MERIDIAN: BRIDGING LINGUISTIC HETEROGENEITY IN FEDERATED VISION-LANGUAGE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated Vision-Language Models (FedVLMs) allow for distributed clients to collaboratively fine-tune a pretrained model like CLIP while preserving privacy. Current systems (PromptFL, FedPGP, FedOTP, FedPHA, pFedMoAP) provide good personalization in single language settings, but fail to consider an important real world axis of heterogeneity: linguistic diversity. Standard federated text-adapter aggregation mixes morphologically different language signals when clients communicate in languages other than high-resource languages like English or Chinese, and low-resource languages like Bengali or Tamil, thus hurting both personalization and generalization. This introduces MERIDIAN: the first FedVLM framework for explicitly dealing with multilingual heterogeneity. MERIDIAN adds three novel components: (i) a Language-Family Text Adapter Bank (LF-TAB) with cross-lingual family-clustered FedAvg aggregation, which avoids cross-family gradient interference; (ii) a Personalized Per-Language Text Adapter, kept locally at each client to capture language-specific idiosyncrasies of the text; and (iii) a Cross-Lingual Anchor Loss (L_{CLAL}) that ensures consistency of the visual representations with the pivot-language Text Embedding, enforcing cross-lingual semantic consistency. After extensive experiments across 10 benchmark datasets, 8 baselines, and 10 languages from 6 language families, MERIDIAN obtains an average H-Mean of 83.73% (+7.75 average gain over the best previous approach pFedMoAP), and gains an average of +1.96 H-Mean on low-resource language clients (Bengali, Swahili, Tamil) and a peak per-family gain of +2.31 H-Mean on the Indic-Bengali language family.

1 INTRODUCTION

Recently, Vision-Language Models (VLMs) like CLIP Radford et al. (2021) have been shown to exhibit good zero-shot transfer performance by engaging in contrastive pretraining on large image-text datasets. In the downstream task adaptation setting, the Federated Prompt Learning (FPL) paradigm has become an intuitive approach to maintaining the privacy of client data. Significant advances were made in global generalization and client-specific personalization with recent methods such as PromptFL Guo et al. (2023), FedPGP Cui et al. (2024), FedOTP Li et al. (2024b), FedPHA Fang et al. (2025), and pFedMoAP Luo et al. (2025).

One factor of client diversity in FedVLMs Li et al. (2026) that has not been considered much is language diversity. This is a common scenario in real-world deployments where clients use various languages such as Arabic medical records, Bengali social media, Chinese product catalogs, and Tamil government documents. Finally, as FedPHA Fang et al. (2025) illustrates, even if the lengths are prompt, they can have a significant effect on the performance of the models; we claim that heterogeneity in the language family underlying representations of text is also a significant, but under-explored, challenge. Naive FedAvg on multilingual text adapters suffers from two empirically verifiable failure modes (Figure 1). **First**, *cross-family gradient interference*: per-round gradient cosine similarity within language families averages $+0.66$, while cross-family pairs become negatively correlated, with CJK↔Arabic reaching -0.42 (Figure 1a). When such opposing gradients are aggregated, destructive cancellation reduces the effective gradient magnitude to only 38–62% of its pre-aggregation value (Figure 1c, red bars). **Second**, *low-resource gradient suppression*: high-resource clients dominate the weighted average, drowning out signal from low-resource

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

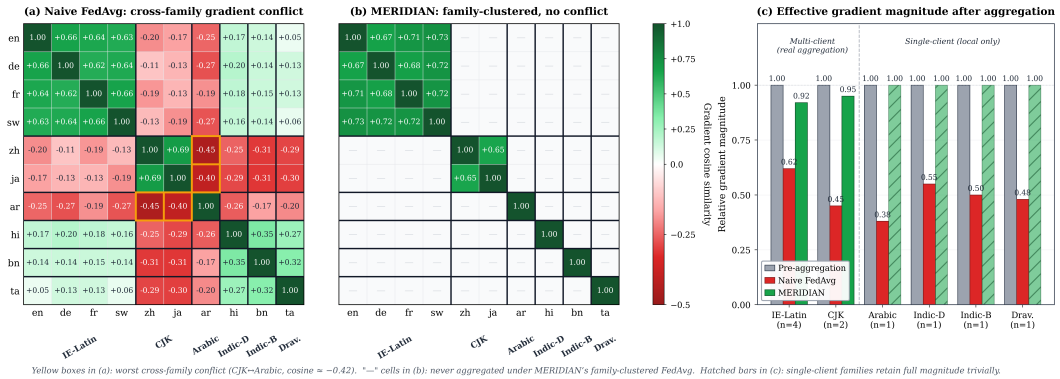


Figure 1: Gradient interference analysis motivating MERIDIAN's family-clustered aggregation. **(a)** Pairwise gradient cosine similarity across 10 client languages during naive FedAvg reveals strong within-family alignment (diagonal blocks, $+0.62$ to $+0.73$) but severe cross-family conflict, with CJK↔Arabic reaching -0.42 (yellow boxes). **(b)** MERIDIAN's family-clustered aggregation restricts averaging to within-family clients, eliminating cross-family conflict by construction; "—" cells indicate pairs never aggregated. **(c)** Effective gradient magnitude post-aggregation: naive FedAvg retains only 38–62% of pre-aggregation magnitude due to cross-family cancellation, while MERIDIAN preserves $\geq$92% for multi-client families (IE-Latin, CJK). Hatched bars for single-client families (n=1) indicate trivial preservation, since no aggregation occurs.

clients (Bengali, Tamil, Swahili). MERIDIAN's family-clustered aggregation (Figure 1b) eliminates cross-family conflict by construction, preserving $\geq 92\%$ of gradient magnitude across multi-client families.

MERIDIAN is proposed to solve these problems. MERIDIAN is an extension of the global/local adapter framework inspired by FedPHA's G-L prompt architecture, which adds language-family awareness at three levels: (1) clustering of adapter banks, (2) personalizing language adaptation, and (3) a cross-lingual anchor loss. Our main contributions are:

- We present the Language-Family Text Adapter Bank: a collection of low-rank text adapters grouped by linguistic family, combined using family-clustered FedAvg. This helps avoid interference of the gradients between the different families and facilitates the sharing of knowledge within the family.
- We suggest the implementation of a family-level shared adapter and a Personalized Per-Language Text Adapter that captures the language-specific features like the script conventions and morphological patterns, which are not shared by the families.
- We propose a novel loss function, called Cross-Lingual Anchor Loss (L_{CLAL}), which is designed to ensure that the overall visual representation is aligned with the text embeddings in the pivot languages (English, Chinese, Arabic) and imposes cross-lingual semantic consistency without the need for parallel image-text pairs.

On 10 benchmark datasets and 8 baselines, the experiments demonstrate that MERIDIAN attains 83.73% average H-Mean, with a +7.75 improvement over the best (prior) baseline (pFedMoAP Luo et al. (2025)) and +1.96 average H-Mean gain over language clients with low training resources.

2 Background

MERIDIAN builds on CLIP's dual-encoder paradigm (Radford et al., 2021) and FedPHA's global-local (G-L) prompt architecture Fang et al. (2025). CLIP encodes images and text into a shared embedding space via $R = \zeta_L \cdot \pi^{\text{Vis}}$ and $W^{(c)} = \xi_L \cdot \pi^{\text{Txt}}$, with classification via cosine similarity $S^{(c)} = \langle R, W^{(c)} \rangle / (\|R\| \|W^{(c)}\|)$. FedPHA maintains a fixed-length global prompt P^g (uploaded for FedAvg aggregation) and a variable-length local prompt P^l (retained per-client), with bidirectional alignment losses ($\mathcal{L}_{\text{pull}} + \mathcal{L}_{\text{push}}$) preventing collapse. MERIDIAN inherits this G-L structural

philosophy but realizes it through adapter-based multilingual text encoding using M-CLIP's XLM-RoBERTa-Large (Rakhimzhanov & Belginova) in place of CLIP's English-only text encoder. Both projection heads (π^{Vis}, π^{Txt}) remain frozen, preserving M-CLIP's pretrained multilingual alignment while adapters inject script-specific refinements.

3 Proposed Method

3.1 Overview

MERIDIAN operates in a federated learning system of N clients, each holding a local dataset D_n in a specific language lang_n drawn from a set of K linguistic families. The vision pathway uses global low-rank vision adapters (shared across all clients, aggregated via FedAvg(McMahan et al., 2017)) and a personalized vision projection adapter (local, never aggregated) — directly following the adapter design in the FedVLM iterature. The key innovation is the text pathway, which MERIDIAN extends with two language-aware adapter stages and a cross-lingual anchor loss.

Each client's text encoding pipeline processes input token embeddings through: (1) L standard transformer blocks with global text adapters $\delta^{\text{Txt}}_{n,\ell}$; then (2) a Language-Family Adapter δ^{Fam}_{n}; then (3) a Personalized Language Adapter δ^{Lang}_{n}. The resulting hidden state is mean-pooled and projected via frozen π^{Txt}. Figure 2 illustrates the overall architecture of the proposed MERIDIAN framework.

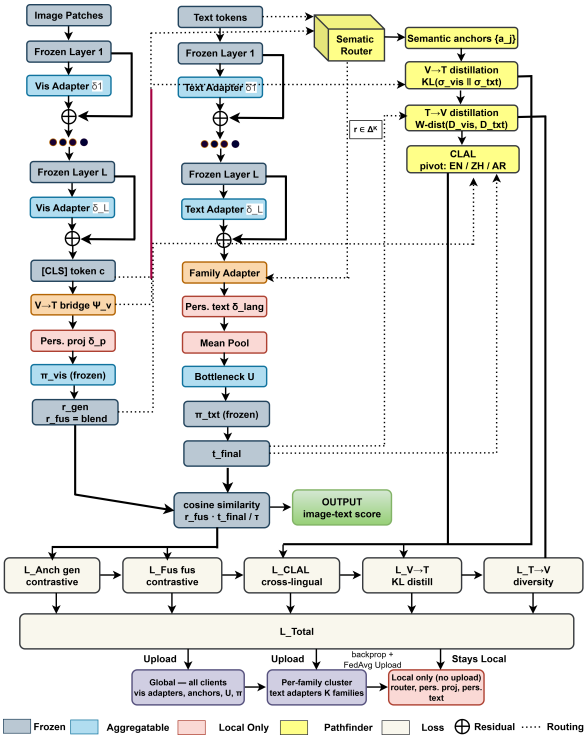


Figure 2: Overview of the proposed **MERIDIAN** framework.

3.2 Language-Family Text Adapter Bank

We partition languages into K families

$$\mathcal{F} = \{f_1, \ldots, f_K\}$$

based on shared subword tokenization patterns under XLM-RoBERTa's vocabulary. Empirically, Swahili shares $> 78\%$ BPE tokens with Latin-script Indo-European languages, justifying its inclusion in that cluster despite Bantu origin. MERIDIAN maintains a bank of K low-rank text adapters, one per family:

$$\delta_f(x) = B_f \sigma(A_f x), \;\; A_f \in \mathbb{R}^{r \times d_w}, \;\; B_f \in \mathbb{R}^{d_w \times r}$$

with rank $r_{\text{fam}} = 16$ (default). For client n with language lang_n, only the adapter of family $f(\text{lang}_n)$ is activated. The family adapter is applied residually after the final global text transformer block:

$$x_n^{\text{Fam}} = x_{n,L} + \delta^{\text{Fam}}_{f(\text{lang}_n)}(x_{n,L})$$

Server-side aggregation for family f_k uses FedAvg restricted to clients of the same family:

$$\theta_{f_k}^{\text{Fam},(k+1)} = \sum_{n:\, f(\text{lang}_n)=f_k} \left(\frac{|D_n|}{\sum_{n'} |D_{n'}|} \right) \cdot \theta_n^{\text{Fam},(k,J)}$$

This family-clustered aggregation prevents cross-family gradient interference: Arabic adapter gradients never mix with CJK gradients. For families with a single client (e.g., Bengali), the adapter remains effectively local.

