This CVPR Findings paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version; the final published version of the proceedings is available on IEEE Xplore.

Revisiting Model Inversion Evaluation: From Misleading Standards to Reliable Privacy Assessment

Sy-Tuyen Ho[♣*] Koh Jun Hao[♢*] Ngoc-Bao Nguyen[♢*]
Alexander Binder[♡] Ngai-Man Cheung[♢]
[♣] University of Maryland College Park
[♢] Singapore University of Technology and Design (SUTD) [♡] Leipzig University
{sytuyen_ho,ngaiman_cheung}@sutd.edu.sg

Abstract

Model Inversion (MI) attacks aim to reconstruct information from private training data by exploiting access to a target model. Nearly all recent MI studies evaluate attack success using a standard framework that computes attack accuracy through a secondary evaluation model trained on the same private data and task design as the target model. In this paper, we present the ***first in-depth analysis*** *of this dominant evaluation framework and reveal a fundamental issue: many reconstructions deemed "successful" under the existing framework are in fact false positives that do not capture the visual identity of the target individual.*

As our first contribution, we show that ***these MI false positives satisfy the same formal conditions as Type I adversarial examples****. In addition, through controlled experiments, we demonstrate extremely high false-positive transferability, an empirical signature characteristic of adversarial behavior, indicating that many MI false positives likely contain Type I adversarial features. This adversarial transferability significantly inflates reported attack accuracy and leads to an overstatement of privacy leakage in existing MI work.*

To address this issue, as our second contribution, we introduce ***a new evaluation framework based on Multimodal Large Language Models (MLLMs)****, whose general-purpose visual reasoning avoids the shared-task vulnerability and reduces Type-I adversarial transferability of current evaluation framework. We propose systematic design principles for MLLM-based evaluation. Using this framework, we reassess 27 MI attack setups across diverse datasets, target models, and priors, and find consistently high false-positive rates under the conventional approach. Our results call for a reevaluation of progress in MI research and establish MLLM-based evaluation as a more reliable standard for assessing privacy risks in machine learning systems.*

To support future work, we release our MLLM-based MI evaluation framework and benchmarking suite.[1]

1. Introduction

Model Inversion (MI) attacks pose a significant privacy threat by attempting to reconstruct confidential information from sensitive training data through exploiting access to machine learning models. Recent state-of-the-art (SOTA) MI attacks [4, 21, 27, 28, 32, 33, 41, 46, 47] have shown considerable advancements, reporting attack success rates exceeding 90%. This vulnerability is particularly alarming for security-sensitive applications such as face recognition [12, 15, 23, 34], medical diagnosis [6, 7, 44], or speech recognition [3, 18]. The model under an MI attack is referred to as the *target model* T.

Research gap. Recently, there are many studies on improving MI attacks [1, 4, 8, 13, 16, 27, 28, 33, 46, 47] and MI defenses [14, 17, 31, 38, 42]. To assess the effectiveness of these MI attacks/defenses, MI Attack Accuracy (AttAcc) is a standard and the most important metric. To measure AttAcc, almost all recent MI studies adopt the evaluation framework introduced by [47], which we denote as $\mathcal{F}_{Curr}$. Under $\mathcal{F}_{Curr}$, an *evaluation model* E is used to predict the identities of individuals based on the MI-reconstructed images. Model E is trained on *the same private dataset and follows the same training task design* as in target model T. In particular, consider the scenario where the adversary targets a specific identity y. From this point onward, we use y exclusively to denote the target class (label) in the MI attack, rather than a label predicted by any model. The adversary produces reconstructed images x_y^r of the target label y by exploiting access to the target model T. To measure AttAcc, these x_y^r samples are then passed through an evaluation model E. According to $\mathcal{F}_{Curr}$, the attack is deemed successful if E classifies x_y^r as y. **Even though such MI evaluation framework becomes the de facto standard and**

