GuidedLatent: Defending VAEs against Membership Inference Attacks via Distribution-Guided Privacy

Chengze Du[a,d,✉], Guangzhen Yao[b], Jibin Shi[c], Ying Zhang[d], Renda Han[e]
[a] School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing, China
[b] College of Science, National University of Defense Technology, Changsha, China
[c] School of Artificial Intelligence, Xidian University, Xi'an, China
[d] AI Department, Zhipu Huazhang Technology, Beijing, China
[e] School of Computer Science and Technology, Hainan University, Hainan, China
ducz0338@bupt.edu.cn, yaoguangzhen@nenu.edu.cn, zround@stu.xidian.edu.cn,
ying.zhang01@aminer.cn, hanrenda@hainanu.edu.cn

***Abstract*—Variational autoencoders (VAEs) have been deployed in many privacy-sensitive domains, and their vulnerability to membership inference attacks (MIAs) poses giant privacy risks. While some existing privacy protection methods like differential privacy often compromise generative models' utility, we present GuidedLatent, a novel mechanism that enhances membership privacy and preserves their generative performance. GuidedLatent allows the model to adjust latent representations dynamically based on distribution similarities, coupled with a two-phase training strategy that gradually incorporates privacy constraints. We also establish bounds on the privacy-utility trade-off theoretically and prove our mechanism reduces the performance of membership inference attacks compared to other baseline approaches. Extensive experiments demonstrate that our method maintains high-quality generation capabilities while minimizing degradation in quality metrics. Our method performs effectively across various VAE variants and architectures, providing a practical solution for privacy-preserving generative models.** [1]

***Index Terms*—Membership Inference Attack, Variational Inference, Variational Autoencoder (VAE), Privacy-Preserving Machine Learning**

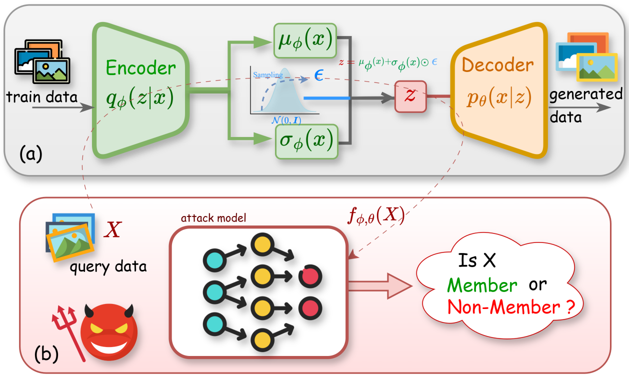

Fig. 1. Overview of VAE architecture and membership inference attack. (a) The standard VAE framework consists of an encoder $q_\phi(z|x)$ that maps input data to latent distributions parameterized by mean $\mu_\phi(x)$ and variance $\sigma_\phi(x)$, followed by sampling through the reparameterization trick to get the z, and a decoder $p_\theta(x|z)$ that generates output data. (b) The membership inference attack framework, where an adversary uses an attack model to determine whether a query sample X was used in training the VAE by analyzing the model's behavior $f_{\phi,\theta}(X)$ on the sample.

I. Introduction

Deep generative models have become powerful tools for data synthesis and representation learning, making increasing differences to sensitive domains such as healthcare imaging, biometric authentication , and personal data processing [16]. However, their deployment has raised significant privacy concerns regarding the potential leakage of sensitive training data through membership inference attacks (MIA) [22], where adversaries attempt to determine whether individual data samples have been used for training models.

The vulnerability of generative models to MIA originated from their learned representations and generation patterns. During training, these models naturally develop statistical patterns for processing and generating data that are similar to training samples, thus creating exploitable differences between how the model handles training versus on-training samples [18]. This behavior is essential for high-quality generation but also provides reliable signals for membership inference attack, posing significant risks in privacy-sensitive applications where the confirmation of an individual's data being used in training [21].

Recent advances in privacy-preserving machine learning have shown various defense mechanisms against membership inference attacks [19]. Some methods [2], [7], [8] based on differential privacy have become a theoretical cornerstone, offering mathematical guarantees through scaleful noise injection during training. Parallel developments in regularization-based approaches [17] leverage additional constraints to mitigate overfitting, and some adversarial training frameworks [30] optimize models explicitly against membership inference through discriminator networks. Knowledge distillation [14] techniques further expand these defenses by transferring representations from

[1] This work was done during Chengze's internship at Zhipu.

private teacher models to public student models, obscuring individual training instances by distillation. Despite these methods having advanced in general privacy preservation, the special characteristics of VAEs present additional challenges that require unique solutions.

As seen in Fig 1, present methods on privacy-preserving VAEs struggle to preserve important latent space characteristics. Traditional privacy mechanisms often fail to address these unique requirements of variational models. Differential privacy techniques, while protecting individual samples, but also degrade latent representation quality [3].Likewise, regularization techniques can break the vital balance between ELBO goals, and adversarial strategies could compromise variational training. This creates an unnecessary trade-off between privacy protection and generative quality in applications requiring both, undermining the fundamental advantages of VAE architectures.

To address these challenges, we propose a novel mechanism (**GuidedLatent**) for privacy-preserving VAEs improves membership privacy while maintaining generative power. And our approach consists of three main components. Firstly, we introduce a controlled distribution mechanism that adjusts latent representations based on cluster-wise similarities, which is used to obscure individual sample characteristics while maintaining semantic structure.Secondly, we develop a two-phase training strategy that the model first learns basic data representations before gradually incorporating privacy-preserving modifications, which helps to maintain training stability. Thirdly, we provide theoretical analysis of the privacy-utility trade-off for our mechanism, establishing bounds on membership inference attack success rates using Hellinger distance related to training and test sample distributions.Unlike previous approaches that rely primarily on noise injection or generic regularization, our mechanism leverages the inherent semantic relationships between data samples to create more robust privacy-preserving latent representations. At the same time, our method also have better performance in defending against membership inference attacks while maintaining high-quality generation capabilities across various domains.

