Vision Language Model Helps Private Information De-Identification in Vision Data

Anonymous ACL submission

Abstract

Visual Language Models (VLMs) have gained significant popularity due to their remarkable ability. While various methods exist to enhance privacy in text-based applications, privacy risks associated with visual inputs remain largely overlooked such as Protected Health Information (PHI) in medical images. To tackle this problem, two key tasks: accurately localizing sensitive text and processing it to ensure privacy protection should be performed. To address this issue, we introduce VisShield (Vision Privacy Shield), an end-to-end framework designed to enhance the privacy awareness of VLMs. Our framework consists of two key components: a specialized instruction-tuning dataset OPTIC (Optical Privacy Text Instruction Collection) and a tailored training methodology. The dataset provides diverse privacy-oriented prompts that guide VLMs to perform targeted Optical Character Recognition (OCR) for precise localization of sensitive text, while the training strategy ensures effective adaptation of VLMs to privacy-preserving tasks. Specifically, our approach ensures that VLMs recognize privacy-sensitive text and output precise bounding boxes for detected entities, allowing for effective masking of sensitive information. Extensive experiments demonstrate that our framework significantly outperforms existing approaches in handling private information, paving the way for privacy-preserving applications in vision-language models.

1 Introduction

Vision Language Models (VLMs) (Alayrac et al., 2022; Liu et al., 2024b; Bai et al., 2023), which are developed following the impressive success of LLMs, show a remarkable ability to solve image-related tasks. Similar to text-only Large Language Models (LLMs) (Dubey et al., 2024; Abdin et al., 2024), which pose potential privacy risks by memorizing and outputting sensitive information from training data (Mireshghallah et al., 2022; Huang et al., 2022; Carlini et al., 2021), VLMs also suffer from privacy risks because VLMs share the generation part with LLMs (Liu et al., 2024c).

Figure 1: An illustrative example of medical imaging containing protected health information (PHI), shown in the top-left region, adapted from Rutherford et al. (2021). The displayed information is synthetic and thus remains unmasked for demonstration purposes.

To mitigate the privacy risks of text-only LLMs, several methods are proposed. For example, Jang et al. (2022) utilized knowledge editing to make LLMs forget the private information. Moreover, Zeng et al. (2024) proposed privacy restoration to remove the private information in the input and Yang et al. (2024a) leveraged an auxiliary LLM to remove the sensitive information in the training data. However, most of them focus on the text while neglecting the potentially sensitive information in visual input. For example, medical images often contain protected health information (PHI), which is considered sensitive information. We also show an example of PHI in Fig. 1.

To tackle privacy issues arising from vision data, one promising solution is data de-identification (Ribaric et al., 2016). De-identification is the process of removing or masking personally identifiable information (PII) from datasets to ensure privacy. However, previous

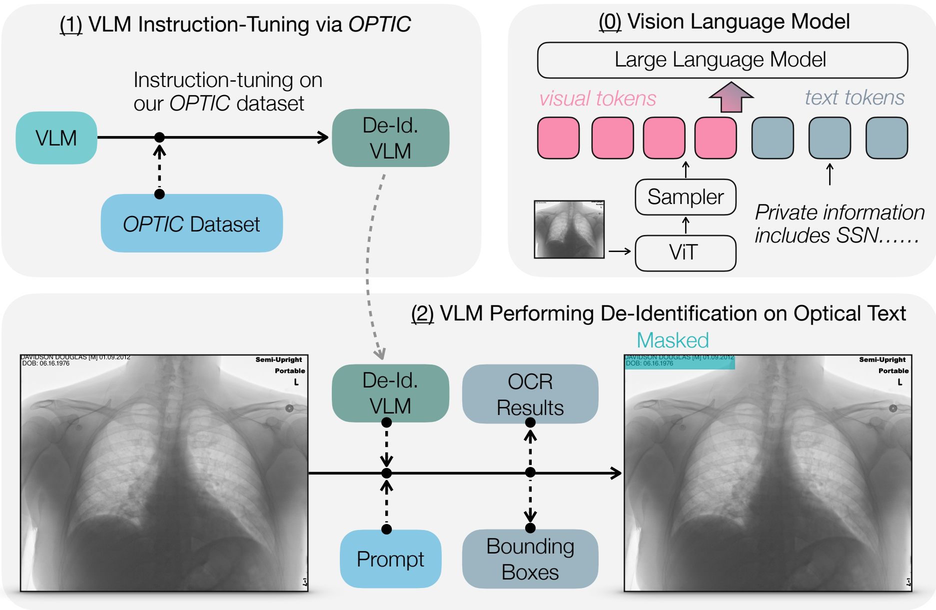

Figure 2: The proposed de-identification pipeline. Our approach leverages instruction-tuned VLMs to first perform targeted OCR on privacy-sensitive regions, followed by selective masking of identified confidential information.

works on image de-identification mainly focus on faces, which aim at obscuring identifiable facial features using generative models (Brkic et al., 2017; Cao et al., 2021). There is a lack of work focusing on textual private information in vision data. To the best of our knowledge, only Presidio (Microsoft, 2023) attempts to de-identify such information. However, Presidio lacks the flexibility to define what constitutes private information and demonstrates suboptimal performance in our experiments.

To address the lack of methods for de-identifying textual private information in vision data, two key tasks are required: accurately localizing sensitive text and processing it to ensure privacy protection. Therefore, in this paper, we propose an end-to-end framework named VisShield (Vision Privacy Shield), which leverages a Vision Language Model to assist in the de-identification of vision data. Our framework includes two components:
1) A specialized instruction-tuning dataset OPTIC (Optical Privacy Text Instruction Collection) designed to teach VLMs how to handle privacy-sensitive textual elements. This dataset includes diverse, privacy-oriented instructions that guide VLMs to perform OCR-based localization of private text. We generate synthetic image-text pairs with embedded fake private information, covering both natural and medical image scenarios, ensuring robust generalization. Our dataset comprises 50M samples, providing a rich training resource for localizing sensitive text.
2) A tailored training methodology that enables a VLM to accurately understand customized definitions of private information and apply de-identification mechanisms effectively. We fine-tuned a pre-trained VLM, `Kosmos-2.5` (Lv et al., 2023) on the OPTIC dataset to enable the VLM to process sensitive text accurately.

