Rethinking Membership Inference Attacks for CLIP

Lluis Gomez

Computer Vision Center – Universitat Autònoma de Barcelona
lgomez@cvc.uab.cat

Abstract

Membership Inference Attacks (MIAs) test whether a model has memorized training data, and are a key tool for auditing privacy risks in machine learning. Recent papers report near-perfect MIA success against large vision-language models such as CLIP, but almost all evaluations train on one web-scale corpus (e.g. LAION-400M) and treat samples from a different corpus (e.g. COCO or CC12M) as non-members – thereby turning the task into out-of-distribution (OOD) detection rather than true membership testing, introducing spurious signals unrelated to true memorization.

We revisit the problem with a distribution-matched benchmark built from the CommonPool-L corpus of DataComp. A ViT-B/16 CLIP trained on 400 M pairs is accompanied by two 26-shard, i.i.d. splits that serve as member and non-member sets, sharing the exact same acquisition and preprocessing pipeline. Under this strictly in-distribution setting, every published MIA baseline collapses to chance (≈51% AUC). To explain this collapse, we derive a scaling-law upper bound for similarity-based attacks showing that the expected member vs. non-member similarity gap decays as $\mathcal{O}(T/N)$ for contrastive learning with T epochs over N samples. Empirically, as we vary the training set size while holding all hyperparameters fixed, the gap follows the predicted linear trend in log–log space, and Cosine Similarity Attack AUC drops from 94% to 51%. Finally, we propose a simple, white-box, gradient-based MIA that outperforms prior attacks for CLIP without relying on OOD cues. We release code, checkpoints, and data to foster comprehensive and reproducible privacy research on multimodal CLIP-like foundation models.

Code, Models, and Data — https://clipmiabench.github.io

Introduction

Membership inference attacks (MIAs) aim to determine whether a given data sample was part of a model's training set. While MIAs have been extensively investigated for conventional classifiers (Shokri et al. 2017; Carlini et al. 2022), their application to large-scale, multi-modal architectures like CLIP (Radford et al. 2021) remains relatively underexplored. CLIP models, which align image and text representations through contrastive training, introduce new attack surfaces: an adversary might detect whether a specific image or caption was used during training—even without access to paired samples. These models are widely deployed in retrieval, captioning, and generative pipelines, making their privacy risks particularly relevant.

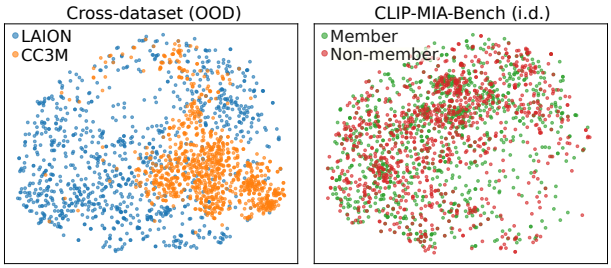

Figure 1: **MIA or OOD Detection?** Left: member and non-member caption-embeddings form distinct clusters in cross-dataset evals – MIAs can exploit spurious signals unrelated to true memorization. Right: member vs. non-member captions in *CLIP-MIA-Bench* are nearly inseparable.

Recent studies have introduced MIAs for CLIP (Ko et al. 2023; Hintersdorf et al. 2024) and reported strong performance – often approaching perfect accuracy. However, in this paper, we show that such results are misleading. Many MIAs evaluation protocols use models trained with small datasets, and/or draw member samples from one dataset (e.g., LAION-400M) and non-member samples from another (e.g., COCO or CC3M), effectively turning the task into an out-of-distribution (OOD) detection problem (see Figure 1). In these settings, models can succeed by exploiting distributional artifacts rather than true memorization (Das, Zhang, and Trantèr 2025), thus overestimating privacy leakage. Our thesis is that without proper data scale, and controlling for distributional shifts between member and non-member samples, it is impossible to disentangle genuine membership signals from dataset-level differences.

To address these issues, we first perform a systematic re-evaluation of published MIAs on CLIP. When member and non-member examples are drawn from the *same Internet-scale crawl and identical preprocessing pipeline*, all prior attacks drop to near-random performance. We then derive

a simple scaling-law bound showing that the member vs. non-member cosine similarity gap decays as $\mathcal{O}(T/N)$ for contrastive learning over N samples and T epochs, and confirm this law empirically when training set grows from 4M to 400M image-text pairs.

To support the community in developing more reliable privacy evaluations, we introduce CLIP-MIA-Bench, a new benchmark consisting of: (i) three ViT-based CLIP models trained from scratch on a curated large-scale dataset, (ii) two balanced subsets of member and non-member samples drawn from the same source and processed identically, and (iii) a full evaluation suite.

Contributions.

- We identify a critical confound in existing MIA evaluations for CLIP, showing that prior results are inflated due to unacknowledged OOD signals and/or small-scale training datasets.
- We empirically re-evaluate several state-of-the-art MIA methods under a realistic, in-distribution setting and show that their performance collapses in the absence of distributional cues. We show, both theoretically and empirically, that this is mainly due to the member/non-member similarity gap decaying with data scale.
- We propose a deliberately simple white-box gradient-based MIA that increases membership signal and outperforms prior attacks without relying on OOD cues.
- We introduce CLIP-MIA-Bench, the first benchmark specifically designed for reproducible, in-distribution MIA evaluation on CLIP models, and release all checkpoints, code, and data splits to facilitate future research.

