Membership is Ownership: A Robust Ownership Verification Framework for Diffusion Models

Feng Jiang*, Zuobin Xiong†, An Huang†, Zhipeng Cai*, and Yingshu Li*
*Department of Computer Science, Georgia State University, Atlanta, USA
†Department of Computer Science, University of Nevada Las Vegas, Las Vegas, USA
* *fjiang4@student.gsu.edu, {yili, zcai}@gsu.edu*; † *{zuobin.xiong, an.huang}@unlv.edu*

***Abstract*—Large-scale diffusion models have fueled numerous profitable downstream applications for AI-related businesses, including visual editing and content creation. Meanwhile, due to the huge amount of resource consumption (e.g., computation and high-quality data) during training, such diffusion models are deemed valuable intellectual property (IP) for tech companies like OpenAI and Google. Yet, the IP assets are vulnerable to various unauthorized uses by adversaries seeking to steal models for customized, usually commercial applications. Some existing approaches explored IP protection for AI models; however, they mostly face structural limitations in common — using a training-time watermarking by injecting artifacts in the model, which can impose a measurable utility cost and can be weakened by post-hoc fine-tuning. To address these challenges, this work investigates the IP protection (i.e., model ownership verification) for diffusion models in a realistic commercial scenario with minimal model utility loss. Specifically, the proposed method builds a framework for model ownership verification, termed as "Membership is Ownership (MiO)", based on a population-level hypothesis test on a private member evidence dataset. The verification is upheld by two central criteria: model attribution through membership inference attacks and the model separation against public references, which are rigorously verified by hypothesis tests with $p < 10^{-6}$. Via extensive experiments, the proposed MiO framework outperforms existing baselines in verification success, false positive rate, and mode utility loss. Furthermore, MiO stays stable under different post-theft fine-tuning and weight perturbation in adversarial scenarios, reflecting better robustness compared to the watermarking methods.**

***Index Terms*—Ownership Verification, Diffusion Models, Intellectual Property Protection, Model Watermarking**

I. Introduction

Large-scale generative diffusion models have become the critical infrastructure for image generation [1], producing numerous profitable downstream applications. However, the impressive capability of these models is bought at considerable costs: the intensive computational resources and high-quality training data, which make the resulting generative AI models valuable intellectual property (IP) for the institutions and companies that own them. Yet these IP assets are a vulnerable mining target to malicious theft, an adversary who aims to steal the model, which can then fine-tune and redeploy the model for other purposes without contribution, sabotaging the IP copyright and ownership. Establishing *model ownership verification* framework, therefore, is a pressing issue of practical economic and legal significance.

Ownership verification is to assert whether a suspect model/checkpoint originates from a particular owner's (i.e., victim) base model. Existing approaches to diffusion model IP protection address this problem along two complementary lines. (1) *Training-time watermarking* embeds an explicit signature into model parameters or output behavior [2]–[4], e.g., adding invisible noise into parameter space or output space for later identification, which are effective in some controlled settings, but imposes a measurable utility cost and can be progressively weakened by post-hoc fine-tuning, a common strategy used by adversaries. (2) *Data memorization verification* for diffusion models [5]–[7] exploit the model's intrinsic memorization gap – reconstruction error on training samples is systematically lower than on unseen samples – and can verify used training data. However, the memorization verification is fundamentally a per-sample classifier where a low reconstruction error on a **single sample** cannot certify **a model ownership** claim, because some individual samples may be easy to reconstruct.

The challenges above suggest another approach: conducting model ownership verification at the population level while minimizing impact on model utility. In this paper, we present **Membership is Ownership** (MiO), a verification framework built on a principled observation: although the per-sample memorization signal is weak to claim ownership, the *population-level* memorization pattern over a privately-curated evidence set is both robust and statistically auditable. Particularly, the model trainer/owner reserves a private subset $\mathcal{W}$ of the training data, which is never disclosed to the public, and serves as a non-invasive watermark that leaves model parameters and output quality entirely intact. To extract a robust signal from a suspect model for ownership verification, we introduce a multi-timestep t-error score via forward noising and single-step reconstruction by using different statistics (e.g., the 25th percentile) to suppress noise while retaining the memorization. To convert this per-sample information into a calibrated decision, we use Gaussian quantile-regression networks with closed-form false positive rate (FPR) thresholds at any target FPR budget.

We validate the framework across two evaluation scenarios in naive diffusion models, like DDIMs trained from scratch and commercialized diffusion models, like Stable Diffusion fine-tuned from the production-scale datasets. Across both scenarios, MiO can provide an accurate ownership claim at a controlled false-accusation rate.

In summary, the contributions of this work are as follows.

- We reframe the model ownership verification into a population-level hypothesis test problem, which leads to

a statistically auditable framework.
- The ownership verification process introduces minimal impact (i.e., 0 change in FID score) on the generation quality of original models by using a non-invasive evidence dataset.
- The proposed two-point verification protocol applies Gaussian quantile regression for attribution comparison and Cohen distance for reference separation, providing a closed-form and controllable FPR for different applications.
- We validate the framework on two scenarios across multiple datasets with different configurations, showing that MiO outperforms existing methods in verification accuracy, costs, utility loss, and post-theft robustness.

II. Related Works

Model Watermarking. Watermarking embeds an owner-specific signal that can be recovered later to prove provenance. The methods differ mainly in where the signal lives, which decides how hard it is to remove. One line places it in the generation process. Gaussian Shading [8] hides the message in the initial latent while keeping its distribution Gaussian, so training and image quality are less affected. Newer variants, Gaussian Shading++ [9], turn repeated latent embeddings into a per-bit likelihood ratio for soft decoding, the pseudorandom-code watermark of Gunn et al. [10] makes the mark cryptographically undetectable in the sampled noise, ROBIN [11] moves the mark to an intermediate denoising state, and GaussMarker [12] adds a spatial-domain channel. Because the signal rides on the seed and the sampling path, a user who controls generation can drop it by resampling the initial latent.

A second line binds the signal to the model. The watermark diffusion process [2] and the trigger-based recipe of Zhao et al. [3] install it at training time; Stable Signature [13] fine-tunes the latent decoder so every output carries a fixed bitstring; ProMark [14] marks the training images and trains the model to reproduce them. SleeperMark [4] carries this to diffusion models by separating watermark knowledge from semantic knowledge during training, so the trigger is not forgotten when the stolen model is fine-tuned for a new task. Those model-related marks resist resampling, but they cost measurable utility and can be mitigated by removal attacks that weaken such watermarks while preserving image quality [15]–[20].

Data Memorization Verification. The memorization of training data in model parameters can be used to check model or data attribution [21], [22]. Specifically, membership inference attack (MIA), an attack to decide whether an input is part of a model's training data [23]–[25], has been used in literature to verify models. For diffusion models, recent works investigate MIA using reconstruction-based signals [5], [26], [27], loss/likelihood-based criteria [6], [7], or scalable quantile-regression formulations [28], [29]. These methods are primarily evaluated as per-example membership decision (i.e., a binary classification of member or non-member) and are typically reported via attack metrics. Ownership verification, however, requires comparative, population-level evidence: the ownership claim must be supported by the aggregate evidence on a designated evidence set via comparison with calibrated reference models to show separation, and tested against the suspect model via a statistically interpretable hypothesis rather than single-sample predictions. One recent work, CDI [22], adopted a similar idea to ours by aggregating several per-sample MIA, but they focused on data ownership, not the model ownership. MiO differs by curating a private evidence set as the ownership anchor and calibrating a closed-form FPR before the population test, so the verification carries an explicit false-accusation guarantee.

