Securing Representations via Latent Disruption and Private Decoding

Anonymous authors
Paper under double-blind review

Abstract

Pre-trained encoders facilitate efficient data sharing through semantically rich latent embeddings, which, however, pose privacy risks under malicious inference or exploitation. We propose **SEAL**, an attack-agnostic framework that secures latent spaces by disrupting semantic dependencies based on information-theoretic principles. It prevents potential misuse while enabling *selective reconstruction* for trusted users. **SEAL** learns to encode controlled perturbations by minimizing the *Matrix Norm-based Quadratic Mutual Information* (MQMI) functional between original and secured embeddings within a hyperspherical latent space. Meanwhile, a private decoder, jointly trained with the **SEAL** encoder, ensures accurate reconstruction that is accessible only to authorized users. Extensive experiments on vision and text datasets demonstrate that **SEAL** effectively mitigates latent leakage, defends against inference attacks, and preserves reconstruction utility.

1 Introduction

Large-scale pretrained encoders have transformed data-sharing workflows through latent representations that are semantically rich, transmission-efficient, yet not human-interpretable (Awais et al., 2025; Zhang et al., 2024; Huang et al., 2024; Oquab et al., 2024; Nobi et al., 2022; Dosovitskiy et al., 2021). Meanwhile, in scenarios that require human understanding, such as clinical diagnosis based on medical records or commercial analysis based on textual inputs, the information contained in the original signal often remains crucial (Dey et al., 2022; Chung et al., 2021). This raises the question of how to secure embeddings against misuse without compromising their usefulness for legitimate human use.

Latent embeddings have been increasingly recognized as potential sources of privacy leakage (Huang et al., 2024; Morris et al., 2023; Samuel et al., 2023; Kwon et al., 2023; Fredrikson et al., 2015). Leaked embeddings, much like original data, can be exploited to train models (Zhu et al., 2024; Liu et al., 2024; He et al., 2023) or attacked through membership and attribute inference (Wen et al., 2024; Kim et al., 2024; Ko et al., 2023; Jia & Gong, 2018). In practice, data-driven workflows increasingly rely on continuously acquired and shared information (Meyer et al., 2018; Li et al., 2010). As such information propagates across platforms and evolves over time, the risk of unintended exposure naturally increases. Ensuring its security is therefore essential not only to prevent misuse, but also to preserve its legitimate usability for trusted parties.

To safeguard data against potential misuse in the event of information leakage, *data poisoning* methods deliberately introduce imperceptible perturbations that preserve the visual appearance while preventing the data from being exploited for malicious purposes (Fowl et al., 2024; Huang et al., 2021). Early studies focused on input-level poisoning (Fang et al., 2024; He et al., 2023; Liu et al., 2023), introducing *availability poisons*, for example, to hinder unauthorized model training (Zhu et al., 2024; Liu et al., 2024; Meng et al., 2024; Biggio & Roli, 2018). However, these approaches require per-dataset optimization and depend on *specific* objective functions (He et al., 2023; Huang et al., 2021), limiting their scalability and practicality under attack-agnostic threat models. Furthermore, they operate primarily in the input space, leaving *latent embeddings* vulnerable to privacy leakage.

We propose **SEAL** (*Secured Embedding via Adversarial Learning*), which introduces controlled perturbations in the latent space while enabling *selective reconstruction* for authorized users. **SEAL** consists of a disruption

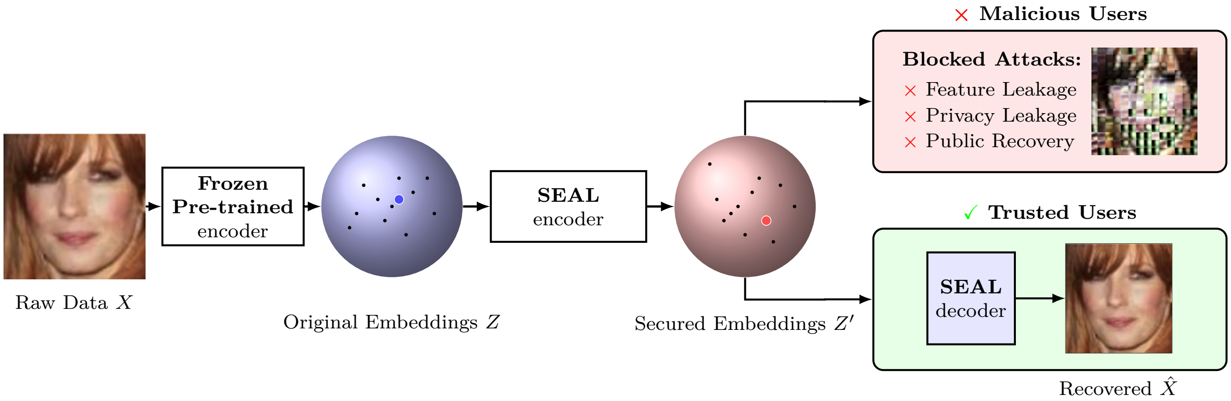


Figure 1: Overview of the proposed **SEAL** framework. A pretrained encoder, including ViT-based models (Siméoni et al., 2025; Oquab et al., 2024; Caron et al., 2021b) and encoder-only Transformer models (Warner et al., 2024; Liu et al., 2019; Devlin et al., 2019), maps sensitive data X to latent embeddings Z on a hyperspherical manifold (illustrated here as a 2D sphere for clarity), where the blue dot represents a sample among other data embeddings (black dots). A *geodesic transformation*, learned by the **SEAL** encoder, perturbs Z into the secured embedding Z', shown as the red dot. The **SEAL** encoder can be deployed as part of a secure service that only outputs Z', while trusted users employ a co-trained private decoder to reconstruct the original input from Z'. This design mitigates the risk of unauthorized reconstruction or inference attacks under potential exposure to Z'.

encoder and a private decoder, trained jointly to prevent misuse and to preserve accurate input reconstruction for those with access to the decoder. Our method minimizes surrogate mutual information functional to reduce dependence between original and perturbed embeddings, offering a principled and task-independent approach to limiting potential exploitation.

The main contributions of our work are:

- We propose **SEAL**, a framework that learns to apply controlled perturbations to latent embeddings, providing *protection against malicious exploitation* while enabling *selective reconstruction* for trusted users through a jointly trained private decoder.
- **SEAL** introduces the *Matrix Norm-based Quadratic Mutual Information (MQMI)*, a functional defined over the eigenspectrum of kernel matrices that quantifies the dependency between original and secured embeddings, enabling interpretable control over the disruption process.
- Experiments show that **SEAL** effectively disrupts malicious exploitation of embeddings, such as training poisoned classifiers or k-NN models, while reducing risks of membership and attribute inference. It simultaneously preserves high-quality reconstruction for trusted users through the private decoder.

2 Related Work

2.1 Data Protection through Controlled Corruption

Data poisoning attacks (Zhu et al., 2024; Liu et al., 2024; Meng et al., 2024; He et al., 2023; Zhang et al., 2023), also referred to as data availability poisons (Fang et al., 2024; Liu et al., 2023), aim to protect data from malicious use by deliberately corrupting datasets, thereby preventing models trained on them from generalizing effectively. For support vector machines (SVMs), even a single poisoned data point could degrade model performance significantly (Biggio & Roli, 2018; Demontis et al., 2019; Biggio et al., 2012). These attacks are generated through bi-level optimization (Mei & Zhu, 2015; Xiao et al., 2015; Biggio et al., 2012).

In deep learning, the aforementioned optimization becomes computationally infeasible, motivating data poisoning attacks like *unlearnable examples* (Zhu et al., 2024; Liu et al., 2024; Meng et al., 2024; Zhang et al., 2023; Huang et al., 2021) and adversarial poisoning (Fowl et al., 2024) that makes strong poisons. These methods typically use Projected Gradient Descent (PGD) (Fowl et al., 2024; Huang et al., 2021; Madry et al., 2017) to inject perturbations in the input space. By adding carefully crafted perturbations imperceptible to humans, these methods produce datasets that are unexploitable for model training.

Recent research has also extended data poisoning attacks to self-supervised frameworks like SimCLR (Chen et al., 2020), MoCo (He et al., 2019), and BYOL (Grill et al., 2020), targeting contrastive learning objectives. Despite the strong representation learning capability of contrastive frameworks, data poisoning can still produce datasets unexploitable for contrastive training and significantly degrade performance on downstream tasks (He et al., 2023).

2.2 The Risk of Malicious Exploitation in Latent Spaces

Data poisoning methods mainly perturb source data, such as images (Liu et al., 2024; Fang et al., 2024; Huang et al., 2021). However, latent representations extracted by publicly available encoders, such as a DINO family (Siméoni et al., 2025; Oquab et al., 2024; Caron et al., 2021b) and a BERT family (Warner et al., 2024; Liu et al., 2019; Devlin et al., 2019), also retain rich semantic information and have emerged as a critical source of privacy leakage (Huang et al., 2024; Morris et al., 2023; Samuel et al., 2023; Kwon et al., 2023; Fredrikson et al., 2015).

In practical deployments such as healthcare or enterprise collaboration (Schneider et al., 2024; Khalid et al., 2023), data-driven workflows increasingly rely on continuously acquired and shared information (Bhatta, 2024; Borra, 2024; Zou et al., 2020). As such information propagates across services and evolves over time, latent embeddings that encode high-level semantics become particularly vulnerable to unintended exposure.

Recent studies show that embeddings from pretrained models retain substantial semantic content, making them exploitable even without access to original inputs (Huang et al., 2024; Morris et al., 2023; Fredrikson et al., 2015). Feature inversion attacks (Fredrikson et al., 2015; Mahendran & Vedaldi, 2014) pose a major threat by reconstructing approximate inputs or inferring sensitive attributes directly from latent spaces. These threats motivate the need for protection mechanisms against malicious exploitation of latent representations.

3 Method

The geometry of latent embeddings is determined by the input data distribution and the encoder architecture. In typical practical settings, e.g., transformers with LayerNorm (Wu et al., 2024; Ba et al., 2016), embeddings exhibit high-probability concentration near a hypersphere $\|z\|_2 \simeq R$ with radius R. This can be explained by concentration of norms in high dimensions (Vershynin, 2018) and is observed across pretrained models and datasets, as shown in Figure 8 in Appendix F.

3.1 Setup and Objective

We assume $\|z\|_2 \simeq R$, and train a **SEAL** encoder $h_\theta : \mathbb{R}^d \to \mathbb{R}^d$ to map each embedding z to its secured counterpart $z' = h_\theta(z)$, while preserving the hyperspherical constraint. We define the *chordal shift* $\delta = z' - z$, as illustrated in Figure 2, and denote by Z and Z' the corresponding random variables. We aim to minimize the statistical dependence between Z and Z', such that the **SEAL** encoder learns to introduce controlled disruption over the hyperspherical embedding manifold.