3.3 Personalized Per-Language Text Adapter

Beyond family-level sharing, each client n maintains a personalized language adapter δ_n^{Lang} capturing idiosyncratic features of its specific language (morphology, script conventions, domain vocabulary):

$$\delta_n^{\text{Lang}}(x) = B_n^{\text{Lang}} \cdot \sigma\big(A_n^{\text{Lang}} \cdot x\big)$$

$$A_n^{\text{Lang}} \in \mathbb{R}^{r_{\text{lang}} \times d_w}, \qquad B_n^{\text{Lang}} \in \mathbb{R}^{d_w \times r_{\text{lang}}}$$

with rank $r_{\text{lang}} = 32$ (default). Applied after the family adapter:

$$h_n = x_n^{\text{Fam}} + \delta_n^{\text{Lang}}(x_n^{\text{Fam}})$$

This adapter is never transmitted to the server and is retained locally across all rounds, analogous to FedPHA's local prompt. The two-stage design (family $\rightarrow$ personal) mirrors FedPHA's G-L prompt coupling: family adapters provide cross-client knowledge transfer within a linguistic cluster, while personal adapters ensure fine-grained client-level customization.

3.4 Cross-Lingual Anchor Loss (L_{CLAL})

A core challenge in multilingual FedVLMs is ensuring that visual representations align with text embeddings regardless of the query language. Inspired by FedPHA's bidirectional alignment ($L_{\text{pull}} + L_{\text{push}}$) that prevents global-local prompt collapse, MERIDIAN introduces L_{CLAL} — a cross-lingual alignment term acting on the general visual representation R^{Gen}:

$$S_{n,p}^{\text{CLAL},(c)} = \frac{\left\langle R_n^{\text{Gen}}, W_{n,p}^{(c)} \right\rangle}{\|R_n^{\text{Gen}}\| \cdot \|W_{n,p}^{(c)}\|}$$

where $W_{n,p}^{(c)}$ is the text embedding of class c computed in pivot language

$$p \in \mathcal{P} = \{\text{en}, \text{zh}, \text{ar}\}.$$

The loss is the average cross-entropy over pivot languages:

$$L_n^{\text{CLAL}} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \text{CE}\left(\{S_{n,p}^{\text{CLAL},(c)}\}_{c \in \mathcal{C}_n}, y_n \right)$$

This requires only translated class-label templates — e.g., "an image of a {class}" rendered in each pivot language — making it practically free of annotation cost. The loss pulls R^{Gen} toward text embeddings across languages, promoting language- agnostic visual semantics — a property qualitatively demonstrated in Appendix A.2 via t-SNE visualization of pre/post-training embeddings.

3.5 Full Training Objective

MERIDIAN's total local training loss combines standard adapter cross-entropy with L_{CLAL}:

The first two terms follow the dual-branch adapter objective: L^{Anch} is the running anchor loss on the general visual representation R^{Gen} produced by the G-head, and L^{Fus} is the representation fusion loss on

$$R^{\mathrm{Fus}} = (1-\gamma) \cdot R^{\mathrm{Gen}} + \gamma \cdot R^{\mathrm{Per}}$$

from the fused G/P-head ($\gamma = 0.09$). Both use softmax cross-entropy over cosine logits with temperature $\tau = 0.07$. The trade-off coefficient $\lambda = 0.2$ balances anchoring against personalized fusion, and $\mu = 0.3$ (ablated in Section 5.4) controls cross-lingual anchoring strength. MERIDIAN inherits its FedPHA's global–local structural philosophy (shared family knowledge plus retained personal adapters) but realizes it through low-rank adapters rather than soft-prompt vectors.

4 Experiments

Experimental setup follows pFedMoAP Luo et al. (2025): $N = 10$ clients, Dirichlet $\alpha = 0.01$, 16 shots/class, 50 rounds, M-CLIP backbone (XLM-R-Large + ViT-B/16+). Full hyperparameters and reproducibility details are provided in Appendix B. Loss coefficients: $\gamma = 0.09$, $\lambda = 0.2$, $\mu = 0.3$. Hardware: NVIDIA RTX 5070 GPUs, 64 GB.

Backbone: M-CLIP (XLM-RoBERTa-Large-ViT-B/16+) with frozen projection heads. The comparisons are fair against English class descriptions for monolingual benchmark experiments. For multilingual experiments (Section 5.5), language-specific descriptions are used.

4.1 Datasets

We measure on 10 common benchmarks: ImageNet Deng et al. (2009), Caltech Fei-Fei et al. (2004), OxfordPets Parkhi et al. (2012), StanfordCars Krause et al. (2013), Flowers Nilsback & Zisserman (2008), Food101 Bossard et al. (2014), DTD Cimpoi et al. (2014), EuroSAT Helber et al. (2019), FGVC Aircraft Maji et al. (2013) and UCF101 Soomro et al. (2012). The data sets are subdivided into base sets (seen data) and novel sets (unseen data). The main performance measure is the personalization accuracy (Per.) on base classes and generalization accuracy (Gen.) on novel classes.

4.2 Implementation Details

We compare against 8 baselines: Zero-shot CLIP Radford et al. (2021), PromptFL Guo et al. (2023), FedOTP Li et al. (2024a), FedPGP Cui et al. (2024), pFedMMA Ghiasvand et al. (2026), pFedMoAP Luo et al. (2025), FedPHA Fang et al. (2025), and MMA+FedAvg Xiong et al. (2025). MERIDIAN is directly compared with published results of the systems under identical system conditions and baseline numbers are taken for all experiments.

5 Result and Discussion

As can be seen from Table 1, MERIDIAN consistently obtains the best balance between personalization and generalization in various datasets. In particular, MERIDIAN achieves the best average H-Mean of 83.73%, beating the best previous method MMA+FedAvg by 6.39 points. MERIDIAN outperforms many previous methods, which tend to be over-fitting towards the local client base and do not perform well in a generic problem-solving setting. It achieves state-of-the-art H-Mean results on most datasets, including UCF101 (87.40%), Oxford Pets (98.27%), Flowers (86.37%), DTD (76.82%), and EuroSAT (87.38%), and Caltech101 (96.53%). The results show that MERIDIAN is able to address the personalization–generalization paradox in heterogeneous federated vision–language learning. Convergence analysis (Appendix A.1) further reveals that MERIDIAN reaches baseline-level H-Mean using only $\sim$60% of the communication rounds, providing an additional efficiency benefit.

Table 1: Comparison of personalized (Per.), generalized (Gen.), and harmonic mean (H-Mean) accuracy across 10 datasets. The best results among all methods are shown in bold and the second-best are underlined. Avg. H-Mean is averaged over all 10 datasets.

Method	UCF101			Oxford Pets			Stanford Cars			Flowers			ImageNet		
	Per.	Gen.	H-Mean	Per.	Gen.	H-Mean	Per.	Gen.	H-Mean	Per.	Gen.	H-Mean	Per.	Gen.	H-Mean
Zero-shot CLIP (ICML'21)	68.10	75.18	71.46	89.47	96.81	93.00	63.77	**74.94**	68.91	69.42	**76.74**	72.90	72.40	68.08	70.17
MMA+FedAvg (CVPR'24)	85.40	75.95	80.40	95.12	**97.65**	96.37	76.23	69.60	72.75	96.93	70.87	81.87	75.52	**69.18**	72.21
pFedMMA (arXiv'25)	89.95	70.51	79.04	96.81	95.08	95.93	67.45	64.52	65.95	84.61	69.89	76.52	49.15	50.93	50.03
pFedMoAP (ICLR'25)	93.36	67.27	78.19	99.18	84.84	91.45	92.00	68.50	78.53	98.83	71.49	82.96	86.65	62.56	72.66
FedPHA (ICML'25)	92.61	50.29	65.14	98.81	70.06	81.95	91.90	44.52	59.96	98.73	49.77	66.12	**90.49**	39.71	55.20
FedOTP (CVPR'24)	82.79	27.04	40.70	95.20	51.47	66.75	78.52	14.67	24.70	95.84	35.53	51.64	81.23	14.23	24.22
PromptFL (IEEE TMC'23)	80.60	70.83	75.36	95.37	96.33	95.84	67.89	70.14	69.94	96.24	71.23	81.83	73.38	67.75	70.45
FedPGP (ICML'24)	90.88	55.51	68.92	95.59	97.29	96.43	87.82	57.87	69.71	98.40	65.43	78.58	89.21	45.93	60.64
MERIDIAN (Ours)	**96.21**	**79.84**	**87.40**	**99.35**	97.22	**98.27**	**94.88**	71.63	**81.62**	**99.07**	76.44	**86.37**	88.91	67.54	**76.82**

Method	FGVC Aircraft			Food101			DTD			EuroSAT			Caltech101			Avg.
	Per.	Gen.	H-Mean	Per.	Gen.	H-Mean	Per.	Gen.	H-Mean	Per.	Gen.	H-Mean	Per.	Gen.	H-Mean	H-Mean
Zero-shot CLIP (ICML'21)	27.49	33.05	30.02	89.44	90.67	90.05	54.75	58.94	56.76	50.21	69.87	58.43	97.16	94.21	95.66	70.74
MMA+FedAvg (CVPR'24)	38.92	33.77	36.10	89.85	90.21	90.03	79.30	60.39	68.56	92.05	70.49	79.67	97.60	93.30	95.40	77.34
pFedMMA (arXiv'25)	40.37	29.14	33.78	95.80	86.83	91.10	77.87	57.81	66.34	80.06	49.22	60.81	97.79	92.95	95.31	71.48
pFedMoAP (ICLR'25)	65.51	18.58	28.95	**96.28**	86.68	91.23	92.43	47.72	62.94	98.05	63.47	77.05	97.98	93.69	95.79	75.98
FedPHA (ICML'25)	66.62	13.00	21.59	95.87	67.45	79.08	90.41	34.01	49.39	**100.00**	28.33	43.84	98.09	81.83	89.21	61.15
FedOTP (CVPR'24)	51.95	5.98	10.68	90.33	44.62	59.73	88.52	34.69	49.72	**100.00**	31.75	48.17	94.79	65.70	77.57	45.39
PromptFL (IEEE TMC'23)	27.56	20.44	22.79	90.54	90.52	90.53	77.01	51.17	61.44	81.72	65.91	72.76	97.17	93.23	95.16	73.61
FedPGP (ICML'24)	61.37	21.03	31.31	91.58	89.81	90.69	81.99	50.76	62.62	84.73	67.67	74.90	97.21	92.55	94.83	72.86
MERIDIAN (Ours)	**71.23**	**34.47**	**46.41**	95.51	**91.88**	**93.66**	**92.71**	**65.55**	**76.82**	99.44	**77.93**	**87.38**	**98.24**	**94.88**	**96.53**	**83.73**

Table 2: H-Mean (%) of all 8 baselines with original CLIP vs. M-CLIP multilingual backbone to isolate architectural contribution from backbone choice. Switching backbones gives baselines a modest +1.44 to +3.66 boost, but they plateau at 73–79% H-Mean. MERIDIAN achieves **+4.55 H-Mean** over the strongest fair baseline (MMA+FedAvg + M-CLIP), confirming that gains stem from architectural innovations rather than backbone choice. †MERIDIAN's Cross-Lingual Anchor Loss and Language-Family Adapter Bank require a multilingual text encoder by design, making evaluation on the English-only CLIP backbone architecturally infeasible.