The main contributions of this paper are as follows:

1) **Novel Privacy-Preserving mechanism**: We propose a mechanism for VAEs that enhances privacy through controlled modification of variational posterior distributions. Our approach uses cluster-based distances in the latent space to adjust representations while maintaining the model's generative capabilities.
2) **Technical Innovations**: We introduce three key technical components: (a) a controlled distribution mechanism that dynamically adjusts latent representations based on semantic similarities, (b) a phase-wise training strategy that ensures stable learning of privacy-preserving features, and (c) theoretical bounds that quantify the privacy-utility trade-off in the variational framework.
3) **Comprehensive Evaluation**: We perform extensive experiments on several datasets to show that our method: decreases MIA performance by as much as 18.2% under baseline VAEs while preserving high reconstruction quality, and generalizes across various VAE variants and architectures.

II. Related Work

Attacks A major privacy risk in machine learning is membership inference attacks, which were firstly come up with was in the groundbreaking work of Shokri et al. [22]. Black-box settings [21], [23], where attackers only access model predictions, and white-box settings [18], [20], where model parameters and intermediate computations are available to attackers. All of them are the attack scenarios that were established by early research [9], [28], [31].

The landscape of membership inference attacks evloves with the rise of generative models, particularly in the context of VAEs and their variants [10], [11]. Recent studies [5], [6], [32] have revealed that generative models exhibit unique vulnerabilities to MIA, stemming from their fundamental objective of learning and reproducing data distributions. Groundbreaking work [10] demonstrated that attackers can exploit the reconstruction quality and latent space characteristics of VAEs to achieve high membership inference accuracy. In white-box settings [1], attackers can directly analyze latent representations to identify training samples, as these tend to form distinctive clusters with measurable statistical properties. Black-box attacks [32] have become increasingly sophisticated with methods such as multi-query usage for aggregating model outputs, making use of reconstruction errors as measures of confidence, and exploiting generated sample consistency. These improvements make it critical that privacy protection processes are robust, being specially adapted to generative architecture.

Defense Machine learning privacy preservation has been realized by a series of methods, with differential privacy (DP) as an essential framework that offers formal privacy protection guarantees [2], [7], [8]. Conventional DP techniques add calibrated noise to model gradients or parameters during training time, which has the effect of obscuring the contribution of individual training examples [16]. Aside from DP, other researchers [12], [17], [24] have also discussed other regularization mechanisms that avoid overfitting and memorization like dropout and weight decay, that indirectly impose privacy by making the models generalize well. Knowledge distillation [14], [27] mechanisms try to map learned distributions into a student model while hiding particular training samples, and adversarial training mechanisms [30] actually hamper the effect of any membership inference attacks by adding an adversarial discriminator as part of the training.

III. Background

A. Variational Autoencoders

Variational Autoencoders (VAEs) represent a powerful class of deep generative models that combine variational inference with deep learning to learn complex data distributions. Two primary components of the VAE architecture cooperate: a decoder network reconstructing the input from the latent coding and an encoder network mapping input data to a probabilistic latent representation.

For the encoding process, given an input sample x, the encoder network $q_\phi(z \mid x)$ maps it to a probability distribution in the latent space, parameterized as a multivariate Gaussian distribution $z \sim q_\phi(z|x) = \mathcal{N}(z; \mu_\phi(x), \sigma^2_\phi(x))$. Then the encoder outputs both mean $\mu_\phi(x)$ and variance $\sigma^2_\phi(x)$ parameters. The latent code z is then sampled from this distribution using the reparameterization trick: $z = \mu_\phi(x) + \sigma^2_\phi(x) \bigodot \varepsilon$, where $\varepsilon \sim N(0, I)$, and then this sampled latent code is passed to the decoder network $p_\phi(x \mid z)$, which reconstructs the input by learning to map the latent representation back to the original data space.

VAEs are trained by using the evidence lower bound (ELBO), which consists of two terms: a reconstruction term that ensures high-quality data reconstruction, and a regularization term that encourages the learned latent distribution to match a prior distribution p(z), typically chosen as a standard normal distribution:

$$\mathcal{L}_{VAE}(\theta, \phi) = \underbrace{\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]}_{\text{reconstruction term}} - \underbrace{D_{KL}(q_\phi(z|x)||p(z))}_{\text{regularization term}} \tag{1}$$

B. Membership Inference Attack

Membership inference attacks threaten VAE privacy by identifying whether specific samples were used in training, exploiting the more accurate reconstruction and consistent latent representations of training data. Formally, given a query sample x, a model $\mathcal{V}$ with parameters θ, and an attacker with knowledge level $\mathcal{K}$, the attack aims to classify x as belonging to either the training set $\mathcal{D}_{\text{train}}$ or non-training set $\mathcal{D}_{\text{non-mem}}$. Attack success varies based on the attacker's access level, from complete model parameters to only input-output observations.

