Multi-modal Identity Extraction

Ryan Webster   Teddy Furon
Inria/IRISA
Centre Inria de l’Université de Rennes, France
ryan.webster@inria.fr

Abstract

The success of multi-modal foundational models is partly attributed to their diverse, billions-scale training data. By nature, web data contains human faces and descriptions of individuals. Thus, these models pose potentially widespread privacy issues. Recently, identity membership inference attacks (IMIAs) against the CLIP model showed that membership of an individual’s name and image within training data can be reliably inferred.

This work formalizes the problem of identity extraction, wherein an attacker can reliably extract the names of individuals given their images only. We provide the following contributions (i) we adapt a previous IMIA to the problem of selecting the correct name among a large set and show that the method scales to millions of names (ii) we design an attack that outperforms the adapted baseline (iii) we show that an attacker can extract names via optimization only. To demonstrate the interest of our framework, we show how identity extraction can be used to audit model privacy. Indeed, a family of prominent models that advertise blurring faces before training to protect privacy is still highly vulnerable to attack.

1. Introduction

Multi-modal models have become a crucial component in many advanced systems due to their versatility, performance, and availability as open-sourced models. Likewise, even bigger models such as GPT-4o or Claude Sonnet are accessible with prices of around a few cents per image/prompt query. These models are trained on a massive amount of data scraped from the Internet. This raises privacy concerns. Indeed, *publicly available* does not imply *non-private* [36]. Individuals often post data online that is implicitly not intended for any usage.

Major AI service providers are aware of this privacy concern. Indeed, most of them prohibit the ability to retrieve someone’s name from a photo. For instance, Google’s reverse image search now usually returns “*image results for people are limited*” when queried with a human face picture. Major providers like Anthropic or OpenAI expressly prohibit people identification and assert their vision-language models will refuse to do so. One can read in the Anthropic Usage Policy [2]: “*Do Not Compromise Someone’s Privacy or Identity ... If the shared image happens to contain a human face, Claude never identifies or names any humans in the image, nor does it imply that it recognizes the human*”

Likewise, many recent academic works advertise privacy as a top priority when releasing multi-modal datasets or models, such as explicitly blurring faces [16], removing personally identifiable information [41], or employing differentially private pre-training [31]. However, less work has been done to audit the model privacy post hoc to verify privacy claims. D. Hintersdorf *et al*. propose a novel attack for CLIP-based vision-language models to assess their privacy risk by answering the question “Does CLIP know my face?” [17]. They introduce **Identity Membership Inference (IMI)**, which is the problem of determining whether a face/name pair was used for training. However, this attack depends on the possession of the ground truth name for each person’s facial image. This work aims at broadening the efficiency of the privacy audit by asking the more challenging questions:

How many faces does CLIP know?
Who are these people?

To answer these questions, we introduce **Identity Extraction (IE)**, which is the problem of *extracting* the ground truth name given only facial images. In other words, our attack doesn’t just reveal membership information; it yields a person’s previously unknown identity. We summarize our outline and our contributions below.

Contributions

- Section 3 formalizes the problem of Identity Extraction that is solved in two ways in Sect. 4: Our Extraction by Selection Attack (IESA) extracts correct names from a large set of candidate names, whereas our Extraction by Generation (IEGA) directly generates the candidate set.

- Section 5 details a more challenging evaluation set than [17], with more than 10x identities and significantly more diversity in the queries. As a minor contribution, we also propose a more query-efficient IMIA attack to cope with this hard setting.
- Section 6 shows that IESA can extract roughly a third of the identities in our query set with extremely high precision. When IESA is used in tandem with IEGA, names can be extracted directly from CLIP, or VLMs like GPT4o, without access to a candidate name set.
- Finally, in Section 6.3, we show the power of IE as an auditing tool. Datacomp made privacy a top priority during its design through facial blurring. The family of models trained on Datacomp are still vulnerable to our attack. We conclude that the blurring was probably too weak.

2. Related work

The CLIP Model The CLIP network [19, 29, 35] is a model that scores image/text pairs on their relevance to one another. CLIP is amongst the first foundational models, trained on a large corpus of public web data. CLIP's general knowledge has been integral to creating other web-datasets [4, 15, 32], or as a direct component of other multi-modal systems [26, 30, 37]. Foundational models currently face legal and ethical issues around the use of public web data beyond merely privacy concerns [12, 14, 36]. Regardless of guidelines, data providers lack sound training data proofs, in particular those which minimize the chance of making a false accusation (i.e., a false positive [42]).

Membership Inference Membership Inference (MI) is the problem of discerning which individual datum were used to train a model post-hoc [7, 33]. Recent studies have attempted MI versus foundational vision models [13, 22]. Normally, MI is studied "in the lab," where a researcher controls the training set. For foundational models, retraining is not worthwhile and if the training set is known, a set of non-member samples must be constructed. Recent works propose collecting new data with later timestamps, or using similar datasets collected at the same time [13, 22]. [11] remarks this hinders the trustworthiness of the attack due to distribution shift and [42] remarks MI may not be feasible when the training set is not known, as an attacker cannot guarantee a false accusation will not be made [42]. This works deviates from the standard MI setup, in which we shift from sample level to identity level inference.

