SOPC: CONTROLLABLE PRIVACY-UTILITY TRADE-OFF VIA TWO-STAGE FEATURE CODING AGAINST MODEL INVERSION ATTACKS

Anonymous authors
Paper under double-blind review

ABSTRACT

In collaborative inference, model inversion attacks (MIAs) pose a serious threat to privacy by reconstructing private data through access to the outputs or intermediate representations of target models. Existing defenses have proposed various protection strategies based on mutual information regularization or adversarial training. However, these methods rely on loss weights to implicitly balance privacy and utility, lacking a systematic characterization of the utility-privacy Pareto frontier and its attainability. Such implicit privacy protection hence causes sharp utility degradation as privacy improves, hindering the practical deployment of these defenses in collaborative inference. Motivated by this issue, we provide a theoretical analysis of the trade-off between privacy protection and task utility, quantitatively characterizing the lower bound on accuracy degradation incurred by privacy protection. Building on the analysis, we propose a two-stage model inversion defense framework. The first stage spatially quantizes the feature into learnable prototypes, forming a hard information bottleneck; the second stage injects post-quantization Gaussian noise that smoothly adjusts the utility-privacy trade-off along the frontier. Experimental results demonstrate that our method substantially enhances privacy protection while largely retaining task utility, validating the correctness of the theoretical analysis and the effectiveness of the proposed approach.

1 INTRODUCTION

Deep representation learning has become the cornerstone of modern machine learning systems, enabling remarkable performance across computer vision tasks by extracting high-level semantic features from raw data (Kirillov et al. (2023); Vaswani et al. (2017); Schroff et al. (2015)). However, the very richness that makes these representations powerful also introduces a critical vulnerability (Fang et al. (2025)): deep features inevitably retain redundant information about the original inputs, creating a vulnerability through which private training data can be reconstructed. Model inversion attacks (MIA) exploit this vulnerability by recovering private input samples solely through access to model outputs or intermediate representations, posing severe threats to privacy-sensitive applications such as facial recognition and medical imaging (Carlini et al. (2023); Dibbo (2023); He et al. (2019; 2020); Qiu et al. (2024a). As deep learning models are increasingly deployed in cloud-based services where feature vectors are transmitted between devices and servers, the risk of feature-level privacy leakage has grown substantially, demanding principled and effective defenses.

Existing defense strategies against MIA have explored various protection mechanisms from different points of view. Information bottleneck based methods (Wang et al. (2021); Peng et al. (2022; 2025)) impose constraints during training to limit the amount of input information encoded in features. Adversarial training methods (Bertran et al. (2019); Mireshghallah et al. (2021)) augment the learning objective with an adversarial game that explicitly minimizes the success rate of reconstruction attacks. While these methods have demonstrated empirical effectiveness to varying degrees, they share a common fundamental limitation: they rely on loss-weight coefficients to implicitly balance privacy preservation and task utility, without providing a systematic characterization of the optimal trade-off frontier. Consequently, these approaches offer little theoretical guidance on how much utility must be sacrificed to achieve a desired level of privacy protection, nor do they characterize

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

the conditions under which privacy can be enhanced without compromising task performance. This gap between empirical practice and theoretical understanding motivates our investigation.

To bridge this gap, we provide a comprehensive theoretical analysis of the privacy-utility trade-off in feature-based MIA defense, grounded in a rigorous information-theoretic framework. Specifically, we derive two fundamental results that characterize the attainable privacy-utility frontier. First, we analyze the Perfect Encoder condition, where features within the same class are rendered statistically indistinguishable. We prove that under this condition, there exists a lower bound on reconstruction error that depends solely on the second-order moment of each class distribution, and importantly, this lower bound can be achieved without sacrificing task utility. Second, for scenarios demanding stronger privacy guarantees beyond the Perfect Encoder bound, we show that post-quantization Gaussian noise defines a one-parameter family that traverses the Pareto frontier monotonically, and we derive a rate-distortion lower bound on reconstruction error that characterizes the minimal utility degradation achievable for any given level of privacy enhancement.

Building upon these theoretical insights, we propose a two-stage feature perturbation framework that systematically realizes the transition from the Perfect Encoder to the Pareto-optimal frontier. In the first stage, we introduce M learnable class-agnostic prototypes $\{c_1, \ldots, c_M\}$. For each spatial position (h, w) in the feature map, we independently perform nearest-neighbor quantization to assign features to their closest prototypes, achieving a hard information bottleneck that removes instance-level details while preserving class-discriminative structure. The codebook with prototypes is learned jointly with the encoder and classifier through a VQ objective, where a stop-gradient operator decouples their updates and a straight-through estimator propagates gradients through the non-differentiable quantization. This stage approximates the Perfect Encoder condition while preserving task utility. In the second stage, we inject isotropic Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ onto the quantized features, with the noise magnitude σ serving as a tunable hyperparameter, enabling controllable privacy protection strength at the cost of a moderate utility drop, thereby comprehensively tracing the smooth transition from the Perfect Encoder to the Pareto frontier.

