Your Prompts Are Not Safe: Output-Free Membership Inference via Prompt Vectors in Vision-Language Tuning

Yuran Bian[1,2], Xiaohan Zhang[1,2], Zhiyuan Yu[4], Changqing Li[1,2,3], Li Pan[1,2,3]*

[1]School of Computer Science, Shanghai Jiao Tong University, Shanghai 200240
[2]Shanghai Key Laboratory of Integrated Administration Technologies for Information Security, Shanghai 200240, China
[3]Zhangjiang Institute for Advanced Study, Shanghai, 201203, China
[4]Washington University in St. Louis
bianyrrr@sjtu.edu.cn, xhzhang1@sjtu.edu.cn, yu.zhiyuan@wustl.edu,
stari1nk@sjtu.edu.cn, panli@sjtu.edu.cn

Abstract

Prompt tuning enables Vision-Language Models (VLMs) to efficiently adapt to new tasks through learnable prompt vectors. This naturally raises a question: do these prompts leak private information about their training data? While Membership Inference Attacks (MIAs) can quantify this risk, current methods rely on access to model outputs or internal gradients. This limitation prevents a clear assessment of a prompt's standalone privacy leakage, particularly in deployment scenarios where such information is inaccessible. In this paper, we propose Prompt Intrinsic Privacy Risk Analyzer (PIPRA) to address this gap. As the first output-free MIA, PIPRA leverages open-source pre-trained VLMs to extract features from both prompts and samples within a shared cross-modal semantic space. By employing a contrastive learning-based feature projector to enhance these representations, PIPRA enables a subsequent discriminator to effectively perform membership inference. Extensive experiments across nine benchmark datasets and multiple VLMs show PIPRA achieves an average AUC of 87.58%, significantly outperforming traditional output-dependent methods (77.05%). These findings reveal that prompts pose a substantially greater privacy risk than previously recognized, highlighting the urgent need for prompt-level privacy protection.

Introduction

In recent years, prompt tuning has rapidly emerged as a lightweight and efficient alternative to traditional fine-tuning for adapting large vision-language models (VLMs) to downstream tasks (Zhou et al. 2022b,a; Sun et al. 2023). By introducing a small and learnable context—known as the prompt—this technique enables service providers to leverage powerful pre-trained models while flexibly and efficiently supporting diverse, user-specific applications, all without updating the entire model (Zhou et al. 2022b; Wang et al. 2023). This paradigm shift has greatly improved the scalability and accessibility of deploying customized AI services across domains (Wang et al. 2024; Yu et al. 2023; Alayrac et al. 2022).

Despite these advantages, prompt tuning introduces new and underexplored privacy risks (Li et al. 2024a; Wu et al. 2024; Liu et al. 2025). In real-world deployments, prompts are often transmitted via cloud APIs or shared through frameworks like PEFT on platforms such as Hugging Face Hub. However, these prompts—treated as standalone assets—face risks of interception during transmission or storage. Moreover, prompts are usually tuned on small, information-rich datasets that may contain sensitive or proprietary data (Li et al. 2024a; Samson et al. 2024). If prompt parameters unintentionally encode such data, they could expose users' intellectual property or personal information. A notable example is the membership inference attack (MIA), which determines whether a sample was used in prompt tuning (Shokri et al. 2017), potentially revealing user data and intentions and posing serious privacy threats.

Recent studies (Wu et al. 2024) demonstrate that by exploiting the output predictions or gradient signals of a VLM after prompt tuning, an adversary can infer whether a given sample was used to optimize the prompt. However, these approaches rely on the strong assumption that adversaries have access to the target model's output predictions, which is often unrealistic in real-world scenarios where outputs may be suppressed (Rezaei and Liu 2021; Hu et al. 2025), privatized, or entirely inaccessible. Meanwhile, in many commercial and open platforms that provide prompt tuning as a service, prompts are often directly exposed to users or service providers (Yu et al. 2023; Li et al. 2024a), making them a natural attack surface. Surprisingly, the privacy risks of directly analyzing prompt vectors for MIA have not been systematically investigated in existing literature.

To bridge this gap, we propose PIPRA (Prompt Intrinsic Privacy Risk Analyzer), the first MIA framework designed to assess the inherent privacy risks of prompts in vision-language prompt tuning. Unlike previous approaches, PIPRA leverages only the semantic relationship between the prompt vector and given samples, inferring membership status by analyzing their distributional patterns in the aligned cross-modal semantic space. Specifically, it leverages open-source pre-trained VLMs to extract feature vectors of the intercepted prompt and given samples in the shared seman-

*Corresponding author.

tic space. These features are then passed through a feature projector based on contrastive learning, which is optimized using positive and negative sample pairs constructed from multiple shadow prompts. It effectively enlarges or narrows the distance between non-member and member sample pairs, thereby facilitating accurate membership inference by a downstream discriminator.

Even when traditional black-box MIAs fail due to output suppression defenses (Chen and Pattabiraman 2024), PIPRA remains effective by analyzing the geometric structure of the semantic space, consistently outperforming the best baseline by over 5% AUC on most datasets.

Overall, our contributions can be summarized as follows:

- We propose an output-free MIA framework that, unlike prior methods relying on the overfitting responses of the target model to its training data, performs membership inference by analyzing the distributional characteristics of prompt vectors and given samples in the cross-modal semantic space.
- We propose a novel approach for systematically evaluating whether a prompt alone can leak information about its training set. Our findings demonstrate that the prompt itself poses a significant privacy risk, highlighting the urgent need for protective measures at both the prompt and semantic representation levels.
- Across nine public datasets, our output-free scheme consistently outperforms traditional methods, notably achieving 90.37% accuracy on Caltech101. We further validate the robustness of PIPRA to the choice of different open-source pre-trained models and shadow datasets.