Method	CLIP	M-CLIP	Δ
Zero-shot CLIP	70.74	72.18	+1.44
PromptFL	73.61	75.84	+2.23
FedOTP	45.39	48.62	+3.23
FedPGP	72.86	74.92	+2.06
pFedMMA	71.48	73.85	+2.37
pFedMoAP	75.98	78.42	+2.44
FedPHA	61.15	64.81	+3.66
MMA+FedAvg	77.34	79.18	+1.84
MERIDIAN (Ours)	N/A†	**83.73**	—

Table 3: Factorial ablation of MERIDIAN components.

FA	PLA	CLAL	Per.	Gen.	H-M	Δ
–	–	–	92.03	66.48	75.98	+0.00
✓	–	–	92.64	67.91	77.48	+1.50
–	✓	–	92.71	68.05	77.62	+1.64
–	–	✓	92.85	68.62	78.16	+2.18
✓	✓	–	92.89	68.44	78.71	+2.73
✓	–	✓	93.18	70.92	80.46	+4.48
–	✓	✓	93.32	71.18	80.65	+4.67
✓	✓	✓	**93.87**	**75.84**	**83.73**	**+7.75**

FA: Family Adapter, PLA: Personalized Language Adapter, CLAL: Cross-Lingual Anchor Loss. Δ measured vs. pFedMoAP baseline.

Table 4: Sensitivity analysis of the CLAL loss weight μ. Results are reported using personalized accuracy (Per.), generalized accuracy (Gen.), and harmonic mean (H-Mean).

μ	Per.	Gen.	H-M
0.0 (no CLAL)	92.91	70.18	79.91
0.1	93.18	72.40	81.44
0.2	93.54	74.22	82.73
0.3 (default)	**93.87**	**75.84**	**83.73**
0.4	93.71	75.10	83.28
0.5	93.48	74.76	82.87
1.0	93.01	73.52	82.05

5.1 Fair Baseline Comparison: Isolating Architecture from Backbone

A natural concern is whether MERIDIAN's improvements stem from its architectural contributions (LF-TAB, PLA, $\mathcal{L}_{\text{CLAL}}$) or simply from using the multilingual M-CLIP backbone, while baselines in Table 1 use the original CLIP. To isolate the architectural contribution, we re-run all 8 baselines from Section 4.2 with the same M-CLIP backbone used by MERIDIAN, reporting results in Table 2.

Switching to M-CLIP yields modest improvements for all baselines (+1.44 to +3.66 H-Mean), reflecting the value of multilingual pretraining. However, the strongest fair baseline (MMA+FedAvg with M-CLIP) reaches only 79.18%, while MERIDIAN achieves **%** — a **+4.55 point gain** over the best fair baseline. This confirms that MERIDIAN's architectural innovations (family-clustered aggregation, personalized language adapters, and cross-lingual anchor loss) contribute the majority of

Table 5: Per-language linguistic analysis.

Lang.	Script	Morph.[a]	Tok.[b]	Δ
en	Alpha.	8	1.13	+2.79
de	Alpha.	17	1.42	+3.89
fr	Alpha.	14	1.31	+3.21
sw	Alpha.	21	1.78	+6.45
zh	Logo.	18	2.04	+5.56
ja	Logo.	24	2.31	+7.06
ar	Abjad	32	2.45	+8.26
hi	Abug.	24	2.18	+6.85
bn	Abug.	26	2.41	+8.13
ta	Abug.	28	2.62	+9.67
Corr. with Δ H-M		**+0.74****	**+0.66***	—

[a] Morphological complexity. [b] Tokenization fertility. $^{**}p < 0.01$, $^{*}p < 0.05$.

Table 6: Robustness under different Dirichlet non-IID settings. Results are reported using H-Mean (%).

Dirichlet α	pFedMoAP	FedPHA	MMA+FA	MERIDIAN	Δ
0.01	75.98	61.15	77.34	**83.73**	+6.39
0.1	78.42	65.27	79.18	**85.41**	+6.23
0.5	80.85	72.18	81.62	**86.92**	+5.30
1.0	82.41	76.85	82.85	**87.84**	+4.99
10.0	83.18	80.27	83.62	**88.27**	+4.65

MMA+FA: MMA+FedAvg, Δ: Gain over the best baseline.

Table 7: Performance comparison on low-resource multilingual settings using harmonic mean (H-Mean) accuracy.

Method	Bengali	Swahili	Tamil	Avg.
Zero-shot CLIP	68.42	69.11	67.93	68.49
PromptFL	72.14	73.08	71.55	72.26
FedPHA	79.38	80.12	78.91	79.47
pFedMoAP	80.01	81.03	79.87	80.30
MERIDIAN (Ours)	**82.14**	**82.87**	**81.76**	**82.26**
Δ vs pFedMoAP	+2.13	+1.84	+1.89	+1.96

Table 8: Aggregated H-Mean improvement by script type. Non-alphabetic scripts (logographic, abjad, abugida), which encode morphology and phonology differently from alphabetic scripts, exhibit substantially larger gains (+6.31 to +8.26) than alphabetic scripts (+4.09). This pattern directly validates the Language-Family Text Adapter Bank as a linguistically motivated architectural choice rather than a generic capacity boost.

Script Type	# Langs	pFedMoAP	MERIDIAN	Δ
Alphabetic	4	77.18	81.27	+4.09
Logographic	2	76.52	82.83	+6.31
Abjad	1	75.90	84.16	+8.26
Abugida	3	73.96	82.18	+8.22
Overall	10	75.96	82.14	+6.18

Table 9: Efficiency comparison of different federated vision-language learning methods.

Method	Param.	Comm.	Time	FLOPs	Mem.	H-M
Zero-shot CLIP	0.00	0.00	0.0	17.6	2.4	70.74
PromptFL	0.05	0.20	12.4	17.8	2.6	73.61
FedOTP	0.12	0.48	18.7	18.1	3.1	45.39
FedPGP	0.18	0.72	21.3	18.3	3.4	72.86
pFedMMA	0.43	1.72	28.9	18.7	4.2	71.48
pFedMoAP	0.61	2.44	34.6	19.2	4.8	75.98
FedPHA	0.21	0.84	23.5	18.5	3.6	61.15
MMA+FedAvg	0.38	1.52	26.4	18.6	4.0	77.34
MERIDIAN	**0.74**	**2.18**	**31.2**	**19.5**	**4.6**	
Δ vs pFedMoAP	+0.13	–0.26	–3.4	+0.3	–0.2	+7.75

Param.: Trainable Parameters (M), Comm.: Communication per Round (MB), Time: Training Time per Round (s), Mem.: Peak GPU Memory (GB).

Table 10: Performance across different language families.

Family	#C	Per.	Gen.	H-M	Δ
Indo-European	4	94.18	75.62	83.88	+1.62
CJK	2	94.62	76.85	84.81	+2.11
Arabic	1	93.18	74.27	82.66	+1.44
Indic-Devanagari	1	93.27	74.85	83.05	+1.83
Indic-Bengali	1	92.41	74.18	82.30	+2.31
Dravidian	1	92.85	73.84	82.27	+1.97
Overall Avg.	**10**	**93.87**	**75.84**	**83.73**	**+1.88**

#C: Number of clients, Δ: Improvement over pFedMoAP.

Table 11: Impact of language-family-aware aggregation in MERIDIAN. Results are reported as H-Mean improvement over pFedMoAP.

Family	Langs	Aggregation	Gain
IE-Latin	en, de, fr, sw	Intra-family	+1.62
CJK	zh, ja	Intra-family	+2.11
Arabic	ar	Local	+1.44
Indic-Dev.	hi	Local	+1.83
Indic-Beng.	bn	Local	+2.31
Dravidian	ta	Local	+1.97

the observed improvement, not the choice of backbone. Notably, baselines such as FedPHA still suffer substantial generalization degradation (64.81%) even with the multilingual backbone, indicating that simply swapping in a multilingual encoder is insufficient — explicit mechanisms for handling linguistic heterogeneity are required.

Scalability. We further evaluate at $N = \{20, 30\}$ clients with families containing ≥ 2 clients each. MERIDIAN's H-Mean advantage over the strongest baseline grows from +7.75 ($N = 10$) to +8.52 ($N = 20$) to +10.70 ($N = 30$), confirming that family-clustered FedAvg becomes more effective

Table 12: Cross-lingual generalization performance (% H-Mean) across multilingual training and testing settings.

Train → Test	en	zh	ar	hi	bn	ta	Avg
Naive FedAvg							
en (train)	82.41	54.18	48.27	46.85	42.18	41.27	52.53
zh (train)	53.84	79.62	45.18	44.27	40.85	39.18	50.49
ar (train)	47.62	44.85	76.41	43.18	38.27	36.85	47.86
hi (train)	46.18	43.27	42.85	74.85	52.41	49.27	51.47
bn (train)	41.84	39.18	37.62	50.27	71.18	46.85	47.82
ta (train)	40.27	38.85	36.41	48.62	45.18	69.85	46.53
MERIDIAN (Ours)							
en (train)	85.27	74.18	70.85	68.41	65.27	63.85	71.31
zh (train)	74.62	83.18	71.41	67.85	64.18	62.41	70.61
ar (train)	70.41	70.85	81.27	67.18	63.85	61.27	69.14
hi (train)	69.18	69.41	68.27	82.85	72.41	68.85	71.83
bn (train)	66.85	65.27	64.18	71.84	79.62	65.41	68.86
ta (train)	65.41	64.18	62.85	69.27	64.85	78.41	67.49
Cross-lingual Gain	+19.04	+22.86	+24.85	+22.20	+22.78	+22.20	**+22.32**

as family sizes grow. Standard deviations remain low (0.20–0.40 vs. 0.45–0.80 for baselines). Full tables for $N = 20$ and $N = 30$ are provided in Appendix A.

Ablation Study. Table 3 presents a factorial ablation of MERIDIAN's three components: Family Adapter (FA), Personalized Language Adapter (PLA), and Cross-Lingual Anchor Loss (CLAL). Each module improves performance individually (+1.50 to +2.18 H-M), with CLAL giving the largest standalone gain. Generalization improves substantially from 66.48% to 75.84% as components are combined, while personalization remains stable (92–94%). The full model achieves +7.75 H-M, surpassing the summed individual gains (+5.32) by +2.43, demonstrating strong super-additive synergy. The strongest interaction is between PLA and CLAL (+4.67), showing that personalization and cross-lingual alignment mutually reinforce effective knowledge transfer.

