ALLOCATE BEFORE YOU PERTURB: ADAPTIVE SHARED RELEASE SPACES FOR CLIENT-LEVEL DIFFERENTIALLY PRIVATE FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated learning enables collaboration without centralizing raw data. Client-level differential privacy (DP) protects each client's entire dataset. However, at a fixed per-coordinate noise scale, expected isotropic Gaussian noise energy grows linearly with perturbation dimension, degrading utility in high dimensions. Independent client supports can expand during aggregation, weakening control over this dimension. We propose FedEASE, an entropy adaptive framework for shared sparse release that allocates before perturbation. From a fixed public partition and historical DP statistics, FedEASE samples a fixed number of blocks shared across clients before current updates form. This controls aggregation dimension. Entropy regularization balances historical scores and continued block selection. Inverse marginal correction adjusts sparse updates before clipping and compensates for unequal observation frequencies in the history. Noise aware feedback updates history using only released DP aggregates, without querying unprivatized client statistics. We establish client-level privacy, support coverage, and nonconvex convergence guarantees. Extensive experiments on four image classification datasets and IMDB demonstrate FedEASE's accuracy advantages over the compared DP baselines. These advantages are more pronounced under stronger noise.

1 INTRODUCTION

Federated learning (FL) trains models across devices by aggregating client updates without centralizing raw data (McMahan et al., 2017). Client-level differential privacy (DP) treats each client's entire dataset as one privacy unit (McMahan et al., 2018; Zhang et al., 2022). Client updates are clipped before aggregation and randomized with Gaussian noise. At a fixed per-coordinate noise scale, expected Gaussian noise energy grows with perturbation dimension. High dimensional updates therefore amplify noise in private optimization. Utility thus depends on both client sensitivity and the perturbation space used at aggregation.

To improve utility under client-level privacy, differentially private FL (DPFL) methods follow three directions. **Client update constraints** use clipping analysis or local regularization to reduce distortion from sensitivity control (Zhang et al., 2022; Cheng et al., 2022). **Sparse and low dimensional release** reduces the randomized representation through sparsification, projection, or low dimensional parameterization (Cheng et al., 2022; Hu et al., 2024; Li et al., 2025a). **Server side signal recovery** uses postprocessing or reconstruction to recover informative structure from noisy aggregates (Li et al., 2025b; Malekmohammadi et al., 2024).

Despite these advances, adaptive sparse DPFL faces three challenges in controlling the perturbation space and using private history, as illustrated in Figure 1. First, aggregation must preserve the dimension reduction achieved by sparse release. Independent client supports expand in union as participation grows. Local sparsity then fails to ensure a smaller aggregate perturbation space. Second, fixed supports cannot track changing optimization needs. History driven selection can repeatedly favor high scoring directions, persistently underselecting other blocks. Third, nonuniform selection reweights update contributions and skews historical accumulation toward more frequently observed directions. Adaptive selection must therefore use previously released DP information.

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

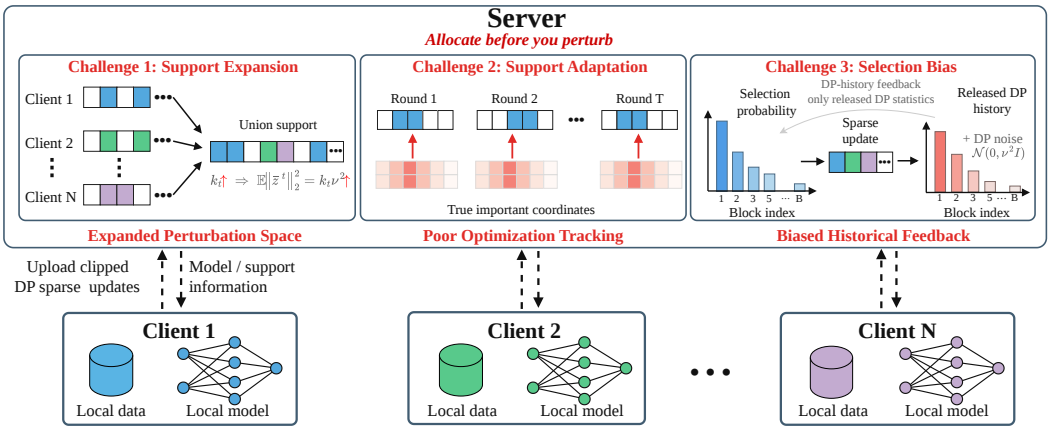

Figure 1: Challenges in adaptive sparse DPFL. Independent client supports expand the aggregate perturbation space. Fixed or overly concentrated supports fail to track changing optimization needs. Unequal selection probabilities distort update contributions and historical feedback. FedEASE addresses these challenges through shared perturbation spaces, entropy adaptive allocation, marginal correction, and noise aware feedback.

To address these challenges, we propose **FedEASE** (**Fed**erated **E**ntropy **A**daptive **S**parse R**e**lease). Its core principle is to *allocate before you perturb*. FedEASE organizes the release space into fixed public blocks and adapts selection at the block level. Historical DP aggregates guide an entropy-regularized allocation of exactly L blocks shared across clients before current updates form. Aggregation remains within this shared space. Entropy regularization favors high scoring blocks while preserving a positive inclusion probability for every block. Inverse marginal correction removes update reweighting in conditional expectation before clipping. It also corrects historical accumulation for unequal observation frequencies. Noise aware feedback subtracts expected Gaussian energy and weights released statistics to guide subsequent allocation.

The main contributions of this paper are summarized as follows:

• We propose FedEASE, an adaptive block-level release framework for client-level DPFL. Entropy adaptive allocation selects exactly L shared blocks from public history before current updates form. This controls aggregation dimension while preserving opportunities to revisit every block.

• We design marginally corrected sparse updates and noise aware feedback. Inverse marginal correction removes selection reweighting before clipping and adjusts historical accumulation for unequal observation frequencies. Noise aware feedback subtracts expected Gaussian energy and weights released statistics to guide subsequent allocation without new private queries.

• Client-level (ε, δ)-DP is established for FedEASE's released noisy aggregate transcript, alongside support coverage and nonconvex convergence guarantees. The analysis quantifies how active dimension, support probability, clipping distortion, and Gaussian noise jointly affect private optimization.

• Extensive experiments on four image classification datasets and IMDB evaluate FedEASE across noise levels, data partitions, and model architectures. FedEASE demonstrates accuracy advantages over the compared DP baselines, with larger gains under stronger noise.

2 Related Work

2.1 Client-Level Differential Privacy

Client-level DPFL balances protection of each client's contribution against model utility (McMahan et al., 2018; Zhang et al., 2022). Clipping analysis quantifies the effects of sensitivity control on private optimization (Zhang et al., 2022), while Bounded Local Update Regularization (BLUR) constrains local trajectories to reduce the resulting utility loss (Cheng et al., 2022). At aggregation, Noise-Aware DPFL handles heterogeneous client noise (Malekmohammadi et al., 2024), and

FedCEO recovers signals through collaborative low rank structure across clients (Li et al., 2025b). These methods improve contributions or noisy signals, but their perturbation spaces typically follow the original update space or a predefined structure.

2.2 Sparse and Low-Dimensional Private Release

Sparse and low dimensional mechanisms reduce the randomized representation to mitigate high dimensional Gaussian noise. Local Update Sparsification (LUS) selects coordinates before privatization (Cheng et al., 2022). FedSMP preserves sparse aggregation with shared random or public-data-derived coordinate masks (Hu et al., 2024). In nonprivate FL, SparsyFed adapts sparse training masks to heterogeneous data (Guastella et al., 2025). PREAMBLE provides efficient private aggregation for block sparse vectors (Asi et al., 2025). The Sketched Gaussian Mechanism compresses the randomized space and analyzes the privacy utility tradeoff with sketch dimension (Li et al., 2025a). Adapting a fixed budget across public blocks from historical DP signals further requires coverage control and correction for nonuniform inclusion.

2.3 Adaptive Differentially Private Optimization

Adaptive DP optimization uses training history to match subsequent mechanisms to evolving optimization states. Adaptive Clipping adjusts thresholds using historical update norms (Andrew et al., 2021). Automatic Clipping reduces manual threshold tuning (Bu et al., 2023). SlaClip adapts clipping using gradient norm slack (Zou et al., 2026). AdaSig uses adaptive sigmoid clipping to balance distortion in update direction and magnitude (Moradi Kalarde et al., 2025). Beyond clipping, DP^2 uses adaptive preconditioning (Li et al., 2023), and DOPPLER filters privacy noise (Zhang et al., 2024). These mechanisms adapt clipping, preconditioning, or noise processing. How historical DP statistics can guide the perturbation space across rounds remains insufficiently explored.

3 Preliminaries

3.1 Federated Learning

Consider N federated clients (McMahan et al., 2017). Client i holds a local dataset $\mathcal{D}_i$. The objective weights clients uniformly:

$$\min_{w\in\mathbb{R}^d} F(w), \qquad F(w)=\frac{1}{N}\sum_{i=1}^{N} F_i(w), \qquad F_i(w)=\frac{1}{|\mathcal{D}_i|}\sum_{\xi\in\mathcal{D}_i}\ell(w;\xi), \tag{1}$$

$w\in\mathbb{R}^d$ is the global model with d trainable parameters, F_i is client i's empirical risk, and $\ell(w;\xi)$ is the loss on sample ξ. At round $t=0,\dots,T-1$, where T is the training horizon, the server uniformly samples $\mathcal{S}_t$ without replacement, with $|\mathcal{S}_t|=K$ and participation rate $q_c=K/N$. Each selected client starts at $w_i^{t,0}=w^t$ and takes Q local stochastic gradient descent (SGD) steps:

$$w_i^{t,q+1}=w_i^{t,q}-\eta_l g_i(w_i^{t,q};\xi_i^{t,q}), \qquad q=0,\dots,Q-1, \tag{2}$$

where $w_i^{t,q}$ is client i's model at local step q, and $\eta_l>0$ is the local learning rate (lr). The stochastic gradient $g_i(\cdot;\xi_i^{t,q})$ uses mini-batch $\xi_i^{t,q}$. The server averages the participating clients' model updates to obtain the next global model.

3.2 Client-Level Differential Privacy

We adopt client replacement adjacency. Federated datasets $\mathcal{D}=(\mathcal{D}_i)_{i=1}^N\simeq\mathcal{D}'$ have identical client identities and differ only in one client's entire local dataset (Mironov, 2017; Birrell et al., 2024).

Definition 3.1 (Client-Level Differential Privacy). A randomized mechanism $\mathcal{M}$ satisfies client-level (ε,δ)-DP if, for all $\mathcal{D}\simeq\mathcal{D}'$ and measurable output sets $\mathcal{O}$,

$$\Pr[\mathcal{M}(\mathcal{D})\in\mathcal{O}]\le e^{\varepsilon}\Pr[\mathcal{M}(\mathcal{D}')\in\mathcal{O}]+\delta, \tag{3}$$

where $\varepsilon > 0$ controls output distinguishability and $\delta \in (0, 1)$ is the additive relaxation parameter (Dwork & Roth, 2014).

Definition 3.2 (Rényi Differential Privacy). For order $\alpha > 1$ and $\mathcal{D} \simeq \mathcal{D}'$, the Rényi divergence of $\mathcal{M}$ is

$$D_\alpha\big(\mathcal{M}(\mathcal{D})\|\mathcal{M}(\mathcal{D}')\big) = \frac{1}{\alpha - 1} \log \int \left(\frac{\mathrm{d}\mathcal{M}(\mathcal{D})}{\mathrm{d}\mathcal{M}(\mathcal{D}')}\right)^{\alpha} \mathrm{d}\mathcal{M}(\mathcal{D}'), \tag{4}$$

where $\mathcal{M}(\mathcal{D})$ denotes the output distribution on $\mathcal{D}$, and the density ratio is its Radon–Nikodym derivative with respect to $\mathcal{M}(\mathcal{D}')$. Mechanism $\mathcal{M}$ satisfies Rényi differential privacy (RDP) with parameters $(\alpha, \varepsilon_\alpha)$ if $D_\alpha(\mathcal{M}(\mathcal{D})\|\mathcal{M}(\mathcal{D}')) \le \varepsilon_\alpha$ for all adjacent datasets (Mironov, 2017). We use Gaussian RDP, adaptive composition, RDP to DP conversion, and postprocessing invariance; see Appendix C.4.

3.3 Threat Model

We study client replacement privacy for an observer of the released shared supports, noisy aggregates, and postprocessed models. The server samples K clients uniformly without replacement and combines their updates through secure aggregation (Bonawitz et al., 2017). Participant identities and sampling randomness remain internal to the coordinating server. All K clients follow the protocol, complete each round, and contribute independent Gaussian noise shares that remain private. Supports and feedback depend on previous DP releases and public randomness. This release protocol admits fixed-size subsampling amplification and adaptive composition (Wang et al., 2019).

4 Method

4.1 Shared Perturbation Space

Independent client supports can expand in union as participation grows. Selecting supports from current private updates also makes them data dependent. FedEASE therefore partitions the flattened parameter vector into $B \le d$ fixed public contiguous blocks, independently of private data. Before current updates form, FedEASE selects exactly L shared blocks to define the active perturbation space:

$$\Omega_t = \bigcup_{b \in \mathcal{B}_t} \mathcal{G}_b, \qquad |\mathcal{B}_t| = L, \qquad k_t = |\Omega_t| = \sum_{b \in \mathcal{B}_t} m_b, \tag{5}$$

where $\mathcal{G}_b = \{s_{b-1}+1, \dots, s_b\}$ with boundaries $s_b = \lfloor bd/B \rfloor$ for $b = 0, \dots, B$. These blocks partition $[d] = \{1, \dots, d\}$, with sizes $m_b \in \{\lfloor d/B \rfloor, \lceil d/B \rceil\}$. The set $\mathcal{B}_t$ indexes the selected blocks, Ω_t contains their coordinates, and k_t is the perturbation dimension. Exact-L selection fixes the block count, while $L\lfloor d/B \rfloor \le k_t \le L\lceil d/B \rceil$. Let $\mathcal{H}_t$ contain prior DP aggregates, released supports, their postprocessed statistics, and current public protocol parameters, before current updates form. Scores, inclusion probabilities, and supports use only $\mathcal{H}_t$ and data independent randomness. Appendix C.4.1 gives the conditional privacy analysis. To quantify the growth in the union of client supports, suppose each of K clients independently selects L blocks uniformly, with $p_{\mathrm{uni}} = L/B$. Then

$$\mathbb{E}\left|\bigcup_{i \in \mathcal{S}_t} \mathcal{B}_i^t\right| = B\left[1 - (1 - p_{\mathrm{uni}})^K\right] \ge B\left(1 - e^{-Kp_{\mathrm{uni}}}\right),$$

where $\mathcal{B}_i^t$ is client i's independently sampled support. The union grows with K, whereas a shared $\mathcal{B}_t$ keeps aggregation within L public blocks. Appendices B.1 and C.1 give the partition implementation and union bound proof, respectively.

4.2 Entropy Adaptive Allocation

Deterministic selection favors high scoring blocks. Rarely selected blocks accumulate fewer statistics and can remain persistently underselected. For $1 \le L < B$, FedEASE allocates marginal probabilities with entropy regularization under a fixed block budget:

$$\max_{\pi^t \in (0,1)^B} \sum_{b=1}^{B} H(\pi_b^t) + \beta_t \sum_{b=1}^{B} h_b^t \pi_b^t, \qquad \text{s.t.} \quad \sum_{b=1}^{B} \pi_b^t = L, \tag{6}$$

where $\pi^t = (\pi_1^t, \dots, \pi_B^t)$ collects block inclusion probabilities and $H(p) = -p \log p - (1-p)\log(1-p)$ is binary entropy. The historical score $h_b^t \in [0,1]$ guides block b's inclusion probability π_b^t, with concentration controlled by β_t. Equation (6) has the unique solution $\pi_b^t = \mathrm{sigmoid}(\beta_t h_b^t - \tau_t)$, where $\mathrm{sigmoid}(x) = (1+e^{-x})^{-1}$ and the scalar τ_t enforces $\sum_b \pi_b^t = L$. Entropy prevents premature concentration; the score term favors high scoring blocks.

Accumulated history supports stronger exploitation, while greater privacy noise calls for broader selection. FedEASE balances these effects through the concentration parameter:

$$\beta_t = \sqrt{\frac{\rho_{t-1}}{1+\rho_{t-1}}} \frac{1}{\sqrt{1+\sigma_t^2}}, \qquad t \ge 1, \tag{7}$$

where ρ_{t-1} is the cumulative history weight from Sec. 4.4, and σ_t is the current Gaussian noise multiplier. The history factor saturates as evidence accumulates. The noise factor limits concentration, giving $0 \le \beta_t \le 1$. Dependent rounding preserves $|\mathcal{B}_t| = L \quad \text{and} \quad \Pr(b \in \mathcal{B}_t \mid \mathcal{H}_t) = \pi_b^t$. With no initial history, $h_b^0 = 1/2$, $\rho_{-1} = 0$, $\beta_0 = 0$, and $\pi_b^0 = L/B$. For $L = B$, set $\pi_b^t = 1$ and select all blocks.

