FEDSALT: SUBSPACE-DRIFT-GUIDED ADAPTIVE LOCAL TRAINING FOR FEDERATED LORA FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated low-rank adaptation (LoRA) enables efficient fine-tuning of large language models (LLMs) over decentralized data by communicating only compact low-rank adapters. However, heterogeneous client tasks naturally produce task-specific LoRA updates, making it difficult to learn a single global model through aggregation. We observe an empirical asymmetry in this heterogeneity: task differentiation is substantially more pronounced in the output-side geometry of LoRA updates than in the input-side geometry. Motivated by this observation, we propose FedSALT, a geometry-aware closed-loop optimization framework that maintains a momentum-based consensus subspace, measures client-module disagreement using a factorization-invariant output-subspace score computed directly from the realized LoRA update, and converts the score into adaptive proximal regularization for subsequent local optimization. Unlike existing methods that apply uniform regularization or intervene only during aggregation, FedSALT selectively guides drifting modules toward a momentum-lookahead consensus anchor while retaining standard factor-wise FedAvg aggregation. Experiments on three heterogeneous federated LLM benchmarks show that FedSALT consistently achieves the best overall accuracy among representative baselines. Mechanism studies further validate the motivating output-side asymmetry and demonstrate the effectiveness of adaptive geometry-aware control.

1 INTRODUCTION

Low-rank adaptation (LoRA) has become a practical parameterization for federated fine-tuning of large language models (LLMs) because clients optimize and communicate only compact adapter factors (Hu et al., 2022; Zhang et al., 2024; Babakniya et al., 2023; Sun et al., 2024). The standard objective is to learn a single global model from decentralized client data. This objective becomes increasingly challenging when clients represent different tasks or domains. Although every client starts each communication round from the same global adapter, local optimization naturally steers the adapter toward task-dependent solutions, producing heterogeneous low-rank updates before aggregation. The resulting task-dependent drift not only complicates optimization but also makes it increasingly difficult for a single global LoRA model to capture knowledge shared across heterogeneous clients. An effective federated LoRA algorithm therefore requires not only efficient aggregation, but also a mechanism for identifying and regulating task-dependent drift during local training.

To understand how such drift should be regulated, we first examine where it appears in the geometry of LoRA updates. Our empirical analysis (Figure 1 and Figure 2) reveals a clear asymmetry between the two sides of the realized LoRA update. When measured separately, the output-side subspace exhibits substantially larger variation across clients and tasks, whereas the input-side remains comparatively consistent. Moreover, clients belonging to the same task retain much higher output-side similarity than unrelated task pairs. This asymmetry is consistent with recent studies of heterogeneous federated LoRA, which report that the input factor tends to capture shared representations while the output factor carries stronger client-specific information (Zhu et al., 2024; Guo et al., 2025; Wang et al., 2025). Together, these findings suggest that task-dependent drift is primarily reflected in the output subspace of LoRA updates.

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.

However, existing federated optimization methods do not exploit this characterization of task-dependent drift. FedProx (Li et al., 2020a) applies a uniform proximal regularization, while SCAFFOLD (Karimireddy et al., 2020) and related control-variate methods reduce optimization drift through correction states. When applied to federated LoRA, these methods remain agnostic to which client-module pairs exhibit stronger output-subspace divergence. Federated LoRA methods that freeze or share adapter factors similarly reduce the optimization flexibility of all clients, regardless of whether individual modules actually require stronger regularization (Sun et al., 2024; Guo et al., 2025). Alignment-aware and exact aggregation methods improve how completed updates are combined at the server, but they intervene only after local trajectories have already diverged (Wang et al., 2024; et al., 2025; Zhang et al., 2026).

This limitation becomes important because aggregation itself cannot distinguish benign variation within a shared output subspace from persistent departures into task-specific subspaces. Consequently, when heterogeneous client updates occupy different output subspaces, averaging may attenuate or even cancel directions that are locally beneficial, degrading the learning of a single global model. The challenge is therefore not merely gradient heterogeneity, but the emergence of client-dependent low-rank geometries before aggregation. A desirable optimization strategy should identify where persistent task-dependent drift occurs, accumulate this information over time to suppress transient fluctuations, and selectively regulate only the affected client-module pairs before the next aggregation step.

To address this challenge, we propose **FedSALT** (Federated Subspace-drift-guided Adaptive Local Training), a geometry-aware closed-loop optimization framework for heterogeneous federated LoRA. For each adapted module, the server maintains a momentum-based estimate of historical consensus updates and extracts its dominant output subspace. Each incoming client update is then evaluated according to the fraction of its energy lying within this consensus subspace, producing a factorization-invariant measure of output-side alignment. The alignment score is temporally smoothed and converted into a client-specific, module-specific proximal strength. During the following communication round, modules exhibiting persistent disagreement are pulled toward a momentum-lookahead consensus update, while well-aligned modules remain only lightly constrained. FedSALT therefore establishes a closed-loop optimization framework that senses, smooths, allocates, and corrects task-specific drift before it propagates to the next aggregation step.

Our contributions are summarized below.

- We formulate heterogeneous federated LoRA drift as disagreement between client updates and a historical consensus output subspace, linking task differentiation to a measurable low-rank update geometry.
- We introduce a factorization-invariant and scale-invariant output-subspace alignment score that is computed directly from the LoRA factors without explicitly constructing the full weight update.
- We propose FedSALT, a closed-loop optimization framework that converts temporally smoothed alignment scores into client-specific and module-specific proximal regularization using a momentum-lookahead consensus anchor.
- We evaluate FedSALT on heterogeneous federated language-model benchmarks and demonstrate the effectiveness of output-subspace sensing, adaptive regularization, and momentum-lookahead guidance through extensive ablation studies.

2 Problem Formulation and Observations

Consider a federated setting with I clients. Each client $i \in \{1, \dots, I\}$ holds a local dataset $\mathcal{D}_i$ that may originate from a distinct downstream task. All clients share a frozen pretrained LLM with L layers, where the weight of layer l is denoted $W_l^0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ and d_{in}, d_{out} are the input and output dimensions. For parameter-efficient fine-tuning, each adapted module l carries a LoRA update (Hu et al., 2022)

$$\Delta W_l = \frac{\alpha}{r} B_l A_l, \qquad B_l \in \mathbb{R}^{d_{\text{out}} \times r}, \quad A_l \in \mathbb{R}^{r \times d_{\text{in}}}, \tag{1}$$

where $r \ll \min(d_{\text{in}}, d_{\text{out}})$ is the LoRA rank and α is the scaling coefficient. The adapted weight of layer l is therefore $W_l = W_l^0 + \Delta W_l$.

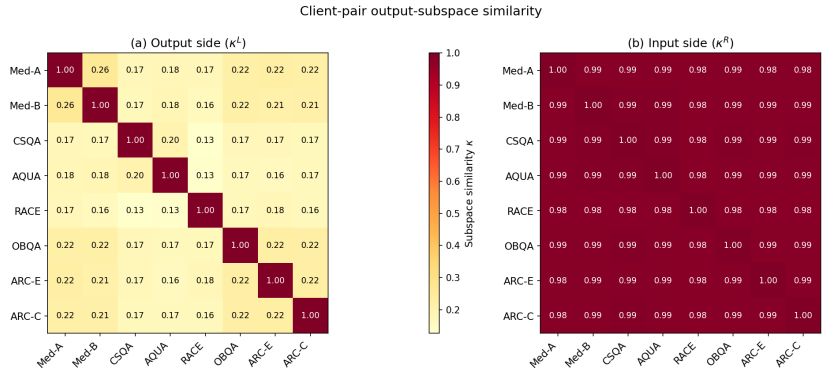


Figure 1: Client-pair LoRA subspace similarity under task heterogeneity. Output-side subspaces exhibit substantial and pair-dependent differentiation, while input-side subspaces remain nearly identical across clients. The two MedMCQA clients, which share the same underlying task, are most similar on the output side, but the overall structure remains continuous rather than forming discrete client clusters.

The federated fine-tuning objective is to learn global adapters $\{A_l, B_l\}_{l=1}^{L}$ that minimize the data-size-weighted client objectives,

$$\min_{\{A_l,B_l\}_{l=1}^{L}} \sum_{i=1}^{I} p_i \, \mathcal{L}_i\big(\{A_l, B_l\}_{l=1}^{L}; \mathcal{D}_i\big), \qquad p_i = \frac{|\mathcal{D}_i|}{\sum_{j=1}^{I} |\mathcal{D}_j|}, \tag{2}$$

where $\mathcal{L}_i$ is the local task loss of client i and p_i is its data-size weight.

At communication round t, the server broadcasts the global factors $\{A_l^t, B_l^t\}_{l=1}^{L}$. Each client i performs K local optimization steps and returns $\{A_{i,l}^{t+1}, B_{i,l}^{t+1}\}_{l=1}^{L}$. After collecting the local factors from all clients, the server applies standard data-size-weighted FedAvg:

$$A_l^{t+1} = \sum_{i=1}^{I} p_i A_{i,l}^{t+1}, \qquad B_l^{t+1} = \sum_{i=1}^{I} p_i B_{i,l}^{t+1}. \tag{3}$$

2.1 Empirical Observations

We first examine how task heterogeneity is reflected in the geometry of realized LoRA updates. For each client and LoRA module, we compute the dominant output- and input-side subspaces of the realized update. We define the *subspace similarity* between two clients as the overlap of their corresponding dominant subspaces, where a larger value indicates that the two updates occupy more similar directions.

Figure 1 reveals a clear asymmetry between the two sides of the LoRA update. The input-side subspaces remain highly aligned across all client pairs, suggesting that they primarily capture shared structure. In contrast, output-side similarity varies substantially across clients. Notably, the two MedMCQA clients exhibit considerably higher similarity than most cross-task pairs, indicating that task-specific information is largely reflected in the output-side geometry. Moreover, the similarities change continuously across client pairs rather than forming a small number of isolated clusters, suggesting that task differentiation is gradual instead of discrete.

The pairwise analysis identifies where client differentiation occurs, but does not directly provide a client-specific signal for online control. We therefore compare each update with the historical consensus subspace using the alignment scores $s_{i,l}^{L,t}$ and $s_{i,l}^{R,t}$ defined in Eq. (10). Figure 2 shows that the resulting output-side score remains substantially more variable across clients than its input-side counterpart.

These empirical observations directly motivate the design of FedSALT. First, output-subspace alignment provides an informative signal of task-dependent drift because it captures substantial client-

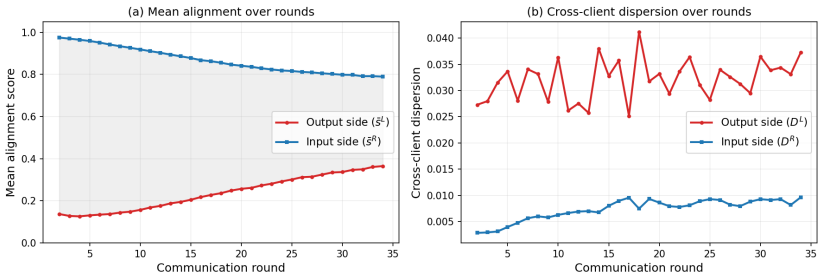


Figure 2: Alignment with the historical consensus during federated training. The left panel reports mean output-side and input-side alignment over communication rounds. The right panel reports their cross-client dispersion. Output-side alignment is consistently lower and substantially more dispersed, whereas the input-side score remains high and comparatively homogeneous.

and task-specific variation, whereas the consistently high input-side alignment contributes little additional information for distinguishing heterogeneous clients. Second, the continuous distribution of output-side alignment scores indicates that drift varies gradually across client-module pairs, motivating adaptive client- and module-specific regularization rather than a uniform coefficient or hard clustering. Third, the variation of alignment over communication rounds suggests that the signal should be temporally smoothed before it is used to regulate subsequent local optimization. Together, these observations determine what should be measured and how the regularization strength should be allocated. The remaining question is how identified drifting updates should be guided toward a shared optimization trajectory, which motivates the adaptive optimization framework introduced in the next subsection.

3 METHOD

To instantiate these requirements, we propose **FedSALT** (Federated Subspace-drift-guided Adaptive Local Training). FedSALT addresses two complementary questions: *how strongly* each client-module update should be regulated and *toward what* it should be guided. The output-subspace disagreement identified above determines the adaptive regularization strength. For the correction target, a zero-centered proximal penalty is insufficient because it suppresses all local movement, including directions that agree with recent global progress. FedSALT therefore adopts a low-rank momentum of historical consensus updates as a momentum-lookahead reference.

The resulting method forms a one-round-delayed feedback loop. At communication round t, the server updates the historical consensus reference, evaluates each client-module update by its output-subspace alignment, and converts the disagreement into a temporally smoothed proximal strength. During round $t+1$, each client uses this strength to regulate local optimization toward the momentum-lookahead reference. Algorithm 1 summarizes the complete procedure.

3.1 CONSENSUS-DERIVED REFERENCE SUBSPACES

FedSALT first constructs a compact historical reference that summarizes recent global optimization directions. This reference not only defines the subspaces used for drift sensing but also provides the momentum-lookahead anchor for subsequent local optimization.

Let the realized product-space update of client i at LoRA module l be

$$\Delta_{i,l}^{t} = \frac{\alpha}{r}\left(B_{i,l}^{t+1}A_{i,l}^{t+1} - B_{l}^{t}A_{l}^{t}\right), \tag{4}$$

where the product difference is used because the effective model update is determined by the change in BA rather than by either factor individually. The server then forms the data-size-weighted consensus displacement

$$\bar{\Delta}_{l}^{t} = \sum_{i=1}^{m} p_i \Delta_{i,l}^{t}, \qquad p_i = \frac{|\mathcal{D}_i|}{\sum_j |\mathcal{D}_j|}, \tag{5}$$

while continuing to aggregate the LoRA factors with standard FedAvg as in Eq. equation 3. The product-space consensus is therefore used exclusively for sensing and local control, whereas the

factor-space aggregation determines the deployed global model. The distinction is deliberate because FedSALT regulates update geometry before aggregation, while any residual mismatch introduced by factor averaging is handled by the aggregation procedure itself.

Since a single communication round provides only a noisy estimate of the global optimization direction, the server maintains a low-rank exponential moving average (EMA):

$$M_l^t = \Pi_{r_M}\left[\rho M_l^{t-1} + (1-\rho)\bar{\Delta}_l^t\right], \qquad M_l^0 = 0. \tag{6}$$

Here $\rho \in [0, 1)$ controls the temporal horizon of the consensus reference and Π_{r_M} denotes the best rank-r_M approximation. Smaller values of ρ adapt more quickly to recent consensus updates, whereas larger values provide a smoother but less responsive reference. Because the EMA is constructed solely from client updates already available at the server, no auxiliary server data are required.

Let

$$M_l^t = U_{M,l}^t \Sigma_l^t (V_{M,l}^t)^\top \tag{7}$$

be the truncated singular value decomposition maintained by the server. The dominant output-side and input-side reference subspaces are defined as

$$U_l^t = U_{M,l}^t[:, 1 : r_{\text{ref}}], \qquad V_l^t = V_{M,l}^t[:, 1 : r_{\text{ref}}], \tag{8}$$

and the rank-r_{ref} momentum used by local optimization is

$$\check{M}_l^t = \Pi_{r_{\text{ref}}}(M_l^t) = \check{F}_l^t \check{G}_l^t. \tag{9}$$

The reference subspaces are used for drift sensing in the next subsection, while the truncated momentum serves as the lookahead reference in the local objective.

3.2 Output-Side Subspace-Drift Sensing

With the reference subspaces of Eq. equation 8 in hand, the next step is to quantify how closely each client update aligns with the recent consensus geometry. To avoid evaluating an update against a reference that already contains itself, client i at communication round t is always compared with the historical reference constructed from rounds up to $t-1$. Following the empirical observations in Section 2.1, we retain separate output-side and input-side alignment scores to characterize the asymmetric geometry of LoRA updates.