III. Problem Formulation

A. Problem Statement

1) Diffusion Models.: A diffusion model corrupts a clean sample x via a forward noising process,

$$x_t = \sqrt{\bar{\alpha}_t}\, x + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, x_t is the noised image at timestep $t \in \{1, \ldots, T\}$ and trains a network $\epsilon_\theta(x_t, t)$ to predict the added noise. The network ϵ_θ is optimized by minimizing the denoising objective with t drawn uniformly from $\{1, \ldots, T\}$.

$$\mathcal{L}(\theta) = \mathbb{E}_{x,\, t,\, \epsilon} \left\| \epsilon - \epsilon_\theta(x_t, t) \right\|_2^2, \tag{2}$$

New samples are produced by iteratively denoising from $x_T \sim \mathcal{N}(0, I)$; we adopt the deterministic DDIM sampler, which drops the stochastic term and admits a closed-form estimate $\hat{x}_0$ of the clean signal from any (x_t, t) through ϵ_θ. This $\hat{x}_0$ estimate is precisely the quantity on which our reconstruction score e_t is built (Section IV-A).

Latent diffusion models (LDMs), e.g. Stable Diffusion [1], run the diffusion process in a compressed latent space, where a pretrained autoencoder maps an image x to $z = \mathcal{E}(x)$ and decodes it as $\mathcal{D}(z)$. The forward process and training objective (Eqs. (1)–(2)) then act on z in place of x, and the denoiser $\epsilon_\theta(z_t, t, c)$ is conditioned on a text embedding c injected through the U-Net cross-attention layers, which are the same layers LoRA adapters modify. For LDMs, the reconstruction error e_t can be evaluated in either latent space or pixel space, and the ablation study is evaluated in Section V-F.

Both diffusion models systematically memorize portions of their training data [26], [30] – a memorization gap our framework exploits in Section IV.

2) Membership Inference vs. Ownership Verification: MIA leverages the memorization gap to determine whether a single query sample belongs to the training set of a target model [5], [27], while the ownership verification is different. We state this problem formally:

Definition 1 (Membership Inference Attack)**.** *Given a target model M trained from $\mathcal{D}_{\text{train}}$ and a query sample x, MIA is a binary classifier $\mathcal{A} : (M, x) \to \{0, 1\}$ that predicts whether $x \in \mathcal{D}_{\text{train}}$.*

However, ownership verification requires a population-level test: significant memorization on a *designated evidence set* relative to independent reference models.

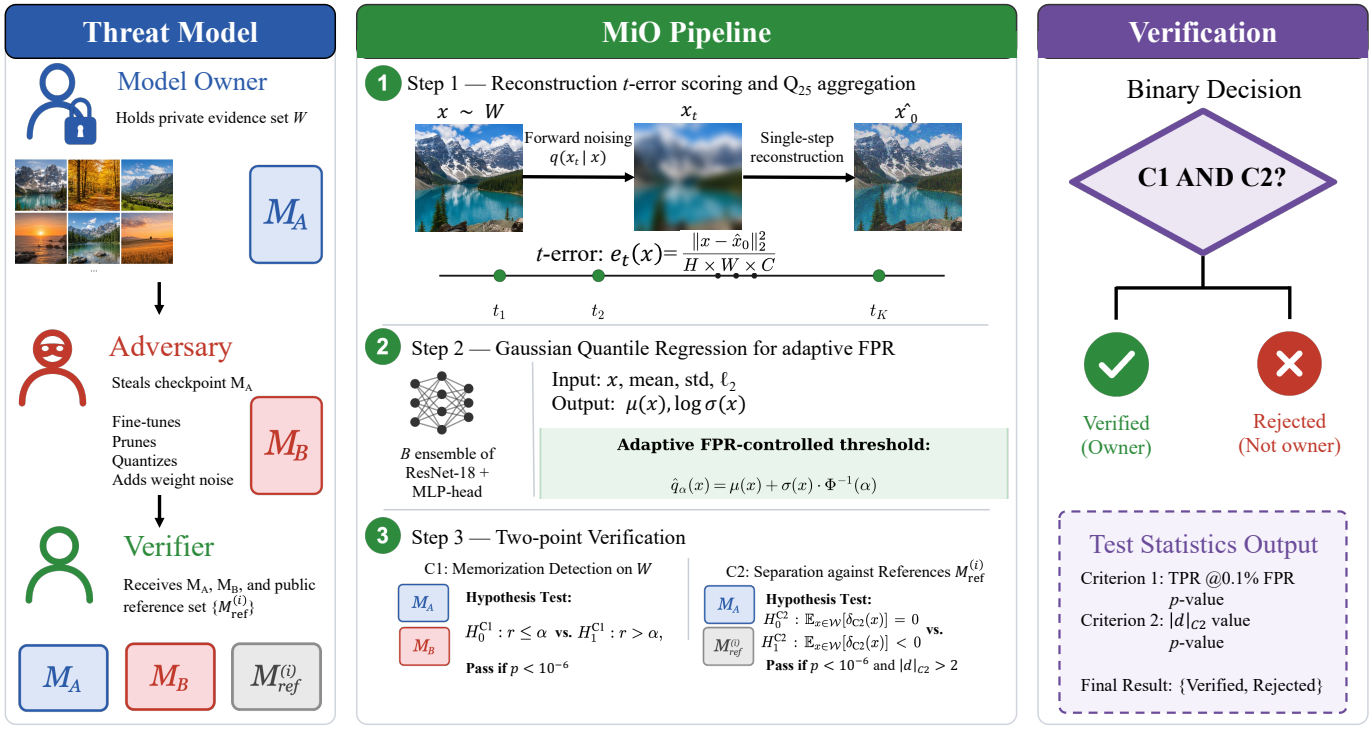

Fig. 1. The framework and operation steps in the proposed MiO method.

Definition 2 (Model Ownership Verification)**.** *Let $\mathcal{M}_A$ be the owner's base model, $\mathcal{M}_B$ be a suspect model to be verified, and $\mathcal{W} \subset \mathcal{D}_{\text{train}}(\mathcal{M}_A)$ be a private evidence set. An ownership verification protocol is a hypothesis test that decides whether $\mathcal{M}_B$ is stolen from $\mathcal{M}_A$ under a controlled false-positive rate α.* $\mathcal{V} : (\mathcal{M}_B, \mathcal{W}, \{M_{ref}^{(i)}\}) \to \{\text{VERIFIED}, \text{REJECTED}\}$

Reference Models. The reference model set $\{\mathcal{M}_{\text{ref}}^{(i)}\}_{i=1}^{|\mathcal{R}|}$ in the definition above should satisfy three conditions: (i) its training data is disjoint from the private set $\mathcal{W}$; (ii) it has independent provenance from $\mathcal{M}_A$; and (iii) it is type-compatible with the t-error computation procedure of Section IV-A – i.e., a diffusion model on which a per-sample reconstruction score on $\mathcal{W}$ is well-defined.