4.3 Marginally Corrected Sparse Updates

Unequal inclusion probabilities give frequently selected directions greater weight in the expected aggregate. This distorts the contribution of local updates before clipping or noise. Therefore, FedEASE applies inverse marginal correction on the shared support. After full dimensional local training (Sec. 3.1), client i forms the displacement $c_i^t = w^t - w_i^{t,Q}$ and corrects it as

$$u_i^t = \mathrm{Diag}(\widetilde{\pi}^t)^{-1} D_t c_i^t, \tag{8}$$

Here $D_t : \mathbb{R}^d \to \mathbb{R}^{k_t}$ selects active coordinates; $D_t^\top$ embeds them by zero filling. The vector $\widetilde{\pi}^t \in \mathbb{R}^{k_t}$ assigns each active coordinate its block's inclusion probability. The operator $\mathrm{Diag}(\cdot)$ forms a diagonal matrix. Block b enters jointly with probability π_b^t; selected coordinates are scaled by $1/\pi_b^t$. With $\mathbb{I}_b^t = \mathbf{1}\{b \in \mathcal{B}_t\}$, $\mathbb{E}[\mathbb{I}_b^t/\pi_b^t \mid \mathcal{H}_t] = 1$. The support is sampled before current updates form. Given the public history, local training is independent of the sampled support. Thus, conditional expectation over support randomness gives

$$\mathbb{E}\left[D_t^\top u_i^t \mid \mathcal{H}_t, c_i^t\right] = c_i^t. \tag{9}$$

Equation (9) removes selection reweighting before clipping. To bound client sensitivity, FedEASE clips the entire corrected update: $x_i^t = u_i^t / \max\{1, \|u_i^t\|_2 / C_t\}$, where C_t is the round-t ℓ_2 threshold. The uniformly weighted mean has sensitivity at most $\Delta_t = \frac{2C_t}{K}$. Set $\nu_t = \sigma_t \Delta_t = \frac{2\sigma_t C_t}{K}$. Secure aggregation implements the Gaussian mechanism with per-coordinate standard deviation ν_t:

$$\widetilde{y}^t = \frac{1}{K} \sum_{i \in \mathcal{S}_t} x_i^t + \bar{z}^t, \qquad \mathbb{E}\left[\|\bar{z}^t\|_2^2 \mid \mathcal{H}_t, \mathcal{B}_t\right] = k_t \nu_t^2, \tag{10}$$

where $\widetilde{y}^t \in \mathbb{R}^{k_t}$ is the server-observed private aggregate. The aggregate noise satisfies $\bar{z}^t \mid (\mathcal{H}_t, \mathcal{B}_t) \sim \mathcal{N}(0, \nu_t^2 I_{k_t})$, where I_{k_t} is the identity matrix. The server updates the full model by $w^{t+1} = w^t - \eta_g D_t^\top \widetilde{y}^t$, where $\eta_g > 0$ is the server lr.

4.4 Noise Aware Feedback

Released DP aggregates provide historical block information, but Gaussian noise distorts their energy statistics. FedEASE constructs noise aware feedback by correcting expected noise energy and reusing the released aggregates through postprocessing. For each active block $b \in \mathcal{B}_t$, the noise corrected statistic is

$$S_b^t = \|\widetilde{y}_b^t\|_2^2 - m_b \nu_t^2, \tag{11}$$

where $\widetilde{y}_b^t$ is the block-b subvector of $\widetilde{y}^t$. Let $x_{i,b}^t$ denote block b of x_i^t and $\mu_b^t = \frac{1}{K}\sum_{i\in\mathcal{S}_t} x_{i,b}^t$ its clipped aggregate. Conditional on the pre-noise state, Eq. (10) gives $\mathbb{E}_z[S_b^t] = \|\mu_b^t\|_2^2$, where $\mathbb{E}_z$ averages over aggregate Gaussian noise. Equation (11) removes the expected Gaussian energy $m_b\nu_t^2$; it retains total block signal energy. The full conditional moments are given in Appendix C.2. However, nonuniform selection still gives frequently observed blocks greater weight in the accumulated history. FedEASE further corrects the projected statistics for unequal inclusion probabilities:

$$\widehat{E}_b^t = \begin{cases} \dfrac{1}{\pi_b^t}\Pi_{[0,1]}\left(\dfrac{S_b^t}{C_t^2}\right), & b \in \mathcal{B}_t, \\ 0, & b \notin \mathcal{B}_t \end{cases}, \tag{12}$$

where $\Pi_{[0,1]}(\cdot)$ projects onto $[0,1]$. The factor $1/\pi_b^t$ compensates for unequal observation frequencies, with zero contribution from inactive blocks. Projection bounds each observed score before marginal weighting. The resulting $\widehat{E}_b^t$ tracks historical block importance under support-dependent clipping. Across rounds, Gaussian noise energy can also vary, even after the block statistics are corrected. Therefore, FedEASE weights each round using average normalized Gaussian noise energy:

$$V_t = \frac{1}{L}\sum_{b\in\mathcal{B}_t}\frac{m_b\nu_t^2}{C_t^2} = \frac{k_t\nu_t^2}{LC_t^2}, \qquad r_t = \frac{1}{1+V_t}, \tag{13}$$

where V_t is normalized Gaussian noise energy per block, and $r_t \in (0,1]$ is the noise aware weight. The weighted history is

$$\rho_t = \sum_{s=0}^{t} r_s, \qquad \widehat{E}_b^{\le t} = \frac{\sum_{s=0}^{t} r_s \widehat{E}_b^s}{\rho_t}, \tag{14}$$

where ρ_t is the cumulative weight, and $\widehat{E}_b^{\le t}$ is block b's weighted historical statistic. Finally, the weighted history is normalized across blocks to form the scores for the next allocation:

$$h_b^{t+1} = \frac{1}{2}\left[1 + \tanh\left(\frac{\widehat{E}_b^{\le t} - \bar{E}^{\le t}}{\sqrt{s_{E,t}^2 + \epsilon_{\text{num}}}}\right)\right], \tag{15}$$

where $\bar{E}^{\le t} = \frac{1}{B}\sum_b \widehat{E}_b^{\le t}$ is the cross-block mean. The variance is $s_{E,t}^2 = \frac{1}{B}\sum_b(\widehat{E}_b^{\le t} - \bar{E}^{\le t})^2$. The constant $\epsilon_{\text{num}} > 0$ ensures numerical stability. Section 4.2 uses this history weight and the next round's noise level to set its allocation concentration. The complete procedure is provided in Algorithm 1 in Appendix A.

5 Theoretical Analysis

5.1 Client-Level Privacy Guarantee

Clipping bounds replacement sensitivity by $\Delta_t = 2C_t/K$, with aggregate Gaussian scale $\nu_t = \sigma_t\Delta_t$. Fixed-size subsampling and adaptive RDP composition protect the aggregate transcript specified in Section 3.3 (Mironov, 2017; Wang et al., 2019).

Theorem 5.1 (Client-Level Privacy Guarantee). *Under the release protocol in Section 3.3 and the amplification conditions in Appendix C.4.2, FedEASE's released supports, aggregates, and postprocessed models satisfy client-level (ε, δ)-DP after T rounds if $\sigma_t \geq \sigma$ and*

$$\sigma \geq \frac{c_2 q_c \sqrt{T \log(1/\delta)}}{\varepsilon}. \tag{16}$$

Here $q_c = K/N$, and $c_2 = \sqrt{8c_0}$ for the fixed amplification constant c_0 defined in the appendix. Equations (27)–(29) give finite-order privacy bounds for arbitrary positive noise multipliers. Reducing k_t lowers Gaussian energy $k_t \nu_t^2$ at the same privacy calibration. The detailed proof is provided in Appendix C.4.

5.2 Support Coverage Guarantee

Coverage requires a persistent chance to reselect every block. Bounded scores and concentration give entropy allocation a uniform positive probability lower bound.

Theorem 5.2 (Support Coverage Guarantee). *Suppose $1 \leq L < B$, $h_b^t \in [0,1]$, and $0 \leq \beta_t \leq \beta_{\max} < \infty$. Let $u = L/B$ and $p_{\min} = u/[e^{\beta_{\max}} - (e^{\beta_{\max}} - 1)u]$. Then $\pi_b^t \geq p_{\min} > 0$ for every block and round. Let $\mathcal{U}_{b,t_0}^{(H)}$ be the event that block b is unselected for H consecutive rounds from t_0. Then*

$$\Pr\left(\mathcal{U}_{b,t_0}^{(H)} \mid \mathcal{H}_{t_0}\right) \leq (1 - p_{\min})^H, \qquad H \geq 1. \tag{17}$$

Equation (17) bounds each block's nonselection probability geometrically in H, including under history-dependent selection across rounds. The detailed proof is provided in Appendix C.5.

5.3 Convergence Analysis

Shared sparse release reduces Gaussian energy. Support randomization, inverse marginal correction, and clipping introduce additional optimization errors. The following bound quantifies these effects for nonconvex objectives.

Theorem 5.3 (Convergence of FedEASE). *Under Assumptions C.1–C.4 in Appendix C.6.1, if $0 < \eta_l \leq 1/(QL_F)$ and $0 < \eta_g \leq 1/(48 L_F \eta_l Q H_K)$, then*

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(w^t)\|_2^2 \leq & \underbrace{\frac{2(F(w^0) - F_*)}{T\eta_g\eta_l Q} + \eta_l^2 Q^2 L_F^2 G^2 + 12 L_F \eta_g \eta_l Q \mathcal{V}_K}_{\text{FedAvg optimization terms}} \\ & + \underbrace{\frac{6\bar{\Gamma}^2}{(\eta_l Q)^2} + \frac{3 L_F \eta_g}{\eta_l Q}\left(\bar{\Phi}^2 + \overline{k\nu^2}\right)}_{\text{FedEASE sparse private release terms}}.\end{aligned} \tag{18}$$

Here L_F is the smoothness constant, G^2 bounds gradient second moments, and $F_* = \inf_w F(w)$. The factors H_K and $\mathcal{V}_K$ quantify sampling heterogeneity and aggregate gradient variance. The terms $\bar{\Gamma}^2$, $\bar{\Phi}^2$, and $\overline{k\nu^2}$ average clipping, support, and noise errors (Appendix C.6.1). The bound $\pi_b^t \geq p_{\min} > 0$ controls inverse-correction variance, while $k_t \nu_t^2$ controls Gaussian energy. The detailed proof is provided in Appendix C.6.3.

6 Experiments

We evaluate FedEASE's utility and efficiency across datasets, data partitions, and noise levels. A framework overview, privacy calibration, experimental details, and additional analyses are provided in Appendices A–B. Accuracy differences are reported in percentage points (pp).

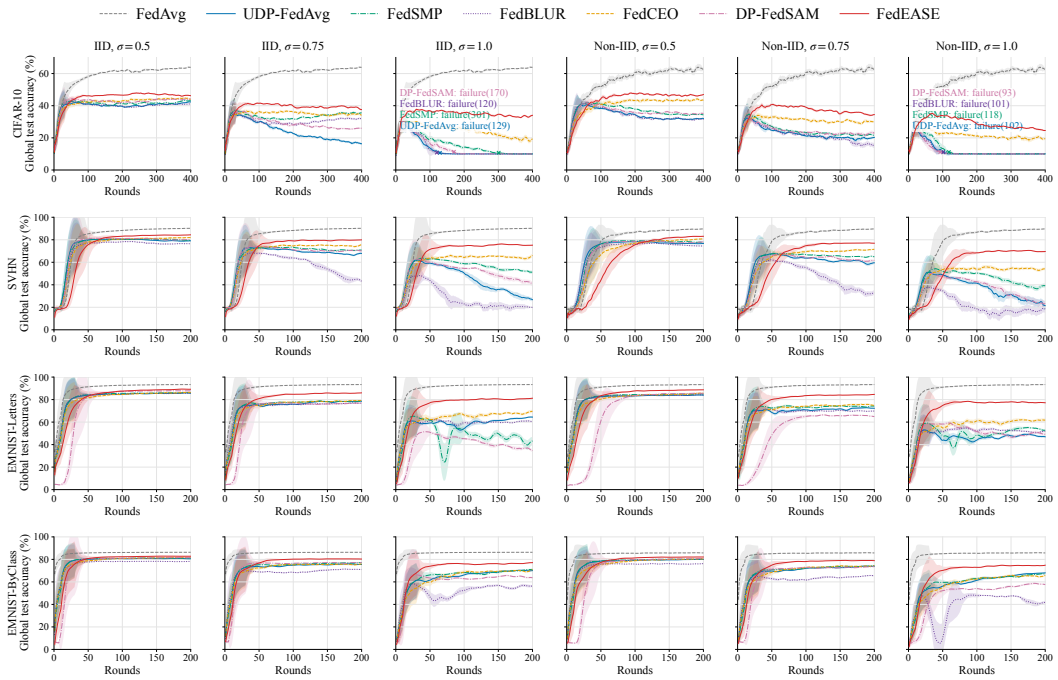

Figure 2: Training dynamics on four image classification datasets. FedEASE retains higher late-stage accuracy under stronger privacy noise. Rows show CIFAR-10, SVHN, EMNIST-Letters, and EMNIST-ByClass. The first and last three columns show IID and Non-IID partitions, respectively, with $\sigma = 0.5, 0.75, 1.0$ in each group. Shading shows the standard deviation across three independent runs; failure (r) marks the onset of training instability at round r.

6.1 Experimental Setup

Datasets and Models. We use LeNet-5 on CIFAR-10 (Krizhevsky, 2009), SVHN (Netzer et al., 2011), EMNIST-Letters, and EMNIST-ByClass (Cohen et al., 2017). Each dataset uses independent and identically distributed (IID) partitions and Dirichlet non-IID partitions, with Dirichlet concentration $\beta = 0.5$ by default. Each round uniformly samples $K = 50$ of $N = 500$ clients without replacement. Local training uses 5 epochs and batch size 64. All DP methods share the task-specific training settings, clipping radius $C = 1$, $q_c = 0.1$, $\delta = 10^{-5}$, and constant nominal noise multiplier $\sigma_t = \sigma \in \{0.5, 0.75, 1.0\}$, with method-specific hyperparameters detailed in Appendix B.1.

Baselines. We compare FedEASE with FedAvg (McMahan et al., 2017), UDP-FedAvg (McMahan et al., 2018; Zhang et al., 2022), FedSMP (Hu et al., 2024), FedBLUR (Cheng et al., 2022), FedCEO (Li et al., 2025b), and DP-FedSAM (Shi et al., 2023). The baselines cover nonprivate FL, full space client-level DP, sparse perturbation, local regularization, server recovery, and optimization for flatter solutions.

6.2 Overall Performance

Table 1 compares accuracy and mean complete wall clock time per round on the four image classification datasets. FedEASE achieves the highest mean final and peak accuracy in all 24 private settings. Gains generally increase with privacy noise, especially under Non-IID data. On Non-IID SVHN, the final-accuracy gain over the strongest DP baseline is 2.26 pp at $\sigma = 0.5$. It rises to 16.71 pp at $\sigma = 1.0$. On Non-IID EMNIST-Letters, the corresponding gain rises from 2.97 to 17.58 pp.

FedEASE also retains the highest mean final accuracy under weaker noise. On IID EMNIST-ByClass at $\sigma = 0.5$, it achieves 82.83%, exceeding DP-FedSAM by 1.32 pp and FedCEO, the strongest DP baseline, by 1.11 pp. This trend is consistent with shared exact-L aggregation: at fixed Gaussian calibration, expected noise energy scales with the active dimension.

Table 1: Mean final test accuracy / peak test accuracy, in %, over three independent runs, and complete wall-clock time per round (s) on four image tasks. Metrics follow Appendix B.1. FedAvg is nonprivate; bold marks the best DP accuracy for each metric and setting.

