SEARCH WHAT YOU MUST, SOLVE WHAT YOU CAN: FEDERATED CONTINUOUS PROMPT ADAPTATION FOR BLACK-BOX VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated adaptation of pretrained vision–language models (VLMs) commonly relies on lightweight prompts, yet existing approaches typically require gradient or feature access to the frozen backbone. Such white-box access may be unavailable when VLMs are exposed through restricted inference interfaces or opaque runtimes, while distributed users still need to adapt them to private local data. We therefore study federated adaptation of black-box VLMs, where clients can interact with the model through continuous multimodal prompts and class-score feedback but cannot access model internals. Rather than treating black-box adaptation as a single optimization problem, we ask a simpler question: *what must be searched, and what can be solved?* We propose FedMUSE, which decomposes federated black-box VLM adaptation into two complementary components: multimodal prompt search for what must be optimized through queries, and global decision refinement for what can be solved explicitly from the returned score space. FedMUSE first introduces *mean–contrast prompt search*, which parameterizes textual and visual prompts through common and differential latent kernels and performs block-wise zeroth-order estimation to reduce cross-block interference. It then freezes the selected prompt predictor and performs *global residual refinement* in the resulting shared score space. Clients communicate additive sufficient statistics rather than locally fitted heads, allowing the server to recover the pooled residual solution exactly without collecting individual scores or labels. Across 13 classification benchmarks, FedMUSE achieves 70.15% mean accuracy, improving its prompt-only predictor by 5.16 percentage points while adding less than 0.34% training queries, with compact communication and single-call VLM inference. Code is available at https://anonymous.4open.science/r/FedMUSE-04C1.

1 INTRODUCTION

Federated learning (FL) allows multiple data holders to learn a shared predictor without centralizing their training examples (McMahan et al., 2017). Pretrained vision–language models (VLMs), such as CLIP and ALIGN (Radford et al., 2021; Jia et al., 2021), provide this collaboration with transferable representations and a useful starting point when labels are scarce. Parameter-efficient adaptation makes their use practical: PromptFL and pFedPrompt communicate lightweight prompts (Guo et al., 2024; 2023), while FedCLIP adapts pretrained representations through small external modules (Lu et al., 2023a). Subsequent work improves task transfer, personalization, and multimodal cooperation through prompt generation, low-rank adaptation, and grouped text–visual prompts (Qiu et al., 2024; Cui et al., 2024; Bo et al., 2025).

These gains depend on what clients can access. Freezing the backbone reduces the number of trainable parameters, but training input prompts by backpropagation still requires derivatives through the encoders. A locally hosted, opaque runtime may instead accept inputs and return class scores while withholding those derivatives and its internal features. The need for collaboration remains: training data are still distributed, and a client's small, class-skewed sample is still insufficient to define a broadly useful predictor. The resulting problem is therefore to preserve the benefits of federated adaptation when its white-box update mechanism is unavailable. We study this transition under a continuous soft-prompt and class-score interface, rather than assuming compatibility with

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.

every text-only service. To our knowledge, continuous multimodal prompt adaptation for federated black-box VLMs under this interface has not been studied before.

Derivative-free optimization restores a route to local learning (Sun et al., 2022; 2024; Lu et al., 2023b), but a finite query budget makes its design materially different from replacing an exact gradient with an arbitrary estimator. *The first difficulty is the geometry of the searched coordinates.* Equivalent prompt families can induce different directional estimates: the coordinates determine which variables are perturbed together and how one block's gradient contributes to another block's estimation noise. In a multimodal federation, the relative sensitivities of text and visual prompts can also vary with the classes observed by different clients. A modality-wise allocation of a fixed number of queries can therefore devote a substantial part of the budget to estimating an imbalanced pair of signals. This makes the parameterization itself a design variable of the black-box optimizer, even when it introduces no additional expressivity.

The second difficulty is separating what must be searched from what can be solved globally. Once a client receives class scores, a linear decision correction becomes an explicit supervised problem; spending further black-box queries on its coefficients is unnecessary. Yet the correction cannot simply be fitted locally: label-skewed clients observe different score second moments, so averaging their heads does not minimize the pooled objective (Karimireddy et al., 2020; Fanì et al., 2024). Solving the head while prompts continue to move is also unstable as an organization principle, because every prompt update changes both its coordinates and residual targets. This motivates the second stage directly: freeze the shared score map selected after federated prompt search, aggregate additive objective statistics, and solve one global residual rule in that fixed coordinate system.

We propose FedMUSE to address these two difficulties with complementary mechanisms. First, *mean–contrast prompt search* constructs text and visual prompts from common-mode and differential-mode kernels through fixed isometric random lifts. It preserves the latent prompt family while organizing block-wise perturbations around coordinated and opposing changes in the two modalities. The resulting estimator confines each block's local-linear estimation error to its own gradient signal. Its energy identities explain when a near-equal, fixed query allocation is appropriate, motivating a common allocation that we use across datasets without task-specific retuning.

Second, *global residual refinement* reserves explicit global decision learning for a separate stage. The federation freezes a validation-selected prompt predictor, and every client evaluates its reference images in that same predictor's score coordinates. Clients communicate additive objective statistics instead of locally optimized heads, allowing the server to recover the pooled residual solution without collecting individual score vectors or labels. Validation controls the residual strength through an identity path that returns the trained reference predictor's decisions at zero correction. Thus prompt-coordinate aggregation establishes the shared score map, while statistic aggregation combines the heterogeneous class evidence expressed in that map.

Our main contributions are summarized as follows:

- We formulate federated black-box VLM adaptation under continuous multimodal prompts and score-only access, separating query-based prompt search from explicitly solvable global decision learning.
- We develop complementary mechanisms for this separation: mean–contrast coordinates organize finite-budget multimodal estimation, while global residual refinement pools distributed evidence in a frozen score system. We analyze their estimation, aggregation, drift, and deployment properties.
- Across 13 benchmarks, we compare federated text, visual, and multimodal baselines, isolate the main design choices, and account for query, communication, and single-call inference costs.

