# Do Vision-Language Models Leak What They Learn? Adaptive Token-Weighted Model Inversion Attacks

Ngoc-Bao Nguyen[1] Sy-Tuyen Ho[1,2] Koh Jun Hao[1] Ngai-Man Cheung[1]
[1]Singapore University of Technology and Design (SUTD) [2] University of Maryland, College Park
{thibaongoc_nguyen, ngaiman_cheung}@sutd.edu.sg

## Supplementary material

In this supplementary material, we provide additional experiments, analysis, ablation study, and details that are required to reproduce our results. These are not included in the main paper due to space limitations.

## Contents



## 1. Research Reproducibility Details

### 1.1. Hyperparameters

To fine-tune the VLMs, we follow the standard hyperparameters provided in the official implementations of LLaVA-v1.6-Vicuna-7B[1] [14], Qwen2.5-VL-7B[2] [3], MiniGPT-v2[3] [4], and InternVL2.5 [4] [6, 7]. Fine-tuning is conducted on the VQA-FaceScrub, VQA-CelebA, and VQA-StanfordDogs datasets.

For the attacks, we use $N = 70$ inversion steps for all experiments. The inversion update rate $\beta = 0.05$.

To compute the regularization term $f_{reg}$ in Eqn. 3, we follow [16] by using 2,000 images from a public dataset $\mathcal{D}_{pub}$ to estimate the mean and variance of the penultimate layer activations of the VLMs.

### 1.2. Computational Resources

All experiments were conducted on NVIDIA RTX A6000 Ada GPUs running Ubuntu 20.04.2 LTS, equipped with AMD Ryzen Threadripper PRO 5975WX 32-core processors. The environment setup for each model is provided in the official implementations of the VLMs, including: LLaVA-v1.6-Vicuna-7B [14], Qwen2.5-VL-7B [3], MiniGPT-v2 [4], and InternVL2.5 [6, 7].

To evaluate $AttAcc_M$, we strictly follow the protocol in [11], using the Gemini 2.0 Flash API. In total, we evaluate nearly 100,000 MI-reconstructed images for our main experiments (main paper).

## 2. Additional results

### 2.1. Extended Evaluation on Publicly Released VLM

In this section, we extend our analysis to the publicly available LLaVA-v1.6-7B model [14] and MiniGPTv2 [4], aiming to reconstruct training images from accessing the model only.

Figure S.1 and Figure S.2 show the results of our best setup of MI attack, SMI-AW using the logit maximization loss $\mathcal{L}_{LOM}$. The target is to reconstruct images of celebrities that appear in the training dataset of the LLaVA-v1.6-

[1] https://github.com/haotian-liu/LLaVA
[2] https://github.com/QwenLM/Qwen2.5-VL
[3] https://github.com/Vision-CAIR/MiniGPT-4
[4] https://github.com/OpenGVLab/InternVL

7B and MiniGPTv2 model. To reconstructed images from the model, we use the textual input $t$ = "What is the person's name in the image? Return only their name" and the target textual answer is a celebrity's name, i.e $y$ = "Beyoncé".

We visualize image pairs: in each pair, the right image is the reconstruction generated from the publicly available model, and the left image is a training image of an individual. We emphasize that the training dataset is fully unknown and inaccessible for the inversion attack. The visual similarity between the pairs indicates that the pre-trained VLM may reveal identifiable information from its training data, exposing a vulnerability to model inversion attacks.

### 2.2. Additional Qualitative Results

Reconstructed images from the FaceScrub dataset using four VLMs, LLaVA-v1.6-7B, MiniGPT-v2, Qwen2.5-VL, and InternVL2.5 are shown in Figure S.3, Figure S.4, Figure S.5, and Figure S.6, respectively. For the CelebA and Stanford Dogs datasets, reconstructed images using LLaVA-v1.6-7B are presented in Figure S.7 and Figure S.8. All reconstructions are generated using SMI-AW with the logit maximization loss $\mathcal{L}_{LOM}$.

For each pair, the left column shows images from the private training dataset, while the right column presents the reconstructed images corresponding to each individual in the left column. Qualitative results demonstrate the effectiveness of our method. This strong visual similarity highlights the ability of our model inversion approach to recover identifiable features from the training data.

### 2.3. Additional Attention Map Analysis

Additional attention map of four models including LLaVa-1.6-7B, MiniGPTv2, Qwen2.5-VL, and InternVL2.5 are visualized in Figure S.9, Figure S.10, Figure S.11, and Figure S.12. We visualize the cross-attention map between the reconstructed image and each output token during inversion. Different tokens exhibit markedly different attention maps: visually grounded tokens show strong attention, while others produce weak responses, indicating limited reliance on the image. Moreover, attention patterns evolve over inversion steps, as a token's dependence on visual input changes when the reconstructed image becomes more consistent with the target output. These observations reveal that token-level gradients vary substantially in visual informativeness both across tokens and over time. This motivates our SMI-AW method, which dynamically reweights token contributions based on their visual attention strength.


(a) Clint Eastwood

(b) Eva Longoria

(c) Chris Evans

(d) Jennifer Aniston

(e) George Clooney

Figure S.1. Reconstructed images using our SMI-AW with $\mathcal{L}_{LOM}$ on the publicly available LLaVA-v1.6-7B model. Each pair consists of a reconstructed image (right) and a corresponding training image (left) in the training dataset of LLaVA-v1.6-7B model. We emphasize that the training dataset is fully unknown and inaccessible for the inversion attack. The strong similarity suggests the pre-trained VLM may leak identifiable training data, exposing it to model inversion attacks.

## 3. Ablation Study

### 3.1. Ablation Study on input prompt $y$

In this section, we further evaluate SMI-AW using a more diverse set of input prompts $y$. The results are summarized in Table S.1. It shows that SMI-AW maintains consistently strong attack performance across different prompt choices, demonstrating its robustness to prompt variation.

### 3.2. Error Bar

We repeat each experiment three times using different random seeds and report the results in Table S.2. Specifically,

Table S.1. We evaluate SMI-AW using a more diverse set of input prompts $y$. Here, we use $M$ = LLaVa-v1.6-7B, $\mathcal{D}_{priv}$ = Facescrub and logit maximization loss $\mathbf{L}_{LOM}$.

| Input question | $AttAcc_M \uparrow$ | $AttAcc_D \uparrow$ | | $\delta_{face} \downarrow$ | $\delta_{eval} \downarrow$ |
|---|---|---|---|---|---|
| | | $Top1$ | $Top5$ | | |
| Who is the person in the image? | 61.01% | 37.62% | 66.16% | 0.7265 | 134.94 |
| What is the person's name in the image? | 59.08% | 37.10% | 64.62% | 0.7318 | 135.28 |
| Who is the man/woman in the photo? | 59.98% | 37.78% | 64.25% | 0.7348 | 135.65 |

we use $M$ = LLaVA-v1.6-7B, $\mathcal{D}_{priv}$ = Facescrub. The results demonstrate that our attacks have low standard deviation.