Related Work

Privacy Risks in Prompt Tuning for VLMs With the rapid development of prompt tuning in VLMs (Zhou et al. 2022b,a), numerous privacy concerns have also emerged, including adversarial image attacks via learnable prompts and the injection of malicious instructions into medical images (Zhang et al. 2024; Clusmann et al. 2025). Wu et al. highlighted that the prompt tuning process poses privacy risks of leaking membership or attribute information (Wu et al. 2024). However, they overlook the inherent privacy risks embedded in the prompt vectors themselves, instead providing a coarse-grained assessment of the system as a whole.

Membership Inference Attack MIA is one of the privacy inference attacks that can quantify the privacy risks of target models (Shokri et al. 2017; Wu et al. 2024; Huang et al. 2025). They are typically divided into white-box MIAs and black-box MIAs. White-box MIAs infer membership status by analyzing internal signals of the target model, such as gradients or training trajectories (Nasr, Shokri, and Houmansadr 2019; Leino and Fredrikson 2020; Rezaei and Liu 2021). However, they are impractical in real-world systems where internals are not exposed to external users (Leemann, Pawelczyk, and Kasneci 2023; Liu et al. 2022a; Duan et al. 2023), severely restricting its applicability in commercial or API-based platforms (Yu et al. 2023). To bypass the limitations of white-box MIAs, black-box MIAs infer membership using outputs like confidence scores (Shokri et al. 2017; Salem et al. 2018), prediction entropy (Liu et al. 2022b), or loss values (Carlini et al. 2022). Recent strategies include decision boundary distance (Li and Zhang 2021), adversarial robustness (Hu et al. 2022), and explainability-based signals (Liu et al. 2024). However, they are ineffective as real-world defenses increasingly restrict access to model outputs (Carlini et al. 2019). They also lack the capability to assess the privacy risks inherent to the prompts themselves due to their reliance on the overall model's output predictions. Moreover, these attacks exploit overfitting, which performs poorly under regularization or few-shot training, leading to high false positive and negative rates (Zarifzadeh, Liu, and Shokri 2024; Li et al. 2024b). In contrast, PIPRA leverages the distributional patterns between prompts and given samples in the shared semantic space of VLMs, thereby independently assessing the inherent privacy threats of the prompt vectors.

System Model and Problem Formulation

This section describes the process of prompt tuning for vision-language model as a service and establishes the threat model under which our attack is conducted.

Prompt-Tuning-as-a-Service Framework

The target system is designed to provide prompt-based inference services built upon a frozen vision-language backbone (Yu et al. 2023; Zhou et al. 2022b). The user specifies the task (e.g., diagnostic classification of COVID-19 CT images) and uploads a small-scale fine-tuning dataset to the service provider. On the backend, the service provider hosts a pre-trained, parameter-frozen VLM g_δ (e.g., a CLIP architecture). Using the fine-tuning dataset, the provider optimizes a task-specific prompt vector $\mathbf{p}^\star$, which is given to the user for future inference as shown in Figure 1(a). The

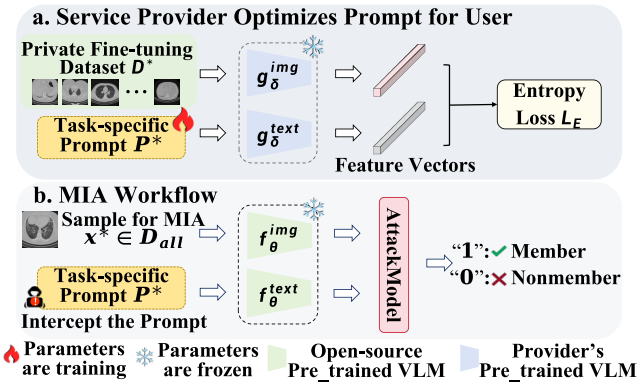


Figure 1: attack scenario. (a) illustrates the normal process of a service provider optimizing a prompt for a user. (b) shows the workflow of an adversary performing MIA using the intercepted prompt.

user then submits the task dataset to be classified. By using the frozen model g_δ together with the optimized prompt $\mathbf{p}^\star$, the service provider performs inference and returns the prediction results to the user.

Vision-Language Prompt Tuning Preliminaries

To clarify the optimization process of the prompt, it is represented as a set of learnable continuous vectors:

$$\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_M], \tag{1}$$

where M denotes the prompt length. The prompt vector is prepended to the input text tokens and fed into the frozen pre-trained model $m_\alpha(\cdot)$. Given an input sample $\mathbf{x}$ with label y, the model computes the class probability as:

$$p(y \mid \mathbf{x}; \mathbf{P}) = \mathrm{softmax}\big(m_\alpha([\mathbf{P}; \mathbf{x}])\big), \tag{2}$$

where $[\mathbf{P}, \mathbf{x}]$ denotes the concatenation of the prompt vectors and the input tokens. The prompt vectors $\mathbf{P}$ are optimized by minimizing the cross-entropy loss over the training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$:

$$\min_{\mathbf{P}} \ \mathcal{L}(\mathbf{P}) = -\frac{1}{N}\sum_{i=1}^{N} \log p(y_i \mid \mathbf{x}_i; \mathbf{P}). \tag{3}$$

During training, the model parameters α remain fixed, and only the prompt vectors $\mathbf{P}$ are updated. This approach enables efficient adaptation to downstream tasks with significantly fewer trainable parameters.

Attack Scenario and Threat Model

Attack Scenario In the scenario of Prompt-Tuning-as-a-Service, the prompt is directly exposed to the users and is vulnerable to interception by adversaries. If the prompt itself leaks the privacy of its training data, it can result in significant privacy breaches of the training set, even in service systems where output access is restricted as Figure 1(b) shows.

Adversary's Goal Given a target prompt vector $\mathbf{p}^\star$ itself, the adversary aims to infer whether a given sample $\mathbf{x}^\star$ was included in the training dataset $\mathbf{D}^\star$ used to optimize this prompt.