B. Threat Model

We consider three parties in the verification framework: the model **owner** who trained $\mathcal{M}_A$ and holds $\mathcal{W}$, the **adversary** A who obtained $\mathcal{M}_A$ through theft (e.g., reverse engineering attacks or insider attacks) and produced a derivative model $\mathcal{M}_B$ via fine-tuning, and the **verifier** V who judges whether $\mathcal{M}_B$ is derived from $\mathcal{M}_A$.

We assume the adversary A has white-box access to $\mathcal{M}_A$ (due to theft) and may possess similar domain data disjoint from $\mathcal{W}$. The adversary can apply adaptations, such as fine-tuning, weight perturbation, quantization, or pruning, to the stolen model weights to escape from the verification protocol. Knowledge distillation and full retraining from scratch lie out of the scope of this paper, as they will completely remove the connection between $\mathcal{M}_A$ and $\mathcal{M}_B$. Nonetheless, we still examine distillation as a threat-model boundary case in Section V-D (Table IV). The verifier V (i.e., copyright judge or court) has white-box access to both $\mathcal{M}_A$, $\mathcal{M}_B$, and a set of public reference models $\{\mathcal{M}_{\text{ref}}^{(i)}\}_{i=1}^{|\mathcal{R}|}$, representing the null hypothesis.

The verification framework is established on the core fact that the adversary cannot access the owner's private evidence set $\mathcal{W}$, which should be the confidential information of a company. In the worst case, even if $\mathcal{D}_{\text{train}}$ were leaked, selecting the exact subset $\mathcal{W}$ from training samples $\mathcal{D}_{\text{train}}$ is computationally impossible with probability at most $P(\text{exact match}) = \frac{1}{\binom{N}{|\mathcal{W}|}} \leq \left(\frac{|\mathcal{W}|}{N}\right)^{|\mathcal{W}|}$, which is a 0.1^{5000} under the the small CIFAR-10 setting ($N = 50{,}000$, $|\mathcal{W}| = 5{,}000$).

IV. MiO: Ownership Verification Framework

The MiO is a three-step verification framework that combines the following parts: (1) a reconstruction error $e_t(x)$ to quantify model memorization on each timestep t; (2) a Gaussian quantile regression-based function to derive a closed-form member distribution at any chosen false positive level; and (3) a two-point verification criterion for ownership verification. The framework of MiO is described in Fig. 1.

A. Reconstruction t-error Scoring

Diffusion models reconstruct their training data with lower error than unseen data [26] and thus, the *memorization gap* can be used to quantify membership. Specifically, given a clean sample x, we first corrupt it via the forward diffusion schedule as shown in Eq. (1) to obtain x_t. The model then predicts the

Algorithm 1 t-error Computation

Require: Diffusion model ϵ_θ, sample $x \in \mathbb{R}^{H\times W\times C}$, timesteps $\mathcal{T} = \{t_1, \ldots, t_K\}$, noise schedule $\{\bar{\alpha}_t\}$
Ensure: Aggregated score $s(x)$
  errors $\leftarrow []$
  for $t \in \mathcal{T}$ **do**
    $\epsilon \sim \mathcal{N}(0, I)$
    $x_t \leftarrow \sqrt{\bar{\alpha}_t} \cdot x + \sqrt{1-\bar{\alpha}_t} \cdot \epsilon$
    $\hat{\epsilon} \leftarrow \epsilon_\theta(x_t, t)$
    $\hat{x}_0 \leftarrow (x_t - \sqrt{1-\bar{\alpha}_t} \cdot \hat{\epsilon})/\sqrt{\bar{\alpha}_t}$
    $e_t \leftarrow \|x - \hat{x}_0\|_2^2/(H \times W \times C)$
    errors.append(e_t)
  end for
  return $s(x) \leftarrow Q_{25}$(errors)

added noise $\hat{\epsilon}_\theta(x_t, t)$, from which we can recover x_0 from x_t via $\hat{x}_0 = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\cdot\hat{\epsilon}_\theta(x_t,t)}{\sqrt{\bar{\alpha}_t}}$.

We define the reconstruction t-*error*, $e_t(x)$, as the pixel-level mean squared error at timestep t between the original and reconstructed data as follows:

$$e_t(x) = \frac{\|x - \hat{x}_0\|_2^2}{H \times W \times C} \tag{3}$$

where H, W, and C denote the height, width, and number of channels, respectively. For latent diffusion models such as Stable Diffusion, the reconstruction error calculation is applied in the VAE latent space without modification. The empirical pixel-vs-latent comparison is in Section V-F. The full procedure is summarized in Algorithm 1, distinct from the single training timestep t in Eq. (2), the timestep set $\mathcal{T} = \{t_1, \ldots, t_K\}$ collects K timesteps drawn independently and uniformly from $\{1, \ldots, T\}$.

B. Gaussian Quantile Regression for Adaptive Threshold

A single timestep t-error provides a noisy estimate of reconstruction quality, since the magnitude of $e_t(x)$ varies substantially with t. To obtain a more stable statistic $s(x)$, we aggregate t-errors over K uniformly sampled timesteps via different percentiles Q_d. That is

$$s(x) = Q_d\big(\{e_{t_i}(x)\}_{i=1}^K\big). \tag{4}$$

We use $K = 50$ in the experiments, and an ablation study in Section V-F shows that $K \geq 25$ will produce a stable regime. For the percentile Q_d, a low percentile is preferred because the memorization gap between members and non-members is more pronounced at the timesteps where reconstruction is more accurate on members. The 25th percentile (Q_{25}) is selected based on our empirical evaluation in Section V-F, which retains the best performance while suppressing the upper tail.

After calculating the statistic $s(x)$, what remains is to derive a calibrated, FPR-controlled threshold for $s(x)$ to detect the ownership signal for various inputs, because the threshold in real applications should not be a fixed value but data-dependent. We therefore model the conditional distribution of $s(x)$ for adaptive threshold construction. Concretely, we apply a log transform to the score and parameterize the result as a conditional Gaussian distribution whose parameters are learned from x:

$$y = \log(1 + s(x)) \sim \mathcal{N}\big(\mu(x),\, \sigma^2(x)\big). \tag{5}$$

The conditional Gaussian parameters are produced by a learned predictor f_ψ with parameters ψ that takes sample x and a 3-dimensional summary $(\text{mean}, \text{std}, \ell_2)$ of the t-error sequence $\{e_{t_i}(x)\}_{i=1}^K$, and outputs the mean and the *log* standard deviation, $f_\psi : (x, summary) \mapsto (\mu(x), \log\sigma(x))$. Predicting $\log\sigma$ rather than σ leaves the head output unconstrained while guaranteeing a positive scale and avoiding the $\sigma \to 0$ blow-up of the likelihood. We then recover $\sigma(x) = \exp(\log\sigma(x))$. In the experiment design, the Gaussian learner f_ψ is a ResNet-18 backbone, and the feature is concatenated and passed through an MLP head. f_ψ is trained on an auxiliary set $\mathcal{D}_{\text{aux}}$, drawn from the same domain as $\mathcal{M}_A$'s training data but *disjoint* from it, by minimizing the Gaussian negative log-likelihood shown in Eq. (6). This disjointness matters because $\mathcal{D}_{\text{aux}}$ defines the *non-member* score distribution that calibrates the threshold. In practice, we use a held-out split of the same dataset or public data of the same domain and exclude both $\mathcal{W}$ and the rest of $\mathcal{D}_{\text{train}}(\mathcal{M}_A)$.

