COMMUNICATING LEARNING DIRECTIONS THROUGH A SHARED FOUNDATION MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Adapting a large target model to private client data typically requires the client to receive and optimize that model. We study a target-decoupled setting where the client instead has access to a smaller shared reference model, while the server holds the target model and a PUBLIC repository. We introduce GRADHOP, which uses the private data to form learning directions under the reference model and communicates them once to the server. The server uses these directions to retrieve client-relevant examples from the PUBLIC repository and then adapts the target model directly on the selected examples. This avoids requiring parameter-wise correspondence between the reference and target models. We analyze this cross-model re-grounding and provide conditions under which reference-space retrieval identifies PUBLIC examples whose target gradient approximates the inaccessible private target gradient. Experiments on Banking77 and CLINC150 show improved target adaptation over the unadapted model and alternatives that use only PUBLIC data, while the target-space analysis confirms stronger alignment with the private target gradient. The same reference-side information also supports adaptation across different target families without exposing the target model to the client.

1 INTRODUCTION

Federated learning (FL) enables learning from distributed datasets without centralizing raw data (McMahan et al., 2017). In standard FL, clients keep their data local but participate in learning by receiving and optimizing a common model. This design decouples the *data* from the server, but not the *model* from the clients. For large foundation models, this model-side requirement can be restrictive: A data owner may not have sufficient memory or computation to host and differentiate through the server's target model, or may not have access to the target model at all. This raises a natural question: *Can private client data be used to adapt a target model without requiring the client to receive or optimize that model?*

We present GRADHOP, a target-decoupled adaptation framework for this setting. A data owner holds private examples and has access to a relatively small shared reference model, while the server holds the target model and a PUBLIC repository. The PUBLIC repository is not assumed to follow the client's data distribution. The client uses its private data and the reference model to form a small set of learning directions and communicates them to the server. The server uses these directions to identify client-relevant examples from the PUBLIC repository and adapts the target model directly on the resulting PUBLIC support. The client therefore never receives or optimizes the target model, and raw private examples remain local. Reference-space information determines which PUBLIC examples are used for adaptation, while optimization is performed entirely in the target model space. This design does not require a parameter-wise correspondence between the reference and target models, and their parameterizations and dimensions may differ. The client communicates once, after which the server can optimize the target without further access to the private data.

Several existing approaches address related operational constraints. One-shot federated learning reduces repeated client participation (Li et al., 2021; Jhunjhunwala et al., 2024; Liu et al., 2025), and recent methods extend one-shot adaptation to large pretrained models (Xu et al., 2025; Zhang et al., 2026). Communication-efficient federated fine-tuning compresses information computed with respect to the target model, for example through seeded zeroth-order directions or low-dimensional projections (Qin et al., 2024; Shu et al., 2025). Model-heterogeneous and proxy-based methods in-

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.

stead exchange predictions, prototypes, distilled models, or target-derived emulators (Li & Wang, 2019; Lin et al., 2020; Tan et al., 2022; Xiao et al., 2023; Fan et al., 2026). These approaches reduce repeated communication, communication volume, or the cost of client-side adaptation. GRADHOP instead considers adaptation without client access to the target model or its gradient representation.

The central analytical question is whether learning information expressed under the reference model remains effective after it is re-grounded through PUBLIC examples under the target model. To study this question, we view each PUBLIC example as a paired learning anchor, with one representation in the reference gradient space and another in the target gradient space. Our analysis characterizes when reference-space learning directions identify PUBLIC support that approximates a client-relevant target-space direction. The resulting conditions do not require a parameter-wise correspondence between the reference and target models or equal dimensions of their gradient spaces, nor do they require a global correspondence between their example rankings.

In the experiments, we examine the potential of GRADHOP for language-model adaptation across reference-model scales, target-model families, and datasets. We compare GRADHOP with the unadapted target, PUBLIC-only adaptation, and baseline methods that operate under less restrictive client-side assumptions. Our results demonstrate that GRADHOP can improve a larger target model using learning information derived from a smaller reference model without requiring the client to host or optimize the target itself.

2 RELATED WORK

Model-asymmetric and heterogeneous adaptation. Model heterogeneity has motivated federated methods that exchange information other than directly aggregatable model parameters. FedMD communicates predictions among independently designed client models (Li & Wang, 2019), FedDF distills an ensemble of client models into a server model using unlabeled data (Lin et al., 2020), and FedProto communicates class prototypes to accommodate heterogeneous client architectures (Tan et al., 2022). Related work on foundation models considers settings where the data owner cannot efficiently use the full server model. Offsite-Tuning provides the data owner with a lightweight adapter and a lossy emulator derived from the target model (Xiao et al., 2023). FedProxy instead compresses a proprietary LLM into a smaller proxy model, adapts this proxy federatively, and transfers the learned information back to the larger model (Fan et al., 2026). In GRADHOP, the shared reference model need not be derived from or parameter-compatible with the server target. PUBLIC examples instead connect learning directions formed under the reference model to adaptation of the independently held target model.

One-shot and communication-efficient federated learning. One-shot federated learning reduces repeated client participation by restricting client-server interaction to a single communication round (Li et al., 2021; Liu et al., 2025). Existing approaches transfer locally learned knowledge or combine locally trained models using quantities such as Fisher information (Jhunjhunwala et al., 2024), and recent work considers one-shot adaptation of large pretrained or multimodal models (Xu et al., 2025; Zhang et al., 2026). A related direction reduces communication or client cost when adapting language models. FedKSeed represents zeroth-order directions for the target model through shared random seeds and scalar directional derivatives (Qin et al., 2024), while Ferret projects first-order target updates into a low-dimensional space and reconstructs them at the server (Shu et al., 2025). These approaches reduce the cost of communicating information computed with respect to the target model. GRADHOP considers a complementary setting in which clients do not compute updates with respect to the target model.

Gradient-based data attribution and selection. Gradient representations have also been used for data attribution and data selection. TRAK uses projected gradient features for scalable data attribution, tracing model behavior back to training examples (Park et al., 2023), while DataInf estimates the influence of individual training examples in parameter-efficiently tuned generative models (Kwon et al., 2024). LESS uses low-dimensional, optimizer-aware gradient similarity to select instruction-tuning examples that are relevant to a desired capability (Xia et al., 2024). In contrast, GRADHOP uses learning directions formed under a reference model to identify PUBLIC anchors for adaptation of an independently held target model. The reference and target models need not share a parameterization.