Dataset	Method	IID			Time (s/round)	Non-IID ($\beta = 0.5$)			Time (s/round)
		Accuracy (%)				Accuracy (%)			
		$\sigma = 0.5$	$\sigma = 0.75$	$\sigma = 1.0$		$\sigma = 0.5$	$\sigma = 0.75$	$\sigma = 1.0$	
CIFAR-10	FedAvg (non-DP)	63.83 / 64.22			7.06	63.10 / 65.68			6.97
	UDP-FedAvg	42.87 / 42.90	17.08 / 35.82	9.97 / 30.16	8.14	31.83 / 42.07	20.15 / 33.82	10.00 / 27.40	8.06
	FedSMP	43.51 / 44.11	35.28 / 38.25	9.99 / 33.15	8.03	34.88 / 43.51	21.67 / 35.73	10.07 / 28.51	7.86
	FedBLUR	40.85 / 43.47	31.55 / 36.97	9.96 / 31.24	9.73	31.19 / 43.13	17.13 / 35.15	9.99 / 27.30	9.57
	FedCEO	43.62 / 45.66	33.71 / 37.85	19.18 / 31.39	8.19	42.61 / 45.26	28.77 / 35.54	18.81 / 28.46	8.17
	DP-FedSAM	44.86 / 44.86	26.25 / 37.26	10.00 / 30.62	8.39	35.74 / 42.98	23.19 / 33.33	9.97 / 27.43	8.15
	FedEASE	**47.40 / 48.11**	**37.72 / 41.95**	**34.32 / 38.55**	7.92	**47.25 / 48.50**	**34.24 / 41.55**	**24.91 / 36.06**	7.88
SVHN	FedAvg (non-DP)	90.14 / 90.46			11.31	89.51 / 90.11			11.62
	UDP-FedAvg	79.27 / 80.83	67.33 / 74.10	24.71 / 63.90	12.48	76.63 / 79.25	58.62 / 68.09	23.00 / 52.35	12.63
	FedSMP	79.51 / 81.43	70.50 / 74.31	49.79 / 64.89	12.30	77.95 / 79.58	64.22 / 68.63	36.99 / 56.24	12.61
	FedBLUR	76.16 / 78.87	42.56 / 69.25	19.79 / 50.61	14.57	74.55 / 77.34	33.92 / 63.24	11.08 / 40.07	15.05
	FedCEO	82.15 / 82.30	75.46 / 76.48	66.18 / 68.30	12.41	80.99 / 81.27	71.10 / 72.25	53.56 / 57.25	12.72
	DP-FedSAM	79.90 / 81.61	70.92 / 75.02	38.54 / 63.56	12.08	78.35 / 79.59	61.31 / 68.93	23.39 / 51.86	12.19
	FedEASE	**84.35 / 84.59**	**80.15 / 80.15**	**75.39 / 76.65**	11.86	**83.25 / 83.42**	**77.25 / 77.68**	**70.27 / 71.16**	11.77
EMNIST-ByClass	FedAvg (non-DP)	86.30 / 86.37			110.64	86.14 / 86.14			110.59
	UDP-FedAvg	80.45 / 81.00	75.68 / 76.22	70.20 / 70.96	112.89	79.99 / 80.15	73.92 / 74.20	67.99 / 68.28	114.93
	FedSMP	81.30 / 81.40	76.77 / 77.29	71.79 / 71.79	116.24	80.70 / 80.70	74.50 / 74.52	66.64 / 67.52	116.90
	FedBLUR	78.35 / 78.48	71.50 / 71.81	56.35 / 59.58	143.02	76.61 / 76.61	66.10 / 66.16	42.08 / 50.00	145.09
	FedCEO	81.72 / 81.72	76.71 / 76.71	70.74 / 71.41	113.32	80.56 / 80.56	74.53 / 74.81	65.32 / 66.53	113.53
	DP-FedSAM	81.51 / 81.71	76.82 / 77.41	62.91 / 66.95	119.36	80.83 / 80.92	73.23 / 73.65	56.30 / 59.17	121.23
	FedEASE	**82.83 / 82.97**	**80.39 / 80.65**	**77.20 / 77.26**	107.77	**82.05 / 82.26**	**78.96 / 79.29**	**75.57 / 75.57**	108.77
EMNIST-Letters	FedAvg (non-DP)	93.36 / 93.51			17.34	93.13 / 93.59			19.42
	UDP-FedAvg	85.65 / 86.22	78.13 / 79.11	64.30 / 65.08	18.73	83.79 / 84.37	72.00 / 73.48	44.63 / 54.68	20.66
	FedSMP	85.84 / 86.53	77.91 / 79.43	45.62 / 67.27	18.51	84.69 / 85.27	74.15 / 75.11	53.43 / 60.34	20.72
	FedBLUR	86.14 / 86.14	77.14 / 77.48	59.52 / 63.12	22.43	84.40 / 84.63	69.21 / 72.07	51.18 / 55.56	25.35
	FedCEO	86.27 / 86.86	78.65 / 80.09	68.22 / 71.22	18.62	85.81 / 86.05	74.58 / 76.74	59.32 / 64.00	20.74
	DP-FedSAM	87.58 / 88.07	76.68 / 77.59	33.79 / 52.38	18.54	85.67 / 85.85	65.10 / 67.13	49.41 / 60.66	20.99
	FedEASE	**89.30 / 89.43**	**86.22 / 86.22**	**81.43 / 81.52**	17.79	**88.78 / 88.80**	**84.85 / 84.90**	**76.90 / 78.86**	19.71

FedEASE remains competitive in runtime. On CIFAR-10, complete rounds take 7.92 s (IID) and 7.88 s (Non-IID), comparable to FedSMP and UDP-FedAvg. It runs faster than every DP baseline on SVHN, EMNIST-Letters, and EMNIST-ByClass. Additional server computation occupies only a small fraction of each round (Appendix B.12).

Training dynamics. Figure 2 shows how privacy noise affects accuracy throughout training. At $\sigma = 1.0$, several CIFAR-10 baselines exhibit a chance-level plateau near 10%, while FedEASE preserves higher late-stage accuracy. On SVHN, FedEASE improves more slowly early on but reaches higher late-stage accuracy, especially under stronger noise.

7 Conclusion

We present FedEASE, an entropy adaptive framework for shared sparse release in client-level DPFL. FedEASE selects a shared support from historical DP statistics before current updates form. This controls the perturbation space at aggregation. Entropy balances history exploitation and continued selection. Marginal correction adjusts updates and historical accumulation for nonuniform selection. Noise aware feedback updates history from released DP aggregates. The analysis establishes client-level privacy for the released aggregate transcript, support coverage, and nonconvex convergence under the stated conditions. Experiments on four image classification tasks and IMDB demonstrate FedEASE's accuracy advantages and competitive training cost.

AI use statement

Generative artificial intelligence (AI) tools were used to assist with translation and LaTeX formatting and editing. The authors reviewed all AI-assisted material and take responsibility for the final content of this work.

REFERENCES

Galen Andrew, Om Thakkar, H. Brendan McMahan, and Swaroop Ramaswamy. Differentially private learning with adaptive clipping. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17455–17466, 2021.

Hilal Asi, Vitaly Feldman, Hannah Keller, Guy N. Rothblum, and Kunal Talwar. PREAMBLE: Private and efficient aggregation via block sparse vectors. In *Advances in Neural Information Processing Systems*, volume 38, pp. 20093–20137, 2025.

Jeremiah Birrell, Reza Ebrahimi, Rouzbeh Behnia, and Jason Pacheco. Differentially private stochastic gradient descent with fixed-size minibatches: Tighter RDP guarantees with or without replacement. In *Advances in Neural Information Processing Systems*, volume 37, pp. 11087–11131, 2024.

Kallista A. Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1175–1191, 2017.

Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. Automatic clipping: Differentially private deep learning made easier and stronger. In *Advances in Neural Information Processing Systems*, volume 36, pp. 41727–41764, 2023.

Anda Cheng, Peisong Wang, Xi Sheryl Zhang, and Jian Cheng. Differentially private federated learning with local regularization and sparsification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10122–10131, 2022.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. EMNIST: Extending MNIST to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 2921–2926, 2017. doi: 10.1109/IJCNN.2017.7966217.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Adriano Guastella, Lorenzo Sani, Alex Iacob, Alessio Mora, Paolo Bellavista, and Nic Lane. SparsyFed: Sparse adaptive federated learning. In *International Conference on Learning Representations*, pp. 14781–14813, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Rui Hu, Yuanxiong Guo, and Yanmin Gong. Federated learning with sparsified model perturbation: Improving accuracy under client-level differential privacy. *IEEE Transactions on Mobile Computing*, 23(8):8242–8255, 2024. doi: 10.1109/TMC.2023.3343288.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Qiaobo Li, Zhijie Chen, and Arindam Banerjee. Sketched gaussian mechanism for private federated learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 131952–132017, 2025a.

Tian Li, Manzil Zaheer, Ken Ziyu Liu, Sashank J. Reddi, H. Brendan McMahan, and Virginia Smith. Differentially private adaptive optimization with delayed preconditioners. In *International Conference on Learning Representations*, 2023.

Yuecheng Li, Lele Fu, Tong Wang, Jian Lou, Bin Chen, Lei Yang, Jian Shen, Zibin Zheng, and Chuan Chen. Clients collaborate: Flexible differentially private federated learning with guaranteed improvement of utility-privacy trade-off. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 34333–34354. PMLR, 2025b.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150. Association for Computational Linguistics, 2011.

Saber Malekmohammadi, Yaoliang Yu, and Yang Cao. Noise-aware algorithm for heterogeneous differentially private federated learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34461–34498. PMLR, 2024.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017.

H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *International Conference on Learning Representations*, 2018.

Ilya Mironov. Rényi differential privacy. In *Proceedings of the 30th IEEE Computer Security Foundations Symposium*, pp. 263–275, 2017.

Faeze Moradi Kalarde, Ali Bereyhi, Ben Liang, and Min Dong. Adaptive sigmoid clipping for balancing the direction–magnitude mismatch trade-off in differentially private learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 56116–56152, 2025.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Yifan Shi, Yingqi Liu, Kang Wei, Li Shen, Xueqian Wang, and Dacheng Tao. Make landscape flatter in differentially private federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24552–24562, 2023.

Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled Rényi differential privacy and analytical moments accountant. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pp. 1226–1235. PMLR, 2019. URL `https://proceedings.mlr.press/v89/wang19b.html`.

Xinwei Zhang, Xiangyi Chen, Mingyi Hong, Steven Wu, and Jinfeng Yi. Understanding clipping for federated learning: Convergence and client-level differential privacy. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 26048–26067. PMLR, 2022.

Xinwei Zhang, Zhiqi Bu, Mingyi Hong, and Meisam Razaviyayn. DOPPLER: Differentially private optimizers with low-pass filter for privacy noise reduction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 41826–41851, 2024.

Shuyan Zou, Shaowei Wang, Zhanxing Zhu, Jin Li, Changyu Dong, Vladimiro Sassone, and Han Wu. SlaClip: Gradient norm slacks can be indicator for adaptive clipping in DP-SGD. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

A Overview of the FedEASE Framework

FedEASE follows **allocate before you perturb**, as summarized in Algorithm 1. The server initializes the global model and public block histories. At each round, entropy adaptive allocation uses previously released statistics to assign block inclusion probabilities. Dependent rounding selects exactly L shared blocks before current local updates form. Each selected client trains the full model independently of the realized support, applies inverse marginal correction on the selected coordinates, and clips the corrected update. Each of the K participants submits $y_i^t = x_i^t/K + z_i^t$ through secure aggregation, where z_i^t is its independent Gaussian noise share. The operator $\mathrm{SecureSum}$ reveals their sum. Conditional on the public history and support, the noise shares sum to $\bar{z}^t \sim \mathcal{N}(0, \nu_t^2 I_{k_t})$, yielding Eq. (10). The server embeds the noisy aggregate into the full parameter space and updates the global model. Noise aware feedback then removes expected noise energy, corrects unequal observation frequencies, and downweights rounds with greater normalized noise energy. Finally, the weighted history is normalized to guide the next allocation. Theorem 5.1 accounts for the released transcript through fixed-size subsampling and adaptive composition.

Algorithm 1 FedEASE

Input: Public model w^0, clients N, participants K, public partition $\{\mathcal{G}_b\}_{b=1}^B$, active blocks L, rounds T, local steps Q, local lr η_l, server lr η_g, public schedules $\{C_t, \sigma_t\}_{t=0}^{T-1}$, numerical constant $\epsilon_{\mathrm{num}} > 0$, RDP orders $\mathcal{A} \subset \{2, 3, \ldots\}$, target $\delta \in (0, 1)$.
Output: Global model w^T and privacy bound ε for the released transcript.
Server
Set $q_c = K/N$, $h_b^0 = 1/2$ for every block, and $\mathcal{R}_\alpha = 0$ for $\alpha \in \mathcal{A}$.
Initialize the history numerator A_b^{-1} and weight ρ_{-1} to zero.
for $t = 0, \ldots, T-1$ **do**
  Set $\beta_t = 0$ if $t = 0$; otherwise compute β_t by Eq. (7).
  Set $\pi^t = \mathbf{1}$ if $L = B$; otherwise solve Eq. (6).
  Select all blocks if $L = B$; otherwise draw $\mathcal{B}_t$ by dependent rounding of π^t.
  Construct Ω_t, D_t, $\widetilde{\pi}^t$, and k_t from the shared support.
  Sample K clients uniformly without replacement as $\mathcal{S}_t$, independently of private data.
  $\nu_t \leftarrow 2\sigma_t C_t/K$.
  Broadcast w^t, Ω_t, $\widetilde{\pi}^t$, C_t, and ν_t.
  for all $i \in \mathcal{S}_t$ **in parallel do**
    Client i executes ClientUpdate$(w^t, \Omega_t, \widetilde{\pi}^t, C_t, \nu_t, i)$.
  end for
  Receive only $\widetilde{y}^t = \mathrm{SecureSum}_{i \in \mathcal{S}_t}(y_i^t)$ after all K clients complete.
  $w^{t+1} \leftarrow w^t - \eta_g D_t^\top \widetilde{y}^t$.
  Compute $\widehat{E}_b^t$ for every block by Eqs. (11)–(12), with zeros on inactive blocks.
  Compute the noise aware weight r_t by Eq. (13).
  $A_b^t \leftarrow A_b^{t-1} + r_t \widehat{E}_b^t$ for every block.
  $\rho_t \leftarrow \rho_{t-1} + r_t$; $\widehat{E}_b^{\le t} \leftarrow A_b^t/\rho_t$, implementing Eq. (14).
  Update each block score h_b^{t+1} by Eq. (15).
  $\mathcal{R}_\alpha \leftarrow \mathcal{R}_\alpha + a_{\mathrm{WOR}}(\alpha, q_c, \sigma_t)$ for $\alpha \in \mathcal{A}$, using Eq. (27).
end for
Compute ε from $\{\mathcal{R}_\alpha\}_{\alpha \in \mathcal{A}}$ and δ by Eq. (29).
return (w^T, ε).
ClientUpdate$(w^t, \Omega_t, \widetilde{\pi}^t, C_t, \nu_t, i)$
Set $k_t = |\Omega_t|$ and construct D_t in the public coordinate order of Ω_t.
Set $w_i^{t,0} = w^t$.
for $q = 0, \ldots, Q-1$ **do**
  Draw a local mini-batch $\xi_i^{t,q}$ independently of the support realization.
  $w_i^{t,q+1} \leftarrow w_i^{t,q} - \eta_l g_i(w_i^{t,q}; \xi_i^{t,q})$ in the full parameter space.
end for
$c_i^t \leftarrow w^t - w_i^{t,Q}$.
$u_i^t \leftarrow \mathrm{Diag}(\widetilde{\pi}^t)^{-1} D_t c_i^t$.
$x_i^t \leftarrow u_i^t / \max\{1, \|u_i^t\|_2 / C_t\}$.
Independently sample $z_i^t \sim \mathcal{N}(0, \nu_t^2 I_{k_t}/K)$.
Submit $y_i^t = x_i^t/K + z_i^t$ through secure aggregation.

B Additional Experiments

Implementation Environment. Experiments run on Ubuntu 24.04 LTS with an AMD Ryzen Threadripper 3960X CPU (24 cores), 64 GB of system memory, and two NVIDIA GeForce RTX 3090 GPUs (24 GB each). The implementation uses Python 3.10.18, PyTorch 2.1.2, and CUDA 12.1. Each training trial uses both GPUs, with at most one selected client trained on each device at a time; remaining clients wait for an available device. Experimental configurations run sequentially, with one configuration completed before the next begins.

Privacy Calibration. We compare methods at a common nominal noise multiplier σ, with fixed-size participation $q_c = 0.1$ and target $\delta = 10^{-5}$. Privacy budgets depend on each method's release mechanism and accountant. For FedEASE, Eqs. (27)–(29) bound the released transcript using $\mathcal{A} = \{2, 3, \ldots, 128\} \cup \{160, 192, 256, 320, 384, 448, 512\}$. Table A8 reports these bounds for one run at each default training horizon. Privacy costs compose across models released from the same private clients.

B.1 Experimental Protocol and Model Architecture

We follow the common training settings in Section 6.1, with a constant task-specific lr and $C_t = C$. CIFAR-10 uses 400 rounds with local lr $\eta_l = 0.03$; SVHN uses 200 rounds with local lr $\eta_l = 0.05$. Both EMNIST tasks use 200 rounds with local lr $\eta_l = 0.03$. The server lr is $\eta_g = 1.0$. For each run, Final is the test accuracy at the last training round, and Peak is the maximum observed across all rounds. Table 1 averages each metric over three independent runs, including the recorded last-round and peak accuracies of runs marked failure (r) in Figure 2.

Hyperparameter Settings. We use a common SGD configuration across methods, with momentum 0.9, weight decay 5×10^{-4}, and the task-specific lr given above. FedAvg and UDP-FedAvg use this configuration without additional method-specific hyperparameters. For FedSMP (Hu et al., 2024), the random-k (rand-k) retention ratio is set to 0.20, so each client uploads 20% of the model coordinates. For FedBLUR (Cheng et al., 2022), the BLUR coefficient is $\lambda_{\mathrm{BLUR}} = 0.4$, and tensorwise LUS retains the highest-utility 20% of coordinates, with at least one retained per tensor and threshold ties included. For DP-FedSAM (Shi et al., 2023), the sharpness-aware minimization (SAM) radius is $\rho_{\mathrm{SAM}} = 0.2$ in all image and text experiments. For FedCEO (Li et al., 2025b), the initial smoothing coefficient is $\lambda_{\mathrm{CEO}} = 0.6$, the geometric scaling factor is $r_{\mathrm{CEO}} = 1.04$, and the smoothing interval is 10 rounds. Its tensor nuclear norm (TNN) solver uses a tolerance of 10^{-10} and at most 1000 iterations. For FedEASE, the nominal active-block ratio is $\gamma = 0.20$, and the numerical stability constant is $\epsilon_{\mathrm{num}} = 10^{-12}$. These settings apply throughout unless varied in a sensitivity study.

Baseline Implementation. FedBLUR and DP-FedSAM jointly clip each client displacement and add full-dimensional Gaussian noise before averaging. FedBLUR applies tensorwise LUS before clipping and uploads the resulting full-dimensional noisy vector. At smoothing rounds after initialization, FedCEO starts each client from its cached local weights when available, falling back to the global model; other rounds start from the global model. FedEASE uses secure aggregation on a shared support, with realized block and coordinate ratios L/B and k_t/d.

