FiMMIA: scaling semantic perturbation-based membership inference across modalities

Anton Emelyanov
SberAI
login-const@mail.ru

Sergei Kudriashov
Sber, HSE University
sakudryashov@hse.ru

Alena Fenogenova
SberAI
alenush93@gmail.com

Abstract

Membership Inference Attacks (MIAs) aim to determine whether a specific data point was included in the training set of a target model. Although there are have been numerous methods developed for detecting data contamination in large language models (LLMs), their performance on multimodal LLMs (MLLMs) falls short due to the instabilities introduced through multimodal component adaptation and possible distribution shifts across multiple inputs. In this work, we investigate multimodal membership inference and address two issues: first, by identifying distribution shifts in the existing datasets, and second, by releasing an extended baseline pipeline to detect them. We also generalize the perturbation-based membership inference methods to MLLMs and release **FiMMIA** — a modular **F**ramework for **M**ultimodal **MIA**.[1] We propose to train a neural networks to analyze the target model's behavior on perturbed inputs, capturing interactions between semantic domains and loss values on members and non-members in the local neighborhood of each sample. Comprehensive evaluations on various fine-tuned multimodal models demonstrate the effectiveness of our perturbation-based membership inference attacks in multimodal settings.

1 Introduction

The development of MLLMs has exceeded expectations (Liu et al., 2023a; Lin et al., 2023), showcasing extraordinary performance on various multimodal benchmarks (Chervyakov et al., 2025; Lu et al., 2022; Liu et al., 2023b; Song et al., 2024), even surpassing human performance. However, due to the partial obscurity associated with MLLMs training or fine-tuning (OpenAI, 2023; Reid et al., 2024), it remains challenging to definitively ascertain the impact of training data on model performance, despite some works showing the employment of the training set of certain datasets (Liu et al., 2023a; Chen et al., 2023; Bai et al., 2023). The issue of data contamination occurs when training or test data of benchmarks is exposed during the model training or fine-tuning phase (Xu et al., 2024) and could potentially instigate inequitable performance comparisons among models.

Although numerous works in the field of LLMs have proposed methods for detecting data contamination (Mozaffari and Marathe, 2024; Hu et al., 2022a; Song et al., 2025; Li et al., 2024b), MLLMs, due to their various modalities that, in most implementations, lack corresponding target tokens for multimodal inputs, while multiple training phases, common for MLLM training, complicate an inference when one tries to apply these methods directly. Therefore, there is a necessity in a multimodal contamination detection framework specifically tailored for MLLMs. Our main contributions can be summarized as follows:

- We extended the work of Das et al. (2024) to multimodal data and assessed image as well as recent text MIA benchmarks (Fu et al., 2025; Hallinan et al., 2025) for distribution shifts. We have found that even the *most recent proposed benchmarks are subject to distribution shifts between member and non-member data*.
- We *release a baseline attack pipeline for text, image, video and audio data*, that collects various statistics from the dataset distribution and trains a classifier on top to distinguish members from non-members without any signal from the target model.
- We *extend perturbation-based MIA methods to MLLMs*, revealing their effectiveness and transferability even at the scale of billion-parameter models.
- We *release a modular framework* **FiMMIA** supporting diverse datasets, modalities, and

[1] The source code and framework have been made publicly available under the MIT license via link.The video demonstration is available on YouTube.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 3: System Demonstrations, pages 139–153
March 24-29, 2026 ©2026 Association for Computational Linguistics

neighbor generation methods. Our pipelines support MIA in multiple settings: when only text, multimodal or both parts are assumed to be leaked.

2 Related Work

2.1 Data contamination and distribution shifts hinder reliable evaluations

Preserving training data confidentiality is critical for LLMs, as their datasets can contain sensitive private information and tests (Yeom et al., 2018; Hu et al., 2022b). Additionally, data contamination between training and test sets undermines benchmark reliability and complicates model comparison (Balloccu et al., 2024; Sainz et al., 2023), driving recent adoption of dynamically updated benchmarks (White et al., 2025).

Distribution shifts pose significant risks as neural networks' ability to extract subtle correlations makes them vulnerable to adversarial examples (Moayeri et al., 2022), spurious correlations in explanations (Ribeiro et al., 2016), and data poisoning (Souly et al., 2025). Recent studies have also found that modern LLMs are capable of intensional *sandbagging*, i.e., strategically underperforming during the evaluations in the presence of an incentive to do so (van der Weij et al., 2024). In other words, capable LLMs can intensionally manipulate their logprobs, which poses an additional challenge both for capability elicitation and loss-based MIA attacks [2].

2.2 Membership inference attacks aim to solve the problem

Membership Inference Attacks (MIAs) determine whether a data sample was part of a model's training set (Shokri et al., 2017) or originates from the general distribution. As noted by (Carlini et al., 2022), this constitutes a hypothesis testing task that crucially relies on the i.i.d. assumption.

Membership Inference Attacks have been the subject of considerable research across a variety of machine learning models, including classification models (Long et al., 2018; Song et al., 2019; Choquette-Choo et al., 2021), generative models (Hayes et al., 2017; Hilprecht et al., 2019; Chen et al., 2020), and embedding models (Song and Raghunathan, 2020; Mahloujifar et al., 2021). The appearance of LLMs has likewise led to numerous studies investigating membership inference attacks against them (Mireshghallah et al., 2022; Fu et al., 2023; Shi et al., 2024; Mattern et al., 2023). However, the field of MIAs for multimodal models is still in its nascent stages and requires further exploration, facing challenges due to the absence of targets for modality-related tokens, instabilities from multimodal adaptation etc. Several methods (Ko et al., 2023; Hu et al., 2022d) proposed to conduct MIAs based on the similarity between an image and its associated text label. However, this technique is limited to the presence of a paired entry (pair image/text), not the presence of a solitary image or text sequence.

	Dataset / task	Best reported(%)	Our baseline(%)
text	WikiMIA-hard	64.0 (Hallinan et al., 2025)	57.7 ± 2.5
	WikiMIA-24	99.8 (Fu et al., 2025)	99.9 ± 0.1
	VL-MIA-Text (32 tok.)	96.2 (Li et al., 2024c)	84.9 ± 4.0
	VL-MIA-Text (64 tok.)	99.3 (Li et al., 2024c)	95.5 ± 0.9
image	VL-MIA-Flickr	94.2 (Yin et al., 2025)	99.1 ± 0.4
	VL-MIA-Flickr-2k	74.0 (Li et al., 2024c)	98.6 ± 0.4
	VL-MIA-Flickr-10k	NA	99.3 ± 0.1
	VL-MIA-DALL-E	84.0 (Yin et al., 2025)	99.9 ± 0.1
	LAION-MI*	2.42 (Dubiński et al., 2023)	1.11 ± 0.1

Table 1: AUC-ROC Evaluations of image and text MIA datasets for the occurrence of distribution shifts between members and non-members data. * corresponds to TPR@1FPR instead. Datasets with no distribution shifts between members and non-members should display values of **50%** for AUC-ROC and **0** for TPR@1FPR.