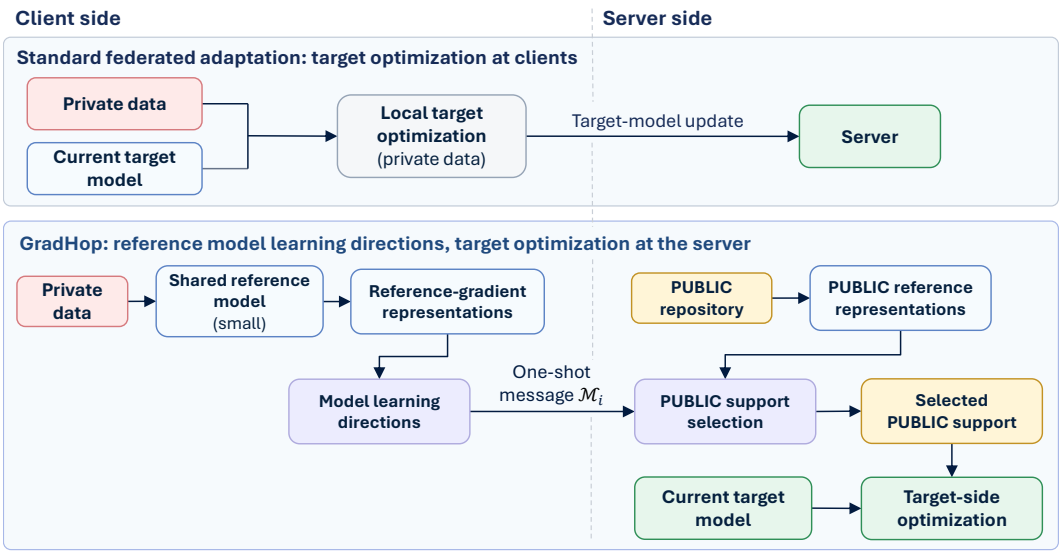

Figure 1: Overview of GRADHOP. The client uses its private data and the shared reference model to form an aggregate learning direction and several mode directions. The server uses these directions to select a PUBLIC support and adapts the independently held target model directly on the selected PUBLIC examples. The client does not receive or differentiate through the target model.

PUBLIC surrogates and data distillation. Another related direction communicates a surrogate for a client's local data or objective. FedDM constructs compact synthetic datasets by aligning local distributions so that the server can learn from transmitted distilled data (Xiong et al., 2023). DistDD aggregates locally distilled data into a reusable global surrogate using gradient-based objectives (Wang & Wang, 2024). GRADHOP takes a different approach. The client neither synthesizes examples nor attempts to reconstruct its private dataset at the server. Instead, the communicated learning directions are used to identify anchors from an existing PUBLIC repository, and the server adapts the target model directly on the selected PUBLIC support. The PUBLIC examples therefore serve as reusable candidate anchors rather than as a distilled surrogate of the client's private data.

3 TARGET-DECOUPLED ADAPTATION ALGORITHM

GRADHOP operates in two stages. The client uses its private data and the reference model to form a small set of learning directions derived from gradients. The server uses these directions to select PUBLIC examples aligned with them and adapts the target model directly on the selected support. Figure 1 illustrates the overall procedure.

3.1 PROBLEM SETUP

We consider a data owner (client) i with private dataset $\mathcal{D}_i$ and a server that holds a target model $F_{\boldsymbol{\theta}}$. The server also has access to a PUBLIC repository $\mathcal{P} = \{z_1, \ldots, z_M\}$ containing M examples, and both sides have access to a relatively small reference model $R_{\boldsymbol{\phi}}$. The PUBLIC repository may follow a distribution different from that of the private client data. The target and reference models are parameterized by $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, respectively. These vectors may represent the full model parameters or selected trainable subsets, and their dimensions need not be equal. We denote the losses for individual examples of the reference and target models by $\ell_R(z, \boldsymbol{\phi})$ and $\ell_T(z, \boldsymbol{\theta})$, respectively. The client may evaluate and differentiate through the reference model, but it does not receive, evaluate, or optimize the target model.

Our objective is to convey information relevant to adapting the target to the client's private data through a single message from the client to the server, without transferring the raw private examples or requiring access to the target model at the client. We write the client message as

$$\mathcal{M}_i = \Psi_R(\mathcal{D}_i, \boldsymbol{\phi}),$$

where Ψ_R uses only the private data and the reference model. This interface is summarized as

$$\mathcal{D}_i \xrightarrow{R_{\boldsymbol{\phi}}} \mathcal{M}_i, \quad (\mathcal{M}_i, \mathcal{P}, F_{\boldsymbol{\theta}}) \xrightarrow{\text{GradHop}} \boldsymbol{\theta}_i^+,$$

where $\boldsymbol{\theta}_i^+$ is the adapted target parameter vector. The message is computed entirely using the reference model and contains neither a gradient nor an update of the target model. The PUBLIC repository supplies the examples through which this information is re-grounded from the reference model to the target.

3.2 Client learning directions

For a private or PUBLIC example z, let

$$\boldsymbol{r}(z) = \boldsymbol{S}_R \nabla_{\boldsymbol{\phi}} \ell_R(z, \boldsymbol{\phi}) \in \mathbb{R}^{d_R} \tag{1}$$

be its reference gradient representation. Here, $\boldsymbol{S}_R$ is a fixed deterministic linear map applied to the gradient of the reference model, yielding a representation in $\mathbb{R}^{d_R}$. The same map is used for all private and PUBLIC examples. These representations are used to construct the aggregate and mode learning directions communicated by the client.

The aggregate learning direction of client i is

$$\boldsymbol{q}_i^R = \frac{1}{|\mathcal{D}_i|} \sum_{z \in \mathcal{D}_i} \boldsymbol{r}(z), \tag{2}$$

which averages the reference gradient representations of the private examples. This direction is used for retrieval rather than as a model update. However, averaging can obscure distinct directional structure among the private examples. We therefore supplement it with K mode directions.

For each private example with $\boldsymbol{r}(z) \neq \mathbf{0}$, we normalize its representation as $\boldsymbol{u}_R(z) = \frac{\boldsymbol{r}(z)}{\|\boldsymbol{r}(z)\|}$ and partition these normalized representations using deterministic spherical k-means, yielding clusters $\{C_{i1}, \ldots, C_{iK}\}$. Each cluster defines a mode direction:

$$\boldsymbol{q}_{ik}^R = \frac{1}{|C_{ik}|} \sum_{z \in C_{ik}} \boldsymbol{r}(z), \quad w_{ik} = \frac{|C_{ik}|}{|\mathcal{D}_i|}. \tag{3}$$

The client communicates

$$\mathcal{M}_i = \{(w_{ik}, \boldsymbol{q}_{ik}^R)\}_{k=1}^K.$$

The aggregate direction can be recovered from the same message as

$$\boldsymbol{q}_i^R = \sum_{k=1}^K w_{ik} \boldsymbol{q}_{ik}^R,$$

since examples with zero reference gradient representation do not contribute to the aggregate. Thus, a single client message carries both the aggregate and mode information.

3.3 Construction of the PUBLIC support

Given the client learning directions, the server identifies a small set of PUBLIC examples, called the *PUBLIC support*, that are relevant to the client's adaptation need. The server precomputes $\boldsymbol{r}(z_j)$ for each PUBLIC example z_j using the same reference model and map $\boldsymbol{S}_R$. PUBLIC examples with zero reference gradient representation are excluded from directional retrieval. For a nonzero aggregate learning direction, the cosine score is

$$s_{ij}^{\text{glob}} = \frac{\langle \boldsymbol{q}_i^R, \boldsymbol{r}(z_j) \rangle}{\|\boldsymbol{q}_i^R\| \, \|\boldsymbol{r}(z_j)\|}, \tag{4}$$

which induces an aggregate ranking π_i^{glob} of the PUBLIC repository. Each nonzero mode direction $\boldsymbol{q}_{ik}^R$ similarly induces a ranking π_{ik}. These mode rankings allow PUBLIC examples associated with distinct directional modes of the private data to contribute to the selected support. We combine the

mode rankings into a deterministic candidate list π_i^{mode}. The weights w_{ik} determine the relative quota of each mode in this list, and each PUBLIC example is retained at most once.

