Privacy and Accuracy-Aware AI/ML Model Deduplication Extended Technical Report

Hong Guan, Akshay Mhatre, Lulu Xie, Lei Yu, Lixi Zhou,
Li Xiong, Deepti Gupta, Kanchan Chowdhury, Xusheng Xiao, Jia Zou

***Abstract*—With the growing adoption of privacy-preserving machine learning algorithms, such as Differentially Private Stochastic Gradient Descent (DP-SGD), training or fine-tuning models on private datasets has become increasingly prevalent. This shift has led to the need for models offering varying privacy guarantees and utility levels to satisfy diverse user requirements. Managing numerous versions of large models introduces significant operational challenges, including increased inference latency, higher resource consumption, and elevated costs. Model deduplication is a technique widely used by many model serving and database systems to support high-performance and low-cost inference queries and model diagnosis queries. However, none of the existing model deduplication works have considered privacy, leading to unbounded aggregation of privacy costs for certain deduplicated models and inefficiencies when applied to deduplicate DP-trained models.**

In this work, we formalize the problem of deduplicating DP-trained models as two variants with different objectives: (1) to minimize the storage cost and privacy loss with the constraints on privacy loss and accuracy, and (2) to maximize the accuracy of the deduplicated model (the primary objective) and minimize the storage cost (the secondary objective) with the constraints on privacy loss and accuracy. The first problem variant aims to reduce storage and operational costs associated with managing a large number of models, while the second problem variant seeks to create high-quality models with storage benefits. To address these problems, we developed novel greedy strategies to partition models into *base* models and *target* models, and assign *base* models to *target* models to minimize storage costs while minimizing privacy costs and maximizing accuracy, respectively. When deduplicating a *target* model, we designed two different dynamic approaches for scheduling accuracy validations. For the first problem variant, compared to baselines, our approach improved the compression ratio by up to $35\times$ for individual models (including large language models and vision transformers). We also observed up to $43\times$ inference speedup due to the reduction of I/O operations. For the second problem variant, we observed up to 1.3% accuracy improvement and around 90% overall C.R. achieved by our proposed algorithm for representative scenarios.

***Index Terms*—deduplication, privacy, accuracy, DP-SGD**

I. Introduction

AI/ML systems that support model marketplace [1], [2], [3], ML-as-a-Service (MLaaS) [4], [5], [6], and multi-tasking models on edge devices [7], [8] require efficient management of many models, including large language models (LLMs). These models are often trained or fine-tuned on private datasets using privacy-preserving algorithms [9], [10], such as DP-SGD [11]. Such algorithms are widely used to protect training data from attacks, such as re-identification [12], membership inference [13], and model inversion [14]. The management of many DP models poses new challenges:

- Model marketplaces [1], [3] train many versions of a model using different privacy budgets (ϵ). Then, these models are sold to different buyers so that models with higher ϵ (less noise) and higher accuracy will require a higher price, providing more options to potential buyers, while ensuring arbitrage freeness that no model buyers can take advantage of the marketplace by combining multiple lower-quality models sold at lower prices into a higher-quality model that has a designated higher price [1], [2]. In addition, each buyer or buyer group should be assigned a maximal accumulative ϵ over a private dataset based on what they pay or their privilege. This is widely adopted in many systems such as Google BigQuery [15], Snowflake [16], and DProvDB [17]. Usually, a model needs to be stored after being sold for customer service and auditing, incurring operational costs.
- MLaaS platforms [4], [5] have similar problems, serving many models trained with various ϵ values [18], [19], while each customer's total privacy cost is limited by her/his privacy budget. In addition, model serving is usually subject to stringent latency requirements [20]. When the models to be served exceed available memory, heavy I/O operations are required to swap models between disk and memory, making it even more challenging to meet the latency requirements.
- Multi-tasking applications on edge devices [7], [8] deploy multiple DP-protected models (purchased from model marketplaces or trained using ML-as-a-service) in a resource-constrained environment, facing difficulties in balancing response latency and resource (memory/battery) limitations.

Model deduplication is promising to resolve these problems by sharing similar weight blocks across multiple models to dramatically reduce memory footprint and I/O operations, and thus shorten the latency for serving multiple models in resource-constrained environments [7], multi-tenant environments [21], and database systems that support inference queries [22] and model diagnosis queries [23]. In addition, we observed that the block-wise disparity score [7] is small for models trained with different ϵ using DP-SGD, which also suggests that model deduplication is promising for addressing the target problems. Unlike *single-model* compression techniques such as quantization [24], pruning [25], and knowledge distillation [26], deduplication is a *multi-model* technique and can work together with single-model compression techniques by deduplicating multiple compressed models or compressing deduplicated models [22].

Two Target Problem Variants. In this work, we consider two correlated problem variants: (1) How to deduplicate multiple models with DP guarantees to optimize the compression ratio and privacy loss while maintaining the privacy and accuracy constraints of the models? and (2) How to optimize the accuracy and the compression ratio of the deduplicated models

while maintaining the privacy and accuracy constraints of the models? The first problem variant aims to optimize the trade-offs between storage (determining operational costs) and privacy loss for managing a large number of models, while the second problem variant aims to balance the trade-offs between accuracy and storage costs. Both variants are faced with the following challenges.

Challenge 1. Privacy Constraint. Existing approaches allow blocks from a target model to be replaced by blocks from any other models called base models, causing the composed privacy loss (Sec. IV) of the deduplicated model unbounded. Our first problem variant requires a careful selection of base models to minimize overall storage and privacy costs, while our second problem variant requires this process to maximize the accuracy of deduplicated models. (Both problem variants have to meet the privacy and accuracy constraints.) Base model selection is a challenging optimization problem, since it is hard to estimate the compression ratio of deduplicating a pair of base and target models under an accuracy constraint.

Challenge 2. Accuracy Validation. Existing accuracy-aware model deduplication requires frequent assessment of the accuracy of each modified model in downstream tasks, to ensure that any accuracy decline remains within acceptable limits. This validation process presents a trade-off: frequent accuracy checks will delay the processing, while infrequent validations lead to more deduplication failures and hurt the compression ratio. In addition, while existing works [22] focus on problem variant 1 that accepts minimal accuracy above a threshold, there exists no work discussing how to maximize the accuracy of the deduplicated model.

Novel Privacy and Accuracy-Aware Model Deduplication. To address Challenge 1, we quantify different types of privacy cost composition caused by deduplication, and formalize the base model selection problem for each problem variant. We develop a greedy strategy that first clusters models based on the similarity of model metadata, e.g., model architecture and training dataset identifier, and then select base models with high quality (measured by privacy loss and storage benefits for problem variant 1 and by accuracy and storage benefits for problem variant 2) to deduplicate the rest of the models in their cluster. (See Sec. V-A.)

To address Challenge 2, we propose a dynamic validation strategy coupled with gradient-based saliency analysis to reduce the number of accuracy evaluations during deduplication. Instead of validating after every block replacement, we adaptively deduplicate a variable number of blocks at each step to efficiently distinguish deduplicable and non-deduplicable blocks. Based on the different objectives, we design distinct deduplication strategies for each problem variant. For problem variant 1, blocks are ordered by ascending saliency so that low-impact blocks are deduplicated first to maximize the compression ratio with minimal impact to accuracy. For problem variant 2, we adopt an iterative two-phase strategy with a gradually increased accuracy constraint. Each iteration has two steps. In the first step, blocks with high saliency are deduplicated first while identifying "*high potential*" blocks that failed to be deduplicated due to a near miss of the accuracy threshold. Then, in the second step, these high-potential blocks are deduplicated first with an attempt to accumulate positive impacts on the model accuracy, while avoiding the negative accuracy impacts from low-potential blocks. (See Sec. V-B)

The key contributions of our work include:

1) We are the first to formalize the privacy loss composition of deduplicated models and two problem variants.
2) We are also the first to design and develop end-to-end model deduplication pipelines to optimize and balance privacy, accuracy, and compression ratio, including base model selection, and dynamic block deduplication, for each problem variant, respectively.
3) We implement the proposed system and conduct comprehensive evaluations. The results show that, for the first problem variant, compared to SOTA model deduplication methods (with our base model selection to ensure privacy), our approach can improve the compression ratio for deduplicating multiple large language models, vision transformers, and ResNet models, by up to $35\times$ for individual models with privacy costs reduced by up to $3\times$ compared to alternative base model selection designs. The multi-model inference speed is also accelerated by $43\times$ by avoiding swapping models between memory and disk. For the second problem variant, we demonstrate that it is possible to obtain models with better accuracy and smaller storage at the same time, while satisfying privacy and accuracy constraints.

This paper is an extension of our earlier work published in SIGMOD 2025 [27]. However, the earlier version only focuses on the problem variant 1, while this work systematically explores the problem variant 2. The newly added material is all highlighted in blue.

II. Preliminaries

Differential Privacy [28] (DP) is a mathematical framework for quantifying privacy risks in data analysis, and is increasingly adopted in real-world applications by companies like Google and Apple.

Definition 2.1 (Differential Privacy [28]): A randomized algorithm $\mathcal{M}$ is (ϵ,δ)-differentially private if for all neighboring datasets (i.e. differ by exactly one element) $\mathcal{D}$ and $\mathcal{D}'$, and for $\forall \mathcal{S} \subseteq \text{Range}(\mathcal{M})$, $Pr[\mathcal{M}(\mathcal{D}) \in \mathcal{S}] \leq e^{\epsilon} * Pr[\mathcal{M}(\mathcal{D}') \in \mathcal{S}] + \delta$. ϵ, also called **privacy budget**, is a metric of privacy loss at a differential change in data. δ denotes the probability of the privacy guarantee being failed.

DP is characterized by three key properties. *Sequential Composition* illustrates how privacy costs accumulate when multiple analyses are performed on the same dataset. When analyses are conducted on disjoint subsets of data, *Parallel Composition* [29] considers the maximum privacy cost on each dataset as the privacy cost. *Post-processing* ensures that any data-independent processing of differentially private outputs will not incur additional privacy costs. The formal statements of these properties are:

Theorem 2.1 (Sequential Composition [28]): If $\mathcal{M}_1$ is (ϵ_1, δ_1)-DP and $\mathcal{M}_2$ is (ϵ_2, δ_2)-DP, then their sequential composition $\mathcal{M}(x) = (\mathcal{M}_1(x), \mathcal{M}_2(x))$ is $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$-DP.

Theorem 2.2 (Parallel Composition[29]): Let $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$ be disjoint subsets of the database $\mathcal{D}$. If algorithm $\mathcal{M}_i$ is (ϵ, δ)-DP for each $i \in [1, k]$, then the parallel composition $M(x) = (\mathcal{M}_1(\mathcal{D}_1), \ldots, \mathcal{M}_k(\mathcal{D}_k))$ is $(\max_i \epsilon_i, \max_i \delta_i)$-DP.

Theorem 2.3 (Post-processing [28]): If $\mathcal{M}$ is (ϵ, δ)-DP and f is any data independent function, then $\mathrm{f} \circ \mathcal{M}$ is also (ϵ, δ)-DP.

The post-processing property ensures that the privacy budget of a model will not increase with the number of inferences. In addition, given models trained on a dataset D, deduplicating them will not affect the overall privacy budget on D.

In the context of DP, sensitivity measures the maximum change in the output of a function when a single data point in the input dataset is modified. Formally, we have Def. 2.2.

Definition 2.2 (Sensitivity[28]): For a given function f, the sensitivity Δf is defined as $\Delta f = \max_{\mathcal{D},\mathcal{D}'} \|f(\mathcal{D}) - f(\mathcal{D}')\|$ where $\mathcal{D}$ and $\mathcal{D}'$ are any two neighboring datasets.

Differentially Private Stochastic Gradient Decent (DP-SGD) [11]. In DP-SGD, privacy is maintained through two primary modifications to the traditional SGD algorithm: gradient clipping and the addition of noise. Mathematically, the update rule for the model parameters θ at step t in DP-SGD can be expressed as:

$$\theta_{t+1} \leftarrow \theta_t - \eta \left(\frac{1}{B} \sum_{i=1}^{B} \text{clip}(g_i(\theta_t), C) + \mathcal{N}(0, \sigma^2 I) \right)$$

where η is the learning rate, B is the batch size, $g_i(\theta_t)$ is the gradient of the loss function for the i-th data point in the mini-batch, $\text{clip}(g_i(\theta_t), C)$ denotes the gradient clipping operation, and $\mathcal{N}(0, \sigma^2 I)$ denotes the Gaussian noise with variance σ^2.

Note in this paper, for simplicity, we focus on ϵ, which dominates the privacy budget. However, all our analysis can be easily extended to δ by applying the same constraints and computations on δ as on ϵ since they follow the same composition rules (Theorem 2.1, 2.2, 2.3).

III. Privacy Derivation and Problem Formalization

No existing works have formalized privacy budget derivation and quantification for the deduplicated models. Therefore, in Sec. IV, we are the first to quantify the privacy loss of a target model M_t deduplicated by a base model $\hat{M}$, while M_t and $\hat{M}$ could be trained on the same, disjoint, and even overlapping datasets. Then, in Sec. III-B1 and Sec. III-B2, we formalize both problem variants with different objectives.

A. Privacy Budget Derivation

The privacy budget of M'_t in the above deduplication process can be determined by Theorem 3.1, of which the intuition is as follows. Consider the case where two similar blocks b_t and $\hat{b}$ are from model M_t with ϵ_t and $\hat{M}$ with $\hat{\epsilon}$ that are trained on dataset D_t and $\hat{D}$ respectively. Using $\hat{b}$ to replace b_t will change ϵ_t into ϵ'_t in three situations: (1) **Intra-Model** ($M_t = \hat{M}$): If b_t and $\hat{b}$ are from the same model, they can be safely deduplicated with $\epsilon'_t = \epsilon_t$. (2) **Inter-Model Intra-Data** ($M_t \neq \hat{M} \wedge D_t = \hat{D}$): If b_t and $\hat{b}$ are from different models that were trained with the same dataset, we have $\epsilon'_t = \epsilon_t + \hat{\epsilon}$ by applying DP's sequential composition property [28], [29], detailed in Sec. II. (3) **Inter-Model Inter-Data** ($(M_t \neq \hat{M}) \wedge (D_t \cap \hat{D} = \Phi)$): If b_t and $\hat{b}$ are from different models that were trained on disjoint datasets, we have $\epsilon'_t = \max(\epsilon_t, \hat{\epsilon})$ by applying DP's parallel composition [30], [29], detailed in Sec. II.

Theorem 3.1: Suppose that d disjoint training datasets are used for training all models. Let $\hat{M}_t$ be a (ϵ_t, δ_t)-DP model to be deduplicated, and let $\hat{\mathcal{M}} = \{\hat{M}_1, ..., \hat{M}_k\}$ be a set of k distinct differentially private models, each satisfying $(\hat{\epsilon}_i, \hat{\delta}_i)$-DP, serving as potential block providers. Any two models in $\{M_t\} \cup \hat{\mathcal{M}}$ are trained on either the same dataset or completely disjoint datasets from each other. Let $\hat{\mathcal{M}}_j = \{\hat{M}_{j_1}, ...\hat{M}_{j_l}\}$ be the group of models trained on the dataset D_j. Then, on D_j, the resulting deduplicated model, denoted as M'_t, satisfies $(\epsilon_t \cdot \mathbb{1}_{M_t, D_j} + \sum_{i=1}^{l} \hat{\epsilon_{j_i}}, \quad \delta_t \cdot \mathbb{1}_{M_t, D_j} + \sum_{i=1}^{l} \hat{\delta_{j_i}})$-DP, where the indicator function $\mathbb{1}_{m, D_j}$ is equal to 1 if model m is trained on D_j, otherwise 0. If $\hat{\mathcal{M}} = \emptyset$, then M'_t satisfies (ϵ_t, δ_t)-DP.

