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

1. Additional results

1.1. Additional results on designing and implementing our evaluation framework

1.1.1. Our evaluation framework aligns well with human evaluation

In $\mathcal{F}_{MLLM}$, we employ Gemini to assess the success of MI attacks given a query. In the main paper, we demonstrate that Gemini is effective in recognizing samples from the private dataset. Such experiment is conducted with natural images (i.e., training set of Facescrub). In this Supp, we further demonstrate this with MI reconstructed images.

Table 1. We conduct an experiment to demonstrate Gemini's effectiveness in recognizing samples from the private dataset. This results establish that Gemini can serve as a reliable evaluator in MI attack setups. We collect samples for these data from a comprehensive set of MI setups spanning 5 different MI attacks: PPA [16], IFGMI [15], LOMMA [13], KEDMI [1], and PLGMI [20], 3 $\mathcal{D}_{pub}$, 2 $\mathcal{D}_{priv}$, and 8 T. The details of annotation can be found in Sec. 1.1.1.

	"Yes" Rate	"No" Rate
Positive Pair	95.16%	4.84%
Negative Pair	22.88%	77.12%

Setup. To establish positive and negative pairs for MI reconstructed images, we leverage human annotation for them. Since human annotation is costly and time-consuming, we sample 30 images per attack setup across 10 setups spanning 5 different MI attack methods. This results in a total of 300 images. Each image is independently evaluated by four human participants. To mitigate the subjectivity of human evaluation, we retain only the images with high inter-annotator agreement, defined as at least 3 out of 4 consistent annotations. The final label for each retained image is the majority vote among the consistent annotations. After filtering, our human-annotated dataset includes 215 images, which we treat as ground truth to assess the reliability of $\mathcal{F}_{\text{MLLM}}$.

Table 2. We conduct an experiment to demonstrate Gemini's effectiveness in recognizing samples from the private dataset. This results establish that Gemini can serve as a reliable evaluator in MI attack setups.

	Dataset	"Yes" rate (%)	"No" rate (%)
Positive pairs	CelebA	94.88	5.12
	FaceScrub	93.84	3.16
Negative pairs	CelebA	8.25	91.75
	FaceScrub	4.41	95.59

Results. The results are presented in Tab. 1. We observe consistently high "Yes" rates for positive pairs and high "No" rates for negative pairs across datasets. This indicates that Gemini is effective at recognizing samples from the private dataset in both natural and MI-reconstructed images. These results further demonstrate that Gemini serves as a reliable evaluator for our MI setups.

1.1.2. ChatGPT-5 refuses to MI queries

Despite ChatGPT-5 is a powerful closed-source model, it refuses to assess MI queries with high "Refuse" rates. Some examples are provided in Fig. 1.

1.1.3. Our evaluation framework is robust to MI evaluation across datasets

The results in Tab. 3 show that $\mathcal{F}_{MLLM}$ is robust to MI evaluation across commonly used dataset in MI research