MIAs are commonly categorized into metric-based and shadow model-based approaches (Hu et al., 2022b). Metric-based MIAs (Yeom et al., 2018; Salem et al., 2018; Song and Mittal, 2021; Shi et al., 2024) compare model output statistics against a threshold, while shadow model-based methods (Shokri et al., 2017; Salem et al., 2018) require computationally expensive model replication. Recent work has introduced semantic MIAs (Koike et al., 2025; Mozaffari and Marathe, 2024) that exploit local model properties through sample perturbations. We extend this semantic approach to image, audio, video, and text modalities.

3 FiMMIA

3.1 Overview

The system is the first collection of models and pipelines for membership inference attacks against LLMs, built and evaluated initially on the Russian language, and extendable to any other language or MMLM dataset. The pipeline supports differ-

[2] Such behavior is only possible if the evaluation data or environment presents enough evidence to distinguish it from the training environment, even due to subtle cues.

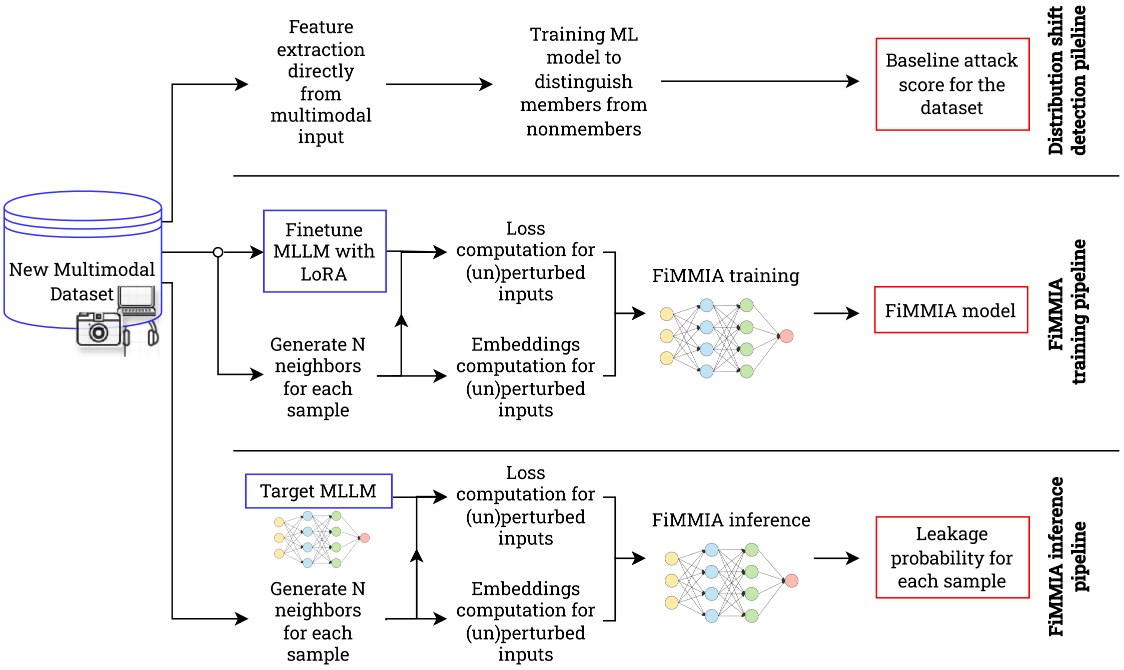
New Multimodal Dataset
Feature extraction directly from multimodal input
Training ML model to distinguish members from nonmembers
Baseline attack score for the dataset
Distribution shift detection pipeline
Finetune MLLM with LoRA
Generate N neighbors for each sample
Loss computation for (un)perturbed inputs
Embeddings computation for (un)perturbed inputs
FiMMIA training
FiMMIA model
FiMMIA training pipeline
Target MLLM
Generate N neighbors for each sample
Loss computation for (un)perturbed inputs
Embeddings computation for (un)perturbed inputs
FiMMIA inference
Leakage probability for each sample
FiMMIA inference pipeline

Figure 1: Overview of FiMMIA Inference pipeline for MLLMs. Inputs to the pipeline are shown in blue. Outputs of the pipeline are shown in red.

ent modalities: text, image, audio and video and is fully open source[3]. In order to allow for membership inference in cases, when only the text or multimodal part is assumed to be leaked, we support separate neighbor and embedding generation for both parts of the input, thus providing an option to disentangle their contribution to the final MIA score.

We release pretrained FiMMIA models to promote experiments within the community[4]. Although in our experiments we focus on MERA datasets (Chervyakov et al., 2025) to ensure independence in the split between members and nonmembers, *the presented pipeline is built with the idea of supporting modular extension and is intended to be easily adopted.*

Overall, the system is a set of models and Python scripts in a GitHub repository that supports three major functionalities: 1) a baseline attack based on distribution statistics, intended to ensure the reliability of multimodal MIA baselines; 2) inference scripts for the FiMMIA model; 3) a training pipeline for new datasets. Main system components are shown at Figure 1. We describe the general pipeline for multimodal MIA in subsection 3.4.

3.2 Multimodal membership inference benchmarks suffer from distribution shifts

Recently, (Das et al., 2024) have evaluated common textual membership inference benchmarks using blind statistical methods, and have found that they suffer from distribution shifts, with baseline methods independent of any target model's output outperforming best membership inference attacks on these datasets. An introduction of embedding model into the pipeline (Mozaffari and Marathe, 2024; Hu et al., 2022c) obviously makes the matter even worse, as they shine in tasks related to the separation of different distributions. This fact has, e.g. been recently utilized by (Miyamoto et al., 2025), who have also acknowledged the problem, and used a DINO-V2 (Oquab et al., 2023) to extract image features to show that VL-MIA member and non-member data suffer from a distributional mismatch introduced by the generative nature of non-member samples with an AUC-ROC of **94.9%** using their method. There are reasons for us to argue against this approach. Foremost, the usage of advanced deep learning model still poses threats alike the ones outlined above. Thus, we extend the work of (Das et al., 2024) to multimodal data and, to our surprise, find that attacks that directly

[3] https://github.com/ai-forever/data_leakage_detect

[4] https://huggingface.co/collections/ai-forever/fimmia

use features obtained from the dataset samples in absence of any information from the target model outperform best known attacks on most multimodal MIA benchmarks.

3.3 Distribution shift detection & baseline attacks

Essentially, for each input sample from the dataset with specified members and non-members we extract common heuristic (e.g. SIFT, LBP histogram) or spectral features, and them as inputs to a shallow ML model (e.g. logistic regression or gradient boosting)[5]. The model is trained on 5-fold cross-validation splits with the final attack score for each dataset taken as an average of ones obtained across folds. We assume that if both members and non-members come from the same distribution, i.e. the assumption of i.i.d. samples is valid, then this type of attacks should fail, showing AUC-ROC around 50%. Otherwise, if data collection method was biased (e.g. due to temporal differences, different data generation processes or other factors), these baseline attacks should serve as a lower bound for the proposed membership inference approaches.

We evaluated recently proposed MIA benchmarks in text (Fu et al., 2025; Hallinan et al., 2025) and image (Li et al., 2024c) modalities using the proposed method, and found that most of them suffer from severe distribution shifts, making them hardly useful to evaluate MIAs, with only LAION-MI (Dubiński et al., 2023) being mostly unaffected. See Table 1. Thus, in order to ensure credible results, we aim to use random splits of recently open-sourced multimodal datasets for Russian language (Chervyakov et al., 2025) in our further experiments. Although we are unaware of any common MIA benchmarks for audio or video data, we release both image and audio pipelines and encourage the community to use them prior to the release of new MIA datasets.