Related Work

MIAs aim to determine whether a given sample was part of a model's training set. The earliest approaches (Shokri et al. 2017) used shadow models to simulate the target model's behavior and train an attack classifier. Although effective, shadow modeling requires substantial computational resources and knowledge of the target data distribution. Simpler alternatives compare metrics like confidence or loss between members and non-members (Yeom et al. 2018). More refined methods like LiRA (Carlini et al. 2022) use statistical tests on confidence scores across shadow models, offering better sample efficiency and robustness in classification settings. A parallel line of work exploits access to the model parameters in white-box MIAs. Gradient-norm and influence-function tests (Koh and Liang 2017; Liu et al. 2022; Carlini et al. 2023) directly measure how much an input affects the loss landscape, often outperforming confidence-based black-box attacks. Their applicability to large-scale multi-modal encoders has not been yet studied; we close this gap with a simple gradient-based MIA for CLIP.

MIAs in Multi-Modal and Vision-Language Models. Recent works have extended MIAs to multi-modal models, particularly CLIP (Radford et al. 2021). (Liu et al. 2021) proposed a black-box membership inference attack for contrastively-pre-trained encoders: by measuring the cosine similarity between two random augmentations of the same image, the attack exploits higher intra-sample consistency for training points. (Ko et al. 2023) introduced a Weakly-Supervised Attack (WSA) that leverages one-sided knowledge of non-members to find pseudo-member samples by cosine similarity thresholding and train a better member detector. An attacker may obtain a set of samples guaranteed not to be in training – e.g. images collected after the target CLIP model's release (assuming the training data cutoff is known). Using many such known non-members, the attacker trains a classifier (or model) to distinguish pseudo-members from non-members based on features like the CLIP embeddings and similarity score. (Hintersdorf et al. 2024) proposed identity inference attacks that leverage CLIP's ability to link faces with names, exposing risks related to memorized personal data. (Jayaraman, Guo, and Chaudhuri 2024) shift the focus from membership to attribute inference: two otherwise-identical CLIP models are trained on disjoint image–caption splits, and for each caption the authors compare how well each model's k-NN retrieval over a public image pool recovers the ground-truth objects. Target–reference precision/recall gaps then quantify leakage, revealing non-trivial memorization. All of these attacks operate in a black-box setting, relying solely on CLIP embeddings and/or similarity scores returned by the model.

Other than CLIP, (Li et al. 2024) and (Hu et al. 2025) have explored MIAs on instruction-tuned vision-language models such as LLaVA, investigating both modality-specific and token-level signals under varying adversarial capabilities.

Despite this growing interest, nearly all of these works evaluate MIAs on small-scale data and/or using training data from one source (e.g., LAION-400M) and non-members from a different dataset (e.g., COCO or CC12M). This setting introduces strong distributional differences between member and non-member sets – violating core assumptions of a valid MIA and confounding evaluation results.

(Das, Zhang, and Trantèr 2025) recently exposed this problem for several text-only MIA evaluation datasets. They showed that "blind" classifiers – trained only on the data distribution, without any access to the model – can often achieve high accuracy by distinguishing dataset origin alone. These results suggest that many previously reported MIA successes on CLIP may stem not from model memorization but from spurious cues such as image resolution, face blurring, or caption style.

Although these critiques are important, no prior work provides a systematic fix. Existing MIA benchmarks for CLIP and other VLMs still conflate membership with OOD detection. Our work fills this gap by introducing the first benchmark – **CLIP-MIA-Bench** – that ensures a strict in-distribution evaluation. We provide a CLIP model trained from scratch with controlled preprocessing and define two sample sets (members and non-members) from the same dataset pool. This allows us to isolate genuine membership leakage and offers a foundation for future MIA research in vision-language models. We also demonstrate that similarity-based MIAs for CLIP evaluated on models trained with a small dataset can not generalize to foundational large-scale pretrained models.

Background

Membership inference setting

Let $\mathcal{D}_{\text{tr}}$ be the (unknown) distribution over image–text pairs used to train a vision–language model $f_\theta = (f_\theta^{\text{img}}, f_\theta^{\text{text}})$. Given a query pair (x, t), a *membership–inference attack (MIA)* outputs a binary decision $m \in \{0, 1\}$ indicating whether the sample was contained in the training set $\mathcal{S}_{\text{tr}} \sim \mathcal{D}_{\text{tr}}$. A sound evaluation *must* draw both **members** and **non-members** i.i.d. from $\mathcal{D}_{\text{tr}}$; otherwise the task is polluted by out-of-distribution (OOD) cues that an attacker can exploit without revealing true memorisation.

Most prior CLIP MIAs ignore this principle and train the target model f_θ on one dataset, e.g. LAION-400M, and treat samples from another dataset, e.g. COCO, CC3M or CC12M, as non-members. Those corpora differ in image resolution, caption style and domain statistics, so the reported MIA metrics largely reflects dataset discrimination rather than privacy leakage. Figure 1 contrasts this flawed protocol with our in-distribution benchmark.