Given any disjoint D_j $(1 \leq j \leq d)$, if the derived model M'_t satisfies (ϵ^j, δ^j)-DP, the derived model M'_t satisfies $(\max_j \epsilon^j, \max_j \delta^j)$-DP on the union of datasets $\cup_{j=1}^{d} D_j$.

Proof. Because any two models are trained on either the same dataset or the datasets completely disjoint from one another, we can divide the models in $\hat{\mathcal{M}}$ into groups by their training datasets. $\hat{\mathcal{M}}_j = \{\hat{M}_{j_1}, ...\hat{M}_{j_l}\}$ is a group of models using training data D_j. By DP composability, the collection of blocks from this group of models satisfies $(\sum_{i=1}^{l} \hat{\epsilon_{j_i}}, \sum_{i=1}^{l} \hat{\delta_{j_i}})$-DP. Because the deduplication algorithm iterates over all blocks to decide replacements, by DP post-processing property, the derived model M'_t satisfies $(\sum_{i=1}^{l} \hat{\epsilon_{j_i}}, \sum_{i=1}^{l} \hat{\delta_{j_i}})$-DP on D_j when the original model M_t is not trained on D_j. If M_t is trained on D_j, M'_t satisfies $(\epsilon_t + \sum_{i=1}^{l} \hat{\epsilon_{j_i}}, \delta_t + \sum_{i=1}^{l} \hat{\delta_{j_i}})$-DP by DP sequential composition. Given any disjoint dataset D_j $(1 \leq j \leq d)$, if the deduplicated model M'_t satisfies (ϵ^j, δ^j)-DP, it satisfies $(\max_j \epsilon^j, \max_j \delta^j)$-DP on the union of disjoint datasets $\cup_{j=1}^{d} D_j$, by the parallel composition property [29].

Generalization. We can easily remove the disjoint dataset assumption by abstracting the problem as a graph. Assume that for graph G, each node represents a base model, and two nodes $\hat{M}_p$ and $\hat{M}_q$ have an edge if and only if their training datasets overlap. Regarding each connected component C_j of k nodes as a model group $\{\hat{M}_{j_1}, ...\hat{M}_{j_k}\}$, similarly, the collection of blocks from it satisfies $\epsilon^j = \sum_{i=1}^{k} \epsilon_{j_i}$ -DP. Suppose G has d connected components, then the derived model M' satisfies $(\max_{1 \leq j \leq d} \epsilon^j)$-DP.

B. Overall Problem Formalization

Given a set of DP-trained models managed by a model marketplace, an MLaaS platform, or an edge server, the deduplication process can reduce the storage costs, the memory costs, and the multi-model serving latency, because a smaller number of blocks need to be stored and loaded into the memory than in the case without deduplication. This process would also desire better accuracy (i.e., utility) and lower privacy loss (i.e., ϵ) achieved for the resulting model (i.e., the target model after deduplication). In addition, this process

must satisfy the *fairness rule* that ensures models with higher ϵ on the same dataset must maintain higher accuracy, and vice versa, models trained with better accuracy must be associated with more privacy loss [2]. While there exist many different variants of the problem definitions with different optimization objectives and constraints, we focus on two variants that we believe have the most practical impacts.

1) Problem Variant 1—Optimize Storage and Privacy Costs: In the first variant, we aim to minimize the storage costs (i.e., number of blocks after deduplication, as represented in Eq. 1) and the overall privacy loss (i.e., Eq. 2, where privacy loss on the same dataset are summed up for sequential compositions, and the maximum of summed privacy on each dataset is obtained for parallel composition). At the same time, it must ensure that the privacy loss increase and the accuracy (i.e., utility) drop for each model after deduplication is within pre-defined constraints, as well as the fairness rule. The formalized definition for the problem is as follows.

Given:

- A set of n models $\{M_1, M_2, \ldots, M_n\}$ trained on k datasets $D_1, \ldots, D_k$. Model M_i is composed of a set of unique blocks B_i, and is trained with utility u_i and privacy budget ϵ_i
- We let $B = \bigcup_{i=1}^{n} B_i$.
- A set of models $\{M'_1, M'_2, \ldots, M'_n\}$ are obtained after deduplicating $\{M_1, M_2, \ldots, M_n\}$, where B'_i, the set of blocks in model M'_i, satisfies $B'_i \subseteq B$. Model M'_i has utility u'_i and privacy budget ϵ'_i.
- $B' = \bigcup_{i=1}^{n} B'_i$ is the set of all unique blocks used after deduplication, i.e., $B' \subseteq B$.
- Each deduplicated model preserves the architecture, and thus we have $|M'_i| = |M_i|$ for all i.
- $\Delta\epsilon_i = \epsilon'_i - \epsilon_i$ is the privacy budget increase from M_i to M'_i after deduplication.
- $\Delta u_i = u_i - u'_i$ is the utility drop from M_i to M'_i after deduplication.
- Thresholds ϵ^*_i and u^*_i are the maximum allowable privacy increase and utility drop for each model.
- For similar models $M_j \sim M_k$ trained on the same dataset, if $u'_j < u'_k$, then $\epsilon'_j \leq \epsilon'_k$ (fairness constraint).

Objectives:

$$\min_{M'_1,\ldots,M'_n} |B'| \tag{1}$$

$$\min_{M'_1,\ldots,M'_n} \max_i \sum_{M_j \text{ trained on } D_i} \epsilon'_j \tag{2}$$

Subject to:

$$\epsilon'_i - \epsilon_i \leq \epsilon^*_i \quad \forall i \in \{1, \ldots, n\} \tag{3}$$

$$u_i - u'_i \leq u^*_i \quad \forall i \in \{1, \ldots, n\} \tag{4}$$

$$M_j \sim M_k,\ u'_j < u'_k \Leftrightarrow \epsilon'_j \leq \epsilon'_k \tag{5}$$

2) Formalization of Problem Variant 2—Optimize Accuracy: In the second variant, we attempt to maximize the accuracy of models as the primary objective and minimize the storage costs after deduplication as the secondary objective, while ensuring that the privacy loss increase is within pre-defined constraints and that the fairness rule is also satisfied. Solving this problem will bring high-accuracy models with reduced storage costs while meeting the privacy constraint and the fairness constraint. This may generate a significant impact on the model management platforms so that they can create new, higher-quality models without conducting more training rounds on the sensitive data from the data owners. The problem formalization shares the same setup, including constraints, with problem variant 1 (Sec. III-B1), but has different objectives, detailed as follows:

Objectives:

$$\max_{M'_1,\ldots,M'_n} \sum_{i=1}^{n} u'_i \quad \text{(primary objective)} \tag{6}$$

$$\min_{M'_1,\ldots,M'_n} |B'| \quad \text{(secondary objective)} \tag{7}$$

Subject to: Eq. 3, Eq. 4, and Eq. 5.

IV. Problem Analysis

Here, we will summarize the problems with the SOTA deduplication procedure, and an overview of our solution.

A. Existing Model Deduplication Procedure

The state-of-the-art (SOTA) accuracy-aware model deduplication technique [22], termed **Dedup**, assumes that models arrive in order. Starting from the second model, it will repeat the following steps: **Step 1.** To deduplicate a target model M_t, use all previously arrived models to serve as potential block providers, called base models. **Step 2.** Order all blocks from M_t by some saliency measures (e.g., the 3rd-quantile of weight values in the block) ascendingly. **Step 3.** For each block $b \in M_t$ in the ordered list, run the following sub-steps: (3-1) Select a block b' that is most similar to b from the collection of blocks in all base models (e.g., using locality sensitive hashing [22]). (3-2) Substitute b with b', resulting in a modified model M'_t. (3-3) Evaluate the accuracy of the modified target model using a validation dataset. (This step could be performed once every N iterations, naively reducing the validation frequency.) (3-4) If the accuracy drop exceeds a threshold, undo the block replacement and stop early, otherwise (3-5) Move to the next block $b \in M_t$.

The SOTA method has the following limitations for our targeting problem:

(L1) Step 1 naively takes all available models as base models, without checking whether combining the target model with blocks from all base models will cause a privacy budget increase that violates the privacy budget constraint.

(L2) Step 2 uses simple saliency measures, which we found less effective for models trained with DP.

(L3) Step 2 orders the blocks ascendingly by block saliency measurements. Deduplication following such ordering can help avoid an accuracy drop of the target model, while deduplicating as many blocks as possible, which is consistent with problem variant 1. However, it cannot help improve the

accuracy of the deduplicated model, which is required by problem variant 2.
(L4) Step 3 adopts a static strategy to validate the accuracy of a deduplicated model. (As mentioned, the expensive accuracy validations compose a major bottleneck of the latency.)

B. Methodology Overview

An overview of our work is as follows.
Base Model Selection (Sec. V-A). To address L1, we first formalize the base model selection as a combined set partitioning and generalized assignment problem, which partitions the models into base models and target models and assigns base models to target models. However, there is no easy way to efficiently estimate the compression ratio for deduplicating each pair of models without incurring additional privacy costs. Therefore, we developed a greedy algorithm to prioritize high-quality models that have lower privacy costs and higher storage benefits to serve as the base models for problem variant 1. We further extend the greedy algorithm to problem variant 2 by considering the accuracy of candidate base models.
Deduplication with Saliency Analysis and Dynamic Validation (Sec. V-B). L2 and L3 are correlated. We compared several existing weight saliency measurements [31], [32], [33], [34] and found the gradient magnitude [34] well balances saliency profiling latency and effectiveness. We also propose our dynamic deduplication scheme for problem variant 1, where dynamic ranges of non-salient blocks are grouped and deduplicated together to avoid accuracy validation failures. We further extend the algorithm to problem variant 2 by adopting an iterative two-step strategy that gradually increases the accuracy constraint, and in each iteration, it first prioritizes the blocks measured as high saliency for deduplication in Step 1, and then in Step 2, it prioritizes the high-potential blocks identified in Step 1 for deduplication.

V. Novel Automatic Deduplication

In this section, we will present our solutions in detail.

A. Base Model Selection

1) Problem Variant 1: A base model is a block provider, e.g., M_1 in Fig. 7.

According to the privacy analysis in Sec. IV, the more base models used, the more the privacy loss (ϵ) of the target model increases, which is the **key** to optimize the privacy loss of the target model. The research question is *how to select base models and assign target models to base models to minimize storage and privacy costs while meeting the accuracy and privacy constraints*. We start from two assumptions: (Assumption-1) each target model can only use one base model for deduplication, which practically reduces the privacy loss of the resulting models and the deduplication overheads, while many target models can share one base model to maximize their utilization; and (Assumption-2) We do not allow a model $\hat{M}$ to be assigned as a base model and a target model at the same time, because deduplicating $\hat{M}$ as a target model may cause parameter and accuracy changes in all models deduplicated using $\hat{M}$ as a base model.

Problem Definition. Given a set of models trained on the same dataset or disjoint datasets, $\mathcal{M} = \{M_1, ..., M_n\}$, satisfying $\epsilon_1, \ldots, \epsilon_n$-DP respectively. These models have utilities (accuracy) $u_1, \ldots, u_n$ respectively. Each model is split into (tensor) blocks of equal size. Each model $M_j (j = 1, \ldots, n)$ has an ϵ increase threshold ϵ^* and a utility drop threshold u_j^*. The deduplication process must satisfy (1) M_j's ϵ increase should be bounded by ϵ^*, (2) M_j's utility drop should be bounded by u_j^*, and (3) the *fairness rule* ensuring models trained with higher ϵ on the same dataset must maintain higher accuracy than those trained with lower ϵ.

We suppose n_{ij} in an $n \times n$ matrix $\mathcal{N} = \{n_{ij}\}$ represents the number of blocks from M_j (as a target model) that can be replaced by blocks from M_i (as a base model) with M_j's utility drop bounded by u_j^*. $\mathcal{A} = \{a_{ij}\}$ represents the partitioning and assignment matrix that partitions models into target models and base models and assigns target models to base models. $a_{ij} = 1$ means M_j is assigned to the base model M_i. The objective function and constraints are formalized below, where $f_\epsilon(M_i, M_j)$ denotes the privacy budget (ϵ_i') of the target model M_j using M_i as a base model based on the derivation in Sec. IV. λ is a regularization factor to balance the two objectives (1) To maximize the total number of blocks that can be deduplicated (replaced) ($\sum_{i=1}^{n}\sum_{j=1}^{n}(n_{ij})$) and (2) to minimize the target models' overall ϵ increase caused by deduplication ($\sum_{i=1}^{n}\sum_{j=1}^{n}(f_\epsilon(M_i, M_j))$). The optimization process must satisfy the adjusted privacy constraint ϵ_j^{adjust} (Eq. 9) explained later, the Assumption-1 (Eq. 10), and Assumption-2 (Eq. 11).

$$\max_{a_{ij}=0,1(1\le i,j\le n)} \sum_{i=1}^{n}\sum_{j=1}^{n}(n_{ij} - \lambda \times f_\epsilon(M_i, M_j)) \times a_{ij}, \quad s.t. \tag{8}$$

$$f_\epsilon(M_i, M_j) - \epsilon_j < \epsilon_j^{adjusted} \quad \text{if} \quad a_{ij} = 1 \tag{9}$$

$$\sum_{i=1}^{n} a_{ij} = 1 \tag{10}$$

$$\forall M_i, \nexists M_k, M_j (i \neq k \neq j) \quad \text{satisfying} \quad a_{ik} = 1 \wedge a_{ji} = 1 \tag{11}$$

Adjusted utility decrease threshold u_i^{adjust} and privacy cost increase threshold ϵ_j^{adjust}. To meet the fairness rule, we enforce that the orderings of these models by the ϵ and by *utility* do not change after deduplication. Then, if all models satisfy the fairness rule before deduplication, these models still adhere to the fairness rule after deduplication. Suppose M_i and M_{i-1} are two consecutive models in the sequence of k models trained on the same dataset ordered ascendingly by utility ($u_i > u_{i-1}$). If the user-specified utility ranges of the deduplicated models M_i' and M_{i-1}' are $[u_i - u_i^*, u_i]$ and $[u_{i-1} - u_{i-1}^*, u_{i-1}]$ and they overlap, then $u_i' < u_{i-1}'$ is possible, which violates the fairness order by utilities. To eliminate such overlapping ranges, we adjust u_i^* to be $u_i^{adjusted} = min(u_i^*, u_i - u_{i-1}')$ for $i = 2, \ldots, k$ so that the deduplication algorithm always ensures $u_i' > u_{i-1}'$. Similarly, we set ϵ_i^* to be $\epsilon_i^{adjusted} = min(\epsilon_i^*, \epsilon_i - \epsilon_{i-1}')$ for $i = 2, \ldots, k$.

Analysis. If we fix a subset of models $\mathcal{B} \subset \mathcal{M}$ to be base models, the problem is a generalized assignment problem (GAP) [35], where one or more target models (i.e., tasks in the classical GAP problem) are assigned to each base model (i.e., agents in the classical GAP). The GAP problem was proven to be NP-hard [35]. However, instead of having a fixed $\mathcal{B}$, our problem additionally requires searching for a set-partitioning scheme that divides $\mathcal{M}$ into $\mathcal{B}$ and $\mathcal{M} - \mathcal{B}$.

Moreover, the problem is a black-box optimization problem [36], since it is expensive to estimate n_{ij}. Performing actual deduplication to determine n_{ij} is expensive. Using weight-level similarity to estimate n_{ij} is also impractical since the similarity function will increase privacy loss. Therefore, we developed a greedy strategy to address the problem.