# 4. Experimental setting

## 4.1. Inversion Loss Design for VLMs

In this section, we present the adaptation of the inversion loss from conventional unimodal MI to VLMs. Specifically, the inversion loss in traditional MI typically consists of two components: $\mathcal{L}_{inv} = \mathcal{L}_{id} + \mathcal{L}_{prior}$, where the identity loss $\mathcal{L}_{id}$ guides the generator $G(w)$ to produce images that induce the label $y$ from the target model $M_{DNN}$, and $\mathcal{L}_{prior}$ is a regularization or prior loss. To extend this to VLMs, we focus on adapting the identity loss $\mathcal{L}_{id}$. We categorize it into two main types: cross-entropy-based and logit-based losses.

**Cross-entropy-based.** This loss is widely used in MI attacks [5, 19, 28] to optimize $w$ such that the reconstruction has the highest likelihood for the target class under the model $M$. For VLMs, we adapt the cross-entropy loss $\mathcal{L}_{CE}$ for each target token $y_i$ as follows:

$$\mathcal{L}_{CE}(M(\mathbf{t}, G(w), y_{<i}), y_i) = -\log \mathbb{P}_M(y_i|\mathbf{t}, G(w), y_{<i}) \tag{1}$$

$\mathbb{P}_M(y_i|\mathbf{t}, G(w), y_{<i})$ denotes the predicted probability of token $y_i$, computed over the tokenizer vocabulary of the VLM (e.g., LLaVa-v1.6 uses a vocabulary of 32,000 tokens).

**Logit-based.** Prior work shows that using cross-entropy loss in MI can lead to gradient vanishing [27] or sub-optimal results [16]. To address this, Yuan et al. [27] and Nguyen et al. [16] propose optimizing losses directly over logits of a target class. We adopt two such logit-based losses for VLMs: the Max-Margin Loss $\mathcal{L}_{MML}$ [27] and the Logit-Maximization Loss $\mathcal{L}_{LOM}$ [16] for a target token $y_i$:

$$\begin{aligned}\mathcal{L}_{MML}(M(\mathbf{t}, G(w), y_{<i}), y_i) = &-l_{y_i}(\mathbf{t}, G(w), y_{<i}) \\ &+ \max_{k \neq y_i} l_k(\mathbf{t}, G(w), y_{<i})\end{aligned} \tag{2}$$

$$\begin{aligned}\mathcal{L}_{LOM}(M(\mathbf{t}, G(w), y_{<i}), y_i) = &-l_{y_i}(\mathbf{t}, G(w), y_{<i}) \\ &+ \lambda \|f_{y_i} - f_{reg}\|_2^2\end{aligned} \tag{3}$$

Here, $l_{y_i}$ is the logit corresponding to the target token $y_i$, $\lambda$ is a hyperparameter, $f_{y_i} = M^{pen}(\mathbf{t}, G(w), y_{<i})$ where $M^{pen}()$ denotes the function that extracts the penultimate layer representations for a given input, and $f_{reg}$ is a sample activation from the penultimate layer $M^{pen}()$ computed using public images from $\mathcal{D}_{pub}$. Following [16], the distribution of $f_{reg}$ is estimated over 2000 input pairs $(\mathbf{t}, \mathbf{x}_{pub})$, where $\mathbf{x}_{pub} \in \mathcal{D}_{pub}$. $\mathcal{L}_{MML}$ maximizes the logit of the correct token $y_i$ while penalizing the highest incorrect logit to mitigate gradient vanishing. On the other hand, $\mathcal{L}_{LOM}$ also maximizes the correct token's logit to avoid sub-optimality, while additionally penalizing deviations in the penultimate activations to prevent unbounded logits problem.

## 4.2. Evaluation metrics

In this section, we provide a detailed implementation for five metrics used in our work to access MI attacks.

- **Attack accuracy.** Attack accuracy measures the success rates of MI attacks. Following existing literature, we compute attack accuracy via three frameworks:
  - **Attack accuracy evaluated by conventional evaluation framework** $\mathcal{F}_{DNN}$ ($AttAcc_D \uparrow$) [5, 16, 19, 21, 28]. Following [21, 22], we use InceptionNet-v3 [23] as the evaluation model. For a fair comparison, we use the identical checkpoints of InceptionNet-v3 for Facescrubs, CelebA and Stanford Dogs from [21] for evaluation of each dataset. We report *Top-1* and *Top-5* Accuracy.
  - **Attack accuracy evaluated by MLLM-based evaluation framework** $\mathcal{F}_{MLLM}$ ($AttAcc_M \uparrow$). [11] demonstrate that $\mathcal{F}_{MLLM}$ can achieve better alignment with human evaluation than $\mathcal{F}_{DNN}$ ($AttAcc_D \uparrow$ by mitigating Type-I adversarial transferability. The evaluation involves presenting a reconstructed image (image A) and a set of private reference images (set B) to an MLLM (e.g., Gemini 2.0 Flash), and prompting it with the question: "Does image A depict the same individual as images in set B?" If the model responds "Yes", the attack is considered successful. An example query is shown in Fig. S.13.
  - **Attack accuracy evaluated by human** $\mathcal{F}_{Human}(AttAcc_H \uparrow)$. Following existing studies [1, 16], we conduct the user study on Amazon Mechanical Turk. Participants are asked to evaluate the success of MI-reconstructed by referencing the corresponding private images. Similar to $\mathcal{F}_{MLLM}$, it

Table S.2. Error bars for our two model inversion strategies SMI and SMI-AW. Each experiment was repeated 3 times, and we report the mean and standard deviation of the attack performance. Here, we use $M$ = LLaVa-v1.6-7B, $\mathcal{D}_{priv}$ = Facescrub. All inversion strategies are combined with logit maximization loss $\mathbf{L}_{LOM}$.

| Method | $AttAcc_M \uparrow$ | $AttAcc_D \uparrow$ | | $\delta_{face} \downarrow$ | $\delta_{eval} \downarrow$ |
|---|---|---|---|---|---|
| | | $Top1$ | $Top5$ | | |
| SMI | $57.83 \pm 1.18\%$ | $33.50 \pm 0.19\%$ | $61.56 \pm 0.30\%$ | $0.7473 \pm 0.0006$ | $137.89 \pm 2.62$ |
| SMI-AW | $59.53 \pm 0.93\%$ | $37.76 \pm 0.32\%$ | $66.18 \pm 0.13\%$ | $0.7265 \pm 0.0038$ | $134.94 \pm 0.64$ |

involves presenting an image A and a set of images B. They are asked to answer "Yes" or "No" to indicate whether image A depicts the same identity as images in set B (see Fig. S.13). Each image pair is shown in a randomized order and displayed for up to 60 seconds. Each user study involves 4,240 participants for the FaceScrub dataset and 8,000 participants for the CelebA dataset.