3.2 Matrix Norm-based Quadratic Rényi's Entropy Functional

Estimating mutual information is notoriously difficult, particularly in high-dimensional settings (Czyz et al., 2025; 2023; Goldfeld & Greenewald, 2021). While variational approaches based on neural networks require auxiliary model training, we adopt recent spectral methods (Skean et al., 2024; Yu et al., 2019; Giraldo et al., 2014) that define entropy functional directly over the eigenspectrum of kernel matrices. This avoids

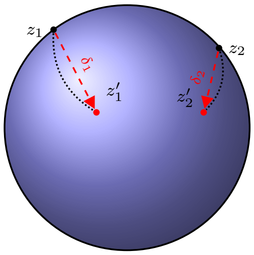


Figure 2: Visualization of hyperspherical disruption. Two example points z_1 and z_2 and their perturbed versions z'_1, z'_2 (red dots) are shown on a 2D sphere. Red dashed lines denote chord distances δ_1 and δ_2, while black dotted curves indicate geodesic paths on the hypersphere.

the need for probabilistic modeling in high-dimensional spaces, satisfies Rényi's axioms for entropy (Rényi, 1961), and allows for efficient, training-free computation. Given a batch of N embedding samples $\{z_i\}$ from the hyperspherical random variable Z, we construct the Gram matrix $K^Z \in \mathbb{R}^{N\times N}$ to capture pairwise similarities.

For hyperspherical embeddings, the von Mises–Fisher (vMF) kernel is commonly used to model directional data (Trosten et al., 2023; Wang & Isola, 2022). Specifically, the vMF kernel entry between two embeddings z_i and z_j is given by $K^Z_{ij} = \exp\big(\kappa\, z_i^\top z_j/(\|z_i\|_2\|z_j\|_2)\big)$, where $\kappa > 0$ is a concentration hyperparameter controlling the sensitivity of the similarity measure.

In the following, we denote a normalized positive definite (NPD) matrix A, obtained from the original Gram matrix K and evaluated with the vMF kernel, as

$$A = \frac{K}{\mathrm{Tr}(K)}, \tag{1}$$

following the convention in Yu et al. (2019); Giraldo et al. (2014).

3.2.1 Matrix Norm-based Quadratic Entropy Functional

The *Matrix Norm-based Quadratic (MQ)* entropy functional of Z is given by:

$$\mathbf{S}_2(Z) \;=\; -\,\log \big\|A^Z\big\|_F^2, \tag{2}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The formulation follows the original eigenvalue-based definition of the matrix-based Rényi's entropy functional, but by setting the order to $\alpha = 2$, the expression naturally avoids the eigen decomposition of the normalized Gram matrix, thereby reducing the computational complexity from $O(N^3)$ to $O(N^2)$ (Skean et al., 2025; 2024; Bach, 2022; Yu et al., 2019).

For Δ, where each chord $\delta_i = z'_i - z_i$ connects corresponding points z_i and z'_i on the hypersphere for $i \in \{1, \ldots, N\}$, we compute the Gram matrix K^Δ over samples $(\delta_1, \delta_2, \ldots, \delta_N)$ using a radial basis function (RBF) kernel. Specifically, the entries are given by $K^\Delta_{ij} = \exp\big(-\|\delta_i - \delta_j\|_2^2/(2\sigma^2)\big)$, where σ is the kernel width parameter. The MQ entropy functional of Δ is then defined as:

$$\mathbf{S}_2(\Delta) \;=\; -\,\log \big\|A^\Delta\big\|_F^2, \tag{3}$$

which evaluates the MQ entropy functional on the chordal disruptions in the latent space.

3.2.2 MQ Joint Entropy Functional

We define the *joint Gram matrix* $K^{Z,Z'}$ as the element-wise (Hadamard) product of the individual Gram matrices (Yu et al., 2019), that is, $K^{Z,Z'} = K^Z \odot K^{Z'}$. The MQ joint entropy functional of Z, Z' is given by:

$$\mathbf{S}_2(Z, Z') \;=\; -\,\log \big\|A^{Z,Z'}\big\|_F^2. \tag{4}$$

3.3 Matrix Norm-based Quadratic Mutual Information Functional

We aim to quantify the dependency between the original embeddings Z and their perturbed counterparts Z'. In analogy with Shannon's definition, we adopt a matrix-based functional of *Quadratic Mutual Information* (Yu et al., 2019):

$$\mathbf{I}_2(Z, Z') = \mathbf{S}_2(Z) + \mathbf{S}_2(Z') - \mathbf{S}_2(Z, Z'). \tag{5}$$

While the exact form of $\mathbf{S}_2(Z')$ depends jointly on both the original embeddings and the perturbations, it is desirable to separate the effect of the perturbation in order to explicitly regulate it.

To enable direct control over how perturbations affect the representations, we introduce the chordal shift variable $\Delta := Z' - Z$, which captures how each embedding moves on the hypersphere. Instead of modeling the marginal entropy of Z' directly, we regulate the perturbation statistics through the MQ entropy of Δ, allowing **SEAL** to shape the perturbation and its interaction with the original embeddings.

This yields the following objective:

$$\mathrm{MQMI}(Z, Z') = -\beta \underbrace{\log \|A^{\Delta}\|_F^2}_{\text{entropy}} \; + \; \underbrace{\log \|A^{Z,Z'}\|_F^2}_{\text{disruption}}, \tag{MQMI}$$

where $\beta \in \mathbb{R}$ acts as a Lagrange multiplier balancing:

- **Entropy:** the MQ marginal entropy of the perturbations Δ;
- **Disruption:** the MQ joint entropy between Z and Z'.

A geometric justification for substituting $\mathbf{S}_2(Z')$ with $\mathbf{S}_2(\Delta)$ is provided in Appendix D. We adopt this formulation as a practical surrogate because the perturbations Δ directly capture how Z' deviates from Z, allowing the objective to isolate and control this effect in a tunable way. The empirical behavior of this surrogate is further validated in our ablation study.

3.4 Joint Training with Private Decoder

By minimizing MQMI, we decrease the shared information between Z and Z', which disrupts malicious exploitation of embeddings. Concurrently, we aim to ensure that **only** trusted users have access to reconstruct the data from the secured embeddings.

We jointly optimize the **SEAL** encoder h_θ, which generates secured embeddings Z', and the private decoder g_ω, which enables selective reconstruction $\hat{X} = g_\omega(Z')$, as illustrated in Figure 1. The overall objective is:

$$\min_{\theta,\omega} \Big[\mathrm{MQMI}(Z, Z') \; + \; \gamma\, \mathcal{L}_{\mathrm{recon}}(\hat{X}, X)\Big], \tag{\textbf{SEAL}}$$

where minimizing $\mathrm{MQMI}(Z, Z')$ reduces the dependency between Z and Z', introducing controlled disruption, while $\mathcal{L}_{\mathrm{recon}}(\hat{X}, X)$ ensures reconstruction utility through the private decoder using secured embeddings. The hyperparameter γ balances the two objectives.

3.5 Trade-off Between Disruption and Utility

The **SEAL** framework is built upon a trade-off between *disruption* and *utility*, where *utility* denotes the remaining information within the secured embeddings that is usable for reconstruction by an authorized private decoder. The encoder is trained to reduce statistical dependencies in the latent representation, while the private decoder reconstructs the input from the latent embeddings. This design deliberately disrupts latent information to safeguard data under potential latent-space leakage, preventing malicious exploitation while preserving sufficiency for authorized reconstruction.

3.6 Geometric Motivation for the Perturbation Term

To understand why the perturbation-based regularization in our objective naturally arises on the hypersphere, we provide a brief geometric motivation. Since both the original embeddings Z and the perturbed embeddings Z' lie on $\mathbb{S}^{d-1}$, any change can be represented as a *chordal shift*

$$\Delta := Z' - Z,$$

which captures how each embedding moves under the perturbation.

Because z_i and $z_i' = z_i + \delta_i$ lie on the unit sphere, their relationship satisfies a simple constraint:

Proposition 3.1 (Chord–Sphere Constraint)**.** *Let $z \in \mathbb{S}^{d-1}$ and $z' = z + \delta$ with $\|z'\|_2 = 1$. Then*

$$z^\top \delta \;=\; -\tfrac{1}{2}\|\delta\|_2^2.$$

This identity simply states that, on the unit sphere, the inner product between z and the perturbation δ is completely determined by the norm of δ. The matrix–based quadratic entropy functional $\mathbf{S}_2(\cdot)$ captures pairwise interactions through kernel evaluations. Replacing the marginal entropy of Z' with the entropy of the chordal shifts $\Delta = Z' - Z$ allows us to directly regulate the amount of perturbation introduced.

In practice, this leads to the objective in Equation (MQMI), where the perturbation term $\mathbf{S}_2(\Delta)$ is controlled by a coefficient β. A brief derivation and supporting details are provided in Appendix D.

STL-10					
Metric	**Baseline**	**MMD**	**CS-Div**	**CS-QMI**	**MQMI**
LR ↓	95.45	86.16	78.91	36.10	**24.59**
R@1 ↓	84.06	80.35	80.16	15.90	**1.34**
R@5 ↓	96.79	95.61	95.73	47.43	**4.79**
F1@5 ↓	86.47	84.18	84.10	23.90	**1.88**

Tiny-ImageNet					
Metric	**Baseline**	**MMD**	**CS-Div**	**CS-QMI**	**MQMI**
LR ↓	77.78	77.74	76.82	78.78	**74.70**
R@1 ↓	74.62	74.92	72.62	64.84	**31.62**
R@5 ↓	87.84	88.10	87.10	82.88	**45.66**
F1@5 ↓	78.16	79.06	77.37	70.84	**32.21**

20 Newsgroup					
Metric	**Baseline**	**MMD**	**CS-Div**	**CS-QMI**	**MQMI**
LR ↓	61.47	28.41	5.57	12.02	**3.09**
R@1 ↓	41.10	21.79	4.02	11.14	**3.62**
R@5 ↓	72.12	50.42	16.70	39.02	**15.38**
F1@5 ↓	47.04	29.31	6.96	18.90	**5.46**

AG News					
Metric	**Baseline**	**MMD**	**CS-Div**	**CS-QMI**	**MQMI**
LR ↓	90.53	4.52	**2.70**	68.84	5.04
R@1 ↓	86.25	38.24	**4.67**	41.73	11.50
R@5 ↓	97.38	66.53	**14.38**	81.54	37.70
F1@5 ↓	89.77	44.07	**7.58**	53.24	17.82

Table 1: **Feature Leakage Evaluation.** We evaluate by training models on secured embeddings and testing on clean embeddings. The baseline uses clean embeddings for both training and testing. Lower values (↓) indicate stronger disruption. Classification and retrieval performance are measured using logistic regression and k-NN across vision (STL-10 with ViT-Base (MAE), Tiny-ImageNet with ViT-Base (DINO)) and text (20 Newsgroup, AG News with BERT-Base) datasets. Metrics include classification accuracy (LR), recall at k (R@1, R@5), F1-score. We compare the proposed MQMI with other kernel-based methods: MMD, CS divergence (CS-Div), and CS-QMI.

