AGGREGATION-FAITHFUL VALUE ADAPTATION FOR FEDERATED VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated adaptation of large vision-language models (VLMs) is appealing for privacy-sensitive applications but must contend with client data heterogeneity, strict communication budgets, and on-device inference constraints. Recent methods predominantly adopt lightweight prompt or adapter interfaces, yet averaging their parameters may fail to preserve the functional changes learned across heterogeneous clients. We therefore ask: can robust federated VLM adaptation be achieved through an aggregation-faithful intrinsic interface? We answer this question with **TriVA** (Tri-State Value Adaptation), a module-free framework that fine-tunes only native value biases in both visual and textual encoders. TriVA operates on existing backbone parameters and introduces no additional inference-time modules. Our analysis shows that this intrinsic interface provides substantially higher aggregation fidelity than the tested prompt- and adapter-based alternatives. To make this aggregation-friendly interface effective under heterogeneous clients without excessive synchronization, we further introduce a one-shot calibration mechanism that assigns value-bias groups to three persistent states: Frozen groups preserve pretrained values, Shared groups participate in federated aggregation, and Private groups retain client-specific adaptation locally. The resulting assignments remain fixed throughout training, with only shared groups exchanged for aggregation. Extensive experiments across heterogeneous federated benchmarks demonstrate strong adaptation performance with a compact trainable parameter set and substantially reduced communication cost. Semantic segmentation experiments further demonstrate the applicability of TriVA beyond image classification to dense prediction.

1 INTRODUCTION

Adapting large vision-language models (VLMs) in federated settings is increasingly important for privacy-sensitive applications, yet remains challenging under heterogeneous data and limited client resources. Foundation models such as CLIP (Radford et al., 2021) provide transferable representations and strong zero-shot generalization, but full-parameter fine-tuning is costly for bandwidth- and resource-constrained clients. Federated VLM adaptation must therefore balance adaptation effectiveness, communication efficiency, and deployment practicality.

Beyond efficiency, heterogeneous client data creates a coordination problem: useful local adaptations must remain meaningful after server-side parameter averaging. Learnable prompts steer representations through interactions between trainable context tokens and input content (Zhou et al., 2022), while bottleneck adapters encode feature corrections through jointly learned projections and nonlinear activations (Gao et al., 2024). Although both interfaces support effective local adaptation, their functional effects depend on interactions among the trainable components and the features they process. Averaging prompt embeddings or adapter weights therefore need not preserve the average correction learned by individual clients. The key issue is not only what each client can learn, but how its learned adaptation is represented and combined.

Motivated by this distinction, we ask: *can client-specific adaptations be expressed through a more directly composable intrinsic interface?* Native value biases provide a direct additive correction to attention outputs (Huang et al., 2026). For fixed inputs to an attention layer, value-bias updates introduce feature offsets without directly changing that layer's query–key attention weights. The

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.

frozen output projection maps these offsets into a common pretrained feature space, so averaging value biases averages their directly injected corrections at this layer. Clients can thus adjust internal representations without jointly learning additional transformation matrices. This structure motivates native visual and textual value biases as an aggregation-friendly adaptation interface.

To examine whether this structural property carries over to end-to-end predictions, we quantify *aggregation fidelity*: the agreement between the functional change induced by parameter averaging and the weighted average of client-wise functional changes. As illustrated in Figure 1, native value-bias adaptation achieves a functional merge error of 0.0148, compared with 0.4123 and 0.2085 for the tested prompt- and adapter-based interfaces, respectively. These results support the suitability of value biases for preserving client-wise functional changes under federated aggregation.

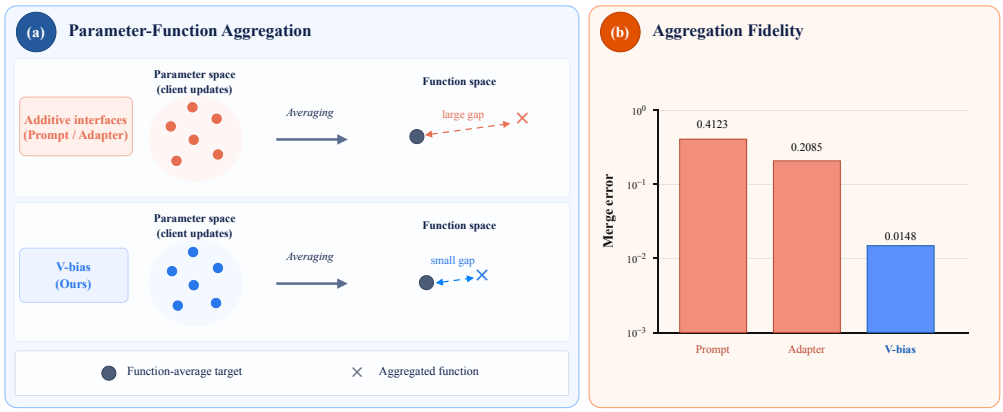

Figure 1: **Motivation of aggregation-faithful value adaptation.** Native value-bias adaptation exhibits substantially lower functional merge error than the tested prompt- and adapter-based interfaces, providing a more aggregation-friendly adaptation space.

An aggregation-friendly interface, however, does not by itself solve the synchronization problem. Some value-bias groups may be better left at their pretrained values, while useful local adaptations may still be unsuitable for synchronization. Our empirical analysis reveals *over-synchronization*: full sharing can underperform local retention, while eliminating sharing altogether is not consistently optimal. A robust federated adaptation method must therefore answer a second question: *how should adaptation capacity be preserved, synchronized, or retained locally?*

To address both challenges, we propose **TriVA** (**Tri**-State **V**alue **A**daptation), a module-free framework that fine-tunes native value biases in both visual and textual encoders. To make this interface compatible with heterogeneous clients, TriVA uses a one-shot client-conditioned calibration to evaluate adaptation utility and synchronization cost, assigning value-bias groups to three persistent states: **Frozen** groups preserve pretrained values, **Shared** groups participate in local optimization and federated aggregation, and **Private** groups retain locally optimized values across rounds without server overwriting. These assignments remain fixed throughout subsequent training. By communicating only Shared groups, TriVA makes communication efficiency a natural consequence of its synchronization structure, without introducing additional inference-time modules.

Our contributions are summarized as follows:

- We identify aggregation fidelity as an important property of federated adaptation interfaces and show that native value-bias adaptation provides substantially stronger parameter-to-function aggregation fidelity than the tested prompt- and adapter-based alternatives.
- We propose TriVA, which addresses over-synchronization through a static Frozen/Shared/Private structure that balances pretrained preservation, cross-client collaboration, and persistent local adaptation.