2) A Greedy Strategy.: The main intuition is that the quality of the base models is more important than the quantity. A high-quality base model should satisfy the following requirements: (**R1**) **Similar to target models.** Using a base model with greater similarity in model architecture, training dataset, utility, and privacy budget to the target model usually leads to a better compression ratio (with the same utility constraint). (**R2**) **Low privacy budget.** Usually, a base model with a lower privacy loss leads to a smaller increase in privacy cost in the deduplicated model. (**R3**) **Low chance to be deduplicated (Low opportunity cost)**. A model having fewer qualified base models than qualified target models is more suitable to serve as a base model than a target model. Based on the intuition, our key idea is to cluster models by similarity of model metadata, and in each cluster, we select the models satisfying R1, R2, and R3, to serve as base models for the rest of the cluster:

- **Partitioning Models Based on Similarity.** We first cluster models from $\mathcal{M}$ based on public metadata, such as model architecture, public training dataset, and the anonymous identifier of the private training dataset. We employed a simple hierarchical clustering strategy and ensured that all models in the same cluster shared the same model architecture and similar datasets.
- **Assigning Dangling Models as Base Models.** We define dangling models to be models that do not have any qualified base models in their clusters (i.e., M_j cannot find any M_i satisfying Eq. 9). When ϵ^* is set to be smaller than the models' lowest ϵ, there must be at least one dangling model in each cluster–the model with the smallest ϵ. We only allow these dangling models to serve as the base models to deduplicate other models in the cluster since they are similar to other models (i.e., can deduplicate a good number of blocks from other models) and zero blocks can be deduplicated from them if assigned as target models, thus maximizing the storage benefits in its local cluster. The process is described in Alg. 1.
- **Intra-Group and Inter-Group Base Model Selection.** For each non-dangling model, it must be able to find at least one qualified base model in its cluster. If it has two or more qualified base models in its cluster, selecting any of them leads to similar compression ratios. Therefore, we will select the one that brings minimum privacy cost increase as its base model (line 6-9 in Alg. 3). Then, for those dangling models that are never used as base models (i.e., $count == 0$), we search for qualified base models from other clusters to deduplicate them (line 10-15 in Alg. 3).

Algorithm 1 Identify dangling models in a group $\mathcal{G} = \{M_1, ..., M_k\}$

```
1: procedure CandidateBaseModelGen(G)
2:     B ← {}; T ← G
3:     for M_i in T do                ▷ Identify dangling models and add them to B
4:         if ε_j ≥ ε_i^adjusted for all j ∈ {1, ..., k} and j ≠ i then
5:             B = B ∪ {M_i}; T = T − {M_i};
6:         end if
7:     end for
8:     return
9: end procedure
```

Algorithm 2 Select a base model $\hat{M}$ for a target model M_t in a Group $\mathcal{G}$ (, aware of dangling models)

```
1: procedure BaseModelSelectStep(M_t, G)
2:     B ← The set of candidate base models in G; (Dangling models identified by Alg. 1)
3:     Find base model M̂ ∈ B with the smallest ε̂ to satisfy f_ε(M̂, M_t) − ε_t < ε_t^* to serve as the base model for M_t; M̂.count++
4:     return M̂
5: end procedure
```

Algorithm 3 Base model selection for all models in $\mathcal{M}$

```
1: procedure BaseModelSelectMain(M)
2:     C = {G_1, ..., G_l} ← Cluster(M)
3:     for G_i ∈ C do CandidateBaseModelGen(G_i)
4:     end for
5:     for G_i ∈ C do
6:         sort (C' = C − G_i) based on similarity to G_i
7:         for M_j ∈ G_i do
8:             ret ← BaseModelSelectStep(M_j, G_i)
9:             record that M_j's base model is ret
10:        end for
11:        for M_k ∈ B_i do
12:            if M_k.count == 0 then
13:                while ret = NULL ∧ next_cluster(C') ≠ NULL do
14:                    G_k ← next_cluster(C')
15:                    ret ← BaseModelSelectStep(M_k, G_k)
16:                end while
17:                record that M_k's base model is ret
18:            end if
19:        end for
20:    end forreturn
21: end procedure
```

3) Extension to Problem Variant 2: In this section, we first examine whether the greedy strategy proposed in Sec. V-A2 is suitable for problem variant 2. We consider two different situations as follows:

(a) The privacy constraints are relaxed. Note that deduplication will not incur the privacy loss from the data owner's side due to the post-processing properties; however, the privacy loss (information leakage) to each model buyer would be increased through deduplication. A relaxed privacy constraint for deduplication means the broker sets a high bound of privacy loss to model buyers, which means the information leakage to each model buyer is less of a concern. This setting raises an interesting question: when privacy constraints are sufficiently relaxed, is it possible to obtain higher-accuracy models through deduplication rather than retraining with more data and more epochs? Reducing the training epochs not only reduces training/finetuning overhead but also reduces the direct depletion of data owners' privacy budgets.

With relaxed privacy constraints, there are no dangling models anymore, rendering the base model selection strategy proposed for problem variant 1 ineffective. To maximize the

accuracy of deduplicated models, within each cluster, we always select the model with the highest accuracy (largest ϵ), which provides the best-quality blocks, to serve as the base model for the rest of the models in the cluster. Therefore, we replace Alg. 2 by a new base model selection step as illustrated in Alg. 4. The overall base model selection algorithm is thus slightly different from Alg. 3 by removing Line 3 and changing Line 8 and Line 15 by replacing `BaseModelSelectStep` with `BaseModelSelectStep1`.

Algorithm 4 Select a base model $\hat{M}$ for a target model M_t in a Group $\mathcal{G}$ (, NOT aware of dangling models)

1: **procedure** BASEMODELSELECTSTEP1(M_t, $\mathcal{G}$)
2: $\mathcal{B} \leftarrow \mathcal{G}$;
3: Find a base model $\hat{M} \in \mathcal{B}$ with the largest $\hat{\epsilon}$ to satisfy $f_\epsilon(\hat{M}, M_t) - \epsilon_t < \epsilon_t^{adjusted}$ to serve as the base model for M_t; $\hat{M}.count++$
4: **return** $\hat{M}$
5: **end procedure**

(b) The privacy constraints are strict, where each cluster has at least one dangling model. In this case, we shall follow Alg. 3, which allows the maximal number of models to participate in the deduplication process as target models by fully exploiting dangling models to serve as base models to reduce storage costs. However, to optimize accuracy as required by problem variant 2, a small change should be made in Alg. 2 by changing Line 3 to find base model $\hat{M} \in \mathcal{B}$ with the **largest** $\hat{\epsilon}$ that satisfies $f_\epsilon(\hat{M}, M_t) - \epsilon_t < \epsilon_t^{adjusted}$ to serve as the base model for M_t.

B. Deduplication Algorithms

1) Problem Variant 1: Once we select the base model for each target model, the privacy budget increase becomes a constant and the privacy and fairness constraints are ensured. Then, the model deduplication problem is reduced to a set of two-model deduplication problems. Each problem involves a target model M_j and a base model M_i. The optimization objective is to maximize $n_{ij} = |M_i \cap M'_j|$, i.e., the number of unique blocks in the deduplicated model M'_j that are from M_i (i.e., minimize the compression ratio) while meeting the utility drop constraint formalized below.

$$\max n_{ij} \text{ with}$$
$$M'_j = \text{deduplicate}(M_j; M_i \text{ as base model}) \; s.t.$$
$$u'_j > max(u_j - u^{adjusted}, u_i)$$

2) Problem Analysis: Given any block from M_j, it can be replaced by any block from M_i (or no replacement), leading to $(|M_i| + 1)^{|M_j|}$ possible deduplication plans. Depending on the block size, deep learning models in our experiments may have hundreds to thousands of blocks. Therefore, it is computationally inhibiting to exhaustively search and evaluate the storage costs and accuracy drop of all deduplication plans. Following the existing deduplication framework presented in Sec. IV-A to perform two-model deduplication, it performs a validation after deduplicating every N blocks. When N is set to a small number, the frequent utility validation operations will bottleneck the deduplication process. However, if we set N to a large number, it often deduplicates too many blocks to keep the utility drop within the constraint, due to the lack of sophisticated saliency analysis to order the blocks. A poor saliency measurement may mix a few salient blocks, which will significantly impact the model's utility if deduplicated, with many non-salient blocks in one batch. Then, the salient blocks will cause a validation failure, which will further cause *all* deduplications in the batch, including the deduplications of non-salient blocks, to be rolled back. This will lead to missed opportunities and wasting of resources. However, saliency analysis is a challenging task in nature and made even more difficult by the noises introduced by DP [37].

3) Our Dynamic Deduplication Algorithms: We resolve the problem in two steps: (1) We carefully examine and compare the block saliency measurements for our target problem. (2) We design new algorithms that dynamically adjust the size of each group to balance the frequency of validations and the number of rollbacks (validation failures). For example, at the beginning, a large group of the least salient blocks should be deduplicated together. As we progress through the ordered list, the group size gradually decreases, ensuring that non-salient blocks are checked in a fine-grained manner.

Block Saliency Measures. Weights that significantly impact the prediction are called "salient weights" [38], [33]. There are many ways to measure the saliency of weights for model explanability [39], compression [40], [31], [41], [34]. The most popular ones include the weight magnitude [22], activation magnitude [42], and the Fisher information that involves the square of gradients [32]. We find that the Fisher information metric is the most accurate among them. However, it is also the most computationally intensive. In addition, the squared gradients are usually very small and may introduces unstable numerical results. Therefore, we adopted the absolute value of gradient (called gradient magnitude) as our saliency measure. The saliency of a block is defined as the mean of the gradient magnitude of each weight within a block, which is obtained by one forward and backward pass of a dataset. This dataset could be a public dataset or a disjoint partition from the private training/validation dataset. In the latter case, we added noise to the gradient, and this one-time privacy cost is usually smaller than the model's ϵ and thus gets absorbed due to DP's parallel composition.

Group Deduplication Algorithms. Our idea is to dynamically determine the group size and re-evaluate some blocks in a group that fail to be deduplicated. We designed a recursive dynamic-range deduplication (`DRD`) algorithm, as formalized in Alg. 5. `DRD` begins by dividing the input of blocks (i.e., input range) into two halves. It then attempts to deduplicate the left half. If this deduplication fails (i.e., the utility drop exceeds the bound), the algorithm recursively runs on the left half (Line 14) so that at every recursion level, the number of blocks that are deduplicated in a batch will be reduced by half. After deduplicating the left half, the algorithm recursively deduplicates the right half (Line 15). The base case for this recursion occurs when the input range has fewer than L blocks, and it skips such a small range because such remaining blocks are likely to be salient blocks. We also proposed a variant of `DRD`, called Dynamic-Range-Expansion Deduplication (`DRED`), slightly different from `DRD` in Line 17. `DRED` will recursively deduplicate the blocks from $m+1$

Algorithm 5 Dynamic-Range Deduplication (DRD)

```
1: Input: Target and base models represented by a list of block identifiers B, B̂, sorted by block saliency ascendingly; utility drop threshold T; minimum #blocks in a batch L.
2: procedure DEDUPLICATION(B, B̂, T)
3:     Initialize l = 0; r = |B| − 1;
4:     Recursion(B, B̂, l, r, T);
5:     return ;
6: end procedure
7: procedure RECURSION(B, B̂, l, r, T)
8:     if r − l < L then                                   ▷ Base case, skip a small batch
9:         return ;
10:    end if
11:    m = (r + l)//2;
12:    for block j in l, l + 1, ..., m do                  ▷ Try deduplicate the left half
13:        Replace block B[j] by the most similar block in B ∪ B̂;
14:    end for
15:    Evaluate model on validation set, compute utility drop Δu;
16:    if Δu >= T then                                     ▷ If validation fails
17:        Rollback the deduplication of block l, l + 1, ..., m;
18:        if l < r then
19:            Recursion(B, B̂, l, m, T);                   ▷ Recursively deduplicate the range
20:        end if
21:    end if
22:    Recursion(B, B̂, m + 1, r, T);                       ▷ Recursively deduplicate the right half
23:    return ;
24: end procedure
```

to $|B+1|$, rather than $m+1$ to r (i.e., changing line 17 to Recursion$(B, \hat{B}, m+1, |B|-1, T)$). It allows for more aggressive (coarser-grained) deduplication attempts.

Given a block b in each deduplication batch, a block from the base model most similar to block b is chosen to replace b. To measure similarity, we compared several distance measures, such as `l1-norm`, `l2-norm`, and `cosine` and chose to directly use `l2-norm` (i.e., Euclidean distance). We use pairwise comparison for best accuracy, since it is not the bottleneck of our target scenario (less than 10% of overheads). If needed, it can be easily replaced by a faster but more error-prone locality-sensitive hashing [43] technique.

C. Extension to Problem Variant 2

Problem Definition. The two-model deduplication process for our problem variant 2 is defined as a multi-objective optimization problem in linear form as follows:

$$\max \alpha * u_{ij} + \beta * n_{ij}/|B_j| \text{ with}$$
$$M'_j = \text{deduplicate}(M_j; M_i \text{ as base model}) \; s.t.$$
$$u'_j > max(u_j - u^{adjusted}, u_i)$$

Problem Analysis. To deduplicate two models for problem variant 2, our first observation is that ordering blocks by their saliency in ascending order, as adopted in Sec. V-B3 is not helpful for problem variant 2. That is because deduplicating the least-salient blocks can avoid a drop in accuracy, but cannot improve the accuracy. Therefore, we propose *reversing the order and visiting the most salient blocks first*.

Our second observation is that the value of $max(u_j - u^{adjusted}, u_i)$ will increase with the base model's utility u_i when $u_i > u_j - u^{adjusted}$, which is usually the case since we always prioritize high-utility models with large ϵ to serve as base models for problem variant 2. This shrinks the search space compared to using a low-ϵ base model. Therefore, we run the first iteration with the accuracy constraint (i.e., the lowest allowed accuracy of the deduplicated model) to be

Algorithm 6 Iterative Accuracy-Optimized Deduplication (IAOD))

```
1: Input: Target and base models represented by a list of block identifiers B, B̂, sorted by block saliency descendingly; utility drop threshold T; number of blocks m; a positive small value δ.
2: procedure DEDUPLICATION(B, B̂, T)
3:     t ← max(u_j − u^adjusted, u_i);                     ▷ Set initial accuracy drop threshold, a negative value means the M'_j has better accuracy than M_j
4:     while t < T do
5:         S ← ∅
6:         opt ← 0
7:         for i = 1, ..., m do
8:             Replace block B[i] by the most similar block in B ∪ B̂;
9:             Evaluate model on validation set, compute and record utility drop Δu, if not computed before;
10:            if Δu > t then
11:                Rollback;
12:                if Δu < 0 then
13:                    add B[j] to S;
14:                end if
15:            end if
16:        end for
17:        S' ← top_k(S)
18:        Deduplication(S', B̂, t)                         ▷ Applying Alg. 5
19:        if α * u_ij + β * n_ij/|M_j| < opt then
20:            return ;
21:        else
22:            opt ← α * u_ij + β * n_ij/|M_j|
23:            t ← t − δ                                   ▷ Tighten the accuracy constraint
24:        end if
25:    end while
26:    return ;
27: end procedure
```

$max(u_j - u^{adjusted}, u_i)$, and gradually increase the constraint. (However, due to the shrunk search space, we observed that a single iteration suffices to optimize the objective for common α and β values, such as $\alpha = \beta = 0.5$.)

Our third observation is that if replacing a block in M_j by a block in M_i leads to a drop of accuracy exceeding the current accuracy constraint but it significantly improves the target model's performance, this block replacement may still have a positive impact to accuracy, and accumulating the replacement of such blocks (while avoiding the effects brought by other blocks for which the replacement leads to significant accuracy drops) may finally lead to a satification of the current accuracy constraint. Therefore, each iteration has two steps.: (1) Track those blocks that have a positive accuracy impact on the target model but cannot satisfy the current accuracy threshold on their own, and (2) Prioritize those high-potential blocks to be deduplicated.