Specifically, the alignment scores are defined as

$$s_{i,l}^{L,t} = \frac{\|U_l^{t-1}(U_l^{t-1})^\top \Delta_{i,l}^t\|_F^2}{\|\Delta_{i,l}^t\|_F^2}, \quad s_{i,l}^{R,t} = \frac{\|\Delta_{i,l}^t V_l^{t-1}(V_l^{t-1})^\top\|_F^2}{\|\Delta_{i,l}^t\|_F^2}, \tag{10}$$

where $s_{i,l}^{L,t}$ and $s_{i,l}^{R,t}$ denote the fractions of update energy lying in the historical consensus output and input subspaces, respectively. Both scores lie in $[0, 1]$ whenever $\Delta_{i,l}^t \neq 0$. If the update norm falls below a small numerical threshold, both scores are set to one because no meaningful directional disagreement can be inferred.

Although both scores are computed throughout training, FedSALT deploys only the output-side score

$$s_{i,l}^t = s_{i,l}^{L,t} \tag{11}$$

as the control signal. This design follows the empirical findings of Section 2.1, which show that task-dependent differentiation is primarily reflected in the output-side geometry. The input-side score is retained for mechanism analysis and ablation. For completeness, we also consider a two-sided variant

$$s_{i,l}^{LR,t} = \sqrt{s_{i,l}^{L,t} s_{i,l}^{R,t}}, \tag{12}$$

which is evaluated experimentally but is not used by the proposed controller.

The deployed score admits a simple geometric interpretation. Let $P_{U,l}^{t-1} = U_l^{t-1}(U_l^{t-1})^\top$, then

$$1 - s_{i,l}^{L,t} = \frac{\|(I - P_{U,l}^{t-1})\Delta_{i,l}^t\|_F^2}{\|\Delta_{i,l}^t\|_F^2} \tag{13}$$

can be used to measure the normalized update energy lying outside the historical consensus subspace. Because the score depends only on the realized update, it is invariant to equivalent LoRA reparameterizations: replacing (B, A) by $(BR, R^{-1}A)$ for any invertible R leaves the score unchanged. The normalization also makes the score invariant to nonzero rescaling of the update.

The alignment scores can be computed without explicitly forming the dense update matrix. Define

$$\Phi_{i,l}^t = \frac{\alpha}{r}\left[B_{i,l}^{t+1}, -B_l^t\right], \qquad \Psi_{i,l}^t = \begin{bmatrix} A_{i,l}^{t+1} \\ A_l^t \end{bmatrix}, \tag{14}$$

so that $\Delta_{i,l}^t = \Phi_{i,l}^t \Psi_{i,l}^t$. The numerator of $s_{i,l}^{L,t}$ depends only on $(U_l^{t-1})^\top \Phi_{i,l}^t$, while its denominator is computed from the small Gram matrices $(\Phi_{i,l}^t)^\top \Phi_{i,l}^t$ and $\Psi_{i,l}^t(\Psi_{i,l}^t)^\top$. The corresponding computation for $s_{i,l}^{R,t}$ uses $\Psi_{i,l}^t V_l^{t-1}$. Consequently, all computations remain within the low-rank LoRA representation.

Subspace energy alone does not distinguish whether an update follows or opposes the historical optimization direction. To resolve this ambiguity, FedSALT augments the alignment score with an orientation gate,

$$\tilde{s}_{i,l}^t = s_{i,l}^t \cdot \mathbb{I}\left(\langle \Delta_{i,l}^t, M_l^{t-1} \rangle_F > 0\right), \tag{15}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. If the historical momentum is zero, we set $\tilde{s}_{i,l}^t = 1$, reducing the first controlled communication round to standard FedAvg behavior.

3.3 Adaptive Client-Module Controller

The alignment score quantifies the output-subspace disagreement at each communication round, but individual observations may fluctuate because of stochastic optimization and client sampling. FedSALT therefore smooths the sensed disagreement over time to obtain a stable client-module regularization strength:

$$\mu_{i,l}^{t+1} = \gamma \mu_{i,l}^t + (1-\gamma)\mu_{\max}\left(1 - \tilde{s}_{i,l}^t\right), \qquad \mu_{i,l}^0 = 0, \tag{16}$$

where $\gamma \in [0, 1)$ controls the temporal smoothing horizon and $\mu_{\max}$ specifies the maximum regularization strength. Persistent disagreement gradually increases $\mu_{i,l}$, whereas sustained alignment causes it to decay toward zero. The resulting controller suppresses transient fluctuations and prevents individual noisy communication rounds from dominating subsequent local optimization.

Unlike FedProx, which applies a fixed proximal coefficient to every client and module, FedSALT allocates regularization according to the observed output-subspace geometry. Client-module pairs exhibiting persistent disagreement receive stronger regularization, while well-aligned updates remain close to standard federated LoRA training. The server maintains the controller states and broadcasts only the corresponding L scalar strengths to each client, introducing negligible communication and storage overhead.

3.4 Momentum-Lookahead Local Objective

Given the adaptive regularization strengths, each client solves a momentum-guided local optimization problem. Specifically, client i minimizes

$$\mathcal{L}_i^t = f_i + \sum_{l=1}^{L} \frac{\mu_{i,l}^t}{2} \left\| \Delta W_{i,l} - \beta \check{M}_l^{t-1} \right\|_F^2, \tag{17}$$

where $f_i = \mathcal{L}_i\left(\{A_{i,l}, B_{i,l}\}_{l=1}^L; \mathcal{D}_i\right)$ is the local task loss and

$$\Delta W_{i,l} = \frac{\alpha}{r}\left(B_{i,l}A_{i,l} - B_l^t A_l^t\right) \tag{18}$$

is the current local product update, with $\beta \geq 0$ controlling the magnitude of the momentum-lookahead reference.

Unlike a conventional zero-centered proximal objective, which discourages all deviations from the broadcast model, Eq. equation 17 encourages drifting client-module updates to remain close to

Algorithm 1 FedSALT at communication round t

1: **Server broadcasts** global factors $\{A_l^t, B_l^t\}$, client-specific strengths $\{\mu_{i,l}^t\}$, and anchor factors $\{\check{F}_l^{t-1}, \check{G}_l^{t-1}\}$
2: **for** each participating client i in parallel **do**
3: Optimize $\mathcal{L}_i^t$ in Eq. (17) for K local steps
4: Return $\{A_{i,l}^{t+1}, B_{i,l}^{t+1}\}_{l=1}^L$
5: **end for**
6: Aggregate factors using Eq. (3)
7: Construct product updates $\Delta_{i,l}^t$ and consensus updates $\bar{\Delta}_l^t$
8: Compute $s_{i,l}^{L,t}$ and $s_{i,l}^{R,t}$ using Eq. (10)
9: Set the deployed score by Eq. (11) and apply Eq. (15)
10: Update $\mu_{i,l}^{t+1}$ using Eq. (16)
11: Update M_l^t, U_l^t, V_l^t, and $\check{M}_l^t$ using Eqs. (6)-(9)

the recent global optimization trajectory represented by $\check{M}_l^{t-1}$. Consequently, globally consistent progress is preserved while client-specific drift is selectively suppressed. Setting $\beta = 0$ recovers an adaptive product-space proximal objective, isolating the contribution of the momentum-lookahead reference in the ablation study.

The proximal term is evaluated entirely in low-rank form. Let $Z_l^t = \beta \check{M}_l^{t-1}$ and write $\check{M}_l^{t-1} = \check{F}_l^{t-1}\check{G}_l^{t-1}$. Then

$$\frac{\mu_{i,l}^t}{2}\|\Delta W_{i,l} - Z_l^t\|_F^2 = \frac{\mu_{i,l}^t}{2}\|\Delta W_{i,l}\|_F^2 - \mu_{i,l}^t\langle \Delta W_{i,l}, Z_l^t\rangle_F + \frac{\mu_{i,l}^t}{2}\|Z_l^t\|_F^2. \tag{19}$$

The last term is constant with respect to the local factors. The cross-term gradient with respect to $B_{i,l}$ is

$$\frac{\partial}{\partial B_{i,l}}\left\langle \frac{\alpha}{r} B_{i,l}A_{i,l}, \check{F}_l^{t-1}\check{G}_l^{t-1}\right\rangle_F = \frac{\alpha}{r}\check{F}_l^{t-1}\left(\check{G}_l^{t-1}A_{i,l}^\top\right), \tag{20}$$

and the derivative with respect to $A_{i,l}$ is obtained analogously. Consequently, the entire proximal objective is evaluated using only low-rank factors without explicitly forming dense $d_{\text{out}} \times d_{\text{in}}$ matrices.

3.5 Overall Algorithm

Algorithm 1 summarizes one communication round of FedSALT. The controller operates with a one-round delay: updates returned at round t are evaluated against the historical reference from round $t-1$, while the resulting controller strengths and momentum reference are used during local optimization at round $t+1$. This prevents current updates from influencing the reference used to evaluate themselves.

FedSALT introduces little additional overhead beyond standard federated LoRA. The uplink remains unchanged because clients transmit only the LoRA factors after local optimization. The additional downlink consists of one shared low-rank momentum reference together with L client-specific controller scalars. Server-side sensing, controller updates, and momentum recompression operate entirely on low-rank factors and small Gram matrices.

4 Theoretical Analysis

We analyze an idealized realized-adapter-space version of FedSALT in order to isolate the effect of adaptive local control from the non-Euclidean LoRA factorization. The analysis assumes exact aggregation of realized adapter updates and therefore does not directly cover the tangent geometry of factor-space SGD or the mismatch induced by factor-wise averaging. Full statements, auxiliary quantities, and proofs are provided in the supplementary material.

Let $\mathcal{H} = \bigoplus_{l=1}^L \mathbb{R}^{d_{\text{out},l}\times d_{\text{in},l}}$ denote the realized-adapter space, and let $F(W) = \sum_i p_i F_i(W)$, where $p_i > 0$ and $\sum_i p_i = 1$.

Assumption 1 (Smooth stochastic objectives) *Each F_i is L_F-smooth and F is bounded below by $F_\star$. The stochastic gradient error satisfies*

$$\mathbb{E}[\xi_i^{t,k} \mid \mathcal{F}_{t,k}] = 0, \qquad \mathbb{E}[\|\xi_i^{t,k}\|^2 \mid \mathcal{F}_{t,k}] \le \sigma^2.$$

Conditional on the round-start history, the client noise sequences are independent across clients.

Assumption 2 (Ideal realized-space training) *All clients participate. Client i starts round t from $U_i^{t,0} = 0$ and performs*

$$U_i^{t,k+1} = U_i^{t,k} - \eta\Big[\nabla F_i(W^t + U_i^{t,k}) + \xi_i^{t,k} + \Lambda_i^t(U_i^{t,k} - Z^t)\Big], \tag{21}$$

where block l of Λ_i^t equals the adaptive strength $\mu_{i,l}^t$ and $Z_l^t = \beta \check{M}_l^{t-1}$ is the lookahead anchor. The server aggregates realized updates exactly as $W^{t+1} = W^t + \sum_i p_i U_i^{t,K}$. The controller states, reference projectors, and anchors used in round t are measurable at the beginning of that round.

We first state the mechanism underlying the controller. Let P_l^{t-1} project onto the historical consensus output subspace and let $Q_l^{t-1} = I - P_l^{t-1}$.

Proposition 1 (Off-subspace response) *Fix a round and let $X_{i,l}^{t,K}$ be the K-step local update obtained by freezing the gradient at $\nabla_l F_i(W^t)$ and removing stochastic noise. For $\tau = \eta K$,*

$$\|Q_l^{t-1} X_{i,l}^{t,K}\|_F^2 = \tau^2 a_K(\mu_{i,l}^t)^2 \|Q_l^{t-1} \nabla_l F_i(W^t)\|_F^2,$$

where $a_K(\mu) = \frac{1}{K}\sum_{h=0}^{K-1}(1-\eta\mu)^h$. When $K \ge 2$ and $\eta\mu < 1$, $a_K(\mu)$ is strictly decreasing. Therefore, stronger control directly contracts the leading off-subspace component of local drift.

The supplementary material shows that the deployed gated score estimates this off-subspace energy up to reference, local-update, and temporal tracking errors. Under the sufficient calibration, dispersion, and response-curvature conditions stated in the supplementary material, adaptive allocation strictly reduces the leading off-subspace response relative to uniform control with the same weighted mean strength. The comparison is made at a common round state with fixed gradients, projector, and anchor; it is not a trajectory-wise dominance claim.

To keep the main statement concise, let $\mathcal{B}_{\mathrm{sel}}$ denote the selection bias caused by nonuniform client-module responses, let $\Omega_\parallel^2 + \Omega_\perp^2$ denote client-gradient heterogeneity, and let $\mathcal{A}_{\mathrm{path}}$ and $\mathcal{E}_{\mathrm{anc}}$ denote the two lookahead-anchor contributions. Their exact definitions are given in the Appendix.

Theorem 1 (Nonconvex convergence of idealized FedSALT) *Under Assumptions 1-2, suppose $K \ge 2$, $0 \le \mu_{i,l}^t \le \mu_{\max}$, and $\eta\mu_{\max} < 1$. Define $c_- = (1-\eta\mu_{\max})^{K-1} > 0$. There exist universal constants $c_{\mathrm{stab}} \le 1/4$ and $C > 0$ such that, if $\tau L_F \le c_{\mathrm{stab}} c_-$, then*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(W^t)\|^2 \le \frac{C(F(W^0) - F_\star)}{c_-\tau T} + \frac{CL_F\eta(\sum_i p_i^2)\sigma^2}{c_-} + \frac{C}{c_-^2}\left(\mathcal{B}_{\mathrm{sel}} + \mathcal{E}_{\mathrm{anc}}\right) + \frac{C\tau^2 L_F^2}{c_-^2}\left(\Omega_\parallel^2 + \Omega_\perp^2 + \mathcal{A}_{\mathrm{path}} + \frac{\sigma^2}{K}\right). \tag{22}$$

The bound separates ordinary optimization and stochastic terms from three FedSALT-specific effects: nonuniform-control selection bias, client heterogeneity, and anchor error. When $\mu_{i,l}^t \equiv 0$, the adaptive-control terms disappear and the result reduces to a standard local-SGD-type guarantee in realized adapter space. When $\beta = 0$, both anchor terms vanish.

5 Experiments

Experimental setup. We evaluate FedSALT in three heterogeneous federated language-model settings. MC-8 contains eight clients drawn from multiple-choice question-answering tasks, including two disjoint MedMCQA partitions that provide a same-task reference. CS8-Dir partitions a common dataset across eight clients using a Dirichlet distribution with concentration parameter

Table 1: Overall accuracy (%)

Method	MC-8	CS8-Dir ($\alpha = 0.1$)	MMLU8
FedSALT (ours)	**58.5 ± 0.1**	**74.5 ± 0.7**	**46.5 ± 0.3**
FedRot-LoRA	58.2 ± 0.0	73.0 ± 0.6	45.6 ± 0.3
FedProx	58.2 ± 0.2	73.6 ± 0.5	45.7 ± 0.5
SCAFFOLD	58.0 ± 0.1	73.7 ± 0.8	45.7 ± 0.2
FedEx-LoRA	57.9 ± 0.1	73.3 ± 1.5	45.8 ± 0.5
FedAvg	57.9 ± 0.1	74.0 ± 0.7	45.8 ± 0.8
FedSA-LoRA	57.2 ± 0.2	72.2 ± 0.4	46.0 ± 0.5
FFA-LoRA	56.8 ± 0.1	71.3 ± 0.6	46.2 ± 0.4

Table 2: Client-level accuracy (%) on MC-8

Client partition	FedSALT (ours)	FedAvg	FedProx	SCAFFOLD
MedMCQA-A	43.0 ± 0.5	43.4 ± 0.6	**43.5 ± 0.5**	43.4 ± 1.3
MedMCQA-B	**43.1 ± 0.4**	41.9 ± 0.4	42.3 ± 0.3	42.1 ± 0.8
CSQA	**71.8 ± 0.5**	**71.8 ± 0.5**	70.9 ± 0.6	71.5 ± 0.1
AQUA	**32.8 ± 2.1**	29.3 ± 0.8	31.0 ± 1.1	31.5 ± 2.0
RACE	**66.8 ± 0.9**	66.1 ± 1.2	66.1 ± 0.9	65.0 ± 1.1
OBQA	69.6 ± 1.5	69.5 ± 1.1	**69.7 ± 0.6**	69.6 ± 0.2
ARC-E	79.1 ± 0.4	79.0 ± 0.7	**79.2 ± 0.1**	79.0 ± 0.3
ARC-C	62.0 ± 1.0	62.3 ± 0.7	**62.7 ± 0.6**	61.6 ± 0.5
Client average	**58.5 ± 0.1**	57.9 ± 0.1	58.2 ± 0.2	58.0 ± 0.1

0.1, while MMLU8 contains eight task-separated clients from MMLU. We fine-tune Llama-3.2-1B using LoRA with rank $r = 16$ and scaling $\alpha = 32$ on the attention and MLP projections. Each communication round uses $K = 40$ local steps with batch size 16, and the optimizer is reset at the beginning of each round. All results are reported over three random seeds. Full experiment settings are provided in Appendix D.

