VLMs) is an emerging field that intersects privacy concerns with multimodal machine learning systems. Below is a detailed summary of various studies that explore sensitive data leakage, including methodologies for detection, the datasets used, and the key findings.

[1] **"Privacy-Aware Visual Language Models"**: One of the most directly relevant works is "Privacy-Aware Visual Language Models", which introduces dedicated benchmarks such as PRIVBENCH and PRIVBENCH-HARD for evaluating the privacy capabilities of Vision-Language Models (VLMs). The study demonstrates that VLMs like MiniGPT-v2 and TinyLLaVa can be fine-tuned using a privacy-specific dataset (PRIVTUNE) to accurately distinguish between sensitive and non-sensitive visual-textual inputs, such as passports versus public scenes. This aligns closely with the goals of our project, which also focuses on training VLMs to detect sensitive information in multimodal content. Their approach validates the potential of fine-tuning VLMs for privacy-focused tasks and highlights methodologies for building effective privacy benchmarks, both of which inform and support our own work.

[2] **"Extracting Training Data from Document-Based VQA Models"**: The study "Extracting Training Data from Document-Based VQA Models" investigates how Vision-Language Models (VLMs) used in Visual Question Answering (VQA) systems can unintentionally leak sensitive information memorized during training. The authors propose techniques such as counterfactual analysis and Extraction Blocking (EB) to reduce these risks by ensuring that answers are only generated when sufficient contextual evidence exists. Although the focus is on preventing leakage at the model behavior level, the work directly relates to our project, which aims to proactively detect sensitive information in multimodal inputs. Both studies operate in the domain of document-based VLMs and address the broader issue of privacy in models that process visual-textual data, reinforcing the importance of developing safeguards around sensitive data in real-world systems.

[3] **"Private Attribute Inference from Images with Vision-Language Models"**: One closely related study is "Private Attribute Inference from Images with Vision-Language Models", which highlights a critical privacy concern in modern multimodal systems. The authors demonstrate that Vision-Language Models (VLMs), such as GPT-4V, can infer sensitive personal attributes like income level or geographic location from seemingly non-sensitive images. By using techniques such as advanced prompting and zooming into image details, the models achieved high inference accuracy, raising serious concerns about unintentional privacy violations. This work is highly relevant to our project, as it underscores the need not only to detect explicitly sensitive content but also to consider how VLMs may leak sensitive information through visual or contextual inference. Our project aims to address this issue by building detection models that can proactively identify and flag content that poses a risk of sensitive data exposure, either directly or indirectly.

[4] **"Analyzing Leakage of Personally Identifiable Information in Language Models"**: Another relevant study is "Analyzing Leakage of Personally Identifiable Information in Language Models", which examines how large language models (LLMs) like GPT-2 can unintentionally reveal sensitive data, such as names or phone numbers, even when trained with privacy-preserving techniques like Differential Privacy (DP). The research shows that larger models and repeated exposure to PII in training data significantly increase the risk of leakage. While this work focuses on text-only models, it highlights an important concern that extends to Vision-Language Models (VLMs) as well - especially when dealing with multimodal content where sensitive information may be encoded in both text and images. This paper reinforces the need for robust detection and mitigation strategies, supporting our project's goal of proactively identifying and preventing sensitive data leakage in VLMs.

IV. ARCHITECTURE

1) **System Overview:** We frame image-level PII detection as structured generation with a vision-language model. For each document image, the model receives a short audit prompt and must generate a JSON object containing Boolean labels for a fixed set of PII classes. We fine-tune Qwen2.5-VL-7B-Instruct and BLIP-2 with LoRA adapters while keeping the base weights frozen. The same pipeline supports training, validation, and a held-out test evaluation.
2) **Data Layer:** Images are stored under a dataset root with two subfolders, "sensitive" and "non_sensitive". Records come from a compact JSONL file with fields:
   - **image:** path to the image
   - **is_sensitive:** global binary flag
   - **types:** subset of PII classes