- Extensive experiments across heterogeneous classification benchmarks demonstrate strong adaptation and communication efficiency, while semantic segmentation on CamVid further demonstrates applicability beyond image-level recognition.

2 Related Work

2.1 Parameter-Efficient Adaptation of Vision-Language Models

Pretrained vision-language models such as CLIP (Radford et al., 2021) are commonly adapted through parameter-efficient fine-tuning (PEFT), which updates only a small subset of parameters. Existing approaches broadly follow additive and intrinsic paradigms. Additive methods introduce learnable prompts (Zhou et al., 2022; Khattak et al., 2023) or lightweight adapters (Gao et al., 2024), while keeping most backbone parameters frozen. Intrinsic approaches instead update existing parameters without inserting new modules. Related studies in language models include BitFit (Ben Zaken et al., 2022), which explores bias-only tuning, and BEFT (Huang et al., 2026), which analyzes value-projection biases in pretrained transformers. However, parameter efficiency and local adaptation effectiveness do not directly characterize how independently learned parameters behave when averaged. TriVA studies this complementary property through *aggregation fidelity*, motivating native value biases as a federated adaptation interface.

2.2 Federated Adaptation of Vision-Language Models

Federated VLM adaptation commonly combines lightweight tuning with mechanisms for handling client heterogeneity. Prompt-based approaches employ shared prompts and text-conditioned prompt generation (Guo et al., 2024; Qiu et al., 2024), or coordinate global and personalized prompts through optimal transport and structured personalization (Li et al., 2024; Cui et al., 2024; Fang et al., 2025). Adapter-based approaches such as pFedMMA (Ghiasvand et al., 2026) instead organize multi-modal adaptation through shared and client-specific adapter components. These methods primarily address how a chosen prompt or adapter interface should be coordinated across clients. TriVA complements this direction by examining the aggregation fidelity of the interface itself, then controlling synchronization within the selected native value-bias space.

2.3 Personalized Federated Optimization and Partial Synchronization

Personalized federated learning mitigates heterogeneity by separating shared representations from client-specific components or retaining selected parameters locally (Collins et al., 2021; Li et al., 2021). These approaches relax the requirement that every parameter follow a common global state. TriVA studies this synchronization problem within the compact value-bias subspace of a pretrained VLM. Beyond distinguishing shared and local adaptation, it explicitly separates **Frozen** pretrained values from **Private** parameters that continue to adapt locally, while **Shared** parameters participate in federated aggregation. The central question is therefore how to organize native adaptation capacity among pretrained preservation, cross-client collaboration, and persistent local retention, rather than simply which parameters to communicate.

3 Methodology

We propose **TriVA**, a module-free federated adaptation framework for vision-language models. TriVA rests on two coupled design principles. First, we restrict adaptation to native value biases in both encoders, using aggregation fidelity to assess this compact intrinsic interface. Second, to accommodate heterogeneous clients, we organize value-bias groups into **Frozen**, **Shared**, and **Private** states through one-shot calibration. Shared groups are optimized and aggregated, Private groups retain local adaptation, and Frozen groups preserve pretrained values. Figure 2 summarizes TriVA.

3.1 Problem Setup

We consider K clients with private datasets $\{\mathcal{D}_k\}_{k=1}^{K}$, where $n_k = |\mathcal{D}_k| > 0$ and $p_k = n_k / \sum_j n_j$. The pretrained VLM comprises visual and textual encoders with frozen parameters θ and candidate

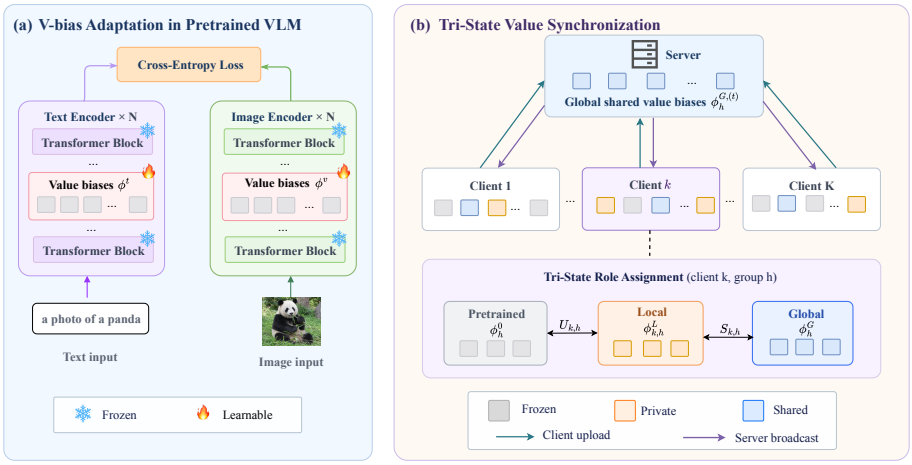

Figure 2: **Overview of TriVA.** One-shot calibration assigns native visual and textual value-bias groups to Frozen, Shared, or Private states. The assignments remain fixed, and only Shared groups participate in federated aggregation.

adaptation parameters $\phi = (\phi_v, \phi_t)$. Predictions use temperature-scaled visual–textual cosine similarities, and local optimization minimizes standard cross-entropy. TriVA groups native value biases by attention head, with h indexing groups across encoders and layers. Each client assigns group h a persistent role $r_{k,h} \in \{F, S, P\}$.

