HIERARCHICAL BAYESIAN MEMBERSHIP INFERENCE: A GENERAL FRAMEWORK WITH APPLICATIONS TO DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Membership inference attacks often reduce a model's response to a single score and treat the parameters of class-conditional signal distributions estimated from finitely many reference models as known, without propagating estimation uncertainty or separating variation across stochastic queries, trained models, data points, and groups. We propose a general hierarchical Bayesian framework that models these sources of variation under the member and non-member hypotheses, propagates uncertainty from finite reference-model data, and yields a posterior-predictive likelihood-ratio test. We instantiate the framework for denoising diffusion and flow matching models, whose randomized objectives yield repeated loss observations. For record-level inference, our attack outperforms state-of-the-art methods. For group-level inference, it infers an individual's presence without access to their training images and outperforms adapted baselines with a mixed DDPM and flow-matching reference pool, demonstrating robustness to model-type mismatch. Beyond diffusion, the framework applies to deterministic, repeated stochastic, and grouped observations.

1 INTRODUCTION

Membership inference attacks (MIAs) test training-set membership and support empirical privacy auditing (Shokri et al., 2017; Yeom et al., 2018; Sablayrolles et al., 2019; Carlini et al., 2022). Statistic-based MIAs cast membership as a hypothesis test (Ye et al., 2022). LiRA and RMIA build on likelihood-ratio testing (Carlini et al., 2022; Zarifzadeh et al., 2024), while BASE approximates the Bayes-optimal decision rule of Sablayrolles et al. (2019) (Lassila et al., 2025). All three use *reference models* (also called *shadow models*) to estimate how a point's score varies across independently trained models and compare the target-model score with that variation.

In a growing number of settings of practical interest, two further sources of variation are present. First, the score is itself random: querying the same model twice on the same data point returns two different values, so the evidence is a *sample* from a distribution. Second, the subject of membership can be a *group* of data points, such as the images contributed by one individual or the records of one patient, and their scores vary within the group. LiRA (Carlini et al., 2022) and BASE (Lassila et al., 2025) describe variation across reference models but treat quantities estimated from them as fixed. Neither represents the randomness of a repeated query or the variation among data points belonging to a single subject. This is the gap we address.

In this work, we develop a hierarchical Bayesian framework for membership inference from repeated, stochastic, and grouped observations. Rather than reducing the evidence to a point statistic, we specify a generative model for how the observed query values arise, sample by sample, model by model, query by query, and evaluate a likelihood ratio between the two hypotheses under that model. The framework separates several sources of uncertainty and propagates them by marginalization: the randomness between repeated queries to a fixed model; the variability between independently trained models; the intrinsic difficulty of the target sample; the uncertainty left by having only finitely many reference models; and, at the group level, the variability among the records in a group.

We consider two instantiations of the framework for diffusion models: record-level and group-level membership inference. Diffusion models are a natural choice for two reasons. First, they are now the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

state of the art in high-resolution image generation (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021a; Rombach et al., 2022; Lipman et al., 2023), and are trained on corpora that are frequently sensitive or copyright-protected (Lohanimit et al., 2025; Jiang et al., 2023). Second, their training objective is an expectation over a randomized noising process, and in practice the loss is evaluated by sampling a timestep and a noise vector. A loss query is thus a *draw from a distribution*, and the hierarchy *query* $\rightarrow$ *model* $\rightarrow$ *data point* reflects how the adversary's observations are generated.

One further practical uncertainty arises because the adversary does not know the target model exactly: reference models may differ in architecture, training procedure, or training data. Our framework handles model-to-model variation naturally: the target model has its own latent signal parameters, and the test marginalizes over these and the latent parameters shared across models.

Our contribution. Our key contributions include:

- We develop a hierarchical Bayesian framework for membership inference from repeated, stochastic, and grouped observations. Query noise, model-to-model variability and sample difficulty are separate levels of a generative model, and the test statistic marginalizes over all of them, so the uncertainty left by a finite set of reference models is carried into the score.
- We instantiate it for record-level membership inference against diffusion models. The model is chosen to be conditionally conjugate, so every conditional is available in closed form; this yields an efficient Gibbs sampler and a Rao-Blackwellized posterior predictive density, and we derive moment-based empirical-Bayes estimators for the prior parameters. The attack outperforms LiRA and BASE, at times substantially.
- We instantiate it at the group-level and show that an individual's presence can be inferred confidently *without access to any of their training images*, and we lift LiRA and BASE to this setting as baselines.
- Our framework naturally captures the adversary's uncertainty about the target model, its parameters, training procedure and architecture. Therefore, it is more robust than state-of-the-art attacks to the mismatch between reference models and target model, which is unavoidable in practice.
- We show that LiRA (Carlini et al., 2022) and BaVarIA (Brännvall, 2026) arise as particular cases of the framework.
- We evaluate prior reference-free MIAs specifically designed for diffusion models and show that they vastly underestimate privacy risks compared to the reference-based state-of-the-art.

2 Related Work

MIAs. Shokri et al. (2017) introduced shadow-model attacks, training a classifier on queries with known membership labels. Yeom et al. (2018) proposed a simpler threshold on the target loss, and Sablayrolles et al. (2019) formulated membership inference as a Bayesian decision problem. Reference-model attacks calibrate the decision to the target point (Watson et al., 2022; Ye et al., 2022). LiRA uses a parametric likelihood-ratio test and emphasizes low false-positive rates (Carlini et al., 2022); RMIA gives a robust non-parametric test with few reference models (Zarifzadeh et al., 2024); BASE approximates the Bayesian decision rule and relates closely to RMIA (Lassila et al., 2025). BaVarIA uses Bayesian inference to stabilize fitted signal distributions (Brännvall, 2026).

MIAs against diffusion models. Several works have studied membership inference against diffusion models (Duan et al., 2023; Kong et al., 2023; Pang and Wang, 2025; Li et al., 2025; Rao et al., 2026). Most of these attacks do not use shadow models, despite evidence of the value of shadow-model calibration (Carlini et al., 2022; Watson et al., 2022). SecMI and PIA use estimation errors along the noise trajectory (Duan et al., 2023; Kong et al., 2023), while Rao et al. (2026) use the ℓ_4 norm of the denoising network output. Shadow models have also been used: Carlini et al. (2023) apply LiRA to diffusion losses, and Pang and Wang (2025) use them in a black-box attack. Identity-level membership inference has been studied for GANs and metric-embedding models (Webster et al., 2021; Li et al., 2022) and recently for diffusion models (Vora et al., 2025). A related problem has also been considered in the context of copyright auditing for diffusion models (Dubiński et al., 2025).

3 Preliminaries and Setup

Notation. We denote the set of integers $\{1, 2, \ldots, n\}$ by $[n]$, the indicator function by $\mathbb{1}[\cdot]$, and the set of indexed objects $\{x_1, \ldots, x_n\}$ by $\{x_i\}_{i=1}^n$ or $x_{1:n}$. See App. A.1 for more notation.

Membership inference. A MIA determines if a target subject z was used to train a model $\boldsymbol{\theta}$. The subject may be a single data point (record-level membership) or a collection of data points associated with an individual or entity (group-level membership). Formally, membership inference distinguishes

$$H_0 : \boldsymbol{\theta} \sim P_z, \qquad H_1 : \boldsymbol{\theta} \sim Q_z, \tag{1}$$

where P_z and Q_z denote the distributions of models trained without and with z, respectively. Statistic-based attacks perform this test using a signal obtained by querying the target model. Reference models whose membership status is known provide observations from the corresponding signal distributions (Carlini et al., 2022; Zarifzadeh et al., 2024; Lassila et al., 2025).

Stochastic loss observations. We explicitly consider MIAs based on randomized loss signals, which arise in generative models such as diffusion models and variational autoencoders whose objectives are evaluated using sampled auxiliary variables. Let $\mathcal{L}_{\boldsymbol{\theta},z}$ denote the loss distribution induced by this randomness. For fixed model $\boldsymbol{\theta}$ and data point z,

$$y_1, \ldots, y_J \mid \boldsymbol{\theta}, z \overset{\text{i.i.d.}}{\sim} \mathcal{L}_{\boldsymbol{\theta},z}. \tag{2}$$

Diffusion losses. We instantiate this setting using denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020) and conditional optimal-transport flow-matching models (Lipman et al., 2023). We write $\ell(\boldsymbol{\theta}, z; t, \boldsymbol{\epsilon})$ for the pointwise loss, where t is a diffusion time and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is forward-process corruption noise for DDPMs and a base-distribution sample for conditional flow matching. The precise loss depends on the model parameterization.

For the attack, we fix $t = t_*$ at a time for which the membership signal is strong and resample only $\boldsymbol{\epsilon}$. Our attacks use the log-loss observations

$$x_j = \log \ell(\boldsymbol{\theta}, z; t_*, \boldsymbol{\epsilon}_j), \qquad j \in [J]. \tag{3}$$

Here each $\boldsymbol{\epsilon}_j$ is sampled independently from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Thus, conditional on $(\boldsymbol{\theta}, z, t_*)$, the x_j are i.i.d. log-loss observations. The framework developed next applies equally to other attack signals with repeated stochastic or grouped observations.

4 A Hierarchical Bayes Framework for Membership Inference

In this section, we propose a hierarchical Bayesian framework for membership inference. For each membership class, it jointly models attack observations and latent variables associated with repeated queries, independently trained shadow models, target subjects, and data points belonging to a group. Marginalizing these variables yields a class-conditional posterior-predictive density and hence a log-likelihood-ratio (LLR) membership score. This propagates uncertainty at every modeled level, including from a finite set of shadow models. The framework is independent of model architecture and attack signal; Sec. 5 specializes it to stochastic diffusion losses.

4.1 Posterior-Predictive Membership Test

Let $c \in \{0, 1\}$ denote the membership class, with $c = 1$ for members and $c = 0$ for non-members, and let $y^{(c)}$ denote a generic membership signal. Let $\mathcal{S}$ index the dimensions of the hierarchy and $\mathcal{I}_\mathcal{S} = \{i_s\}_{s \in \mathcal{S}}$ be the corresponding tuple. Its entries may index candidate subjects, models, and repeated queries; when a subject is a collection of data points, an additional entry may index the data points within that subject. We write $\mathcal{Y}^{(c)} = \{y^{(c)}_{\mathcal{I}_\mathcal{S}}\}$ for the observations. We introduce R latent-variable types: type r has instances $\phi^{(r,c)}_{\mathcal{I}_{\mathcal{S}_r}}$, indexed by the dimensions $\mathcal{S}_r \subseteq \mathcal{S}$. Let $\Phi^{(c)} = \{\phi^{(r,c)}_{\mathcal{I}_{\mathcal{S}_r}}\}_{r=1}^R$ collect them. Let $\mathsf{Pa}(v)$ denote the parents of node v. We require $\mathcal{S}_q \subseteq \mathcal{S}_r$ for $q \in \mathsf{Pa}(r)$, so a parent varies over no more dimensions than its child. Finally, let $\eta^{(c)}$ collect the class-specific hyperparameters. We define the framework through the following local conditional distributions:

$$\begin{aligned} y^{(c)}_{\mathcal{I}_\mathcal{S}} \mid \{\phi^{(r,c)}_{\mathcal{I}_{\mathcal{S}_r}} : r \in \mathsf{Pa}(y)\} &\sim \mathcal{Q}_y\Big(\{\phi^{(r,c)}_{\mathcal{I}_{\mathcal{S}_r}} : r \in \mathsf{Pa}(y)\}, \eta^{(c)}\Big), \\ \phi^{(r,c)}_{\mathcal{I}_{\mathcal{S}_r}} \mid \{\phi^{(q,c)}_{\mathcal{I}_{\mathcal{S}_q}} : q \in \mathsf{Pa}(r)\} &\sim \mathcal{Q}_r\Big(\{\phi^{(q,c)}_{\mathcal{I}_{\mathcal{S}_q}} : q \in \mathsf{Pa}(r)\}, \eta^{(c)}\Big), \qquad r \in [R]. \end{aligned} \tag{4}$$

The hierarchy is a directed acyclic graph. Its edges encode the conditional-dependence assumptions of the chosen model, and the local conditionals in Eq. (4) jointly define $p(\mathcal{Y}^{(c)}, \Phi^{(c)} \mid \eta^{(c)})$.

To construct the predictive distribution for class c, the adversary queries M shadow models for which the subject's membership status is known and equal to c. Group the resulting signals as $\bar{\mathcal{Y}}^{(c)} = \{\mathcal{Y}_m^{(c)}\}_{m=1}^M$. Let $\bar{\Phi}^{(c)}$ contain the latent variables shared across models and $\Phi_m^{(c)}$ those specific to shadow model m. We assume that the blocks $(\mathcal{Y}_m^{(c)}, \Phi_m^{(c)})$ are conditionally independent across models given $\bar{\Phi}^{(c)}$. These signals update the shared variables through the posterior $p(\bar{\Phi}^{(c)} \mid \bar{\mathcal{Y}}^{(c)}, \eta^{(c)})$.

We treat the target model as another draw from the same class-conditional model population. Let $\Phi_*^{(c)}$ denote its model-specific latent variables and let $\boldsymbol{y}_*$ collect the membership signals returned when the adversary queries the target model on the target subject. The class-c joint posterior-predictive density of $\boldsymbol{y}_*$ is

$$p_c(\boldsymbol{y}_* \mid \mathcal{Y}^{(c)}, \eta^{(c)}) = \int p(\boldsymbol{y}_* \mid \Phi_*^{(c)}, \bar{\Phi}^{(c)}, \eta^{(c)}) p(\Phi_*^{(c)} \mid \bar{\Phi}^{(c)}, \eta^{(c)}) p(\bar{\Phi}^{(c)} \mid \mathcal{Y}^{(c)}, \eta^{(c)}) \, \mathrm{d}\Phi_*^{(c)} \, \mathrm{d}\bar{\Phi}^{(c)}. \tag{5}$$

The three factors in (5) respectively model these query signals, treat the target-model variables as a new draw from the shared population, and represent what the shadow-model signals reveal about that population. Integrating over both sets of latent variables propagates their uncertainty into the predictive density. Following the Neyman–Pearson principle, we use the posterior-predictive LLR

$$\Lambda(\boldsymbol{y}_*) = \log \frac{p_1(\boldsymbol{y}_* \mid \mathcal{Y}^{(1)}, \eta^{(1)})}{p_0(\boldsymbol{y}_* \mid \mathcal{Y}^{(0)}, \eta^{(0)})} \tag{6}$$

as the scalar membership score assigned to $\boldsymbol{y}_*$.

The prior parameters $\eta^{(c)}$ may be specified a priori or estimated from calibration data using empirical Bayes. Depending on the chosen hierarchy, the predictive density may be available in closed form or evaluated using, e.g., Markov chain Monte Carlo (MCMC) or variational inference.

4.2 Existing Attacks as Special Cases

LiRA. For a target data point z_i, LiRA (Carlini et al., 2022) obtains one deterministic signal $y_{im}^{(c)}$ from each shadow model m and fits a Gaussian for each class c:

$$y_{im}^{(c)} \mid \mu_i^{(c)}, \sigma_i^{(c)} \sim \mathcal{N}\Big(\mu_i^{(c)}, (\sigma_i^{(c)})^2\Big), \qquad \mu_i^{(c)} \sim \delta_{\hat{\mu}_i^{(c)}}, \qquad \sigma_i^{(c)} \sim \delta_{\hat{\sigma}_i^{(c)}}. \tag{7}$$

Here δ_a is a point mass at a, while $\hat{\mu}_i^{(c)}$ and $\hat{\sigma}_i^{(c)}$ are the sample mean and standard deviation across the shadow models of class c. Within our hierarchical framework, this corresponds to $\mathcal{Y}_m^{(c)} = \{y_{im}^{(c)}\}$, $\bar{\Phi}^{(c)} = (\mu_i^{(c)}, \sigma_i^{(c)})$, $\Phi_m^{(c)} = \varnothing$, and $\boldsymbol{y}_* = (y_{i*})$. The point masses yield LiRA's Gaussian predictive density with plug-in parameters.