4 Experiments

We present a comprehensive evaluation of our proposed **SEAL** framework, demonstrating its ability to disrupt latent representations against malicious exploitation while preserving faithful reconstruction for authorized users. We use STL-10 (Deng et al., 2009), Tiny-ImageNet (Deng et al., 2009), CelebA-Smiles (Liu et al., 2015), and UTKFace (Zhang et al., 2017) for vision tasks, and Emotion (Saravia et al., 2018), AG News (Zhang et al., 2015), and 20 Newsgroups (Mitchell, 1997) for NLP tasks. Vision data uses ViT-base (MAE) (He et al., 2021) and ViT-base (DINO) (Caron et al., 2021a) as the encoder, while text data uses a BERT-base model (Devlin et al., 2019). Full experimental details are provided in Appendices A and B.

UTKFace					
Metric	**Base.**	**MMD**	**CS-Div**	**CS-QMI**	**MQMI**
M-Acc↓	50.15	50.05	**49.93**	50.15	50.96
M-Prec↓	80.16	80.05	**79.94**	80.16	80.96
M-AUC↓	0.51	0.50	0.50	0.51	0.52
$\mathrm{A_G}$-Acc↓	91.31	90.59	90.51	91.25	**81.97**
$\mathrm{A_G}$-Prec↓	91.32	90.60	90.55	91.25	**80.96**
$\mathrm{A_G}$-AUC↓	0.97	0.97	0.97	0.97	**0.90**
$\mathrm{A_E}$-Acc↓	77.47	78.51	**77.16**	77.39	77.24
$\mathrm{A_E}$-Prec↓	68.04	70.22	69.32	**67.57**	68.01
$\mathrm{A_E}$-AUC↓	0.91	0.92	0.91	0.91	0.90

CelebA-Smiles					
Metric	**Base.**	**MMD**	**CS-Div**	**CS-QMI**	**MQMI**
M-Acc↓	50.02	49.82	49.94	**49.62**	49.87
M-Prec↓	80.02	79.82	79.94	**79.62**	79.87
M-AUC↓	0.50	0.50	0.50	0.50	0.50
$\mathrm{A_M}$-Acc↓	96.93	96.63	96.73	96.54	**92.01**
$\mathrm{A_M}$-Prec↓	96.80	96.49	96.60	96.38	**91.79**
$\mathrm{A_M}$-AUC↓	0.99	0.99	0.99	0.99	**0.97**
$\mathrm{A_A}$-Acc↓	80.95	80.83	80.88	80.68	**79.39**
$\mathrm{A_A}$-Prec↓	80.98	80.85	80.91	80.69	**79.39**
$\mathrm{A_A}$-AUC↓	0.90	0.89	0.90	0.89	**0.87**

20 Newsgroup					
Metric	**Base.**	**MMD**	**CS-Div**	**CS-QMI**	**MQMI**
M-Acc↓	56.79	52.78	**52.33**	57.53	57.59
M-Prec↓	66.83	62.81	**62.37**	67.57	67.62
M-AUC↓	0.63	**0.54**	**0.54**	0.63	0.62
A-Acc↓	58.86	58.76	56.70	56.72	**55.56**
A-Prec↓	58.51	57.12	55.17	55.99	**54.42**
A-AUC↓	0.94	0.93	**0.92**	0.93	0.93

AG News					
Metric	**Base.**	**MMD**	**CS-Div**	**CS-QMI**	**MQMI**
M-Acc↓	50.06	50.30	50.46	50.04	**49.99**
M-Prec↓	94.10	94.34	94.50	94.08	**94.03**
M-AUC↓	0.50	0.52	0.53	0.50	0.50
A-Acc↓	90.32	87.89	**84.67**	89.47	89.26
A-Prec↓	90.31	87.90	**84.70**	89.45	89.27
A-AUC↓	0.98	0.98	**0.96**	0.98	0.98

Table 2: **Privacy Evaluation.** We evaluate membership (M) and attribute (A) inference attacks under various disruption methods. In each attack, a classifier is trained and evaluated on secured embeddings to simulate latent leakage. Metrics include accuracy, precision, and AUC, where lower values (↓) indicate stronger privacy preservation. Vision benchmarks use DINO-Base ViT encoders and test for gender ($\mathrm{A_G}$, $\mathrm{A_M}$), ethnicity ($\mathrm{A_E}$), and attractiveness ($\mathrm{A_A}$); text benchmarks use BERT-Base and evaluate for category-level attribute prediction.

STL-10	**LR ↓**	**R@1 ↓**	**R@5 ↓**	**F1@5 ↓**
$\beta = 0.0$	45.45	41.10	71.14	46.98
$\beta = -0.1$	34.94	40.85	71.07	46.50
$\beta = -1.0$	**24.59**	**1.34**	**4.79**	**1.88**
CS-QMI	36.10	15.90	47.43	23.90

Emotion	**LR ↓**	**R@1 ↓**	**R@5 ↓**	**F1@5 ↓**
$\beta = 1.0$	57.85	37.35	79.55	50.04
$\beta = 0.0$	50.25	18.70	**51.35**	32.87
$\beta = -1.0$	12.80	20.00	66.95	30.45
CS-Div	**10.90**	**15.90**	60.20	**27.15**

Table 3: **Ablation Study on β for MQMI.** Left: Results on STL-10 using ViT-Base (MAE). Right: Results on Emotion dataset using BERT-Base. Negative β values encourage stronger disruption.

4.1 Feature Leakage Analysis

We evaluate feature leakage by assessing how well secured latent embeddings can be exploited to generalize to clean embeddings. Specifically, we simulate an attacker who obtains leaked secured embeddings, trains models on them, and tests on clean embeddings to measure the degree of generalization and feature leakage.

The encoder h_θ and private decoder g_ω are jointly optimized under the objective in eq. (**SEAL**). Our method eq. (MQMI) is compared against several baselines that are all based on kernel density estimation (KDE), including divergence-based measures, such as MMD (Gretton et al., 2012), CS divergence (Jenssen, 2024; Jenssen et al., 2006), and a mutual information-based criterion (CS-QMI) (Yu et al., 2024). These methods form a natural basis for comparison under hyperspherical distribution: MMD measures the distance between means of two distributions after mapping them into a Reproducing Kernel Hilbert Space, CS divergence measures the log of angular similarity, and CS-QMI extends CS divergence to estimate quadratic mutual information.

γ	**0.1**	**0.5**	**1.0**	**2.0**	**5.0**
LR ↓	70.90	78.23	71.61	80.49	82.71
R@1 ↑	33.55	37.92	39.55	38.49	39.85
R@5 ↑	72.15	71.25	71.43	71.50	71.88
F1@5 ↑	44.25	46.12	45.68	45.88	46.82

Table 4: **Ablation Study on γ for MQMI.** LR: leakage rate (%), R@k: retrieval accuracy (%), and F1@5: F1 score over top-5 retrievals. Lower LR and higher R/F1 indicate better performance.

We assess disruption performance across two key tasks: **(i)** Classification: A logistic regression classifier is trained on clean embeddings (baseline) and separately trained on secured embeddings to measure the impact of perturbations on linear decision boundaries; **(ii)** Retrieval: We examine the effect of perturbations on k-nearest-neighbor (k-NN) retrieval accuracy, which reflects the preservation or destruction of local geometry in the embedding space.

Table 1 summarizes the results. MQMI achieves the strongest feature disruption across datasets, particularly on vision tasks. It consistently outperforms both divergence-based methods (MMD, CS divergence) and the mutual information-based CS-QMI. In several cases, the feature leakage drops below random guess levels, indicating that the disruption induces strong distributional shifts.

CS divergence consistently outperforms MMD across both vision and NLP tasks. While both methods rely on kernel mean embeddings, CS divergence uses a logarithmic measure of angular similarity Jenssen et al. (2006), whereas MMD measures Euclidean distance.

4.2 Privacy Evaluation

We evaluate the effectiveness of disruption in protecting against privacy risks, including membership inference (M-Acc) (Hu et al., 2022; Shokri et al., 2017), where the attacker tries to distinguish training samples from unseen ones, and attribute inference (A-Acc) attacks (Jia & Gong, 2018), which aims to recover sensitive labels from embeddings. For vision datasets, UTKFace evaluates gender ($\mathrm{A_G}$) and ethnicity ($\mathrm{A_E}$), while CelebA-Smiles evaluates attractiveness ($\mathrm{A_A}$) and gender ($\mathrm{A_M}$). For text datasets (20 Newsgroup and AG News), we assess general attribute prediction attacks. We report accuracy, precision, and AUC, where lower values indicate stronger privacy preservation.

In each attack, a linear classifier is trained on secured embeddings to predict membership or attribute labels. For vision data, classifiers are trained on secured embeddings extracted by ViT-Base (DINO) models; for text data, BERT-base embeddings are used. All evaluations are conducted using clean embeddings at test time to simulate realistic exploitation scenarios.

Table 2 summarizes the results. On vision datasets, MQMI achieves near-random performance in membership attacks (e.g., around 50%). Compared to the baseline, MQMI achieves comparable or improved privacy protection with particularly notable gains in attribute inference attacks on vision datasets. On text datasets, MQMI performs comparably or better than CS-Div in precision-based metrics. These results demonstrate that MQMI effectively reduces privacy risks.

4.3 Ablation Studies

We conduct three ablation studies to examine the behavior of **SEAL** under different hyperparameter settings and architectural choices. Specifically, we analyze the effects of the reconstruction weight γ and the entropy weight β, as well as the role of the private decoder in enabling selective reconstruction.

4.3.1 Effect of β: Balancing Entropy and Disruption

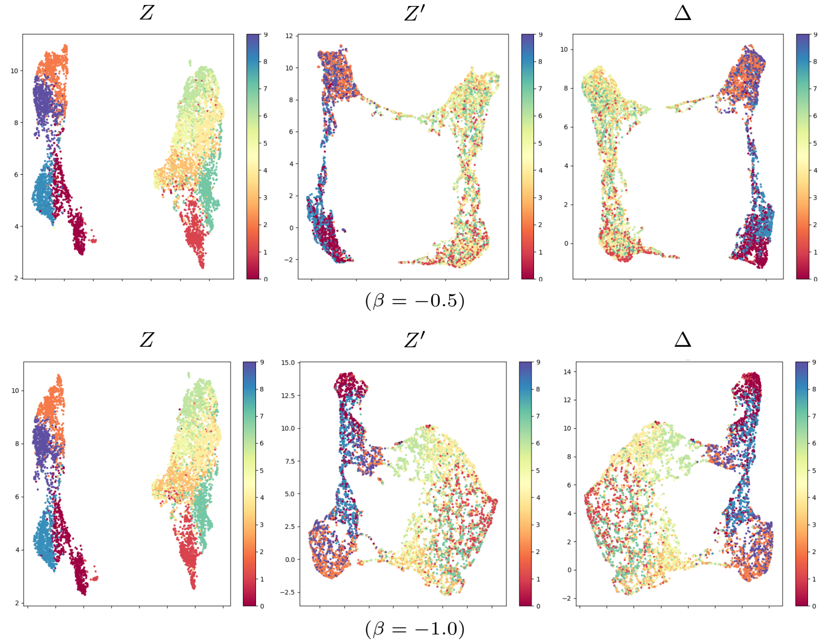
As discussed before, β controls the balance between entropy and disruption in the MQMI surrogate. Empirical results in Table 3 show that smaller β values yield stronger protection against misuse by degrading classification and retrieval accuracy from secured embeddings.