The attack mechanism can be formalized using a discrimination threshold τ to separate member from non-member samples. The decision rule for the attack is as follows:

$$\mathcal{M}(x, \mathcal{V}_\theta, \mathcal{K}) = \begin{cases} x \in \mathcal{D}_{\text{train}}, & \text{if } \Phi(x, \mathcal{K}) > \tau \\ x \in \mathcal{D}_{\text{non-mem}}, & \text{otherwise} \end{cases} \tag{2}$$

$\Phi(x, \mathcal{K})$ represents the attack score, which varies depending on the attacker's knowledge K. In white-box scenarios, where the attacker has full access to the model parameters θ, the attack can exploit all aspects of the VAE's behavior. The score function in this case can be expressed as:

$$\begin{aligned} \Phi\text{white}(x, \mathcal{K}) = h\left(||x - \mathcal{V}_\theta(x)||_2, \mathbb{E}_{z\sim q\phi(z|x)}[||z||],\right. \\ \left. D_{\text{KL}}(q\phi(z|x)||p(z)); \theta\right) \end{aligned} \tag{3}$$

where h is a scoring function that combines reconstruction error, expected latent norm, and KL divergence to assess membership likelihood. In contrast, black-box attacks, where the attacker only has access to input-output observations, must rely solely on the reconstruction characteristics of the model. The attack score in this case is:

$$\Phi\text{black}(x, \mathcal{K}) = g\left(||x - \mathcal{V}_\theta(x)||_2\right) \tag{4}$$

where g is a monotonic function that maps reconstruction error to membership likelihood, typically implemented as a simple thresholding or scaling function.

IV. Proposed Method and Analysis

A. Privacy-Preserving Distribution Modification

We propose a novel approach to enhance privacy in VAEs by systematically modifying the learned latent distributions while preserving generation capabilities. Our method builds upon the observation that VAEs naturally learn semantically meaningful representations where similar data samples are mapped to similar latent codes [13]. This property allows us use the natural cluster structures that develop during early training phases to guide our privacy-preserving changes.

The foundation of our approach lies in the controlled modification of the approximate posterior distribution produced by the encoder network $q'\phi(z|x)$. Rather than using the standard encoded distribution parameters directly, we introduce a modified distribution:

$$q'\phi(z|x) = \mathcal{N}(z; \mu'\phi(x), \sigma'\phi(x)) \tag{5}$$

where the mean parameter $\mu'\phi(x)$ is adjusted based on the cluster structures identified during early training:

$$\mu'\phi(x) = \mu_\phi(x) + \lambda \sum_{c\in\mathcal{C}} w_c(\mu_c - \mu_\phi(x)) \tag{6}$$

The cluster influence is weighted by similarity-based coefficients using softmax normalization:

$$w_c = \frac{\exp(-||\mu_\phi(x) - \mu_c||^2)}{\sum_{c'\in\mathcal{C}} \exp(-||\mu_\phi(x) - \mu_{c'}||^2)} \tag{7}$$

where $\mathcal{C}$ represents the set of cluster centers identified in the latent space during the initial training phase, and μ_c denotes the mean of cluster c. To further enhance privacy protection, we introduce controlled variance expansion through a hyperparameter α:

$$\sigma'_\phi(x) = \alpha \cdot \sigma_\phi(x) \tag{8}$$

The modified training objective maintains the core VAE structure while incorporating our distribution adjustments:

$$\mathcal{L}_{total} = -\mathbb{E}_{q'_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q'_\phi(z|x)||p(z)) \tag{9}$$

This formulation allows us to balance privacy protection and generation quality through parameters λ and α. The training process follows a two-phase approach (Algorithm 1): first identifying cluster structures, then gradually applying privacy-preserving modifications to ensure stable training.

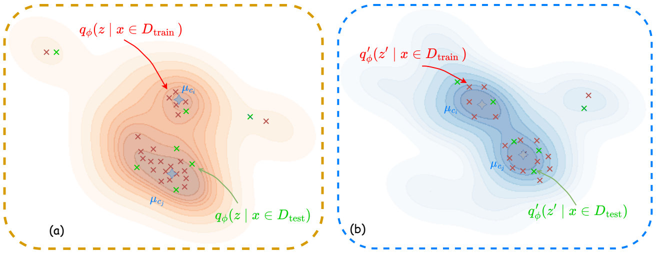
$q_\phi(z \mid x \in D_{\text{train}})$
μ_{c_i}
μ_{c_j}
$q_\phi(z \mid x \in D_{\text{test}})$
(a)
$q'_\phi(z' \mid x \in D_{\text{train}})$
μ_{c_i}
μ_{c_j}
$q'_\phi(z' \mid x \in D_{\text{test}})$
(b)

Fig. 2. Comparison of latent space distributions. (a) Original VAE latent space shows clear separation between training(Member) ($q_\phi(z|x \in D_{\text{train}})$) and test(Non-member) samples ($q_\phi(z|x \in D_{\text{test}})$). (b) Our privacy-enhanced VAE produces more dispersed latent distributions with increased overlap between training and test samples, making membership inference more difficult.

B. Theoretical Privacy Guarantees

We begin by establishing the formal mechanism for analyzing membership inference attacks against Variational Autoencoders (VAEs), building upon theoretical frameworks established in prior literature [4].

Following established approaches in privacy analysis, membership inference attacks can be formalized as a binary hypothesis testing problem:

Definition 1. *(Membership Inference). Given an input sample x_i, the membership inference problem is defined as:*

$$H_0: \quad x_i \in \mathcal{D}_{train} \quad (member)$$
$$H_1: \quad x_i \in \mathcal{D}_{non\text{-}mem} \quad (non\text{-}member)$$

The attack model $\mathcal{M} : (x, f(\theta)) \rightarrow \{out, in\}$ outputs *'in'* if the query sample x is inferred to be in the training set, and *'out'* otherwise. The performance of this binary classifier can be evaluated through its true positive rate (TPR), true negative rate (TNR), false positive rate (FPR), and false negative rate (FNR):

$$\text{TPR} = \mathbb{E}_{x_i}[P(\mathcal{M}(x_i, f(\theta)) = in | x_i \in \mathcal{D}_{\text{train}})]$$
$$\text{TNR} = \mathbb{E}_{x_i}[P(\mathcal{M}(x_i, f(\theta)) = out | x_i \in \mathcal{D}_{\text{non-mem}})]$$

Assumption 1. Due to the KL divergence term in the VAE loss function, the latent representations $z_i = E_\phi(x_i)$ of training samples $x_i \in \mathcal{D}_{\text{train}}$ are constrained to lie within specific regions of the latent space. On the other hand, test samples $x_i \in \mathcal{D}_{\text{test}}$ may map to regions outside these learned clusters. Let P denote the distribution of distances between the training sample encodings and their respective cluster centers, and let Q denote the distribution for the test samples.