BaVarIA. BaVarIA (Brännvall, 2026) uses the same one-signal-per-model structure and assigns a conjugate normal–Gamma prior to the shared Gaussian parameters:

$$\begin{aligned} y_{im}^{(c)} \mid \mu_i^{(c)}, \beta_i^{(c)} &\sim \mathcal{N}\Big(\mu_i^{(c)}, (\beta_i^{(c)})^{-1}\Big), \\ \mu_i^{(c)} \mid \beta_i^{(c)} &\sim \mathcal{N}\Big(\xi^{(c)}, (\kappa^{(c)}\beta_i^{(c)})^{-1}\Big), \qquad \beta_i^{(c)} \sim \Gamma\Big(a^{(c)}, b^{(c)}\Big). \end{aligned} \tag{8}$$

Within our framework, this corresponds to $\eta^{(c)} = (\xi^{(c)}, \kappa^{(c)}, a^{(c)}, b^{(c)})$, $\bar{\Phi}^{(c)} = (\mu_i^{(c)}, \beta_i^{(c)})$, $\Phi_m^{(c)} = \varnothing$, and $\boldsymbol{y}_* = (y_{i*})$. Marginalizing the shared parameters gives a closed-form Student-t posterior-predictive density. LiRA and BaVarIA both operate on one deterministic signal per trained model.

Diffusion losses instead yield multiple stochastic signals per model, motivating an additional query level in the hierarchy. Sec. 5 develops this specialization, and the subsequent group-level attack adds a data-point level within each group.

5 Record-level MIAs Against Diffusion Models

We now instantiate our hierarchical Bayesian framework for record-level MIAs against denoising diffusion and flow-matching models. The resulting hierarchy captures variation across target data points, shadow models, and repeated stochastic loss queries.

5.1 Threat Model and Shadow-Model Signals

Following standard practice (Ye et al., 2022; Carlini et al., 2022; Zarifzadeh et al., 2024), we formalize record-level membership inference as a game between a challenger and an adversary; we give the full game in App. B.1. The subject z in (1) is a single data point. The adversary receives z and gray-box access to a model $\boldsymbol{\theta}$ trained on an unknown dataset $\mathcal{D}$, through a query interface that exposes the loss. At the fixed diffusion time t_*, independent draws $\boldsymbol{\epsilon}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I})$ yield $x_j = \log \ell(\boldsymbol{\theta}, z; t_*, \boldsymbol{\epsilon}_j)$ for $j \in [J]$. Thus, conditional on $(\boldsymbol{\theta}, z, t_*)$, the query-signal vector $x_{1:J} = (x_1, \ldots, x_J)$ consists of i.i.d. log-loss observations. The adversary may additionally have side knowledge $\mathcal{H}$, such as the model architecture, training procedure, and hyperparameters.

To learn the class-conditional signal distributions, set $z_1 = z$ and choose auxiliary points $z_2, \ldots, z_N$ from the same data distribution. For every z_n and class $c \in \{0, 1\}$, the adversary trains M shadow models[1] $\{\boldsymbol{\theta}^{(c)}_{nm}\}_{m=1}^{M}$ under the conditions specified by $\mathcal{H}$, including or excluding z_n according to c. From each shadow model, the adversary obtains K conditionally i.i.d. log-loss observations as in (3) and collects $\mathcal{X}^{(c)} = \{x^{(c)}_{nmk} : n \in [N], m \in [M], k \in [K]\}$. This gives a hierarchy over data points, shadow models, and repeated stochastic loss evaluations; App. B.2 details its construction and likelihood-ratio test.

5.2 Hierarchical Model

We model the randomized log-loss observations as Gaussian distributed. For $c \in \{0, 1\}$, $n \in [N]$, $m \in [M]$, and $k \in [K]$, our hierarchy is

$$x^{(c)}_{nmk} \mid \mu^{(c)}_{nm}, \beta^{(c)}_{nm} \sim \mathcal{N}\Big(\mu^{(c)}_{nm}, (\beta^{(c)}_{nm})^{-1}\Big), \tag{9a}$$

$$\mu^{(c)}_{nm} \mid \nu^{(c)}_{n}, \tau^{(c)}_{n} \sim \mathcal{N}\Big(\nu^{(c)}_{n}, r^{(c)}_{n}(\tau^{(c)}_{n})^{-1}\Big), \qquad \beta^{(c)}_{nm} \sim \Gamma\Big(a^{(c)}_{\beta}, b^{(c)}_{\beta}\Big), \tag{9b}$$

$$\nu^{(c)}_{n} \sim \mathcal{N}\Big(\xi^{(c)}, (\lambda^{(c)})^{-1}\Big), \qquad \tau^{(c)}_{n} \sim \Gamma\Big(a^{(c)}_{\tau}, b^{(c)}_{\tau}\Big). \tag{9c}$$

Here $(\mu^{(c)}_{nm}, \beta^{(c)}_{nm})$ and $(\nu^{(c)}_{n}, \tau^{(c)}_{n})$ are shadow-model- and data-point-specific latent variables, respectively. The class-specific hyperparameters are $a^{(c)}_{\beta}, b^{(c)}_{\beta}, \xi^{(c)}, \lambda^{(c)}, a^{(c)}_{\tau}$, and $b^{(c)}_{\tau}$; the point-specific prior multiplier $r^{(c)}_{n} > 0$ allows the between-model variance to vary across data points.

This specializes our general framework with $\mathcal{I}_{\mathcal{S}} = (n, m, k)$, $\Phi^{(c)}_{m} = (\mu^{(c)}_{nm}, \beta^{(c)}_{nm})$, $\bar{\Phi}^{(c)} = (\nu^{(c)}_{n}, \tau^{(c)}_{n})$, and $(\Phi^{(c)}_{*}, \boldsymbol{y}_*) = ((\mu^{(c)}_{*}, \beta^{(c)}_{*}), x_{1:J})$. The hierarchy is motivated by three considerations. **Expressivity:** Eq. (9a)–(9c) separately represent variation across repeated loss evaluations, independently trained shadow models, and data points. **Tractability:** the Gaussian–Gamma specification yields conjugate full conditional distributions and hence efficient Gibbs inference. **Empirical evidence:** we evaluated the Gaussian assumption using 512 lightweight CIFAR-10 models (256 per membership class) and 128 loss evaluations per model and data point; the empirical log-loss distributions are broadly consistent with the Gaussian observation model. App. B.7 provides additional discussion. Because the two classes are treated identically, we suppress c below when unambiguous.

5.3 Inference and Posterior-Predictive Membership Score

Given the shadow-model observations $\mathcal{X}^{(0)}$ and $\mathcal{X}^{(1)}$, the attack proceeds in three stages.

1. Empirical-Bayes calibration. We first fit the prior parameters and point-specific multipliers separately for each membership class using all observations in $\mathcal{X}^{(c)}$.

Proposition 5.1 (Moment-based prior calibration)**.** *For each class, moment matching gives* $\hat{a}_{\zeta} = 2 + \hat{q}^2_{\zeta,1}/(\hat{q}_{\zeta,2} - \hat{q}^2_{\zeta,1})$ *and* $\hat{b}_{\zeta} = \hat{q}_{\zeta,1}(\hat{a}_{\zeta} - 1)$ *for* $\zeta \in \{\beta, \tau\}$*, together with* $\hat{\xi} = \bar{x}$*,* $\hat{\lambda} = (\hat{\sigma}^2_x - \hat{q}_{\tau,1}/M - \hat{q}_{\beta,1}/(KM))^{-1}$*, and* $\hat{r}_n = N e^{\hat{\gamma} y_n} / \sum_{h=1}^{N} e^{\hat{\gamma} y_h}$*. The definition of* $\hat{q}_{\zeta,1}$*,* $\hat{q}_{\zeta,2}$*,* $\bar{x}$*,* $\hat{\sigma}^2_x$*,* $\hat{\gamma}$*, and* y_n*, and the derivation are given in App. B.3.*

We collect the resulting estimates in $\hat{\eta}^{(c)} = \{\hat{a}^{(c)}_{\beta}, \hat{b}^{(c)}_{\beta}, \hat{\xi}^{(c)}, \hat{\lambda}^{(c)}, \hat{a}^{(c)}_{\tau}, \hat{b}^{(c)}_{\tau}\}$.

[1] Only $2 \times M$ shadow models in total has to be trained to achieve this.

2. Local posterior inference. For the target point z_1, let $\mathcal{X}_1^{(c)} = \{x_{1mk}^{(c)}\}_{m\in[M],k\in[K]}$. Holding $\hat{\eta}^{(c)}$ fixed, we use Gibbs sampling to infer the local latent variables $(\mu_{1:M}^{(c)}, \beta_{1:M}^{(c)}, \nu_1^{(c)}, \tau_1^{(c)})$ from $\mathcal{X}_1^{(c)}$. Prop. B.2 in App. B.4 gives the full conditionals and establishes convergence of the sampler.

3. Posterior-predictive scoring. We treat the target model as a new draw from the same class-conditional population as the shadow models. The posterior $p(\nu, \tau \mid \mathcal{X}_1)$ captures what the shadow-model signals reveal about the target-point population, while $p(\mu \mid \nu, \tau)p(\beta)$ generates the signal parameters of a new model. Suppressing c and the fitted hyperparameters, marginalization gives

$$p(x_{1:J} \mid \mathcal{X}_1) = \int p(x_{1:J} \mid \mu, \beta)p(\mu \mid \nu, \tau)p(\beta)p(\nu, \tau \mid \mathcal{X}_1)\,\mathrm{d}\mu\,\mathrm{d}\beta\,\mathrm{d}\nu\,\mathrm{d}\tau, \tag{10}$$

where the integral is over the whole domain of μ, β, ν, and τ. Restoring c, this posterior predictive approximates the class-conditional likelihood $p(x_{1:J}|c, z, \mathcal{H})$ required by the LLR test. The J observations share (μ, β) and are thus evaluated jointly rather than through separate marginal likelihoods.

Proposition 5.2 (Posterior-predictive density)**.** *Let* $\bar{x}_* = J^{-1}\sum_{j=1}^J x_j$ *and* $S_* = (J-1)^{-1}\sum_{j=1}^J (x_j - \bar{x}_*)^2$. *Integrating* β *over* $p(x_{1:J} \mid \mu, \beta)p(\beta)$ *gives*

$$g(x_{1:J}\,;\,\mu) := \frac{\Gamma(a_\beta + J/2)b_\beta^{a_\beta}}{(2\pi)^{J/2}\Gamma(a_\beta)}\left(b_\beta + \frac{1}{2}[(J-1)S_* + J(\bar{x}_* - \mu)^2]\right)^{-a_\beta - J/2}. \tag{11}$$

Let $U = \sum_{m=1}^M \mu_m$ *and generate* N_s *samples* $\{(U^{(s)}, \tau^{(s)})\}_{s=1}^{N_s}$ *from the Gibbs sampler in Prop. B.2. For Gauss–Hermite nodes and weights* $(t_q, w_q)_{q=1}^{N_q}$, *the Rao–Blackwellized estimator*

$$\hat{p}(x_{1:J} \mid \mathcal{X}_1) = \frac{1}{\sqrt{\pi}N_s}\sum_{s=1}^{N_s}\sum_{q=1}^{N_q} w_q\, g\left(x_{1:J}\,;\,\frac{\tau^{(s)}U^{(s)} + \lambda\xi}{M\tau^{(s)} + \lambda} + \sqrt{\frac{2}{\tau^{(s)}} + \frac{2}{M\tau^{(s)} + \lambda}}\,t_q\right) \tag{12}$$

converges to (10) *as* $N_s, N_q \to \infty$.

App. B.5 gives the derivation and a stable log-domain computation for $\hat{p}(x_{1:J} \mid \mathcal{X}_1)$.

Applying this procedure to the two classes gives the final membership score

$$\widehat{\Lambda}(x_{1:J}) = \log \hat{p}_1(x_{1:J} \mid \mathcal{X}_1^{(1)}) - \log \hat{p}_0(x_{1:J} \mid \mathcal{X}_1^{(0)}). \tag{13}$$

Thus, $\mathcal{X}^{(c)}$ calibrates the class-level prior, $\mathcal{X}_1^{(c)}$ updates the target-point latent variables, and the posterior-predictive density evaluates the target model's J observations jointly while retaining uncertainty from finitely many shadow models. Appendix B.6 discuss a class-shared τ adaptation that is useful with access to few shadow models.

6 Group-Level Membership Inference

A group-level MIA asks whether any data from a group (an individual, entity, or collection of semantically similar points) was used in training. An individual may contribute multiple data points to the training set. The adversary typically queries other points from the same group that were not used to train the target model. In this section, we instantiate our hierarchical framework for this setting, emphasizing reference-model uncertainty, and extend LiRA (Carlini et al., 2022) and BASE (Lassila et al., 2025) to group-level inference in App. D.1.

6.1 Threat Model and Observable Signals

Let $\boldsymbol{\theta}_*$ be the target model, $u \in [U]$ index groups, and $n \in [N]$ index data points, with $u(n)$ the group of point n and $\mathcal{I}_u = \{n \in [N] : u(n) = u\}$ the group u. The adversary queries points in $\mathcal{S}_u \subseteq \mathcal{I}_u$, which need not overlap the target model's training set. App. D.2 formalizes the membership inference game. Let P_u and Q_u denote the distributions of models trained with group u excluded and included, respectively. We test

$$H_{0u} : \boldsymbol{\theta}_* \sim P_u, \qquad H_{1u} : \boldsymbol{\theta}_* \sim Q_u. \tag{14}$$

Under H_{1u}, a queried point may be a training member or an unseen point from a member group; under H_{0u}, it is an unseen point from a non-member group. To simplify the analysis, we model only H_{0u} and use a one-sided test.

The adversary has M independently trained reference models sampled from P_u for the evaluated groups. For point n, model m, and stochastic evaluation $k \in [K]$, let $\ell_{nmk} > 0$ be the loss. We use $y_{nmk} = -\log \ell_{nmk}$ as the membership signal and model it as Gaussian. Larger signal values correspond to lower losses.

6.2 Pointwise Non-Member Model

Let $\bar{y}_{nm} = K^{-1} \sum_{k=1}^{K} y_{nmk}$ and $\hat{e}_{nm} = [K(K-1)]^{-1} \sum_{k=1}^{K} (y_{nmk} - \bar{y}_{nm})^2$, the estimated sampling variance of this mean. For point n and reference model m, our null model is

$$\bar{y}_{nm} \mid \theta_n, v_n \sim \mathcal{N}(\theta_n, v_n + \hat{e}_{nm}) \ , \quad \log v_n \sim \mathcal{N}(\eta, \sigma_v^2) \, . \tag{15}$$

Here θ_n is the pointwise mean and v_n captures between-model variation. We use $p(\theta_n \mid v_n) \propto 1$.

Integrating out θ_n at fixed v_n gives a local marginal likelihood $L_n(v_n)$. We estimate the shared hyperparameters by marginal maximum likelihood:

$$(\hat{\eta}, \hat{\sigma}_v) = \arg\max_{\eta, \sigma_v} \sum_n \log \int_0^\infty L_n(v) p(v \mid \eta, \sigma_v) \, \mathrm{d}v \, . \tag{16}$$

We use this estimate to fit a posterior density of v_n and compute the predictive mean μ_n and variance s_n^2. This is more efficient than retaining the full predictive distribution, at little cost to accuracy. See App. D.3 for technical details.

To capture model-dependent shifts in signal values across data points, including possible target–reference mismatch, we fit the rank-one factor model

$$\bar{y}_{nm} = \mu_n + \lambda_n z_m + \epsilon_{nm} \, , \qquad \epsilon_{nm} \sim \mathcal{N}(0, s_n^2) \, . \tag{17}$$

Here z_m is a one-dimensional model coordinate encoding a systematic shift in model m's signal values across data points. The coefficient λ_n measures point n's sensitivity to this shift. For identification, we impose $M^{-1} \sum_m z_m = 0$ and $M^{-1} \sum_m z_m^2 = 1$, plus a fixed sign convention due to the symmetry $(z_m, \lambda_n) \mapsto (-z_m, -\lambda_n)$. We estimate (λ_n, z_m) by initializing each z_m and alternating regularized MAP (ridge-type) updates of λ_n and z_m until convergence. Because the local model was fitted before introducing this factor, its s_n^2 also includes variation that the factor captures. We subtract $\lambda_n z_m$ from $\bar{y}_{nm}$, recompute (μ_n, s_n^2), and refit the factor once; one pass suffices empirically. The resulting $(\mu_n, s_n^2, \lambda_n, z_m)$ are plug-in quantities for downstream inference. See App. D.4 for refitting details.

6.3 Cross-Fitted Target Model Adaptation and Group-Level Score

Let $\bar{y}_{*n}$ denote the sample average of random loss observations from the target model, and let z_* denote its coordinate. Because the reference coordinates are standardized to empirical mean zero and unit variance, we use the prior $z_* \sim \mathcal{N}(0, 1)$. Under the null, this treats the target model as a new draw from the same standardized coordinate population as the reference models.