Adversary's Knowledge and Capabilities

- **Intercepting target prompt.** The adversary is assumed to intercept the target prompt $\mathbf{p}^\star$ during its transmission from the service provider to the user.
- **Utilizing open-source VLMs.** The adversary can invoke publicly available open-source pre-trained VLMs f_θ with similar architectures. This approach allows for the extraction of distributional features in the aligned cross-modal semantic space, eliminating the high cost of training shadow models from scratch, which is a major bottleneck in conventional reference-based MIAs.
- **Shadow dataset $\mathcal{D}_{\text{shadow}}$.** Unlike other reference-based attacks, PIPRA does not require a perfectly matching shadow dataset $\mathcal{D}_{\text{shadow}}$. We find that a dataset sharing the same domain and visual style is sufficient.
- **Acting as a legitimate user.** An adversary can operate as a legitimate customer, uploading a shadow dataset to the service provider to obtain corresponding shadow prompts. These special shadow prompts can then be used to select the most suitable open-source pre-trained model for maximizing the effectiveness of the MIA.

Our Proposal

In this section, we propose an output-free MIA that enables quantification of the inherent privacy risks that prompt vectors pose to their training data.

Overview

Our motivation stems from a key observation: an optimized prompt vector's semantic and geometric distribution in the embedding space is closer to its training data. This insight allows us to assess a prompt's privacy risks by detecting this latent signature without relying on the model's output. We divide the entire process of quantifying the inherent privacy risks of prompts into **two stages: attack model training and membership inference**. As Figure 2 shows, the first stage involves three steps to generate an attack model for membership inference, while the second stage aim to use this attack model to infer the membership status of given samples as shown in Figure 1(b).

In the **first step** of the stage for training the attack model, PIPRA partitions the shadow dataset into multiple subsets, each further divided into training and testing splits that correspond to genuine member and non-member samples. In the **second step**, each shadow prompt vector is trained using the member samples from its corresponding subset. Subsequently, feature representations of the prompts and all samples are extracted within a shared embedding space using publicly available pre-trained models. Feature vectors of member samples and prompts are paired as positive samples, while those of non-member samples and prompts are paired as negative samples. These pairs are then used to train a contrastive learning–based feature projector and a membership discriminator in the **third step**. In the final inference stage, the attack model requires only the target prompt vector. By analyzing their distributional characteristics in the shared semantic space, it can efficiently infer the membership status of given samples.

Stage1: Training Process for the Attack Model

Step1: Shadow Dataset Splitting To facilitate the attack model's learning of distributional differences and associations among multiple shadow prompts and samples, we first partition the shadow dataset into multiple subsets in this step, thereby generating multiple exemplars for the attack model's training. We randomly partitioning the shadow dataset $\mathcal{D}_{\text{shadow}}$ into k disjoint subsets:

$$\mathcal{D}_{\text{shadow}} = \bigcup_{i=1}^{k} \mathcal{D}_i \tag{4}$$

$$\mathcal{D}_i \cap \mathcal{D}_j = \emptyset \quad (i \neq j) \tag{5}$$

For each subset $\mathcal{D}_i$, half of the samples are randomly selected as the member subset $\mathcal{D}_i^{\text{in}}$, which serves as the training set for optimizing the corresponding shadow prompt $\mathbf{p}_i$ later. The remaining half are taken as the non-member subset $\mathcal{D}_i^{\text{out}}$. These well-organized labeled samples can simulate genuine member and non-member samples, thereby providing the attack model with the necessary knowledge to effectively distinguish between member and non-member samples.

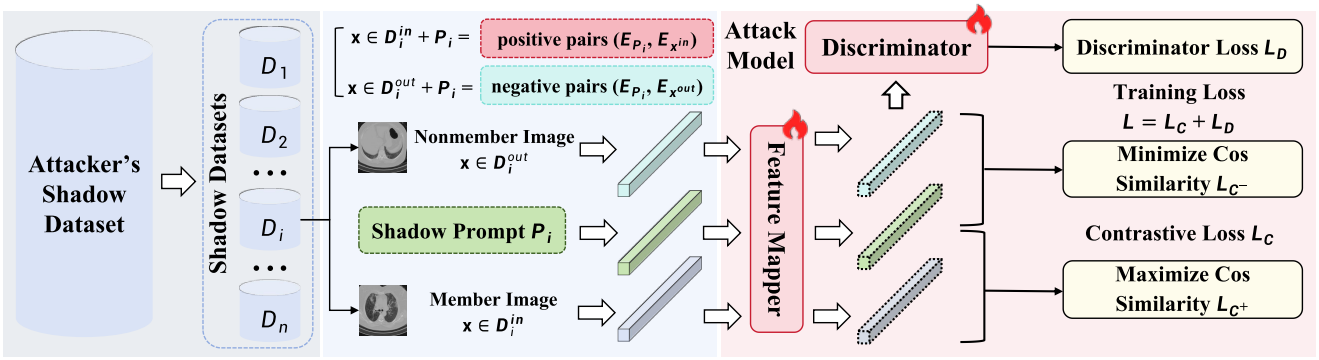


Figure 2: Overview of the Training Process for the Attack Model.

Step2: Shadow Prompt Training and Sample Pair Preparing In the second step, we train each shadow prompt $\mathbf{p}_i$ using the member set of its corresponding subset from the aforementioned partition, and construct positive and negative sample pairs based on the member and non-member labels annotated in the first step.