Baselines. We compare with FedAvg (McMahan et al., 2017), FedProx (Li et al., 2020a), and SCAFFOLD (Karimireddy et al., 2020). Federated LoRA baselines include FFA-LoRA (Sun et al., 2024), FedSA-LoRA (Guo et al., 2025), FedEx-LoRA (et al., 2025), and FedRot-LoRA (Zhang et al., 2026). All methods use the same backbone, client partitions, LoRA rank, local training budget, and evaluation protocol.

5.1 Overall Performance

Table 1 shows that FedSALT achieves the highest average accuracy across all three federations. Its advantage over FedProx indicates that proximal control benefits from being allocated according to the observed LoRA geometry rather than applied uniformly. The comparison with FedEx-LoRA and FedRot-LoRA further distinguishes FedSALT from server-side aggregation repair: controlling local trajectories before aggregation addresses a complementary source of error.

Table 2 decomposes the MC-8 result across clients. The improvement is supported by gains on multiple heterogeneous tasks rather than uniform dominance on every partition. FedSALT therefore improves the global client average while remaining competitive across the federation.

5.2 Validating the Closed-Loop Controller

Output-side sensing. Figures 1 and 2 establish the empirical premise of FedSALT. Output-side subspaces exhibit substantially stronger client-dependent variation than input-side subspaces, while the two MedMCQA clients remain comparatively similar. The same asymmetry appears relative to the historical consensus: output-side alignment is lower and more dispersed across clients, providing the contrast needed for selective control. Figure 3 in Appendix D shows that output-side scores retain greater client-level variation than input-side scores. This variation is preserved by the one-round-delayed controller, which assigns stronger control to more strongly disagreeing client-module pairs.

Component ablations. Table 3 in Appendix D.1 isolates the sensing signal, control allocation, temporal smoothing, and correction target. Matched-mean uniform and permuted controls show that the benefit depends on signal-guided client-module allocation rather than the overall regularization strength alone.

Efficiency. Table 4 in Appendix D.4 shows that FedSALT is efficient relative to local drift-control methods, while its communication cost is in the middle of them.

6 Conclusion

We presented FedSALT, a geometry-aware closed-loop optimization framework for heterogeneous federated LoRA. Motivated by the observation that task-dependent drift is concentrated in the output-side geometry of LoRA updates, FedSALT combines output-subspace drift sensing, adaptive client-module regularization, and a momentum-guided local objective while remaining compatible with standard factor-wise aggregation. Experiments across three heterogeneous federated benchmarks demonstrated consistent improvements over existing methods, supporting both the effectiveness of the proposed optimization framework and the underlying geometric insight.

Reproducibility Statement

The complete algorithm is provided in Algorithm 1. Experimental protocols, ablations, system measurements, and dataset details are provided in Appendix D, and complete theoretical proofs are provided in Appendix B.

Ethics Statement

This work uses existing language-model benchmark datasets and follows the ICLR Code of Ethics.

AI Use Statement

Generative AI tools were used to assist with language editing and manuscript organization.

References

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N. Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *ICLR*, 2021. URL https://arxiv.org/abs/2111.04263.

Sara Babakniya, Ahmed Roushdy Elkordy, Yahya H. Ezzeldin, Qingfeng Liu, Kee-Bong Song, Mostafa El-Khamy, and Salman Avestimehr. SLoRA: Federated parameter efficient fine-tuning of language models, 2023. URL https://arxiv.org/abs/2308.06522. arXiv:2308.06522.

Jiamu Bai, Daoyuan Chen, Bingchen Qian, Liuyi Yao, and Yaliang Li. Federated fine-tuning of large language models under heterogeneous tasks and client resources. In *NeurIPS*, 2024. URL https://arxiv.org/abs/2402.11505.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Short Papers*, pp. 1–9, 2022. URL https://aclanthology.org/2022.acl-short.1/.

Jieming Bian, Lei Wang, Letian Zhang, and Jie Xu. FedALT: Federated fine-tuning through adaptive local training with rest-of-world LoRA. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026a. URL https://ojs.aaai.org/index.php/AAAI/article/view/39054.

Jieming Bian, Lei Wang, Letian Zhang, and Jie Xu. FedTreeLoRA: Reconciling statistical and functional heterogeneity in federated LoRA fine-tuning. In *ICML*, 2026b. URL https://arxiv.org/abs/2603.13282.

Shuangyi Chen, Yuanxin Guo, Yue Ju, Harik Dalal, and Ashish Khisti. Robust federated finetuning of LLMs via alternating optimization of LoRA. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2502.01755.

Yae Jee Cho, Luyang Liu, Zheng Xu, Aldi Fahrezi, and Gauri Joshi. Heterogeneous LoRA for federated fine-tuning of on-device foundation models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024. URL https://aclanthology.org/2024.emnlp-main.717/.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 2089–2099, 2021. URL https://proceedings.mlr.press/v139/collins21a.html.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract-Conference.html.

Canh T. Dinh, Nguyen H. Tran, and Tuan Dung Nguyen. Personalized federated learning with moreau envelopes. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f4f1f13c8289ac1b1ee0ff176b56fc60-Abstract.html.

Singhal et al. FedEx-LoRA: Exact aggregation for federated and efficient fine-tuning of large language models. In *ACL*, pp. 1316–1336, 2025. doi: 10.18653/v1/2025.acl-long.67. URL https://aclanthology.org/2025.acl-long.67/.

Wenzhi Fang, Dong-Jun Han, Liangqi Yuan, Seyyedali Hosseinalipour, and Christopher G. Brinton. Federated sketching LoRA: A flexible framework for heterogeneous collaborative fine-tuning of LLMs. In *ICML*, 2026. URL https://arxiv.org/abs/2501.19389.

Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/e32cc80bf07915058ce90722ee17bb71-Abstract.html.

Pengxin Guo, Shuang Zeng, Yanran Wang, Huijie Fan, Feifei Wang, and Liangqiong Qu. Selective aggregation for low-rank adaptation in federated learning. In *ICLR*, 2025. URL https://openreview.net/forum?id=iX3uESGdsO.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2790–2799, 2019.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification, 2019. URL https://arxiv.org/abs/1909.06335. arXiv preprint.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. URL https://arxiv.org/abs/2106.09685.

Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. Personalized cross-silo federated learning on non-iid data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Divyansh Jhunjhunwala, Shiqiang Wang, and Gauri Joshi. FedExP: Speeding up federated averaging via extrapolation. In *ICLR*, 2023. URL https://arxiv.org/abs/2301.09604.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *ICML*, 2020. URL https://arxiv.org/abs/1910.06378.

Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. Breaking the centralized barrier for cross-device federated learning. In *NeurIPS*, 2021. URL https://arxiv.org/abs/2008.03606.

Geeho Kim, Jinkyu Kim, and Bohyung Han. Communication-efficient federated learning with accelerated client gradient. In *CVPR*, pp. 12385–12394, 2024. doi: 10.1109/CVPR52733.2024.01177. URL https://openaccess.thecvf.com/content/CVPR2024/html/Kim_Communication-Efficient_Federated_Learning_with_Accelerated_Client_Gradient_CVPR_2024_paper.html.

Dawid J. Kopiczko, Tijmen Blankevoort, and Yuki M. Asano. VeRA: Vector-based random matrix adaptation. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.11454.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, 2021. URL https://aclanthology.org/2021.emnlp-main.243/.

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *CVPR*, 2021a. URL https://arxiv.org/abs/2103.16257.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems (MLSys)*, 2020a. URL https://arxiv.org/abs/1812.06127.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *ICLR*, 2020b. URL https://arxiv.org/abs/1905.10497.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 6357–6368, 2021b. URL https://proceedings.mlr.press/v139/li21h.html.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pp. 4582–4597, 2021. URL https://aclanthology.org/2021.acl-long.353/.

Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. FedBN: Federated learning on non-iid features via local batch normalization. In *International Conference on Learning Representations*, 2021c. URL https://openreview.net/forum?id=6YEQUn0QICG.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 32100–32121, 2024. URL https://proceedings.mlr.press/v235/liu24bn.html.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, pp. 1273–1282, 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *ICML*, 2019. URL https://arxiv.org/abs/1902.00146.

Jaehoon Oh, SangMook Kim, and Se-Young Yun. FedBABU: Toward enhanced representation for federated image classification. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=HuaYQfggn5u.

Krishna Pillutla, Kshitiz Malik, Abdel-Rahman Mohamed, Mike Rabbat, Maziar Sanjabi, and Lin Xiao. Federated learning with partial model personalization. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 17716–17758, 2022. URL https://proceedings.mlr.press/v162/pillutla22a.html.

Zhe Qu, Xingyu Li, Rui Duan, Yao Liu, Bo Tang, and Zhuo Lu. Generalized federated learning via sharpness aware minimization. In *ICML*, 2022. URL https://arxiv.org/abs/2206.02618.

Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and H. Brendan McMahan. Adaptive federated optimization. In *ICLR*, 2021. URL https://arxiv.org/abs/2003.00295.

Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. Improving LoRA in privacy-preserving federated learning. In *ICLR*, 2024. URL https://arxiv.org/abs/2403.12313.

Saeed Vahidian, Mahdi Morafah, Weijia Wang, Vyacheslav Kungurtsev, Chen Chen, Mubarak Shah, and Bill Lin. Efficient distribution similarity identification in clustered federated learning via principal angles between client data subspaces. In *AAAI*, volume 37, pp. 10043–10052, 2023. doi: 10.1609/aaai.v37i8.26197. URL https://ojs.aaai.org/index.php/AAAI/article/view/26197.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *NeurIPS*, 2020. URL https://arxiv.org/abs/2007.07481.

Lei Wang, Jieming Bian, Letian Zhang, and Jie Xu. Adaptive LoRA experts allocation and selection for federated fine-tuning. In *Advances in Neural Information Processing Systems*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/6df1b2b45e64d402588746f79b68b82c-Abstract-Conference.html.

Ziyao Wang, Zheyu Shen, Yexiao He, Guoheng Sun, Hongyi Wang, Lingjuan Lyu, and Ang Li. FLoRA: Federated fine-tuning large language models with heterogeneous low-rank adaptations. In *NeurIPS*, 2024. URL https://arxiv.org/abs/2409.05976.

Feijie Wu, Zitao Li, Yaliang Li, Bolin Ding, and Jing Gao. FedBiOT: LLM local fine-tuning in federated learning without full model. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3345–3355, 2024. URL https://arxiv.org/abs/2406.17706.

Jing Xu, Sen Wang, Liwei Wang, and Andrew Chi-Chih Yao. FedCM: Federated learning with client-level momentum, 2021. URL https://arxiv.org/abs/2106.10874. arXiv preprint.

Yuxuan Yan, Qianqian Yang, Shunpu Tang, and Zhiguo Shi. FeDeRA: Efficient fine-tuning of language models in federated learning leveraging weight decomposition, 2024. URL https://arxiv.org/abs/2404.18848. arXiv:2404.18848.

Haoran Zhang, Dongjun Kim, Seohyeon Cha, and Haris Vikalo. FedRot-LoRA: Mitigating rotational misalignment in federated LoRA. In *ICML*, 2026. URL https://arxiv.org/abs/2602.23638.

Jianyi Zhang, Saeed Vahidian, Martin Kuo, Chunyuan Li, Ruiyi Zhang, Tong Yu, Guoyin Wang, and Yiran Chen. Towards building the federated GPT: Federated instruction tuning. In *ICASSP*, pp. 6915–6919, 2024. URL https://arxiv.org/abs/2305.05644.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=lq62uWRJjiY.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. GaLore: Memory-efficient LLM training by gradient low-rank projection. In *ICML*, 2024. URL https://arxiv.org/abs/2403.03507.

Jiacheng Zhu, Kristjan Greenewald, Kimia Nadjahi, Haitz Sáez De Ocáriz Borde, Rickard Brüel Gabrielsson, Leshem Choshen, Marzyeh Ghassemi, Mikhail Yurochkin, and Justin Solomon. Asymmetry in low-rank adapters of foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 62369–62385, 2024. URL https://proceedings.mlr.press/v235/zhu24c.html.

A Related Work

A.1 Federated Optimization under Heterogeneity

FedAvg learns a single global model by alternating local optimization and server averaging (McMahan et al., 2017). Under statistically heterogeneous client data, multiple lines of work modify either the local objective or the server update to reduce optimization drift. FedProx introduces a uniform proximal penalty, SCAFFOLD uses control variates, FedDyn applies dynamic regularization, and FedNova normalizes heterogeneous local updates (Li et al., 2020a; Karimireddy et al., 2020; Acar et al., 2021; Wang et al., 2020). Other methods employ adaptive server optimization, client or server momentum, extrapolation, transferred optimizer states, contrastive objectives, or sharpness-aware local training (Reddi et al., 2021; Hsu et al., 2019; Xu et al., 2021; Jhunjhunwala et al., 2023; Karimireddy et al., 2021; Kim et al., 2024; Li et al., 2021a; Qu et al., 2022). These methods provide general-purpose corrections for heterogeneous federated optimization. FedSALT instead exploits the realized geometry of LoRA updates to determine both where control is required and how strongly each client-module pair should be regulated.

A.2 Personalized and Clustered Federated Learning

Another response to client heterogeneity is to relax the requirement of learning a single shared model. Agnostic and fairness-aware formulations modify the global objective to improve robustness across client distributions (Mohri et al., 2019; Li et al., 2020b). Personalized federated learning methods instead regularize local models toward shared parameters, as in pFedMe and Ditto, or separate shared representations from client-specific components, as in FedRep (Dinh et al., 2020; Li et al., 2021b; Collins et al., 2021). Related approaches personalize normalization layers, classifiers, or subsets of model parameters (Li et al., 2021c; Oh et al., 2022; Pillutla et al., 2022). Clustered methods such as IFCA, FedAMP, and PACFL learn multiple models or induce client groups according to their statistical or representation-level similarity (Ghosh et al., 2020; Huang et al., 2021; Vahidian et al., 2023). FedSALT addresses a different setting where one globally shared adapter is retained and client differences are treated as a continuous signal for optimization control rather than producing personalized or cluster-specific models.

A.3 Parameter-Efficient and Low-Rank Adaptation

Parameter-efficient fine-tuning (PEFT) adapts a frozen pretrained model using a small set of trainable parameters. Representative approaches include adapters, prefix tuning, prompt tuning, and bias-only updates (Houlsby et al., 2019; Li & Liang, 2021; Lester et al., 2021; Ben Zaken et al., 2022). LoRA parameterizes each weight update as a low-rank product and has inspired extensions that allocate rank adaptively, quantize the frozen backbone, share random projections, or decouple weight magnitude and direction (Hu et al., 2022; Zhang et al., 2023; Dettmers et al., 2023; Kopiczko et al., 2024; Liu et al., 2024). Recent analyses also study the optimization dynamics of low-rank adaptation and the asymmetric roles of its two factors (Zhu et al., 2024; Zhao et al., 2024). FedSALT builds on this structure but operates on the realized product update rather than either factor individually. This yields a sensing signal that is invariant to equivalent LoRA factorizations and can be evaluated without materializing the dense weight update.