3.2 Aggregation-Faithful Value Adaptation

A central question is which compact adaptation interface supports meaningful aggregation of client updates. For fixed token features X, head h has value projection $V_h = XW_{V,h} + \mathbf{1}b_{V,h}^\top$. Under standard softmax attention without attention dropout, the attention matrix satisfies $A_h(X)\mathbf{1} = \mathbf{1}$. Consequently, changing only $b_{V,h}$ produces a direct additive correction (Huang et al., 2026). Let δb_h denote an increment to $b_{V,h}$ and $W_{O,h}$ the frozen output-projection block associated with head h. The resulting change in this head's projected attention output is

$$\Delta H_h(X; \delta b_h) = A_h(X)\mathbf{1}\delta b_h^\top W_{O,h} = \mathbf{1}\delta b_h^\top W_{O,h}. \tag{1}$$

where $\mathbf{1}$ is the all-ones vector over tokens. Thus, for a common layer input, averaging value-bias updates exactly averages their directly injected corrections. This provides a structural rationale for using native value biases as an aggregation interface. Since joint adaptation across layers and encoders remains nonlinear, we assess prediction-level aggregation fidelity through a controlled interface comparison. Within each interface, clients adapt from a common reference ϕ^{ref} to obtain states ϕ_k, whose parameter average is $\bar{\phi} = \sum_k p_k \phi_k$. Using the same ordered class vocabulary, let $\widetilde{z}(x; \phi)$ denote logits with their class-wise mean subtracted. Define $\Delta z_k(x) = \widetilde{z}(x; \phi_k) - \widetilde{z}(x; \phi^{\text{ref}})$, $\Delta z_{\text{func}}(x) = \sum_k p_k \Delta z_k(x)$, and $\Delta z_{\text{agg}}(x) = \widetilde{z}(x; \bar{\phi}) - \widetilde{z}(x; \phi^{\text{ref}})$. The normalized functional merge error is

$$\mathcal{E}_{\text{merge}} = \frac{\mathbb{E}_{x \sim \mathcal{D}_{\text{diag}}} \left[\| \Delta z_{\text{agg}}(x) - \Delta z_{\text{func}}(x) \|_2^2 \right]}{\mathbb{E}_{x \sim \mathcal{D}_{\text{diag}}} \left[\sum_k p_k \| \Delta z_k(x) \|_2^2 \right] + \epsilon}, \tag{2}$$

where $\mathcal{D}_{\text{diag}}$ is the diagnostic input distribution and $\epsilon > 0$. Lower error indicates closer agreement between parameter averaging and functional composition. Figure 1 shows lower error for value biases than for the tested prompt and adapter interfaces. This motivates the adaptation space, while leaving open which local updates should be synchronized.

3.3 TRI-STATE VALUE SYNCHRONIZATION

One-shot calibration. To distinguish adaptation utility from synchronization suitability, we first obtain local and global references with all value-bias groups temporarily trainable:

$$\phi_k^L = \mathrm{LocalAdapt}(\phi^0; \mathcal{D}_k), \qquad \phi^G = \sum_{k=1}^{K} p_k \phi_k^L, \tag{3}$$

where ϕ^0 is the pretrained value-bias state.

Counterfactual role assignment. Let $\widehat{L}_k(\phi) = \mathcal{L}_k(\phi; \mathcal{P}_k)$ be cross-entropy on a fixed local probe $\mathcal{P}_k$. We evaluate each group by replacing only its local value, keeping all other groups at $\phi_{k,-h}^L$:

$$U_{k,h} = \widehat{L}_k(\phi_{k,-h}^L, \phi_h^0) - \widehat{L}_k(\phi_k^L), \tag{4}$$

$$S_{k,h} = \widehat{L}_k(\phi_{k,-h}^L, \phi_h^G) - \widehat{L}_k(\phi_k^L). \tag{5}$$

Positive $U_{k,h}$ indicates that restoring the pretrained value increases probe loss. We form an active set $\mathcal{A}_k$ from positive-utility groups, ranked by $U_{k,h}$ when the active-ratio cap $\rho_{\max}$ is binding. For active groups, $S_{k,h}$ measures the loss change caused by global replacement. The assignment is

$$r_{k,h} = \begin{cases} F, & h \notin \mathcal{A}_k, \\ S, & h \in \mathcal{A}_k,\ S_{k,h} \le 0, \\ P, & h \in \mathcal{A}_k,\ S_{k,h} > 0. \end{cases} \tag{6}$$

Thus, Frozen groups preserve pretrained values, Shared groups use global replacements with non-positive probe cost, and Private groups retain beneficial local adaptation; roles remain fixed.

3.4 ROLE-AWARE FEDERATED OPTIMIZATION

After calibration, the three roles determine initialization and synchronization. At round t,

$$\phi_{k,h}^{(t,0)} = \begin{cases} \phi_h^0, & r_{k,h} = F, \\ \phi_h^{G,(t)}, & r_{k,h} = S, \\ \phi_{k,h}^{(t-1,E)}, & r_{k,h} = P. \end{cases} \tag{7}$$

Here $\phi_h^{G,(t)}$ is the server state and $\phi_{k,h}^{(t-1,E)}$ is the client's previous local state after E epochs. For the first post-calibration round, Shared and Private groups use ϕ^G and ϕ_k^L, respectively. Clients optimize Shared and Private groups for E epochs with cross-entropy, while Frozen groups stay fixed.

Only Shared groups are exchanged. For $\mathcal{K}_h^S = \{k : r_{k,h} = S\} \neq \varnothing$, the server updates

$$\phi_h^{G,(t+1)} = \sum_{k \in \mathcal{K}_h^S} \frac{n_k}{\sum_{j \in \mathcal{K}_h^S} n_j} \phi_{k,h}^{(t,E)}. \tag{8}$$

Groups with no sharing clients require no server update. Private groups are neither uploaded nor overwritten. This static synchronization structure directly reduces communication.

3.5 THEORETICAL JUSTIFICATION

We distinguish aggregation fidelity from synchronization safety. For fixed client states with the same ordered class vocabulary, write

$$g_k(x) = \widetilde{z}(x; \phi_k), \qquad g_{\mathrm{par}}(x) = \widetilde{z}\left(x; \sum_k p_k \phi_k\right), \qquad g_{\mathrm{fun}}(x) = \sum_k p_k g_k(x).$$

Let $\mathcal{Q}_k$ denote client k's evaluation distribution and define

$$R_k(g) = \mathbb{E}_{(x,y)\sim\mathcal{Q}_k}\left[\ell(g(x), y)\right].$$

For evaluation weights $\omega_k \geq 0$ with $\sum_k \omega_k = 1$, define

$$R(g) = \sum_k \omega_k R_k(g), \qquad R_{\mathrm{loc}} = \sum_k \omega_k R_k(g_k),$$

and

$$\mathcal{M} = \mathbb{E}_{x \sim Q_{\omega,X}} \left[\|g_{\mathrm{par}}(x) - g_{\mathrm{fun}}(x)\|_2^2 \right], \qquad \mathcal{N} = R(g_{\mathrm{fun}}) - R_{\mathrm{loc}}, \tag{9}$$

where $Q_{\omega,X}$ is the input marginal of $Q_\omega = \sum_k \omega_k \mathcal{Q}_k$. Client-macro evaluation uses $\omega_k = 1/K$. Evaluating Eq. (2) on $Q_{\omega,X}$ makes $\mathcal{M}$ its unnormalized numerator; Appendix A.2 gives their relation.