According to our intuition, the prompt vector $\mathbf{p}_i \in \mathbb{R}^d$ shares more similar patterns of semantic feature distribution with samples from $\mathcal{D}_i^{\text{in}}$ in the feature space. Therefore, constructing positive and negative sample pairs corresponding to the generated shadow prompts enables the contrastive learning module in the attack model to better capture the representational differences in the feature space, further distinguishing the distributional characteristics between member and non-member samples. Specifically, we first train a prompt vector $\mathbf{p}_i$ for each member subset $\mathcal{D}_i^{\text{in}}$ using few-shot learning methods by minimizing the loss function:

$$\mathbf{p}_i = \arg\min_{\mathbf{p}} \mathcal{L}\big(f_\theta(\mathbf{p}, \mathcal{D}_i^{\text{in}}), y_i\big),$$

where f_θ denotes the pre-trained model with parameters θ, and y_i represents the corresponding labels of the samples in $\mathcal{D}_i^{\text{in}}$. Then, for each sample $\mathbf{x}$ and its corresponding shadow prompt $\mathbf{p}_i$, we input them into the pre-trained model $f_\theta(\cdot)$ respectively to obtain their feature representations in the aligned multi-modal semantic space, and form sample pairs accordingly: given a sample $\mathbf{x}$ and its corresponding prompt vector $\mathbf{p}_i$, we use $f_\theta(\cdot)$ to form sample feature pairs.

$$\mathbf{E}_x = f_\theta^{\text{img}}(\mathbf{x}) \tag{6}$$

$$\mathbf{E}_{p_i} = f_\theta^{\text{text}}(\mathbf{p}_i + cls) \tag{7}$$

where cls denotes the ground-truth label of the sample $\mathbf{x}$. For member samples $\mathbf{x} \in \mathcal{D}_i^{\text{in}}$, we define positive pairs $(\mathbf{E}p_i, \mathbf{E}x^{in})$, and for non-member samples $\mathbf{x} \in \mathcal{D}_i^{\text{out}}$, we define negative pairs $(\mathbf{E}p_i, \mathbf{E}x^{out})$.

Step3: Attack Model Training In the third step, we train the attack model consisting of a feature projector and a membership discriminator using the positive and negative sample pairs prepared in the second step. The contrastive learning-based feature projector is designed to narrow the distributional characteristic differences between prompts and member samples while expanding the semantic distinctions between prompts and non-member samples. The membership discriminator, a binary classifier, is responsible for making the final membership decision.

We input the positive and negative sample pairs into the feature projector, and minimize the distance between prompts and member samples in the feature space while maximizing the distance between prompts and non-member samples by utilizing the contrastive learning loss function $\mathcal{L}_{\text{C}}$:

$$S_{\text{pos}} = \exp\left(\frac{\text{sim}(\mathbf{E}_{p_i}, \mathbf{E}_{x^{in}})}{\tau}\right) \tag{8}$$

$$\mathcal{L}_{\text{C}} = -\log \frac{S_{\text{pos}}}{S_{\text{pos}} + \sum\limits_{x^{out}} \exp\left(\frac{\text{sim}(\mathbf{E}_{p_i}, \mathbf{E}_{x^{out}})}{\tau}\right)} \tag{9}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, τ is the temperature parameter, and S_{pos} represents the similarity score for a positive sample pair. Here, x^{in} is a member sample, and x^{out} is a non-member sample.

This enables the attack model to predict membership status based on the re-encoded feature vectors.

The output vectors of the feature projector are used as inputs to the membership discriminator, which is trained based on binary cross-entropy:

$$\mathcal{L}_{\text{D}} = -y \log \hat{y} - (1-y)\log(1-\hat{y}) \tag{10}$$

where $y \in \{0,1\}$ indicates membership status and $\hat{y} = \sigma(A(\mathbf{E}_{\mathbf{p_i}}, \mathbf{E}_x))$ is the prediction from the attack model $A(\cdot)$ with sigmoid activation $\sigma(\cdot)$.

Stage2: Membership Inference

Our objective is to investigate whether the prompt itself can lead to privacy leakage of its training data. To this end, membership inference should not rely on the output of the target model, but instead infer the membership status of a given sample solely based on the target prompt.

When it comes to the final inference stage, the adversary first encodes the target prompt $\mathbf{p}^*$, the given image $\mathbf{x}^*$, and its ground-truth label y^* using a pre-selected pre-trained model $f_\theta(\cdot)$. And the final prediction is obtained via:

$$E_{P^*} = f_\theta^{\text{text}}(\mathbf{P}^* + y^*) \tag{11}$$

$$\hat{z}^* = A(E_{P^*}, f_\theta^{\text{img}}(\mathbf{x}^*)) \tag{12}$$

If $\hat{z}^* > \gamma$, the sample is inferred as a member sample.

It is important to note that the ground-truth label y^* here refers to the true annotation of the input image, rather than the predicted output of the target model. This distinction reflects practical adversarial settings, where access to model outputs is typically restricted and query budgets are limited. By utilizing ground-truth labels instead of predicted outputs, PIPRA more accurately captures the semantic distribution of the samples, making it better suited for real-world scenarios with limited output access.

Experiments

In this section, we present experiments to evaluate the effectiveness of PIPRA. Specifically, we aim to address the following questions:

- **RQ1:** How effective is PIPRA as a membership inference attack scheme?
- **RQ2:** How do the selection of shadow datasets and pre-trained models affect membership inference?
- **RQ3:** Does the contrastive learning–based feature projector really work for MIA?

Experiment Setup

Datasets We evaluate our framework on nine datasets include Caltech101, CIFAR100, ImageNet, OxfordPets, StanfordCars, Flowers102, Food101, SUN397 and EuroSAT. These datasets are widely used in VLM research and are capable of effectively evaluating the performance of PIPRA in various tasks and data distributions.

Baselines We evaluate the most recent and representative baselines, including three black-box MIAs and one white-box MIA, to comprehensively assess the effectiveness of PIPRA. **Confidence Score-based** (Bertran et al. 2023), this black-box attack leverages the target model's confidence scores and uses quantile regression to model the distribution of non-member predictions and then infers membership by detecting deviations from this distribution. **Threshold-based** (Song and Mittal 2021), a widely adopted MIA baseline distinguishes members from non-members by comparing their loss values against a fixed threshold, under the assumption that member samples tend to incur lower losses due to memorization during training. **Adversarial Robustness-based** (Del Grosso et al. 2022), this method assesses a sample's robustness to adversarial perturbations, based on the observation that member samples often exhibit greater robustness due to overfitting to the training data. **Gradient-based** (Nasr, Shokri, and Houmansadr 2019), this white-box method infers membership by analyzing the gradients of the loss with respect to model parameters, based on the insight that member samples often yield smaller or more stable gradients due to their alignment with the learned decision boundary.