*Equal Contribution

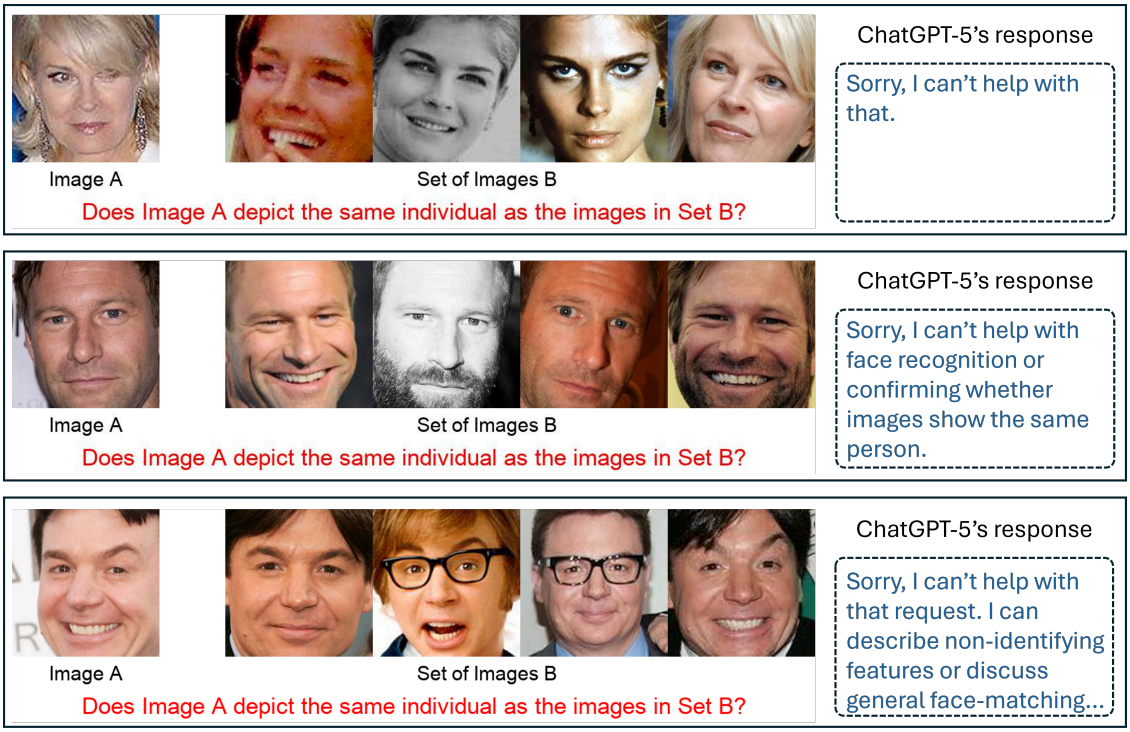

Figure 1. Examples of ChatGPT-5 refusing to evaluate MI-related queries

Table 3. $\mathcal{F}_{MLLM}$ is robust to MI evaluation across prompts

	Dataset	"Yes" rate (%)	"No" rate (%)
Positive pairs	CelebA	94.88	5.12
	FaceScrub	93.84	3.16
Negative pairs	CelebA	8.25	91.75
	FaceScrub	4.41	95.59

including Facescrub and CelebA

1.1.4. Our evaluation framework is robust to MI evaluation across prompts

In this section, we provide an analysis of the variance in our proposed framework with respect to: **(1) the choice of reference images**, and **(2) different questions**, as shown in Fig. 2. For different questions, We run our evaluation framework three times with three different questions: "Does Image A depict the same individual as the images in Set B?", "Does Image A show the same person as those in Set B?", "Is the person in Image A the same as the one(s) shown in Set B?". For different choices of reference images, we run our evaluation framework three times, each with a different random selection of reference images. we show that our $\mathcal{F}_{MLLM}$ is robust to MI evaluation across prompts.

1.2. Evaluation results on MI defenses

Our main focus in this work is MI attacks, where we highlight that the previously reported success rates using the $\mathcal{F}_{Curr}$ are problematic. In fact, the threat of MI attacks has been overestimated and the amount of leaked information is considerably less than previously assumed. As recent MI defenses also use $\mathcal{F}_{Curr}$ to compute MI success rates, we aim to assess the effectiveness of these defenses using our MLLM-annotated dataset.

In this section, we focus on high-resolution setups with PPA [16]. Specifically, we include the latest SOTA MI defenses, such as RoLSS [9], TL [4], LS [17], and TTS [9]. The MI setups strictly follow the configurations in these MI defense studies. The results can be found in Tab. 4.

In general, similar to our observations on MI attacks in the main manuscript, $\mathcal{F}_{Curr}$ may inaccurately assess the effectiveness of SOTA MI defenses. For example, we observe a mismatch between AttAcc comparisons via $\mathcal{F}_{Curr}$ and AttAcc measured by $\mathcal{F}_{MLLM}$. For example, AttAcc via $\mathcal{F}_{Curr}$ suggests that TL [4] outperforms RoLSS and TTS [9]. However, AttAcc via $\mathcal{F}_{MLLM}$ indicates that RoLSS and TTS are more effective defenses. In what follows, we further discuss these results.

These MI defenses result in a reduction in FP rates due to the degradation of the transferability of adversarial characteristics from T to E. Specifically, under TL defense [4], only the later layers of T are fine-tuned on $\mathcal{D}_{priv}$, while earlier

Table 4. **Our investigation on the effectiveness of MI defenses using our MLLM-annotated dataset of MI attack samples.** We present the results of the latest MI defenses including RoLSS [9], TL [4], LS [17], and TTS [9]. We observe a mismatch between AttAcc comparisons via $\mathcal{F}_{Curr}$ and actual AttAcc measured by $\mathcal{F}_{MLLM}$. Overall, consistent with our findings on MI attacks, this suggests that $\mathcal{F}_{Curr}$ may have issues in evaluating MI defenses.

T	Model Acc	$\mathcal{F}_{MLLM}$	$\mathcal{F}_{Curr}$					
		AttAcc	E	**AttAcc**	**FP rate**	**FN rate**	**TP rate**	**TN rate**
ResNet101	94.86%	28.68%	InceptionNetV3	84.69%	82.71%	10.36%	89.64%	17.29%
ResNet101-RoLSS	92.98%	19.46%		43.47%	40.70%	45.09%	54.91%	59.30%
ResNet101-TL	92.51%	25.09%		34.17%	31.27%	57.14%	42.86%	68.73%
ResNet101-TTS	94.16%	18.44%		42.52%	39.39%	43.61%	56.39%	60.61%
ResNet101-LS	92.21%	10.54%		16.56%	14.90%	69.35%	30.65%	85.10%

layers are frozen from the pre-trained backbone. Hence, later layers of T capture $\mathcal{D}_{priv}$ features, while earlier layers of T capture $\mathcal{D}_{pretrain}$ features. In contrast, E captures $\mathcal{D}_{priv}$ features across all layers since all layers of E are fine-tuned on $\mathcal{D}_{priv}$. This mismatch in feature representations between T and E under TL is likely to reduce adversarial transferability [6, 11, 14], thereby reducing FP rates. Under LS defense [17], negative label smoothing (LS) is employed to improve MI robustness. LS slightly reduces label dominance and weakens gradient alignment between surrogate and target models [22]. Negative LS amplifies this effect, further degrading gradient similarity. Therefore, training T with negative LS diminishes gradient alignment with E (trained on standard labels), reducing adversarial transferability [2, 22]. Under RoLSS and TTS defenses [9], removing certain skip connections improves resilience to MI attacks. Skip connections are known to improve adversarial transferability [19]. By modifying T to remove some skip connections, adversarial examples generated by T transfer less effectively to E.