$$\mathcal{L}(\psi) = \mathbb{E}_{x\sim\mathcal{D}_{\text{aux}}}\left[\frac{1}{2}\log\sigma^2(x) + \frac{\big(y - \mu(x)\big)^2}{2\sigma^2(x)}\right] \tag{6}$$

Members produce systematically lower scores $s(x)$ than non-members, so the membership signal lies in the *lower tail* of the non-member score distribution. To declare x a member at false-positive rate α, we therefore set an adaptive threshold $\hat{q}_\alpha(x)$ equal to the α-quantile of the non-member distribution of $\log(1+s(x))$ conditional on x. Under the conditional Gaussian model, this quantile admits the closed form

$$\hat{q}_\alpha(x) = \mu(x) + \sigma(x)\,\Phi^{-1}(\alpha), \tag{7}$$

where Φ^{-1} is the standard-Gaussian quantile function. The threshold gives a per-sample, FPR-calibrated instruction for membership claim: we flag x as a likely member if $\log(1 + s(x)) \leq \hat{q}_\alpha(x)$, or equivalently by the threshold:

$$s(x) \leq \exp\big(\hat{q}_\alpha(x)\big) - 1. \tag{8}$$

Therefore, based on the threshold, only an α fraction of non-members fall below $\hat{q}_\alpha(x)$, delivering an expected false-positive rate at level α.

To reduce the predictor variance, we replace the single f_ψ with a bagging ensemble of $B = 50$ predictors trained on 80% bootstrap resamples of $\mathcal{D}_{\text{aux}}$, and use the averaged $\hat{q}_\alpha(x)$ at test time.

C. Two-Point Verification Criteria

Built on the quantile score $s(x)$ from Section IV-B, this section defines the model ownership verification criterion. Given the owner model $\mathcal{M}_A$, the suspect model $\mathcal{M}_B$, a set of public references $\{\mathcal{M}_{\text{ref}}^{(i)}\}$, and the private evidence set

Algorithm 2 Gaussian Quantile Regression Training

Require: Auxiliary non-member data $\mathcal{D}_{\text{aux}}$, ensemble size B, bootstrap ratio r, max epochs E
Ensure: Trained ensemble $\{f_{\psi^{(b)}}\}_{b=1}^{B}$
  for $b = 1$ to B **do**
    $\mathcal{D}_b \leftarrow \text{BootstrapSample}(\mathcal{D}_{\text{aux}}, r)$ $\{r = 0.8$, with replacement$\}$
    $\mathcal{D}_b^{\text{fit}}, \mathcal{D}_b^{\text{val}} \leftarrow \text{Split}(\mathcal{D}_b, 0.9)$
    Initialize $f_{\psi^{(b)}}$ with ResNet-18 backbone
    for epoch $= 1$ to E **do**
      for batch $(x, s(x), \phi(x))$ in $\mathcal{D}_b^{\text{fit}}$ **do**
        $y \leftarrow \log(1 + s(x))$
        $(\mu, \log\sigma) \leftarrow f_{\psi^{(b)}}(x, \phi(x))$
        $\sigma \leftarrow \exp(\log\sigma)$ {positive scale}
        $\mathcal{L} \leftarrow \frac{1}{2}\log\sigma^2 + (y-\mu)^2/(2\sigma^2)$
        Update $\psi^{(b)}$ by gradient descent on $\mathcal{L}$
      end for
      if validation loss on $\mathcal{D}_b^{\text{val}}$ has not improved for 10 epochs **then**
        break {early stopping}
      end if
    end for
  end for
  return $\{f_{\psi^{(b)}}\}_{b=1}^{B}$

$\mathcal{W}$, the following criteria must be satisfied to declare verified ownership.

Criterion 1: Memorization Detection on $\mathcal{W}$. Applying the per-sample, FPR-α threshold $\hat{q}_\alpha(x)$ from Section IV-B to the suspect $\mathcal{M}_B$, we flag each $x \in \mathcal{W}$ and compute the empirical hit rate $\hat{r}$ on the evidence set population:

$$\hat{r} = \frac{1}{|\mathcal{W}|} \sum_{x \in \mathcal{W}} \mathbf{1}\big[s_{\mathcal{M}_B}(x) \le \exp(\hat{q}_\alpha(x)) - 1\big]. \tag{9}$$

Let r denote the population hit rate $\Pr_{x\in\mathcal{W}}\big[s_{\mathcal{M}_B}(x) \le \exp(\hat{q}_\alpha(x)) - 1\big]$ that $\hat{r}$ estimates, and test the one-sided pair

$$H_0^{\text{C1}} : r \le \alpha \qquad \text{vs.} \qquad H_1^{\text{C1}} : r > \alpha,$$

where the null hypothesis H_0^{C1} states that $\mathcal{M}_B$ has no memorization of $\mathcal{W}$ beyond the calibrated false-positive rate, whereas H_1^{C1} states that $\mathcal{M}_B$ retains memorization of $\mathcal{W}$ beyond it. At the null boundary $r = \alpha$, the count $|\mathcal{W}| \cdot \hat{r}$ follows $\text{Binomial}(|\mathcal{W}|, \alpha)$, so a one-sided exact binomial test can be applied. The criterion 1 is declared passed when its p-value falls below 10^{-6}.

Criterion 2: Separation against Reference Models. The owner model $\mathcal{M}_A$ must reconstruct evidence-set samples significantly better than any public reference model, verifying the unique memorization of $\mathcal{M}_A$ on the private dataset $\mathcal{W}$. For each reference model $\mathcal{M}_{\text{ref}}^{(i)}$, we first compute the per-sample difference $\delta_{\text{C2}^{(i)}}(x) = s_{\mathcal{M}_A}(x) - s_{\mathcal{M}_{\text{ref}}^{(i)}}(x)$, and aggregate it across the evidence set into Cohen's distance d as follows:

$$|d|_{\text{C2}} = \frac{|\,\text{mean}(\delta_{\text{C2}}(\mathcal{W}))\,|}{\text{std}(\delta_{\text{C2}}(\mathcal{W}))}. \tag{10}$$

Then we test the null hypothesis $H_0^{\text{C2}} : \mathbb{E}_{x\in\mathcal{W}}[\delta_{\text{C2}}(x)] = 0$ against the one-sided alternative $H_1^{\text{C2}} : \mathbb{E}_{x\in\mathcal{W}}[\delta_{\text{C2}}(x)] < 0$. H_0^{C2} says the owner reconstructs $\mathcal{W}$ no better than a public model that never saw it. H_1^{C2} says the owner's t-error on $\mathcal{W}$ is strictly lower than the reference's, a gap that only training on $\mathcal{W}$ can produce. Criterion 2 passes when a one-sided t-test rejects H_0^{C2} at $p < 10^{-6}$ and $|d|_{\text{C2}} > 2$.

The verifier declares VERIFIED only if *both* the memorization criterion (C1) and the separation criterion (C2) pass, and REJECTED otherwise. The two criteria target different null hypotheses and different risks. C1 asks whether the suspect $\mathcal{M}_B$ has actually memorized $\mathcal{W}$; C2 asks whether the owner $\mathcal{M}_A$ genuinely stands apart from public references that never saw $\mathcal{W}$. A VERIFIED verdict therefore requires both tests to reject their respective nulls, and each rejection is calibrated at the 10^{-6} level.