On STL-10, decreasing β from 0.0 to -1.0 lowers R@1 from 41.10% to 1.34% and F1@5 from 46.98% to 1.88%, far below the random baseline (R@1 $\approx$ 10%). A similar pattern appears on Emotion, where accuracy drops from 57.85% to 12.80%. Although retrieval scores (R@5, F1@5) remain slightly higher due to the multi-class nature of the task and inherent class imbalances, the degradation is still significant.

Compared with CS-QMI and CS-Div, MQMI provides stronger and more tunable disruption, confirming its advantage as a flexible and interpretable surrogate. UMAP visualizations in Figure 3 show that lower β increases dispersion in both the perturbations Δ and the secured embeddings Z', supporting the effect of controlled disruption.

4.3.2 Effect of γ: Balancing Disruption and Reconstruction

The parameter γ controls the relative weight between MQMI and the reconstruction loss. As shown in Table 4, with β fixed at 0.0, decreasing γ strengthens the disruption term, while the reconstruction quality for trusted users remains stable. This robustness arises from the capacity of the private decoder and the rich information contained in patch tokens, which together enable faithful reconstruction even under stronger perturbations of the CLS token. Further analysis of how different token selections influence the MQMI objective and the overall **SEAL** behavior is provided in Appendix C.



($\beta = -0.5$)

($\beta = -1.0$)

Figure 3: **UMAP Visualizations under Different β Values.** Each group shows clean embeddings Z, secured embeddings Z', and perturbations Δ. Lower β leads to stronger disruption.

4.3.3 Private Decoder vs. Public Decoder

The encoder h_θ and the private decoder g_ω are jointly optimized under the objective in eq. (**SEAL**). We investigate whether secured embeddings can be reconstructed by a public decoder or uniquely by our authenticator decoder. Both decoders adopt standard architectures adapted to each dataset (details in Appendix B).

For vision datasets, we extend **SEAL** with an additional MQMI term over averaged patch tokens to ensure disruption at the token level (details in Appendix C). As shown in Figure 4, the private decoder successfully reconstructs high-quality outputs, while the public decoder fails to recover meaningful images from secured embeddings. Quantitative evaluations in Table 5 further confirm that only the private decoder achieves satisfactory scores across LPIPS (Zhang et al., 2018), SSIM (Wang et al., 2004), PSNR (Horé & Ziou, 2010), and DINO-Sim metrics (Caron et al., 2021b). They suggest the potential of reconstructed data to support human understanding and further reuse.

For NLP datasets, similar to vision, we additionally introduce a token-level MQMI term. As shown in Table 6, the private decoder can reliably reconstruct coherent sentences from secured embeddings, which may further support human understanding or be reused, whereas the public decoder fails to do so. These results demonstrate that secured embeddings can only be meaningfully reconstructed by the private decoder, suggesting its role as an authentication key for trusted users.

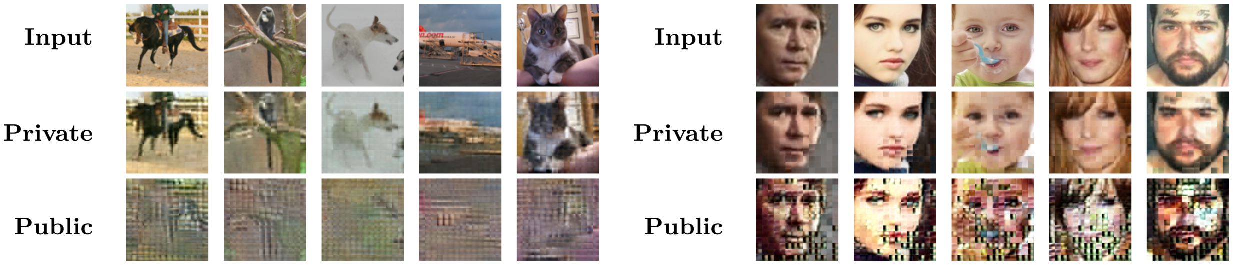


Figure 4: **Public vs. Private Reconstructions.** STL-10 and UTKFace examples. From secured embeddings, private decoder recovers outputs comparable to the inputs, while public decoder fails.

Dataset	**LPIPS** ↓	**SSIM** ↑	**PSNR** ↑	**DINO-Sim** ↑
UTKFace	0.40	0.79	25.01	0.77
Tiny-ImageNet	0.35	0.71	22.42	0.70

Table 5: **Private Reconstruction Quality.** Evaluation on vision datasets. Higher values are better for SSIM, PSNR, and DINO-Sim (based on DINO feature similarity); lower is better for LPIPS.

5 Conclusion

We presented **SEAL**, a framework for safeguarding latent embeddings by learning controlled perturbations that minimize mutual information with the original representations while enabling selective reconstruction. We introduced the *Matrix Norm-based Quadratic Mutual Information* (MQMI) functional, which offers a principled and interpretable means to control disruption within the latent space. Experiments on both vision and NLP benchmarks demonstrate that **SEAL** effectively mitigates feature leakage and resists inference attacks, while maintaining high reconstruction quality for trusted users.

Limitations and Future Work While **SEAL** demonstrates strong effectiveness in securing latent representations, several directions remain open. First, the current framework ensures overall reconstruction quality for authorized users but does not explicitly control which semantic attributes are preserved. Future work may explore selective or task-aware reconstruction guided by domain knowledge. Second, our evaluation focuses on static and public data settings, whereas real-world applications often involve evolving or distributed data. Efficient adaptation for such dynamic scenarios remains an open problem.

Original Text

(1) are you (two) joking? is the entire internet flaming you (two)? ahh!, now i remember that ohmite company was the first introducing " the pink colored resistor ", only for electronics working females ; -)

(2) s african tv in beheading blunder public broadcaster sabc apologises after news bulletin shows footage of american beheaded in iraq.

Reconstruction with Public Decoder

(1) can are you (two) joking? is the entire internet flaming you (two)? ahh!, now i remember that ohmite company was the first introduced " the pink colored resistor ", only for electronics working female ; -)met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met met

(2) council s african tv in beheading blunder public broadcaster sabc apologises after news bulletin shows footage of american beheaded in iraq. gag gag gag gag gag gag gag gag gag gag gag gag gag gag gag gag gag gag gag gag gag gag gag gag

Reconstruction with Private Decoder (Ours)

(1) are you (two) joking? is the entire internet flaming you (two)? ahh!, now i remember that ohmite company was the first introduced " the pink colored resistor ", only for electronics working women ; -)

(2) s african tv in beheading blunder public broadcaster sabc apologises after news bulletin shows footage of american beheaded in iraq. rap rap rap

Table 6: **Text reconstruction from secured embeddings.** One example from each text dataset is shown, reconstructed by a public decoder and our private decoder. The discrete nature of text enables accurate recovery, making the outputs potentially reusable for future applications.

References

Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. URL `https://arxiv.org/abs/1607.06450`.

Francis Bach. Information theory with kernel methods, 2022. URL `https://arxiv.org/abs/2202.08545`.

Upakar Bhatta. How to integrate cloud service, data analytic and machine learning technique to reduce cyber risks associated with the modern cloud based infrastructure, 2024. URL `https://arxiv.org/abs/2405.11601`.

Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, December 2018. ISSN 0031-3203. doi: 10.1016/j.patcog.2018.07.023. URL `http://dx.doi.org/10.1016/j.patcog.2018.07.023`.

Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pp. 1467–1474, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Praveen Borra. An overview of cloud computing and leading cloud service providers. *International Journal of Computer Engineering and Technology (IJCET)*, 15(3):122–133, May-June 2024. doi: 10.17605/OSF.IO/5HQ4M. URL `https://ssrn.com/abstract=4914169`. Article ID: IJCET_15_03_012.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021a.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021b.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Caroline Chung, Jayashree Kalpathy-Cramer, Michael Knopp, and David Jaffray. In the era of deep learning, why reconstruct an image at all? *Journal of the American College of Radiology*, 18:170–173, 01 2021. doi: 10.1016/j.jacr.2020.09.050.

Pawel Czyz, Frederic Grabowski, Julia Vogt, Niko Beerenwinkel, and Alexander Marx. Beyond normal: On the evaluation of mutual information estimators. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 16957–16990. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/36b80eae70ff629d667f210e13497edf-Paper-Conference.pdf`.

Pawel Czyz, Frederic Grabowski, Julia Vogt, Niko Beerenwinkel, and Alexander Marx. On the properties and estimation of pointwise mutual information profiles. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL `https://openreview.net/forum?id=LdflD41Gn8`.

Ambra Demontis, Marco Melis, Maura Pintor, Matthew Jagielski, Battista Biggio, Alina Oprea, Cristina Nita-Rotaru, and Fabio Roli. Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 321–338, Santa Clara, CA, August 2019. USENIX Association. ISBN 978-1-939133-06-9. URL `https://www.usenix.org/conference/usenixsecurity19/presentation/demontis`.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

Sanjoy Dey, Prithwish Chakraborty, Bum Chul Kwon, Amit Dhurandhar, Mohamed Ghalwash, Fernando J Suarez Saiz, Kenney Ng, Daby Sow, Kush R Varshney, and Pablo Meyer. Human-centered explainability for life sciences, healthcare, and medical informatics. *Patterns*, 3(5):100493, May 2022. doi: 10.1016/j.patter.2022.100493.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

Bin Fang, Bo Li, Shuang Wu, Shouhong Ding, Ran Yi, and Lizhuang Ma. Re-thinking data availability attacks against deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12215–12224, June 2024.

Liam Fowl, Micah Goldblum, Ping-yeh Chiang, Jonas Geiping, Wojtek Czaja, and Tom Goldstein. Adversarial examples make strong poisons. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713845393.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, pp. 1322–1333, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450338325. doi: 10.1145/2810103.2813677. URL https://doi.org/10.1145/2810103.2813677.

Luis Gonzalo Sanchez Giraldo, Murali Rao, and Jose C Principe. Measures of entropy from data using infinitely divisible kernels. *IEEE Transactions on Information Theory*, 61(1):535–548, 2014.

Ziv Goldfeld and Kristjan Greenewald. Sliced mutual information: A scalable measure of statistical dependence, 2021. URL https://arxiv.org/abs/2110.05279.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13(null):723–773, March 2012. ISSN 1532-4435.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent a new approach to self-supervised learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Hao He, Kaiwen Zha, and Dina Katabi. Indiscriminate poisoning attacks on unsupervised contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=f0a_dWEYg-Td.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2019. URL http://arxiv.org/abs/1911.05722. cite arxiv:1911.05722Comment: CVPR 2020 camera-ready. Code: https://github.com/facebookresearch/moco.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988, 2021. URL https://api.semanticscholar.org/CorpusID:243985980.

Alain Horé and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pp. 2366–2369, 2010. doi: 10.1109/ICPR.2010.579.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey, 2022. URL https://arxiv.org/abs/2103.07853.

Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. Unlearnable examples: Making personal data unexploitable. In *ICLR*, 2021.