The aggregate ranking and the mode candidate list have complementary roles. The former emphasizes examples aligned with the client's overall learning direction, while the latter preserves examples associated with distinct directional modes. Let B be the total support budget and $B_g \leq B$ the portion reserved for the aggregate ranking. The server constructs the selected PUBLIC support $\widehat{S}_i$ by first taking the top B_g examples from π_i^{glob} and then adding the highest ranked examples not already selected from π_i^{mode} until B examples are obtained. If the mode candidate list does not fill the remaining positions, the server continues through π_i^{glob}, excluding examples already selected.

3.4 TARGET ADAPTATION

After constructing $\widehat{S}_i$, the learning directions are no longer used. The server adapts the target model directly on the selected PUBLIC examples by minimizing

$$L_{\widehat{S}_i}(\boldsymbol{\theta}) = \frac{1}{|\widehat{S}_i|} \sum_{j \in \widehat{S}_i} \ell_T(z_j, \boldsymbol{\theta}), \tag{5}$$

yielding the adapted target parameter vector $\boldsymbol{\theta}_i^+$. The client does not participate in this optimization and never supplies a gradient of the target model. GRADHOP therefore neither transforms a learning direction from the reference model into an update of the target model nor requires a correspondence between the parameters of the reference and target models. Instead, PUBLIC examples serve as anchors between the two models: They are selected using learning directions formed under the reference model and optimized directly under the target model. The complete procedure is summarized in Algorithm 1.

4 ANALYSIS OF CROSS-MODEL RE-GROUNDING

GRADHOP uses learning directions from the reference model to select PUBLIC support, while adaptation is performed under the target model. We analyze when the target gradient induced by this support approximates the inaccessible private target gradient by asking whether 1) suitable support exists in the PUBLIC repository and 2) retrieval under the reference model can identify it. The complete proofs are provided in Section B.

4.1 PUBLIC SUPPORT AND TARGET COVERAGE

Throughout this section, normalization is applied only to nonzero reference vectors, and PUBLIC indices refer to examples with $\boldsymbol{r}(z_j) \neq \mathbf{0}$. We assume $\boldsymbol{q}_i^R \neq \mathbf{0}$ and that at least B such PUBLIC examples are available. Statements involving the separation margin additionally assume that at least one eligible PUBLIC example lies outside the oracle support defined below. For each PUBLIC example z_j, we associate its normalized reference representation with its target gradient:

$$\boldsymbol{r}_j = \frac{\boldsymbol{r}(z_j)}{\|\boldsymbol{r}(z_j)\|} \in \mathbb{R}^{d_R}, \quad \boldsymbol{t}_j(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \ell_T(z_j, \boldsymbol{\theta}) \in \mathbb{R}^{d_T}. \tag{6}$$

Here, d_T is the dimension of the target parameter vector used for optimization. The two representations may have different dimensions and are linked only through the shared PUBLIC example z_j. For retrieval under the reference model, the normalized aggregate direction is $\boldsymbol{u}_i^R = \boldsymbol{q}_i^R / \|\boldsymbol{q}_i^R\|$, and PUBLIC example j receives cosine score $s_{ij} = \langle \boldsymbol{u}_i^R, \boldsymbol{r}_j \rangle$.

To assess the quality of the selected support for target adaptation, consider the private target objective

$$F_i(\boldsymbol{\theta}) = \frac{1}{|\mathcal{D}_i|} \sum_{z \in \mathcal{D}_i} \ell_T(z, \boldsymbol{\theta}), \quad \boldsymbol{g}_i^{\mathrm{tar}}(\boldsymbol{\theta}) = \nabla F_i(\boldsymbol{\theta}). \tag{7}$$

The gradient $\boldsymbol{g}_i^{\mathrm{tar}}$ is not available to GRADHOP since the target model is never evaluated on the private data. For any PUBLIC support S with $|S| = B$, its average target gradient is

$$\boldsymbol{g}_S^{\mathrm{tar}}(\boldsymbol{\theta}) = \frac{1}{B} \sum_{j \in S} \boldsymbol{t}_j(\boldsymbol{\theta}). \tag{8}$$

At the current target parameters, let S_i^* be an arbitrary fixed PUBLIC support of size B whose average target gradient is closest to the private target gradient:

$$S_i^* \in \operatorname*{arg\,min}_{S:\,|S|=B} \left\| \boldsymbol{g}_i^{\mathrm{tar}}(\boldsymbol{\theta}) - \boldsymbol{g}_S^{\mathrm{tar}}(\boldsymbol{\theta}) \right\|. \tag{9}$$

The corresponding coverage error

$$\epsilon_i^{\mathrm{cov}}(\boldsymbol{\theta}) = \|\boldsymbol{g}_i^{\mathrm{tar}}(\boldsymbol{\theta}) - \boldsymbol{g}_{S_i^*}^{\mathrm{tar}}(\boldsymbol{\theta})\| \tag{10}$$

measures how well the PUBLIC repository can represent the private target gradient using an equal-weight support of exactly B examples. Under the reference direction $\boldsymbol{q}_i^R$, the separation margin of S_i^* is

$$\Gamma_i^R = \min_{j \in S_i^*} \left\langle \boldsymbol{u}_i^R, \boldsymbol{r}_j \right\rangle - \max_{j \notin S_i^*} \left\langle \boldsymbol{u}_i^R, \boldsymbol{r}_j \right\rangle. \tag{11}$$

Since $|S_i^*| = B$, a positive Γ_i^R means that every example in S_i^* scores above every example outside it, so top-B retrieval selects exactly S_i^*. A positive Γ_i^R also quantifies the separation available to tolerate perturbations in the reference learning direction. The coverage error captures a limitation of the PUBLIC repository under the size-B, equal-weight surrogate class considered here. If $\epsilon_i^{\mathrm{cov}}$ is large, even the best surrogate in this class cannot approximate the private target gradient closely. A simple counterexample is given in Section B.1.

4.2 Retrieval stability and target approximation

Let $\mathcal{R}_B(\boldsymbol{q})$ be the B PUBLIC examples with the largest cosine scores with respect to a reference learning direction $\boldsymbol{q}$. The margin also characterizes the robustness of aggregate retrieval to changes in the reference learning direction. Consider a perturbed direction $\widetilde{\boldsymbol{q}}_i^R$ and define

$$\widetilde{\boldsymbol{u}}_i^R = \frac{\widetilde{\boldsymbol{q}}_i^R}{\|\widetilde{\boldsymbol{q}}_i^R\|}, \quad \eta_i = \left\| \widetilde{\boldsymbol{u}}_i^R - \boldsymbol{u}_i^R \right\|. \tag{12}$$

By the Cauchy-Schwarz inequality, the cosine score of each PUBLIC example changes by at most η_i, so the separation between S_i^* and its complement decreases by at most $2\eta_i$. Hence, if

$$\eta_i < \frac{\Gamma_i^R}{2}, \tag{13}$$

then $\mathcal{R}_B(\widetilde{\boldsymbol{q}}_i^R) = S_i^*$. This exact selection is sufficient but not necessary for effective re-grounding.

