

DeltaMask: Minimizing Communication in Federated Fine-Tuning of Vision Foundation Models

Vasileios Tsouvalas
Eindhoven University of Technology
The Netherlands
v.tsouvalas@tue.nl

Yuki M. Asano
University of Technology Nuremberg
Germany
yuki.asano@utn.de

Aaqib Saeed
Eindhoven University of Technology
The Netherlands
a.saeed@tue.nl

Abstract—Foundation Models (FMs) have revolutionized machine learning with their adaptability and high performance across tasks; yet, their integration into Federated Learning (FL) is challenging due to substantial communication overhead from their extensive parameterization. Current communication-efficient FL strategies, such as gradient compression, reduce bitrates to around 1 bit-per-parameter (bpp). However, these approaches fail to harness the characteristics of FMs, with their large number of parameters still posing a challenge to communication efficiency, even at these bitrate regimes. In this work, we present **DeltaMask**, a novel method that efficiently fine-tunes FMs in FL at an ultra-low bitrate, well below 1 bpp. **DeltaMask** employs stochastic masking to detect highly effective subnetworks within FMs and leverage stochasticity and sparsity in client masks to compress updates into a compact grayscale image using probabilistic filters, deviating from traditional weight training approaches. Our comprehensive evaluations across various datasets and architectures demonstrate **DeltaMask** efficiently achieves bitrates as low as 0.09 bpp, enhancing communication efficiency while maintaining FMs performance, as measured on 8 datasets and 5 pre-trained models of various network architectures.

Index Terms—federated learning, fine-tuning, probabilistic masking, probabilistic filters, foundation models

I. INTRODUCTION

Federated learning (FL) enables collaborative training of neural network models directly on edge devices (referred to as clients), locally with on-device data [1]. FL comprises multiple federated rounds, which involve server-to-client model updates dispatch, local training by clients, and server-side aggregation (e.g., via FedAvg [2]) of clients’ model updates, iteratively performed until model convergence. Despite its appealing properties for users’ privacy, FL requires constant models’ updates transfer between server and clients, which poses a challenge in terms of communication efficiency. This becomes even more critical when the clients are resource-constrained edge devices, which operate under limited transmission bandwidth and strict energy constraints.

Recent advances in FL have led to a variety of methods aimed at enhancing communication efficiency, particularly by reducing the data volume exchanged in each federated round. These strategies often employ gradient compression techniques, including sparsification [3, 4], quantization [5, 6, 7], and low-rank approximation [8, 9], which are pivotal in streamlining data transmission.

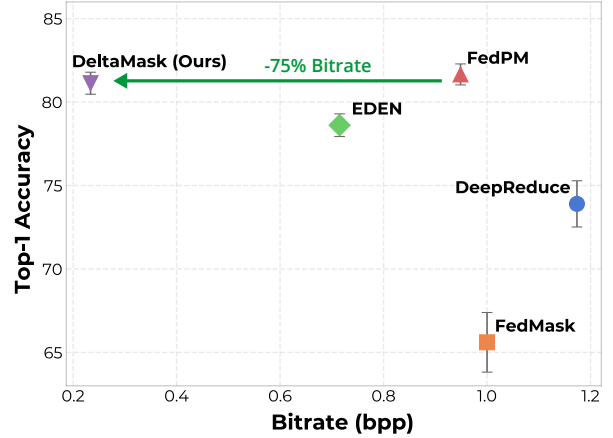


Fig. 1: **DeltaMask** (Ours) vs. state-of-the-art communication-efficient FL techniques when using CLIP pre-trained ViT-B/32, results are averaged over 8 datasets.

Similarly, the “Lottery Ticket Hypothesis” [10] has paved the way for FL training regimes that diverge from traditional weight updates. Here, the focus has shifted toward identifying and cultivating high-potential subnetworks within randomly initialized neural models [11, 12, 13, 14]. Such subnetworks demonstrate good performance without the need for extensive weight adjustments, offering a viable path to minimize FL communication overhead. FedMask [11] and FedPM [14], which learn binary masks on top of random dense networks, are shown to reduce bitrate from 32 to 1 bit-per-parameter (bpp). However, jointly learning effective subnetworks in large, randomly initialized models, can severely affect training duration and model convergence. Furthermore, all aforementioned approaches primarily focus on training models from initial random weights; yet, when fine-tuning pre-trained large models, they do not exploit their unique characteristics to enable FL training at extremely low bitrates (e.g., $\ll 1$ bpp).

Leveraging the advancements in self-supervised learning, vision Foundation Models (vFMs) have brought significant advancement across various machine learning domains with their remarkable representation quality. Models like CLIP [15] and DINOv2 [16] demonstrate rapid adaptability to diverse tasks, achieving unmatched performance in several downstream applications. Notably, recent developments have seen vision FMs, such as the ViT models, expand to billions of param-

ters [17], exemplifying the scale and complexity of modern FMs. In turn, and as an alternative to traditional fine-tuning, masking strategies have emerged [18, 19] in a centralized setting, where selective binary masks are learned on top of frozen pre-trained weights, matching the performance of full fine-tuning. Nevertheless, the high parameter count of FMs inhibits their straightforward expansion into decentralized settings due to the substantial communication overhead [20], even at a bitrate of 1 bpp, thereby limiting their potential to tap into the valuable data available in distributed environments.

To bridge the gap between the high-performance potential of Foundation Models (FMs) and the practical constraints of federated settings, we introduce *DeltaMask*, an approach designed to fine-tune FMs to various downstream tasks in federated settings with significantly reduced bitrate requirements (see Figure 1). Inspired by the sparse mask updates between subsequent federated rounds, which naturally occur due to the rapid adaptability of FMs, our approach combines stochastic masking with probabilistic filters to find high-performing subnetworks within pre-trained FM, while operating in an ultra-low bitrate regime. Moreover, unlike FedPM [14], *DeltaMask* enables an effective control over the bitrate-accuracy trade-off by simply adjusting the bits-per-element (bpe) in probabilistic filters. This paves the way for fine-tuning FMs in federated settings without the massive communication burden caused by their large number of parameters, crucial for scenarios where niche, sensitive data must remain local yet is immensely valuable for learning, such as in healthcare applications. Concisely, the main contributions of our work are as follows:

- We present a simple, yet effective, method termed *DeltaMask*, to allow fine-tuning of FMs in federated settings in a highly communication-efficient manner.
- We combine stochastic binary masks with probabilistic filters to compactly communicate mask updates and reduce bitrate bpp below 0.1.
- Our evaluation across 8 datasets and 5 pre-trained models of various network architectures demonstrate strong effectiveness of *DeltaMask* to fine-tune FMs compared to existing FL techniques.

II. RELATED WORK

Communication-Efficient FL. Enhancing communication efficiency in FL can be achieved through fast adaptation to downstream tasks by utilizing methods such as adaptive optimizers [21] or efficient client-sampling processes [22] that accelerate the convergence rate and, consequently, minimize the data transmission requirement. Alternatively, model updates can be compressed using gradient compression strategies like sparsification [3, 4], quantization [6, 7, 23], and low-rank approximation [8, 9], which reduce the volume of data transmitted during the learning process by shrinking the size of updates communicated in each round. Likewise, training masks over densely initialized networks has been explored to provide communication efficiency between client and server [14, 12]. In HideNseek [12] binary masks were extracted using the

piece-wise sign function over client’s mask scores prior to transmission. Alternatively, in FedPM [14], stochastic mask training was utilized to sample binary masks from locally trained mask probabilities using a Bernoulli distribution, which were then transmitted to the server.

To reduce the bitrate below 1 bpp, FedPM [14] employs arithmetic coding [24] to encode masks based on the sparsity level (frequency of activations). However, arithmetic coding is characterized by computational complexity and intricate implementation, while its efficacy is limited due to the stochasticity of mask sampling, which leads to inconsistent mask activations. In contrast, our approach improves stochastic mask training over pre-trained FMs, introducing a novel lightweight communication protocol that employs probabilistic filters by leveraging the sparsity in subsequent mask changes. Alongside [14], we align *DeltaMask* with gradient compression baselines, such as EDEN [6] and DeepReduce [23], which operate in a similar bitrate regime (≈ 1 bpp).

Masking Neural Networks. Masks learned using gradient descent on pre-trained models can yield subnetworks with performance in parallel to conventional fine-tuning for various downstream tasks, as first demonstrated by [18]. A similar observation was evident in [19], where masking was utilized on language models. Following the Lottery Ticket Hypothesis [10], the same idea was expanded to randomly initialized densely connected networks [25, 26, 27], where mask training was shown to act as an alternative form of weight training. These concepts were applied in FL in [11, 12, 14] to deal with different challenges in FL, such as personalization, communication efficiency and privacy. Compared to these works, our masking method focus on pre-trained foundation models and further reduce the required bitrate below the 1 bpp threshold by communicating the minimal information to reconstruct the clients’ masks through probabilistic filters [28].