[*] Equal Contribution

[1] https://github.com/hosytuyen/MI-Eval-MLLM

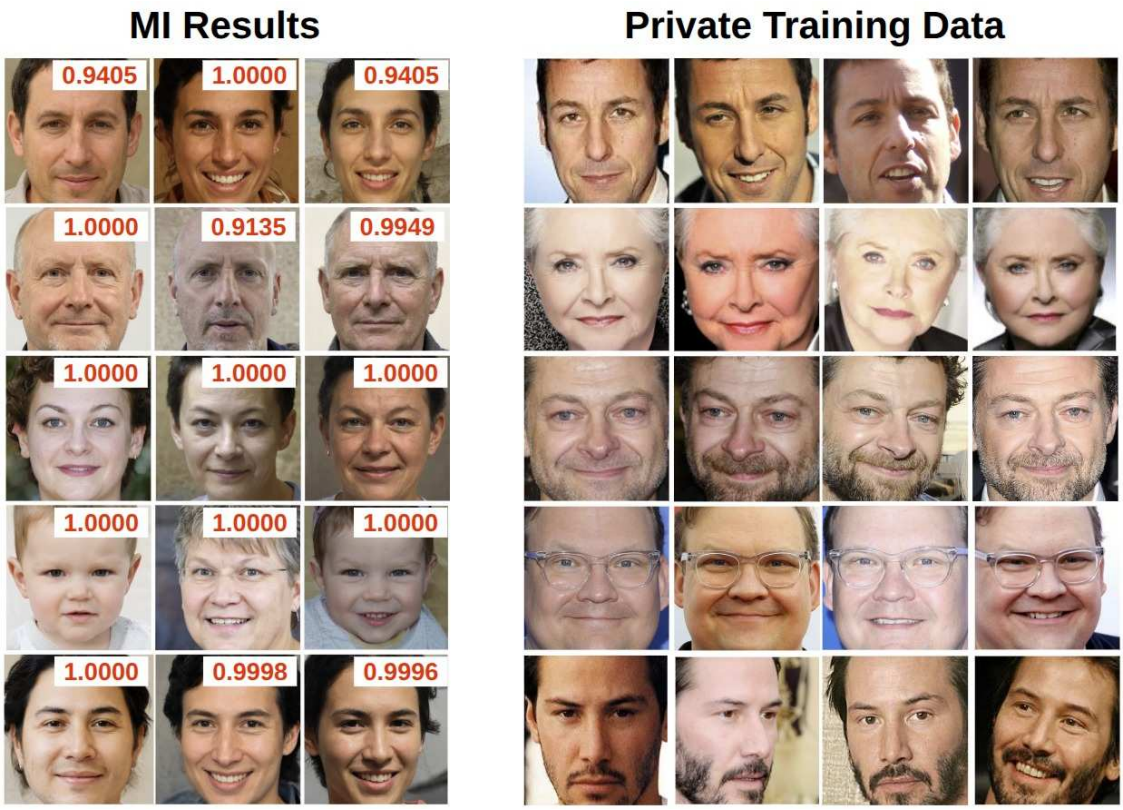


Figure 1. **We present the first and in-depth study on the Model Inversion (MI) evaluation**. Particularly, we investigate the most common MI evaluation framework $\mathcal{F}_{Curr}$ to measure MI Attack Accuracy (AttAcc). $\mathcal{F}_{Curr}$ is introduced in [47] and is utilized to assess almost all recent MI attacks/defenses. However, we find that $\mathcal{F}_{Curr}$ suffers from a significant number of false positives. These false positive MI reconstructed samples do not capture visual identity features of the target individual in the private training data, but they are still deemed successful attacks according to $\mathcal{F}_{Curr}$ with a high confidence (indicated in red text). Extensive false positives are in the Supp.

the accuracy measured by $\mathcal{F}_{Curr}$ has been the most important metric in gauging MI research progress in almost all recent MI studies, there has not been any in-depth and comprehensive study to understand the accuracy of $\mathcal{F}_{Curr}$ and its limitations.

In this work, we conduct the first in-depth study of the standard MI evaluation framework $\mathcal{F}_{Curr}$. For a truly successful attack, the reconstructed images x_y^r should capture the visual identity features of y. However, we find that there exists a considerable number of MI reconstructed images that lack visual identity features of y, yet both target model T and evaluation model E under $\mathcal{F}_{Curr}$ still assign high likelihood to y, i.e., high values of $P_E(y|x_y^r)$. Some examples are illustrated in Fig. 1. **These *false positives* potentially inflate the reported success rate of recent SOTA MI attacks.**

To shed light on the causes of these false positives, we systematically discover the impact of Type I Adversarial Features [26, 40] in Model Inversion, highlighting a connection between two previously distinct research areas. The optimization processes in MI attacks and Type I adversarial attacks are similar: both maximizing the likelihood with respect to (w.r.t.) the input under a fixed model. We show that false positives in MI and Type I adversarial examples are mathematically equivalent: **MI false positives and Type I adversarial examples are essentially the same construct mathematically, only arising under different problem contexts, MI versus adversarial attacks.** Moreover, due to the well-documented phenomenon of *adversarial transferability* [26], these false positives can transfer to the MI evaluation model E. This transferability is especially pronounced when the evaluation model E in $\mathcal{F}_{Curr}$ shares the same task design as the target model T [22, 29]. Ultimately, we demonstrate a fundamental issue in $\mathcal{F}_{Curr}$, potentially leading to unreliable assessment and inflated success rates reported for recent MI attacks under this evaluation framework.

To mitigate this issue, we propose a new MI evaluation framework $\mathcal{F}_{MLLM}$ based on Multimodal Large Language Models (MLLMs). We propose systematic design principles for $\mathcal{F}_{MLLM}$. Leveraging the general-purpose visual reasoning of MLLMs and avoiding the reliance on the same training task design used in T, $\mathcal{F}_{MLLM}$ minimizes Type-I transferability and offers a more faithful assessment of MI attacks. Using data annotated by $\mathcal{F}_{MLLM}$, we empirically benchmark the reliability of the common MI evaluation framework $\mathcal{F}_{Curr}$ and reveal consistently high false positive

rates across 27 diverse MI setups. Our findings challenge the reliability of the standard MI evaluation framework and underscore the importance of adopting our MLLM-based approach for more reliable assessments. Our main contributions are summarized below:

- We present the first in-depth study on the most common evaluation framework $\mathcal{F}_{Curr}$ to compute MI AttAcc. Our study identify the surprising effect of Type I Adversarial Features, as well as adversarial transferability [26, 40], on existing MI attack accuracy evaluation, highlighting a relationship between two previously distinct research areas (see Sec. 3). Our findings explain numerous false positives in $\mathcal{F}_{Curr}$.
- To mitigate this issue, we propose an MLLM-based MI evaluation framework $\mathcal{F}_{MLLM}$ (see Sec. 4.1). Our framework leverages the powerful general-purpose visual understanding capabilities of advanced MLLMs to evaluate MI reconstructions, replacing the traditional evaluation model E. We propose systematic design principles for $\mathcal{F}_{MLLM}$. Without relying on the same training task design used in T, $\mathcal{F}_{MLLM}$ minimizes Type-I transferability and provides a more faithful assessment of MI attacks.
- We systematically and comprehensively demonstrate that there are considerable number of false positives under $\mathcal{F}_{Curr}$. Ultimately, our findings challenge the validity of this dominant evaluation framework and underscore the importance of adopting our proposed MI evaluation framework for more reliable assessments (see Sec. 4.2).

