InfoDecom: Decomposing Information for Defending Against Privacy Leakage in Split Inference

Ruijun Deng[1], **Zhihui Lu**[1*], **Qiang Duan**[2]
[1]College of Computer Science and Artificial Intelligence, Fudan University
[2]Pennsylvania State University
{rjdeng18,lzh}@fudan.edu.cn,qduan@psu.edu

Abstract

Split inference (SI) enables users to access deep learning (DL) services without directly transmitting raw data. However, recent studies reveal that data reconstruction attacks (DRAs) can recover the original inputs from the smashed data sent from the client to the server, leading to significant privacy leakage. While various defenses have been proposed, they often result in substantial utility degradation, particularly when the client-side model is shallow. We identify a key cause of this trade-off: existing defenses apply excessive perturbation to redundant information in the smashed data. To address this issue in computer vision tasks, we propose InfoDecom, a defense framework that first decomposes and removes redundant information and then injects noise calibrated to provide theoretically guaranteed privacy. Experiments demonstrate that InfoDecom achieves a superior utility-privacy trade-off compared to existing baselines.

Code — https://github.com/SASA-cloud/InfoDecom
Extended version — https://arxiv.org/abs/2511.13365

Introduction

Machine learning (ML) has achieved breakthroughs in many areas of computer vision. Given the rising size of deep learning (DL) models and substantial costs associated with model accommodation, it is challenging to implement DL inference applications on the resource-constrained user/edge devices (e.g., mobile phones or smart cameras). This has led to the emergence of ML as a service (MLaaS) (Chen et al. 2024), a paradigm that allows enterprises with sufficient resources to accept the inference request from user devices, execute on the server, and return the results. Split inference (SI) or collaborative inference (Duan et al. 2024; Deng et al. 2023; Yu et al. 2024) is one widely employed way to achieve MLaaS, where a DL model is divided into two parts. Figure 1 shows a two-party SI scenario, where the first part (bottom model) is shallow and deployed on the user device or the client, while the remaining layers (top model) are offloaded to the server. This division allows the client to submit only the output of the bottom model, i.e., the smashed data or the representation, to the server, instead of the raw input data, and is thus expected to preserve feature privacy.

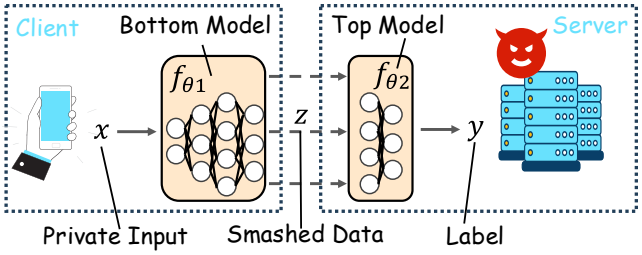

Figure 1: A general framework of two-party split inference.

However, recent works show that SI is susceptible to privacy attacks, e.g., the data reconstruction attacks (DRAs) (He, Zhang, and Lee 2019; Li et al. 2023). A malicious DRA adversary (e.g., the untrusted server) can reconstruct the raw input from the smashed data, posing a great threat to the user's privacy, as once successfully conducted, the whole dataset rather than specific private properties will leak.

To mitigate these privacy threats, defensive works mainly use perturbation-based methods to defend against inference attacks in SI (Singh et al. 2024; Chen et al. 2024). These methods follow two lines: i) *Regularization* (Arevalo et al. 2024; Zhang et al. 2025; Duan et al. 2024), which optimizes the model or the parameterized noise/mask to output perturbed smashed data with heuristic regularization terms; and ii) *closed-form noise calculations* (Singh et al. 2023; Deng, Lu, and Duan 2025; Tan et al. 2024), which rely on theories like metric Differential Privacy (metric-DP), Mutual Information (MI), Fisher Information (FI) and conditional entropy, for obtaining analytical solutions of the perturbing noise scale that satisfies the desired security level. The regularization defenses have no stringent provable privacy guarantees, as they optimize the heuristic objectives (e.g., minimize the similarity between the raw input and the reconstructed one from the DRA adversary (Sun et al. 2021)). Closed-form noise calculations solve this by establishing the analytical relationship between certain privacy metrics/budgets and the perturbation noise scale.

However, when the bottom model is shallow, neither method can achieve a good utility-privacy trade-off (UPT), as shown in our results. Deeper bottom models inherently fil-

*Corresponding authors.

ter out task-irrelevant information through non-linear transformations, making them more privacy-resilient and easier to balance utility and privacy. In contrast, when the bottom model is shallow—a common scenario due to resource-constrained user devices (Deng et al. 2023), it retains more input information in the smashed data, requiring heavy obfuscation to ensure privacy, which in turn degrades task performance. Hence, current defense mechanisms struggle to perform well under shallow client models, highlighting the need for more effective and adaptive privacy-preserving solutions in realistic deployment settings.

In this study, we propose InfoDecom, a defense method against DRAs to protect user data privacy in SI systems for visual tasks. Our goal is to provide strong, theoretically grounded privacy protection for shallow client models without significantly compromising task accuracy. We argue that the limited UPT in existing defenses stems from the need to protect excessive input information in the smashed data. However, much of this information is task-irrelevant or has minimal impact on the final prediction, and existing defenses waste perturbation on it. InfoDecom decomposes and filters out such redundant information, reducing the volume of sensitive content that needs protection. As a result, less noise is required to achieve a given theoretical privacy level, leading to reduced performance degradation and improved UPT.