- **Feature distance.** We compute the $l_2$ distance between the feature representations of the reconstructed and the private training images [21]. Lower values indicate higher similarity and better inversion quality.
  - $\delta_{eval}$. Features are extracted by the evaluation model as used in $\mathcal{F}_{DNN}$.
  - $\delta_{face}$. Features are extracted by a pre-trained FaceNet model [20].

### 4.3. Initial Candidate Selection

Following the method from [21], we perform an initial selection to identify promising candidates for inversion. We begin by sampling 2000 latent vectors, denoted as $\{w\}_{i=1}^{2000}$, from the prior distribution. For each $w$, we evaluate the target VLMs loss. We then select the top $n$ vectors with the lowest loss to serve as our initialization candidates. In our experiments, we set $n = 16$ to create 16 candidates for attacks.

### 4.4. Final Selection

To select the final reconstructed image, we perform a final selection step, also following the method from [21]. This step aims to identify the reconstructed images that have the highest confidence. For each of the $n$ initialization candidates, we apply 10 random data augmentations and re-evaluate the target VLMs loss. We calculate the average loss for each candidate across these augmentations and select the $n/2$ candidates with the lowest average loss as the final attack outputs.

## 5. Related Work

**Model Inversion.** Model Inversion (MI) seeks to recover information about a model's private training data via pretrained model. Given a target model $M$ trained on a private dataset $\mathcal{D}_{\text{priv}}$, the adversary aims to infer sensitive information about the data in $\mathcal{D}_{\text{priv}}$, despite it being inaccessible after training. MI attacks are commonly framed as the task of reconstructing an input that the model $M$ would classify as belonging to a particular label $y$. The foundational MI method is introduced in [8], demonstrating that machine learning models could be exploited to recover patients' genomic and demographic data.

**Model Inversion in Unimodal Vision Models.** Model Inversion (MI) has been extensively studied to reconstruct private training images in unimodal vision models. For example, in the context of face recognition, MI attacks attempt to recover facial images that the model would likely associate with a specific individual.

Building on the foundational work of [8], early MI attacks targeting facial recognition are proposed in [9, 26], demonstrating the feasibility of reconstructing recognizable facial images from the outputs of pretrained models. However, performing direct optimization in the high-dimensional image space is challenging due to the large search space. To address this, recent advanced generative-based MI attacks have shifted the search to the latent space of deep generative models [5, 16, 19, 21, 24, 26–28].

Specifically, GMI [28] and PPA [21] employ WGAN [2] and StyleGAN [12], respectively, trained on an auxiliary public dataset $\mathcal{D}_{\text{pub}}$ that similar to the private dataset $\mathcal{D}_{\text{priv}}$. The pretrained GAN is served as prior knowledge for the inversion process. To improve this prior knowledge, KEDMI [5] trains inversion-specific GANs using knowledge extracted from the target model $M$. PLGMI [27] introduces pseudo-labels to enhance conditional GAN training. IF-GMI [19] utilizes intermediate feature representations from pretrained GAN blocks. Most recently, PPDG-MI [18] improves the generative prior by fine-tuning GANs on high-quality pseudo-private data, thereby increasing the likelihood of sampling reconstructions close to true private data. Beyond improving GAN-based priors, several studies focus on improving the MI objective including max-margin loss [27] and logit loss [16] to better guide the inversion process. Additionally, LOMMA [16] introduces the concept of augmented models to improve the generalizability of MI attacks.

Unlike MI attacks, MI defenses aim to reduce the leakage of private training data while maintaining strong predictive performance. Several approaches have been proposed


(a) Beyoncé

(b) Barack Obama

(c) Jennifer Lopez

(d) Donald Trump

(e) Kristen Stewart

Figure S.2. Reconstructed images using our SMI-AW with $\mathcal{L}_{LOM}$ on the publicly available MiniGPTv2 model. Each pair consists of a reconstructed image (right) and a corresponding training image (left) in the training dataset of MiniGPTv2 model. We emphasize that the training dataset is fully unknown and inaccessible for the inversion attack. The strong similarity suggests the pre-trained VLM may leak identifiable training data, exposing it to model inversion attacks.