Regarding FN rates, although this is not the main focus of our study, we observe that FN rates tend to increase under MI defenses compared to MI attacks. FN rates depend on the classification accuracy and generalization capability of E. SOTA MI defenses introduce various strategies (e.g., fixing earlier layers trained on public data [4], perturbing labels [17], and removing skip connections [9]) to encode less information in the predictions of T. These approaches may encourage T to learn more generalized features. As a result, reconstructed images based on these generalized features of T may differ more from the seen training data. However, in the prevalent MI setups, E in $\mathcal{F}_{Curr}$ is often trained with standard training procedures and architectures. This could limit its generalization capacity, making it less capable of accurately classifying these reconstructed images via the target models T under MI defenses.

1.3. Additional results on the effect of Type I adversarial attacks in MI on false positive rates

In Sec. 4.2 in the main manuscript, we provide an analysis to demonstrate the effect of Type I adversarial attacks in MI on false positive rates. In this Supp, we provide results on additional setups. The results are presented in Tab. 11. These additional results are consistent with our observation in the main manuscript demonstrating the effect of Type I Adversarial features in MI evaluation resulting in a significant number of false positives.

1.4. Additional extended MI evaluation results

To further broaden the scope of evaluation, we include additional results covering prompt refinement for ChatGPT-5, generic-domain setups, cross-MLLM consistency, label-only API-access MI, and cross-domain validation.

1.4.1. Prompt refinement for ChatGPT-5 evaluation

ChatGPT-5 can exhibit high refusal rates under prompts that directly mention identity-sensitive phrasing. We therefore evaluate a prompt variant with identity-related wording removed. The results in Tab. 5 show that this refinement substantially reduces refusal rates while preserving strong discrimination between positive and negative pairs.

1.4.2. Agreement with human judgment on false positives

To compare evaluator behavior against human judgment, we measure false-positive rates under $\mathcal{F}_{\text{MLLM}}$ and $\mathcal{F}_{\text{Curr}}$. As shown in Tab. 6, $\mathcal{F}_{\text{MLLM}}$ yields substantially fewer false positives.

1.4.3. Generic-domain setup: CIFAR-100

Beyond face recognition, we evaluate a generic-domain setup by adapting PPA with T=ResNet18, $\mathcal{D}_{priv}$=CIFAR100, and $\mathcal{D}_{pub}$=CIFAR10. Results in Tab. 7 show that $\mathcal{F}_{\text{Curr}}$ still exhibits a high FP rate.

Table 5. ChatGPT-5 evaluation outcomes under original and identity-removed prompts.

Prompt	Pos. Yes	Pos. No	Pos. Refusal	Neg. Yes	Neg. No	Neg. Refusal
Original	17.50%	2.67%	79.83%	0.09%	23.04%	76.86%
Identity removal	95.19%	2.87%	1.94%	5.57%	93.32%	1.11%

Table 6. False-positive rates measured with human judgment as reference.

	FP rate
$\mathcal{F}_{\text{MLLM}}$	15.13%
$\mathcal{F}_{\text{Curr}}$	43.70%

1.4.4. Consistency across eligible MLLMs

We evaluate consistency across two eligible MLLMs. As shown in Tab. 8, the evaluation outcomes are stable across model versions.

In addition, sensitivity to the number of reference images remains low: varying from 2 to 5 reference images gives 29.16 ± 1.82, indicating stable evaluation behavior.

1.4.5. Label-only API-access MI setup (BREP-MI)

We also include a label-only API-access MI setup following BREP-MI with T=VGG16 and $\mathcal{D}_{priv}$=CelebA. Tab. 9 again shows high FP rates under $\mathcal{F}_{\text{Curr}}$.

1.4.6. Cross-domain evaluator validation: StanfordDogs

To validate the evaluator outside face recognition, we evaluate a dog-recognition setup on StanfordDogs. The results in Tab. 10 show high yes/no agreement and zero refusal rates.

2. Detailed experimental reproducibility

2.1. Detailed Implementation

Our implementation of $\mathcal{F}_{MLLM}$ is illustrated in Fig. 2. To evaluate whether a reconstructed image is a successful or unsuccessful attack, we employ Gemini 2.0 Flash API (see the main manuscript for our justification for choosing Gemini) for the evaluation.