Proposition 3.1 (Aggregation fidelity and synchronization risk)**.** *For fixed client models, assume $R_{\mathrm{loc}} < \infty$, $R(g_{\mathrm{fun}}) < \infty$, $R(g_{\mathrm{par}}) < \infty$, and $\mathcal{M} < \infty$. Then*

$$|R(g_{\mathrm{par}}) - R_{\mathrm{loc}} - \mathcal{N}| \leq \sqrt{2\mathcal{M}}. \tag{10}$$

Hence, full synchronization increases risk whenever $\mathcal{N} > \sqrt{2\mathcal{M}}$. A two-client affine-logit construction further shows that $\mathcal{M} = 0$ can coexist with $R(g_{\mathrm{par}}) > R_{\mathrm{loc}}$.

Proof. See Appendix A.1. □

Proposition 3.1 separates two effects: $\mathcal{M}$ measures how faithfully parameter averaging reproduces functional averaging, whereas $\mathcal{N}$ measures the effect of replacing client-specific predictions with their shared composition. Thus, even perfect aggregation fidelity does not by itself guarantee beneficial synchronization. This motivates TriVA to combine an aggregation-friendly value-bias interface with explicit synchronization control.

We further analyze the joint tri-state assignment in Appendix A. Proposition A.1 shows that its calibration-loss change decomposes into the individual Frozen and Shared counterfactual effects plus a bounded cross-group interaction term, providing a theoretical interpretation of the adaptation-utility and synchronization-cost probes.

4 Experiments

4.1 Experimental Setup

Datasets. Following prior federated VLM adaptation studies, we evaluate TriVA on four groups of benchmarks. For single-domain CLIP transfer, we use Caltech101 (Fei-Fei et al., 2007), Food101 (Bossard et al., 2014), UCF101 (Soomro et al., 2012), Stanford Cars (Krause et al., 2013), and SUN397 (Xiao et al., 2010), using 16 shots per class and a Dirichlet partition ($\beta = 0.5$) across 10 clients. For multi-domain shift, we use OfficeHome (Venkateswara et al., 2017) and DomainNet (Peng et al., 2019), with each domain divided into two clients; DomainNet uses a fixed subset of 100 classes across six domains. We further evaluate CIFAR-10/100 (Krizhevsky, 2009) over 100 clients and semantic segmentation on CamVid (Brostow et al., 2009).

Baselines. We compare TriVA against zero-shot CLIP (Radford et al., 2021), PromptFL (Guo et al., 2024), FedTPG (Qiu et al., 2024), FedOTP (Li et al., 2024), FedPHA (Fang et al., 2025), and pFedMMA (Ghiasvand et al., 2026). All baselines use the same backbone, data partitions, and federated protocol. CamVid uses the same dense-prediction pipeline for all methods.

Implementation details. We use CLIP ViT-B/16 as the common backbone and freeze parameters outside each method's adaptation interface. TriVA uses $\tau_U = \tau_S = 0$, an active cap $\rho_{\max} = 0.75$, and three calibration probe batches. We set $E_{\mathrm{cal}} = E_{\mathrm{train}} = 5$ and $R = 50$, except on CIFAR-10/100 where $R = 25$. For classification, performance is reported as the average client-macro test accuracy over the last 10 communication rounds. Communication cost denotes the steady-state bidirectional FP16 parameter payload per client per communication round. All experiments are conducted on a single NVIDIA A100 GPU with 40GB memory.

4.2 PERFORMANCE COMPARISON

Single-domain CLIP transfer under non-IID data. We first compare TriVA with representative federated prompt- and adapter-tuning baselines on five single-domain benchmarks under the 16-shot setting. Table 1 reports the classification results. TriVA achieves the highest accuracy on all five datasets, with particularly clear gains on UCF101 and Stanford Cars. These results demonstrate the effectiveness of native value-bias adaptation with tri-state synchronization across diverse recognition tasks. Table 2 further compares trainable parameters and communication cost. TriVA locally optimizes a compact value-bias subspace while exchanging only Shared groups, enabling communication-efficient adaptation without additional inference-time modules.

Table 1: **Comparison with state-of-the-art (SOTA) methods on single-domain datasets across 10 clients.** All methods are evaluated under the same 16-shot setting.

Methods	Caltech101	Food101	UCF101	Stanford Cars	SUN397	Avg.
CLIP	93.16	85.24	65.08	65.53	62.58	74.32
PromptFL	95.15	83.76	80.83	77.54	73.54	82.16
FedTPG	95.06	84.97	79.95	72.49	71.02	80.70
FedOTP	93.03	93.87	71.03	62.51	69.98	78.08
FedPHA	93.38	96.06	78.01	77.10	72.46	83.40
pFedMMA	94.71	95.67	79.33	72.94	71.15	82.76
TriVA (Ours)	**96.26**	**96.44**	**82.76**	**79.34**	**74.34**	**85.83**

Impact of the number of shots. We further investigate the effect of varying the number of shots per class from 1 to 16 on the single-domain CLIP transfer datasets. As shown in Figure 3, TriVA consistently delivers strong performance and maintains the highest average accuracy across all shot settings. These results demonstrate its effectiveness under varying amounts of local supervision.

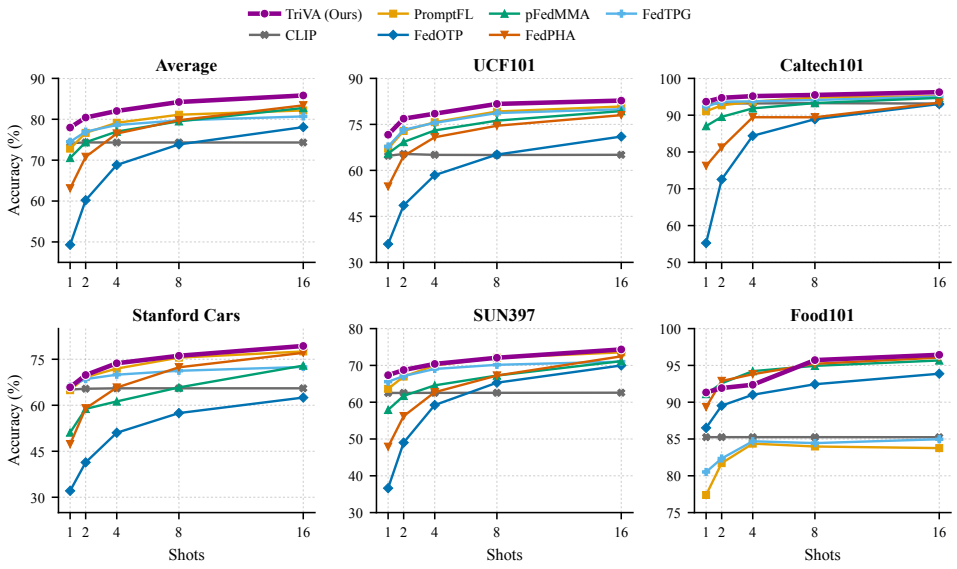

Figure 3: **Impact of the number of shots.** Accuracy comparison on five single-domain CLIP transfer datasets.