Our framework pipeline as shown in Fig. 2 enables the VLM to understand customized definitions of private information and extract private information through OCR, which can then be masked to ensure privacy. Extensive experiments demonstrate that our VisShield achieves superior privacy-aware OCR performance and leads to potential new applications of VLMs. Overall, we summarize our contribution below:

- To the best of our knowledge, we are the first to address the problem of de-identification with customized definitions of textual private information in vision data.
- We collect a diverse instruction-tuning dataset, which contains both text and image parts. This dataset comprises up to 50M image-text pairs, enabling VLMs to output OCR results for identifying private information in images.
- We fine-tune `Kosmos-2.5` to demonstrate that even a small portion of our dataset suffices for fine-tuning a pre-trained VLM to assist with de-identification.

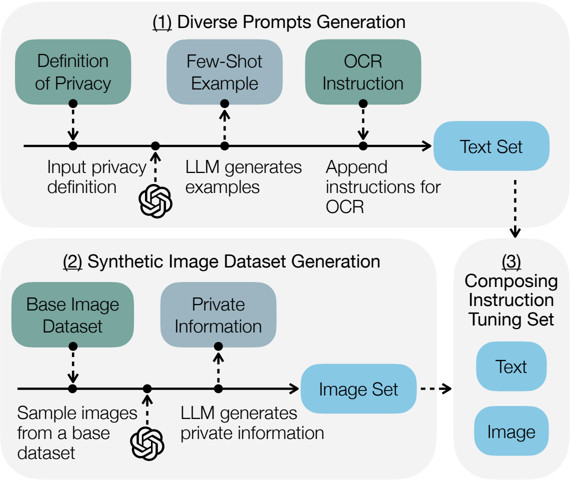

Figure 3: Overview of our three-stage dataset generation pipeline: (1) leveraging large language models (LLMs) to synthesize diverse instruction prompts, (2) creating synthetic images containing private information through controlled generation, and (3) producing aligned instruction-label pairs by combining the generated prompts with the synthetic image dataset.

2 Related Work

Vinson Language Models With the help of LLMs' powerful reasoning abilities, Vision Language Models (VLMs) have achieved significant success in recent days. Different models, including Llava (Liu et al., 2024b), BLIP2 (Li et al., 2023), Flamingo (Alayrac et al., 2022), Qwen2-VL (Wang et al., 2024), mini-GPT4 (Zhu et al., 2023) have shown their impressive results among different vision-related tasks, which contains but not limited to Visual question answering (Biten et al., 2022; Guo et al., 2023; Özdemir and Akagündüz, 2024; Hu et al., 2024), image captioning (Rotstein et al., 2024; Yang et al., 2024b) or visual grounding (Peng et al., 2023; Yu et al., 2025). Among all tasks, document OCR (Wei et al., 2025; Lv et al., 2023) and its application, which outputs the bounding box for texts in the images and answers the question based on the texts, are the task most similar to ours, where our task is based on the bounding boxes for texts. However, none of the previous works have utilized VLMs for de-identification to protect the privacy of vision data. Our collected dataset and model not only address this gap but also expand the application scope of VLMs.

Instruction Tuning Instruction tuning is used to make language models follow natural language instructions and complete more complex tasks (Ouyang et al., 2022; Wang et al., 2022; Wei et al., 2021; Zhang et al., 2023a). Instruction tuning improves the zero- and few-shot generalization abilities of LLMs for both text-only LLMs, which include ChatGPT (Achiam et al., 2023; OpenAI, 2023), Llama family (Touvron et al., 2023; Dubey et al., 2024) and Flan family (Longpre et al., 2023; Chung et al., 2024), to VLMs (Liu et al., 2024b,a) with diverse vision prompts as additional inputs.

The quality of instruction tuning is highly dependent on the quality of the tuning dataset (Zhou et al., 2024). Therefore, previous works like Llava (Liu et al., 2024b,a) leverage LLMs to expand the existing image dataset (Lin et al., 2014) to various instruction-following datasets. In this work, we use a similar pipeline based on the flickr30k dataset (Plummer et al., 2015) and medical images (Rutherford et al., 2021).

De-identification De-identification is the process of removing or obfuscating personal information from data to prevent the identification of individuals (Ribaric et al., 2016). For image de-identification, most current methods aim at face images, where replacing faces in images to protect privacy (Gross et al., 2006; Brkic et al., 2017; Cao et al., 2021). However, to the best of our knowledge, there is no previous work focused on de-identifying burn-in pixels (texts in the images), especially with the help of VLMs. Therefore, our model fills the gap and extends the application range of VLMs.

3 Methodology

3.1 De-identification Pipeline

As shown in Fig. 2, our full de-identification pipeline contains prompting fine-tuned VLMs to output OCR results. Then, we mask out the text using the top-left color of every bounding box in the output. To achieve a successful de-identification as shown in the pipeline, two key tasks: 1) accurately localizing sensitive text and 2) processing it to ensure privacy protection are required. To perform these two tasks, we propose a framework called VisShield and introduce two components of VisShield: 1) a specialized dataset OPTIC for instruction tuning and 2) a training methodology.