Given a reconstructed image (Image A), we construct an evaluation query image by pairing it with a set of private training images (Set B) that includes the target identity. We then formulate a natural language textual prompt along with the evaluation query image and pass it to Gemini. The textual prompts are shown in the table below and are fixed across evaluation queries for a fair comparison.

For each reconstructed image, the model outputs a categorical response ("Yes" or "No"). A "Yes" answer is interpreted as a successful attack. By evaluating a large number of such queries and computing the proportion of correct identifications, $\mathcal{F}_{MLLM}$ provide an automated and faithful evaluation of MI.

2.2. The detailed prompt

The detailed textual prompts in our MI evaluation framework can be found in Tab. 12.

2.3. Error Bar of Evaluation results

As mentioned in the main manuscript, we provide an error bar of evaluation results with $\mathcal{F}_{MLLM}$ to further demonstrate the robustness of our proposed MI evaluation framework. The results can be found in Tab. 13

2.4. Detailed MI setup

To ensure the reproducibility, we strictly follow previous studies [1, 4, 9, 13, 15, 16, 21] for MI setups.

MI attacks. Our study focuses on SOTA GAN-based MI attack that achieve strong performance in computer vision domain. These attacks optimize the GAN latent space rather than directly optimize the image space.

KEDMI [1] Introduces an MI-specific GAN that incorporates knowledge from the target classifier. The discriminator performs dual tasks: distinguishing real and fake samples and predicting class-wise labels.

LOMMA [13] Improves MI attacks using a novel logit loss and model augmentation to mitigate overfitting.

PLGMI [20] Leverages conditional GANs to isolate class-specific search spaces and uses Max-Margin Loss to address vanishing gradients in MI optimization.

PPA [16] Utilizes powerful StyleGAN for high-resolution image MI attacks, emphasizing a modular design adaptable to different architectures and datasets.

IFGMI [15] Proposes Intermediate Features Generative Model Inversion, extending optimization from latent codes to intermediate features, enhancing the attack's expressive capability.

MI defense. Our study focuses on SOTA MI defenses. Differ from MI attacks, MI defenses aim to minimize the disclosure of training samples during the MI optimization process.

TL [4] Leverages Transfer Learning to limit sensitive information encoding in earlier layers, degrading MI attack performance.

LS [17] Introduces label smoothing with negative factors to impede class-related information extraction.

Table 7. Results on the CIFAR-100 setup with adapted PPA.

$\mathcal{F}_{\text{MLLM}}$		$\mathcal{F}_{\text{Curr}}$				
AttAcc	E	**AttAcc**	**FP rate**	**FN rate**	**TP rate**	**TN rate**
45.00%	InceptionNetV3	66.00%	57.73%	23.89%	76.11%	42.27%

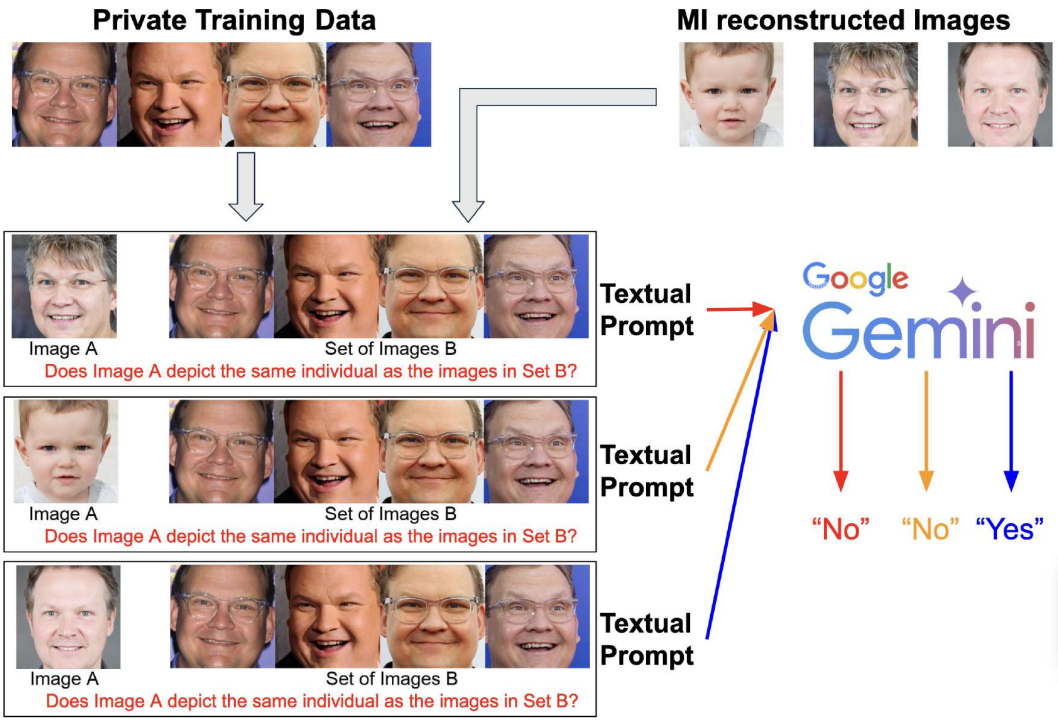

Figure 2. **Our detailed implementation of MLLM-based MI Evaluation Framework** $\mathcal{F}_{MLLM}$. For each reconstructed image, we pair with a set of private training data to construct an evaluation query image. Then, each evaluation query image is passed to Gemini with a textual prompt. The detailed of textual prompt can be found in Sec. 2.2. The final attack accuracy is computed based on Gemini's responses.