III. METHODOLOGY

A. Preliminaries

In this section, we introduce fundamental concepts utilized in *DeltaMask*: (i) probabilistic filters for efficient encoding of clients’ mask updates, and (ii) stochastic mask training, employed to train probabilistic masks on distributed clients and aggregate them on server-side.

Probabilistic Filters. Probabilistic filters are data structures that map a universe of keys, denoted as $\mathcal{U}$, of varying bit lengths, to fixed-size bit values, thereby compacting real-world data representations effectively. They achieve this by using hash functions to transform and store data in a uniformly distributed array, known as the fingerprints $\mathcal{H}$. This compact representation $\mathcal{H}$ facilitates efficient membership checking, with an adjustable rate of false positives — where a non-member might be incorrectly identified as a member — while ensuring zero false negatives. There are multiple variations of probabilistic filters, we focus on *binary fuse filters* (BFuse) [28], which are known for their exceptional space efficiency and computational

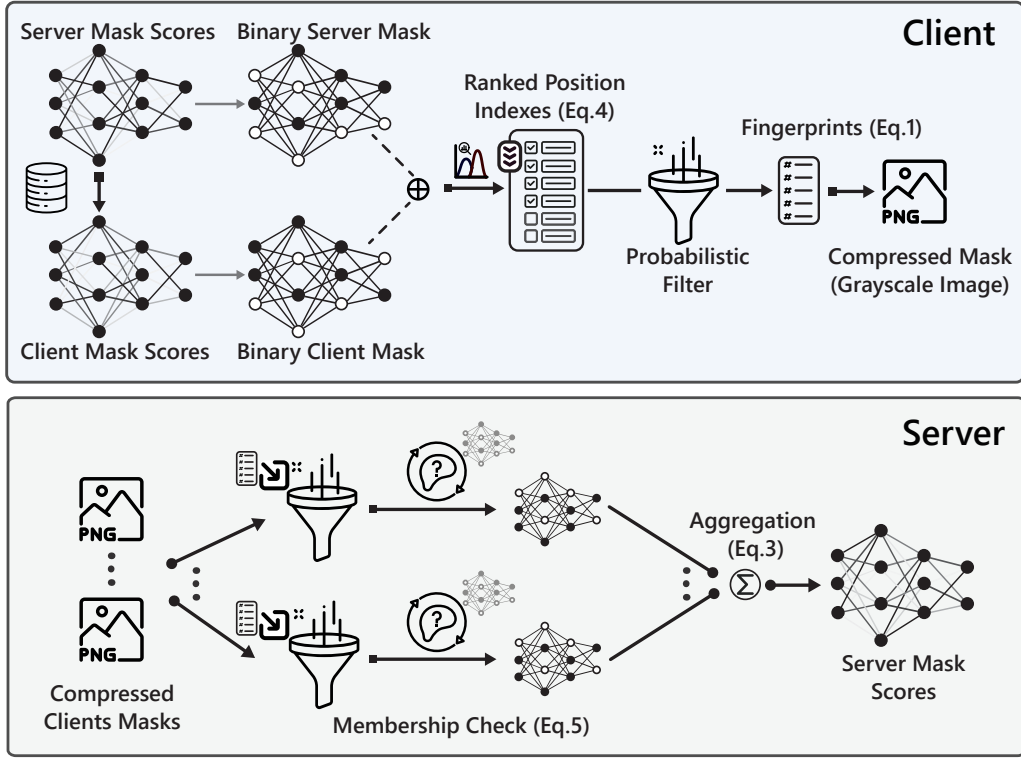


Fig. 2: Overview of DeltaMask: Discrepancies between received and learned masks are ranked using relative entropy. Essential updates are selected via top_{κ} , hashed, and compressed into a grayscale image. The server reconstructs clients' masks through membership queries and updates the global mask using Bayesian aggregation.

effectiveness. These filters offer a space efficiency of 8.62 bits per entry and a low false positive rate (up to 2^{-32}).

Formally, an μ -wise BFuse utilizes m distinct hash functions $h_j : \{0, 1, \dots, 2^n - 1\} \rightarrow \{1, 2, \dots, l\}$, for $j = 1, \dots, \mu$, where l denotes the size of the fingerprints array, $\mathcal{H}$. Let $f : \mathbb{N} \rightarrow \{0, 1, \dots, 2^n - 1\}$ be the fingerprint generation function, mapping each key to an n -bit value. For a given set of keys $\mathcal{U}$, we can compute the fingerprint array $\mathcal{H}$ as:

$$\mathcal{H} = \bigcup_{k \in \mathcal{U}} \phi(k) = \bigcup_{k \in \mathcal{U}} \left(\bigcup_{j=1}^m \{h_j(f(k))\} \right) \quad (1)$$

Here, $\phi(k)$ computes the set of m locations in $\mathcal{H}$ for each key k in $\mathcal{U}$. Once $\mathcal{H}$ is constructed, we can perform a membership check as:

$$\text{Member}(x) = \begin{cases} \text{true}, & \bigoplus_{j=1}^m \mathcal{H}[h_j(f(x))] = f(x) \\ \text{false}, & \text{otherwise} \end{cases} \quad (2)$$

where, $\bigoplus_{j=1}^m \mathcal{H}[\cdot]$ represents the bitwise XOR operation performed on the array values of $\mathcal{H}$, indicated by the hash functions $h_j(f(x))$. The $\text{Member}(\cdot)$ function returns true if the result of the XOR operation over $\mathcal{H}$ matches the fingerprint $f(x)$, suggesting that x is likely a member of the set, and false in all other occasions. Note that while computing a large number of hashes may seem daunting, not all hashing algorithms are computationally intensive. For example, BFuse

use MurmurHash3 [29], which is computationally efficient and exhibits exceptional properties for hashing large data structures into space-efficient arrays (e.g., uniform hash distribution and randomness).

Stochastic Mask Training. Unlike the thresholding mechanisms [11, 12, 9] that creates binary masks by clipping mask scores $s \in \mathbb{R}^d$, stochastic mask training [14], involves drawing a binary mask $m \in \{0, 1\}^d$ from the underlying mask's probability θ using the Bernoulli distribution (noted as $m \sim \text{Bern}(\theta)$). To generate θ from the *unbounded* mask scores s , a sigmoid transformation is applied (i.e., $\theta = \text{Sigmoid}(s)$). Hence, m is used during the forward pass to compute the loss $\mathcal{L}(\cdot)$, and θ is subsequently updated through back-propagation. As the values of s remain *unbounded*, it allows for an unbiased estimation of the true aggregate of the local clients' mask probabilities through Bayesian aggregation [30]. Specifically, it refines the global model at round t in federated setting by treating the stochastic mask's probability $\theta^{g,t}$ as a Beta distribution $\text{Beta}(\alpha_{g,t}, \beta_{g,t})$, with $\alpha_{g,t}$ and $\beta_{g,t}$ initialized to λ_0 . These parameters are updated with the aggregated local binary masks from participating clients (denoted as $\bar{\theta}^{g,t}$), computed as $\alpha_{g,t} = \alpha_{g,t-1} + \bar{\theta}^{g,t}$ and $\beta_{g,t} = \beta_{g,t-1} + K \cdot \mathbf{1}_d - \bar{\theta}^{g,t}$. The aggregated probability mask is then calculated by:

$$\theta^{g,t} = \frac{\alpha_{g,t} - 1}{\alpha_{g,t} + \beta_{g,t} - 2}, \quad (3)$$

where the division is performed element-wise division. For best performance, α and β are periodically reset to $\lambda_0 = 1$ at a rate inverse to the participation rate p [14]. It's important to note that while the model's weight values remain unchanged, the binary mask m selectively activates neurons by element-wise multiplication with the initialized model weights w_{init} , denoted as $w_{k,t} = m^{k,t} \odot w_{\text{init}}$.

B. Federated Masked Fine Tuning of Self-Supervised Foundation Models

We now present the `DeltaMask` training pipeline (see Figure 2). First, clients initialize a neural network $p_{w_{\text{init}}}$ with the weight vector $w_{\text{init}} = (w_{\text{init},1}, w_{\text{init},2}, \dots, w_{\text{init},d}) \in \mathbb{R}^d$, from the pre-trained foundation model. The weight vector w_{init} is kept fixed and never modified during training. `DeltaMask` collaboratively learns a probabilistic mask $\theta \in [0, 1]^d$ using stochastic mask training, such that the function $\mathcal{L}_{\tilde{W}}$ minimize its error rate on a given downstream task ($\tilde{W} = m \odot w_{\text{init}}$). Specifically, at every federated round t , the server samples a set K_t participants ($|K_t|=K$ out of N clients), which individually train their local probability masks $\theta^{k,t}$ on their locally stored datasets D^k , each composed of $|D_k|$ samples.