2. Related Work

Model Inversion Attacks. Model Inversion (MI) attacks aim to extract information about private training data from a trained model. Given a target model T trained on a private dataset $\mathcal{D}priv$, an adversary seeks to infer properties or reconstruct samples from $\mathcal{D}priv$ without direct access to it. MI attacks are commonly formulated as reconstructing an input x_y^r that T is likely to classify as a label y. For example, in facial recognition settings, MI attacks correspond to reconstruct images that are likely to be identified as belonging to a particular person. Early MI work [8, 9] showed that trained models can leak sensitive genomic, demographic, or facial information. Subsequent approaches such as adversarial MI [45] treat T as an encoder and train an auxiliary network to reconstruct inputs from model outputs. Recent generative MI methods instead reduce the search to a latent space by training a deep generator [4, 27, 33, 37, 41, 45–47]. Methods like GMI [47] and PPA [37] use pretrained GANs (e.g., WGAN, StyleGAN) on auxiliary datasets resembling $\mathcal{D}_{priv}$ and optimize over latent vectors. Recent improvements refine MI attacks from multiple angles such as prior knowledge, where KEDMI [4] trains inversion-specific GANs and Pseudo-Label MI [46] uses pseudo-labels; latent feature guidance, as in IF-GMI [33]; objective design, introducing max-margin [46] and logit losses [27] beyond Cross-Entropy; and overfitting mitigation, where LOMMA [27] uses augmented models to improve generalization. The Eq. 1 represents the general step of SOTA MI attacks:

$$w^* = \arg\min_w(-\log P_T(y|G(w)) + \lambda\mathcal{L}_{prior}(w)) \quad (1)$$

Here, $-\log \mathcal{P}_T(y|G(w))$ represents the identity loss, guiding the reconstruction of $x_y^r = G(w)$ that is most likely to be classified as target class y by target classifier T. The $\mathcal{L}_{prior}$ is a prior loss, using public information to establish a distributional prior via GANs, thereby guiding the inversion process towards meaningful reconstructions.

Model Inversion Defenses. In contrast to MI attacks, MI defenses aim to minimize the leakage of private training samples during the MI optimization process. The goal is to train the target classifier T on $\mathcal{D}_{priv}$ such that it minimizes information disclosure about specific labels while maintaining good performance. Several defenses have been proposed. MID [42] adds a regularizer that penalizes the mutual information between inputs x and outputs $T(x)$. BiDO [31] introduces a bilateral regularizer that suppresses input information in feature representations z while preserving label information. Beyond regularization-based approaches, TL-DMI [14] improves robustness via transfer learning, and LS [38] applies Negative Label Smoothing. More recently, architectural factors influencing MI vulnerability have also been studied [17].

Model Inversion Evaluation Metrics. To evaluate MI attacks and defenses, almost all existing studies rely on the standard MI evaluation framework from [47], denoted as $\mathcal{F}_{Curr}$, which computes attack accuracy and serves as the main metric for monitoring progress in MI research. Suppose an adversary targets a class y and reconstructs images x_y^r using access to the target model T. These x_y^r are then classified by an evaluation model E (trained on the same $\mathcal{D}_{priv}$ but distinct from T). Under $\mathcal{F}_{Curr}$, an attack is successful if E predicts x_y^r as y. Several complementary metrics have been introduced. KNN distance [4, 27, 37, 46, 47] measures feature proximity between reconstructions and private samples. FID [31, 33, 37, 46] assesses perceptual realism. Knowledge Extraction Score [38] evaluates discriminative content by training a surrogate classifier on reconstructed images. While these provide additional insights, $\mathcal{F}_{Curr}$ and attack accuracy remain the dominant evaluation tools. Although these complementary metrics provide additional perspectives, $\mathcal{F}_{Curr}$ remains the dominant evaluation framework and attack accuracy the most widely used measure of progress in MI research. However, despite its prevalence, there has not yet been a comprehensive study examining the reliability and limitations of $\mathcal{F}_{Curr}$. *In this work, we take the first step toward such an investigation.*

3. Revealing a Hidden Connection: Issues in MI Evaluation and Type I Adversarial Attacks

We discover, for the first time, the strong connection between Type I Adversarial Attacks and MI Attacks. Due to this strong connection, the adversarial type I examples can be generated during MI attacks. Additionally, due to the well-documented phenomenon of adversarial transferability [26], these adversarial type I examples can transfer to E in $\mathcal{F}_{Curr}$. Ultimately, these phenomenons results in unreliable assessment for $\mathcal{F}_{Curr}$.

3.1. An Overview of Adversarial Attacks

An adversarial attack on machine learning models is an intentional manipulation of input data to cause incorrect predictions, highlighting potential vulnerabilities of the model. Adversarial attacks aim to create inputs that deceive machine learning classifiers into making errors while humans do not. There are two main types of adversarial attacks: Type I and Type II. Type II Adversarial Attacks [2, 10, 19, 24, 30, 35, 39] are commonly studied and aim to produce false negatives. In this attack, minor and imperceptible perturbations are added to the input data x to generate an adversarial example $x^{\text{adv-II}}$, which is incorrectly classified by the model. Mathematically, this is represented as:

$$f(x^{\text{adv-II}}) \neq f(x) \text{ and } f_{oracle}(x^{\text{adv-II}}) = f_{oracle}(x) \quad (2)$$

Here, f is the model under attack, f_{oracle} is an oracle or hypothetical, idealized classifier. In addition to Type II attack, Type I Adversarial Attacks are designed in [26, 40] to generate false positives by creating examples that are significantly different from the original input but are still classified as the same class by the model. This involves producing an adversarial example $x^{\text{adv-I}}$ that, despite being significantly different from input x, the target model f mis-classifies as the original class. **Formally, Type I adversarial attack sample is defined as follows** [26, 40]:

$$f(x^{\text{adv-I}}) = f(x) \text{ and } f_{oracle}(x^{\text{adv-I}}) \neq f_{oracle}(x) \quad (3)$$

Type I adversarial samples can be produced by optimizing the input by iteratively updating it to maximize the likelihood under a fixed target model classifier [26]. This process optimizes the input to remain within a targeted decision boundary, however, it is different from training data by human perception. This phenomenon can also be viewed as over-confidence phenomenon of machine learning models [11, 43].