Contrastive training in CLIP

CLIP learns two encoders that map images x and texts t into a common, ℓ_2-normalised embedding space:

$$z^{\text{img}} = f_\theta^{\text{img}}(x) \in \mathbb{S}^{d-1}, \qquad z^{\text{text}} = f_\theta^{\text{text}}(t) \in \mathbb{S}^{d-1}.$$

For a batch of N matched pairs $\{(x_i, t_i)\}_{i=1}^N$ the symmetric InfoNCE loss is

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2}(\mathcal{L}_{t\to x} + \mathcal{L}_{x\to t}), \tag{1}$$

with

$$\mathcal{L}_{t\to x} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(g_{ii}/\tau)}{\sum_{j=1}^{N}\exp(g_{ij}/\tau)} \tag{2}$$

and analogous $\mathcal{L}_{x\to t}$, where $g_{ij} = g(x_j, t_i; \theta) = \langle z_j^{\text{img}}, z_i^{\text{text}}\rangle$, is the cosine similarity of the image and text embeddings (z_j^{img}, z_i^{text}), and τ the temperature.

Similarity-based MIAs

Because optimisation increases g_{ii} for training pairs, a simple black-box attack declares membership when the observed similarity exceeds a threshold (Ko et al. 2023):

$$\widehat{m}(x, t) \;=\; \mathbb{1}\big[g(x, t; \theta) > \kappa\big], \tag{3}$$

mirroring confidence-score MIAs in classification models (Yeom et al. 2018). In the next section we analyze how the *expected* gap $\Delta = \mathbb{E}[\, g_{\text{member}} - g_{\text{non-member}}\,]$ decays with training dataset size and why, under an in-distribution split, threshold attacks alone become ineffective at web scale.

Theoretical Analysis of Similarity–Based MIAs on CLIP

We now formalise why threshold-style, *similarity-based* attacks (3) inevitably degrade as the training set grows. The analysis is deliberately minimal: we assume the standard InfoNCE objective, ℓ_2-normalised embeddings and GD optimization. Under these conditions we prove that the **expected similarity gap** $\Delta_N = \mathbb{E}\big[g_{\text{member}} - g_{\text{non}}\big]$ decays as $\mathcal{O}(T/N)$, where T is the number of epochs and N the number of unique training pairs. Because a threshold attack's AUC is a sigmoidal function of Δ_N, the attack quickly approaches random guessing once N enters the hundreds of millions. For our analysis we make the following assumptions:

1. *Unit-sphere embeddings.* Both encoders map to $\mathbb{S}^{d-1}$: $\|f_\theta^{\text{img}}(x)\| = \|f_\theta^{\text{text}}(t)\| = 1$.
2. *Lipschitz encoder family.* There exists $L_{\text{enc}} > 0$ such that $\|\nabla_\theta g(x, t; \theta)\| \le L_{\text{enc}}$ for all (x, t) and θ. This holds for transformers with bounded weight matrices.
3. *Full-batch gradient descent (GD).* Each epoch performs a single update with step size η using the complete dataset of N pairs. Each positive pair in a batch competes against $N - 1$ negatives.

Theorem 1 (Similarity-gap upper bound)**.** *Under (A1)–(A3) and after T epochs of training,*

$$\boxed{\Delta_N \;\le\; \frac{2\,\eta\,T\,L_{\text{enc}}}{\tau\,N} \;=\; \mathcal{O}\big(T/N\big)}$$

where τ is the InfoNCE temperature.

In other words, as the dataset grows, black-box threshold attacks lose power in inverse proportion to the number of samples N. The expected cosine-similarity advantage enjoyed by any training pair over a fresh, unseen pair decays at most linearly with $1/N$.

Proof sketch. By Assumption (A2) we have $\|\nabla_\theta g(x, t; \theta)\| \le L_{\text{enc}}$. For the InfoNCE loss in Eq. (2), the derivative of a per-sample loss ℓ_i with respect to its positive logit is bounded in magnitude by $1/\tau$, since softmax probabilities lie in $[0, 1]$. By the chain rule this gives $\|\nabla_\theta \ell_i(\theta)\| \le 2L_{\text{enc}}/\tau$ (accounting for the two directions $t \to x$ and $x \to t$). Under full-batch gradient descent with step size η and N training pairs, the update after one epoch is $\theta_{k+1} = \theta_k - \eta\,\frac{1}{N}\sum_{i=1}^{N}\nabla_\theta \ell_i(\theta_k)$, so the contribution of a single example to the parameter change is at most $2\eta L_{\text{enc}}/(\tau N)$. Summing over T epochs and using the triangle inequality yields $\|\theta_T - \theta_0\| \le 2\eta T L_{\text{enc}}/(\tau N)$. Since $\|\nabla_\theta g(x, t; \theta)\| \le L_{\text{enc}}$, the similarity is L_{enc}-Lipschitz in parameter space. Therefore, for a fixed member pair, we have $|g(x, t; \theta_T) - g(x, t; \theta_0)| \le L_{\text{enc}}\|\theta_T - \theta_0\|$. Meanwhile, the expected similarity of a fresh non-member remains unchanged. This implies the claimed bound on Δ_N. The tightness of this bound is confirmed empirically in our experimental section.

From similarity gap to attack AUC

If member and non-member similarities are Gaussian with common variance σ^2 (empirically a good fit), then the ROC–AUC of any threshold attack is

$$\text{AUC}_N = \Phi\Big(\frac{\Delta_N}{\sqrt{2}\,\sigma}\Big),$$

where Φ is the standard normal CDF. Combined with Theorem 1 this yields $\text{AUC}_N \to \frac{1}{2}$ (a random classifier, ≈50% AUC) once $N \gg (2\eta T L_{\text{enc}}/\tau)/\sigma$, up to constant factors.