Robustness to label distribution shift. We further examine robustness on CIFAR-10 and CIFAR-100, partitioned across 100 clients using a Dirichlet distribution ($\beta = 0.5$). As reported in Table 3, TriVA achieves the highest accuracy on CIFAR-10 and remains competitive on CIFAR-100, where pFedMMA performs best. These results indicate that the proposed value-bias framework remains effective under label heterogeneity and a larger federation scale.

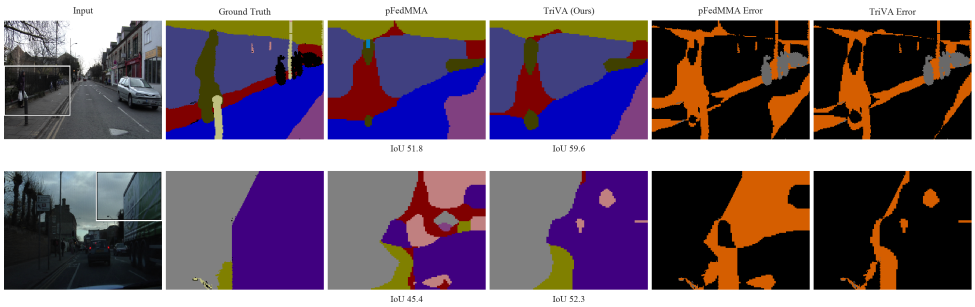

Figure 4: **Qualitative segmentation results on CamVid.** Representative examples comparing pFedMMA and TriVA. Orange regions denote prediction errors and gray regions denote ignored pixels.

Table 2: **Trainable parameters and communication cost.**

Methods	Trainable Parameters	Communication Cost
PromptFL	8.192 K	32.768 KB
FedTPG	4.206 M	16.824 MB
FedOTP	16.384 K	32.768 KB
FedPHA	16.384 K	32.768 KB
pFedMMA	758.880 K	76.032 KB
TriVA (Ours)	11.520 K	12.037 KB

Table 3: **SOTA comparison on CIFAR-10 and CIFAR-100 across 100 clients.**

Methods	CIFAR-10	CIFAR-100
CLIP	87.91	64.81
PromptFL	92.76	74.40
FedTPG	91.89	70.37
FedOTP	94.08	76.71
FedPHA	94.00	75.21
pFedMMA	96.07	**81.97**
TriVA (Ours)	**96.25**	80.78

Cross-domain robustness. We next evaluate cross-domain robustness on OfficeHome and DomainNet-100, where each domain is divided into two clients. Table 4 reports both per-domain and average accuracies. TriVA achieves the highest average accuracy on both benchmarks, supporting its effectiveness under feature-distribution heterogeneity.

Table 4: **Comparison with SOTA methods on OfficeHome and DomainNet-100.** Each domain is divided into two clients. DomainNet-100 is evaluated on a fixed subset of 100 classes.

Methods	OfficeHome					DomainNet-100						
	A	C	P	R	Avg.	C	I	P	Q	R	S	Avg.
CLIP	84.51	66.05	89.63	89.81	82.50	82.42	64.34	78.31	23.00	90.52	78.37	69.49
PromptFL	86.10	76.45	94.27	92.32	87.28	84.56	68.90	79.89	34.82	90.77	79.44	73.06
FedTPG	85.66	73.02	93.73	92.87	86.32	86.31	71.81	82.82	27.75	91.15	81.99	73.64
FedOTP	82.29	79.31	94.14	89.81	86.39	87.96	72.95	84.66	56.53	91.97	85.11	79.86
FedPHA	87.34	81.38	95.89	92.72	89.33	90.60	81.27	89.75	62.48	94.85	88.64	84.60
pFedMMA	85.29	83.35	**95.95**	91.95	89.13	91.23	78.52	89.64	**78.16**	94.67	88.85	86.85
TriVA (Ours)	**88.48**	**84.16**	95.87	**93.32**	**90.46**	**91.83**	**82.11**	**90.87**	71.91	**94.94**	**89.79**	**86.91**

Transfer to semantic segmentation. We further evaluate TriVA on CamVid using the same dense-prediction pipeline for all methods. TriVA achieves 52.27% mIoU, compared with 50.91% for pFedMMA. Figure 4 further shows reduced segmentation errors in representative examples. These results extend TriVA beyond classification to dense prediction.

4.3 TriVA as an Add-on to Existing Federated Adaptation

We further evaluate TriVA as an add-on to pFedMMA under the non-IID 16-shot setting. Table 5 shows consistent accuracy improvements across all five datasets. These results indicate that native value-bias adaptation can complement an existing adapter-based federated framework.

Table 5: **TriVA as an add-on to pFedMMA under non-IID 16-shot learning.**

Methods	Caltech101	Food101	UCF101	Stanford Cars	SUN397
pFedMMA	94.71	95.67	79.33	72.94	71.15
pFedMMA + TriVA	**94.82**	**95.94**	**79.63**	**74.75**	**72.16**

4.4 Ablation Study

Effect of visual and textual value adaptation. Table 6 compares visual-only, textual-only, and joint value-bias adaptation under static tri-state allocation, together with an All-Shared variant. Joint adaptation achieves the highest average accuracy of 85.83%, compared with 84.37% and 85.45% for the visual-only and textual-only variants, respectively. Although All-Shared performs best on UCF101, TriVA achieves higher accuracy on the other four datasets. These results support joint visual–textual adaptation with selective synchronization.

Table 6: **Value-bias adaptation ablation.** Visual-only, textual-only, and joint adaptation use static tri-state allocation; All-Shared fully synchronizes both encoders.