3.2. The Strong Connection Between Model Inversion Attacks and Type I Adversarial Attacks

In the following, we analyze the connection between MI attacks and Type I adversarial attacks. The general inversion step used in SOTA MI attacks is described in Eq. 1. As a result, reconstructed images often exhibit high likelihood under the target classifier T. However, not all reconstructed images successfully capture the visual identity features of the target individual from the private training data. Some examples are illustrated in Fig. 1. We refer to these cases as **false positives (under T) in MI**. Specifically, with target classifier T and f_{oracle} denoting the oracle, **false positive in MI attack is mathematically represented as**:

$$\begin{aligned} T(x_y^r) = T(x_y) = y \\ \text{and } f_{oracle}(x_y^r) \neq f_{oracle}(x_y) \end{aligned} \quad (4)$$

Here, x_y^r does not resemble the visual identity feature of x_y. However, T classifies x_y^r as the target label y. Hence, x_y^r is a false positive in MI. In Eq. 4, an alternative interpretation is that the false positive x_y^r is the example that can deceive T to classify it as the target label y while oracle classifier can not recognize x_y^r as y.

Table 1. Mathematical equivalence of False Positive (FP) in MI (Eq. 4) and Type I adversarial attack (Eq. 3) [26, 40].

	False Positive in MI	**Type I Adversarial Attack**
Fixed Model Under Attack	T	f
Private/Original Sample	x_y	x
Attack Sample	x_y^r	$x^{\text{adv-I}}$
Formulation	$T(x_y^r) = T(x_y)$	$f(x^{\text{adv-I}}) = f(x)$
	$f_{oracle}(x_y^r) \neq f_{oracle}(x_y)$	$f_{oracle}(x^{\text{adv-I}}) \neq f_{oracle}(x)$

Critically, by comparing Eq. 4 and Eq. 3, we reveal the mathematical equivalence of MI false positives and Type I adversarial examples: both describe attack samples (x_y^r, $x^{\text{adv-I}}$ resp.) optimized under a fixed model (T, f resp.) that preserve the model's prediction while deviating from human-perceived identity. This equivalence uncovers a key insight: MI false positives and Type I adversarial examples are essentially the same construct mathematically, only arising under different problem contexts, MI versus adversarial attacks. Tab. 1 summarizes the equivalence.

3.3. Adversarial Transferability Can Lead to Critical Issues in the Common Model Inversion Evaluation Framework

In $\mathcal{F}_{\text{Curr}}$, an evaluation model E is used to predict the identities of individuals based on MI-reconstructed images. The model E is trained on the same private dataset and follows the same training task design (i.e., an n-way classification task). Prior work has shown that adversarial examples crafted for one model can often transfer and mislead other models, even those with different architectures, as

Table 2. MI-Generated Negatives produce abnormally high false-positive rates across models, indicative of Type I Adversarial Features. Additional results are in Supp.

Attack	$\mathcal{D}_{priv}$	E	Input	FP under E
PPA	FaceScrub	InceptionV3	Neg x_y^r	89.04%
			Neg $x_y^{natural}$	0.94%
		MaxViT	Neg x_y^r	73.95%
			Neg $x_y^{natural}$	0.22%
PLGMI	CelebA	FaceNet112	Neg x_y^r	97.55%
			Neg $x_y^{natural}$	0.00%

long as they are trained on the similar dataset and task design [22, 26, 29]. Therefore, the similarity between E and the target model T may enable the transferability of Type I adversarial examples from T to E. This phenomenon can be expressed mathematically as:

$$T(x_y^r) = T(x_y) = y \qquad E(x_y^r) = E(x_y) = y$$
$$\text{and } f_{oracle}(x_y^r) \neq f_{oracle}(x_y) \tag{5}$$

3.4. MI-Generated Samples Exhibit Abnormally High Transferability Characteristic of Adversarial Features

After establishing that MI false positives and Type I adversarial examples are mathematically equivalent, we further design a study to support the claim that *many MI false positives contain Type I adversarial features*. We do so by analyzing the transferability of MI-generated samples and examining whether they exhibit the defining empirical behavior of adversarial examples: abnormally high transferability across models [22, 26, 29].

Specifically, we evaluate *MI-generated negative samples* that are generated under MI on the target model T for the target label y, but fail under the oracle. We measure how often they are also misclassified as the target identity y by a separate evaluation model E. By comparing this transferability against that of *natural negative samples*, we assess whether MI-generated samples display adversarial-like behavior. A significantly higher false-positive rate for MI-generated negatives would indicate that MI-generated samples exhibit abnormally high transferability characteristic of adversarial features.

Setups. We examine two representative MI scenarios: (1) Setup 1: target model T = ResNet18, private dataset $\mathcal{D}_{priv}$ = FaceScrub, attack = PPA, and evaluation models E = InceptionNetV3 and MaxViT; (2) Setup 2: target model T = VGG16, private dataset $\mathcal{D}_{priv}$ = CelebA, attack = PLGMI, and evaluation model E = FaceNet112.

MI-generated negative samples. We first perform MI attacks on T to obtain reconstructed samples x_y^r. We then identify all *MI-generated negative samples*, denoted as Neg x_y^r, using oracle-based annotation: $f_{\text{oracle}}(\text{Neg } x_y^r) \neq f_{\text{oracle}}(x_y)$. Let $n = |\text{Neg } x_y^r|$ denote the number of such samples.

Natural negative samples. To construct a controlled baseline, we create a dataset of *natural negative samples*, Neg $x_y^{natural}$. We randomly select n FFHQ images (with no identity overlap with FaceScrub) and assign each image a randomly chosen FaceScrub identity, forming n negative samples satisfying $f_{\text{oracle}}(\text{Neg } x_y^{natural}) \neq f_{\text{oracle}}(x_y)$. Because these images are randomly-sampled natural images and not optimized with respect to the target classifier, they are free from Type I adversarial characteristics and serve as a control dataset.

Evaluation. We pass both Neg x_y^r and Neg $x_y^{natural}$ through the evaluation model E and count how many are falsely predicted as the target label y, resulting in false-positive rates for each group.

Results. The false-positive rates are reported in Table 2. Across all setups, MI-generated negatives Neg x_y^r exhibit an abnormally higher false-positive rate compared to natural negatives Neg $x_y^{natural}$. This large discrepancy indicates that MI-generated samples transfer across models in a manner characteristic of adversarial examples, supporting the conclusion that many MI false positives contain Type I adversarial features.

