# Membership and Dataset Inference Attacks on Large Audio Generative Models

**Jakub Proboszcz**
Warsaw University of Technology

**Paweł Kochański**
Warsaw University of Technology

**Karol Korszun**
Warsaw University of Technology

**Donato Crisostomi**
Sapienza University of Rome

**Giorgio Strano**
Sapienza University of Rome

**Emanuele Rodolà**
Sapienza University of Rome

**Kamil Deja**
Warsaw University of Technology
IDEAS Research Institute

**Jan Dubiński**
Warsaw University of Technology
NASK National Research Institute
jan.dubinski.dokt@pw.edu.pl

## Abstract

Generative audio models, based on diffusion and autoregressive architectures, have advanced rapidly in both quality and expressiveness. This progress, however, raises pressing copyright concerns, as such models are often trained on vast corpora of artistic and commercial works. A central question is whether one can reliably verify if an artist's material was included in training, thereby providing a means for copyright holders to protect their content. In this work, we investigate the feasibility of such verification through *membership inference attacks* (MIA) on open-source generative audio models, which attempt to determine whether a specific audio sample was part of the training set. Our empirical results show that membership inference alone is of limited effectiveness at scale, as the per-sample membership signal is weak for models trained on large and diverse datasets. However, artists and media owners typically hold collections of works rather than isolated samples. Building on prior work in text and vision domains, in this work we focus on *dataset inference* (DI), which aggregates diverse membership evidence across multiple samples. We find that DI is successful in the audio domain, offering a more practical mechanism for assessing whether an artist's works contributed to model training. Our results suggest DI as a promising direction for copyright protection and dataset accountability in the era of large audio generative models.

## 1 Introduction

Generative audio models have undergone rapid advances in recent years, driven largely by diffusion (DMs) [1, 2, 3, 4] and autoregressive architectures (ARMs) [5, 6]. These models are capable of producing highly realistic soundscapes, speech, and music. While this progress opens exciting opportunities in areas such as creative expression, accessibility, and interactive media, it also raises urgent concerns about privacy, copyright, and data governance. In particular, the vast datasets required to train such systems can contain artistic or commercial audio without transparent disclosure, leaving creators uncertain about whether their works contributed to a model's capabilities [7]. A central

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Generative and Protective AI for Content Creation.

question in this context is whether one can reliably determine if a specific artist's recordings were included in training of a generative model. Addressing this question is critical both for protecting intellectual property and for enabling accountability in machine learning practice. Similar challenges have been investigated in computer vision [8, 9, 10] and natural language processing [11, 12, 13], where MIAs [14] attempt to determine if a given sample was used in training, and DI [15, 11, 16] extends this idea to entire collections. However, the effectiveness of these techniques for large generative models in the audio domain remains unclear.

In this paper, we conduct a study of membership and dataset inference attacks on large audio generative models. We begin by evaluating the effectiveness of existing MIA strategies when applied to open-source ARMs and DMs. Our findings reveal that single-sample membership inference is weak in this setting, offering limited evidence of training set inclusion. Motivated by the observation that artists and rights-holders typically possess collections of works rather than isolated samples, we shift focus to DI. By aggregating diverse membership signals across multiple samples, DI achieves substantially higher effectiveness, enabling more reliable detection of training set participation. Our contributions are threefold:

- We benchmark existing MIAs on large audio ARMs and DMs, highlighting their limitations.
- We extend the existing DI methodology to audio generative models, assessing its effectiveness in the audio domain.
- We provide an extensive empirical evaluation across several state-of-the-art audio models, demonstrating that DI can succeed where single-sample attacks fail, and thus suggest it as a promising mechanism for copyright protection and dataset accountability.

Our work aims to initiate a discussion on existing methods for protecting copyrighted audio samples in large-scale generative models, while also laying the groundwork for auditing methods that empower creators to assert control over their intellectual property.

## 2 Background

### 2.1 Identifying Training Data

**Membership Inference** (MIA). MIAs [14] aim to decide whether a given sample was part of a model's training set. They exploit overfitting: training samples typically yield lower losses than unseen ones. Formally, the attacker constructs an attack function $A_{f_\theta} : \mathcal{X} \to 0, 1$ that predicts membership. A standard approach is the threshold attack [17], which classifies $x$ as a member if the chosen metric is below a threshold: $A_{f_\theta}(x) = \mathbb{1}![\mathcal{M}(f_\theta, x) < \gamma]$, where $\mathcal{M}$ is the metric and $\gamma$ the decision threshold.