Algorithm Formalization. A formalized algorithm, called Iterative Accuracy-Optimized Deduplication (IAOD), is illustrated in Alg. 6. In Line 3, the accuracy drop threshold t is initialized to $max(u_j - u^{adjusted}, u_i)$. Line 4 starts the iterations. In each iteration, we first initialize $\mathcal{S}$, the list of blocks that failed to be deduplicated with the current utility drop threshold but have positive impacts on accuracy. Then, the iteration is split into two steps. The first step (Line 8 to Line 16) will perform Greedy-1 deduplication while tracking the blocks' impact on accuracy for failed deduplications that were rolled back and add those blocks with positive accuracy impact to $\mathcal{S}$. Then, the second step (Line 17 to Line 18) will first get the top k blocks from $\mathcal{S}$, and then deduplicate those blocks using Alg. 5. If the objective function is not improved, the algorithm terminates. Otherwise, the accuracy drop threshold t is tightened by deducting a small positive value δ (Line 23), and it continues the iterations.

D. Implementation

In our implementation, a model includes metadata information and parameter tensors. The former includes pointers to the training and validation datasets, and hyper-parameters used in the training process, such as privacy budget, training accuracy, and validation accuracy. Such information is used for clustering models. A model may contain many parameter tensors, e.g., embedding vectors, weight matrices, or convolutional filters. In our implementation, each parameter tensor is flattened into a one-dimensional array and then partitioned into blocks of fixed size. The last block will be padded with zeros. Some parameter tensors, such as biases and layer norms, are much smaller than the block size, so they are not partitioned and are managed separately.

A deduplicated model is stored in three parts : (1) A 2-D array that contains the blocks for all models, where each row represents a block flattened into a 1-D array. (2) A dictionary for each model containing tensor shapes and extra weights, such as biases and norms that are much smaller than the block size and are handled separately. (3) A list of block indices for each model. For example, a model with K blocks would have a K-way array, with the i-th element specifying the i-th block's row index in the 2-D array.

Serving a deduplicated model requires reconstructing the model from the blocks and the dictionaries. Whenever a new query requires a model that is not reconstructed, if there is insufficient memory, a victim model will be evicted following the Least Recently Used (LRU) policy to free up memory space for reconstructing the new model. The system then retrieves the model ID, looks up the corresponding model constitution and extra weights, and reconstructs the queried model. The process is accelerated by reusing a reconstructed model with the same architecture and leveraging our array-based indexing to access blocks for replacement.

VI. Experiments

We conducted comprehensive evaluations to demonstrate the effectiveness of our proposed methods.

A. Experimental Settings

1) Workloads: To demonstrate the broad applicability of our proposed approaches, we experimented with a diverse range of model architectures and tasks[1]:

1) Roberta-Base [47] (called Roberta): A Transformer encoder for natural language inference tasks. It contains 277 blocks with a block size of $58,982$ floating points.
2) Vision Transformer-Large (called ViT) [48]: A Transformer model for image classification tasks. It has 288 blocks with a block size of $1,048,576$ floating points.
3) ResNet152 [49] (Called ResNet): A convolutional neural network, with 238 blocks of a size of $262,144$ floating points.

[1]Most of the models in this work are obtained by fine-tuning a pre-trained model on (private) datasets. We focuses on full-finetuning (FFT) that will update all model parameters during the fine-tuning process. The accuracy of FFT outperformed parameter-efficient fine-tuning (PFT) such as Low-Rank Adaptation (LoRA) [44] for scenarios involving large-scale fine-tuning data, low privacy budget, and complex tasks [45], [46], [44]. Our approach can also be applied to deduplicate models with LoRA weights.

These models encompass all major layer types in modern deep learning architectures, ensuring that our proposed algorithms apply to a wide variety of models beyond those explicitly tested. The datasets were pre-partitioned into training, validation, and test sets, which are disjoint with each other.

The *choice of privacy budgets* is contingent upon the data and model architecture. Existing works employ varied ranges. For instance, [1.0, 8.0] and (0, 2] for both natural language and image classification tasks [50], [51], [9], [52], [53], [54], [0.08, 4.6] for image and tabular data classification tasks [55], and [1,16] for image processing tasks [56]. In our work, we adopted ranges (0.0,10.0] for the Roberta-base models and (0, 5.5] for ViT and ResNet, respectively.

Our evaluation for problem variant 1 involves eight scenarios of models shown as A1-A5 and B1-B3 in Tab. I, designed to be diverse and representative. Each scenario involves one cluster of models. The clusters differ in the numbers of $(5-20)$ models, model architectures, training datasets, distribution of epsilons, privacy loss increase threshold ϵ^*, and utility drop threshold u^*. (The latter two constraints are shared by all models in each cluster for simplicity of experiment settings.) Following our base model selection algorithm proposed in Sec. V-A, the first model is selected as the base model for all scenarios except A2. All models in A2 are dangling models, and they will use the model from A3 with $\epsilon = 0.2$ as their base model. Scenarios C1-C4 in Tab. I involve subsets of A1, B2, B3, and A3, respectively, for an ablation study of the base model selection algorithm.

Our evaluation for problem variant 2 involves two representative scenarios as shown in Tab. II, with more scenarios described and evaluated in Sec. J in the Appendix.

TABLE I: The Deduplication Scenarios for Problem Variant 1

	Model	Data	Epsilons	ϵ^*	u^*
A1	Roberta	QNLI [57]	[1.0, 2.0, 4.0, 6.0, 7.0]	1.0	0.015
A2	Roberta	SST2 [58]	[0.3, 0.4, 0.6, 0.8, 1.0]	0.05	0.015
A3	Roberta	5 MNLI parts	[0.2, 0.4, 0.8, 1.6, 2.0]	0.2	0.015
A4	ViT	CIFAR100 [59]	[0.5, 0.6, 0.75, 1.0, 2.0]	0.5	0.020
A5	ResNet	CelebA [60]	[0.4, 0.6, 0.8, 1.0, 2.0]	0.5	0.020
B1	Roberta	QNLI	[1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0]	1.0	0.015
B2	ViT	CIFAR100	[0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95]	0.5	0.020
B3	ResNet	CelebA	[0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2, 2.1]	0.2	0.020
C1	Roberta	QNLI	[5.0, 6.0, 7.0, 8.0, 9.0]	5.0	0.015
C2	ViT	CIFAR100	[0.5, 0.6, 0.75, 1.0, 2.0]	0.5	0.020
C3	ResNet	CelebA	[0.7 0.8, 0.9, 1.0, 2.0]	0.7	0.020
C4	Roberta	4 MNLI parts	[0.4, 0.8, 1.6, 2.0]	0.3	0.015

TABLE II: The Deduplication Scenarios for Problem Variant 2

	Model	Data	Epsilons	ϵ^*	u^*
D1	ViT	CIFAR100	[1, 2, 3, 4, 5]	5.0	0.0000
D2	Roberta	QNLI	[3, 4, 5, 6, 7]	7.0	0.0000

2) Comparison: None of the existing model deduplication mechanisms has considered privacy. Therefore, **to evaluate the effectiveness of our overall deduplication and accuracy validation approach**, we considered the following algorithms, which used our base model selection mechanism to ensure privacy. Unless explicitly noted, they all use gradient magni-

tude for saliency measurement and L2 (Euclidean) distance as similarity measurement:

1) Our **DRD** as formalized in Alg. 5 and its variant, **DRED**.
2) The SOTA **Dedup** [22] algorithm as described in Sec. IV-A.
3) **Mistique** [23] uses MinHash [61] to identify and deduplicate similar model weights without considering accuracy. We improve it to use Dedup's saliency measurement and accuracy validation steps described in Sec. IV-A to provide an accuracy guarantee.
4) The **Greedy-N** algorithm, which we developed as a static variant of **DRD** and **DRED**. Unlike Dedup, it does not stop after a failure of accuracy validation; instead, it will roll back and then move on to process the next batch.
5) We also developed a **Monte Carlo Tree Search (MCTS)** [62] algorithm, where each state represents a deduplication scheme, and an action selects a group of N blocks from the target model to be deduplicated using the most similar blocks from the base model, and run an accuracy validation. If the validation fails, the reward (i.e., storage costs saving) is computed and back-propagated. It repeats the process until a time budget is achieved.
6) Our **IAOD** algorithm for problem variant 2.

In addition, we also considered the following reference baselines.

1) **Original**, which does not perform any deduplication.
2) **Retrain**, which simply finetunes the models using DP-SGD with the new privacy budget achieved by our deduplication approach. This approach provides the upper-bound accuracy with the same privacy budget increase without deduplication.
3) **Model Merge**, which merges two models using different strategies to improve accuracy. In this work, we consider four popular model merging strategies: (1) Logits Ensemble [63], (2) Model Soups [64], (3) Task Arithmetic (TA) [65], and (4) TIES [66]. We mainly use those model merging algorithms as baselines for evaluating our solutions for problem variant 2, which focuses on accuracy improvement. For each test case, we fine-tune the weights of each model in the merging processes and report the best results.

To evaluate our base model selection method, we extended Alg. 2 to allow multiple base models for one target model by iteratively adding qualified base models until the ϵ increase exceeds the limit, called **Multi**. We further extend **Multi** to allow a deduplicated target model to serve as a base model, called **Cross**.

3) Measurements.: We considered the following metrics.

Compression ratio (C.R.). We measure the compression ratio for both individual models and clusters of models. Let $M_1, ..., M_n$ denote a cluster of models including the base model, and $M'_1, ..., M'_n$ represent these models after deduplication. Each model M_i or M'_i represents a set of unique blocks, and $|M_i|$ or $|M'_i|$ represents the number of blocks in the set. The overall compression ratio is computed as $|M'_1 \cup ... \cup M'_n|/|M_1 \cup ... \cup M_n|$. For a deduplicated model M'_j with its base model M_i , the compression ratio is defined as $|M'_j - M_i|/|M'_j|$, *which is the smaller the better.*

Accuracy. We measure the actual accuracy of individual models. In the ablation studies, we use $\max(\Delta u)$ to represent the maximal accuracy drop of individual models in the group.
Model Inference Latency. A great benefit brought by the reduction of the memory footprint of models is the saving of the latency for serving these models in resource-constrained environments where not all models can fit into memory, which is measured in seconds.
Privacy Budget. The privacy budget for fine-tuning a model on a dataset using DP-SGD contains both ϵ and δ. This paper focuses on meeting the constraints on ϵ, which dominates the privacy budget. These delta values are sufficiently small, set to the inverse of the training dataset size, ranging from 6.1×10^{-6} (CelebA) to 2.0×10^{-5} (CIFAR100). For deduplicated models, the epsilon and delta values are computed according to Theorem 3.1.

4) Experimental Environments: The deduplication experiments are run on a machine with an Intel(R) Xeon(R) Silver 4310 CPU (2.10 Hz) with 128 GB memory and one NVIDIA A10 GPU with 24 GB GPU memory. To simulate a resource-constrained environment, the model serving experiment for the multi-user many-model scenario is run on a c5a.2xlarge instance (8 CPU cores, 16 GB main memory). All other model-serving experiments are run on a c5a.xlarge instance (4 CPU cores, 8 GB main memory).

B. Evaluation for Problem Variant 1

1) Overall Results: We first compared the **compression ratio** and the required **number of validations** of various baselines on scenarios A1-A5 and B1-B3. The results, presented in Tab. III, showed that our `DRD` and `DRED` achieved 1.9 to 3.7× better compression ratio than Mistique, and 1.3 to 3.3× better than Dedup. Dedup and Mistique sometimes achieved poor compression ratios (e.g., $> 95\%$ for A2 and A3) with very small validation numbers. That is because of their early stop mechanism (i.e., immediately stop in case of a validation failure). We also found that MCTS requires numerous validation steps yet fails to achieve a low compression ratio due to the large search space. For Greedy-N, it is challenging to automatically tune the value of N. When we tune N to achieve a competitive compression ratio, the validation number (a constant $|M_t|/N$) is usually worse than `DRD` and `DRED`. In conclusion, our `DRD` and `DRED` achieved better trade-offs between the model compression ratio and the validation times.
Deduplication Overhead. Note that most of the deduplication time is spent on accuracy validation. For example, each validation takes 85.6 seconds for Roberta on QNLI, 74.4 seconds for Roberta on SST2, 58.3 seconds for Roberta on one MNLI partition, 84.7 seconds for ViT on CIFAR100, and 84.8 seconds for ResNet on CelebA. It highlights the importance of balancing both objectives.
Impact on Individual Models. Fig. 1 illustrated each model's **compression ratio** and **accuracy** after applying various deduplication approaches to individual models in A1-A5 of Tab. I. The results confirmed the storage benefits achieved by our proposed `DRD` and `DRED` methods. Considering individual models, DRED outperformed Dedup-20 and Mistique-20 by

TABLE III: Comparison of Different Deduplication Algorithms. Compression Ratios (C.R.) are in %.

	A1		A2		A3		A4		A5		B1		B2		B3	
	C.R.	#Val.	C.R.	#Val.	C.R.	#Val.	C.R.	#Val.	C.R.	#Val.	C.R.	#Val.	C.R.	#Val.	C.R.	#Val.
Dedup-20	48.1	42	95.7	18	**56.4**	38	**32.1**	60	**87.4**	12	40.8	98	34.5	116	85.4	57
Dedup-30	48.1	**30**	95.7	**17**	56.9	**28**	33.4	**43**	94.9	**9**	**39.7**	**70**	**32.6**	**85**	**85.2**	57
Mistique-20	84.4	19	100.0	10	**77.3**	26	63.5	56	97.6	**9**	**76.2**	108	90.4	126	91.9	38
Mistique-30	84.4	**15**	100.0	10	78.0	**21**	63.5	**38**	**94.9**	10	88.6	**72**	**88.8**	**81**	**87.6**	38
MCTS-20 (Ours)	**39.9**	257	82.3	**246**	36.9	391	33.4	909	**35.9**	743	23.3	1280	20.4	2266	**39.7**	2628
MCTS-30 (Ours)	41.1	**107**	**58.4**	307	**35.4**	**331**	**32.0**	**571**	42.2	**447**	**20.1**	**1031**	**20.2**	**1417**	40.7	**1031**
Greedy-20 (Ours)	28.5	44	**29.4**	55	**24.4**	44	**30.0**	60	**33.3**	48	**16.1**	99	**18.6**	135	**37.2**	228
Greedy-30 (Ours)	**25.1**	**32**	30.4	**40**	29.9	**32**	32.0	**40**	36.9	**32**	16.2	**72**	20.8	**90**	40.8	**152**
DRD (Ours)	**29.7**	**18**	**27.3**	**30**	24.4	**16**	34.1	**19**	35.5	**21**	**17.0**	**55**	24.7	**50**	**37.6**	**114**
DRED (Ours)	31.6	26	35.6	36	**23.9**	20	**33.4**	26	**34.2**	25	17.7	59	**24.6**	56	37.7	149

up to 35×. In addition, the Retrain baseline (i.e., full fine-tuning) achieved the best accuracy within the same privacy constraint. However, it does not perform any deduplication. Other baselines achieved similar accuracy to ours since we are using the same accuracy constraints, except that in a few cases where, Dedup and Mistique stopped early due to an accuracy validation failure, leading to high accuracy accompanied by a significantly worse compression ratio (i.e., close to 100%, which means zero compression) than our approaches.

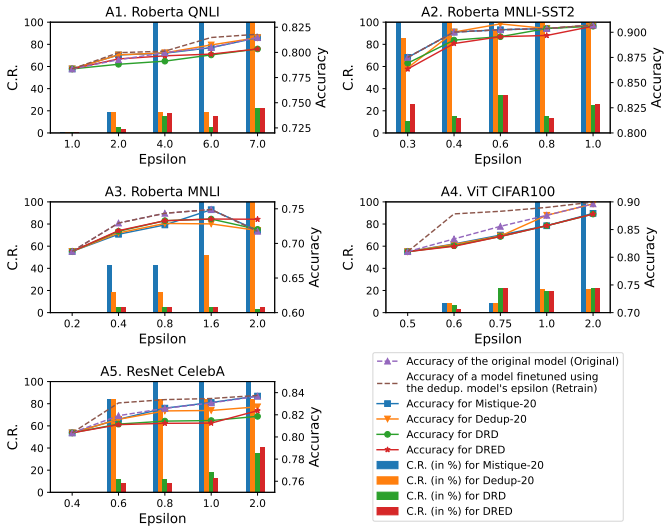


Fig. 1: Compression ratio (C.R.) for individual models in A1 to A5. C.R. of the base model is not affected by deduplication and is thus not shown. A2's base model is the first model in A3.