> ***Finding 1:*** Because MI false positives satisfy exactly the same formal conditions as Type-I adversarial examples, and because our controlled experiments reveal abnormally high cross-model transferability, a defining empirical signature of adversarial behavior, we conclude that many MI false positives indeed contain Type-I adversarial features. MI false positives with Type-I adversarial features ultimately transfer to evaluation models, especially when both evaluation and target models share the same task design.

4. Our Proposed MLLM-Based Model Inversion Evaluation Framework

In this section, we introduce a novel and faithful evaluation framework, $\mathcal{F}_{\text{MLLM}}$. To mitigate the undesirable effects of Type I adversarial transferability under T, our key idea is to employ a model trained under a fundamentally different learning paradigm. We argue that MLLMs are ideal candidates because they are trained on broad, general-purpose tasks using data and optimization pipelines entirely distinct from those of the target model in MI. In fact, MLLMs are increasingly used for automated data labeling, offering robustness and scalability in both academia and industry [5, 20, 36, 48]. However, not every MLLM is suitable for



Figure 2. **An example of evaluation query in our $\mathcal{F}_{\text{MLLM}}$.** The task is to determine whether "Image A" depicts the same individual as those in "Image B". We have two setups: (1) "Image A" and "Image B" consist of private images and (2) "Image A" is an MI-reconstructed image x_y^r of the target label y while four real images of y are randomly selected as "Image B". MLLM is tasked with responding either "Yes" or "No" to indicate whether "Image A" matches the identity in "Image B".

Table 3. Following our selection principle, the experiment results indicate that Gemini-2.0 serves as a reliable MLLM for $\mathcal{F}_{\text{MLLM}}$ by demonstrating (1) Strong capability in understanding interleaved image-text inputs and (2) have no usage restrictions on MI tasks.

MLLM		**"Yes" rate ↑**	**"No" rate ↓**	**"Refuse" rate ↓**		**"Yes" rate ↓**	**"No" rate ↑**	**"Refuse" rate ↓**
Gemini 2.0	**Positive Pairs**	**93.84%**	**3.16%**	**0%**	**Negative Pairs**	**4.41%**	**95.59%**	**0%**
ChatGPT-5		17.50%	2.67%	79.83%		0.09%	23.04%	76.86%
Qwen2.5VL		88.51%	11.49%	0%		5.50%	94.55%	0%

Table 4. Detailed textual prompts in our MI evaluation framework $\mathcal{F}_{MLLM}$. They are carefully designed to guide the MLLM toward reliable assessment. For face recognition, the prompt explicitly instructs the model to handle variations (e.g., aging, lighting, hairstyle, accessories), and for dog breed recognition, the prompt instructs the model to account for viewpoint and lighting differences to reduce false positives and false negatives.

Domain	$\mathcal{D}_{priv}$	Textual Prompt
Face Recognition	Facescrub / CelebA	You are an expert in face recognition. Taking into account the face aging, lighting different hair styles, wearing and not wearing of eye glasses or other accessory, do the task in the image. Only answer yes or no
Dog Recognition	Stanford Dogs	You are an expert in dog breed recognition. Taking into account the lighting and viewpoint, do the task in the image. Only answer yes or no

$\mathcal{F}_{\text{MLLM}}$ when evaluating MI problems, even the most SOTA models. To systematically determine an ideal candidate, we propose a principle for designing and selecting MLLMs in Sec. 4.1. Then, we then re-evaluate MI attack accuracy using $\mathcal{F}_{\text{MLLM}}$ in Sec. 4.2.

4.1. Our MLLM-based evaluation framework

$\mathcal{F}_{\text{MLLM}}$ design. Our $\mathcal{F}_{\text{MLLM}}$ leverages a MLLM to evaluate whether a MI reconstructed image is a successful or unsuccessful attack. For each reconstructed sample, we construct an evaluation query (see Fig. 2 for an example) by pairing it with a set of private training images that include the target identity. We then combine this query image with a natural language textual prompt and provide both as input to the MLLM. The detailed prompts (shown in Tab. 4) are fixed across all evaluation queries to ensure fairness. For each reconstructed image, the model outputs a categorical response ("Yes" or "No"), where "Yes" indicates a successful attack. By evaluating many such queries and computing the proportion of correct identifications, $\mathcal{F}_{\text{MLLM}}$ provides an automated and faithful evaluation of MI.

We propose following criteria in selecting an MLLM for a faithful MI evaluation:

- **Strong capability in understanding interleaved image–text inputs.** MI evaluation requires the MLLM to jointly reason over visual content and task-specific prompts. Without robust visual-semantic grounding, the model may misinterpret identity cues and produce inflated false positives or false negatives due to prompt misinterpretation or hallucination.
- **No usage restrictions on MI tasks.** Many commercial MLLMs refuse queries involving human faces or identity comparison, resulting in incomplete or biased evaluation. To ensure full coverage and reliability, the chosen MLLM

Table 5. **Our investigation on MI evaluation framework using our comprehensive dataset of MI attack samples.** We study 27 standard MI setups covering SOTA MI studies (PPA [37], LOMMA [33], KEDMI [4], PLGMI [46], IFGMI [33], TL [14], TTS [17], RoLSS [17], LS [38]), spanning 9 target classifiers T, 3 private datasets $\mathcal{D}_{priv}$, and 4 public datasets $\mathcal{D}_{pub}$. Under $\mathcal{F}_{Curr}$, we find high false positive (FP) rates, indicating that prior work overestimates MI threats. For example, while SOTA attacks (e.g., IFGMI, LOMMA, PLGMI, PPA) claim over 90–100% AttAcc in some setups, actual privacy leakage remains below 80% across all setups, with some attacks falling under 60%. **Further details and results on other architectures, attack scenario are in the Supp.**