**Dataset Inference** (DI). DI [15] asks whether an entire dataset was used during training. Unlike MIAs, which evaluate individual samples, DI aggregates membership signals (often based on MIAs) across multiple points into a dataset-level statistic, thereby amplifying weak per-sample evidence. This makes DI effective for large models and datasets where single-sample inference is unreliable. Initially proposed for supervised models, DI extracts per-sample features, aggregates them into a dataset score, and applies a statistical test [18, 19]. Recent work has extended DI to generative models, including large language models (LLMs) [11, 20], DMs [16], and autoregressive image models [21]. Formally, DI compares scores from a suspected member set and a non-member set via Welch's $t$-test at $\alpha = 0.01$ with the null hypothesis $H_0$ : mean(member scores) $\leq$ mean(non-member scores). Rejecting $H_0$ implies the dataset was part of training. Correctness requires both sets be i.i.d.; otherwise, distributional mismatch can bias the test. The strength of DI depends on the number of available samples. To quantify leakage, we define $P$ as the minimum number of samples needed to reject $H_0$. Smaller $P$ indicates stronger leakage.

### 2.2 Audio Generative Models

We experiment on 4 models described in Table 1 representing both AR and DM famielies. Currently, DMs dominate high-quality audio generation. **AudioLDM2** [4] unifies text-to-audio, text-to-music, and text-to-speech within one framework. It leverages a "Language of Audio" representation, mapped from AudioMAE [22] features through GPT-2 [23], to condition a UNet [24] diffusion model over mel-spectrogram latents. With 29.5k hours of diverse training data, it establishes a common semantic space that supports multiple generative tasks. **TANGO** [1] follows a simpler design. A frozen

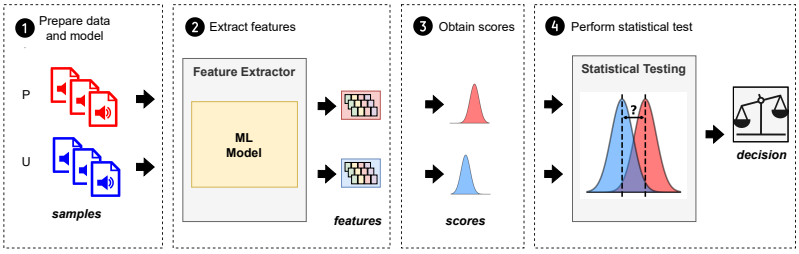


Figure 1: **Dataset Inference Procedural Steps.** The process consists of four main steps: ❶ *Data Preparation:* Prepare the data to verify whether the (suspected) member samples **P** were used to train the model. The (confirmed) nonmember samples **U**, from the same distribution as **P**, serve as the validation set. ❷ *Feature Extraction:* Run each individual MIA on all inputs from $\{\mathbf{P}, \mathbf{U}\}$ to extract membership features for all data samples. We use our MIAs tailored to ARMs and DMs for the respective model types. ❸ *Score Computation:* We map extracted feature vectors into scalar membership scores for each sample. We use a scoring model for DMs, following [16], and feature summation [21] for ARMs (see Appendix for more details. ❹ *Statistical Testing:* Apply a statistical t-test to verify whether the scores obtained for the public suspect data points **P** are statistically significantly higher than those for **U**. If so, **P** is marked as being part of the model's training set. Otherwise, the test is inconclusive and the model's training set is considered independent of **P**.

Table 1: **Comparison of audio generative models on which we experiment in our work.**

| Model | Training steps | Dataset | Size | Params | Type | Input |
|---|---|---|---|---|---|---|
| AudioGen | 200k | AudioCaps + other | 4k h | 285M | AR | Audio tokens |
| FIGARO | 100k | LakhMIDI | 176k files | 88M | AR | MIDI |
| AudioLDM2 | 800k | AudioCaps + other | 29.5k h | 1.5B | DM | Mel-spec |
| TANGO | 230k | AudioCaps | 46k clips | 866M | DM | Mel-spec |

FLAN-T5 [25] encoder provides text embeddings that guide a latent diffusion model trained on the AudioCaps [26] dataset. Augmentation is based on pressure-level mixing of sounds, ensuring balanced exposure. Despite training on only 46k clips, TANGO achieves competitive generation quality by relying on strong instruction-tuned text features.