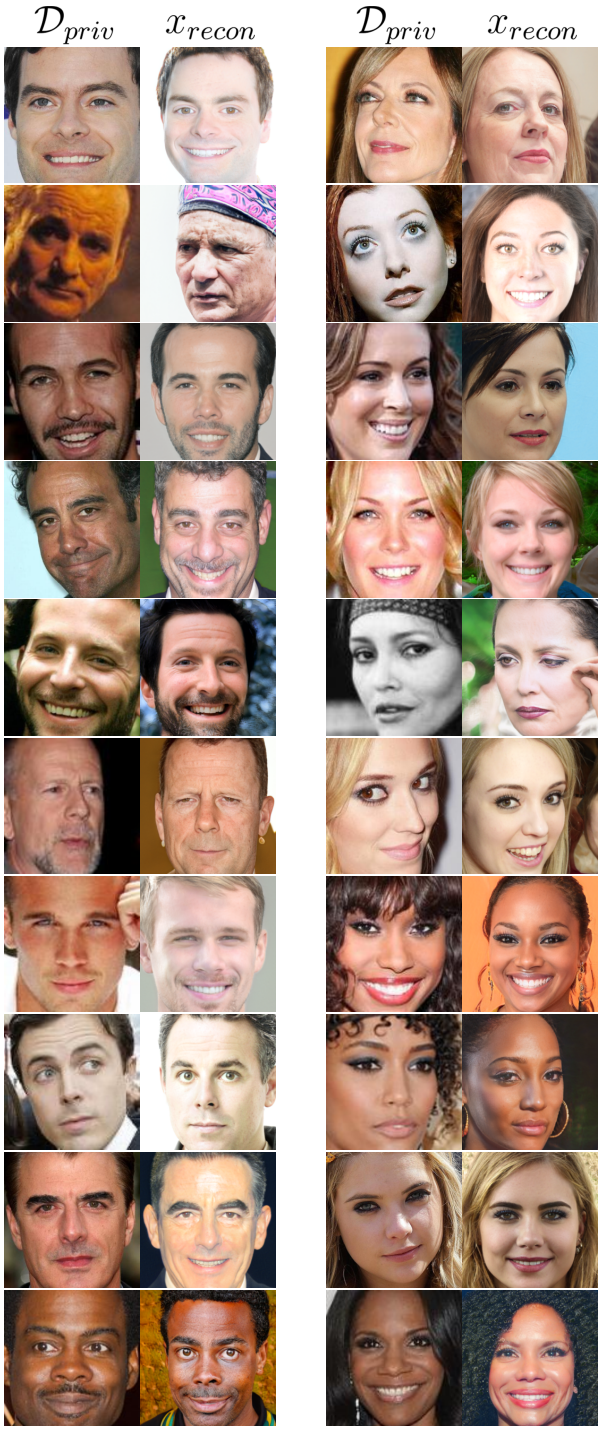



Figure S.3. Qualitative results on Facescrub dataset using the SMI-AW and $\mathcal{L}_{LOM}$, $M$ = LLaVA-v1.6-7B. For each pair, the left column shows images from the private training dataset, while the right column presents the reconstructed images corresponding to each individual in the left column.

to defend against MI attacks. MID [25] and BiDO [17] introduce regularization-based defenses that include the term of regularization in the training objective. The crucial drawback of these approaches is that the regularizers often conflict with the training objective resulting in a significant degradation in model's utility. Beyond regularization-based strategies, TL-DMI [10] leverages transfer learning to improve MI robustness, and LS [22] applies Negative Label Smoothing to mitigate inversion risks. Architectural ap-

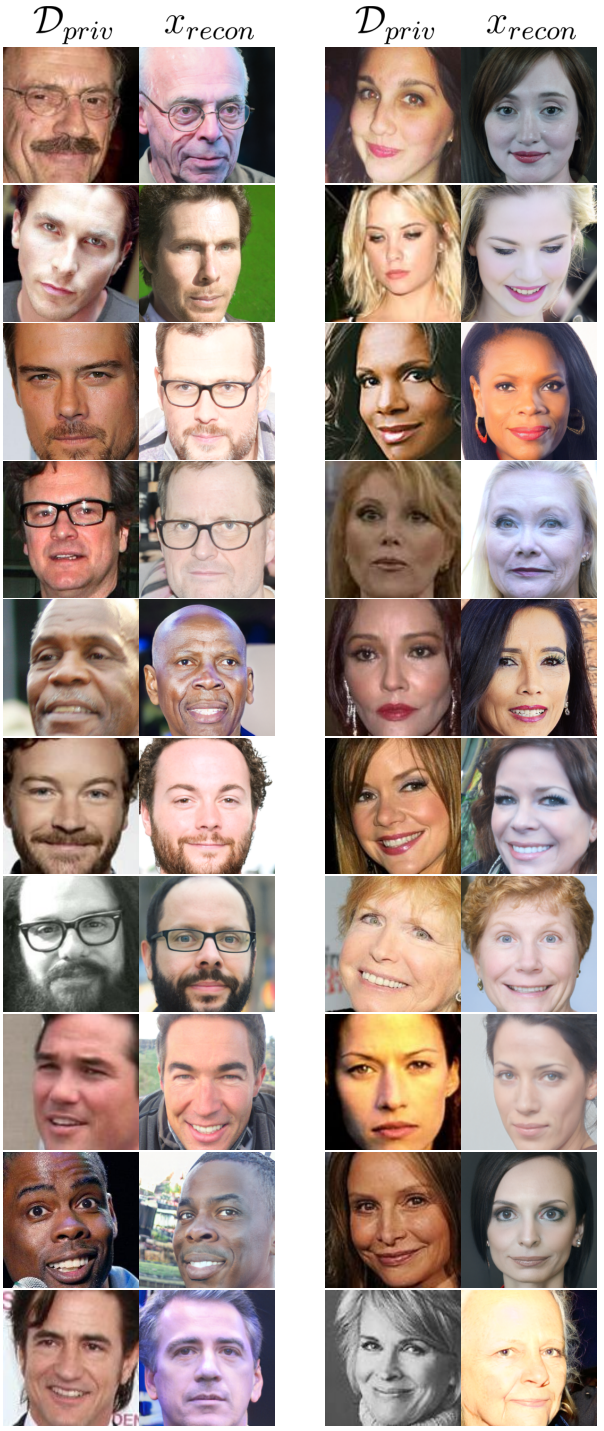



Figure S.4. Qualitative results on Facescrub dataset using the SMI-AW and $\mathcal{L}_{LOM}$, $M$ = MiniGPT-v2. For each pair, the left column shows images from the private training dataset, while the right column presents the reconstructed images corresponding to each individual in the left column.