Evaluation Metrics Following previous works, we evaluate the performance of PIPRA and the four baselines with three metrics: **Accuracy** (Shokri et al. 2017), The percentage of correctly predicted member and nonmember samples. In the context of MIA, this metric reflects the overall effectiveness of the attack in distinguishing training data (members) from non-training data (nonmembers). **AUC** (Salem et al. 2018), The Area Under the Receiver Operating Characteristic Curve (AUC) is a widely used metric that provides an overall assessment of MIA performance. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various decision thresholds. **TPR@1%FPR** and **TPR@0.1%FPR** (Carlini et al. 2022), The true positive rates when the false positive rate is constrained to 1% and 0.1%, respectively. These metrics emphasize precision in identifying member samples and reflect the adversary's ability to confidently infer membership while minimizing false alarms.

Settings Our experiments are conducted using 2 NVIDIA A6000 GPUs. We follow the default training settings to train all prompts for 200 epochs for 4 shots, and the number of context tokens M is set to 16. We conduct a grid search to determine the discrimination threshold γ for each attack, and all attack models are trained for 100 epochs.

PIPRA's Effectiveness of Membership Inference (RQ1)

PIPRA achieves superior performance in membership inference. We report the AUC, TPR@1%FPR and TPR@0.1%FPR results of PIPRA across nine datasets. In terms of AUC values, as shown in Table 1, we find that PIPRA outperforms all baselines. When the adversary is completely barred from any model outputs, PIPRA still achieves an average AUC of 87.6%, significantly outperforming the strongest baseline Gradient-based attack with an average AUC of 79.7%. On general datasets such as Caltech101 and CIFAR100, the AUC of PIPRA exceeds 92%, while even on the harder case of EuroSAT, it maintains an AUC above 79%.

PIPRA shows stable and reliable performance in membership inference. Beyond achieving strong overall accuracy, it exhibits minimal performance variance across repeated runs under identical experimental settings. Specifically, the standard deviations consistently remain below 4.4%, highlighting the robustness of our method across domains with diverse visual and semantic characteristics.

Importantly, PIPRA operates under a highly restricted threat model without requiring access to the target model's outputs or internal states, yet still achieves state-of-the-art inference. This robustness is attributed to our contrastive inference strategy in semantic space, which captures fine-grained semantic proximity between samples and prompt. These results underscore the robustness of PIPRA in realistic black-box scenarios and further validate that **prompt vectors themselves can leak private information pertaining to their training data**.

Method	Caltech101	CIFAR100	ImageNet	OxfordPets	SUN397	StanfordCars	Food101	Flowers102	EuroSAT
ConfScore-based	69.42%	74.43%	67.89%	73.07%	77.66%	76.21%	71.34%	81.89%	72.97%
Threshold-based	71.27%	77.52%	<u>78.53%</u>	<u>82.94%</u>	<u>82.54%</u>	<u>83.52%</u>	74.64%	79.30%	<u>76.85%</u>
AdvRob-based	74.61%	72.53%	73.34%	77.35%	80.34%	83.28%	72.20%	74.64%	76.47%
Gradient-based	<u>78.54%</u>	<u>82.42%</u>	76.21%	82.65%	82.33%	81.36%	<u>75.24%</u>	<u>82.61%</u>	75.63%
PIPRA	**92.37%**	**92.23%**	**89.58%**	**84.66%**	**90.94%**	**90.81%**	**82.44%**	**86.00%**	**79.21%**

Table 1: AUC for different membership inference methods across various datasets. **Bold** and <u>underline</u> indicate the best and the runner-up for each dataset, respectively.

	Caltech101		CIFAR100		ImageNet		OxfordPets		StanfordCars		Flowers102		Food101		SUN397		EuroSAT	
Method	1%	0.1%	1%	0.1%	1%	0.1%	1%	0.1%	1%	0.1%	1%	0.1%	1%	0.1%	1%	0.1%	1%	0.1%
ConfScore-based	11.3%	2.7%	14.2%	2.4%	9.9%	1.9%	13.0%	2.9%	11.2%	1.9%	11.9%	3.1%	9.4%	1.2%	8.9%	2.1%	7.3%	1.7%
Threshold-based	19.1%	<u>6.7%</u>	19.7%	<u>7.3%</u>	18.2%	5.7%	<u>17.9%</u>	6.2%	<u>15.3%</u>	4.6%	**18.4%**	4.2%	12.2%	3.4%	15.1%	4.4%	13.0%	<u>3.0%</u>
AdvRob-based	16.1%	4.9%	17.3%	5.3%	18.3%	4.4%	14.5%	3.2%	13.8%	4.8%	15.4%	3.8%	13.7%	3.3%	<u>16.6%</u>	3.5%	<u>13.4%</u>	2.7%
Gradient-based	<u>21.5%</u>	5.3%	<u>22.2%</u>	4.6%	<u>19.1%</u>	<u>8.5%</u>	17.3%	<u>7.6%</u>	14.7%	**6.9%**	17.5%	<u>5.0%</u>	<u>14.2%</u>	<u>3.8%</u>	15.1%	<u>4.7%</u>	9.3%	2.9%
PIPRA	**26.7%**	**9.7%**	**24.1%**	**8.7%**	**31.6%**	**11.3%**	**21.8%**	**10.9%**	**16.3%**	<u>6.2%</u>	<u>18.2%</u>	**7.2%**	**14.3%**	**8.2%**	**21.9%**	**8.4%**	**16.1%**	**5.3%**

Table 2: TPR@1%FPR and TPR@0.1%FPR for different membership inference methods across various datasets. **Bold** and <u>underline</u> respectively denote the best and the runner-up.