Algorithm 1: Privacy-Preserving VAE Training

```
Input  : Dataset $\mathcal{X}$, initial VAE parameters $\theta$, $\phi$,
         hyperparameters $\alpha$, $\lambda$
Output: Privacy-enhanced VAE with parameters
         $\theta'$, $\phi'$
/* Phase 1: Initial Training and Cluster Analysis */
for $epoch = 1$ to $N_{init}$ do
    Update $\theta$, $\phi$ using standard VAE objective;
    if $epoch = N_{init}$ then
        Extract latent representations
         $\mathcal{Z} = \{\mu_\phi(x) | x \in \mathcal{X}\}$;
        Identify cluster centers $\mathcal{C} = \{\mu_c\}$ from $\mathcal{Z}$;
/* Phase 2: Privacy Enhancement Training */
for $epoch = N_{init} + 1$ to $N_{total}$ do
    for each batch $\mathcal{B} \subset \mathcal{X}$ do
        Compute initial encodings $\mu_\phi(x)$, $\sigma_\phi(x)$ for
         $x \in \mathcal{B}$;
        /* Distribution Modification */
        Compute similarity weights:
         $w_c = \text{softmax}(-||\mu_\phi(x) - \mu_c||^2)$;
        Update mean:
         $\mu'_\phi(x) = \mu_\phi(x) + \lambda \sum_c w_c(\mu_c - \mu_\phi(x))$;
        Update variance: $\sigma'_\phi(x) = \alpha \cdot \sigma_\phi(x)$;
        /* Model Update */
        Sample $z \sim \mathcal{N}(\mu'_\phi(x), \sigma'_\phi(x)^2)$;
        Compute
         $\mathcal{L} = \mathbb{E}[\log p_\theta(x|z)] - D_{KL}(q'_\phi(z|x)||p(z))$ ;
        Update $\theta$, $\phi$ using gradient of $\mathcal{L}$;
return $\theta$, $\phi$
```

The Area Under the Receiver Operating Characteristic Curve (AUC) is used to evaluate the attack performance, as it provides a threshold-independent measure of the attacker's ability to distinguish between the distributions P and Q. A higher AUC indicates that the attacker is more effective at separating member from non-member samples based on their distances in latent space.

Theorem 1. *(Privacy Guarantee). For any membership inference attack against our privacy-preserving VAE, the attacker's AUC is bounded by:*

$$AUC \leq -D_H^2(P', Q') + \sqrt{2}D_H(P', Q') + \frac{1}{2} \tag{10}$$

where P' and Q' represent the modified distributions under our privacy mechanism.

Proof 1. *According to Lin et al. [15], we have:*

$$AUC \leq -\frac{1}{2}D_{TV}(P', Q')^2 + D_{TV}(P', Q') + \frac{1}{2} \tag{11}$$

Using the result from [25], we obtain:

$$D_{TV}(P', Q') \leq \sqrt{2}D_H(P', Q') \tag{12}$$

For Gaussian distributions, the Hellinger distance is given by:

$$D_H^2(P', Q') = 1 - \sqrt{\frac{2\sigma_1'\sigma_2'}{{\sigma_1'}^2 + {\sigma_2'}^2}} \exp\left(-\frac{1}{4}\frac{(\mu_1' - \mu_2')^2}{{\sigma_1'}^2 + {\sigma_2'}^2}\right) \tag{13}$$

where μ_1', σ_1' represent the mean and standard deviation of the modified training sample distribution P', and μ_2', σ_2' represent those of the modified test sample distribution Q'. Substituting this into the inequality yields the desired bound. ■

Corollary 1. *(Privacy Enhancement). Under our distribution modification mechanism, we have:*

$$D_H^2(P', Q') \le D_H^2(P, Q)$$

where the Hellinger distance bounded between 0 and 1 [25].

Proof 2. *(Privacy Enhancement) For Gaussian distributions, recall that:*

$$D_H^2(P, Q) = 1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}} \exp\left(-\frac{1}{4}\frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}\right)$$

Under the action of our mechanism, we make the value of μ_1 decrease and the value of $\sigma_1 - \sigma_2$ increase (detailed in Section V-B). Given that $|\sigma_2' - \sigma_2| \ll |\sigma_1' - \sigma_1|$, we can approximate $\sigma_2' \approx \sigma_2$. Using $\mu_1' - \mu_2' = (1-\gamma)(\mu_1 - \mu_2)$ and $\sigma_1' = \omega\sigma_1$ ($\omega > 1$, $0 < \gamma \le 1$) to present this change, then we have

$$\begin{aligned} D_H^2(P', Q') &= 1 - \sqrt{\frac{2\sigma_1\frac{\sigma_2}{\omega}}{\sigma_1^2 + \frac{\sigma_2}{\omega}^2}} \exp\left(-\frac{1}{4}\cdot\frac{(1-\gamma)^2}{\omega^2}\cdot\frac{(\mu_1-\mu_2)^2}{\sigma_1^2 + \frac{\sigma_2}{\omega}^2}\right) \\ &\le 1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}} \exp\left(-\frac{1}{4}\cdot\frac{(1-\gamma)^2}{\omega^2}\cdot\frac{(\mu_1-\mu_2)^2}{\sigma_1^2 + {\sigma_2}^2}\right) \\ &\le 1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}} \exp\left(-\frac{1}{4}\frac{(\mu_1-\mu_2)^2}{\sigma_1^2 + \sigma_2^2}\right) \\ &= D_H^2(P, Q) \end{aligned}$$

where the inequality holds since $\frac{(1-\gamma)^2}{\omega^2} < 1$ for $\omega > 1$ and $0 \le \gamma < 1$, and $\exp(-cx) > \exp(-x)$ for $x > 0$ and $0 < c < 1$. Therefore, combining with Theorem 1, we obtain:

$$\begin{aligned} AUC &\le -D_H^2(P', Q') + \sqrt{2}D_H(P', Q') + \frac{1}{2} \\ &\le -D_H^2(P, Q) + \sqrt{2}D_H(P, Q) + \frac{1}{2} \end{aligned}$$