Model Loading and Inference Speedup. As illustrated in Fig. 2, we compared the latency with and without applying our proposed deduplication approach (based on DRD) for A2, A3, and B1-B3, in AWS c5a.xlarge instance with EBS gp2 SSD and EBS magnetic HDD, as described in Sec. VI-A4. (A1, A4, and A5 are subsets of B1, B2, and B3, and their performance in this experiment is similar to A2 and A3, which are omitted due to space limits.) For this experiment, we measured and broke down the overall latency into loading (including reconstruction) and inference latency for serving 100 inference queries for each scenario. Each query involves a model randomly sampled from the models in the corresponding scenario. The results showed that our approach achieved significantly better speedups in inference for scenarios that involve more models. It achieved 5.6× and 14.1× speedup of the overall latency in SSD and HDD respectively for serving 10 Roberta models in B1, 4.3× to 31.0× for serving 10 ViT models in B2, and 2.2× to 18.8× for serving 20 ResNet models in B3. Even for smaller scenarios such as A2 and A3, our deduplication approach achieved around 1.3× speedup for both SSD and HDD. The results demonstrated that by deduplicating the models' weights, the overall memory footprint required to serve multiple models can be significantly reduced, bringing lower I/O overheads and lower cache misses.

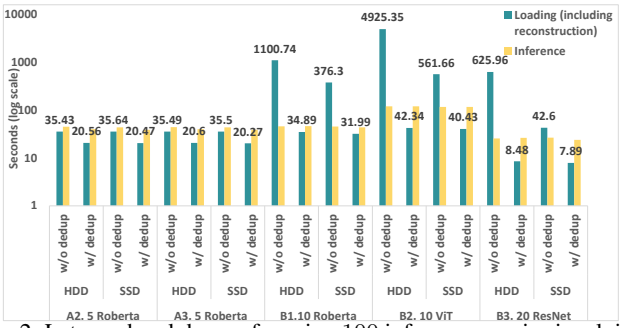


Fig. 2: Latency breakdown of serving 100 inference queries involving multiple models randomly w/o and w/ deduplication (DRD is used, and latency is represented in log scale)

2) Base Model Selection: We compared the overall compression ratio and increased privacy costs of our base model selection algorithm (Alg. 3) to the Multi and Cross baselines on five distinct scenarios A2, A3, and B1 to B3, as shown in Fig. 3. To allow Multi and Cross, we relax the ϵ^* of each model to be the sum of the first three models in its scenario, while u^* is kept the same. Our Alg. 3 achieved the lowest overall increase in privacy costs, which is 1.7 to 4.3× lower than Cross and 1.0 to 3.5× lower than Multi. Our approach also achieved 1 to 1.7× and 1 to 1.8× better compression ratios than Multi and Cross respectively. Taking A3 as an example, it achieved 1.7× better compression ratio and 1.5× and 1.75× lower privacy costs compared to Multi and Cross.

C. Evaluation for Problem Variant 2

1) Deduplicating individual pairs of models: We first compare the resulting compression ratio and accuracy of deduplicating individual <base, target> pairs using our proposed IAOD methodology to various baselines, including Original, Greedy-1, and model merging. As illustrated in Fig. 4 and Fig. 5, **IAOD** provides a better trade-off between accuracy and storage efficiency. Fig. 4 showed that **IAOD** can achieve up to 1.334% accuracy improvement compared to Original and 1.292% accuracy improvement compared to Greedy-1. Fig. 5

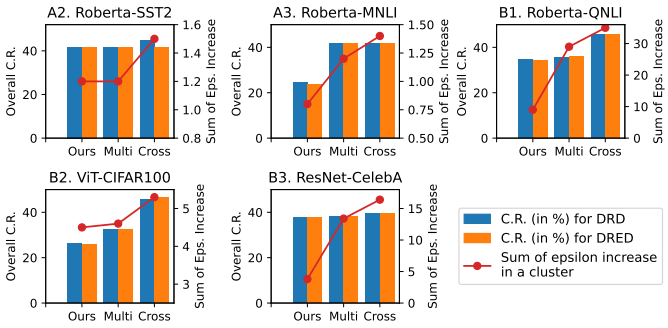


Fig. 3: Comparison of Base Model Selection Strategies. We need an accuracy comparison for this figure.

showed that **IAOD** can achieve 77% to 96% compression ratio for varying target models. **IAOD** achieves significantly better accuracy compared to Original and Greedy-1, and significantly better compression ratio compared to model merging. It demonstrates that **it is possible to use model deduplication to achieve accuracy improvement and storage compression for target models at the same time.**

In addition, for merging these ViT models, among various model merging strategies, we found that TIES underperforms in all cases. Model Soups and TA achieved similar performance, which outperformed Ensemble and TIES in most cases. However, for those challenging cases where target models are trained with large privacy budgets, Ensemble slightly outperformed all other model merging strategies. Details are plotted in the Appendix, Sec. I. For those challenging cases, **IAOD** achieves the best accuracy among all baselines.

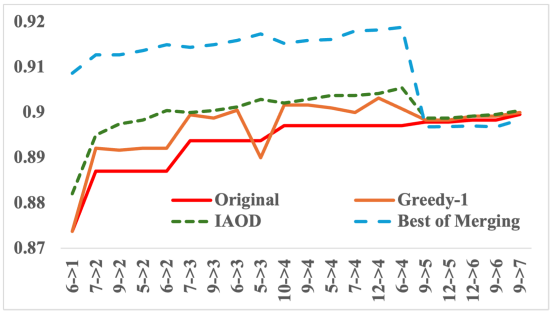


Fig. 4: Comparison of the target model's accuracy for processing 20 randomly sampled pairs of ViT models (sorted by Original's accuracy), with each pair labeled in the form of $\epsilon_{base} \rightarrow \epsilon_{target}$

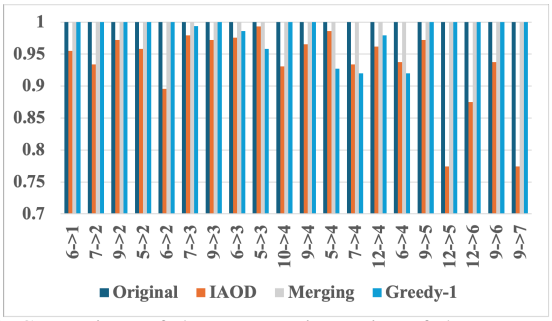


Fig. 5: Comparison of the compression ratios of the target models for processing 20 randomly sampled pairs of ViT models, with each pair labeled in the form of $\epsilon_{base} \rightarrow \epsilon_{target}$

2) Deduplicating Representative Scenarios: We further investigate two representative scenarios, D1 and D2 (Tab. II). The results are illustrated in Tab. IV and Tab. V.

For D1, we observed that **IAOD** improved the accuracy of models M1, M2, M3, and M4 by up to 1.13% compared to Origin and Greedy-1, while achieving an overall compression ratio of 0.91 for these models, outperforming all other baselines as shown in Tab. IV. Among those models, M1, M2, M3, and M4 have 6, 12, 18, and 65 blocks deduplicated, respectively. The compression ratio of individual models increases with the model's privacy budgets, because the model with a higher privacy budget is more similar to the base model. Model merging generally achieves better accuracy, but it does not perform any compression: each target model is replaced by a model merged from the target model and the base model, without block replacement.

For D2, **IAOD** improved the accuracy of models M1, M2, M3, and M4 by up to 0.4% compared to Origin and Greedy-1. **IAOD** also achieved the second best accuracy for M2, among all baselines, and the third best accuracy for M1, M3, and M4, while gaining a compression ratio of 0.89, significantly better than Greedy-1. M1, M2, M3, and M4 have 6, 2, 10, 100 blocks deduplicated, respectively. Finetuning achieved the best accuracy, however, it will incur an increase in the overall privacy loss from the data owner's perspective, since more training rounds need to run over the training dataset.

3) Base Model Selection: Through our experiments, we found that if we use low-epsilon models as the base model, e.g., the model with $\epsilon = 1$ or $\epsilon = 3$ from D1 or the model with $\epsilon = 3$ or $\epsilon = 5$ from D2 as base models, the deduplication of these scenarios achieved significantly less benefit: it either hardly deduplicates any block or is difficult to achieve any accuracy improvement while ensuring the fairness rule. For D1, the achieved compression ratios are 0.9609, 0.9757, and 0.9123 when using models trained with ϵ being 1, 3, and 5 (our selected) as base models, respectively. For D2, the corresponding compression ratios are 0.9738, 0.9621, and 0.8953 using models trained with ϵ being 3, 5, and 7 (our selected) as base models, respectively. We also observed that the accuracy of deduplicated models also improved with the increase of the base model's ϵ. This demonstrates the effectiveness of our model selection strategy for problem variant 2, which prioritizes high-ϵ (i.e., high-accuracy) models as base models. To further investigate the base model selection strategy to analyze the impacts of high-accuracy base models, in D1, we replaced the base model having $\epsilon = 5$ with models having $\epsilon = 6$ and $\epsilon = 7$, respectively. We found that in both cases, the accuracy and compression ratio have improved. However, using the base model with $\epsilon = 6$ achieved a better compression ratio than $\epsilon = 7$ (89.8% vs. 94.6%) and slightly better accuracy. It is possibly because the model with $\epsilon = 7$ is more disimilar with the models with small privacy budgets and leading to fewer deduplication opportunities. This result may indicate that the assumed model with $\epsilon = 7$ should not belong to the same cluster as the models in D1. It also implies that a good clustering mechanism can ensure that selecting the model with the highest accuracy as the base model for each cluster works well for problem variant 2 in most cases.

More detailed results about this experiment can be found in the Appendix (Sec. J).

The experimental results about the accuracy-focused problem variant 2 showed some promising directions for applications: Using finetuning or merging for creating new models when the overall privacy loss and/or storage costs are not major concerns, and using **IAOD** for other cases.

D. Summary of Findings

Overall, for storage-focused problem variant 1, our proposed work brings 17% to 38% C.R. (compression ratio) for eight representative scenarios, leading to a 31× model serving speedup. In addition, in a realistic scenario that involves 50 models, our work achieved 23.7% C.R., resulting in 28× serving speedup. Even in a challenging scenario that runs multiple heterogeneous models required by a social robot in a resource-constrained environment, our approach achieved an impressive 43× inference speedup brought by a C.R. of 74%. In addition, our dynamic deduplication algorithms DRD and DRED have improved the compression ratio by up to 3.3× compared to the best of existing deduplication algorithms (with our base model selection to ensure privacy). Moreover, our base model selection strategy has reduced privacy costs by up to 3× compared to alternative base model selection designs.

For accuracy-focused problem variant 2, our proposed work achieved up to 1.3% accuracy improvement and up to 10.5% overall C.R. for representative scenarios that involve five RoBERTa and ViT models, balancing better the storage and accuracy objectives than other baselines. We found that deduplication can not only improve model accuracy and achieve storage compression at the same time, while avoiding the overall privacy loss increase that would be incurred by finetuning. Although the achieved improvement in accuracy and storage costs is moderate, it shows a new direction for producing new models with better quality from existing models.

A detailed ablation study is provided in Sec. K. Our other observations include:

Block Size Selection. A smaller block size usually leads to a better compression ratio with higher deduplication overhead. However, a very small block size not only slows down the deduplication, but also overlooks the synergy among adjacent blocks, thus not benefiting the compression ratio.

Saliency Measure. Weight magnitude is easy to measure but its accuracy is sub-optimal. Wanda [33] based on weights and activation is only applicable to linear layers. Fisher information [32] suffers from the value vanishing problem caused by the square of small gradient values. Our proposed gradient magnitude measure is more effective in our target problems than the above approaches.

VII. Conclusions

This work proposed privacy-centric and accuracy-centric redesigns of model deduplication techniques to (1) compress multiple models and alleviate the inference and storage costs of multi-tenant and multi-tasking model serving platforms in resource-constrained environments, while meeting accuracy drop constraints, or (2) improving model's accuracy using deduplication to avoid training overhead and reduce overall privacy depletion caused by training, while reducing the storage costs. To address the problems, we proposed novel techniques, including greedy base model selection strategies and novel deduplication strategies (e.g., DRD/DRED/IAOD).

References

[1] J. Liu, J. Lou, J. Liu, L. Xiong, J. Pei, and J. Sun, "Dealer: an end-to-end model marketplace with differential privacy," *Proceedings of the VLDB Endowment*, vol. 14, no. 6, 2021.

[2] B.-R. Lin and D. Kifer, "On arbitrage-free pricing for general data queries," *Proceedings of the VLDB Endowment*, vol. 7, no. 9, pp. 757–768, 2014.

[3] L. Chen, P. Koutris, and A. Kumar, "Towards model-based pricing for machine learning in a data marketplace," in *Proceedings of the 2019 international conference on management of data*, 2019, pp. 1535–1552.

[4] T. Luo, M. Pan, P. Tholoniat, A. Cidon, R. Geambasu, and M. Lécuyer, "Privacy budget scheduling," in *15th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 21)*, 2021, pp. 55–74.

[5] T. Hunt, C. Song, R. Shokri, V. Shmatikov, and E. Witchel, "Chiron: Privacy-preserving machine learning as a service," *arXiv preprint arXiv:1803.05961*, 2018.

[6] W. Yin, M. Xu, Y. Li, and X. Liu, "Llm as a system service on mobile devices," *arXiv preprint arXiv:2403.11805*, 2024.

[7] S. Lee and S. Nirjon, "Fast and scalable in-memory deep multitask learning via neural weight virtualization," in *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*, 2020, pp. 175–190.

[8] L. Zhou, K. S. Candan, and J. Zou, "Deepmapping: Learned data mapping for lossless compression and efficient lookup," in *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 2024, pp. 1–14.

[9] D. Yu, H. Zhang, W. Chen, J. Yin, and T.-Y. Liu, "Large scale private learning via low-rank reparametrization," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 208–12 218.

[10] D. Yu, H. Zhang, W. Chen, and T.-Y. Liu, "Do not let privacy overbill utility: Gradient embedding perturbation for private learning," *arXiv preprint arXiv:2102.12677*, 2021.

[11] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *CCS*, 2016, pp. 308–318.

[12] K. Benitez and B. Malin, "Evaluating re-identification risks with respect to the hipaa privacy rule," *Journal of the American Medical Informatics Association*, vol. 17, no. 2, pp. 169–177, 2010.

[13] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang, "Membership inference attacks on machine learning: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 11s, pp. 1–37, 2022.

[14] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 1322–1333.

[15] G. Cloud, "Restrict data access using analysis rules." [Online]. Available: https://cloud.google.com/bigquery/docs/analysis-rules

[16] snowflake, "Differential privacy in snowflake," https://docs.snowflake.com/en/user-guide/diff-privacy/differential-privacy-overview#label-diff-privacy-loss-budgets.

[17] S. Zhang and X. He, "Dprovdb: Differentially private query processing with multi-analyst provenance," *Proceedings of the ACM on Management of Data*, vol. 1, no. 4, pp. 1–27, 2023.

[18] Z. Xu and Y. Zhang, "Advances in private training for production on-device language models," 2017.

[19] D. Desfontaines, "A list of real-world uses of differential privacy," *Ted is writing things*, 2021.

[20] D. Crankshaw, "The design and implementation of low-latency prediction serving systems," Ph.D. dissertation, UC Berkeley, 2019.

[21] D. Crankshaw, X. Wang, G. Zhou, M. J. Franklin, J. E. Gonzalez, and I. Stoica, "Clipper: A low-latency online prediction serving system," in *14th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 17)*, 2017, pp. 613–627.