A.4 Federated Fine-Tuning of Language and Foundation Models

Federated PEFT combines local adapter optimization with communication-efficient server aggregation. Early systems such as FedIT and SLoRA establish federated instruction tuning and federated LoRA training for language models (Zhang et al., 2024; Babakniya et al., 2023). Subsequent work considers resource constraints, initialization, heterogeneous adapter ranks, limited model access, and communication-efficient low-rank representations through FFA-LoRA, FeDeRA, FedBiOT, HetLoRA, FlexLoRA, and Federated Sketching LoRA (Sun et al., 2024; Yan et al., 2024; Wu et al., 2024; Cho et al., 2024; Bai et al., 2024; Fang et al., 2026).

A separate line changes which adapter components are shared across clients. FedSA-LoRA shares the input-side factor, while FedALT, FedLEASE, and FedTreeLoRA introduce personalized components, expert structures, or hierarchical sharing (Guo et al., 2025; Bian et al., 2026a; Wang et al.,

2025; Bian et al., 2026b). RoLoRA instead changes the optimization schedule by alternating updates to the two LoRA factors (Chen et al., 2025). These methods primarily modify adapter capacity, sharing structure, personalization, or factor optimization. FedSALT retains a single globally shared adapter with both factors trainable and regulates its local trajectory using online client-module disagreement.

A.5 Aggregation in Federated LoRA

Factor-wise averaging of LoRA parameters does not generally equal averaging their realized product updates. FLoRA removes this mismatch through factor stacking, FedEx-LoRA transfers residual aggregation error into the frozen weight, and FedRot-LoRA aligns equivalent factor representations before averaging (Wang et al., 2024; et al., 2025; Zhang et al., 2026). These methods improve how completed local updates are combined at the server.

FedSALT instead changes the local trajectories that generate those updates. Output-subspace disagreement observed after round t determines client- and module-specific control in round $t+1$, while server aggregation remains standard factor-wise FedAvg. FedSALT is therefore complementary to aggregation-side correction.

B Proofs of the Theoretical Results

B.1 Scope, notation, and response functions

Let $W \in \mathcal{H}$ denote the collection of realized LoRA adapter matrices. We equip $\mathcal{H}$ with the block-wise Frobenius inner product and write $\|\cdot\|$ for its induced norm. Throughout this section, Assumptions 1 and 2 and the local recursion (21) are in force. The analysis is formulated directly in realized-adapter space. Relating this Euclidean recursion to unconstrained SGD on factors $W = BA$ would additionally require controlling the state-dependent factor-space tangent metric. The exact realized-space consensus update used below corresponds to the product-space consensus displacement $\bar{\Delta}_l^t$ in Eq. equation 5.

Fix a round t when no confusion is possible. Write

$$G_i = \nabla F_i(W^t), \qquad P_l = P_l^{t-1},$$
$$Q_l = I - P_l, \qquad Z_l = Z_l^t.$$

The block-diagonal operator $\Lambda_i = \Lambda_i^t$ has block value $\mu_{i,l}^t$. Since $0 \le \eta\mu_{i,l}^t < 1$,

$$S_i := I - \eta\Lambda_i \quad \text{satisfies} \quad \|S_i\| \le 1. \tag{23}$$

All conditional expectations below are with respect to the round-start sigma-field $\mathcal{F}_{t,0}$ unless a finer filtration is displayed.

Assumption 3 (Stable local regime) *We assume* $K \ge 2$, $\eta > 0$, $\mu_{\max} > 0$, $\eta\mu_{\max} < 1$, $\tau = \eta K$, $\tau L_F \le \frac{1}{4}$. *These conditions ensure that the local proximal recursion remains stable and that the accumulated smoothness error within each communication round is controlled.*

For $k \ge 1$ and $0 \le \mu \le \mu_{\max}$, define

$$a_k(\mu) = \frac{1}{k}\sum_{h=0}^{k-1}(1-\eta\mu)^h, \qquad \kappa_k(\mu) = 1 - (1-\eta\mu)^k. \tag{24}$$

Set $a_0 = \kappa_0 = 0$. We use two responses:

$$\phi_K(\mu) = a_K(\mu)^2, \qquad \psi_K(\mu) = \frac{1}{K}\sum_{k=0}^{K-1}\left(\frac{k}{K}\right)^2 a_k(\mu)^2, \tag{25}$$

and the anchor-path coefficient

$$\chi_K(\mu) = \frac{1}{K}\sum_{k=0}^{K-1}\kappa_k(\mu)^2. \tag{26}$$

Here ϕ_K is the final-step off-reference energy response, while ψ_K is the average local-path response that enters the nonconvex proof. Both are nonincreasing and twice continuously differentiable on $[0, \mu_{\max}]$. The response ϕ_K is strictly decreasing for $K \geq 2$; $\psi_2 \equiv 1/8$, while ψ_K is strictly decreasing for $K \geq 3$.

B.2 Proof roadmap and dependencies

The theoretical development consists of two complementary layers. The first layer establishes the main optimization guarantees: the off-subspace response identity and the nonconvex stationarity bound. The second layer develops the geometric, statistical, and dynamical results needed to justify the score, the reference process, and the adaptive allocation rule. The dependencies are summarized below.

Main optimization results. The proof of Proposition 1 begins with the closed-form frozen-gradient recursion established in Lemma 7. Projecting this recursion onto the orthogonal complement of the historical reference subspace removes the anchor component and isolates the off-subspace response. The monotonicity of a_K then shows how increasing the proximal strength contracts this component.

The proof of Theorem 1 proceeds at the level of one communication round. Lemma 14 first expresses the aggregated frozen-gradient update as a preconditioned global-gradient step together with the selection and anchor terms. Lemma 15 controls the discrepancy between the frozen-gradient recursion and the actual nonlinear stochastic recursion, while Lemma 16 converts this decomposition into a one-round descent inequality. Summing over communication rounds yields a response-sensitive stationarity bound. The remaining path-dependent term is controlled through the weighted variance decomposition across clients, which closes the bound in Theorem 1.

Corollary 1 applies the predictive form of the round decomposition and expresses the anchor contribution in terms of its prediction error. Corollary 2 specializes the general result to the cases of zero proximal control, a zero lookahead anchor, and deterministic local gradients.

Supporting analytical results. The score-related results establish the geometric interpretation of the signal used by the controller. Proposition 2 identifies the alignment score with normalized subspace energy, and Lemma 1 relates this quantity to directional alignment with the reference update. Proposition 3 establishes factorization and scaling invariance, while Lemma 2 gives the corresponding rank-one specialization.

The reference analysis quantifies how accurately the historical momentum tracks the current consensus geometry. Lemmas 3– 5 control the momentum-tracking error, the resulting projector perturbation, and the induced perturbation of the normalized score. These estimates lead to Theorem 2 and the probabilistic guarantee in Corollary 3. The score error is subsequently propagated through the EMA controller recursion in Lemma 6.

The local-response analysis separates the exact effect of proximal control from the errors introduced by nonlinear and stochastic local optimization. Lemma 7 gives the exact frozen-gradient solution, Lemma 8 controls the deviation along the local trajectory, and Lemma 9 bounds the resulting remainder. Together, these results yield Theorem 3.

The comparison between adaptive and uniform allocation relies on the response properties established in Lemma 10 and the covariance relations in Lemmas 11 and 12. These results provide the conditions used in Theorem 4, while Remark 2 clarifies the consistency requirement between the controller and the score from which its strengths are constructed.

Finally, Theorem 5 characterizes the representation bias induced by heterogeneous proximal strengths, and Corollary 4 derives a calibration rule for the associated surrogate objective. Proposition 4 describes the sensitivity to the lookahead anchor, and Lemma 13 analyzes the auxiliary confidence statistic under signal and noise.

B.3 Proof of Proposition 1

Set $k = K$ in equation 52. Because $Z_l^t = \beta \check{M}_l^{t-1}$ lies in the reference output subspace, $Q_l^{t-1} Z_l^t = 0$. Hence

$$Q_l^{t-1} X_{i,l}^{t,K} = -\tau a_K(\mu_{i,l}^t) Q_l^{t-1} \nabla_l F_i(W^t),$$

and squaring gives the displayed identity in Proposition 1. Moreover,

$$a_K'(\mu) = -\frac{\eta}{K} \sum_{h=1}^{K-1} h(1-\eta\mu)^{h-1} < 0$$

for $K \geq 2$ and $0 \leq \eta\mu < 1$. Thus a_K is strictly decreasing. □

B.4 Proof of Theorem 1

For block l, write

$$a_{i,l}^t = a_K(\mu_{i,l}^t),\ \kappa_{i,l}^t = \kappa_K(\mu_{i,l}^t),\ \bar{a}_l^t = \sum_i p_i a_{i,l}^t,\ \bar{\kappa}_l^t = \sum_i p_i \kappa_{i,l}^t.$$

Define the client-gradient deviation

$$\delta G_{i,l}^t := \nabla_l F_i(W^t) - \nabla_l F(W^t).$$

The auxiliary quantities are

$$\mathcal{B}_{\text{sel}} = \frac{1}{T} \sum_{t,l} \mathbb{E}\left[\left(\sum_i p_i (a_{i,l}^t - \bar{a}_l^t)^2 \right) \left(\sum_i p_i \|\delta G_{i,l}^t\|_F^2 \right) \right], \tag{27}$$

$$\mathcal{H}_{\text{path}} = \frac{1}{T} \sum_{t,i,l} p_i \mathbb{E}\left[\psi_K(\mu_{i,l}^t) \|Q_l^{t-1} \nabla_l F_i(W^t)\|_F^2 \right], \tag{28}$$

$$\Omega_{\|}^2 = \frac{1}{T} \sum_{t,i,l} p_i \mathbb{E}\left[\|P_l^{t-1} \delta G_{i,l}^t\|_F^2 \right], \tag{29}$$

$$\Omega_{\perp}^2 = \frac{1}{T} \sum_{t,i,l} p_i \mathbb{E}\left[\|Q_l^{t-1} \delta G_{i,l}^t\|_F^2 \right], \tag{30}$$

$$\mathcal{A}_{\text{path}} = \frac{1}{T} \sum_{t,i,l} p_i \mathbb{E}\left[\chi_K(\mu_{i,l}^t) \frac{\|Z_l^t\|_F^2}{\tau^2} \right], \tag{31}$$

$$\mathcal{E}_{\text{anc}} = \frac{1}{T\tau^2} \sum_{t,l} \mathbb{E}\left[(\bar{\kappa}_l^t)^2 \|Z_l^t\|_F^2 \right]. \tag{32}$$

The anchor quantity is deliberately defined using The definition of $\mathcal{E}_{\text{anc}}$ uses Z_l^t rather than $Z_l^t + \tau \nabla_l F(W^t)$. Consequently, it introduces no additional stationarity-dependent term and vanishes when $\beta = 0$. The response-sensitive quantity $\mathcal{H}_{\text{path}}$ is retained to expose the effect of adaptive control; it is eliminated from the final closed bound below by an explicit absorption argument.

The proof of Theorem 1 proceeds through the response-sensitive estimate

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(W^t)\|^2 \leq \frac{C(F(W^0) - F_\star)}{c_- \tau T} + \frac{C L_F \eta \chi_p \sigma^2}{c_-} + \frac{C}{c_-^2} (\mathcal{B}_{\text{sel}} + \mathcal{E}_{\text{anc}}) \\ + \frac{C\tau^2 L_F^2}{c_-^2} \left(\mathcal{H}_{\text{path}} + \Omega_{\|}^2 + \mathcal{A}_{\text{path}} + \frac{\sigma^2}{K} \right), \end{aligned} \tag{33}$$

and the path closure

$$\begin{aligned} \mathcal{H}_{\text{path}} \leq \psi_0 \left(\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(W^t)\|^2 + \Omega_{\perp}^2 \right), \\ \psi_0 = \frac{(K-1)(2K-1)}{6K^2} < \frac{1}{3}. \end{aligned} \tag{34}$$

Here $\chi_p = \sum_i p_i^2$. The stability condition in Theorem 1, with a sufficiently small universal $c_{\mathrm{stab}} \leq 1/4$, implies Assumption 3.

Let $G_{i,l}^t = \nabla_l F_i(W^t)$ and $G_l^t = \sum_i p_i G_{i,l}^t$. Define

$$b_l^t = \sum_i p_i (a_{i,l}^t - \bar{a}_l^t)(G_{i,l}^t - G_l^t), \tag{35}$$

$$e_{\mathrm{anc},l}^t = \bar{\kappa}_l^t Z_l^t, \tag{36}$$

$$e_{\mathrm{pred},l}^t = \bar{\kappa}_l^t (Z_l^t + \tau G_l^t). \tag{37}$$

Apply Lemma 16 to equation 81 with $\mathcal{F} = \mathcal{F}_{t,0}$ and $e = e_{\mathrm{anc},t} + e_{\mathrm{loc},t}$. Use

$$\|e_{\mathrm{anc},t} + e_{\mathrm{loc},t}\|^2 \leq 2\|e_{\mathrm{anc},t}\|^2 + 2\|e_{\mathrm{loc},t}\|^2,$$

Lemma 14, Lemma 15, and equation 82. The selection-vector contribution is bounded by equation 79. The part of equation 83 proportional to $\tau^2 L_F^2 \|\nabla F(W^t)\|^2$ enters the one-round descent inequality with coefficient of order $\tau^3 L_F^2 / c_-$. It is absorbed into the principal $c_- \tau$ descent term whenever $\tau L_F \leq c_{\mathrm{stab}} c_-$ for a sufficiently small universal c_{stab}.

Sum over t, use $F(W^T) \geq F_\star$, take total expectation, and divide by a constant multiple of $c_- \tau T$. The definitions equation 27, equation 28, equation 29, equation 31, and equation 32 give the response-sensitive bound equation 33, after enlarging a universal constant C if necessary.

It remains to close the path term. Since ψ_K is nonincreasing, $0 \leq \psi_K(\mu_{i,l}^t) \leq \psi_0$. For every t, l, the weighted variance identity gives

$$\sum_i p_i \|Q_l^{t-1} G_{i,l}^t\|_F^2 = \|Q_l^{t-1} G_l^t\|_F^2 + \sum_i p_i \|Q_l^{t-1}(G_{i,l}^t - G_l^t)\|_F^2.$$

Therefore

$$\mathcal{H}_{\mathrm{path}} \leq \psi_0 \left(\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(W^t)\|^2 + \Omega_\perp^2 \right),$$

which is equation 34. Substitution into equation 33 creates a multiple of the averaged gradient norm with coefficient at most $C\psi_0(\tau L_F)^2 / c_-^2$. Shrink the same universal c_{stab}, if necessary, so this coefficient is at most $1/2$, move it to the left, and enlarge C once more. This proves equation 22. Its right-hand side contains only the initial gap, stochastic variance, gradient-dispersion quantities, selection bias, and observable anchor terms, so the bound is closed. □

B.5 Corollaries and Proofs

Here, we present two corollaries along with their proofs. Corollary 1 applies the predictive form of the round decomposition and expresses the anchor contribution in terms of its prediction error. Corollary 2 specializes the general result to the cases of zero proximal control, a zero lookahead anchor, and deterministic local gradients.

Corollary 1 (Prediction refinement) *Define*

$$\mathcal{E}_{\mathrm{pred}} = \frac{1}{T\tau^2} \sum_{t,l} \mathbb{E}\left[(\bar{\kappa}_l^t)^2 \|Z_l^t + \tau \nabla_l F(W^t)\|_F^2 \right]. \tag{38}$$

The alternate exact round decomposition in the appendix gives equation 33 with $\mathcal{E}_{\mathrm{anc}}$ replaced by $\mathcal{E}_{\mathrm{pred}}$. Applying the same path-term absorption used in Theorem 1 first yields the closed non-anchor terms of equation 22, plus $C\mathcal{E}_{\mathrm{pred}}/c_-^2$. In particular, if for that universal constant C

$$\mathcal{E}_{\mathrm{pred}} \leq \frac{c_-^2}{2C} \left(\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(W^t)\|^2 \right) + \varepsilon_{\mathrm{pred}},$$

then the stationarity bound is closed after absorption, with the non-anchor terms multiplied by at most two and with an additional term $2C\varepsilon_{\mathrm{pred}}/c_-^2$. This corollary formalizes the benefit of a predictive anchor only under an explicit prediction-quality condition.