Note that in our VAE setting, empirical analysis shows $D_H(P, Q) < 0.2$, and the AUC bound function is monotonically increasing in this region. Therefore, our privacy-preserving mechanism effectively reduces the upper bound of the membership inference attack's AUC, thereby providing stronger privacy guarantees. ■

V. Experiments

A. Experiments Setup

1) Dataset and Models: We conduct extensive experiments across multiple standard image datasets: **MNIST**, **Fashion-MNIST**, and CIFAR variants including **CIFAR-10** and **ImageNet-10**. We implement CNN-based and MLP-based VAE architectures. Additionally, we evaluate our method on advanced VAE variants including Vector Quantized VAE (**VQ-VAE**) [29] and Batch Normalization VAE (**BN-VAE**) [26] to demonstrate the generalizability of our approach across different architectural choices.

2) Defense Baselines: For comprehensive comparison, we implement several baseline defense methods: Differential Privacy (**DP-SGD**), **Dropout**, and **Early Stop**. The hyperparameter selection for all defend methods is discussed in Section V-D (see **Note**). All models are trained for 100 epochs with batch size 64, with specific optimization settings for each VAE variant.

B. Analytical Insights

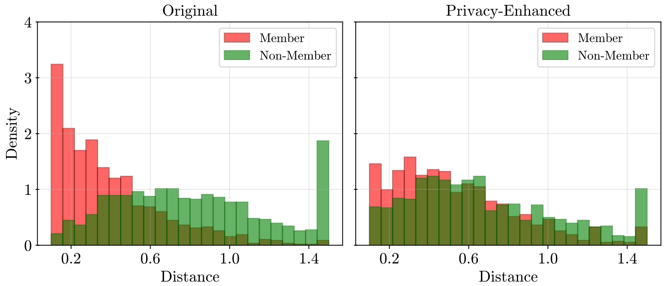

Fig. 3. Comparison of latent space distance distributions before and after applying our privacy-preserving mechanism.

To better understand the effectiveness of our privacy-preserving mechanism, we visualize the distribution of distances between samples and their cluster centers in the latent space. Figure 3 presents a comparative analysis of these distributions before and after applying our method. In the original distribution (left), we observe a clear distinction between training(Member) and test(Non-Member) samples, with training samples exhibiting a notably lower mean distance ($\mu = 0.39$) and tighter spread ($\sigma = 0.05$) compared to test samples ($\mu = 0.82$, $\sigma \approx 0.15$). This separation makes the model vulnerable to membership inference attacks, as an adversary can exploit these distinctive patterns. Following our privacy-preserving approach (right), the mean distances of training and test samples align better; training samples show $\mu = 0.54$ ($\sigma = 0.10$) whereas test samples show $\mu = 0.68$ ($\sigma \approx 0.15$). This reduced separation and increased overlap between the distributions demonstrates that our method effectively obscures membership information while maintaining the underlying structure of the latent space.

TABLE I
MEMBERSHIP INFERENCE ATTACK PERFORMANCE (AUC SCORES) AGAINST DIFFERENT VAE ARCHITECTURES UNDER VARIOUS DEFENSE METHODS.

	Model	MNIST								Fashion-MNIST							
		VAE-MLP		VAE-CNN		BN-VAE		VQ-VAE		VAE-MLP		VAE-CNN		BN-VAE		VQ-VAE	
	Defense	$\mathcal{W}_0/\mathcal{B}_0$	$\mathcal{W}_1/\mathcal{B}_1$	$\mathcal{W}_0/\mathcal{B}_0$	$\mathcal{W}_1/\mathcal{B}_1$	$\mathcal{W}_0/\mathcal{B}_0$	$\mathcal{W}_1/\mathcal{B}_1$	$\mathcal{W}_0/\mathcal{B}_0$	$\mathcal{W}_1/\mathcal{B}_1$	$\mathcal{W}_0/\mathcal{B}_0$	$\mathcal{W}_1/\mathcal{B}_1$	$\mathcal{W}_0/\mathcal{B}_0$	$\mathcal{W}_1/\mathcal{B}_1$	$\mathcal{W}_0/\mathcal{B}_0$	$\mathcal{W}_1/\mathcal{B}_1$	$\mathcal{W}_0/\mathcal{B}_0$	$\mathcal{W}_1/\mathcal{B}_1$
White-box	No-defense	0.552	0.920	0.548	0.935	0.564	0.931	0.559	0.957	0.568	0.945	0.549	0.956	0.561	0.940	0.546	0.959
	DP-SGD	0.539	0.771	0.520	0.789	0.531	0.790	**0.520**	0.781	0.522	0.790	**0.511**	0.780	**0.518**	0.791	**0.510**	0.744
	Dropout	0.544	0.789	0.521	0.793	0.542	0.774	0.535	0.795	0.569	0.763	0.559	0.794	0.551	0.786	0.530	0.760
	Early Stop	0.539	0.780	0.531	0.792	0.540	0.764	0.550	0.791	0.532	0.754	0.544	0.764	0.550	0.775	0.561	0.782
	Ours	**0.519**	**0.732**	**0.513**	**0.752**	**0.509**	**0.764**	**0.520**	**0.720**	**0.519**	**0.691**	0.512	**0.702**	**0.518**	**0.711**	0.514	**0.715**
Black-box	No-defense	0.522	0.770	0.523	0.752	0.518	0.760	0.525	0.773	0.529	0.795	0.525	0.763	0.515	0.782	0.526	0.783
	DP-SGD	0.518	0.682	0.518	0.683	**0.509**	0.680	0.508	0.661	0.511	0.674	0.508	0.640	0.517	0.650	0.506	0.621
	Dropout	0.520	0.701	0.520	0.713	0.511	0.725	0.510	0.734	0.518	0.713	0.511	0.728	0.521	0.731	0.513	0.742
	Early Stop	0.511	0.742	0.513	0.752	0.514	0.764	0.515	0.773	0.523	0.753	0.514	0.765	0.515	0.772	0.519	0.784
	Ours	**0.517**	**0.702**	**0.503**	**0.712**	**0.509**	**0.724**	**0.502**	**0.735**	**0.503**	**0.693**	**0.502**	**0.681**	**0.507**	**0.715**	**0.504**	**0.725**