2 Related Work

2.1 Federated parameter-efficient VLM adaptation

Prompt learning and visual prompting adapt frozen VLMs with compact input parameters (Zhou et al., 2022b;a; Bahng et al., 2022; Jia et al., 2022; Han et al., 2023). Federated extensions exchange shared or personalized prompts, pretrained features, prompt generators, and grouped text–visual

prompts (Guo et al., 2024; 2023; Lu et al., 2023a; Qiu et al., 2024; Cui et al., 2024; Bo et al., 2025). They establish the value of lightweight federated adaptation, but optimize through gradients or internal features. We study a shared predictor trained through continuous input prompts and returned class scores alone.

2.2 Federated learning with black-box models

Zeroth-order estimation supplies the optimization basis for learning from function values (Spall, 1992; Nesterov & Spokoiny, 2017; Liu et al., 2020; Larson et al., 2019; Hansen, 2023). Federated analyses improve convergence, dimension-dependent communication, or curvature use (Ling et al., 2024; Li et al., 2025; 2026; Loyola-González, 2019). FedMUSE builds on these estimators while removing an explicitly solvable decision rule from stochastic search.

Methods compatible with score-only access. Text-side methods include low-dimensional BBT, federated FedBPT, and personalized Fed-DISC (Sun et al., 2022; 2024; Li et al., 2024). Visual methods include BlackVIP and its compact-subspace extension EffiBlackVIP (Oh et al., 2023; Liu et al., 2025). BPTVLM and ZIP provide multimodal or compact prompt search controls (Yu et al., 2023; Park et al., 2025). These mechanisms adapt inputs from returned predictions, but locally optimized updates still face label-skewed client objectives and leave the returned score vector itself underused. ZooPFL adds private embeddings and local score remapping for personalized prediction (Lu et al., 2023b), whereas we require one client-independent rule over the complete vocabulary. FedOne and FedPOB instead optimize discrete prompts (Wang et al., 2025; Lu et al., 2025). FedMUSE complements these search mechanisms with an explicitly solved global decision stage.

Black-box settings with different outputs. FedOT trains over returned feature embeddings, which our score-only interface does not expose (Kong et al., 2026). FDPT uses textual VLM responses and language-agent feedback rather than a fixed class-score vector (Wu et al., 2025). Reprogramming and intrinsic-subspace methods offer related optimization foundations (Tsai et al., 2020; Aghajanyan et al., 2021; Yu et al., 2025), but do not address how heterogeneous clients form a shared rule from class scores.

2.3 Decision refinement in returned score spaces

CraFT motivates residual refinement after centralized black-box prompt adaptation (Wang et al., 2024). We extend this perspective to heterogeneous clients by freezing a common score system and pooling objective statistics instead of averaging local heads. Fed3R establishes statistic-based global ridge fitting for returned features (Fanì et al., 2024); FedMUSE applies the aggregation principle to residual targets constructed only from class scores. Its identity path preserves the selected prompt predictor, and validation controls the deployed correction. Closed-form adaptation in R2D2 and interpolation in WiSE-FT provide complementary precedents (Bertinetto et al., 2019; Wortsman et al., 2022). These ideas support our separation of query-based prompt search from communication-efficient global decision learning.

3 Method

FedMUSE combines *mean–contrast prompt search* with *global residual refinement*. The first component organizes multimodal prompts for block-wise zeroth-order estimation. The second pools heterogeneous client supervision in the fixed score coordinates of the learned predictor. Figure 1 and Algorithm 1 specify the complete procedure.

3.1 Problem setting

Client i holds a private dataset $\mathcal{D}_i$ over a common ordered vocabulary of K classes. The frozen VLM returns $z_\theta(x) \in \mathbb{R}^K$ for image x and continuous input prompts θ. Clients can supply soft prompts and receive class scores, but cannot access backbone gradients, feature embeddings, or weights. We learn a single shared predictor under this interface, using

$$F(\theta) = \sum_{i=1}^{M} w_i \, \mathbb{E}_{(x,y)\sim\mathcal{D}_i} \ell_{\mathrm{CE}}(z_\theta(x), y), \qquad w_i = \frac{|\mathcal{D}_i|}{\sum_j |\mathcal{D}_j|}. \tag{1}$$