Before observing the target model signals, fix an integer $F \geq 2$ and an assignment $\phi : [U] \to [F]$ of groups to disjoint folds. For fold $f \in [F]$, define $\mathcal{U}_{-f} = \{u \in [U] : \phi(u) \neq f\}$ and let $\mathcal{Y}_{*,-f}$ denote the target model signals from groups in $\mathcal{U}_{-f}$. Group u is scored using $f = \phi(u)$ and we use the coordinate posterior $p(z_* \mid \mathcal{Y}_{*,-f})$ to compute the score for that group. In this way, the fold cross-fitting construction ensures target signals for group u does not influence the posterior of z_* used for that score, preventing the shift adjustment to "explain away" the group being tested. We derive an exact Gaussian form of the coordinate posterior in App. D.5.

Conditioning on the fitted reference model quantities $\widehat{\Psi} = \{(\mu_n, s_n^2, \lambda_n) : n \in [N]\}$, the moment-matched null model for a new target observation is

$$\bar{y}_{*n} \mid z_*, \widehat{\Psi} \sim \mathcal{N}(\mu_n + \lambda_n z_*, s_n^2) \, . \tag{18}$$

We take these pointwise predictive variables to be conditionally independent given z_* and $\widehat{\Psi}$.

Our one-sided test-statistic is based on the group aggregated loss signals, $t_{*u} = \sum_{n \in \mathcal{S}_u} \bar{y}_{*n}$. Specifically, we compute a score based on the posterior predictive upper-tail probability

$$S_u = -\log \Pr\Big(T_u \geq t_{*u} \mid D_{*,-f}, \widehat{\Psi}\Big) \ , \tag{19}$$

where $f = \phi(u)$. Large S_u indicates unusually small target model losses under the adapted non-member predictive model. We give an expression for the upper-tail probability in terms of a Gaussian survival function in App. D.6.

7 Experiments

We evaluate our record-level and group-level hierarchical Bayesian MIA on DDPM (Ho et al., 2020; Song et al., 2021b) and flow matching models (Lipman et al., 2023) trained on a variety of image datasets: CIFAR10 (Krizhevsky, 2009), CelebA faces (Liu et al., 2015), ImageNet (Deng et al., 2009; Russakovsky et al., 2015), and VGGFace2 (Cao et al., 2018).

Baseline attacks. We compare our attacks to two state-of-the-art MIAs, LiRA (Carlini et al., 2022) and BASE (Lassila et al., 2025). Since they were originally developed for confidence values from classification models, we adapt them to generative models with stochastic objective functions in App. C. We also compare with three MIAs designed for diffusion models in particular: SecMI (Duan et al., 2023), PIA (Kong et al., 2023), and SimA-MC (Rao et al., 2026).

Architecture. We use a U-net architecture for both our DDPM and flow matching models with a similar architecture as in the seminal work (Ho et al., 2020), including several of the improvements of (Dhariwal and Nichol, 2021). For further details about architecture and hyperparamter settings, see App. E.2.

Online and offline mode. MIAs based on the likelihood ratio test and reference models can naturally support both *online* and *offline* operation modes. The online mode uses both reference models trained on the target point and reference models not trained on the target point, which we refer to as *in* and *out* models, respectively. In offline mode, only out models are used. The likelihood ratio test can only be evaluated in online mode. In offline mode, one can replace the two-sided likelihood ratio test with a one-sided test where the distribution can be modeled solely by out models. A natural one-sided test statistic is the cumulative distribution function (CDF) of the target score under the out distribution (Carlini et al., 2022), which we use in the offline modes of our hierarchical Bayesian MIAs.

Data splits. To obtain an equal number of in and out models for each data point in each benchmark, we follow (Carlini et al., 2022; Lassila et al., 2025) and train reference models such that each audit point is included in the training data of exactly half of the reference models. For the group-level scenario, an audit point is a particular group of data points of which several could be included in a training set. In this case, a group is represented in the training dataset of exactly half the reference models, and the group is represented by a subset of its total data points. See App. E.1 for details.

Loss normalization. Regression objectives used to train diffusion models are typically not normalized. We normalize the loss values by standardizing them over all data points. This allows us to directly compare loss samples from DDPM and flow matching models, enabling us to use a mixture of them as reference models, representing adversarial uncertainty about the target model.

Evaluation. We evaluate MIA performance primarily in the low false positive rate (FPR) regime, following (Carlini et al., 2022). More specifically, we measure true positive rate (TPR) at low and fixed FPR, and standardized partial AUC at 1% FPR.

Record-level experiments. Table 1 compares our record-level hierarchical Bayesian MIA against prior state-of-the-art attacks on a variety of dataset and model type combinations. Our attack performs better overall, and in the few cases where it is not the top performing attack, the difference to the top performance is insignificant. Notably, the performance of all attacks degrade with increasing dataset size. CIFAR10, CelebA, and ImageNet contain 60k, 200k, and 1.3M records, respectively, out of which half are used as training data. Fig. 1 shows ROC curves in log-scale for our attack, the reference-based baselines, and the reference-free baselines. The performance difference between reference-based and reference-free MIAs are striking, with reference-free attacks struggling to improve upon random-guessing. The trend of a performance degradation with increasing dataset size seems likely also for reference-free approaches, which already at the scale of CelebA are only

Table 1: Record-level audits. Values are percentages: mean ± population standard deviation over ten targets. pAUC is standardized at 1% FPR. Bold means are highest within online and offline attacks separately for each benchmark and metric.

	FM / CIFAR10			FM / CelebA			
Attack	pAUC@1%	TPR@1%	TPR@0.1%	pAUC@1%	TPR@1%	TPR@0.1%	TPR@0.01%
BASE	83.95 ± 1.51	80.10 ± 2.56	51.45 ± 4.19	80.35 ± 2.13	73.88 ± 4.25	43.35 ± 3.95	18.42 ± 2.12
LiRA	83.36 ± 1.88	79.04 ± 2.79	48.94 ± 4.92	76.93 ± 2.94	69.83 ± 5.58	32.59 ± 5.39	9.58 ± 2.80
Ours	**86.89** ± 0.93	**81.51** ± 1.77	**64.72** ± 2.04	**82.20** ± 1.71	**75.15** ± 3.44	**50.51** ± 3.46	**27.63** ± 2.63
BASE (off)	72.63 ± 1.01	61.88 ± 2.58	21.80 ± 2.21	73.01 ± 1.53	60.39 ± 4.15	26.54 ± 1.79	7.49 ± 1.86
LiRA (off)	66.09 ± 2.52	49.17 ± 5.93	12.92 ± 3.38	61.31 ± 2.34	37.90 ± 6.47	7.17 ± 2.00	0.86 ± 0.37
Ours (off)	**79.69** ± 1.34	**69.76** ± 2.68	**46.00** ± 2.89	**77.01** ± 2.25	**65.61** ± 4.53	**39.36** ± 4.03	**13.69** ± 3.47

	FM / ImageNet			
Attack	pAUC@1%	TPR@1%	TPR@0.1%	TPR@0.01%
BASE	51.38 ± 0.10	5.67 ± 0.28	0.94 ± 0.11	0.13 ± 0.06
LiRA	51.02 ± 0.08	4.51 ± 0.20	0.71 ± 0.06	0.09 ± 0.03
Ours	**51.97** ± 0.09	**6.80** ± 0.24	**2.03** ± 0.21	**0.66** ± 0.17

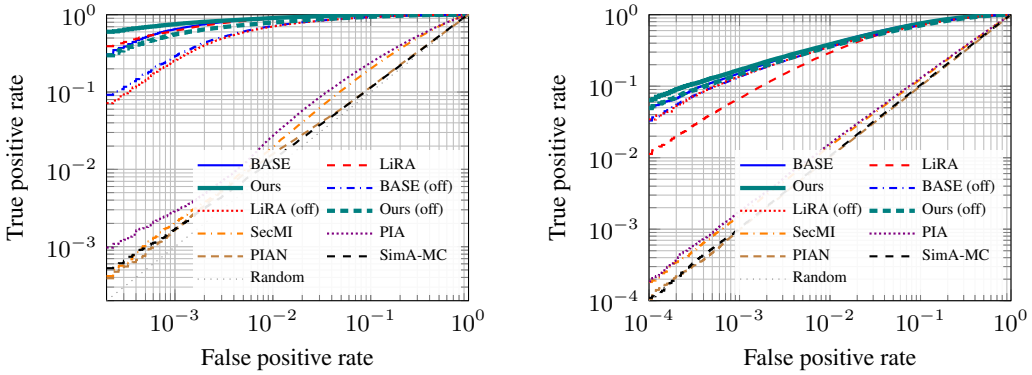


Figure 1: Target-averaged ROC curves for record-level DDPM audits. (Left) CIFAR10. (Right) CelebA. BASE, LiRA, and Ours are shown in both online and offline mode. SecMI uses a step size of 10, and SimA-MC uses the ℓ_4-norm signal. Curves are averaged over ten target models.

Table 2: Group-level audits with mixed DDPM and flow-matching reference models. Values are percentages: mean ± population standard deviation over ten targets. pAUC standardized at 1% FPR, and TPR reported down to 0.1% FPR. 5–15 images per celebrity are audited. Column A represents the case where no target train images was included in the audit set. Column B shows results when at most 1 training image was included. The inclusion of a single training record raises the overall MIA performance significantly, demonstrating the challenge of making group-level inferences in the most realistic scenario where the adversary does not have access to training records. Bold means are highest for each benchmark and metric.

	DDPM / CelebA: A			DDPM / CelebA: B		
Attack	pAUC@1%	TPR@1%	TPR@0.1%	pAUC@1%	TPR@1%	TPR@0.1%
BASE	54.84 ± 0.43	15.69 ± 0.84	4.51 ± 1.05	68.33 ± 0.87	51.99 ± 1.27	19.37 ± 5.11
LiRA	55.90 ± 0.58	16.89 ± 1.02	6.54 ± 1.79	77.35 ± 1.49	68.77 ± 1.55	30.87 ± 8.49
Ours	**61.71** ± 0.86	**31.22** ± 1.84	**15.27** ± 1.86	**82.68** ± 1.25	**78.24** ± 1.33	**43.82** ± 5.80
	FM / CelebA: A			FM / CelebA: B		
Attack	pAUC@1%	TPR@1%	TPR@0.1%	pAUC@1%	TPR@1%	TPR@0.1%
BASE	51.20 ± 0.28	5.33 ± 0.87	0.63 ± 0.31	52.92 ± 0.64	11.13 ± 1.88	1.27 ± 0.74
LiRA	54.60 ± 0.70	12.54 ± 1.57	6.67 ± 1.50	61.93 ± 1.88	29.19 ± 3.86	18.53 ± 4.57
Ours	**57.13** ± 1.12	**19.40** ± 2.43	**9.55** ± 2.05	**66.21** ± 2.30	**40.92** ± 4.83	**23.14** ± 3.83

a slight improvement on random-guessing. This motivates the need for reference-based MIAs for accurate privacy auditing of diffusion models.

Group-level experiments. Table 2 compares our group-level MIA to group-level adaptations of BASE and LiRA, using 16 mixed reference models; 8 DDPM and 8 flow matching. This represents adversarial uncertainty about the target model. Our attack significantly outperforms the baselines in this setting. App. F.1 provides additional experimental results, including on the VGGFace2 dataset.

Conclusion We introduced a hierarchical Bayesian framework that propagates uncertainty across queries, models, data points, and groups. Integrating model-level uncertainty makes it robust to target-reference mismatch. On diffusion and flow-matching models, it outperforms leading record-level attacks and group-level adaptations, even without access to an individual's training images.

AI use statement. In this work, we used generative AI to develop theoretical models, formulate and prove mathematical claims and assist in the writing of them, propose hypotheses, provide feedback on methodology, implement parts of the code, and assist in the writing. We have reviewed all AI-assisted work, in part manually and with the help of AI tools. We take responsibility for the final content of submission.

Ethics statement. This work studies membership inference as a privacy-auditing tool. The same methods can reveal participation in training data, so operational use should follow applicable authorization and disclosure rules. Experiments use synthetic data.

Reproducibility statement. The anonymous supplement (`https://anonymous.4open.science/r/eb-mia-55D9`) provides source code to reproduce all results in the paper.

REFERENCES

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

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *IEEE Comput. Secur. Found. Symp. (CSF)*, pages 268–282, July 2018. doi: 10.1109/CSF.2018.00027.

Alexandre Sablayrolles, Matthijs Douze, Yann Ollivier, Cordelia Schmid, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning (ICML)*, 2019.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *IEEE Symposium on Security and Privacy (SP)*, 2022.

Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *ACM CCS*, 2022.

Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference attacks. In *International Conference on Machine Learning (ICML)*, 2024.

Marcus Lassila, Johan Östman, Khac-Hoang Ngo, and Alexandre Graell i Amat. Practical bayes-optimal membership inference attacks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, 2015.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021a.

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

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Rawisara Lohanimit, Yankun Wu, Amelia Katirai, Yuta Nakashima, and Noa Garcia. Privacy in image datasets: A case study on pregnancy ultrasounds. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2025.

Harry H. Jiang, Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. Ai art and its impact on artists. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2023.

Rickard Brännvall. Exponential-family membership inference: From lira and rmia to bavaria, 2026. URL `https://arxiv.org/abs/2603.11799`.

Lauren Watson, Chuan Guo, Graham Cormode, and Alex Sablayrolles. On the importance of difficulty calibration in membership inference attacks. In *International Conference on Learning Representations (ICLR)*, 2022.

Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? In *International Conference on Machine Learning (ICML)*, 2023.

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. In *International Conference on Learning Representations (ICLR)*, 2023.

Yan Pang and Tianhao Wang. Black-box membership inference attacks against fine-tuned diffusion models. In *NDSS*, 2025.

Jingwei Li, Jing Dong, Tianxing He, and Jingzhao Zhang. Towards black-box membership inference attack for diffusion models. In *International Conference on Machine Learning (ICML)*, 2025.

Mingxing Rao, Bowen Qu, and Daniel Moyer. Score-based membership inference on diffusion models, 2026. URL https://arxiv.org/abs/2509.25003.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *USENIX Conference on Security Symposium*, 2023.

Ryan Webster, Julien Rabin, Loïc Simon, and Frédéric Jurie. This person (probably) exists: Identity membership attacks against gan generated faces. *arXiv preprint arXiv:2107.06018*, 2021.

Guoyao Li, Shahbaz Rezaei, and Xin Liu. User-level membership inference attack against metric embedding learning. In *ICLR PAIR^{2}struct Workshop*, 2022.

Jayneel Vora, Nader Bouacida, Aditya Krishnan, Prabhu R. V. Shankar, and Prasant Mohapatra. Identity-focused inference and extraction attacks on diffusion models. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, 2025.

Jan Dubiński, Antoni Kowalczuk, Franziska Boenisch, and Adam Dziedzic. CDI: Copyrighted data identification in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021b.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A Dataset for Recognising Faces across Pose and Age . In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Jan R. Magnus. The moments of products of quadratic forms in normal variables. *Statistica Neerlandica*, 32(4):201–210, 1978. doi: https://doi.org/10.1111/j.1467-9574.1978.tb01399.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9574.1978.tb01399.x.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Machine Learning (ICML)*, 2019.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2017.

A Notation and Mathematical Preliminaries

A.1 Notation

We denote scalars by italic letters, e.g., x and X, vectors by boldface lower-case letters, e.g., $\boldsymbol{x}$, and matrices by boldface upper-case letters, e.g., $\boldsymbol{X}$. We use $\mathbf{1}_n$, $\mathbf{0}_n$, and $\mathbf{I}_n$ to denote the $n \times 1$ all-one vector, $n \times 1$ all-zero vector, and the $n \times n$ identity matrix, respectively, and we omit the subscript n if the dimensionality is obvious. For a vector $\boldsymbol{x}$, we denote its p-norm by $\|\boldsymbol{x}\|_p$. The subscript $^\top$ denotes the transpose. We denote the set of integers $\{1, 2, \ldots, n\}$ by $[n]$, the indicator function by $\mathbb{1}[\cdot]$, and the set of indexed objects $\{x_1, \ldots, x_n\}$ by $\{x_i\}_{i=1}^n$ or $x_{1:n}$.

For a random variable X, $\mathbb{E}[X]$ and $\mathrm{Var}(X)$ denote its mean and variance, respectively. We denote the n-dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and its probability density function evaluated at $\boldsymbol{x}$ by

$$\mathcal{N}(\boldsymbol{x}\,;\,\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}\det(\boldsymbol{\Sigma})^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right). \tag{20}$$

We denote the Gamma distribution with scale a and rate b by $\Gamma(a, b)$, and its probability density function evaluated at x by

$$\Gamma(x\,;\,a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}, \tag{21}$$

where $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}\mathrm{d}t$ is the Gamma function. We denote by χ^2_n the chi-squared distribution with K degrees of freedom. We use the convention $0/0 = 0$; $\log(\cdot)$ is the natural logarithm.