Extraction Attacks Extraction attacks are a special case of (MI), where an attacker fully reconstructs a training sample [6, 8, 28, 38]. In [6], large excerpts of training text are extracted verbatim from a language model and Nasr *et al.* succeed to do so even on aligned models like ChatGPT [28]. In [6] pixel-for-pixel copies of training data can be extracted from image generators. The prompts which trigger extraction are short and non-de-script, which suggests successful extraction implies a low false positive rate [42].

Identity Membership Inference Identity membership inference (IMI) seeks to infer membership of identities, rather than individual datum. Identity data typically composes a set of samples that share identifying information specific to an individual, e.g. faces, names, pseudonyms, or other personal information. If the first attacks studied unconditional models trained on image or audio data [9, 39], recent work studies conditional models such as CLIP or its audio counterpart [10, 17, 25]. Notably, IMI against CLIP uses the assumption that facial images share little a-priori information with names and in particular, a model not trained on an identity should associate names and faces randomly (up to "cultural similarity [17]). This does allow the attacker to have some level of confidence that they will not make a false accusation, even without access to ground truth membership information [17, 42].

3. Problem Formulation

Notations Let us denote by z a given identity, element of a set $\mathcal{Z}$. This identity has the name $n_z \in \mathcal{N}$. We denote images by X and text (caption) by C. The set $\mathcal{X}_z$ denotes all images that depict the identity z and $\mathcal{C}_z$ as captions including the name n_z. We denote by X_z an image randomly picked in $\mathcal{X}_z$ and by x_z a particular image of this set. Superscript A denotes the sets in the hand of the attacker (e.g. $\mathcal{Z}^A$) and superscript T the sets used to train the model under scrutiny (e.g. $\mathcal{Z}^T$).

Attack Assumptions The algorithms in this paper are called *attacks*, and the person running these algorithms *the attacker*. The model under attack is denoted as M. We assume a black-box scenario for all settings, i.e., the attacker can only query model M. Specifically, if the attacker queries M, he receives the scalar similarity "CLIP" score.

Identity Membership An Identity Membership Inference Attack (IMIA) aims to determine whether a person's data was used to train a model M. It is a hypothesis test formulated about identity z. The null hypothesis $\mathcal{H}_0$ is that no data related to z was used to train M. The alternative hypothesis $\mathcal{H}_1$ is that the training used some data about individual z, like pairs of image/caption (X_z, C_z). These are denoted by capital letters to outline that we do not know which images and captions exactly are part of the training set. An IMIA is thus a generalization of the standard Membership Inference Attack (MIA). A MIA is a test about the presence of a specific piece of data (x_z, c_z) in the training set. An IMIA involves other samples $\{(X_{z,i}, C_{z,i})\}_i$ associated to that individual z. The adversary has his own set $\mathcal{Z}^A$, which may share identities with the set $\mathcal{Z}^T$ of the identities whose data were used during the training.

Identity Extraction In the Identity Extraction (IE) problem, the attacker has a set of images $\mathcal{X}^A$ which is the union of sets $\mathcal{X}_z^A$. Each image $X \in \mathcal{X}_z^A$ is associated to a unique entity z but the attacker does not know her name. One possibility, so-called Extraction by Selection, is to run an IMIA on each set $\mathcal{X}_z^A$ with name set $\mathcal{N}^A$ with the hope that $n_z \in \mathcal{N}^A$. Another possibility, so-called Extraction by Generation, is to make a name generator successfully guess the correct name n_z. The key difference with an IMIA is that a successful IE attack finally identifies z by its previously unknown name.

4. Attack Details

This section first reviews the IMIA[1] versus a CLIP model presented in [17]. Then, it explains our new IE attacks.

4.1. Identity Membership against CLIP

Given identity z, the attacker first gathers in the set $\mathcal{X}_z^A$ pictures depicting this person. He also has K templates $\{\mathsf{C}^k\}_{k=1}^K$ to derive captions from a name. This results in the caption $C_z = \mathsf{C}^k(n_z)$ in the style of 'A photo of John Doe', 'John doe in a suit', ... The attacker submits a correct face / text pair $(X_z, \mathsf{C}^k(n_z))$ together with distractor non-matching pairs $(X_z, \mathsf{C}^k(n_{z'}))$, with $n_{z'} \in \mathcal{N}^A$. The key assumption is that the correct pair will return a high score if $z \in \mathcal{Z}^T$ and a low score otherwise. The attack then identifies the person depicted in X_z by taking the maximum CLIP score (1). This prediction is per image and caption template. It is strengthened by a majority vote over multiple images from $\mathcal{X}_z^A$, yielding an identification $\hat{z}\,(k)$ per caption template (2). The probability of a successful identification is estimated as the frequency at which the correct identity is chosen over multiple caption templates (3).
$\forall X_z \in \mathcal{X}_z^A, \forall k, 1 \leq k \leq K$:

$$\hat{n}_z(X_z, k) = \arg \max_{n \in \mathcal{N}^A \cup \{n_z\}} M\left(X_z, \mathsf{C}^k(n)\right), \quad (1)$$

$$\hat{n}_z(k) = \mathsf{MajorityVote}_{X \in \mathcal{X}_z^A}\left(\hat{n}_z(X, k)\right), \quad (2)$$

$$f_z = \mathsf{Average}_{1 \leq k \leq K}\left(\mathbb{1}[\hat{n}_z(k) = n_z]\right). \quad (3)$$

Identity z is finally deemed as a training identity if $f_z > \tau$, where τ is the threshold chosen for the attack. The number of caption templates K and the number of distractors are fixed beforehand, however, the number of images per identity $|\mathcal{X}_z^A|$ may vary.