Because the probability that both tests reject at once cannot exceed the smaller of their two individual rejection probabilities, the type-I error of the joint decision is itself at most 10^{-6}, and this inequality holds without assuming that the two tests are independent. The 10^{-6} thresholds here are the criterion-level significance applied to the C1 and C2 p-values; they should not be confused with the per-sample operating point α that fixes $\hat{q}_\alpha(x)$ inside C1.

The conjunction is in fact conservative. In the false-accusation scenario the guarantee must protect against, where $\mathcal{M}_B$ is not derived from $\mathcal{M}_A$ and was never trained on $\mathcal{W}$, the null H_0^{C1} already holds, so the exact binomial test of C1 alone caps the false-positive rate at 10^{-6}; C2 then acts as an orthogonal safeguard rather than a condition the bound relies on. Algorithm 3 states the complete procedure.

V. EXPERIMENTS

A. Experimental Setup

We use two evaluation scenarios in our experiments: (1) DDIM trained from scratch provides a controlled measurement of the protocol's quantitative properties. (2) Stable Diffusion models with a fine-tuning dataset instantiate the production-scale threat surface.

Datasets. We evaluate the DDIM model on four datasets spanning different resolutions and scales: CIFAR-10 and CIFAR-100 (both 32×32, 50k training images with 5k reserved as the evidence set), STL-10 (96×96, 5k training images with 1k as the evidence set), and CelebA (64×64, 162k training images with 5k as the evidence set).

For the Stable Diffusion experiments, the owner's private training set consists of 1,000 images from COCO 2014 while the adversaries fine-tune on a disjoint COCO subset (also 1,000 images) plus a synthetic-image set.

Baselines and Reference Models. For the baseline comparison in Section V-E, we compare MiO against three diffusion-model IP protection methods: **WDM** [2], a training-time watermarking method that learns a secondary watermark diffusion process; **Zhao et al.** [3], which embeds StegaStamp fingerprints into training images before EDM training; and **CDI** [22], an inference-time MIA-aggregation method. For Stable Diffusion,

Algorithm 3 Ownership Verification Protocol

Require: Models $\mathcal{M}_A, \mathcal{M}_B$, references $\{\mathcal{M}_{\text{ref}}^{(i)}\}_{i=1}^{|\mathcal{R}|}$, evidence set $\mathcal{W}$, calibrated quantile $\hat{q}_\alpha(\cdot)$ from Section IV-B with $\alpha = 10^{-3}$

Ensure: VERIFIED or REJECTED

$S_A \leftarrow \{s_{\mathcal{M}_A}(x) : x \in \mathcal{W}\}$; $S_B \leftarrow \{s_{\mathcal{M}_B}(x) : x \in \mathcal{W}\}$

C1 (memorization detection, binomial):

$k \leftarrow \sum_{x \in \mathcal{W}} \mathbf{1}\big[s_{\mathcal{M}_B}(x) \le \exp(\hat{q}_\alpha(x)) - 1\big]$

$p_1 \leftarrow \Pr\big[X \ge k \,\big|\, X \sim \text{Binomial}(|\mathcal{W}|, \alpha)\big]$

if $p_1 > 10^{-6}$ **then**

return REJECTED

end if

C2 (separation, paired): for each reference $i \in \{1, \ldots, |\mathcal{R}|\}$:

$S_{\text{ref}}^{(i)} \leftarrow \{s_{\mathcal{M}_{\text{ref}}^{(i)}}(x) : x \in \mathcal{W}\}$

$\delta_{\text{C2}}^{(i)} \leftarrow S_A - S_{\text{ref}}^{(i)}$ {element-wise on $\mathcal{W}$}

$p_2^{(i)} \leftarrow \text{OneSidedTTest}_<\big(\delta_{\text{C2}}^{(i)}, \mu_0 = 0\big)$ {lower-tail: H_1^{C2} : $\mathbb{E}[\delta_{\text{C2}}] < 0$}

$|d|_{\text{C2}}^{(i)} \leftarrow \big|\,\text{mean}(\delta_{\text{C2}}^{(i)})\,\big| / \text{std}(\delta_{\text{C2}}^{(i)})$

if $\exists i : p_2^{(i)} > 10^{-6}$ **or** $|d|_{\text{C2}}^{(i)} < 2.0$ **then**

return REJECTED

end if

return VERIFIED

we additionally compare against **SleeperMark** [4], a trigger-prompt-coupled, LoRA-resistant watermark method.

As public references for the verification protocol, we use pretrained checkpoints from HuggingFace: `google/ddpm-cifar10-32` serves as the reference for both CIFAR-10 and CIFAR-100 datasets, `google/ddpm-ema-bedroom-256` (resized) is for STL-10, and `google/ddpm-celebahq-256` is for CelebA. These references are trained on disjoint data from our evidence sets and represent the null hypothesis in the two-point verification protocol.

B. RQ1: How Reliable is MIA in Criteria 1 at Restricted FPR?

The calibrated quantile $\hat{q}_\alpha(x)$ turns the per-sample t-error into an FPR-α decision: we flag x as a member when $\log(1 + s_{\mathcal{M}_A}(x)) \le \hat{q}_\alpha(x)$. With the Gaussian quantile regression ensemble in Section IV-B, the threshold is calculated once per dataset on an auxiliary non-member partition. We then evaluate this indicator on held-out members and non-members for each owner, reporting ROC-AUC and TPR at the restricted FPR budgets $\alpha \in \{10^{-3}, 10^{-4}\}$. Because an ownership verification in practice must be held at a controlled false-accusation rate, we treat TPR at a fixed low FPR as the primary reliability metric.

Table I reports per-sample MIA performance for the four DDIM owner models and the three different Stable Diffusion v1.4 owner configurations. First, the AUC is uniformly high where every owner model attains at least 0.965 AUC. Second, and more informative, the signal stays non-trivial once the FPR budget is tightened. Specifically, for both DDIM and SD, the MIA achieves a TPR of at least 0.845 at 10^{-3} FPR and still reaches 0.793 at the stricter 10^{-4} FPR. Consequently, even at these restricted-FPR operating points, the detection rate remains far above the corresponding chance level, indicating that the t-error signal underlying the MIA is sufficiently reliable in the low-FPR regime.

TABLE I
PER-SAMPLE MIA PERFORMANCE UNDER GAUSSIAN QR CALIBRATION.

Owner	AUC	TPR@10^{-3}	TPR@10^{-4}
DDIM (from scratch)			
CIFAR-10	0.997	0.859	0.819
CIFAR-100	0.965	0.845	0.798
STL-10	0.989	0.935	0.899
CelebA	0.998	0.985	0.953
Stable Diffusion v1.4			
LoRA r=64	0.991	0.857	0.793
LoRA r=256	0.996	0.911	0.838
Full FT	0.995	0.879	0.799

These results answer RQ1 in the affirmative for the case we test: the t-error MIA retains usable true-positive rates at false-positive budgets as low as 10^{-4}. This per-sample reliability is the foundation that Criterion 1 rests on – the population-level hit rate $\hat{r}$ that C1 tests against the null H_0^{C1}: $r \le \alpha$ is meaningful even if the underlying per-sample decision genuinely controls FPR at small α levels.