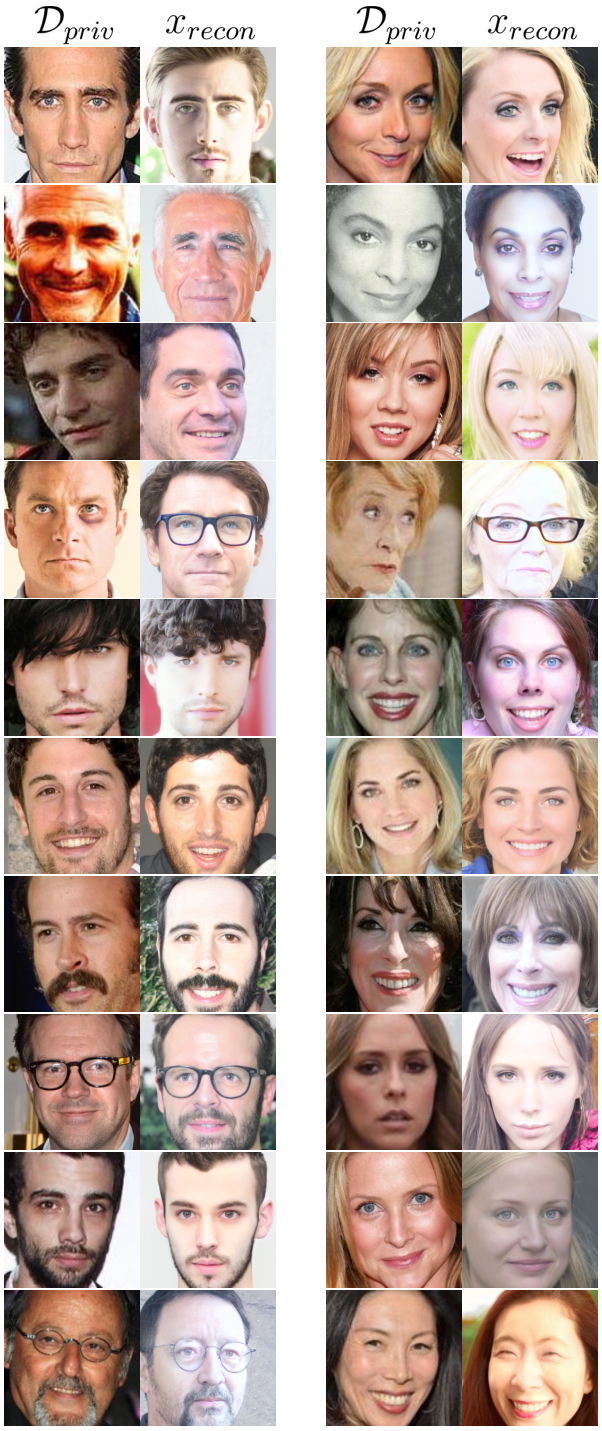



Figure S.5. Qualitative results on Facescrub dataset using the SMI-AW and $\mathcal{L}_{LOM}$, $M$ = Qwen2.5-VL. For each pair, the left column shows images from the private training dataset, while the right column presents the reconstructed images corresponding to each individual in the left column.

proaches to improve MI robustness have also been explored in [13]. More recently, Trap-MID [15] introduces a novel defense by embedding trapdoor signals into $M$. These signals act as decoys that mislead MI attacks into reconstruct-

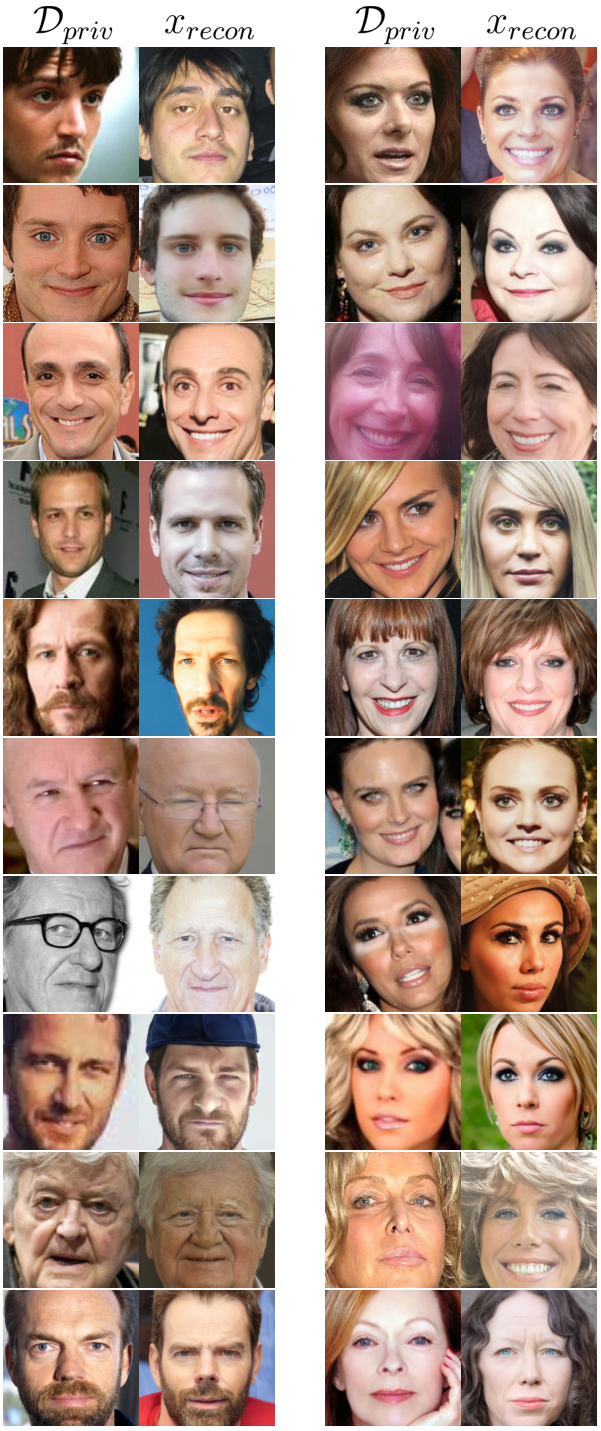



Figure S.6. Qualitative results on Facescrub dataset using the SMI-AW and $\mathcal{L}_{LOM}$, $M$ = InternVL2.5. For each pair, the left column shows images from the private training dataset, while the right column presents the reconstructed images corresponding to each individual in the left column.