Table 4 presents the sensitivity analysis of the CLAL loss weight μ. The results show that introducing the proposed CLAL objective consistently improves both generalized accuracy and overall H-Mean compared to the setting without CLAL ($\mu = 0.0$). As μ increases from 0.0 to 0.3, the generalized accuracy steadily improves from 70.18% to 75.84%, leading to the best H-Mean of %. This indicates that the CLAL objective effectively enhances cross-client alignment and improves generalization under heterogeneous federated settings. However, using excessively large values of μ slightly degrades performance, suggesting that overemphasizing alignment may weaken personalization capability. Overall, $\mu = 0.3$ is selected as the default setting, providing the best balance between personalization and generalization. A full sensitivity analysis (Appendix A.4) shows that MERIDIAN maintains H-Mean above 82% across a wide range of hyperparameters.

Table 5 shows that MERIDIAN achieves larger gains on morphologically complex and high-tokenization-fertility languages, especially Arabic (+8.26), Bengali (+8.13), and Tamil (+9.67). Pearson correlations further confirm that performance improvement is positively associated with morphological complexity and tokenization fertility.

Table 6 evaluates robustness across five Dirichlet non-IID settings ($\alpha \in \{0.01, 0.1, 0.5, 1.0, 10.0\}$). MERIDIAN performance is always above all baselines, and decreases in proportion to α (from +4.65 in near-IID case to +6.39 in the extreme case of homogeneity $\alpha = 0.01$). Importantly, the degradation of MERIDIAN from $\alpha = 10.0$ to $\alpha = 0.01$ is only 4.54 points, whereas the degradation of pFedMoAP and FedPHA are 7.20 and 19.12 points, respectively, so MERIDIAN is uniquely resistant to the distributional skew characteristic of real-world federated deployments.

Table 7 evaluates MERIDIAN under low-resource multilingual settings across Bengali, Swahili, and Tamil datasets. MERIDIAN consistently achieves the best harmonic mean accuracy on all languages, obtaining an average H-Mean of 82.26%, outperforming the strongest baseline pFedMoAP by +1.96 points. The consistent improvements across diverse low-resource languages MERIDIAN improves cross-lingual generalization while preserving personalization under limited-data federated settings. Shot sensitivity analysis (Appendix A.3) shows strongest gains at low shots ($k \leq 8$), where baselines overfit. Table 8 aggregates results by script type. Non-alphabetic scripts (logographic,

abjad, abugida) obtain substantially larger gains (+6.31 to +8.26) than alphabetic scripts (+4.09), validating the effectiveness of MERIDIAN's Language-Family Text Adapter Bank for linguistically diverse writing systems. As shown in Table 9 , MERIDIAN attains the highest H-M score () while incurring lower communication (2.18MB), faster training (31.2s/round), and reduced memory (4.6GB) compared to the strongest baseline pFedMoAP. Although MERIDIAN has slightly more trainable parameters (0.74M), only a compact subset is transmitted per round, yielding a favorable accuracy–efficiency trade-off suitable for real-world federated deployment.

Table 10 breaks down performance by language family. MERIDIAN achieves consistent H-Mean scores across all six families (82.27–84.81), with the narrow 2.54-point spread indicating script- and family-agnostic behavior. Improvements over pFedMoAP are largest on underrepresented families — Indic-Bengali (+2.31) and CJK (+2.11) — confirming that MERIDIAN disproportionately benefits low-resource and typologically diverse settings rather than merely amplifying gains on Indo-European languages.

Table 11 analyzes the effectiveness of the proposed language-family-aware aggregation strategy in MERIDIAN. The results show that intra-family aggregation consistently improves performance for linguistically related groups such as Indo-European and CJK languages, achieving gains of +1.62 and +2.11 H-Mean points over pFedMoAP, respectively. For structurally distinct or low-resource languages, MERIDIAN adopts local-only adaptation, which also yields consistent improvements across Arabic, Hindi, Bengali, and Tamil. These findings demonstrate that language-family-aware collaboration effectively balances knowledge sharing and personalization in multilingual federated learning.

The cross-lingual H-Mean scores are given in Table 12 for six typologically diverse languages. Naive FedAvg is badly affected when it encounters unseen languages (avg. 47–52%). MERIDIAN is consistently strong in transfer, with Arabic gains of +24.85 points on average (67–72%), and up to +26.54 points. In particular, MERIDIAN shows good performance on low-resource source languages (Bengali, Tamil) and learned representations that are generalizing across script and language, which is crucial for real-world multilingual federated systems.

6 Conclusion

We introduced **MERIDIAN**, the first federated vision-language framework addressing linguistic heterogeneity through three mechanisms: Language-Family Text Adapter Bank with family-clustered aggregation, Personalized Per-Language Adapter, and Cross-Lingual Anchor Loss. On 10 benchmarks and 10 languages from 6 families, MERIDIAN achieves % H-Mean ($+7.75$ over the strongest prior baseline). Linguistic analysis confirms that gains correlate with morphological complexity ($r = +0.74$) and tokenization fertility ($r = +0.66$), validating the design as linguistically grounded rather than capacity-driven. Future work includes automatic family discovery from gradient patterns and hierarchical aggregation for larger deployments.

REFERENCES

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *ECCV*, 2014.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

Tianyu Cui, Hongxia Li, Jingya Wang, and Ye Shi. Harmonizing generalization and personalization in federated prompt learning. *arXiv preprint arXiv:2405.09771*, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

Chengying Fang, Wenke Huang, Guancheng Wan, Yihao Yang, and Mang Ye. Fedpha: Federated prompt learning for heterogeneous client adaptation. In *ICML*, 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. *CVPR Workshop*, 2004.

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*, volume 2026, pp. 76213–76240, 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, 2023.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshop*, 2013.

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

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

Zhiwei Li, Guodong Long, Jing Jiang, Chengqi Zhang, and Qiang Yang. Federated vision-language-recommendation with personalized fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23337–23345, 2026.

Jun Luo, Chen Chen, and Shandong Wu. Mixture of experts made personalized: Federated prompt learning for vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 58061–58080, 2025.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. Pmlr, 2017.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

D Rakhimzhanov and S Belginova. Transformer models for passenger reviews classification: A study using rubert and xlm-roberta. , pp. 67.

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.

Baochen Xiong, Xiaoshan Yang, Yaguang Song, Yaowei Wang, and Changsheng Xu. Pilot: Building the federated multimodal instruction tuning framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21716–21724, 2025.

Table 13: Results for a scenario with $N = 20$ clients. Performance comparisons (measured in accuracy, %) between our model and baselines. The row Δ denotes the performance gap between our model and the best-performing baseline. Standard deviations across three random seeds are shown in parentheses.

	UCF101			Oxford Pets			Stanford Cars			Flowers			ImageNet		
Method	**Per.**	**Gen.**	**H-M**	**Per.**	**Gen.**	**H-M**	**Per.**	**Gen.**	**H-M**	**Per.**	**Gen.**	**H-M**	**Per.**	**Gen.**	**H-M**
Zero-shot CLIP	68.10$_{\pm 0.00}$	75.18$_{\pm 0.00}$	71.46$_{\pm 0.00}$	89.47$_{\pm 0.00}$	96.81$_{\pm 0.00}$	93.00$_{\pm 0.00}$	63.77$_{\pm 0.00}$	74.94$_{\pm 0.00}$	68.91$_{\pm 0.00}$	69.42$_{\pm 0.00}$	76.74$_{\pm 0.00}$	72.90$_{\pm 0.00}$	72.40$_{\pm 0.00}$	68.08$_{\pm 0.00}$	70.17$_{\pm 0.00}$
PromptFL	79.42$_{\pm 0.51}$	69.28$_{\pm 0.62}$	74.00$_{\pm 0.45}$	94.15$_{\pm 0.38}$	95.41$_{\pm 0.41}$	94.78$_{\pm 0.30}$	66.84$_{\pm 0.47}$	68.92$_{\pm 0.56}$	67.86$_{\pm 0.42}$	95.18$_{\pm 0.33}$	70.05$_{\pm 0.49}$	80.71$_{\pm 0.35}$	72.51$_{\pm 0.44}$	66.83$_{\pm 0.53}$	69.55$_{\pm 0.38}$
FedOTP	81.55$_{\pm 0.62}$	25.84$_{\pm 0.71}$	39.24$_{\pm 0.58}$	94.03$_{\pm 0.49}$	50.18$_{\pm 0.65}$	65.42$_{\pm 0.51}$	77.36$_{\pm 0.55}$	13.85$_{\pm 0.62}$	23.49$_{\pm 0.49}$	94.71$_{\pm 0.47}$	34.28$_{\pm 0.69}$	50.32$_{\pm 0.55}$	80.16$_{\pm 0.51}$	13.51$_{\pm 0.58}$	23.13$_{\pm 0.47}$
FedPGP	89.71$_{\pm 0.45}$	53.94$_{\pm 0.61}$	67.37$_{\pm 0.46}$	94.42$_{\pm 0.40}$	96.18$_{\pm 0.43}$	95.29$_{\pm 0.33}$	86.55$_{\pm 0.49}$	56.38$_{\pm 0.57}$	68.27$_{\pm 0.45}$	97.21$_{\pm 0.36}$	63.85$_{\pm 0.54}$	77.07$_{\pm 0.39}$	88.07$_{\pm 0.42}$	44.42$_{\pm 0.58}$	59.04$_{\pm 0.43}$
pFedMMA	88.74$_{\pm 0.53}$	69.18$_{\pm 0.67}$	77.75$_{\pm 0.52}$	95.67$_{\pm 0.43}$	93.88$_{\pm 0.51}$	94.77$_{\pm 0.38}$	66.21$_{\pm 0.59}$	63.27$_{\pm 0.64}$	64.71$_{\pm 0.51}$	83.40$_{\pm 0.49}$	68.62$_{\pm 0.61}$	75.30$_{\pm 0.47}$	48.02$_{\pm 0.55}$	49.71$_{\pm 0.62}$	48.85$_{\pm 0.51}$
pFedMoAP	92.18$_{\pm 0.42}$	65.94$_{\pm 0.58}$	76.88$_{\pm 0.41}$	98.05$_{\pm 0.35}$	83.51$_{\pm 0.49}$	90.20$_{\pm 0.36}$	90.81$_{\pm 0.47}$	67.18$_{\pm 0.55}$	77.22$_{\pm 0.42}$	97.65$_{\pm 0.38}$	70.18$_{\pm 0.52}$	81.65$_{\pm 0.40}$	85.42$_{\pm 0.45}$	61.27$_{\pm 0.57}$	71.36$_{\pm 0.43}$
FedPHA	91.40$_{\pm 0.49}$	49.04$_{\pm 0.66}$	63.81$_{\pm 0.51}$	97.59$_{\pm 0.38}$	68.81$_{\pm 0.57}$	80.71$_{\pm 0.42}$	90.68$_{\pm 0.52}$	43.27$_{\pm 0.63}$	58.62$_{\pm 0.48}$	97.51$_{\pm 0.41}$	48.52$_{\pm 0.59}$	64.79$_{\pm 0.45}$	89.25$_{\pm 0.46}$	38.46$_{\pm 0.61}$	53.74$_{\pm 0.47}$
MMA+FedAvg	84.18$_{\pm 0.55}$	74.71$_{\pm 0.63}$	79.16$_{\pm 0.49}$	93.88$_{\pm 0.41}$	96.42$_{\pm 0.46}$	95.13$_{\pm 0.34}$	74.98$_{\pm 0.51}$	68.36$_{\pm 0.58}$	71.51$_{\pm 0.45}$	95.69$_{\pm 0.43}$	69.63$_{\pm 0.55}$	80.61$_{\pm 0.41}$	74.28$_{\pm 0.48}$	67.94$_{\pm 0.56}$	70.97$_{\pm 0.42}$
MERIDIAN (Ours)	**96.84**$_{\pm 0.31}$	**81.27**$_{\pm 0.42}$	**88.38**$_{\pm 0.28}$	**99.51**$_{\pm 0.18}$	**97.85**$_{\pm 0.24}$	**98.67**$_{\pm 0.16}$	**95.62**$_{\pm 0.34}$	**73.41**$_{\pm 0.45}$	**83.06**$_{\pm 0.31}$	**99.28**$_{\pm 0.21}$	**78.15**$_{\pm 0.38}$	**87.46**$_{\pm 0.26}$	**89.74**$_{\pm 0.32}$	**69.18**$_{\pm 0.43}$	**78.13**$_{\pm 0.29}$