   At start, the code optionally downloads and extracts a dataset ZIP file from Dropbox and repairs broken paths. Each record is normalized to a dense label dictionary (all classes present as booleans). The loader builds balanced splits by sampling from sensitive and non_sensitive subfolders to configurable sizes (e.g., train 4k/4k, val 1k/1k, test 1k/1k).
3) **Vision preprocessing:**
   - We open each image in RGB and shrink it so the longest side is at most 896 pixels.
   - No task-specific augmentations are applied. Additional processing is handled by the model processor (Qwen or BLIP-2).
4) **Prompt & target format (structured generation):** For each sample we give the model a short instruction:
   - Role: "You are a PII auditor."
   - Scope: the exact class list to consider
   - Output: return only JSON that contain "labels" - ground truth booleans (true/false for each class), and "evidence_text" - an empty list (kept for format consistency).

   Both models receive the image + prompt as input and the correct JSON as the expected output. Qwen uses its chat template, while BLIP-2 uses a strict JSON skeleton in the prompt to enforce structure.
5) **Model & adaptation:**
   - Backbones: Qwen2.5-VL-7B-Instruct and BLIP-2 (blip2-flan-t5-xl).
   - Running in bf16 on GPU for speed/memory efficiency.
   - Gradient checkpointing enabled to save memory.
   - Fine-tuning with LoRA adapters (rank=32, dropout=0.05).
   - LoRA is applied to the main attention layers and also to the vision-to-text projector.
   - The base model weights stay frozen - only the small LoRA adapters are trained.
6) **Collation & loss:**
   - Each batch is prepared twice:
     a) Prompt only (image + question) $\rightarrow$ to know how long the prompt is.
     b) Prompt + correct JSON answer $\rightarrow$ the real training input.
   - We copy the tokens but hide the prompt and padding so the model is only graded on the JSON answer part.
   - This way, the model learns only to generate the correct JSON labels, not to repeat the question.
7) **Training:**
   - We train with the AdamW optimizer.
   - The LoRA layers use a learning rate of 2e-4.
   - The vision projector uses a smaller learning rate of 1e-4.
   - We also use:
     – Weight decay = 0.01
     – Warmup = 5% of steps
     – Gradient accumulation so the batch size is effectively larger

   Checkpoints are saved regularly and can be resumed safely.
8) **Inference:**
   - At test time the model writes JSON text with its predictions.
   - We read that JSON and turn it into true/false labels for each class.
   - If parsing fails, we just return all false.
9) **Metrics:**
   - For each PII type: check precision, recall, F1, and how many examples there are.
   - Average all types $\rightarrow$ Macro-F1.
   - Also check a simple yes/no PII case (any class positive): report F1, precision, recall, accuracy.

V. DATA EXPLORATION

1) **Data sources:**
   To support the training and evaluation of Vision-Language Models (VLMs) for sensitive data detection, we created a comprehensive dataset consisting of labeled multimodal examples, combining synthetic samples that we generated with additional samples collected from the web. The dataset is divided into two main categories: sensitive and non-sensitive information, with a total of approximately 42,000 samples.
2) **Dataset Structure:**
   The Train dataset contain 42,000 samples:
   Sensitive data:
   - Credit Card images
   - Driver License images
   - Images of Medical documents
   - Images of Phone Bill documents
   - Images of PIN code letters from the bank

- Mix of PII images

These images contain the following sensitive information: Credit card number, SSN, Driver license number, Personal id, Pin code, Name, Address, Email, Phone, Bank account number.

Non-Sensitive data:

- Images of Advertisements
- Images of Budget documents
- Random Email documetns
- Form documents
- Handwritten letters
- Formal letters
- Images of memos
- Resume documents
- Scientific reports

Due to hardware limitations - just 12,000 from the whole dataset used for training the Qwen2.5-VL-7B-Instruct model. We used the same process for the BLIP-2 model to ensure a fair comparison under identical training conditions and dataset constraints.

The Evaluation dataset contain 2000 unseen samples :
Sensitive data:

- Credit Card images (unseen images)
- Images of PIN code letters from the bank (unseen images)
- Personal letters containing PII
- Images of PIN-code letters from the mobile carrier.

Non-Sensitive data:

- Handwritten letters (unseen images)
- Formal letters (unseen images)
- Customer reviews
- Receipts images

These images contain the following sensitive information: Credit card number, SSN, Personal id, Pin code, Name, Address, Phone, Bank account number.