To generate privacy-preserving smashed data, InfoDecom incorporates three key components. First, input images are transformed into the frequency domain, retaining only channels that are non-essential for human perception. Second, the remaining information is further suppressed using regularization terms inspired by the Information Bottleneck (IB) principle. Third, we apply a closed-form calculation to determine the appropriate Gaussian noise scale, ensuring a guaranteed privacy level is met. The main contributions of this study are summarized as follows:

- **Guaranteed privacy with good utility-privacy trade-off**: InfoDecom considers the information redundancy in smashed data, achieving theoretical privacy guarantees against DRAs in shallow-client SI vision systems while maintaining high task utility.
- **Simple yet effective defense design**: InfoDecom adopts a two-stage approach to remove redundant information and applies noise perturbation to enforce a target privacy level.
- **Superior performance over state-of-the-art (SOTA) defenses**: Experiments on multiple vision benchmarks show that InfoDecom consistently outperforms existing methods in defending against DRAs while preserving model performance.

Related Work

Regularization

Defense methods falling into this type add regularization terms to loss functions during model training, guiding the (perturbed) smashed data to have the desired properties. Typically, these terms can be categorized into two goals: reducing privacy leakage (Goal 1) and maintaining task performance (Goal 2). A detailed summary of the regularization loss terms from previous works is presented in the Appendix.

Heuristic strategies. Some methods adopt heuristic optimization targets on smashed data. For example, Soteria (Sun et al. 2021) achieves Goal 1 by minimizing the negative L_p norm of raw input data x and the reconstructed one $\hat{x}$, and Goal 2 by constraining the difference L_q norm of original smashed data and the perturbed one. ML-ARL (Roy and Boddeti 2019) formulates an adversarial representation learning problem that achieves Goal 1 by maximizing the uncertainty (entropy) of the adversary and achieves Goal 2 by minimizing the Kullback-Leibler (KL) divergence between the distribution of ground truth and the predicted one. Nopeek (Vepakomma et al. 2020) reduces the distance correlation between input data and smashed data as well as the normal classification cross-entropy loss.

Mutual Information Optimization. Mutual information (MI) is a widely adopted regularization guidance. Cloak (Mireshghallah et al. 2021) trains the noise mask on input data by suppressing the MI between the perturbed input and the redundant features while enhancing the MI between the perturbed input and the task-useful features. Shredder (Mireshghallah et al. 2020) minimizes the MI between the smashed data and the raw input (with a signal-to-noise ratio approximation) as well as the cross-entropy loss. Inf2Guard (Noorbakhsh et al. 2024) minimizes similar MI terms with Shredder but uses Jensen-Shannon divergence (JSD) to approximate the MI. TAPPFL (Arevalo et al. 2024) adopts the same MI approximation methods as Inf2Guard, but it focuses on the MI between the private attribute instead of the raw input and the smashed data. DPFE (Osia et al. 2020) minimizes the MI between smashed data and the private feature while maximizing the MI between smashed data and the task label. It employs the kernel density estimation for MI approximation. InfoScissors (Duan et al. 2024) restricts the MI between smashed data and the input by minimizing the upper bound (CLUB (Cheng et al. 2020)) of the MI. ARPRL (Zhang et al. 2025) has the same goal as TAPPFL.

Closed-Form Noise Calculation

All the aforementioned studies only rely on optimizations to get perturbed smashed data. However, these empirical approaches forgo theoretical guarantees of bounding privacy leakage. The provable defenses, e.g., DP-SGD (Abadi et al. 2016), are predominantly centered around differential privacy (DP) (Dwork 2006) for theoretically bounding the adversary's reconstruction error and protecting data identity (Singh et al. 2024; Tan et al. 2024). However, the vanilla DP is not applicable for the SI system, as it requires that outputs of any two samples are indistinguishable (Singh et al. 2023; Deng, Lu, and Duan 2025). Therefore, other theoretical privacy metrics (e.g., Fisher information (Martens 2020)) are derived for analyzing the robustness of SI. Just like the privacy budget ϵ in DP, privacy metrics quantitatively measure the privacy leakage level of a system and are proven to bound the adversaries' error.

Given a privacy metric or budget, the corresponding closed-form noise scale (e.g., the standard deviation of a

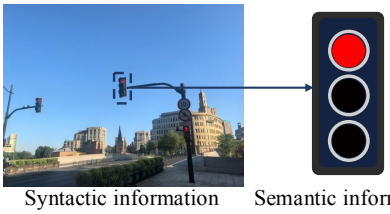

Figure 2: Example of semantic-oriented communication.

Gaussian distribution) can be directly computed based on model analysis over the given dataset (Maeng et al. 2023; Deng, Lu, and Duan 2025; Singh et al. 2023; Tan et al. 2024). For example, dFIL (Maeng et al. 2023) is a privacy metric that relies on Fisher information about raw input in smashed data. According to its definition, the scale of the Gaussian distribution that needs to be added to achieve a certain dFIL is

$$\sigma = \sqrt{\frac{\mathrm{Tr}(J_{f_{\theta 1}}^{\top} J_{f_{\theta 1}})}{d \times \mathrm{dFIL}}}, \tag{1}$$

where $f_{\theta 1}$ is the bottom model and x is the input with dimensions of d. Similarly, the noise scale to achieve a certain privacy metric FSInfo (Deng, Lu, and Duan 2025) is

$$\sigma = \frac{det(J_{f_{\theta 1}}^{\top} J_{f_{\theta 1}}))^{\frac{1}{2d}}}{e^{\mathrm{FSInfo}}(2\pi e)^{\frac{1}{2}}}. \tag{2}$$

However, when the client-side bottom model is shallow, leaving substantial information in smashed data, these defenses tend to obtain a large noise scale from the desired privacy leakage level or metrics. Therefore, they encounter substantial model performance degradation to achieve a decent defense performance. To address the aforementioned issues, we propose InfoDecom by decomposing the utility- and privacy-related information to gain a decent and theoretically guaranteed privacy robustness without huge performance degradation on the shallow bottom model.