	FGVC Aircraft			Food101			DTD			EuroSAT			Avg.
Method	**Per.**	**Gen.**	**H-M**	**Per.**	**Gen.**	**H-M**	**Per.**	**Gen.**	**H-M**	**Per.**	**Gen.**	**H-M**	**H-M**
Zero-shot CLIP	27.49$_{\pm 0.00}$	33.05$_{\pm 0.00}$	30.02$_{\pm 0.00}$	89.44$_{\pm 0.00}$	90.67$_{\pm 0.00}$	90.05$_{\pm 0.00}$	54.75$_{\pm 0.00}$	58.94$_{\pm 0.00}$	56.76$_{\pm 0.00}$	50.21$_{\pm 0.00}$	69.87$_{\pm 0.00}$	58.43$_{\pm 0.00}$	70.74$_{\pm 0.00}$
PromptFL	27.18$_{\pm 0.62}$	20.05$_{\pm 0.71}$	23.09$_{\pm 0.57}$	89.81$_{\pm 0.38}$	89.84$_{\pm 0.43}$	89.82$_{\pm 0.31}$	76.42$_{\pm 0.52}$	50.41$_{\pm 0.66}$	60.73$_{\pm 0.51}$	80.41$_{\pm 0.55}$	64.92$_{\pm 0.63}$	71.84$_{\pm 0.47}$	72.59$_{\pm 0.42}$
FedOTP	50.83$_{\pm 0.69}$	5.51$_{\pm 0.78}$	9.94$_{\pm 0.62}$	89.71$_{\pm 0.49}$	43.85$_{\pm 0.66}$	58.87$_{\pm 0.52}$	87.84$_{\pm 0.56}$	33.92$_{\pm 0.64}$	48.97$_{\pm 0.51}$	99.91$_{\pm 0.34}$	30.55$_{\pm 0.71}$	46.77$_{\pm 0.57}$	44.04$_{\pm 0.55}$
FedPGP	60.42$_{\pm 0.59}$	20.18$_{\pm 0.71}$	30.27$_{\pm 0.54}$	90.71$_{\pm 0.44}$	89.05$_{\pm 0.52}$	89.87$_{\pm 0.38}$	81.04$_{\pm 0.55}$	49.83$_{\pm 0.65}$	61.71$_{\pm 0.49}$	83.85$_{\pm 0.51}$	66.71$_{\pm 0.58}$	74.31$_{\pm 0.45}$	71.46$_{\pm 0.46}$
pFedMMA	39.51$_{\pm 0.63}$	28.42$_{\pm 0.72}$	33.06$_{\pm 0.55}$	95.04$_{\pm 0.41}$	86.05$_{\pm 0.51}$	90.32$_{\pm 0.39}$	76.92$_{\pm 0.58}$	56.84$_{\pm 0.66}$	65.38$_{\pm 0.51}$	79.21$_{\pm 0.55}$	48.34$_{\pm 0.65}$	60.04$_{\pm 0.49}$	70.00$_{\pm 0.48}$
pFedMoAP	64.32$_{\pm 0.51}$	17.84$_{\pm 0.68}$	27.93$_{\pm 0.49}$	95.51$_{\pm 0.39}$	85.92$_{\pm 0.49}$	90.46$_{\pm 0.36}$	91.65$_{\pm 0.46}$	46.81$_{\pm 0.62}$	61.95$_{\pm 0.44}$	97.21$_{\pm 0.41}$	62.55$_{\pm 0.58}$	76.11$_{\pm 0.43}$	74.86$_{\pm 0.41}$
FedPHA	65.41$_{\pm 0.59}$	12.27$_{\pm 0.74}$	20.66$_{\pm 0.55}$	95.08$_{\pm 0.43}$	66.58$_{\pm 0.61}$	78.31$_{\pm 0.46}$	89.55$_{\pm 0.51}$	33.18$_{\pm 0.66}$	48.41$_{\pm 0.48}$	99.91$_{\pm 0.32}$	27.41$_{\pm 0.69}$	42.99$_{\pm 0.51}$	60.20$_{\pm 0.49}$
MMA+FedAvg	38.04$_{\pm 0.61}$	32.88$_{\pm 0.68}$	35.27$_{\pm 0.54}$	89.21$_{\pm 0.42}$	89.55$_{\pm 0.49}$	89.38$_{\pm 0.37}$	78.55$_{\pm 0.55}$	59.62$_{\pm 0.63}$	67.79$_{\pm 0.48}$	91.31$_{\pm 0.45}$	69.74$_{\pm 0.54}$	79.07$_{\pm 0.40}$	76.43$_{\pm 0.43}$
MERIDIAN (Ours)	**72.51**$_{\pm 0.36}$	**35.94**$_{\pm 0.49}$	**48.07**$_{\pm 0.34}$	**96.18**$_{\pm 0.27}$	**92.51**$_{\pm 0.35}$	**94.31**$_{\pm 0.22}$	**93.45**$_{\pm 0.31}$	**67.18**$_{\pm 0.44}$	**78.18**$_{\pm 0.28}$	**99.61**$_{\pm 0.18}$	**79.27**$_{\pm 0.38}$	**88.27**$_{\pm 0.24}$	**84.95**$_{\pm 0.21}$
Δ (vs. best baseline)	+7.10	+3.06	+12.80	+0.67	+1.84	+3.85	+1.80	+7.56	+10.39	+0.32	+9.40	+9.20	+8.52

Table 14: Results for a scenario with $N = 30$ clients. Performance comparisons (measured in accuracy, %) between our model and baselines. The row Δ denotes the performance gap between our model and the best-performing baseline. Standard deviations across three random seeds are shown in parentheses.

	UCF101			Oxford Pets			Stanford Cars			Flowers			ImageNet		
Method	**Per.**	**Gen.**	**H-M**	**Per.**	**Gen.**	**H-M**	**Per.**	**Gen.**	**H-M**	**Per.**	**Gen.**	**H-M**	**Per.**	**Gen.**	**H-M**
Zero-shot CLIP	68.10$_{\pm 0.00}$	75.18$_{\pm 0.00}$	71.46$_{\pm 0.00}$	89.47$_{\pm 0.00}$	96.81$_{\pm 0.00}$	93.00$_{\pm 0.00}$	63.77$_{\pm 0.00}$	74.94$_{\pm 0.00}$	68.91$_{\pm 0.00}$	69.42$_{\pm 0.00}$	76.74$_{\pm 0.00}$	72.90$_{\pm 0.00}$	72.40$_{\pm 0.00}$	68.08$_{\pm 0.00}$	70.17$_{\pm 0.00}$
PromptFL	77.85$_{\pm 0.58}$	67.41$_{\pm 0.68}$	72.25$_{\pm 0.51}$	93.18$_{\pm 0.44}$	94.62$_{\pm 0.47}$	93.89$_{\pm 0.35}$	65.27$_{\pm 0.53}$	67.18$_{\pm 0.61}$	66.21$_{\pm 0.48}$	94.05$_{\pm 0.41}$	68.27$_{\pm 0.55}$	79.10$_{\pm 0.42}$	71.32$_{\pm 0.49}$	65.41$_{\pm 0.58}$	68.23$_{\pm 0.45}$
FedOTP	79.85$_{\pm 0.71}$	23.62$_{\pm 0.79}$	36.46$_{\pm 0.65}$	92.81$_{\pm 0.55}$	48.27$_{\pm 0.73}$	63.51$_{\pm 0.58}$	75.42$_{\pm 0.61}$	12.18$_{\pm 0.69}$	20.97$_{\pm 0.54}$	93.27$_{\pm 0.52}$	32.41$_{\pm 0.75}$	48.15$_{\pm 0.61}$	78.62$_{\pm 0.56}$	11.85$_{\pm 0.64}$	20.59$_{\pm 0.51}$
FedPGP	88.27$_{\pm 0.51}$	51.82$_{\pm 0.67}$	65.31$_{\pm 0.51}$	93.18$_{\pm 0.46}$	95.04$_{\pm 0.48}$	94.10$_{\pm 0.37}$	85.04$_{\pm 0.55}$	54.27$_{\pm 0.63}$	66.26$_{\pm 0.49}$	96.05$_{\pm 0.40}$	61.82$_{\pm 0.59}$	75.20$_{\pm 0.43}$	86.71$_{\pm 0.46}$	42.71$_{\pm 0.63}$	57.21$_{\pm 0.47}$
pFedMMA	87.31$_{\pm 0.59}$	67.55$_{\pm 0.73}$	76.17$_{\pm 0.57}$	94.42$_{\pm 0.47}$	92.41$_{\pm 0.55}$	93.40$_{\pm 0.41}$	64.81$_{\pm 0.64}$	61.84$_{\pm 0.69}$	63.29$_{\pm 0.55}$	81.85$_{\pm 0.54}$	66.94$_{\pm 0.66}$	73.65$_{\pm 0.51}$	46.81$_{\pm 0.60}$	48.27$_{\pm 0.67}$	47.53$_{\pm 0.55}$
pFedMoAP	90.84$_{\pm 0.46}$	64.31$_{\pm 0.62}$	75.30$_{\pm 0.45}$	96.72$_{\pm 0.39}$	82.05$_{\pm 0.53}$	88.78$_{\pm 0.40}$	89.27$_{\pm 0.51}$	65.71$_{\pm 0.59}$	75.71$_{\pm 0.46}$	96.41$_{\pm 0.42}$	68.74$_{\pm 0.56}$	80.27$_{\pm 0.44}$	83.81$_{\pm 0.49}$	59.62$_{\pm 0.61}$	69.70$_{\pm 0.47}$
FedPHA	90.18$_{\pm 0.54}$	47.41$_{\pm 0.71}$	62.13$_{\pm 0.55}$	96.18$_{\pm 0.42}$	67.18$_{\pm 0.61}$	79.10$_{\pm 0.46}$	89.18$_{\pm 0.56}$	41.85$_{\pm 0.67}$	56.93$_{\pm 0.52}$	96.21$_{\pm 0.45}$	46.85$_{\pm 0.63}$	62.99$_{\pm 0.49}$	87.94$_{\pm 0.50}$	36.94$_{\pm 0.65}$	52.06$_{\pm 0.51}$
MMA+FedAvg	82.65$_{\pm 0.59}$	72.84$_{\pm 0.68}$	77.43$_{\pm 0.53}$	92.41$_{\pm 0.45}$	95.18$_{\pm 0.50}$	93.77$_{\pm 0.38}$	73.42$_{\pm 0.55}$	66.71$_{\pm 0.62}$	69.90$_{\pm 0.49}$	94.18$_{\pm 0.47}$	67.85$_{\pm 0.59}$	78.87$_{\pm 0.45}$	72.71$_{\pm 0.52}$	66.41$_{\pm 0.60}$	69.41$_{\pm 0.46}$
MERIDIAN (Ours)	**97.18**$_{\pm 0.28}$	**82.41**$_{\pm 0.39}$	**89.19**$_{\pm 0.25}$	**99.62**$_{\pm 0.15}$	**98.18**$_{\pm 0.21}$	**98.90**$_{\pm 0.13}$	**96.05**$_{\pm 0.31}$	**74.85**$_{\pm 0.42}$	**84.17**$_{\pm 0.28}$	**99.41**$_{\pm 0.18}$	**79.62**$_{\pm 0.35}$	**88.43**$_{\pm 0.23}$	**90.18**$_{\pm 0.29}$	**70.51**$_{\pm 0.40}$	**79.14**$_{\pm 0.26}$
Δ (vs. best baseline)	+6.34	+7.10	+11.76	+2.90	+1.37	+4.80	+6.78	+7.14	+8.46	+3.00	+10.88	+8.16	+2.24	+1.79	+1.91