3.4 Methodology

Membership inference attacks (MIAs) against LLMs aim to determine whether given a target model $\mathcal{M}$ and a given data point was part of the training dataset used to train the target model. Given a multimodal sample $x = (t, s)$ from the dataset $D \sim \mathcal{P}(\mathcal{T} \times \mathcal{S})$ where $s \in \mathcal{S}$ is some modality (image/video/audio), $t \in \mathcal{T}$ is the text, estimate $\mathbb{P}(x \in D|\mathcal{M})$, probability that a target model was trained on x.

In accordance with the original article (Mozaffari and Marathe, 2024), we divided the training algorithm into the following subsequent steps with some modifications:

1. Neighbor generation
2. Embedding generation
3. Loss computation
4. Training the attack model

3.4.1 Neighbor and embedding generation

For each original data point (t, s) we generate $K = 24$ perturbed "neighbors" $(t_{\prime}^{k}, s_{\prime}^{k})$. Recently, there have been increasing attempts to link adversarial theory of neural networks to membership inference, arguing for the special local properties of the loss function in the neighborhood of each input (Xue et al., 2025; Ali et al., 2023). However, there have been several reasons for us to refrain from this approach: generating adversarial examples in discrete domains faces challenges due to non-differentiability (Yang et al., 2020) and generally necessitates to assume a white-box access to the target model, which was against our design principles. Moreover, recently (Gupta et al., 2025) have shown that adverarial examples for MMLMs are not generally transferable, which would additionally limit the applicability of our framework and its transferability across models. Instead, we've performed 4 different structured perturbations to the untokenized input string t:

1. Random masking and sampling masked words with Fred-T5 model [6]
2. Removing random words
3. Duplication of random words
4. Swapping random words

Each technique is applied to the each text sample t 6 times, resulting in totally 24 "neighbors" per sample. Although, in our experiments we fix $s = s_{\prime}^{k}$, $\forall s \in D$, so the modality data remains unchanged, the pipeline can be modified to support neighbors from different modalities as well.

Then for each original text t and its neighbors $t_{\prime}^{k}$ we extract their text embeddings using a fixed encoder:

$$e = \mathcal{E}(t), \quad e'_k = \mathcal{E}(t'_k)$$

where $\mathcal{E}$ is `intfloat/e5-mistral-7b-instruct` [7].

[5] Details on the design of distribution shift detection pipeline and features extracted are available at A.6

[6] ai-forever/FRED-T5-1.7B, (Zmitrovich et al., 2024)

[7] intfloat/e5-mistral-7b-instruct in our experiments. It used

3.4.2 Loss computation

We compute the multimodal loss for both models $\mathcal{M}$ and $\mathcal{M}_{leak}$ on both the original and neighbor data points:

$$\mathcal{L} = \mathcal{L}(\mathcal{M}, t, s), \quad \mathcal{L}'_k = \mathcal{L}(\mathcal{M}, t^k_{\prime}, s^k_{\prime})$$

Text input t is provided to each model, accompanied by the corresponding modality s (image, video, or audio data in its original, unchanged form).

3.4.3 Attack model training

The core of FiMMIA is a binary neural network classifier trained to distinguish between models that have and have not seen the data. For each neighbor k we create two training examples by computing feature differences[8]:

$$\Delta\mathcal{L} = \mathcal{L} - \mathcal{L}^k_{\prime}, \quad \Delta e = e - e^k_{\prime}$$

These feature vectors are paired with labels $y \in \{0, 1\}$ indicating whether the losses came from $\mathcal{M}$ (non-leaked) or $\mathcal{M}_{leak}$ (leaked). However, absolute values of these statistics may vary across datasets and models. To make the system more stable, we apply the z-score normalization technique (Wikipedia, 2025). The values mean μ and standard deviation σ of the models' loss differences $\Delta\mathcal{L}$, used to normalize input features during training and evaluation are obtained from disjoint train/test splits to mimic real-world scenarios.

$$\Delta\mathcal{L}_{norm} = \frac{\Delta\mathcal{L} - \mu}{\sigma}$$

.

This process yields random batch training triplets $(\Delta\mathcal{L}_{norm}, \Delta e, y)$ per original data point. The FiMMIA detector, f_{FiMMIA} is trained to predict the probability $p = f_{FiMMIA}(\Delta\mathcal{L}_{norm}, \Delta e)$ that the input features originate from a model that has been trained on the target data. We provide the details of the architecture for FiMMIA model in subsection A.1 and the hyperparameters for training the FiMMIA model in subsection A.2.

It should be noted, that although we suppose a grey-box access to the MLLM in our experiments, i.e. an attacker has full access to the model's logprobs for loss computation, our setup can be extended to the black-box scenario in presence of compatible APIs, with e.g. only top-k logprobs being released, using approaches from (Finlayson et al., 2024; Bao et al., 2025). We plan to implement such functionality in future releases.

3.4.4 Inference

To infer if a target model $\mathcal{M}'$ has been trained on a specific data point (t, s), we compute the loss and embedding differences for this model. We then compute the leakage score A for the data point by taking the average probability output by the detector over all K neighbors:

$$A(t, m) = \frac{1}{K} \sum_{k=1}^{K} f_{FiMMIA}(\Delta\mathcal{L}^k_{norm}, \Delta e^k)$$

4 Experiment setup

4.1 Data

We evaluate our method on the MERA benchmark (Chervyakov et al., 2025), which comprises 18 audio, video, and image datasets. All tasks in the benchmark are multimodal, taking both a modality input and an instruction, and requiring a text output in a constrained format (e.g., multiple-choice or short-answer). For training phase we fine-tune MLLM $\mathcal{M}_{leak}$ on each modality separately. Each sample in the training data for the MLLM can be represented as $x = (s, q, a)$, a concatenation of the question and the answer as the textual part t, along with the multimodal input s (image, video, or audio). In order to ensure credible evaluation of FiMMIA model we split each dataset into train and test parts randomly. The size of the test part is 10% of original dataset. Normalization parameters $\mu_{D,\mathcal{M}}$ and $\sigma_{D,\mathcal{M}}$ are calculated from the train part of each of the splitted datasets for each model.

The detailed overview of the benchmark is presented in Table 2.

4.2 Models

We evaluate 9 publicly available multimodal models from the most trending model families on HuggingFace, varying in size from 3B to 12B parameters. See Appendix A.3 for detailed model descriptions.

4.3 Cross-lingual transfer

This section presents our experimental evaluation, extending the pipeline to English image datasets

to be SoTA on the MTEB benchmark (Muennighoff et al., 2022) at the time of the model experiments

[8] Similar ideas has been already explored e.g. in (He et al., 2024) where the authors explored both utilizing shadow models and perturbed datasets as calibration data, and found that they are, to a large degree, interchangeable. The idea of using embedding differences as a proxy for difficulty calibration serves as another intuition for our method.