Instead of communicating the stochastic binary mask m , we significantly reduce the required bits-per-parameter (bpp) during the training process by solely communicating the subsequent key updates (indicated as position indexes set Δ) between the received and trained mask. In particular, we efficiently represent Δ using probabilistic filters and transmit the fingerprint set $\mathcal{H}$ (see Eq. 1) to the server via means of a single gray-scaled image. On server-side reconstruction of clients masks $m^{k,t}$ is feasible via fast membership check using the probabilistic filter (see Eq. 2). These local masks are then aggregated by the server (i.e., by using Eq. 3) to complete the t_{th} round.

Compressing Mask Updates. Our approach utilizes a local training scheme for probability masks, where clients aim to learn a binary mask via stochastic mask training. In brief, clients receive a global probability mask $\theta^{g,t-1}$ at round t , where each client k performs local training and updates the mask via back-propagation. To satisfy $\theta^{k,t} \in [0, 1]^d$ without clipping, we apply a sigmoid operation over the mask's *unbounded* mask scores $s \in \mathbb{R}^d$. Then, clients can utilize a binary mask $m^{k,t}$, i.e., sampled from $\text{Bern}(\theta^{k,t})$ and aim to minimize $\mathcal{L}(p_{w_{k,t}}, D^k)$ over their locally stored data, D^k , after which they back-propagate to update their mask scores.

To enable ultra-low bitrate levels, `DeltaMask` leverages the inherent sparsity in consecutive mask updates in subsequent federated rounds. For a given round t , we deterministically sample a binary mask $m^{g,t-1}$ from the received mask distribution, $\text{Bern}(\theta^{g,t-1})$ using a publicly shared *seed*. This ensures uniformity of the generated binary mask among all clients ($m_i^{g,t-1} = m_j^{g,t-1}$ for any $i, j \in K$). Instead of communicating $m^{k,t}$ (or a compressed version of it), we catalog the index positions of differences between $m^{g,t-1}$ and $m^{k,t}$ to create a set of index differences, denoted as $\Delta^{k,t}$. As the mask updates

sparsity progressively increase during training, we introduce a *top $_{\kappa}$* ranking that selects $\kappa\%$ of $\Delta^{k,t}$ based on their relative entropy between the received and the updated probability masks. As a result, we bring a notion of importance sampling into the communication scheme, similar to [31, 32, 33], helping minimize the distributed mean estimation error under low bitrate regimes. It provides essential updates early in training and conveys detailed information of $m^{k,t}$ without significantly increasing bitrate due to the increasing sparsity of mask differences. Using Kullback–Leibler (KL) divergence as a measure of entropy, $\Delta^{k,t}$ is formally defined as:

$$\Delta^{k,t} = \underset{\text{KL}(\theta^{k,t}, \theta^{g,t-1})}{\text{Sort}} \left\{ i \mid m_i^{g,t-1} \neq m_i^{k,t}, \forall i \in d \right\} [1 : K], \quad (4)$$

where d is the dimension of the probability mask m , and K represents the number of elements to be retained in the sorted set, determined as $\kappa\%$ of $|D_k|$.

Next, we utilize an 4-wise *binary fuse filters* with 8-bit per entry (noted as `BFuse8`) to extract a fingerprint array $\mathcal{H}^{k,t}$ from $\Delta^{k,t}$, following Eq. 1. By doing so, we essentially transition from 32-bit indexes to ≈ 8 -bit hashed entries. In comparison with other probabilistic data structures (e.g., Bloom filters), *binary fuse filters* offer an exceptional balance between space efficiency and false positive rate, while their considerable computational benefits make them ideal for resource-constrained devices often used in FL. We further encode the fingerprints set $\mathcal{H}^{k,t}$ into a pseudo gray-scale image using lossless image compression techniques $\Psi(\cdot)$. Specifically, we employ PNG-like compression (i.e., DEFLATE) due to its wide optimization for edge devices — both the compression algorithm itself and the associated transmission techniques. This approach leverages possible non-uniform distributions of entries across the fingerprints locations to further reduce the bitrate. Note that while other lossless compression schemes could be used here; yet, their impact on bitrate is minor since the primary reduction comes from transitioning from $\Delta^{k,t}$ to $\mathcal{H}^{k,t}$. The resulting image, denoted as $A_{k,t}$, now efficiently encapsulates the mask updates in a visual and compressed format, suitable for transmission to the server.

Bayesian Aggregation of Compressed Masks. Once the local training at round t is completed, the server needs to form a new global probability mask from the received clients' gray-scale images, $A_{k,t}$. Specifically, for each client k , server first decompress the gray-scale image to the extract $\mathcal{H}^{k,t}$ using $\Psi^{-1}(A_{k,t})$ and reconstruct the client's *original* probabilistic filter, `BFuse8k`. The indexes of updates from prior server mask $m^{g,t-1}$ for client k can now be estimated via a membership query across all possible indexes of $m^{g,t-1}$, as follows:

$$\hat{\Delta}^{k,t} = \{i \mid \text{Member}(i) = \text{true}, \forall i \in d\} \quad (5)$$

The clients' stochastic binary sample mask $m^{k,t}$ from $\text{Bern}(\theta^{k,t})$ can be constructed by a simple “*bit-flip*” of $m^{g,t-1}$ in the positions derived from $\hat{\Delta}^{k,t}$. Now, server can compute the estimated aggregated probability mask $\theta^{g,t} = \frac{1}{K} \sum_{k \in K} m^{k,t}$, which is an unbiased estimation of

Algorithm 1 DeltaMask algorithm. BayesAgg refers to Eq. 3.

```

1: Server initialize global model  $G$  with pretrained model weights  $w_{\text{init}}$ .
2: Server initialize mask weights  $\theta^{g,0} \in \mathbf{R}^d$  and Beta priors  $\alpha^{g,0}=\beta^{g,0}=\lambda_0$ .
3: for  $r = 1, \dots, R$  do
4:   Randomly select  $K$  clients to participate in round  $t$ 
5:   for each client  $k \in K$  in parallel do
6:     Sample binary server mask  $m^{g,t-1} \sim \text{Bern}_{t-1}(\theta^{g,t-1})$ 
7:      $\theta^{k,t} \leftarrow \text{ClientUpdate}(\theta^{g,t-1})$ 
8:     Sample binary mask  $m^{k,t} \sim \text{Bern}(\theta^{k,t})$ 
9:      $\Delta'^{k,t} \leftarrow \text{Sort} \{i \mid m^{k,t} \neq m^{g,t-1}\}_{i \in d} [1 : K]$   $\triangleright // \text{ See Equation 4}$ 
10:     $\mathcal{H}^{k,t} \leftarrow \bigcup_{i \in \Delta'^{k,t}} \phi(i)$   $\triangleright // \text{ See Equation 1}$ 
11:     $\text{PNG}^{k,t} \leftarrow \Psi(\mathcal{H}^{k,t})$ 
12:  end for
13:  for each client  $k \in K$  do
14:     $\mathcal{H}^{k,t} \leftarrow \Psi^{-1}(\text{PNG}^{k,t})$ 
15:     $\hat{\Delta}^{k,t} \leftarrow \{i \mid \text{Member}(i) = \text{true}\}_{i \in d}$   $\triangleright // \text{ See Equation 5}$ 
16:     $m^{k,t} \leftarrow m^{g,t-1} \text{XOR } \mathcal{F}$   $\triangleright // \mathcal{F} \text{ is 1 in all positions of } \hat{\Delta}^{k,t} \text{ and 0 otherwise}$ 
17:  end for
18:   $\theta^{g,t} \leftarrow \text{BayesAgg}(\{m^{k,t}\}_{k \in K}, t, \rho)$ 
19: end for
20: procedure CLIENTUPDATE( $\theta$ )
21:   for epoch  $e = 1, 2, \dots, E$  do
22:     for batch  $b \in \mathcal{D}^k$  do
23:       Sample a binary mask  $m \sim \text{Bern}(\theta)$ 
24:        $\theta \leftarrow \theta - \eta \nabla_{\theta} (\mathcal{L}_{CE}(y, P_{m \odot w_{\text{init}}}(y|x_b)))$ 
25:     end for
26:   end for
27:   return  $\theta$ 
28: end procedure

```

the underlying probability mask $\theta^{g,t} = \frac{1}{K} \sum_{k \in K} \theta^{k,t}$ using [30] or a similar strategy. Furthermore, DeltaMask has a bounded estimation error (see Appendix B) defined as:

$$\mathbb{E}_{M^{k,t} \sim \text{Bern}(\theta^{k,t}) \forall k \in K} \left[\|\theta^{g,t} - \bar{\theta}^{g,t}\|_2^2 \right] \leq \frac{d}{4K} \quad (6)$$

In contrast to related approaches, such as FedPM [14] and HideNseek [12], our method enables better control over the model generalization versus bitrate bpp during FL training stage. Specifically, HideNseek transmits raw mask updates, and FedPM uses arithmetic encoding to further reduce bitrate; thus, both methods' control over the bitrate vs. model generalization trade-off depends on the clients' mask structure (i.e., the frequency of ones and zeros). In contrast, DeltaMask allows finer control of this trade-off by adjusting κ and bits per element (bpe) of probabilistic filters. While we primarily focus on 8-bit per entry (bpe) in our probabilistic filters, we provide analysis of bpe in Section V-E. We also provide a complete algorithm in Algorithm 1.