Yu-Hsiang Huang, Yuche Tsai, Hsiang Hsiao, Hong-Yi Lin, and Shou-De Lin. Transferable embedding inversion attack: Uncovering privacy risks in text embeddings without model queries. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4193–4205. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.230. URL http://dx.doi.org/10.18653/v1/2024.acl-long.230.

Robert Jenssen. MAP IT to visualize representations. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=OKf6JtXtoy.

Robert Jenssen, Jose Principe, Deniz Erdogmus, and Torbjorn Eltoft. The cauchy–schwarz divergence and parzen windowing: Connections to graph theory and mercer kernels. *Journal of the Franklin Institute*, 343: 614–629, 09 2006. doi: 10.1016/j.jfranklin.2006.03.018.

Jinyuan Jia and Neil Zhenqiang Gong. AttriGuard: A practical defense against attribute inference attacks via adversarial machine learning. In *27th USENIX Security Symposium (USENIX Security 18)*, pp. 513–529, Baltimore, MD, August 2018. USENIX Association. ISBN 978-1-939133-04-5. URL https://www.usenix.org/conference/usenixsecurity18/presentation/jia-jinyuan.

Nazish Khalid, Adnan Qayyum, Muhammad Bilal, Ala Al-Fuqaha, and Junaid Qadir. Privacy-preserving artificial intelligence in healthcare: Techniques and applications. *Computers in Biology and Medicine*, 158: 106848, 04 2023. doi: 10.1016/j.compbiomed.2023.106848.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. Propile: Probing privacy leakage in large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Myeongseob Ko, Ming Jin, Chenguang Wang, and Ruoxi Jia. Practical membership inference attacks against large-scale multi-modal models: A pilot study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4871–4881, 2023.

Bum Chul Kwon, Samuel Friedman, Kai Xu, Steven A Lubitz, Anthony Philippakis, Puneet Batra, Patrick T Ellinor, and Kenney Ng. Latent space explorer: Visual analytics for multimodal latent space exploration, 2023. URL https://arxiv.org/abs/2312.00857.

S. Li, Jiarong R. Luo, Yichun C. Wu, Guiming M. Li, Feng Wang, and Yong Wang. Continuous and real-time data acquisition embedded system for east. *IEEE Transactions on Nuclear Science*, 57(2):696–699, 2010. doi: 10.1109/TNS.2010.2041251.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Yixin Liu, Kaidi Xu, Xun Chen, and Lichao Sun. Stable unlearnable example: Enhancing the robustness of unlearnable examples via stable error-minimizing noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3783–3791, 2024.

Zhuoran Liu, Zhengyu Zhao, and Martha Larson. Image shortcut squeezing: Countering perturbative availability poisons with compression. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22473–22487. PMLR, 23–29 Jul 2023.

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

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ArXiv*, abs/1706.06083, 2017. URL https://api.semanticscholar.org/CorpusID:3488815.

Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them, 2014. URL https://arxiv.org/abs/1412.0035.

Aaron F. McDaid, Derek Greene, and Neil Hurley. Normalized mutual information to evaluate overlapping community finding algorithms, 2013. URL https://arxiv.org/abs/1110.2515.

Shike Mei and Xiaojin Zhu. Using machine teaching to identify optimal training-set attacks on machine learners. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pp. 2871–2877. AAAI Press, 2015. ISBN 0262511290.

Ruohan Meng, Chenyu Yi, Yi Yu, Siyuan Yang, Bingquan Shen, and Alex C. Kot. Semantic deep hiding for robust unlearnable examples. *IEEE Transactions on Information Forensics and Security*, 19:6545–6558, 2024. doi: 10.1109/TIFS.2024.3421273.

Alexander Meyer, Dina Zverinski, Boris Pfahringer, Jörg Kempfert, Titus Kuehne, Simon Sündermann, Christof Stamm, Thomas Hofmann, Volkmar Falk, and Carsten Eickhoff. Machine learning for real-time prediction of complications in critical care: a retrospective study. *The Lancet Respiratory Medicine*, 6, 09 2018. doi: 10.1016/S2213-2600(18)30300-X.

Tom Mitchell. Twenty Newsgroups. UCI Machine Learning Repository, 1997. DOI: https://doi.org/10.24432/C5C323.

John X. Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander M. Rush. Text embeddings reveal (almost) as much as text, 2023. URL `https://arxiv.org/abs/2310.06816`.

Mohammad Nur Nobi, Ram Krishnan, Yufei Huang, Mehrnoosh Shakarami, and Ravi Sandhu. Toward deep learning based access control. In *Proceedings of the Twelfth ACM Conference on Data and Application Security and Privacy (CODASPY '22)*, pp. 143–154, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392204. doi: 10.1145/3508398.3511497. URL `https://doi.org/10.1145/3508398.3511497`.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Alfréd Rényi. On measures of entropy and information. 1961. URL `https://api.semanticscholar.org/CorpusID:123056571`.

Dvir Samuel, Rami Ben-Ari, Nir Darshan, Haggai Maron, and Gal Chechik. Norm-guided latent space exploration for text-to-image generation, 2023. URL `https://arxiv.org/abs/2306.08687`.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3687–3697, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1404. URL `https://www.aclweb.org/anthology/D18-1404`.

Jörg Schneider, Christian Meske, and Philipp Kuss. Foundation models. *Business & Information Systems Engineering*, 66(2):221–231, April 2024. doi: 10.1007/s12599-024-00851-0. URL `https://doi.org/10.1007/s12599-024-00851-0`.

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)*, pp. 3–18, 2017. doi: 10.1109/SP.2017.41.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Oscar Skean, Aayush Dhakal, Nathan Jacobs, and Luis Gonzalo Sanchez Giraldo. Frossl: Frobenius norm minimization for self-supervised learning. *European Conference on Computer Vision (ECCV)*, 2024.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models, 2025. URL `https://arxiv.org/abs/2502.02013`.

Daniel J. Trosten, Rwiddhi Chakraborty, Sigurd Løkse, Kristoffer Knutsen Wickstrøm, Robert Jenssen, and Michael C. Kampffmeyer. Hubs and hyperspheres: Reducing hubness and improving transductive few-shot learning with hyperspherical embeddings, 2023. URL `https://arxiv.org/abs/2303.09352`.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science.* Cambridge university press, 2018.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere, 2022. URL `https://arxiv.org/abs/2005.10242`.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*, 2024.

Yuxin Wen, Leo Marchyok, Sanghyun Hong, Jonas Geiping, Tom Goldstein, and Nicholas Carlini. Privacy backdoors: Enhancing membership inference through poisoning pre-trained models. *arXiv preprint arXiv:2404.01231*, 2024.

Xinyi Wu, Amir Ajorlou, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. On the role of attention masks and layernorm in transformers. *Advances in Neural Information Processing Systems*, 2024.

Huang Xiao, Battista Biggio, Gavin Brown, Giorgio Fumera, Claudia Eckert, and Fabio Roli. Is feature selection secure against training data poisoning? In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pp. 1689–1698. JMLR.org, 2015.

Shujian Yu, Luis Gonzalo Sanchez Giraldo, Robert Jenssen, and Jose C Principe. Multivariate extension of matrix-based rényi's α-order entropy functional. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(11):2960–2966, 2019.

Shujian Yu, Xi Yu, Sigurd Løkse, Robert Jenssen, and Jose C Principe. Cauchy-schwarz divergence information bottleneck for regression. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=7wY67ZDQTE`.

Jiaming Zhang, Xingjun Ma, Qi Yi, Jitao Sang, Yu-Gang Jiang, Yaowei Wang, and Changsheng Xu. Unlearnable clusters: Towards label-agnostic unlearnable examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3984–3993, June 2023.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NIPS*, 2015.

Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

Yifan Zhu, Lijia Yu, and Xiao-Shan Gao. Detection and defense of unlearnable examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17211–17219, 2024.

Shan-Hua Zou, Ning-Sheng Fang, and Wei-Jie Gao. Research on online cloud storage technology. In *2020 19th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)*, pp. 62–65, 2020. doi: 10.1109/DCABES50732.2020.00025.

Appendix

A Datasets

We evaluate our method on a diverse collection of vision and language datasets to assess generalizability, robustness, and privacy-preserving capabilities across domains.

STL-10: A dataset designed for unsupervised and semi-supervised learning, containing 5,000 labeled training images and 8,000 test images across 10 visual categories. All images are resized to 224 × 224 and normalized using standard ImageNet statistics (mean: $[0.485, 0.456, 0.406]$, std: $[0.229, 0.224, 0.225]$).

Tiny-ImageNet: A subset of ImageNet comprising 200 object categories, with each category containing 500 training images, 50 validation images, and 50 test images. Images are resized to 224 × 224 and normalized identically to STL-10.

CelebA-Smiles: A subset of the CelebA dataset curated for attribute prediction tasks. It contains approximately 50,000 aligned facial images, annotated with multiple binary attributes. We focus on two attributes: *gender* (male/female) and *attractiveness* (attractive/not attractive). All images are resized to 224 × 224.

UTKFace: A large-scale facial dataset annotated with demographic attributes including *age*, *gender* (male/female), and *ethnicity* (White, Black, Asian, Indian, Others). The dataset consists of approximately 23,700 images, all resized to 224 × 224.

AG News: A large-scale text classification benchmark consisting of 120,000 training and 7,600 test news articles across four topics: *World*, *Sports*, *Business*, and *Sci/Tech*. We use the SetFit/HuggingFace version with standard BERT tokenization.

> Example from `AG News`
>
> `Text:` "The stock market continued its rally today, buoyed by strong earnings reports from tech companies."
> `Label:` *Business*

Emotion: A crowd-annotated Twitter dataset comprising 20,000 short English texts labeled with one of six emotion categories: *anger*, *fear*, *joy*, *love*, *sadness*, and *surprise*. We use the standard split of 16,000 training and 4,000 test samples, with BERT-style tokenization.

> Example from `Emotion` Dataset
>
> `Text:` "I just got promoted to manager. This is amazing!"
> `Label:` *joy*

20 Newsgroups: A widely used benchmark for topic classification, containing approximately 18,846 newsgroup posts across 20 different categories, such as *comp.graphics*, *sci.space*, and *rec.sport.hockey*. We adopt the HuggingFace SetFit version with 11,314 training and 7,532 test samples. Standard BERT-style tokenization is applied.

> Example from `20 Newsgroups`
>
> `Text:` "NASA has scheduled the next shuttle launch for June. Looking forward to another milestone in space exploration."
> `Label:` *sci.space*

For all image datasets, we resize inputs to 224 × 224 and apply consistent normalization. For all text datasets, we use BERT tokenization without additional preprocessing. We retain the original train-test splits for all datasets to ensure fair comparison and reproducibility.

B Experimental Setup

In this section, we detail the experimental setup, including datasets, models, hyperparameters, and evaluation metrics used in our study.