Variants	Caltech101	Food101	UCF101	Stanford Cars	SUN397	Avg.
Image-V (Static)	95.70	96.20	81.10	75.30	73.57	84.37
Text-V (Static)	95.87	96.38	82.33	78.72	73.94	85.45
All-Shared	95.32	86.86	**83.59**	77.79	73.43	83.40
TriVA (Ours)	**96.26**	**96.44**	82.76	**79.34**	**74.34**	**85.83**

Effect of tri-state synchronization. Table 7 evaluates the contribution of the three synchronization states on Stanford Cars. All-Shared yields lower accuracy and substantially higher communication cost than TriVA. Removing Frozen also reduces accuracy, while removing Shared causes a modest performance drop. Although removing Private slightly improves accuracy, it increases communication cost from 11.264 KB to 36.096 KB. The complete tri-state structure therefore provides a favorable balance between adaptation accuracy and communication efficiency.

Table 7: **Tri-state synchronization ablation on Stanford Cars.**

Variants	Accuracy (%)	Communication Cost
All-Shared	77.79	61.440 KB
w/o Frozen	78.72	11.264 KB
w/o Private	79.50	36.096 KB
w/o Shared	79.07	0 KB
TriVA (Ours)	79.34	11.264 KB

5 Conclusion

We presented TriVA, a module-free framework for federated VLM adaptation that fine-tunes native value biases in both visual and textual encoders. Rather than introducing external prompts or adapters, TriVA operates on a compact, aggregation-friendly intrinsic parameter space without adding inference-time modules. To address over-synchronization under client heterogeneity, TriVA uses one-shot calibration to assign value-bias groups to persistent Frozen, Shared, and Private states, enabling selective communication while preserving pretrained values and client-specific adaptation. Experiments across diverse federated classification benchmarks demonstrate strong adaptation performance with reduced communication cost, while semantic segmentation on CamVid further demonstrates applicability beyond image-level recognition. These results highlight aggregation fidelity and structured synchronization as complementary principles for efficient and deployment-friendly federated VLM adaptation. Future work will explore broader VLM architectures and adaptive synchronization under evolving client distributions.

AI USE STATEMENT

AI tools were used for manuscript preparation only. Their role was confined to language editing—improving clarity and correcting grammar—and to assistance with the LaTeX source, including table transcription and scripts used to cross-check the compiled manuscript against the numerical results recorded in our experimental artifacts. They were not used to formulate the research question of aggregation-faithful federated adaptation, to design the native visual–textual value-bias interface, to develop the one-shot calibration or the Frozen/Shared/Private synchronization mechanism in TriVA, to derive the theoretical analysis, to implement or modify the experimental pipeline, to generate experimental data, or to produce or interpret any result reported in this paper. All reported numbers are traceable to the authors' experimental records, and the authors take full responsibility for all methods, claims, analyses, and conclusions.

REPRODUCIBILITY STATEMENT

The complete method and training protocol are described in Sections 3 and 4. All assumptions and proofs for the theoretical results are provided in Appendix A. The paper reports the datasets, client partitions, training rounds, calibration settings, evaluation protocol, and communication accounting used in our experiments.

REFERENCES

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 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 (Volume 2: Short Papers)*, pp. 1–9, 2022. doi: 10.18653/v1/2022.acl-short.1.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, pp. 446–461, 2014. doi: 10.1007/978-3-319-10599-4_29.

Gabriel J. Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009. doi: 10.1016/j.patrec.2008.04.005.

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.

Tianyu Cui, Hongxia Li, Jingya Wang, and Ye Shi. Harmonizing generalization and personalization in federated prompt learning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 9646–9661, 2024.

Chengying Fang, Wenke Huang, Guancheng Wan, Yihao Yang, and Mang Ye. Fedpha: Federated prompt learning for heterogeneous client adaptation. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 15960–15975, 2025.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007. doi: 10.1016/j.cviu.2005.09.012.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024. doi: 10.1007/s11263-023-01891-x.

Sajjad Ghiasvand, Mahnoosh Alizadeh, and Ramtin Pedarsani. pfedmma: Personalized federated fine-tuning with multi-modal adapter for vision-language models. In *International Conference on Learning Representations*, 2026.

Tao Guo, Song Guo, Junxiao Wang, Xueyang Tang, and Wenchao Xu. Promptfl: Let federated participants cooperatively learn prompts instead of models – federated learning in age of

foundation model. *IEEE Transactions on Mobile Computing*, 23(5):5179–5194, 2024. doi: 10.1109/TMC.2023.3302410.

Baichuan Huang, Ananth Balashankar, and Amir Aminifar. Beft: Bias-efficient fine-tuning of language models in low-data regimes. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 38833–38851, 2026. doi: 10.18653/v1/2026.acl-long.1799.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19113–19122, 2023.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013. doi: 10.1109/ICCVW.2013.77.

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

Hongxia Li, Wei Huang, Jingya Wang, and Ye Shi. Global and local prompts cooperation via optimal transport for federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12151–12161, 2024.

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*, 2021.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1406–1415, 2019. doi: 10.1109/ICCV.2019.00149.

Chen Qiu, Xingyu Li, Chaithanya Kumar Mummadi, Madan Ravi Ganesh, Zhenzhen Li, Lu Peng, and Wan-Yi Lin. Federated text-driven prompt generation for vision-language models. In *International Conference on Learning Representations*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5018–5027, 2017.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010. doi: 10.1109/CVPR.2010.5539970.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. doi: 10.1007/s11263-022-01653-1.

A Additional Theoretical Analysis

This appendix provides the complete proof of Proposition 3.1, establishes its connection to the normalized functional merge error in Eq. (2), and develops the joint tri-state analysis summarized in Section 3.5.

A.1 Proof of Proposition 3.1

Proof. For a logit vector $z \in \mathbb{R}^C$ and label $y \in \{1, \dots, C\}$, the cross-entropy loss is

$$\ell(z, y) = \log \left(\sum_{c=1}^{C} e^{z_c} \right) - z_y. \tag{11}$$

Adding a common scalar to all logits leaves the cross-entropy loss unchanged, so the same expression applies to the class-centered logits used in our analysis.