Competing with DMs across different modalities, autoregressive approaches have also been applied to audio generation. **AudioGen** [6] treats audio as a sequence of discrete codec tokens and learns a text-conditioned Transformer decoder to generate them. The model is trained on roughly 4k hours from ten heterogeneous audio–text corpora, and augmentations based on mixing sound sources are used to expose it to overlapping events. **FIGARO** [5] addresses symbolic music with controllable generation. It introduces description-to-sequence learning, combining interpretable features such as chords, instrumentation, and rhythm with learned latent codes as conditioning. Training on the 176k-file LakhMIDI dataset [27] allows the model to reconstruct bar-level sequences and to provide global and fine-grained control.

## 3 Method

**Membership Inference.** We begin by evaluating MIAs on large audio generative models. For DMs, we apply the attack suite explored in [16], which exploits denoising dynamics to distinguish training samples from non-members. For ARMs, we use the approach from Kowalczuk et al. [21], which leverages token-level log-likelihoods and related statistics. We give more details on individual MIAs used in our work in the Appendix. In both cases, we use the train split of each model's dataset as the source of *members* and the held-out test split as the source of *non-members*. This ensures that attacks are evaluated under a realistic and controlled setting where the attacker's candidate pool contains both genuine training data and independent test samples.

**Dataset Inference.** To facilitate the task, we extend our studies to DI. We follow the methodology introduced by [16] for DMs and [21] for ARMs. Each candidate dataset consists of a collection of suspected member samples $\mathcal{P}$ and an equal number of non-member samples $\mathcal{U}$ drawn from the test split. We extract membership features for each sample using multiple MIAs, aggregate them into scalar scores, and then apply Welch's $t$-test to compare $\mathcal{P}$ and $\mathcal{U}$. The null hypothesis states that

the mean score of $\mathcal{P}$ is no greater than that of $\mathcal{U}$, and we reject it at $\alpha = 0.01$ if sufficient evidence is found. Following standard practice, we report the minimum number of samples $P$ required to reject the null hypothesis, with smaller $P$ indicating stronger information leakage. Our approach is demonstrated on Figure 1.

## 4 Results

In our experiments on MIAs, we report the Area Under the Curve (AUC) and the True Positive Rate at a False Positive Rate of 1% (TPR@FPR = 1%). For AUC, a value of 0.50 corresponds to random guessing, while for TPR@FPR = 1%, the baseline for random guessing is 0.01. For DI, we report the minimal number of samples in $\mathcal{P}$ required to successfully reject the null hypothesis $H_0$, *i.e.,*, to flag the audio samples in $\mathcal{P}$ as having been used in training a given model.

Table 2 presents the AUC values and TPR@1% for MIAs on ARMs. For both AudioGen and FIGARO, the results remain close to 50% AUC, indicating chance-level performance. This suggests that single-sample membership inference is ineffective against SOTA audio ARMs, trained on larger datasets, as they do not leak strong per-sample signals. Similar observation can be seen with TPR at a fixed FPR of 1%, where the values are consistently low, rarely exceeding 1%, which confirms that MIAs provide limited evidence for distinguishing members from non-members in SOTA audio ARMs.

Table 2: **MIA results for Autoregressive Models.** We report AUC and TPR@FPR=1%.

| | **AUC** | | **TPR@1%** | |
|---|---|---|---|---|
| **Attack** | AudioGen | FIGARO | AudioGen | FIGARO |
| Loss [17] | 52.85±0.00 | 50.28±0.61 | 0.68±0.00 | 1.18±0.11 |
| Zlib [28] | 50.46±0.00 | 49.37±0.61 | 0.74±0.00 | 1.11±0.22 |
| Hinge [29] | 54.42±0.00 | 50.05±0.60 | 1.42±0.00 | 1.12±0.21 |
| Min-K% [13] | 55.34±0.00 | 50.28±0.58 | 1.13±0.00 | 1.13±0.20 |
| Min-K%$^{++}$ [30] | 50.86±0.00 | 49.65±0.56 | 0.97±0.00 | 1.01±0.19 |
| CAMIA [31] | 51.86±0.00 | 51.68±0.53 | 1.35±0.00 | 1.07±0.21 |

For DMs, Table 3 shows AUC and TPR@1% values for a range of MIA strategies. AudioLDM2 again yields performance near chance, with AUCs around 50–55%. In contrast, TANGO shows a detectable membership signal, with AUC values reaching nearly 70%. This difference likely arises from the smaller scale of TANGO's training set, which makes overfitting more apparent. Results with TPR@FPR=1%, makes this distinction even clearer: for AudioLDM2 the detection rate is essentially zero across all attacks, while for TANGO it reaches 16–17% for the best-performing MIAs are feasible only for smaller DMs, but scale poorly to models trained on larger datasets.