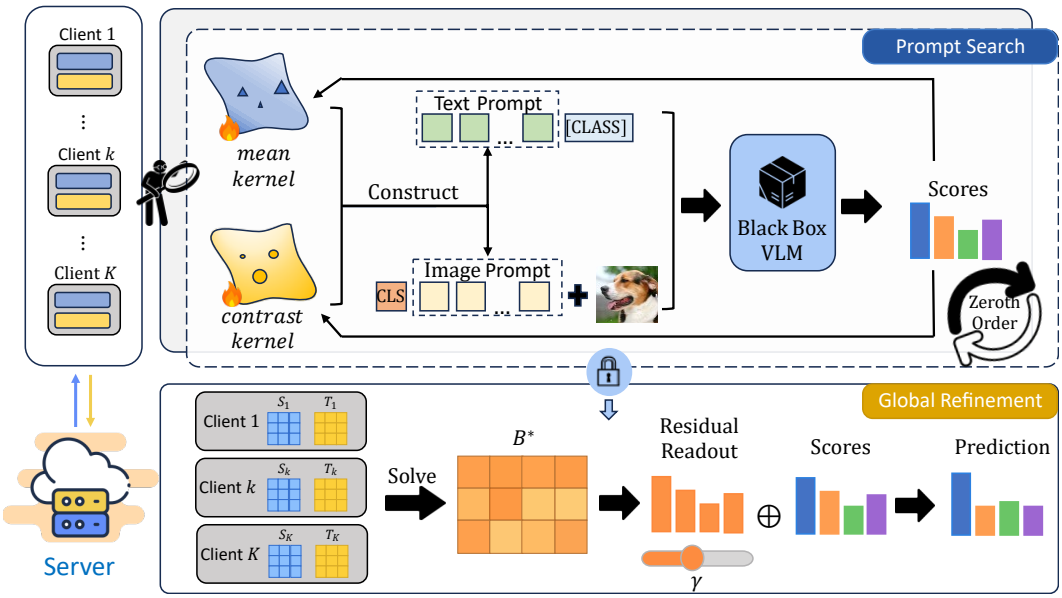


Figure 1: **Overview of FedMUSE.** Clients federatively optimize mean–contrast prompts C, D from black-box scores (top). After freezing the selected predictor, the shared server pools client statistics S_i, T_i, solves $B^\star$, and returns a residual with validation-selected strength γ. Final prediction adds the scaled residual to the reference scores (bottom).

Examples remain local; prompt updates and sufficient statistics are communicated. This data flow does not itself provide a formal privacy guarantee.

3.2 Mean–Contrast Prompt Search

Prompt geometry. Common and differential kernels generate both input prompts:

$$P_T = U(C + D)V_T^\top, \qquad P_V = U(C - D)V_V^\top, \qquad C, D \in \mathbb{R}^{5\times 64}. \tag{2}$$

The fixed, weight-independent orthogonal lifts are $U \in \mathbb{R}^{5\times 5}$, $V_T \in \mathbb{R}^{512\times 64}$, and $V_V \in \mathbb{R}^{768\times 64}$. Their isometry preserves displacement norms across towers of different widths. With $Z_T = C + D$ and $Z_V = C - D$, the invertible transformation preserves the prescribed latent prompt family. It changes the coordinates probed by the optimizer: C induces common changes and D differential changes across modalities.

Zeroth-order estimation. Clients independently perturb each block $b \in \{C, D\}$:

$$\widehat{g}_b = \frac{1}{m_b}\sum_{j=1}^{m_b} \frac{\ell(\theta + c_t\Delta_{b,j}) - \ell(\theta - c_t\Delta_{b,j})}{2c_t}\Delta_{b,j}, \qquad (m_C, m_D) = (3, 2). \tag{3}$$

Each direction has independent equiprobable entries in $\{-1, +1\}$, and directions are independent across probes and blocks. Both estimates are computed at the same parameter point and followed by a joint momentum update. This block-wise form of SPSA (Spall, 1992) uses ten calls per step. The server aggregates the 640 kernel coordinates; validation selects the trained reference predictor $\theta_\star$.

Proposition 1 (Block estimation and allocation)**.** *For a local linear loss, let g_T, g_V be the modality-coordinate gradients, $g_C = g_T + g_V$, $g_D = g_T - g_V$, and $d_b = 320$. The block estimates obey*

$$\mathbb{E}\widehat{g}_b = g_b, \qquad \mathbb{E}\|\widehat{g}_b - g_b\|^2 = \frac{d_b - 1}{m_b}\|g_b\|^2. \tag{4}$$

A joint Rademacher estimator with m directions instead has block error $[(d_b - 1)\|g_b\|^2 + d_b\|g_{b'}\|^2]/m$, where $b' \neq b$. Moreover,

$$\|g_C\|^2 + \|g_D\|^2 = 2(\|g_T\|^2 + \|g_V\|^2), \quad \|g_C\|^2 - \|g_D\|^2 = 4\langle g_T, g_V\rangle. \tag{5}$$

For a fixed total number of directions, the continuous optimal allocation is $m_b \propto \sqrt{d_b - 1}\|g_b\|$.

Appendix A.2 derives these SPSA identities for the prompt coordinates in Eq. (2). Block estimation removes the other block's additive variance term, preventing its gradient from injecting noise into a stationary block. Equation (5) motivates a near-equal allocation when the matched-coordinate inner product is small, including when one modality dominates in norm. We fix the 3/2 allocation across tasks. Its near-optimality is conditional on this energy balance; random lifts alone do not establish it along an adaptive training trajectory. Nonlinear finite differences additionally introduce bias, as discussed in the proof.

3.3 Global residual refinement

Shared score coordinates. After freezing $\theta_\star$, every client evaluates references under the same predictor and projection:

$$u = z_{\theta_\star}(x) - \frac{\mathbf{1}^\top z_{\theta_\star}(x)}{K}\mathbf{1}, \qquad v = \frac{u}{\sqrt{\|u\|^2 + 10^{-12}}}, \qquad \psi = \begin{bmatrix} Rv \\ 1 \end{bmatrix}. \tag{6}$$

Here $R \in \mathbb{R}^{r \times K}$ has orthonormal rows and $r = \min(32, K)$. Normalization preserves class ordering, and the fixed projection bounds readout size. With rows of Φ_i, V_i, Y_i containing $\psi^\top, v^\top$ and one-hot labels, respectively, the global residual objective is

$$J(B) = \tfrac{1}{2}\sum_i \|\Phi_i B - E_i\|_F^2 + \tfrac{\lambda}{2}\|B\|_F^2, \qquad E_i = Y_i - V_i, \quad \lambda = 1. \tag{7}$$

Global pooling. The intercept is regularized. Clients transmit the residual objective's additive sufficient statistics:

$$S_i = \Phi_i^\top \Phi_i, \quad T_i = \Phi_i^\top E_i, \qquad A = \sum_i S_i + \lambda I, \quad T = \sum_i T_i, \quad AB^\star = T. \tag{8}$$

Summing raw statistics preserves the sample weights of the pooled objective. The server solves a system of order $r + 1$ without receiving individual scores or labels.