Mini-batch SGD in practice. The proof sketch above assumed a full batch of size N, yet CLIP training uses mini-batches of size $B \ll N$ with in-batch negatives. Standard analyses of reshuffled SGD show that, in expectation, the per-epoch update coincides with that of full-batch GD, so the $O(T/N)$ dependence in Theorem 1 still governs the similarity gap up to constants.

Dependence on optimiser and schedule. (A3) can be generalised to cosine decay or AdamW by integrating the actual effective step size, changing only the constant factor.

Limitations. The bound assumes smooth CLIP encoders and applies only to attacks that look at *global* image–text similarity. It does not constrain white-box gradient-based attacks nor patch- or token-level "micro-memory" probes – e.g. detecting whether a face crop or rare word embedding was memorised as in (Hintersdorf et al. 2024).

White-Box Loss & Grad-Norm Attack

Following (Ko et al. 2023) Weakly Supervised Attack (WSA), we assume the attacker holds a small set $\mathcal{D}_{\text{no}}$ of samples *guaranteed* not to be in training (e.g. scraped after the model's public release). This "one-sided" knowledge is realistic – model owners rarely reveal training data, but an auditor can always collect fresh samples. We use $\mathcal{D}_{\text{no}}$ to set a loss+grad-norm threshold that mines a pool of *pseudo-members* $\tilde{\mathcal{D}}_{\text{mem}}$ from an unlabelled corpus. The attacker then trains a logistic regressor on a dataset with positives $\tilde{\mathcal{D}}_{\text{mem}}$ and negatives $\mathcal{D}_{\text{no}}$.

We use the symmetric CLIP loss in Eq. 1 and denote its *per-sample loss* value for a given image-text pair (x, t) as $L(x, t)$. Let $g(x, t)$ be the gradient of this loss with respect to the ℓ_2-normalised image and text embeddings, and $G(x, t) = \|g(x, t)\|_2$ its norm. Then, we form z-scores:

$$z_L(x,t) = \frac{\mu_L - L(x,t)}{\sigma_L}, \quad z_G(x,t) = \frac{\mu_G - G(x,t)}{\sigma_G},$$

where μ_L, σ_L, μ_G, σ_G are respectively the empirical mean and standard deviation of the loss and gradient norm in $\mathcal{D}_{no}$. Finally, we define a joint score $s(x, t) = \frac{1}{2}(z_L(x, t) + z_G(x, t))$ that is used to label pseudo-members from the unlabeled corpus, and train a logistic regressor with $[L(x, t); G(x, t)]$ as per sample features.

This MIA is intentionally minimal, aiming at a transparent baseline that reveals memorization without OOD cues. The intuition is that SGD drives the *feature-space gradient* at each training pair towards zero: once the pair's similarity is "good enough" its local loss stops changing, whereas an unseen pair still exerts non-zero gradient pressure, the loss itself follows a similar pattern. As mentioned before, these gradient-based features are not subject to the similarity-gap upper bound derived in Theorem 1.

A Realistic Benchmark for CLIP Membership Inference

To address the conflation of membership inference and out-of-distribution (OOD) detection in prior work, we introduce **CLIP-MIA-Bench** – a new benchmark that supports reliable, reproducible evaluation of membership inference attacks on vision–language models. Our benchmark includes:

- Three ViT-B/16 CLIP models trained from scratch on {4 M, 40 M, 400 M} image–text pairs with identical hyper-parameters.
- A member set and a non-member set drawn from the same data distribution, ensuring that no distributional cues can be exploited by attacks.
- Full code, data splits, and model checkpoints to support transparent and future-proof comparisons.

Training Data: CommonPool from DataComp

We use the large-scale CommonPool corpus from the DataComp benchmark (Gadre et al. 2023), which contains metadata for approximately 1.3 billion image–text pairs. While the corpus provides only metadata (image URLs, alt-text, CLIP similarity scores, etc.), we recover ∼80% of the samples using the DataComp distributed downloader. As expected, a portion of samples are unavailable due to dead links or rate-limiting – an inherent limitation when working with archived internet-scale datasets.

To ensure our setup matches real-world CLIP training, we follow DataComp's default preprocessing pipeline: each image is resized so that its longest side is at most 512 pixels, and faces are blurred using automated face detectors to protect privacy. This is in contrast to LAION-400M, which does not apply face blurring. We apply CLIPScore-based filtering to retain the top 40% of samples, resulting in a final dataset of ∼400 million high-quality image–text pairs. This scale mirrors that of LAION-400M and the original CLIP model (Radford et al. 2021), ensuring compatibility.

CLIP Model Training

We train three ViT-B/16 encoders – the DataComp standard for our data scale – using `OpenCLIP`: one on 4M pairs, one on 40M pairs, and one on 400M pairs. All hyper-parameters (batch=8192, LR=5e-4, 32 epochs, cosine decay) are held constant to isolate the effect of data scale. Training takes approximately 0.7, 7, and 70 hours respectively using 128 NVIDIA H100 GPUs.

Member and Non-Member Subsets

DataComp organizes downloaded samples into 10,000-sample shards. We collected a total of 43,026 shards and build the evaluation sets as follows:

- We hold out 26 shards *before training* to serve as the **non-member set**.
- The remaining 40,000 shards are used for training. From these, we select 26 different shards *post-training* to define the **member set**.