Table 3: **MIA results for Diffusion Models.** We report AUC and TPR@FPR=1%.

| | **AUC** | | **TPR@1%** | |
|---|---|---|---|---|
| **Attack** | AudioLDM2 | Tango | AudioLDM2 | Tango |
| Loss [10] | 52.29±0.54 | 70.52±0.87 | 0.00±0.00 | 16.03±2.21 |
| Gradient Masking [16] | 49.15±0.56 | 51.06±0.99 | 0.12±0.29 | 3.18±0.78 |
| Multiple Loss [16] | 54.83±0.53 | 69.47±0.83 | 3.49±0.26 | 16.68±2.17 |
| NoiseOpt [16] | 52.44±0.55 | 51.69±0.98 | 0.00±0.00 | 1.06±0.29 |
| PIA [8] | 50.09±0.56 | 52.29±0.93 | 0.00±0.00 | 1.78±0.35 |
| PIAN [8] | 51.69±0.55 | 50.90±0.95 | 0.00±0.01 | 2.18±0.48 |

Finally, Table 4 reports the number of samples required for DI to reach statistical significance. Here, the advantage of DI over MIA becomes clear. AudioGen requires around 900 samples to reject the null hypothesis, while FIGARO and AudioLDM2 require only 300 samples. Most strikingly, TANGO requires just 20 samples, showing that DI can detect training set usage with very small collections. Overall, these results highlight that single-sample MIAs are limited when applied to models trained on larger audio datasets, but DI provides strong and scalable evidence, making it a more practical tool for auditing the training sets of generative audio models. However, for most models, DI requires more samples than an individual artist is likely to possess, especially given the fact that the lenght of AudioCaps sample is 10 secods. These requirements remain attainable for media owners, but highlight the need for further methodological development.

Table 4: **results for Dataset inference.** Minimum number of samples required to achieve mean $p \leq 0.01$.

| **Model** | **# Samples** |
|---|---|
| AudioGen | 900 |
| FIGARO | 300 |
| AudioLDM2 | 300 |
| Tango | 20 |

## 5 Discussion and Conclusions

Our study demonstrates that membership inference alone is not a reliable mechanism for verifying whether individual audio samples contributed to the training of large audio generative models. However, when aggregated via dataset inference, the weak per-sample signals accumulate to provide statistically significant evidence of training set inclusion, even with relatively small collections. This highlights DI as a promising tool for creators and auditors seeking to verify copyright misuse.

An important implication of our findings is the responsibility of model providers in enabling meaningful auditing. Currently, many released models do not disclose clear train–test splits or maintain accessible held-out evaluation sets. This lack of transparency makes it challenging to fairly assess privacy leakage, data governance, and copyright compliance. We argue that providers of generative models should adopt the practice of reporting well-defined train/test partitions and reserving clean held-out sets that remain unused during training. Such held-out data would allow independent researchers to systematically study privacy risks, monitor overfitting, and develop robust detection techniques without ambiguity about data provenance.

## Acknowledgments

This work is partly supported by the MUR FIS2 grant n. FIS-2023-00942 "NEXUS" (cup B53C25001030001), and by Sapienza University of Rome via the Seed of ERC grant "MINT.AI" (cup B83C25001040001).

## References

[1] Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-audio generation using instruction tuned llm and latent diffusion model. *arXiv preprint arXiv:2304.13731*, 2023.

[2] Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization, 2024.

[3] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. *Proceedings of the International Conference on Machine Learning*, 2023.

[4] Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024.

[5] Dimitri von Rütte, Luca Biggio, Yannic Kilcher, and Thomas Hofmann. Figaro: Controllable music generation using learned and expert features. In *The Eleventh International Conference on Learning Representations*, 2023.

[6] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*, 2022.

[7] Tim W. Dornis and Sebastian Stober. Generative ai training and copyright law, 2025.

[8] Fei Kong, Jinhao Duan, RuiPeng Ma, Hengtao Shen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. An efficient membership inference attack for the diffusion model by proximal initialization. *arXiv preprint arXiv:2305.18355*, 2023.

[9] Jan Dubiński, Antoni Kowalczuk, Stanisław Pawlak, Przemyslaw Rokita, Tomasz Trzciński, and Paweł Morawiecki. Towards more realistic membership inference attacks on large diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4860–4869, 2024.

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

[11] Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. Llm dataset inference: Did you train on my dataset?, 2024.

[12] Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schoelkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership inference attacks against language models via neighbourhood comparison. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11330–11343, Toronto, Canada, July 2023. Association for Computational Linguistics.

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

[14] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, Los Alamitos, CA, USA, may 2017. IEEE Computer Society.