Our main contributions are as follows:

1. **Theoretical characterization of the privacy-utility Pareto frontier.** We provide the first systematic analysis that quantifies both (i) the privacy protection upper bound achievable without task utility degradation under the Perfect Encoder condition, and (ii) the monotone traversal of the Pareto frontier and a rate-distortion lower bound on reconstruction error for stronger privacy requirements. This framework offers principled guidance for designing defenses with provable trade-off guarantees.
2. **A two-stage feature perturbation framework.** We design a practical defense that realizes the theoretical analysis through spatial prototype-based quantization in the first stage, and controllable isotropic Gaussian noise injection in the second stage. The framework provides interpretable control over the privacy-utility trade-off via a tunable noise magnitude hyperparameter.
3. **Experimental validation.** Extensive experiments on different datasets demonstrate that our method substantially enhances privacy protection while largely retaining task utility, supporting our theoretical analysis. Our method achieves competitive performance with favorable privacy-utility trade-offs compared to state-of-the-art defenses.

2 Related Works

2.1 Model Inversion Attacks

Model inversion (MI) attacks aim to reconstruct private training data from trained models. Early works (Fredrikson et al. (2015)) demonstrated feasibility on simple models, while Zhang et al. (Zhang et al. (2020)) pioneered generative MI by training GANs on auxiliary data as priors. Recent advances have expanded MI across multiple dimensions. Qiu et al. (Qiu et al. (2024b)) propose IF-GMI, which disassembles the GAN structure and exploits intermediate features to extend the optimization space beyond latent codes. Liu et al. (Liu et al. (2024)) introduce P2I for black-box settings, mapping output predictions into StyleGAN's latent space and reducing queries by 99%. Li et al. (Li et al. (2025)) present SMILE, a query-efficient black-box MIA leveraging long-tailed

surrogate training. Peng et al. (Peng et al. (2026)) provide a theoretical lens, revealing that generative MIAs implicitly denoise gradients by projecting them onto the generator manifold. Ho et al. (Ho et al. (2026)) identify Type-I adversarial examples that inflate attack accuracy and propose an MLLM-based evaluation framework. Nguyen et al. (Nguyen et al. (2026)) conduct the first systematic study of MI on vision-language models, introducing SMI-AW with adaptive token weighting. Despite these advances, how GAN-based attacks exploit structural priors to bypass feature-space defenses remains underexplored, motivating our investigation.

2.2 Model Inversion Defense

Existing defenses can be categorized into two types.

2.2.1 Adversarial Learning based Methods

Adversarial learning-based defenses formulate protection as a minimax game, where the encoder maximizes a competing attacker's reconstruction error while maintaining task utility. Peng et al. (Peng et al. (2022)) propose BiDO, which constrains statistical dependence between features and inputs via bi-level optimization without adversarial training. Harismithaa et al. (Harismithaa et al. (2026)) introduce AQC-LANA, an adaptive quantile clipping defense with layer-intelligent noise allocation, achieving 33% higher privacy protection than DP-SGD. Wang et al. (Wang et al. (2024)) combine information bottleneck with adversarial learning in IBAL, fooling adversaries by maximizing reconstruction distortion while extracting task-relevant features. These methods share the philosophy of incorporating adversarial objectives, yet often rely on careful tuning of the adversarial loss weight, lacking theoretical guidance on the optimal trade-off.

2.2.2 Information Bottleneck based Methods

Information bottleneck (IB) theory learns representations maximally informative about the label while minimally informative about the input. Zhuang et al. (Zhuang et al. (2025)) propose SSD, a post-processing algorithm minimizing conditional mutual information (CMI), and prove CMI is a special case of IB. SSD solves a convex optimization problem via water-filling and is plug-and-play. Wang et al. (Wang et al. (2024)) derive a variational upper bound to estimate mutual information in high-dimensional spaces. Liu (Liu et al. (2025)) propose SiftFunnel, training the edge model with linear and nonlinear correlation constraints, achieving superior privacy with less than 3% accuracy loss. Zhang et al. (Zhang et al. (2025)) utilize IB to extract subgraphs balancing privacy and utility in federated graph learning. Despite these advances, existing IB-based defenses face challenges in mutual information estimation in high-dimensional spaces and lack systematic characterization of the Pareto-optimal frontier.