C. RQ2: Is Memorization Unique to the Owner?

In Criteria 2 of the two-point verification, we aim to separate the owner's model from public reference models by checking if the memorization on the private evidence set is only unique to the owner's model. Table II reports $|d|_{\text{C2}}$ (Eq. (10)) for seven owner models: four DDIM models on CIFAR-10, CIFAR-100, STL-10, CelebA trained from scratch and three Stable Diffusion v1.4 models fine-tuned on a 1,000-image COCO subset with LoRA at rank 64, LoRA at rank 256, full UNet fine-tuning. Each model is evaluated against three reference models from different settings: semantic-similar, cross-domain, and random initialization.

Semantic-similar means the reference model is trained/fine-tuned on data of the same type and distribution as the owner, but has never seen the owner's private evidence set $\mathcal{W}$. Take DDIM as an example: for the CIFAR-10, CIFAR-100, and STL-10 owners, the semantic-similar reference models are uniformly `google/ddpm-cifar10-32` (trained on CIFAR-10, likewise 32×32 natural images); for the CelebA owner, it is `google/ddpm-celebahq-256` (trained from the face domain). Cross-domain means the reference model is trained/fine-tuned on data from an entirely different category. Trained on indoor bedroom scenes, LSUN's `google/ddpm-bedroom-256` is the cross-domain reference for CIFAR-10, CIFAR-100, and CelebA. For STL-10, we adopt `google/ddpm-church-256` to avoid a domain-match confound with some of STL-10's classes since its training dataset is church images. Random initialization reference

TABLE II
PAIRED $|d|_{C2}$ ACROSS DDIM AND SD OWNERS AGAINST THREE REFERENCE ROLES.

Owner	Semantic-similar	Cross-domain	Random init
DDIM (from scratch)			
CIFAR-10	✓ (23.9)	✓ (31.6)	✓ (45.1)
CIFAR-100	✓ (16.3)	✓ (25.2)	✓ (44.1)
STL-10	✓ (87.3)	✓ (22.2)	✓ (41.5)
CelebA	✓ (20.5)	✓ (23.5)	✓ (63.0)
Stable Diffusion v1.4			
LoRA r=64	✓ (2.54)	✓ (2.49)	✓ (19.0)
LoRA r=256	✓ (2.61)	✓ (2.65)	✓ (19.1)
Full FT	✓ (2.36)	✓ (2.44)	✓ (19.1)

model is a network at the dataset's native resolution with fully random weights and no training at all.

Stable Diffusion is analogous, except that the models are not trained from scratch but fine-tuned from SD v1.4. Its semantic-similar reference is therefore the un-fine-tuned base `CompVis/stable-diffusion-v1-4`: the owner model is fine-tuned from it on a 1,000-image COCO subset, so this reference shares the same SD v1.4 backbone as the owner yet was never fine-tuned on the owner's evidence set $\mathcal{W}$. The cross-domain reference is `hakurei/waifu-diffusion-v1-3` (anime style), whose fine-tuning data is entirely disjoint from COCO. The random-init reference is a randomly initialized UNet that reuses only the SD v1.4 VAE.

In the DDIM part of Table II, all separations are significant. Across all 12 (owner, reference) cells, $|d|_{C2}$ ranges from a minimum of 16.3 to a maximum of 87.3, far beyond the threshold 2.0 in Criterion 2. The reason is that the owner model $\mathcal{M}_A$ is trained entirely from scratch by us, so its memorization of $\mathcal{W}$ is way stronger than that of any reference model naturally. For the latent-space SD owners, any reference that shares the SD v1.4 backbone reconstructs the same image to a comparable error, so the gap is compressed and $|d|_{C2}$ falls in a smaller range (e.g, between 2 and 3). Only the randomly initialized UNet, which does not share the backbone, drives the separation up to about 19. Yet, across all 9 cells, $|d|_{C2}$ remains greater than 2, so the separation still holds under the threshold. It is worth noting that no matter for DDIM or SD, the random-init reference gives the largest separation, further showing that the owner's memorization of its dedicated evidence set $\mathcal{W}$ is clearly stronger than that of any model not trained on it.

In summary, across both architecture families, all owners' models are separate from the corresponding reference models, reflecting the uniqueness of $\mathcal{M}_A$'s memorization on $\mathcal{W}$ that Criterion 2 demands.

D. RQ3: Post-Theft Robustness

In the real scenario, the adversary can apply various post-theft modifications to the stolen model, trying to evade IP detection methods. Therefore, an ownership verification protocol is only useful when its signal survives realistic post-theft attacks. We test whether MIA memorization from Section IV-A can continue to flag members of the private evidence set $\mathcal{W}$ after the adversary mutates the stolen $\mathcal{M}_A$ into $\mathcal{M}_B$ via five attacks: MMD-FT, SGD-FT, gaussian perturbation with noise scale of $\sigma = 10^{-3}, 10^{-2}$, and LoRA update.

TABLE III
POST-THEFT ROBUSTNESS OF THE MIO FRAMEWORK.

Attack on $\mathcal{M}_A$	AUC	TPR@10^{-3}	TPR@10^{-4}
None (positive control: $\mathcal{M}_A$)	0.997	0.859	0.819
MMD-FT	0.996	0.860	0.813
SGD-FT	0.994	0.843	0.795
Noise $\sigma = 10^{-3}$	0.997	0.860	0.813
Noise $\sigma = 10^{-2}$	0.956	0.796	0.684
SD LoRA	0.906	0.841	0.758

We first describe the attack methods and experimental settings. For MMD-FT, we use 500 iterations with a cubic-polynomial kernel over a frozen CLIP ViT-B/32 to train MMD-FT, an aggressive method that reshapes the owner's model output distribution by matching CLIP features to the training distribution. SGD-FT serves as a controlled version for MMD-FT, which is trained with exactly the same budget, but on clean data and under its own loss. Since both methods receive the same amount of training, any difference in their results must come from the type of attack rather than from unequal computing. Gaussian weight noise has no training, but the isotropic perturbations of standard deviation $\sigma \in \{10^{-3}, 10^{-2}\}$ are added in place, standing in for a quantization or noise-injection adversary. Stable Diffusion LoRA adversary covers the latent-space case most likely in practice, where a single 2,000-step cross-attention LoRA fine-tune is applied to SD v1.4 with rank 256, using 1,000 disjoint COCO images.

Table III shows that the membership signal survives every attack we test. The un-attacked positive control, i.e., the CIFAR-10 owner model of Table I, attains TPR@10^{-4} of 0.819. Each adversarial attack on $\mathcal{M}_B$ degrades this only partially. Across all five attack attempts, AUC stays at or above 0.906, and TPR@10^{-4} never falls below 0.684, still well above the 10^{-4} chance rate. Overall, the ownership signal persists under MMD and SGD fine-tuning, weight noise at both magnitudes, and the latent-space LoRA attack, which retains a TPR@10^{-4} of 0.758. Such post-theft robustness is what lets Criterion 1 anchor an

TABLE IV
DISTILLATION ATTACK PERFORMANCE ON CIFAR-10

Model	AUC	TPR@10^{-3}	TPR@10^{-4}
Owner $\mathcal{M}_A$ (positive control)	0.997	0.859	0.819
Distilled student	0.492	0.0014	0.0004

ownership claim against a suspect $\mathcal{M}_B$ that has been fine-tuned or perturbed after theft.

Furthermore, we tested the robustness of MiO against a stronger attack. All attacks above start from the stolen model weights and modify them, while a more aggressive adversary could instead sample a large synthetic dataset from the stolen model and train a fresh model on it, which is a knowledge distillation-based attack.