Let $\pi = \mathrm{softmax}(z)$ and let e_y denote the y-th standard basis vector. Then

$$\nabla_z \ell(z, y) = \pi - e_y.$$

Since $\pi_c \geq 0$ and $\sum_c \pi_c = 1$,

$$\begin{aligned} \|\nabla_z \ell(z, y)\|_2^2 &= (1 - \pi_y)^2 + \sum_{c \neq y} \pi_c^2 \\ &\leq (1 - \pi_y)^2 + \left(\sum_{c \neq y} \pi_c \right)^2 \\ &= 2(1 - \pi_y)^2 \leq 2. \end{aligned} \tag{12}$$

Therefore, cross-entropy is $\sqrt{2}$-Lipschitz with respect to the logit vector under the Euclidean norm:

$$|\ell(z, y) - \ell(z', y)| \leq \sqrt{2}\, \|z - z'\|_2. \tag{13}$$

By the definition

$$\mathcal{N} = R(g_{\mathrm{fun}}) - R_{\mathrm{loc}},$$

we have

$$R(g_{\mathrm{par}}) - R_{\mathrm{loc}} - \mathcal{N} = R(g_{\mathrm{par}}) - R(g_{\mathrm{fun}}). \tag{14}$$

Applying Eq. (13) to each client evaluation distribution gives

$$\begin{aligned} &|R(g_{\mathrm{par}}) - R(g_{\mathrm{fun}})| \\ &\leq \sum_k \omega_k \mathbb{E}_{(x,y) \sim \mathcal{Q}_k} \left[|\ell(g_{\mathrm{par}}(x), y) - \ell(g_{\mathrm{fun}}(x), y)| \right] \\ &\leq \sqrt{2}\, \mathbb{E}_{x \sim Q_{\omega,X}} \left[\|g_{\mathrm{par}}(x) - g_{\mathrm{fun}}(x)\|_2 \right] \\ &\leq \sqrt{2\, \mathbb{E}_{x \sim Q_{\omega,X}} \left[\|g_{\mathrm{par}}(x) - g_{\mathrm{fun}}(x)\|_2^2 \right]} \\ &= \sqrt{2\mathcal{M}}, \end{aligned} \tag{15}$$

where the third inequality follows from Cauchy–Schwarz. Combining Eqs. (14) and (15) yields

$$|R(g_{\mathrm{par}}) - R_{\mathrm{loc}} - \mathcal{N}| \leq \sqrt{2\mathcal{M}}, \tag{16}$$

which proves the stated bound. Equivalently,

$$\mathcal{N} - \sqrt{2\mathcal{M}} \leq R(g_{\mathrm{par}}) - R_{\mathrm{loc}} \leq \mathcal{N} + \sqrt{2\mathcal{M}}. \tag{17}$$

Hence $\mathcal{N} > \sqrt{2\mathcal{M}}$ is sufficient for $R(g_{\mathrm{par}}) > R_{\mathrm{loc}}$.

To show that perfect aggregation fidelity does not by itself guarantee beneficial synchronization, consider a two-client affine-logit construction with

$$p_1 = p_2 = \omega_1 = \omega_2 = \frac{1}{2}.$$

Both clients observe the same deterministic input x_0, while their labels are always 1 and 2, respectively. Consider

$$g(x_0; \phi) = \left(\frac{\phi}{2}, -\frac{\phi}{2}\right), \qquad \phi_1 = a, \quad \phi_2 = -a, \quad a > 0. \tag{18}$$

The parameter average is $\bar{\phi} = 0$. By linearity,

$$g_{\mathrm{par}}(x_0) = g_{\mathrm{fun}}(x_0) = (0, 0),$$

and therefore $\mathcal{M} = 0$.

Each client's local model incurs cross-entropy loss $\log(1 + e^{-a})$, whereas the synchronized model incurs $\log 2$ on both clients. Hence

$$R(g_{\mathrm{par}}) - R_{\mathrm{loc}} = \log 2 - \log(1 + e^{-a}) > 0. \tag{19}$$

Thus, perfect aggregation fidelity does not generally imply synchronization safety. □

A.2 Connection to the Normalized Functional Merge Error

We next connect $\mathcal{M}$ in Proposition 3.1 to the normalized functional merge error in Eq. (2).

Let g_{ref} be a common reference function and define

$$\Delta g_k = g_k - g_{\mathrm{ref}}, \qquad \Delta g_{\mathrm{par}} = g_{\mathrm{par}} - g_{\mathrm{ref}}.$$

Since $\sum_k p_k = 1$,

$$\begin{aligned} \Delta g_{\mathrm{par}}(x) - \sum_k p_k \Delta g_k(x) &= g_{\mathrm{par}}(x) - g_{\mathrm{ref}}(x) - \sum_k p_k \big(g_k(x) - g_{\mathrm{ref}}(x)\big) \\ &= g_{\mathrm{par}}(x) - g_{\mathrm{fun}}(x). \end{aligned} \tag{20}$$

When the normalized merge error is evaluated on $Q_{\omega,X}$, define

$$D_{\mathrm{shift}} = \mathbb{E}_{x \sim Q_{\omega,X}} \left[\sum_k p_k \|\Delta g_k(x)\|_2^2 \right] + \epsilon, \qquad \epsilon > 0. \tag{21}$$

Assume $D_{\mathrm{shift}} < \infty$. Since $\epsilon > 0$, the normalization is well defined. Then

$$\mathcal{E}_{\mathrm{merge}}(Q_{\omega,X}) = \frac{\mathcal{M}}{D_{\mathrm{shift}}}, \tag{22}$$

or equivalently,

$$\mathcal{M} = D_{\mathrm{shift}}\, \mathcal{E}_{\mathrm{merge}}(Q_{\omega,X}). \tag{23}$$

Substituting this identity into Proposition 3.1 gives

$$|R(g_{\mathrm{par}}) - R_{\mathrm{loc}} - \mathcal{N}| \le \sqrt{2 D_{\mathrm{shift}}\, \mathcal{E}_{\mathrm{merge}}(Q_{\omega,X})}. \tag{24}$$

For a fixed normalization scale D_{shift}, lower normalized functional merge error tightens the bound on the risk difference between parameter and functional averaging, while $\mathcal{N}$ separately captures the effect of replacing client-specific predictions with their shared function.

A.3 JOINT EFFECT OF TRI-STATE SYNCHRONIZATION

We now analyze how the individual counterfactual probes compose when multiple value-bias groups are assigned simultaneously.