	Dataset / task	Size	Answer
audio	ruEnvAQA	596	MC
	RuSLUn	741	OE
	*AQUARIA	738	MC
	*ruTiE-Audio	1500	MC
image	ruCLEVR	1148	OE
	ruCommonVQA	3015	OE
	ruNaturalScienceVQA	363	MC
	WEIRD	814	MC
	*LabTabVQA	339	MC
	*RealVQA	773	OE
	*ruHHH-Image	595	MC
	*ruMathVQA	502	OE
	*ruTiE-Image	1500	MC
	*SchoolScienceVQA	4227	MC
	*UniScienceVQA	7432	OE
video	CommonVideoQA	907	MC
	*RealVideoQA	671	MC
	*ruHHH-Video	911	MC

Table 2: Overview of datasets in MERA benchmark. Those marked with an asterisk were collected from scratch by Chervyakov et al. (2025), while the others are *public datasets* compiled from open-source datasets. **Size** column shows the number of samples in the dataset, and **Answer** column is the task format (MC and OE stand for multiple-choice and open-ended, respectively).

and models. Following the paper by (Song et al., 2025), our analysis leverages two multi-choice datasets: ScienceQA (Lu et al., 2022) and MM-Star (Chen et al., 2024), along with caption dataset: COCO-Caption2017 (Lin et al., 2015). We randomly selected 2000 samples from ScienceQA's test set, respectively, with 1000 samples from the other datasets. We select Qwen2.5-VL-3B-Instruct as a target fine-tuned MLLM and train FiMMIA as described in section subsection 3.4 only on MERA benchmark (Chervyakov et al., 2025) without fine-tuning or using any English data. We evaluate 4 publicly available multimodal models similar to the paper (Song et al., 2025) that presents MM-DETECT method (see Table 9 for model descriptions). That method calculates Δ score for the dataset and if $\Delta < 0$, dataset leakage is presumed. In order to make a comparison with this method we calculate % of leaked samples from the dataset, guided by our pipeline.

5 Results

We report AUC-ROC for binary classification (leaked vs. clean) as shown in Tables 3, 5, 4. Also we report TPR with low FPR in Tables 12, 10, 11 .In order to evaluate the transferability of the trained attack model we also report scores when the origin and test models differ. The $\mathcal{M}_{\texttt{origin}}$ is the model used to train FiMMIA, while $\mathcal{M}_{\texttt{test}}$ is the model whose losses are used to test FiMMIA (predict whether the data sample was used to train $\mathcal{M}_{\texttt{test}}$ or not).

$\mathcal{M}_{\texttt{origin}}$	$\mathcal{M}_{\texttt{test}}$	AUC
Qwen2.5-VL-3B-Instruct	Qwen2.5-VL-3B-Instruct	**96.2**
Qwen2.5-VL-3B-Instruct	Qwen2-VL-7B-Instruct	86.0
Qwen2.5-VL-3B-Instruct	Qwen2.5-VL-7B-Instruct	88.0
Qwen2.5-VL-3B-Instruct	Llava-Next-8b-hf	*90.2*
Qwen2.5-VL-3B-Instruct	Gemma-3-4B-it	65.8
Qwen2.5-VL-3B-Instruct	Gemma-3-12b-it	67.9
Qwen2-VL-7B-Instruct	Qwen2.5-VL-3B-Instruct	78.0
Qwen2-VL-7B-Instruct	Qwen2-VL-7B-Instruct	**96.2**
Qwen2-VL-7B-Instruct	Qwen2.5-VL-7B-Instruct	*80.5*
Qwen2-VL-7B-Instruct	Llama3-llava-next-8b-hf	78.0
Qwen2-VL-7B-Instruct	Gemma-3-4b-it	77.7
Qwen2-VL-7B-Instruct	Gemma-3-12b-it	73.7
Qwen2.5-VL-7B-Instruct	Qwen2.5-VL-3B-Instruct	92.8
Qwen2.5-VL-7B-Instruct	Qwen2-VL-7B-Instruct	93.1
Qwen2.5-VL-7B-Instruct	Qwen2.5-VL-7B-Instruct	**98.1**
Qwen2.5-VL-7B-Instruct	Llama3-llava-next-8b-hf	*95.8*
Qwen2.5-VL-7B-Instruct	Gemma-3-4b-it	95.4
Qwen2.5-VL-7B-Instruct	Gemma-3-12b-it	94.5
Llama3-llava-next-8b-hf	Qwen2.5-VL-3B-Instruct	94.6
Llama3-llava-next-8b-hf	Qwen2-VL-7B-Instruct	90.0
Llama3-llava-next-8b-hf	Qwen2.5-VL-7B-Instruct	96.6
Llama3-llava-next-8b-hf	Llama3-llava-next-8b-hf	*97.7*
Llama3-llava-next-8b-hf	Gemma-3-4b-it	99.1
Llama3-llava-next-8b-hf	Gemma-3-12b-it	**99.5**
Gemma-3-4b-it	Qwen2.5-VL-3B-Instruct	76.0
Gemma-3-4b-it	Qwen2-VL-7B-Instruct	71.5
Gemma-3-4b-it	Qwen2.5-VL-7B-Instruct	85.2
Gemma-3-4b-it	Llama3-llava-next-8b-hf	86.5
Gemma-3-4b-it	Gemma-3-4b-it	**99.4**
Gemma-3-4b-it	Gemma-3-12b-it	*98.7*
Gemma-3-12b-it	Qwen2.5-VL-3B-Instruct	84.1
Gemma-3-12b-it	Qwen2-VL-7B-Instruct	81.3
Gemma-3-12b-it	Qwen2.5-VL-7B-Instruct	91.2
Gemma-3-12b-it	Llama3-llava-next-8b-hf	93.3
Gemma-3-12b-it	Gemma-3-4b-it	*99.4*
Gemma-3-12b-it	Gemma-3-12b-it	**99.7**

Table 3: AUC-ROC FiMMIA performance metrics for various evaluated **Image** MLLMs.

$\mathcal{M}_{\texttt{origin}}$	$\mathcal{M}_{\texttt{test}}$	AUC
Qwen2.5-VL-3B-Instruct	Qwen2.5-VL-3B-Instruct	*95.9*
Qwen2.5-VL-3B-Instruct	Qwen2.5-VL-7B-Instruct	**99.5**
Qwen2.5-VL-3B-Instruct	LLaVA-NeXT-Video	91.7
Qwen2.5-VL-3B-Instruct	LLaVA-NeXT-Video-DPO	91.2
Qwen2.5-VL-7B-Instruct	Qwen2.5-VL-3B-Instruct	*98.7*
Qwen2.5-VL-7B-Instruct	Qwen2.5-VL-7B-Instruct	**100.0**
Qwen2.5-VL-7B-Instruct	LLaVA-NeXT-Video	96.5
Qwen2.5-VL-7B-Instruct	LLaVA-NeXT-Video-DPO	95.7
LLaVA-NeXT-Video	Qwen2.5-VL-3B-Instruct	63.7
LLaVA-NeXT-Video	Qwen2.5-VL-7B-Instruct	*71.5*
LLaVA-NeXT-Video	LLaVA-NeXT-Video	**100.0**
LLaVA-NeXT-Video	LLaVA-NeXT-Video-DPO	**100.0**
LLaVA-NeXT-Video-DPO	Qwen2.5-VL-3B-Instruct	53.6
LLaVA-NeXT-Video-DPO	Qwen2.5-VL-7B-Instruct	*56.2*
LLaVA-NeXT-Video-DPO	LLaVA-NeXT-Video	**100.0**
LLaVA-NeXT-Video-DPO	LLaVA-NeXT-Video-DPO	**100.0**

Table 4: AUC-ROC FiMMIA performance metrics for various evaluated **Video** MLLMs.