Table 8. Evaluation outcomes across eligible MLLMs.

Gemini 2.0	**Gemini 2.5**
28.37%	29.08%

RoLSS [9] Demonstrates that removing skip connections in the last stage significantly reduces MI attack accuracy, offering a better MI robustness trade-off.

TTS [9] Buliding on top of RoLSS. Particularly, in the first stage, the model T with full skip-connections architecture is trained on private dataset. Then in the stage 2, the skip connection removed architecture, i.e. RoLSS, is fine-tuned on private dataset. The pre-trained parameters in Stage 1 serves as initialization for the stage 2, thereby improve the convergence of model in stage 2.

Private training data $\mathcal{D}_{priv}$. Following previous works [1, 4, 9, 13, 15, 16, 21], we focus on reconstruction of images and use the face recognition as a running example including FaceScrub [12] and CelebA [10].

FaceScrub [12]: FaceScrub provides cropped facial images for 530 identities. The dataset publicly a total of 37,878 images. After train/test splitting, this resulted in 34,090 training samples and 3,788 test samples.

CelebA [10]: CelebA is a dataset of celebrity facial images available for non-commercial research. Following previous works [1, 4, 9, 13, 15, 16, 21], we select the top 1,000 identities with the most samples from 10,177 available identities, resulting in 27,034 training samples and 3,004 test samples.

Public data for GAN $\mathcal{D}_{pub}$. Following the data preparation in previous works [1, 4, 9, 13, 15, 16, 21], we use $\mathcal{D}_{pub}$ ensuring that the dataset $\mathcal{D}_{priv}$ and $\mathcal{D}_{pub}$ with no class intersection. $\mathcal{D}_{priv}$ is used to train the target classifier T, while $\mathcal{D}_{pub}$ is used to train GAN to extract general features

Table 9. Results on a label-only API-access BREP-MI setup.

$\mathcal{F}_{\text{MLLM}}$	$\mathcal{F}_{\text{Curr}}$					
AttAcc	E	**AttAcc**	**FP rate**	**FN rate**	**TP rate**	**TN rate**
72.91%	FaceNet112	80.94%	79.01%	18.35%	81.65%	20.99%

Table 10. Cross-domain evaluator validation on StanfordDogs.

Prompt	**Pos. Yes**	**Pos. No**	**Pos. Refusal**	**Neg. Yes**	**Neg. No**	**Neg. Refusal**
Original prompt	90.08%	9.92%	0.00%	1.25%	98.75%	0.00%

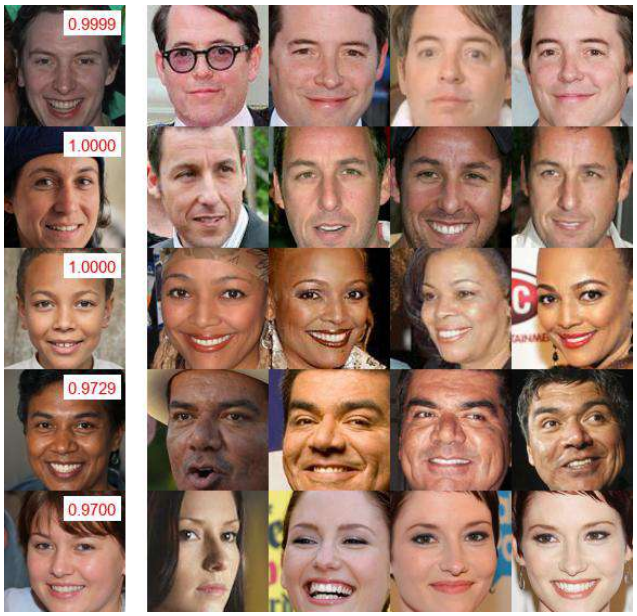

Figure 3. Additional visualization of false positives. These MI false positives do not capture visual identity features of the target individual in the private training data, but they are still deemed successful attacks according to $\mathcal{F}_{Curr}$ with a high confidence (indicated in red text). Here, T=MaxViT [18], $\mathcal{D}_{priv}$=FaceScrub [12], $\mathcal{D}_{pub}$=FFHQ [7], E=InceptioNetV3 under PPA attack [16].

Figure 4. Additional visualization of false positives. These MI false positives 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). Here, T=DenseNet121 [5], $\mathcal{D}_{priv}$=FaceScrub [12], $\mathcal{D}_{pub}$=FFHQ [7], E=InceptioNetV3 under PPA attack [16].

only.