3.2 OPTIC Dataset

Our instruction-tuning approach aims to enable VLMs to analyze and extract private information precisely through OCR. In order to achieve this

```
Prompt Used to Generate Instruction Prompts
You need to generate the instruction that guides MLLMs to do OCR for private information, your instruction should have:
1.   Define these private information:
     You should use 1-2 sentences to define what private information is, and you should randomly choose     one or more
     information including the following categories:
          [name, DOB, SSN, address, phone, email, medical record numbers, disease name]
     You should directly define what is private information like 'private information stands for names'. And you should the exact name
     I list here. Do not use the full name of information here. Please use diverse sentences to demonstrate the same meaning.
 2. Generate few-shot examples of the information:
      - Generate a random example with the information you choose
      - Use the the generated example as few-shot examples
      - one example for every information you choose

3. Contain the instruction:
      - must include a special token '' so that my model knows it should do OCR job.
      - You should not re-define what is private information here.
      - Please make the sentences as diverse as possible.
Format the response without anything else:
` ` ` INSTRUCTION
 [The full prompt including the defined sentence of private information, few-shot examples and instruction]
INFORMATION
 [Types of information you choose in the step 1 store in python list format like]` ` `
```

Figure 4: Template prompt utilized for instruction generation, implemented with GPT-4 and Claude-3.5 Sonnet. This prompt guides the LLMs to synthesize diverse task-specific instruction prompts.

goal, the OPTIC dataset contains in total of 50M sample sizes with various instruction prompts and images with private information.

3.2.1 Instruction Prompts

Config	Numbers	Options
Font	6	Arial, Times_New_Roman, Verdana, courbi, DejaVuSans, NotoSansMono
Font Size	N/A	3%-9% of the whole image
Font Color	9	White, Black, yellow, cyan, orange, pink, lightgreen, red, blue

Table 1: Detailed options of different generation configurations. During generation, we will random sample each configuration to ensure a diverse generation.

The instruction set encompasses four distinct contextual categories, which we detail in the following sections.

Definition of Private Information The notion of private information is inherently context-dependent and domain-specific. For instance, numerical sequences in medical contexts may represent confidential medical record identifiers, while similar numerical patterns in other domains might have no privacy implications. We explicitly incorporate contextual definitions within each instruction prompt to enable VLMs to identify and process private information across diverse scenarios accurately. These definitions follow a precise format (e.g., "Private information encompasses names and email addresses") to eliminate ambiguity and ensure consistent interpretation by the model.

Few-shot Examples Providing abstract definitions of private information alone is often insufficient for optimal VLM performance, as the format and structure of sensitive data vary significantly across contexts. For instance, medical record numbers follow institution-specific formats, while phone number structures differ across national boundaries. To enhance the instruction-following capabilities of VLMs and improve OCR accuracy for targeted information, we leverage in-context learning (Dong et al., 2022; Zhang et al., 2023b) by incorporating carefully curated few-shot examples into our instructions. These examples are specifically designed to align with and contextualize the provided definitions, enabling more robust recognition of diverse data formats.

Instruction The critical component of our instruction prompts is a targeted directive that guides VLMs to extract OCR results exclusively from private information. We leverage a specialized token *<ocr>* for OCR tasks. This token is consistently incorporated across all instructions, serving as a standardized trigger that signals the fine-tuned VLM to initiate OCR processing for privacy-relevant content within the prompted region.

Generation Building upon established methodologies (Liu et al., 2024b,a), we employ state-of-the-art large language models to generate diverse instruction prompts. Specifically, we utilize GPT-4 (OpenAI) and Claude-3.5 Sonnet (Anthropic), which represent the current frontier of language model capabilities. Our framework encompasses eight distinct categories of sensitive information, ranging from personally identifiable information (PII), such as email addresses and Social Security Numbers (SSN), to protected health information, including disease classifications. A comprehensive taxonomy of these information types is presented in Table 2. We developed structured prompts that direct these LLMs to randomly sample from these information categories, generate few-shot examples, and produce diverse task-specific instructions. The complete prompt template used for instruction generation is illustrated in Fig. 4, with a representative example of a generated instruction prompt shown in Appendix Fig. 6. We have a total of 2500 different instruction prompts, with 1250 generated by GPT-4o and 1250 generated by Claude-3.5-Sonnet.

Type of Information	Number	Example
Name	16300	Joe Dohn
DOB	16276	18 Jun 1983
SSN	16350	071-30-5000
Phone Number	16271	555-304-8389
Address	16270	086 Holt Summit, CT 58671
Email	16149	54jmz@hotmail.com
Medical Numbers	16243	MRN93987011
Disease Name	16274	Migraine

Table 2: Examples of information types we consider in this paper. We consider 8 types with balanced numbers of size in each type. All the information is fake.

3.2.2 Synthetic Images

To fine-tune the VLMs, we need images containing private information and bounding box annotations for the private information in images. However, since we are the first to address the challenge of textual private information in images, there is a lack of existing image datasets. In order to obtain the dataset, we create images with private information based on the base image datasets.

Base Image Dataset We overlay private information onto the base image dataset to generate vision data, where the base image dataset plays an important role. We hope the base image dataset includes diverse images to enhance generalization ability. Therefore, we first utilize the existing dataset that already has diverse images from image caption domains. In detail, we use the flickr30k dataset (Plummer et al., 2015) as the first part of the base image dataset. Additionally, we include the medical images in our base image dataset since the medical area is the most important application area for de-identification. Specifically, we use a public medical dataset containing various types of medical images from Rutherford et al. (2021).

Generation For the generation of our synthetic dataset, we first sample one base image from our base image datasets and then overlay the private information on the sampled image. In detail, after sampling the image, we determine the amount of private information to be overlaid on the sampled image by randomly selecting an integer between four and ten. Then for each piece of information, we randomly decide the type of the information and generate fake information using the Faker package (Joke and contributors, 2024). Then, we print the generated fake information on the sampled image using PIL package (Clark and contributors, 2024), which also provides the ground truth bounding box information for the text. While overlaying the information on the sampled image, we use different fonts, font sizes, and colors to ensure the diversity of generated text. The details of the generation configuration can be found at Table 1. In total, we generate 20,000 images with more than 130,000 bounding boxes.