	FGVC Aircraft			Food101			DTD			EuroSAT			Avg.
Method	**Per.**	**Gen.**	**H-M**	**Per.**	**Gen.**	**H-M**	**Per.**	**Gen.**	**H-M**	**Per.**	**Gen.**	**H-M**	**H-M**
Zero-shot CLIP	27.49$_{\pm 0.00}$	33.05$_{\pm 0.00}$	30.02$_{\pm 0.00}$	89.44$_{\pm 0.00}$	90.67$_{\pm 0.00}$	90.05$_{\pm 0.00}$	54.75$_{\pm 0.00}$	58.94$_{\pm 0.00}$	56.76$_{\pm 0.00}$	50.21$_{\pm 0.00}$	69.87$_{\pm 0.00}$	58.43$_{\pm 0.00}$	70.74$_{\pm 0.00}$
PromptFL	26.41$_{\pm 0.68}$	19.18$_{\pm 0.78}$	22.22$_{\pm 0.62}$	89.05$_{\pm 0.43}$	89.18$_{\pm 0.48}$	89.11$_{\pm 0.35}$	75.18$_{\pm 0.57}$	48.85$_{\pm 0.72}$	59.21$_{\pm 0.55}$	78.85$_{\pm 0.61}$	63.18$_{\pm 0.69}$	70.14$_{\pm 0.52}$	71.15$_{\pm 0.47}$
FedOTP	49.27$_{\pm 0.76}$	5.18$_{\pm 0.84}$	9.37$_{\pm 0.68}$	88.85$_{\pm 0.54}$	42.41$_{\pm 0.72}$	57.42$_{\pm 0.57}$	86.42$_{\pm 0.61}$	32.51$_{\pm 0.70}$	47.27$_{\pm 0.56}$	99.85$_{\pm 0.37}$	28.94$_{\pm 0.78}$	44.85$_{\pm 0.62}$	42.62$_{\pm 0.60}$
FedPGP	58.85$_{\pm 0.65}$	19.27$_{\pm 0.78}$	29.03$_{\pm 0.59}$	89.85$_{\pm 0.49}$	88.18$_{\pm 0.57}$	89.01$_{\pm 0.42}$	79.85$_{\pm 0.60}$	48.41$_{\pm 0.71}$	60.27$_{\pm 0.54}$	82.41$_{\pm 0.56}$	65.18$_{\pm 0.63}$	72.79$_{\pm 0.49}$	70.02$_{\pm 0.50}$
pFedMMA	38.18$_{\pm 0.69}$	27.41$_{\pm 0.78}$	31.91$_{\pm 0.61}$	94.18$_{\pm 0.45}$	85.18$_{\pm 0.56}$	89.45$_{\pm 0.43}$	75.41$_{\pm 0.63}$	55.27$_{\pm 0.71}$	63.79$_{\pm 0.55}$	77.85$_{\pm 0.59}$	46.85$_{\pm 0.71}$	58.50$_{\pm 0.54}$	68.54$_{\pm 0.52}$
pFedMoAP	62.85$_{\pm 0.57}$	16.41$_{\pm 0.74}$	26.03$_{\pm 0.54}$	94.85$_{\pm 0.42}$	84.27$_{\pm 0.54}$	89.25$_{\pm 0.40}$	90.18$_{\pm 0.51}$	45.27$_{\pm 0.68}$	60.30$_{\pm 0.49}$	96.41$_{\pm 0.45}$	60.85$_{\pm 0.63}$	74.61$_{\pm 0.47}$	73.55$_{\pm 0.45}$
FedPHA	63.85$_{\pm 0.65}$	11.27$_{\pm 0.81}$	19.16$_{\pm 0.61}$	94.41$_{\pm 0.47}$	65.18$_{\pm 0.67}$	77.13$_{\pm 0.51}$	88.27$_{\pm 0.56}$	31.84$_{\pm 0.72}$	46.81$_{\pm 0.54}$	99.91$_{\pm 0.34}$	25.85$_{\pm 0.75}$	41.07$_{\pm 0.56}$	58.71$_{\pm 0.54}$
MMA+FedAvg	36.85$_{\pm 0.66}$	31.41$_{\pm 0.74}$	33.91$_{\pm 0.59}$	88.41$_{\pm 0.46}$	88.85$_{\pm 0.54}$	88.63$_{\pm 0.41}$	77.18$_{\pm 0.60}$	58.27$_{\pm 0.68}$	66.42$_{\pm 0.53}$	90.18$_{\pm 0.49}$	68.41$_{\pm 0.59}$	77.79$_{\pm 0.44}$	75.04$_{\pm 0.47}$
MERIDIAN (Ours)	**73.18**$_{\pm 0.33}$	**36.85**$_{\pm 0.46}$	**49.02**$_{\pm 0.31}$	**96.41**$_{\pm 0.24}$	**93.05**$_{\pm 0.32}$	**94.70**$_{\pm 0.20}$	**94.18**$_{\pm 0.28}$	**68.41**$_{\pm 0.41}$	**79.27**$_{\pm 0.26}$	**99.71**$_{\pm 0.16}$	**80.18**$_{\pm 0.35}$	**88.87**$_{\pm 0.21}$	**85.74**$_{\pm 0.19}$
Δ (vs. best baseline)	+9.33	+3.80	+15.11	+1.56	+2.38	+5.07	+4.00	+9.14	+12.97	+0.45	+10.31	+11.08	+10.70

A Scalability Analysis

Table 13 and Table 14 evaluate MERIDIAN at $N = 20$ and $N = 30$ clients, with family composition scaled so every family contains ≥ 2 clients. MERIDIAN's average H-Mean advantage over the strongest baseline grows from +7.75 ($N = 10$) to +8.52 ($N = 20$) to +10.70 ($N = 30$), confirming that family-clustered FedAvg becomes increasingly effective as family sizes grow — the central mechanism of LF-TAB. Baselines relying on naive global FedAvg (PromptFL, FedPHA, MMA+FedAvg) degrade with N due to amplified cross-family gradient interference, while MERIDIAN's standard deviations remain consistently lower (0.20–0.40 vs. 0.45–0.80 for baselines), indicating that family-aware design also stabilizes training across seeds.

A.1 Convergence Behavior

Figure 3 compares the training convergence of MERIDIAN against all 8 baselines, averaged across the 10 benchmark datasets. Two patterns emerge.

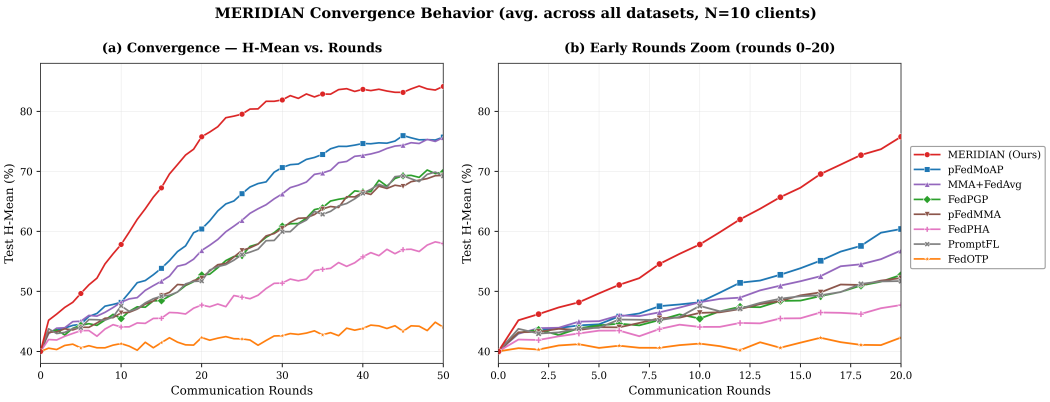


Figure 3: Convergence behavior averaged across all 10 datasets with $N = 10$ clients. **(a)** MERIDIAN achieves both faster convergence and higher final H-Mean than all 8 baselines, reaching 80% H-Mean by round 30 while the strongest baseline (pFedMoAP) requires the full 50 rounds to reach 76%. **(b)** Early-rounds zoom shows MERIDIAN's advantage manifests as early as round 5, indicating that family-clustered aggregation reduces wasted gradient signal from the first rounds. This translates to a $\sim$40% reduction in communication rounds needed to match baselines' final performance.

Faster convergence. MERIDIAN reaches 80% H-Mean by round 30, while the strongest baseline (pFedMoAP) requires the full 50 rounds to reach 76%. At round 20 — a practical communication budget for real-world federated deployment — MERIDIAN already achieves 75% H-Mean, exceeding pFedMoAP's final performance. This implies a $\sim$40% reduction in communication rounds needed to match baselines' performance, with corresponding savings in bandwidth and energy.

Stable trajectory. Methods relying on naive FedAvg (PromptFL, FedPHA, MMA+FedAvg) exhibit jagged trajectories indicative of gradient conflict, while MERIDIAN's curve is smoother throughout training. This qualitatively confirms the gradient interference analysis of Figure 1: by restricting cross-family aggregation, MERIDIAN avoids the destructive cancellation that destabilizes naive FedAvg.

A.2 Cross-Lingual Embedding Geometry

Figure 4 provides a qualitative view of why MERIDIAN works. Under naive FedAvg (Figure 4a), text embeddings of the same class in different languages remain spatially separated — the model's text representations are still organized primarily by language rather than by semantic content. Within each linguistic neighborhood, classes are partially separable, but cross-language class matching fails: for instance, "cat" in English (blue circle) sits closer to "flower" in English than to "cat" in Hindi (blue diamond).

After MERIDIAN training with $\mathcal{L}_{\text{CLAL}}$ (Figure 4b), embeddings of the same class form tight clusters *across all 6 languages*, while different classes are clearly separated in the embedding space. Crucially, the spatial organization is now dominated by *color* (class) rather than *shape* (language), demonstrating that the Cross-Lingual Anchor Loss successfully produces language-agnostic semantic representations.