C. Weight Initialization

The neural network $p_{w_{\text{init}}}$ is initialized using weights $w_{\text{init}} = (w_{\text{init},1}, w_{\text{init},2}, \dots, w_{\text{init},d}) \in \mathbf{R}^d$ derived from a pre-trained foundation model, yet, the classification head for downstream tasks is randomly initialized. This means that while the pre-trained backbone offers high-quality features useful across various tasks, the randomly initialized classifier head significantly influences the model's overall performance. Prior research has sampled weights from a uniform distribution around the Kaiming initialization to find highly-performing

subnetworks on randomly initialized network [14, 25, 26, 25]. However, as we focus on pre-trained models, we allow the classification head to adapt during a single round of linear probing, where the rest of the model remains frozen. This yields more stable results and rapid convergence. For a fair comparison, we employ identical weights initialization methods across all considered baselines.

IV. EXPERIMENT SETUP

Datasets and Models. We conduct experiments across 8 diverse image classification datasets, namely CIFAR-10 [34], CIFAR-100 [34], SVHN [35], EMNIST [36], Fashion-MNIST [37], EuroSAT [38], Food-101 [39], and Cars196 [40]. Here, we solely focus on classification tasks, enabling direct comparisons with similar approaches [23, 14, 11, 6, 7]; however it is crucial to note that DeltaMask does not impose restrictions on the underlying downstream task.

We utilize popular ViT architectures pre-trained in self-supervised manner, such as CLIP [15], DINOv2 [16], where, we learn a mask for the last five transformer blocks and keep all prior blocks frozen, similar to [19]. In all experiments with CLIP, we use CLIP ViT-B/32, unless stated otherwise. Apart from ViT architectures, we also evaluate convolution-based model, namely ConvMixer-768/32 [41], where mask is learned for the last five convolutional blocks, similar to ViT. We use pre-trained weights from HuggingFace for each backbone architecture and report results averaged over three runs.

FL Setup. We performed experiments using Flower [42]. In our data splitting procedure, we utilized Dirichlet distribution over classes, denoted as $Dir(a)$, where a refers to the distribution concentration, following [43]. For IID experiments, we fix $a=10$, essentially setting the class distribution per client (referred to as C_p) to ≈ 1.0 (examples from all classes are available across clients), while for non-IID settings, we set $a=0.1$, which results in a $C_p \approx 0.2$. In each round, clients perform a single local training step ($E=1$) and we use a cosine scheduler for top_{κ} mechanism starting from $\kappa=0.8$.

Baselines. We evaluate DeltaMask in terms of accuracy, bitrate (bits-per-parameters), computational complexity and total volume of communicated data between client and server. As baselines, we include both fine-tuning and linear probing, where the latter involves training a classifier on top of a frozen backbone architecture. This establishes an upper performance bound for all remaining baselines. From the domain of gradient compression techniques, we incorporate EDEN [6], DRIVE [7], QSGD [5], and FedCode [44] into our evaluation, which operate in a close bitrate regime (≈ 1 bpp). Additionally, we consider DeepReduce [23] as a baseline owing to its analogous use of bloom filter-based compressor — enabling a direct comparison with BFuse filters performance. From masking strategies within FL, we assess DeltaMask by comparing it against the threshold-based FedMask (without initial pruning) and the stochastic masking FedPM. We use a fixed number of rounds across all baselines to facilitate a direct comparison of data transfer volumes for fine-tuning FMs, as inferred from the reported bitrates (lower is better).

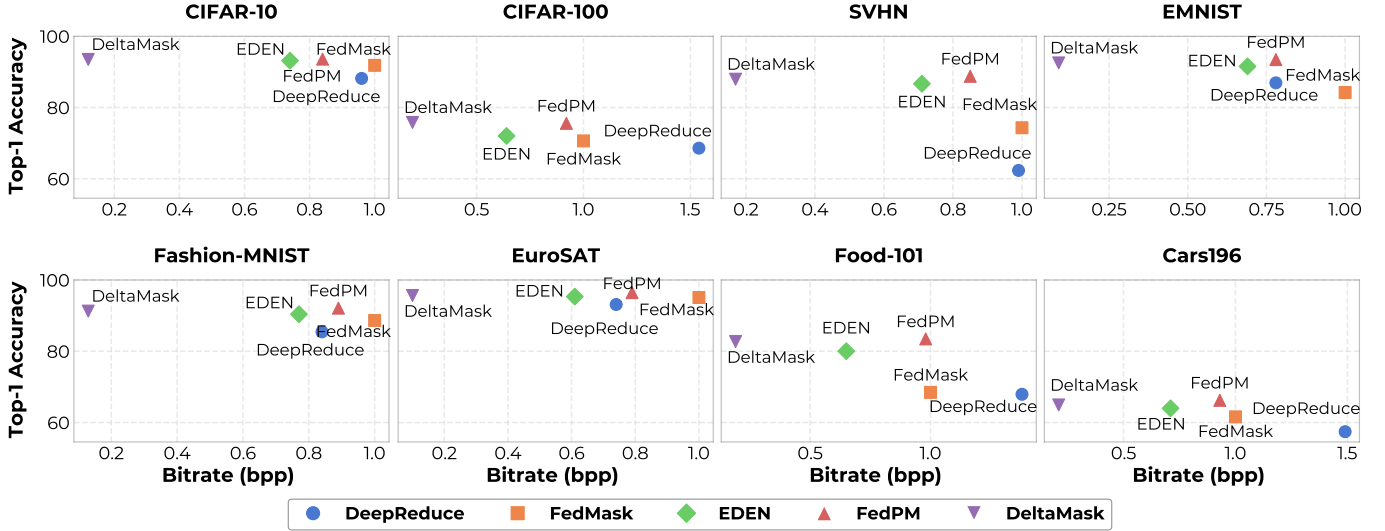


Fig. 3: Performance evaluation of **DeltaMask (Ours)** in terms of average bitrate (bits-per-parameter) during FL training using $Dir(10)$ over classes ($C_p \approx 1.0$ / **IID settings**) for CLIP ViT-B/32. Federated parameters are set to $N=30$, $R=100$, $\rho=1$, and $E=1$. Detailed performance metrics can be found in Table II.

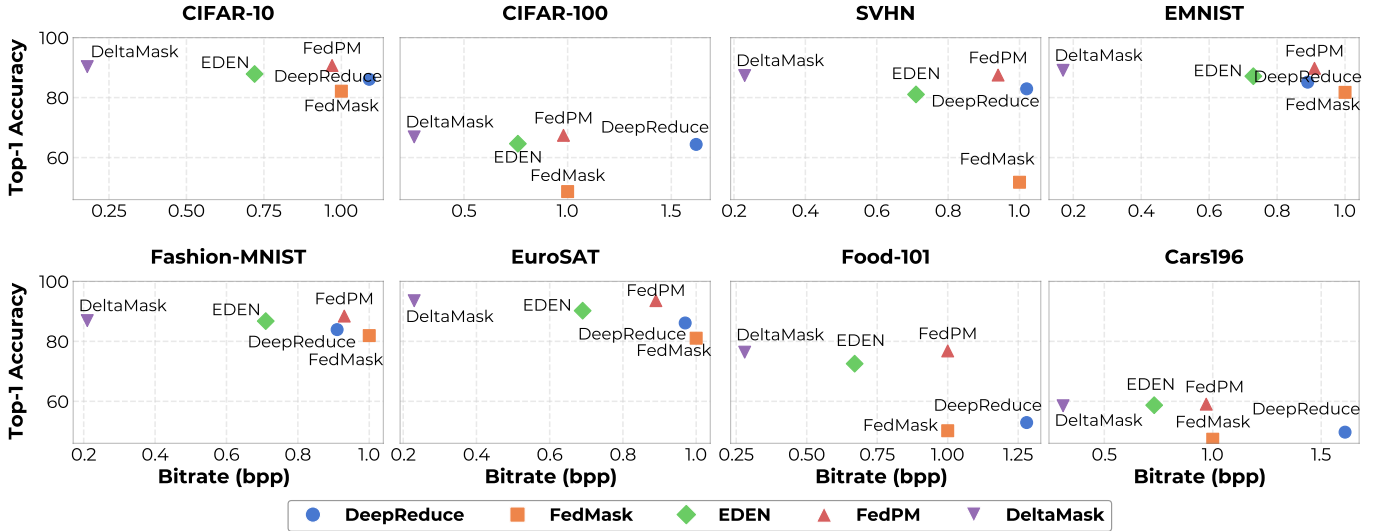


Fig. 4: Performance evaluation of **DeltaMask (Ours)** in terms of average bitrate (bits-per-parameter) during FL training using $Dir(0.1)$ over classes ($C_p \approx 0.2$ / **non-IID settings**) for CLIP ViT-B/32. Federated parameters are set to $N=30$, $R=300$, $\rho=0.2$, and $E=1$. Detailed performance metrics can be found in Table III.