Public Block Partition. All FedEASE experiments use the same construction from the public model architecture. We flatten trainable tensors in their registered parameter order, using `reshape(-1)` within each tensor, and concatenate them into a vector of dimension d. The block boundaries and active-block budget are

$$B = \lceil \sqrt{d} \rceil, \qquad s_b = \lfloor bd/B \rfloor \quad (b = 0, \ldots, B), \qquad L = \max\{1, \lfloor B\gamma \rfloor_{\mathrm{fp}}\}.$$

Here, the subscript fp denotes flooring the floating-point budget product. The active ratio is evaluated as the complement of the configured inactive fraction before multiplication by B. For ResNet-18, nominal $\gamma = 0.20$ with $B = 3345$ yields $L = 668$ under this rule. Block b uses the zero-based slice $[s_{b-1}, s_b)$, equivalent to the one-based coordinates in Section 4.1; blocks may cross tensor boundaries. The partition and coordinate order remain fixed across rounds and FedEASE ablations for each architecture. For $1 \le L < B$, the implementation solves $\sum_b \mathrm{sigmoid}(\beta_t h_b^t - \tau_t) = L$ by 60 float64 bisection steps on $[\min_b(\beta_t h_b^t) - 60, \max_b(\beta_t h_b^t) + 60]$, using a stable sigmoid; it handles $\beta_t = 0$ as $\pi_b^t = L/B$. It checks the residual against $10^{-10}\max(1, \sqrt{B})$. Dependent rounding

starts from π^t and updates two current fractional marginals (p_{b_1}, p_{b_2}), with $a_+ = \min(1 - p_{b_1}, p_{b_2})$ and $a_- = \min(p_{b_1}, 1 - p_{b_2})$. With probability $a_-/(a_+ + a_-)$, it uses $(p_{b_1} + a_+, p_{b_2} - a_+)$; otherwise, it uses $(p_{b_1} - a_-, p_{b_2} + a_-)$. Each step preserves their sum and, in exact arithmetic, each marginal in expectation, while making at least one entry integral. The float64 implementation uses a boundary tolerance of 10^{-12} and verifies a final support size of exactly L. Block sizes differ by at most one, with the larger blocks distributed by the boundary rule above. Shared selection keeps all client uploads and their aggregate within the same k_t coordinates. Equation (11) subtracts expected Gaussian energy from the total energy of each block. Nearly equal block sizes make these total-energy statistics comparable across blocks. Figure A1 illustrates the construction and the resulting coordinate budget for LeNet-5 on CIFAR-10 and SVHN.

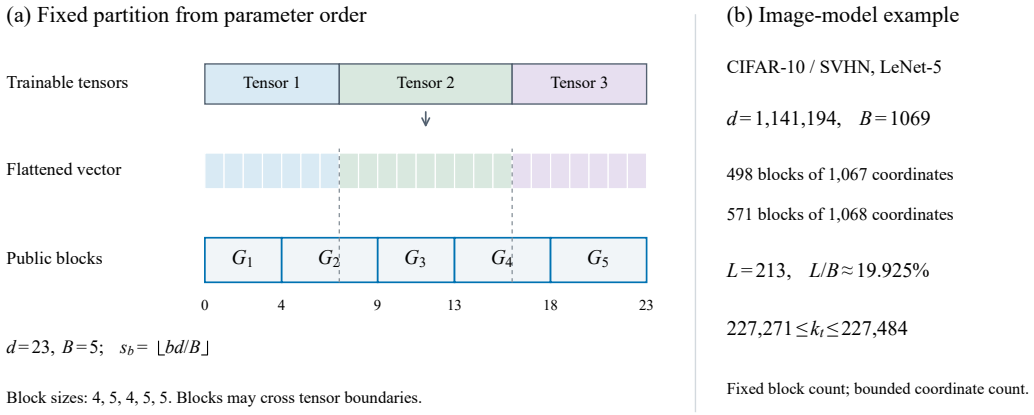

Figure A1: Public block construction. A small example ($d = 23$, $B = 5$) illustrates registered parameter order, fixed boundaries, and blocks that cross tensor boundaries. The LeNet-5 example shows the block sizes and resulting coordinate-count range for $L = 213$.

LeNet-5. All four image tasks use the same LeNet-5 implementation, with the output dimension of the final classification layer determined by the number of classes in each dataset. The network contains two 5×5 convolutional layers, two 2×2 max pooling layers, and three fully connected layers. Rectified linear unit (ReLU) activations are applied after both convolutional layers and the first two fully connected layers. For both EMNIST tasks, each 28×28 grayscale image is resized to 32×32 and replicated across three channels. After scaling pixels to $[0, 1]$, each channel is normalized with mean 0.1307 and standard deviation 0.3081. Training and test data use the same deterministic preprocessing, without data augmentation. The output dimension $\mathrm{num_classes}$ is 10 for CIFAR-10 and SVHN, 26 for EMNIST-Letters, and 62 for EMNIST-ByClass. The architecture is:

1. $\mathrm{Conv2d}(3, 32, \text{kernel size} = 5, \text{stride} = 1, \text{padding} = 0)$
2. ReLU
3. $\mathrm{MaxPool2d}(\text{kernel size} = 2, \text{stride} = 2)$
4. $\mathrm{Conv2d}(32, 64, \text{kernel size} = 5, \text{stride} = 1, \text{padding} = 0)$
5. ReLU
6. $\mathrm{MaxPool2d}(\text{kernel size} = 2, \text{stride} = 2)$
7. $\mathrm{Flatten}$
8. $\mathrm{Linear}(1600, 512)$
9. ReLU
10. $\mathrm{Linear}(512, 512)$
11. ReLU
12. $\mathrm{Linear}(512, \mathrm{num_classes})$

Runtime is reported as the average per communication round. Image-task timings in Tables 1 and A7 use $\sigma = 1.0$; IMDB timings use $\sigma = 3.0$. FedAvg is timed without DP noise. The main paper uses complete per-round wall clock time to measure end-to-end training efficiency. Appendix B.12 further reports client-side and server-side runtime, together with their sum as the total computation time.

B.2 Effect of Data Heterogeneity

Table A1 evaluates FedEASE across label distributions with Dirichlet concentration $\beta \in \{0.1, 1.0\}$. FedEASE achieves the highest final and peak accuracy among the compared DP methods in all eight settings. On SVHN at $\sigma = 1.0$, it attains 59.02% final accuracy under the more heterogeneous partition ($\beta = 0.1$) and 71.51% at $\beta = 1.0$. The method thus maintains its relative advantage across both levels of heterogeneity.

On CIFAR-10 at $\beta = 0.1$, the final-accuracy margin over the strongest DP baseline rises from 3.30 to 11.66 pp as σ increases. On SVHN at the same heterogeneity level, the corresponding margin rises from 3.74 to 17.22 pp. The larger margins under stronger noise are consistent with the reduced perturbation dimension of shared sparse release.

Table A1: Effect of data heterogeneity with LeNet-5 and $C = 1$; FedEASE uses nominal ratio $\gamma = 0.20$. Entries report **Final / Peak**: final test accuracy / peak test accuracy, in %. Smaller Dirichlet β indicates stronger heterogeneity. **Bold** marks the highest DP accuracy for each metric and setting. FedAvg is a nonprivate method.

Dataset	Method	$\beta = 0.1$		$\beta = 1.0$	
		$\sigma = 0.5$	$\sigma = 1.0$	$\sigma = 0.5$	$\sigma = 1.0$
CIFAR-10	UDP-FedAvg	29.34 / 36.78	10.01 / 21.09	34.69 / 42.31	9.97 / 27.91
	FedSMP	30.42 / 38.66	9.95 / 23.25	37.06 / 42.93	10.08 / 30.07
	FedBLUR	24.79 / 37.11	9.94 / 21.72	32.04 / 43.16	10.28 / 29.01
	FedCEO	39.19 / 41.22	10.57 / 21.99	43.33 / 44.94	16.99 / 28.89
	DP-FedSAM	30.20 / 37.69	9.97 / 20.74	37.13 / 44.04	10.00 / 28.02
	FedEASE	**42.49 / 43.84**	**22.23 / 29.67**	**46.70 / 48.48**	**26.30 / 37.40**
	FedAvg (non-DP)	61.74 / 62.33		62.52 / 64.53	
SVHN	UDP-FedAvg	70.07 / 72.49	19.58 / 27.75	77.83 / 80.35	23.80 / 58.28
	FedSMP	71.89 / 73.07	20.74 / 36.82	78.52 / 80.48	42.91 / 60.28
	FedBLUR	52.50 / 66.04	19.22 / 20.59	76.83 / 78.60	19.62 / 48.10
	FedCEO	69.66 / 72.52	41.80 / 43.54	81.85 / 82.37	57.16 / 62.81
	DP-FedSAM	72.10 / 72.17	19.58 / 24.46	78.91 / 80.16	28.16 / 56.98
	FedEASE	**75.84 / 77.06**	**59.02 / 59.78**	**83.53 / 84.03**	**71.51 / 73.85**
	FedAvg (non-DP)	88.15 / 88.15		90.44 / 90.44	

B.3 Effect of Clipping Radius

Table A2 examines how clipping controls the balance between update retention and privacy noise. We fix $\gamma = 0.20$ and vary $C \in \{0.1, 0.2, 0.3, 0.5, 1.0, 1.5\}$ under non-IID data. Since the aggregate noise scale is $\nu = 2\sigma C/K$, decreasing C reduces Gaussian energy quadratically but also clips more of the corrected update. With q_c and σ fixed, the privacy bound in Eq. (27) is constant across this sweep. Clipping radius and release ratio control perturbation magnitude and dimension, respectively.

For CIFAR-10, final accuracy peaks at $C = 0.5$ for $\sigma = 0.5$ and at $C = 0.2$ for $\sigma = 1.0$. At the stronger noise level, decreasing C from 1.0 to 0.2 raises final accuracy from 24.91% to 45.98%; peak accuracy is highest at the nearby radius $C = 0.3$. SVHN favors $C = 1.0$ and $C = 0.5$ at the two noise levels, respectively, with the latter increasing final accuracy from 70.27% to 78.05%. The shift toward smaller radii under stronger noise highlights the value of controlling perturbation magnitude,

with the SVHN results favoring radii that retain sufficient update information. Task-specific clipping thus offers further gains over the common $C = 1$ setting used in the main comparison.

Table A2: Clipping sensitivity of FedEASE under non-IID ($\beta = 0.5$) data with nominal ratio $\gamma = 0.20$. Entries report **Final / Peak**: final test accuracy / peak test accuracy, in %. **Bold** marks each metric's maximum over C for the same dataset and noise level.

Dataset	C	$\sigma = 0.5$	$\sigma = 1.0$
CIFAR-10	0.1	38.05 / 38.21	38.67 / 38.88
	0.2	46.57 / 46.58	**45.98** / 45.98
	0.3	50.37 / 50.50	45.01 / **46.63**
	0.5	**51.06 / 51.07**	39.58 / 43.48
	1.0	47.25 / 48.50	24.91 / 36.06
	1.5	39.81 / 44.02	16.41 / 30.73
SVHN	0.1	21.48 / 22.27	24.32 / 24.32
	0.2	37.44 / 37.44	49.56 / 49.56
	0.3	62.12 / 62.12	69.23 / 69.23
	0.5	79.10 / 79.30	**78.05 / 78.05**
	1.0	**83.25 / 83.42**	70.27 / 71.16
	1.5	80.30 / 80.92	56.95 / 63.04

B.4 Effect of Release Ratio

Table A3 varies the nominal active-block ratio γ. Smaller budgets reduce perturbation dimension and Gaussian noise energy. Excessive sparsification can discard informative updates and leave useful blocks rarely observed. As γ approaches 1, release approaches the full parameter space. The ratio therefore controls the tradeoff between information retention and perturbation energy.

The preferred release ratio depends on the task and noise level. CIFAR-10 favors smaller release ratios at both noise levels, whereas SVHN performs best at $\gamma = 0.20$ when $\sigma = 0.5$. The two EMNIST tasks are less sensitive to γ under weaker noise, but also benefit clearly from lower dimensional release under stronger noise. The main experiments use a common nominal $\gamma = 0.20$ across datasets and noise levels.

Table A3: Sensitivity of FedEASE to the nominal active-block ratio γ under Non-IID ($\beta = 0.5$) data with $C = 1$. Each entry reports **Final / Peak**: final test accuracy / peak test accuracy, in %. **Bold** marks each metric's maximum over γ for the same dataset and noise level.

Dataset	σ	$\gamma = 0.05$	$\gamma = 0.10$	$\gamma = 0.20$	$\gamma = 0.40$	$\gamma = 0.50$	$\gamma = 0.80$	$\gamma = 1.00$
CIFAR-10	0.5	**51.22 / 51.22**	50.25 / 50.83	47.25 / 48.50	44.36 / 45.86	39.45 / 44.51	37.02 / 43.40	36.34 / 41.60
	1.0	**39.70 / 42.22**	33.40 / 41.55	24.91 / 36.06	18.52 / 32.59	16.17 / 31.10	10.01 / 29.06	10.04 / 27.43
SVHN	0.5	72.72 / 72.72	81.62 / 82.01	**83.25 / 83.42**	82.69 / 82.90	80.05 / 81.17	77.74 / 79.94	77.20 / 79.33
	1.0	75.27 / 76.25	**75.64 / 76.88**	70.27 / 71.16	58.72 / 64.49	58.20 / 65.09	40.42 / 56.23	29.31 / 54.61
EMNIST-Letters	0.5	88.48 / 88.69	**88.95 / 89.05**	88.78 / 88.80	88.00 / 88.00	87.24 / 87.49	85.75 / 86.16	83.96 / 85.06
	1.0	**83.96 / 84.36**	81.97 / 81.97	76.90 / 78.86	70.28 / 71.98	68.22 / 69.25	63.13 / 64.75	55.99 / 58.22
EMNIST-ByClass	0.5	82.01 / 82.01	**82.70 / 82.70**	82.05 / 82.26	81.61 / 81.70	81.46 / 81.46	80.60 / 80.60	80.37 / 80.52
	1.0	**77.51 / 78.35**	77.20 / 77.96	75.57 / 75.57	71.10 / 71.54	71.33 / 71.43	69.56 / 69.92	67.54 / 67.94

B.5 Component Ablation

Table A4 evaluates the three components introduced in Section 4 by removing each in turn. Replacing entropy adaptive allocation with uniform shared selection sets $\beta_t = 0$ while retaining the exact-L support budget (Sec. 4.2). The variant *w/o Marginal Correction* replaces Eq. (8) with $u_i^t = D_t c_i^t$, while retaining the historical correction in Eq. (12). The variant *w/o Noise Aware Feedback* omits $m_b \nu_t^2$ from Eq. (11) and sets $r_t = 1$ in Eq. (13); historical accumulation and score normalization remain unchanged.

FedEASE achieves the highest final and peak accuracy in all eight settings. Entropy adaptive allocation contributes gains of 0.64–2.14 pp in final accuracy over its ablated variant. The corresponding gains are 1.14–2.39 pp for marginal correction and 0.85–2.75 pp for noise aware feedback. The gains show that adaptive selection, marginal correction, and noise aware feedback contribute complementary utility improvements across partitions and noise levels.

Table A4: Component ablation on CIFAR-10 and SVHN. Entries report **Final / Peak**: final test accuracy / peak test accuracy, in %. Each variant removes one component; FedEASE retains all three. **Bold** marks the highest accuracy in each setting.

Dataset	Partition	σ	w/o Entropy Adaptive Allocation	w/o Marginal Correction	w/o Noise Aware Feedback	FedEASE
CIFAR-10	IID	0.5	45.26 / 47.75	45.01 / 47.67	44.65 / 47.62	**47.40 / 48.11**
		1.0	33.27 / 37.58	32.65 / 37.60	33.15 / 37.19	**34.32 / 38.55**
	Non-IID	0.5	46.12 / 47.28	45.93 / 47.57	46.05 / 47.53	**47.25 / 48.50**
		1.0	23.35 / 34.74	23.07 / 34.38	23.52 / 34.81	**24.91 / 36.06**
SVHN	IID	0.5	83.71 / 83.85	83.05 / 83.18	83.36 / 83.43	**84.35 / 84.59**
		1.0	74.42 / 75.03	74.25 / 75.80	74.00 / 76.09	**75.39 / 76.65**
	Non-IID	0.5	82.13 / 83.10	81.93 / 82.49	82.11 / 83.37	**83.25 / 83.42**
		1.0	68.95 / 70.39	68.03 / 69.27	69.42 / 70.03	**70.27 / 71.16**

B.6 Effect of Support Selection

To separate support sharing from adaptation, we compare four rules at the same nominal block ratio $\gamma = 0.20$ and $K = 50$. Independent random draws an exact-L support for each client. Uniform shared redraws one common support per round, whereas shared fixed reuses a common support sampled at initialization. FedEASE draws a shared support from its adaptive marginals. The comparison uses the default non-IID protocol in Section 6.1. The uniform shared and FedEASE entries are the corresponding results from the component ablation; independent random and shared fixed provide the additional support controls.

All four rules perturb the aggregate on its realized coordinate support with the same variance $\nu^2 = (2\sigma C/K)^2$ per coordinate. Independent updates are embedded with zeros before averaging over the union. Uniform shared uses inclusion probability L/B for marginal correction, shared fixed uses 1 on its selected blocks, and FedEASE uses its current marginals. Correction precedes joint norm clipping. These rules characterize support design through aggregate dimension, temporal coverage, and the corrected updates passed to clipping.

Figure A2 shows higher final accuracy for both fixed and resampled shared supports than for independent supports in all four settings. Resampling a uniform shared support further improves over fixing it, consistent with broader coverage across rounds. FedEASE achieves the highest final accuracy, exceeding uniform shared by 1.12–1.56 pp across the four $K = 50$ settings. Adapting selection from historical DP statistics therefore improves utility at the same shared block budget.