CelebA [10]: Following previous works [1, 4, 13, 21], we select 30,000 images from identities distinct from the 1,000 identities in $\mathcal{D}_{priv}$.

FFHQ [7]: This dataset contains 70,000 high-quality human face images sourced from Flickr, offering significant diversity in age, ethnicity, and backgrounds.

MetFaces [8]. This dataset includes 1,336 high-quality artistic renderings of human faces, covering various art styles. The images exhibit significant diversity and uniqueness.

Target Classifier T. Following previous works [1, 4, 9, 13, 15, 16, 21], we include a wide ranges of architectures as T in our study including ResNet18/101/152 [3], DenseNet121[5], MaxViT [18], FaceNet [1], and VGG16 [3]. To ensure the reproducibility, we utilize the checkpoints of these target classifier in the previous works.

2.5. Computing resources

We conducted all experiments on NVIDIA RTX A5000 GPUs running Ubuntu 20.04.2 LTS, with AMD Ryzen Threadripper PRO 5975WX 32-Core processors. The environment setup includes CUDA 12.2, Python 3.8.18, and PyTorch 1.12.0 with Torchvision 0.14.1. For high-resolution tasks, [15, 16], we use model architectures and pre-trained ImageNet backbone weights from Torchvision. For the low-resolution setup, following [1, 4, 13], we employed VGG architecture with pre-trained ImageNet weights from Torchvision, while we utilize IR152 and FaceNet architectures with

Table 11. **Our controlled experiment to show the effect of Type I adversarial attacks in MI on false positive rates.** We provide results on this experiment on additional setups in Appx. 2.4

Attack	E	$\mathcal{D}_{priv}$	$\mathcal{D}_{pub}$	T		**FP rates under** E
PPA	InceptionV3	Facescrub	FFHQ	Resnet101	Neg x_y^r	82.71%
					Neg $x_y^{natural}$	0.94%
				Resnet152	Neg x_y^r	85.09%
					Neg $x_y^{natural}$	0.94%
				MaxViT	Neg x_y^r	79.48%
					Neg $x_y^{natural}$	0.94%
				DenseNet121	Neg x_y^r	72.41%
					Neg $x_y^{natural}$	0.94%
IFGMI			MetFaces	Resnet18	Neg x_y^r	72.71%
					Neg $x_y^{natural}$	0.94%
PLGMI	FaceNet112	CelebA	FFHQ	VGG16	Neg x_y^r	88.49%
					Neg $x_y^{natural}$	0.00%
LOMMA			CelebA	FaceNet64	Neg x_y^r	93.73%
					Neg $x_y^{natural}$	0.00%
				IR152	Neg x_y^r	86.80%
					Neg $x_y^{natural}$	0.00%
				VGG16	Neg x_y^r	90.70%
					Neg $x_y^{natural}$	0.00%
			FFHQ	FaceNet64	Neg x_y^r	69.73%
					Neg $x_y^{natural}$	0.00%
				IR152	Neg x_y^r	77.85%
					Neg $x_y^{natural}$	0.00%
				VGG16	Neg x_y^r	61.55%
					Neg $x_y^{natural}$	0.00%
KEDMI			CelebA	FaceNet64	Neg x_y^r	78.40%
					Neg $x_y^{natural}$	0.00%
				IR152	Neg x_y^r	74.55%
					Neg $x_y^{natural}$	0.00%
				VGG16	Neg x_y^r	69.80%
					Neg $x_y^{natural}$	0.00%
			FFHQ	FaceNet64	Neg x_y^r	52.24%
					Neg $x_y^{natural}$	0.00%
				IR152	Neg x_y^r	51.02%
					Neg $x_y^{natural}$	0.00%
				VGG16	Neg x_y^r	41.33%
					Neg $x_y^{natural}$	0.00%

Table 12. Detailed textual prompts in our MI evaluation framework $\mathcal{F}_{MLLM}$

Domain	$\mathcal{D}_{priv}$	Textual Prompt
Face Recognition	Facescrub	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
	CelebA	
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

pre-trained backbones from face.evoLVe[1].

We employ the Gemini 2.0 Flash API in $\mathcal{F}_{\text{MLLM}}$ and emphasize that our implementation is both reliable and cost efficient. Particularly, in our implementation, each evaluation query costs \$0.0002886 (see the official Gemini API documentation[2] for cost estimation). This cost is reasonable for large-scale evaluations. For example, in our study involving larger-scale 26 experimental setups and a total of 71,880 MI-reconstructed images, the overall cost is around \$20.75, making our evaluation framework scalable and accessible for future research.

3. Additional visualization of false positives

In the main paper, we provide some visualizations of MI false positives. In this Supp., we provide more extensive visualizations of MI false positives in Fig. 3, 4, 5, 6, 7.

[1] https://github.com/ZhaoJ9014/face.evoLVe

[2] `https://ai.google.dev/gemini-api/docs`

Table 13. **Our investigation on MI evaluation framework using our comprehensive dataset of MI attack samples.** We run the evaluations with our $\mathcal{F}_{MLLM}$ threes times and report mean $\pm$ std.