Overall, the results of the FiMMIA detection capabilities are presented in Table 6. All models show significant success within their own family; however, the success of the attack may decrease when testing on a model from a different family. Nev-

$\mathcal{M}_{\texttt{origin}}$	$\mathcal{M}_{\texttt{test}}$	AUC
Qwen2-Audio-7B-Instruct	Qwen2-Audio-7B-Instruct	**87.7**
Qwen2-Audio-7B-Instruct	Qwen-Audio-Chat	*76.0*
Qwen-Audio-Chat	Qwen2-Audio-7B-Instruct	*61.3*
Qwen-Audio-Chat	Qwen-Audio-Chat	**100.0**

Table 5: AUC-ROC FiMMIA performance metrics for various evaluated **Audio** MLLMs.

ertheless, the metric score for each experiment exceeds 65.0, which indicates the promising transferability of the proposed method. Moreover, average metrics for each modality are quite high, ranging from 80 to 90% AUC-ROC.

Modality	AUC
Image	88.658
Video	88.388
Audio	81.250

Table 6: Average AUC-ROC of FiMMIA per modality. Averaging over the models used for training and evaluating FiMMIA.

Evaluations on the transferability of the model to a different language inputs are presented in Table 7. The results indicate that our method is almost entirely in agreement with those presented in the paper (Song et al., 2025). If $\Delta < 0$ the amount of samples predicted by FiMMIA as leaked is more than 0.1 in most cases, which corresponds to at least 10% of the dataset. However, if the task allows, we suggest to train FiMMIA for particular dataset and language from scratch to obtain more accurate and reliable results.

Dataset	Model	FiMMIA	MM-DETECT Δ
COCO	Phi-3-vision-128k-instruct	0.00	0.5
	Qwen-VL-Chat	0.00	-1.9
	LLaVA-1.5-7B	0.58	-0.6
	fuyu-8b	0.22	1.0
MMStar	Phi-3-vision-128k-instruct	0.06	3.2
	Qwen-VL-Chat	0.00	3.3
	LLaVA-1.5-7B	0.13	2.8
	fuyu-8b	0.011	-1.2
ScienceQA	Phi-3-vision-128k-instruct	0.10	0.7
	Qwen-VL-Chat	0.00	0.1
	LLaVA-1.5-7B	0.21	1.3
	fuyu-8b	0.19	-0.5

Table 7: Comparison FiMMIA % leakage samples detected of MLLMs on English datasets with MM-DETECT score for image modality.

6 Conclusion

This paper introduces FiMMIA, a novel framework that leverages input semantics and strategic perturbations to train a highly effective neural network for data leakage detection in MLLMs. Our key contribution is a language-agnostic system capable of training robust leakage detection models for any dataset. Designed for extensibility, the framework natively supports neighbor generation across multiple modalities paving the way for future research.

Limitations

Scope of the Method When training FiMMIA, we only target a fine-tuning scenario for the MLLM using a low-rank adapter. The results for pretraining and full fine-tuning may be different due to the capacity scaling laws (Morris et al., 2025), and other factors. We leave these evaluations for further work.

Determinism and Reproducibility Even our fine-tuned models' losses are subject to stochasticity, as the entire hardware–software stack affects inference: GPU model, drivers/CUDA/cuDNN, PyTorch, vLLM/transformers (and commit hashes), flash-attention kernels, tokenizers/checkpoints, precision/quantization, and batching – some of which are non-deterministic or can vary between environments. However, in general, the variance that these factors contribute to evaluation metrics is not substantial.

Speed and Computational Complexity In our experiments the inference process took appx. 10 hours on a single GPU for one dataset. Generally, the time complexity of our algorithm scales as $\mathcal{O}(|D|N(M+E+G))$, where $|D|$ is the number of samples in the dataset, N is the number of neighbors, and M, E, G are time complexities of the target, embedding and neighbor generation models.

Model Assumption Dependencies The method relies on per-sample loss access (a gray-box assumption) and depends on an external model for generating embeddings. The applicability of the method in a strict black-box setting, where such access is unavailable, is not addressed in this work, despite the existence of relevant prior research.

Ethical consideration

Use of Public Data All experiments and evaluations in this study rely exclusively on openly accessible public datasets. No proprietary, confidential, or otherwise sensitive information was involved. This choice supports transparency, facilitates independent verification, and avoids any infringement on data-privacy protections.

Defensive and Constructive Purpose Our work reconceptualizes membership-inference analysis as a diagnostic and privacy-protecting tool rather than a privacy-threat vector. The method is designed to:

- By identifying cases in which benchmark samples have been inadvertently memorized during training, the approach helps prevent benchmark saturation and dataset contamination, thereby supporting fair and meaningful model comparison.
- The technique offers researchers a practical mechanism for auditing training pipelines to ensure that performance improvements stem from genuine advances rather than overfitting to widely used evaluation sets.
- As competitive leaderboard dynamics can unintentionally encourage data leakage and undermine the long-term value of public benchmarks, our framework contributes to more resilient evaluation standards that promote steady, reliable scientific progress.

Acknowledgments

The authors would like to express their sincere gratitude to Dmitry Gorbetsky, Yaroslav Grebnyak, Oleg Yangalichin and Artem Chervyakov for their valuable contributions and support in this work.

References

Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S. Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *CoRR*, abs/2404.14219.

Hassan Ali, Adnan Qayyum, Ala Al-Fuqaha, and Junaid Qadir. 2023. Membership inference attacks on dnns using adversarial perturbations. *arXiv preprint arXiv: 2307.05193*.

Junxing Bai, Shanshan Bai, Shiqi Yang, Shi Wang, Shoujie Tan, Panpan Wang, Jiawei Lin, Chaozhe Zhou, and Junrui Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

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. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Guangsheng Bao, Yanbin Zhao, Juncai He, and Yue Zhang. 2025. Glimpse: Enabling white-box methods to use proprietary models for zero-shot llm-generated text detection. *Preprint*, arXiv:2412.11506.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. Membership inference attacks from first principles. *Preprint*, arXiv:2112.03570.

Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. 2020. GAN-Leaks: A taxonomy of membership inference attacks against generative models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pages 343–362.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and et al. 2024. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Zequn Chen, Jiannan Wu, Wenqian Wang, Weijiang Su, Guangda Chen, Shen Xing, Mingyang Zhong, Qing Zhang, Xin Zhu, Lei Lu, Bo Li, Peihao Luo, Tong Lu, Yi Qiao, and Jifeng Dai. 2023. Internvl: Scaling up vision foundation models and aligning for generic visuo-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Artem Chervyakov, Ulyana Isaeva, Anton Emelyanov, Artem Safin, Maria Tikhonova, Alexander Kharitonov, Yulia Lyakh, Petr Surovtsev, Denis Shevelev, Vildan Saburov, Vasily Konovalov, Elisei Rykov, Ivan Sviridov, Amina Miftakhova, Ilseyar Alimova, Alexander Panchenko, Alexander Kapitanov, and Alena Fenogenova. 2025. Multimodal evaluation of Russian-language architectures. *Preprint*, arXiv:2511.15552.