3.2.3 Label Generation

So far, we have introduced the input part of our dataset. However, to fine-tune VLMs, we also need labels to optimize the loss function. Our target is to make VLMs output the OCR results for the defined private information. The labels should differ based on the same instruction prompt with different images or for different instruction prompts applied to the same image. Therefore, we first randomly sample one prompt from instruction prompts and one image from the synthetic image dataset to form the full input and then generate the label corresponding to the full input. We provide bounding boxes only for the private information types that are used to define private information in the instruction to generate labels. For example, if the instruction prompt specifies that 'private information only stand for names', then we will only provide bounding box for names in the given image as the label. If there is no such information in the image, the answer will be 'No private information'. If there is such information, the answer will be the concatenation of each bounding box which is expressed as <*bbox*> $< x_{tl} > < y_{tl} > < x_{br} > < y_{br} >$ </*bbox*>. The coordinates denote the top-left and bottom-right corners of the bounding box.

Model	Name		DOB		SSN		Email		Phone Number		Address		Medical Number		Disease Name	
	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU
Evaluation Set Generated by Training Base Image Dataset																
Full	**0.9733**	0.9134	0.9849	0.8984	**0.9781**	0.9103	**0.9719**	**0.9482**	0.9736	0.9045	0.9809	0.9615	**0.9762**	0.8626	0.9426	**0.8920**
LoRA	0.9728	**0.9194**	0.9849	**0.9196**	0.9714	**0.9205**	0.9601	0.9419	**0.9801**	**0.9144**	**0.9849**	**0.9690**	0.9714	**0.8898**	**0.9501**	0.8782
Presidio	N/A	0.0085	N/A	0.0074	N/A	0.0067	N/A	0.0119	N/A	0.0072	N/A	0.0141	N/A	0.0074	N/A	0.0067
Evaluation Set Generated by COCO																
Full	0.9708	0.9058	**0.9903**	**0.9472**	0.9767	0.8997	**0.9693**	0.9338	**0.9838**	0.9017	0.9703	0.9632	0.9637	0.8706	0.9565	0.8805
LoRA	**0.9713**	**0.9075**	0.9818	0.9083	**0.9859**	**0.9157**	0.9679	**0.9369**	0.9772	**0.9097**	**0.9802**	**0.9657**	**0.9818**	**0.8995**	**0.9661**	**0.8764**
Presidio	N/A	0.0067	N/A	0.0060	N/A	0.0054	N/A	0.0085	N/A	0.0057	N/A	0.1201	N/A	0.0057	N/A	0.0052
Evaluation Set Generated by ADE-20K																
Full	**0.9499**	**0.9075**	**0.9842**	**0.8849**	0.9576	**0.8918**	**0.9718**	0.9252	**0.9481**	**0.9200**	**0.9564**	**0.9508**	**0.9818**	0.8633	0.9606	0.8863
LoRA	0.9300	0.8921	0.9769	0.9025	**0.9740**	0.8913	0.9496	**0.9282**	0.9412	0.8984	0.9513	0.9453	0.9725	**0.8655**	**1.0000**	**0.8905**
Presidio	N/A	0.0027	N/A	0.0024	N/A	0.0021	N/A	0.0033	N/A	0.0022	N/A	0.0048	N/A	0.0023	N/A	0.0021
Evaluation Set Generated by RITE																
Full	0.9836	0.9251	0.9633	0.9093	**0.9863**	0.9149	**0.9842**	0.9449	**0.9911**	0.9176	**0.9910**	**0.9751**	**0.9902**	0.8777	**1.0000**	0.9058
LoRA	**0.9938**	**0.9723**	**0.9851**	**0.9785**	0.9843	**0.9953**	0.9689	**0.9669**	0.9109	**0.9304**	0.9266	0.9491	0.9210	**0.9760**	0.8966	**0.9118**
Presidio	N/A	0.0077	N/A	0.0070	N/A	0.0066	N/A	0.0096	N/A	0.0073	N/A	0.0126	N/A	0.0068	N/A	0.0062

Table 3: Comparative analysis of model performance across information categories, model architectures, and evaluation datasets. We evaluate using randomly sampled instruction prompts from the training set. Results demonstrate that our fine-tuned models achieve strong generalization capabilities, with full model fine-tuning consistently outperforming other adaptation strategies.

3.3 Training on OPTIC

While the OPTIC dataset provides a rich foundation for training privacy-aware VLMs, effectively leveraging it to improve the model's capability remains a significant challenge. To address this challenge, we introduce our training strategy and our strategy is built upon three key principles:

Efficiency While our dataset contains 50M samples, training on the full dataset is computationally expensive and unnecessary. Instead, we demonstrate that training on a **small subset of 100K samples** is sufficient to significantly enhance the model's de-identification capabilities. This approach allows us to reduce resource requirements.

Knowledge Transfer Instead of training a VLM from scratch, we fine-tune `Kosmos-2.5` (Lv et al., 2023), a pre-trained multimodal model that inherently supports OCR extraction from images. However, to make it privacy-aware, our fine-tuning process could improve its ability to selectively extract only privacy-relevant text rather than all OCR content, and refine its bounding box localization for privacy-sensitive elements.

Adaptation Strategies We explore two fine-tuning strategies to integrate privacy-awareness into the model. The first is **full fine-tuning**, where the entire model is fine-tuned on privacy-sensitive OCR tasks, while the second is **LoRA** (Hu et al., 2021), a parameter-efficient approach that updates only a limited set of trainable parameters, reducing memory consumption.

With our training strategy, we ensure that our end-to-end framework learns to effectively identify, localize, and process private textual information.

4 Experiments

In this section, we provide our experimental results to show the robustness of fine-tuned models. We start with the experimental setting at first.