Under hypothesis $\mathcal{H}_0$, entity z is not part of the training, so that the guessed name $\hat{n}_z(k)$ (2) is random and, on expectation, $f_z = 1/|\mathcal{N}^A|$. This statistical model holds if $\mathcal{N}^A$ gathers identities "*that are culturally similar to*" z to avoid biases in the distribution of $\hat{n}_z(k)$ under $\mathcal{H}_0$ [17].

4.2. Identity Extraction against VLMs

In an Identity Extraction attack, the attacker has a set of images and names, but doesn't know their association, if any. We consider two settings: in Extraction by Selection, the attacker has a candidate name set containing some overlap with the ground truth (e.g. gathered from common names). Then, we describe Extraction by Generation, wherein the attacker first generates a candidate name set using the visual information in $\mathcal{X}_z^A$, and then does Extraction by Selection.

4.2.1. Identity Extraction by Selection Attack- IESA

As the attacker no longer knows the true name n_z, he cannot exploit this information to compute (1) and (3). We thus resort to a frequentist attack: the attacker computes the frequency (3) for any suspect and identifies the person depicted in $\mathcal{X}_z^A$ as the most frequent one, if confident enough:

$$\hat{z} = \begin{cases} \hat{z}^\star := \arg\max_{z' \in \mathcal{Z}^A} f_{z'} & \text{if } f_{\hat{z}^\star} > \tau, \\ \emptyset & \text{otherwise.} \end{cases} \quad (4)$$

We consider this as our baseline attack. In our initial experiments, successful IESA attacks typically have one dominant frequency corresponding to the ground truth name, and a long tail of small frequencies for the other names. Incorrect identifications are due to high values for several competing names, typically those sharing first or last names. To adjust for this, we propose an adaptive variant taking into account the second largest frequency. The attack is confident only if the gap between the max and the second max is large enough:

$$\hat{z} = \begin{cases} \hat{z}^\star & \text{if } \frac{f_{\hat{z}^\star}}{\max_{z' \in \mathcal{Z}^A \setminus \{z^\star\}} f_{z'}} > \tau, \\ \emptyset & \text{otherwise.} \end{cases} \quad (5)$$

The success of this attack shows that a CLIP model can play the role of a Face Recognition System not only for verification but also for identification in a closed-set scenario.

4.2.2. Identity Extraction by Generation Attack - IEGA

This setting explores whether it is possible to generate a set of names with the visual information in $\mathcal{X}_z^A$, guided by the CLIP score. We consider two methods for generating a candidate set of names: query a pre-trained captioning model (VLM) or via optimizing a large language model (LLM) on the CLIP score. For the VLM approach, the attacker queries the VLM with $\mathcal{X}_z^A$ and extracts any names from the returned captions. Name generation in this way assumes the VLM itself has knowledge of the person; in the case when the VLM does not recognize the person, the method is susceptible to hallucination. Thus, the attacker must perform IESA on the generated name set to reject hallucinated names. The same verification holds for the second approach described next.

[1] Referred to as IdIA in [17].

Optimize LLM with REINFORCE We search a language model $P_{\theta^\star(z)}$ that can produce captions suitable for all the images depicting individual z:

$$\theta^*(z) = \arg\max_{\theta \in \Theta} \mathbb{E}_{C \sim P_\theta} \left[\sum_{X \in \mathcal{X}_z^A} M(X, C) \right] \tag{6}$$

Solving Eq. (6) presents several challenges. First, image captioning VLMs typically inject CLIP features directly into a language model [23, 24, 26]. We retain a black-box setting throughout and thus do not consider access to CLIP features. Furthermore, we do not have access to the CLIP score gradients, nor the ground truth captions. Finally, we consider an attacker with modest resources who utilizes only the set of images within $\mathcal{X}_z^A$ to extract the name n_z, rather than a large multi-modal dataset typically used for training. For this, we propose using the REINFORCE algorithm [40], which has recently been revisited in the context of language model fine-tuning [3]. We fine-tune a language model to maximize the expectation of the CLIP score for a single image by estimating its gradient with REINFORCE. We initialize θ with an open source model like Mistral-7B [20]. At each training iteration, the language model generates a batch of captions whose CLIP scores with the images in $\mathcal{X}_z^A$ are computed and summed as in (6). The REINFORCE algorithm estimates the gradient that updates the model parameters θ. We also propose pushing generated captions away from other identities by also adding a negative CLIP score for a small set of images of other identities $\{\mathcal{X}_{z'}\}_{z'}$. We found this to work significantly better in practice for identity extraction. A full outline of the algorithm can be found in the supplementary.

5. Evaluation protocol and datasets

This section explains the constitution of the ground truth and the auxiliary set (of identities, face images, and names) from the following public datasets.

5.1. Datasets

LAION-2B (L2B) and DataComp (DC) L2B was amongst the first public billion-scale text and image datasets, containing roughly two billion text/image pairs. It has been used to train a variety of generative models, CLIP models and VLMs [15, 24, 34]. More recently, DataComp is a dataset of image and text pairs with sizes varying from 1B to over 10B [15]. DataComp carefully filtered subsets of larger pools of image/text samples which yielded more performant CLIP models, even with no improvements on the architecture side. Relevant to this work, both the curation and training of the DataComp models were done with face blurring to protect privacy.