Christopher A Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. 2021. Label-only membership inference attacks. In *International conference on machine learning*, pages 1964–1974. PMLR.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *Preprint*, arXiv:2407.10759.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Debeshee Das, Jie Zhang, and Florian Tramèr. 2024. Blind baselines beat membership inference attacks for foundation models. *arXiv preprint arXiv: 2406.16201*.

Jan Dubiński, Antoni Kowalczuk, Stanisław Pawlak, Przemysław Rokita, Tomasz Trzciński, and Paweł Morawiecki. 2023. Towards more realistic membership inference attacks on large diffusion models. *Preprint*, arXiv:2306.12983.

Matthew Finlayson, Xiang Ren, and Swabha Swayamdipta. 2024. Logits of api-protected llms leak proprietary information. *arXiv preprint arXiv: 2403.09539*.

Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. 2023. Practical membership inference attacks against fine-tuned large language models via self-prompt calibration. *arXiv preprint arXiv:2311.06062*.

Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. 2025. MIA-tuner: Adapting large language models as pre-training text detector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Philadelphia, Pennsylvania, USA.

Isha Gupta, Rylan Schaeffer, Joshua Kazdan, Ken Ziyu Liu, and Sanmi Koyejo. 2025. Understanding adversarial transfer: Why representation-space attacks fail where data-space attacks succeed. *arXiv preprint arXiv: 2510.01494*.

Skyler Hallinan, Jaehun Jung, Melanie Sclar, Ximing Lu, Abhilasha Ravichander, Sahana Ramnath, Yejin Choi, Sai Praneeth Karimireddy, Niloofar Mireshghallah, and Xiang Ren. 2025. The surprising effectiveness of membership inference with simple n-gram coverage. *arXiv preprint arXiv: 2508.09603*.

Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. 2017. Logan: Membership inference attacks against generative models. *arXiv preprint arXiv:1705.07663*.

Yu He, Boheng Li, Yao Wang, Mengda Yang, Juan Wang, Hongxin Hu, and Xingyu Zhao. 2024. Is difficulty calibration all we need? towards more practical membership inference attacks. *arXiv preprint arXiv: 2409.00426*.

Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. 2019. Monte carlo and reconstruction membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies*.

Haoyang Hu, Zoran Salcic, Lingyu Sun, Gillian Dobbie, Philip S. Yu, and Xinhui Zhang. 2022a. Membership inference attacks on machine learning: A survey. *ACM Comput. Surv.*, 54(11s):1–37.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. 2022b. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37.

Pingyi Hu, Zihan Wang, Ruoxi Sun, Hu Wang, and Minhui Xue. 2022c. M$\hat{4}$i: Multi-modal models membership inference. *arXiv preprint arXiv: 2209.06997*.

Pingyi Hu, Zihan Wang, Ruoxi Sun, Hu Wang, and Minhui Xue. 2022d. M^{4}i: Multi-modal models membership inference. In *Advances in Neural Information Processing Systems*, volume 35, pages 1867–1882.

Myeongseob Ko, Ming Jin, Chenguang Wang, and Ruoxi Jia. 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*, pages 4871–4881.

Ryuto Koike, Liam Dugan, Masahiro Kaneko, Chris Callison-Burch, and Naoaki Okazaki. 2025. Machine text detectors are membership inference attacks. *Preprint*, arXiv:2510.19492.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024a. Llava-next: Stronger llms supercharge multimodal capabilities in the wild.

Zhan Li, Yongtao Wu, Yihang Chen, Francesco Tonin, Elias Abad Rocamora, and Volkan Cevher. 2024b. Membership inference attacks against large vision-language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 98645–98674. Curran Associates, Inc.

Zhan Li, Yongtao Wu, Yihang Chen, Francesco Tonin, Elias Abad Rocamora, and Volkan Cevher. 2024c. Membership inference attacks against large vision-language models. *arXiv preprint arXiv: 2411.02902*.

Jiawei Lin, Haoqi Yin, Weiran Ping, Yu Lu, Pavlo Molchanov, Andrew Tao, Huiyu Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. 2023. Vila: On pre-training for visual language models. *Preprint*, arXiv:2312.07533.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft COCO: Common Objects in Context. *CoRR*, abs/1405.0312.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *CoRR*, abs/2310.03744.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

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

Yang Liu, Hongfan Duan, Yan Zhang, Binbin Li, Shaohan Zhang, Wei Zhao, Yonglong Yuan, Jinmao Wang, Chuxiong He, Zhongying Liu, Kechen Chen, and Dahua Lin. 2023b. Mmbench: Is your multi-modal model an all-around player? *Preprint*, arXiv:2307.06281.

Yunhui Long, Vincent Bindschaedler, Lei Wang, Diyue Bu, Xiaofeng Wang, Haixu Tang, Carl A Gunter, and Kai Chen. 2018. Understanding membership inferences on well-generalized learning models. *arXiv preprint arXiv:1802.04889*.

Pengcheng Lu, Sudip Mishra, Tianyi Xia, Leqi Qiu, Kai-Wei Chang, Scott Cheng-Hsin Zhu, Oyvind Tafjord, Peter Clark, and Anirudh Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Saeed Mahloujifar, Huseyin A Inan, Melissa Chase, Esha Ghosh, and Marcello Hasegawa. 2021. Membership inference on word embedding and beyond. *arXiv preprint arXiv:2106.11384*.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. 2023. Cross-entropy loss functions: Theoretical analysis and applications. *arXiv preprint arXiv:2304.07288*. Published in ICML 2023.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schoelkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. 2023. Membership inference attacks against language models via neighbourhood comparison. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11330–11343.

Fatemehsadat Mireshghallah, Archit Uniyal, Tianhao Wang, David K Evans, and Taylor Berg-Kirkpatrick. 2022. An empirical analysis of memorization in fine-tuned autoregressive language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1816–1826.

Ryoto Miyamoto, Xin Fan, Fuyuko Kido, Tsuneo Matsumoto, and Hayato Yamana. 2025. Openlvlm-mia: A controlled benchmark revealing the limits of membership inference attacks on large vision-language models. *arXiv preprint arXiv: 2510.16295*.

Mazda Moayeri, Kiarash Banihashem, and Soheil Feizi. 2022. Explicit tradeoffs between adversarial and natural distributional robustness. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

John X. Morris, Chawin Sitawarin, Chuan Guo, Narine Kokhlikyan, G. Edward Suh, Alexander M. Rush, Kamalika Chaudhuri, and Saeed Mahloujifar. 2025. How much do language models memorize? *arXiv preprint arXiv: 2505.24832*.

Hamid Mozaffari and Virendra Marathe. 2024. Semantic membership inference attack against large language models. In *Neurips Safe Generative AI Workshop 2024*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814. Omnipress.

OpenAI. 2023. GPT-4 technical report. CoRR, abs/2303.08774.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv: 2304.07193*.

Maxwell Reid, Nikita Savinov, Dmitry Teplyashin, Danil Lepikhin, Timothy Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. 2018. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, volume 80, pages 4596–4604. PMLR.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18.

Congzheng Song and Ananth Raghunathan. 2020. Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pages 377–390.

Dandan Song, Shuai Chen, Guanhua Chen, Fan Yu, Xiuyue Wan, and Bo Wang. 2024. Milebench: Benchmarking mllms in long context. *arXiv preprint arXiv:2404.18532*.