4.1 Experimental Setting

Dataset To evaluate the robustness and generalization ability of the fine-tuned model, we test the fine-tuned models with five different datasets: 1) Images generated from the same base image dataset and the same instruction prompts in the training set, 2) Images from the same base image dataset and different instruction prompts from the training set, 3) Images from different base image dataset and different instruction prompts from the training set, 4) Images from different base image dataset with extra private information (not in 8 types of private information considered in training) and different instruction prompts from the training set, and 5) real-world images, which is annotated by human as described in (Orekondy et al., 2018). We will provide a more detailed introduction to these datasets in the following section.

Training Parameters For full fine-tuning, we use an epoch of 5, learning rate 2e-5 with batch size 16. For LoRA, following previous work (Sun et al., 2023), we use a larger learning rate 3e-4 and a larger epoch 10 with the same batch size. For both trainings, we use AdamW (Loshchilov, 2017) as

Model	Name		DOB		SSN		Email		Phone Number		Address		Medical Number		Disease Name	
	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU
Instruction Prompts Generated by Gemma1.5																
Full	0.9493	0.9008	0.9636	0.9013	0.9842	0.9075	0.9537	0.9290	0.9114	0.9080	0.9591	0.9644	0.9760	0.8586	0.9247	0.8973
LoRA	0.9561	0.9791	0.9764	0.9491	0.9721	0.9798	0.9669	0.9767	0.8960	0.9121	0.9177	0.9429	0.9130	0.9721	0.8815	0.8948
Presidio	N/A	0.0085	N/A	0.0074	N/A	0.0067	N/A	0.0119	N/A	0.0072	N/A	0.0141	N/A	0.0074	N/A	0.0067
Instruction Prompts Generated by Human																
Full	0.9420	0.9247	0.9943	0.9094	0.9723	0.9211	0.9129	0.9353	0.9842	0.9010	0.9823	0.9613	0.9511	0.8749	0.9746	0.9210
LoRA	0.9758	0.9667	0.9847	0.9499	0.9799	0.9560	0.9414	0.9877	0.9196	0.9251	0.9247	0.9447	0.9333	0.9675	0.8751	0.8911
Presidio	N/A	0.0085	N/A	0.0074	N/A	0.0067	N/A	0.0119	N/A	0.0072	N/A	0.0141	N/A	0.0074	N/A	0.0067

Table 4: Performance comparisons for different types of information, different models, and different instruction prompts. The evaluation image set is chosen for the evaluation set generated by the training base image dataset.

Model	Name		DOB		SSN		Email		Phone Number		Address		Medical Number		Disease Name	
	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU
Instruction Prompts Generated by Gemma1.5																
Full	0.9483	0.9062	0.9625	0.8985	0.9771	0.9000	0.9309	0.8990	0.9245	0.9090	0.9782	0.9625	0.9464	0.8673	0.8586	0.8942
LoRA	0.9852	0.9689	0.9851	0.9636	0.9576	0.9751	0.9635	0.9749	0.9017	0.9078	0.9105	0.9309	0.9100	0.9669	0.8915	0.8906
Presidio	N/A	0.0067	N/A	0.0060	N/A	0.0054	N/A	0.0085	N/A	0.0057	N/A	0.1201	N/A	0.0057	N/A	0.0052
Instruction Prompts Generated by Human																
Full	0.9586	0.9027	0.9928	0.9042	0.9636	0.9153	0.9234	0.9389	0.9697	0.9132	0.9129	0.9626	0.9391	0.8786	0.9139	0.8902
LoRA	0.9761	0.9826	0.9879	0.9621	0.9602	0.9564	0.9695	0.9727	0.9026	0.9094	0.9139	0.9337	0.9225	0.9668	0.8980	0.9004
Presidio	N/A	0.0067	N/A	0.0060	N/A	0.0054	N/A	0.0085	N/A	0.0057	N/A	0.1201	N/A	0.0057	N/A	0.0052

Table 5: Performance comparisons for different types of information, different models, and different instruction prompts. The evaluation image set is chosen for the evaluation set generated by COCO.

the optimizer. All training methods are conducted on a single Nvidia Tesla A100 80GB GPU.

Metrics In this paper, we mainly consider two different metrics to measure the quality. Following previous works (Olejniczak and Šulc, 2022; Ren et al., 2016), we use F1 to evaluate the quality of OCR results for defined private information and use the Intersection over Union (IoU) to evaluate the quality of detection, which are both important for the following mask out procedure.

Research Questions In this section, we mainly focus on three different research questions about the generalization ability of the fine-tuned Model: 1) Whether fine-tuned VLM is stable for different images, 2) Whether fine-tuned VLM is stable for various instructions and 3)Whether the fine-tuned VLM is stable for new information types. Besides, Our experimental results also show that our fine-tuned VLM performs well even in real-world data and we put the detailed results in Appendix.

4.2 RQ1: Whether Fine-tuned VLM is Stable for Different Images

To answer this research question, we use different base image datasets to generate the evaluation set. We only provide the results for our method in most cases. In detail, we consider using: 1) our training base image dataset, 2) COCO (Lin et al., 2014), 3) ADE20K (Zhou et al., 2017), and 4) RITE (Hu et al., 2013) to generate evaluation image datasets, ensuring comprehensive scenarios from city scene to medical images considered in the experiments. We generate 1500 images for each dataset with the same generation methods but more generation configurations. We compare our model with Presidio (Microsoft, 2023) and the results are shown in Table 3. The F1 score for Presidio is N/A because it cannot output OCR results. We have the following observations:

1) The previous tool Presidio shows a bad performance. Since we cannot customize the private definition for Presidio, the performance of Presidio is highly random for different types of information.
2) Our fine-tuned model shows a very good performance with a mean IoU larger than 0.9. And this good performance remains for various image datasets, showing the robustness of our method.
3) There is no clear winner for full fine-tuning and LoRA. Though the LoRA model wins more times, this winning is marginal given the good performance of both models.