The Attacker Set - VGGFace2 For both attacks, the attacker needs a set of faces and names. The work [17] uses the FaceScrub dataset [1] containing roughly 500 identities. Not only is this an extremely small number of identities compared to L2B, but also FaceScrub people are largely white, American celebrities. However, biases in terms of race and gender have a strong effect on the behavior of multi-modal models [27]. We thus seek a more diverse attacker set for a comprehensive privacy analysis.

We analyzed the race and sex distributions of several datasets containing ground truth names using FairFace [21]. We chose VGGFace2, which is a dataset with roughly eight thousand people, including hundreds of diverse samples across age and pose [5]. We found that VGGFace2 had roughly a quarter of non-white individuals, yielding several thousand non-white individuals, with a roughly equal sex distribution. On the other hand, we found that the Facescrub dataset used in [17] contains only 46 non-white individuals. We use VGGFace2 (VGGF2) as our attacker set in all experiments. We detail more analysis of the individuals present in VGGF2 as well as some post-processing of the data set we did in the supplementary.

English Wikipedia Names Our protocol requires distractor names. In [17], they are gathered from common names in census data in [17]. We propose using the set of all names from Wikipedia, to form a more extensive and more diverse set. We collect all Wikipedia pages under the category "Living People" and use the title as the name. We do some processing on these names, including removing names with specifiers when the name has been disambiguated. Doing so results in roughly a million names. As this is the distractor set, we also remove names shared with VGGFace2.

OpenCLIP All models are from the OpenCLIP [19] repository. We abbreviate the models via their architecture and their training set after an underscore, e.g. ViT-H-14_L2B for LAION-2B or ViT-H-14_DC for DataComp.

5.2. Constitution of the ground truth

In the experimental protocol of [17], the ground truth set of members comprises names from FaceScrub that also appear within the captions of the LAION-400M dataset. We extend this process by looking for names of VGGFace2 in the LAION-2B dataset. Names are first normalized to only contain lowercase characters that can be encoded in ASCII, which removes accents and converts non-English characters. Then, we check whether each name is a substring within normalized captions of LAION-2B. We enforce that the name is preceded and followed by spaces or punctuation, i.e. the name is not a substring within a word. Finally, we consider names occurring more than 10 times to be positive labels for membership inference.

For Identity Membership Inference (Sect. 4.1), the names of VGGFace2 constitute the attacker's set $\mathcal{Z}^A$ partitioned into $|\mathcal{Z}^M| = 7k$ members (i.e. present in LAION-2B as explained above) and $|\mathcal{Z}^{NM}| = 1k$ non-members. This partition of $\mathcal{Z}^A$ is used as the ground truth to measure the true positive and false positive rates of IMIA (Sect. 4.1). Note that in [17], only 8 non-members were used for the experiments involving LAION-400M. The set $\mathcal{X}^A$ corresponds to the VGGFace2 face images of the individuals in $\mathcal{Z}^A$. In order to measure low false positive rates for the Identity Extraction problem (Sect. 4.2), we need to include more non-members. The name set $\mathcal{N}^A$ for the Extraction by Selection (Sect. 4.2.1) is the same as above but augmented by all identities from the English Wikipedia Names (Sect. 5.1).

5.3. Metrics

The IMIA of Sect. 4.1 returns a binary output indicating whether identity z is deemed as a training identity. This result is compared to the ground truth member/non-member sets defined in Sect. 5.2. This is done over all the identities of VGGFace2 to estimate the true positive (under $\mathcal{H}_1$, $z \in \mathcal{Z}^M$) and false positive (under $\mathcal{H}_0$, $z \in \mathcal{Z}^{NM}$) rates.

For the IESA of Sect. 4.2.1, the output of the attack $\hat{n}_z$ is a name among a set of suspects $\mathcal{N}^A$ or nobody ($\hat{n}_z = \emptyset$). The set $\mathcal{N}^A$ is the VGGFace2 set of names with the English Wikipedia Names appended (Sect. 5.2). Under $\mathcal{H}_0$ where $z \in \mathcal{Z}^{NM}$, a false positive occurs if $\hat{n}_z \neq \emptyset$. Under $\mathcal{H}_1$ where $z \in \mathcal{Z}^M$, a true positive occurs if $\hat{n}_z = n_z$.

For the IEGA of Sect.4.2.2, the output of the attack is a generated name. Under $\mathcal{H}_1$, the attack is successful if the extracted name is nearly identical to the ground truth one. Two names are declared nearly identical if their Levenshtein edit distance is lower than ℓ modifications. For instance, Irish last names such as "O'Donnell" are sometimes generated with or without the apostrophe, which are both considered true positives under this distance. We measure how many names are extracted, versus the percentage that are correct up to edit distance $\ell = 1$.

For all attacks, we emphasize performance at low error rates. For membership inference, it is imperative that an attacker is certain of its accusation of membership, rather than making as many membership claims as possible [7]. Likewise, for extraction, it is more important to be certain extracted names are identical to the ground truth, rather than extracting as many names as possible.