Dingjie Song, Sicheng Lai, Mingxuan Wang, Shunian Chen, Lichao Sun, and Benyou Wang. 2025. Both text and images leaked! a systematic analysis of data contamination in multimodal LLM. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10527–10542, Suzhou, China. Association for Computational Linguistics.

Liwei Song and Prateek Mittal. 2021. Systematic evaluation of privacy risks of machine learning models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2615–2632.

Liwei Song, Reza Shokri, and Prateek Mittal. 2019. Membership inference attacks against adversarially robust deep learning models. In *2019 IEEE Security and Privacy Workshops (SPW)*, pages 50–56. IEEE.

Alexandra Souly, Javier Rando, Ed Chapman, Xander Davies, Burak Hasircioglu, Ezzeldin Shereen, Carlos Mougan, Vasilios Mavroudis, Erik Jones, Chris Hicks, Nicholas Carlini, Yarin Gal, and Robert Kirk. 2025. Poisoning attacks on llms require a near-constant number of poison samples. *arXiv preprint arXiv: 2510.07192*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Teun van der Weij, Felix Hofstätter, Ollie Jaffe, Samuel F. Brown, and Francis Rhys Ward. 2024. Ai sandbagging: Language models can strategically underperform on evaluations. *arXiv preprint arXiv: 2406.07358*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Venkat Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. 2025. Livebench: A challenging, contamination-free LLM benchmark. In *The Thirteenth International Conference on Learning Representations*.

Wikipedia. 2025. Standard score — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Standard_score. [Online; accessed 17-November-2025].

Rui Xu, Ze Wang, Ren-Zhang Fan, and Pengfei Liu. 2024. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*.

Jing Xue, Zhishen Sun, Haishan Ye, Luo Luo, Xiangyu Chang, Ivor Tsang, and Guang Dai. 2025. Privacy leaks by adversaries: Adversarial iterations for membership inference attack. *arXiv preprint arXiv: 2506.02711*.

Puyudi Yang, Jianbo Chen, Cho-Jui Hsieh, Jane-Ling Wang, and Michael I Jordan. 2020. Greedy attack and gumbel attack: Generating adversarial examples for discrete data. *Journal of Machine Learning Research*, 21(43):1–36.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282.

Jinhua Yin, Peiru Yang, Chen Yang, Huili Wang, Zhiyang Hu, Shangguang Wang, Yongfeng Huang, and Tao Qi. 2025. Black-box membership inference attack for lvlms via prior knowledge-calibrated memory probing. *arXiv preprint arXiv: 2511.01952*.

Dmitry Zmitrovich, Aleksandr Abramov, Andrey Kalmykov, Vitaly Kadulin, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergei S. Markov, Vladislav Mikhailov, and Alena Fenogenova. 2024. A family of pretrained transformer language models for Russian. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 507–524, Torino, Italia. ELRA and ICCL.

A Appendix

A.1 Attack model neural network architecture

The detailed architecture of the FiMMIA is provided below.

1. **Input Data:**
    - `loss_input`: A tensor fed into the `loss_component`.
    - `embedding_input`: A tensor fed into the `embedding_component`.
2. **`loss_component`:**
    - A Linear layer: 1 input feature → `projection_size` output features.
    - Dropout(0.2) and ReLU (Nair and Hinton, 2010) activation.
3. **`embedding_component`:**
    - A Linear layer: `embedding_size` → `embedding_size // 2`.
    - Dropout(0.2) and ReLU (Nair and Hinton, 2010) activation.
    - A Linear layer: `embedding_size // 2` → 512.
    - Dropout(0.2) and ReLU (Nair and Hinton, 2010) activation.
4. **Concatenation (`torch.hstack`):**
    - The outputs from the `loss_component` (`projection_size`) and the `embedding_component`(512) are concatenated into a single vector of size `2 * projection_size`.
5. **`attack_encoding`:**

- A series of 6 fully connected Linear layers with Dropout(0.2) and ReLU (Nair and Hinton, 2010) activations between them: `2 * projection_size` → 512 → 256 → 128 → 64 → 32.
- The final Linear layer: 32 → 2 (output logits for classification).

6. **Output:**

- The model returns the logits (size 2).
- If labels are provided, it also calculates and returns the cross-entropy loss (Mao et al., 2023).

A.2 Attack model hyperparameters

To construct the neighbor datasets, we generate $k = 24$ neighbors for each data point. We employ the adafactor optimizer (Shazeer and Stern, 2018) to train the network on our training data over 10 epochs. The batch size is set to 64, meaning each batch contains random triplets. For experiments, we use a learning rate of 2×10^{-6}.

A.3 Models Details

Table 8 contains information about multimodal LLMs used for the experiments. As the number of MLLMs trained with a focus on russian is limited, we evaluate our method using known open-source models. Although it may contribute to higher ROC-AUC scores we observe in our experiments due to the models being adapted to vastly new domain, it also helps us alleviate possible effects related to the possibility of our evaluation datasets' traces being already present in models' training data.

A.4 English Models Details

Table 9 contains information about multimodal LLMs used for the language transfer experiments. All models are selected from the following paper (Song et al., 2025).

A.5 TPR at low FPR (FPR=5%) results

Here we report the True Positive Rate (TPR) at a low False Positive Rate (FPR), which measures the detection rate at a meaningful threshold. The modality of image is presented in Table 12, the video in Table 10 and the audio accordingly in Table 11.

A.6 Description of the distribution shift detection pipelines

For the information on the features extracted from image and audio data see Table 13.

Model	Parameters	Context length	Hugging Face Hub link	Citation
Qwen2-VL-7B-Instruct	7B	32K	Qwen/Qwen2-VL-7B-Instruct	Wang et al. (2024)
Qwen2.5-VL-3B-Instruct	3B	128K	Qwen/Qwen2.5-VL-3B-Instruct	Bai et al. (2025)
Qwen2.5-VL-7B-Instruct	7B	128K	Qwen/Qwen2.5-VL-7B-Instruct	
gemma-3-4b-it	4B	128K	google/gemma-3-4b-it	Team et al. (2025)
gemma-3-12b-it	12B	128K	google/gemma-3-12b-it	
llama3-llava-next-8b-hf	8B	128K	llava-hf/llama3-llava-next-8b-hf	Li et al. (2024a)
LLaVA-NeXT-Video	7B	4K	llava-hf/LLaVA-NeXT-Video-7B-hf	Liu et al. (2024b)
LLaVA-NeXT-Video-DPO	7B	4K	llava-hf/LLaVA-NeXT-Video-7B-DPO-hf	
Qwen2-Audio-7B-Instruct	7B	32K	Qwen/Qwen2-Audio-7B-Instruct	Chu et al. (2024)
Qwen/Qwen-Audio-Chat	7B	32K	Qwen/Qwen-Audio-Chat	Chu et al. (2023)

Table 8: General information about used multimodal LLMS for experiments.

Model	Parameters	Context length	Hugging Face Hub link	Citation
Phi-3-vision-128k-instruct	8B	128K	microsoft/Phi-3-vision-128k-instruct	(Abdin et al., 2024)
LLaVA-1.5-7B	7B	16K	llava-hf/llava-1.5-7b-hf	(Liu et al., 2024a)
Qwen-VL-Chat	7B	8K	Qwen-VL-Chat	(Bai et al., 2023)
fuyu-8b[9]	8B	16K	adept/fuyu-8b	

Table 9: General information about used multimodal LLMS used for the language transfer experiments.