	Model	ImageNet-10								CIFAR-10							
		VAE-MLP		VAE-CNN		BN-VAE		VQ-VAE		VAE-MLP		VAE-CNN		BN-VAE		VQ-VAE	
	Defense	$\mathcal{W}_0/\mathcal{B}_0$	$\mathcal{W}_1/\mathcal{B}_1$	$\mathcal{W}_0/\mathcal{B}_0$	$\mathcal{W}_1/\mathcal{B}_1$	$\mathcal{W}_0/\mathcal{B}_0$	$\mathcal{W}_1/\mathcal{B}_1$	$\mathcal{W}_0/\mathcal{B}_0$	$\mathcal{W}_1/\mathcal{B}_1$	$\mathcal{W}_0/\mathcal{B}_0$	$\mathcal{W}_1/\mathcal{B}_1$	$\mathcal{W}_0/\mathcal{B}_0$	$\mathcal{W}_1/\mathcal{B}_1$	$\mathcal{W}_0/\mathcal{B}_0$	$\mathcal{W}_1/\mathcal{B}_1$	$\mathcal{W}_0/\mathcal{B}_0$	$\mathcal{W}_1/\mathcal{B}_1$
White-box	No-defense	0.623	0.954	0.631	0.935	0.634	0.937	0.695	0.970	0.630	0.950	0.643	0.960	0.635	0.972	0.656	0.983
	DP-SGD	0.562	0.823	0.572	0.832	0.563	0.834	**0.554**	0.853	0.582	0.834	0.574	0.826	0.574	0.854	**0.565**	0.861
	Dropout	0.593	0.849	0.591	0.892	0.573	0.863	0.603	0.850	0.580	0.853	0.593	0.879	0.595	0.933	0.615	0.883
	Early Stop	0.581	0.879	0.593	0.884	0.594	0.897	0.616	0.870	0.603	0.858	0.584	0.864	0.575	0.875	0.595	0.884
	Ours	**0.537**	**0.772**	**0.547**	**0.781**	**0.527**	**0.783**	0.558	**0.773**	**0.542**	**0.793**	**0.523**	**0.783**	**0.533**	**0.784**	0.558	**0.805**
Black-box	No-defense	0.574	0.982	0.572	0.995	0.563	0.962	0.550	0.970	0.532	0.962	0.545	0.963	0.542	0.963	0.563	0.984
	DP-SGD	**0.517**	0.840	0.539	0.859	0.543	0.868	**0.522**	0.874	0.511	0.853	0.522	0.861	0.521	0.876	0.523	0.883
	Dropout	0.534	0.897	0.535	0.883	0.538	0.884	0.533	0.895	0.518	0.861	0.536	0.863	0.538	0.885	0.544	0.860
	Early Stop	0.552	0.873	0.537	0.876	0.541	0.836	0.540	0.863	0.511	0.843	0.537	0.850	0.527	0.832	0.533	0.846
	Ours	0.521	**0.826**	**0.518**	**0.804**	**0.502**	**0.822**	0.530	**0.808**	**0.509**	**0.795**	**0.514**	**0.803**	**0.512**	**0.799**	**0.520**	**0.783**

C. Defend Attack

We evaluate the effectiveness of different defense methods against membership inference attacks using the Area Under the Curve (AUC) scores as our primary metric. Our evaluation framework encompasses two distinct attack scenarios: white-box ($\mathcal{W}_0$, $\mathcal{W}_1$) and black-box ($\mathcal{B}_0$, $\mathcal{B}_1$). Specifically, $\mathcal{W}_0$ and $\mathcal{B}_0$ represent the attack approaches proposed by Shokri et al. [22] and Chen et al. [5] respectively, and $\mathcal{W}_1$ and $\mathcal{B}_1$ correspond to attack methods introduced by Azadmanesh et al. [1] and Zhang et al. [32], which have demonstrated superior performance in some datasets such as CIFAR-10 and ImageNet-10, as well as in generating models. Additionally, we use Auxiliary data samples to increase their attack performance. Table I presents our experimental results across various VAE architectures and benchmark datasets, providing a comprehensive comparison of defense effectiveness under these different attack scenarios.

1) White Attacks: In white-box settings, our comprehensive experiments reveal several significant insights regarding defense effectiveness against membership inference attacks. First, our novel defense method consistently outperforms the traditional defense mechanisms of DP-SGD, Dropout, and Early Stop with a significant margin in all settings. This superiority is even clearer when the attackers have access to the Auxiliary data samples [32], where our method achieves considerably lower AUC scores compared to baseline methods.