A.2 Mathematical Preliminaries

We present some mathematical preliminary results that will be used in our later appendices. First, the product of two Gaussian density functions is proportional to a third Gaussian density functions:

$$\mathcal{N}(\boldsymbol{x}\,;\,\boldsymbol{a}, \boldsymbol{A})\,\mathcal{N}(\boldsymbol{x}\,;\,\boldsymbol{b}, \boldsymbol{B}) \propto \mathcal{N}\Big(\boldsymbol{x}\,;\,\boldsymbol{A}^{-1}+\boldsymbol{B}^{-1})^{-1}(\boldsymbol{A}^{-1}\boldsymbol{a}+\boldsymbol{B}^{-1}\boldsymbol{b}), (\boldsymbol{A}^{-1}+\boldsymbol{B}^{-1})^{-1}\Big). \tag{22}$$

For a zero-mean Gaussian vector $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ and a symmetric matrix $\boldsymbol{P}$, Magnus (1978, Lemma 2.3) gives

$$\mathbb{E}[(\boldsymbol{x}^\top \boldsymbol{P}\boldsymbol{x})^2] = 2\,\mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{P}\boldsymbol{\Sigma}) + \{\mathrm{tr}(\boldsymbol{P}\boldsymbol{\Sigma})\}^2. \tag{23}$$

Next, we state that the empirical variance estimate from uncorrelated samples is unbiased for the average variance.

Lemma A.1. *For uncorrelated random variables $\{X_m\}_{m=1}^M$ with common mean ν and variances $\{\sigma_m^2\}_{m=1}^M$, define $\bar{X} = \frac{1}{M}\sum_{m=1}^M X_m$ and $S = \frac{1}{M-1}\sum_{m=1}^M (X_m - \bar{X})^2$. It holds that*

$$\mathbb{E}[S] = \frac{1}{M}\sum_{j=1}^M \sigma_m^2. \tag{24}$$

Proof. Using $\sum_{m=1}^M (X_m - \bar{X})^2 = \sum_{m=1}^M (X_m - \nu)^2 - M(\bar{X} - \nu)^2$, we have

$$\mathbb{E}\left[\sum_{m=1}^M (X_m - \bar{X})^2\right] = \sum_{m=1}^M \sigma_m^2 - M\mathrm{Var}(\bar{X}) \tag{25}$$

$$= \sum_{m=1}^M \sigma_m^2 - M\left(\frac{1}{M^2}\sum_{m=1}^M \sigma_m^2\right) \tag{26}$$

$$= \frac{M-1}{M}\sum_{m=1}^M \sigma_m^2, \tag{27}$$

which proves (24). In (26), we used the assumption that the variables are uncorrelated and thus $\mathrm{Var}(\bar{X}) = \frac{1}{M^2}\sum_{m=1}^M \sigma_m^2$. □

B Derivations for the Record-Level Hierarchical Model

B.1 Detailed Record-Level Membership Inference Game

Following standard practice in the MIA literature (Ye et al., 2022; Carlini et al., 2022; Zarifzadeh et al., 2024), we formulate record-level membership inference as a game between a challenger and an adversary.

Definition 1. (Record-level membership inference game) *Let π denote the data distribution.*

1. *The challenger draws a training dataset $\mathcal{D} = \{z_1, \dots, z_{|\mathcal{D}|}\}$ i.i.d. from π and trains $\boldsymbol{\theta} \leftarrow \mathcal{T}(\mathcal{D})$.*
2. *The challenger draws $c \sim \mathrm{Bernoulli}(1/2)$. If $c = 1$, it draws z uniformly from $\mathcal{D}$; otherwise, it draws $z \sim \pi$ conditioned on $z \notin \mathcal{D}$.*
3. *The challenger gives the adversary z and query access to the target model through a possibly randomized interface $f_{\boldsymbol{\theta}}$.*
4. *Using optional side knowledge $\mathcal{H}$, the adversary outputs $\hat{c} \leftarrow \mathrm{MIA}(f_{\boldsymbol{\theta}}, z, \mathcal{H})$.*
5. *The attack succeeds if $\hat{c} = c$.*

The side knowledge $\mathcal{H}$ may include the model architecture, training procedure, and hyperparameters. The interface $f_{\boldsymbol{\theta}}$ may expose model parameters, losses, or only outputs. We study gray-box access to the randomized loss in (3). At the fixed time t_*, the adversary draws

$$\boldsymbol{\epsilon}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}), \qquad \tilde{x}_j = \log \ell(\boldsymbol{\theta}, z; t_*, \boldsymbol{\epsilon}_j), \qquad j \in [J]. \tag{28}$$

Conditional on $(\boldsymbol{\theta}, z, t_*)$, the raw observations $\tilde{x}_{1:J}$ are i.i.d. We apply the model-wise standardization and sign reversal described in the main text and write the resulting evidence as $x_{1:J}$. The transformation is fixed when the attack is evaluated, so conditional independence is preserved and the derivations below apply directly to the transformed evidence.

B.2 Likelihood-Ratio Test and Shadow-Model Construction

For the game in Definition 1, the ideal score is

$$\Lambda(x_{1:J}) = \log \frac{p(x_{1:J} \mid c = 1, z, \mathcal{H})}{p(x_{1:J} \mid c = 0, z, \mathcal{H})}. \tag{29}$$

Let $z_1 = z$ and choose auxiliary points $z_2, \dots, z_N$. For every $n \in [N]$, $c \in \{0, 1\}$, and $m \in [M]$, construct

$$\mathbb{1}[z_n \in \mathcal{D}_{nm}^{(c)}] = c, \qquad \boldsymbol{\theta}_{nm}^{(c)} \leftarrow \mathcal{T}(\mathcal{D}_{nm}^{(c)}), \tag{30}$$

and collect conditionally i.i.d. observations

$$x_{nmk}^{(c)} = \log \ell\Big(\boldsymbol{\theta}_{nm}^{(c)}, z_n; t_*, \boldsymbol{\epsilon}_{nmk}^{(c)}\Big), \qquad \boldsymbol{\epsilon}_{nmk}^{(c)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}), \qquad k \in [K]. \tag{31}$$

Writing $\mathcal{X}_n^{(c)} = \{x_{nmk}^{(c)}\}_{m,k}$ and $\mathcal{X}^{(c)} = \{\mathcal{X}_n^{(c)}\}_{n=1}^N$, the fitted hierarchy gives

$$\widehat{\Lambda}(x_{1:J}) = \log \frac{p(x_{1:J} \mid \mathcal{X}_1^{(1)}, \hat{\eta}^{(1)})}{p(x_{1:J} \mid \mathcal{X}_1^{(0)}, \hat{\eta}^{(0)})}. \tag{32}$$

The auxiliary points may be reused across audits; only the class-conditional observations for a new target point must be added.

B.3 Empirical-Bayes Moment Calibration

We suppress the class label because the derivation first applies to either class separately. Define

$$\bar{x}_{nm} = \frac{1}{K}\sum_{k=1}^{K} x_{nmk}, \qquad V_{nm} = \frac{1}{K-1}\sum_{k=1}^{K}(x_{nmk} - \bar{x}_{nm})^2, \tag{33}$$

$$\bar{x}_n = \frac{1}{M}\sum_{m=1}^{M} \bar{x}_{nm}, \qquad B_n = \frac{1}{M-1}\sum_{m=1}^{M}(\bar{x}_{nm} - \bar{x}_n)^2, \tag{34}$$

$$x = \frac{1}{N}\sum_{n=1}^{N} \bar{x}_n, \qquad \bar{u}_n = B_n - \frac{1}{KM}\sum_{m=1}^{M} V_{nm}. \tag{35}$$

The correction in u_n removes the contribution of finite-K query noise from the across-model sample variance.

Proposition B.1 (Moment-based prior calibration)**.** *Let*

$$\hat{q}_{\beta,1} = \frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M} V_{nm}, \qquad \hat{q}_{\beta,2} = \frac{K-1}{NM(K+1)}\sum_{n=1}^{N}\sum_{m=1}^{M} V_{nm}^2, \tag{36}$$

$$\hat{\sigma}_x^2 = \frac{1}{N-1}\sum_{n=1}^{N}(\bar{x}_n - \bar{x})^2, \qquad y_n = \frac{\bar{x}_n - \bar{x}}{\max(\hat{\sigma}_x, \sqrt{\epsilon})}, \tag{37}$$

$$\tilde{u}_n = \max(u_n, \epsilon), \qquad \hat{\gamma} = \frac{\sum_{n=1}^{N} y_n \left(\log \tilde{u}_n - N^{-1}\sum_{h=1}^{N} \log \tilde{u}_h\right)}{\max(\sum_{n=1}^{N} y_n^2, \epsilon)}, \tag{38}$$

$$\hat{r}_n = \frac{\exp(\hat{\gamma} y_n)}{N^{-1}\sum_{h=1}^{N}\exp(\hat{\gamma} y_h)}, \qquad \hat{q}_{\tau,1} = \frac{1}{N}\sum_{n=1}^{N}\frac{u_n}{\hat{r}_n}. \tag{39}$$

The second inverse-τ moment used by the attack is

$$\hat{q}_{\tau,2} = \frac{M-1}{M+1}\left(\frac{1}{N}\sum_{n=1}^{N}\frac{u_n^2}{\hat{r}_n^2} - \frac{4\hat{q}_{\tau,1}\hat{q}_{\beta,1}}{K(M-1)} - \frac{2\hat{q}_{\beta,2}}{K(K-1)M} - \frac{2\hat{q}_{\beta,1}^2}{K^2M(M-1)}\right). \tag{40}$$

For $\zeta \in \{\beta, \tau\}$, admissible inverse moments give

$$\hat{a}_\zeta = 2 + \frac{\hat{q}_{\zeta,1}^2}{\hat{q}_{\zeta,2} - \hat{q}_{\zeta,1}^2}, \qquad \hat{b}_\zeta = \hat{q}_{\zeta,1}(\hat{a}_\zeta - 1). \tag{41}$$

Finally,

$$\hat{\xi} = \bar{x}, \qquad \hat{\lambda} = \left[\max\left(\hat{\sigma}_x^2 - \frac{\hat{q}_{\tau,1}}{M} - \frac{\hat{q}_{\beta,1}}{KM}, \epsilon\right)\right]^{-1}. \tag{42}$$

The normalization in (39) *ensures $N^{-1}\sum_n \hat{r}_n = 1$.*

Gamma inverse moments. If $Q \sim \Gamma(a, b)$ in shape–rate form and $a > 2$, then

$$q_1 = \mathbb{E}[Q^{-1}] = \frac{b}{a-1}, \qquad q_2 = \mathbb{E}[Q^{-2}] = \frac{b^2}{(a-1)(a-2)}. \tag{43}$$

Solving these equations gives (41).

Query-precision moments. Conditional on β_{nm}, Cochran's theorem gives

$$(K-1)\beta_{nm}V_{nm} \sim \chi^2_{K-1}. \tag{44}$$

Consequently, $\mathbb{E}[V_{nm}] = q_{\beta,1}$ and $\mathbb{E}[V_{nm}^2] = (K+1)q_{\beta,2}/(K-1)$, which yields the estimates of $q_{\beta,1}$ and $q_{\beta,2}$ in (36).

Point-specific dispersion. After integrating μ_{nm},

$$\bar{x}_{nm} \mid \nu_n, \tau_n, \beta_{nm}, r_n \sim \mathcal{N}\left(\nu_n, \frac{r_n}{\tau_n} + \frac{1}{K\beta_{nm}}\right). \tag{45}$$

The model means are conditionally independent across m. Lemma A.1 and the conditional unbiasedness of V_{nm} therefore imply

$$\mathbb{E}[u_n \mid \tau_n, r_n] = \frac{r_n}{\tau_n}, \qquad \mathbb{E}[u_n \mid r_n] = r_n q_{\tau,1}. \tag{46}$$

Thus $\log \tilde{u}_n \approx \alpha + \gamma y_n$. Ordinary least squares with an intercept gives the slope in (39); exponentiation followed by mean-one normalization gives $\hat{r}_n$.

Second inverse-τ moment. Let $C_n = (KM)^{-1} \sum_m V_{nm}$, so $u_n = B_n - C_n$, and write $w_{nm} = r_n/\tau_n + \beta_{nm}^{-1}/K$. Conditional Gaussian sample means are independent of their within-query sample variances, hence B_n is independent of C_n given $(\tau_n, \boldsymbol{\beta}_n, r_n)$. With $\boldsymbol{w}_n = (w_{n1}, \ldots, w_{nM})$,

$$\mathbb{E}[B_n^2 \mid \tau_n, \boldsymbol{\beta}_n, r_n] = \frac{2(M-2)}{M(M-1)^2}\|\boldsymbol{w}_n\|_2^2 + \frac{M^2 - 2M + 3}{M^2(M-1)^2}\|\boldsymbol{w}_n\|_1^2, \tag{47}$$

$$\mathbb{E}[C_n^2 \mid \boldsymbol{\beta}_n] = \frac{1}{K^2M^2}\left[\left(\sum_m \beta_{nm}^{-1}\right)^2 + \frac{2}{K-1}\sum_m \beta_{nm}^{-2}\right]. \tag{48}$$

Expanding $\mathbb{E}[(B_n - C_n)^2]$ and averaging over τ_n and $\boldsymbol{\beta}_n$ gives

$$\mathbb{E}[u_n^2 \mid r_n] = \frac{M+1}{M-1} r_n^2 q_{\tau,2} + \frac{4 r_n q_{\tau,1} q_{\beta,1}}{K(M-1)} + \frac{2 q_{\beta,2}}{K(K-1)M} + \frac{2 q_{\beta,1}^2}{K^2 M(M-1)}. \tag{49}$$

For fixed fitted scales, define $\rho_j = N^{-1} \sum_n r_n^{-j}$. Dividing (49) by r_n^2, averaging over n, and solving yields the exact moment identity

$$q_{\tau,2} = \frac{M-1}{M+1}\left(\mathbb{E}_n\left[\frac{u_n^2}{r_n^2}\right] - \frac{4 q_{\tau,1} q_{\beta,1}}{K(M-1)}\rho_1 - \frac{2 q_{\beta,2}}{K(K-1)M}\rho_2 - \frac{2 q_{\beta,1}^2}{K^2 M(M-1)}\rho_2\right). \tag{50}$$

Equation (40) is the estimator used in our experiments: it retains the pointwise r_n^{-2} adjustment in the leading empirical term and uses the mean-one approximation $\rho_1 = \rho_2 = 1$ in the finite-K corrections. This approximation is exact for $r_n \equiv 1$ and treats the smaller query-noise corrections as homoscedastic after the leading pointwise adjustment.

Center prior. Conditional on $(\nu_n, \tau_n, \boldsymbol{\beta}_n, r_n)$,

$$\mathrm{Var}(\bar{x}_n \mid \nu_n, \tau_n, \boldsymbol{\beta}_n, r_n) = \frac{r_n}{M\tau_n} + \frac{1}{KM^2}\sum_{m=1}^{M} \beta_{nm}^{-1}. \tag{51}$$

The law of total variance and $N^{-1} \sum_n r_n = 1$ give

$$\mathbb{E}_n[\mathrm{Var}(\bar{x}_n)] = \lambda^{-1} + \frac{q_{\tau,1}}{M} + \frac{q_{\beta,1}}{KM}, \tag{52}$$

which gives (42); $\hat{\xi} = \bar{x}$ follows from the marginal mean.

Numerical regularization. The exact Gamma fit requires $q_1 > 0$ and $q_2 > q_1^2$. If empirical moments violate these conditions, we floor the first moment at ϵ, use the finite fallback shape $a = 2.05$, and set $b = q_1(a-1)$. Extremely concentrated valid fits are capped at $a = 10^6$. These conventions ensure proper priors but are numerical regularization rather than additional moment identities.