The distillation attack is evaluated in Table IV. 50,000 synthetic CIFAR-10 images are sampled from the stolen model $\mathcal{M}_A$ with 10-step DDIM, and then a student model is trained from scratch for 100,000 iterations at batch 128. On the student model, the per-sample MIA collapses toward a random chance, where the student's hit rate on $\mathcal{W}$ stays at the α floor, so Criterion 1's binomial test fails to reject H_0^{C1} and the protocol returns REJECTED. The reason is that although distilled from $\mathcal{M}_A$, the student model starts from an independent initialization and never sees $\mathcal{W}$, so the signal that per-example memorization relies on does not transfer. However, distillation from scratch costs about as much as retraining a new model, which forfeits the very cost savings that motivate theft. Therefore, in our threat model in Section III-B, distillation is not considered as a common model stealing method.

In summary, the proposed MiO's signal withstands every popular adversary that modifies the stolen weights in our threat model, which makes the ownership verification robust against post-theft change.

E. RQ4: Controlled Baseline Comparison

Existing ownership verification methods are difficult to compare head-to-head on verification effectiveness, because each defines "verification" through a different statistical instrument, evaluated at a different operating point and sample-size regime. For example, SSIM-based watermark similarity is used in WDM [2], bit-accuracy on a fixed message is used in Zhao et al. [3] and SleeperMark [4], and Welch's t-test on aggregated MIA features is used in CDI [22].

Since these protocols share no common operating framework, we instead report each method under its own native detection score in Table VI. Under its own metric, every baseline attains near-perfect verification on a clean protected model: WDM extracts its watermark at an SSIM of 0.997, Zhao and SleeperMark recover their embedded message at 0.999 and 0.99 bit-accuracy, and CDI identifies its dataset with over 99% confidence. MiO is on par with these dedicated schemes, with an AUC of 0.997. However, these scores cannot be collapsed into a single comparable number, such as TPR at a fixed FPR.

TABLE V
COMPUTATIONAL COST BREAKDOWN ON CIFAR-10 DATASET.

Component	Time	Parameters
DDIM training (owner model)	$\sim$34 h	30.5M
Gaussian QR ($\times$1 model)	$\sim$2 h	11.57M
Gaussian QR ensemble (B=50)	$\sim$100 h†	579M total
Inference (per verification query)		
t-error scoring (5k samples)	$\sim$10 min	—
Ensemble prediction	$<$1 min	—
Two-point criteria	$<$1 s	—
Total query time	**$\sim$11 min**	—

We therefore organize the baseline methods not by a single effectiveness score but along two *cost* axes in Table VI: (a) whether the protocol modifies the owner model's weights, which risks degrading generation quality (FID); and (b) whether verification requires an auxiliary component coupled to the deployed model, as opposed to being performed offline on an unmodified model. MiO sits on the favorable end of both axes: (1) It never modifies the owner model's weights because the ownership signal is injected by a private dataset in a non-invasive way, and (2) its Gaussian quantile regression ensemble is constructed offline on the verifier side, so the ownership verification operates on an unmodified model without auxiliary components. As shown in Table VI, MiO requires zero weight modification, adds no overhead to the owner's training pipeline, and incurs no FID cost on the owner model, whereas WDM, Zhao et al., and SleeperMark each embed information into the trained model and consequently add FID cost under their default configurations.

Specifically, CDI is the only baseline on the membership-based side amongst baselines, and the only prior method like MiO that requires zero weight modification or any deployment-time component. However, it is a *dataset-ownership* verification method rather than a *model-ownership* verification method.

Furthermore, we tested the training overhead of the MiO method in Table V. Even though MiO does not need extra computation in embedding watermarks, building the Gaussian quantile regression ensemble offline on the verifier's side also takes a certain amount of cost. The first part of the cost is a one-time, upfront training cost. Note that, although training the ensemble takes $\approx$100 V100-hours for B=50 when run serially on a single V100, it is parallelable across the B=50 models and can be reused for every query. The second part of the cost is the actual cost when an auditor judges the ownership. It covers $\approx$10 minutes to score the 5,000 evidence-set samples on a single V100, plus one minute for ensemble quantile prediction, and an immediate two-point verification. Unlike watermarking-based methods, which spend much computation on re-training the diffusion model under a watermark objective, our MiO instead moves the verification to the verifier in an offline manner.

TABLE VI
COMPARISON ACROSS DIFFUSION-MODEL OWNERSHIP-VERIFICATION METHODS.

Method	Modifies weights	Training overhead	ΔFID	Detection score
WDM	✓	High (extra WDP)	+0.6 (CIFAR-10)	SSIM 0.997
Zhao	✓	High (encoder + DM)	+4.87 (CIFAR-10)	Bit-acc 0.999
CDI	✗	None	N/A	>99% conf. (dataset)
SleeperMark	✓	Medium (LoRA backdoor)	+0.48 (SD v1.4)	Bit-acc ≈0.99
MiO (ours)	✗	**None**	**N/A**	**AUC** 0.997

TABLE VII
AGGREGATION ABLATION ON CIFAR-10.

Aggregation	$\lvert d\rvert_{C2}$
Mean	18.2
Q10	21.5
Q25 (ours)	**23.93**
Median	19.8

TABLE VIII
EFFECT OF ENSEMBLE SIZE B ON QUANTILE PREDICTION STABILITY

Ensemble B	Std($\hat{q}_\tau$)	TPR @ FPR 0.1%
1	0.142	0.83 ± 0.05
10	0.068	0.87 ± 0.03
20	0.041	0.88 ± 0.03
50 (ours)	0.019	0.89 ± 0.03
100	0.016	0.89 ± 0.02

F. Hyperparameters

There are a few hyperparameters in our MiO framework. In this section, we conduct an ablation study for each of them and test the best combination of them.

t-error Aggregation Strategy. Table VII compares four strategies for collapsing the t-error sequence into a single membership score: mean, median, and two lower quantiles Q10 and Q25. We rank them by $|d|_{C2}$ as IV-C defined, so a larger value is preferable. Q25 yields the strongest separation ($|d|_{C2} = 23.93$), surpassing the mean (18.2), the median (19.8), and the more aggressive Q10 (21.5). The same observation holds in the latent-space diffusion setting, so we adopt Q25 in this setting.

The Size of Ensemble B. Varying the bagging ensemble of Gaussian quantile regression from $B = 1$ to 100, we measure the standard deviation of the predicted quantile $\hat{q}_\tau(x)$ over five runs on CIFAR-10 at FPR 0.1%. Table VIII shows that the prediction variance drops sharply from $B = 1$ to 20, keeps falling through $B = 50$, and plateaus when beyond. So, we adopt $B = 50$ as a favorable stability and cost trade-off.

The Number of Sampled Timesteps. We vary K from 10 to 100 and find out that $|d|$ plateaus once $K \geq 25$ and changes negligibly beyond $K = 50$. Therefore, we use the fixed $K = 50$ throughout.