B.7 Effect of Client Participation

We evaluate participation at two levels: public support simulations measure the aggregate dimension, and training experiments assess the resulting accuracy on non-IID CIFAR-10 and SVHN. The geometry study uses the LeNet-5 parameter count and the default release ratio, with 2,000 public simulations at each marked participation level in Figure A3. We compare independent client supports with a uniformly sampled shared support under the training noise calibration.

The training experiments vary $K \in \{20, 40, 50, 100\}$ at fixed $N = 500$, $C = 1$, noise multiplier σ, and training horizon. This gives participation rates $q_c \in \{0.04, 0.08, 0.10, 0.20\}$. Under the distributed-noise implementation, the per-coordinate standard deviation on each unaveraged client update is $2\sigma C/\sqrt{K}$, and averaging gives aggregate scale $\nu = 2\sigma C/K$. Increasing participation

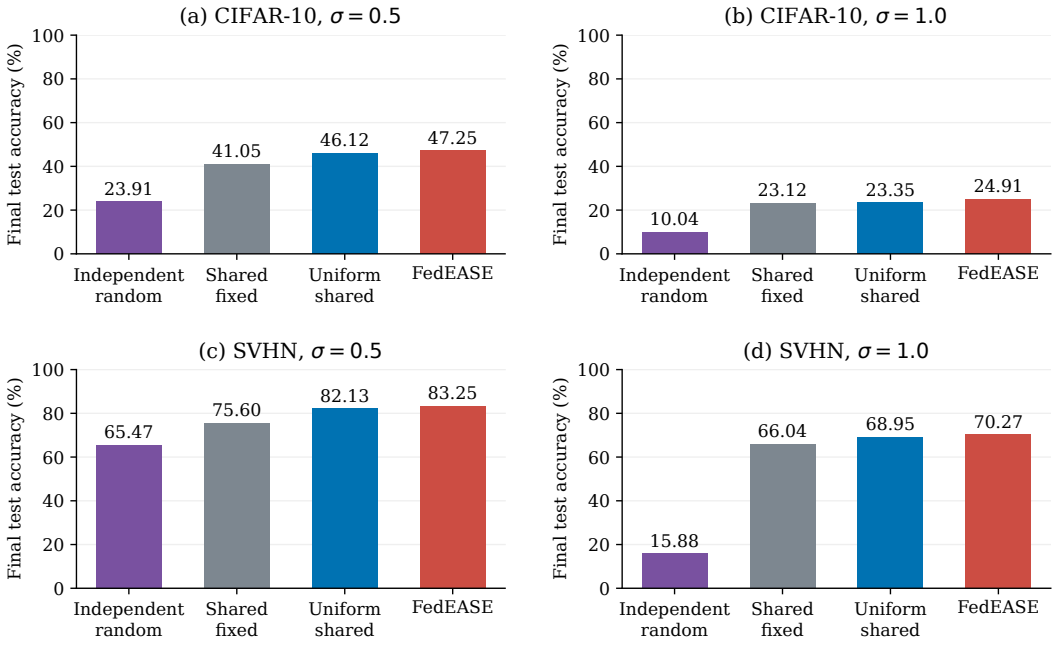

Figure A2: Final test accuracy under four support rules on non-IID CIFAR-10 and SVHN, with $K = 50$, $C = 1$, and nominal $\gamma = 0.20$. Independent random uses client-specific supports; the remaining rules share one support across clients. Shared fixed preserves it across rounds, uniform shared resamples it uniformly, and FedEASE adapts its marginals.

reduces both noise scales. We account for FedEASE's released transcript at each $q_c = K/N$ using Eqs. (27)–(29).

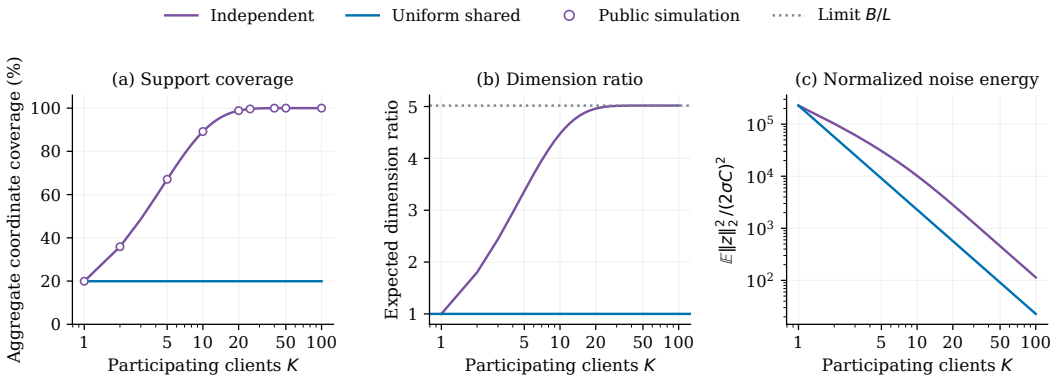

Figure A3: Participation and aggregate dimension for $d = 1{,}141{,}194$, $B = 1069$, and $L = 213$. (a) Expected coordinate coverage; markers average 2,000 public simulations. (b) Ratio of the expected dimensions, approaching B/L. (c) Expected Gaussian energy divided by $(2\sigma C)^2$, equal to $\mathbb{E}[D]/K^2$, where D is the aggregate coordinate count under the corresponding support rule. Noise is isotropic on the whole aggregate support. The energy curves are calculated from public support geometry and the specified Gaussian variance.

Figure A3 shows independent supports approaching full coordinate coverage, while uniform shared coverage remains near 20%. Their expected dimension ratio approaches $B/L \approx 5.02$. Absolute expected noise energy decreases as K grows, because $\nu^2 = 4\sigma^2C^2/K^2$ at fixed σ and C. Shared selection maintains the smaller perturbation space throughout this participation range.

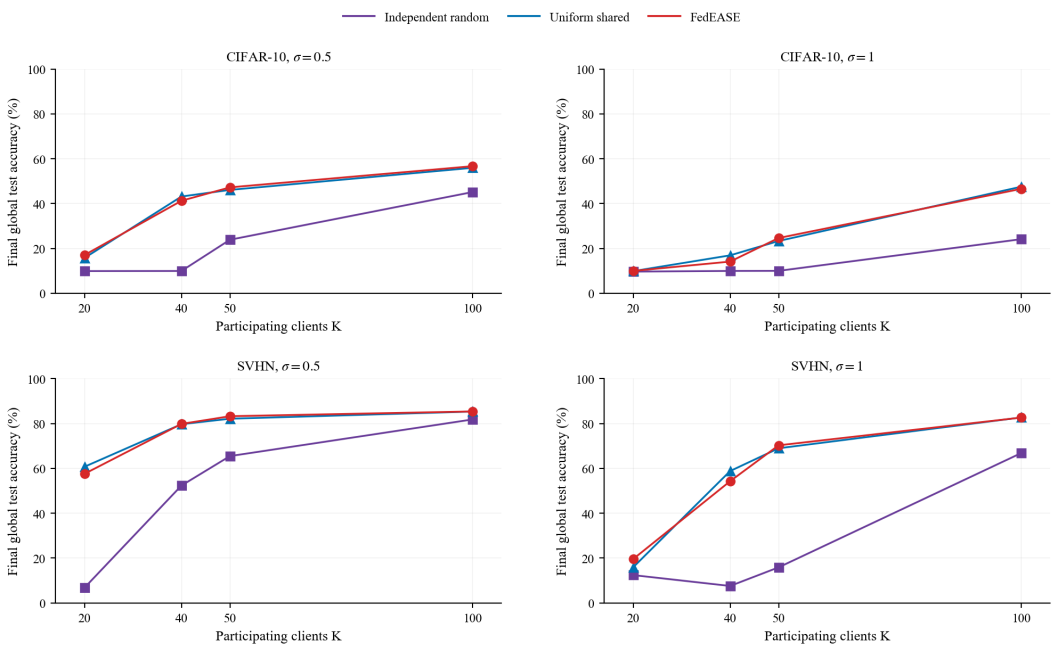

Figure A4: Final test accuracy as client participation varies on non-IID CIFAR-10 and SVHN. Independent random, uniform shared, and FedEASE use $K \in \{20, 40, 50, 100\}$, $N = 500$, $C = 1$, and nominal $\gamma = 0.20$. Each panel fixes σ, with aggregate noise scale $\nu = 2\sigma C/K$. The participation rate is $q_c = K/N$; FedEASE's aggregate transcript is accounted for by Eqs. (27)–(29).

Figure A4 connects this geometric comparison to model utility. Shared selection generally improves final accuracy over independent supports across the participation sweep. On CIFAR-10 at $\sigma = 0.5$, uniform shared reaches 43.17% versus 9.98% for independent random at $K = 40$, and 55.96% versus 45.14% at $K = 100$. The relative ordering of uniform and adaptive sharing varies with K, reflecting the interaction between client participation and allocation. At the default $K = 50$, FedEASE leads both alternatives in all four settings. Larger K combines more client updates with lower aggregate noise, while shared selection maintains a compact perturbation space throughout the sweep.

B.8 Effect of Allocation Concentration

To assess the concentration schedule in Eq. (7), we compare uniform shared allocation ($\beta_t = 0$), fixed concentrations $\beta_t \in \{0.25, 0.50, 0.75\}$, and FedEASE at $K = 50$. All variants share an exact-L support; the fixed values refer to allocation concentration, with the data partition held fixed. Figure A5 shows their training trajectories on CIFAR-10 and SVHN.

FedEASE's final accuracies are 47.25% and 24.91% on CIFAR-10, and 83.25% and 70.27% on SVHN, for $\sigma = 0.5$ and 1.0, respectively. They exceed uniform shared in all four settings, with an average absolute gain of 1.28 pp. FedEASE also exceeds all tested fixed concentrations in three settings. Averaged across settings, FedEASE comes within 0.10 pp of selecting the best fixed concentration separately for each setting. The bounded adaptive schedule therefore provides a practical way to use historical feedback while approaching the accuracy of per-setting concentration selection.

B.9 Extension to Deeper Model Architectures

We evaluate ResNet-18 (He et al., 2016) on CIFAR-10 and SVHN to test whether the gains extend beyond LeNet-5. CIFAR-10 uses 400 rounds with local lr $\eta_l = 0.03$; SVHN uses 200 rounds with local lr $\eta_l = 0.05$. Both retain $N = 500$, $K = 50$, 5 local epochs, batch size 64, $C = 1$, and $\gamma = 0.20$. We compare IID and non-IID partitions with $\beta = 0.5$; the SVHN model uses a CIFAR-style stem with group normalization.

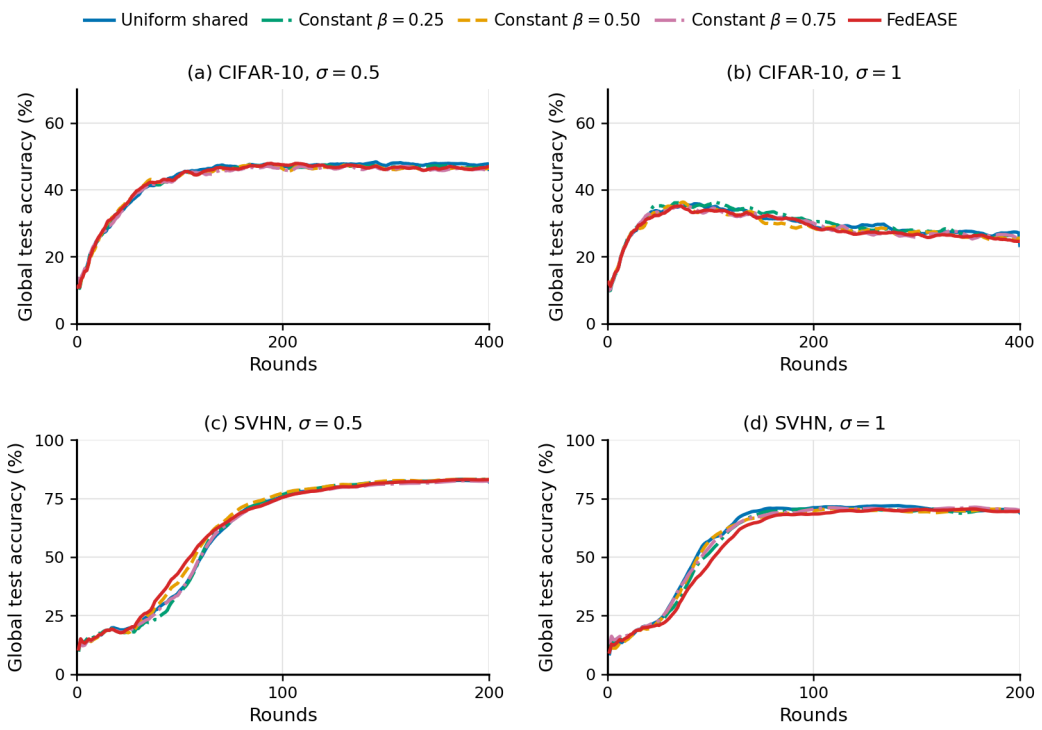

Figure A5: Training dynamics under uniform, fixed-concentration, and adaptive shared allocation on non-IID CIFAR-10 and SVHN, with $K = 50$, $C = 1$, and nominal $\gamma = 0.20$. The constant values specify β_t in Eq. (6); FedEASE uses Eq. (7). All variants retain shared exact-L selection.

Table A5 shows that FedEASE achieves the highest final accuracy in 11 of 12 settings and the highest peak accuracy in 11 of 12 settings. On CIFAR-10 IID at $\sigma = 1.0$, FedEASE reaches 36.70% final and 40.29% peak accuracy; DP-FedSAM reaches 36.81% and 37.86%, respectively. On SVHN IID at $\sigma = 0.75$, FedEASE reaches 80.24% in both metrics, compared with 78.97% final and 80.31% peak accuracy for FedBLUR.

Under stronger noise, the SVHN IID final-accuracy gain reaches 5.99 pp at $\sigma = 1.0$ (72.84% versus 66.85%). Peak accuracy also improves from 69.73% to 73.23%. At the same noise level under non-IID data, the final and peak gains are 1.91 and 2.23 pp, respectively. FedEASE thus retains its accuracy advantage with deeper models across data partitions and noise levels.

B.10 Evaluation on the IMDB Dataset

We evaluate text sentiment classification on IMDB (Maas et al., 2011) using the deep averaging multilayer perceptron (MLP) architecture of SlaClip (Zou et al., 2026), with 16-dimensional token embeddings, a 16-unit ReLU hidden layer, and a two-class output. We use IID and Dirichlet non-IID partitions with $\beta = 0.5$, 400 rounds, local lr $\eta_l = 0.1$, training and evaluation batch sizes of 32 and 100, and $\sigma \in \{3, 4, 5\}$. FedEASE uses the nominal active-block ratio $\gamma = 0.20$. Text is tokenized with the case-preserving bert-base-cased WordPiece vocabulary of 28,996 tokens using BertTokenizerFast. Reviews are right-truncated to 254 text tokens, enclosed by [CLS] and [SEP], and right-padded to 256 tokens with [PAD] (ID 0). Mean pooling excludes padding positions.

Table A6 compares accuracy under both data partitions. FedEASE achieves the highest final and peak accuracy among the compared DP methods in all six settings. Its final-accuracy gains over the strongest baseline range from 5.36 to 6.97 pp under IID data and from 4.99 to 7.53 pp under non-IID data. At $\sigma = 5$, FedEASE retains 61.46% final accuracy under IID data and 60.37% under non-IID data. The gains demonstrate the applicability of adaptive block-level release to text classification under both data partitions.

Table A5: ResNet-18 results on CIFAR-10 and SVHN. Entries report **Final / Peak**: final test accuracy / peak test accuracy, in %. **Bold** marks the highest DP accuracy separately for each metric and setting. FedAvg is a nonprivate method.

Dataset	Method	IID			Non-IID ($\beta = 0.5$)		
		$\sigma = 0.5$	$\sigma = 0.75$	$\sigma = 1.0$	$\sigma = 0.5$	$\sigma = 0.75$	$\sigma = 1.0$
CIFAR-10	UDP-FedAvg	41.32 / 42.85	34.05 / 36.68	32.91 / 33.81	43.47 / 45.25	36.14 / 37.36	32.38 / 33.59
	FedSMP	44.12 / 44.64	35.50 / 37.99	31.94 / 34.07	44.21 / 44.92	38.20 / 39.48	31.97 / 34.03
	FedBLUR	42.14 / 43.85	36.49 / 37.23	31.52 / 33.55	43.67 / 43.67	34.56 / 37.10	30.45 / 32.74
	FedCEO	42.78 / 45.00	35.28 / 38.08	32.83 / 33.71	35.65 / 38.10	30.86 / 32.86	27.49 / 29.64
	DP-FedSAM	41.48 / 42.47	40.66 / 42.02	**36.81** / 37.86	35.81 / 35.81	33.61 / 35.03	28.20 / 30.95
	FedEASE	**45.88 / 46.71**	**42.39 / 43.38**	36.70 / **40.29**	**45.88 / 45.88**	**40.77 / 41.30**	**37.21 / 37.30**
	FedAvg (non-DP)		57.08 / 59.65			56.55 / 56.89	
SVHN	UDP-FedAvg	82.40 / 84.42	78.11 / 79.59	65.04 / 65.48	73.21 / 74.03	66.70 / 67.36	59.46 / 59.46
	FedSMP	84.23 / 84.23	76.47 / 77.92	66.85 / 68.09	74.16 / 74.17	67.05 / 67.32	60.73 / 62.10
	FedBLUR	84.70 / 85.07	78.97 / **80.31**	64.25 / 69.73	74.25 / 74.63	67.05 / 67.77	61.25 / 61.25
	FedCEO	71.87 / 72.77	70.10 / 72.10	41.91 / 42.85	57.10 / 58.29	51.81 / 56.13	24.48 / 25.76
	DP-FedSAM	84.01 / 84.35	74.60 / 75.17	65.93 / 67.56	71.81 / 72.07	66.95 / 67.14	54.05 / 54.22
	FedEASE	**86.00 / 86.09**	**80.24** / 80.24	**72.84 / 73.23**	**75.42 / 75.47**	**67.58 / 68.53**	**63.16 / 64.33**
	FedAvg (non-DP)		93.20 / 93.39			91.20 / 91.28	

FedEASE combines this accuracy advantage with competitive computation time. On non-IID data, it requires 3.210 s per round, 10.18% less than the fastest DP baseline, DP-FedSAM. On IID data, it takes 5.745 s per round, compared with 5.433 s for DP-FedSAM.