MI Attack	$\mathcal{D}_{pub}$	$\mathcal{D}_{pub}$	T	$\mathcal{F}_{MLLM}$	$\mathcal{F}_{Curr}$					
				AttAcc	E	**AttAcc**	**FP rate**	**FN rate**	**TP rate**	**TN rate**
PPA	FaceScrub	FFHQ	ResNet18	28.37%	InceptionNetV3	91.39%	90.09%	5.32%	94.58%	9.91%
			ResNet101	28.68%		84.69%	82.71%	10.36%	89.64%	17.29%
			ResNet152	30.26%		86.84%	85.09%	9.12%	90.88%	14.91%
			MobileNet-V2	47.18%		83.37%	80.39%	13.30%	86.70%	19.61%
			DenseNet121	27.43%		72.41%	70.13%	21.58%	78.42%	29.87%
			MaxViT	30.19%		79.48%	77.16%	15.16%	84.84%	22.84%
	Stanford Dogs	AFHQ	ResNest101	74.58%	InceptionNetV3	81.98%	61.07%	10.89%	89.11%	38.93%
IFGMI	FaceScrub	FFHQ	ResNet18	34.46%	InceptionNetV3	95.85%	94.60%	1.78%	98.22%	5.40%
		Metfaces	ResNet18	1.56%		72.50%	72.21%	9.09%	90.91%	27.79%
PLGMI	CelebA	CelebA	VGG16	73.73%	FaceNet112	98.73%	99.49%	1.54%	98.46%	0.51%
		FFHQ	VGG16	48.47%		88.67%	88.49%	11.14%	88.86%	11.51%
LOMMA	CelebA	CelebA	IR152	79.80%	FaceNet112	90.40%	86.80%	8.69%	91.31%	13.20%
			FaceNet64	78.73%		92.00%	93.73%	8.47%	91.53%	6.27%
		FFHQ	VGG16	79.93%		90.13%	90.70%	10.01%	89.99%	9.30%
			IR152	44.93%		77.73%	77.85%	22.40%	77.60%	30.27%
			FaceNet64	46.27%		72.13%	69.73%	25.07%	74.93%	22.15%
			VGG16	55.27%		63.07%	61.55%	35.71%	64.29%	38.45%
KEDMI	CelebA	CelebA	IR152	66.73%	FaceNet112	79.27%	74.55%	18.38%	81.62%	24.45%
			FaceNet64	65.73%		80.53%	78.40%	18.36%	81.64%	21.60%
		FFHQ	VGG16	69.53%		73.13%	69.80%	25.41%	74.59%	30.20%
			IR152	37.67%		52.20%	51.02%	45.84%	54.16%	48.98%
			FaceNet64	36.07%		54.60%	52.24%	41.22%	58.78%	47.76%
			VGG16	38.07%		42.47%	41.33%	55.69%	44.31%	58.67%

must consistently process MI-related prompts, especially identity verification, without triggering refusal behavior.

To quantify these criteria, we test whether the MLLM can accurately recognize samples from a private dataset [37, 47]. As shown in Fig. 2, we design two test sets: • Positive pairs: "Image A" is a real image of the same individual present in "Images B," with the expected answer "Yes." • Negative pairs: "Image A" is a real image of a different individual than those in "Images B," with the expected answer "No". An MLLM is considered a reliable evaluator if it consistently yields high "Yes" rates for Positive pairs and high "No" rates for Negative pairs. Moreover, we expect MLLMs to have a low "Refusal" rate when assessing queries. We conduct this experiment on the widely used MI dataset FaceScrub [25]. We evaluate several SOTA MLLMs, including ChatGPT-5, Gemini-2.0, and Qwen2.5VL-72B within our $\mathcal{F}_{\text{MLLM}}$ framework.

Results. The results are reported in Tab. 3. We observe that Gemini-2.0 achieves high "Yes" rates for Positive pairs and high "No" rates for Negative pairs. Qwen2.5VL-72B may have limited capability of current open-source MLLMs to understand interleaved image-text inputs compared to closed-source commercial models. Despite ChatGPT-5 is a powerful closed-source model, it refuses to assess MI queries with high "Refuse" rates (see examples in the Supp.). Overall, our principle recommends Gemini-2.0 as a reliable MLLM for $\mathcal{F}_{\text{MLLM}}$. Furthermore, $\mathcal{F}_{\text{MLLM}}$ powered by Gemini-2.0 is robust to randomness, additional MI dataset, and aligns well with human evaluation (See Supp.). Note that while Gemini-2.0 is a strong choice in our study, it is not the only MLLM option for $\mathcal{F}_{\text{MLLM}}$. As MLLMs continue to evolve, we can adopt other models as long as they satisfiy our selection principles.

4.2. Reassessing MI Attack Accuracy Using Our MLLM-Based Evaluation Framework

We empirically reassess SOTA MI attacks using our proposed evaluation framework $\mathcal{F}_{\text{MLLM}}$. Importantly, we quan-

titatively show that there are many Type I adversarial examples, which are classified as successful by $\mathcal{F}_{\text{Curr}}$ but do not to capture true visual identity. The false positive rates is consistently high and up to 99% (See Tab. 5). This demonstrates that many SOTA MI methods report inflated attack accuracy, indicating that actual privacy leakage is significantly lower than previously believed.

Experimental Setups. Using our evaluation framework $\mathcal{F}_{\text{MLLM}}$, we reassess 27 SOTA MI attacks across 5 attacks, 4 defenses, 3 private datasets, 4 public datasets, and 9 target models $\boldsymbol{T}$, following the original setups. Detailed settings are in Supp. We will release the MI-reconstructed image collection publicly upon publication.