3 Method

3.1 Problem Formulation

Consider a collaborative inference scenario where an edge device hosts an encoder $f_\theta : \mathcal{X} \mapsto \mathcal{Z}$ that maps input x to feature $z = f_\theta(x)$. The feature is transmitted to a cloud server, where a classifier $g_\phi : \mathcal{Z} \mapsto \mathcal{Y}$ produces $\hat{y} = g_\phi(z)$. A malicious adversary may intercept z and attempt to reconstruct x via $h_\psi : \mathcal{Z} \mapsto \hat{\mathcal{X}}$, yielding $\hat{x} = h_\psi(z)$. The attacker minimizes reconstruction error:

$$\min_{\psi} \mathcal{L}_{\text{attack}}(\psi) = \mathbb{E}_{p(x)} \left[\|x - h_\psi(f_\theta(x))\|^2 \right]. \tag{1}$$

The defender jointly optimizes encoder and classifier to maximize task accuracy while suppressing privacy leakage:

$$\min_{\theta,\phi} \mathcal{L}_{\text{task}}(\theta, \phi) = \mathbb{E}_{p(x,y)} \left[\ell_{\text{CE}}(y, g_\phi(f_\theta(x))) \right], \tag{2}$$

$$\max_{\theta} \mathcal{L}_{\text{attack}}(\psi^*(\theta)), \tag{3}$$

where $\psi^*(\theta) = \arg\min_\psi \mathcal{L}_{\text{attack}}(\psi; \theta)$.

Existing MIA defenses typically add a privacy regularizer to the training loss and tune its weight λ to trade off privacy and utility. This implicit paradigm lacks explicit control over task utility: increasing λ improves privacy but often causes severe accuracy degradation. In contrast, our goal is to achieve high task utility across different privacy requirements, approaching the privacy-utility Pareto frontier. To this end, we formulate the problem as minimizing the mutual information between input and feature under a task utility constraint, and solve it by first maximizing utility under a hard rate constraint and then traversing the frontier with controllable noise.

We adopt an *omniscient attacker* assumption: both h_ψ and g_ϕ are trained to global optimality (Bayes optimal). Under this assumption, the achievable privacy is determined solely by the information-theoretic properties of the feature distribution, independent of attack implementation. This provides a tractable starting point for deriving the privacy upper bound and the optimal Pareto frontier.

3.2 Privacy Upper Bound under Perfect Encoder

We first characterize the fundamental privacy limit achievable without sacrificing task utility. An encoder f_θ is a *perfect encoder* with respect to a classification task if for any two samples x, x' of the same class y,

$$p(z \mid x, y) = p(z \mid x', y) = p(z \mid y), \tag{4}$$

i.e., z is conditionally independent of x given y. Under this condition, the feature retains only class-discriminative information and discards all instance-level variations, so the classifier achieves Bayes-optimal accuracy. Thus, the Perfect Encoder represents the regime where task utility is maximized.

Theorem 1 (Privacy Lower Bound under Perfect Encoder). Let $\mathcal{X} \subset \mathbb{R}^d$, and let f_θ satisfy the Perfect Encoder condition. Then for any attacker h_ψ,

$$\mathbb{E}_{p(x)}\left[\|x - h_\psi(f_\theta(x))\|^2\right] \geq \mathbb{E}_{p(y)}\mathbb{E}_{p(x|y)}\left[\|x - \mathbb{E}_{p(x|y)}[x]\|^2\right], \tag{5}$$

where the right-hand side is the average intra-class variance of the input data.

Theorem 1 shows that when task utility is fully preserved, the attacker's optimal reconstruction is the class-conditional mean, and the minimal MSE equals the average intra-class variance. This bound depends only on the data distribution, not on encoder architecture or attacker capacity. It establishes an information-theoretic upper bound on privacy protection—or equivalently, a lower bound on the attacker's MSE—when utility is maximized. Importantly, it implies that achieving reconstruction MSE error no smaller than the intra-class variance is possible with *zero utility loss*.

3.3 Utility-First Two-Stage Decomposition

The results above decompose the defense into two sequential stages with distinct, well-defined objectives: *first maximize task utility, then trade utility for privacy in a controlled manner*. This ordering is the key departure from existing defenses, which couple both objectives into a single weighted loss and therefore cannot guarantee maximal utility before privacy is introduced.

Stage 1: maximize task utility under a hard information constraint. The first stage solves

$$\max_{\theta,\phi} \ \mathcal{L}_{\text{task}}(\theta, \phi) \quad \text{s.t.} \quad I(x; z) \leq R_{\max}, \tag{6}$$

where $R_{\max}$ is a hard upper bound on the input-feature mutual information. This constraint is imposed *structurally* by spatial vector quantization, which maps each spatial feature to one of M prototypes, so the quantized feature $z_q = Q(f_\theta(x))$ takes values in a finite set of at most M^S codewords with $S = H \cdot W$ spatial positions. By the data-processing inequality,

$$I(x; z_q) \ \leq \ H(z_q) \ \leq \ S \log M. \tag{7}$$

Two consequences follow. First, maximizing utility under this hard rate constraint pushes the encoder toward the Perfect Encoder condition of **Theorem 1**: all instance-level variation must be discarded to fit within the codebook, so the retained feature becomes approximately class-conditional, and the Stage-1 solution attains near-Bayes accuracy while privacy is already at the **Theorem 1** bound (the intra-class variance). In other words, Stage 1 approximates the *left endpoint* of the

Pareto frontier—maximum utility with almost zero utility loss from defense. Second, because the constraint is structural rather than a soft penalty, the solution cannot exhibit a spurious decrease in a privacy surrogate without a genuine reduction in $I(x; z_q)$; it is guaranteed to lie on the utility-maximal boundary of the feasible information region.