Table A6: IMDB results under IID and non-IID ($\beta = 0.5$) partitions. Accuracy entries report **Final / Peak**: final test accuracy / peak test accuracy, in %. Time reports mean client and server computation per round at $\sigma = 3$, excluding global evaluation (Appendix B.12). **Bold** marks the highest DP accuracy for each metric and setting and the lowest DP time. FedAvg is nonprivate.

Dataset	Method	IID				Non-IID ($\beta = 0.5$)			
		Accuracy (%)			Time (s/round)	Accuracy (%)			Time (s/round)
		$\sigma = 3$	$\sigma = 4$	$\sigma = 5$		$\sigma = 3$	$\sigma = 4$	$\sigma = 5$	
IMDB	FedAvg (non-DP)	77.39 / 78.20			4.582	74.61 / 75.48			3.242
	UDP-FedAvg	58.52 / 60.98	57.81 / 61.12	51.36 / 59.30	6.775	57.96 / 59.99	54.45 / 59.30	52.48 / 56.30	4.162
	FedSMP	61.55 / 61.84	58.81 / 60.22	53.79 / 57.57	5.871	57.26 / 57.81	54.19 / 55.75	52.21 / 54.60	3.656
	FedBLUR	60.99 / 63.08	56.85 / 59.11	53.64 / 57.93	7.524	59.25 / 60.19	55.45 / 58.70	55.38 / 57.52	4.511
	FedCEO	58.65 / 58.86	56.90 / 57.41	54.49 / 56.04	6.971	56.23 / 57.22	55.35 / 57.42	54.54 / 55.91	4.307
	DP-FedSAM	49.97 / 54.28	49.82 / 54.36	50.05 / 53.94	**5.433**	49.80 / 53.94	49.85 / 54.35	49.84 / 51.48	3.574
	FedEASE	**66.91 / 66.91**	**65.26 / 65.26**	**61.46 / 61.74**	5.745	**64.42 / 66.49**	**62.98 / 63.38**	**60.37 / 61.53**	**3.210**

B.11 Communication Efficiency

We evaluate the cumulative uplink payload required to reach 40% test accuracy on CIFAR-10 and 82% on SVHN. Both tasks use ResNet-18, IID partitions, and $\sigma = 0.5$, with $N = 500$, $K = 50$, and FedEASE release ratio $\gamma = 0.20$. The target round is the first crossing in the raw per-round test accuracy.

We measure expected cumulative uplink payload per client as $U_m(t) = 4q_c 10^{-6} \sum_{\tau=1}^{t} d_{m,\tau}$ MB, where $q_c = K/N = 0.1$, m indexes methods, t is the number of completed rounds, and $d_{m,\tau}$ counts transmitted update values in round τ. Each value uses 32-bit floating point, and 1 MB is 10^6 bytes; downlink, coordinate indices, and protocol overhead are excluded. UDP-FedAvg, FedBLUR, FedCEO, and DP-FedSAM upload full vectors of $d = 11{,}185{,}578$ values. FedSMP uploads 20% of the coordinates, whereas FedEASE transmits the k_τ coordinates in its shared release support.

On CIFAR-10, FedEASE reaches the target at round 59 with 52.72 MB, the lowest payload among the compared methods (Figure A6). FedSMP, the most communication-efficient baseline, requires 69 rounds and 61.74 MB. FedEASE thus reaches the same accuracy 10 rounds earlier with 14.61%

less payload. Relative to full-vector UDP-FedAvg at 277.40 MB, the reduction is 81.00%. FedEASE combines sparse uploads with fewer rounds to the target.

On SVHN, FedEASE reaches the target at round 130 with 116.16 MB (Figure A7). FedSMP again has the lowest baseline payload, requiring 153 rounds and 136.91 MB; FedEASE reduces the payload by 15.16% and reaches the target 23 rounds earlier. FedBLUR reaches the target sooner, at round 113, but its full-vector uploads require 505.59 MB, giving FedEASE a 77.02% payload reduction. FedCEO remains below the target after 200 rounds and 894.85 MB. Together, these results show that FedEASE combines compact uploads with efficient progress toward the target accuracy, yielding the lowest uplink payload on both tasks.

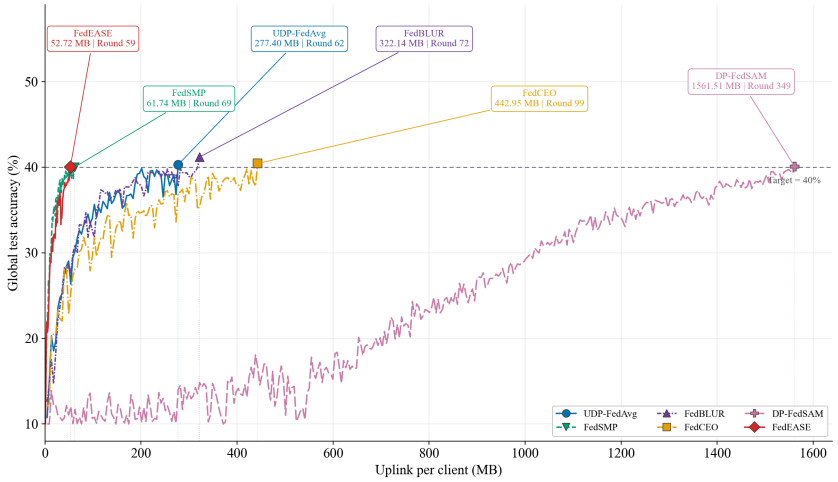

Figure A6: Expected cumulative uplink payload per client to reach 40% accuracy on CIFAR-10 with ResNet-18 under IID data and $\sigma = 0.5$. FedEASE reaches the target with the lowest payload. Markers indicate the first raw-accuracy crossing; annotations report cumulative payload and round. Curves end at that crossing.

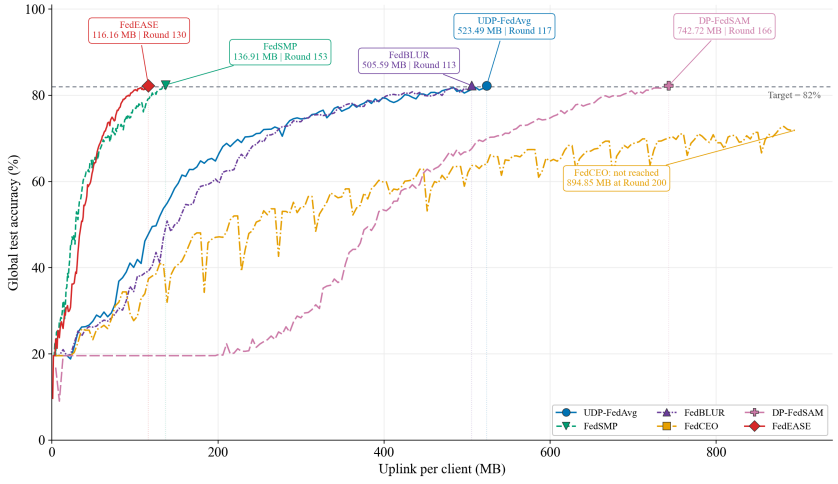

Figure A7: Expected cumulative uplink payload per client to reach 82% accuracy on SVHN with ResNet-18 under IID data and $\sigma = 0.5$. FedEASE uses the lowest payload among methods that reach the target. Markers indicate the first raw-accuracy crossing; annotations report cumulative payload and round. Curves end at that crossing; FedCEO remains below the target, so its curve extends to the 200-round training horizon.

B.12 RUNTIME OVERHEAD

Table A7 decomposes mean per-round computation into the client phase and server algorithm components. Total is their sum, excluding global evaluation. Table 1 reports complete wall-clock time per round, including evaluation.

Table A7 shows that FedEASE has the lowest total runtime among the DP methods in all eight image settings. On non-IID IMDB, FedEASE takes 3.210 s per round, compared with 3.574 s for the fastest DP baseline, DP-FedSAM, a reduction of 10.18%. Its client phase and server computation take 3.174 s and 0.036 s, respectively. On IID IMDB, the corresponding total times are 5.745 s for FedEASE and 5.433 s for DP-FedSAM. Shared allocation and feedback account for a small portion of the total computation time.

Table A7: Client-side, server-side, and total runtime (s/round) under IID and Non-IID settings. Total equals client-side plus server-side runtime. Bold marks the lowest client-side and total time among DP methods in each setting. FedAvg is a nonprivate method.

Dataset	Method	IID			Non-IID		
		Client-side	Server-side	Total	Client-side	Server-side	Total
CIFAR-10	UDP-FedAvg	6.408	0.014	6.422	6.333	0.014	6.347
	FedSMP	6.256	0.052	6.308	6.126	0.044	6.170
	FedBLUR	8.019	0.026	8.045	7.852	0.025	7.876
	FedCEO	6.476	0.013	6.489	6.432	0.014	6.446
	DP-FedSAM	6.877	0.026	6.903	6.647	0.026	6.673
	FedEASE	**6.160**	0.050	**6.209**	**6.107**	0.049	**6.157**
	FedAvg (non-DP)	5.333	0.012	5.345	5.277	0.012	5.289
SVHN	UDP-FedAvg	7.966	0.014	7.980	8.166	0.014	8.180
	FedSMP	7.701	0.049	7.750	8.005	0.046	8.051
	FedBLUR	9.978	0.026	10.005	10.503	0.026	10.529
	FedCEO	7.912	0.014	7.926	8.196	0.013	8.209
	DP-FedSAM	8.080	0.026	8.106	8.224	0.026	8.250
	FedEASE	**7.271**	0.049	**7.320**	**7.180**	0.048	**7.228**
	FedAvg (non-DP)	6.760	0.012	6.772	7.034	0.012	7.046
EMNIST-Letters	UDP-FedAvg	13.830	0.014	13.844	15.764	0.014	15.778
	FedSMP	13.605	0.044	13.649	15.761	0.044	15.804
	FedBLUR	17.531	0.026	17.557	20.421	0.026	20.446
	FedCEO	13.725	0.013	13.738	15.834	0.014	15.848
	DP-FedSAM	14.162	0.026	14.187	16.590	0.026	16.616
	FedEASE	**12.879**	0.045	**12.924**	**14.739**	0.048	**14.787**
	FedAvg (non-DP)	12.424	0.012	12.436	14.459	0.012	14.471
EMNIST-ByClass	UDP-FedAvg	85.655	0.014	85.669	87.324	0.014	87.338
	FedSMP	88.630	0.044	88.674	89.341	0.045	89.386
	FedBLUR	115.800	0.026	115.825	117.570	0.025	117.595
	FedCEO	86.059	0.014	86.073	86.270	0.013	86.283
	DP-FedSAM	94.622	0.026	94.648	96.345	0.026	96.370
	FedEASE	**80.029**	0.045	**80.074**	**81.148**	0.045	**81.194**
	FedAvg (non-DP)	83.041	0.012	83.053	82.996	0.012	83.008
IMDB	UDP-FedAvg	6.749	0.026	6.775	4.155	0.006	4.162
	FedSMP	5.824	0.047	5.871	3.631	0.025	3.656
	FedBLUR	7.485	0.039	7.524	4.492	0.019	4.511
	FedCEO	6.954	0.017	6.971	4.302	0.006	4.307
	DP-FedSAM	**5.414**	0.019	**5.433**	3.556	0.018	3.574
	FedEASE	5.703	0.042	5.745	**3.174**	0.036	**3.210**
	FedAvg (non-DP)	4.511	0.070	4.582	3.237	0.005	3.242

C Theoretical Proofs

C.1 Proof for the Union of Independent Client Supports

Proof. Consider any public block b. If each participating client independently includes this block with the common marginal probability $p_{\text{uni}} = L/B$, then the probability that the block appears in at least one client support is $1 - (1 - p_{\text{uni}})^K$. Therefore,

$$\begin{aligned}\mathbb{E}\left|\bigcup_{i\in\mathcal{S}_t}\mathcal{B}_i^t\right| &\overset{(1)}{=} \sum_{b=1}^{B}\Pr\left(b\in\bigcup_{i\in\mathcal{S}_t}\mathcal{B}_i^t\right)\\ &\overset{(2)}{=} B\left[1-(1-p_{\text{uni}})^K\right]\\ &\overset{(3)}{\geq} B\left(1-e^{-Kp_{\text{uni}}}\right),\end{aligned} \tag{19}$$

where (1) expands the union size using indicator variables and linearity of expectation; (2) uses independent support selection across clients; and (3) uses $1 - x \leq e^{-x}$. If all clients share the same $\mathcal{B}_t$, the number of active blocks at aggregation is always L, while the number of active coordinates remains $k_t = \sum_{b\in\mathcal{B}_t} m_b$. This gives the independent support union bound used in Sec. 4.1. This completes the proof. □

C.2 Moments of the Gaussian Debiased Block Energy

Fix the public history, realized support, participating clients, and their clipped contributions at round t, and let $\mathcal{J}_t$ denote this pre-noise state. For an active block b, define $\mu_b^t = K^{-1}\sum_{i\in\mathcal{S}_t} x_{i,b}^t \in \mathbb{R}^{m_b}$. Then

$$\widetilde{y}_b^t = \mu_b^t + \bar{z}_b^t, \qquad \bar{z}_b^t \mid \mathcal{J}_t \sim \mathcal{N}(0, \nu_t^2 I_{m_b}).$$

Here, $\bar{z}_b^t$ is the block-b component of the aggregate noise. Expectations and variances are conditional on $\mathcal{J}_t$ and taken only over this newly generated Gaussian noise.

Proof. From Eq. (11) in the main text,

$$\begin{aligned}S_b^t &= \|\mu_b^t + \bar{z}_b^t\|_2^2 - m_b\nu_t^2\\ &\overset{(1)}{=} \|\mu_b^t\|_2^2 + 2\mu_b^{t\top}\bar{z}_b^t + \|\bar{z}_b^t\|_2^2 - m_b\nu_t^2.\end{aligned} \tag{20}$$

Equality (1) expands the squared norm. Since $\mathbb{E}[\bar{z}_b^t \mid \mathcal{J}_t] = 0$ and $\mathbb{E}[\|\bar{z}_b^t\|_2^2 \mid \mathcal{J}_t] = m_b\nu_t^2$, we obtain

$$\mathbb{E}[S_b^t \mid \mathcal{J}_t] \overset{(2)}{=} \|\mu_b^t\|_2^2, \tag{21}$$

where (2) uses the first and second moments of a centered isotropic Gaussian. Furthermore,

$$\begin{aligned}\mathrm{Var}(\mu_b^{t\top}\bar{z}_b^t \mid \mathcal{J}_t) &\overset{(3)}{=} \nu_t^2\|\mu_b^t\|_2^2,\\ \mathrm{Var}(\|\bar{z}_b^t\|_2^2 \mid \mathcal{J}_t) &\overset{(4)}{=} 2m_b\nu_t^4,\end{aligned} \tag{22}$$

where (3) uses the variance of a linear Gaussian projection; (4) uses the second moment of a χ^2 variable. The cross covariance between the two terms is zero because the corresponding expansion contains only odd order moments of a centered Gaussian. Therefore,

$$\begin{aligned}\mathrm{Var}(S_b^t \mid \mathcal{J}_t) &\overset{(5)}{=} 4\,\mathrm{Var}(\mu_b^{t\top}\bar{z}_b^t \mid \mathcal{J}_t) + \mathrm{Var}(\|\bar{z}_b^t\|_2^2 \mid \mathcal{J}_t)\\ &\overset{(6)}{=} 4\nu_t^2\|\mu_b^t\|_2^2 + 2m_b\nu_t^4,\end{aligned} \tag{23}$$

where (5) uses the zero cross covariance; (6) substitutes Eq. (22). This completes the proof. □

C.3 Sensitivity of Uniform Aggregation

Proof. Fix the same public history, shared support, and participating client set $\mathcal{S}_t$. Suppose adjacent datasets $\mathcal{D} \simeq \mathcal{D}'$ differ only in the complete local data of client i_0. If $i_0 \notin \mathcal{S}_t$, the mean query at the current round is unchanged. If $i_0 \in \mathcal{S}_t$, define $\bar{x}^t = K^{-1} \sum_{i \in \mathcal{S}_t} x_i^t$ and let $\bar{x}'^t$ denote the corresponding mean under $\mathcal{D}'$. Then

$$\begin{aligned} \|\bar{x}^t - \bar{x}'^t\|_2 &\overset{(1)}{=} \frac{1}{K} \|x_{i_0}^t - x_{i_0}'^t\|_2 \\ &\overset{(2)}{\leq} \frac{1}{K} \left(\|x_{i_0}^t\|_2 + \|x_{i_0}'^t\|_2 \right) \\ &\overset{(3)}{\leq} \frac{2C_t}{K}, \end{aligned} \tag{24}$$

where (1) uses uniform client aggregation; (2) uses the triangle inequality; and (3) uses $\|x_i^t\|_2 \leq C_t$. Therefore, the ℓ_2 sensitivity of the mean query at the current round can be bounded by $\Delta_t = 2C_t/K$. Since marginal correction is completed before clipping, the final sensitivity depends only on the bound on clipped client contributions. If local optimization contains data independent randomness, the adjacent executions can first be coupled using the same random seed, for which the bound holds pointwise. Marginalizing over this randomness cannot increase the subsequent RDP upper bound. This completes the proof. □

C.4 Proof of Theorem 5.1

The release at round t consists of $(\mathcal{B}_t, \widetilde{y}^t)$; model states and public feedback are postprocessing of these releases. The sampled-client set and its sampling randomness remain internal to the coordinator. We condition on a common released history $\mathcal{H}_t$ and apply fixed-size subsampling amplification (Wang et al., 2019).