Theorem 1 (Global solution and heterogeneity gap)**.** *For fixed reference observations, predictor, and projection with* $\lambda > 0$, $B^\star = A^{-1}T$ *is the unique pooled solution and is invariant to client repartition. Let* $N_i > 0$ *count reference observations,* $N = \sum_i N_i$, $\rho_i = N_i/N$, $A_i = S_i + \lambda\rho_i I$, $B_i = A_i^{-1}T_i$, *and* $\bar{B} = \sum_i \rho_i B_i$. *Then*

$$\Xi = \sum_i (A_i - \rho_i A)(B_i - \bar{B}), \tag{9}$$

$$J(\bar{B}) - J(B^\star) = \tfrac{1}{2}\|A^{1/2}(\bar{B} - B^\star)\|_F^2 = \tfrac{1}{2}\|A^{-1/2}\Xi\|_F^2. \tag{10}$$

For $\mathcal{E} = \operatorname{tr}(T^\top A^{-1} T)$ *and* $\gamma \in [0, 1]$,

$$J(\gamma B^\star) = J(0) - \gamma(1 - \gamma/2)\mathcal{E}. \tag{11}$$

The theorem identifies the federated mismatch that parameter averaging leaves unresolved: local predictor disagreement interacts with differences in score second moments. Aggregating the objective eliminates that gap, without assuming homogeneous clients. This extends the pooled-ridge construction (Fanì et al., 2024) to residual targets; Appendix A.4 gives the proof, and Section 4.3 tests the corresponding local-readout replacement. Fixed coordinates also remove prompt-induced changes in the head's design matrix and targets; their perturbation bound is derived in Appendix A.5.

Residual readout and strength selection. The reference contributes an identity path:

$$f_\gamma(x) = v(z_{\theta_\star}(x)) + \gamma(B^\star)^\top \psi(z_{\theta_\star}(x)), \qquad \Gamma = \{0, .25, .5, .75, 1\}. \tag{12}$$

If $\widehat{G}(\gamma)$ denotes validation accuracy gain over $\gamma = 0$ on n examples, we select

$$\widetilde{\gamma} = \min \arg\max_{\gamma \in \Gamma} \widehat{G}(\gamma), \qquad \widehat{\gamma} = \begin{cases} \widetilde{\gamma}, & \widehat{G}(\widetilde{\gamma}) \geq \tau, \\ 0, & \text{otherwise}, \end{cases} \quad \tau = 2/n. \tag{13}$$

Zero strength preserves $\theta_\star$'s class decisions; all strengths reuse cached validation scores.

Theorem 2 (Selection risk)**.** *Let a finite family of L checkpoints and training-fitted readouts be independent of the n independent validation observations. Define the average expected classification risk $\mathcal{R}_{\ell,\gamma}$ and, for $\delta \in (0,1)$,*

$$\varepsilon_n = \sqrt{2\log(2L(|\Gamma|-1)/\delta)/n}. \tag{14}$$

With probability at least $1-\delta$, any validation-selected checkpoint $\widehat{\ell}$ and Eq. (13) *satisfy*

$$\mathcal{R}_{\widehat{\ell},\widehat{\gamma}} \le \min_{\gamma\in\Gamma} \mathcal{R}_{\widehat{\ell},\gamma} + 2\varepsilon_n + \tau. \tag{15}$$

For $\widehat{\gamma} > 0$, $\mathcal{R}_{\widehat{\ell},\widehat{\gamma}} - \mathcal{R}_{\widehat{\ell},0} \le \varepsilon_n - \tau$; for $\widehat{\gamma} = 0$ the difference is zero.

The finite-family argument (Maillard et al., 2021) quantifies the trade-off: the threshold demands evidence for adopting a correction but can reject useful small gains. It accounts for checkpoint reuse through L; it does not certify test non-degradation at a fixed two-example threshold. The target is the validation distribution, with class-prior shift and adaptive method development outside the stated bound. The proof is in Appendix A.6, and the gating ablations assess its empirical cost. Refinement adds $(r+1)K$ learned coefficients without enlarging the zeroth-order search; inference requires one VLM call and an affine readout.

Separation as an organization choice. One possible organization interleaves the two forms of adaptation: refit the pooled residual each round, then continue prompt search under the corrected predictor's square loss. We call this strategy INTERLEAVED. Its score map and decision rule evolve together. FedMUSE instead completes cross-entropy prompt search before freezing the score map, pooling residual statistics, and selecting deployment strength. Appendix A.8 specifies the interleaved objective; Section 4.3 compares these organization choices.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and metrics. We evaluate CLIP ViT-B/16 (Radford et al., 2021) on 13 classification datasets: Caltech101 (Fei-Fei et al., 2004) and Mini-ImageNet (Vinyals et al., 2016) for object recognition; Oxford Pets (Parkhi et al., 2012), Flowers102 (Nilsback & Zisserman, 2008), Food101 (Bossard et al., 2014), FGVC Aircraft (Maji et al., 2013), and Stanford Cars (Krause et al., 2013) for fine-grained recognition; DTD (Cimpoi et al., 2014) and SUN397 (Xiao et al., 2010) for textures and scenes; EuroSAT (Helber et al., 2019) and RESISC45 (Cheng et al., 2017) for remote sensing; and UCF101 images (Soomro et al., 2012) and CLEVR counting (Johnson et al., 2017) for actions and synthetic recognition. The main setting is 16-shot pathological partitioning over 20 clients. We report three-seed mean top-1 accuracy and an equal-weight dataset average; appendix tables include sample standard deviations. Training and selection details are in Appendix B.1.