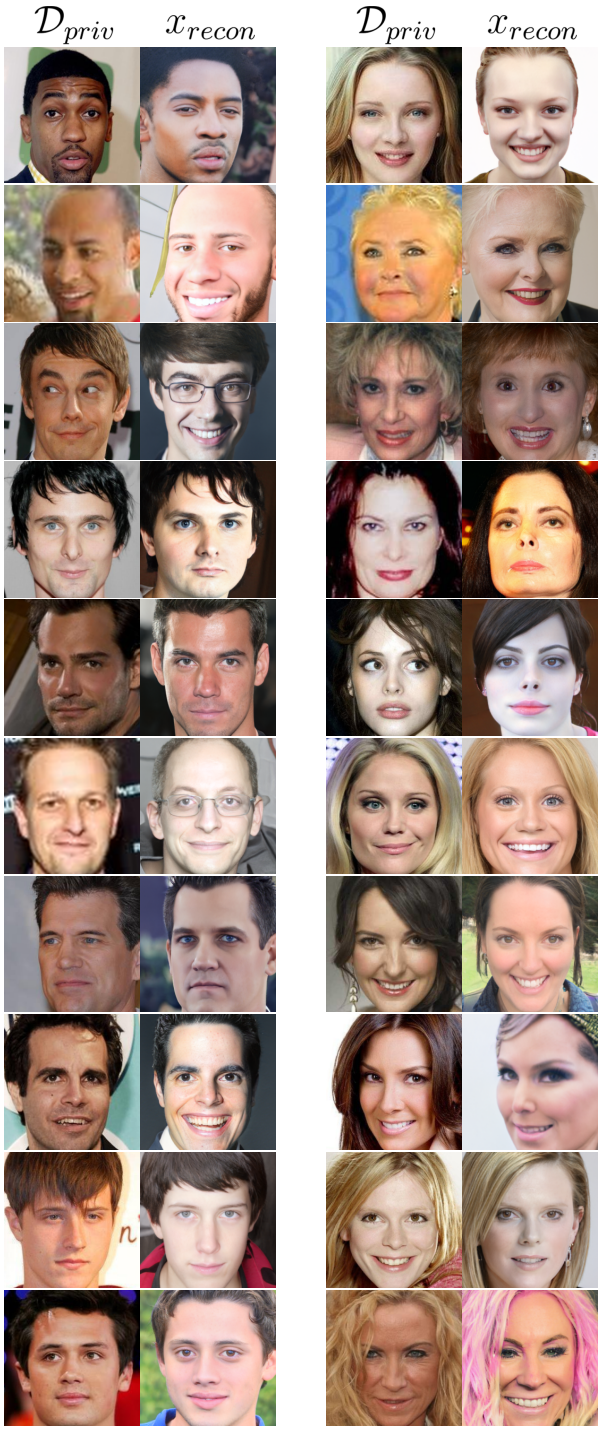



Figure S.7. Qualitative results on CelebA dataset using the SMI-AW and $\mathcal{L}_{LOM}$, $M$ = LLaVA-v1.6-7B. For each pair, the left column shows images from the private training dataset, while the right column presents the reconstructed images corresponding to each individual in the left column.

ing trapdoor triggers instead of actual private data.

**Model Inversion in Multimodal Large Vision-Language Models.** Large Vision-Language Models (VLMs) are increasingly deployed in many real-world ap-

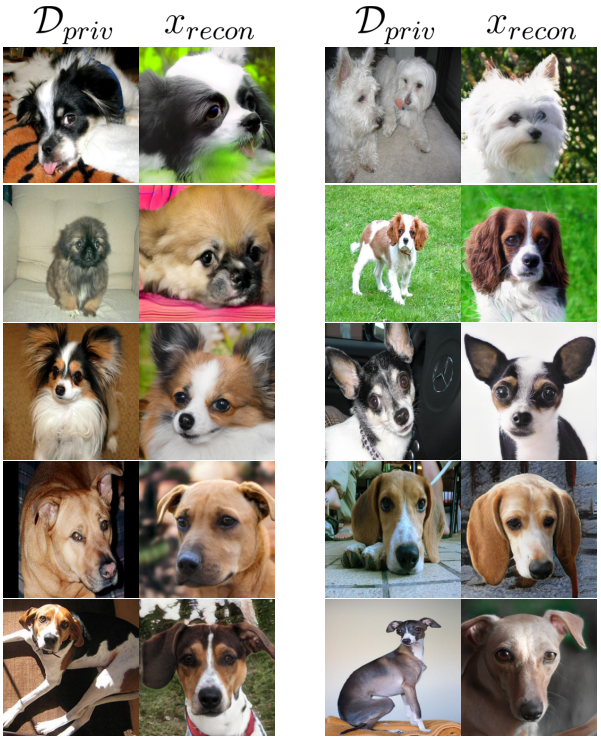



Figure S.8. Qualitative results on the Stanford Dogs dataset using the SMI-AW and $\mathcal{L}_{LOM}$, $M$ = LLaVA-v1.6-7B. For each pair, the left column shows images from the private training dataset, while the right column presents the reconstructed images corresponding to each dog breed in the left column.

plications across diverse domains, including sensitive areas [3, 4, 6, 7, 14]. Unlike unimodal vision models, VLMs are designed to process both image and text inputs and generate text responses. A typical VLM architecture includes a text tokenizer to encode textual inputs into text tokens, a vision encoder to extract image features as image tokens, and a lightweight projection layer that maps image tokens into the text token space. These tokens are then concatenated and passed through a LLM to produce the final response. This multimodal processing pipeline fundamentally distinguishes VLMs from traditional unimodal vision models.

As VLMs are being adopted more widely, including in privacy-sensitive scenarios, understanding their potential vulnerability to data leakage via MI attacks becomes critical. **However, while MI attacks have been extensively studied in unimodal vision models, to the best of our knowledge, there has been no prior work investigating MI attacks on multimodal VLMs. To fill this gap, *we conduct the first study on MI attacks targeting VLMs and propose a novel MI attack framework specifically tailored to the multimodal setting of VLMs.***

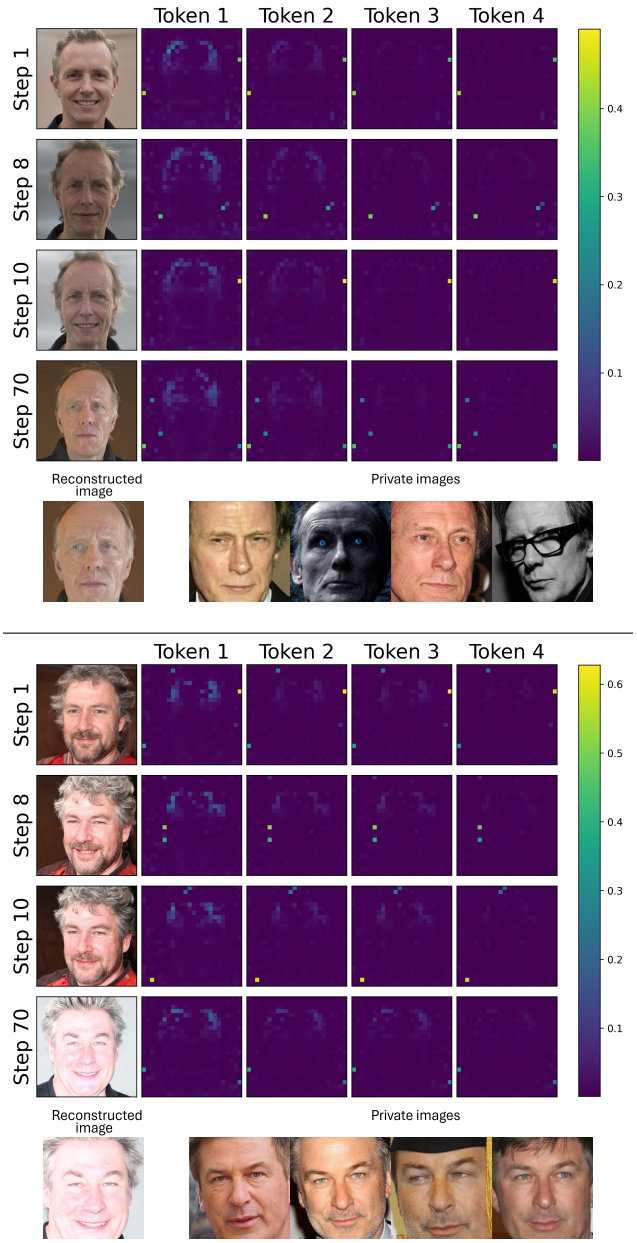
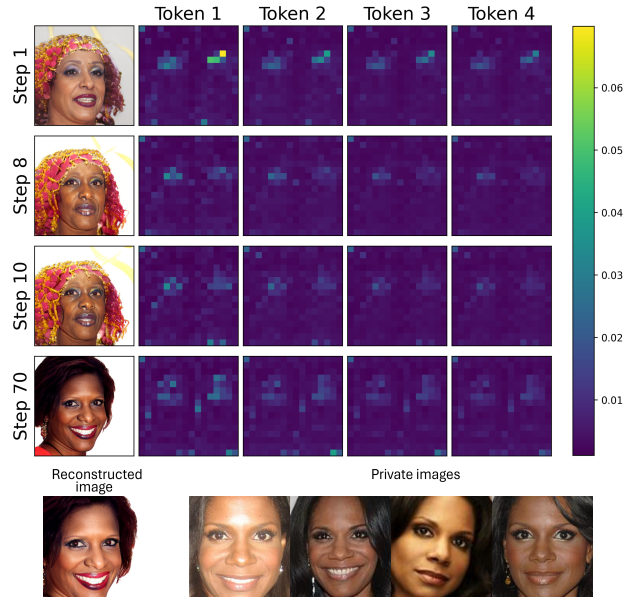