C.4.1 Single-Round Subsampled Gaussian Mechanism

Proof. Conditional on $\mathcal{H}_t$, the shared support $\mathcal{B}_t$ is generated from public parameters and fresh support randomness. It is independent of the current client sample and local training randomness. Fix a support realization. The coordinator draws a uniform size-K subset of the N clients, independently of private data. All selected clients complete secure aggregation, and their Gaussian shares give

$$\bar{z}^t = \sum_{i \in \mathcal{S}_t} z_i^t \sim \mathcal{N}(0, \nu_t^2 I_{k_t}), \qquad \nu_t = \frac{2\sigma_t C_t}{K}. \tag{25}$$

For any two size-K collections differing in one client contribution, joint clipping gives

$$\left\| K^{-1} \sum_{i=1}^{K} x_i^t - K^{-1} \sum_{i=1}^{K} x_i^{t\prime} \right\|_2 \leq \frac{2C_t}{K}, \qquad \varepsilon_{\mathrm{base},t}^{\mathrm{RDP}}(\alpha) \leq \frac{\alpha}{2\sigma_t^2}. \tag{26}$$

The base mechanism therefore has noise-to-sensitivity ratio σ_t, uniformly over the fixed history and support. Local randomization can be coupled before applying this bound.

Specializing Theorem 9 of Wang et al. (2019) to the Gaussian base mechanism yields, for every integer $\alpha \geq 2$,

$$\begin{aligned} b_2(\sigma) &= \min\{4(e^{1/\sigma^2} - 1),\, 2e^{1/\sigma^2}\}, \\ Z_\alpha(q, \sigma) &= 1 + q^2 \binom{\alpha}{2} b_2(\sigma) + 2 \sum_{j=3}^{\alpha} \binom{\alpha}{j} q^j e^{j(j-1)/(2\sigma^2)}, \\ a_{\mathrm{WOR}}(\alpha, q, \sigma) &= \min \left\{ \frac{\alpha}{2\sigma^2},\, \frac{\log Z_\alpha(q, \sigma)}{\alpha - 1} \right\}. \end{aligned} \tag{27}$$

Here q denotes the participation rate, set to q_c in FedEASE; b_2 is the second-order coefficient, Z_α is the moment-bound expression, and the sum is empty for $\alpha = 2$. The Gaussian base RDP at

order j is $j/(2\sigma^2)$, and its infinite-order bound is unbounded, giving the coefficients above. The first term also bounds the release: coupling the same sampled set under adjacent datasets bounds the conditional Gaussian divergence by $\alpha/(2\sigma^2)$, and marginalizing the sample preserves this upper bound. The replacement bound holds in both adjacency directions. The bound is uniform over supports. Since their conditional distribution is data independent, releasing the support alongside the aggregate preserves the bound. This establishes the single-round RDP guarantee for the observer in Section 3.3.

C.4.2 Composition and Noise Calibration

Adaptive composition gives the following RDP upper bound at each integer order $\alpha \geq 2$:

$$\mathcal{R}_\alpha = \sum_{t=0}^{T-1} a_{\mathrm{WOR}}(\alpha, q_c, \sigma_t). \tag{28}$$

The noise schedule is public and fixed before training. RDP-to-DP conversion gives the privacy bound

$$\varepsilon_{\mathcal{A}}(\delta) = \min_{\alpha \in \mathcal{A}} \left\{ \mathcal{R}_\alpha + \frac{\log(1/\delta)}{\alpha - 1} \right\}. \tag{29}$$

Each support and its inclusion probabilities depend on the preceding released history and fresh public randomness, so the conditional bounds compose throughout training. Noise aware feedback in Section 4.4, subsequent supports, and model states are postprocessing of the released DP aggregates. Each new aggregate contributes the per-round cost in Eq. (28), and Eq. (29) bounds the full transcript. We next derive the noise condition in Theorem 5.1.

Quadratic Amplification Regime. Equation (27) provides a finite-order bound for every positive noise multiplier. Let $\lambda_\delta = \log(1/\delta)$, $0 < \varepsilon \leq \lambda_\delta$, and $\alpha_\varepsilon = \lceil 1 + 2\lambda_\delta/\varepsilon \rceil$. Fix an amplification constant $c_0 > 0$, independently of the sampling rate and noise scale, and require

$$a_{\mathrm{WOR}}(\alpha_\varepsilon, q_c, \sigma) \leq c_0 \frac{q_c^2 \alpha_\varepsilon}{\sigma^2}. \tag{30}$$

For example, fixing $c_0 = 4$ defines the regime by substitution into Eq. (27). For a target budget ε, we use an order set $\mathcal{A}$ containing α_ε. If $\sigma_t \geq \sigma$ and Eq. (16) holds with $c_2 = \sqrt{8c_0}$, then

$$\mathcal{R}_{\alpha_\varepsilon} \overset{(1)}{\leq} c_0 \frac{T q_c^2 \alpha_\varepsilon}{\sigma^2} \overset{(2)}{\leq} \frac{\alpha_\varepsilon \varepsilon^2}{8\lambda_\delta} \overset{(3)}{\leq} \frac{\varepsilon}{2},$$

$$\mathcal{R}_{\alpha_\varepsilon} + \frac{\lambda_\delta}{\alpha_\varepsilon - 1} \overset{(4)}{\leq} \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Here (1) uses Eq. (30), composition, and monotonicity in σ; (2) uses Eq. (16); (3) uses $\alpha_\varepsilon \leq 2 + 2\lambda_\delta/\varepsilon \leq 4\lambda_\delta/\varepsilon$; and (4) uses $\alpha_\varepsilon - 1 \geq 2\lambda_\delta/\varepsilon$ and RDP-to-DP conversion. This proves Theorem 5.1. □

Table A8 evaluates the experimental settings directly through Eq. (29).

Table A8: Per-run upper bounds on FedEASE's aggregate-transcript privacy budget at $q_c = 0.1$ and $\delta = 10^{-5}$, computed from the subsampled Gaussian RDP bound and conversion in Eqs. (27)–(29).

Tasks	T	Noise multiplier σ: privacy bound ε		
CIFAR-10	400	0.5: 306.75	0.75: 56.25	1: 32.69
SVHN / EMNIST	200	0.5: 159.13	0.75: 33.88	1: 22.10
IMDB	400	3: 9.04	4: 7.18	5: 6.36

C.5 Proof of Theorem 5.2

Proof. First consider $1 \leq L < B$. Let $u = L/B$, $\pi^t_{\min} = \min_b \pi^t_b$, and define $a_{\mathrm{ent}} = \exp(\beta_{\max})$. From the sigmoid solution in Sec. 4.2, for any blocks b, b',

$$\log \frac{\pi_b^t}{1-\pi_b^t} - \log \frac{\pi_{b'}^t}{1-\pi_{b'}^t} = \beta_t(h_b^t - h_{b'}^t) \le \beta_{\max}. \tag{31}$$

Choose b' as a block attaining the minimum marginal inclusion probability. Since both the logit and sigmoid functions are monotonically increasing,

$$\begin{aligned} \frac{\pi_b^t}{1-\pi_b^t} &\overset{(1)}{\le} a_{\text{ent}} \frac{\pi_{\min}^t}{1-\pi_{\min}^t}, \\ \pi_b^t &\overset{(2)}{\le} \frac{a_{\text{ent}}\pi_{\min}^t}{1+(a_{\text{ent}}-1)\pi_{\min}^t}, \end{aligned} \tag{32}$$

where (1) exponentiates Eq. (31); (2) follows by algebraic rearrangement. Averaging over blocks and using $\sum_b \pi_b^t = L$ yields

$$\begin{aligned} u = \frac{1}{B}\sum_{b=1}^{B} \pi_b^t &\overset{(3)}{\le} \frac{a_{\text{ent}}\pi_{\min}^t}{1+(a_{\text{ent}}-1)\pi_{\min}^t} \\ &\overset{(4)}{\Longrightarrow} \pi_{\min}^t \ge \frac{u}{a_{\text{ent}}-(a_{\text{ent}}-1)u} = p_{\min}, \end{aligned} \tag{33}$$

where (3) uses the uniform upper bound from Eq. (32); (4) directly rearranges the inequality. This proves that every block has a uniform positive inclusion probability at every round.

Let t_0 denote the starting round, and define $A_j = \bigcap_{r=0}^{j-1}\{\mathbb{I}_b^{t_0+r} = 0\}$, where A_0 is the certain event. Since A_j is determined by the public history before support generation at round $t_0 + j$,

$$\begin{aligned} \Pr(A_{j+1} \mid \mathcal{H}_{t_0}) &\overset{(5)}{=} \mathbb{E}\left[\mathbf{1}_{A_j} \Pr\left(\mathbb{I}_b^{t_0+j} = 0 \mid \mathcal{H}_{t_0+j}\right) \middle| \mathcal{H}_{t_0}\right] \\ &\overset{(6)}{\le} (1-p_{\min}) \Pr(A_j \mid \mathcal{H}_{t_0}), \end{aligned} \tag{34}$$

where (5) uses the tower property and measurability of A_j; (6) uses $\Pr(b \in \mathcal{B}_{t_0+j} \mid \mathcal{H}_{t_0+j}) \ge p_{\min}$. Iterating the recurrence for H rounds and setting $\mathcal{U}_{b,t_0}^{(H)} = A_H$ gives Eq. (17) in the main text. The per-round conditional lower bound accommodates dependencies across training rounds. When $L = B$, every block has inclusion probability one. This completes the proof. □

C.6 Convergence Analysis of FedEASE

C.6.1 Assumptions and Proof Notation

Let $\mathcal{H}_t$ denote the public history before sampling the support, sampling clients, drawing local stochastic gradients, and generating Gaussian noise at round t. It contains previously released DP outputs and supports, their postprocessed statistics, and the current public protocol parameters. Define $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{H}_t]$. Assume $N \ge 2$, $1 \le K \le N$, $Q \ge 1$, and $\eta_l, \eta_g > 0$.

Assumption C.1 (Smoothness and Lower Bound). There exists $L_F > 0$ such that, for all $w, v \in \mathbb{R}^d$, every F_i is L_F smooth:

$$\|\nabla F_i(w) - \nabla F_i(v)\|_2 \le L_F \|w - v\|_2. \tag{35}$$

In addition, the global objective admits a finite lower bound

$$F_* = \inf_w F(w) > -\infty. \tag{36}$$

Let $\mathcal{F}^{t,q}$ collect the round history, sampled clients, shared support, and all selected clients' local iterates and randomness before local step q. For Assumption C.2, set $\mathcal{F}_i^{t,q} = \mathcal{F}^{t,q}$.

Assumption C.2 (Conditionally Independent, Unbiased Stochastic Gradients with Bounded Variance). Given $\mathcal{F}^{t,q}$, the current mini-batch draws are independent across selected clients. The stochastic gradients satisfy

$$\mathbb{E}\left[g_i(w_i^{t,q};\xi_i^{t,q}) \mid \mathcal{F}_i^{t,q}\right] = \nabla F_i(w_i^{t,q}), \qquad \mathbb{E}\left[\|g_i(w_i^{t,q};\xi_i^{t,q}) - \nabla F_i(w_i^{t,q})\|_2^2 \mid \mathcal{F}_i^{t,q}\right] \le V_g. \quad (37)$$

Here $V_g \ge 0$ bounds the conditional variance of each client's stochastic gradient.

Assumption C.3 (Bounded Second Moment). There exists $G \ge 0$ such that

$$\mathbb{E}\left[\|g_i(w_i^{t,q};\xi_i^{t,q})\|_2^2 \mid \mathcal{F}_i^{t,q}\right] \le G^2. \quad (38)$$

Assumption C.4 (Bounded Client Heterogeneity). There exist $\beta_h^2 \ge 1$ and $\kappa_h^2 \ge 0$ such that, for every w,

$$\frac{1}{N}\sum_{i=1}^{N}\|\nabla F_i(w)\|_2^2 \le \beta_h^2\|\nabla F(w)\|_2^2 + \kappa_h^2. \quad (39)$$

Here β_h^2 and κ_h^2 bound multiplicative and additive client heterogeneity, respectively; neither is the data-partition concentration β. Since $\nabla F(w) = N^{-1}\sum_i \nabla F_i(w)$, it further follows that

$$\frac{1}{N}\sum_{i=1}^{N}\|\nabla F_i(w) - \nabla F(w)\|_2^2 \le (\beta_h^2 - 1)\|\nabla F(w)\|_2^2 + \kappa_h^2.$$

Define the finite population coefficient for fixed size sampling without replacement and two aggregate quantities:

$$\chi_K = \frac{N-K}{K(N-1)}, \qquad H_K = 1 + \chi_K(\beta_h^2 - 1), \qquad \mathcal{V}_K = \frac{V_g}{K} + \chi_K\kappa_h^2. \quad (40)$$

Here χ_K is the finite-population sampling factor, H_K is the resulting gradient-moment factor, and $\mathcal{V}_K$ combines stochastic-gradient and client-sampling variance. We reuse the active coordinate selection operator from Sec. 4. Let $\widetilde{\pi}^t \in \mathbb{R}^{k_t}$ be obtained by copying each block level inclusion probability to the corresponding active coordinates, and define

$$\mathcal{Q}_t(v) = D_t^\top\left(\mathrm{Diag}(\widetilde{\pi}^t)^{-1}D_t v\right), \qquad c^t = \frac{1}{K}\sum_{i\in\mathcal{S}_t} c_i^t. \quad (41)$$

Here $\mathcal{Q}_t : \mathbb{R}^d \to \mathbb{R}^d$ is the corrected support operator and c^t is the mean uncompressed client displacement. The diagonal scaling divides each active coordinate by its block's inclusion probability. Support selection remains governed by the block marginals π_b^t. We further define

$$\Gamma_t = \frac{1}{K}\sum_{i\in\mathcal{S}_t}(\|u_i^t\|_2 - C_t)_+, \qquad \Phi_t^2 = \mathbb{E}_t\|\mathcal{Q}_t(c^t) - c^t\|_2^2. \quad (42)$$

Here $(a)_+ = \max\{a, 0\}$, Γ_t bounds the clipping-error norm, and Φ_t^2 is the conditional mean squared support error. The cross-round averages used in the main text are

$$\bar{\Gamma}^2 = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\Gamma_t^2], \qquad \bar{\Phi}^2 = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\Phi_t^2], \qquad \overline{k\nu^2} = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[k_t\nu_t^2]. \quad (43)$$

C.6.2 Second Moment of Marginal Correction

For $v \in \mathbb{R}^d$, let v_b be its block-b subvector and let $\iota_b : \mathbb{R}^{m_b} \to \mathbb{R}^d$ insert that subvector into block b and zeros elsewhere. Equation (41) becomes

$$\mathcal{Q}_t(v) = \sum_{b=1}^{B} \frac{\mathbb{I}_b^t}{\pi_b^t} \iota_b(v_b), \qquad \mathbb{I}_b^t = \mathbf{1}\{b \in \mathcal{B}_t\}.$$

Fix v, or let a random v be conditionally independent of the current support given $\mathcal{H}_t$. By marginal preserving rounding,

$$\begin{aligned} \mathbb{E}[\mathcal{Q}_t(v) \mid \mathcal{H}_t, v] &\overset{(1)}{=} \sum_{b=1}^{B} \frac{\mathbb{E}[\mathbb{I}_b^t \mid \mathcal{H}_t, v]}{\pi_b^t} \iota_b(v_b) \\ &\overset{(2)}{=} \sum_{b=1}^{B} \iota_b(v_b) = v, \end{aligned} \tag{44}$$

where (1) expands $\mathcal{Q}_t$ over public blocks and uses linearity of conditional expectation; (2) uses $\mathbb{E}[\mathbb{I}_b^t \mid \mathcal{H}_t, v] = \pi_b^t$ together with the fact that the public blocks partition the full coordinate space.

Furthermore, because different public blocks occupy disjoint coordinates in the full Euclidean space,

$$\begin{aligned} &\mathbb{E}\left[\|\mathcal{Q}_t(v) - v\|_2^2 \mid \mathcal{H}_t, v\right] \\ &\overset{(3)}{=} \sum_{b=1}^{B} \mathbb{E}\left[\left(\frac{\mathbb{I}_b^t}{\pi_b^t} - 1\right)^2 \middle| \mathcal{H}_t, v\right] \|v_b\|_2^2 \\ &\overset{(4)}{=} \sum_{b=1}^{B} \left(\frac{1}{\pi_b^t} - 1\right) \|v_b\|_2^2, \end{aligned} \tag{45}$$

where (3) uses orthogonality of nonoverlapping blocks; (4) uses $(\mathbb{I}_b^t)^2 = \mathbb{I}_b^t$ and $\mathbb{E}[\mathbb{I}_b^t \mid \mathcal{H}_t, v] = \pi_b^t$. Block orthogonality reduces the variance calculation to marginal inclusion probabilities, so Eq. (45) applies to exact-L dependent rounding.