Baselines. PromptFL (Guo et al., 2024) is the gradient-based white-box reference. FedBBT (Sun et al., 2022) optimizes text prompts by evolutionary search; FedZIP (Park et al., 2025) provides a zeroth-order text-prompt alternative. FedBlackVIP (Oh et al., 2023) learns input-dependent visual reprogramming, and FedEffiBlackVIP (Liu et al., 2025) restricts visual search to a subspace. FedBPTVLM (Yu et al., 2023) jointly optimizes visual and textual prompts. These federated transfers control the local adaptation mechanism. Native federated alternatives include FedBPT (Sun et al., 2024) and Fed-DISC (Li et al., 2024). Together with the centralized-method transfers, they cover text, visual, and multimodal search as well as federated optimization. This comparison tests whether a different local prompting mechanism alone can combine class evidence distributed across clients.

4.2 MAIN RESULTS

FedMUSE achieves 70.15% mean accuracy, exceeding the unified FedBPT result by 7.58 percentage points (pp) and its prompt-only reference by 5.16 pp. It comes within 1.39 pp of white-box PromptFL's 71.54%, closing approximately 79% of the gap between its prompt reference and that gradient-based comparison. Thus restricted score access need not leave a large adaptation gap when prompt search is complemented by global decision learning.

Table 1: **Main comparison.** Three-seed mean accuracy (%) on 13 datasets. V/T denote visual/text prompts; ZS/WB denote zero-shot/white-box references. Bold/underline denote best/second-best listed black-box results. Gain is the final method minus each baseline. Protocols are detailed in Appendix B.1.

Method	Modality	Caltech	Pets	CLEVR	Flowers	Food	Aircraft	SUN397	DTD	EuroSAT	Cars	RESISC	UCF	Mini-IN	Avg.	Gain
CLIP (ZS)	-	92.37	88.17	14.69	67.03	83.76	23.91	62.57	43.09	48.65	65.51	57.94	63.66	85.76	61.32	—
PromptFL (WB)	T	94.32	92.52	25.01	87.95	86.58	29.61	69.13	61.80	76.21	69.60	76.44	69.85	91.01	71.54	—
FedBlackVIP	V	92.24	88.85	17.60	67.33	83.95	23.38	63.66	42.59	52.48	65.46	59.46	63.96	85.85	62.06	+8.09
FedEffiBlackVIP	V	92.32	87.14	15.48	66.61	83.34	22.87	61.98	41.74	44.60	64.51	57.07	63.64	85.40	60.52	+9.63
FedBBT	T	88.72	87.31	13.87	70.02	85.03	22.90	62.74	40.07	51.26	64.17	57.90	60.99	83.26	60.64	+9.51
Fed-DISC	T	88.78	87.83	18.94	68.43	85.88	22.28	61.62	40.74	54.00	65.55	56.64	62.50	82.68	61.22	+8.93
FedZIP	T	89.32	84.23	16.58	59.78	83.62	14.17	62.67	38.34	50.04	62.55	53.06	57.42	82.95	58.06	+12.09
FedBPT	T	90.99	87.37	25.52	70.19	85.48	23.45	64.70	39.20	53.77	63.80	58.59	64.59	85.67	62.56	+7.58
FedBPTVLM	V&T	88.61	85.50	14.66	68.28	84.63	23.29	58.79	38.22	38.29	63.26	52.95	61.21	81.86	58.43	+11.72
Ours (FedMUSE)	V&T	**92.67**	**92.58**	**26.03**	**85.44**	**86.41**	**31.24**	**68.96**	**58.49**	**68.37**	**68.89**	**75.11**	**68.88**	**88.82**	**70.15**	Ref.

The gains concentrate on Flowers, DTD, EuroSAT, and RESISC45, where the reference leaves substantial classification errors to correct. Twelve dataset means improve and CLEVR ties. Food101 changes little because its reference is already strong; refinement serves as a correction mechanism rather than uniformly replacing the learned predictor. A task-level transition breakdown confirms that the large gains arise from a surplus of corrected errors rather than indiscriminate prediction changes (Appendix B.10).

The full method remains effective with fewer labels and different client partitions. Four- and eight-shot accuracy reaches 67.53% and 68.84%, respectively, compared with 70.15% at 16 shots. Under IID and Dirichlet $\alpha = 0.1/0.5$, accuracy reaches 70.70% and 70.51/70.87%, within 0.73 pp of the main partition. Appendix B resolves these averages into task-level supervision and partition effects.

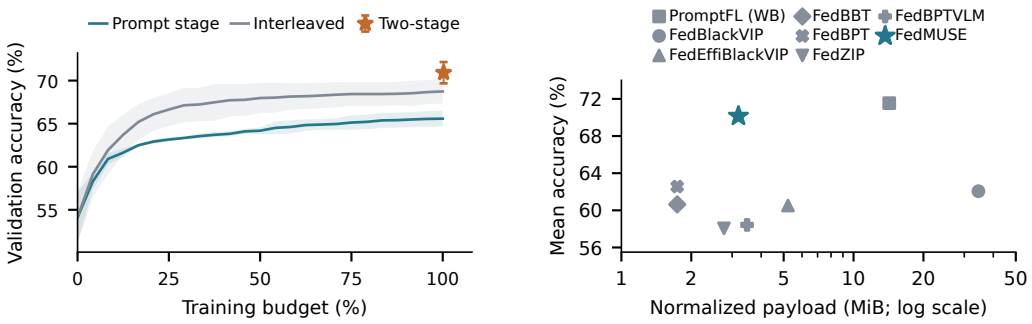
4.3 Ablation

Component ablation. The component comparison in Table 2 separates the two principal components from the choices internal to residual refinement. Removing prompt search applies the identical global residual stage to zero-shot CLIP scores and reaches 69.16%, 0.99 pp below FedMUSE. Thus explicit score-space learning accounts for much of the adaptation gain, while the learned prompt map supplies a further improvement. Removing the residual readout instead returns the 64.99% learned prompt predictor. Replacing global statistics with independently fitted local heads gives 66.02%, 4.12 pp below the complete method. Because the local penalties are sample-weight matched, this deficit supports pooling objective statistics before solving; the corresponding objective mismatch is analyzed against its classification consequence in Appendix B.7.