PIPRA's strong capability under extremely low FPR. In many practical scenarios, achieving a high true positive rate (TPR) at an extremely low false positive rate (FPR) is essential for accurately identifying members while minimizing costly false positives. As shown in Table 2, PIPRA demonstrates excellent performance at very low FPRs (0.1%), significantly improving the TPR with relative gains exceeding 30% compared to the best existing baselines. Even at more relaxed FPR thresholds (1%), PIPRA continues to maintain substantial advantages, showcasing its strong membership distinction ability.

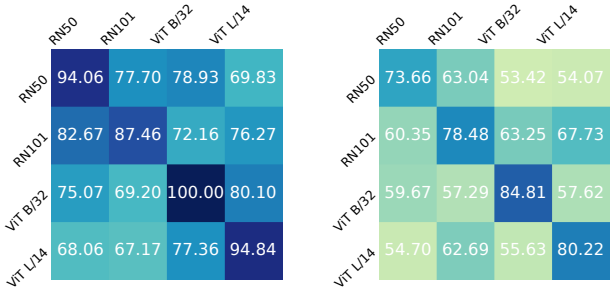
These results demonstrate that suppressing conventional model output predictions—such as logits, loss values, or gradients—is **insufficient** to prevent MIAs. Prompt vectors themselves can leak latent training membership signals, particularly when optimized on small and informative datasets. Notably, PIPRA achieves state-of-the-art performance under an output-free threat model: with only a intercepted prompt, a moderate shadow dataset, and a public pre-trained VLM, it consistently outperforms both black-box and white-box baselines across all datasets. Our findings underscore the urgent need for **stronger, representation-level defenses**, and highlight a critical challenge for the secure deployment of prompt tuning paradigms.

Robustness of PIPRA (RQ2)

PIPRA's robustness to the selection of pre-trained model In realistic black-box settings where the adversary has no access to the target system's pre-trained model, they resort to using surrogate models for MIA. To evaluate the robustness and generalizability of PIPRA to the choice of pre-trained models, we compare its performance across four distinct pretrained VLMs: **ResNet-50** (He et al. 2016), **ResNet-101** (He et al. 2016), **ViT-B/32** (Dosovitskiy et al. 2021), and **ViT-L/14** (Dosovitskiy et al. 2021). Specifically, we assess the attack performance when PIPRA and the target system are instantiated with different pre-trained models, quantifying the impact of architectural mismatches between the surrogate model used to train shadow prompts and the actual target model. We also benchmark PIPRA against three representative black-box methods that similarly rely on surrogate models, analyzing their sensitivity to surrogate model selection.

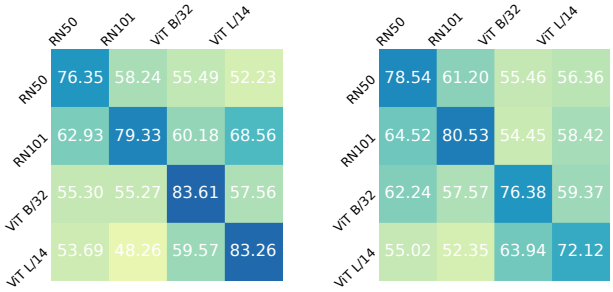
Figure 3 reports the attack accuracy on Caltech101 using four pre-trained VLMs. The results demonstrate that PIPRA consistently outperforms all baselines across diverse VLM architectures, showcasing superior generalization and practical applicability. While stronger model alignment yields better performance, even architecturally dissimilar VLMs retain sufficient representational structure to support effective membership inference. The relative degradation across pre-trained VLMs remains within approximately 30%, underscoring that PIPRA relies on the degree of semantic alignment between the image and text encoders within a VLM, rather than on the architectural similarity between the surrogate VLM and the target VLM. This fundamentally distinguishes it from prior MIAs, which depend on surrogate model responses closely resembling the target.

PIPRA's robustness to the selection of shadow dataset We assess how varying degrees of similarity between shadow and target datasets impact attack performance. As shown in Figure 4, the accuracy drop is minimal. This resilience is because PIPRA focuses on semantic distributional differences, not target model responses. We find that a shadow dataset with a similar domain and visual style is sufficient, and the data volume required for MIA is significantly



(a) PIPRA

(b) ConfScore-based



(c) Threshold-based

(d) AdvRob-based

Figure 3: The attack accuracy of our method and three baselines with different pre-trained models for the target system and adversary. The x-axis denotes the target model, the y-axis the surrogate model.

reduced in few-shot scenarios.

The robustness of PIPRAunder domain and architecture mismatch stems from the inherent semantic alignment capabilities of pre-trained VLMs, whose cross-modal embeddings preserve semantic proximity even under moderate shifts due to their training on large and diverse datasets. Since PIPRArelies on the relative distributional patterns between prompts and samples rather than absolute features, it remains effective across heterogeneous VLMs, which can still encode consistent geometric relationships—an effect theoretically grounded in domain adaptation and invariant representation learning (Du et al. 2024).

Ablation Study (RQ3)

We perform an ablation study comparing the baseline attack model trained with standard binary cross-entropy loss with a variant augmented by a contrastive learning objective.

Table 3 presents the accuracy and AUC across nine benchmark datasets. The average improvement across the nine datasets is approximately 6.7% in accuracy and 7.3% in AUC after incorporating contrastive learning. This confirms that contrastive learning significantly enhances the attack model's ability to distinguish member and non-member samples in representation space. By aligning member pairs and distancing non-member samples within the shared semantic space, the attack model captures subtle representational cues overlooked by purely supervised methods. This underscores the escalating threat posed by representation-aware adversaries in vision-language tuning.