Let c_b^t be the block-b subvector of c^t. Client sampling and local training are conditionally independent of the current support given $\mathcal{H}_t$; setting $v = c^t$ and applying iterated expectation yields

$$\begin{aligned} \Phi_t^2 &\overset{(5)}{=} \mathbb{E}_t \sum_{b=1}^{B} \left(\frac{1}{\pi_b^t} - 1\right) \|c_b^t\|_2^2 \\ &\overset{(6)}{\leq} \left(\frac{1}{\pi_{\min}^t} - 1\right) \mathbb{E}_t \|c^t\|_2^2, \end{aligned} \tag{46}$$

where (5) follows from Eq. (45) and the tower property; (6) uses $\pi_{\min}^t = \min_b \pi_b^t$. The same conditional unbiasedness in Eq. (44) also gives $\mathbb{E}_t\langle c^t, \mathcal{Q}_t(c^t) - c^t\rangle = 0$. These identities describe marginal correction before clipping. The following decomposition accounts for the nonlinear clipping error.

C.6.3 Proof of Theorem 5.3

Local Multi-step Updates. Let $g_i^{t,q} = g_i(w_i^{t,q}; \xi_i^{t,q})$ and $g^t = \nabla F(w^t)$. From Sec. 3.1 and Sec. 4.3,

$$c_i^t = \eta_l \sum_{q=0}^{Q-1} g_i^{t,q}, \qquad c^t = \frac{1}{K} \sum_{i \in \mathcal{S}_t} c_i^t. \tag{47}$$

Let $\bar{g}^{t,q} = K^{-1} \sum_{i \in \mathcal{S}_t} g_i^{t,q}$. By successively adding and subtracting the exact local gradients of participating clients at $w_i^{t,q}$ and w^t,

$$\bar{g}^{t,q} - g^t = \underbrace{\frac{1}{K}\sum_{i\in\mathcal{S}_t}\left(g_i^{t,q} - \nabla F_i(w_i^{t,q})\right)}_{\xi^{t,q}} + \underbrace{\frac{1}{K}\sum_{i\in\mathcal{S}_t}\left(\nabla F_i(w_i^{t,q}) - \nabla F_i(w^t)\right)}_{d^{t,q}} + \underbrace{\frac{1}{K}\sum_{i\in\mathcal{S}_t}\nabla F_i(w^t) - \nabla F(w^t)}_{s_{\text{cli}}^t}, \tag{48}$$

where $\xi^{t,q}$, $d^{t,q}$, and s_{cli}^t denote stochastic gradient noise, local model drift, and client sampling error, respectively. By Assumption C.2 and uniform client sampling without replacement, $\mathbb{E}_t[\xi^{t,q}] = 0$ and $\mathbb{E}_t[s_{\text{cli}}^t] = 0$.

Expanding the local SGD iterates and using Assumptions C.1 and C.3,

$$\begin{aligned}
\mathbb{E}_t\|w_i^{t,q} - w^t\|_2^2 &\overset{(1)}{=} \eta_l^2\mathbb{E}_t\left\|\sum_{r=0}^{q-1} g_i^{t,r}\right\|_2^2 \\
&\overset{(2)}{\leq} \eta_l^2 q\sum_{r=0}^{q-1}\mathbb{E}_t\|g_i^{t,r}\|_2^2 \\
&\overset{(3)}{\leq} \eta_l^2 q^2 G^2, \\
\mathbb{E}_t\|d^{t,q}\|_2^2 &\overset{(4)}{\leq} \frac{L_F^2}{K}\mathbb{E}_t\left[\sum_{i\in\mathcal{S}_t}\|w_i^{t,q} - w^t\|_2^2\right] \\
&\overset{(5)}{\leq} L_F^2\eta_l^2 q^2 G^2,
\end{aligned} \tag{49}$$

where (1) expands the local SGD recursion; (2) uses $\|\sum_{r<q} a_r\|_2^2 \leq q\sum_{r<q}\|a_r\|_2^2$; (3) uses Assumption C.3; (4) uses Jensen's inequality and L_F smoothness; and (5) substitutes the preceding bound.

Fixed size sampling without replacement further gives

$$\begin{aligned}
\mathbb{E}_t\|\xi^{t,q}\|_2^2 &\overset{(6)}{\leq} \frac{V_g}{K}, \\
\mathbb{E}_t\|s_{\text{cli}}^t\|_2^2 &\overset{(7)}{=} \chi_K\frac{1}{N}\sum_{i=1}^{N}\|\nabla F_i(w^t) - g^t\|_2^2 \\
&\overset{(8)}{\leq} \chi_K\left[(\beta_h^2 - 1)\|g^t\|_2^2 + \kappa_h^2\right],
\end{aligned} \tag{50}$$

where (6) conditions first on $\mathcal{F}^{t,q}$: Assumption C.2 eliminates cross-client noise inner products and bounds the sum of K variances by KV_g; division by K^2 and iterated expectation give V_g/K; (7) is the finite population variance identity for uniform sampling without replacement; and (8) uses Assumption C.4.

Lemma C.6.1 (Local Update Bound). Under Assumptions C.1–C.4,

$$
\begin{aligned}
\mathbb{E}_t\langle g^t, c^t\rangle &\overset{(9)}{=} \eta_l \sum_{q=0}^{Q-1} \mathbb{E}_t\langle g^t, \bar{g}^{t,q}\rangle \\
&\overset{(10)}{\geq} \eta_l Q\|g^t\|_2^2 - \eta_l \sum_{q=0}^{Q-1}\left(\frac{1}{4}\|g^t\|_2^2 + \mathbb{E}_t\|d^{t,q}\|_2^2\right) \\
&\overset{(11)}{\geq} \frac{3}{4}\eta_l Q\|g^t\|_2^2 - \frac{1}{3}\eta_l^3 Q^3 L_F^2 G^2,
\end{aligned}
\tag{51}
$$

where (9) uses Eq. (47); (10) substitutes Eq. (48), uses the conditional zero means of $\xi^{t,q}$ and s_{cli}^t, and applies $-\langle x, y\rangle \leq \|x\|_2^2/4 + \|y\|_2^2$ to the drift inner product; and (11) uses Eq. (49) together with $\sum_{q=0}^{Q-1} q^2 \leq Q^3/3$.

On the other hand,

$$
\begin{aligned}
\mathbb{E}_t\|c^t\|_2^2 &\overset{(12)}{\leq} \eta_l^2 Q \sum_{q=0}^{Q-1} \mathbb{E}_t\|\bar{g}^{t,q}\|_2^2 \\
&\overset{(13)}{\leq} 4\eta_l^2 Q \sum_{q=0}^{Q-1}\left(H_K\|g^t\|_2^2 + \mathcal{V}_K + L_F^2\eta_l^2 q^2 G^2\right) \\
&\overset{(14)}{\leq} 4\eta_l^2 Q^2\left(H_K\|g^t\|_2^2 + \mathcal{V}_K\right) + \frac{4}{3}\eta_l^4 Q^4 L_F^2 G^2,
\end{aligned}
\tag{52}
$$

where (12) uses Cauchy–Schwarz; (13) expands $\bar{g}^{t,q} = g^t + \xi^{t,q} + d^{t,q} + s_{\text{cli}}^t$ into four terms, applies $\|a+b+c+d\|_2^2 \leq 4(\|a\|_2^2 + \|b\|_2^2 + \|c\|_2^2 + \|d\|_2^2)$, and substitutes Eqs. (49)–(50) and Eq. (40); and (14) uses $\sum_{q=0}^{Q-1} q^2 \leq Q^3/3$. This completes the proof of Lemma C.6.1. □

Mechanism Error Decomposition. Let $\bar{x}^t = K^{-1}\sum_{i\in\mathcal{S}_t} x_i^t$ and define $s^t = \mathcal{Q}_t(c^t) - c^t$, $r^t = D_t^\top \bar{x}^t - \mathcal{Q}_t(c^t)$, and $z^t = D_t^\top \bar{z}^t$. Here s^t, r^t, and z^t are the support error, clipping error, and embedded Gaussian noise; r^t differs from the scalar weight r_t. The full space update applied by the server is therefore $\Delta^t = D_t^\top \widetilde{y}^t = c^t + s^t + r^t + z^t$. By Appendix C.6.2, clipping, and the Gaussian mechanism,

$$
\mathbb{E}_t\langle g^t, s^t\rangle = 0, \qquad \mathbb{E}_t\|s^t\|_2^2 = \Phi_t^2, \qquad \mathbb{E}_t[z^t] = 0, \qquad \mathbb{E}_t\|z^t\|_2^2 = \mathbb{E}_t[k_t\nu_t^2].
$$

The Gaussian second moment is first conditioned on the realized support and then averaged over the support, allowing public blocks to have different sizes. Writing $\mathrm{Clip}_{C_t}(v) = v/\max\{1, \|v\|_2/C_t\}$, the clipping error satisfies

$$
\begin{aligned}
\|r^t\|_2 &\overset{(15)}{\leq} \frac{1}{K}\sum_{i\in\mathcal{S}_t} \|\,\mathrm{Clip}_{C_t}(u_i^t) - u_i^t\|_2 \\
&\overset{(16)}{=} \frac{1}{K}\sum_{i\in\mathcal{S}_t}(\|u_i^t\|_2 - C_t)_+ = \Gamma_t,
\end{aligned}
\tag{53}
$$

where (15) uses the triangle inequality and $D_t^\top \bar{x}^t - \mathcal{Q}_t(c^t) = K^{-1}\sum_i D_t^\top[\mathrm{Clip}_{C_t}(u_i^t) - u_i^t]$; (16) uses the exact excess norm under Euclidean clipping together with the isometric embedding property of $D_t^\top$.

From Eq. (51) and Young's inequality,

$$
\begin{aligned}
\mathbb{E}_t\langle g^t, \Delta^t\rangle &\overset{(17)}{=} \mathbb{E}_t\langle g^t, c^t\rangle + \mathbb{E}_t\langle g^t, r^t\rangle \\
&\overset{(18)}{\geq} \frac{5}{8}\eta_l Q\|g^t\|_2^2 - \frac{1}{3}\eta_l^3 Q^3 L_F^2 G^2 - \frac{2\mathbb{E}_t[\Gamma_t^2]}{\eta_l Q},
\end{aligned}
\tag{54}
$$

where (17) uses the conditional zero means of the support error and Gaussian noise; (18) applies $-\langle x, y\rangle \le \frac{\eta_l Q}{8}\|x\|_2^2 + \frac{2}{\eta_l Q}\|y\|_2^2$, to $\langle g^t, r^t\rangle$ and then uses Eq. (53).

On the other hand,

$$\begin{aligned}\mathbb{E}_t\|\Delta^t\|_2^2 &\overset{(19)}{\le} 3\mathbb{E}_t\|c^t + s^t\|_2^2 + 3\mathbb{E}_t\|r^t\|_2^2 + 3\mathbb{E}_t\|z^t\|_2^2 \\ &\overset{(20)}{=} 3\mathbb{E}_t\|c^t\|_2^2 + 3\Phi_t^2 + 3\mathbb{E}_t\|r^t\|_2^2 + 3\mathbb{E}_t[k_t\nu_t^2] \\ &\overset{(21)}{\le} 3\mathbb{E}_t\|c^t\|_2^2 + 3\Phi_t^2 + 3\mathbb{E}_t[\Gamma_t^2] + 3\mathbb{E}_t[k_t\nu_t^2],\end{aligned} \tag{55}$$

where (19) uses the three term squared norm inequality; (20) expands $\|c^t + s^t\|_2^2$ and uses $\mathbb{E}_t\langle c^t, s^t\rangle = 0$; and (21) uses Eq. (53).

One Round Descent. By L_F smoothness and $w^{t+1} = w^t - \eta_g\Delta^t$,

$$\begin{aligned}\mathbb{E}_t[F(w^{t+1}) - F(w^t)] &\overset{(22)}{\le} -\eta_g\mathbb{E}_t\langle g^t, \Delta^t\rangle + \frac{L_F\eta_g^2}{2}\mathbb{E}_t\|\Delta^t\|_2^2 \\ &\overset{(23)}{\le} -\left(\frac{5}{8}\eta_g\eta_l Q - 6L_F\eta_g^2\eta_l^2Q^2H_K\right)\|g^t\|_2^2 \\ &\quad + \frac{1}{3}\eta_g\eta_l^3Q^3L_F^2G^2 + 6L_F\eta_g^2\eta_l^2Q^2\mathcal{V}_K + 2L_F^3\eta_g^2\eta_l^4Q^4G^2 \\ &\quad + \frac{2\eta_g}{\eta_l Q}\mathbb{E}_t[\Gamma_t^2] + \frac{3L_F\eta_g^2}{2}\left(\Phi_t^2 + \mathbb{E}_t[\Gamma_t^2] + \mathbb{E}_t[k_t\nu_t^2]\right),\end{aligned} \tag{56}$$

where (22) is the smoothness descent lemma; (23) substitutes Eqs. (52), (54), and (55), and groups the resulting terms according to local drift, client sampling, clipping, support randomization, and Gaussian noise.

Under the step size conditions of Theorem 5.3,

$$0 < \eta_l \le \frac{1}{QL_F}, \qquad 0 < \eta_g \le \frac{1}{48L_F\eta_l Q H_K}$$

we have

$$6L_F\eta_g^2\eta_l^2Q^2H_K \le \frac{1}{8}\eta_g\eta_l Q.$$

Moreover, since $H_K \ge 1$, we have $L_F\eta_g\eta_l Q \le 1/48$. Therefore, the higher order local drift term satisfies

$$2L_F^3\eta_g^2\eta_l^4Q^4G^2 \le \frac{1}{24}\eta_g\eta_l^3Q^3L_F^2G^2,$$

and the second order clipping term satisfies

$$\frac{3L_F\eta_g^2}{2}\mathbb{E}_t[\Gamma_t^2] \le \frac{1}{32}\frac{\eta_g}{\eta_l Q}\mathbb{E}_t[\Gamma_t^2].$$

Applying conservative relaxations to these terms, Eq. (56) gives

$$\begin{aligned}\mathbb{E}_t[F(w^{t+1}) - F(w^t)] &\overset{(24)}{\le} -\frac{1}{2}\eta_g\eta_l Q\|g^t\|_2^2 + \frac{1}{2}\eta_g\eta_l^3Q^3L_F^2G^2 \\ &\quad + 6L_F\eta_g^2\eta_l^2Q^2\mathcal{V}_K + \frac{3\eta_g}{\eta_l Q}\mathbb{E}_t[\Gamma_t^2] \\ &\quad + \frac{3L_F\eta_g^2}{2}\left(\Phi_t^2 + \mathbb{E}_t[k_t\nu_t^2]\right),\end{aligned} \tag{57}$$

where (24) uses the step size constraints above and conservatively relaxes the constants. Summing over $t = 0, \ldots, T-1$, taking full expectation, and using $F(w^T) \geq F_*$ yields

$$\begin{aligned}\frac{1}{2}\eta_g\eta_l Q \sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(w^t)\|_2^2 \overset{(25)}{\leq}\ & F(w^0) - F_* + \frac{1}{2}T\eta_g\eta_l^3 Q^3 L_F^2 G^2 \\ & + 6TL_F\eta_g^2\eta_l^2 Q^2 \mathcal{V}_K + \frac{3T\eta_g}{\eta_l Q}\bar{\Gamma}^2 \\ & + \frac{3TL_F\eta_g^2}{2}\left(\bar{\Phi}^2 + \overline{k\nu^2}\right),\end{aligned}$$

where (25) uses telescoping summation and the three cross-round averages in Eq. (43). Dividing both sides by $T\eta_g\eta_l Q/2$ gives

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(w^t)\|_2^2 \leq\ & \frac{2(F(w^0) - F_*)}{T\eta_g\eta_l Q} + \eta_l^2 Q^2 L_F^2 G^2 \\ & + 12L_F\eta_g\eta_l Q\mathcal{V}_K + \frac{6\bar{\Gamma}^2}{(\eta_l Q)^2} \\ & + \frac{3L_F\eta_g}{\eta_l Q}\left(\bar{\Phi}^2 + \overline{k\nu^2}\right).\end{aligned} \tag{58}$$

This matches Eq. (18) in the main text. This completes the proof. □

Control of Support Variance by the Entropy Lower Bound. From $\pi_b^t \geq p_{\min} > 0$ in Theorem 5.2 and Eq. (46),

$$\Phi_t^2 \overset{(26)}{\leq} \left(\frac{1}{p_{\min}} - 1\right)\mathbb{E}_t\|c^t\|_2^2. \tag{59}$$

Inequality (26) follows from the uniform lower bound on inclusion probability. Further, by the conditional unbiasedness in Eq. (44),

$$\begin{aligned}\mathbb{E}_t\|\mathcal{Q}_t(c^t)\|_2^2 &\overset{(27)}{=} \mathbb{E}_t\|c^t\|_2^2 + \Phi_t^2 \\ &\overset{(28)}{\leq} \frac{1}{p_{\min}}\mathbb{E}_t\|c^t\|_2^2,\end{aligned} \tag{60}$$

where (27) expands $\mathcal{Q}_t(c^t) = c^t + [\mathcal{Q}_t(c^t) - c^t]$ and uses conditional unbiasedness to eliminate the cross term; (28) substitutes Eq. (59). Marginal preservation removes first-order support bias before clipping, while the entropy lower bound controls the second-moment amplification of inverse marginal correction.