For the final PUBLIC support $\widehat{S}_i$ selected by GradHop, its mismatch with S_i^* is

$$m_i = |S_i^* \setminus \widehat{S}_i| = |\widehat{S}_i \setminus S_i^*|. \tag{14}$$

Define the maximum PUBLIC target-gradient norm as

$$B_T(\boldsymbol{\theta}) = \max_j \|\boldsymbol{t}_j(\boldsymbol{\theta})\|. \tag{15}$$

Proposition 1 (Cross-model re-grounding error)**.** *For the final support $\widehat{S}_i$ of size B,*

$$\|\boldsymbol{g}_i^{\mathrm{tar}}(\boldsymbol{\theta}) - \boldsymbol{g}_{\widehat{S}_i}^{\mathrm{tar}}(\boldsymbol{\theta})\| \le \epsilon_i^{\mathrm{cov}}(\boldsymbol{\theta}) + \frac{2m_i}{B} B_T(\boldsymbol{\theta}). \tag{16}$$

The bound decomposes the target approximation error into repository coverage and selection mismatch. The first term is the best achievable error with a PUBLIC support of size B, while the second accounts for selecting $\widehat{S}_i$ instead of S_i^*. Neither requires a correspondence between the parameter coordinates of the reference and target models.

For simplicity of notation, we suppress the common dependence on $\boldsymbol{\theta}$ and define

$$\widehat{\boldsymbol{g}}_i^{\mathrm{tar}} = \boldsymbol{g}_{\widehat{S}_i}^{\mathrm{tar}}(\boldsymbol{\theta}), \quad E_i = \epsilon_i^{\mathrm{cov}}(\boldsymbol{\theta}) + \frac{2m_i}{B} B_T(\boldsymbol{\theta}). \tag{17}$$

Then proposition 1 implies $\|\widehat{\boldsymbol{g}}_i^{\mathrm{tar}} - \boldsymbol{g}_i^{\mathrm{tar}}\| \le E_i$. We next assess the alignment between the selected PUBLIC surrogate $\widehat{\boldsymbol{g}}_i^{\mathrm{tar}}$ and the private target gradient $\boldsymbol{g}_i^{\mathrm{tar}}$.

Table 1: Target-space re-grounding on Banking77. Cosine similarity and relative error (mean $\pm$ standard deviation) compare the target gradient induced by the selected PUBLIC examples with the private target gradient at the same initialization. Positive cosine is the percentage of evaluations across runs and tasks with cosine similarity above zero.

SUPPORT	COSINE ↑	REL. ERROR ↓	POSITIVE COSINE (%)
Random PUBLIC-128	0.157 ± 0.014	1.519 ± 0.008	72.5
Balanced PUBLIC-128	0.258 ± 0.010	1.294 ± 0.004	85.0
GRADHOP	0.636 ± 0.001	1.075 ± 0.010	100.0

Corollary 1 (Directional alignment)**.** *If* $E_i < \|\boldsymbol{g}_i^{\mathrm{tar}}\|$*, then* $\widehat{\boldsymbol{g}}_i^{\mathrm{tar}} \neq \mathbf{0}$ *and*

$$\cos\big(\boldsymbol{g}_i^{\mathrm{tar}}, \widehat{\boldsymbol{g}}_i^{\mathrm{tar}}\big) \geq \sqrt{1 - \left(\frac{E_i}{\|\boldsymbol{g}_i^{\mathrm{tar}}\|}\right)^2} > 0. \tag{18}$$

Therefore, sufficiently small re-grounding error ensures positive alignment with the private target gradient. Consequently, $-\widehat{\boldsymbol{g}}_i^{\mathrm{tar}}$ is a descent direction for the private objective, and a sufficiently small step along this direction decreases F_i. The bound in Eq. (18) is tight under the stated relative error condition.

The perturbation result in Eq. (13) characterizes the stability of a single cosine ranking. In contrast, proposition 1 and corollary 1 depend only on the final support $\widehat{S}_i$ and apply directly to the hybrid support construction in Section 3. The coverage error is independent of the retrieval procedure, while multiple directions may improve the final support by reducing the selection mismatch m_i.

Since target adaptation may include multiple optimization steps, its final displacement need not coincide with a single step along $\widehat{\boldsymbol{g}}_i^{\mathrm{tar}}$. Section B.3 gives a sufficient condition relating this final displacement to improvement of the private objective.

5 EXPERIMENTS

Experimental setup. For the main experiments, we use Qwen2.5-1.5B-Instruct as the reference model and Qwen2.5-7B-Instruct as the target. Both use rank-4 LoRA with scaling parameter $\alpha = 8$ on the q_{proj} and v_{proj} modules, and reference gradients are represented in $d_R = 32{,}768$ dimensions. We use Banking77 and CLINC150 for evaluation. For each dataset, we construct eight client tasks. For the main evaluation, we consider these tasks independently, with one private client participating in each adaptation run. We also study the extension to multiple clients in Section 5.3. Each task contains eight focus intents and 24 background intents, for a total of 32 allowed labels. The client's private data are restricted to the eight focus intents, with 32 private training examples per intent, while the server's PUBLIC repository spans all 32 task labels. GRADHOP uses $K = 4$ mode-specific directions and a PUBLIC support budget of $B = 128$, with $B_g = 96$ examples retained from the aggregate ranking and up to 32 additions from the mode rankings. Target adaptation uses AdamW for four epochs with learning rate 10^{-3}, zero weight decay, gradient clipping at 1.0, and default Adam coefficients.

We evaluate performance using exact-match accuracy for generated labels and negative log-likelihood (NLL). The main comparisons examine whether conditioning the PUBLIC selection on client information improves target adaptation. We compare with BASE, which denotes the unadapted target, Random PUBLIC-128, which samples 128 examples from the task PUBLIC repository independently of the client message, Balanced PUBLIC-128, which allocates the same budget approximately uniformly across the 32 allowed labels, and All PUBLIC, which adapts the target using the entire task PUBLIC repository without client conditioning. Random PUBLIC-128 and Balanced PUBLIC-128 use the same support budget as GRADHOP, while All PUBLIC tests whether adaptation on the complete PUBLIC repository is sufficient. All adaptation methods use the same target-side optimizer. We repeat the main adaptation experiments with five different target initializations and report the mean and standard deviation. For each initialization, results are averaged over the eight client tasks. Complete experimental details are provided in Section C.

Table 2: Results of target-model adaptation to client-specific data on Banking77 and CLINC150, evaluated using exact-match accuracy (%) and negative log-likelihood (NLL).