4.3 RQ2: Whether Fine-tuned VLM is Stable for Various Instructions

To answer the research question related to various instructions, we generate instruction prompts that are different from our training set by involving human writers and Gemini (Team et al., 2023), and then pair the new prompts with three image datasets we used before with one-shot examples. We generate 1500 text-image pairs for model evaluation, and the results are shown in Table 4 and Table 5. We have the following observations:

1) Compared with the results in Table 3, the performance of both full fine-tuning and LoRA exhibits a slight decrease. However, this decrease is minimal, and the fine-tuned models continue to deliver strong performance.

2) Even when using a different image dataset and Instruction Prompts together, our models still achieve strong performance for the de-identification task.

4.4 RQ3: Whether Fine-tuned VLM is Stable for New Information Type.

Now, we conduct experiments to test the performance of fine-tuned VLM on new information types. Here, we focus on two new types of information: 1) phone numbers with a format of 11 digits and 2) passport number that begins with a letter and ends with eight numbers. We use a similar method to generate the evaluation set and we regenerate the instruction prompts with the one-shot prompt to ask models to output OCR results for new types of information. We present our results in Table 2. We find that:

1) Overall, our fine-tuned models continue to demonstrate strong performance when incorporating new types of information, further highlighting their robustness and reliability.

2) Compared to 11-digit phone numbers, the performance on passport numbers is lower because our models had not previously encountered the format of passport numbers. In contrast, earlier phone numbers share a similar pattern with the new ones, aiding the model's performance.

4.5 Ablation Study

In this section, we provide a comparison of the performance of one-shot prompts and zero-shot prompts. More ablation study results can be found in the Appendix. Here, we consider the 11-digit Phone Number and Passport Number as in Section 4.4, and the results for various datasets are presented in Fig. 5. We found that:

1) Compared with the one-shot prompt, using the zero-shot prompt can lead to better performance across different datasets, highlighting the importance of few-shot examples.

Model	11-Digit Phone Number		Passport Number	
	F1	IoU	F1	IoU
Evaluation Set Generated by Training Base Image Dataset				
Full	0.9803	0.8724	0.8887	0.8596
LoRA	0.9803	0.8887	0.8725	0.8597
Presidio	N/A	0.0071	N/A	0.0064
Evaluation Set Generated by COCO				
Full	0.9796	0.8679	0.8920	0.8625
LoRA	0.9023	0.8167	0.8776	0.8583
Presidio	N/A	0.0086	N/A	0.0054
Evaluation Set Generated by RITE				
Full	0.9910	0.8761	0.9271	0.8758
LoRA	0.8678	0.7463	0.8892	0.8700
Presidio	N/A	0.0075	N/A	0.0069

Table 6: Performance comparisons for new types of information, different models, and different evaluation image sets.

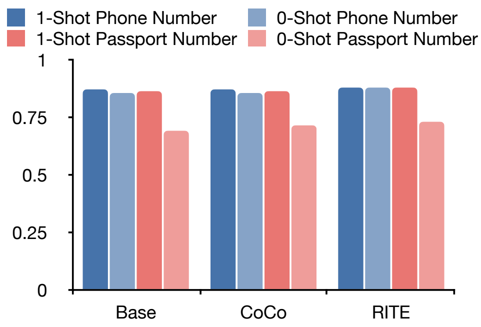

Figure 5: IoU performance comparison with different Dataset on 11-digit Phone Number and Passport Number. The experiments are on the full fine-tuned model.

2) The performance gap between two prompts is larger when we consider passport numbers. This is because the model has seen similar phone numbers during training, but it never encountered anything similar to passport numbers before. This highlights the importance of few-shot examples.

5 Conclusion

In conclusion, this work presents a novel approach to de-identify textual information in visual data by leveraging the power of VLMs. We generate a comprehensive instruction-tuning dataset with diverse images and instruction prompts. By fine-tuning `Kosmos-2.5` with this comprehensive instruction-tuning dataset, we demonstrated that VLMs can effectively identify and mask private information. Our results show strong generalization and robustness across different datasets and real-world scenarios, laying a foundation for safer integration of VLMs into privacy-sensitive applications.

Limitation

While our approach demonstrates strong performance, it has two key limitations. First, the model's effectiveness depends on the quality of the instruction-tuning dataset, and while we have ensured diversity, rare or highly domain-specific private information formats may still pose challenges. Second, our method relies on OCR accuracy for text extraction, meaning that errors in detecting or recognizing text in low-quality or distorted images could affect de-identification performance.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, et al. 2023. 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. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Anthropic. Claude 3.5: A Sonnet. https://www.anthropic.com/news/claude-3-5-sonnet. Accessed: 2024-11-10.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Ali Furkan Biten, Ron Litman, Yusheng Xie, Srikar Appalaraju, and R Manmatha. 2022. Latr: Layout-aware transformer for scene-text vqa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16548–16558.

Karla Brkic, Ivan Sikiric, Tomislav Hrkac, and Zoran Kalafatic. 2017. I know that person: Generative full body and face de-identification of people in images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1319–1328. IEEE.

Jingyi Cao, Bo Liu, Yunqian Wen, Rong Xie, and Li Song. 2021. Personalized and invertible face de-identification by disentangled identity information manipulation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3334–3342.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Jeffrey A. Clark and contributors. 2024. Pillow. A friendly fork of the Python Imaging Library (PIL).

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ralph Gross, Latanya Sweeney, Fernando De la Torre, and Simon Baker. 2006. Model-based face de-identification. In *2006 Conference on computer vision and pattern recognition workshop (CVPRW'06)*, pages 161–161. IEEE.

Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. 2023. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10867–10877.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Qiao Hu, Michael D Abràmoff, and Mona K Garvin. 2013. Automated separation of binary overlapping trees in low-contrast color retinal images. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II 16*, pages 436–443. Springer.

Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. 2024. Bliva: A simple multimodal llm for better handling of text-rich visual questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2256–2264.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2022. Are large pre-trained language models leaking your personal information? *arXiv preprint arXiv:2205.12628*.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2022. Knowledge unlearning for mitigating privacy risks in language models. *arXiv preprint arXiv:2210.01504*.

Edén Joke and contributors. 2024. Faker: Python package. Version 15.3.4.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

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

Zheyuan Liu, Guangyao Dou, Mengzhao Jia, Zhaoxuan Tan, Qingkai Zeng, Yongle Yuan, and Meng Jiang. 2024c. Protecting privacy in multimodal large language models with mllmu-bench. *arXiv preprint arXiv:2410.22108*.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Tengchao Lv, Yupan Huang, Jingye Chen, Yuzhong Zhao, Yilin Jia, Lei Cui, Shuming Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, et al. 2023. Kosmos-2.5: A multimodal literate model. *arXiv preprint arXiv:2309.11419*.

Microsoft. 2023. Presidio - open source data protection and privacy engineering platform. https://microsoft.github.io/presidio/. Accessed: 2023-11-14.

Fatemehsadat Mireshghallah, Archit Uniyal, Tianhao Wang, David Evans, and Taylor Berg-Kirkpatrick. 2022. Memorization in nlp fine-tuning methods. *arXiv preprint arXiv:2205.12506*.

Krzysztof Olejniczak and Milan Šulc. 2022. Text detection forgot about document ocr. *arXiv preprint arXiv:2210.07903*.

OpenAI. GPT-4 Turbo System Card. https://openai.com/index/gpt-4o-system-card/. Accessed: 2024-11-10.

OpenAI. 2023. Chatgpt. Accessed: 2023-11-10.

Tribhuvanesh Orekondy, Mario Fritz, and Bernt Schiele. 2018. Connecting pixels to privacy and utility: Automatic redaction of private information in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8466–8475.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Övgü Özdemir and Erdem Akagündüz. 2024. Enhancing visual question answering through question-driven image captions as prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1562–1571.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149.

Slobodan Ribaric, Aladdin Ariyaeeinia, and Nikola Pavesic. 2016. De-identification for privacy protection in multimedia content: A survey. *Signal Processing: Image Communication*, 47:131–151.

Noam Rotstein, David Bensaïd, Shaked Brody, Roy Ganz, and Ron Kimmel. 2024. Fusecap: Leveraging large language models for enriched fused image captions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5689–5700.

Michael Rutherford, Seong K Mun, Betty Levine, William Bennett, Kirk Smith, Phil Farmer, Quasar Jarosz, Ulrike Wagner, John Freyman, Geri Blake, et al. 2021. A dicom dataset for evaluation of medical image de-identification. *Scientific Data*, 8(1):183.

Xianghui Sun, Yunjie Ji, Baochang Ma, and Xiangang Li. 2023. A comparative study between full-parameter and lora-based fine-tuning on chinese instruction data for instruction following large language model. *arXiv preprint arXiv:2304.08109*.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022. Benchmarking generalization via in-context instructions on 1,600+ language tasks. *arXiv preprint arXiv:2204.07705*, 2.

Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2025. Vary: Scaling up the vision vocabulary for large vision-language model. In *European Conference on Computer Vision*, pages 408–424. Springer.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Tianyu Yang, Xiaodan Zhu, and Iryna Gurevych. 2024a. Robust utility-preserving text anonymization based on large language models. *arXiv preprint arXiv:2407.11770*.

Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Xin Geng. 2024b. Exploring diverse in-context configurations for image captioning. *Advances in Neural Information Processing Systems*, 36.

En Yu, Liang Zhao, Yana Wei, Jinrong Yang, Dongming Wu, Lingyu Kong, Haoran Wei, Tiancai Wang, Zheng Ge, Xiangyu Zhang, et al. 2025. Merlin: Empowering multimodal llms with foresight minds. In *European Conference on Computer Vision*, pages 425–443. Springer.

Ziqian Zeng, Jianwei Wang, Junyao Yang, Zhengdong Lu, Huiping Zhuang, and Cen Chen. 2024. Privacyrestore: Privacy-preserving inference in large language models via privacy removal and restoration. *arXiv preprint arXiv:2406.01394*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023a. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. 2023b. What makes good examples for visual in-context learning? *Advances in Neural Information Processing Systems*, 36:17773–17794.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Example of Instruction prompt

B More Experiments

In this section, we provide more experimental results to support our conclusion.

B.1 mAP Results

Here, we provide the results for mean Average Precision (mAP) to further demonstrate the results of our experiments. Following previous works in detection, we consider a correction if IoU > 0.5. And the results for different images are provided in Table 8 and Table 9. The results in both experiments show that our fine-tuned models also have a very good mAP result, which is reasonable since our IoU results are very high.

B.2 Experiments on Real-world Data

In this section, we use real-world data to test the robustness of the fine-tuned models. In detail, we use images from (Orekondy et al., 2018), which contains real-world images from different scenarios. And human annotators will annotate the images with private information and the corresponding bounding box information. More specifically, we focus on names and phone numbers. Then, we use instructions that define private information as names and phone numbers to test the performance on real-world data. Our results can be found in Table 7. Our experimental results show that even though the performance drops, our full fine-tuned model can also perform well in real-world data, showing good robustness of the model fine-tuned with our dataset.

Model	Phone Number		Name	
	F1	mAP	F1	mAP
Full	0.7001	0.5439	0.7229	0.6037
Presidio	N/A	0.0002	N/A	0.0003

Table 7: Performance comparisons for different types of information, different models on real-world dataset

B.3 More ablation studies

In this section, we provide more results of our ablation studies. In detail, we provide the results for the different number of few-shot examples and different training sizes.