6. Experimental results

6.1. IMIA Results

We evaluate IMIA under three settings. We examine the original proposal of [17] and the usage of a much larger set $\mathcal{N}^A$ of distractors in (1) taken from English Wikipedia Names. To fairly compare the attacks, we use a fixed per-identity query budget of $Q_z = 20\,000 * |\mathcal{X}^A_z|$ in all three settings, where $|\mathcal{X}^A_z|$ is the number of images for identity z in the hand of the attacker.

Clip Network	S1[17]	S2 (ours)	S3 (ours)
TPR@FPR=0.25%			
ViT-G-14_L2B	16.7	22.9	**34.6**
ViT-B-32_L2B	13.5	24.5	**25.9**
Convnext-xxlarge_L2B	21.4	23.9	**25.1**
ViT-H-14_L2B	19.6	36.0	**36.4**
TPR@FPR=1%			
ViT-G-14_L2B	55.4	60.1	**60.8**
ViT-B-32_L2B	54.5	**57.2**	52.8
Convnext-xxlarge_L2B	54.0	55.8	**59.6**
ViT-H-14_L2B	44.6	56.1	**57.9**

Table 1. IMIA results with TPR@FPR=0.25% and TPR@1% for the three settings outlined. S1 is the reproduction of [17], on the more challenging VGGFace2 attack set / LAION-2B training set.

Setting 1: (Hintersdorf) We first reproduce the exact attack parameters in [17]: $K = 20$ caption templates and a set of $|\mathcal{N}^A| = 999$ distractors. Each set of images $\mathcal{X}^A_z$ gathers all images available in VGGFace2 for identity z (on average several hundred). We then compute the CLIP score for every image, ground truth name plus distractor names and caption template combination. Thus, the total queries required per identity is $Q_z = |\mathcal{X}^A_z| \times K \times (|\mathcal{N}^A| + 1)$.

Setting 2: Larger name set (ours) We diverge from [17] by using more distractor name queries rather than caption templates. The attacker uses a single generic caption template for each name (e.g. $K = 1$) as "A photo of John Doe", and then accordingly increases the number of distractors to $|\mathcal{N}^A| = 19\,999$.

Setting 3: Full English Wikipedia Name set (ours) We extend setting 2 to use every name available in English Wikipedia Names as the distractor set. Given this set has roughly a million samples, we reduce the number of images per identity to $|\mathcal{X}^A_z| = 10$.

Comparison Table 1 shows the performance for the settings outlined above on a selection of CLIP models trained on LAION 2B [32]. Setting 3 outperforms the others for nearly every network at both .25% and 1% FPR. We believe setting 1 is more prone to false positives due to the diversity present in VGGFace2. If an identity is unique in terms of race with respect to the distractors, Eq.(3) may choose the correct name purely based on attributes. Once sufficiently diverse distractors are added, the chance of guessing just via the race or gender of the individual is reduced. To investigate this further, we tried generating names based on the race, gender, and nationality of individuals in VGGFace2.

Clip Network	$\lvert\mathcal{X}_z^A\rvert$	Extraction E@99.5 (baseline)	E@99.5 (ours)	E@95 (baseline)	E@95 (ours)
Evae-14_L2B [35]	all	0.68	**21.78**	**45.21**	44.27
ViT-B-32_L2B [19]	all	14.41	**23.24**	41.88	**41.91**
ViT-G-14_L2B [19]	all	30.58	**32.53**	**52.04**	47.42
ViT-H-14_L2B [19]	all	0.35	**0.65**	33.24	**38.55**
Convnext-xxlarge_L2B [32]	all	0.70	**15.85**	39.30	**41.79**
ViT-H-14_DC [15]	all	5.04	**16.45**	**30.22**	29.80
ViT-L-14_DC [15]	all	0.03	**1.90**	21.46	**24.92**
Evae-14_L2B [35]	10	0.03	0.03	40.63	**41.05**
ViT-B-32_L2B[19]	10	**4.74**	3.02	34.17	**34.20**
ViT-G-14_L2B [19]	10	16.92	**18.65**	**44.13**	41.68
ViT-H-14_L2B [19]	10	0.15	**0.61**	28.35	**29.96**
Convnext-xxlarge_L2B [32]	10	0.09	**1.39**	31.67	**34.07**
ViT-H-14_DC [15]	10	**6.40**	5.05	25.18	**25.33**
ViT-L-14_DC [15]	10	0.32	**0.95**	**13.53**	13.23

Table 2. Performance comparison of different models. E@99 refers to the percentage of names extracted at 99% precision (higher is better). $|\mathcal{X}_z^A|$=all means all images in VGGFace2 associated with the tested identity are used by the attacker, where $|\mathcal{X}_z^A| = 10$ only 10 are used. We compare the baseline attack (4) inspired from [17] with our proposed adaptive attack (5). The attack with superior performance for each setting is emboldened. See more details in Sec. 6.2.