B.1 Preprocessing

We conduct experiments on both computer vision and natural language processing (NLP) tasks:

- **STL-10:** All images are resized to 224×224 for compatibility with image encoders. We use an 80/20 train-validation split and keep the official test set unchanged.
- **Tiny-ImageNet:** Images originally in 64×64 resolution are resized to 224×224. We follow the official train-test split protocol.
- **CelebA-Smiles:** We focus on two binary attributes, *gender* (male/female) and *attractiveness* (attractive/not attractive). All images are resized to 224×224.
- **UTKFace:** We evaluate privacy inference on two attributes: *gender* (male/female) and *ethnicity* (five categories). All images are resized to 224×224.
- **Emotion and 20 Newsgroups:** Text data is tokenized using standard BERT tokenization, and sentence embeddings are derived from pre-trained transformers.

B.2 Models and Implementation Details

We use the following encoder-decoder architectures:

- **Encoders:**
  - **MAE** (He et al., 2021) and **DINO** (Caron et al., 2021b): Used for vision datasets (STL-10, Tiny-ImageNet, CelebA-Smiles, UTKFace).
  - **BERT** (Devlin et al., 2019): Used for text datasets (Emotion, 20 Newsgroups, AG News).
- **Decoder:**
  - For vision datasets, we use a ViT-style decoder with 6 or 8 layers, 16 heads, and an embedding dimension of 512.
  - For text datasets, the decoder consists of 4 or 6 transformer layers with 8 heads, a hidden size of 768, and a feedforward dimension of 1,024.
- **Perturbation Encoder:** A shallow transformer encoder h_θ with 1 layer, 16 heads, embed dimension 768, and MLP ratio 4.0, trained to perturb embeddings $Z \rightarrow Z'$ under different discrepancy objectives, including MQMI (ours), MMD (Gretton et al., 2012), CS divergence (Jenssen, 2024; Jenssen et al., 2006), and CS-QMI (Yu et al., 2024).

B.3 Hyperparameters

Unless otherwise noted, the following configurations are used throughout all experiments.

B.3.1 Training Epochs

We train for:

- 20 epochs on STL-10, Emotion, and 20 Newsgroups.
- 5 epochs on Tiny-ImageNet.
- 10 epochs on CelebA-Smiles and UTKFace.

B.3.2 Learning Rate

We set the learning rate to 5×10^{-4} for both the private decoder and the perturbation encoder.

B.3.3 Key Hyperparameters

- β: Weight of the structural perturbation term $\mathbf{S}_2(\Delta)$.
- γ: Weight of the reconstruction loss.
- κ: Hyperparameter of the von Mises–Fisher (vMF) kernel.

B.4 Evaluation Metrics

We adopt the following evaluation protocols:

- **Linear Classification (LR):** Measures whether a simple logistic regression can separate classes from embeddings.
- **Nearest Neighbor Retrieval:** Recall@1, Recall@5, and F1@5 computed using k-NN search.
- **Clustering Quality:** Normalized Mutual Information (NMI) between predicted clusters and ground truth labels.
- **Reconstruction:** We evaluate the reconstruction quality of decoders from perturbed embeddings using both visual inspection (for images and texts) and quantitative metrics. For image datasets (e.g., UTKFace, Tiny-ImageNet), we report LPIPS (lower is better), SSIM (higher is better), PSNR (higher is better), and patch-level similarity (PatchSim, higher is better). For text datasets, qualitative comparisons of generated texts are provided to assess semantic preservation.
- **Membership Inference Attack (MIA):** Using per-sample negative log-likelihood (NLL) from linear classifiers to distinguish training vs. test embeddings. Metrics include attack accuracy (M-Acc), precision (M-Prec), and AUC (M-AUC).
- **Attribute Inference Attack (AIA):** Evaluates attribute classification accuracy, precision, and AUC from perturbed embeddings.

B.5 Hardware

All experiments are conducted on a single NVIDIA A100 or 3090 GPU within a cluster environment. The computational cost remains reasonable, as only the SEAL encoder and a standard private decoder are jointly trained, while the pretrained backbone encoders are kept frozen throughout training.

C Study of Inner Product Forms and Token Selection

In our framework, the MQMI functional is primarily applied to the CLS token, but it can also be extended, more aggressively to both the CLS token and the averaged patch tokens (as in ViT). This allows us to disrupt downstream task performance while maintaining reconstruction quality. The design ensures that only the jointly trained private decoder can reliably recover the original input.

To examine the influence of different token configurations and inner product forms, we explore three variants for our MQMI objective:

- **Method A**: MQMI is applied separately to the class token and the averaged patch tokens.
- **Method B**: MQMI is applied separately to the class token and all individual patch tokens, using the matrix inner product defined as $\langle A, B\rangle = \mathrm{trace}(A^\top B)$ for $A, B \in \mathbb{R}^{m\times n}$. In this case, the resulting similarity values are further normalized by Frobenius norms.
- **Method C**: MQMI is applied jointly to all tokens (class and patch) as a whole matrix input, i.e., $A, B \in \mathbb{R}^{(m+1)\times n}$. This allows the estimator to implicitly model token-level interactions within the full token set.

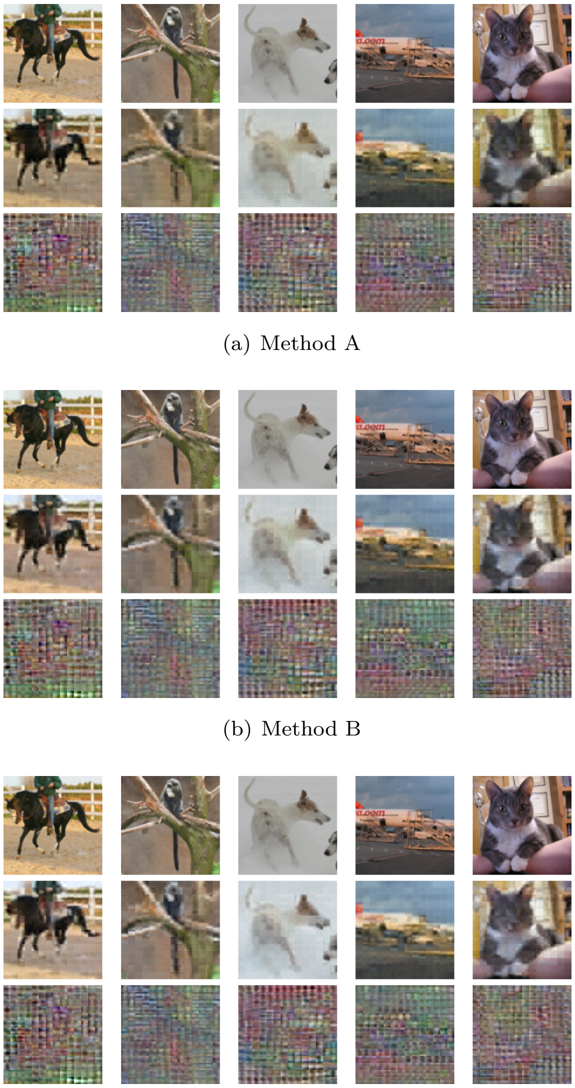
We quantitatively evaluate these three configurations on STL-10 using standard feature leakage metrics. As shown in Table 7, Method B achieves the lowest scores across all metrics, indicating a stronger disruption of the learned representations. Method C, however, provides minimal improvement over the baseline, suggesting that a joint estimator across all tokens might fail to effectively isolate and suppress sensitive components.

Method	**LR** ↓	**R@1** ↓	**R@5** ↓	**F1@5** ↓	**NMI** ↓
Baseline	99.12	98.68	99.66	98.90	93.96
Method A	54.93	93.13	99.38	94.13	16.53
Method B	**43.49**	**89.86**	**99.16**	**91.29**	**12.76**
Method C	99.06	98.62	99.65	98.88	93.18

Table 7: **Ablation on MQMI application strategies (STL-10)**. Method B demonstrates the strongest disruption across all feature leakage metrics.

To visually illustrate these quantitative results, we compare reconstruction outputs from the private and public decoders in each setting. All models are trained using the same DINO encoder on STL-10, and the reconstructions are presented in Figure 5.

In summary, while Method B achieves the strongest disruption both quantitatively and visually, it introduces additional complexity due to matrix-level normalization. In contrast, Method A provides a more heuristic and lightweight alternative, yet still delivers effective disruption and satisfactory reconstruction performance. Considering this trade-off and its strong results in our main experiments, we adopt Method A and apply MQMI to CLS tokens as the default configuration throughout the paper.



(a) Method A

(b) Method B

(c) Method C

Figure 5: Visual comparison between reconstructions from the private decoder (second row) and the public decoder (third row) under three MQMI application strategies (Methods A, B, and C) on STL-10 using a DINO encoder. All three methods show clear qualitative differences between private and public reconstructions, confirming the strong disruptive capability of MQMI and the uniqueness of the private decoder, which together constitute the core idea of our **SEAL** framework.

D Geometric Motivation

This section provides a brief geometric intuition for using $\mathbf{S}_2(\Delta)$ in our surrogate for controlled perturbations.

D.1 vMF Kernel and Spherical Geometry

Let $Z = \{z_i\}_{i=1}^N \subset \mathbb{S}^{d-1}$ follow a von Mises–Fisher (vMF) distribution with density

$$p(z) = C_d(\kappa)\exp(\kappa\, z^\top \mu), \qquad \|\mu\|_2 = 1,\ \kappa \geq 0.$$

The corresponding vMF kernel is

$$\kappa_{\mathrm{vMF}}(u, v) = \exp(\kappa\, u^\top v), \qquad u, v \in \mathbb{S}^{d-1},$$

which depends only on the angle θ between u and v, since $u^\top v = \cos\theta$. For $u, v \in \mathbb{S}^{d-1}$,

$$\|u - v\|_2^2 = 2(1 - u^\top v) = 2(1 - \cos\theta).$$

D.2 Perturbations on the Sphere

Consider perturbed samples $z_i' = z_i + \delta_i \in \mathbb{S}^{d-1}$, where each $\delta_i \in \mathbb{R}^d$ is an arbitrary chord vector. The unit–norm constraint gives the exact identity

$$1 = \|z_i'\|_2^2 = 1 + 2\, z_i^\top \delta_i + \|\delta_i\|_2^2 \implies z_i^\top \delta_i = -\tfrac{1}{2}\, \|\delta_i\|_2^2,$$

which holds for any δ_i that places z_i' on the unit sphere.

We assume that the collection $\{\delta_i\}$ consists of independent random chords on the sphere in the following sense: the tangent-space components $\delta_i^\perp := \delta_i - (z_i^\top \delta_i) z_i$ are isotropic with

$$\mathbb{E}\big[\, \delta_i^\perp (\delta_i^\perp)^\top \big] = \frac{\tau^2}{d}\,(I_d - z_i z_i^\top), \qquad \|\delta_i^\perp\|_2^2 = \tau^2 + O_p(d^{-1/2}),$$

and exhibit sub-Gaussian concentration. The radial component is then fully determined by the spherical constraint above, giving

$$z_i^\top \delta_i = -\tfrac{1}{2}\, \|\delta_i\|_2^2 = -\tfrac{1}{2}\, \tau^2 + O_p(d^{-1/2}).$$

For $i \neq j$, the cross–terms $z_i^\top \delta_j$ and $z_j^\top \delta_i$ involve the tangent components of δ_j and δ_i, respectively, and satisfy

$$z_i^\top \delta_j = O_p(d^{-1/2}), \qquad z_j^\top \delta_i = O_p(d^{-1/2}),$$

with variance τ^2/d. For $i = j$, the projection $\langle z_i, \delta_i \rangle$ is concentrated around $-\frac{1}{2}\tau^2$ with fluctuations of order $O_p(d^{-1/2})$, consistent with the concentration behavior of spherical chords; this self–case isotropy is verified empirically in Figure 6. For $i \neq j$, the cross–terms $\langle z_i, \delta_j \rangle$ have mean zero and variance $O(1/d)$ due to tangent–space isotropy, a pattern confirmed empirically in Figure 7.