For the different number of few-shot examples, we consider using instruction prompts as well as few-shot examples written by human. We focus on the Medical Numbers and Email using CoCo as base image dataset. And the results are shown in Fig. 7. We can see that using few-shot examples can boost the performance. However, without using few-shot examples, we can still get a decent result.

In Fig. 8, we present our results for different sizes of training datasets for using CoCo as the base image dataset and instructions from the training set. From the figure we can observe that using 100k training pairs is more than enough to get a good result, showing the potential ability to use VLMs to de-identify data.

B.4 Example on Real-world Dataset

In Fig. 9, we present an example of applying our fine-tuned model on the real-world dataset. From the figure, we can see that the names and the phone number are correctly masked by our de-identification pipeline.

Generated Instruction Prompt

INSTRUCTION
Private information includes SSN, address, and medical record numbers, as they are sensitive and often used for identity verification or medical purposes.

Examples:
- SSN: 123-45-6789
- Address: 456 Elm Street, Apt. 12B, Springfield, IL 62704
- Medical Record Number: MRN-9876543210

<ocr> Extract and capture any visible private information in the image, focusing on elements like the specified codes, addresses, or identifiers.

INFORMATION
["SSN", "address", "medical record numbers"]

Figure 6: One instruction prompt example generated by GPT-4o.

Model	Name	DOB	SSN	Email	Phone Number	Address	Medical Number	Disease Name
Evaluation Set Generated by Training Base Image Dataset								
Full	0.9478	0.9479	0.9482	0.9482	0.9480	0.9484	0.9478	0.9492
Presidio	0.0007	0.0006	0.0005	0.0006	0.0007	0.0012	0.0004	0.0004
Evaluation Set Generated by COCO								
Full	0.9470	0.9472	0.9472	0.9472	0.9473	0.9470	0.9468	0.9467
Presidio	0.0006	0.0005	0.0005	0.0006	0.0006	0.0011	0.0005	0.0004
Evaluation Set Generated by ADE-20K								
Full	0.9196	0.9196	0.9198	0.9198	0.9200	0.9199	0.9197	0.9196
Presidio	0.0002	0.0002	0.0001	0.0002	0.0002	0.0003	0.0001	0.0001
Evaluation Set Generated by RITE								
Full	0.9394	0.9388	0.9398	0.9396	0.9399	0.9397	0.9398	0.9400
Presidio	0.0003	0.0003	0.0003	0.0003	0.0003	0.0007	0.0003	0.0003

Table 8: Comparative analysis of model performance across information categories, model architectures, and evaluation datasets using mAP as the metric.

Model	Name	DOB	SSN	Email	Phone Number	Address	Medical Number	Disease Name
Instruction Prompts Generated by Gemini1.5								
Full	0.8933	0.8932	0.8932	0.8930	0.8931	0.8929	0.8928	0.8933
Presidio	0.0007	0.0006	0.0005	0.0006	0.0007	0.0012	0.0004	0.0004
Instruction Prompts Generated by Human								
Full	0.9221	0.9229	0.9234	0.9224	0.9231	0.9233	0.9223	0.9233
Presidio	0.0006	0.0005	0.0005	0.0006	0.0006	0.0011	0.0005	0.0004

Table 9: Performance comparisons for different types of information, different models, and different instruction prompts. The evaluation image set is chosen to evaluation set generated by the training base image dataset using mAP as the metric.

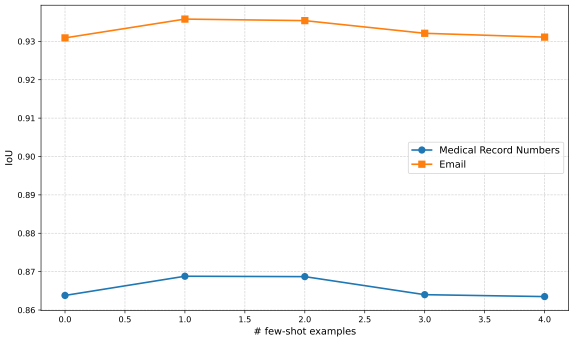

Figure 7: IoU performance comparison with different numbers of few shot examples.

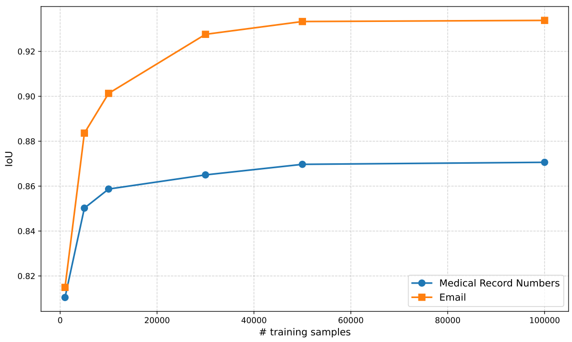

Figure 8: IoU performance comparison with different sizes of training dataset

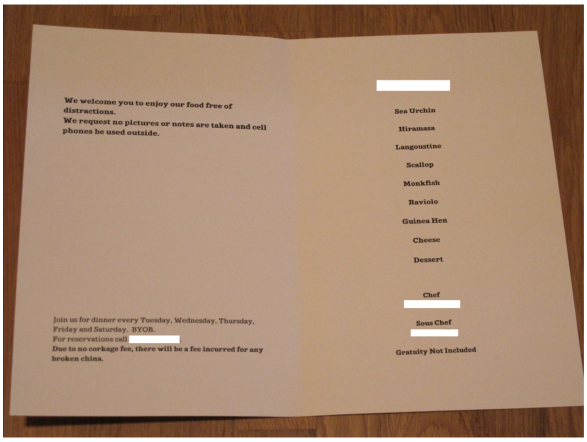

Figure 9: A real-world image example that de-identified by our pipeline.