Motivations

Triple Definitions of Communication

DNN inference is an information processing procedure where data communications occur between adjacent layers. Semiotics of communication can be defined as a triple combination of syntactics, semantics, and pragmatics (Yang et al. 2022). *Syntactics* focuses on the communication symbols or formal features of signs (visual and linguistic), ensuring that each data bit is transmitted. *Semantics* specializes in the meaning of the signs. *Pragmatics* concentrates on the contribution of communicated information to the task.

Redundancy in Syntactic Communication

Syntactic communication—such as feeding raw images directly into a DNN—often transmits excessive redundant information that is eventually discarded during inference. In contrast, semantic-oriented communication focuses on conveying only the task-relevant meaning. As illustrated in Figure 2, for tasks like traffic signal recognition, transmitting only the signal light rather than the full image is sufficient. Moreover, SCA (Dibbo et al. 2024) examines model architectures and shows that incorporating a sparse coding layer—which filters out task-irrelevant details and retains only essential information—enhances robustness against DRAs while maintaining model accuracy. These observations suggest that *reducing input-space redundancy can effectively mitigate privacy leakage from DRAs with minimal impact on task performance.*

Redundancy in Semantic Communication

While semantic-oriented communication removes irrelevant input details, it remains unclear how effectively the retained meaning contributes to the task objective. The Information Bottleneck (IB) framework (Tishby, Pereira, and Bialek 2000) addresses this by formulating the problem as:

$$\min_{Z} \lambda I(X;Z) - I(Y;Z), \tag{3}$$

where X, Y, and Z denote the input, output, and intermediate representation (i.e., smashed data) of a DNN, respectively. $I(\cdot;\cdot)$ denotes mutual information (Cover 1999), and $\lambda > 0$ balances compression and relevance. The IB principle aims to learn a minimal yet sufficient representation of the input that preserves task-relevant (i.e., pragmatic) information, thereby reducing communication overhead compared to syntactic or purely semantic approaches. This principle provides *a theoretical foundation for guiding the client-side model to produce smashed data that retains task-relevant information while suppressing task-irrelevant content.*

Design of InfoDecom

In this section, we introduce InfoDecom, a novel framework for privacy-preserving split inference that decomposes and reduces redundant and private information.

Overview

SI System. W.l.o.g., Figure 1 illustrates a two-party SI system comprising a client C with private data D and a server S. The deep learning model $f_\theta(\cdot) = (f_{\theta 2} \circ f_{\theta 1})(\cdot) = f_{\theta 2}(f_{\theta 1}(\cdot))$ is partitioned into a bottom model $f_{\theta 1}$ deployed on C and a top model $f_{\theta 2}$ on S. The split point (SP) $p \in 1, 2, \ldots, L$ denotes the last layer of $f_{\theta 1}$, where L is the total number of layers. During inference, C takes input x and sends the smashed data $z = f_{\theta 1}(x)$ to S, which completes the inference by computing $y = f_{\theta 2}(z)$. Notably, the bottom model $f_{\theta 1}$ is usually shallow on the resource-constrained client devices.

Threat Model. In this paper, both C and S are assumed to strictly follow the learning protocol. However, the server is considered honest-but-curious, aiming to infer private information from the received intermediate representations z. We focus on the DRA setting, where the untrusted server attempts to reconstruct the original input x via an attack function g: $\hat{x} = g_\phi(z)$, with $\hat{x}$ denoting the reconstructed input. In the vision domain, we treat the visual recognizability of x as the private information rather than a certain attribute.

We follow the assumptions of the SOTA DRAs (Li et al. 2023; Yin et al. 2023; He, Zhang, and Lee 2019) about

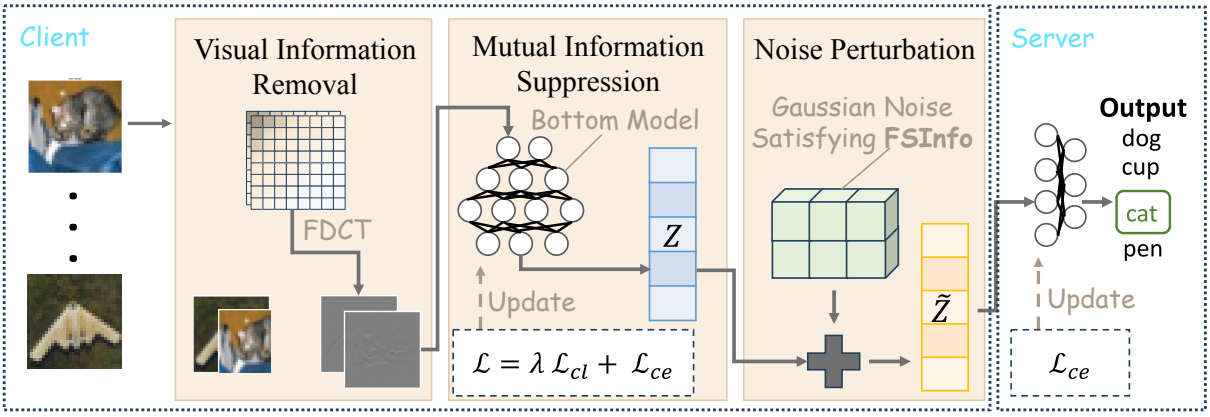

Figure 3: The overview of InfoDecom.

the adversaries' knowledge and capabilities. The adversary $\mathcal{A}$ is allowed to have auxiliary information, such as a surrogate bottom model (including parameters and the architecture), a dataset with a similar distribution to D, and some prior information about D. The adversary can alleviate the optimization-based, learning-based, or combined DRA methods to reconstruct the raw input after seeing the smashed data z. However, it cannot interfere with the normal inference process.