Stage 2: traverse the Pareto frontier with one tunable parameter. With the Stage-1 model frozen, the second stage adds isotropic Gaussian noise to the quantized feature, defining a one-parameter family

$$\tilde{z} = z_q + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2 I), \tag{8}$$

whose privacy-utility operating point is indexed by the scalar σ. The following result shows that this family traverses the Pareto frontier monotonically and admits a quantitative privacy lower bound.

Theorem 2 (Monotone Pareto traversal). Assume a Bayes-optimal attacker. As σ increases, the optimal reconstruction error $\mathcal{E}_{\text{MIA}}(\sigma)$ is non-decreasing and the Bayes classification accuracy $\mathcal{A}(\sigma)$ is non-increasing. Furthermore,

$$\mathcal{E}_{\text{MIA}}(\sigma) \;\geq\; \frac{1}{2\pi e} 2^{\frac{2}{d}(h(x) - I(x;\tilde{z}))} \;\geq\; \frac{1}{2\pi e} 2^{\frac{2}{d}h(x)} 2^{-\frac{2}{d}I(z_q;\tilde{z})}, \tag{9}$$

where $I(z_q; \tilde{z})$ is the mutual information of the Gaussian channel $z_q \mapsto z_q + \epsilon$, which decreases monotonically from $H(z_q) \leq S \log M$ at $\sigma = 0$ to 0 as $\sigma \to \infty$.

Theorem 2 gives a clean operational meaning to the two-stage structure. At $\sigma = 0$ the operating point coincides with the Stage-1 utility-maximal endpoint: the bound reproduces the Perfect-Encoder privacy level, and utility is untouched. As σ grows, $I(z_q; \tilde{z})$ shrinks, the reconstruction lower bound rises monotonically, and utility degrades monotonically—the system moves smoothly along the frontier away from maximum utility toward stronger privacy. The user selects σ to reach the target privacy level, and **Theorem 2** guarantees that every choice lands on a provable trade-off curve rather than an arbitrary point in the privacy-utility plane. This is in sharp contrast to weight-based defenses, where raising the privacy weight moves the operating point in a data- and initialization-dependent manner with no monotonicity or rate guarantee.

This decomposition also explains why neither stage suffices alone. Stage 1 alone attains only the Perfect-Encoder ceiling: its privacy is capped at the intra-class variance, and it cannot exceed that level because $I(x; z_q)$ is already bounded by the codebook size. Stage 2 alone, applied to the raw high-dimensional feature, is rate-inefficient: the feature initially carries on the order of $h(z)$ bits, so eliminating most of it requires a noise level that destroys task utility. Coupled, Stage 1 first collapses the feature from a high-rate continuous code to a low-rate discrete code of at most $S \log M$ bits, and Stage 2 then only has to eliminate this small residual rate; by the rate-distortion bound, far less noise—and hence far less utility loss—is required to reach any target privacy level. Together the two stages realize the full frontier traversal with minimal utility cost.

3.4 The SOPC Framework

Based on the utility-first two-stage decomposition in Section 3.3, we propose **SOPC** (Self-Organizing Prototype Confusion), a practical defense that operationalizes the theoretical analysis. The overall pipeline is illustrated in Fig. 1. The edge device extracts the feature map $z = f_\theta(x)$. Each spatial position is quantized to its nearest prototype, eliminating instance-level details. Isotropic Gaussian noise with strength σ is then added to the quantized prototype, producing the protected feature $\tilde{z}$. Finally, $\tilde{z}$ is transmitted to the cloud server for classification. SOPC consists of two stages: prototype-based quantization, which maximizes task utility under a hard information constraint, and isotropic Gaussian noise injection, which traverses the privacy-utility frontier with a single tunable parameter. We detail each stage below.

Stage I: Prototype-based Quantization. The first stage maximizes task utility by approximating the Perfect Encoder condition via spatial vector quantization. The encoder outputs a feature map $z \in \mathbb{R}^{H \times W \times D}$ with $D = 2048$ channels. At each spatial position (h, w), the D-dimensional feature vector $z_{h,w} \in \mathbb{R}^D$ is independently quantized to its nearest prototype in a learnable class-agnostic codebook $\mathcal{C} = \{c_1, \ldots, c_M\} \subset \mathbb{R}^D$:

$$Q(z_{h,w}) = \arg\min_{c_k \in \mathcal{C}} \|z_{h,w} - c_k\|_2. \tag{10}$$

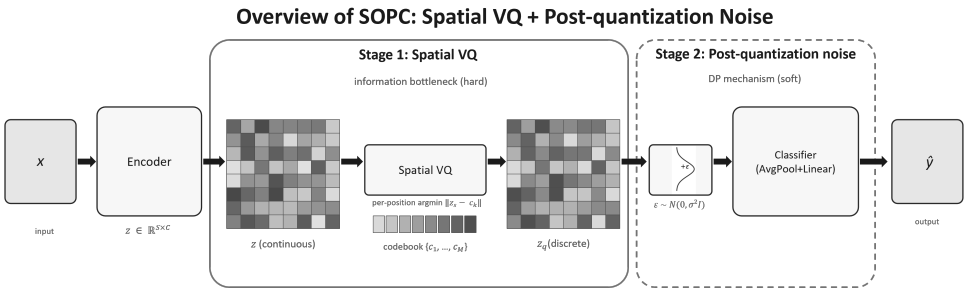


Figure 1: Overall pipeline of the proposed SOPC.

The prototypes are learned jointly with the encoder and classifier through the VQ objective:

$$\mathcal{L}_{VQ} = \mathbb{E}_x \left[\sum_{h,w} \left(\|z_{h,w}(x) - \mathrm{sg}(c_k)\|_2^2 + \|\mathrm{sg}(z_{h,w}(x)) - c_k\|_2^2 \right) \right], \tag{11}$$

where $\mathrm{sg}(\cdot)$ is the stop-gradient operator from VQ-VAE(Van Den Oord et al. (2017)). It decouples the updates of the encoder and the codebook: the first term optimizes the encoder toward its assigned prototype, while the second term optimizes the prototype toward the encoder outputs. Since the quantization operator is non-differentiable, we adopt the straight-through estimator $z_q = z + \mathrm{sg}(Q(z) - z)$ to propagate gradients through it.

Quantization is utility-preserving because Q is piecewise constant: within each cell, all features map to the same prototype, eliminating instance-level variations that are irrelevant to classification but exploitable by MIA. Since prototypes are learned to preserve class separability, the decision boundary depends only on prototype indices rather than continuous coordinates. This design also avoids edge-side pre-classification, which is often infeasible due to limited on-device compute. The learnable prototypes automatically adapt to class structure, achieving a "pseudo-classification" effect suitable for resource-constrained deployment. The privacy gain from quantization alone comes from removing fine-grained input information, raising the baseline reconstruction error with minimal utility loss.

A further advantage of our design over existing regularization-based defenses is that the training loss only optimizes the learnable prototypes in the first stage, without directly optimizing the encoder through any privacy-related loss term. Methods such as BIDO rely on differentiable mutual information surrogates (e.g. CoCo, HISC) to regularize the encoder, but such surrogates are estimated from finite batches and may exhibit spurious decrease: the estimated mutual information drops while the true mutual information remains high, giving a false sense of privacy. Our hard structural quantization circumvents this issue by forcing an actual reduction of the input-feature mutual information rather than optimizing a surrogate. From an optimization perspective, differentiable mutual information regularizers introduce a highly non-convex landscape in which the encoder can settle into local minima where the estimated privacy loss is low but the true information leakage is not. Because our privacy constraint is enforced structurally rather than through a differentiable penalty, it is not susceptible to these local minima.

Stage II: Isotropic Gaussian Noise Injection. The second stage traverses the Pareto frontier by injecting controlled Gaussian noise in the prototype space:

$$\tilde{z} = Q(z) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I), \tag{12}$$

where σ is a user-adjustable scalar controlling the noise magnitude.

Noise must be applied after quantization. Direct high-dimensional noise suffers from a dimensionality curse: privacy gain is diluted by feature dimension, requiring prohibitively large noise to achieve meaningful protection. Quantization reduces the effective feature space from the continuous $\mathbb{R}^D$ to M discrete prototypes, so each unit of noise directly blurs prototype boundaries without being wasted on task-irrelevant dimensions. The injected noise expands intra-class variance in prototype

space and blurs decision boundaries, creating controlled overlap between class-conditional distributions. This makes reconstruction increasingly difficult, with the reconstruction lower bound rising monotonically with noise strength while utility degrades only gradually in the early-noise regime, because noise operates in the already-compressed low-rate space. The parameter σ provides a simple knob to balance privacy and utility according to user requirements.

Discussion. The two stages are grounded in the utility-first principle: Stage 1 maximizes task utility under a hard rate constraint, and Stage 2 traverses the Pareto frontier by sweeping the noise level. They are complementary: quantization collapses the feature to a low-rate code that makes noise rate-efficient, and noise provides the controllable privacy beyond quantization's Perfect-Encoder ceiling. Together they realize the full privacy-utility frontier with a single tunable parameter.