Proof of Corollary 1. Use the alternate exact decomposition equation 78. Its block preconditioner is $\bar{a}_l^t + \bar{\kappa}_l^t \in [c_-, 2]$, and its anchor remainder is $e_{\mathrm{pred},l}^t$. Repeating the response-sensitive part of the proof of Theorem 1, and then applying equation 34 with the same stability condition, gives

$$\overline{G}_T \le B_T + \frac{C}{c_-^2}\mathcal{E}_{\mathrm{pred}}, \qquad \overline{G}_T = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(W^t)\|^2,$$

where B_T is the right-hand side of the closed bound equation 22 with the anchor term removed. Under the stated prediction-quality condition,

$$\overline{G}_T \le B_T + \frac{1}{2}\overline{G}_T + \frac{C}{c_-^2}\varepsilon_{\mathrm{pred}}.$$

Move the middle term to the left to obtain the claimed closed bound. □

Corollary 2 (Degenerate cases) *If $\mu \equiv 0$, then $a_K \equiv 1$, $\kappa_K \equiv 0$, $\mathcal{B}_{\mathrm{sel}} = \mathcal{A}_{\mathrm{path}} = \mathcal{E}_{\mathrm{anc}} = 0$, and Theorem 1 reduces to a standard local-SGD-style stationarity bound in realized adapter space. If $\beta = 0$, then $Z^t = 0$ and both $\mathcal{A}_{\mathrm{path}}$ and $\mathcal{E}_{\mathrm{anc}}$ vanish, giving score-adaptive proximal local training toward zero displacement. If $\sigma = 0$, the stochastic terms vanish.*

Proof of Corollary 2. If $\mu_{i,l}^t \equiv 0$, then $a_K(0) = 1$ and $\kappa_K(0) = 0$ for every client and block. Therefore all strengths are identical, so $\mathcal{B}_{\mathrm{sel}} = 0$, and both anchor quantities vanish. If $\beta = 0$, then $Z_l^t = 0$, which directly gives $\mathcal{A}_{\mathrm{path}} = \mathcal{E}_{\mathrm{anc}} = 0$. If $\sigma = 0$, every term originating from stochastic-gradient noise is zero. Substitution into Theorem 1 proves the claims. □

Remark 1 (Scope of the theoretical guarantees) *The local responses ϕ_K and ψ_K, the controller comparison, and the stationarity bound are established for the ideal realized-space recursion equation 21 with exact realized-update aggregation. The resulting guarantee is therefore formulated in the Euclidean adapter space; an extension to ordinary LoRA factor SGD would require additional control of the state-dependent tangent metric and the factor-averaging mismatch. The reference analysis separates the optional consensus-estimation discrepancy E_l^t from the recompression residual $R_{c,l}^t$, with $E_l^t = 0$ in the exact ideal model. The quantities that may offset the reduction of off-reference path drift appear explicitly in the bounds: controller calibration and historical gate failures in Lemma 6, selection-induced reweighting $\mathcal{B}_{\mathrm{sel}}$, orthogonal and parallel gradient dispersion $\Omega_\perp^2 + \Omega_\parallel^2$, and the anchor terms. The predictive refinement additionally requires the quantitative prediction-quality condition in Corollary 1.*

C Supporting Lemmas and Auxiliary Results

This section presents the supporting results that relate FedSALT's sensing, temporal smoothing, adaptive allocation, and lookahead control to the main optimization framework. The subsections follow the causal pathway from score geometry and reference tracking to controller dynamics, local responses, adaptive allocation, and round-level descent. Additionally, they include auxiliary representations and anchor results.

C.1 Exact score geometry

This subsection establishes the exact geometric properties of the alignment scores in Eq. equation 10: their identification with normalized subspace energy, their one-sided soundness relation to directional alignment with the reference update, and the factorization and scaling invariances used in the main text.

Proposition 2 (Subspace-energy identities) *For orthogonal projectors $P_U = UU^\top$, $P_V = VV^\top$, and every nonzero matrix Δ,*

$$1 - s^L = \frac{\|(I - P_U)\Delta\|_F^2}{\|\Delta\|_F^2},$$
$$1 - s^R = \frac{\|\Delta(I - P_V)\|_F^2}{\|\Delta\|_F^2}.$$

For $s = \sqrt{s^L s^R}$,

$$1 - s \geq \frac{(1 - s^L) + (1 - s^R)}{2}.$$

Proof of Proposition 2. Since $P_U\Delta$ and $(I - P_U)\Delta$ are Frobenius-orthogonal,

$$\|\Delta\|_F^2 = \|P_U\Delta\|_F^2 + \|(I - P_U)\Delta\|_F^2.$$

Division by $\|\Delta\|_F^2$ gives the left identity. The right identity is the analogous orthogonal decomposition on the right. Finally, $\sqrt{s^L s^R} \leq (s^L + s^R)/2$ by AM–GM, which is equivalent to the stated combined inequality. □

Lemma 1 (Score soundness) *If* $\check{M} = P_U \check{M} P_V$, *then*

$$\cos_F^2(\Delta, \check{M}) \leq \min\{s^L, s^R\} \leq \sqrt{s^L s^R},$$
$$\cos_F^2(\Delta, \check{M}) \leq s^L.$$

Thus a low deployed output-side score certifies directional misalignment. A high output-side score does not rule out right-side drift.

Proof. Using $\check{M} = P_U \check{M}$ and Cauchy–Schwarz,

$$|\langle \Delta, \check{M} \rangle_F| = |\langle P_U\Delta, \check{M} \rangle_F| \leq \|P_U\Delta\|_F \|\check{M}\|_F.$$

After normalization and squaring, $\cos_F^2(\Delta, \check{M}) \leq s^L$. The identity $\check{M} = \check{M} P_V$ gives the right-sided inequality. Since $0 \leq s^L, s^R \leq 1$, their minimum does not exceed their geometric mean. □

Proposition 3 (Gram identities and invariance) *If* $X = \Phi\Psi$, *then*

$$\|X\|_F^2 = \mathrm{tr}((\Phi^\top \Phi)(\Psi\Psi^\top)),$$
$$\|PX\|_F^2 = \mathrm{tr}((\Phi^\top P \Phi)(\Psi\Psi^\top)),$$
$$\|XQ\|_F^2 = \mathrm{tr}((\Phi^\top \Phi)(\Psi Q \Psi^\top)).$$

Consequently, the score is invariant under nonzero scaling and under every invertible gauge transform $(\Phi, \Psi) \mapsto (\Phi R, R^{-1}\Psi)$.

Proof. All identities follow from cyclic invariance of the trace. Scaling multiplies numerator and denominator by the same squared factor, while a gauge transform leaves $\Phi\Psi$ unchanged. □

Lemma 2 (Rank-one recovery) *For nonzero rank-one matrices* $M = uv^\top$ *and* $X = xy^\top$, $\sqrt{s^L s^R} = |\cos_F(X, M)|$. *After the orientation gate, the two-sided score equals* $[\cos_F(X, M)]_+$.

Proof. Direct calculation gives $s^L = \cos^2(u, x)$, $s^R = \cos^2(v, y)$, and $\cos_F(X, M) = \cos(u, x)\cos(v, y)$. □

C.2 Momentum and score consistency

This subsection quantifies how accurately the historical momentum and its induced projector track a slowly varying consensus reference. The tracking and perturbation bounds culminate in Theorem 2, which controls the deviation of the deployed gated score from the frozen off-reference energy.

Assumption 4 (Reference process) *For each module, define the exact realized consensus displacement*

$$C_l^t := \sum_i p_i U_{i,l}^{t,K}.$$

For robustness, let the server-side quantity entering the momentum be $\widehat{C}_l^t = C_l^t + E_l^t$, *where* $\|E_l^t\|_2 \leq \delta_e$. *In the exact realized-space model of Assumption 2,* $E_l^t = 0$; *the extra term is retained*

only to separate optional consensus-estimation error from recompression error. Let $M_l^{-1} = 0$ and, for $t \geq 0$,

$$\widetilde{M}_l^t = \rho M_l^{t-1} + (1-\rho)\widehat{C}_l^t, \qquad M_l^t = \Pi_{r_M}(\widetilde{M}_l^t).$$

and define the recompression residual $R_{c,l}^t := \widetilde{M}_l^t - M_l^t$, with $\|R_{c,l}^t\|_2 \leq \delta_c$. Equivalently,

$$M_l^t = \rho M_l^{t-1} + (1-\rho)C_l^t + D_l^t, \qquad D_l^t = (1-\rho)E_l^t - R_{c,l}^t, \tag{39}$$

so $\|D_l^t\|_2 \leq \delta_M := (1-\rho)\delta_e + \delta_c$.

Fix a best rank-r_{ref} approximation $C_{l,r}^t = \Pi_{r_{\mathrm{ref}}}(C_l^t)$. Its smallest retained singular value is at least $g_0 > 0$. Moreover, for $t \geq 0$ and $t \geq 1$, respectively,

$$\|C_l^t\|_2 \leq G_C, \qquad \|C_l^t - C_l^{t-1}\|_2 \leq v_C.$$

Write $\tau_{r,l}^t = \|C_l^t - C_{l,r}^t\|_2$. All these bounds are understood pathwise on the reference process under study.

Let $P_l^{\star,t}$ be the column projector of the chosen $C_{l,r}^t$. For $t \geq 0$, let P_l^t be a top-r_{ref} left singular projector of M_l^t and set $Q_l^t = I - P_l^t$. We choose P_l^t and $\check{M}_l^t = \Pi_{r_{\mathrm{ref}}}(M_l^t)$ from the same truncated SVD, so that $P_l^t \check{M}_l^t = \check{M}_l^t$. For the zero pre-round reference used only to initialize the recursion, define $P_l^{-1} = 0$ and $Q_l^{-1} = I$.

Theorem 2 (Score consistency) *Under Assumption 4, define*

$$b_{M,l}^t := \rho^{t+1} G_C + \frac{\rho}{1-\rho} v_C + \delta_e + \frac{\delta_c}{1-\rho}. \tag{40}$$

Then

$$\|P_l^t - P_l^{\star,t}\|_2 \leq \min\left\{1, \frac{2}{g_0}\left(b_{M,l}^t + \tau_{r,l}^t\right)\right\} =: \epsilon_{P,l}^t. \tag{41}$$

Whenever $b_{M,l}^t + \tau_{r,l}^t < g_0/2$, the top-$r_{\mathrm{ref}}$ left singular subspace of M_l^t is unique.

Fix an active client–module pair in round $t \geq 1$. Let $X_{i,l}^{t,K}$ be its nonzero frozen-gradient update defined in Theorem 3, and suppose

$$\|U_{i,l}^{t,K} - X_{i,l}^{t,K}\|_F \leq \epsilon_{u,i,l}^t \|X_{i,l}^{t,K}\|_F, \qquad 0 \leq \epsilon_{u,i,l}^t < 1. \tag{42}$$

For generality, allow a reference threshold $\varepsilon_M \geq 0$; the deployed specification uses $\varepsilon_M = 0$. Assume that the threshold conventions are not activated, as guaranteed by

$$\|M_l^{t-1}\|_F > \varepsilon_M, \qquad (1 - \epsilon_{u,i,l}^t)\|X_{i,l}^{t,K}\|_F > \varepsilon_\Delta, \tag{43}$$

and assume the positive orientation margin

$$\langle X_{i,l}^{t,K}, M_l^{t-1} \rangle_F \geq \nu_{i,l}^t \|X_{i,l}^{t,K}\|_F \|M_l^{t-1}\|_F, \qquad \epsilon_{u,i,l}^t < \nu_{i,l}^t. \tag{44}$$

Then the hard gate is one for both $X_{i,l}^{t,K}$ and $U_{i,l}^{t,K}$, and the small-update convention is not invoked. For the deployed one-sided score, let $q_{i,l}^t = 1 - \tilde{s}_{i,l}^t$ and

$$\theta_{i,l}^t = \frac{\|Q_l^{t-1} X_{i,l}^{t,K}\|_F^2}{\|X_{i,l}^{t,K}\|_F^2}, \qquad \theta_{i,l}^{\star,t} = \frac{\|(I - P_l^{\star,t-1}) X_{i,l}^{t,K}\|_F^2}{\|X_{i,l}^{t,K}\|_F^2}.$$

Then

$$|q_{i,l}^t - \theta_{i,l}^t| \leq \frac{4\epsilon_{u,i,l}^t}{1 - \epsilon_{u,i,l}^t}, \tag{45}$$

$$|q_{i,l}^t - \theta_{i,l}^{\star,t}| \leq \frac{4\epsilon_{u,i,l}^t}{1 - \epsilon_{u,i,l}^t} + \epsilon_{P,l}^{t-1}. \tag{46}$$

No calibration claim is made when the orientation or norm-margin conditions fail; such rounds are handled explicitly in the controller bound below.

Lemma 3 (EMA reference tracking) *Under Assumption 4,*

$$\|M_l^t - C_l^t\|_2 \leq \rho^{t+1} G_C + \frac{\rho}{1-\rho} v_C + \delta_e + \frac{\delta_c}{1-\rho} = b_{M,l}^t. \tag{47}$$

Proof. From equation 39 and $M_l^{-1} = 0$,

$$M_l^t = (1-\rho)\sum_{s=0}^{t}\rho^{t-s}C_l^s + \sum_{s=0}^{t}\rho^{t-s}D_l^s.$$

Subtracting C_l^t gives a missing-mass term $-\rho^{t+1}C_l^t$, a temporal variation term, and the accumulated residual. Therefore

$$\begin{aligned}\|M_l^t - C_l^t\|_2 &\le \rho^{t+1}G_C + (1-\rho)\sum_{s=0}^{t}\rho^{t-s}\|C_l^s - C_l^t\|_2 + \sum_{s=0}^{t}\rho^{t-s}\|D_l^s\|_2 \\ &\le \rho^{t+1}G_C + v_C(1-\rho)\sum_{h=0}^{\infty}h\rho^h + \frac{\delta_M}{1-\rho} \\ &= \rho^{t+1}G_C + \frac{\rho}{1-\rho}v_C + \delta_e + \frac{\delta_c}{1-\rho}.\end{aligned}$$

The last equality uses $\delta_M = (1-\rho)\delta_e + \delta_c$. □

Lemma 4 (Projector perturbation) *Let C_r have rank r and $\sigma_r(C_r) \ge g_0 > 0$. If $M = C_r + E$, let P be the column projector of C_r. If $\|E\|_2 < g_0/2$, the top-r left singular subspace of M is unique; denote its projector by $\widehat{P}$. In that case*

$$\|\widehat{P} - P\|_2 \le \frac{2\|E\|_2}{g_0}.$$

For arbitrary E, if $\widehat{P}$ is any top-r left singular projector of M, then

$$\|\widehat{P} - P\|_2 \le \min\{1, 2\|E\|_2/g_0\}.$$

Proof. Put $e = \|E\|_2$. If $e < g_0/2$, Weyl's inequality gives

$$\sigma_r(M) \ge \sigma_r(C_r) - e \ge g_0 - e > e \ge \sigma_{r+1}(M),$$

where the final inequality follows from Eckart–Young because C_r has rank r. Hence the top-r left singular subspace of M is unique. For any unit $y \in \mathrm{col}(C_r)$, let $v = C_r^\dagger y$. Then $C_r v = y$ and $\|v\| \le 1/g_0$. Since $\|(I-\widehat{P})M\|_2 = \sigma_{r+1}(M) \le e$,

$$\begin{aligned}\|(I-\widehat{P})y\| &= \|(I-\widehat{P})(M-E)v\| \\ &\le \|(I-\widehat{P})M\|_2\|v\| + \|E\|_2\|v\| \\ &\le \frac{2e}{g_0}.\end{aligned}$$