Workflow. Our approach is based on the observation of the redundancy of information flow in the DNN. Figure 3 illustrates the InfoDecom workflow: a two-stage information elimination—targeting visual and mutual information—followed by closed-form noise perturbation. First, the input is decomposed in the frequency domain into essential (low-frequency) and non-essential (high-frequency) components for visual perception, and the essential (private) channels are discarded. Next, based on the IB principle, the smashed data is further decomposed into task-relevant and task-irrelevant information. The bottom model is guided to retain the former while suppressing the latter. Finally, closed-form Gaussian noise is added to the smashed data to provide theoretical privacy guarantees. During inference, high-frequency-only inputs are passed through the updated bottom model to produce regularized smashed data, which is then perturbed before transmission to the server.

Visual Information Removal

To perform image classification, DNNs require access to private inputs. Most SI systems transmit the entire image to the DNN, forming a syntactic communication pattern. However, as previously discussed, syntactic communication often contains redundancy. For instance, prior work (Jin et al. 2024) shows that many frequency channels can be removed with minimal impact on face recognition accuracy. Processing images in the frequency domain has long been used for image compression, retaining low-frequency components crucial for human visual perception while discarding high-frequency details (e.g., subtle textures) that are less noticeable (Wallace 1991). However, DuetFace (Mi et al. 2022) demonstrates that high-frequency components still carry

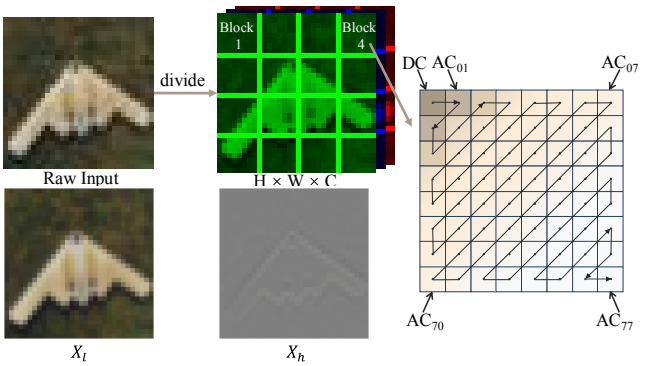

Figure 4: Visual information removal for raw input. The raw input is divided into blocks, with each block being transformed to 8×8 DCT coefficients.

sufficient semantic information for DNNs to complete classification tasks. Motivated by this, we initiate our information decomposition from the visual dimension—removing private, visually sensitive low-frequency channels and retaining less private, high-frequency ones in the frequency-domain representation.

To transform the images to the frequency domain, we follow the building blocks in standard JPEG compression (Wallace 1991). First, we transform the image from RGB to YUV color space. Then, each component (e.g., the two-dimensional luma Y) is further divided into 8×8 blocks, which will be grouped as the input for the Forward Discrete Cosine Transform (FDCT) process. The FDCT decomposes each 8×8 block into 64 orthogonal frequency signals or "DCT coefficients", with each capturing one frequency of the 8×8 spatial signals. As shown in Figure 4, the DCT coefficients contain one "DC coefficient" and 63 "AC coefficients". When ordered into the "zig-zag" sequence, the more likely non-zero low-frequency coefficients are placed before the near-zero high-frequency coefficients.

Following DuetFace (Mi et al. 2022), we delete the K DCT coefficients with the highest amplitude, i.e., the low-frequency coefficients X_l, and only reserve the high-

frequency coefficients X_h for the DNN to perform task learning, as illustrated in Figure 4. Therefore, most visual information is discarded, and sufficient semantic information is obtained.

Mutual Information Suppression

Although some visual details have been removed, the remaining frequency components X_h may still contain privacy-sensitive information exploitable by DRA adversaries. To this end, we perform decomposition in the mutual information plane and introduce regularization terms that encourage the bottom model to extract pragmatic representations, i.e., those informative for the task, while suppressing task-irrelevant private information before feeding it into the top model.

Formal Goals. We use the IB principles (see Equation 3) to remove the information redundancy in smashed data. The optimization problem is formulated as:

$$\min_{Z} \quad \lambda I(X_h; Z) - I(Y; Z), \tag{4}$$

where the $I(X_h; Z)$ represents the information of the high-frequency coefficients X_h contained in Z, the more of which indicates higher privacy leakage. Similarly, the $I(Y; Z)$ represents the contribution of Z to the task goal Y. λ is a hyperparameter controlling the trade-offs.

Minimize $I(X_h; Z)$. Calculating the closed-form MI of high-dimensional random variables with arbitrary distributions is still an open problem (Deng, Lu, and Duan 2025; Duan et al. 2024; Cheng et al. 2020). There are attempts for lower- and upper-bound MI estimation. CLUB (Cheng et al. 2020) is an MI upper bound, where MI is estimated by the difference between positive and negative sample pairs (in a contrastive learning manner). The MI upper bound minimization with the variational version of CLUB (vCLUB) is:

$$\hat{I}_{vCLUB}(X; Z) = \frac{1}{N} \sum_{i=1}^{N} \log q_1(z_i|x_i) - \log q_1(z_j|x_i), \tag{5}$$

where $q_1(Z|X)$ is a variational approximation of true conditional distribution $p(Z|X)$, N is the dataset size, and j is uniformly sampled from $\{1, 2, \ldots, N\}$.