Since the metadata in CommonPool is globally shuffled, the holdout and training shards represent i.i.d. samples from the same underlying distribution. Both sets undergo the exact same preprocessing pipeline, including face blurring and image resizing, eliminating spurious visual or textual cues that could lead to false positive detections. Table 1 provides a

	Member Set	Non-Member Set
Source Dataset	CommonPool-Large (DataComp)	CommonPool-Large (DataComp)
Sampling Method	From training shards	Held out before training
Image Preprocessing	Resize to max 512px, face blurring	Resize to max 512px, face blurring
Text Format	Alt-text only	Alt-text only
Eval Sample Size	27,000	27,000
Other Sample Size	$\sim$200,000	$\sim$200,000

Table 1: Summary of member and non-member subsets in CLIP-MIA-Bench.

summary of key statistics of member and non-member subsets and test/other splits. The balanced test split is designed for MIA evaluation, while the rest of the samples are kept aside for other fair uses, such as methods that leverage one-sided knowledge of non-members to find pseudo-members and then train a MIA classifier.

Release Details

To promote transparency and reproducibility, we release:

- The three ViT-B/16 checkpoints (4M, 40M, 400M).
- The member and non-member samples metadata (CommonPool IDs, image URLs + captions) plus their precomputed image-text embeddings.
- All code for reproducing our experiments.

The benchmark is designed to be plug-and-play: future MIA methods can be tested using our fixed model and splits, ensuring consistency in evaluation and removing distributional confounds. Public data and code allow future extensions to larger CLIP backbones, fine-tuned encoders, and alternative VLMs.

Model Validation

To validate our training setup and the quality of our models, we evaluate the zero-shot performance of the 400M model on a broad suite of downstream tasks. Specifically, we compare our ViT-B/16 CLIP model to the original OpenAI CLIP (Radford et al. 2021) and the OpenCLIP variant trained on LAION-400M (Ilharco et al. 2021), using the standard set of 38 evaluation benchmarks proposed in (Gadre et al. 2023). These include classification tasks across a range of domains, image-text retrieval datasets, and robustness tests.

Model	Pretraining Dataset	Avg. (38 tasks)
OpenAI CLIP	Private (400M)	56.26
OpenCLIP	LAION-400M	56.21
Ours	CommonPool (400M)	**56.34**

Table 2: Zero-shot average accuracy across 38 tasks. Our model matches or slightly exceeds the performance of prior CLIP ViT-B/16 variants trained at similar scale.

These results confirm that our model achieves competitive generalization performance relative to widely used public CLIP checkpoints trained on similar data scale. It thus serves as a realistic and representative target for evaluating membership inference attacks for CLIP.

Experiments

In this section, we systematically evaluate membership inference attacks (MIAs) targeting CLIP models. We consider a diverse suite of prior art, alongside several baselines, under two evaluation regimes:

LAION400M vs. {CC12M $\cup$ CC3M $\cup$ MSCOCO}. Following Ko *et al.* (Ko et al. 2023), we replicate their approach of attacking a CLIP-like model trained on LAION400M (Schuhmann et al. 2021), while using three other datasets as the non-member set: CC3M (Sharma et al. 2018), CC12M (Changpinyo et al. 2021), and MSCOCO (Lin et al. 2014).

CLIP-MIA-Bench. Our new evaluation benchmark, where member and non-member samples come from the same underlying distribution (CommonPool) and share identical preprocessing.

Evaluation Metrics

Since CLIP-MIA-Bench is intended as a reproducible and fair benchmark for membership inference attacks (MIAs), we adopt metrics that reflect both overall attack effectiveness and real-world applicability. We report two complementary metrics: **ROC-AUC** and **TPR@1%FPR**, both of which are standard in the MIA literature (Carlini et al. 2022; Nasr, Shokri, and Houmansadr 2018).

ROC-AUC. The Area Under the Receiver Operating Characteristic curve summarizes how well an attack separates member and non-member examples across all decision thresholds. It measures the probability that a randomly chosen member will receive a higher membership score than a randomly chosen non-member. A perfect attack has AUC = 1.0; a random guess yields AUC = 0.5. This metric is useful for global comparisons and sanity checks.

TPR@1%FPR. The True Positive Rate at a fixed False Positive Rate of 1% captures how well an attack performs under strict adversarial constraints. In realistic threat models, an attacker often seeks to identify members with high confidence while minimizing false alarms. This setting reflects privacy-sensitive use cases such as individual data exposure auditing, where even a small number of false positives can be costly. TPR@1%FPR is also more sensitive to

LAION400M vs. {CC12M ∪ CC3M ∪ MSCOCO}						
	ViT-B/32		ViT-B/16		ViT-L/14	
	AUC	TPR@1%FPR	AUC	TPR@1%FPR	AUC	TPR@1%FPR
Blind Attack (Das et al. 2024)	98.42	96.01	98.42	96.01	98.42	96.01
CSA Baseline	74.60	7.23	75.93	7.58	78.76	4.87
AEA (Liu et al. 2021)	76.25	9.40	78.97	9.20	79.50	8.36
WSA (Ko et al. 2023)	<u>91.99</u>	72.12	<u>93.49</u>	73.81	<u>94.13</u>	76.11
MCDropout	**97.13**	**89.52**	**98.27**	**90.31**	**99.53**	**92.15**
Loss-Grad	86.68	<u>74.65</u>	88.84	<u>83.67</u>	88.91	<u>85.50</u>

Table 3: Attack performance evaluation on the *LAION400M vs.* $\{$*CC12M* $\cup$ *CC3M* $\cup$ *MSCOCO*$\}$ MIA setting for three different model architectures: ViT-B/32, ViT-B/16, and ViT-L/14. We report both AUC and TPR@1%FPR. Bold and underline numbers indicate best and second best results respectively.

overfitting or artifact exploitation than AUC, and thus serves as a robustness diagnostic.