Taking the supremum over unit $y \in \mathrm{col}(C_r)$ yields $\|(I-\widehat{P})P\|_2 \le 2e/g_0$. By the CS decomposition of two r-dimensional subspaces, the singular values of $(I-\widehat{P})P$ are the sines of their principal angles, and the nonzero eigenvalues of $P - \widehat{P}$ are the corresponding positive and negative sines. Thus

$$\|P - \widehat{P}\|_2 = \|(I-\widehat{P})P\|_2.$$

This proves the nontrivial case. If $e \ge g_0/2$, then $2e/g_0 \ge 1$, while the difference of two orthogonal projectors has operator norm at most one, proving the stated global bound. □

Lemma 5 (Normalized score perturbation) *For orthogonal projectors $P, \widehat{P}$ and nonzero matrices $X, \widehat{X} = X + \Xi$ satisfying $\|\Xi\|_F \le \epsilon_u\|X\|_F$ with $\epsilon_u < 1$,*

$$\left|\frac{\|\widehat{P}\widehat{X}\|_F^2}{\|\widehat{X}\|_F^2} - \frac{\|PX\|_F^2}{\|X\|_F^2}\right| \le \frac{4\epsilon_u}{1-\epsilon_u} + \|\widehat{P} - P\|_2.$$

Proof. For fixed normalized X, replacing P by $\widehat{P}$ changes the quadratic form by at most $\|\widehat{P}-P\|_2$. Put $x = X/\|X\|_F$ and $\widehat{x} = \widehat{X}/\|\widehat{X}\|_F$. Since $\|\widehat{X}\|_F \geq (1-\epsilon_u)\|X\|_F$,

$$\|\widehat{x} - x\|_F \leq \frac{2\epsilon_u}{1-\epsilon_u}.$$

For any orthogonal projector H and unit vectors x, y,

$$|\langle x, Hx\rangle_F - \langle y, Hy\rangle_F| \leq 2\|x-y\|_F,$$

so the normalized update perturbation contributes at most $4\epsilon_u/(1-\epsilon_u)$. Adding the projector perturbation proves the claim. □

Proof of Theorem 2. Write

$$M_l^t = C_{l,r}^t + \big(M_l^t - C_l^t\big) + \big(C_l^t - C_{l,r}^t\big).$$

Lemma 3 gives

$$\|M_l^t - C_{l,r}^t\|_2 \leq b_{M,l}^t + \tau_{r,l}^t.$$

Applying Lemma 4 proves equation 41, including the uniqueness statement in the nontrivial regime.

For the score statement, the relative update bound implies

$$\|U_{i,l}^{t,K}\|_F \geq (1-\epsilon_{u,i,l}^t)\|X_{i,l}^{t,K}\|_F > \varepsilon_\Delta,$$

while equation 43 gives $\|M_l^{t-1}\|_F > \varepsilon_M$. Hence the implementation does not invoke its small-norm convention. Moreover,

$$\begin{aligned}\langle U_{i,l}^{t,K}, M_l^{t-1}\rangle_F &\geq \langle X_{i,l}^{t,K}, M_l^{t-1}\rangle_F - \|U_{i,l}^{t,K} - X_{i,l}^{t,K}\|_F\|M_l^{t-1}\|_F \\ &\geq (\nu_{i,l}^t - \epsilon_{u,i,l}^t)\|X_{i,l}^{t,K}\|_F\|M_l^{t-1}\|_F > 0.\end{aligned}$$

Thus the hard gate equals one. Apply Lemma 5 first with the same deployed projector on both sides, and then with P_l^{t-1} and $P_l^{\star,t-1}$. □

Corollary 3 (Probabilistic score calibration) *Fix a nonzero frozen block update $X_{i,l}^{t,K}$ and suppose $\|M_l^{t-1}\|_F > \varepsilon_M$. Assume the deterministic orientation margin*

$$\langle X_{i,l}^{t,K}, M_l^{t-1}\rangle_F \geq \nu\|X_{i,l}^{t,K}\|_F\|M_l^{t-1}\|_F$$

holds for some $\nu > 0$. For every

$$0 < \varepsilon < \min\left\{1, \nu, 1 - \frac{\varepsilon_\Delta}{\|X_{i,l}^{t,K}\|_F}\right\},$$

where the interval is nonempty only when $\|X_{i,l}^{t,K}\|_F > \varepsilon_\Delta$, the score-calibration conclusion of Theorem 2 holds with $\epsilon_{u,i,l}^t = \varepsilon$ except on an event of conditional probability at most

$$\frac{2\mathbb{E}[\|R_i^t\|^2 \mid \mathcal{F}_{t,0}] + 2\eta^2 K\sigma^2}{\varepsilon^2\|X_{i,l}^{t,K}\|_F^2}.$$

For a finite family of blocks satisfying their corresponding deterministic norm, orientation, and reference assumptions, the failure probabilities may be summed by a union bound.

Proof. The event $\|U_{i,l}^{t,K} - X_{i,l}^{t,K}\|_F \leq \varepsilon\|X_{i,l}^{t,K}\|_F$ implies both the norm margin and positive orientation required by Theorem 2. Moreover,

$$\|U_{i,l}^{t,K} - X_{i,l}^{t,K}\|_F \leq \|R_i^t + \zeta_i^t\|,$$

and

$$\mathbb{E}[\|R_i^t + \zeta_i^t\|^2 \mid \mathcal{F}_{t,0}] \leq 2\mathbb{E}[\|R_i^t\|^2 \mid \mathcal{F}_{t,0}] + 2\eta^2 K\sigma^2.$$

Markov's inequality proves the one-block statement, and the last claim is the ordinary union bound. □

C.3 THE CAUSAL EMA CONTROLLER

This subsection analyzes the controller recursion in Eq. equation 16. Lemma 6 shows that the smoothed strength tracks a slowly varying drift target up to score errors, uncalibrated rounds, and an initialization transient. This bound provides the controller-calibration error used in the allocation comparison of Theorem 4.

Lemma 6 (Robust causal EMA tracking) *The controller recursion in equation 16 satisfies, for* $t \geq 1$,

$$\frac{\mu_{i,l}^t}{\mu_{\max}} = (1-\gamma)\sum_{r=0}^{t-1}\gamma^{t-1-r}q_{i,l}^r. \tag{48}$$

Let $\theta_{i,l}^r \in [0,1]$ *be a target drift sequence. On rounds for which Theorem 2 is applicable, take it to be the frozen off-reference energy in that theorem. Let* $b_{i,l}^r \in \{0,1\}$ *indicate an uncalibrated round, including an anti-oriented or thresholded round. Suppose*

$$|q_{i,l}^r - \theta_{i,l}^r| \leq (1-b_{i,l}^r)\epsilon_{s,i,l}^r + b_{i,l}^r, \qquad |\theta_{i,l}^r - \theta_{i,l}^{r-1}| \leq v_\theta.$$

Define

$$\overline{\epsilon}_{s,i,l,t} = (1-\gamma)\sum_{r=0}^{t-1}\gamma^{t-1-r}(1-b_{i,l}^r)\epsilon_{s,i,l}^r, \tag{49}$$

$$\Gamma_{i,l,t} = (1-\gamma)\sum_{r=0}^{t-1}\gamma^{t-1-r}b_{i,l}^r. \tag{50}$$

Then

$$\left|\frac{\mu_{i,l}^t}{\mu_{\max}} - \theta_{i,l}^t\right| \leq \gamma^t + \overline{\epsilon}_{s,i,l,t} + \Gamma_{i,l,t} + \frac{v_\theta}{1-\gamma} =: \epsilon_{\mu,i,l,t}. \tag{51}$$

In particular, if all relevant past rounds are calibrated and $\epsilon_{s,i,l}^r \leq \epsilon_s$, *then* $\Gamma_{i,l,t} = 0$ *and* $\overline{\epsilon}_{s,i,l,t} \leq \epsilon_s$. *If* θ *is constant and the score errors are independent, zero mean, and have variance at most* σ_q^2 *on every round, the stationary variance of* $\mu/\mu_{\max}$ *around* θ *is at most* $(1-\gamma)\sigma_q^2/(1+\gamma)$.

Proof of Lemma 6. Unrolling equation 16 from $\mu^0 = 0$ gives equation 48. Adding and subtracting the target sequence,

$$\left|\frac{\mu^t}{\mu_{\max}} - \theta^t\right| \leq \gamma^t|\theta^t| + (1-\gamma)\sum_{r=0}^{t-1}\gamma^{t-1-r}|q^r - \theta^r| + (1-\gamma)\sum_{r=0}^{t-1}\gamma^{t-1-r}|\theta^r - \theta^t|.$$

The first term is at most γ^t. The assumed score-error decomposition makes the second term exactly bounded by $\overline{\epsilon}_{s,t} + \Gamma_t$. Temporal regularity gives $|\theta^r - \theta^t| \leq (t-r)v_\theta$, and therefore

$$(1-\gamma)\sum_{r=0}^{t-1}\gamma^{t-1-r}(t-r)v_\theta \leq (1-\gamma)v_\theta\sum_{h=1}^{\infty}h\gamma^{h-1} = \frac{v_\theta}{1-\gamma}.$$

This proves equation 51. For constant θ and no uncalibrated rounds, the centered stationary recursion is a linear filter with squared weights $(1-\gamma)^2\gamma^{2h}$, whose sum is $(1-\gamma)/(1+\gamma)$. □

C.4 LOCAL K-STEP AND PATH RESPONSES

This subsection characterizes the local K-step trajectory under adaptive proximal control. The frozen-gradient recursion admits exact final-step and path responses, and the remaining lemmas bound the deviation of the actual nonlinear stochastic recursion from this idealization. The resulting remainder estimates enter the proof of Theorem 1.

Theorem 3 (Local response and perturbation) *Fix a round and write* $G_i = \nabla F_i(W^t)$. *For the frozen-gradient recursion obtained from equation 21 by replacing* $\nabla F_i(W^t + U_i^{t,k})$ *with* G_i *and setting* $\xi_i^{t,k} = 0$, *the block-*l *solution after* k *steps is*

$$X_{i,l}^{t,k} = -\eta k\, a_k(\mu_{i,l}^t)G_{i,l}^t + \kappa_k(\mu_{i,l}^t)Z_l^t. \tag{52}$$

In particular, since $P_l^{t-1} Z_l^t = Z_l^t$,

$$\|Q_l^{t-1} X_{i,l}^{t,K}\|_F^2 = \tau^2 \phi_K(\mu_{i,l}^t) \|Q_l^{t-1} G_{i,l}^t\|_F^2. \tag{53}$$

Moreover,

$$\frac{1}{K} \sum_{k=0}^{K-1} \|Q_l^{t-1} X_{i,l}^{t,k}\|_F^2 = \tau^2 \psi_K(\mu_{i,l}^t) \|Q_l^{t-1} G_{i,l}^t\|_F^2. \tag{54}$$

For the actual stochastic recursion, there exist a martingale term ζ_i^t and a nonlinear remainder R_i^t such that

$$\begin{aligned} U_i^{t,K} &= X_i^{t,K} + R_i^t + \zeta_i^t, \\ \mathbb{E}[\zeta_i^t \mid \mathcal{F}_{t,0}] &= 0, \\ \mathbb{E}[\|\zeta_i^t\|^2 \mid \mathcal{F}_{t,0}] &\le \eta^2 K \sigma^2. \end{aligned}$$

The supporting lemmas below give the explicit bound

$$\mathbb{E}[\|R_i^t\|^2 \mid \mathcal{F}_{t,0}] \le C_{\mathrm{loc}} \tau^4 L_F^2 \left(\mathcal{H}_{i,t}^{\mathrm{path}} + \mathcal{P}_{i,t}^{\mathrm{path}} + \mathcal{A}_{i,t}^{\mathrm{path}} + \frac{\sigma^2}{K} \right). \tag{55}$$

where the three path quantities are defined in Subsection C.4. Consequently,

$$\mathbb{E}[\|Q_l^{t-1} U_{i,l}^{t,K}\|_F^2 \mid \mathcal{F}_{t,0}] \le 2\tau^2 \phi_K(\mu_{i,l}^t) \|Q_l^{t-1} G_{i,l}^t\|_F^2 + 4\mathbb{E}[\|R_i^t\|^2 \mid \mathcal{F}_{t,0}] + 4\eta^2 K \sigma^2. \tag{56}$$

Thus the direct stochastic term is of order $\tau^2 \sigma^2 / K$ and does not disappear when $L_F = 0$.

For one client, define the frozen recursion

$$X_i^{k+1} = S_i X_i^k - \eta G_i + \eta \Lambda_i Z, \qquad X_i^0 = 0. \tag{57}$$

For later use, set

$$\mathcal{H}_{i,t}^{\mathrm{path}} = \sum_l \psi_K(\mu_{i,l}^t) \|Q_l G_{i,l}^t\|_F^2, \tag{58}$$

$$\mathcal{P}_{i,t}^{\mathrm{path}} = \sum_l \psi_K(\mu_{i,l}^t) \|P_l G_{i,l}^t\|_F^2, \tag{59}$$

$$\mathcal{A}_{i,t}^{\mathrm{path}} = \sum_l \chi_K(\mu_{i,l}^t) \frac{\|Z_l^t\|_F^2}{\tau^2}. \tag{60}$$

Lemma 7 (Frozen solution and path energy) *The solution of equation 57 is*

$$X_{i,l}^k = -\eta k a_k(\mu_{i,l}) G_{i,l} + \kappa_k(\mu_{i,l}) Z_l.$$

Moreover,

$$\sum_{k=0}^{K-1} \|X_i^k\|^2 \le K\tau^2 \left(\mathcal{H}_{i,t}^{\mathrm{path}} + 2\mathcal{P}_{i,t}^{\mathrm{path}} + 2\mathcal{A}_{i,t}^{\mathrm{path}} \right). \tag{61}$$

Proof. Unroll the affine recursion and sum the geometric series. Applying Q_l kills the anchor and gives

$$Q_l X_{i,l}^k = -\eta k a_k(\mu_{i,l}) Q_l G_{i,l},$$

which proves the exact final and path identities. In the reference subspace,

$$\|P_l X_{i,l}^k\|_F^2 \le 2\eta^2 k^2 a_k(\mu_{i,l})^2 \|P_l G_{i,l}\|_F^2 + 2\kappa_k(\mu_{i,l})^2 \|Z_l\|_F^2.$$

Summing over k and l proves equation 61. □

Let $E_i^k = U_i^k - X_i^k$ and $\delta_i^k = \nabla F_i(W^t + U_i^k) - G_i$.

Lemma 8 (Perturbed local path) *Under Assumptions 1 and 3,*

$$\mathbb{E}[\|E_i^k\|^2 \mid \mathcal{F}_{t,0}] \le 2\eta^2 k L_F^2 \sum_{h=0}^{k-1} \mathbb{E}[\|U_i^h\|^2 \mid \mathcal{F}_{t,0}] + 2\eta^2 k \sigma^2. \tag{62}$$

Consequently,

$$\sum_{k=0}^{K-1} \mathbb{E}[\|U_i^k\|^2 \mid \mathcal{F}_{t,0}] \le 4K\tau^2 \left(\mathcal{H}_{i,t}^{\mathrm{path}} + 2\mathcal{P}_{i,t}^{\mathrm{path}} + 2\mathcal{A}_{i,t}^{\mathrm{path}} \right) + 4\tau^2 \sigma^2. \tag{63}$$

Proof. Subtract equation 57 from the actual recursion and unroll:

$$E_i^k = -\eta \sum_{h=0}^{k-1} S_i^{k-1-h}(\delta_i^h + \xi_i^h).$$

Cauchy–Schwarz bounds the deterministic sum. Conditional martingale orthogonality with respect to $\mathcal{F}_{t,h}$ bounds the stochastic sum. Using $\|S_i\| \le 1$ and $\|\delta_i^h\| \le L_F\|U_i^h\|$ gives equation 62.