6.2. IESA results

We evaluate our extraction attacks on how many samples are marked as extracted, versus the percentage of marked extractions matching the ground truth. Figure 1 shows the extraction curve for the Convnext-xxlarge_L2B model [32]. The left figure compares the results for our adaptive attack (5) versus the baseline attack (4) inspired by [17]. We consider two settings here, the first is where we use all images available in VGGFace2 per identity ($|\mathcal{X}_z^A|$ =all), which is on average several hundred unique samples per name, versus the low sample regime using only $|\mathcal{X}_z^A| = 10$ images per identity. First, the adaptive attack extracts significantly more names in the high precision regime for the $|\mathcal{X}_z^A|$ =all setting. It can extract roughly 500 names with no errors. Second, the low sample regime shows more similar results between the two attacks. This is not surprising; the adaptive attack exploits the fact that the histogram of CLIP selections for extracted names tends to have low entropy and it is difficult to estimate this histogram with few face image samples. Figure 1 (right) shows the adaptive attack in the $|\mathcal{X}_z^A|$=all setting under various relaxed ground truths. For instance, we check whether the ground truth name exists within the top-5 extracted names (orange curve), or whether the extracted last name matches the ground truth (green curve). At 99% precision, around double the last names are extracted than if both first and last name are considered.

Table 2 evaluates extraction performance across several state-of-the-art CLIP networks. It reports the percentage of extracted identities at 99.5% precision (E@99.5) as well as 95% precision (E@95). Our attack significantly outperforms the baseline at E@99 for the majority of CLIP networks and for every model in the $|\mathcal{X}_z^A|$=all setting. We do not see a clear relationship between network size and susceptibility to extraction. For instance, the ViT-H-14_L2B and Evae-14_L2B models [19] are the largest and most performant models and yet give worse E@99.5 than the smallest ViT-B-32_L2B model. However, the ViT-G-14_L2B model ranks amongst the top performing models on OpenCLIP and is all around the most susceptible to the attack, with $\approx$ 33% extraction (i.e. about three thousand names) with less than .5% error.

6.3. IEGA results

First we detail how names are generated with either REINFORCE or a VLM, and then analyze the results after performing an IESA.

Training Details for REINFORCE We implement the attack described in Sec. 4.2.2 by adapting the language model Mistral-7B [20] with LORA fine-tuning [18] for each tested identity. We use 50 images for each identity and 10 images from 100 other identities randomly selected for the distractor identities. We sample the language model with a relatively high temperature of $T = 1.25$ to promote diversity in generation, enforce generations of 24 output tokens and run 50 iterations of training.

VLM Extraction Our second extraction by generation attack queries several VLMs via an API. Some of them, such as GPT-4o, Gemini-1.5-Pro and Pixtral-12B reject naive at-

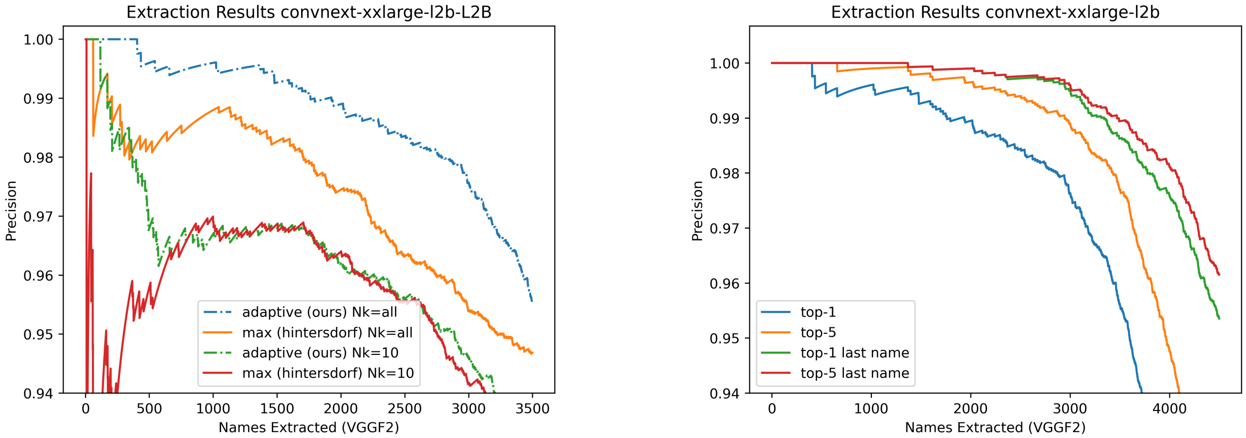

Figure 1. Precision of the attack vs. the number of identities extracted from Convnext-xxlarge_L2B. Left: Our selection attack (5) vs. the baseline attack (4) inspired from [17]. Right: All curves are our adaptive attack, with various relaxed ground truths, including only extracting last names.