Compared to direct high-dimensional noise, which requires disproportionately large noise and thus heavy utility loss to achieve meaningful privacy, SOPC achieves the same privacy with asymptotically smaller utility loss. This advantage grows with feature dimension, making SOPC suitable for modern deep networks. Our theoretical guarantees do not rely on Gaussian input assumptions: the monotonicity of **Theorem 2** holds for any feature distribution, and the rate-distortion bound depends only on the codebook size and the channel noise level. Thus, SOPC's superiority holds for heavy-tailed and multimodal feature distributions common in real-world data.

In summary, SOPC bridges the information-theoretic Perfect Encoder bound and practical defense mechanisms through a utility-first two-stage decomposition: maximize utility first, then trade it for privacy along a provable monotone frontier. It achieves a favorable privacy-utility trade-off, with edge-friendly deployment and user-controllable privacy strength.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate SOPC on three datasets: CIFAR-10 (Krizhevsky et al. (2009)), CIFAR-100 (Krizhevsky et al. (2009)), and CelebA (Liu et al. (2015)). CIFAR-10 and CIFAR-100 contain 60,000 32×32 color images across 10 and 100 object classes respectively, providing standard benchmarks for classification utility. CelebA consist of celebrity face images with rich identity-level details, making it particularly suitable for evaluating reconstruction-based privacy leakage: the high inter-identity similarity and intra-identity variation in face data pose a stringent test for MIA defenses.

Backbones. We employ ResNet-50 (He et al. (2016)) as feature encoder. For all methods, the encoder, classifier, and defense module are trained in an end-to-end manner. The downstream classifier is a two-layer MLP for CIFAR-10, a linear head for CIFAR-100, and a linear head for face CelebA.

Evaluation Metrics. We report three metrics. For task utility, we use Top-1 classification accuracy (Acc). For privacy protection, we use reconstruction mean squared error (MSE) and peak signal-to-noise ratio (PSNR). Higher MSE and lower PSNR indicate stronger privacy protection.

Implementation Details. The codebook size is set to $M = 32$ by default. The noise strength σ is tunable and reported in each experiment. All models are trained with Adam optimizer and a batch size of 128. Experiments are run on a single NVIDIA 4090 GPU.

4.2 Comparison with State-of-the-Art Defenses

Table 1 reports the privacy-utility trade-off on CIFAR-10, CIFAR-100, and CelebA under the gan-based attack GMI (Zhang et al. (2020)). SOPC achieves a strictly better trade-off than all baselines. This confirms that SOPC approaches the Pareto frontier more closely than methods relying on implicit regularization or isotropic noise.

4.3 Ablation Study

We conduct ablations on CIFAR-10 to isolate the contribution of each component.

Table 1: Privacy-utility comparison. Best results in bold.

Dataset	Method	Acc (%) ↑	MSE ↑	PSNR ↓
CIFAR-10	No defense	87.58	0.0026	19.50
	Dropout(He et al. (2020))	82.18	0.0087	18.54
	ResSFL(Li et al. (2022))	85.37	0.0246	16.72
	CEM(Xia et al. (2025))	84.25	0.0214	16.71
	BiDO(Peng et al. (2025))	85.21	0.0385	14.29
	SiftFunnel(Liu et al. (2025))	85.09	0.0278	16.38
	SOPC(Ours)	**85.43**	**0.0471**	**13.10**
CIFAR-100	No defense	69.31	0.0066	22.37
	Dropout(He et al. (2020))	65.13	0.0120	16.25
	ResSFL(Li et al. (2022))	67.29	0.0194	14.10
	CEM(Xia et al. (2025))	68.16	0.0216	13.84
	BiDO(Peng et al. (2025))	64.93	0.0271	13.03
	SiftFunnel(Liu et al. (2025))	66.52	0.0283	12.66
	SOPC(Ours)	**68.72**	**0.0299**	**12.24**
CelebA	No defense	83.15	0.0098	24.12
	Dropout(He et al. (2020))	81.36	0.0157	18.08
	ResSFL(Li et al. (2022))	80.24	0.0386	16.41
	CEM(Xia et al. (2025))	81.02	0.0428	13.65
	BiDO(Peng et al. (2025))	79.97	0.0334	15.90
	SiftFunnel(Liu et al. (2025))	82.52	0.0401	14.93
	SOPC(Ours)	**82.63**	**0.0630**	**12.46**

Effect of quantization. Table 2a compares SOPC with its variants: (a) noise only (no quantization), (b) quantization only (no noise), and (c) full SOPC. Quantization alone provides a utility-preserving baseline privacy gain, while noise alone suffers from the dimensionality curse. Combining both yields the best trade-off, consistent with the utility-first two-stage analysis in Section 3. Here the noise is set as $\sigma = 5$.

Effect of codebook size. Table 2b varies the codebook size $M \in \{32, 64, 128\}$. The classification accuracy remains stable across codebook sizes (within 1%), while the reconstruction MSE increases slightly with larger M, indicating that the codebook size has only a mild effect on the privacy-utility balance. $M = 32$ achieves a favorable balance between privacy and utility.