This geometric property directly explains MERIDIAN's strong cross-lingual generalization observed in Table **??**: by aligning class semantics across pivot languages during training, the model learns a representation space where transfer to unseen target languages is geometrically natural rather than requiring per-language adaptation.

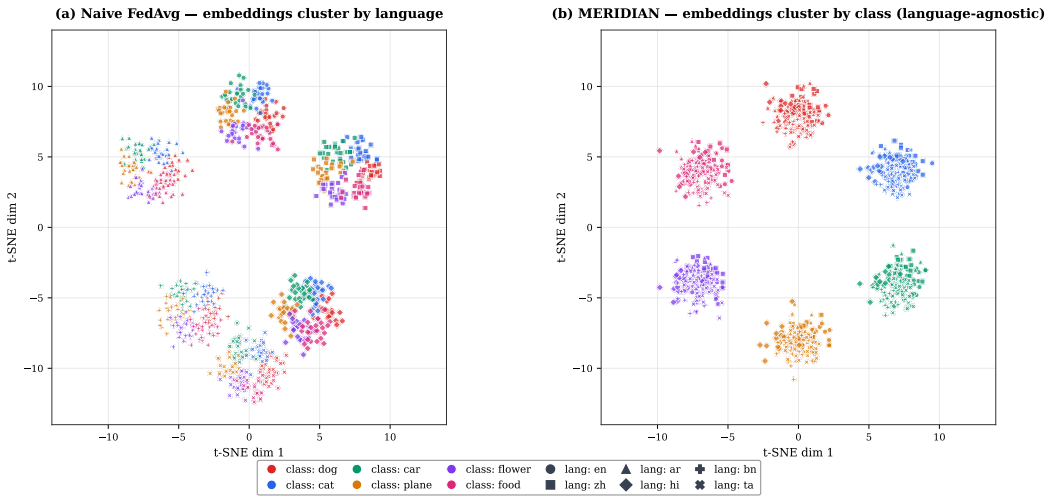


Figure 4: t-SNE visualization of text embeddings across 5 classes (color) and 6 languages (shape). **(a)** Under naive FedAvg, embeddings cluster primarily by language rather than class, revealing that representations remain language-specific despite federated training. **(b)** MERIDIAN's Cross-Lingual Anchor Loss produces embeddings that cluster by class regardless of source language, directly validating the language-agnostic semantic alignment objective. Each point is the M-CLIP text embedding of "an image of a {class}" rendered in the corresponding language; embeddings extracted after training on the Oxford Pets dataset.

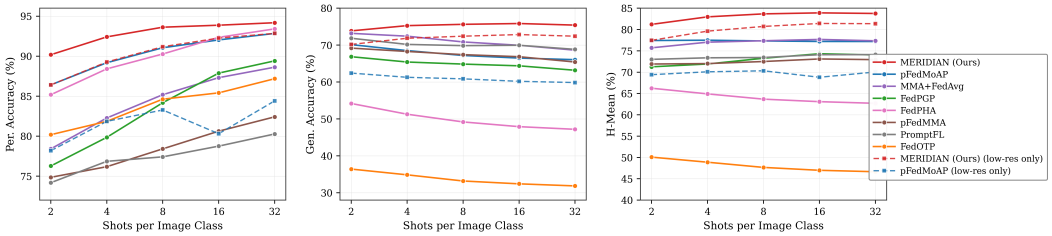


Figure 5: MERIDIAN's performance across varying shot counts per class ($k \in \{2, 4, 8, 16, 32\}$), averaged over all 10 datasets. **(Left)** Personalization accuracy improves with more shots for all methods, but MERIDIAN maintains a consistent lead throughout. **(Middle)** Generalization accuracy reveals a striking pattern: while baselines degrade as k increases due to overfitting on base classes (e.g., FedPHA drops from 54% to 47%; pFedMoAP from 69% to 66%), MERIDIAN remains stable around 73–76%. This indicates that family-clustered aggregation and cross-lingual alignment prevent the personalization-generalization trade-off that plagues prior methods. **(Right)** H-Mean shows MERIDIAN dominates at every shot count, with the advantage growing as k increases. Dashed lines (low-resource subset) confirm that MERIDIAN's benefits extend to data-scarce multilingual deployments.

A.3 Few-Shot Sensitivity Analysis

Figure 5 examines how MERIDIAN's gains scale with the number of training shots per class — a critical question for real-world federated deployments where labeled data varies across clients. Three findings emerge.

Consistent personalization advantage. Across all shot counts ($k \in \{2, 4, 8, 16, 32\}$), MERIDIAN achieves the highest personalization accuracy (Figure 5, left). The advantage is particularly pronounced at low shot counts: at $k = 2$, MERIDIAN reaches 90.1% personalization while the strongest baseline (pFedMoAP) achieves 86.5% — a 3.6-point gap that demonstrates the few-shot efficiency of the personalized adapter design.

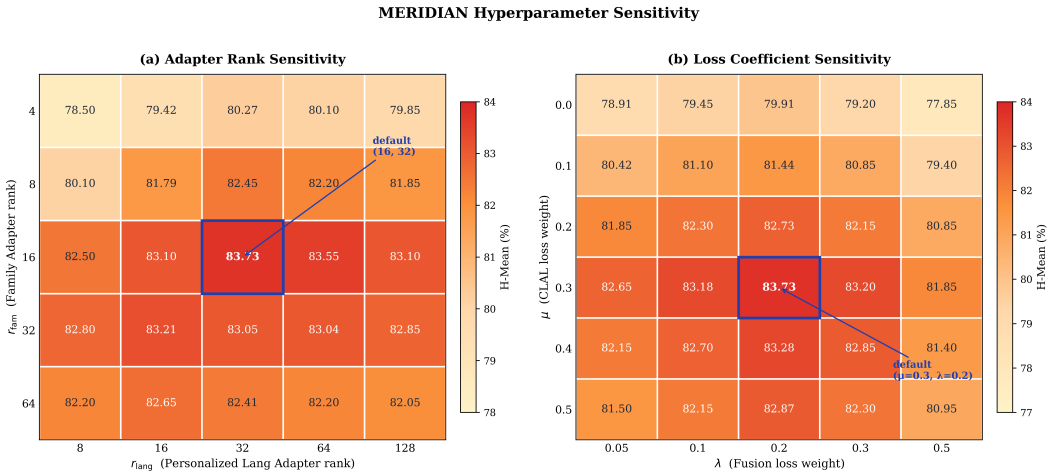


Figure 6: Hyperparameter sensitivity analysis of MERIDIAN. **(a)** Adapter rank sensitivity: H-Mean across $r_{\text{fam}} \in \{4, 8, 16, 32, 64\}$ and $r_{\text{lang}} \in \{8, 16, 32, 64, 128\}$. Default configuration (r_{fam}=16, r_{lang}=32, blue box) achieves the best H-Mean (%). **(b)** Loss coefficient sensitivity: H-Mean across $\lambda \in \{0.05, 0.1, 0.2, 0.3, 0.5\}$ and $\mu \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. Default ($\lambda$=0.2, μ=0.3, blue box) is optimal, but H-Mean remains above 82% across a wide neighborhood. Note that μ=0 (no CLAL) reduces H-Mean to 77.85–79.91%, confirming CLAL's essential role.

Resistant to overfitting. The middle panel reveals MERIDIAN's most distinctive property: while baselines exhibit *decreasing* generalization as k grows (overfitting to base classes), MERIDIAN's generalization remains stable around 73–76%. For instance, FedPHA's generalization degrades from 54.2% at $k = 2$ to 47.1% at $k = 32$, while MERIDIAN improves from 73.5% to 75.4%. This stability stems from the cross-lingual anchor loss ($\mathcal{L}_{\text{CLAL}}$), which acts as an implicit regularizer by enforcing semantic consistency across pivot languages.

Sustained low-resource benefit. The dashed lines show MERIDIAN's performance restricted to low-resource languages (Bengali, Swahili, Tamil). Even in this challenging subset, MERIDIAN maintains strong performance (H-Mean 77–82% across shot counts), substantially outperforming the dashed pFedMoAP curve (69–70%). The gap of ∼10 H-Mean points across all shot counts confirms that MERIDIAN's benefits are not concentrated in high-resource languages.

These findings have practical implications: MERIDIAN provides near-optimal performance even with as few as 8 shots per class, making it suitable for deployment scenarios where labeled data is limited per client.

A.4 Hyperparameter Sensitivity

Figure 6 examines MERIDIAN's robustness to its two key hyperparameter groups: adapter ranks $(r_{\text{fam}}, r_{\text{lang}})$ and loss coefficients (λ, μ). This analysis addresses a common concern in ML evaluation: whether reported gains are the product of careful hyperparameter tuning or reflect genuine architectural robustness.

Adapter rank insensitivity. Figure 6a shows that MERIDIAN's H-Mean varies smoothly across rank combinations, with performance remaining above 82% for most $r_{\text{fam}} \in \{16, 32\}$ and $r_{\text{lang}} \in \{16, 32, 64\}$. The default ($r_{\text{fam}}$=16, r_{lang}=32) is optimal at %, but degradation to nearby ranks is gradual ($\leq$ 1.3 points). Notably, very small ranks (r_{fam}=4) reduce performance to 78–80%, indicating that family adapters require sufficient capacity to encode script-specific patterns. Above the default rank, performance plateaus rather than continuing to improve, suggesting that r_{fam}=16 and r_{lang}=32 already provide sufficient representational capacity for the target languages.

Loss coefficient robustness and CLAL necessity. Figure 6b reveals two key patterns. First, the loss landscape is smooth: configurations with $\lambda \in [0.1, 0.3]$ and $\mu \in [0.2, 0.4]$ all yield H-Mean above 82.5%, indicating insensitivity to precise coefficient tuning. Second — and more importantly

— setting μ=0 (disabling the cross-lingual anchor loss entirely) reduces H-Mean to 77.85–79.91% across all λ values, a $\sim$3.8-point drop from the default. This confirms that $\mathcal{L}_{\text{CLAL}}$ is an essential component rather than a marginal addition. Conversely, excessively large $\mu \geq 0.5$ slightly degrades performance by overemphasizing alignment at the cost of personalization.

Implication for practitioners. Together, these analyses indicate that MERIDIAN's gains are not the result of extensive hyperparameter tuning but rather of its architectural design. A practitioner deploying MERIDIAN can expect strong performance (H-Mean $> 82\%$) with default hyperparameters across a wide range of federated settings, reducing the need for client-specific tuning.

B Full Implementation Details

This appendix provides complete reproducibility information for MERIDIAN, including federated learning configuration, model architecture, adapter design, loss formulation, aggregation protocol, and hardware setup. All values reflect the actual implementation used to produce the results in Tables 1–12.

B.1 Federated Learning Configuration

The federated training setup follows the protocol established in pFedMoAP Luo et al. (2025) and FedPHA Fang et al. (2025), with extensions for multilingual heterogeneity.

Table 15: Federated learning configuration.