[22] L. Zhou, J. Chen, A. Das, H. Min, L. Yu, M. Zhao, and J. Zou, "Serving deep learning models with deduplication from relational databases," *Proc. VLDB Endow.*, vol. 15, no. 10, p. 2230–2243, jun 2022. [Online]. Available: https://doi.org/10.14778/3547305.3547325

TABLE IV: Deduplication Results for Scenario D1, using model M5 ($\epsilon = 5$), as the base model.

	Origin	Greedy-1	IAOD	Ensemble	Model Soup	TA	TIES	Finetune
M1 ($\epsilon = 1$) Acc.	0.8738	0.8738	0.8821	0.9009	**0.9095**	**0.9095**	0.8863	0.8957
M2 ($\epsilon = 2$) Acc.	0.8871	0.8871	0.8983	0.9061	**0.9127**	**0.9127**	0.8953	0.8979
M3 ($\epsilon = 3$) Acc.	0.8938	0.8938	0.9029	0.9080	**0.9144**	**0.9144**	0.8980	0.8974
M4 ($\epsilon = 4$) Acc.	0.8971	0.8971	0.9038	0.9099	**0.9153**	**0.9153**	0.8995	0.8983
Overall C.R.	1	0.9714	**0.9123**	1	1	1	1	1

TABLE V: Deduplication Results for Scenario D2, using model M5 ($\epsilon = 7$), as the base model.

	Origin	Greedy-1	IAOD	Ensemble	Model Soup	TA	TIES	Finetune
M1 ($\epsilon = 3$) Acc.	0.8623	0.8623	0.8655	0.8686	0.8651	0.8651	0.8620	**0.8752**
M2 ($\epsilon = 4$) Acc.	0.8653	0.8653	0.8693	0.8689	0.8666	0.8666	0.8644	**0.8755**
M3 ($\epsilon = 5$) Acc.	0.8657	0.8657	0.8682	0.8700	0.8682	0.8682	0.8655	**0.8761**
M4 ($\epsilon = 6$) Acc.	0.8666	0.8691	0.8699	0.8717	0.8688	0.8688	0.8660	**0.8763**
Overall C.R.	1	0.9729	**0.8953**	1	1	1	1	1

[23] M. Vartak, J. M. F. da Trindade, S. Madden, and M. Zaharia, "Mistique: A system to store and query model intermediates for model diagnosis," in *Proceedings of the 2018 International Conference on Management of Data*, 2018, pp. 1285–1300.

[24] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer, "A survey of quantization methods for efficient neural network inference," in *Low-Power Computer Vision*. Chapman and Hall/CRC, 2022, pp. 291–326.

[25] D. Blalock, J. J. Gonzalez Ortiz, J. Frankle, and J. Guttag, "What is the state of neural network pruning?" *Proceedings of machine learning and systems*, vol. 2, pp. 129–146, 2020.

[26] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[27] H. Guan, L. Yu, L. Zhou, L. Xiong, K. Chowdhury, L. Xie, X. Xiao, and J. Zou, "Privacy and accuracy-aware ai/ml model deduplication," *Proceedings of the ACM on Management of Data*, vol. 3, no. 3, pp. 1–28, 2025.

[28] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3–4, p. 211–407, aug 2014. [Online]. Available: https://doi.org/10.1561/0400000042

[29] F. D. McSherry, "Privacy integrated queries: An extensible platform for privacy-preserving data analysis," in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '09. New York, NY, USA: Association for Computing Machinery, 2009, p. 19–30. [Online]. Available: https://doi.org/10.1145/1559845.1559850

[30] C. Dwork, "Differential privacy: A survey of results," in *International conference on theory and applications of models of computation*. Springer, 2008, pp. 1–19.

[31] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning filters for efficient convnets," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=rJqFGTslg

[32] L. Liu, S. Zhang, Z. Kuang, A. Zhou, J.-H. Xue, X. Wang, Y. Chen, W. Yang, Q. Liao, and W. Zhang, "Group fisher pruning for practical network compression," in *International Conference on Machine Learning*. PMLR, 2021, pp. 7021–7032.

[33] M. Sun, Z. Liu, A. Bair, and J. Z. Kolter, "A simple and effective pruning approach for large language models," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=PxoFut3dWW

[34] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, "Gptq: Accurate post-training quantization for generative pre-trained transformers," *arXiv preprint arXiv:2210.17323*, 2022.

[35] R. Cohen, L. Katzir, and D. Raz, "An efficient approximation for the generalized assignment problem," *Information Processing Letters*, vol. 100, no. 4, pp. 162–166, 2006.

[36] B. Doerr, J. Lengler, T. Kötzing, and C. Winzen, "Black-box complexities of combinatorial problems," in *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, 2011, pp. 981–988.

[37] P. Rust and A. Søgaard, "Differential privacy, linguistic fairness, and training data influence: Impossibility and possibility theorems for multilingual language models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 29 354–29 387.

[38] K. Huang, B. Yang, and W. Gao, "Elastictrainer: Speeding up on-device training with runtime elastic tensor selection," in *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*, 2023, pp. 56–69.

[39] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *CoRR*, vol. abs/1312.6034, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:1450294

[40] N. Lee, T. Ajanthan, S. Gould, and P. H. S. Torr, "A signal propagation perspective for pruning neural networks at initialization," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=HJeTo2VFwH

[41] E. Frantar and D. Alistarh, "Optimal brain compression: A framework for accurate post-training quantization and pruning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 4475–4488, 2022.

[42] J. Lin, J. Tang, H. Tang, S. Yang, W.-M. Chen, W.-C. Wang, G. Xiao, X. Dang, C. Gan, and S. Han, "Awq: Activation-aware weight quantization for llm compression and acceleration," in *MLSys*, 2024.

[43] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in *Proceedings of the twentieth annual symposium on Computational geometry*, 2004, pp. 253–262.

[44] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[45] D. Biderman, J. G. Ortiz, J. Portes, M. Paul, P. Greengard, C. Jennings, D. King, S. Havens, V. Chiley *et al.*, "Lora learns less and forgets less," *arXiv preprint arXiv:2405.09673*, 2024.

[46] Z. Bu, Y.-X. Wang, S. Zha, and G. Karypis, "Differentially private bias-term fine-tuning of foundation models," in *Forty-first International Conference on Machine Learning*.

[47] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[48] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[50] F. Tramer and D. Boneh, "Differentially private learning needs better features (or much more data)," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YTWGvpFOQD-

[51] Z. Bu, Y.-X. Wang, S. Zha, and G. Karypis, "Differentially private bias-term fine-tuning of foundation models," in *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*, 2022.

[52] B. Ghazi, N. Golowich, R. Kumar, P. Manurangsi, and C. Zhang, "Deep learning with label differential privacy," *Advances in neural information processing systems*, vol. 34, pp. 27 131–27 145, 2021.

[53] J. Fu, Q. Ye, H. Hu, Z. Chen, L. Wang, K. Wang, and R. Xun, "Dpsur: Accelerating differentially private stochastic gradient descent using selective update and release," *arXiv preprint arXiv:2311.14056*, 2023.

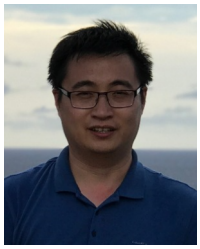
[54] P. Nanayakkara, M. A. Smart, R. Cummings, G. Kaptchuk, and E. M. Redmiles, "What are the chances? explaining the epsilon parameter in differential privacy," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 1613–1630.

[55] D. Bernau, G. Eibl, P. W. Grassal, H. Keller, and F. Kerschbaum, "Quantifying identifiability to choose and audit o in differentially private deep learning," in *Proceedings of the Conference on Very Large Databases*, 2021.
[56] M. Jagielski, J. Ullman, and A. Oprea, "Auditing differentially private machine learning: How private is private sgd?" *Advances in Neural Information Processing Systems*, vol. 33, pp. 22 205–22 216, 2020.
[57] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," *arXiv preprint arXiv:1804.07461*, 2018.
[58] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 1631–1642.
[59] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.
[60] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.
[61] A. Z. Broder, "On the resemblance and containment of documents," in *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*. IEEE, 1997, pp. 21–29.
[62] G. M. J.-B. C. Chaslot, "Monte-carlo tree search," 2010.
[63] R. Gontijo-Lopes, Y. Dauphin, and E. D. Cubuk, "No one representation to rule them all: Overlapping features of training methods," *arXiv preprint arXiv:2110.12899*, 2021.
[64] M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith *et al.*, "Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time," in *International conference on machine learning*. PMLR, 2022, pp. 23 965–23 998.
[65] G. Ilharco, M. T. Ribeiro, M. Wortsman, S. Gururangan, L. Schmidt, H. Hajishirzi, and A. Farhadi, "Editing models with task arithmetic," *arXiv preprint arXiv:2212.04089*, 2022.
[66] P. Yadav, D. Tam, L. Choshen, C. A. Raffel, and M. Bansal, "Ties-merging: Resolving interference when merging models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 7093–7115, 2023.

Hong Guan received the M.S. degree from Arizona State University, USA, in 2020. He is currently pursuing the Ph.D. degree in Computer Science at Arizona State University. His research interests include privacy-aware data management, safe artificial intelligence, and machine learning systems.

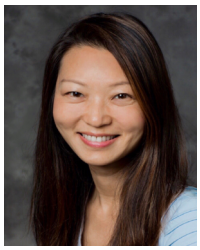
Akshay Mhatre is currently pursuing the M.S. degree in Information Systems at Texas AM University - Central Texas. He has 7+ years of industry experience building and shipping production-grade, end-to-end software systems. His research interests include federated learning, privacy-aware model deduplication, LLM fine-tuning, multimodal learning for healthcare data, and anomaly detection for unreliable or adversarial client behavior.

Lulu Xie is a PhD student working on synthetic data generation and privacy-preserving machine learning. The research focuses on training models under differential privacy guarantees and developing high-fidelity synthetic datasets to enable safe data sharing. Research interests include differential privacy, synthetic data generation and evaluation, and efficient privacy-preserving model training.



Lei Yu is an Assistant Professor in the Department of Computer Science at Rensselaer Polytechnic Institute. Prior to joining RPI, he was a Research Staff Member at IBM Research. He received his Ph.D. in Computer Science from the Georgia Institute of Technology. His research focuses on data privacy and security, trustworthy AI, machine learning systems, and mobile/cloud computing.

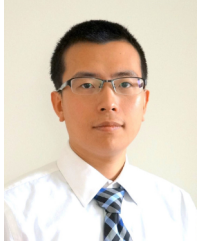
Lixi Zhou received the Ph.D. degree from Arizona State University, USA, in 2025. He is currently working at Google. His research interests focus on AI-driven query optimization, plan representation, and high-performance data processing frameworks.



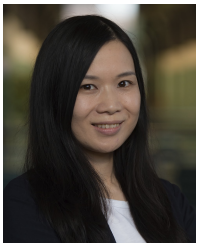
Li Xiong Li Xiong is a Samuel Candler Dobbs Professor of Computer Science and Professor of Biomedical Informatics at Emory University. She held a Winship Distinguished Research Professorship from 2015-2018. She has a Ph.D. from Georgia Institute of Technology. She conducts research in the intersection of data management, machine learning, and data privacy and security. She has received six best paper or runner-up awards. She is an IEEE fellow, an ACM fellow, and an AAAS fellow.

Deepti Gupta is currently an Assistant Professor with Texas AM University-Central Texas. She worked on developing novel security mechanisms, models, and architectures for next-generation smart medical healthcare, smart home, and smart farming. Her research has been funded by U.S. National Science Foundation (NSF) and U.S. Department of Defense (DoD). She also developed novel anomaly detection models and fine-grained access control models to develop secure infrastructure for IoT.

Kanchan Chowdhury is currently an Assistant Professor of Computer Science at Marquette University, Milwaukee, WI, USA, where he directs the DAISS Lab. His research interests include AI for geospatial data, geospatial data analytics, AI for databases, and databases for AI. He published peer reviewed publications on developing deep learning framework for geospatial and temporal datasets, predicting next SQL query during an interaction between a user and a database system, and optimizing machine learning inference queries on in-database machine learning systems.



Xusheng Xiao is an Associate Professor in the School of Computing and Augmented Intelligence at Arizona State University. He is leading the Reliable, Intelligent, Secure, and Efficient (RISE) Software and System lab at ASU. His general research interests span between software engineering and computer security, with the emphasis on AI-enhanced software and system analysis approaches that synergistically combine software analysis and artificial intelligence to improve the reliability and the security of software and computer systems.



Jia Zou is an assistant professor from Arizona State University. She received her Ph.D. degree from Tsinghua University. She has worked in IBM Research as a Research Staff Member. Her research focuses on database systems, AI/ML in databases, applied AI/ML to databases, and privacy-preserving data management. She is a recipent of NSF CAREER award, and has published 30+ papers in top conferences including VLDB, SIGMOD, and ICDE. She has filed 15 patents.

APPENDIX

A. Motivation

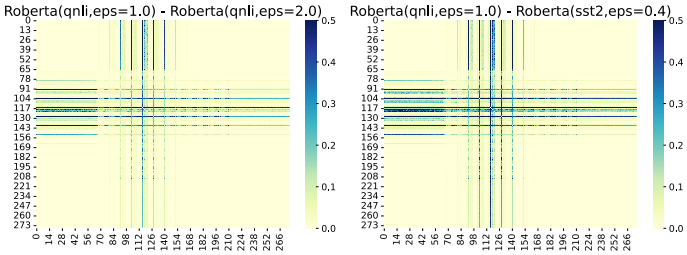


Fig. 6: Weight disparity between DP-SGD-trained models. Left: RoBERTa finetuned on QNLI, with $\epsilon = 1.0$ vs. $\epsilon = 2.0$; right: RoBERTa finetuned on QNLI with $\epsilon = 1.0$ vs fine-tuned on SST2 with $\epsilon = 0.4$). The x- and y-axis represent blocks of two models respectively. Each point represents the disparity score [7] between these blocks. The lower the disparity score, the more similar the two blocks.

In a block-based model deduplication scenario, as illustrated in Fig. 7, the models' weights are partitioned into uniform-sized blocks, e.g., 20 blocks in Fig. 7 ❶. Blocks from a base model (i.e., M_1) are identified to replace blocks in other models (i.e., M_2, M_3, and M_4). As a result, only 8 distinct blocks are needed to reconstruct the models (without significantly degrading models' utilities), which alleviated the broker's operational cost challenges (Optimization1(O1)) in Fig. 7 ❷). However, none of the existing model deduplication works considered the privacy of the models, which poses significant new issues. If any of a model's blocks are replaced (i.e., deduplicated) by blocks from other models trained on the same dataset, the overall privacy budget incurred on the training dataset remains unchanged due to the post-processing rule (Sec. II), which means the total privacy loss on the training datasets will not change, and thus model deduplication will not affect data owners. However, the privacy loss (ϵ) of a deduplicated model may increase (e.g., ϵ_2 increases to ϵ_2' in Fig. 7 ❷) due to the composition of DP (Sec. II), which should be bounded based on its target user's privilege and trust-level on the dataset [17], [1], for the following reasons: (1) The private information leakage to a specific user may increase, which should be bounded by the user's authorized overall ϵ. For example, the increased privacy cost of M_2 (ϵ_2') causes more information leakage to Buyer-1 in Fig. 7 ❸, while Buyer-1 has an overall ϵ limit of 5.5. It will further affect User-2 in Fig. 7 ❹, the end user of M_2, with an ϵ limit of 4. (2) The deduplication may affect both the model ϵ and accuracy, which should satisfy *the (arbitrage-free) fairness rule* that a model trained on the same dataset having a higher accuracy must have a higher ϵ [1]. We further consider **minimizing** the ϵ increase of a model caused by deduplication to prevent the broker from selling high ϵ models to buyers.