The efficacy of our defense method exhibits fascinating trends under different datasets and model architectures. On simpler datasets like MNIST and Fashion-MNIST, all defense methods work better compared to more challenging datasets like ImageNet-10 and CIFAR-10, which reflects that the complexity of the inherent data distribution plays a major role in defense efficacy. Our method, however, maintains its relative superiority across all dataset complexities, proving itself highly adaptable to different data properties.

2) Black Attacks: In the case of black-box attacks, our experimental results reveal several important observations about the performance of various defense strategies. Although black-box attacks are less successful than white-box attacks in general, the vulnerability of undefended models is still considerable. Our proposed defense strategy shows excellent robustness against both standard black-box attacks ($\mathcal{B}_0$) and its advanced versions ($\mathcal{B}_1$), surpassing traditional defense strategies under all tested settings.

A very interesting outcome is achieved in evaluating the performance on various datasets. In terms of less complex datasets such as MNIST and Fashion-MNIST, the distinction between our proposed method and conventional baseline defenses (DP-SGD, Dropout, and Early Stop) becomes increasingly evident, particularly when it comes to B_1 attacks. This indicates that our defense method has an outstanding performance in defending against state-of-the-art black-box attacks in image recognition tasks of high importance. However, as the complexity of the dataset increases (e.g., ImageNet10 and CIFAR-10), there is a general decrease in the efficacy of defenses for all methods, though our approach maintains its relative advantage.

VAE-CNN, BN-VAE, and VQ-VAE, which implies that the proposed method is architecture-agnostic.

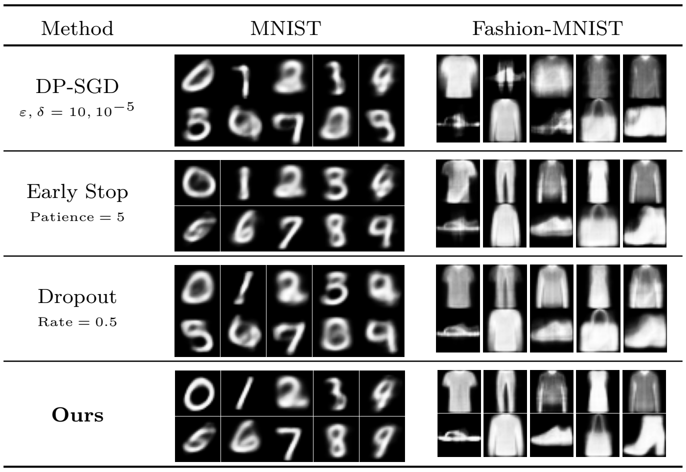

Fig. 4. Comparison of generated samples using different privacy protection methods on MNIST and Fashion-MNIST datasets(Based on VAE-MLP). Our method maintains high visual quality comparable to other defense approaches.

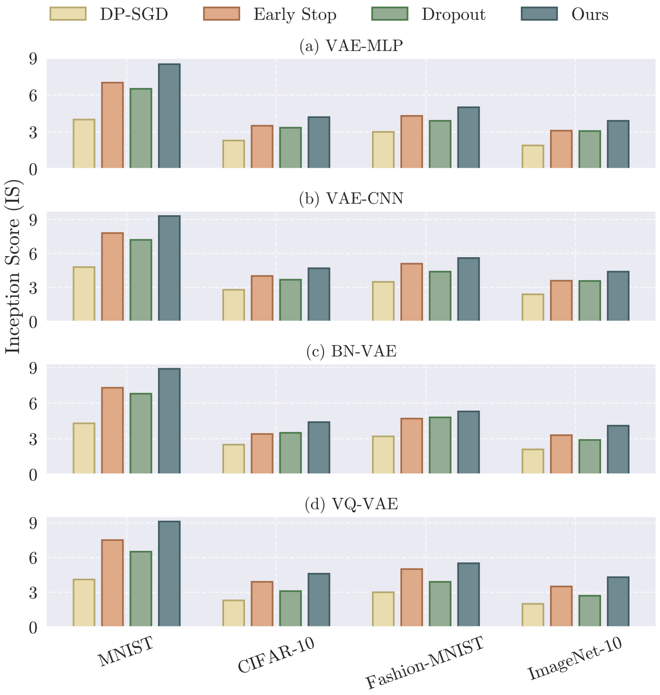

Fig. 5. Inception Score (IS) comparison of different defense methods across VAE architectures (VAE-MLP, VAE-CNN, BN-VAE, and VQ-VAE) on MNIST, CIFAR-10, Fashion-MNIST, and ImageNet-10 datasets. Higher IS indicates better generation quality.

D. Utility

We evaluate the utility of different defense methods by assessing their impact on the generative capabilities of VAE models. Figure 4 presents a qualitative comparison of generated samples across different defense methods on MNIST and Fashion-MNIST datasets. Our method produces visually sharper and more coherent images compared to other defense approaches, particularly DP-SGD which shows noticeable degradation in image quality. The samples generated by our method maintain clear digit structures for MNIST and distinct clothing patterns for Fashion-MNIST, demonstrating that our privacy-enhancing modifications do not significantly compromise the model's generative capabilities.