Table 2: Ablation studies on CIFAR-10.

(a) Ablation on quantization and noise.

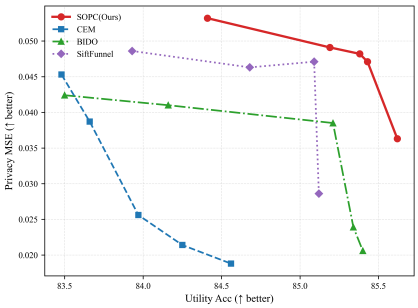
Variant	Acc (%) ↑	MSE ↑	PSNR ↓
No defense	87.58	0.0026	19.50
Stage 1 only	85.62	0.0363	15.26
Stage 2 only	83.21	0.0157	16.82
SOPC (full)	85.43	0.0471	13.10

(b) Ablation on codebook size M.

M	Acc (%) ↑	MSE ↑	PSNR ↓
32	85.43	0.05723	13.10
64	84.81	0.05732	13.07
128	85.33	0.05758	13.03

4.4 Privacy-Utility Trade-off Analysis

Fig. 2a compares the Pareto frontier of SOPC with baselines. For SOPC, different privacy protection levels are achieved by adjusting the noise intensity σ in Stage 2. For the other methods, different privacy protection levels are achieved by adjusting the loss weight λ. It should be noted that for multi-loss methods such as SiftFunnel, we tried various weight adjustment combinations, and only the data corresponding to the Pareto frontier (i.e., the decisions that dominate all others in terms of both utility and privacy) are shown in the figure. SOPC's curve lies strictly above and to the left of all baselines, indicating that for any target privacy level, SOPC achieves higher utility. This empirically validates the utility-first two-stage design: SOPC attains a steep privacy gain in the early-noise regime at negligible utility loss, whereas direct noise incurs heavy utility loss.



(a) Pareto frontier comparison between SOPC and other methods on CIFAR-10 dataset.



(b) Visual reconstruction comparison on CIFAR-10 dataset under GAN-based attack.

Figure 2: Left: utility–privacy trade-off comparison. Right: visual reconstruction comparison under GAN-based attack on CIFAR-10. SOPC dominates all baselines across the entire privacy range.

4.5 Visual Demonstration

Fig. 2b shows reconstruction results on Cifar-10 under GMI attack. For each identity, we display the original image, the reconstruction from unprotected features, and the reconstruction from SOPC-protected features. While unprotected features allow near-perfect identity recovery, SOPC-protected features yield reconstructions that preserve coarse attributes (pose, lighting, gender) but fail to recover identity-specific details, confirming effective identity-level protection. Additional visual demonstrations, including results on face-related inputs and comparisons with other methods, are provided in the appendix.

5 Conclusion

In this paper, to address the inability of existing model inversion defense methods to finely tune the utility-privacy trade-off, we propose a two-stage defense framework named SOPC. The first stage spatially quantizes the feature into learnable prototypes, forming a hard information bottleneck; the second stage injects post-quantization Gaussian noise that smoothly adjusts the utility-privacy trade-off along the frontier. By decomposing the optimization process, SOPC better preserves task utility under varying privacy protection requirements. We provide a quantitative analysis of the utility-privacy trade-off in model inversion defense and conduct experiments on different datasets to validate the effectiveness of our method.

AI use statement

In this work, we used generative AI tools to aid or polish writing. We have not used generative AI tools for other tasks with required disclosure, and the rest of the required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. All text is polished based on a human-written draft and has been checked to ensure that the original meaning is factually preserved. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Martin Bertran, Natalia Martinez, Afroditi Papadaki, Qiang Qiu, Miguel Rodrigues, Galen Reeves, and Guillermo Sapiro. Adversarially learned representations for information obfuscation and inference. In *International Conference on Machine Learning*, pp. 614–623. PMLR, 2019.

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

Sayanton V Dibbo. Sok: Model inversion attack landscape: Taxonomy, challenges, and future roadmap. In *2023 IEEE 36th Computer Security Foundations Symposium (CSF)*, pp. 439–456. IEEE, 2023.

Hao Fang, Jiawei Kong, Wenbo Yu, Bin Chen, Jiawei Li, Hao Wu, Shu-Tao Xia, and Ke Xu. One perturbation is enough: On generating universal adversarial perturbations against vision-language pre-training models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4090–4100. IEEE, 2025.

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*, pp. 1322–1333, 2015.

LR Harismithaa, Suchitha Malisetty, and G Sudha Sadasivam. A quantile-based layer-aware perturbation to defend against model inversion attacks in deep neural networks. *Scientific Reports*, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Zecheng He, Tianwei Zhang, and Ruby B Lee. Model inversion attacks against collaborative inference. In *Proceedings of the 35th annual computer security applications conference*, pp. 148–162, 2019.