Fix client k. Let ϕ_k^L denote its calibrated local state and define the deterministic calibration-probe loss

$$\widehat{L}_k(\phi) = \mathcal{L}_k(\phi; \mathcal{P}_k).$$

For each disjoint value-bias group h, define the group-supported replacement increment

$$\delta_{k,h} = \begin{cases} \phi_h^0 - \phi_{k,h}^L, & h \in F_k, \\ \phi_h^G - \phi_{k,h}^L, & h \in S_k, \\ 0, & h \in P_k, \end{cases} \tag{25}$$

where F_k, S_k, and P_k denote the Frozen, Shared, and Private group sets, respectively.

The state obtained after applying all assigned replacements is

$$\phi_k^r = \phi_k^L + \sum_h \delta_{k,h}. \tag{26}$$

Proposition A.1 (Counterfactual composition of tri-state assignment)**.** *Suppose $\widehat{L}_k$ is twice continuously differentiable on a neighborhood of*

$$\Omega_k = \left\{ \phi_k^L + \sum_h t_h \delta_{k,h} : t_h \in [0,1] \right\}, \tag{27}$$

and its mixed Hessian blocks satisfy

$$\left\| \nabla^2_{h\ell} \widehat{L}_k(\phi) \right\|_{\mathrm{op}} \le \gamma_{k,h\ell} \tag{28}$$

for every $\phi \in \Omega_k$ and every $h < \ell$. Then

$$\begin{aligned} \widehat{L}_k(\phi_k^r) - \widehat{L}_k(\phi_k^L) &= \sum_{h \in F_k} U_{k,h} + \sum_{h \in S_k} S_{k,h} + \mathcal{R}_k, \\ |\mathcal{R}_k| &\le \Xi_k = \sum_{h<\ell} \gamma_{k,h\ell} \|\delta_{k,h}\|_2 \|\delta_{k,\ell}\|_2. \end{aligned} \tag{29}$$

Proof. Fix client k and write

$$b = \phi_k^L, \qquad f = \widehat{L}_k, \qquad \delta_h = \delta_{k,h}$$

for brevity.

For each group h, define its individual replacement effect

$$c_h = f(b + \delta_h) - f(b).$$

By the definitions of the adaptation utility and synchronization cost,

$$c_h = \begin{cases} U_{k,h}, & h \in F_k, \\ S_{k,h}, & h \in S_k, \\ 0, & h \in P_k. \end{cases} \tag{30}$$

Let

$$s_0 = 0, \qquad s_{h-1} = \sum_{j<h} \delta_j.$$

A telescoping decomposition gives

$$f\left(b + \sum_h \delta_h\right) - f(b) = \sum_{h=1}^{H} \left[f(b + s_{h-1} + \delta_h) - f(b + s_{h-1}) \right]. \tag{31}$$

Subtracting the sum of individual effects gives

$$\mathcal{R}_k = \sum_{h=1}^{H} B_h, \tag{32}$$

where

$$B_h = f(b + s_{h-1} + \delta_h) - f(b + s_{h-1}) - f(b + \delta_h) + f(b). \tag{33}$$

Applying the fundamental theorem of calculus first along δ_h and then along s_{h-1} gives

$$\begin{aligned} B_h &= \int_0^1 \delta_h^\top \left[\nabla f(b + s_{h-1} + t\delta_h) - \nabla f(b + t\delta_h)\right] dt \\ &= \int_0^1 \int_0^1 \delta_h^\top \nabla^2 f\left(b + u s_{h-1} + t\delta_h\right) s_{h-1}\, du\, dt \\ &= \sum_{j<h} \int_0^1 \int_0^1 \delta_h^\top \nabla^2 f\left(b + u s_{h-1} + t\delta_h\right) \delta_j\, du\, dt. \end{aligned} \tag{34}$$

Every integration point

$$b + u s_{h-1} + t\delta_h$$

belongs to Ω_k: the coefficients of the previously applied increments equal $u \in [0,1]$, the coefficient of the current increment equals $t \in [0,1]$, and all remaining coefficients equal zero.

Since δ_h and δ_j are supported on disjoint parameter groups, the corresponding mixed Hessian block satisfies

$$\left|\delta_h^\top \nabla^2 f(\phi) \delta_j\right| \le \gamma_{k,jh} \|\delta_h\|_2 \|\delta_j\|_2, \qquad j < h. \tag{35}$$

Therefore,

$$\begin{aligned} |\mathcal{R}_k| &\le \sum_{h=1}^{H} \sum_{j<h} \gamma_{k,jh} \|\delta_h\|_2 \|\delta_j\|_2 \\ &= \sum_{h<\ell} \gamma_{k,h\ell} \|\delta_{k,h}\|_2 \|\delta_{k,\ell}\|_2 \\ &= \Xi_k. \end{aligned} \tag{36}$$

Combining Eq. (30) with the telescoping decomposition gives

$$\widehat{L}_k(\phi_k^r) - \widehat{L}_k(\phi_k^L) = \sum_{h \in F_k} U_{k,h} + \sum_{h \in S_k} S_{k,h} + \mathcal{R}_k, \qquad |\mathcal{R}_k| \le \Xi_k, \tag{37}$$

which proves the proposition. □

A.4 Effect of the Active-Ratio Cap

Under TriVA's zero-threshold synchronization rule, every Shared group satisfies

$$S_{k,h} \le 0.$$

However, because the active set is constrained by $\rho_{\max}$, a Frozen group may still have positive adaptation utility if it is excluded when the active-ratio cap becomes binding.

To separate this effect, define

$$\Gamma_k = \sum_{h \in F_k} \max\{U_{k,h}, 0\}, \tag{38}$$

and

$$G_k = -\sum_{h \in F_k} \min\{U_{k,h}, 0\} - \sum_{h \in S_k} S_{k,h}. \tag{39}$$

Both Γ_k and G_k are nonnegative. Moreover,

$$\sum_{h \in F_k} U_{k,h} + \sum_{h \in S_k} S_{k,h} = \Gamma_k - G_k. \tag{40}$$

Combining Eq. (40) with Proposition A.1 yields

$$\widehat{L}_k(\phi_k^r) - \widehat{L}_k(\phi_k^L) \leq \Gamma_k - G_k + \Xi_k. \tag{41}$$

Therefore,

$$G_k \geq \Gamma_k + \Xi_k \tag{42}$$

is sufficient for the joint tri-state assignment to preserve or reduce the calibration loss relative to the fully locally calibrated state.

This decomposition clarifies the role of the active-ratio cap. Γ_k measures the potential cost of freezing useful groups that are excluded from the active set, G_k collects the gains from restoring non-useful Frozen groups and from Shared replacements with non-positive synchronization cost, and Ξ_k accounts for cross-group interactions.