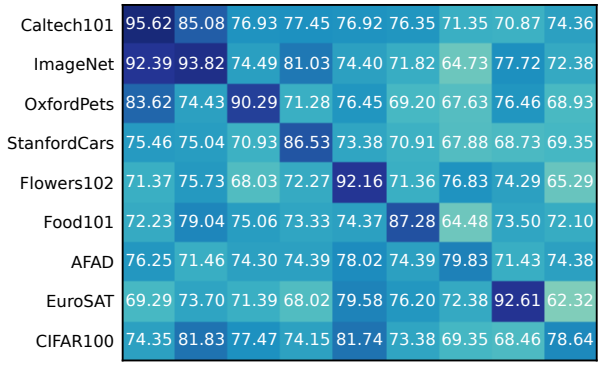


Figure 4: The attack accuracy when the adversary selects different shadow datasets. The y-axis denotes the training datasets of the target system, while the x-axis denotes the attacker's shadow datasets, ordered identically to the y-axis from top to bottom.

	Baseline		+ Contrastive Learning	
Dataset	**Acc (%)**	**AUC (%)**	**Acc (%)**	**AUC (%)**
Caltech101	83.5	88.0	90.4 (**+6.9**)	92.4 (**+4.4**)
CIFAR100	78.3	84.8	87.9 (**+9.6**)	92.2 (**+7.4**)
ImageNet	80.2	85.1	92.7 (**+12.5**)	89.6 (**+4.5**)
OxfordPets	77.1	79.3	82.9 (**+5.8**)	84.7 (**+5.4**)
StanfordCars	80.3	82.3	85.2 (**+4.9**)	90.8 (**+8.5**)
Flowers102	75.8	78.4	76.2 (**+0.4**)	84.7 (**+6.3**)
Food101	72.5	74.6	79.9 (**+7.4**)	86.0 (**+11.4**)
SUN397	76.3	83.2	81.4 (**+5.1**)	90.9 (**+7.7**)
EuroSAT	75.6	72.7	80.0 (**+4.4**)	79.2 (**+6.5**)

Table 3: AUC and accuracy comparison between the baseline discriminator and the discriminator with contrastive learning-based projector across nine datasets. Improvements (%) over the baseline are shown in **bold** within parentheses.

Conclusion

This paper introduces PIPRA, the first method for assessing privacy leakage risks from prompt vectors. Unlike previous approaches, PIPRA doesn't need outputs or internal parameters. It infers membership by analyzing the alignment of leaked prompt vectors and given samples in a shared semantic space. It reveals that prompt vectors can unintentionally expose sensitive training data, challenging existing defenses that solely suppress model outputs. Extensive experiments show PIPRA outperforms existing baselines, even in restricted scenarios, highlighting significant and previously overlooked privacy risks associated with prompt vectors. This is critical for real-world VLMs, especially in healthcare and finance. As prompt tuning grows, accidental data leakage via prompt vectors risks severe social and ethical consequences.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant Nos. 62572316 and 62302303, the Natural Science Foundation of Shanghai under Grant Nos. 25ZR1402279 and 23ZR1434000, and the Shanghai Pujiang Program under Grant No. 24PJD043.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Zisserman, A.; and Kavukcuoglu, K. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems (NeurIPS) 2022*, 23716–23736. New Orleans, LA: Curran Associates, Inc.

Bertran, M.; Tang, S.; Kearns, M.; Morgenstern, J.; Roth, A.; and Wu, Z. S. 2023. Scalable Membership Inference Attacks via Quantile Regression. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*. New Orleans, LA: NeurIPS.

Carlini, N.; Liu, C.; Úlfar Erlingsson; Kos, J.; and Song, D. 2019. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. In *Proceedings of the 28th USENIX Security Symposium (USENIX Security 19)*, 267–284. Santa Clara, CA: USENIX Association. ISBN 978-1-939133-06-9.

Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, U.; Oprea, A.; and Raffel, C. 2022. Membership Inference Attacks from First Principles. In *Proceedings of the 2022 IEEE Symposium on Security and Privacy (S&P)*, 1897–1914. San Francisco, CA: IEEE.

Chen, Z.; and Pattabiraman, K. 2024. Overconfidence is a Dangerous Thing: Mitigating Membership Inference Attacks by Enforcing Less Confident Prediction. In *Proceedings of the Network and Distributed System Security Symposium (NDSS) 2024*, 1–14. San Diego, CA: Internet Society.

Clusmann, J.; Ferber, D.; Wiest, I. C.; Schneider, C. V.; Brinker, T.; Foersch, S.; Truhn, D.; and Kather, J. N. 2025. Prompt Injection Attacks on Vision Language Models in Oncology. *Nature Communications*, 16: 1239.

Del Grosso, G.; Jalalzai, H.; Pichler, G.; Palamidessi, C.; and Piantanida, P. 2022. Leveraging Adversarial Examples to Quantify Membership Information Leakage. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10389–10399. New Orleans, LA: IEEE Computer Society.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*. Virtual Conference: OpenReview.net.

Du, Z.; Li, X.; Li, F.; Lu, K.; Zhu, L.; and Li, J. 2024. Domain-Agnostic Mutual Prompting for Unsupervised Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23375–23384. Seattle, WA: IEEE.

Duan, J.; Kong, F.; Wang, S.; Shi, X.; and Xu, K. 2023. Are Diffusion Models Vulnerable to Membership Inference Attacks? In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 8717–8730. Honolulu, HI: PMLR.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. Las Vegas, NV, USA: IEEE.

Hu, H.; Salčić, Z.; Dobbie, G.; Chen, J.; Sun, L.; and Zhang, X. 2022. Membership Inference via Backdooring. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*, 3832–3838. Vienna, Austria: IJCAI Organization.

Hu, Y.; Li, Z.; Liu, Z.; Zhang, Y.; Qin, Z.; Ren, K.; and Chen, C. 2025. Membership Inference Attacks Against Vision-Language Models. In *Proceedings of the 34th USENIX Security Symposium (USENIX Security '25)*. Seattle, WA, USA: USENIX Association.