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	28.22±0.30%	InceptionNetV3	91.39%	90.03±0.09%	4.82±0.51%	94.82±0.32%	9.97±0.09%
			ResNet101	28.48±0.36%		84.69%	82.79±0.09%	10.52±0.16%	89.48±0.16%	17.21±0.09%
			ResNet152	30.20±0.09%		86.84%	85.13±0.14%	9.21±0.34%	90.79±0.34%	14.87±0.14%
			DenseNet121	27.44±0.27%		72.41%	70.11±0.05%	21.52±0.07%	78.48±0.07%	29.89±0.05%
			MaxViT	30.30±0.13%		79.48%	77.32±0.16%	15.54±0.33%	84.46±0.34%	22.68±0.14%
IFGMI	FaceScrub	FFHQ	ResNet18	34.14±0.29%	InceptionNetV3	95.85%	94.61±0.03%	1.75±0.07%	98.25±0.07%	5.39±0.03%
		Metfaces	ResNet18	1.57±0.07%		72.50%	72.24±0.05%	11.39±3.74%	88.61±3.74%	27.76±0.05%
PLGMI	CelebA	CelebA	VGG16	73.51±0.75%	FaceNet112	98.73%	99.33±0.14%	1.48±0.05%	98.52±0.05%	0.67±0.14%
		FFHQ	VGG16	48.59±0.57%		88.67%	88.51±0.07%	11.16±0.06%	88.84±0.06%	11.47±0.07%
LOMMA	CelebA	CelebA	IR152	79.02±0.30%	FaceNet112	92.00%	92.47±1.22%	8.13±0.33%	91.87±0.33%	7.53±1.22%
			FaceNet64	79.76±0.27%		90.40%	87.71±0.80%	8.92±0.20%	91.08±0.20%	12.29±0.08%
			VGG16	80.73±0.77%		90.13%	90.29±0.95%	9.91±0.22%	90.09±0.21%	9.71±0.95%
		FFHQ	IR152	45.60±0.58%		77.73%	77.37±0.45%	21.84±0.52%	78.16±0.52%	22.63±0.45%
			FaceNet64	45.49±0.74%		72.13%	69.92±0.18%	25.21±0.17%	74.79±0.17%	30.08±0.18%
			VGG16	56.09±1.20%		63.07%	61.33±0.31%	35.58±0.18%	64.42±0.18%	38.67±0.31%
KEDMI	CelebA	CelebA	IR152	67.24±0.83%	FaceNet112	79.27%	74.97±0.37%	18.64±0.25%	81.36±0.25%	24.70±0.23%
			FaceNet64	66.15±0.73%		80.53%	77.27±1.04%	17.81±0.49%	82.19±0.49%	30.15±1.56%
			VGG16	69.38±1.04%		73.13%	69.85±1.56%	25.44±0.62%	74.56±0.62%	30.15±1.56%
		FFHQ	IR152	36.96±0.62%		52.20%	50.24±0.75%	44.42±1.34%	55.58±1.34%	49.76±0.75%
			FaceNet64	35.96±0.14%		54.60%	52.08±0.62%	40.91±1.13%	59.09±1.13%	47.92±0.62%
			VGG16	38.85±0.80%		42.47%	41.24±0.36%	55.50±0.58%	44.40±0.58%	58.76±0.36%

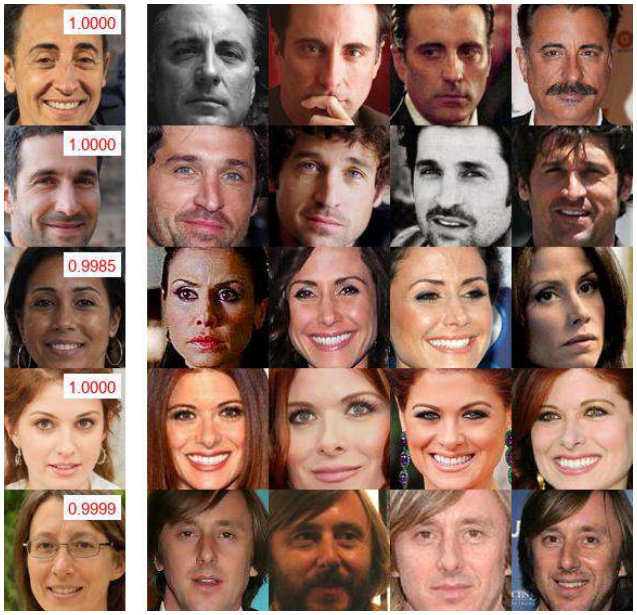

Figure 5. Additional visualization of false positives. These MI false positives do not capture visual identity features of the target individual in the private training data, but they are still deemed successful attacks according to $\mathcal{F}_{Curr}$ with a high confidence (indicated in red text). Here, T=ResNet101 [3], $\mathcal{D}_{priv}$=FaceScrub [12], $\mathcal{D}_{pub}$=FFHQ [7], E=InceptioNetV3 under PPA attack [16].