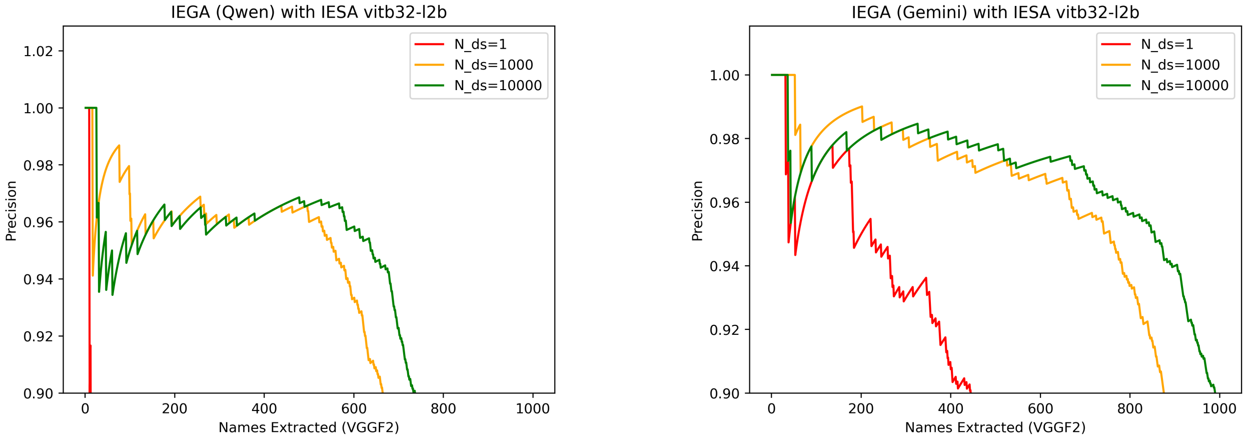

Figure 2. Importance of IESA for IEGA generations. Here we show names generated by Qwen (left) and Gemini 1.5 (right) given a single image of the target person, followed by an IESA attack using varying amounts of distractor names. Without IESA, an attack has low precision as it cannot filter hallucinated names.

tempts to name individuals in pictures, so we designed a jailbreak prompt to mislead the model into identification. However, the prompts are generic, such as "Describe the subject of this artwork," and provide no information about the individual. We submit only one image per subject and request the VLM returns a json structured output. See the supplementary for details and examples.

Results Table 3 shows the extraction by generation results. As both attacks are expensive, we use a subset of 4 000 identities from VGGFace2. The total extraction rate is the percentage of identities where the ground truth name is found in any generation. VLMs are queried to return roughly 10 names, whereas the REINFORCE based optimization returns around 50. All IEGA attacks are followed by an IESA attack for which we append 10k additional distractor names to perform the attack. The is critical for attack precision, and we explore what happens when only the generated names are used in Fig. 2, described in the subsequent section.

Importance of IESA step in IEGA Fig. 2 demonstrates the importance of IESA for IEGA generations. Here we show names generated by two different VLMs, followed by an IESA attack using varying amounts of distractor names. Using only these names results in very low precision, as the VLM will still hallucinate names, or generate generic texts. With few names, the IESA attack is unable to distinguish

Model	Total Extraction Rate	E@95
REINFORCE		
ViT-L-14_L2B	12.68	7.53
ViT-H-14_L2B	11.48	7.45
Convnext-xxlarge_L2B	11.8	7.15
ViT-H-14_DC	9.17	5.24
ViT-B-32_L2B	11.48	7.50
Vision-Language Models (VLMs)		
GPT-4o	52.51	22.25
Qwen2-72B	22.83	13.06
Pixtral-12B	8.40	4.93
Gemini-1.5-Pro	33.48	19.34

Table 3. IEGA attack performance. Total extraction rate denotes the total ground truth names present in any generation and E95 precision with IESA.

real extractions from generated names (red curves). Interestingly, the attack continues to perform better with more distractor names appended, despite having more names to select from. The IEGA attack is the most precise with 10k distractor names (green curve).

Qualitative Analysis During the optimization with REINFORCE, the LLM consistently generates increasingly relevant captions and names along the iterations. The generation begins to include relevant professions or settings even if the guessed names are incorrect. For instance, for "John Kerry," captions at early iterations described other American politicians. At later iterations, captions typically stabilize on the relevant profession, e.g. sports, politics, journalism, etc., and differ only in the names guessed. Some Examples of training caption trajectories are shown in the supplementary.

VLM-based extraction is surprisingly efficient. For instance, GPT-4o finds the ground truth for roughly 50% of queries, despite only being provided a single image. Of course, VLMs require vastly more compute and training than our REINFORCE procedure. Still, even the relatively large open source VLM Pixtral-12B, performed worse than our REINFORCE. We didn't try to improve the VLM-based extraction as we wanted to explore a baseline. Both REINFORCE based and VLM based attacks expose simultaneously the vulnerabilities of both the CLIP model and the language model/VLM used. For REINFORCE, the language model's knowledge of the individual, and other like individuals, is necessary for refining relevant captions. Ultimately though, we are exploiting the CLIP model's vulnerability: the extraction by selection on the generated names allows the overall attack to be precise and gives the ability to sift out the unlikely generations.

Model	Raw	Crop	Bbox Blur
ViT-G-14_L2B	44.78	25.91	0.03
ViT-L-14_DC [15]	36.25	2.11	0.25
ViT-B-16-quickgelu_DC [15]	6.14	6.14	0.51

Table 4. Models trained on DataComp were trained with face blurring, and are still vulnerable to name extraction (first column). Once sufficient blurring is performed, the attacks are mostly nullified (last column).

Privacy Auditing The ViT-L-14_DC and ViT-B-16_DC models were trained with face blurring to protect privacy [15]. The procedure blurs images in a bounding box over the face [15]. Surprisingly, it is still possible to extract names with high precision from these models (see Table 2). To verify whether it is possible to extract names from CLIP using only non-facial information, we examine several post processing of VGGFace2:

- *Strict facial cropping.* The samples used for Tab. 4 are VGGFace2 images, which are a loose crop containing a small area around the head. We use a strict crop of the face only for the attacker image set.
- *Bbox blurring.* We keep the original sample and heavily blur the facial crop region.