V. RESULTS

A. Main Empirical Findings

For ease of reading, we provide a concise list of our empirical findings below, each linked to its related subsection.

- **Section V-B:** DeltaMask offers significant bitrate reductions — up to 9-fold and 6-fold decreases in IID and non-IID settings, respectively — while maintaining accuracy in par to prior approaches.
- **Section V-C:** DeltaMask excels in environments with strict data transmission limits, achieving the highest accuracy per total bit transmitted compared to all considered methods. Furthermore, its computational efficiency

rivals traditional gradient compression schemes, exhibiting similar encode and decode times.

- **Section V-D:** Our approach is model-agnostic, effective across various architectures, and performs exceptionally well with large models, such as ViT-L/14.
- **Section V-E:** By adjusting κ and the bpe of filters, DeltaMask enables a simple, yet effective control over bitrate versus model generalization trade-off.

B. Bitrate-Accuracy Trade-Off

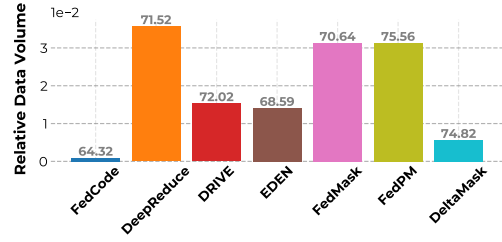
IID Data Split. Here, we focus on IID data distribution, with the number of clients (N) set to 30 and full client participation ($\rho=1$). As depicted in Figure 3, DeltaMask achieves

significant reductions in communication costs compared to the considered baselines - consistently across all datasets. Among the baselines, EDEN requires the least bandwidth, while FedPM attains the highest accuracy; nevertheless, *DeltaMask* reliably matches the accuracy of FedPM. This notable improvement in bitrate over FedPM indicates that mask updates entail significant overhead. Transmitting only essential information via binary fuse filters leads to considerable reductions in bpp (up to approximately $9\times$ less) without compromising on model accuracy. Compared to DeepReduce — which utilizes Bloom filters to transmit the updates — our method underscores the importance of accurate mask reconstruction, as Bloom filters are prone to a higher false positive rate for the same number of hash functions and bits per entry. Further experimental results under conditions of low client participation in IID settings are presented in Appendix A-A.

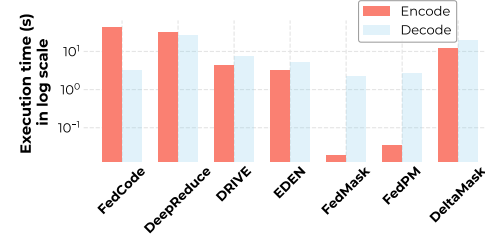
Non-IID Data Split. We now evaluate in a more realistic federated setting, where clients’ data follow a non-IID distribution using $Dir(0.1)$ over classes ($C_p \approx 0.2$). Furthermore, the number of clients (N) set to 30, while we consider partial participation with $\rho=0.2$ (meaning that in each round $\rho \cdot N = 6$ clients are randomly selected). This represents a challenging and realistic scenario, where clients have limited network resources and their data generation processes differ drastically, leading to different data distribution. From Figure 4, we notice similar gains over the baselines as the IID experiments presented in Figure 3; in that *DeltaMask* can maintain a learning procedure that results in better generalizable model, despite having up to a 6-fold reduction in the communication cost. Furthermore, the Bayesian aggregation mechanism, as presented in Eq. 3, is pivotal in achieving high accuracy when $\rho < 1$ under non-IID settings; evident from the increase in performance of DeepReduce compared to FedMask that performs poorly across all datasets. Note that in IID setting with full client participation, this behavior was reversed. We provide additional experiments under a similar non-IID setting with full client participation in Appendix A-B, where the superiority of *DeltaMask* in terms of bitrate reduction is evident.

C. Data Volume & Computational Cost Improvements

We evaluate the impact of *DeltaMask* on computational resources at both client and server levels, as well as the communication efficiency relative to the total data volume transmitted. For this, we perform experiments using CLIP on CIFAR-100 with $N=10$, and measure the encoding and decoding times for various gradient compression schemes: EDEN, DRIVE, and FedCode. Additionally, DeepReduce is included for comparison with *DeltaMask* in terms of efficiency against Bloom-based compression. From masking approaches, we include FedMask and FedPM, omitting their arithmetic encoding step to simplify computational complexity and ensure comparable execution times. All tests are conducted on a CPU, excluding aggregation in decoding time measurements. Data volume is normalized against full fine-tuning size, and we report the necessary volume to reach within 1% of peak accuracy, effectively combining communication efficiency with convergence speed analysis.



(a) Relative Data Volume against fine-tuning size.



(b) Encode/Decode run time on CPU.

Fig. 5: Performance evaluation of **DeltaMask (Ours)** in terms of (a) data volume and (b) encoding/decoding time (with baselines), required for CLIP ViT-B/32 to reach within 1% of peak performance on CIFAR-100. Data volume is normalized over full fine-tuning data size.

Among the evaluated methods in Figure 5, FedCode is the most communication-efficient in terms of data volume; yet, it suffers from longer encoding times and has the lowest model performance across baselines. Additionally, we notice that DeepReduce, utilizing a Bloom-based compressor, struggles with scalability due to longer execution times; in contrast, while *DeltaMask* offer significant improvements in filter construction and query times. FedMask and FedPM offer a compromise between data volume and execution time, with FedPM leading in accuracy among all approaches. Surprisingly, *DeltaMask*, while using slightly more data than FedCode, provides quicker encoding, critical for devices with limited resources, and matches the high accuracy of FedPM with significantly less data communicated. This positions *DeltaMask* as an effective choice for environments with computational and communication constraints.

D. Generalization across Neural Architectures and Pre-training Strategies

Next, we evaluate *DeltaMask* ability to work across various neural architectures pre-trained in a different self-supervised manner. We train masks for downstream task adaptation in a communication-constrained FL environment. For this, we perform experiments with $N=10$ on additional (larger) ViT architectures, namely CLIP-Large and DINOv2-Large, as well as a pure convolution-based architecture, ConvMixer-768/32 on CIFAR-100 as a downstream classification task. In all experiments, we mask the last 5 blocks, as discussed in Section IV. From Table I, *DeltaMask* demonstrates robust adaptability across diverse pre-trained architectures in a FL setup with communication constraints. Notably, *DeltaMask* performance on large ViT architectures yield accuracies near those of fine-

TABLE I: DeltaMask using $Dir(10.0)$ over classes ($C_p \approx 1.0$ / IID settings) across architectures and pre-training strategies.

Metric	CLIP ViT-B/32	CLIP ViT-L/14	DINOv2-Base	DINOv2-Small	ConvMixer-768/32
Fine-tuning	77.35 $\pm$ 0.009	89.07 $\pm$ 0.012	75.01 $\pm$ 0.007	65.55 $\pm$ 0.019	78.52 $\pm$ 0.009
DeltaMask (Ours)	75.82 $\pm$ 0.023	89.48 $\pm$ 0.031	73.36 $\pm$ 0.027	63.01 $\pm$ 0.033	75.31 $\pm$ 0.021
Avg. bpp	0.207 $\pm$ 0.001	0.225 $\pm$ 0.002	0.197 $\pm$ 0.001	0.214 $\pm$ 0.001	0.251 $\pm$ 0.001

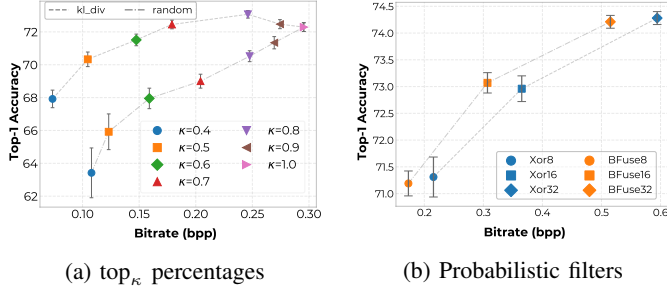


Fig. 6: Impact of top_κ mechanism and probabilistic filter choice in DeltaMask performance. Experiments performed in CIFAR-100 using $Dir(10)$ over classes ($C_p \approx 1.0$ / IID settings). Federated parameters are set to $N=10$, $R=100$, $\rho=1$ and $E=1$.

tuning, with CLIP ViT-L/14 slightly surpassing it. This is significant, considering the communication efficiency depicted by the average bitrate, which remains close to 0.2 bpp across all architectures. ConvMixer-768/32 also adapts well with DeltaMask, showing a modest accuracy reduction while meeting communication constraints. These results reinforce our method’s suitability across diverse architectures, allowing for communication-efficient downstream task adaptation of FMs in a federated setting.