B.4 GIBBS SAMPLER

For the target point, suppress the point and class labels and let

$$\kappa = \frac{\tau}{r}, \qquad \bar{x}_m = \frac{1}{K}\sum_{k=1}^{K} x_{mk}, \qquad Q_m = \sum_{k=1}^{K}(x_{mk} - \bar{x}_m)^2. \tag{53}$$

Proposition B.2 (Full conditionals and convergence)**.** *For fixed empirical-Bayes parameters and $r > 0$, the full conditionals are*

$$\mu_m \mid \cdot \sim \mathcal{N}\left(\frac{K\beta_m \bar{x}_m + \kappa\nu}{K\beta_m + \kappa}, (K\beta_m + \kappa)^{-1}\right), \tag{54a}$$

$$\beta_m \mid \cdot \sim \Gamma\left(a_\beta + K/2, b_\beta + \frac{1}{2}[Q_m + K(\bar{x}_m - \mu_m)^2]\right), \tag{54b}$$

$$\nu \mid \cdot \sim \mathcal{N}\left(\frac{\kappa\sum_{m=1}^{M}\mu_m + \lambda\xi}{M\kappa + \lambda}, (M\kappa + \lambda)^{-1}\right), \tag{54c}$$

$$\tau \mid \cdot \sim \Gamma\left(a_\tau + M/2, b_\tau + \frac{1}{2r}\sum_{m=1}^{M}(\mu_m - \nu)^2\right). \tag{54d}$$

A fixed systematic scan through these conditionals is Harris ergodic and converges in total variation to $p(\mu_{1:M}, \beta_{1:M}, \nu, \tau \mid \mathcal{X}_1)$ from every valid initial state.

Algorithm 1 Gibbs sampling from the local latent posterior.

Require: Local reference observations $\mathcal{X}_1$, fitted parameters $\hat{\eta}$, and scale r
1: Initialize $\mu_{1:M}$, $\beta_{1:M} > 0$, ν, and $\tau > 0$
2: **for** $s = 1, \ldots, S_{\text{warm}} + S$ **do**
3: Draw every μ_m using (54a)
4: Draw every β_m using (54b)
5: Draw ν using (54c)
6: Draw τ using (54d)
7: **if** $s > S_{\text{warm}}$ **then**
8: Retain $U = \sum_m \mu_m$ and τ
9: **end if**
10: **end for**

Derivation of the conditionals. The joint density factorizes as

$$\left\{\prod_{m=1}^{M} p(x_{m,1:K} \mid \mu_m, \beta_m)p(\mu_m \mid \nu, \tau, r)p(\beta_m)\right\} p(\nu)p(\tau). \tag{55}$$

For μ_m, completing the square combines precisions $K\beta_m$ and κ. For β_m, the likelihood decomposition

$$\sum_{k=1}^{K}(x_{mk} - \mu_m)^2 = Q_m + K(\bar{x}_m - \mu_m)^2 \tag{56}$$

combines with its Gamma prior. The M Gaussian factors for μ_m combine with the prior precision λ to give the ν conditional. Finally, the terms involving τ have kernel

$$\tau^{a_\tau + M/2 - 1}\exp\left[-\tau\left(b_\tau + \frac{1}{2r}\sum_{m}(\mu_m - \nu)^2\right)\right], \tag{57}$$

which gives (54d).

Convergence proof. The state space is $\mathbb{R}^M \times (0,\infty)^M \times \mathbb{R} \times (0,\infty)$. All priors are proper, and integrating each observation likelihood against its Gamma precision prior gives a finite function bounded uniformly in μ_m; hence the posterior is proper. Every full conditional above has a continuous, strictly positive density on its support. A complete systematic sweep therefore has a continuous transition density $k(s,s') > 0$ for every pair of states. It follows that the chain is irreducible with respect to Lebesgue measure. On any compact rectangle contained in the state space, continuity and positivity give a one-step minorization, so the chain is aperiodic. Each coordinate update leaves the posterior invariant, and therefore so does a complete sweep. An irreducible chain with an invariant probability distribution is positive Harris recurrent; together with aperiodicity this proves Harris ergodicity, uniqueness of the invariant posterior, and total-variation convergence. The corresponding ergodic theorem justifies Monte Carlo averages of integrable functions of the retained states.

B.5 Posterior-Predictive Evaluation

For a new model under one fitted class,

$$\beta_* \sim \Gamma\big(a_\beta, b_\beta\big)\,, \qquad \mu_* \mid \nu,\tau,r \sim \mathcal{N}\big(\nu, r/\tau\big)\,, \qquad x_j \mid \mu_*,\beta_* \sim \mathcal{N}\Big(\mu_*, \beta_*^{-1}\Big)\,. \tag{58}$$

Let $\bar{x}_* = J^{-1}\sum_j x_j$ and $Q_* = \sum_j (x_j - \bar{x}_*)^2$. Integrating β_* gives

$$g(x_{1:J}\,;\,\mu) = \frac{\Gamma(a_\beta + J/2) b_\beta^{a_\beta}}{(2\pi)^{J/2}\Gamma(a_\beta)} \left[b_\beta + \frac{1}{2}\Big(Q_* + J(\bar{x}_* - \mu)^2\Big)\right]^{-a_\beta - J/2}. \tag{59}$$

Given a retained Gibbs state $(U^{(s)}, \tau^{(s)})$, define

$$\kappa^{(s)} = \frac{\tau^{(s)}}{r}, \qquad m^{(s)} = \frac{\kappa^{(s)} U^{(s)} + \lambda\xi}{M\kappa^{(s)} + \lambda}, \qquad v^{(s)} = \frac{r}{\tau^{(s)}} + \frac{1}{M\kappa^{(s)} + \lambda}. \tag{60}$$

The Gaussian ν conditional in (54c) and the new-model distribution in (58) imply

$$\mu_* \mid U^{(s)}, \tau^{(s)}, \mathcal{X}_1 \sim \mathcal{N}\Big(m^{(s)}, v^{(s)}\Big)\,. \tag{61}$$

Therefore the Rao–Blackwellized online predictive density is

$$p(x_{1:J} \mid \mathcal{X}_1) = \mathbb{E}_{U,\tau\mid\mathcal{X}_1}\left[\int_{-\infty}^{\infty} g(x_{1:J}\,;\,\mu)\mathcal{N}(\mu\,;\,m,v)\;\mathrm{d}\mu\right]. \tag{62}$$

For Gauss–Hermite nodes and weights $(t_q, w_q)_{q=1}^{N_q}$, this becomes

$$\hat{p}(x_{1:J} \mid \mathcal{X}_1) = \frac{1}{\sqrt{\pi}N_s}\sum_{s=1}^{N_s}\sum_{q=1}^{N_q} w_q\, g\Big(x_{1:J}\,;\, m^{(s)} + \sqrt{2v^{(s)}}t_q\Big)\,. \tag{63}$$

The online score is the difference between the IN and OUT log predictive densities. Numerically, each quadrature sum and the outer Gibbs average are evaluated with log-sum-exp:

$$h^{(s)} = \mathrm{LogSumExp}_q\left[\log w_q + \log g\Big(x_{1:J}\,;\, m^{(s)} + \sqrt{2v^{(s)}}t_q\Big)\right] - \tfrac{1}{2}\log\pi, \tag{64}$$

$$\log\hat{p}(x_{1:J} \mid \mathcal{X}_1) = \mathrm{LogSumExp}_s(h^{(s)}) - \log N_s. \tag{65}$$

Offline score. With only OUT reference models, the attack uses the target query mean. Integrating β_* in $\bar{X}_* \mid \mu_*, \beta_* \sim \mathcal{N}\big(\mu_*, (J\beta_*)^{-1}\big)$ gives a Student-t distribution with $2a_\beta$ degrees of freedom, location μ_*, and scale

$$s_\beta = \sqrt{\frac{b_\beta}{a_\beta J}}. \tag{66}$$

Writing F_{t_ν} for the standard Student-t CDF, the implemented one-sided score is

$$L_{\mathrm{off}}(x_{1:J}) = \log\mathbb{E}_{U,\tau\mid\mathcal{X}_1^{(0)}}\left[\int F_{t_{2a_\beta}}\left(\frac{\bar{x}_* - \mu}{s_\beta}\right)\mathcal{N}(\mu\,;\,m,v)\;\mathrm{d}\mu\right]. \tag{67}$$

The inner Gaussian expectation uses the same Gauss–Hermite rule as (63), and the outer expectation averages over OUT-class Gibbs states. Signals are oriented so that larger values indicate smaller losses; hence larger lower-tail probabilities indicate stronger membership evidence.

B.6 Sharing the Between-Model Precision Across Classes

The hierarchy above gives IN and OUT separate point-specific precisions $\tau_n^{(1)}$ and $\tau_n^{(0)}$. In the online setting, an alternative is to share one τ_n and one r_n while retaining class-specific centers, query precisions, and hyperparameters:

$$\mu_{nm}^{(c)} \mid \nu_n^{(c)}, \tau_n, r_n \sim \mathcal{N}\Big(\nu_n^{(c)}, r_n/\tau_n\Big), \qquad c \in \{0,1\}. \tag{68}$$

All class-specific conditionals remain as above with $\kappa_n = \tau_n/r_n$. With M reference models per class, the shared precision conditional is

$$\tau_n \mid \cdot \sim \Gamma\left(a_\tau + M, b_\tau + \frac{1}{2r_n}\sum_{c=0}^{1}\sum_{m=1}^{M}(\mu_{nm}^{(c)} - \nu_n^{(c)})^2\right). \tag{69}$$

For empirical-Bayes calibration, let $u_n^{(c)}$ be the corrected variance proxy from (35). We estimate r_n from the mean of the two class centers and the pooled local variance, and define $v_n^{(c)} = u_n^{(c)}/r_n$. The shared inverse moments are estimated by

$$\hat{q}_{\tau,1} = \frac{1}{N}\sum_{n=1}^{N}\max\left(\frac{v_n^{(0)} + v_n^{(1)}}{2}, \epsilon\right), \qquad \hat{q}_{\tau,2} = \frac{1}{N}\sum_{n=1}^{N} v_n^{(0)} v_n^{(1)}. \tag{70}$$

The cross-class product estimates the second inverse moment without squaring either class's noisy local variance proxy. The resulting Gamma prior is shared, whereas $(\xi^{(c)}, \lambda^{(c)}, a_\beta^{(c)}, b_\beta^{(c)})$ remain class-specific.

Sharing τ_n pools the IN and OUT residuals in every Gibbs update. It can therefore reduce dispersion-estimation noise when only a small number of reference models is available, while preserving different IN and OUT locations and query-noise distributions. The tradeoff is the additional assumption that the two membership classes have the same between-model dispersion at a fixed point. This construction requires both classes and is therefore used only for online inference; offline inference retains the class-specific OUT model.

B.7 Modeling Considerations

Alignment with the attack signals. The hierarchy mirrors their construction. (9a) models randomized evaluations of a fixed shadow model and data point; (9b) allows the signal distribution to vary across independently trained shadow models; and (9c) models variation across data points. Thus, each source of variation has a corresponding level rather than being pooled into a single signal distribution.

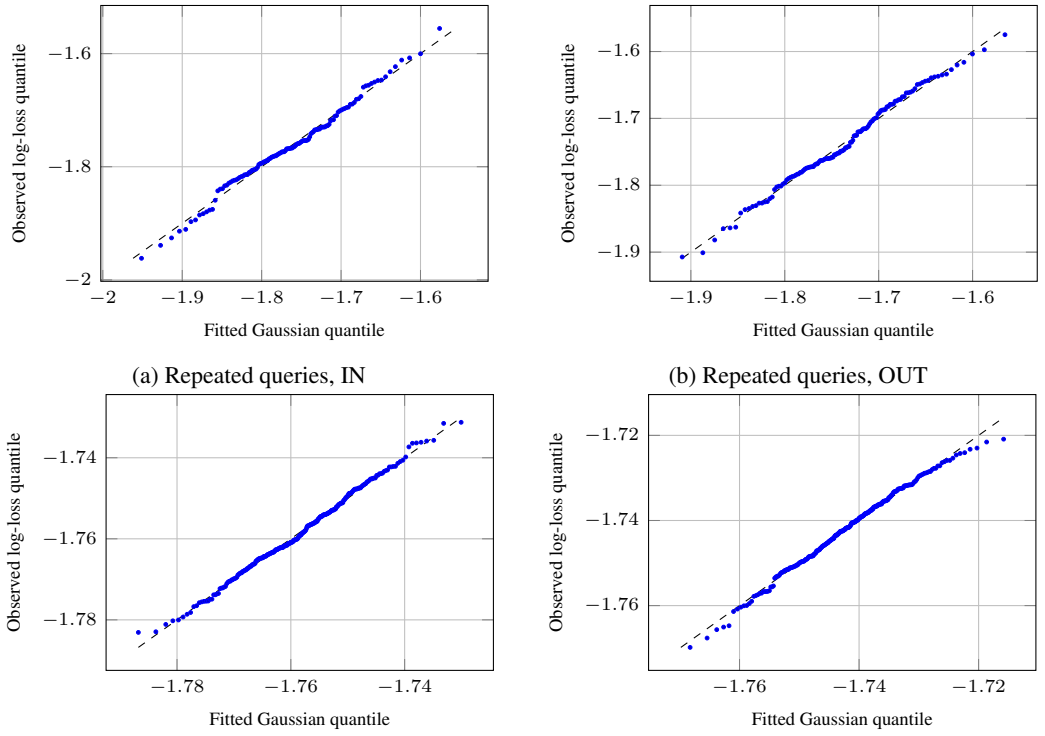
Tractability. The Gaussian–Gamma structure yields conjugate full conditional distributions for every latent variable. This supports the Gibbs sampler in App. B.4, while analytic marginalization and Gauss–Hermite quadrature reduce the Monte Carlo variance of the posterior-predictive estimator in App. B.5.

Empirical support. We first assess the Gaussian assumptions directly on the unstandardized log-loss observations in (3). We use 512 lightweight CIFAR-10 models and 200 randomly selected data points, with 256 models in each membership class and 128 loss evaluations per model and data point. For each model and point, we test the repeated log-loss observations appearing in (9a). For each point and membership class, we additionally test the distribution of the 256 model-specific sample means. These means provide observable estimates of the latent model means in (9b).

Table 3: Normality diagnostics for unstandardized log loss. Entries in the final two columns give the percentage of distributions for which normality is rejected at the 5% level.

Level	Class	Tests	Shapiro–Wilk	Anderson–Darling
Repeated queries	IN	51,200	5.2%	4.8%
Repeated queries	OUT	51,200	5.4%	4.9%
Across-model means	IN	200	5.0%	5.0%
Across-model means	OUT	200	9.0%	4.0%

At the repeated-query level, both tests reject at approximately their nominal rate. The same holds for the IN model means. For OUT model means, the Shapiro–Wilk rejection rate is somewhat elevated, while the tail-weighted Anderson–Darling test remains near its nominal rate. Representative Q–Q plots in Fig. 2 likewise show close agreement with the fitted Gaussian distributions, with limited deviations at the extremes. These diagnostics support the conditional Gaussian observation model and are consistent with the across-model Gaussian approximation, while not implying that either distribution is exactly Gaussian.



(a) Repeated queries, IN

(b) Repeated queries, OUT

(c) Across-model means, IN

(d) Across-model means, OUT

Figure 2: Representative Gaussian Q–Q plots for raw log loss at a randomly selected data point, using randomly selected models for the repeated-query panels.

Effect of model-wise standardization. The experiments use the transformation $z_{nmj} = (x_{nmj} - a_m)/b_m$, where a_m and $b_m > 0$ are model-specific location and scale parameters. Within a fixed model, this is a common positive affine transformation. It therefore preserves Gaussianity and leaves the fitted Q–Q shape, Shapiro–Wilk statistic, and Anderson–Darling statistic unchanged. Accordingly, the repeated-query rejection rates are identical before and after standardization.

Across models, the transformation parameters vary with m, so preservation of Gaussianity does not follow from affine invariance alone. We therefore repeat the across-model diagnostics after standardization. The Shapiro–Wilk and Anderson–Darling rejection rates remain 5% and 5% for IN means and change only from 9% and 4% to 8% and 6% for OUT means. Thus, in these diagnostics, model-wise standardization does not materially disturb the approximate normality observed for the raw log loss.

C Record-Level Baseline Adaptations

LiRA (Carlini et al., 2022) and BASE (Lassila et al., 2025) assume that querying a model on a data point produces one deterministic scalar. The diffusion and flow-matching objectives considered here instead produce a stochastic loss through the sampled auxiliary variable ϵ. We adapt the baselines by first reducing the repeated loss observations to one scalar for each model–record pair and then applying their original record-level scoring rules. This construction allows the baselines

to use repeated queries, but, unlike our hierarchical model, does not retain or propagate query-level uncertainty after this reduction.