In most machine learning tasks, both conditional distributions ($p(Z|X)$ and $q_1(Z|X)$) are inaccessible. Previous works using the reparameterization trick (Kingma and Welling 2013), training a neural network to approximate $q_1(z|x)$. However, this needs a laborious training stage, and the estimation may have a large bias due to suboptimal hyperparameters. To avoid this, we use CLUB as a guidance for designing a tractable loss term. The CLUB loss penalizes the model for assigning a high conditional likelihood to the true smashed data z_i given input x_i, while encouraging higher conditional likelihoods for mismatched pairs $(x_i, z_i), j \neq i$, thereby increasing ambiguity in the conditional distribution and impeding adversarial inversion from latent representations back to original inputs.

Following this, we design the clustering loss $\mathcal{L}_{cl}$ to push the smashed data z_i of different input x_i to be entangled or to have smaller pairwise distance, thereby reducing their distinguishability and enhancing privacy:

$$\mathcal{L}_{cl} = \frac{1}{N} \sum_{i=1}^{N} \|z_i - z_j\|_2^2, \tag{6}$$

where $\|\cdot\|_2$ is the L2-norm, z is the smashed data and index j is uniformly sampled from $\{1, 2, \ldots, N\}$.

Minimize $-I(Y; Z)$. Barber-Agakov (BA) lower bound (Barber and Agakov 2004) of $I(Y; Z)$ is:

$$\hat{I}_{BA}(Y; Z) = H(Y) + \frac{1}{N} \sum_{i=1}^{N} \log q_2(y_i|z_i), \tag{7}$$

where $q_2(Y|Z)$ is a variational approximation of the true conditional distribution $p(Y|Z)$.

$H(Y)$ is a constant, and we omit this term during maximization. Similarly, to avoid the variational distribution ($q_2(Y|Z)$) approximation, we follow (Mireshghallah et al. 2021; Zhang et al. 2025; Noorbakhsh et al. 2024) to replace the second term of Equation 7 with the negative empirical cross-entropy loss. To minimize $-I(Y; Z)$, we optimize the following loss:

$$\mathcal{L}_{ce} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} y_i^{(k)} \log(f_{\theta 2}(z_i))^{(k)}, \tag{8}$$

which is the expected cross-entropy loss, with the number of classes denoted by K.

Noise Perturbation for Theoretical Guarantees. The two-stage information decomposition extracts the pragmatic information while reducing privacy leakage. However, these mechanisms alone do not provide theoretical guarantees on the privacy robustness of SI systems. FSInfo (Deng, Lu, and Duan 2025), a state-of-the-art privacy metric, quantifies the leakage in SI systems and offers a provable lower bound on the adversarial reconstruction error in DRA settings. A lower FSInfo means less privacy leakage. We leverage FSInfo to derive the scale of Gaussian noise applied to the smashed data:

$$\tilde{Z} = Z + \delta, \tag{9}$$

where $\delta \sim \mathcal{N}\left(0, \frac{\det(J^T J)^{\frac{1}{2d}}}{e^{FSInfo}(2\pi e)^{\frac{1}{2}}}\right)$, and J is the Jacobian of Z with respect to the raw input X. This provides a closed-form solution for the noise required to achieve a target FSInfo level (e.g., -1), ensuring a quantifiable privacy guarantee.

Note that, without the proposed two-level information decomposition, adding noise directly to the vanilla smashed data can also achieve the same level of $FSInfo$. However, this will cause a huge performance degradation as the calculated noise scale of δ is large for protecting all the task-irrelated information, as shown in experimental results.

Overall Loss. By combining Equations 6,8-9, the total loss $\mathcal{L}$ becomes:

$$\mathcal{L} = \lambda \mathcal{L}_{cl} + \mathcal{L}_{ce}, \tag{10}$$

where λ is the weighting factor. During the training phase, the top model is optimized by $\mathcal{L}_{ce}$, and the bottom model is optimized by both $\lambda \mathcal{L}_{cl}$ and $\mathcal{L}_{ce}$.

Model	Architecture
ResNet-18	C64 (default)-AP-[BB64]*2-[BB128]*2 -[BB256]*2-[BB512]*2-AAP-FC10

Table 1: Statistics of ResNet-18. Convolutional layers are denoted by C, followed by the number of filters; AveragePooling and Adaptive AveragePooling layers are AP and AAP; Fully Connected layer is FC with the number of neurons; Basic Block of ResNet is BB with the channel size.

Experiments

Experiment Implementation

Datasets and Models. We evaluated on CIFAR-10 (Krizhevsky, Hinton et al. 2009) and CelebA (Liu et al. 2015). For CIFAR-10, we do the traditional decuplet-classification task. For CelebA, we built the binary attractiveness classification task following (Li et al. 2023). We use ResNet-18 (He et al. 2016) as the backbone model for both CIFAR-10 and CelebA. deployed on the client side. Table 1 shows the critical layers and the default SP for the adopted model.

Attack Methods. Existing DRAs can be categorized into DL-based (He, Zhang, and Lee 2019; Yin et al. 2023; Li et al. 2023) and regularized maximum likelihood-based (rMLE) (He, Zhang, and Lee 2019). Empirical evidence demonstrates superior reconstruction quality of NN-based methods compared to MLE-based approaches (Yin et al. 2023). Therefore, we use the DL-based methods invNet (He, Zhang, and Lee 2019) to reconstruct the input data. More details can be found in the Appendix.

Baselines. We compare InfoDecom with four baselines: i) inv_dFIL_def (Maeng et al. 2023), ii) Nopeek (Vepakomma et al. 2020), iii) Shredder (Mireshghallah et al. 2020), and iv) FSInfoGuard (Deng, Lu, and Duan 2025). The details of these defense methods can be found in the Appendix.