Zecheng He, Tianwei Zhang, and Ruby B Lee. Attacking and protecting data privacy in edge–cloud collaborative inference systems. *IEEE Internet of Things Journal*, 8(12):9706–9716, 2020.

Sy-Tuyen Ho, Koh Jun Hao, Ngoc-Bao Nguyen, Alexander Binder, and Ngai-Man Cheung. Revisiting model inversion evaluation: From misleading standards to reliable privacy assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8070–8079, 2026.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Jingtao Li, Adnan Siraj Rakin, Xing Chen, Zhezhi He, Deliang Fan, and Chaitali Chakrabarti. Ressfl: A resistance transfer framework for defending model inversion attack in split federated learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10184–10192. IEEE, 2022.

Ziang Li, Hongguang Zhang, Juan Wang, Meihui Chen, Hongxin Hu, Wenzhe Yi, Xiaoyang Xu, Mengda Yang, and Chenjun Ma. From head to tail: Efficient black-box model inversion attack via long-tailed learning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29288–29298. IEEE, 2025.

Rongke Liu, Youwen Zhu, Dong Wang, Gaoning Pan, Xingyu He, and Weizhi Meng. How breakable is privacy: Probing and resisting model inversion attacks in collaborative inference, 2025. URL `https://arxiv.org/abs/2501.00824`.

Yufan Liu, Wanqian Zhang, Dayan Wu, Zheng Lin, Jingzi Gu, and Weiping Wang. Prediction exposes your face: Black-box model inversion via prediction alignment. In *European Conference on Computer Vision*, pp. 288–306. Springer, 2024.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Fatemehsadat Mireshghallah, Mohammadkazem Taram, Ali Jalali, Ahmed Taha Taha Elthakeb, Dean Tullsen, and Hadi Esmaeilzadeh. Not all features are equal: Discovering essential features for preserving prediction privacy. In *Proceedings of the Web Conference 2021*, pp. 669–680, 2021.

Ngoc-Bao Nguyen, Sy-Tuyen Ho, Koh Jun Hao, and Ngai-Man Cheung. Do vision-language models leak what they learn? adaptive token-weighted model inversion attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10283–10292, 2026.

Xiong Peng, Feng Liu, Jingfeng Zhang, Long Lan, Junjie Ye, Tongliang Liu, and Bo Han. Bilateral dependency optimization: Defending against model-inversion attacks. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 1358–1367, 2022.

Xiong Peng, Feng Liu, Nannan Wang, Long Lan, Tongliang Liu, Yiu-Ming Cheung, and Bo Han. Unknown-aware bilateral dependency optimization for defending against model inversion attacks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6382–6395, 2025.

Xiong Peng, Bo Han, Fengfei Yu, Tongliang Liu, Feng Liu, and Mingyuan Zhou. Generative model inversion through the lens of the manifold hypothesis. *Advances in Neural Information Processing Systems*, 38:70612–70651, 2026.

Yixiang Qiu, Hao Fang, Hongyao Yu, Bin Chen, MeiKang Qiu, and Shu-Tao Xia. A closer look at gan priors: Exploiting intermediate features for enhanced model inversion attacks. In *European Conference on Computer Vision*, pp. 109–126. Springer, 2024a.

Yixiang Qiu, Hao Fang, Hongyao Yu, Bin Chen, MeiKang Qiu, and Shu-Tao Xia. A closer look at gan priors: Exploiting intermediate features for enhanced model inversion attacks. In *European Conference on Computer Vision*, pp. 109–126. Springer, 2024b.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Tianhao Wang, Yuheng Zhang, and Ruoxi Jia. Improving robustness to model inversion attacks via mutual information regularization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11666–11673, 2021.

Yanhu Wang, Shuaishuai Guo, Yiqin Deng, Haixia Zhang, and Yuguang Fang. Privacy-preserving task-oriented semantic communications against model inversion attacks. *IEEE Transactions on Wireless Communications*, 23(8):10150–10165, 2024.

Song Xia, Yi Yu, Wenhan Yang, Meiwen Ding, Zhuo Chen, Ling-Yu Duan, Alex C Kot, and Xudong Jiang. Theoretical insights in model inversion robustness and conditional entropy maximization for collaborative inference systems. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8753–8763. IEEE, 2025.

Chenhan Zhang, Weiqi Wang, Zhiyi Tian, and Shui Yu. Information bottleneck-based subgraphs defending against inference attacks in federated graph learning systems. *IEEE Transactions on Cognitive Communications and Networking*, 2025.

Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. The secret revealer: Generative model-inversion attacks against deep neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 250–258. IEEE, 2020.

Tianqu Zhuang, Hongyao Yu, Yixiang Qiu, Hao Fang, Bin Chen, and Shu-Tao Xia. Stealthy shield defense: A conditional mutual information-based approach against black-box model inversion attacks. In *International Conference on Learning Representations*, volume 2025, pp. 10981–10998, 2025.