E. Adjusting Bitrate in DeltaMask

Now, we do ablation of fundamental components in DeltaMask: the mechanism sorting the positions of mask updates indexes and our choice of probabilistic filter, assessing the impact of these components on the model’s final accuracy and bitrate. For this, we conduct experiments using CLIP ViT-B/32 with $N=10$ under full participation. In Figure 6a, we compare our entropy-based top_κ sorting with a naive random sampling mechanism. We notice a consistent gap in performance between these approaches, underscoring the pivotal role of importance sampling in achieving generalization, similar to [31]. Surprisingly, increasing κ does not linearly enhance accuracy, with the best results observed at $\kappa=0.8$, beyond which performance diminishes. This suggests that our top_κ approach effectively filters out noise inherent in the stochastic binary mask sampling mechanism by prioritizing updates with higher certainty (i.e., higher probability), while also benefiting from a reduced bitrate due to transmitting less data.

Lastly, we evaluate the performance of various probabilistic filters, focusing on how variations in bits-per-entry (bpe), ranging from 8 to 32 bits, affect the false positive rate. Our analysis includes binary fuse filters (BFuse) and XoR [45] filters, the latter operating under the same foundational principles but being slightly less space-efficient. Figure 6b demonstrates that BFuse filters generally surpass XoR filters in reducing bitrate

without compromising model accuracy, a consistency observed across all experiments. More importantly, we demonstrate that DeltaMask enables an adjustable bitrate based on the bpe selection of the probabilistic filter, offering a potential solution to the resource heterogeneity among clients in FL.

VI. CONCLUSIONS

We introduce DeltaMask a FL technique for efficiently fine-tuning FMs under low bitrate constraints by utilizing stochastic masking instead of conventional fine-tuning and leveraging probabilistic filters for communicating subsequent clients’ mask updates. Our evaluation demonstrates DeltaMask’s effectiveness across a broad range of datasets and FMs, achieving significant communication reductions with performance similar to traditional fine-tuning. Beyond communication efficiency, DeltaMask can extend to offer personalized models in FL, while it can be expanded to adapt a single FM to multiple tasks, each with its uniquely learned masks. Additionally, the hashing operations of probabilistic filters and their false positive rate — interpreted as a *bit-flipping* error in our masks — can potentially enhance privacy in FL. We believe this is a valuable directions for future research.

ACKNOWLEDGEMENT

V.T’s research is funded by the DAIS project, which has received funding from KDTJU under grant agreement No 101007273.

REFERENCES

- [1] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtarik, A. T. Suresh, and D. Bacon, “Federated learning: Strategies for improving communication efficiency,” in *NIPS Workshop on Private Multi-Party Machine Learning*, 2016. [Online]. Available: <https://arxiv.org/abs/1610.05492>
- [2] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in *International Conference on Artificial Intelligence and Statistics*, 2016. [Online]. Available: <https://api.semanticscholar.org/CorpusID:14955348>
- [3] Y. Lin, S. Han, H. Mao, Y. Wang, and W. J. Dally, “Deep gradient compression: Reducing the communication bandwidth for distributed training,” 2020.
- [4] A. F. Aji and K. Heafield, “Sparse communication for distributed gradient descent,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2017. [Online]. Available: <https://doi.org/10.18653/v1/F17-1045>
- [5] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, “Qsgd: Communication-efficient sgd via gradient quantization and encoding,” 2017.
- [6] S. Vargafitk, R. B. Basat, A. Portnoy, G. Mendelson, Y. Ben-Itzhak, and M. Mitzenmacher, “Eden: Communication-efficient and robust distributed mean estimation for federated learning,” in *International Conference on Machine Learning*, 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID:246411641>

- [7] S. Vargaftik, R. Ben-Basat, A. Portnoy, G. Mendelson, Y. Ben-Itzhak, and M. Mitzenmacher, "Drive: One-bit distributed mean estimation," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 362–377. [Online]. Available: <https://proceedings.neurips.cc/paper/2021/file/0397758f8990c1b41b81b43ac389ab9f-Paper.pdf>
- [8] A. Mohtashami, M. Jaggi, and S. U. Stich, "Masked training of neural networks with partial gradients," 2022.
- [9] H. Mozaffari, V. Shejwalkar, and A. Houmansadr, "Frl: Federated rank learning," 2022.
- [10] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," 2019.
- [11] A. Li, J. Sun, X. Zeng, M. Zhang, H. Li, and Y. Chen, "Fedmask: Joint computation and communication-efficient personalized federated learning via heterogeneous masking," ser. SenSys '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 42–55. [Online]. Available: <https://doi.org/10.1145/3485730.3485929>
- [12] A. K. Vallapuram, P. Zhou, Y. D. Kwon, L. H. Lee, H. Xu, and P. Hui, "Hidenseek: Federated lottery ticket via server-side pruning and sign supermask," 2022.
- [13] A. Li, J. Sun, B. Wang, L. Duan, S. Li, Y. Chen, and H. Li, "Lotteryfl: Personalized and communication-efficient federated learning with lottery ticket hypothesis on non-iid datasets," 2020.
- [14] B. Isik, F. Pase, D. Gunduz, T. Weissman, and M. Zorzi, "Sparse random networks for communication-efficient federated learning," 2023.
- [15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.
- [16] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," 2023.
- [17] M. Dehghani, J. Djolonga, B. Mustafa, P. Padlewski, J. Heek, J. Gilmer, A. Steiner, M. Caron, R. Geirhos, I. Alabdulmohsin, R. Jenatton, B. Beyer, M. Tschannen, A. Arnab, X. Wang, C. Riquelme, M. Minderer, J. Puigcerver, U. Evci, M. Kumar, S. Van Steenkiste, G. F. Elsayed, A. Mahendran, F. Yu, A. Oliver, F. Huot, J. Bastings, M. P. Collier, A. A. Gritsenko, V. Birodkar, C. Vasconcelos, Y. Tay, T. Mensink, A. Kolesnikov, F. Pavetić, D. Tran, T. Kipf, M. Lučić, X. Zhai, D. Keysers, J. Harmsen, and N. Houlsby, "Scaling vision transformers to 22 billion parameters," ser. ICML/23. JMLR.org, 2023.
- [18] A. Mallya, D. Davis, and S. Lazebnik, "Piggyback: Adapting a single network to multiple tasks by learning to mask weights," 2018.
- [19] M. Zhao, T. Lin, F. Mi, M. Jaggi, and H. Schütze, "Masking as an efficient alternative to finetuning for pretrained language models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 2226–2241. [Online]. Available: <https://aclanthology.org/2020.emnlp-main.174>
- [20] W. Zhuang, C. Chen, and L. Lyu, "When foundation model meets federated learning: Motivations, challenges, and future directions," *arXiv preprint arXiv:2306.15546*, 2023.
- [21] S. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečný, S. Kumar, and H. B. McMahan, "Adaptive federated optimization," 2021.
- [22] W. Chen, S. Horvath, and P. Richtarik, "Optimal client sampling for federated learning," 2022.
- [23] K. Kostopoulou, H. Xu, A. Dutta, X. Li, A. Ntoulas, and P. Kalnis, "Deepreduce: A sparse-tensor communication framework for distributed deep learning," 2021.
- [24] J. Rissanen and G. G. Langdon, "Arithmetic coding," *IBM Journal of Research and Development*, vol. 23, no. 2, pp. 149–162, 1979.
- [25] H. Zhou, J. Lan, R. Liu, and J. Yosinski, "Deconstructing lottery tickets: Zeros, signs, and the supermask," 2020.
- [26] V. Ramanujan, M. Wortsman, A. Kembhavi, A. Farhadi, and M. Rastegari, "What's hidden in a randomly weighted neural network?" 2020.
- [27] M. M. Aladago and L. Torresani, "Slot machines: Discovering winning combinations of random weights in neural networks," 2021.
- [28] T. M. Graf and D. Lemire, "Binary fuse filters: Fast and smaller than xor filters," *ACM J. Exp. Algorithmics*, vol. 27, mar 2022. [Online]. Available: <https://doi.org/10.1145/3510449>
- [29] A. Appleby, "Murmurhash3," <https://github.com/aappleby/smhasher/wiki/MurmurHash3>, 2016, accessed: 13/11/2023.
- [30] P. A. Ferreira, P. N. da Silva, V. Gottin, R. Stelling, and T. Calmon, "Bayesian signsgd optimizer for federated learning," *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [31] B. Isik, F. Pase, D. Gunduz, S. Koyejo, T. Weissman, and M. Zorzi, "Communication-efficient federated learning through importance sampling," 2023.
- [32] S. Chatterjee and P. Diaconis, "The sample size required in importance sampling," 2017.
- [33] M. Havasi, R. Peharz, and J. M. Hernández-Lobato, "Minimal random code learning: Getting bits back from compressed model parameters," 2018.
- [34] A. Krizhevsky, "Learning multiple layers of features from tiny images," Tech. Rep., 2009.
- [35] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," 2011.
- [36] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik, "Emnist: an extension of mnist to handwritten letters," 2017.
- [37] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *CoRR*, vol. abs/1708.07747, 2017.
- [38] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," 2017.
- [39] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101 – mining discriminative components with random forests," in *European Conference on Computer Vision*, 2014.
- [40] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *2013 IEEE International Conference on Computer Vision Workshops*, 2013, pp. 554–561.
- [41] A. Trockman and J. Z. Kolter, "Patches are all you need?" 2022.
- [42] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, T. Parcollet, and N. D. Lane, "Flower: A friendly federated learning research framework," *arXiv preprint arXiv:2007.14390*, 2020.
- [43] Q. Li, Y. Diao, Q. Chen, and B. He, "Federated learning on non-iid data silos: An experimental study," 2021.
- [44] S. Khalilian, V. Tsouvalas, T. Ozcelebi, and N. Meratnia, "Fedcode: Communication-efficient federated learning via transferring codebooks," 2023.
- [45] T. M. Graf and D. Lemire, "Xor filters: Faster and smaller than bloom and cuckoo filters," *ACM Journal of Experimental Algorithmics*, vol. 25, p. 1–16, Mar. 2020. [Online]. Available: <http://dx.doi.org/10.1145/3376122>