Put $\mathcal{S}_i = \sum_{k=0}^{K-1} \mathbb{E}[\|U_i^k\|^2 \mid \mathcal{F}_{t,0}]$ and $\mathcal{X}_i = \sum_{k=0}^{K-1} \|X_i^k\|^2$. Since $\|U_i^k\|^2 \le 2\|X_i^k\|^2 + 2\|E_i^k\|^2$, summing equation 62 gives

$$\mathcal{S}_i \le 2\mathcal{X}_i + 4\tau^2 L_F^2 \mathcal{S}_i + 2\tau^2\sigma^2.$$

Assumption 3 implies $4\tau^2 L_F^2 \le 1/4$, hence $\mathcal{S}_i \le 4\mathcal{X}_i + 4\tau^2\sigma^2$. Apply Lemma 7. □

Lemma 9 (Martingale and nonlinear remainder) *The final local update decomposes as*

$$U_i^K = X_i^K + R_i + \zeta_i, \tag{64}$$

where

$$\zeta_i = -\eta \sum_{k=0}^{K-1} S_i^{K-1-k}\xi_i^k, \tag{65}$$

$$R_i = -\eta \sum_{k=0}^{K-1} S_i^{K-1-k}\delta_i^k. \tag{66}$$

They satisfy

$$\mathbb{E}[\zeta_i \mid \mathcal{F}_{t,0}] = 0,$$
$$\mathbb{E}[\|\zeta_i\|^2 \mid \mathcal{F}_{t,0}] \le \eta^2 K\sigma^2, \tag{67}$$
$$\mathbb{E}[\|R_i\|^2 \mid \mathcal{F}_{t,0}] \le 4\tau^4 L_F^2 \left(\mathcal{H}_{i,t}^{\mathrm{path}} + 2\mathcal{P}_{i,t}^{\mathrm{path}} + 2\mathcal{A}_{i,t}^{\mathrm{path}} + \frac{\sigma^2}{K}\right). \tag{68}$$

Proof. The decomposition follows from the same unrolling used above. Conditional martingale orthogonality gives equation 67. For the nonlinear part,

$$\mathbb{E}[\|R_i\|^2 \mid \mathcal{F}_{t,0}] \le \eta^2 K L_F^2 \sum_{k=0}^{K-1} \mathbb{E}[\|U_i^k\|^2 \mid \mathcal{F}_{t,0}].$$

Apply Lemma 8 and use $\eta^2 K = \tau^2/K$. □

Proof of Theorem 3. The exact frozen identities follow from Lemma 7; the remainder statement follows from Lemma 9. Finally,

$$\|Q_l U_{i,l}^K\|^2 \le 2\|Q_l X_{i,l}^K\|^2 + 4\|R_i\|^2 + 4\|\zeta_i\|^2,$$

and conditional expectation gives equation 56. The displayed remainder summary holds, for example, with $C_{\mathrm{loc}} = 8$ after absorbing the coefficients of the three path terms into a single numerical constant. □

C.5 Adaptive allocation

This subsection compares adaptive allocation with uniform control at the same weighted mean strength. Theorem 4 gives sufficient calibration, dispersion, and curvature conditions under which score-guided allocation strictly reduces the corresponding final-step or path-wise off-subspace response.

Index a normalized collection of positively oriented, score-eligible active client–module pairs by j, with weights $\omega_j > 0$ and $\sum_j \omega_j = 1$. For a response r_K, define

$$D_{r,t}(\boldsymbol{\mu}) = \sum_j \omega_j d_j^t r_K(\mu_j), \qquad d_j^t = \|Q_l^{t-1} G_{i,l}^t\|_F^2, \tag{69}$$

and let $\bar{\mu}_t = \sum_j \omega_j \mu_j^t$.

Theorem 4 (Matched-uniform dominance) *Take either $r_K = \phi_K$ with $K \geq 2$, or $r_K = \psi_K$ with $K \geq 3$. Let*

$$X_j^t = X_{i,l}^{t,K},\ \theta_j^t = \frac{\|Q_l^{t-1} X_j^t\|_F^2}{\|X_j^t\|_F^2},\ g_j^t = \frac{\|X_j^t\|_F^2}{\tau^2 a_K(\mu_j^t)^2}.$$

so that $d_j^t = g_j^t \theta_j^t$. Suppose $0 < g_{\min} \leq g_j^t \leq g_{\max}$ and

$$\left| \frac{\mu_j^t}{\mu_{\max}} - \theta_j^t \right| \leq \epsilon_{\mu,t}.$$

Let

$$V_{\theta,t} = \sum_j \omega_j (\theta_j^t - \bar{\theta}_t)^2$$

and define

$$\mathfrak{C}_t = g_{\min} V_{\theta,t} - \frac{(g_{\max} - g_{\min})\bar{\theta}_t}{2} \sqrt{V_{\theta,t}} - g_{\max} \epsilon_{\mu,t}. \tag{70}$$

Let $M_{r,t} = \sup_{0 \leq u \leq \mu_{\max}} |r_K''(u)|$. If $\mathfrak{C}_t > 0$ and

$$-r_K'(\bar{\mu}_t)\mathfrak{C}_t > \frac{M_{r,t}}{2} g_{\max} \mu_{\max}, \tag{71}$$

then

$$D_{r,t}(\boldsymbol{\mu}^t) < D_{r,t}(\bar{\mu}_t \mathbf{1}). \tag{72}$$

For $r_K = \phi_K$, this is a strict reduction of the leading final-step off-reference energy. For $r_K = \psi_K$, it is a strict reduction of the leading off-reference path term that appears in Theorem 1. At $K = 2$, $\psi_2 \equiv 1/8$, so no strict path-response dominance is possible. The comparison is made at the same round state, with the same gradients, projector, anchor, and weighted mean strength; it is not a trajectory-wise ordering of the complete objectives.

Remark 2 (Closed-loop nature of calibration) *When $\beta > 0$, the frozen update $X_j = -\tau a_K(\mu_j) G_j + \kappa_K(\mu_j) Z$ depends on the current strength. Consequently, the normalized quantity θ_j in Theorem 4 also depends on μ_j through its in-reference component. The calibration assumption is therefore a closed-loop controller–sensor consistency condition, not an exogenous measurement identity. When $\beta = 0$ and $G_j \neq 0$,*

$$\theta_j = \frac{\|QG_j\|_F^2}{\|G_j\|_F^2},$$

which is independent of μ_j.

Lemma 10 (Response monotonicity) *On $[0, \mu_{\max}]$, both ϕ_K and ψ_K are nonincreasing and twice continuously differentiable. For $K \geq 2$, ϕ_K is strictly decreasing. Moreover, $\psi_2 \equiv 1/8$, while ψ_K is strictly decreasing for $K \geq 3$.*

Proof. Each a_k is a finite average of nonnegative powers of $1 - \eta\mu \in (0, 1]$, hence is nonincreasing and smooth. The same properties hold after squaring and taking nonnegative weighted sums. For $K \geq 2$, a_K contains a positive multiple of $1 - \eta\mu$, so a_K' and hence ϕ_K' are strictly negative. If $K = 2$, the only nonzero term in ψ_K is $k = 1$, and $a_1 = 1$, giving $\psi_2 = (1/2)(1/2)^2 = 1/8$. If $K \geq 3$, the $k = 2$ term has positive weight and $a_2(\mu) = 1 - \eta\mu/2$ is strictly decreasing and positive, so the full nonnegative sum ψ_K is strictly decreasing. □

Lemma 11 (Matched-uniform covariance bound) *Let r be twice continuously differentiable on $[0, \mu_{\max}]$, $\bar{\mu} = \sum_j \omega_j \mu_j$, and $M_r = \sup_{[0,\mu_{\max}]} |r''|$. Then*

$$\sum_j \omega_j d_j r(\mu_j) - \sum_j \omega_j d_j r(\bar{\mu}) \leq r'(\bar{\mu})\, \mathrm{Cov}_\omega(d, \mu) + \frac{M_r}{2} \sum_j \omega_j d_j (\mu_j - \bar{\mu})^2.$$

Proof. Taylor's theorem around $\bar{\mu}$ gives

$$r(\mu_j) \le r(\bar{\mu}) + r'(\bar{\mu})(\mu_j - \bar{\mu}) + \frac{M_r}{2}(\mu_j - \bar{\mu})^2.$$

Multiply by $\omega_j d_j$ and sum. □

Lemma 12 (Positive covariance) *Suppose $d_j = g_j\theta_j$ and*

$$g_{\min} \le g_j \le g_{\max},$$
$$\left|\frac{\mu_j}{\mu_{\max}} - \theta_j\right| \le \epsilon_\mu.$$

Then

$$\mathrm{Cov}_\omega(d, \mu) \ge \mu_{\max} g_{\min} V_\theta - \mu_{\max} g_{\max} \epsilon_\mu - \frac{\mu_{\max}(g_{\max} - g_{\min})\bar{\theta}}{2}\sqrt{V_\theta}. \tag{73}$$

Proof. First,

$$\mathrm{Cov}_\omega(d, \theta) = \sum_j \omega_j g_j (\theta_j - \bar{\theta})^2 + \bar{\theta} \sum_j \omega_j g_j (\theta_j - \bar{\theta}).$$

The first term is at least $g_{\min} V_\theta$. Center g_j at $(g_{\max} + g_{\min})/2$ and apply Cauchy–Schwarz to obtain

$$\left|\sum_j \omega_j g_j (\theta_j - \bar{\theta})\right| \le \frac{g_{\max} - g_{\min}}{2}\sqrt{V_\theta}.$$

Write $\mu_j = \mu_{\max}(\theta_j + e_j)$ with $|e_j| \le \epsilon_\mu$. Since $0 \le d_j \le g_{\max}$,

$$|\mathrm{Cov}_\omega(d, e)| \le \epsilon_\mu \sum_j \omega_j |d_j - \bar{d}| \le g_{\max}\epsilon_\mu.$$

Combining the bounds proves the claim. □

Proof of Theorem 4. By equation 53, $d_j = g_j\theta_j$. Lemma 12 gives $\mathrm{Cov}_\omega(d, \mu) \ge \mu_{\max}\mathfrak{C}_t$. Moreover, $d_j \le g_{\max}$ and $|\mu_j - \bar{\mu}| \le \mu_{\max}$, so

$$\sum_j \omega_j d_j (\mu_j - \bar{\mu})^2 \le g_{\max}\mu_{\max}^2.$$

Apply Lemma 11. The strict monotonicity conditions required for a potentially nonvacuous strict comparison are exactly those stated in Theorem 4, by Lemma 10. Condition equation 71 makes the resulting upper bound strictly negative. □

C.6 Representation and anchor results

This subsection collects complementary results on the properties and trade-offs of adaptive control: the representation bias induced by heterogeneous proximal strengths, a calibration rule for an explicit surrogate objective, the sensitivity of the proximal solution to the lookahead anchor, and an optional confidence diagnostic. These results further characterize the mechanism but are not used in the proof of Theorem 1.

Theorem 5 (Quadratic representation bias) *Let $F_i(w) = \frac{1}{2}\|w - w_i^\star\|^2, \quad \mu_i \ge 0$. Suppose client i exactly minimizes $F_i(w) + \frac{\mu_i}{2}\|w - w^t\|^2$ to obtain w_i^{t+1}, and the server sets $w^{t+1} = \sum_i p_i w_i^{t+1}$. Assume that the strengths are fixed and the lookahead is zero. The unique fixed point is*

$$\widehat{w} = \sum_i q_i w_i^\star, \qquad q_i = \frac{p_i/(1+\mu_i)}{\sum_j p_j/(1+\mu_j)}.$$

With $\bar{w} = \sum_i p_i w_i^\star$ and $D_\star = \max_i \|w_i^\star - \bar{w}\|$,

$$\|\widehat{w} - \bar{w}\| \le D_\star \frac{\mu_{\mathrm{hi}} - \mu_{\mathrm{lo}}}{1 + \mu_{\mathrm{lo}}} \le D_\star(\mu_{\mathrm{hi}} - \mu_{\mathrm{lo}}). \tag{74}$$

If $|\mu_i/\mu_{\max} - \theta_i| \le \epsilon_\mu$*, then*

$$\|\widehat{w} - \bar{w}\| \le D_\star \mu_{\max} \left(\max_i \theta_i - \min_i \theta_i + 2\epsilon_\mu \right). \tag{75}$$

Thus differential control reweights clients toward those receiving smaller strengths; whether those clients form a numerical majority is an empirical question.

Corollary 4 (Surrogate design rule) *For constants* $A_0, B_0, c_0, q_0 > 0$*, consider the explicit surrogate*

$$\mathcal{J}(\mu) = A_0(1 + c_0 q_0 \mu)^{-2} + B_0 \mu, \qquad \mu \ge 0.$$

Its minimizer is

$$\mu^\star = \left[\frac{(2c_0 q_0 A_0/B_0)^{1/3} - 1}{c_0 q_0} \right]_+, \tag{76}$$

clipped to the permitted interval. This is a calibration rule for the stated surrogate, not a universal optimizer of the full FedSALT bound.

Proposition 4 (Anchor sensitivity) *If* h *is differentiable and* λ*-strongly convex and* $\mu \ge 0$*, consider the unique minimizer of*

$$h(x) + \frac{\mu}{2}\|x - z\|^2.$$

It is $\mu/(\lambda+\mu)$*-Lipschitz in* z*. No global minimizer-sensitivity claim is made for a general nonconvex local objective. Separately, if* $\langle C_t, \check{M}_{t-1}\rangle_F > 0$*, then every*

$$0 < \beta < \frac{2\langle C_t, \check{M}_{t-1}\rangle_F}{\|\check{M}_{t-1}\|_F^2}$$

makes $\beta\check{M}_{t-1}$ *strictly closer to* C_t *than the zero anchor. A persistent additive steering force can create a nonzero stationary gradient for the unforced objective even on a one-dimensional quadratic.*

Lemma 13 (Optional confidence statistic) *For two independent isotropic Gaussian directions in dimension* $d \ge 2$*, the expected rectified cosine is at most* $1/(2\sqrt{d})$*. If* $X = ag + \xi$ *and* $Y = bg + \zeta$ *with* $a, b > 0$*,* $\|g\| = 1$*,* $\|\xi\| \le \epsilon a$*, and* $\|\zeta\| \le \epsilon b$*, then*

$$\frac{\langle X, Y\rangle}{\|X\|\|Y\|} \ge \frac{1 - 2\epsilon - \epsilon^2}{(1+\epsilon)^2},$$

which is positive for $\epsilon < \sqrt{2} - 1$*. Finally, if* $\xi_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\xi^2 I_d)$ *and the pure-noise EMA* $M_t = \rho M_{t-1} + (1-\rho)\xi_t$ *is stationary, then the cosine between* M_t *and* M_{t-1} *converges in probability to* ρ *as* $d \to \infty$*. Thus a fresh-innovation confidence statistic is small in high-dimensional pure noise, whereas a statistic based on consecutive momenta need not be. This optional diagnostic is not used by the deployed orientation-gated controller or by Theorem 1.*

Proof of Theorem 5. The exact local solution is $w_i^{t+1} = (w_i^\star + \mu_i w^t)/(1+\mu_i)$. At a fixed point,

$$\widehat{w} = \sum_i \frac{p_i}{1+\mu_i} w_i^\star + \widehat{w} \sum_i \frac{p_i \mu_i}{1+\mu_i},$$

which gives the displayed q_i. Put $a_i = (1+\mu_i)^{-1}$ and $A = \sum_i p_i a_i$. Then

$$\widehat{w} - \bar{w} = \sum_i p_i \left(\frac{a_i}{A} - 1 \right) (w_i^\star - \bar{w}).$$

Therefore

$$\|\widehat{w} - \bar{w}\| \le D_\star \sum_i p_i \left| \frac{a_i}{A} - 1 \right| \le D_\star \frac{a_{\max} - a_{\min}}{a_{\min}} = D_\star \frac{\mu_{\text{hi}} - \mu_{\text{lo}}}{1 + \mu_{\text{lo}}}.$$

If $|\mu_i/\mu_{\max} - \theta_i| \le \epsilon_\mu$, then

$$\mu_{\text{hi}} - \mu_{\text{lo}} \le \mu_{\max}(\max_i \theta_i - \min_i \theta_i + 2\epsilon_\mu),$$

which proves equation 75. □

Proof of Corollary 4. The derivative of $\mathcal{J}$ is

$$\mathcal{J}'(\mu) = -2c_0q_0A_0(1 + c_0q_0\mu)^{-3} + B_0,$$

which is strictly increasing. The minimizer is either zero or the unique positive root in equation 76. □

Proof of Proposition 4. Let $x(z)$ and $x(z')$ be the two proximal minimizers. Subtract their optimality conditions and take the inner product with $x(z) - x(z')$. Strong monotonicity of ∇h gives

$$(\lambda + \mu)\|x(z) - x(z')\|^2 \le \mu\|x(z) - x(z')\|\|z - z'\|.$$

Cancel the common norm. For the predictive statement, expand

$$\|C - \beta M\|_F^2 - \|C\|_F^2 = \beta^2\|M\|_F^2 - 2\beta\langle C, M\rangle_F.$$

The strict quadratic inequality gives the stated interval. Finally, the iteration $x_{t+1} = x_t - \eta(\lambda x_t - c)$ converges to c/λ for a stable step size. Its gradient for the forced objective is zero, but the gradient λx of the original unforced quadratic has magnitude $|c|$. □

Proof of Lemma 13. For two independent isotropic Gaussian directions, normalization gives two independent uniform directions on the sphere. Their cosine is symmetric about zero, so the expectation of its positive part is half the expectation of its absolute value. Since $\mathbb{E}[\cos^2] = 1/d$, Cauchy–Schwarz gives the first claim. In the signal case,

$$\langle X, Y\rangle \ge ab - a\|\zeta\| - b\|\xi\| - \|\xi\|\|\zeta\| \ge ab(1 - 2\epsilon - \epsilon^2),$$

while $\|X\|\|Y\| \le ab(1 + \epsilon)^2$.