[15] Pratyush Maini, Mohammad Yaghini, and Nicolas Papernot. Dataset inference: Ownership resolution in machine learning. In *Proceedings of ICLR 2021: 9th International Conference on Learning Representationsn*, 2021.

[16] Jan Dubiński, Antoni Kowalczuk, Franziska Boenisch, and Adam Dziedzic. Cdi: Copyrighted data identification in diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18674–18684, 2025.

[17] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282, Los Alamitos, CA, USA, jul 2018. IEEE Computer Society.

[18] Adam Dziedzic, Nikita Dhawan, Muhammad Ahmad Kaleem, Jonas Guan, and Nicolas Papernot. On the difficulty of defending self-supervised learning against model extraction. In *ICML (International Conference on Machine Learning)*, 2022.

[19] Adam Dziedzic, Haonan Duan, Muhammad Ahmad Kaleem, Nikita Dhawan, Jonas Guan, Yannis Cattan, Franziska Boenisch, and Nicolas Papernot. Dataset inference for self-supervised models. In *NeurIPS (Neural Information Processing Systems)*, 2022.

[20] Bihe Zhao, Pratyush Maini, Franziska Boenisch, and Adam Dziedzic. Unlocking post-hoc dataset inference with synthetic data. 2025.

[21] Antoni Kowalczuk, Jan Dubiński, Franziska Boenisch, and Adam Dziedzic. Privacy attacks on image autoregressive models. In *Forty-second International Conference on Machine Learning*.

[22] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. In *NeurIPS*, 2022.

[23] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[25] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022.

[26] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating Captions for Audios in The Wild.

[27] Colin Raffel. *Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching*. PhD thesis, Columbia University, 2016.

[28] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association, August 2021.

[29] Martin Bertran, Shuai Tang, Aaron Roth, Michael Kearns, Jamie H Morgenstern, and Steven Z Wu. Scalable membership inference attacks via quantile regression. *Advances in Neural Information Processing Systems*, 36, 2024.

[30] Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. Min-k%++: Improved baseline for detecting pre-training data from large language models. *arXiv preprint arXiv:2404.02936*, 2024.

[31] Hongyan Chang, Ali Shahin Shamsabadi, Kleomenis Katevas, Hamed Haddadi, and Reza Shokri. Context-aware membership inference attacks against pre-trained large language models. *arXiv preprint arXiv:2409.13745*, 2024.

[32] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[33] Shengfang Zhai, Huanran Chen, Yinpeng Dong, Jiajun Li, Qingni Shen, Yansong Gao, Hang Su, and Yang Liu. Membership inference on text-to-image diffusion models via conditional likelihood discrepancy. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[34] Jean-loup Gailly and Mark Adler. zlib compression library. 2004.

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2020.

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

# A Details on Membership Inference for Autoregressive Models

**Autoregressive Modeling for Audio *and* Symbolic Music.** Autoregressive generators in the audio domain operate on compact *discrete sequences* rather than raw waveforms. Two prevalent instantiations are: (i) *tokenized time–frequency* representations, where a vector-quantized encoder (e.g., VQ-VAE/VQ-GAN) maps a mel-spectrogram $\mathrm{Mel}(x)$ to a low-resolution latent grid and quantizes each cell to a codebook index; and (ii) *symbolic music* representations (e.g., MIDI or piano-roll), where events such as NOTE-ON, NOTE-OFF, pitch, velocity, instrument, bar/beat markers, and control changes are serialized into a discrete token stream. In both cases, the 2D structure (time–frequency for mel tokens; hierarchical/bar–beat–event for MIDI) is linearized into a 1D sequence $\mathbf{c} = (c_1, \ldots, c_N)$ using a fixed, deterministic ordering (typically time-major; optional bar/measure delimiters for music).

The generative objective models next-token probabilities:

$$p(\mathbf{c}) \;=\; \prod_{n=1}^{N} p(c_n \mid c_1, \ldots, c_{n-1}) \,, \tag{1}$$

optimized via maximum likelihood over the training distribution:

$$L_{\mathrm{AR}} \;=\; \mathbb{E}_{x \sim \mathcal{D}_{\mathrm{train}}} \big[ -\log p\big(\mathbf{c}(x)\big) \big] \,, \tag{2}$$

where $\mathbf{c}(x)$ denotes either codebook indices derived from $\mathrm{Mel}(x)$ *or* a serialized symbolic/MIDI event stream for $x$.