Gaussian QR vs. Pinball-loss QR. There are two mainstream approaches to quantile regression: Gaussian QR and pinball-loss QR. Pinball-loss QR estimates a target quantile directly: it trains a regressor under the pinball-loss to calibrate at FPR α, an asymmetric loss whose minimizer is the conditional α-quantile of the score. At the extreme low FPR operating points we report, $\alpha \in \{10^{-3}, 10^{-4}\}$, the two approaches yield comparable TPR. Gaussian QR's advantage is flexibility: one fitted model returns the threshold $\hat{q}_\alpha(x)$ for any α in closed form via $\Phi^{-1}(\alpha)$, whereas pinball-loss QR requires a separate model per quantile level. In our experiments, each model is evaluated across datasets at several FPR levels rather than one fixed operating point, so we need a calibrator that supplies every α from a single training, which leads to the adoption of Gaussian quantile regression.

Latent vs. Pixel Space in Stable Diffusion. As established in Section III, the reconstruction error e_t for Stable Diffusion can be computed in either latent or pixel space. On SD v1.4, latent-space scoring with per-image caption conditioning yields a higher AUC $= 0.9956$, compared to 0.959 for pixel-space scoring since the VAE decode step dilutes the membership signal in pixel space. So all SD experiments in our settings use the latent space configuration.

VI. CONCLUSION

This work studies the ownership verification problem for diffusion models, aiming to provide reliable intellectual property protection without modifying the model or sacrificing generation quality. To overcome the limitations of existing watermarking methods, which often inject artificial signals into model weights or outputs and may degrade utility, this paper proposes Membership is Ownership (MiO), a method that relies on the private member set as a non-invasive watermark. MiO leverages the model's natural memorization behavior on the private member evidence dataset and formulates ownership verification as a population-level hypothesis test. Specifically, MiO verifies ownership through two key criteria: membership-based attribution on the private evidence set and statistical separation from public reference models, while rigorously controlling the false-accusation probability with a significantly small $p < 10^{-6}$. Experimental results show that MiO achieves stable and reliable ownership verification with negligible impact on model utility, and remains robust under post-theft fine-tuning and weight perturbation attacks. Overall, MiO provides a non-invasive, low-cost, and practical alternative for protecting the intellectual property of diffusion models.

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 684–10 695.

[2] S. Peng, Y. Nie, Y. Zhang, and J. Xiong, "Protecting the intellectual property of diffusion models by the watermark diffusion process," *arXiv preprint arXiv:2306.03436*, 2023.

[3] Y. Zhao, T. Pang, C. Du, X. Yang, N.-M. Cheung, and M. Lin, "A recipe for watermarking diffusion models," *arXiv preprint arXiv:2303.10137*, 2023.

[4] Z. Wang, J. Guo, J. Zhu, Y. Li, H. Huang, and M. Chen, "Sleepermark: Towards robust watermark against fine-tuning text-to-image diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[5] F. Kong, J. Duan, R. Ma, H. T. Shen, X. Shi, X. Zhu, and K. Xu, "An efficient membership inference attack for the diffusion model by proximal initialization," in *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

[6] T. Matsumoto, T. Miura, and N. Yanai, "Membership inference attacks against diffusion models," in *IEEE Security and Privacy Workshops (SPW)*, 2023, pp. 77–83.

[7] H. Hu and J. Pang, "Loss and likelihood based membership inference of diffusion models," in *Proceedings of the 26th Information Security Conference (ISC)*, ser. Lecture Notes in Computer Science, vol. 14411. Springer, 2023, pp. 121–141.

[8] Z. Yang, K. Zeng, K. Chen, H. Fang, W. Zhang, and N. Yu, "Gaussian shading: Provable performance-lossless image watermarking for diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 12 162–12 171.

[9] ——, "Gaussian shading++: Rethinking the realistic deployment challenge of performance-lossless image watermark for diffusion models," *arXiv preprint arXiv:2504.15026*, 2025.

[10] S. Gunn, X. Zhao, and D. Song, "An undetectable watermark for generative image models," in *International Conference on Learning Representations (ICLR)*, 2025, pRC watermark; arXiv:2410.07369.

[11] H. Huang, Y. Wu, and Q. Wang, "ROBIN: Robust and invisible watermarks for diffusion models with adversarial optimization," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024, arXiv:2411.03862.

[12] K. Li, Z. Huang, X. Hou, and C. Hong, "Gaussmarker: Robust dual-domain watermark for diffusion models," in *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

[13] P. Fernandez, G. Couairon, H. Jégou, M. Douze, and T. Furon, "The stable signature: Rooting watermarks in latent diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[14] V. Asnani, J. Collomosse, T. Bui, X. Liu, and S. Agarwal, "ProMark: Proactive diffusion watermarking for causal attribution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 10 802–10 811, arXiv:2403.09914.

[15] X. Zhao, K. Zhang, Z. Su, S. Vasan, I. Grishchenko, C. Kruegel, G. Vigna, Y.-X. Wang, and L. Li, "Invisible image watermarks are provably removable using generative AI," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024, regeneration attack; arXiv:2306.01953.

[16] Y. Hu, Z. Jiang, M. Guo, and N. Z. Gong, "A transfer attack to image watermarks," in *International Conference on Learning Representations (ICLR)*, 2025, arXiv:2403.15365.

[17] F. Shamshad, T. Bakr, Y. Shaaban, N. Hussein, K. Nandakumar, and N. Lukas, "First-place solution to neurips 2024 invisible watermark removal challenge," *arXiv preprint arXiv:2508.21072*, 2025.

[18] I. Alam, M. T. Islam, and S. S. Woo, "Saliency-aware diffusion reconstruction for effective invisible watermark removal," in *Companion Proceedings of the ACM Web Conference (WWW Companion)*, 2025, arXiv:2504.12809.

[19] J. Duan, J. Guan, W. Yang, and R. He, "Visual watermarking in the era of diffusion models: Advances and challenges," *arXiv preprint arXiv:2505.08197*, 2025.

[20] M. Pautov, N. Bogdanov, S. Pyatkin, O. Rogov, and I. Oseledets, "Probabilistically robust watermarking of neural networks," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

[21] P. Maini, M. Yaghini, and N. Papernot, "Dataset inference: Ownership resolution in machine learning," in *International Conference on Learning Representations*, 2021.

[22] J. Dubiński, A. Kowalczuk, F. Boenisch, and A. Dziedzic, "CDI: Copyrighted data identification in diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[23] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *IEEE Symposium on Security and Privacy (S&P)*, 2017, pp. 3–18.

[24] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *IEEE Computer Security Foundations Symposium (CSF)*, 2018, pp. 268–282.

[25] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models," in *Network and Distributed System Security Symposium (NDSS)*, 2019.

[26] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramèr, B. Balle, D. Ippolito, and E. Wallace, "Extracting training data from diffusion models," in *Proceedings of the 32nd USENIX Security Symposium (USENIX Security 2023)*, 2023, pp. 5253–5270.

[27] J. Duan, F. Kong, R. Wang, K. Gu *et al.*, "Diffusion model membership inference via reverse process discrepancy," in *International Conference on Machine Learning (ICML)*, 2023.

[28] M. Bertran, S. Tang, M. Kearns, J. Morgenstern, A. Roth, and Z. S. Wu, "Scalable membership inference attacks via quantile regression," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023.

[29] S. Tang, Z. S. Wu, S. Aydöre, M. Kearns, and A. Roth, "Membership inference attacks on diffusion models via quantile regression," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 235, 2024.

[30] G. Somepalli, V. Singla, M. Goldblum, J. Geiping, and T. Goldstein, "Diffusion art or digital forgery? investigating data replication in diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 6048–6058.