APPENDIX A

We present additional experiments conducted under IID/non-IID settings with varying participation rates (ρ). To ensure a fair comparison, we included both Linear Probing, which involves adapting a single linear classifier atop the (frozen) pre-trained model, and full Fine-tuning, wherein only the layers modified in DeltaMask are fine-tuned. In Table II-III, apart from report models' accuracies across tasks, we include the average *bpp* and accuracy across all tasks for a concise comparison.

A. Additional Experiments in IID settings

From Table II, we note that DeltaMask achieves significant reductions in bitrate, while maintaining performance on par with Fine-tuning. This is particularly evident in scenarios with ρ less than 1, where DeltaMask ability to reduce bitrate without compromising on accuracy highlights its effectiveness in federated learning environments with varying levels of client participation.

B. Additional Experiments in non-IID settings

Table III reveals a notable improvement in DeltaMask performance, especially when the participation ratio ρ is less than 1, with only a 2% accuracy difference compared to Fine-tuning. This is a critical observation, since non-IID data distributions coupled with

TABLE II: Evaluation of **DeltaMask (Ours)** in terms of average bitrate (bits-per-parameter) during FL training using *Dir*(10) over classes ($C_p \approx 1.0$ / **IID settings**) for CLIP ViT-B/32. Federated parameters are $N=30$ and $E=1$. For $\rho < 1$, clients are randomly selected.

	Method	CIFAR-10	CIFAR-100	SVHN	EMNIST	Fashion-MNIST	EuroSAT	Food-101	Cars196	Avg. Acc	Avg. bpp
$\rho = 0.2$	Linear Probing	92.12 $\pm$ 0.007	67.23 $\pm$ 0.011	59.70 $\pm$ 0.016	89.89 $\pm$ 0.008	89.05 $\pm$ 0.010	94.81 $\pm$ 0.009	67.58 $\pm$ 0.014	59.87 $\pm$ 0.016	77.51	-
	Fine-tuning	94.38 $\pm$ 0.013	76.12 $\pm$ 0.019	91.88 $\pm$ 0.012	94.02 $\pm$ 0.018	92.54 $\pm$ 0.009	97.61 $\pm$ 0.015	85.73 $\pm$ 0.017	66.98 $\pm$ 0.011	87.48	32
	FedMask	85.32 $\pm$ 0.033	61.38 $\pm$ 0.057	68.71 $\pm$ 0.046	81.32 $\pm$ 0.024	84.32 $\pm$ 0.044	92.01 $\pm$ 0.025	62.28 $\pm$ 0.037	57.12 $\pm$ 0.029	74.06	1.0
	EDEN	87.11 $\pm$ 0.006	65.89 $\pm$ 0.009	79.16 $\pm$ 0.008	86.36 $\pm$ 0.006	85.21 $\pm$ 0.012	91.24 $\pm$ 0.010	69.59 $\pm$ 0.012	62.07 $\pm$ 0.011	78.33	0.703
	DeepReduce	86.71 $\pm$ 0.071	64.98 $\pm$ 0.091	60.32 $\pm$ 0.061	84.42 $\pm$ 0.044	84.09 $\pm$ 0.057	92.37 $\pm$ 0.041	64.91 $\pm$ 0.043	55.72 $\pm$ 0.078	73.61	1.123
	FedPM	90.31 $\pm$ 0.016	74.66 $\pm$ 0.019	87.03 $\pm$ 0.017	91.42 $\pm$ 0.021	89.79 $\pm$ 0.013	95.57 $\pm$ 0.015	74.80 $\pm$ 0.014	62.19 $\pm$ 0.017	83.22	0.946
$\rho = 1.0$	DeltaMask	89.52 $\pm$ 0.021	74.01 $\pm$ 0.033	86.86 $\pm$ 0.024	92.27 $\pm$ 0.027	89.68 $\pm$ 0.014	94.94 $\pm$ 0.019	74.09 $\pm$ 0.029	61.56 $\pm$ 0.030	82.87	0.197
	Linear Probing	93.97 $\pm$ 0.004	74.11 $\pm$ 0.009	59.26 $\pm$ 0.011	89.40 $\pm$ 0.008	89.47 $\pm$ 0.005	95.35 $\pm$ 0.003	76.64 $\pm$ 0.009	61.72 $\pm$ 0.012	79.99	-
	Fine-tuning	94.50 $\pm$ 0.010	77.35 $\pm$ 0.009	92.72 $\pm$ 0.012	94.89 $\pm$ 0.010	92.98 $\pm$ 0.013	98.24 $\pm$ 0.011	86.72 $\pm$ 0.009	67.23 $\pm$ 0.014	88.08	32
	FedMask	90.84 $\pm$ 0.028	70.64 $\pm$ 0.057	74.32 $\pm$ 0.039	84.22 $\pm$ 0.031	88.64 $\pm$ 0.029	95.09 $\pm$ 0.038	68.46 $\pm$ 0.034	61.59 $\pm$ 0.039	79.23	1.0
	EDEN	93.15 $\pm$ 0.009	72.02 $\pm$ 0.010	86.67 $\pm$ 0.007	91.55 $\pm$ 0.009	90.40 $\pm$ 0.012	95.34 $\pm$ 0.010	80.02 $\pm$ 0.004	63.98 $\pm$ 0.008	84.14	0.691
	DeepReduce	88.17 $\pm$ 0.034	68.59 $\pm$ 0.069	62.34 $\pm$ 0.056	86.92 $\pm$ 0.073	85.44 $\pm$ 0.031	94.12 $\pm$ 0.043	67.92 $\pm$ 0.075	58.42 $\pm$ 0.041	76.53	1.089
	FedPM	93.58 $\pm$ 0.014	75.56 $\pm$ 0.011	88.76 $\pm$ 0.013	93.45 $\pm$ 0.015	92.10 $\pm$ 0.009	96.45 $\pm$ 0.019	83.45 $\pm$ 0.013	65.23 $\pm$ 0.014	86.07	0.872
	DeltaMask	93.50 $\pm$ 0.019	74.82 $\pm$ 0.023	87.95 $\pm$ 0.021	92.52 $\pm$ 0.019	91.27 $\pm$ 0.023	95.64 $\pm$ 0.017	82.73 $\pm$ 0.024	64.94 $\pm$ 0.026	85.44	0.151

TABLE III: Evaluation of **DeltaMask (Ours)** in terms of average bitrate (bits-per-parameter) during FL training using *Dir*(0.1) over classes ($C_p \approx 0.2$ / **non-IID settings**) for CLIP ViT-B/32. Federated parameters are $N=30$ and $E=1$. For $\rho < 1$, clients are randomly selected.