This token-first formulation shortens effective sequence length and exposes strong discrete structure (repetition, meter, harmony), enabling high-fidelity generation with tractable context windows. At inference, tokens $\hat{\mathbf{c}}$ are sampled autoregressively from Eq. 1. For mel-token systems, a quantized decoder reconstructs a mel-spectrogram $\widehat{\mathrm{Mel}}$ that is rendered to waveform via a vocoder. For symbolic systems (e.g., FIGARO), the sampled event stream is rendered to audio by a MIDI synthesizer or mapped back to a structured score, optionally honoring controllable conditioning tokens (e.g., chords, instrumentation, rhythm descriptors) embedded in the same sequence.

Kowalczuk et al. [21] introduce the first comprehensive MIA suite for *image* autoregressive models by adapting well-established, token-level attacks from the LLM literature (e.g., Loss, Zlib, Hinge, Min-K%, Min-K%$^{++}$, SURP, CAMIA) to visual next-token prediction. A key observation in their work is that many IARs are trained with *classifier-free guidance* [32], i.e., the forward pass processes each example both *with* conditioning (e.g., a class label or text prompt) and *without* it. Building on CLiD [33], they exploit this extra supervision signal by contrasting the conditional and unconditional paths: instead of feeding raw per-token logits into MIAs, they use the guidance-difference statistic

$$\Delta(x, c) \;=\; p(x \mid c) \;-\; p(x \mid c_{\text{null}}),$$

where $c$ is the conditioning input and $c_{\text{null}}$ denotes the null (unconditional) condition. This replacement amplifies membership signal relative to LLM-style attacks that lack such conditioning, and it avoids relying solely on per-token probabilities.

Because *audio* autoregressive models also operate on discrete token sequences, either time–frequency codes (e.g., codec/VQ tokens) or symbolic music events (e.g., MIDI as in FIGARO), the same LLM-derived MIAs are directly applicable in the audio domain. When audio ARMs are trained with explicit conditioning (e.g., captions, tags, control tokens) and employ classifier-free guidance, the CLiD-style conditional–unconditional contrast $\Delta(x, c)$ can be computed analogously on audio tokens and used as the primary MIA feature.

**Threshold-based attack.** A simple and widely used approach to infer membership is to compare a scalar diagnostic to a fixed cutoff. Let $\mathcal{M}$ be a per-sample metric such as the negative log-likelihood or loss. A sample $x$ is declared a member whenever the metric falls below a threshold $\gamma$:

$$A_{f_\theta}(x) \;=\; \mathbb{1}[\mathcal{M}(f_\theta, x) < \gamma] \,, \tag{3}$$

where $\gamma$ is selected on a validation split. The rationale is that training items typically attain lower loss than points not seen during training.

**MIN-K% PROB metric.** To reduce the influence of highly predictable positions, **(author?)** [13] focus the decision rule on the least likely part of the sequence. For an input $x$ and a fraction $K \in \{10, 20, 30, 40, 50\}$, MIN-K% PROB computes the average negative log-likelihood over the bottom $K\%$ tokens under the model $f_\theta$. Membership is predicted by thresholding this average:

$$A_{f_\theta}(x) \;=\; \mathbb{1}[\text{Min-}K\%(x) < \gamma] \,.$$

Reporting the best result over a small sweep of $K$ makes the attack less sensitive to the choice of this hyperparameter.

**MIN-K% PROB ++**. MIN-K% PROB ++ refines MIN-K% PROB by normalizing token log-probabilities and testing whether low-probability positions behave like local modes of the learned distribution. Given a sequence $x = (x_1, \ldots, x_T)$, define

$$\mathcal{S}_{\text{Min-K\%++}}(x) \;=\; \frac{1}{|S|} \sum_{t \in S} \frac{\log p(x_t \mid x_{<t}) - \mu_{x<t}}{\sigma_{x<t}}, \tag{4}$$

where $S$ is the subset containing the bottom $K\%$ tokens, and $\mu_{x<t}$, $\sigma_{x<t}$ are the mean and standard deviation of token log-probabilities over the entire vocabulary at position $t$. A sample is flagged as a member if

$$A_{f_\theta}(x) \;=\; \mathbb{1}[\mathcal{S}_{\text{Min-K\%++}}(x) \geq \gamma]\,. \tag{5}$$

As with MIN-K% PROB, performance is reported for the best $K \in \{10, 20, 30, 40, 50\}$.

**Zlib ratio attack.** This baseline relates model fit to a model-agnostic compressibility proxy [34]. Let $\mathcal{P}_{f_\theta}(x)$ denote the perplexity (or exponentiated average negative log-likelihood) and $\text{zlib}(x)$ be the compressed size of $x$ under the zlib codec. The statistic

$$\frac{\mathcal{P}_{f_\theta}(x)}{\text{zlib}(x)}$$

tends to be smaller for members, since model perplexity is lower on training data while zlib compression does not benefit from any model-specific memorization. Membership is then inferred by comparing this ratio to a threshold.