Table 2: **Component ablation.** Three-seed mean accuracy scores across all 13 datasets. Unindented rows remove a principal architectural component; prefixed rows modify choices within residual refinement. Performance changes are measured relative to the full method.

Configuration	Mean (%)	Δ (pp)
FedMUSE	70.15	Ref.
w/o residual readout	64.99	-5.16
– w/o global pooling	66.02	-4.12
– w/o strength selection	68.36	-1.79
– w/o two-example threshold	70.51	+0.36
w/o prompt search	65.16	-4.99

Attaching FedMUSE's global residual refinement to FedBPT raises its mean from 62.56% to 68.28%, demonstrating that the proposed decision stage transfers across prompt predictors. The remaining 1.86 pp gap to FedMUSE and a matched four-task coordinate–estimator comparison expose the contribution and limits of prompt search (Appendices B.3 and B.4).



(a) Adaptation organization.

(b) Communication.

Figure 2: **Organization and communication of adaptation.** (a) Cumulative best validation accuracy for the prompt stage and interleaved fitting; the two-stage point adds one final global refinement. (b) Mean test accuracy versus normalized parameter/statistic payload under the common 24-round, 20-client schedule. Bands and whiskers are sample SD over three seed-wise task averages.

Fixing $\gamma = 1$ gives 68.36%, whereas validation-based strength selection gives 70.15%. The fitted residual therefore benefits from deployment control: a correction that fits residual labels can also overturn correct reference decisions. Removing only the two-example threshold raises the mean to 70.51%. We nevertheless retain the threshold as a conservative, pre-test adoption rule: a nonzero residual must correct at least two additional validation examples before it can replace the identity path. This trades 0.36 pp in observed mean accuracy for requiring discrete evidence before changing a trained reference, rather than choosing the protocol that looks best on the test set. The strength trade-off is examined below, while task-level component effects are reported in Appendix B.8.

Search and refinement organization. The INTERLEAVED strategy refits a pooled head during prompt search and optimizes the corrected predictor with square loss. It reaches 66.88%, compared with 70.15% when prompt learning and global decision refinement are separated. The comparison isolates an organization choice: interleaving changes both the search objective and the score system in which the head is fitted. Mechanism controls in Appendix B.5 isolate prompt learning, head tracking, aggregation, and explicit solution; they also retain CLEVR as a counterexample where interleaving is stronger.

The validation trajectories in Figure 2(a) clarify the distinction. Prompt learning forms the first stage of the complete method. Interleaved fitting improves that trajectory, but continually coordinates a head with changing prompts. The separated strategy performs one final pooled solve after checkpoint selection, raising mean validation accuracy from 65.58% to 70.91% with only 0.24% additional training calls.

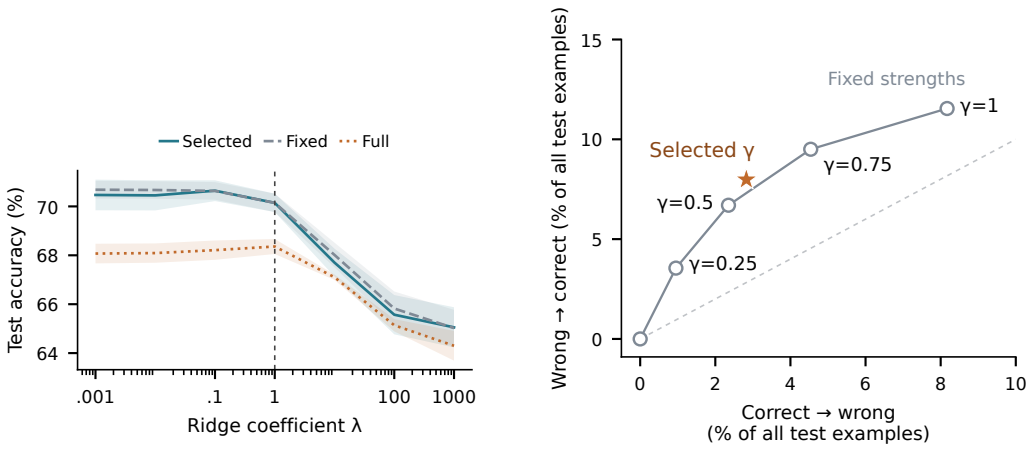
Communication cost. The recurring messages contain 640 float32 prompt coordinates. For R rounds, M broadcasts and A uploads per round, a p-coordinate update has normalized payload

$$\mathcal{C}_{\text{prompt}}(p) = 4R(M + A)p/2^{20} \quad \text{MiB}. \tag{16}$$

With $R = 24$, $M = 20$, $A = 18$, and $p = 640$, prompt search uses 2.23 MiB. Broadcasting the selected reference once, uploading sufficient statistics once, and returning the final head raise the complete cost to 2.30–5.44 MiB across tasks, averaging 3.19 MiB. This yields the accuracy–payload position in Figure 2(b); backbone and control traffic are omitted consistently. Repeating head synchronization every round would instead average 24.24 MiB. Refinement adds only 40–400 batched model calls, or 0.17–0.33% of prompt-training calls, and inference still uses one VLM call. Complete formulas and task-level costs are given in Appendices A.9 and B.11.

4.4 SENSITIVITY ANALYSIS

Ridge regularization. Figure 3(a) varies λ over seven logarithmically spaced values and separates three effects. The selected curve refits the head and reselects γ on validation; the fixed curve retains



(a) Ridge regularization.