Significant False Positive by $\mathcal{F}_{Curr}$ are Type I adversarial examples. An ideally successful attack, $\boldsymbol{x}_y^r$ should capture the visual identity of $\boldsymbol{y}$. However, *for a successful attack as according to* $\mathcal{F}_{Curr}$, $\boldsymbol{x}_y^r$ only needs to be classified as $\boldsymbol{y}$ by an evaluation model $\boldsymbol{E}$. As shown in Fig. 1, we observe that within MI-reconstructed images $\boldsymbol{x}_y^r$, there are cases where the visual identity to $\boldsymbol{y}$ is minimal. Nevertheless, $\boldsymbol{E}$ assigns very high probabilities to $\boldsymbol{y}$ for these examples, i.e., high values of $\boldsymbol{P_E}(\boldsymbol{y}|\boldsymbol{x}_y^r)$. We refer to these cases as false positives under the $\mathcal{F}_{Curr}$ framework. To better understand the extent of this false positive rate, we compare the ground truth success rate (established using our $\mathcal{F}_{MLLM}$) to the success rate as measured by $\mathcal{F}_{Curr}$ framework. Particularly, given the MLLM-annotated labels and the prediction via $\mathcal{F}_{Curr}$, we compute the False Positives (FP) rate, False Negatives (FN) rate, True Positives (TP) rate, and True Negatives (TP) rate for each MI setup. The AttAcc via $\mathcal{F}_{Curr}$, $\text{AttAcc}_{\mathcal{F}_{Curr}} = \frac{FP+TP}{FN+TP+FP+TN}$.

The results in Tab. 5 consistently show that the FP rates are significant high across MI setups. In other words, there are numerous MI reconstructed images that do not capture visual identity of, yet they are deemed success by $\mathcal{F}_{Curr}$. Such high FP rate contributes to the significant inflation in reported AttAcc via $\mathcal{F}_{Curr}$ of latest SOTA MI attack such as PPA, PLGMI, IFGMI, or LOMMA. Notably, in their reported AttAcc using $\mathcal{F}_{Curr}$, these recent attacks report AttAcc values exceeding 90%, or even nearly 100% for certain setups. However, across a wide range of MI setups, the actual success rates never reach 80%. While we focus on high FP rates, FN rates also reveal limitations in the $\mathcal{F}_{Curr}$. Across MI setups, FN rates are consistently lower than FP rates. The FN rates depend on the classification accuracy and generalization capability of $\boldsymbol{E}$. For example, under the PLGMI attack, when $\boldsymbol{E}$ = FaceNet112 is trained on CelebA and evaluated with MI reconstructed images also from CelebA prior ($\boldsymbol{\mathcal{D}_{pub}}$ = CelebA), FN rates are lower. In contrast, if this $\boldsymbol{E}$ is evaluated with MI reconstructed images from FFHQ prior ($\boldsymbol{\mathcal{D}_{pub}}$ = FFHQ), FN rates increase due to distribution shifts.

Furthermore, in certain MI setups, we find that $\mathcal{F}_{Curr}$ *does not align well with* $\mathcal{F}_{MLLM}$ *in evaluating MI attacks.* For example, in the setup of MaxViT as $\boldsymbol{T}$ under the PPA attack, the AttAcc by $\mathcal{F}_{Curr}$ is 11.91% lower than the setup for ResNet18 as $\boldsymbol{T}$ under the same attack. However, the MaxViT as $\boldsymbol{T}$ setup shows a 1.82% higher AttAcc measured by $\mathcal{F}_{MLLM}$ than the ResNet18 as $\boldsymbol{T}$ setup. This suggests that, although less effective, the common MI evaluation framework $\mathcal{F}_{Curr}$ could rate the attack as more successful than it actually is. In conclusion, our analysis shows that *the common MI evaluation framework* $\mathcal{F}_{Curr}$ *is suffered from very high FP rate, significantly affecting the reported results of contemporary MI studies based on* $\mathcal{F}_{Curr}$.

> ***Finding 2:*** In a diverse range of setups, we demonstrates the effect of Type I Adversarial features in standard MI evaluation $\mathcal{F}_{Curr}$ resulting in a significant number of false positives. Our proposed $\mathcal{F}_{MLLM}$ mitigates this issue providing more faithful MI evaluation.

5. Conclusion

This work identifies a critical issue in the standard evaluation framework for Model Inversion attacks: the inflation of attack success due to Type I adversarial examples that do not capture true visual identity. To address this, we propose a reliable MLLM-based MI evaluation framework that minimize the impact of Type I adversarial transferability. Our extensive empirical analysis across 27 MI attack setups demonstrates that false positive rates under the standard evaluation framework can reach up to 99%, severely overstating actual privacy leakage. With our proposed evaluation framework, we offer a more accurate and robust way to measure MI attack success, setting a new standard for evaluating privacy risks in machine learning systems. **Additional analysis and limitation are in Supp.**

Acknowledgement

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programmes (AISG Award No.: AISG2-TC-2022-007); The Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021). This research is supported by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Infocomm Media Development Authority. The work is sponsored by the SUTD Decentralised Gap Funding Grant.

References

[1] Shengwei An, Guanhong Tao, Qiuling Xu, Yingqi Liu, Guangyu Shen, Yuan Yao, Jingwei Xu, and Xiangyu Zhang. Mirror: Model inversion for deep learning network with high fidelity. In *Proceedings of the 29th Network and Distributed System Security Symposium*, 2022.

[2] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

[3] Xuankai Chang, Wangyou Zhang, Yanmin Qian, Jonathan Le Roux, and Shinji Watanabe. End-to-end multi-speaker speech recognition with transformer. In *ICASSP 2020*, pages 6134–6138. IEEE, 2020.

[4] Si Chen, Mostafa Kahla, Ruoxi Jia, and Guo-Jun Qi. Knowledge-enriched distributional model inversion attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16178–16187, 2021.

[5] Google Cloud Community. Simplifying data preparation for gen ai with google's gemini. Google Cloud Community, 2025. https://www.googlecloudcommunity.com/gc/Cloud-Product-Articles/Simplifying-Data-Preparation-for-Gen-AI-with-Google-s-Gemini/ta-p/841925.

[6] Jonas Dippel, Steffen Vogler, and Johannes Höhne. Towards fine-grained visual representations by combining contrastive learning with image reconstruction and attention-weighted pooling. *arXiv preprint arXiv:2104.04323*, 2021.