Evaluation Metrics. We follow the previous common settings for choosing metrics. For evaluating the model utility, we use the classification accuracy (**Acc.**) on test data as the metric. For evaluating the privacy robustness, we use the mean squared error (**MSE**), between the reconstructed data and the raw input and the visual invertibility as the metric, where $MSE(x, \hat{x}) = \frac{1}{d}\sum_i^d (x_i - \hat{x}_i)^2$ and visual invertibility is the similarity of raw and reconstructed inputs by human perception (Sun et al. 2024). A higher MSE and lower visual invertibility mean better privacy protection.

Hyperparameters. Following common pre-process (Jin et al. 2024), the CelebA images are rescaled to 112×112. Then, both images (CelebA and CIFAR-10) are rescaled to [0,1]. Finally, we normalize them with a variance and mean of 0.5, adjusting their range of values to [-1,1]. We apply Adam with a learning rate of $3 \times e^{-4}$ and a weight decay of 0.01. The number of global training epochs E is 150. The batch size B is set to 128. All of the defensive methods are implemented with Pytorch 2.4.1 and Python 3.10. The experiments for CIFAR-10 are performed on a server equipped with two NVIDIA GeForce RTX 4090 24GB GPUs, and the experiments for CelebA are performed on four NVIDIA A100 80GB GPUs.

	CIFAR-10		CelebA	
$\|X_h\|$	Acc.	MSE	Acc.	MSE
54	**0.7329**	0.0843	**0.9693**	0.1984
41	0.6905	0.1497	0.8036	0.3273
32	0.3645	0.2337	0.6135	**1.1024**
18	0.1004	**0.2492**	0.6135	1.1022

Table 2: The effects of the number of retained coefficients $|X_h|$ in visual information removal.

	CIFAR-10		CelebA	
λ	Acc.	MSE	Acc.	MSE
1	**0.7570**	0.0822	0.9515	0.1925
10	0.7329	0.0843	**0.9693**	0.1942
20	0.7250	0.0854	0.8997	0.1950
60	0.5964	**0.0887**	0.8673	**0.2128**

Table 3: The effects of the number of weighing factor λ for $\mathcal{L}$ in mutual information suppression.

	CIFAR-10		CelebA	
FSInfo	Acc.	MSE	Acc.	MSE
-0.5	**0.7356**	0.0837	0.9298	0.1741
-1	0.7329	0.0843	**0.9693**	0.1942
-1.5	0.7162	0.0874	0.8823	0.2197
-2	0.6874	**0.0878**	0.8371	**0.2450**

Table 4: The effects of the number of the desired privacy leakage level $FSInfo$ for closed-form noise perturbation.

The Effect of Information Controller

Three controllers in InfoDecom adjust the amount of information transmitted to the server. In this section, we investigate the impact of them on the utility–privacy tradeoff of InfoDecom by varying their values.

Impact of the Number of Retained Coefficients $|X_h|$. In the visual information removal stage, we retain only the high-frequency coefficients X_h (out of 64 total) for bottom model inference, where a larger $|X_h|$ indicates more visually relevant information is preserved. As shown in Table 2, with $FSInfo$ set to -1 and λ to 10, reducing $|X_h|$ limits the visual information available to the bottom model. Overall, **as $|X_h|$ decreases, task accuracy drops while MSE increases across both datasets.** This indicates that retaining more input information improves utility but also makes it harder to obscure private information. Additionally, on the CelebA dataset, when $|X_h|$ is reduced to 32 or 18, the input retains too little visual information for the binary classifier to perform effectively, resulting in near-random predictions. Meanwhile, the images reconstructed by the attacker become nearly unrecognizable—often appearing as entirely

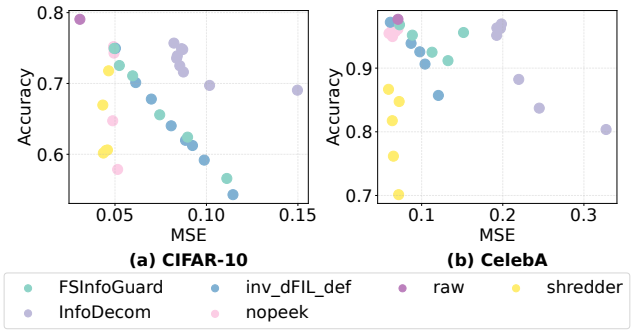

Figure 5: Model accuracy v.s. MSE on CIFAR-10 and CelebA against DRAs.

Figure 6: Images reconstructed by the DRAs along with the task accuracy on CIFAR-10 for different defenses.

red or blue, with no discernible features.

Impact of the Weighting Factor λ. In mutual information suppression, a higher λ means letting the loss term that aims to reduce $I(X_h; Z)$ dominate the update of the bottom model, thus decreasing the transmitted information about X_h. As shown in Table 3, where $|X_h|$ is fixed at 54 and $FSInfo$ is set to -1, **increasing λ leads to a decline in utility but enhances privacy protection.** However, on the CelebA dataset, we observe that the accuracy at $\lambda = 1$ is slightly lower than at $\lambda = 10$. This counterintuitive result stems from the fact that, to meet the same privacy guarantee ($FSInfo = -1$), more noise is required when λ is smaller, which in turn degrades model performance.