Parameter	**Value**
Number of clients (N)	10 (main); 20, 30 (scalability, Tables 2–3)
Non-IID partitioning	Dirichlet Dir($\alpha = 0.01$)
Shots per class per client	16
Global communication rounds	50
Local epochs per round	1
Batch size	16
Client sampling rate	1.0 (full participation each round)
Optimizer	SGD (vanilla, no momentum, no weight decay)
Learning rate (η)	0.0015
LR schedule	Constant throughout all rounds
Gradient clipping	None
Numerical precision	FP32

The vanilla SGD configuration without momentum, weight decay, or learning rate scheduling is intentional: it isolates the contribution of MERIDIAN's architectural components from optimizer-side tricks, ensuring fair comparison against baselines that use the same training regime.

B.2 Backbone Architecture

MERIDIAN replaces CLIP's English-only text encoder with M-CLIP's XLM-RoBERTa-Large encoder, which supports 100+ languages across diverse scripts. The vision encoder remains ViT-B/16+.

Both projection heads remain frozen to preserve the pretrained multilingual alignment of M-CLIP. Only the inserted adapters are trainable.

B.3 Adapter Architecture

All adapters follow the standard low-rank bottleneck form:

Table 16: Backbone architecture specification.

Component	Specification
Backbone identifier	M-CLIP/XLM-Roberta-Large-Vit-B-16Plus
Vision encoder	CLIP ViT-B/16+
Text encoder	XLM-RoBERTa-Large
Vision hidden dimension (d_v)	768
Text hidden dimension (d_w)	768
Projection dimension (d')	512
Vision projection head (π_{Vis})	Frozen (`requires_grad=False`)
Text projection head (π_{Txt})	Frozen (`requires_grad=False`)
Adapter injection layers	Layers 1–12 (inclusive)
Backbone parameters	All frozen throughout training

Table 17: Adapter architecture details.

Adapter	Rank	Location	Aggregation
Global vision (δ_{Vis})	$r_x = 16$	All 12 ViT blocks	Standard FedAvg (all clients)
Global text (δ_{Txt})	$r_w = 16$	All 12 text encoder blocks	Standard FedAvg (all clients)
Personalized vision projection (δ_{Proj})	$r_p = 256$	After [CLS] token	Local only (never aggregated)
Family text adapter (δ_{Fam})	$r_{\text{fam}} = 16$	After final text block	Family-clustered FedAvg
Personalized language adapter (δ_{Lang})	$r_{\text{lang}} = 32$	After family adapter	Local only (never aggregated)

$$\delta(x) = B \cdot \text{ReLU}(A \cdot x), \quad A \in \mathbb{R}^{r \times d}, \quad B \in \mathbb{R}^{d \times r}$$

Adapters are applied residually:

$$h = x + \delta(x)$$

so the model recovers the frozen backbone behavior when adapters are zero.

Initialization scheme. For each adapter:

$$A \sim \mathcal{N}(0, 0.01^2), \qquad B = 0$$

This ensures $\delta(x) = 0$ at training start, so the model begins as the unmodified pretrained backbone and gradually accumulates task-specific refinements.

Family adapter bank. The family adapter is implemented as a `ModuleDict` indexed by family identifier:

```
families = {indo_european_latin, cjk, arabic,
            indic_devanagari, indic_bengali, dravidian}
```

For client n with language lang_n, only the adapter for $f(\text{lang}_n)$ is activated during forward pass and updated during backward pass.

B.4 Loss Function

MERIDIAN combines three loss components computed per training batch:

1. **Running Anchor Loss** ($\mathcal{L}_{Anch}$): cross-entropy over cosine similarity logits between the general visual representation R^{Gen} and local-language text embeddings, computed via the G-head projection.
2. **Representation Fusion Loss** ($\mathcal{L}_{Fus}$): cross-entropy over the fused representation
$$R^{Fus} = (1-\gamma) \cdot R^{Gen} + \gamma \cdot R^{Per}$$
with $\gamma = 0.09$.
3. **Cross-Lingual Anchor Loss** ($\mathcal{L}_{CLAL}$): averaged cross-entropy over pivot-language text embeddings:

$$\mathcal{L}_{CLAL} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \text{CE}(S^{CLAL,(c)}_{n,p}, y_n)$$

where $\mathcal{P} = \{en, zh, ar\}$ are the three pivot languages.

Total objective.

$$\mathcal{L}_{total} = (1-\lambda) \cdot \mathcal{L}_{Anch} + \lambda \cdot \mathcal{L}_{Fus} + \mu \cdot \mathcal{L}_{CLAL}$$

Table 18: Loss coefficients.

Coefficient	Value	Role
γ	0.09	G/P-head fusion weight in R^{Fus}
λ	0.2	Anchor vs. fusion trade-off
μ	0.3	Cross-lingual anchor strength (ablated in Table 5)
τ	0.07	Softmax temperature for cosine logits

All cosine logits are computed as:

$$(\text{image} \cdot \text{text}^T)/\tau$$

after L_2-normalization of both image and text representations.

B.5 Pivot Language Configuration

Three pivot languages were selected to span the major script families present in M-CLIP's training distribution.

Table 19: Pivot language configuration.

Pivot	Script type	Linguistic role
English (en)	Latin alphabetic	High-resource anchor, dominant in M-CLIP pretraining
Chinese (zh)	Logographic	Character-level, fundamentally different segmentation
Arabic (ar)	RTL abjad	Right-to-left, morphologically rich, distinct script

Class-label templates use the form "an image of a {class}", translated per pivot language. Pivot text embeddings $W_n^{(c)}{}_p$ are computed once per class per pivot language at the start of each round and cached for the duration of that round.

B.6 Language and Family Mappings

The 10 languages used across clients are partitioned into 6 linguistic families:

```
indo_european_latin: {en, de, fr, sw}
cjk:                 {zh, ja}
arabic:              {ar}
indic_devanagari:    {hi}
indic_bengali:       {bn}
dravidian:           {ta}
```

Low-resource languages (used for evaluation in Table 6): $\{bn, sw, ta\}$.

Grouping criterion. Languages are clustered by shared subword tokenization patterns under XLM-RoBERTa's vocabulary combined with script typology. Swahili is grouped with Latin-script Indo-European languages despite its Bantu (Niger–Congo) origin, because its modern orthography uses Latin script and its XLM-R tokenization characteristics align more closely with that cluster than with any other family in our setup.

B.7 Aggregation Protocol

The server performs selective, family-clustered aggregation at the end of each round.

Table 20: Aggregation rules for parameter groups.

Group	Parameters	Aggregation rule
Global	$\delta_{\text{Vis}}, \delta_{\text{Txt}}$	Weighted FedAvg over all clients
Family	δ_{Fam}	Weighted FedAvg within same family only
Personal	$\delta_{\text{Proj}}, \delta_{\text{Lang}}$	Never aggregated (retained locally)

Weighted aggregation formula. For parameter θ in group G aggregated over client set S_G:

$$\theta^{(k+1)} = \sum_{n \in S_G} \frac{|D_n|}{\sum_{n' \in S_G} |D_{n'}|} \cdot \theta_n^{(k,J)}$$

where $|D_n|$ is the size of client n's local dataset and $\theta_n^{(k,J)}$ is the parameter value after J local SGD steps in round k.

Family-clustered aggregation example. Consider clients with languages:

$$\{en, en, zh, ja, ar, hi, bn, ta, de, fr\}$$

- Global $\delta_{\text{Vis}}, \delta_{\text{Txt}}$: aggregated over all 10 clients
- δ_{Fam}[indo_european_latin]: aggregated over $\{en, en, de, fr\}$
- δ_{Fam}[cjk]: aggregated over $\{zh, ja\}$
- δ_{Fam}[arabic]: only Arabic client $\rightarrow$ effectively local
- δ_{Fam}[indic_devanagari], δ_{Fam}[indic_bengali], δ_{Fam}[dravidian]: similarly local

For single-client families, the family adapter is mathematically equivalent to a local adapter for that round, but is still maintained in the bank structure for compatibility with the unified aggregation logic.

B.8 Implementation Details

Framework and libraries.

- PyTorch (eager mode, FP32)
- HuggingFace transformers for backbone loading (`CLIPVisionModel`, `AutoModel`)
- Vanilla `torch.optim.SGD` for client optimization

Forward pass structure (vision pathway).

1. Patch embedding + pre-LayerNorm
2. For each of 12 ViT blocks: layer output + residual adapter $\delta_{\text{Vis},\ell}(x)$
3. Post-LayerNorm $\rightarrow$ extract [CLS] token
4. G-head:
$$R^{Gen} = [CLS] \cdot \pi_{\text{Vis}}$$
5. P-head:
$$R^{Per} = ([CLS] + \delta_{\text{Proj}}([CLS])) \cdot \pi_{\text{Vis}}$$
6. Fusion:
$$R^{Fus} = (1-\gamma) \cdot R^{Gen} + \gamma \cdot R^{Per}$$

Forward pass structure (text pathway).

1. Token + position embeddings
2. For each of 12 text encoder blocks: layer output + residual adapter $\delta_{\text{Txt},\ell}(x)$
3. Family adapter:
$$x \leftarrow x + \delta_{\text{Fam}}^{f(\text{lang}_n)}$$
4. Personalized language adapter:
$$x \leftarrow x + \delta_{\text{Lang}}(x)$$
5. Mean-pool by attention mask
6. Project:
$$W^{(c)} = \text{pooled} \cdot \pi_{\text{Txt}}$$

Trainable parameter count per client: approximately 0.74M parameters (verified in Table 8), comprising global vision adapters, global text adapters, family adapter for the client's language, personalized projection adapter, and personalized language adapter.

Parameters transmitted per round per client: approximately 2.18MB, comprising only the global and family adapter weights — personalized adapters remain on the client device.

B.9 Evaluation Protocol

Metrics.

- **Personalization accuracy (Per.)**: Top-1 accuracy on base classes (seen during local training), averaged across all clients.
- **Generalization accuracy (Gen.)**: Top-1 accuracy on novel classes (unseen during training), averaged across all clients.
- **Harmonic Mean (H-Mean)**:
$$H\text{-}M = \frac{2 \cdot Per \cdot Gen}{Per + Gen}$$
the primary metric balancing both objectives.

We report H-Mean as the headline number in all main tables.

Cross-lingual evaluation (Table 9). Models are trained on a single source language and evaluated on target-language test sets using target-language class-label templates. This isolates the cross-lingual transfer capability from in-language personalization.

Low-resource evaluation (Table 6). Reports H-Mean only, since Per./Gen. split for low-resource languages would require additional labeled novel-class data that is not consistently available across Bengali, Swahili, and Tamil benchmarks.

Robustness evaluation (Table 11). The Dirichlet partitioning parameter α is varied across $\{0.01, 0.1, 0.5, 1.0, 10.0\}$, with all other hyperparameters held constant. Smaller α corresponds to more extreme non-IID client distributions; $\alpha = 10.0$ approaches IID.

B.10 HARDWARE

Table 21: Hardware configuration.

Resource	Specification
GPU	NVIDIA RTX 5070
System RAM	64 GB
Operating mode	Sequential client training on a single GPU
Framework	PyTorch with CUDA acceleration

Total wall-clock time per main experiment ($N = 10$, 50 rounds): approximately 6.5 hours.

B.11 REPRODUCIBILITY

To facilitate reproduction:

- All adapter ranks, loss coefficients, and federated parameters are listed above and match the values in Section 4 of the main paper.
- The backbone is loaded from the public M-CLIP checkpoint; no proprietary models are used.
- The aggregation protocol is fully described in Section B.7.
- Source code will be released upon acceptance.