[7] Benoit Dufumier, Pietro Gori, Julie Victor, Antoine Grigis, Michele Wessa, Paolo Brambilla, Pauline Favre, Mircea Polosan, Colm Mcdonald, Camille Marie Piguet, et al. Contrastive learning with continuous proxy meta-data for 3d mri classification. In *MICCAI 2021*. Springer, 2021.

[8] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 17–32, 2014.

[9] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015.

[10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[11] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, pages 1321–1330. PMLR, 2017.

[12] Jianzhu Guo, Xiangyu Zhu, Chenxu Zhao, Dong Cao, Zhen Lei, and Stan Z Li. Learning meta face recognition in unseen domains. In *CVPR*, pages 6163–6172, 2020.

[13] Gyojin Han, Jaehyun Choi, Haeil Lee, and Junmo Kim. Reinforcement learning-based black-box model inversion attacks. In *CVPR*, pages 20504–20513, 2023.

[14] Sy-Tuyen Ho, Koh Jun Hao, Keshigeyan Chandrasegaran, Ngoc-Bao Nguyen, and Ngai-Man Cheung. Model inversion robustness: Can transfer learning help? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12183–12193, 2024.

[15] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2020.

[16] Mostafa Kahla, Si Chen, Hoang Anh Just, and Ruoxi Jia. Label-only model inversion attacks via boundary repulsion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15045–15053, 2022.

[17] Jun Hao Koh, Sy-Tuyen Ho, Ngoc-Bao Nguyen, and Ngai-man Cheung. On the vulnerability of skip connections to model inversion attacks. In *European Conference on Computer Vision*, 2024.

[18] Gautam Krishna, Co Tran, Jianguo Yu, and Ahmed H Tewfik. Speech recognition with no speech or with noisy speech. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1090–1094. IEEE, 2019.

[19] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

[20] Sang Lee and Anisha Patel. Gemini-assisted deep learning classification model for automated medical image annotation. *Journal of Digital Imaging*, 37:123–135, 2024.

[21] Ziang Li, Hongguang Zhang, Juan Wang, Meihui Chen, Hongxin Hu, Wenzhe Yi, Xiaoyang Xu, Mengda Yang, and Chenjun Ma. From head to tail: Efficient black-box model inversion attack via long-tailed learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29288–29298, 2025.

[22] Kaizhao Liang, Jacky Y Zhang, Boxin Wang, Zhuolin Yang, Sanmi Koyejo, and Bo Li. Uncovering the connections between adversarial transferability and knowledge transferability. In *International Conference on Machine Learning*, pages 6577–6587. PMLR, 2021.

[23] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14225–14234, 2021.

[24] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

[25] Hong-Wei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In *2014 IEEE international conference on image processing (ICIP)*, pages 343–347. IEEE, 2014.

[26] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

[27] Ngoc-Bao Nguyen, Keshigeyan Chandrasegaran, Milad Abdollahzadeh, and Ngai-Man Cheung. Re-thinking model inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[28] Ngoc-Bao Nguyen, Keshigeyan Chandrasegaran, Milad Abdollahzadeh, and Ngai man Cheung. Label-only model inversion attacks via knowledge transfer. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[29] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

[30] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE, 2016.

[31] Xiong Peng, Feng Liu, Jingfeng Zhang, Long Lan, Junjie Ye, Tongliang Liu, and Bo Han. Bilateral dependency optimization: Defending against model-inversion attacks. In *KDD*, 2022.

[32] Xiong Peng, Bo Han, Feng Liu, Tongliang Liu, and Mingyuan Zhou. Pseudo-private data guided model inversion attacks. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[33] Yixiang Qiu, Hao Fang, Hongyao Yu, Bin Chen, MeiKang Qiu, and Shu-Tao Xia. A closer look at gan priors: Exploiting intermediate features for enhanced model inversion attacks. In *Proceedings of European Conference on Computer Vision*, 2024.

[34] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[35] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *Advances in neural information processing systems*, 32, 2019.

[36] Laura Smith and Wei Chen. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

[37] Lukas Struppek, Dominik Hintersdorf, Antonio De Almeida Correira, Antonia Adler, and Kristian Kersting. Plug & play attacks: Towards robust and flexible model inversion attacks. In *International Conference on Machine Learning*, pages 20522–20545. PMLR, 2022.

[38] Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. Be careful what you smooth for: Label smoothing can be a privacy shield but also a catalyst for model inversion attacks. In *The Twelfth International Conference on Learning Representations*, 2024.

[39] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[40] Sanli Tang, Xiaolin Huang, Mingjian Chen, Chengjin Sun, and Jie Yang. Adversarial attack type i: Cheat classifiers by significant changes. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):1100–1109, 2019.

[41] Kuan-Chieh Wang, Yan Fu, Ke Li, Ashish Khisti, Richard Zemel, and Alireza Makhzani. Variational model inversion attacks. *Advances in Neural Information Processing Systems*, 34:9706–9719, 2021.

[42] Tianhao Wang, Yuheng Zhang, and Ruoxi Jia. Improving robustness to model inversion attacks via mutual information regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11666–11673, 2021.

[43] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *International conference on machine learning*, pages 23631–23644. PMLR, 2022.

[44] Jiawei Yang, Hanbo Chen, Jiangpeng Yan, Xiaoyu Chen, and Jianhua Yao. Towards better understanding and better generalization of few-shot classification in histology images with contrastive learning. 2022.

[45] Ziqi Yang, Jiyi Zhang, Ee-Chien Chang, and Zhenkai Liang. Neural network inversion in adversarial setting via background knowledge alignment. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 225–240, 2019.

[46] Xiaojian Yuan, Kejiang Chen, Jie Zhang, Weiming Zhang, Nenghai Yu, and Yang Zhang. Pseudo label-guided model inversion attack via conditional generative adversarial network. *Thirty Seventh AAAI Conference on Artificial Intelligence (AAAI 23)*, 2023.

[47] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 253–261, 2020.

[48] Yi Zhou, Cheng Li, and Manish Gupta. A semi-automated approach for crafting outputs with gemini pro. *Advanced Engineering Informatics*, 61:101432, 2024.