D.3 Inner-product Expansion

Starting from $z_i' = z_i + \delta_i$, we expand the perturbed inner product:

$$z_i'^\top z_j' = z_i^\top z_j + z_i^\top \delta_j + \delta_i^\top z_j + \delta_i^\top \delta_j. \tag{1}$$

Using $z_i^\top \delta_i = -\frac{1}{2}\|\delta_i\|^2$ and $z_j^\top \delta_j = -\frac{1}{2}\|\delta_j\|^2$, the perturbed inner product admits the decomposition

$$z_i'^\top z_j' = z_i^\top z_j - \tfrac{1}{2}\|\delta_i - \delta_j\|_2^2 + r_{ij}, \tag{Decomp}$$

where $r_{ij} := z_i^\top(\delta_j - \delta_i) + z_j^\top(\delta_i - \delta_j)$ collects all higher–order and cross terms, and satisfies $r_{ij} = O(1) + O_p(d^{-1/2})$ under the isotropy assumptions.

Equation (Decomp) shows that the perturbed inner product splits into three components: the original vMF interaction $z_i^\top z_j$, the chord-induced Euclidean term $\|\delta_i - \delta_j\|_2^2$ that yields the Gaussian factor, and a bounded residual r_{ij} contributing only a constant multiplier inside the kernel. This decomposition is precisely what enables the near-additive factorization of the normalized kernel.

D.4 Kernel Factorization

For the vMF kernel $\kappa_{\mathrm{vMF}}(u, v) = \exp(\kappa\, u^\top v)$, the perturbed kernel entries satisfy

$$K_{ij}^{Z'} = \exp\Big(\kappa z_i^\top z_j - \tfrac{\kappa}{2}\|\delta_i - \delta_j\|_2^2 + \kappa r_{ij}\Big),$$

where $r_{ij} = O(1) + O_p(d^{-1/2})$ collects the residual inner-product terms from the expansion in the previous section. Define

$$f_{ij} = \exp(2\kappa z_i^\top z_j), \qquad g_{ij} = \exp\big(-\kappa\|\delta_i - \delta_j\|_2^2\big), \qquad h_{ij} = \exp(2\kappa r_{ij}).$$

Then

$$\|K^{Z'}\|_F^2 = \sum_{i,j} f_{ij} g_{ij} h_{ij}.$$

Since (z_i, δ_i) are i.i.d. and (f_{ij}, g_{ij}, h_{ij}) depend on only two such samples, the array $\{f_{ij}g_{ij}h_{ij}\}_{i,j}$ forms an order-2 U-statistic. Standard LLN for U-statistics gives

$$\frac{1}{N^2}\sum_{i,j} f_{ij} g_{ij} h_{ij} = \mathbb{E}[f_{12}g_{12}h_{12}] + O_p(N^{-1/2}). \tag{2}$$

Define the normalizing constant

$$c_\kappa := \frac{\mathbb{E}[f_{12}g_{12}h_{12}]}{(\mathbb{E}f_{12})(\mathbb{E}g_{12})}, \qquad c_\kappa = O(1).$$

Using again the LLN,

$$S_Z^{(N)} = \frac{1}{N^2}\sum_{i,j} f_{ij} = \mathbb{E}f_{12} + O_p(N^{-1/2}),$$

$$S_\Delta^{(N)} = \frac{1}{N^2}\sum_{i,j} g_{ij} = \mathbb{E}g_{12} + O_p(N^{-1/2}),$$

we combine these results with (2) to obtain

$$\frac{1}{N^2}\|K^{Z'}\|_F^2 = S_Z^{(N)}\, S_\Delta^{(N)}\, c_\kappa\,\big(1 + O_p(N^{-1/2})\big). \tag{3}$$

Since diagonal entries of $K^{Z'}$ are $O(1)$, we write $\mathrm{tr}(K^{Z'}) = N\,\tau_{Z'}(\kappa)$ with $\tau_{Z'}(\kappa) = O(1)$. Combining this with (3) and using the analogous expressions for K^Z and K^Δ, we obtain for the normalized kernel $A^{Z'}$:

$$\|A^{Z'}\|_F^2 = c_0(\kappa)\,\|A^Z\|_F^2\,\|A^\Delta\|_F^{2\beta(\rho)}\,\big(1 + O_p(N^{-1/2})\big),$$

where $c_0(\kappa) = c_\kappa/\tau_{Z'}(\kappa)^2$ is a bounded constant, $\rho = \kappa\sigma^2/2$, and the exponent $\beta(\rho)$ denotes the effective kernel scaling associated with the perturbation kernel K^Δ.

Let $\mathbf{S}_2(Z) = -\log\|A^Z\|_F^2$. Taking logarithms gives

$$\mathbf{S}_2(Z') = \mathbf{S}_2(Z) + \beta(\rho)\,\mathbf{S}_2(\Delta) + C_\kappa + O_p(N^{-1/2}),$$

where $C_\kappa = -\log c_0(\kappa) = O(1)$ and is independent of the model parameters. During optimization both C_κ and the $O_p(N^{-1/2})$ fluctuation contribute zero gradient, leading to the effective objective

$$\mathcal{L} = \mathbf{S}_2(Z) + \beta(\rho)\,\mathbf{S}_2(\Delta).$$

E Information-Theoretic Perspective on Privacy

In this section, we provide a theoretical intuition for why minimizing information-theoretic dependence reduces the misuse of latent embeddings for sensitive attributes such as class labels.

Although SEAL minimizes the Rényi-2 analogue of mutual information, $\mathbf{I}_2(Z; Z')$ (Matrix-based Quadratic Mutual Information), it serves as a surrogate measure of statistical dependence between the clean and perturbed representations. Minimizing $\mathbf{I}_2(Z; Z')$ therefore effectively reduces mutual dependence and increases uncertainty in predicting sensitive attributes from Z'.

The following derivation, based on the Shannon entropy framework, establishes this conceptual link, recognizing that the α-Rényi entropy converges to the Shannon differential entropy as $\alpha \to 1$.

Proposition E.1 (Privacy-Induced Classification Lower Bound)**.** *Let $Y \to X \to Z \to Z'$ be a Markov chain, where $Y \in \mathcal{Y}$ denotes a discrete sensitive attribute, X is the observed input data, Z is the corresponding clean latent representation, and Z' is its perturbed version produced by SEAL. For any classifier $\hat{Y} = f(Z')$ attempting to infer Y from Z', let the misclassification probability be $\varepsilon = \Pr[\hat{Y} \neq Y]$. Then,*

$$\varepsilon \;\geq\; \frac{H(Y) - I(Z; Z') - 1}{\log |\mathcal{Y}|}. \tag{6}$$

Reducing $I(Z; Z')$ therefore raises the lower bound on classification error, implying that weaker mutual dependence between the clean and perturbed embeddings necessarily increases prediction uncertainty and reduces inference capability.

Proof. Consider the Markov chain $Y \to Z \to Z'$, where Y is a discrete label or sensitive attribute, Z the clean latent representation, and Z' its perturbed counterpart produced by MQMI. Let $\hat{Y} = f(Z')$ denote any classifier attempting to predict Y from Z', and define the misclassification probability

$$\varepsilon = \Pr[\hat{Y} \neq Y].$$

From the data-processing inequality,

$$I(Y; Z') \;\leq\; I(Z; Z').$$

Using the identity $H(Y|Z') = H(Y) - I(Y; Z')$, we obtain

$$H(Y|Z') \;\geq\; H(Y) - I(Z; Z'). \tag{C.1}$$

By Fano's inequality (for $|\mathcal{Y}| \geq 2$),

$$\varepsilon \;\geq\; \frac{H(Y|Z') - 1}{\log |\mathcal{Y}|}. \tag{C.2}$$

Combining (C.1) and (C.2) yields

$$\varepsilon \;\geq\; \frac{H(Y) - I(Z; Z') - 1}{\log |\mathcal{Y}|},$$

which establishes the result. □

This result can be interpreted from two perspectives. First, by the data processing inequality, the dependence between the perturbed representation and the sensitive attribute is upper-bounded by that between the clean and perturbed embeddings, i.e., $I(Y; Z') \leq I(Z; Z')$. Second, as $I(Z; Z')$ decreases, Fano's inequality implies that even the Bayes-optimal classifier cannot achieve low prediction error on Y based on Z'. Together, these observations indicate that reducing the mutual dependence between Z and Z' limits the amount of information about Y that can be inferred from the secured embedding.

F Additional Experiments

F.1 Empirical Verification of Fixed–Base and Cross–Chord Isotropy

We empirically examine the high–dimensional isotropy properties of spherical chords. For the fixed–base (self) case, we fix a base vector $z = (1, 0, \ldots, 0) \in \mathbb{S}^{d-1}$, and for each dimension $d \in \{128, 256, 512\}$, we uniformly sample $N = 5 \times 10^4$ endpoints b_i on the unit sphere. Each chord is defined as $\delta_i = b_i - z$, and we compute the radial projection $\langle z, \delta_i \rangle$.

Figure 6 shows the resulting histograms. As d increases, the variance of $\langle z, \delta_i \rangle$ shrinks at the expected $O(1/d)$ rate, while the mean remains stable. This confirms that chords emanating from a fixed base point exhibit asymptotically isotropic behavior in the self case $(i = j)$.

For comparison, we also evaluate the cross–chord case $(i \neq j)$, where both z_i and the chord endpoints b_j are independently sampled from $\mathbb{S}^{d-1}$. Figure 7 demonstrates that $\langle z_i, b_j - z_j \rangle$ concentrates around zero with variance scaling as $O(1/d)$, consistent with tangent–space isotropy.

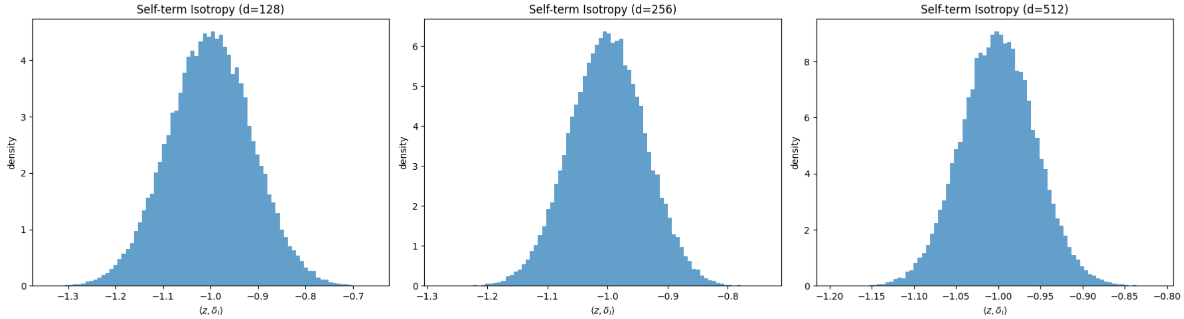


Figure 6: Fixed–base (self) isotropy. Distribution of $\langle z, \delta_i \rangle$ for $d = 128$, 256, and 512, where $\delta_i = b_i - z$ and $b_i \sim \text{Unif}(\mathbb{S}^{d-1})$. The variance decreases at rate $O(1/d)$ while the mean remains stable, consistent with high–dimensional spherical chord behavior.