3) **Coverage & Balance:**
   - **Broad document coverage:** We included many kinds of documents - some with PII (like IDs, bills, bank papers) and some without (like ads, Emails, letters) - so the model learns to tell sensitive from non-sensitive content.
   - **Balanced sensitive vs. non-sensitive:** The core 42k set is split roughly 50/50 between "sensitive" and "non_sensitive" folders. This prevents the model from learning a trivial prior (e.g., "most images are sensitive") and makes the binary "any-PII" metric meaningful.
4) **Resolutions & Formats:**
   - All images are opened in RGB to handle different formats (jpg, png, bmp, tiff, webp).
   - Corrupt files are skipped.
   - We resize the images so the longest side is max 896 px, keeping the aspect ratio.

VI. RESULTS

We evaluated two vision–language models, Qwen2.5-VL-7B-Instruct and BLIP-2 (blip2-flan-t5-xl), each with and without LoRA fine-tuning, on a held-out set of 2,000 unseen images (1,000 sensitive / 1,000 non-sensitive). Both used the same JSON output format for multi-label PII detection and binary any-PII classification.

Base: BLIP-2
Fine-tuned: BLIP-2 with LoRA adapters.

TABLE I
MACRO-F1 PER-CLASS SCORE (HIGHER IS BETTER)

Model	Macro-F1
Base BLIP-2	**0.2227**
BLIP-2 + LoRA	0.1089

TABLE II
ANY-PII (SENSITIVE/NON-SENSITIVE) SCORE (HIGHER IS BETTER)

Model	Any-PII F1
Base BLIP-2	**0.787**
BLIP-2 + LoRA	0.626

TABLE III
PRECISION, RECALL & ACCURACY (HIGHER IS BETTER)

Model	Precision	Recall	Accuracy
Base BLIP-2	**0.780**	0.794	**0.785**
BLIP-2 + LoRA	0.514	**0.800**	0.522

- Both base and fine-tuned BLIP-2 achieve very high recall, with LoRA only slightly higher.
- **Base BLIP-2** is stronger overall:
  - Any-PII F1 = 0.787 vs. 0.626 for LoRA.
  - Better balance across classes, with non-zero F1 for Phone Bill (1.0), Phone (0.667), and Name (0.560), while on LoRA there is just 2 non-zero F1 across classes.
  - Much higher binary precision, reducing false alarms.
- **Fine-tuned BLIP-2 (LoRA)** improves only for credit-card detection:
  - Recall = 1.0 on Credit Card Number, but precision is low (P=0.301).
  - Regarding Name - The base model is very precise (94% precision) but misses many positives (only 40% recall). The LoRA model flips this: it catches 80% of all Name cases (double the recall) but at the

cost of much lower precision (51%). As a result, LoRA slightly improves the F1 score (0.626 vs. 0.560) — but it introduces far more false positives.

- The LoRA's over-recall behavior suggests it overfit to a subset of classes (notably **Name** and **Credit Card Number**) while failing on others it saw less of during training.
- LoRA probably didn't get enough examples of classes like PHONE, PHONE BILL, or ADDRESS. That's why it fails on them. The base model, however, still detects them because it relies on general patterns. For example, it reached F1 = 1.0 on PHONE BILL across 200 items.
- In summary: the base BLIP-2 is better for general PII detection, while the LoRA variant is useful if the focus is specifically on detecting **credit cards** or **names**.
- The full per-class scores are available on GitHub.

Base: Qwen2.5-VL-7B-Instruct
Fine-tuned: Qwen2.5-VL-7B-Instruct with LoRA adapters
Some classes had zero support in this specific evaluation (EMAIL, MEDICAL_LETTER, OTHER_PII), which depresses macro-F1 when included.