Evaluated Methods

We include the following methods in our study:

CSA Baseline. A zero-knowledge attack that uses cosine similarity thresholding.

Augmentation-Enhanced Attack (AEA) (Liu et al. 2021; Ko et al. 2023). AEA extends the CSA baseline with K image transformations, computing the cosine similarity for each transformed image and aggregating the scores. The transformations are chosen without direct knowledge of the training distribution, following techniques in (Kaya and Dumitras 2021; Choquette-Choo et al. 2021).

Weakly Supervised Attack (WSA) (Ko et al. 2023). A black-box attack assuming partial knowledge of the training distribution. Ko *et al.* exploit *one-sided non-member* data collected after the model's release date, and use it to build pseudo-member sets using CSA. They train a binary classifier on concatenated image-text features of these pseudo-member and known non-member samples.

Monte Carlo Dropout (MCD): A standard uncertainty-based technique used in OOD detection (Gal and Ghahramani 2016; Yang et al. 2022; Zhang et al. 2023), included here as a natural baseline to help exposing the confound between OOD detection and MIA. By computing uncertainty statistics over multiple stochastic forward passes, MCD can distinguish members from non-members when data distributions differ. Specifically, we compute the mean cosine similarity across 50 Monte Carlo Dropout samples with dropout rate 0.2. Notice that, contrary to previous methods, MCD is a white-box attack, since it requires access to model weights.

Blind Attack (Das, Zhang, and Trantèr 2025). A simple bag-of-words classifier trained solely on text captions to predict membership, with no model queries. It uses 80% of the labeled member and non-member data for training and 20% for testing, evaluated via 10-fold cross-validation. While not a realistic adversarial method – since it requires direct knowledge of labeled members – this serves as an indicator of inherent distributional differences between member and non-member samples.

CLIP-MIA-Bench		
	ViT-B/16 (400M)	
	AUC	TPR@1%FPR
Blind (Das et al. 2024)	50.09	0.98
CSA Baseline	51.12	1.07
AEA (Ko et al. 2023)	51.30	1.00
WSA (Ko et al. 2023)	50.07	0.88
MCDropout	**52.63**	0.60
Loss-Grad	**53.48**	**1.75**

Table 4: Attack performance evaluation on the *CLIP-MIA-Bench* for ViT-B/16 model trained on 400M image-text pairs. We report both AUC and TPR@1%FPR. In bold — methods whose AUC is significantly above 0.5 at 95% confidence (1,000× bootstrap).

Results and Discussion

Table 3 summarizes the results on the **LAION400M vs. {CC12M ∪ CC3M ∪ MSCOCO}** setting, while Table 4 provides the obtained results on **CLIP-MIA-Bench**. We make the following key observations:

(1) All methods perform well in the cross-dataset setting. Most attacks achieve high AUC and even perfect separation under this evaluation regime. Notably, even the Blind Attack, which does not query the model at all, achieves AUC > 0.98 – exposing the severity of distribution leakage.

(2) MCD shows near-perfect results under cross-dataset conditions. This is consistent with its role in OOD detection benchmarks (e.g., OpenOOD (Yang et al. 2022; Zhang et al. 2023)). Here, the averaged similarity across stochastic draws becomes highly discriminative – not due to true membership, but due to broader confidence margins on unfamiliar (OOD) data. Our gradient-based method and WSA perform competitively.

(3) Performance collapses under large-scale in-distribution evaluation. When tested on CLIP-MIA-Bench, all methods show near random performance in both AUC and TPR@1%FPR.