Huang, Z.; Liu, Y.; He, D.; and Li, Y. 2025. DF-MIA: A Distribution-Free Membership Inference Attack on Fine-Tuned Large Language Models. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI)*, 343–351. Vancouver, Canada: AAAI Press.

Leemann, T.; Pawelczyk, M.; and Kasneci, G. 2023. Gaussian Membership Inference Privacy. In *Advances in Neural Information Processing Systems (NeurIPS 2023)*, 73866–73878. New Orleans, LA: Curran Associates, Inc.

Leino, K.; and Fredrikson, M. 2020. Stolen Memories: Leveraging Model Memorization for Calibrated White-Box Membership Inference. In *Proceedings of the 29th USENIX Security Symposium*, 1605–1622. Boston, MA: USENIX Association.

Li, Z.; Wang, Y.; Chen, X.; and Liu, Y. 2024a. DP-OPT: Make Large Language Model Your Privacy-Preserving Prompt Engineer. In *International Conference on Learning Representations (ICLR) 2024*. Kigali, Rwanda: OpenReview.net.

Li, Z.; Wu, Y.; Chen, Y.; Tonin, F.; Rocamora, E. A.; and Cevher, V. 2024b. Membership Inference Attacks against Large Vision-Language Models. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2024)*. Vancouver, Canada: NeurIPS.

Li, Z.; and Zhang, Y. 2021. Membership Leakage in Label-Only Exposures. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 880–895. Seoul, South Korea: ACM.

Liu, H.; Wu, Y.; Yu, Z.; and Zhang, N. 2024. Please Tell Me More: Privacy Impact of Explainability through the Lens of Membership Inference Attack. In *Proceedings of the 2024 IEEE Symposium on Security and Privacy (S&P)*, 4791–4809. San Francisco, CA: IEEE Computer Society.

Liu, Y.; Zhao, Z.; Backes, M.; and Zhang, Y. 2022a. Membership Inference Attacks by Exploiting Loss Trajectory. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2085–2098. Los Angeles, CA: ACM.

Liu, Z.; Zhang, X.; Chen, C.; Lin, S.; and Li, J. 2022b. Membership Inference Attacks Against Robust Graph Neural Network. In *Proceedings of the 2022 International Conference on Learning Representations (ICLR)*. Vienna, Austria (virtual): OpenReview.net.

Liu, Z.; Zhang, Z.; Xie, Y.; and She, D. 2025. CompressionAttack: Exploiting Prompt Compression as a New Attack Surface in LLM-Powered Agents. In *arXiv preprint arXiv:2510.22963*. https://arxiv.org/abs/2510.22963: arXiv.

Nasr, M.; Shokri, R.; and Houmansadr, A. 2019. Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-Box Membership Inference Attacks. In *Proceedings of the 2019 IEEE Symposium on Security and Privacy (S&P)*, 739–753. San Francisco, CA: IEEE Computer Society.

Rezaei, S.; and Liu, X. 2021. On the Difficulty of Membership Inference Attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7892–7900. Nashville, TN: IEEE/CVF.

Salem, A.; Zhang, Y.; Humbert, M.; Berrang, P.; Fritz, M.; and Backes, M. 2018. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses. In *Proceedings of the 2018 Network and Distributed System Security Symposium (NDSS)*, 267–284. San Diego, CA: The Internet Society.

Samson, L.; Barazani, N.; Ghebreab, S.; and Asano, Y. M. 2024. Little Data, Big Impact: Privacy-Aware Visual Language Models via Minimal Tuning. *arXiv preprint arXiv:2405.17423*.

Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership Inference Attacks Against Machine Learning Models. In *Proceedings of the 2017 IEEE Symposium on Security and Privacy (S&P)*, 3–18. San Jose, CA: IEEE Computer Society.

Song, L.; and Mittal, P. 2021. Systematic Evaluation of Privacy Risks of Machine Learning Models. In *Proceedings of the 30th USENIX Security Symposium (USENIX Security)*, 2615–2632. Vancouver, Canada: USENIX Association.

Sun, J.; Liu, C.; Guo, M.; Fu, Y.; Li, Z.; and Yin, D. 2023. Prompt Tuning based Adapter for Vision-Language Model Adaptation. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, 3922–3931. Ottawa, Canada: ACM.

Wang, Z.; Liang, J.; He, R.; Wang, Z.; and Tan, T. 2024. Collaborative Fine-tuning for Black-Box Vision-Language Models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*. Vienna, Austria: PMLR.

Wang, Z.; Panda, R.; Karlinsky, L.; Feris, R.; Sun, H.; and Kim, Y. 2023. Multitask Prompt Tuning Enables Parameter-Efficient Transfer Learning. In *Proceedings of the 2023 International Conference on Learning Representations (ICLR)*. Online: ICLR.

Wu, Y.; Wen, R.; Backes, M.; Berrang, P.; Humbert, M.; Shen, Y.; and Zhang, Y. 2024. Quantifying Privacy Risks of Prompts in Visual Prompt Learning. In *Proceedings of the 33rd USENIX Security Symposium (USENIX Security)*, 5841–5858. Boston, MA: USENIX Association.

Yu, L.; Chen, Q.; Lin, J.; and He, L. 2023. Black-box Prompt Tuning for Vision-Language Model as a Service. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, 1686–1694. Macau, China: International Joint Conferences on Artificial Intelligence Organization.

Zarifzadeh, S.; Liu, P.; and Shokri, R. 2024. Low-Cost High-Power Membership Inference Attacks. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*. Vienna, Austria: PMLR.

Zhang, J.; Ma, X.; Wang, X.; Qiu, L.; Wang, J.; Jiang, Y.; and Sang, J. 2024. Adversarial Prompt Tuning for Vision-Language Models. In *Proceedings of the European Conference on Computer Vision (ECCV) 2024*, 709–727. Turin, Italy: Springer.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional Prompt Learning for Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16816–16825. New Orleans, LA: IEEE.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision*, 130(9): 2337–2348.