Figure S.9. **Analysis of visual–textual attention across output tokens and inversion steps of LLaVa-1.6-7B model.** We visualize the cross-attention map between the reconstructed image and each output token during inversion. **Our analysis confirms that token-level gradients vary substantially in visual informativeness both across tokens and over time, and this motivates our SMI-AW method with dynamic reweighing.**

## 6. Discussion

### 6.1. Broader Impacts

Our work reveals, for the first time, that VLMs are vulnerable to MI attacks. As VLMs are increasingly deployed in



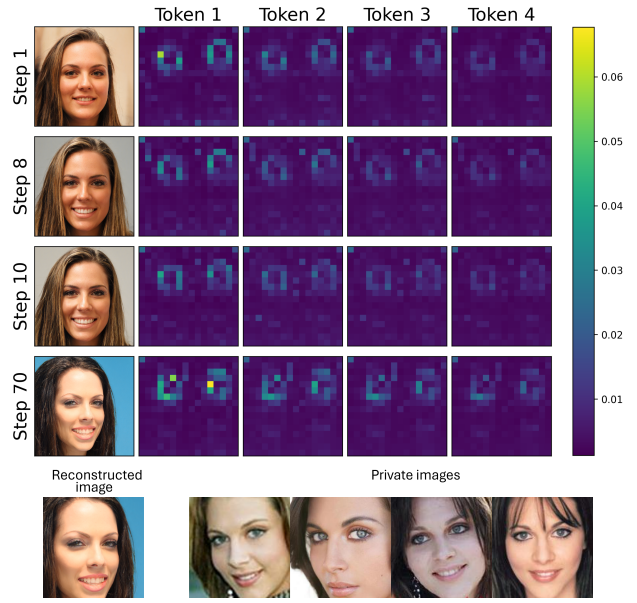
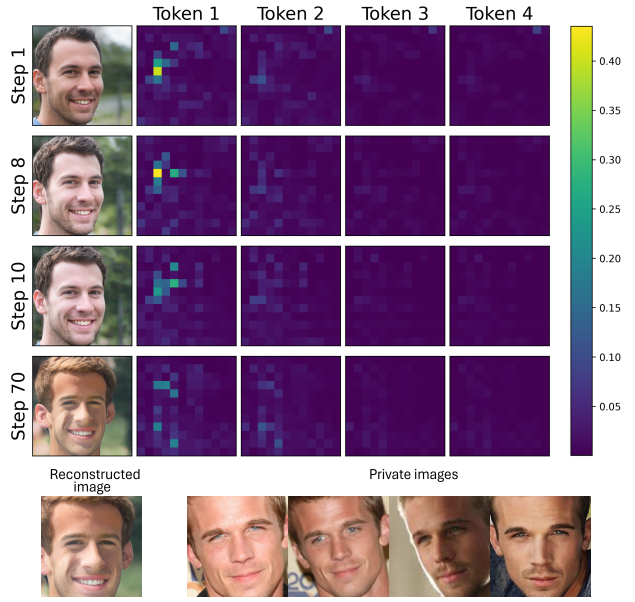




Figure S.10. **Analysis of visual–textual attention across output tokens and inversion steps of MiniGPTv2 model.** We visualize the cross-attention map between the reconstructed image and each output token during inversion. **Our analysis confirms that token-level gradients vary substantially in visual informativeness both across tokens and over time, and this motivates our SMI-AW method with dynamic reweighing.**