C.1 Reduction of Stochastic Loss Observations

Recall that $x_{nmk}^{(c)}$ is the log loss of reference model m on data point z_n under membership class $c \in \{0, 1\}$ and stochastic query $k \in [K]$. Because the scales of the regression objectives can differ across models, we standardize the log losses separately for each model over all evaluated data points and stochastic queries. Let the resulting observation be $\widetilde{x}_{nmk}^{(c)}$. We then define the scalar evidence

$$a_{nm}^{(c)} = -\frac{1}{K} \sum_{k=1}^{K} \widetilde{x}_{nmk}^{(c)}. \tag{71}$$

The target-model evidence is constructed analogously from its J queries:

$$a_{n*} = -\frac{1}{J} \sum_{j=1}^{J} \widetilde{x}_{nj*}. \tag{72}$$

The negative sign makes larger values correspond to smaller-than-typical losses and hence stronger evidence of membership. We use $J = K = 32$ in the experiments.

C.2 BASE

For data point z_n, let $\mathcal{R}_n^{(c)}$ contain the reference models for which its membership class is c. Online BASE uses $\mathcal{R}_n = \mathcal{R}_n^{(0)} \cup \mathcal{R}_n^{(1)}$, whereas offline BASE uses only $\mathcal{R}_n = \mathcal{R}_n^{(0)}$. Writing $M_n = |\mathcal{R}_n|$, the adapted BASE score is

$$B_n = \log \frac{\pi}{1-\pi} + a_{n*} - \log \left(\frac{1}{M_n} \sum_{m \in \mathcal{R}_n} \exp(a_{nm}) \right). \tag{73}$$

Here a_{nm} denotes the evidence from the appropriate membership class for reference model m. We use the uninformative prior $\pi = 1/2$, so the prior log-odds vanish. A logistic transformation of B_n is unnecessary for ROC evaluation because it preserves the ordering of the scores.

C.3 LiRA

LiRA fits Gaussian distributions to the scalar evidence in (71). For each point and membership class, let $M_n^{(c)} = |\mathcal{R}_n^{(c)}|$ and estimate the mean by

$$\widehat{\mu}_n^{(c)} = \frac{1}{M_n^{(c)}} \sum_{m \in \mathcal{R}_n^{(c)}} a_{nm}^{(c)}. \tag{74}$$

Define the corresponding residual sum of squares and degrees of freedom as

$$R_n^{(c)} = \sum_{m \in \mathcal{R}_n^{(c)}} \left(a_{nm}^{(c)} - \widehat{\mu}_n^{(c)} \right)^2, \qquad d_n^{(c)} = M_n^{(c)} - 1. \tag{75}$$

We consider two ways to share variance information across data points. A *local* variance is estimated separately for every z_n, while a *global* variance pools residuals over all N audited points. We also consider either *class-specific* IN and OUT variances or a *shared* variance for both classes. The resulting estimators are

$$\widehat{\sigma}_{n,c,\mathrm{local}}^2 = \frac{R_n^{(c)}}{d_n^{(c)}}, \qquad \widehat{\sigma}_{c,\mathrm{global}}^2 = \frac{\sum_{n=1}^{N} R_n^{(c)}}{\sum_{n=1}^{N} d_n^{(c)}}, \tag{76}$$

$$\widehat{\sigma}_{n,\mathrm{local}}^2 = \frac{R_n^{(0)} + R_n^{(1)}}{d_n^{(0)} + d_n^{(1)}}, \qquad \widehat{\sigma}_{\mathrm{global}}^2 = \frac{\sum_{n=1}^{N} \sum_{c=0}^{1} R_n^{(c)}}{\sum_{n=1}^{N} \sum_{c=0}^{1} d_n^{(c)}}. \tag{77}$$

Thus, local estimation allows the dispersion to depend on the difficulty of the target point, whereas global estimation trades this flexibility for a larger effective sample size. Sharing the variance between classes further reduces estimation noise while retaining separate IN and OUT means.

In the online setting, LiRA assigns the Gaussian LLR

$$L_n^{\mathrm{LiRA}} = \log \frac{\mathcal{N}\left(a_{n*};\, \widehat{\mu}_n^{(1)}, (\widehat{\sigma}_n^{(1)})^2\right)}{\mathcal{N}\left(a_{n*};\, \widehat{\mu}_n^{(0)}, (\widehat{\sigma}_n^{(0)})^2\right)}, \tag{78}$$

where the variances are selected from Eqs. (76) and (77). In the offline setting, only OUT reference models are available. We therefore use the one-sided score

$$L_{n,\mathrm{off}}^{\mathrm{LiRA}} = \log \Phi\left(\frac{a_{n*} - \widehat{\mu}_n^{(0)}}{\widehat{\sigma}_n^{(0)}}\right), \tag{79}$$

where Φ is the standard Gaussian CDF. Larger values indicate that the target evidence is unusually large under the OUT distribution.

C.4 Variance Parameterization

We compare the four online LiRA variance parameterizations using 32 stochastic loss observations per model–record pair. The matched-reference results average over DDPM and flow-matching targets on CIFAR10 and CelebA with reference models of the same type. The mixed-reference results use the same targets but combine DDPM and flow-matching reference models. Each aggregate contains four model–dataset settings and ten target models per setting.

Table 4: Online LiRA variance ablation. Values are percentages averaged over four model–dataset settings and ten targets per setting. pAUC is standardized at 1% FPR.

		Matched references		Mixed references	
Across points	Across classes	pAUC@1%	TPR@0.1%	pAUC@1%	TPR@0.1%
Global	Shared	77.28	33.71	69.09	22.77
Global	Class-specific	**77.41**	34.02	69.18	23.02
Local	**Shared**	77.07	**38.20**	**74.78**	**38.23**
Local	Class-specific	61.36	5.72	70.67	27.09

The local shared estimator gives the highest TPR at 0.1% FPR in both settings and the highest pAUC with mixed reference models. Global class-specific variance improves matched-reference pAUC by only 0.34 percentage points, while performing substantially worse at low FPR and under reference-model mismatch. We therefore use local shared variance for online LiRA. Sharing across classes is not applicable offline, where only the OUT class is observed; there we use a local OUT variance.

D Derivations for the Group-Level Hierarchical Model

D.1 Group-Level Extensions of LiRA and BASE

This subsection specifies the group-level baselines referred to in the main text. The construction is deliberately simple: a record-level attack produces one scalar item of evidence for each queried point, and the group-level baseline adds that evidence over the queried points in the group. This aggregation should be viewed as a baseline extension rather than as an exact likelihood-ratio model for H_{1u}, because under group membership a queried point can itself be either present or absent from the target training set.

For LiRA Carlini et al. (2022), let a_{*n} denote the scalar record-level signal evaluated on the target model at point n, and collect the signals for group u into

$$\mathbf{a}_{*u} = (a_{*n})_{n \in \mathcal{S}_u} \in \mathbb{R}^{d_u}, \qquad d_u = |\mathcal{S}_u|\,. \tag{80}$$

The IN and OUT classes refer here to whether group u, rather than necessarily every queried point in $\mathcal{S}_u$, is represented in the training data of the corresponding reference model. For each class $c \in \{\text{in}, \text{out}\}$, we fit

$$\mathbf{a}_{um}^c \sim \mathcal{N}(\boldsymbol{\mu}_u^c, \boldsymbol{\Sigma}_u^c)\,. \tag{81}$$

We consider four covariance structures: a spherical covariance $\boldsymbol{\Sigma}_u^c = (\sigma_u^c)^2\mathbf{I}$; a diagonal covariance with a separate variance for every queried point; a low-rank-plus-diagonal covariance $\boldsymbol{\Sigma}_u^c = \mathbf{B}_u^c(\mathbf{B}_u^c)^\top + \mathbf{D}_u^c$; and a regularized full covariance. The dispersion parameters can be estimated separately for each group or pooled across groups.

For online LiRA, the group score is the multivariate LLR

$$L_u^{\text{LiRA}} = \log \frac{\varphi_{d_u}(\mathbf{a}_{*u}; \boldsymbol{\mu}_u^{\text{in}}, \boldsymbol{\Sigma}_u^{\text{in}})}{\varphi_{d_u}(\mathbf{a}_{*u}; \boldsymbol{\mu}_u^{\text{out}}, \boldsymbol{\Sigma}_u^{\text{out}})}\,, \tag{82}$$

where $\varphi_d(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the d-dimensional Gaussian density. For the spherical and diagonal covariance models, this density factorizes and (82) reduces to the sum of the pointwise LiRA LLRs,

$$L_u^{\text{LiRA}} = \sum_{n \in \mathcal{S}_u} \log \frac{\varphi(a_{*n}; \mu_n^{\text{in}}, (\sigma_n^{\text{in}})^2)}{\varphi(a_{*n}; \mu_n^{\text{out}}, (\sigma_n^{\text{out}})^2)}\,. \tag{83}$$

If only OUT reference models are available, we use the joint upper-tail probability

$$p_u^{\text{out}} = \Pr\big(\mathbf{A}_u^{\text{out}} \geq \mathbf{a}_{*u}\big)\,, \qquad \mathbf{A}_u^{\text{out}} \sim \mathcal{N}(\boldsymbol{\mu}_u^{\text{out}}, \boldsymbol{\Sigma}_u^{\text{out}}), \tag{84}$$

where the inequality is componentwise, and take $-\log p_u^{\text{out}}$ as the offline score. For spherical and diagonal covariances, this is exactly the sum of the pointwise negative log upper-tail probabilities. For low-rank and full covariances, the multivariate Gaussian upper-orthant probability is evaluated numerically. We found no meaningful difference in attack performance among these covariance models and therefore use the simplest spherical covariance in our comparisons.

For BASE Lassila et al. (2025), let $\bar{\ell}_{*n}$ be the target-model loss used by the record-level attack, let $\bar{\ell}_{nm}$ be the corresponding loss from reference model m, and let π be the nominal membership prior entering each record-level BASE score. If $\mathcal{R}_n$ is the set of reference models used for point n and $M_n = |\mathcal{R}_n|$, its log-odds score is

$$B_n = \log \frac{\pi}{1-\pi} - \bar{\ell}_{*n} - \log\left(\frac{1}{M_n}\sum_{m \in \mathcal{R}_n} e^{-\bar{\ell}_{nm}}\right). \tag{85}$$

The direct additive group generalization is

$$B_u^{\text{BASE}} = \sum_{n \in \mathcal{S}_u} B_n\,. \tag{86}$$

We use the uninformative prior $\pi = 1/2$ throughout, for which the prior log-odds vanish. Thus, in our comparisons, (86) is simply the sum of the record-level BASE evidence terms. The logistic transformation $\{1 + \exp(-B_u^{\text{BASE}})\}^{-1}$ can be used when a posterior-probability scale is desired, but the additive score itself is sufficient because thresholding either quantity gives the same ordering. As with the group LiRA baseline, (86) aggregates record-level evidence and does not explicitly model the mixture of member and non-member points that may occur under H_{1u}.

D.2 Detailed Group-level Membership Inference Game

We model the data as arising from a mixture of group-specific distributions. Let $\mathcal{U} \subseteq \mathbb{N}$ be a set of group indices and let π_i denote the data distribution of group $i \in \mathcal{U}$. The data distribution is then

$$\pi = \sum_{i \in \mathcal{U}} w_i \pi_i\,, \tag{87}$$

where the weights $w_i \geq 0$ satisfy $\sum_{i \in \mathcal{U}} w_i = 1$. A sample from π is generated by first drawing a group index from the categorical distribution $\{w_i\}_{i \in \mathcal{U}}$, then drawing a data point from the corresponding π_i. A dataset is a finite collection of i.i.d. samples from π.

We formulate group-level membership inference as a security game between a challenger and an adversary, following standard practice in the membership inference literature Ye et al. (2022); Carlini et al. (2022); Zarifzadeh et al. (2024).

Definition 2. (Group-level membership inference game)

1. *The challenger samples a subset $\mathcal{U}_1 \subset \mathcal{U}$, not necessarily proportionally to their contribution to π (in practice, $\mathcal{U}_1$ is typically a uniformly random subset of $\mathcal{U}$), and draws a training dataset $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ with $x_i \overset{i.i.d.}{\sim} \sum_{j\in\mathcal{U}_1} w'_j \pi_j$, and where $\{w'_j\}_{j\in\mathcal{U}_1}$ are the weights $\{w_j\}_{j\in\mathcal{U}_1}$ renormalized to sum to one. A target model is then trained via training procedure $\mathcal{A}$,*

$$\boldsymbol{\theta} \leftarrow \mathcal{A}(\mathcal{D})\,.$$

2. *The challenger flips a fair coin to generate a bit $b \in \{0, 1\}$. If $b = 1$, a group index i is drawn uniformly from $\mathcal{U}_1$; otherwise, it is drawn uniformly from $\mathcal{U} \setminus \mathcal{U}_1$.*
3. *The challenger provides the adversary with the group index i, query access to the target model $f_{\boldsymbol{\theta}}$, and query access to each group distribution π_j, $j \in \mathcal{U}$.*
4. *The adversary mounts a membership inference attack to produce a guess $\hat{b}$, potentially using additional side knowledge $\mathcal{H}$,*

$$\hat{b} \leftarrow \mathsf{MIA}(f_{\boldsymbol{\theta}},\, i,\, \{\pi_j\}_{j\in\mathcal{U}},\, \mathcal{H})\,.$$

5. *The attack is successful if $\hat{b} = b$.*

Remarks. The form of the query access to the model $f_{\boldsymbol{\theta}}$ is intentionally left unspecified, so as to capture a broad range of interaction modes. For generative models such as diffusion or flow-matching models, this includes querying the underlying neural network on a data sample to evaluate the associated loss or likelihood. The side knowledge $\mathcal{H}$ represents any auxiliary information available to the adversary that may strengthen the attack, such as knowledge of the model architecture, training procedure, or hyperparameters.

D.3 Local Marginal Likelihood and Predictive Moments

Fix a data point n and abbreviate

$$x_m = \bar{y}_{nm}\,, \qquad e_m = \hat{e}_{nm}\,. \tag{88}$$

For a fixed residual variance $v > 0$, define

$$w_m(v) = \frac{1}{v + e_m}\,, \qquad P(v) = \sum_{m=1}^{M} w_m(v)\,, \qquad h(v) = \sum_{m=1}^{M} w_m(v) x_m\,. \tag{89}$$

The conditional likelihood of the M reference-model means is

$$p(x_{1:M} \mid \theta, v) = (2\pi)^{-M/2} \prod_{m=1}^{M} (v + e_m)^{-1/2} \exp\left[-\frac{1}{2}\sum_{m=1}^{M} w_m(v)(x_m - \theta)^2\right]. \tag{90}$$

To integrate out θ, first expand the weighted residual sum of squares:

$$\sum_{m=1}^{M} w_m(v)(x_m - \theta)^2 = \sum_{m=1}^{M} w_m(v) x_m^2 - 2\theta h(v) + \theta^2 P(v)\,. \tag{91}$$

Completing the square gives

$$\sum_{m=1}^{M} w_m(v)(x_m - \theta)^2 = \sum_{m=1}^{M} w_m(v) x_m^2 - \frac{h(v)^2}{P(v)} + P(v)\left(\theta - \frac{h(v)}{P(v)}\right)^2. \tag{92}$$

With the flat location prior $p(\theta \mid v) \propto 1$, the integral over θ is Gaussian:

$$\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}P(v)\left(\theta - \frac{h(v)}{P(v)}\right)^2\right] \mathrm{d}\theta = \sqrt{\frac{2\pi}{P(v)}}\,. \tag{93}$$

Therefore, up to the arbitrary multiplicative constant of the improper flat prior, the local marginal likelihood is

$$L_n(v) \propto P(v)^{-1/2} \prod_{m=1}^{M} (v + e_m)^{-1/2} \exp\left[-\frac{1}{2} \left\{ \sum_{m=1}^{M} w_m(v) x_m^2 - \frac{h(v)^2}{P(v)} \right\} \right]. \tag{94}$$

The same completion of the square also gives the conditional posterior of the pointwise mean:

$$\theta_n \mid v, D_n \sim \mathcal{N}\Big(m_n(v), P_n(v)^{-1} \Big) , \qquad m_n(v) = \frac{h_n(v)}{P_n(v)} , \tag{95}$$

where $D_n = \{(\bar{y}_{nm}, \hat{e}_{nm})\}_{m=1}^{M}$ and the subscript n restores the point index.