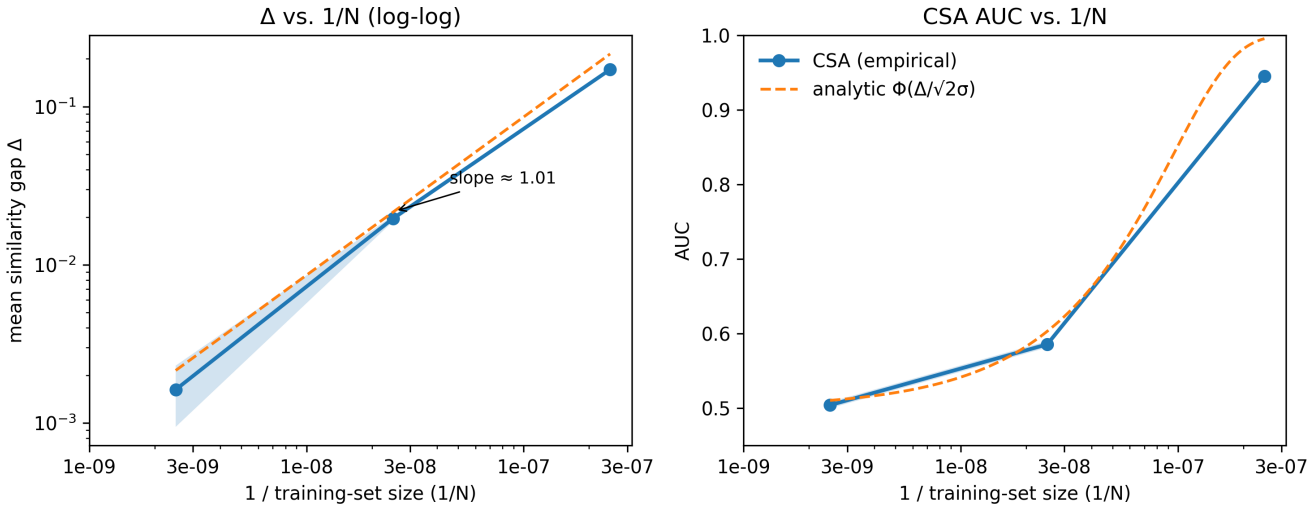

Figure 2: **Similarity-gap scaling**. Left: mean gap $\Delta_N = \mathbb{E}\big[g_{\text{member}} - g_{\text{non}}\big]$ versus $1/N$ in log–log space. Right: CSA AUC versus $1/N$ together with the analytic mapping AUC $= \Phi\big(\Delta_N/\sqrt{2}\sigma\big)$ (dashed orange, σ estimated on 40 M). Error bars show 95% CIs over 1,000 bootstrap draws.

(4) Our loss–grad attack is the *only* method that remains statistically above random in the strict setting. On the 400M–pair CLIP-MIA-Bench model it reaches **AUC = 0.534±0.006** and **TPR@1%FPR = 1.75%±0.35%**. We report 95% confidence intervals obtained from 1,000× stratified bootstrapping for AUC and Clopper–Pearson intervals for TPR. A paired DeLong test rejects the null "AUC=0.5" at $p < 10^{-5}$, and also shows a significant gap over the next-best baseline (MCDropout, $p = 1.3 \times 10^{-4}$). All other attacks' intervals overlap the random-chance line, confirming that they provide no reliable membership signal once OOD cues are removed.

Our results clearly demonstrate that prior CLIP MIA results are not robust to realistic evaluation settings. The dramatic performance collapse across all methods – especially the MCDropout, WSA, and Blind attacks – confirms that earlier gains were due to OOD artifacts, not genuine memorization. These findings underscore the need for reliable benchmarks like CLIP-MIA-Bench to enable future progress in MIA research for foundation models.

Empirical verification of the similarity–gap bound

We trained three ViT-B/16 models on $\{4\,\text{M}, 40\,\text{M}, 400\,\text{M}\}$ pairs drawn from CommonPool, holding *all* hyperparameters (batch = 8 192, 32 epochs, LR = 5e-4, etc.) fixed. For each checkpoint we evaluate the cosine-similarity attack (CSA) on the *same* CLIP-MIA-Bench splits, we compute:

$$\Delta_N = \frac{1}{|\mathcal{D}_{\text{mem}}|} \sum_{(x,t)\in\mathcal{D}_{\text{mem}}} g(x,t) - \frac{1}{|\mathcal{D}_{\text{non}}|} \sum_{(x,t)\in\mathcal{D}_{\text{non}}} g(x,t),$$

and estimate standard deviations via 1,000× stratified bootstrap.

Fig. 2 (left) shows that Δ_N falls almost exactly linearly with $1/N$ (slope 0.99 ± 0.02, $R^2 = 0.997$), matching Theorem 1. Mapping this gap to AUC through the Gaussian approximation yields the dashed curve in Fig. 2 (right), which passes through the empirical CSA points (94.1 % → 62.5 % → 51.1 %). Hence the theory not only bounds but *quantitatively predicts* the collapse of similarity-based MIAs as data scale (N) grows at fixed epoch budget $T = 32$.

The key observation here is that with web-scale datasets ($N \geq 10^8$) and small T the member/non-member cosine gap is inherently too small to exploit, while white-box signals such as our Loss–Grad features might remain informative.

Conclusion

We revisited membership inference for CLIP-style models and showed that previously reported results dissolve once member and non-member samples come from the *same* distribution. To enable rigorous study we release **CLIP-MIA-Bench**: three ViT-B/16 checkpoints (4M, 40M, 400M pairs), matched member/non-member splits, and full code for both methods and evaluation. **CLIP-MIA-Bench** is publicly available at https://clipmiabench.github.io/.

On this benchmark every published black-box attack collapses to chance, in line with our derived bound demonstrating that the cosine-similarity gap decays as $\mathcal{O}(T/N)$. Our simple white-box *Loss–Grad* attack is the only method that remains significantly above random (AUC 0.534±0.006; TPR@1%FPR 1.75%±0.35%) on the 400 M-pair model.

Our white-box Loss-Grad attack requires weight access – realistic for widely used open-weight models. Future work should (i) design stronger white-box or patch-level MIAs, (ii) study fine-tuned CLIP and other VLMs, and (iii) explore defenses that suppress loss/gradient leakage. We hope this work will re-anchor privacy research on vision-language models in realistic settings.