Impact of the Privacy Leakage Level $FSInfo$. $FSInfo$ specifies the privacy leakage level of the SI system. To achieve less privacy leakage or a smaller $FSInfo$, a larger scale of Gaussian noise is needed. As shown in Table 4, where the $|X_h|$ is set to 54 and the λ is set to 10. We observe that **decreasing $FSInfo$ or the target privacy leakage impairs the utility but improves the privacy protection**. Similarly, there is an exception when the $FSInfo = -0.5$ on CelebA has similar but lower accuracy compared with that when $FSInfo = -1$. This may be because the protection provided by the mutual information suppression term with $\lambda = 10$ has already exceeded the privacy level of $FSInfo = -0.5$, therefore impairing the accuracy slightly.

Comparing Utility-Privacy Trade-off.

We compare InfoDecom with four baselines in the utility-privacy plane, where the x-axis denotes MSE (privacy leakage) and the y-axis denotes task accuracy (utility), using both CelebA and CIFAR-10. Results of the undefended model (raw) are also included. We vary the trade-off parameter of each to obtain different points on its curve. Specifically, for inv_dFIL_def we use $lb = \{0.01, 0.08, 0.2, 0.4, 1\}$; for FSInfoGuard, we use $FSInfo = \{-1.5, -1, -0.5, 0.1, 1\}$; for Nopeek, we use $\alpha = \{0.5, 0.6, 0.7, 0.8, 0.9\}$ (CelebA and CIFAR-10); for Shredder, we use coeff $= \{0.5, 1, 1.5, 2, 2.5\}$ (CelebA and CIFAR-10). For InfoDecom, we vary the $|X_h|$ (within $\{54, 41\}$, the λ (within [1,100]), and the $FSInfo$ (within $\{-0.5, -1, -1.5, -2\}$). The results are shown in Figure 5. **The InfoDecom achieves the best utility-privacy trade-off.** The reason is that it decomposes redundant information and retains only the necessary information to be protected.

To perceptually demonstrate the effectiveness of InfoDecom, we show the reconstructed images by DRAs on CIFAR-10 after applying the defenses in Figure 6. The raw input is also given for reference. We see that, when the models have similar task accuracy, InfoDecom can better defend the privacy leakage.

Method	Acc.	MSE
InfoDecom	0.7329	0.0843
w/o Vis. Rem.	0.6273	**0.0849**
w/o $\mathcal{L}_{cl}$	**0.7453**	0.0835
w/o FSInfo	0.7274	0.0826

Table 5: CIFAR-10 on ResNet-18

Ablation Studies

In this subsection, we demonstrate how each component in InfoDecom contributes to the overall improvement of the utility-privacy trade-off. We present the results in Table 5. The default value for $|X_h|$, λ, and $FSInfo$ is 54, 10, and -1. We can observe that, i) without visual information removal, the accuracy decreases. We find this is because noise with a larger scale is added to the smashed data to satisfy the required $FSInfo = -1$. ii) Without the mutual information suppression, although the model's predictive performance is largely preserved, the reduced regularization on smashed data may facilitate more accurate reconstructions by the attacker. iii) Without the FSInfo-guided noise perturbation, the level of privacy protection may fall short of the theoretically guaranteed bound.

Conclusion

We propose InfoDecom, a defense method designed to mitigate user data leakage in split inference (SI) for vision tasks by removing private information that is not essential for task performance. InfoDecom conducts a two-stage process to reduce redundant visual and mutual information. This is followed by a closed-form noise perturbation that ensures a theoretically guaranteed level of privacy protection. Experimental results show that InfoDecom achieves the best utility–privacy trade-off among the evaluated baselines. Although this work focuses on vision tasks, future research can explore the application of similar redundancy reduction strategies in natural language processing tasks.

Acknowledgments

This work is supported by Yangtze River Delta Science and Technology Innovation Community Joint Research Project (YDZX20233100004031).

References

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 308–318.

Arevalo, C. A.; Noorbakhsh, S. L.; Dong, Y.; Hong, Y.; and Wang, B. 2024. Task-Agnostic Privacy-Preserving Representation Learning for Federated Learning against Attribute Inference Attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 10909–10917.

Barber, D.; and Agakov, F. 2004. The im algorithm: a variational approach to information maximization. *Advances in neural information processing systems*, 16(320): 201.

Chen, G.; Qin, Z.; Yang, M.; Zhou, Y.; Fan, T.; Du, T.; and Xu, Z. 2024. Unveiling the vulnerability of private fine-tuning in split-based frameworks for large language models: A bidirectionally enhanced attack. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2904–2918.

Cheng, P.; Hao, W.; Dai, S.; Liu, J.; Gan, Z.; and Carin, L. 2020. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, 1779–1788. PMLR.

Cover, T. M. 1999. *Elements of information theory*. John Wiley & Sons.

Deng, R.; Du, X.; Lu, Z.; Duan, Q.; Huang, S.-C.; and Wu, J. 2023. HSFL: Efficient and Privacy-Preserving Offloading for Split and Federated Learning in IoT Services. In *2023 IEEE International Conference on Web Services (ICWS)*, 658–668. IEEE.

Deng, R.; Lu, Z.; and Duan, Q. 2025. Quantifying Privacy Leakage in Split Inference via Fisher-Approximated Shannon Information Analysis. *arXiv preprint arXiv:2504.10016*.

Dibbo, S. V.; Breuer, A.; Moore, J.; and Teti, M. 2024. Improving robustness to model inversion attacks via sparse coding architectures. In *European Conference on Computer Vision*, 117–136. Springer.

Duan, L.; Sun, J.; Jia, J.; Chen, Y.; and Gorlatova, M. 2024. Reimagining mutual information for enhanced defense against data leakage in collaborative inference. *Advances in Neural Information Processing Systems*, 37: 44479–44500.

Dwork, C. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*, 1–12. Springer.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, Z.; Zhang, T.; and Lee, R. B. 2019. Model inversion attacks against collaborative inference. In *Proceedings of the 35th Annual Computer Security Applications Conference*, 148–162. New York, NY, USA: Association for Computing Machinery.