Table 4 compares the E@95 performance for the above processing. It shows two models from [15] trained with Bbox blurring and one model trained without blurring for comparison. As expected, the attack is ineffective in the heavy blur regime. The ViT-L-14_DC model performs well on the VGGFace2 samples, but performance drops considerably for cropped samples. Given that some blurring was done during training, the model appears to exploit both facial and non-facial features. In any case, it is clear that the blurring was not sufficient to safeguard data privacy.

7. Conclusion

In this work, we provide a more comprehensive privacy audit of multi-model models, by introducing the problem of Identity Extraction. Previous work on IMI yields only training set membership, whilst IE recovers previously unknown names to an attacker. We introduce a more comprehensive attacker set including a set of names gathered from Wikipedia. We showed that CLIP models can still select the correct name amongst millions with our IESA attack, and provided an attack function outperforming a baseline. Finally, we showed that even without a candidate set containing the ground truth, an attacker can still recover names directly from the CLIP model through optimization. We believe IE is a valuable new framework to audit models and demonstrated a dataset which fell short on protecting privacy.

Acknowledgments

This work was supported by the French project Compromis, ANR 22 PECY 00XX 02.

References

[1] Facescrub. https://www.kaggle.com/datasets/rajnishe/facescrub-full. 4

[2] Anthropic usage policy, 2024. 1

[3] Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12248–12267, 2024. 4

[4] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022. 2

[5] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 4

[6] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650, 2021. 2

[7] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, pages 1897–1914. IEEE, 2022. 2, 5

[8] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023. 2

[9] Guangke Chen, Yedi Zhang, and Fu Song. Slmia-sr: Speaker-level membership inference attacks against speaker recognition systems. *ArXiv*, abs/2309.07983, 2023. 2

[10] Ruoxi Cheng, Yizhong Ding, Shuirong Cao, Zhiqiang Wang, and Shitong Shao. Gibberish is all you need for membership inference detection in contrastive language-audio pretraining. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. 2

[11] Debeshee Das, Jie Zhang, and Florian Tramèr. Blind baselines beat membership inference attacks for foundation models, 2025. 2

[12] Cerys Wyn Davies and Gill Dennis. Getty images v stability ai: the implications for uk copyright law and licensing. 2024. 2

[13] Jan Dubinski, Antoni Kowalczuk, Stanislaw Pawlak, Przemyslaw Rokita, Tomasz Trzcinski, and Pawel Morawiecki. Towards More Realistic Membership Inference Attacks on Large Diffusion Models . In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4848–4857, Los Alamitos, CA, USA, 2024. IEEE Computer Society. 2

[14] Giorgio Franceschelli and Mirco Musolesi. Copyright in generative deep learning, 2021. 2

[15] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets, 2023. 2, 4, 6, 8

[16] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems*, pages 27092–27112, 2023. 1

[17] Dominik Hintersdorf, Lukas Struppek, Manuel Brack, Felix Friedrich, Patrick Schramowski, and Kristian Kersting. Does CLIP know my face? *J. Artif. Int. Res.*, 80, 2024. 1, 2, 3, 4, 5, 6, 7

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 6

[19] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP. *Zenodo*, 2021. 2, 4, 6

[20] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, 2023. arXiv 2310.06825. 4, 6

[21] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021. 4

[22] Myeongseob Ko, Ming Jin, Chenguang Wang, and Ruoxi Jia. Practical membership inference attacks against large-scale multi-modal models: A pilot study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4871–4881, 2023. 2

[23] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36:71683–71702, 2023. 4

[24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 4

[25] Songze Li, Ruoxi Cheng, and Xiaojun Jia. Tuni: A textual unimodal detector for identity inference in clip models. *arXiv preprint arXiv:2405.14517*, 2024. 2

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2, 4

[27] Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. In *Advances in Neural Information Processing Systems*, pages 56338–56351. Curran Associates, Inc., 2023. 4

[28] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models, 2023. arXiv 2311.17035. 2

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 2

[31] Tom Sander, Yaodong Yu, Maziar Sanjabi, Alain Durmus, Yi Ma, Kamalika Chaudhuri, and Chuan Guo. Differentially private representation learning via image captioning, 2024. 1

[32] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 2, 5, 6

[33] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017. 2

[34] StabilityAI. Stable diffusion 2.0. `https://github.com/Stability-AI/stablediffusion`, 2022. 4

[35] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: Improved training techniques for CLIP at scale, 2023. arXiv 2303.15389. 2, 6

[36] Florian Tramèr, Gautam Kamath, and Nicholas Carlini. Position: Considerations for differentially private learning with large-scale public pretraining. In *International Conference on Machine Learning*, pages 48453–48467. PMLR, 2024. 1, 2

[37] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2

[38] Ryan Webster. A reproducible extraction of training images from diffusion models. 2023. arXiv 2305.08694. 2

[39] Ryan Webster, Julien Rabin, Loic Simon, and Frederic Jurie. This person (probably) exists. Identity Membership Attacks against GAN generated faces, 2021. arXiv 2107.06018. 2

[40] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256, 1992. 4

[41] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 1

[42] Jie Zhang, Debeshee Das, Gautam Kamath, and Florian Tramer. Position: Membership Inference Attacks Cannot Prove That a Model was Trained on Your Data . In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 333–345, Los Alamitos, CA, USA, 2025. IEEE Computer Society. 2