**CAMIA.** Context-aware MIA [31] augments raw loss features with temporal descriptors of the token-wise loss sequence. Several signals are used: a *slope* feature that captures how quickly losses decline across positions; *approximate entropy*, which measures regularity by the prevalence of repeating patterns; *Lempel–Ziv complexity*, which quantifies diversity in the loss trajectory via the count of distinct substrings; a *count-below* statistic, the fraction of tokens with loss below a preset cutoff; and a *repeated-sequence amplification* feature that measures the reduction in loss when the same input is repeated. Non-members typically display higher irregularity and larger gains from repetition, while members show more stable, low-loss segments.

**Surprising Tokens Attack (SURP).** SURP targets positions where the model is confident overall but assigns low probability to the true token. For each position $t$, let $H_t$ be the Shannon entropy of the predictive distribution and $p(x_t \mid x_{<t})$ the probability of the ground-truth token. Define the surprising set

$$S \;=\; \{\, t \mid H_t < \epsilon_e,\; p(x_t \mid x_{<t}) < \tau_k \,\}\,, \tag{6}$$

where $\epsilon_e \in \{2, 4, 8, 16\}$ controls the entropy threshold and $\tau_k$ is the $k$-th percentile probability with $k \in \{10, 20, 30, 40, 50\}$. The SURP score averages the probabilities on this set:

$$\mathcal{S}_{\text{SURP}}(x) \;=\; \frac{1}{|S|} \sum_{t \in S} p(x_t \mid x_{<t}). \tag{7}$$

Membership is decided by thresholding $\mathcal{S}_{\text{SURP}}(x)$:

$$A_{f_\theta}(x) \;=\; \mathbb{1}[\mathcal{S}_{\text{SURP}}(x) \geq \gamma]\,. \tag{8}$$

In practice, the best-performing pair $(k, \epsilon_e)$ from the specified grids is selected to summarize results.

# B Details on Membership Inference for Diffusion Models

**Diffusion Models** [35] are generative models trained by progressively adding noise to the data and then learning to reverse this corruption. In the forward diffusion process, Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ is added to a clean sample $x$ to obtain a noised sample $x_t \leftarrow \sqrt{\alpha_t}\, x + \sqrt{1 - \alpha_t}\, \epsilon$, where $t \in [0, T]$ is the diffusion timestep and $\alpha_t \in [0, 1]$ is a monotonically decreasing schedule with $\alpha_0 = 1$ and $\alpha_T = 0$. The denoiser $f_\theta$ is trained to predict $\epsilon$ across timesteps by minimizing $\frac{1}{N} \sum_i \mathbb{E}_{t,\epsilon}\, \mathcal{L}(x_i, t, \epsilon; f_\theta)$, where $N$ is the training set size and

$$\mathcal{L}(x, t, \epsilon; f_\theta) = \|\epsilon - f_\theta(x_t, t)\|_2^2\,. \tag{9}$$

Sampling proceeds by iteratively removing predicted noise $f_\theta(x_t, t)$ from $x_t$ for $t = T, T-1, \ldots, 0$, starting from $x_T \sim \mathcal{N}(0, I)$ to obtain a generated sample $x_{t=0}$. For conditional generation, an additional input $y$ (conditioning signal) is provided to $f_\theta$.

Latent diffusion models [36] perform the diffusion process in a learned latent space to improve efficiency. An encoder $\mathcal{E}$ maps $x$ to a latent $z = \mathcal{E}(x)$, and the objective in Eq. 9 becomes

$$\mathcal{L}(z, t, \epsilon; f_\theta) = \|\epsilon - f_\theta(z_t, t)\|_2^2\,. \tag{10}$$

**Denoising Loss.** Early membership inference attacks for diffusion models [10] assess sample membership by directly using the denoising loss as a statistic. The key observation is that the loss at intermediate timesteps

provides the strongest separation between training members and non-members. In particular, $t \approx 100$ often yields the most discriminative signal: very small $t$ makes the task too easy (the noised input remains close to the original), whereas very large $t$ collapses the input toward pure noise, making prediction uniformly hard. A sample is classified as a member if its loss at the chosen timestep falls below a threshold selected on a validation split.