	Banking77		CLINC150	
ALGORITHM	**ACCURACY (%) ↑**	**NLL ↓**	**ACCURACY (%) ↑**	**NLL ↓**
BASE	73.4 ± 0.0	0.557 ± 0.000	80.9 ± 0.0	0.188 ± 0.000
Random PUBLIC-128	74.0 ± 1.5	0.342 ± 0.018	74.3 ± 2.7	0.117 ± 0.021
Balanced PUBLIC-128	78.0 ± 0.9	0.180 ± 0.007	68.1 ± 3.1	0.135 ± 0.027
All PUBLIC	21.2 ± 5.1	1.708 ± 0.293	33.5 ± 2.5	0.370 ± 0.078
GRADHOP	84.9 ± 0.6	0.149 ± 0.010	86.2 ± 0.4	0.114 ± 0.015

Table 3: Ablation of mode-specific directions on Banking77.

CONFIGURATION	**ACCURACY (%) ↑**	**NLL ↓**
Single aggregate	82.8 ± 0.3	0.200 ± 0.017
Full GRADHOP	84.9 ± 0.6	0.149 ± 0.010

5.1 TARGET-SPACE RE-GROUNDING

Section 4 analyzed when PUBLIC examples selected using learning directions formed under the reference model can induce a target gradient that approximates the private target gradient. Here, we empirically examine this relationship by comparing the target gradient induced by the selected PUBLIC examples with the private target gradient at the same initialization. We measure cosine similarity and relative error between these gradients. Unlike the oracle coverage error $\epsilon_i^{\mathrm{cov}}$ in Eq. (10), these measurements are computed for the support actually selected by each method.

Table 1 shows that the PUBLIC examples selected by GRADHOP induce target gradients that are substantially better aligned with the private target gradient than those selected without client information. In particular, the mean cosine increases from 0.258 for Balanced PUBLIC-128 to 0.636 for GRADHOP, while the relative error is also lower than for both PUBLIC baselines. A relative error greater than one does not preclude effective directional alignment, since relative error depends on gradient magnitude as well as direction. Therefore, these results should be interpreted in terms of directional alignment rather than as verification of the sufficient condition in corollary 1.

5.2 MAIN RESULTS

Table 2 summarizes the results. On Banking77, GRADHOP improves average accuracy from 73.4% for the unadapted target to 84.9%, while reducing NLL from 0.557 to 0.149. It also outperforms both Random PUBLIC-128 and Balanced PUBLIC-128, demonstrating the benefit of conditioning PUBLIC selection on client information rather than using client-independent support selection. CLINC150 shows a similar pattern, with GRADHOP achieving the highest mean accuracy and the lowest mean NLL among the listed methods. All PUBLIC performs substantially worse on both datasets despite using the complete PUBLIC repository, indicating that increasing the amount of PUBLIC adaptation data alone is not sufficient.

These comparisons separate PUBLIC data availability from client-conditioned selection. Random PUBLIC-128 and Balanced PUBLIC-128 use the same 128-example support budget as GRADHOP, while All PUBLIC removes the support-budget constraint but does not use the client message.

Ablation of mode-specific directions. We ablate the mode-specific directions on Banking77 by comparing full GRADHOP with a single-aggregate configuration that selects the entire PUBLIC support using only the aggregate direction (Table 3). The full construction increases average accuracy from 82.8% to 84.9% and reduces NLL from 0.200 to 0.149, showing that the mode-specific directions provide an additional benefit beyond the single aggregate direction.

Transfer across target-model families. GRADHOP does not require the reference and target gradient spaces to share coordinates or dimensions. We use the same Qwen2.5-1.5B reference model, client messages, and Banking77 PUBLIC supports across target families while replacing the

Table 4: Cross-model transfer on Banking77. The Qwen2.5-1.5B reference model and selected PUBLIC supports are shared across target families. PUBLIC refers to Balanced PUBLIC-128.

	ACCURACY (%)			NLL		
TARGET	BASE	PUBLIC	GRADHOP	BASE	PUBLIC	GRADHOP
Mistral-7B	59.1 ± 0.0	78.5 ± 2.4	82.7 ± 2.3	0.591 ± 0.000	0.197 ± 0.023	0.163 ± 0.022
Llama-3.1-8B	51.6 ± 0.0	64.5 ± 3.2	63.9 ± 3.5	0.679 ± 0.000	0.215 ± 0.030	0.188 ± 0.021
OLMo-2-7B	55.6 ± 0.0	73.6 ± 1.2	78.4 ± 1.4	0.573 ± 0.000	0.188 ± 0.006	0.175 ± 0.021

Table 5: Multiple-client extension on Banking77.

	GRADHOP		BALANCED PUBLIC	
C	ACCURACY (%) ↑	NLL ↓	ACCURACY (%) ↑	NLL ↓
1	84.3 ± 1.2	0.162 ± 0.011	76.5 ± 1.5	0.189 ± 0.010
2	85.8 ± 0.4	0.121 ± 0.013	77.7 ± 1.4	0.181 ± 0.015
4	86.9 ± 1.1	0.107 ± 0.006	72.7 ± 3.7	0.298 ± 0.030
8	86.0 ± 1.1	0.106 ± 0.007	34.7 ± 12.0	1.206 ± 0.618

server target with Mistral-7B-Instruct-v0.3, Llama-3.1-8B-Instruct, or OLMo-2-1124-7B-Instruct. For each target family, the target-side learning rate is selected using disjoint PUBLIC validation data and is then used for both Balanced PUBLIC-128 and GRADHOP.

Table 4 shows the results. GRADHOP improves substantially over the unadapted target for all three target families. It achieves higher mean accuracy than Balanced PUBLIC-128 for Mistral and OLMo, while Balanced PUBLIC-128 has higher mean accuracy for Llama. GRADHOP has lower mean NLL than Balanced PUBLIC-128 for all three families. These experiments test cross-model re-grounding directly: Client information is formed in the reference gradient space, while the selected PUBLIC examples are optimized under independently parameterized target models. No coordinate-wise map between the reference and target gradients is used. The effect of reducing the client reference model to 0.5B is reported in Section D.2.

5.3 MULTIPLE-CLIENT EXTENSION

We further examine whether GRADHOP extends to multiple clients (Table 5). Each client is re-grounded independently, and the server combines the resulting surrogate losses using equal client weights. Details of the surrogate loss are provided in Section C. We evaluate $C \in \{1, 2, 4, 8\}$ without directly averaging the learning directions communicated by different clients. The accuracy of GRADHOP remains stable as the number of clients increases and is higher on average than that of Balanced PUBLIC for every evaluated C.

6 CONCLUSION

We introduced GRADHOP for target-decoupled model adaptation, where private client data guide a server-held target model through learning directions formed with a shared reference model. Rather than transferring these directions directly between model parameter spaces, GRADHOP uses them to select PUBLIC examples and performs adaptation entirely under the target model. Our analysis relates successful re-grounding to the availability of suitable PUBLIC examples and to their retrieval under the reference model. Across Banking77 and CLINC150, GRADHOP improves target adaptation while the target-space analysis shows that the selected PUBLIC examples better approximate the private target direction. The same reference-side information also supports adaptation across multiple target-model families without requiring a coordinate-wise correspondence between their gradient spaces, and independently re-grounded client objectives can be combined in the multiple-client extension. The approach depends on the PUBLIC repository containing examples that can represent the client-relevant target direction, and the communicated learning directions do not by themselves provide a formal privacy guarantee. Extending the setting to broader tasks, modalities, and stronger privacy mechanisms provides natural directions for future work.