Acknowledgments

This work has been supported by the Ramon y Cajal research fellowship RYC2020-030777-I/AEI/10.13039/501100011033, the European Union under the CERV programme (Call: CERV-2024-CHAR-LITI-CHARTER, Project ID: 101214711), the European Lighthouse on Safe and Secure AI - ELSA funded by European Union's Horizon Europe programme under grant agreement No 101070617, and the Consolidated Research Group 2021 SGR 01559 from the Research and University Department of the Catalan Government. We acknowledge EuroHPC JU for awarding the project ID EHPC-AI-2024A02-040 access to MareNostrum 5 hosted at BSC-CNS.

References

Carlini, N.; Chien, S.; Nasr, M.; Song, S.; Terzis, A.; and Tramer, F. 2022. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, 1897–1914. IEEE.

Carlini, N.; Hayes, J.; Nasr, M.; Jagielski, M.; Sehwag, V.; Tramer, F.; Balle, B.; Ippolito, D.; and Wallace, E. 2023. Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX Security 23)*, 5253–5270.

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3558–3568.

Choquette-Choo, C. A.; Tramer, F.; Carlini, N.; and Papernot, N. 2021. Label-only membership inference attacks. In *International conference on machine learning*, 1964–1974. PMLR.

Das, D.; Zhang, J.; and Trantèr, F. 2025. Blind baselines beat membership inference attacks for foundation models. In *2025 IEEE Security and Privacy Workshops (SPW)*, 118–125. IEEE.

Gadre, S. Y.; Ilharco, G.; Fang, A.; Hayase, J.; Smyrnis, G.; Nguyen, T.; Marten, R.; Wortsman, M.; Ghosh, D.; Zhang, J.; Orgad, E.; Entezari, R.; Daras, G.; Pratt, S.; Ramanujan, V.; Bitton, Y.; Marathe, K.; Mussmann, S.; Vencu, R.; Cherti, M.; Krishna, R.; Koh, P. W.; Saukh, O.; Ratner, A.; Song, S.; Hajishirzi, H.; Farhadi, A.; Beaumont, R.; Oh, S.; Dimakis, A.; Jitsev, J.; Carmon, Y.; Shankar, V.; and Schmidt, L. 2023. DataComp: In search of the next generation of multimodal datasets. arXiv:2304.14108.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International conference on machine learning*.

Hintersdorf, D.; Struppek, L.; Brack, M.; Friedrich, F.; Schramowski, P.; and Kersting, K. 2024. Does clip know my face? *Journal of Artificial Intelligence Research*, 80: 1033–1062.

Hu, Y.; Li, Z.; Liu, Z.; Zhang, Y.; Qin, Z.; Ren, K.; and Chen, C. 2025. Membership Inference Attacks Against Vision-Language Models. *arXiv preprint arXiv:2501.18624*.

Ilharco, G.; Wortsman, M.; Wightman, R.; Gordon, C.; Carlini, N.; Taori, R.; Dave, A.; Shankar, V.; Namkoong, H.; Miller, J.; Hajishirzi, H.; Farhadi, A.; and Schmidt, L. 2021. OpenCLIP. If you use this software, please cite it as below.

Jayaraman, B.; Guo, C.; and Chaudhuri, K. 2024. Déjà vu memorization in vision–language models. *Advances in Neural Information Processing Systems*, 37: 50722–50749.

Kaya, Y.; and Dumitras, T. 2021. When does data augmentation help with membership inference attacks? In *International conference on machine learning*, 5345–5355. PMLR.

Ko, M.; Jin, M.; Wang, C.; and Jia, R. 2023. Practical membership inference attacks against large-scale multi-modal models: A pilot study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4871–4881.

Koh, P. W.; and Liang, P. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*, 1885–1894. PMLR.

Li, Z.; Wu, Y.; Chen, Y.; Tonin, F.; Abad Rocamora, E.; and Cevher, V. 2024. Membership Inference Attacks against Large Vision-Language Models. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 98645–98674. Curran Associates, Inc.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, H.; Jia, J.; Qu, W.; and Gong, N. Z. 2021. Encodermi: Membership inference against pre-trained encoders in contrastive learning. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2081–2095.

Liu, Y.; Zhao, Z.; Backes, M.; and Zhang, Y. 2022. Membership inference attacks by exploiting loss trajectory. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2085–2098.

Nasr, M.; Shokri, R.; and Houmansadr, A. 2018. Comprehensive privacy analysis of deep learning. In *Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP)*, volume 2018, 1–15.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Schuhmann, C.; Kaczmarczyk, R.; Komatsuzaki, A.; Katta, A.; Vencu, R.; Beaumont, R.; Jitsev, J.; Coombes, T.; and Mullis, C. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. In *NeurIPS Workshop Datacentric AI*, FZJ-2022-00923. Jülich Supercomputing Center.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Annual*

Meeting of the Association for Computational Linguistics, 2556–2565.

Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, 3–18. IEEE.

Yang, J.; Wang, P.; Zou, D.; Zhou, Z.; Ding, K.; Peng, W.; Wang, H.; Chen, G.; Li, B.; Sun, Y.; et al. 2022. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35: 32598–32611.

Yeom, S.; Giacomelli, I.; Fredrikson, M.; and Jha, S. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, 268–282. IEEE.

Zhang, J.; Yang, J.; Wang, P.; Wang, H.; Lin, Y.; Zhang, H.; Sun, Y.; Du, X.; Li, Y.; Liu, Z.; et al. 2023. OpenOOD v1. 5: Enhanced Benchmark for Out-of-Distribution Detection. In *NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models*.