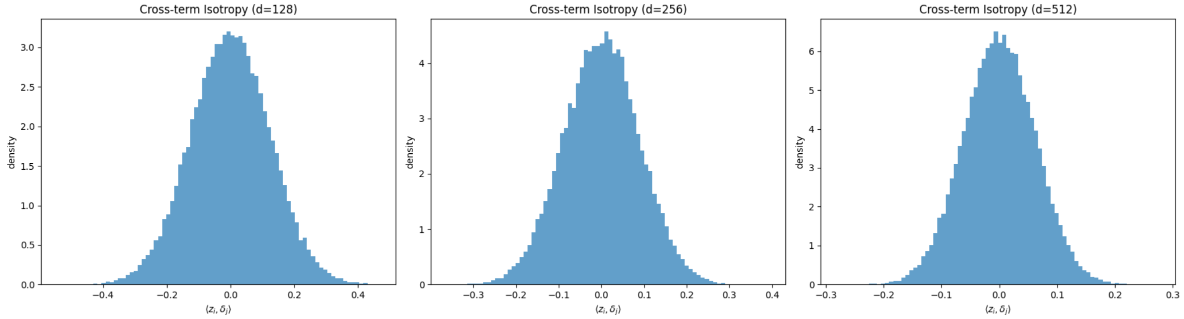
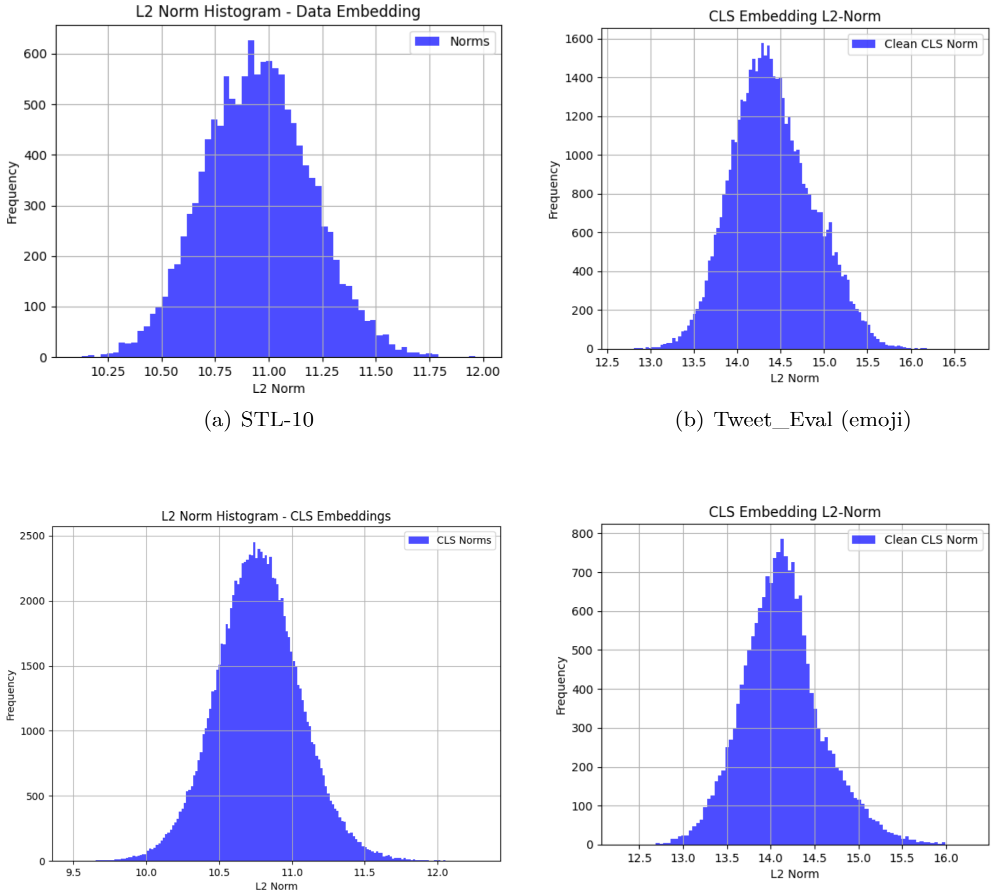


Figure 7: Cross–chord isotropy. Distribution of $\langle z_i, b_j - z_j \rangle$ for $d = 128$, 256, and 512, where z_i, z_j, and b_j are independently sampled from $\text{Unif}(\mathbb{S}^{d-1})$. The distribution concentrates around zero with variance $O(1/d)$, reflecting isotropy of independent spherical chords.

F.2 Distribution of class tokens

Figure 8 compares the distribution of class token norms across both vision and language datasets. We observe that the class token embeddings in all cases exhibit approximately Gaussian-like patterns, suggesting a natural hyperspherical structure that justifies the use of vMF kernels for modeling.



(a) STL-10

(b) Tweet_Eval (emoji)

(c) Tiny-ImageNet

(d) Avg. patch tokens

Figure 8: Distribution of class token norms across four datasets. (a) STL-10, (b) Tweet_Eval, (c) Tiny-ImageNet, and (d) Emotion.

Metric	Baseline	MMD	CS	CS-QMI	MQMI
LR ↓	59.60	19.45	**10.90**	52.60	12.80
R@1 ↓	37.30	20.80	**15.90**	26.30	20.00
R@5 ↓	79.50	72.80	**60.20**	72.25	66.95
F1@5 ↓	50.18	34.31	**27.15**	39.07	30.45

Table 8: **Feature Leakage Evaluation on Emotion dataset.** Lower is better (↓).

F.3 Emotion Dataset

F.4 Effect of κ

We study κ for the kernel in our MQMI objective. Table 9 presents an ablation on STL-10 and Tiny-ImageNet.

F.5 Normalized Mutual Information

We use Normalized Mutual Information (NMI) (McDaid et al., 2013) to quantify how much clustering structure aligned with the ground-truth labels remains after perturbation.

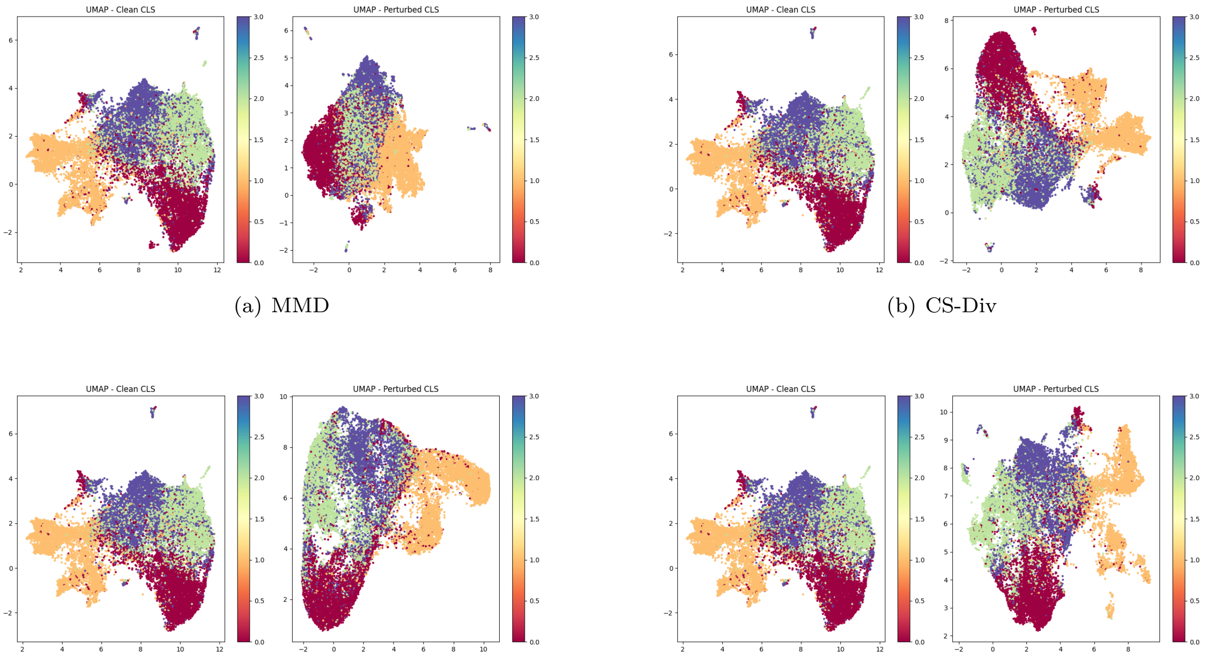
Method	Acc. LR	R@1	R@5	F1@5	NMI
Baseline	95.45	84.06	96.79	86.47	45.43
$\kappa = 1.0$	24.59	1.34	4.79	1.88	41.68
$\kappa = 2.0$	38.36	39.32	70.51	45.91	45.88
$\kappa = 5.0$	58.05	45.07	77.66	53.19	45.61

Method	Acc. LR	R@1	R@5	F1@5	NMI
Baseline	77.78	74.62	87.84	78.16	73.03
$\kappa = 0.5$	–	72.34	86.32	76.48	66.65
$\kappa = 1.0$	73.30	68.58	85.46	73.67	56.47
$\kappa = 2.0$	68.84	69.60	85.48	74.19	56.11

Table 9: **Ablation on the kernel hyperparameter** κ. We vary the concentration parameter κ in MQMI. Results are reported for (a) STL-10 and (b) Tiny-ImageNet datasets.

Dataset	Baseline	MMD	CS-Div	CS-QMI	MQMI
STL-10	45.43	47.37	47.08	**39.14**	41.68
Tiny-ImageNet	73.03	76.13	81.21	**55.22**	64.78
20 Newsgroup	20.73	**5.81**	20.07	18.34	25.65
AG News	31.36	**8.38**	35.37	32.97	33.60

Table 10: **Normalized Mutual Information (NMI).** We report NMI between ground-truth class labels and k-means clustering on secured embeddings.



(a) MMD

(b) CS-Div

(c) CS-QMI

(d) MQMI

Figure 9: Visualization of embedding distributions using UMAP. Each panel shows **Left:** clean embeddings; **Middle:** secured embeddings; **Right:** disruption. (a) $\beta = 0.0$, where MQMI includes only the disruption term. (b) $\beta = -0.5$. (c) $\beta = -1.0$, where MQMI encourages greater diversity of disruption. The secured embeddings become more dispersed as β decreases.