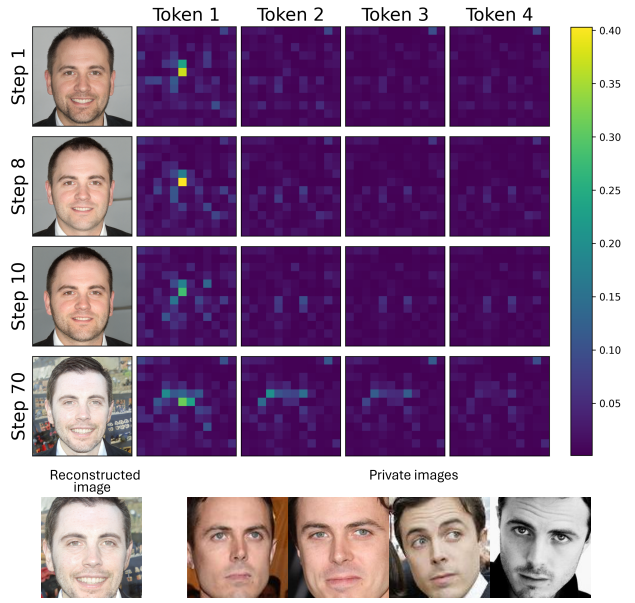
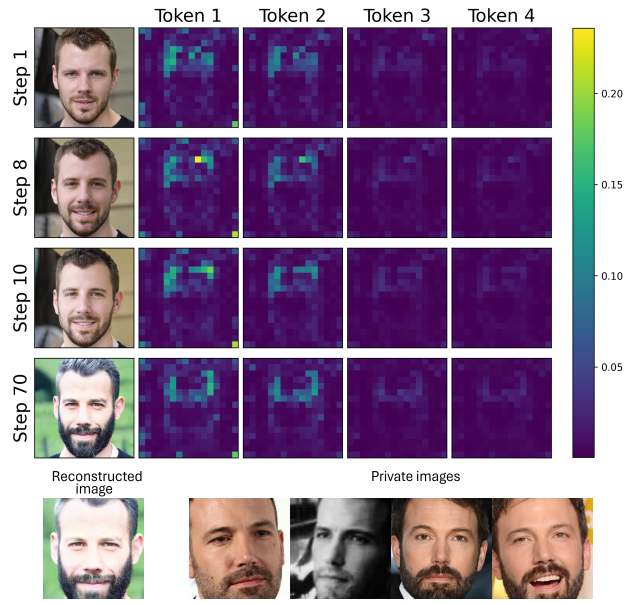




Figure S.11. **Analysis of visual–textual attention across output tokens and inversion steps of Qwen2.5-VL-7B model.** We visualize the cross-attention map between the reconstructed image and each output token during inversion. **Our analysis confirms that token-level gradients vary substantially in visual informativeness both across tokens and over time, and this motivates our SMI-AW method with dynamic reweighing.**

many applications including sensitive domains, this poses serious privacy risks. Although our work focuses on developing a new MI attack for VLMs, we also provide a fundamental understanding for the development of MI defenses in multimodal systems. We hope this work encourages the community to incorporate privacy audits in VLM deployment and to pursue principled model design that mitigates data leakage.

Our methods are intended solely for research and defense development. We strongly discourage misuse and emphasize responsible disclosure when evaluating model vulnerabilities.



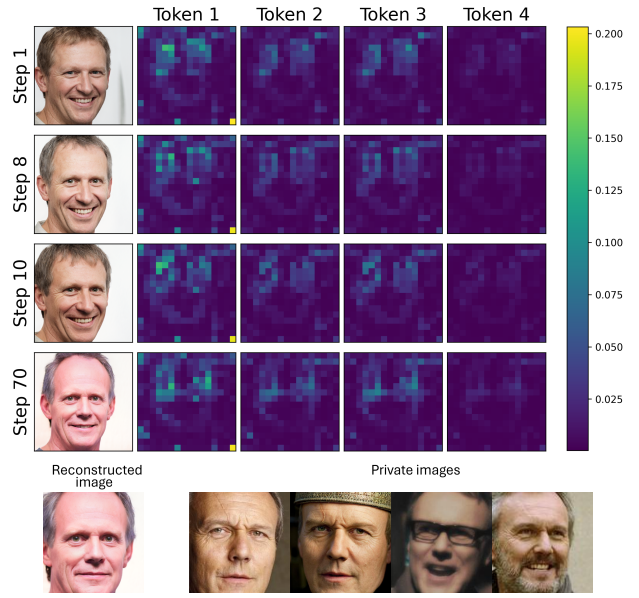




Figure S.12. **Analysis of visual–textual attention across output tokens and inversion steps of InternVL2.5 model.** We visualize the cross-attention map between the reconstructed image and each output token during inversion. **Our analysis confirms that token-level gradients vary substantially in visual informativeness both across tokens and over time, and this motivates our SMI-AW method with dynamic reweighing.**

## 6.2. Limitations

While following conventional MI attacks to focus on facial images and dog breeds, a more diverse domain scenarios, such as natural scenes or medical images, remain an important direction for future research. Moreover, evaluations with more models can further support our claims.

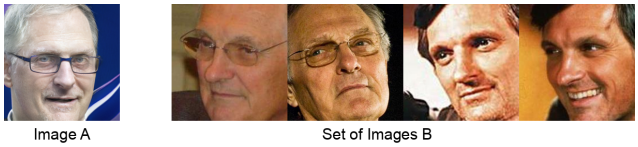



Figure S.13. An example evaluation query in $\mathcal{F}_{MLLM}$ and human evaluation involves determining whether "Image A" depicts the same individual as those in "Image B." "Image A" is a reconstructed image of a target textual answer $y$, while "Image B" contains four real images of the same target textual answer $y$. Gemini or human evaluators respond with "Yes" or "No" to indicate whether "Image A" matches the identity shown in "Image B."

## 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. 3

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017. 4

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 8

[4] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechu Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 1, 8

[5] 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. 3, 4

[6] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 1, 8

[7] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 1, 8

[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. 4

[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. 4

[10] 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

[11] Sy-Tuyen Ho, Koh Jun Hao, Ngoc-Bao Nguyen, Alexander Binder, and Ngai-Man Cheung. Revisiting model inversion evaluation: From misleading standards to reliable privacy assessment, 2025. 1, 3

[12] 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. 4

[13] 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. 6

[14] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 1, 8

[15] ZhenTing Liu and ShangTse Chen. Trap-mid: Trapdoor-based defense against model inversion attacks. *Advances in Neural Information Processing Systems*, 37:88486–88526, 2024. 6

[16] 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*, pages 16384–16393, 2023. 1, 3, 4

[17] Xiong Peng, Feng Liu, Jingfeng Zhang, Long Lan, Junjie Ye, Tongliang Liu, and Bo Han. Bilateral dependency optimization: Defending against model-inversion attacks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1358–1367, 2022. 5

[18] Xiong Peng, Bo Han, Feng Liu, Tongliang Liu, and Mingyuan Zhou. Pseudo-private data guided model inversion attacks. *Advances in Neural Information Processing Systems*, 37:33338–33375, 2024. 4

[19] 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. *Proceedings of European Conference on Computer Vision*, 2024. 3, 4

[20] 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. 4

[21] 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. 3, 4

[22] 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. 3, 5

[23] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 3

[24] 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. 4

[25] 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. 5

[26] 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. 4

[27] Xiaojian Yuan, Kejiang Chen, Jie Zhang, Weiming Zhang, Nenghai Yu, and Yang Zhang. Pseudo label-guided model inversion attack via conditional generative adversarial network. *AAAI 2023*, 2023. 3, 4

[28] 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. 3, 4