AI USE STATEMENT

In this work, we used generative AI tools to improve the text, check mathematical proofs, and supplement the literature review. All AI-assisted work was independently checked and revised by the authors. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Tao Fan, Guoqiang Ma, Yuanfeng Song, Lixin Fan, Kai Chen, and Qiang Yang. FedProxy: Federated fine-tuning of llms via proxy slms and heterogeneity-aware fusion. In *ACL*, 2026.

Divyansh Jhunjhunwala, Shiqiang Wang, and Gauri Joshi. FedFisher: Leveraging fisher information for one-shot federated learning. In *AISTATS*, 2024.

Yongchan Kwon, Eric Wu, Kevin Wu, and James Y. Zou. DataInf: Efficiently estimating data influence in LoRA-tuned LLMs and diffusion models. In *ICLR*, 2024.

Daliang Li and Junpu Wang. FedMD: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*, 2019. Also presented at the NeurIPS 2019 Workshop on Federated Learning.

Qinbin Li, Bingsheng He, and Dawn Song. Practical one-shot federated learning for cross-silo setting. In *IJCAI*, 2021.

Tao Lin, Lingjing Kong, Sebastian U. Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. In *NeurIPS*, pp. 2351–2363, 2020.

Xiang Liu, Zhenheng Tang, Xia Li, Yijun Song, Sijie Ji, Zemin Liu, Bo Han, Linshan Jiang, and Jialin Li. One-shot federated learning methods: A practical guide. In *IJCAI*, 2025.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, 2017.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK: Attributing model behavior at scale. In *ICML*, pp. 27074–27113, 2023.

Zhen Qin, Daoyuan Chen, Bingchen Qian, Bolin Ding, Yaliang Li, and Shuiguang Deng. Federated full-parameter tuning of billion-sized language models with communication cost under 18 kilobytes. In *ICML*, 2024.

Yao Shu, Wenyang Hu, See-Kiong Ng, Bryan Kian Hsiang Low, and Fei Yu. Ferret: Federated full-parameter tuning at scale for large language models. In *ICML*, 2025.

Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. FedProto: Federated prototype learning across heterogeneous clients. In *AAAI*, 2022.

Peiran Wang and Haohan Wang. DistDD: Distributed data distillation aggregation through gradient matching. *arXiv preprint arXiv:2410.08665*, 2024.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *ICML*, pp. 54104–54132, 2024.

Guangxuan Xiao, Ji Lin, and Song Han. Offsite-tuning: Transfer learning without full model. *arXiv preprint arXiv:2302.04870*, 2023.

Yuanhao Xiong, Ruochen Wang, Minhao Cheng, Felix Yu, and Cho-Jui Hsieh. FedDM: Iterative distribution matching for communication-efficient federated learning. In *CVPR*, 2023.

Binqian Xu, Haiyang Mei, Zechen Bai, Jinjin Gong, Rui Yan, Guo-Sen Xie, Yazhou Yao, Basura Fernando, and Xiangbo Shu. You only communicate once: One-shot federated low-rank adaptation of mllm. In *NeurIPS*, 2025.

Li Zhang, Zhongxuan Han, Xiaohua Feng, Jiaming Zhang, Yuyuan Li, Linbo Jiang, Jianan Lin, and Chaochao Chen. TOFA: Training-free one-shot federated adaptation for vision-language models. In *AAAI*, 2026.

Algorithm 1 Target-decoupled adaptation with GRADHOP. For simplicity, handling of degenerate zero-vector cases is omitted. Cosine retrieval uses only nonzero reference representations and nonzero learning directions.

Require: Private client data $\mathcal{D}_i$, PUBLIC repository $\mathcal{P}$, reference model $R_{\boldsymbol{\phi}}$, target model $F_{\boldsymbol{\theta}}$, representation map $\boldsymbol{S}_R$, number of modes K, support budget B, aggregate budget B_g, target optimizer $\mathcal{A}_T$

Compute reference representations $\boldsymbol{r}(z) = \boldsymbol{S}_R \nabla_{\boldsymbol{\phi}} \ell_R(z, \boldsymbol{\phi})$ for $z \in \mathcal{D}_i$
Normalize the nonzero representations and partition them into K spherical-k-means clusters $C_{i1}, \dots, C_{iK}$
for $k = 1, \dots, K$ **do**
    $\boldsymbol{q}_{ik}^R \leftarrow |C_{ik}|^{-1} \sum_{z \in C_{ik}} \boldsymbol{r}(z), \quad w_{ik} \leftarrow |C_{ik}|/|\mathcal{D}_i|$
end for
Client sends $\mathcal{M}_i = \{(w_{ik}, \boldsymbol{q}_{ik}^R)\}_{k=1}^K$ to the server
Recover $\boldsymbol{q}_i^R \leftarrow \sum_{k=1}^K w_{ik} \boldsymbol{q}_{ik}^R$
Compute or retrieve reference representations for the PUBLIC repository
Form the aggregate ranking π_i^{glob} and mode rankings $\{\pi_{ik}\}_{k=1}^K$ by cosine similarity
Combine the mode rankings into a deterministic weighted candidate list π_i^{mode}, removing duplicates
Initialize $\widehat{S}_i$ with the top B_g examples from π_i^{glob}
Add the highest ranked examples from π_i^{mode} not already selected until $|\widehat{S}_i| = B$, using π_i^{glob} as fallback if needed
$\boldsymbol{\theta}_i^+ \leftarrow \mathcal{A}_T(\boldsymbol{\theta}, \widehat{S}_i)$
return $\boldsymbol{\theta}_i^+$

A GRADHOP PROCEDURE

Algorithm 1 summarizes our target-decoupled adaptation procedure.

B ADDITIONAL ANALYSIS AND PROOFS FOR CROSS-MODEL RE-GROUNDING

The results in Section 4 analyze the additional cross-model inference step introduced by target decoupling. They require neither global convexity of neural network training nor equality between the reference and target parameter spaces.

We first illustrate why PUBLIC coverage is necessary. The subsequent analysis proceeds in three steps. First, we separate the target-space error of the selected PUBLIC support into a coverage term and an additional support-selection term. Then, we convert the resulting bound into an angular guarantee with respect to the private target gradient. Finally, since the target optimizer may take multiple update steps (rather than a single gradient step), we relate its final displacement to improvement of the private objective through a local smoothness argument.

B.1 PUBLIC COVERAGE IN TARGET SPACE

The PUBLIC repository must contain an appropriate surrogate for the private target gradient. Consider the one-dimensional case with $M = B = 1$ and $\theta = 0$, and take the single PUBLIC example to have a nonzero reference representation so that retrieval is well defined:

$$F_i(\theta) = \frac{1}{2}(\theta + 1)^2, \quad \ell_T(z_1, \theta) = \frac{1}{2}(\theta - 1)^2. \tag{19}$$

Here, $\boldsymbol{g}_i^{\mathrm{tar}}(0) = 1$, while the only PUBLIC gradient is $\boldsymbol{t}_1(0) = -1$. Retrieval is necessarily exact, yet the PUBLIC surrogate points in the opposite direction. A gradient step on the PUBLIC loss moves to $\theta^+ = \eta$ for $\eta > 0$, giving $F_i(\theta^+) = \frac{1}{2}(1 + \eta)^2 > F_i(0)$. Thus, perfect retrieval under the reference model cannot guarantee useful target adaptation when the PUBLIC repository lacks an appropriate target surrogate. This motivates the coverage error in Eq. (10).