Jin, S.; Wang, H.; Wang, Z.; Xiao, F.; Hu, J.; He, Y.; Zhang, W.; Ba, Z.; Fang, W.; Yuan, S.; et al. 2024. {FaceObfuscator}: Defending deep learning-based privacy attacks with gradient descent-resistant features in face recognition. In *33rd USENIX Security Symposium (USENIX Security 24)*, 6849–6866.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Li, Z.; Yang, M.; Liu, Y.; Wang, J.; Hu, H.; Yi, W.; and Xu, X. 2023. GAN you see me? enhanced data reconstruction attacks against split inference. *Advances in neural information processing systems*, 36: 54554–54566.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Maeng, K.; Guo, C.; Kariyappa, S.; and Suh, G. E. 2023. Bounding the invertibility of privacy-preserving instance encoding using fisher information. *Advances in Neural Information Processing Systems*, 36: 51904–51925.

Martens, J. 2020. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146): 1–76.

Mi, Y.; Huang, Y.; Ji, J.; Liu, H.; Xu, X.; Ding, S.; and Zhou, S. 2022. Duetface: Collaborative privacy-preserving face recognition via channel splitting in the frequency domain. In *Proceedings of the 30th ACM International Conference on Multimedia*, 6755–6764.

Mireshghallah, F.; Taram, M.; Jalali, A.; Elthakeb, A. T. T.; Tullsen, D.; and Esmaeilzadeh, H. 2021. Not all features are equal: Discovering essential features for preserving prediction privacy. In *Proceedings of the Web Conference 2021*, 669–680.

Mireshghallah, F.; Taram, M.; Ramrakhyani, P.; Jalali, A.; Tullsen, D.; and Esmaeilzadeh, H. 2020. Shredder: Learning noise distributions to protect inference privacy. In *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*, 3–18. New York, NY, USA: Association for Computing Machinery.

Noorbakhsh, S. L.; Zhang, B.; Hong, Y.; and Wang, B. 2024. {Inf2Guard}: An {Information-Theoretic} Framework for Learning {Privacy-Preserving} Representations against Inference Attacks. In *33rd USENIX Security Symposium (USENIX Security 24)*, 2405–2422. Philadelphia, PA: USENIX Association.

Osia, S. A.; Taheri, A.; Shamsabadi, A. S.; Katevas, K.; Haddadi, H.; and Rabiee, H. R. 2020. Deep Private-Feature Extraction. *IEEE Transactions on Knowledge & Data Engineering*, 32(01): 54–66.

Roy, P. C.; and Boddeti, V. N. 2019. Mitigating information leakage in image representations: A maximum entropy approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2586–2594.

Singh, A.; Sharma, V.; Sukumaran, R.; Mose, J.; Chiu, J.; Yu, J.; and Raskar, R. 2024. SIMBA: Split Inference—Mechanisms, Benchmarks and Attacks. In *European Conference on Computer Vision*, 214–232. Springer.

Singh, A.; Vepakomma, P.; Sharma, V.; and Raskar, R. 2023. Posthoc privacy guarantees for collaborative inference with modified Propose-Test-Release. *Advances in Neural Information Processing Systems*, 36: 26438–26451.

Sun, J.; Li, A.; Wang, B.; Yang, H.; Li, H.; and Chen, Y. 2021. Soteria: Provable defense against privacy leakage in federated learning from representation perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9311–9319.

Sun, X.; Gazagnadou, N.; Sharma, V.; Lyu, L.; Li, H.; and Zheng, L. 2024. Privacy assessment on reconstructed images: are existing evaluation metrics faithful to human perception? *Advances in Neural Information Processing Systems*, 36: 10223–10237.

Tan, Q.; Li, Q.; Zhao, Y.; Liu, Z.; Guo, X.; and Xu, K. 2024. Defending Against Data Reconstruction Attacks in Federated Learning: An Information Theory Approach. In *33rd USENIX Security Symposium (USENIX Security 24)*, 325–342. Philadelphia, PA: USENIX Association.

Tishby, N.; Pereira, F. C.; and Bialek, W. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

Vepakomma, P.; Singh, A.; Gupta, O.; and Raskar, R. 2020. NoPeek: Information leakage reduction to share activations in distributed deep learning. In *2020 International Conference on Data Mining Workshops (ICDMW)*, 933–942. IEEE.

Wallace, G. K. 1991. The JPEG still picture compression standard. *Communications of the ACM*, 34(4): 30–44.

Yang, W.; Du, H.; Liew, Z. Q.; Lim, W. Y. B.; Xiong, Z.; Niyato, D.; Chi, X.; Shen, X.; and Miao, C. 2022. Semantic communications for future internet: Fundamentals, applications, and challenges. *IEEE Communications Surveys & Tutorials*, 25(1): 213–250.

Yin, Y.; Zhang, X.; Zhang, H.; Li, F.; Yu, Y.; Cheng, X.; and Hu, P. 2023. Ginver: Generative model inversion attacks against collaborative inference. In *Proceedings of the ACM Web Conference 2023*, 2122–2131. New York, NY, USA: Association for Computing Machinery.

Yu, D.; Du, X.; Jiang, L.; Tong, W.; and Deng, S. 2024. EC-SNN: Splitting Deep Spiking Neural Networks for Edge Devices. In Larson, K., ed., *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 5389–5397. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Zhang, B.; Noorbakhsh, S. L.; Dong, Y.; Hong, Y.; and Wang, B. 2025. Learning Robust and Privacy-Preserving Representations via Information Theory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 22363–22371.