The population model for the residual variance is

$$\log v_n \sim \mathcal{N}(\eta, \sigma_v^2) . \tag{96}$$

Its density on the original variance scale is

$$p(v \mid \eta, \sigma_v) = \frac{1}{v \sigma_v \sqrt{2\pi}} \exp\left[-\frac{(\log v - \eta)^2}{2\sigma_v^2} \right] , \qquad v > 0 . \tag{97}$$

Conditional independence across points implies that the marginal likelihood for (η, σ_v) factorizes as

$$p(D_{1:N} \mid \eta, \sigma_v) \propto \prod_{n=1}^{N} \int_0^\infty L_n(v) p(v \mid \eta, \sigma_v) \, \mathrm{d}v . \tag{98}$$

Taking logarithms produces the empirical-Bayes objective used in the main text:

$$(\hat{\eta}, \hat{\sigma}_v) = \arg\max_{\eta, \sigma_v} \sum_{n=1}^{N} \log \int_0^\infty L_n(v) p(v \mid \eta, \sigma_v) \, \mathrm{d}v . \tag{99}$$

After fixing $(\hat{\eta}, \hat{\sigma}_v)$, Bayes' rule yields

$$p_n(v) = \frac{L_n(v) p(v \mid \hat{\eta}, \hat{\sigma}_v)}{\int_0^\infty L_n(t) p(t \mid \hat{\eta}, \hat{\sigma}_v) \, \mathrm{d}t} . \tag{100}$$

We next derive the predictive moments for a new independently trained non-member model. Let $\bar{Y}_{n0}$ denote its mean transformed-loss signal at point n. We use

$$\hat{e}_{n0} = \frac{1}{M} \sum_{m=1}^{M} \hat{e}_{nm} \tag{101}$$

as the plug-in repeated-evaluation variance for a new model. Conditional on (θ_n, v),

$$\bar{Y}_{n0} \mid \theta_n, v \sim \mathcal{N}(\theta_n, v + \hat{e}_{n0}) . \tag{102}$$

Integrating the Gaussian posterior in (95) gives

$$\bar{Y}_{n0} \mid v, D_n \sim \mathcal{N}\big(m_n(v), q_n(v) \big) , \tag{103}$$

where

$$q_n(v) = v + \hat{e}_{n0} + P_n(v)^{-1} . \tag{104}$$

The full local predictive distribution is therefore the one-dimensional Gaussian mixture

$$p(\bar{y} \mid D_n) = \int_0^\infty \varphi\big(\bar{y}; m_n(v), q_n(v)\big) \, p_n(v) \, \mathrm{d}v . \tag{105}$$

Its predictive mean follows from the law of total expectation:

$$\mu_n = \mathbb{E}[\bar{Y}_{n0} \mid D_n] = \int_0^\infty m_n(v) p_n(v) \, \mathrm{d}v . \tag{106}$$

For the variance, first note that

$$\mathbb{E}[\bar{Y}_{n0}^2 \mid v, D_n] = q_n(v) + m_n(v)^2 \,. \tag{107}$$

Hence

$$\mathbb{E}[\bar{Y}_{n0}^2 \mid D_n] = \int_0^\infty \{q_n(v) + m_n(v)^2\} p_n(v)\,\mathrm{d}v \,. \tag{108}$$

Subtracting the square of the predictive mean yields

$$s_n^2 = \int_0^\infty \{q_n(v) + m_n(v)^2\} p_n(v)\,\mathrm{d}v - \mu_n^2 \,. \tag{109}$$

Thus μ_n and s_n^2 are the exact first two moments of the local posterior-predictive mixture in (105). The later replacement of that mixture by a Gaussian with the same mean and variance is a moment-matching approximation; the moment calculations themselves are exact under the local hierarchy.

All remaining integrals are one-dimensional. To evaluate them with Gauss–Hermite quadrature, set $t = \log v$ and then

$$t = \hat{\eta} + \sqrt{2}\hat{\sigma}_v x \,. \tag{110}$$

If $(x_j, a_j)_{j=1}^J$ are Gauss–Hermite nodes and weights for integrals of the form $\int e^{-x^2} g(x)\,\mathrm{d}x$, then for any function $g(v)$,

$$\int_0^\infty g(v) p(v \mid \hat{\eta}, \hat{\sigma}_v)\,\mathrm{d}v \approx \frac{1}{\sqrt{\pi}} \sum_{j=1}^J a_j g\Big(\exp\{\hat{\eta} + \sqrt{2}\hat{\sigma}_v x_j\}\Big) \,. \tag{111}$$

The evidence, posterior normalization, predictive mean, and predictive second moment are all ratios or weighted sums of terms of the form in (111). In implementation, their logarithms are evaluated with a log-sum-exp calculation to avoid numerical underflow.

D.4 Rank-One Factor Fit and Residual Refitting

This subsection makes the alternating ridge updates and the single residual-refitting pass explicit. Suppose an initial local fit has produced (μ_n, s_n^2) for every point. Define

$$r_{nm} = \bar{y}_{nm} - \mu_n \,, \qquad w_n = s_n^{-2} \,. \tag{112}$$

The factor fit minimizes the penalized weighted sum of squares

$$\mathcal{J}(\lambda, z) = \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^M w_n (r_{nm} - \lambda_n z_m)^2 + \frac{\rho_\lambda}{2} \sum_{n=1}^N \lambda_n^2 \,, \tag{113}$$

subject to

$$\sum_{m=1}^M z_m = 0 \,, \qquad \sum_{m=1}^M z_m^2 = M \,. \tag{114}$$

The second constraint is equivalent to empirical unit variance because the first constraint fixes the empirical mean at zero. A fixed sign convention removes the remaining transformation $(z, \lambda) \mapsto (-z, -\lambda)$. A quadratic penalty on z is unnecessary under (114), because $\sum_m z_m^2$ is then constant.

For fixed z, the terms in (113) involving a single loading λ_n are

$$\mathcal{J}_n(\lambda_n) = \frac{w_n}{2} \sum_{m=1}^M (r_{nm} - \lambda_n z_m)^2 + \frac{\rho_\lambda}{2} \lambda_n^2 \,. \tag{115}$$

Differentiating gives

$$\frac{\partial \mathcal{J}_n}{\partial \lambda_n} = -w_n \sum_{m=1}^M z_m r_{nm} + \lambda_n \left(w_n \sum_{m=1}^M z_m^2 + \rho_\lambda \right) \,. \tag{116}$$

Setting the derivative to zero and using $\sum_m z_m^2 = M$ gives the ridge update

$$\lambda_n = \frac{\sum_{m=1}^{M} z_m r_{nm}}{M + \rho_\lambda s_n^2} \,. \tag{117}$$

For fixed λ, define

$$c_m = \sum_{n=1}^{N} w_n \lambda_n r_{nm} \,. \tag{118}$$

Expanding (113) as a function of z gives

$$\mathcal{J}(z) = C + \frac{1}{2} \left(\sum_{n=1}^{N} w_n \lambda_n^2 \right) \sum_{m=1}^{M} z_m^2 - \sum_{m=1}^{M} c_m z_m \,, \tag{119}$$

where C does not depend on z. Under the fixed-norm constraint in (114), the quadratic term is constant. The update therefore maximizes $c^\top z$ subject to zero mean and squared norm M. Let

$$\bar{c} = \frac{1}{M} \sum_{m=1}^{M} c_m \,, \qquad c_m^\circ = c_m - \bar{c} \,. \tag{120}$$

The zero-mean constraint restricts z to the same subspace as c°. By Cauchy–Schwarz,

$$\sum_{m=1}^{M} c_m^\circ z_m \le \left(\sum_{m=1}^{M} (c_m^\circ)^2 \right)^{1/2} \left(\sum_{m=1}^{M} z_m^2 \right)^{1/2} \,. \tag{121}$$

Equality holds when z is proportional to c°. Enforcing $\sum_m z_m^2 = M$ gives

$$z_m = \frac{\sqrt{M}\, c_m^\circ}{\left\{ \sum_{j=1}^{M} (c_j^\circ)^2 \right\}^{1/2}} \,. \tag{122}$$

If the denominator is zero, the current loadings contain no information about a model coordinate and the factor update is left at its previous value. Otherwise, (122) is followed by the fixed sign convention. Alternating Eqs. (117) and (122) decreases the objective until a stationary point of the constrained rank-one problem is reached.

The initial local fit was obtained before introducing $\lambda_n z_m$. Consequently, model-to-model variance generated by the factor can be absorbed into the initial s_n^2. To remove this double counting, let $(\lambda_n^{(0)}, z_m^{(0)})$ denote the first factor fit and form the residualized reference signals

$$\bar{y}_{nm}^{\mathrm{res}} = \bar{y}_{nm} - \lambda_n^{(0)} z_m^{(0)} \,. \tag{123}$$

The complete local empirical-Bayes procedure of the previous subsection is then rerun with $\bar{y}_{nm}^{\mathrm{res}}$ in place of $\bar{y}_{nm}$ and the same $\hat{e}_{nm}$. This produces refitted predictive moments $(\mu_n^{(1)}, (s_n^2)^{(1)})$. Finally, the factor objective is rerun on the original $\bar{y}_{nm}$ using these refitted local moments, yielding $(\lambda_n^{(1)}, z_m^{(1)})$.

To see why the residual pass separates the two sources of variation, suppose temporarily that the fitted factor model were exact and that the reference coordinates had variance one. Then, across independently trained models,

$$\mathrm{Var}(\bar{y}_{nm} \mid n) = \lambda_n^2 + \mathrm{Var}(\epsilon_{nm}) \,. \tag{124}$$

A local model fitted directly to $\bar{y}_{nm}$ cannot distinguish these terms and can therefore attribute part of λ_n^2 to its residual predictive variance. After subtracting $\lambda_n z_m$, the local refit instead targets the residual component. The final factor is then reintroduced explicitly through $\lambda_n z_m$ rather than being counted again inside s_n^2. Repeating these two stages to convergence would more closely approximate a joint fit; the implemented method stops after one residual-refitting pass.

D.5 Cross-Fitted Posterior of the Target-Model Coordinate

We now derive the Gaussian coordinate posterior used for target-model adaptation. Fix a fold f and consider only groups $u \in \mathcal{U}_{-f}$. The quantities $(\mu_n, s_n^2, \lambda_n)$ are treated as fixed plug-in estimates from the reference models.

For coordinate adaptation only, we augment the pointwise working model with an auxiliary group intercept δ_u:

$$\bar{y}_{*n} = \mu_n + \lambda_n z_* + \delta_u + \epsilon_{*n}\,, \qquad \epsilon_{*n} \sim \mathcal{N}(0, s_n^2)\,, \tag{125}$$

for $n \in \mathcal{S}_u$. The intercept is not part of the final null score for group u; its only purpose is to prevent a common shift of all queried signals from another group from being interpreted as evidence about the global model coordinate. We assign the translation-invariant prior $p(\delta_u) \propto 1$ and integrate δ_u analytically.

Define

$$r_n = \bar{y}_{*n} - \mu_n\,, \qquad w_n = s_n^{-2}\,, \qquad W_u = \sum_{n \in \mathcal{S}_u} w_n\,. \tag{126}$$

For any pointwise quantity a_n, define its weighted mean within group u by

$$\langle a \rangle_u^w = \frac{1}{W_u} \sum_{n \in \mathcal{S}_u} w_n a_n\,. \tag{127}$$

For a fixed value of z_*, let

$$a_n(z_*) = r_n - \lambda_n z_*\,. \tag{128}$$

The weighted residual sum of squares for group u is

$$R_u(\delta_u, z_*) = \sum_{n \in \mathcal{S}_u} w_n \{a_n(z_*) - \delta_u\}^2\,. \tag{129}$$

Its weighted mean is

$$\langle a(z_*) \rangle_u^w = \langle r \rangle_u^w - z_* \langle \lambda \rangle_u^w\,. \tag{130}$$

Using the weighted sum-of-squares decomposition,

$$R_u(\delta_u, z_*) = \sum_{n \in \mathcal{S}_u} w_n \{a_n(z_*) - \langle a(z_*) \rangle_u^w\}^2 + W_u \{\delta_u - \langle a(z_*) \rangle_u^w\}^2\,. \tag{131}$$

The Gaussian likelihood contribution of group u is proportional to $\exp\{-R_u(\delta_u, z_*)/2\}$. Integrating the flat-prior intercept gives

$$\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2} W_u \{\delta_u - \langle a(z_*) \rangle_u^w\}^2\right] \mathrm{d}\delta_u = \sqrt{\frac{2\pi}{W_u}}\,. \tag{132}$$

The factor on the right does not depend on z_*. Therefore the marginal likelihood for z_* depends only on the centered residuals. Define

$$\widetilde{r}_n = r_n - \langle r \rangle_u^w\,, \qquad \widetilde{\lambda}_n = \lambda_n - \langle \lambda \rangle_u^w\,. \tag{133}$$

Then

$$a_n(z_*) - \langle a(z_*) \rangle_u^w = \widetilde{r}_n - z_* \widetilde{\lambda}_n\,. \tag{134}$$

The z_*-dependent group contribution is therefore

$$\exp\left[-\frac{1}{2} \sum_{n \in \mathcal{S}_u} w_n (\widetilde{r}_n - z_* \widetilde{\lambda}_n)^2\right]. \tag{135}$$

Expand the quadratic term:

$$\sum_{n \in \mathcal{S}_u} w_n (\widetilde{r}_n - z_* \widetilde{\lambda}_n)^2 = C_u - 2 z_* b_u + z_*^2 I_u\,, \tag{136}$$

where

$$C_u = \sum_{n \in \mathcal{S}_u} w_n \widetilde{r}_n^2\,, \tag{137}$$

$$b_u = \sum_{n \in \mathcal{S}_u} w_n \widetilde{\lambda}_n \widetilde{r}_n \,, \tag{138}$$

and

$$I_u = \sum_{n \in \mathcal{S}_u} w_n \widetilde{\lambda}_n^2 \,. \tag{139}$$

The constants C_u do not depend on z_*. Multiplying the contributions from all groups outside fold f gives the likelihood kernel

$$p(\mathcal{Y}_{*,-f} \mid z_*, \widehat{\Psi}) \propto \exp\left[-\frac{1}{2} \left\{ z_*^2 \sum_{u \in \mathcal{U}_{-f}} I_u - 2 z_* \sum_{u \in \mathcal{U}_{-f}} b_u \right\} \right]. \tag{140}$$

The prior $z_* \sim \mathcal{N}(0,1)$ contributes the kernel $\exp(-z_*^2/2)$. Hence the posterior kernel is

$$p(z_* \mid \mathcal{Y}_{*,-f}, \widehat{\Psi}) \propto \exp\left[-\frac{1}{2} \left\{ \left(1 + \sum_{u \in \mathcal{U}_{-f}} I_u \right) z_*^2 - 2 z_* \sum_{u \in \mathcal{U}_{-f}} b_u \right\} \right]. \tag{141}$$

Define

$$V_f = \left(1 + \sum_{u \in \mathcal{U}_{-f}} I_u \right)^{-1} . \tag{142}$$

The coefficient of the linear term implies

$$m_f = V_f \sum_{u \in \mathcal{U}_{-f}} b_u \,. \tag{143}$$

Completing the square in (141) gives

$$\left(1 + \sum_u I_u \right) z_*^2 - 2 z_* \sum_u b_u = V_f^{-1} (z_* - m_f)^2 - V_f^{-1} m_f^2 \,, \tag{144}$$

where both sums are over $u \in \mathcal{U}_{-f}$. The last term is constant in z_*. Therefore

$$z_* \mid \mathcal{Y}_{*,-f}, \widehat{\Psi} \sim \mathcal{N}(m_f, V_f) \,. \tag{145}$$

This Gaussian form is exact conditional on the moment-matched pointwise working model, the plug-in reference quantities, and the flat auxiliary priors for the group offsets. Cross-fitting enters only through the set $\mathcal{U}_{-f}$: for $f = \phi(u)$, none of the target signals from the scored group u are present in (145).