Figure 6. Additional visualization of false positives. These MI false positives 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). Here, T=ResNet152 [3], $\mathcal{D}_{priv}$=FaceScrub [12], $\mathcal{D}_{pub}$=FFHQ [7], E=InceptioNetV3 under PPA attack [16].

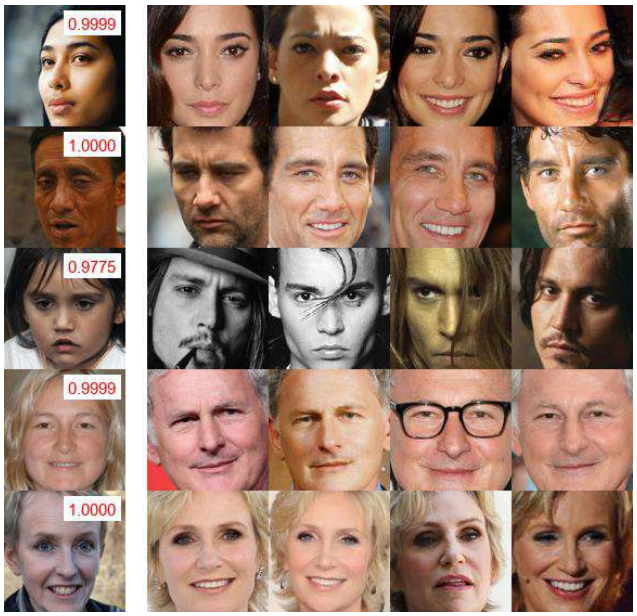

Figure 7. Additional visualization of false positives. These MI false positives do not capture visual identity features of the target individual in the private training data, but they are still deemed successful attacks according to $\mathcal{F}_{Curr}$ with a high confidence (indicated in red text). Here, T=ResNet18 [3], $\mathcal{D}_{priv}$=FaceScrub [12], $\mathcal{D}_{pub}$=FFHQ [7], E=InceptioNetV3 under IFGMI attack [15].

These false positive MI do not capture the visual identity features of the target individual in private training data, but are still considered successful attacks according to $\mathcal{F}_{Curr}$ with high confidence.

4. Limitation

While this study provides valuable insights into the limitations of the MI evaluation framework and propose a more reliable automated MI evaluation framework for future MI study, it is important to acknowledge certain limitations. One such limitation is the focus on specific architectures and datasets. While we strictly follow previous works [1, 4, 9, 13, 15, 16, 21] to includes 26 MI setups, these setups may not include the latest architectures or dataset that are not considered in prevalent MI setups. Future research could expand upon our findings by exploring a wider range of model architectures and datasets. This would further shed the light of MI evaluation and contribute to the development of better MI evaluation frameworks.

5. Ethical Statement

This study examines the limitations of widely used evaluation frameworks for Model Inversion (MI) attacks, which hold critical implications for privacy and data security. Our analysis reveals an overestimation of MI attack success rates, underscoring the need for accurate and reliable evaluation metrics to avoid inflated perceptions of privacy risks. To support the research community, we propose a more reliable and cost-efficient MI evaluation framework based on MLLM. Furthermore, we release the code and a large-scale collection of MI reconstructed images upon publication, advocating for their ethical use to advance privacy protection.

6. LLM Usage

We used a large language model to help polish the grammar, wording, and other minor text issues in this manuscript. The authors are fully responsible for the ideas, analysis and conclusions in this submission.

References

[1] 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.

[2] Ambra Demontis, Marco Melis, Maura Pintor, Matthew Jagielski, Battista Biggio, Alina Oprea, Cristina Nita-Rotaru, and Fabio Roli. Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In *28th USENIX security symposium (USENIX security 19)*, pages 321–338, 2019.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] 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.

[5] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[6] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. 2019.

[7] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[8] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020.

[9] 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.

[10] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of*

the IEEE international conference on computer vision, pages 3730–3738, 2015.

[11] Avery Ma, Amir-massoud Farahmand, Yangchen Pan, Philip Torr, and Jindong Gu. Improving adversarial transferability via model alignment. In *European Conference on Computer Vision*, pages 74–92. Springer, 2024.

[12] 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.

[13] 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.

[14] Zeyu Qin, Yanbo Fan, Yi Liu, Li Shen, Yong Zhang, Jue Wang, and Baoyuan Wu. Boosting the transferability of adversarial attacks with reverse adversarial perturbation. pages 29845–29858, 2022.

[15] 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.

[16] 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.

[17] 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 *ICLR*, 2024.

[18] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *European conference on computer vision*, pages 459–479. Springer, 2022.

[19] Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma. Skip connections matter: On the transferability of adversarial examples generated with resnets. In *International Conference on Learning Representations*, 2020.

[20] 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.

[21] 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.

[22] Yechao Zhang, Shengshan Hu, Leo Yu Zhang, Junyu Shi, Minghui Li, Xiaogeng Liu, Wei Wan, and Hai Jin. Towards understanding adversarial transferability from surrogate training. *IEEE Symposium on Security and Privacy (SP)*, 2024.