B.2 TARGET-SPACE ERROR AND DIRECTIONAL ALIGNMENT

Proof of proposition 1. We compare the selected PUBLIC surrogate with the private target gradient by routing the comparison through the oracle support S_i^*. This separates two sources of error: How well a PUBLIC support of size B can represent the private target gradient, and how much is lost by selecting $\widehat{S}_i$ instead of the oracle support:

$$\begin{aligned}\|\boldsymbol{g}_i^{\mathrm{tar}} - \boldsymbol{g}_{\widehat{S}_i}^{\mathrm{tar}}\| &\le \|\boldsymbol{g}_i^{\mathrm{tar}} - \boldsymbol{g}_{S_i^*}^{\mathrm{tar}}\| + \|\boldsymbol{g}_{S_i^*}^{\mathrm{tar}} - \boldsymbol{g}_{\widehat{S}_i}^{\mathrm{tar}}\| \\ &= \epsilon_i^{\mathrm{cov}} + \|\boldsymbol{g}_{S_i^*}^{\mathrm{tar}} - \boldsymbol{g}_{\widehat{S}_i}^{\mathrm{tar}}\|.\end{aligned} \tag{20}$$

The remaining term depends only on anchors where the two supports disagree. All anchors in their intersection cancel, leaving

$$\boldsymbol{g}_{S_i^*}^{\mathrm{tar}} - \boldsymbol{g}_{\widehat{S}_i}^{\mathrm{tar}} = \frac{1}{B}\left(\sum_{j \in S_i^* \setminus \widehat{S}_i} \boldsymbol{t}_j - \sum_{j \in \widehat{S}_i \setminus S_i^*} \boldsymbol{t}_j\right). \tag{21}$$

Since the two supports have the same cardinality, each difference set contains m_i anchors. Each unmatched anchor can contribute at most B_T in norm, so the triangle inequality and Eq. (15) imply

$$\|\boldsymbol{g}_{S_i^*}^{\mathrm{tar}} - \boldsymbol{g}_{\widehat{S}_i}^{\mathrm{tar}}\| \le \frac{2m_i}{B} B_T. \tag{22}$$

Combining Eqs. (20) and (22) proves Eq. (16). □

Proof of corollary 1. Proposition 1 bounds the selected surrogate inside a ball of radius E_i around the private target gradient. We now translate this norm bound into a bound on their angle. Let

$$\boldsymbol{g} = \boldsymbol{g}_i^{\mathrm{tar}}, \quad \widehat{\boldsymbol{g}} = \widehat{\boldsymbol{g}}_i^{\mathrm{tar}}, \quad \delta = \frac{E_i}{\|\boldsymbol{g}\|} < 1.$$

By proposition 1,

$$\|\widehat{\boldsymbol{g}} - \boldsymbol{g}\| \le \delta \|\boldsymbol{g}\|. \tag{23}$$

Since $\delta < 1$, the reverse triangle inequality implies

$$\|\widehat{\boldsymbol{g}}\| \ge (1 - \delta)\|\boldsymbol{g}\| > 0,$$

hence the angle between the two gradients is well defined.

Let $r = \|\widehat{\boldsymbol{g}}\|/\|\boldsymbol{g}\| > 0$ and let α be the angle between $\boldsymbol{g}$ and $\widehat{\boldsymbol{g}}$. Since the squared distance between the two vectors can be written in terms of their norms and angle, squaring Eq. (23) and dividing by $\|\boldsymbol{g}\|^2$ gives

$$r^2 + 1 - 2r\cos\alpha \le \delta^2, \tag{24}$$

and hence

$$\begin{aligned}\cos\alpha &\ge \frac{r^2 + 1 - \delta^2}{2r} \\ &= \frac{1}{2}\left(r + \frac{1 - \delta^2}{r}\right) \\ &\ge \sqrt{1 - \delta^2},\end{aligned} \tag{25}$$

where the last inequality follows from the arithmetic mean–geometric mean inequality. Therefore, among all vectors satisfying Eq. (23), the least aligned admissible vector has cosine at least $\sqrt{1 - \delta^2}$. This proves Eq. (18). Equality is attained when $r = \sqrt{1 - \delta^2}$ and the error constraint is active, so the bound is tight under the relative error assumption alone. □

B.3 PRIVATE OBJECTIVE IMPROVEMENT

The target optimizer may take multiple steps, so its final displacement need not follow $\widehat{\boldsymbol{g}}_i^{\mathrm{tar}}$ exactly. Let

$$\boldsymbol{d}_i = \boldsymbol{\theta} - \boldsymbol{\theta}_i^+, \quad \text{so that} \quad \boldsymbol{\theta}_i^+ = \boldsymbol{\theta} - \boldsymbol{d}_i. \tag{26}$$

Corollary 2. *Suppose F_i has L-Lipschitz gradient along the line segment from $\boldsymbol{\theta}$ to $\boldsymbol{\theta} - \boldsymbol{d}_i$. If*

$$\langle \widehat{\boldsymbol{g}}_i^{\mathrm{tar}}, \boldsymbol{d}_i \rangle > E_i \|\boldsymbol{d}_i\| + \frac{L}{2}\|\boldsymbol{d}_i\|^2, \tag{27}$$

then

$$F_i(\boldsymbol{\theta} - \boldsymbol{d}_i) < F_i(\boldsymbol{\theta}). \tag{28}$$

The condition allows the final displacement produced by multiple target optimization steps to differ from the selected PUBLIC surrogate, provided its alignment with the surrogate is sufficient relative to the re-grounding error and local curvature.

Proof of corollary 2. Local smoothness controls how much the private objective can deviate from its first-order prediction along this displacement. The descent lemma gives

$$F_i(\boldsymbol{\theta} - \boldsymbol{d}_i) \le F_i(\boldsymbol{\theta}) - \langle \boldsymbol{g}_i^{\mathrm{tar}}, \boldsymbol{d}_i \rangle + \frac{L}{2}\|\boldsymbol{d}_i\|^2. \tag{29}$$

The private gradient in this bound is not available to the server, so we rewrite the inner product using the selected PUBLIC surrogate. By the Cauchy-Schwarz inequality and proposition 1,

$$\begin{aligned} -\langle \boldsymbol{g}_i^{\mathrm{tar}}, \boldsymbol{d}_i \rangle &= -\langle \widehat{\boldsymbol{g}}_i^{\mathrm{tar}}, \boldsymbol{d}_i \rangle - \langle \boldsymbol{g}_i^{\mathrm{tar}} - \widehat{\boldsymbol{g}}_i^{\mathrm{tar}}, \boldsymbol{d}_i \rangle & (30)\\ &\le -\langle \widehat{\boldsymbol{g}}_i^{\mathrm{tar}}, \boldsymbol{d}_i \rangle + E_i\|\boldsymbol{d}_i\|. & (31) \end{aligned}$$

Therefore,

$$F_i(\boldsymbol{\theta} - \boldsymbol{d}_i) \le F_i(\boldsymbol{\theta}) - \langle \widehat{\boldsymbol{g}}_i^{\mathrm{tar}}, \boldsymbol{d}_i \rangle + E_i\|\boldsymbol{d}_i\| + \frac{L}{2}\|\boldsymbol{d}_i\|^2. \tag{32}$$

By Eq. (27), the right-hand side is strictly smaller than $F_i(\boldsymbol{\theta})$, which proves Eq. (28). □

The three terms after $F_i(\boldsymbol{\theta})$ have a direct interpretation: The first is the descent predicted by the selected PUBLIC surrogate, the second is the worst-case penalty for re-grounding error, and the third is the local curvature penalty.