TABLE IV
MACRO-F1 PER-CLASS SCORE (HIGHER IS BETTER)

Model	Macro-F1	Macro-F1 (exclude 0-support)
Base Qwen2.5	0.5481	0.7306
Qwen2.5 + LoRA	**0.6763**	**0.8266**

TABLE V
ANY-PII (SENSITIVE/NON-SENSITIVE) SCORE(HIGHER IS BETTER)

Model	Any-PII F1
Base Qwen2.5	0.676
Qwen2.5 + LoRA	**0.774**

TABLE VI
PRECISION, RECALL & ACCURACY (HIGHER IS BETTER)

Model	Precision	Recall	Accuracy
Base Qwen2.5	0.511	1.000	0.521
Qwen2.5 + LoRA	**0.631**	**1.000**	**0.707**

- Both models catch all sensitive images (Recall = 1).
- Fine-tuned model gives far fewer false alarms:
  - The base model correctly identifies only about 42 of the 1,000 non-sensitive images.
  - The Fine-tuned model correctly identifies about 414 of the 1,000 non-sensitive images.
- Strong numeric types (Credit cards, PIN code, Bank account number) stay near-perfect.
- Because recall is already 1 in both models, the accuracy gain is entirely due to better selectivity on non-sensitive images. This leads to +18.6 percentage-point jump, this is real improvement.
- The full per-class scores are available on GitHub.

VII. CONCLUSIONS

A. Key Findings

1) On Qwen2.5-VL, **LoRA fine-tuning** is clearly beneficial: Macro-F1 improves, precision rises while recall remains 1.0, and accuracy increases substantially. This reduces false positives while preserving perfect coverage of sensitive images—an ideal outcome for safety-critical use cases.
2) On BLIP-2, the **base model** is stronger overall. LoRA slightly raises recall and yields targeted gains for Credit Card Number and a modest F1 lift for Name. However, these come at the cost of many false positives, making the base BLIP-2 the more reliable general detector.

B. Limitations

1) Some classes in the 2k evaluation had zero support (e.g., EMAIL, MEDICAL_LETTER, OTHER_PII), which reduces macro-F1 and under-represents those categories. The restricted training subset and document-centric data distribution may also limit generalization to natural scenes and multilingual content.
2) Scaling training to much more than 12,000 images is both financially expensive and computationally demanding, even when using high-end GPU resources.

C. Implications and Future Work

Our results show that with light fine-tuning, a Vision–Language Model (VLM) can move from being noisy but high-recall to being both **accurate and reliable**, making it more useful for real privacy checks. To make the system even better, future work should:

- Add more examples for classes that don't have enough training data.
- Include real-world photos that may contain PII, not just documents.
- Adjust prompts and outputs to avoid too many false alarms.
- Train on much more images (perform this on appropriate hardware).

D. Summary

In conclusion, our dataset, structured output design, and LoRA fine-tuning pipeline provide a clear and repeatable way to build scalable, privacy-aware multimodal detection systems. From the models tested:

- Qwen2.5-VL with LoRA gave the best mix of accuracy and recall.
- BLIP-2 served as a solid baseline, showing the trade-offs between different model architectures.

VIII. GITHUB

https://github.com/Yuval10Dahan/Final-Project-CS

REFERENCES

[1] L. Samson, N. Barazani, S. Ghebreab, and Y. M. Asano, "Privacy-Aware Visual Language Models," arXiv preprint arXiv:2405.17423, 2024.

[2] F. Pinto, N. Rauschmayr, F. Tramer, P. Torr, and F. Tombari, "Extracting Training Data from Document-Based VQA Models" Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria, PMLR 235, 2024.

[3] B. Tömekçe, M. Vero, R. Staab, and M. Vechev, "Private Attribute Inference from Images with Vision-Language Models," 38th Conference on Neural Information Processing Systems (NeurIPS), Vienna, Austria, 2024.

[4] N. Lukas, A. Salem, R. Sim, S. Tople, L. Wutschitz, and S. Zanella-Béguélin, "Analyzing Leakage of Personally Identifiable Information in Language Models," IEEE Security and Privacy (S&P), 2023.

[5] Md. Faruk Alam, "Top Open-Source Vision-Language Models (VLMs)," LinkedIn Pulse, 2024. Available: https://www.linkedin.com/pulse/top-open-source-vision-language-models-vlms-md-faruk-alam-rugvc/. [Accessed: Nov. 25, 2024].