In light of this scenario, we define a **novel deduplication problem**: to optimize the storage and privacy costs of all DP-trained models while meeting the pre-defined accuracy and privacy constraint of each model, e.g., in Fig. 7, the privacy cost increase $\epsilon_i' - \epsilon_i$ of model M_i is bounded by ϵ_i^*. While the traditional model deduplication problem focuses on addressing the large search space, our identified problem poses new research challenges as follows:

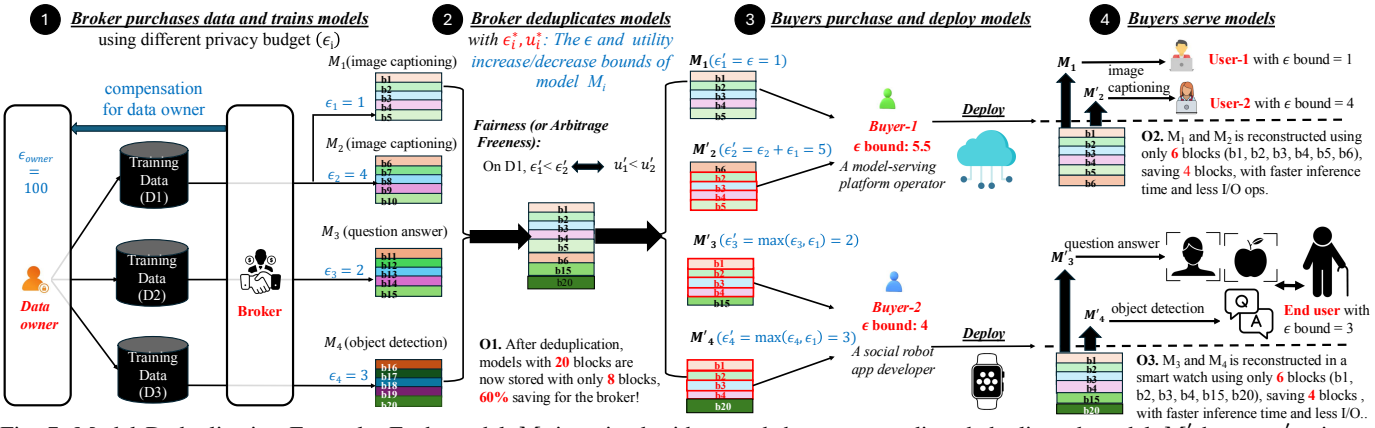


Fig. 7: Model Deduplication Example. Each model M_i is trained with ϵ_i and the corresponding deduplicated model M_i' has an ϵ_i' privacy cost. M_1 (base model) trained on D1 provides blocks for replacing the similar blocks in M_i with privacy and utility bound (e.g., $\epsilon_i' < \epsilon_i + \epsilon_i^*$). ϵ_2' follows sequential composition since M_2 is also trained on D1, while ϵ_3' and ϵ_4' follow parallel composition (see Sec. II). Each dataset (D1, D2, or D3) is a disjoint partition of a logical data collection, for which privacy loss should be assessed collectively.

B. Motivating Example

As shown in Fig. 7, a broker of a model marketplace needs to manage models M_1, M_2, and M_3, M_4 for Buyer-1 and Buyer-2. S/he configures ϵ_i^* (privacy loss increase bound) and u_i^* (utility drop bound) for M_i (e.g., leveraging a privacy provenance system [17]), and deduplicates these models to minimize storage costs and privacy costs (within ϵ_i^* and u_i^* bound). As a result, the number of blocks is reduced from 20 to 8 (**O1** in Fig. 7 2) so that more models can be cached in memory, reducing brokers' operational costs.

Buyer-1 represents a startup MLaaS company, which has an overall ϵ bound of 5.5. In Fig. 7 3, s/he purchases deduplicated models M_1' and M_2' with a total ϵ of 5, and deploys them in a public cloud for serving two users (e.g., User1 and User2 with ϵ bounds of 1 and 4 respectively). MLaaS operators may strive to meet the users' model serving latency requirements as specified in SLOs. Purchasing deduplicated models will reduce the memory footprint for serving both models from 10 blocks to 6 blocks, leading to a significant reduction in model serving latency (**O2** in Fig. 7 4).

Buyer-2 represents an edge application developer, who has an overall ϵ bound of 4. In Fig. 7 3, s/he purchases deduplicated models M_3 and M_4 for deployment on a smart watch for serving a social robot [7] that performs both object recognition task and question answer task at the same time. Assuming User B's edge device can only hold nine blocks, deduplicating models will reduce the number of required blocks from 10 to 6, while maintaining compliance with privacy bounds. It means the deduplication at the broker's side (without significantly impacting the utility) enables the serving of both models in memory, eliminates I/O operations fetching blocks from the disk, and significantly reduces the application response latency (**O3** in Fig. 7 4). After this purchase, Buyer-2 still has a remaining ϵ bound of 1 (i.e., $4 - max(\epsilon_2', \epsilon_3') = 1$).

C. Solution Overview

An overview of our complete solution is illustrated in Fig. 8. In the second box (B2), M1 and M2 are partitioned to the same group because they are trained on the same dataset, while M3 is in a separate group (B2-Step 1). Then, it selects M1 to be the base model, since there is no qualified base model for $M1$ (B2-Step 2). Then, M2 is assigned to M1, while M3, the unused base model in the other group, is also assigned to M1 (B2-Step 3). The next box (B3) shows how M3 deduplicates with its assigned base model M1. M3's blocks are first ordered by saliency ascendingly (B3-Step 1). Then it deduplicates the left half of the blocks by replacing each block using the most similar block from M1, followed by an accuracy validation. If the accuracy drop is within 0.005, it recursively deduplicates the right half (B3-Step 2). Otherwise (B3-Step 3), it rolls back the previous step (B3-Step 4), splits the current group into two, and recursively deduplicates the left half.

D. Proof to Theorem 3.1

E. SVT

We leverage the Sparse Vector Technique (SVT) to reduce the privacy loss when private validation data is used. This approach transforms the validation steps into a series of boolean questions about whether accuracy drops exceed a threshold, maintaining a fixed privacy budget until the number of negative answers reaches a predetermined cut-off value.
(L5) If the validation dataset used in Step 3-3 is private, the accuracy comparison in Step 3-4 will introduce additional privacy costs not considered by existing approaches.
Sparse Vector Technique (SVT) for Validation Using Private Data (Sec. E2). To address L5, we abstract accuracy validation as a boolean question about whether the accuracy drop is above a threshold and apply SVT to provide a fixed budget for a pre-specified number of failed validations (i.e., the boolean questions return negative results).

1) Preliminary: **Sparse Vector Technique (SVT)** is a method in DP for answering a large number of queries while only incurring the privacy cost for those queries that exceed a certain threshold. This is achieved by adding noise to both the threshold and each query output. Let cut-off c denote the maximum number of queries to be answered with result exceeding threshold before halting, Δ represent the sensitivity,

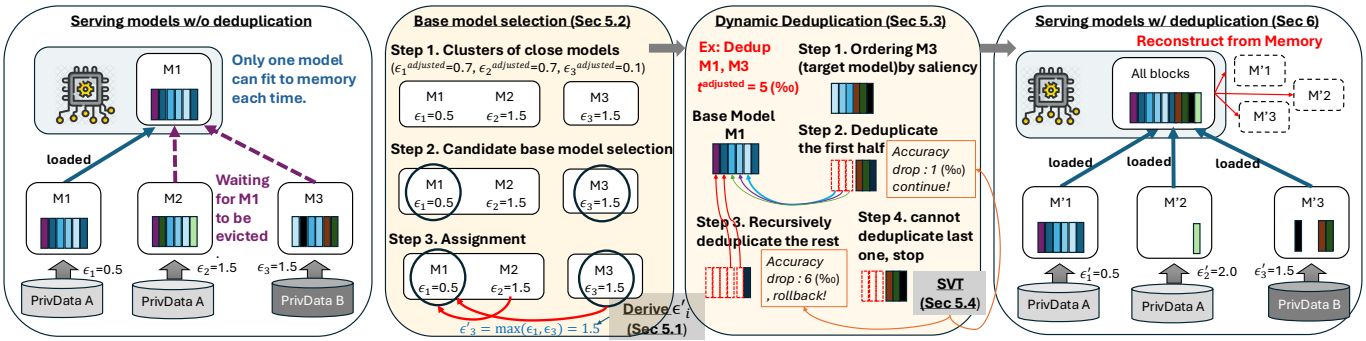


Fig. 8: System Overview.

and ϵ signify the total privacy budget. The privacy budget is divided into two parts: ϵ_1 for the threshold and ϵ_2 for query results, such that $\epsilon_1 + \epsilon_2 = \epsilon$. The algorithm proceeds as follows: noise is sampled from a distribution $\text{Lap}(\frac{\Delta}{\epsilon_1})$ and added to a predetermined threshold. For each query, noise is sampled from a distribution $\text{Lap}(\frac{2c\Delta}{\epsilon_2})$ and added to the query result. If the noisy query result is less than the noisy threshold, it returns a negative answer and the procedure continues; otherwise, it returns a positive answer and increse the counter. The algorithm halts when the counter reaches c. A recommended ratio between the two epsilon values is $\epsilon_1 : \epsilon_2 = 1 : (2c)^{2/3}$.

2) SVT for Validation Using Private Data: Suppose that $f(D, M)$ evaluates the accuracy of a model M on a private validation dataset D, which is disjoint from the training set. Given a classification model, its utility f has a sensitivity of $\frac{1}{|D|}$ because the number of correct predictions changes at most by 1 on two neighbor datasets and the accuracy is calculated as the percentage of the correct predictions. Suppose that M' is the new model after a deduplication step on M. Then, the utility drop, i.e., $f(D, M) - f(D, M')$ has sensitivity $\frac{2}{|D|}$. Each deduplication step is followed by a validation accuracy query on D, and the answer is used to determine whether the deduplication should stop. Accordingly, Sparse Vector Technique (SVT) works by applying Laplace noise and comparing the utility drop with a noisy threshold, as shown in Alg. 7. When the validation dataset is large, the privacy budget required for deduplication can be significantly smaller, as the sensitivity decreases inversely with the size of the dataset.

Alg. 7 employs two distinct privacy parameters: ϵ_1 for noise added to the threshold, ϵ_2 for the validation accuracy comparison. We adopt the recommended ratio of $\epsilon_1 : \epsilon_2 = 1 : (2c)^{2/3}$ as proposed by Lyu et al.. Cut-off c represents the maximum number of times the deduplication process can fail the utility test (i.e., validation), which is pre-specified.

3) SVT for Private Data Validation: We next investigated how using the Sparse Vector Technique (SVT) for model evaluation on private validation datasets affects compression ratio and accuracy. We set the SVT cut-off to 3 and allocated the privacy budget for SVT as the sum of the base and target model budgets, ensuring no additional privacy cost over the entire dataset. We present the overall results for scenarios A1-A5, and B1-B3 in Tab. VI, using Greedy-N, DRD, and DRED. The comparison between Tab. III and Tab. VI showed that

Algorithm 7 ϵ-DP Deduplication with SVT

Input: private validation dataset D; model M to be deduplicated; allowed utility drop threshold T; cut-off c; privacy budget ϵ for validation.
procedure SVT(M, D, T, c, ϵ)
  $\epsilon_1 = \frac{1}{1+(2c)^{2/3}}\epsilon$, $\epsilon_2 = \frac{(2c)^{2/3}}{1+(2c)^{2/3}}\epsilon$;
  $\hat{T} = T + \text{Lap}(\frac{2}{|D|\epsilon_1})$; ▷ Add noise to threshold T
  count $= 0$;
  while true **do**
    Get next set of blocks to be deduplicated $\mathcal{B} = \{b_1, b_2, ..., b_l\}$ proposed by the greedy or dynamic deduplication algorithms.
    if $\mathcal{B} = \emptyset$ **then** Halt;
    end if
    M' = Deduplicate($M, \mathcal{B}$) ;
    $\Delta u = f(D, M) - f(D, M')$; ▷ Compute utility drop
    $v_i = \Delta u + \text{Lap}(\frac{4c}{|D|\epsilon_2})$; ▷ Add noise to utility drop
    if $v_i \geq \hat{T}$ **then** ▷ Deduplication fails
      Output M;
      count $=$ count $+ 1$;
      if count $\geq c$ **then** Halt; ▷ Reach cut-off c
      end if
    else ▷ Deduplication succeeds
      Output M';
      $M = M'$;
    end if
  end while
end procedure

the number of validation steps of the Greedy-N algorithm is more constrained if SVT is used, which worsened the compression ratio, leading to a gap ranging from 0.1% (on A3) to 46.9% (on A5, from 30.4% to 77.3%). However, DRD and DRED are more robust to SVT's cutoff on the validation failures, because of their dynamic range selection without compromising compression ratio, for which the gap between the compression ratios w/o SVT and w/ SVT ranges from -4.0% (on A5) to 4.5% (on B2).

We also showed the impacts to individual models on B2 and B3 scenarios in Fig. 9. The results confirmed that SVT usually worsens the compression ratio by up to 16.7% and in certain cases it improves the compression ratio by up to 3.2% (due to randomness introduced by the noisy utility drop comparison). Observations in other scenarios are similar.

4) Impacts of SVT hyper-parameter tuning.: We also investigated the impacts of cut-off c in SVT-based validation. (Our configuration of privacy budget for SVT-based validation is consistent with Sec. E3.) Taking the deduplication in the A1 scenario as an example, we varied the value of c from 2 to 7, and recorded the compression ratio, the maximum accuracy drop, and the number of validations in Tab. VII. The results showed that when c reaches a point (e.g., 3 in the example), increasing it will not improve the compression ratio, although a larger cut-off allows for more validation failures

TABLE VI: Comparison of Deduplication Algorithms With SVT. Compression ratios (C.R.) are in %.

	A1		A2		A3		A4		A5		B1		B2		B3	
	C.R.	#Val.	C.R.	#Val.	C.R.	#Val.	C.R.	#Val.	C.R.	#Val.	C.R.	#Val.	C.R.	#Val.	C.R.	#Val.
Greedy-20	28.9	44	**72.7**	22	**27.3**	44	**35.9**	56	**42.1**	47	18.0	97	26.6	127	**40.3**	206
Greedy-30	28.9	**32**	77.3	**20**	29.9	**32**	42.8	**38**	50.0	**32**	**16.3**	**72**	**24.3**	**90**	42.8	**148**
DRD	33.6	**24**	70.4	21	24.9	**20**	37.8	24	**38.1**	**21**	**18.0**	51	**28.8**	51	**42.3**	113
DRED	**32.1**	28	**69.9**	**17**	**24.2**	26	**35.3**	**21**	40.4	24	19.0	**49**	29.1	**46**	43.5	**96**

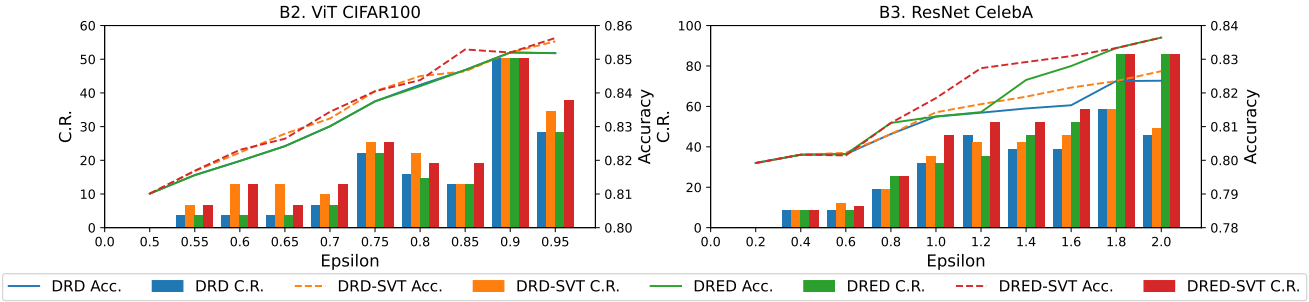


Fig. 9: Deduplication with SVT-based accuracy validation. The compression ratio (C.R.) of the base models, is not not shown.

before it terminates the deduplication process. That's because increasing c will raise the noise level added to the query result, leading to less accurate comparisons.