Figure 5 illustrates a thorough comparison of various defense techniques' effect on generative performance in diverse VAE architectures, employing Inception Score (IS) as the main evaluation metric. Our experimental findings illustrate that our proposed defense technique performs better than baseline methods in all tested settings with especially significant performance on the MNIST dataset where it obtains IS scores of around 9.0 for all architectural variations. This enhanced performance holds through more challenging datasets (CIFAR-10, Fashion-MNIST, and ImageNet-10), although with understandably lower absolute IS values, showing that our method is able to retain generative quality while providing enhanced privacy protection. Interestingly, although VQ-VAE has slightly better performance, especially on MNIST, the relative performance tendencies still hold through different architectural choices (VAE-MLP, VAE-CNN, BN-VAE, and VQ-VAE), demonstrating the robustness of our defense approach. The gap in performance between defense techniques increases on smaller datasets such as MNIST, where our technique demonstrates significant improvement over DP-SGD (which always has the lowest IS scores of about 4.0), while this gap decreases but is still notable on harder datasets such as CIFAR-10 and ImageNet-10.

Note: We observe a trade-off between privacy protection and generation quality controlled by hyperparameters α and λ in GuidedLatent, where larger values provide better defense against membership inference but degrade generation quality. Similar trade-offs exist in baseline methods, with DP-SGD (ε, δ), Early-Stop (patience), and Dropout (rate). Based on empirical evaluation, we select balanced parameters $\alpha = 1.5$ and $\lambda = 0.01$ for GuidedLatent, and $\varepsilon = 10$, $\delta = 10^{-5}$, patience = 5, dropout rate = 0.5 for baselines to achieve an optimal privacy-utility balance.

VI. Conclusion

We present in this paper a mechanism for protecting privacy in VAEs that is more resilient to membership inference attacks at no cost to generative performance. Our approach utilizes cluster-based latent space distances to modify posterior distributions through a two-stage training process that gradually introduces privacy modifications. Experiments demonstrate our approach provides consistently lower AUC scores to membership inference attacks compared to baseline defense strategies across various datasets and VAE architectures. We also provide theoretical bounds with Hellinger distance to study the privacy-utility trade-off, showing that our distribution adjustment mechanism does hinder attackers' inference capability. The method performs particularly well on popular benchmark datasets with decent generation quality, as corroborated by IS scores. The results show that leveraging latent space clustering for distribution adjustment is an effective way to enhance privacy in variational generative models.

References

[1] Maryam Azadmanesh, Behrouz Shahgholi Ghahfarokhi, and Maede Ashouri Talouki. A white-box generator membership inference attack against generative models. In *2021 18th International ISC Conference on Information Security and Cryptology (ISCISC)*, pages 13–17. IEEE, 2021.

[2] Tianshi Cao, Alex Bie, Arash Vahdat, Sanja Fidler, and Karsten Kreis. Don't generate me: Training differentially private generative models with sinkhorn divergence. *Advances in Neural Information Processing Systems*, 34:12480–12492, 2021.

[3] Dingfan Chen, Raouf Kerkouche, and Mario Fritz. A unified view of differentially private deep generative modeling. *Transactions on Machine Learning Research (TMLR)*, 2024. Survey Certification.

[4] Dingfan Chen, Ning Yu, and Mario Fritz. Relaxloss: Defending membership inference attacks without losing utility. In *International Conference on Learning Representations (ICLR)*, 2022.

[5] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. Gan-leaks: A taxonomy of membership inference attacks against generative models. In *ACM Conference on Computer and Communications Security (CCS)*, 2020.

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

[7] Hansle Gwon, Imjin Ahn, Yunha Kim, Hee Jun Kang, Hyeram Seo, Heejung Choi, Ha Na Cho, Minkyoung Kim, JiYe Han, Gaeun Kee, et al. Ldp-gan: Generative adversarial networks with local differential privacy for patient medical records synthesis. *Computers in Biology and Medicine*, 168:107738, 2024.

[8] Conor Hassan, Robert Salomone, and Kerrie Mengersen. Deep generative models, synthetic tabular data, and differential privacy: An overview and synthesis. *arXiv preprint arXiv:2307.15424*, 2023.

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

[10] Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, 3, 2017.

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

[12] Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. Memguard: Defending against black-box membership inference attacks via adversarial examples. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, pages 259–274, 2019.

[13] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: An unsupervised and generative approach to clustering, 2017.

[14] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International conference on machine learning*, pages 1376–1385. PMLR, 2015.

[15] Zinan Lin, Vyas Sekar, and Giulia Fanti. On the privacy properties of gan-generated samples. In *International Conference on Artificial Intelligence and Statistics*, pages 1522–1530. PMLR, 2021.

[16] Yihao Liu, Jinhe Huang, Yanjie Li, Dong Wang, and Bin Xiao. Generative ai model privacy: a survey. *Artificial Intelligence Review*, 58(1):1–47, 2025.

[17] Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, pages 634–646, 2018.

[18] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 739–753, 2019.

[19] Jun Niu, Peng Liu, Xiaoyan Zhu, Kuo Shen, Yuecong Wang, Haotian Chi, Yulong Shen, Xiaohong Jiang, Jianfeng Ma, and Yuqing Zhang. A survey on membership inference attacks and defenses in machine learning. *Journal of Information and Intelligence*, 2024.

[20] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, pages 5558–5567. PMLR, 2019.

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

[22] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

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

[24] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[25] Ton Steerneman. On the total variation and hellinger distance between signed measures; an application to product measures. *Proceedings of the American Mathematical Society*, 88(4):684–688, 1983.

[26] Jianlin Su. Variational autoencoders (v): Vae + bn = better vae, 2021.

[27] Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15731–15740, 2024.

[28] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. Demystifying membership inference attacks in machine learning as a service. *IEEE transactions on services computing*, 14(6):2073–2089, 2019.

[29] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[30] Ling Yang, Haotian Qian, Zhilong Zhang, Jingwei Liu, and Bin Cui. Structure-guided adversarial training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7256–7266, 2024.

[31] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 3093–3106, 2022.

[32] Minxing Zhang, Ning Yu, Rui Wen, Michael Backes, and Yang Zhang. Generated distributions are all you need for membership inference attacks against generative models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4839–4849, 2024.