$\mathcal{M}_{\texttt{origin}}$	$\mathcal{M}_{\texttt{test}}$	AUC	TPR
Qwen2.5-VL-3B-Instruct	Qwen2.5-VL-3B-Instruct	95.9	85.8
Qwen2.5-VL-3B-Instruct	Qwen2.5-VL-7B-Instruct	99.5	98.4
Qwen2.5-VL-3B-Instruct	LLaVA-NeXT-Video	91.7	52.9
Qwen2.5-VL-3B-Instruct	LLaVA-NeXT-Video-DPO	91.2	62.9
Qwen2.5-VL-7B-Instruct	Qwen2.5-VL-3B-Instruct	98.7	95.4
Qwen2.5-VL-7B-Instruct	Qwen2.5-VL-7B-Instruct	100.0	100.0
Qwen2.5-VL-7B-Instruct	LLaVA-NeXT-Video	96.5	80.8
Qwen2.5-VL-7B-Instruct	LLaVA-NeXT-Video-7B-DPO	95.7	82.1
LLaVA-NeXT-Video	Qwen2.5-VL-3B-Instruct	63.7	6.0
LLaVA-NeXT-Video	Qwen2.5-VL-7B-Instruct	71.5	70.0
LLaVA-NeXT-Video	LLaVA-NeXT-Video-7B	100.0	100.0
LLaVA-NeXT-Video	LLaVA-NeXT-Video-7B-DPO	100.0	100.0
LLaVA-NeXT-Video-7B-DPO	Qwen2.5-VL-3B-Instruct	53.6	60.0
LLaVA-NeXT-Video-7B-DPO	Qwen2.5-VL-7B-Instruct	56.2	43.0
LLaVA-NeXT-Video-7B-DPO	LLaVA-NeXT-Video-7B	100.0	100.0
LLaVA-NeXT-Video-7B-DPO	LLaVA-NeXT-Video-7B-DPO	100.0	100.0

Table 10: AUC-ROC and TPR at low FPR (FPR=5%) FiMMIA performance metrics for various evaluated **Video** MLLMs.

$\mathcal{M}_{\texttt{origin}}$	$\mathcal{M}_{\texttt{test}}$	AUC	TPR
Qwen2-Audio-7B-Instruct	Qwen2-Audio-7B-Instruct	87.7	61.9
Qwen2-Audio-7B-Instruct	Qwen-Audio-Chat	76.0	74.5
Qwen-Audio-Chat	Qwen2-Audio-7B-Instruct	61.3	62.7
Qwen-Audio-Chat	Qwen-Audio-Chat	100.0	100.0

Table 11: AUC-ROC and TPR at low FPR (FPR=5%) FiMMIA performance metrics for various evaluated **Audio** MLLMs.

$\mathcal{M}_{\texttt{origin}}$	$\mathcal{M}_{\texttt{test}}$	AUC	TPR
Qwen2.5-VL-3B-Instruct	Qwen2.5-VL-3B-Instruct	96.2	86.1
Qwen2.5-VL-3B-Instruct	Qwen2-VL-7B-Instruct	86.0	39.1
Qwen2.5-VL-3B-Instruct	Qwen2.5-VL-7B-Instruct	88.0	53.0
Qwen2.5-VL-3B-Instruct	llama3-llava-next-8b-hf	90.2	59.9
Qwen2.5-VL-3B-Instruct	gemma-3-4b-it	65.8	6.2
Qwen2.5-VL-3B-Instruct	gemma-3-12b-it	67.9	61.9
Qwen2-VL-7B-Instruct	Qwen2.5-VL-3B-Instruct	78.0	16.5
Qwen2-VL-7B-Instruct	Qwen2-VL-7B-Instruct	96.2	85.1
Qwen2-VL-7B-Instruct	Qwen2.5-VL-7B-Instruct	80.5	35.9
Qwen2-VL-7B-Instruct	llama3-llava-next-8b-hf	78.0	30.6
Qwen2-VL-7B-Instruct	gemma-3-4b-it	77.7	7.2
Qwen2-VL-7B-Instruct	gemma-3-12b-it	73.7	67.8
Qwen2.5-VL-7B-Instruct	Qwen2.5-VL-3B-Instruct	92.8	73.8
Qwen2.5-VL-7B-Instruct	Qwen2-VL-7B-Instruct	93.1	77.0
Qwen2.5-VL-7B-Instruct	Qwen2.5-VL-7B-Instruct	98.1	94.0
Qwen2.5-VL-7B-Instruct	llama3-llava-next-8b-hf	95.8	83.1
Qwen2.5-VL-7B-Instruct	gemma-3-4b-it	95.4	71.8
Qwen2.5-VL-7B-Instruct	gemma-3-12b-it	94.5	66.1
llama3-llava-next-8b-hf	Qwen2.5-VL-3B-Instruct	94.6	78.6
llama3-llava-next-8b-hf	Qwen2-VL-7B-Instruct	90.0	65.7
llama3-llava-next-8b-hf	Qwen2.5-VL-7B-Instruct	96.6	90.9
llama3-llava-next-8b-hf	llama3-llava-next-8b-hf	97.7	93.3
llama3-llava-next-8b-hf	gemma-3-4b-it	99.1	98.2
llama3-llava-next-8b-hf	gemma-3-12b-it	99.5	99.6
gemma-3-4b-it	Qwen2.5-VL-3B-Instruct	76.0	20.2
gemma-3-4b-it	Qwen2-VL-7B-Instruct	71.5	19.6
gemma-3-4b-it	Qwen2.5-VL-7B-Instruct	85.2	42.7
gemma-3-4b-it	llama3-llava-next-8b-hf	86.5	41.7
gemma-3-4b-it	gemma-3-4b-it	99.4	98.0
gemma-3-4b-it	gemma-3-12b-it	98.7	92.7
gemma-3-12b-it	Qwen2.5-VL-3B-Instruct	84.1	49.4
gemma-3-12b-it	Qwen2-VL-7B-Instruct	81.3	50.0
gemma-3-12b-it	Qwen2.5-VL-7B-Instruct	91.2	74.2
gemma-3-12b-it	llama3-llava-next-8b-hf	93.3	77.2
gemma-3-12b-it	gemma-3-4b-it	99.4	97.6
gemma-3-12b-it	gemma-3-12b-it	99.7	98.4

Table 12: AUC-ROC and TPR at low FPR (FPR=5%) FiMMIA performance metrics for various evaluated **Image** MLLMs.

Feature Type	Image Features	Audio Features
Texture/Pattern	• Local Binary Patterns (LBP) histogram • SIFT Bag of Visual Words (BoVW)	• MFCCs (mean coefficients) • Chroma features (mean) • Tonnetz features (mean)
Spectral/Frequency	• DCT coefficients (low-frequency)	• Spectral centroid (mean) • Spectral bandwidth (mean) • Spectral rolloff (mean)
Color/Energy	• HSV histograms (H, S, V channels)	• RMS energy (mean) • Zero-crossing rate (mean)
Temporal/Rhythmic	• —	• Tempogram features (mean)

Table 13: Statistical Features Extracted for Image and Audio Classification