	Method	CIFAR-10	CIFAR-100	SVHN	EMNIST	Fashion-MNIST	EuroSAT	Food-101	Cars196	Avg. Acc	Avg. bpp
$\rho = 0.2$	Linear Probing	84.51 $\pm$ 0.019	49.04 $\pm$ 0.022	43.16 $\pm$ 0.020	82.41 $\pm$ 0.035	86.29 $\pm$ 0.024	91.63 $\pm$ 0.022	51.54 $\pm$ 0.021	47.92 $\pm$ 0.038	67.06	-
	Fine-tuning	92.59 $\pm$ 0.024	70.20 $\pm$ 0.037	87.39 $\pm$ 0.036	92.00 $\pm$ 0.057	88.25 $\pm$ 0.039	95.56 $\pm$ 0.029	79.38 $\pm$ 0.034	60.11 $\pm$ 0.051	83.19	32
	FedMask	83.14 $\pm$ 0.059	51.66 $\pm$ 0.119	51.78 $\pm$ 0.049	83.75 $\pm$ 0.078	85.91 $\pm$ 0.073	90.05 $\pm$ 0.074	53.19 $\pm$ 0.063	51.37 $\pm$ 0.105	68.86	1.0
	EDEN	87.87 $\pm$ 0.037	64.62 $\pm$ 0.106	81.06 $\pm$ 0.081	86.73 $\pm$ 0.050	86.75 $\pm$ 0.055	90.22 $\pm$ 0.062	72.55 $\pm$ 0.056	58.71 $\pm$ 0.034	78.56	0.715
	DeepReduce	86.07 $\pm$ 0.097	64.39 $\pm$ 0.088	82.92 $\pm$ 0.071	85.14 $\pm$ 0.084	83.91 $\pm$ 0.067	86.12 $\pm$ 0.117	52.92 $\pm$ 0.055	49.72 $\pm$ 0.110	73.90	1.173
	FedPM	90.70 $\pm$ 0.045	67.42 $\pm$ 0.095	87.51 $\pm$ 0.079	89.77 $\pm$ 0.095	88.42 $\pm$ 0.092	93.57 $\pm$ 0.067	76.80 $\pm$ 0.076	59.06 $\pm$ 0.098	81.64	0.948
$\rho = 1.0$	DeltaMask	90.32 $\pm$ 0.083	66.90 $\pm$ 0.051	87.36 $\pm$ 0.093	89.09 $\pm$ 0.047	86.91 $\pm$ 0.067	93.54 $\pm$ 0.101	76.39 $\pm$ 0.086	58.52 $\pm$ 0.102	81.13	0.233
	Linear Probing	91.46 $\pm$ 0.026	71.96 $\pm$ 0.025	46.03 $\pm$ 0.017	84.57 $\pm$ 0.031	87.13 $\pm$ 0.018	92.98 $\pm$ 0.016	68.70 $\pm$ 0.028	54.03 $\pm$ 0.032	74.61	-
	Fine-tuning	93.61 $\pm$ 0.048	75.49 $\pm$ 0.052	90.10 $\pm$ 0.063	93.13 $\pm$ 0.037	91.06 $\pm$ 0.041	97.02 $\pm$ 0.034	84.71 $\pm$ 0.013	64.93 $\pm$ 0.063	86.26	32
	FedMask	88.42 $\pm$ 0.051	63.04 $\pm$ 0.081	64.32 $\pm$ 0.073	86.41 $\pm$ 0.039	86.39 $\pm$ 0.044	91.67 $\pm$ 0.031	68.04 $\pm$ 0.050	54.39 $\pm$ 0.089	75.34	1.0
	EDEN	92.14 $\pm$ 0.043	71.65 $\pm$ 0.060	86.28 $\pm$ 0.057	90.87 $\pm$ 0.046	89.94 $\pm$ 0.034	93.26 $\pm$ 0.035	78.79 $\pm$ 0.083	61.18 $\pm$ 0.027	83.01	0.703
	DeepReduce	87.33 $\pm$ 0.052	67.19 $\pm$ 0.061	83.19 $\pm$ 0.048	85.71 $\pm$ 0.082	84.52 $\pm$ 0.075	92.12 $\pm$ 0.060	69.11 $\pm$ 0.092	60.31 $\pm$ 0.094	78.69	1.092
	FedPM	92.99 $\pm$ 0.045	74.34 $\pm$ 0.023	89.35 $\pm$ 0.025	92.65 $\pm$ 0.098	91.33 $\pm$ 0.041	95.37 $\pm$ 0.048	83.69 $\pm$ 0.076	63.65 $\pm$ 0.074	85.42	0.901
	DeltaMask	92.84 $\pm$ 0.083	73.69 $\pm$ 0.051	89.01 $\pm$ 0.085	91.92 $\pm$ 0.089	91.27 $\pm$ 0.055	94.54 $\pm$ 0.103	83.48 $\pm$ 0.081	63.47 $\pm$ 0.096	85.03	0.191

partial client participation closely mirror the conditions of real-world federated settings. Furthermore, our analysis shows that methods using stochastic mask training, such as DeepReduce and FedPM, yield better final model accuracy under non-IID conditions than traditional compression schemes like EDEN or hard-thresholding masking techniques like FedMask. Interestingly, the CLIP ViT-B/32 model excels in non-IID scenarios, underscoring the robust generalization abilities of pre-trained foundation models, which are particularly advantageous in non-IID federated environments. This emphasizes the importance of adapting these models for edge computing, capitalizing on their capability to effectively handle diverse and complex data distributions.

APPENDIX B

Distributed Mean Estimation Error Analysis. We now provide proof of the upper bound on the estimation error of DeltaMask. Recall that we use probabilistic filters to reconstruct clients' binary masks, $m^{k,t} \sim \text{Bern}(\theta^{k,t})$ on server-side, which introduce an independent (across both clients and mask dimensions) “bit-flip” error probability 2^{-p} (p referring to the false positive rate of the filter). We refer to these reconstructed masks as $m'^{k,t}$. Here, our true mean is $\bar{\theta}^{g,t} = \frac{1}{K} \sum_{k \in K_t} \theta_k^t$, while our estimate is $\hat{\theta}^{g,t} = \frac{1}{K} \sum_{k \in K_t} m'^{k,t}$. Furthermore, we use capital letters to refer to random variables, while small letters refer to their deterministic quantities.

We can then compute the error as shown in Eq. 7. We begin by expressing the expected squared L^2 norm of the error $\hat{\theta}^{g,t}$ and $\bar{\theta}^{g,t}$ in (8). From (8) to (9), we use the definitions of $\hat{\theta}^{g,t} = \frac{1}{K} \sum_{k \in K_t} m'^{k,t}$ and $\bar{\theta}^{g,t} = \frac{1}{K} \sum_{k \in K_t} \theta_k^t$. To move from (10) to (11), we use the fact that $M'^{k,t}$ and $M'^{l,t}$ are independent for $k \neq l$; thus the expected value of cross-product terms in the expansion of squared sum of is zero. At (12), we utilize the fact that $M'^{k,t}$ is a Bernoulli variable (meaning

$(M'^{k,t})^2 = M'^{k,t}$) and introduce the “bit-flip” error probability 2^{-p} due to the probabilistic filters; thus its expected value is $E[(M'^{k,t})^2] = (1 - 2^{-p})\theta_k^{k,t} + 2^{-p}\theta_k^{k,t}$. Finally, in (14), given that the variance of a Bernoulli random variable is maximized when the probability of success is 0.5, and that the flipping process does not change this maximum possible variance – as $2^{-p} \ll 1$ given $p \in \{8, 16, 32\}$ – we concluded that the upper bound of the expected squared error is $\frac{d}{4K}$, where d is the number of dimensions and K is the number of clients. It is important to note that our probabilistic filter-based encoding provides the same upper-bound estimation error as [14]; yet, it can achieve significant reductions in terms of required bpp for transmitting masks.

$$\mathbb{E}_{M^{k,t} \sim \text{Bern}(\theta^{k,t}) \forall k \in K} \left[\left\| \bar{\theta}^{g,t} - \hat{\theta}^{g,t} \right\|_2^2 \right] \quad (7)$$

$$= \sum_{i=1}^d \mathbb{E}_{M_i^{k,t} \sim \text{Bern}(\theta_i^{k,t}) \forall k \in K} \left[\left(\bar{\theta}^{g,t} - \hat{\theta}^{g,t} \right)^2 \right] \quad (8)$$

$$= \sum_{i=1}^d \mathbb{E}_{M_i^{k,t} \sim \text{Bern}(\theta_i^{k,t}) \forall k \in K} \left[\left(\frac{1}{K} \sum_{k \in K} (M_i^{k,t} - \theta_i^{k,t}) \right)^2 \right] \quad (9)$$

$$= \frac{1}{K^2} \sum_{i=1}^d \mathbb{E}_{M_i^{k,t} \sim \text{Bern}(\theta_i^{k,t}) \forall k \in K} \left[\left(\sum_{k \in K} (M_i^{k,t} - \theta_i^{k,t}) \right)^2 \right] \quad (10)$$

$$= \frac{1}{K^2} \sum_{i=1}^d \sum_{k \in K} \mathbb{E}_{M_i^{k,t} \sim \text{Bern}(\theta_i^{k,t})} \left[(M_i^{k,t} - \theta_i^{k,t})^2 \right] \quad (11)$$

$$= \frac{1}{K^2} \sum_{i=1}^d \sum_{k \in K} \left(\mathbb{E}_{M_i^{k,t} \sim \text{Bern}(\theta_i^{k,t})} \left[(M_i^{k,t})^2 \right] - 2\theta_i^{k,t} \mathbb{E}_{M_i^{k,t} \sim \text{Bern}(\theta_i^{k,t})} \left[M_i^{k,t} \right] + (\theta_i^{k,t})^2 \right) \quad (12)$$

$$= \frac{1}{K^2} \sum_{i=1}^d \sum_{k \in K} (\theta_i^{k,t} - (\theta_i^{k,t})^2) - 4 \cdot (2^{-p})(\theta_i^{k,t} - (\theta_i^{k,t})^2) + 2^{-p} \quad (13)$$

$$\leq \frac{d}{4K} \quad (14)$$