**Multiple Loss.** A multi-timestep variant aggregates information from several diffusion steps to improve robustness of the signal. This attack evaluates Eq. 10 at a fixed grid of timesteps (e.g., $t \in \{0, 100, \ldots, 900\}$) and combines the resulting losses into a single score, for example by summation or a weighted average. The aggregate loss serves as the decision statistic, again thresholded to yield a membership prediction. Using multiple timesteps reduces variance and can capture complementary difficulty regimes of the denoising task.

**Proximal Initialization Attack (PIA).** The PIA family [8] compares the model's noise predictions when initialized from different proximity states to the data. A canonical instantiation evaluates the prediction error at a clean (or minimally noised) state, such as $t = 0$, and at a moderately noised state, typically around $t = 200$ where separability is reported to be strong. The difference (or ratio) between these errors is used as the attack feature. Intuitively, training samples induce more confident and stable predictions across nearby states of the diffusion process, leading to a lower feature value for members than for non-members.

**PIAN.** An adaptation of PIA, denoted PIAN [8], normalizes the denoiser's output to enforce approximately Gaussian behavior in the predicted noise, thereby reducing scale effects that may confound raw error magnitudes. The membership statistic is computed analogously to PIA after normalization. As with PIA, members are expected to yield smaller scores because the model's predictions align more consistently with the true noise on training data.

**Gradient Masking.** The gradient-masking attack [16] targets semantically critical regions of the latent representation that most influence the denoising loss. For a given $z_t$, the gradient $\mathbf{g} = |\nabla_{z_t} \mathcal{L}(z_t, t, \epsilon; f_\theta)|$ is computed, and a binary mask $\mathbf{M}$ is formed by selecting the top-percentile (e.g., top 20%) entries of $\mathbf{g}$. A perturbed latent $\hat{z}_t = \epsilon \cdot \mathbf{M} + z_t \cdot \neg\mathbf{M}$ is then created by replacing the most influential coordinates with random noise and leaving the remainder unchanged. The attack feature is the reconstruction error restricted to the masked region, $\|(\epsilon - z_t) \cdot \mathbf{M} - f_\theta(\hat{z}_t, t) \cdot \mathbf{M}\|_2^2$, optionally aggregated across multiple timesteps. Because models tend to memorize salient structure in training samples, members exhibit lower masked-region reconstruction error than non-members.

**Noise Optimization.** The central premise of noise-optimisation attack [16] is that stronger (or more effective) perturbations are required to significantly reduce the denoising loss for training members, reflecting higher confidence and tighter fit on seen data. Concretely, starting from $z_t$ at an intermediate timestep (e.g., $t = 100$), an unconstrained optimization seeks a perturbation $\delta$ that minimizes the objective $\min_\delta \|\epsilon - f_\theta(z_t + \delta, t)\|_2^2$, using 5 L-BFGS steps. Two complementary features arise: the minimized prediction error $\|\epsilon - f_\theta(z_t + \delta, t)\|_2^2$ and the perturbation magnitude $\|\delta\|_2^2$. Members typically achieve lower final error yet require larger or more targeted adjustments, producing distinctive signatures relative to non-members.

# C Details on Dataset Inference

DI generalizes MIAs from individual samples to sets. Its central research question is: *was the collection of suspect samples $P$ used to train the model, as opposed to being independent test data?* To answer this, DI compares $P$ against a reference set $U$ drawn from the same distribution but known to be excluded from training. In both DMs and IARs the procedure consists of three steps: (i) extract a suite of per-sample MIA features, (ii) map these features into scalar membership scores, and (iii) perform a statistical test comparing the distributions of scores for $P$ and $U$. The null hypothesis $H_0 : \text{mean}(s(P)) \leq \text{mean}(s(U))$ is tested with Welch's $t$-test at $\alpha = 0.01$.

**Diffusion models.** For DMs, the CDI methodology [16] employs a broad feature set, which we describe in Appendix B. Rather than aggregating these features directly, CDI fits a logistic regression scorer on disjoint control splits of $P$ and $U$, yielding a calibrated mapping from feature vectors to scalar scores. The test is then applied to scores on held-out subsets. To reduce variance, CDI repeats this process across multiple random partitions and averages the resulting $p$-values.

**Image autoregressive models.** For IARs, the approach of [21] follows the same overall structure but makes use of a different feature suite, tailored to token-level modeling. These features, described in Appendix A, capture variations in token probabilities and loss trajectories that arise in autoregressive generation. Each feature is normalized, and the per-sample scalar score is obtained by summing across all features. This lighter-weight procedure that nevertheless suffices in practice for autoregressive token-based models. The resulting scores for $P$ and $U$ are then compared using the same statistical test as above.