(b) Deployment strength.

Figure 3: **Sensitivity of residual deployment.** (a) Test accuracy under validation-reselected, fixed, or full strength as the ridge coefficient varies. (b) Wrong-to-correct versus correct-to-wrong transitions for fixed strengths; the star uses validation selection. The dashed diagonal marks equal correction and disruption.

the strength chosen at $\lambda = 1$; the full curve always uses $\gamma = 1$. Validation reselection changes by less than 0.50 pp over $\lambda \in [0.001, 1]$: $\lambda = 0.1$ reaches 70.64%, while the common unit default reaches 70.15%. This plateau shows that the refinement is not tied to one finely tuned small coefficient. The unit default supplies an eigenvalue floor in normalized-score coordinates and avoids task-specific regularization choices. Beyond this range, stronger penalties shrink useful class corrections and drive all deployment rules toward the reference predictor. Appendix B.12 resolves the macro-average into individual task responses.

Deployment strength. Figure 3(b) follows the prediction transitions induced by $\gamma \in \{0, .25, .5, .75, 1\}$. The horizontal axis counts reference predictions changed from correct to wrong, and the vertical axis counts errors corrected; points above the diagonal yield a positive net gain. Increasing γ initially produces corrections faster than disruptions, but the trajectory bends toward the diagonal as the residual becomes stronger. At $\gamma = 1$, 11.54% of examples are corrected and 8.17% are disrupted, for a 3.37 pp surplus. Dataset-wise validation selection instead yields 7.98% corrections and only 2.83% disruptions, increasing the surplus to 5.16 pp. Strength selection therefore controls how far predictions move along this trade-off, while the two-example threshold makes the separate decision of whether the evidence is sufficient to leave the identity path at all.

5 Limitations

Evidence is limited to CLIP ViT-B/16 and an interface accepting continuous multimodal prompts and complete ordered class scores; discrete prompts, textual feedback, and feature-returning services require different adaptation mechanisms. Global refinement further assumes a shared vocabulary, representative client statistics, and held-out validation. Strength selection cannot guarantee test non-degradation, and exchanging sufficient statistics is not a formal privacy mechanism.

6 Conclusion

FedMUSE separates query-based multimodal prompt search from globally solvable decision refinement. Mean–contrast search organizes a fixed perturbation budget, while frozen-coordinate residual fitting pools heterogeneous client evidence without searching the readout. Across 13 benchmarks, the resulting predictor reaches 70.15% mean accuracy with compact communication and single-call inference. Component results show that global score-space learning supplies most of the gain and prompt search provides a further improvement.

AI USE STATEMENT

Generative AI tools were used to assist in writing portions of proofs, proposing and refining research hypotheses, partially implementing methods, and translating manuscript text. The authors checked AI-assisted mathematical arguments against explicit derivations and numerical identity tests, and reviewed and tested AI-assisted code through controlled runs and result audits. Hypotheses and methodological suggestions were retained only after author review and empirical evaluation. No AI-generated synthetic data were used. The authors reviewed all AI-assisted material and take responsibility for the final text, claims, code, and artifacts in this submission.

REFERENCES

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7319–7328, Online, 2021. Association for Computational Linguistics.

Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models, 2022.

Luca Bertinetto, João F. Henriques, Philip H. S. Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*. ICLR, 2019.

Weihao Bo, Yanpeng Sun, Yu Wang, Xinyu Zhang, and Zechao Li. FedMGP: Personalized federated learning with multi-group text-visual prompts. In *Advances in Neural Information Processing Systems*, volume 38, pp. 51792–51826. Curran Associates, Inc., 2025.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 446–461, Cham, 2014. Springer International Publishing.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3606–3613, Columbus, OH, USA, 2014. IEEE.

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. PMLR, 2024.

Eros Fanì, Raffaello Camoriano, Barbara Caputo, and Marco Ciccone. Accelerating heterogeneous federated learning with closed-form classifiers. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 13029–13048. PMLR, 2024.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pp. 178–178, Washington D.C., USA, 2004. IEEE.

Tao Guo, Song Guo, and Junxiao Wang. pfedprompt: Learning personalized prompt for vision-language models in federated learning. In *Proceedings of the ACM Web Conference 2023*, pp. 1364–1374, New York, NY, USA, 2023. Association for Computing Machinery.

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.

Cheng Han, Qifan Wang, Yiming Cui, Zhiwen Cao, Wenguan Wang, Siyuan Qi, and Dongfang Liu. E2vpt: An effective and efficient approach for visual prompt tuning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17445–17456, Paris, France, 2023. IEEE.

Nikolaus Hansen. The cma evolution strategy: A tutorial, 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*, 12(7):2217–2226, 2019.

Magnus R. Hestenes and Eduard Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49(6):409–436, 1952.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, pp. 4904–4916, Online, 2021. PMLR.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Computer Vision - ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXIII*, pp. 709–727, Berlin, Heidelberg, 2022. Springer-Verlag.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2901–2910, Honolulu, HI, USA, 2017. IEEE.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020.

Eun Gyung Kong, Jewon Yeom, Yonghoon Jeon, and Taesup Kim. Generalized and personalized federated learning with black-box foundation models via orthogonal transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24567–24576. IEEE, 2026.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pp. 554–561, Sydney, Australia, 2013. IEEE.

Jeffrey Larson, Matt Menickelly, and Stefan M Wild. Derivative-free optimization methods. *Acta Numerica*, 28(1):287–404, 2019.

Gang Li et al. Personalized federated learning for text classification with gradient-free prompt tuning. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 4597–4612, Mexico City, Mexico, 2024. ACL.

Zhe Li, Bicheng Ying, Zidong Liu, Chaosheng Dong, and Haibo Yang. Achieving dimension-free communication in federated learning via zeroth-order optimization. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, pp. 23769–23803, Singapore EXPO, 2025.