C Experimental details

Client tasks and PUBLIC repositories. For both Banking77 and CLINC150, each of the eight client tasks contains eight focus intents and 24 background intents. The client has 32 private training examples for each focus intent, giving 256 private examples in total. The resulting task contains 32 allowed labels. The server PUBLIC repository contains candidate examples drawn from these task labels. All methods are evaluated on the same task-specific test construction.

Models, optimization, and communication. The default client reference is Qwen2.5-1.5B-Instruct and the default target is Qwen2.5-7B-Instruct. Both use rank-4 LoRA with $\alpha = 8$ on q_{proj} and v_{proj}. In our experiments, $\boldsymbol{S}_R$ selects a fixed set of 32,768 coordinates from the flattened gradients of the trainable LoRA-B parameters, with the same coordinates used for all private and PUBLIC examples. With $K = 4$, the client communicates four directions and four weights, corresponding to approximately 524 KB in our implementation. This information is sent once. Raw private examples are never transmitted.

The default target-side optimizer is AdamW with learning rate 10^{-3} for four epochs, zero weight decay, gradient clipping at 1.0, coefficients $(0.9, 0.999)$, and $\epsilon = 10^{-8}$. The same target-side procedure is used for all 128-example supports within a target family. For Mistral, Llama, and OLMo, only the learning rate is recalibrated. We select it from $\{1.25 \times 10^{-4}, 2.5 \times 10^{-4}, 5 \times 10^{-4}, 10^{-3}\}$ using disjoint PUBLIC validation data, yielding 2.5×10^{-4} for Mistral and 10^{-3} for Llama and OLMo. The selected learning rate is then used for both Balanced PUBLIC-128 and GradHop, as described in Section 5.2.

Baselines. The primary baselines are defined in Section 5. Random PUBLIC-128 and Balanced PUBLIC-128 use the same 128-example support budget as GRADHOP, while All PUBLIC adapts the target using the complete task PUBLIC repository without the client message.

The ablation using a single aggregate direction sets $B_g = B = 128$, removing the mode-specific contribution while leaving the support budget and target-side optimization unchanged. The exact-private analysis adapts the target directly on private client examples and therefore does not satisfy the target-decoupled setting. It is included only as an analysis reference and should not be interpreted as an accuracy upper bound.

Multiple-client extension. For the experiment with C clients (Section 5.3), each client is re-grounded independently. If $\widehat{S}_c$ is the selected PUBLIC support for client c, the server forms

$$\widehat{F}_c(\boldsymbol{\theta}) = \frac{1}{|\widehat{S}_c|} \sum_{z \in \widehat{S}_c} \ell_T(z, \boldsymbol{\theta}) \tag{33}$$

and optimizes the weighted surrogate

$$\widehat{F}(\boldsymbol{\theta}) = \sum_{c=1}^{C} p_c \widehat{F}_c(\boldsymbol{\theta}), \quad \sum_{c=1}^{C} p_c = 1. \tag{34}$$

We use equal client weights. Client learning directions are not averaged before retrieval. Only the independently re-grounded surrogate losses are combined at the server.

D ADDITIONAL EXPERIMENTS

D.1 DIRECT PRIVATE ADAPTATION

We compare GRADHOP with direct adaptation of the target model on the private client data (Table 6). This does not satisfy the target-decoupled setting and serves as a reference under less restrictive target-access assumptions. Exact-private achieves substantially lower NLL, while its mean accuracy is comparable to that of GRADHOP.

Table 6: Direct private adaptation analysis on CLINC150. Exact-private adapts the target directly on private client examples and therefore does not satisfy the target-decoupled setting.

METHOD	ACCURACY (%) ↑	NLL ↓
Exact-private	85.9 ± 2.1	0.040 ± 0.021
GRADHOP	86.2 ± 0.4	0.114 ± 0.015

D.2 REFERENCE-MODEL CAPACITY

The client reference model size does not need to approach that of the target. We replace Qwen2.5-1.5B-Instruct with Qwen2.5-0.5B-Instruct while keeping the target model, support budget, $K = 4$ construction, representation dimension, and target-side optimizer fixed (Table 7).

Even the 0.5B reference improves substantially over the unadapted 7B target. As shown in Table 8, the smaller reference also reduces peak client GPU memory allocation and private-gradient extraction time. The communication size is the same for both reference models because the number and dimension of the communicated directions are fixed. These measurements concern reference-side extraction only and exclude server-side PUBLIC processing.

E DISCUSSION AND LIMITATIONS

Our analysis gives sufficient conditions for successful re-grounding across models rather than a universal guarantee. A key requirement is that the PUBLIC repository contain examples that can approximate the private target gradient in the target model space. Our target-space experiments show

Table 7: Effect of client reference model size on Banking77.

REFERENCE MODEL	ACCURACY (%) ↑	NLL ↓
BASE	73.4 ± 0.0	0.557 ± 0.000
0.5B, GRADHOP	81.7 ± 0.5	0.244 ± 0.012
1.5B, GRADHOP	84.9 ± 0.6	0.149 ± 0.010

Table 8: Average client-side resource and one-time communication costs on Banking77.

REFERENCE	PEAK GPU MEMORY (GB) ↓	EXTRACTION TIME (s) ↓	COMMUNICATION (KB) ↓
Qwen2.5-0.5B	2.04	32.8	≈ 524
Qwen2.5-1.5B	3.51	37.1	≈ 524

that the PUBLIC supports selected by GRADHOP provide substantially stronger alignment with the private target gradient, demonstrating that useful re-grounding can be achieved in the settings we study. The counterexample in Section B.1 clarifies the boundary of this result: Reference-space retrieval alone cannot compensate for a PUBLIC repository that lacks suitable examples.

The private target gradients used in the analysis of Section 5.1 are computed only for evaluation and are not available to GRADHOP during adaptation. Our experiments focus on intent-classification adaptation of language models using a shared reference model and a PUBLIC repository. The cross-family experiments demonstrate that the same reference-side information can support adaptation across several target-model families, motivating future work on extending the approach to broader tasks and modalities.

In our setting, the private data remain local, but the transmitted learning directions are data dependent and do not by themselves provide a formal privacy guarantee. Differential privacy, secure aggregation, and leakage analysis provide promising directions for future work.