For the final statement, stationarity of the scalar recursion in each coordinate gives $\mathrm{Corr}(M_{t,j}, M_{t-1,j}) = \rho$. The coordinate pairs are iid with finite fourth moments. The weak law of large numbers applied to the empirical inner product and the two squared norms therefore gives convergence of the cosine to ρ. □

C.7 Round-decomposition lemmas

This section outlines the key elements of the convergence proof: the precise breakdown of the combined frozen-gradient update, the limit on the aggregated nonlinear remainder, and the inequality for preconditioned biased descent utilized in the proof of Theorem 1.

Lemma 14 (Exact round decompositions) *The weighted frozen-gradient client update satisfies*

$$\sum_i p_i X_{i,l}^{t,K} = -\tau\bar{a}_l^t G_l^t - \tau b_l^t + e_{\mathrm{anc},l}^t, \tag{77}$$

$$= -\tau(\bar{a}_l^t + \bar{\kappa}_l^t)G_l^t - \tau b_l^t + e_{\mathrm{pred},l}^t. \tag{78}$$

Moreover,

$$\|b_l^t\|_F^2 \le \left(\sum_i p_i(a_{i,l}^t - \bar{a}_l^t)^2\right)\left(\sum_i p_i\|G_{i,l}^t - G_l^t\|_F^2\right). \tag{79}$$

Finally,

$$c_- \le \bar{a}_l^t \le 1, \qquad c_- \le \bar{a}_l^t + \bar{\kappa}_l^t \le 2. \tag{80}$$

Proof. Insert the final frozen solution and use

$$\sum_i p_i a_{i,l}^t G_{i,l}^t = \bar{a}_l^t G_l^t + b_l^t.$$

This gives the first decomposition. Add and subtract $\tau\bar{\kappa}_l^t G_l^t$ to obtain the second. Weighted Cauchy–Schwarz gives equation 79. Since $a_K(\mu) \ge (1 - \eta\mu)^{K-1} \ge c_-$, $a_K(\mu) \le 1$, and $0 \le \kappa_K(\mu) \le 1$, the ranges follow. □

Let A_t be block diagonal with block scalar $\bar{a}_l^t$. From Lemma 9, define

$$e_{\mathrm{loc},t} = \sum_i p_i R_i^t, \qquad \zeta_t = \sum_i p_i \zeta_i^t.$$

Using the first decomposition, the actual round is

$$W^{t+1} = W^t - \tau A_t \nabla F(W^t) - \tau b_t + e_{\mathrm{anc},t} + e_{\mathrm{loc},t} + \zeta_t. \tag{81}$$

Conditional independence across client noise sequences yields

$$\mathbb{E}[\|\zeta_t\|^2 \mid \mathcal{F}_{t,0}] \le \eta^2 K \chi_p \sigma^2. \tag{82}$$

For one round, denote the summands in equation 28, equation 29, and equation 31 by $H_{\mathrm{path},t}$, $\Omega^2_{\parallel,t}$, and $A_{\mathrm{path},t}$, respectively.

Lemma 15 (Aggregated nonlinear remainder) *The numerical constant $C_0 = 8$ satisfies*

$$\frac{1}{\tau^2}\mathbb{E}[\|e_{\mathrm{loc},t}\|^2 \mid \mathcal{F}_{t,0}] \le C_0 \tau^2 L_F^2 \left(\|\nabla F(W^t)\|^2 + H_{\mathrm{path},t} + \Omega^2_{\parallel,t} + A_{\mathrm{path},t} + \frac{\sigma^2}{K} \right). \tag{83}$$

Proof. Jensen's inequality and equation 68 give

$$\mathbb{E}[\|e_{\mathrm{loc},t}\|^2 \mid \mathcal{F}_{t,0}] \le \sum_i p_i \mathbb{E}[\|R_i^t\|^2 \mid \mathcal{F}_{t,0}].$$

The off-reference and anchor contributions are bounded by $4\tau^4 L_F^2 H_{\mathrm{path},t}$ and $8\tau^4 L_F^2 A_{\mathrm{path},t}$, respectively. For the in-reference part, use $0 \le \psi_K \le 1$ and the weighted variance decomposition

$$\sum_i p_i \|P_l G_{i,l}^t\|_F^2 = \|P_l G_l^t\|_F^2 + \sum_i p_i \|P_l (G_{i,l}^t - G_l^t)\|_F^2.$$

Summing over l bounds the first term by $\|\nabla F(W^t)\|^2$ and identifies the second as $\Omega^2_{\parallel,t}$. The largest resulting coefficient is eight. □

Lemma 16 (Preconditioned biased descent) *Let $\mathcal{F}$ be a sigma-field and suppose*

$$W^+ = W - \tau A \nabla F(W) - \tau b + e + \zeta,$$

where W, A, b are $\mathcal{F}$-measurable, A is self-adjoint with spectrum in $[c_-, 2]$, $0 < c_- \le 1$, $\mathbb{E}[\zeta \mid \mathcal{F}] = 0$, and $16 L_F \tau \le c_-$. The random vectors e and ζ need not be independent. Then

$$\begin{aligned}\mathbb{E}[F(W^+) \mid \mathcal{F}] \le F(W) - \frac{c_-\tau}{4}\|\nabla F(W)\|^2 + \frac{4\tau}{c_-}\|b\|^2 \\ + \frac{4}{c_-\tau}\mathbb{E}[\|e\|^2 \mid \mathcal{F}] + 4L_F \mathbb{E}[\|\zeta\|^2 \mid \mathcal{F}].\end{aligned} \tag{84}$$

Proof. Put $g = \nabla F(W)$ and $\Delta = -\tau A g - \tau b + e + \zeta$. Smoothness gives

$$F(W^+) \le F(W) + \langle g, \Delta \rangle + \frac{L_F}{2}\|\Delta\|^2.$$

After conditioning, the linear term involving ζ vanishes. Since $\langle g, Ag \rangle \ge c_- \|g\|^2$, Young's inequality gives

$$-\tau\langle g, b \rangle \le \frac{c_-\tau}{8}\|g\|^2 + \frac{2\tau}{c_-}\|b\|^2,$$

$$\mathbb{E}[\langle g, e \rangle \mid \mathcal{F}] \le \frac{c_-\tau}{8}\|g\|^2 + \frac{2}{c_-\tau}\mathbb{E}[\|e\|^2 \mid \mathcal{F}].$$

For the quadratic term,

$$\|\Delta\|^2 \le 4\tau^2\|Ag\|^2 + 4\tau^2\|b\|^2 + 4\|e\|^2 + 4\|\zeta\|^2.$$

Because $\|A\| \le 2$, its principal contribution is at most $8L_F\tau^2\|g\|^2 \le (c_-\tau/2)\|g\|^2$. Combining this with the two linear allocations leaves at least $c_-\tau/4$ descent. The remaining quadratic terms are absorbed into the displayed coefficients using $L_F\tau \le c_-/16$ and $c_- \le 1$. □

Table 3: Ablations on MC-8.

Variant	Accuracy	$E_\perp \downarrow$
FedSALT (full)	**58.52 ± 0.14**	**0.532 ± 0.004**
Sensing signal		
Input-side score, $s = s^R$	57.84 ± 0.26	0.544 ± 0.001
Two-sided score, $s = \sqrt{s^L s^R}$	57.97 ± 0.33	–
Selective allocation		
Matched-mean uniform control	57.09 ± 0.43	0.712 ± 0.058
Permuted assignment	58.15 ± 0.24	0.565 ± 0.002
Client-only granularity	57.80 ± 0.33	0.544 ± 0.005
Module-only granularity	57.37 ± 0.28	0.556 ± 0.030
Temporal smoothing		
No controller EMA, $\gamma = 0$	58.00 ± 0.27	0.535 ± 0.003
Correction target		
Zero-centered target, $\beta = 0$	58.10 ± 0.39	–

D Additional Experiments

D.1 Component Ablations

Sensing and selective allocation. To quantify the controller's geometric effect, we report the output-side off-subspace energy $E_\perp = \mathrm{mean}_{t,i,l}(1 - s_{i,l}^{L,t})$, averaged over controlled rounds, clients, and adapted modules. We compute $E_\perp$ for each seed and report the mean and standard deviation across three seeds. Table 3 links the sensing signal to the realized update geometry. Output-side sensing yields better alignment and predictive performance than input-side sensing, consistent with the asymmetry observed. Matched-mean uniform control produces substantially greater off-subspace drift despite preserving the average regularization strength, while permuting the coefficients weakens both alignment and accuracy. The benefit therefore depends on assigning control to the client-module pairs that generated the disagreement, rather than on the overall regularization level. Client-only and module-only variants further show that neither axis alone captures the relevant heterogeneity. Resolving both clients and modules yields updates that remain better aligned with the evolving consensus and produces the best global model.

Temporal smoothing and correction. Removing the controller EMA leaves off-subspace energy nearly unchanged but reduces predictive performance, indicating that temporal smoothing provides benefits beyond reducing average drift. Replacing the momentum-lookahead reference with a zero-centered target further weakens performance. Thus, the score determines where intervention is needed, while historical momentum determines how updates are corrected. Together, these findings support the sensing, smoothing, allocation, and correction stages of the FedSALT feedback loop.

D.2 Shared Experimental Protocol

All methods use identical data partitions, evaluation procedures, and training budgets. For each seed, the eight clients share the same train/validation/test splits and fixed client assignments. Evaluation uses the same chat template, answer parsing procedure, and prediction rule across methods.

All methods are trained for the same number of communication rounds with the same client participation schedule, data-size-weighted aggregation, and optimizer-state reset at the beginning of each round. LoRA is applied to `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`, with rank $r = 16$ and scaling $\alpha = 32$. We use batch size 16, $K = 40$ local steps, the same three random seeds, and the same numerical precision for all methods.

All methods use AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, weight decay *0.05*, and learning rate $1e^{-4}$, without learning-rate scheduling or gradient clipping. The optimizer configuration is identical across all methods.

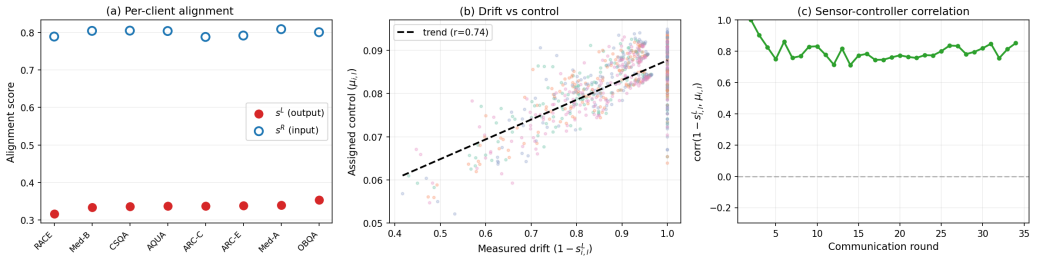


Figure 3: Output-side disagreement and its conversion into adaptive client-module control.

Table 4: Runtime and per-round communication overhead. FFA-LoRA and FedSA-LoRA communicate only B and A, respectively; SCAFFOLD additionally communicates client and server control variables; FedSALT additionally broadcasts the momentum adapter.

Method	Train (s)	Server (s)	Total (s)	Runtime overhead	Upload	Download	Total communication	Ratio
FedAvg	2.92	0.03	2.94	baseline	45.1 MB	45.1 MB	90.2 MB	1.00×
FFA-LoRA	2.80	0.02	2.83	−4.0%	24.1 MB	24.1 MB	48.2 MB	0.53×
FedSALT	2.89	0.24	3.13	+6.3%	45.1 MB	90.2 MB	135.3 MB	1.50×
FedEx-LoRA	2.92	0.20	3.12	+6.0%	45.1 MB	45.1 MB	90.2 MB	1.00×
FedProx	3.34	0.03	3.36	+14.3%	45.1 MB	45.1 MB	90.2 MB	1.00×
FedRot-LoRA	3.34	0.02	3.36	+14.2%	45.1 MB	45.1 MB	90.2 MB	1.00×
SCAFFOLD	4.05	0.02	4.07	+38.2%	90.2 MB	90.2 MB	180.4 MB	2.00×
FedSA-LoRA	4.06	0.01	4.07	+38.3%	21.0 MB	21.0 MB	42.0 MB	0.47×

D.3 Mechanism Visualization

Figure 3 visualizes the complete sensing-to-control pathway on MC-8 dataset. Output-side alignment exhibits substantially greater client-level variation than input-side alignment, and stronger measured disagreement leads to stronger control in the following round. The persistent relationship across communication rounds confirms that the one-round-delayed controller preserves and acts on the geometric signal rather than collapsing to uniform regularization.

D.4 System Overhead

Table 4 shows that FedSALT incurs modest computational cost relative to local drift-control methods, measured on MC-8 with Llama-3.2-1B (bf16) across 8 clients, with a single NVIDIA RTX 6000 Ada GPU. It is faster than FedProx and SCAFFOLD, while its communication cost lies between their $1.0\times$ and $2.0\times$ costs. The additional communication arises from broadcasting one shared momentum adapter; the client upload remains unchanged from FedAvg.

D.5 Dataset Details

MC-8. Eight naturally heterogeneous multiple-choice QA datasets, each assigned to one client: MedMCQA-A and MedMCQA-B (medical, 5-choice), CSQA (commonsense, 5-choice), AQUA (algebraic, 5-choice), RACE (reading comprehension, 4-choice), OBQA (science, 4-choice), ARC-Easy and ARC-Challenge (grade-school science, 4-choice). Heterogeneity arises from domain differences rather than label manipulation.

CS8-Dir (α=0.1)**.** Eight commonsense reasoning tasks—BoolQ, PIQA, SIQA, HellaSwag, WinoGrande, OBQA, ARC-Easy, and ARC-Challenge—are pooled and redistributed across eight clients via a Dirichlet(α=0.1) allocation over four answer labels. The low concentration parameter produces highly skewed label distributions, creating label-distribution heterogeneity with shared domain.

MMLU8. Eight clients drawn from distinct MMLU subject groups and MedMCQA: MedMCQA (medical), STEM-Physics, STEM-Mathematics, Biology-Chemistry, Law, Humanities, Social Science, and Professional. Each client holds examples exclusively from its assigned group, producing domain heterogeneity with a unified four-choice format.