TABLE VII: Impacts of SVT's Cut-off for the A4 cluster

Cut-off c	2	3	4	5	6	7
C.R.(%)	33.6	31.7	31.7	31.7	31.7	31.7
max (Δu)(%)	0.66	1.14	1.14	1.14	1.14	1.14
Num. of val.	18	23	26	28	28	28

Our SVT-based accuracy validation strategy is also demonstrated as effective in minimizing the privacy loss on private validation datasets, with negligible impacts on compression ratio if coupled with DRD and DRED.

SVT Cut-off. A smaller cutoff incurs smaller noise. However, it limits the maximum number of failure during deduplication. A default choice of 3.0 works well in most of our experiments.

F. Additional Experimental Evaluation on Realistic Scenarios

Two Realistic Scenarios. We also introduce two realistic scenarios corresponding to two types of buyers in Fig. 7.

1) *Multi-User Many-Model Scenario,* where a broker trained a total of 50 distinct models, which is a union of models in Tab. I. These models are sold to a buyer who serves those models to multiple independent users (with different trust and payment levels) concurrently on a cloud. (A similar scenario is to directly sell each model to an independent buyer.)

2) *Single-User Heterogeneous-Model Scenario,* where a broker deduplicated seven models to save operational costs, including four ViT models finetuned on CIFAR10 for object detection, CelebA (for face attribution classification), GTSRB (for traffic sign recognition), and SVHN (for streetview house number recognition), respectively, all with $\epsilon = 2$, and three RoBERTa-base models finetuned on SST2 (for sentiment analysis), IMDB (for review classification), and QNLI (for question-answering natural language inference), respectively, all with $\epsilon = 5$. It chose to use the first ViT model as the base model to deduplicate the rest six models. Then, a buyer purchased the six deduplicated models and deployed them in a resource-constrained environment to perform various tasks required by a social robot.

G. Two Realistic Scenarios

1) Multi-User Many-Model Scenario: The overall deduplication effectiveness w/ public validation datasets (w/o SVT-based validation) is shown in Tab. VIII. Greedy-20 achieved the best compression ratio of 21.2%. However, it requires 561 validations, representing a long latency of more than 10 hours; Mistique-30 achieved the lowest number of validations, 222, but its compression ratio is not ideal, 88.6%. We found our DRD method achieves the best trade-off between compression ratio and validation overheads, with a near-optimal compression ratio of 24.5% and a near-optimal validation number of 264. The results on private validation dataset (w/ SVT-based validation) is similar, the DRD and DRED achieved similar effective compression ratios (32.0% and 32.4%) with Greedy-20 and Greedy-30 (31.4% for both), while the former's validation numbers (256 and 234) are significantly lower than the latter (496 and 362).

When serving all 50 models concurrently on a c5a.2xlarge instance, the end-to-end latency (including total model loading and inferences) for all 100 inference queries (each query randomly selects one model for inference) are shown in the left part of Fig. 10. Deduplication achieved a 28× speedup when HDD is used, and 11× speedup when SSD is used.

2) Single-User Heterogeneous-Model Scenario: Tab. IX demonstrates that our deduplication algorithms (w/o SVT-based validation), such as DRD is able to achieve benefits even for multiple heterogeneous models. DRD outperformed all baselines in minimizing the total required number of validations (50). It also achieved the second-best compression ratio (73.6%). While Greedy-1 achieved the best compression ratio (58.8%), it is at the cost of 2,365 validations (i.e., more than 50 hours). On the contrary, DRD only requires 50 validations, which is a 47× speedup.

TABLE VIII: Comparison of different deduplication algorithms for the multi-user many-model scenario including 50 models.

	Dedup-20	Dedup-30	Mistique-20	Mistique-30	MCTS-20 (Ours)	MCTS-30 (Ours)	Greedy-20 (Ours)	Greedy-30 (Ours)	DRD (Ours)	DRED (Ours)
C.R. (%)	57.1	46.7	87.1	88.6	32.2	28.3	**21.2**	23.4	23.7	24.9
#Val	327	257	308	222	6811	4117	561	386	264	320

TABLE IX: Comparison of different deduplication algorithm for the single-user heterogeneous-model scenario.

Datasets	Roberta-SST2 ($\epsilon = 5$)		Roberta-IMDB ($\epsilon = 5$)		Roberta-QNLI ($\epsilon = 5$)		ViT-GTSRB ($\epsilon = 2$)		ViT-CelebA ($\epsilon = 2$)		ViT-SVHN ($\epsilon = 2$)		Overall	
	C.R.(%)	#Val.	C.R.(%)	#Val.	C.R.(%)	#Val.	C.R.(%)	#Val.	C.R.(%)	#Val.	C.R.(%)	#Val.	C.R.(%)	#Val.
Greedy-1	63.4	210	69.5	211	68.6	210	54.2	578	15.9	578	54.8	578	58.8	2365
Greedy-20	90.3	11	90.8	11	90.3	11	73.7	29	23.1	29	77.6	29	73.8	120
Greedy-30	99.8	7	100.0	8	85.6	7	77.6	20	25.1	20	77.6	20	76.0	82
DRD	87.5	5	88.0	5	84.7	8	77.6	11	21.4	8	79.3	13	73.6	50
DRED	87.5	8	88.0	8	82.8	18	78.2	18	23.5	13	78.2	18	73.8	83

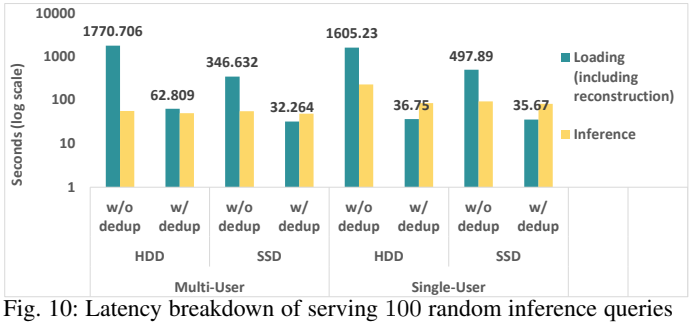


Fig. 10: Latency breakdown of serving 100 random inference queries for the realistic scenarios w/o and w/ deduplication (DRD is used, and latency is represented in log scale)

When serving all 6 models concurrently on a c5a.1xlarge instance, the end-to-end latency for all 100 random inference queries (each query randomly selects one model for inference) is illustrated in the right part of Fig. 10. Deduplicated models achieved a 43× speedup of the overall inference time on HDD storage and a 14× speedup, compared to non-deduplicated models.

H. Further Discussion

In this section, we discuss the scalability of our proposed approach and the dynamic addition/removal of models.

Scalability of the Deduplication Process. First, each pair of base and target models can be deduplicated in parallel with other pairs. Second, the deduplication of a target model (M_j) using a base model (M_i) can be split into two phases: (1) running the deduplication algorithm, where the peak memory is $mem\_dedup = size(M_i) + block\_size$, if the base model M_i is cached in memory; and (2) validating the accuracy, with peak memory being $mem\_valid = size(M_j) + size(feature\_maps) + size(input)$, which mainly sums up the size of the target model, the intermediate feature maps created for inference, and the validation data. Therefore, we have $peak\_mem = \max(mem\ \_dedup, mem\_valid)$.

Model Addition/Removal. Our algorithm can be extended to handle dynamic model addition and removal. A batch of newly arrived models will be dispatched to existing or new clusters based on metadata similarity. Then, it will apply Alg. 2 to each new model. For model removal, the process depends on whether the model is a base model. Non-base models can be safely removed without affecting the system. However, if a base model is deemed no longer useful, two strategies can be employed. The first strategy is to retain its blocks that are in use by other models, minimizing disruption. The second strategy involves archiving the original models in cold storage and re-run the deduplication process for affected models, using an alternative base model.

I. Model Merging Results

Fig. 11 compares the accuracy of merged models using varying model merge algorithms.

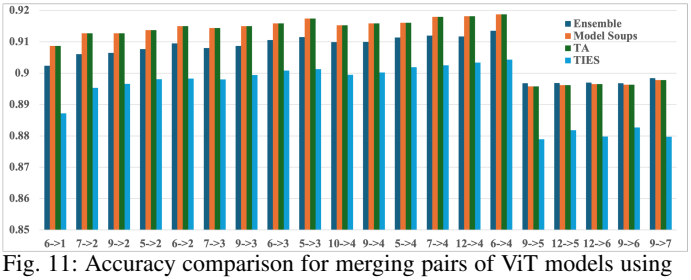


Fig. 11: Accuracy comparison for merging pairs of ViT models using different algorithms

J. More Evaluation Results for Problem Variant 2

In addition to the two representative scenarios presented in Sec. VI-C, we also evaluated more deduplication scenarios for problem variant 2, as listed in Tab. X. Results are illustrated in Tab. XI to Tab. XIII. Overall, in these scenarios, IAOD has achieved significantly better accuracy improvement compared to Original and Greedy-1. Although the accuracy improvement achieved by IAOD is worse than model merging strategies, such as Ensemble, Model Soup, TA, and TIES, IAOD achieved better compression ratios. IAOD has achieved 89.8% to 95.8% compression ratios, representing 4.2% to 10.2% storage cost reduction, outperforming all baselines.

K. Ablation Studies

1) Impact of Deduplication Hyper-Parameters: We compared the impact of block size, saliency measurement, saliency aggregation method, and distance measure. All experiments in this section used the DRED algorithm. We first measured how varying block sizes affect compression ratio, validation steps, and accuracy drop using scenario A4, which consists of five ViT models. We observed that smaller blocks usually resulted in better compression ratios. However, smaller block sizes also lead to more blocks per model, thereby increasing the number of validation steps. In addition, when the block size decreases to a point, the compression ratio will not improve anymore. This is because salient weights exhibit locality and are distributed in multiple small clusters, as verified by Lee et

TABLE X: More Deduplication Scenarios for Problem Variant 2

	Model	Data	Epsilons	ϵ^*	u^*
D3	ViT	CIFAR100	[1, 2, 3, 4, 6]	6.0	0.0000
D4	ViT	CIFAR100	[1, 2, 3, 4, 7]	7.0	0.0000
D5	Roberta	QNLI	[1, 2, 3, 4, 5, 6, 7]	7.0	0.0000

TABLE XI: Deduplication Results for Scenario D3, using model M5 ($\epsilon = 6$), as the base model.

	Origin	Greedy-1	IAOD	Ensemble	Model Soup	TA	TIES
M1 ($\epsilon = 1$) Acc.	0.8738	0.8738	0.8821	0.9009	0.9095	0.9095	0.8863
M2 ($\epsilon = 2$) Acc.	0.8871	0.8921	0.9004	0.9061	0.9127	0.9127	0.8953
M3 ($\epsilon = 3$) Acc.	0.8938	0.9005	0.9013	0.9080	0.9144	0.9144	0.8983
M4 ($\epsilon = 4$) Acc.	0.8971	0.9008	0.9054	0.9099	0.9153	0.9153	0.8995
Overall C.R.	1	0.9748	**0.8984**	1	1	1	1

TABLE XII: Deduplication Results for Scenario D4, using model M5 ($\epsilon = 7$), as the base model.

	Origin	Greedy-1	IAOD	Ensemble	Model Soup	TA	TIES
M1 ($\epsilon = 1$) Acc.	0.87375	0.8775	0.8804	0.9009	0.9095	0.9095	0.8863
M2 ($\epsilon = 2$) Acc.	0.88708	0.89208	0.895	0.9061	0.9127	0.9127	0.8953
M3 ($\epsilon = 3$) Acc.	0.8938	0.8995	0.9000	0.9080	0.9144	0.9144	0.8980
M4 ($\epsilon = 4$) Acc.	0.8971	0.9000	0.9038	0.9099	0.9153	0.9153	0.8995
Overall C.R.	1	0.9783	**0.9462**	1	1	1	1

TABLE XIII: Deduplication Results for Scenario D5, using model M7 ($\epsilon = 7$), as the base model.

	Origin	Greedy-1	IAOD	Ensemble	Model Soup	TA	TIES
M1 ($\epsilon = 1$) Acc.	0.8494	0.8494	0.8517	0.8695	0.8669	0.8669	0.8512
M2 ($\epsilon = 2$) Acc.	0.8598	0.8598	0.8612	0.8699	0.8666	0.8666	0.8607
M3 ($\epsilon = 3$) Acc.	0.8623	0.8642	0.8655	0.8686	0.8651	0.8651	0.8620
M4 ($\epsilon = 4$) Acc.	0.8653	0.8642	0.8693	0.8690	0.8666	0.8666	0.8644
M5 ($\epsilon = 5$) Acc.	0.8657	0.8675	0.8682	0.8700	0.8682	0.8682	0.8654
M6 ($\epsilon = 6$) Acc.	0.8666	0.8691	0.8699	0.8717	0.8688	0.8688	0.86601
Overall C.R.	1	0.9729	**0.9579**	1	1	1	1

al. [7]. When the block size is smaller than the size of these clusters, the compression ratio does not improve.

We further compared the effectiveness of different saliency measures, the saliency aggregation methods, and block similarity measurements, using the deduplication of the A1 scenario as an example. The results in Tab. XIV. It showed that our proposed gradient magnitude measurement outperformed other measurements regarding compression ratio. We also find that profiling of the weight magnitude is the most efficient, taking 113 seconds, while calculating the Wanda score, Fisher information, and gradients using the entire validation dataset takes 145 seconds, 6,576 seconds, and 1,060 seconds, respectively. For saliency aggregation, we found using the l2-norm of weight gradients as the block saliency outperformed l1-norm, l-infinite, and third quartile of weight gradients. For measuring the similarity of two blocks, l2 (Eclidean) distance outperformed other metrics.

2) Impact of Accuracy Threshold: We used various accuracy drop thresholds to deduplicate models in A1, A4, and A5 scenarios, and recorded the compression ratios shown as percentages in Tab. XV. We found that generally, larger accuracy drop thresholds yield better compression ratios. In addition, a negative accuracy drop threshold requires the resulting model to have a higher accuracy than the original one. When the threshold was set to -0.5%, no blocks could be deduplicated in A4 and A5. However, for A1, a few blocks could be deduplicated to improve the accuracy by 0.5%, which indicates that although we focus on constraining the utility drop, deduplication may improve utility in certain cases.

TABLE XIV: Ablation Study of Deduplication Hyper-Parameters

	Values	C.R.(%)	#Val.	$\max(\Delta u)$
A4. block size (#floating points)	4,194,304	53.6	9	0.016
	2,097,152	41.1	16	0.017
	1,048,576	33.4	26	0.018
	524,288	32.5	29	0.020
	262,144	32.12	29	0.020
A1. saliency measurement	Weight Magnitude	63.8	49	0.015
	Wanda	47.3	30	0.015
	Fisher Information	45.9	24	0.015
	Gradient Magnitude	32.8	31	0.019
A1. saliency aggregation method	l1-norm	27.6	26	0.014
	l2-norm	26.1	31	0.013
	l-infinite	30.0	35	0.015
	3rd-quartile	27.1	32	0.015
A1. pairwise distance measurement	l1-distance	25.9	29	0.014
	l2-distance	40.2	31	0.012
	cosine	92.1	63	0.014

TABLE XV: Compression Ratios for Different Accuracy Thresholds

Thresholds(%)	-0.5	0.0	0.5	1.0	1.5	2.0
A1. Roberta	83.7	58.0	23.4	21.9	21.1	20.8
A4. ViT	N/A	84.4	54.1	41.5	35.9	31.0
A5. ResNet	N/A	80.2	44.8	37.8	35.1	31.2