D.6 Posterior-Predictive Group Tail Probability

Fix a group u and let $f = \phi(u)$. Under H_{0u}, conditional on z_* and the plug-in reference quantities, the main-text working model is

$$\bar{Y}_{*n} \mid z_*, \widehat{\Psi} \sim \mathcal{N}(\mu_n + \lambda_n z_*, s_n^2) \,, \qquad n \in \mathcal{S}_u \,, \tag{146}$$

with conditional independence across n. Here $\bar{Y}_{*n}$ denotes the random target-model signal under the null, while $\bar{y}_{*n}$ denotes the observed value. Define the random aggregate and its observed realization by

$$T_u = \sum_{n \in \mathcal{S}_u} \bar{Y}_{*n} \,, \qquad t_{*u} = \sum_{n \in \mathcal{S}_u} \bar{y}_{*n} \,. \tag{147}$$

Also define

$$\mu_u = \sum_{n \in \mathcal{S}_u} \mu_n \,, \tag{148}$$

$$\lambda_u = \sum_{n \in \mathcal{S}_u} \lambda_n \,, \tag{149}$$

and

$$s_u^2 = \sum_{n \in \mathcal{S}_u} s_n^2 \,. \tag{150}$$

A sum of conditionally independent Gaussian variables is Gaussian, so (146) implies

$$T_u \mid z_*, \widehat{\Psi} \sim \mathcal{N}(\mu_u + \lambda_u z_*, s_u^2) \,. \tag{151}$$

The cross-fitted posterior from (145) is

$$z_* \mid \mathcal{Y}_{*,-f}, \widehat{\Psi} \sim \mathcal{N}(m_f, V_f) \,. \tag{152}$$

Because fold f is held out, the stochastic residual in T_u is conditionally independent of the data used to obtain this coordinate posterior. We may therefore represent

$$z_* = m_f + \sqrt{V_f} Z_0 \,, \qquad Z_0 \sim \mathcal{N}(0,1) \,, \tag{153}$$

and

$$T_u = \mu_u + \lambda_u z_* + s_u Z_1 \,, \qquad Z_1 \sim \mathcal{N}(0,1) \,, \tag{154}$$

with Z_0 and Z_1 independent. Substitution gives

$$T_u = \mu_u + \lambda_u m_f + \lambda_u \sqrt{V_f} Z_0 + s_u Z_1 \,. \tag{155}$$

A linear combination of independent standard Gaussian variables is Gaussian. Its mean is

$$\mathbb{E}[T_u \mid \mathcal{Y}_{*,-f}, \widehat{\Psi}] = \mu_u + \lambda_u m_f \,, \tag{156}$$

and its variance is

$$\mathrm{Var}(T_u \mid \mathcal{Y}_{*,-f}, \widehat{\Psi}) = s_u^2 + \lambda_u^2 V_f \,. \tag{157}$$

Consequently,

$$T_u \mid \mathcal{Y}_{*,-f}, \widehat{\Psi}, H_{0u} \sim \mathcal{N}\Big(\mu_u + \lambda_u m_f, s_u^2 + \lambda_u^2 V_f\Big) \,. \tag{158}$$

The one-sided predictive tail probability of the observed statistic is therefore

$$p_u = \Pr\Big(T_u \geq t_{*u} \mid \mathcal{Y}_{*,-f}, \widehat{\Psi}, H_{0u}\Big) = \overline{\Phi}\left(\frac{t_{*u} - \mu_u - \lambda_u m_f}{\sqrt{s_u^2 + \lambda_u^2 V_f}}\right) \,, \tag{159}$$

where Φ is the standard-normal cumulative distribution function and $\overline{\Phi}(x) = 1 - \Phi(x)$ is its survival function. The membership-inference score used in the main text is

$$S_u = -\log p_u \,. \tag{160}$$

Larger S_u corresponds to a target statistic farther into the upper tail of the adapted non-member predictive distribution. The Gaussian calculation from (151) onward is exact within the working model. The approximation relative to the original local hierarchy enters earlier, when the local Gaussian mixture in (105) is replaced by a Gaussian having the same mean μ_n and variance s_n^2, and when the fitted factor quantities are treated as plug-in values rather than integrated over their estimation uncertainty.

E Experimental Details

Here we provide additional details regarding the experiments.

E.1 Data Splitting

In order to allow for an efficient use of reference models, we follow the procedure used in (Carlini et al., 2022) and include each audit point in exactly half of the reference models training data. This is achieved by randomly splitting the data in two halves and training a reference model pair on each half. Here an audit point is a record, (i.e., an image,) in the record-level scenario, and an individual, (i.e., an individual represented in multiple images,) in the group-level scenario. For the record-level case, this means each model is trained on 50% of the dataset. In the group-level case, we further sub-sample records within each group to allow for auditing only images unseen to the target model. Target models are selected from the pool of reference models and when attacking a target model, all other reference models, except the one trained on the complementing half of the target, are used by reference-based attacks. Evaluation metrics reported are averaged over all target models from the reference model pool.

E.2 Target and Reference Model Training

Our flow matching and DDPM models both use a U-net architecture similar to what is used in the original works (Ho et al., 2020; Lipman et al., 2023; Dhariwal and Nichol, 2021) and we use similar hyperparameter settings. We use a Gaussian Fourier embedding (Tancik et al., 2020; Song et al., 2021a) to represent the continuous time dependence of flow matching models and a sinusoidal positional embedding (Vaswani et al., 2017) for conditioning on discrete time in DDPM models. We use an Adam optimizer (Kingma and Ba, 2015) with decoupled weight decay (Loshchilov and Hutter, 2019) and EMA with decay factor 0.9999. On some datasets we use a linear learning rate scheduler with warm-up. To choose the number of epochs, we train one model for a large number of epochs and store checkpoints at intermediate steps. Then we compute a FID score (Heusel et al., 2017) at each intermediate checkpoint and chose the number of epochs that yield the lowest FID score. The FID score is computed primarily on images not seen during training. Although our FID scores are comparable to the FID scores reported in the literature, we stress that our primary goal is not to achieve the highest performing model possible, but to achieve a well-performing representative model to attack.

F Additional Experiments

F.1 Record-level Reference-model Mismatch

Table 5 evaluates robustness to reference-model mismatch. For each target model, the reference pool contains eight DDPM and eight flow-matching models, whereas the target belongs to only one of these model families. Thus, half of the reference models differ from the target in their generative modeling objective. Loss normalization places signals from the two model families on a common scale. The results retain the evaluation protocol and metrics of the main record-level experiments.

Table 5: Record-level audits with mismatched reference models. Values are percentages: mean ± population standard deviation over ten targets. Each attack uses eight DDPM and eight flow-matching reference models. CIFAR10 reports TPR at 0.1% FPR. CelebA reports standardized pAUC at 1% FPR and TPR down to 0.01% FPR. Bold means are highest within online and offline attacks separately for each benchmark and metric. The incomplete flow-matching offline experiments are omitted.

	DDPM / CIFAR10	DDPM / CelebA			
Attack	TPR@0.1%	pAUC@1%	TPR@1%	TPR@0.1%	TPR@0.01%
BASE	48.45 ± 7.25	56.01 ± 0.82	18.85 ± 2.24	5.71 ± 1.09	1.44 ± 0.35
LiRA	64.51 ± 4.40	59.21 ± 1.56	27.95 ± 3.83	8.37 ± 1.93	1.82 ± 0.51
Ours	$\mathbf{65.62} \pm 3.74$	$\mathbf{61.18} \pm 1.34$	$\mathbf{30.06} \pm 3.04$	$\mathbf{13.88} \pm 2.23$	$\mathbf{5.87} \pm 1.24$
BASE (off)	24.08 ± 3.73	55.66 ± 0.73	17.13 ± 1.85	5.93 ± 0.94	1.82 ± 0.24
LiRA (off)	20.73 ± 2.93	54.94 ± 0.65	15.44 ± 1.69	4.91 ± 0.81	1.35 ± 0.23
Ours (off)	$\mathbf{53.77} \pm 5.15$	$\mathbf{61.83} \pm 1.51$	$\mathbf{32.25} \pm 3.37$	$\mathbf{13.77} \pm 2.31$	$\mathbf{4.91} \pm 1.09$
	FM / CIFAR10	FM / CelebA			
Attack	TPR@0.1%	pAUC@1%	TPR@1%	TPR@0.1%	TPR@0.01%
BASE	17.92 ± 3.55	62.69 ± 1.87	37.94 ± 4.54	12.04 ± 2.34	3.52 ± 1.00
LiRA	50.08 ± 3.73	72.53 ± 2.28	57.33 ± 4.38	29.51 ± 4.33	10.77 ± 2.61
Ours	$\mathbf{55.39} \pm 2.75$	$\mathbf{74.23} \pm 1.57$	$\mathbf{59.03} \pm 3.58$	$\mathbf{35.20} \pm 2.34$	$\mathbf{16.45} \pm 2.08$

F.2 ImageNet Record-level Results

Figure 3 reports the online attacks on ImageNet. The larger audit set supports evaluating the ROC curves down to 10^{-5} FPR.

F.3 Group-level Matched Reference Models

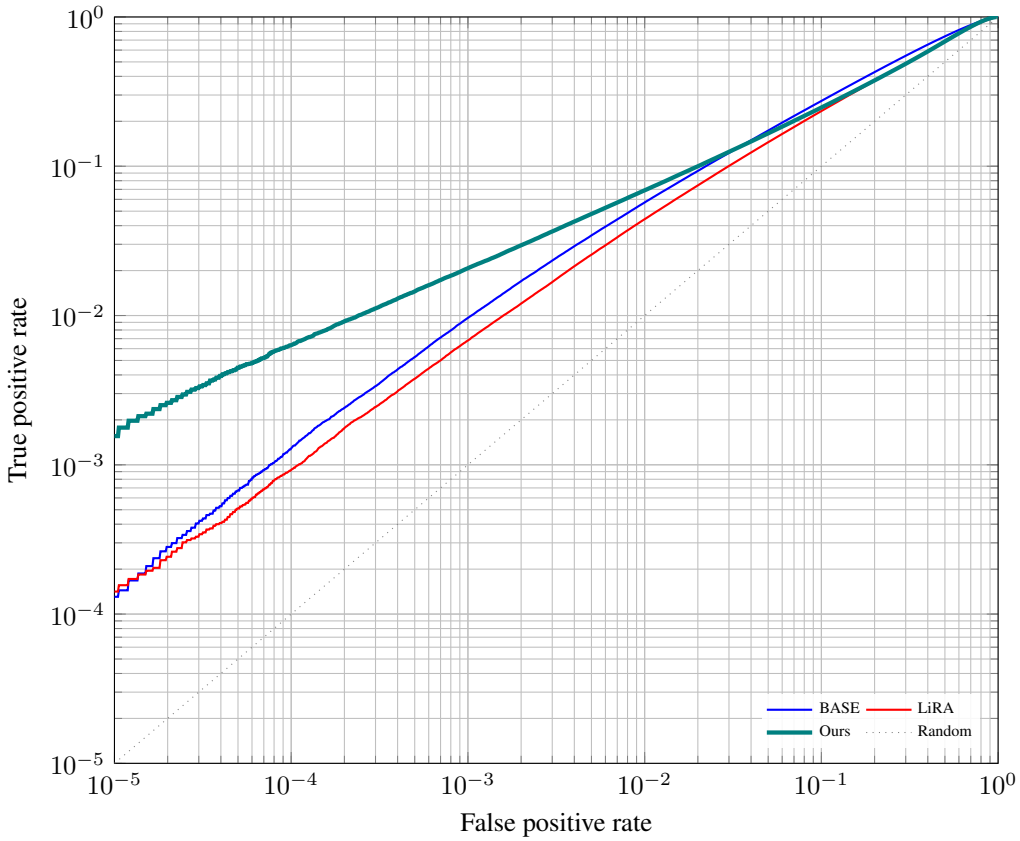


Figure 3: Target-averaged log–log ROC curves for online record-level attacks on flow-matching models trained on ImageNet. BASE, LiRA, and Ours are evaluated on the full audit set and averaged over ten target models.

Table 6: Offline group-level audits with matched reference models. Values are percentages: mean ± population standard deviation over ten targets. pAUC is standardized at 1% FPR, and TPR is reported down to 0.1% FPR. CelebA uses 5–15 probes per entity, while VGGFace2 uses 5–200. Bold means are highest for each benchmark and metric.

DDPM / CelebA: Exclude train			
Attack	pAUC@1%	TPR@1%	TPR@0.1%
BASE	59.28 ± 0.93	27.36 ± 1.68	9.89 ± 2.44
LiRA	60.56 ± 0.45	28.12 ± 0.96	11.27 ± 1.67
Ours	$\mathbf{61.21} \pm 0.82$	$\mathbf{30.08} \pm 2.02$	$\mathbf{14.00} \pm 2.01$
FM / CelebA: Exclude train			
Attack	pAUC@1%	TPR@1%	TPR@0.1%
BASE	56.30 ± 1.26	18.50 ± 2.91	7.49 ± 2.72
LiRA	55.80 ± 0.90	15.97 ± 2.19	7.94 ± 1.68
Ours	$\mathbf{57.34} \pm 1.25$	$\mathbf{20.05} \pm 2.93$	$\mathbf{9.32} \pm 2.29$
FM / VGGFace2: Holdout			
Attack	pAUC@1%	TPR@1%	TPR@0.1%
BASE	$\mathbf{58.27} \pm 0.57$	22.55 ± 1.36	$\mathbf{10.95} \pm 1.32$
LiRA	53.49 ± 0.39	10.41 ± 0.98	4.38 ± 0.64
Ours	58.11 ± 0.52	$\mathbf{22.78} \pm 1.01$	8.60 ± 2.19

F.4 GROUP-LEVEL VGGFACE2 RESULTS

Figure 4 reports group-level offline attacks against flow-matching models trained on VGGFace2. Each target audit contains 4,565 member and 4,565 non-member entities. We therefore restrict the ROC plot to FPR values of at least $10/4565 \approx 2.19 \times 10^{-3}$, for which each target contains at least ten false positives.

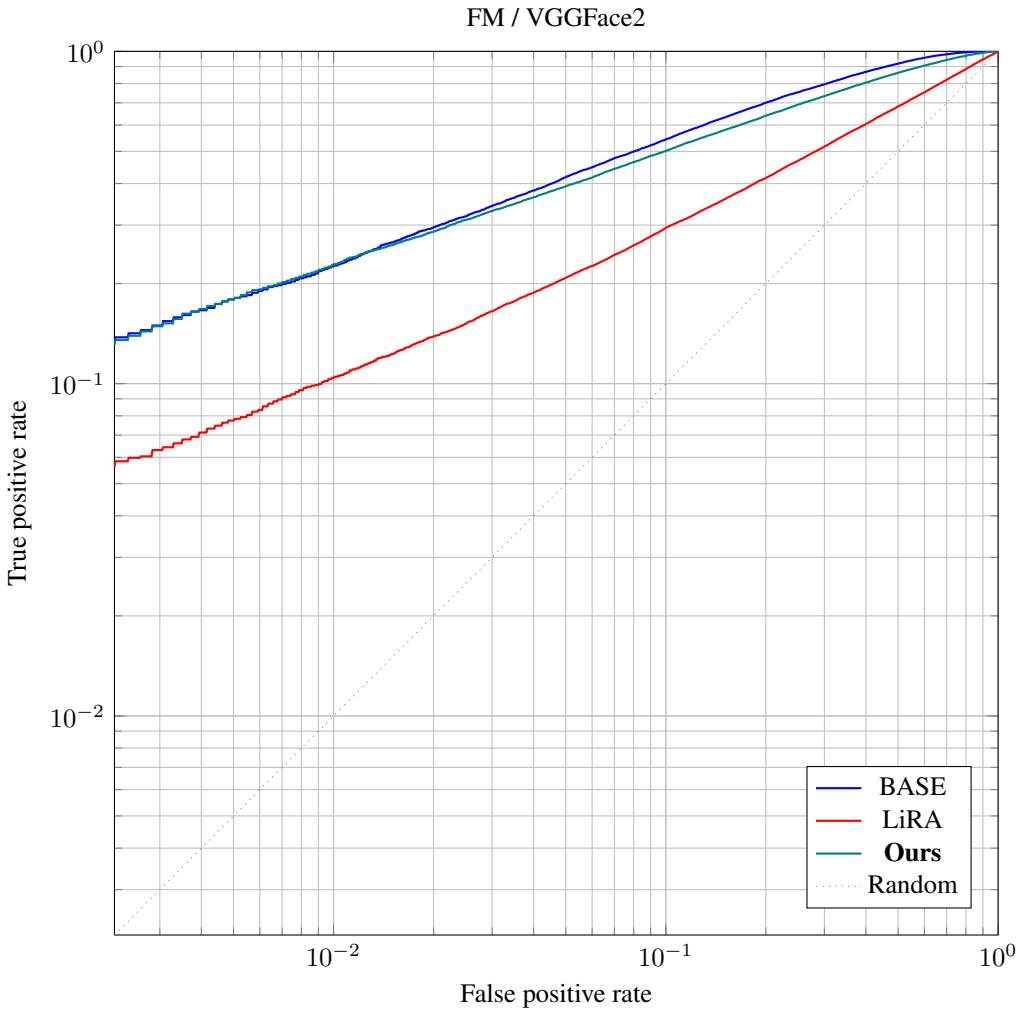


Figure 4: Target-averaged log–log ROC curves for offline group-level attacks on flow-matching models trained on VGGFace2. Curves are averaged over ten target models. The lower FPR limit corresponds to ten false positives among the 4,565 non-member entities in each target audit.