Zhe Li, Bicheng Ying, Zidong Liu, Chaosheng Dong, and Haibo Yang. Converge faster, talk less: Hessian-informed federated zeroth-order optimization. In *International Conference on Learning Representations*, volume 2026, pp. 37066–37099. ICLR, 2026.

Zhenqing Ling, Daoyuan Chen, Liuyi Yao, Yaliang Li, and Ying Shen. On the convergence of zeroth-order federated tuning for large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1827–1838. Association for Computing Machinery, 2024.

Sijia Liu, Pin-Yu Chen, Bhavya Kailkhura, Gaoyuan Zhang, Alfred O. Hero III, and Pramod K. Varshney. A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications. *IEEE Signal Processing Magazine*, 37(5):43–54, 2020.

Zhaogeng Liu, Haozhen Zhang, Hualin Zhang, Xingchen Li, Wanli Shi, Bin Gu, and Yi Chang. Query efficient black-box visual prompting with subspace learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 4322–4331, Nashville, Tennessee, USA, 2025. IEEE.

Octavio Loyola-González. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, 7:154096–154113, 2019. doi: 10.1109/ACCESS.2019.2949286.

Pingchen Lu, Zhi Hong, Zhiwei Shang, Zhiyong Wang, Yikun Ban, Yao Shu, Min Zhang, Shuang Qiu, and Zhongxiang Dai. FedPOB: Sample-efficient federated prompt optimization via bandits. *arXiv preprint arXiv:2509.24701*, 2025.

Wang Lu, Xixu Hu, Jindong Wang, and Xing Xie. Fedclip: Fast generalization and personalization for clip in federated learning, 2023a.

Wang Lu, Hao Yu, Jindong Wang, Damien Teney, Haohan Wang, Yiqiang Chen, Qiang Yang, Xing Xie, and Xiangyang Ji. ZooPFL: Exploring black-box foundation models for personalized federated learning. *arXiv preprint arXiv:2310.05143*, 2023b.

Guillaume Maillard, Sylvain Arlot, and Matthieu Lerasle. Aggregated hold-out. *Journal of Machine Learning Research*, 22(20):1–55, 2021.

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

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 1273–1282, Fort Lauderdale, FL, USA, 2017. PMLR.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729, Bhubaneswar, India, 2008. ACM.

Changdae Oh, Hyeji Hwang, Hyung-Sin Lee, YS Lim, Geon Jung, Jiyoung Jung, Hosik Choi, and Kyungwoo Song. Blackvip: Black-box visual prompting for robust transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24224–24235, Vancouver, Canada, 2023. IEEE.

Seonghwan Park, Jaehyeon Jeong, Yongjun Kim, Jaeho Lee, and Namhoon Lee. Zip: An efficient zeroth-order prompt tuning for black-box vision-language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, pp. 62988–63021, Singapore EXPO, 2025. ICLR.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, Providence, RI, USA, 2012. IEEE.

Chen Qiu, Xingyu Li, Chaithanya Kumar Mummadi, Madan Ganesh, Zhenzhen Li, Lu Peng, and Wan-Yi Lin. Federated text-driven prompt generation for vision-language models. In *International Conference on Learning Representations*, volume 2024, pp. 12997–13012. ICLR, 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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, Online, 2021. PMLR.

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

James C Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992.

Jingwei Sun, Ziyue Xu, Hongxu Yin, Dong Yang, Daguang Xu, Yudong Liu, Zhixu Du, Yiran Chen, and Holger R. Roth. FedBPT: Efficient federated black-box prompt tuning for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 47159–47173. PMLR, 2024.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. Black-box tuning for language-model-as-a-service. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, pp. 20841–20855, Baltimore, Maryland, 2022. PMLR.

Yun-Yun Tsai, Pin-Yu Chen, and Tsung-Yi Ho. Transfer learning without knowing: Reprogramming black-box machine learning models with scarce data and limited resources. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 9614–9624, Online, 2020. JMLR.org.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, Barcelona, Spain, 2016. Curran Associates, Inc.

Ganyu Wang, Jinjie Fang, Maxwell Juncheng Yin, Bin Gu, Xi Chen, Boyu Wang, Yi Chang, and Charles Ling. FedOne: Query-efficient federated learning for black-box discrete prompt learning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, pp. 62857–62889, Vancouver, Canada, 2025. PMLR.

Zhengbo Wang, Jian Liang, Ran He, Zilei Wang, and Tieniu Tan. Connecting the dots: Collaborative fine-tuning for black-box vision-language models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 50931–50943. PMLR, 2024.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7959–7971. IEEE, 2022.

Jiaqi Wu, Simin Chen, Jing Tang, Yuzhe Yang, Yiming Chen, Lixu Wang, Song Lin, Zehua Wang, Wei Chen, and Zijian Tian. FDPT: Federated discrete prompt tuning for black-box visual-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2461–2470. IEEE, 2025.

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

Lang Yu, Qin Chen, Jiaju Lin, and Liang He. Black-box prompt tuning for vision-language model as a service. In Edith Elkind (ed.), *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 1686–1694, Macao, P.R.China, 2023. IJCAI Organization.

Ziming Yu, Chang Yang, Jianshu Wu, and Sijia Liu. Zeroth-order fine-tuning of LLMs in random subspaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4475–4485, Honolulu, HI, USA, 2025. IEEE.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16816–16825, 2022a.

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, 2022b.

A Method details and mathematical properties

A.1 Exact implementation and data flow

The text and visual prompts are inserted before the first frozen Transformer block; no deep prompts are used. The entries of C and D are initialized with standard deviations 0.02 and 0.01. Orthogonal lifts use a dedicated random generator and are shared across clients. The second-stage projection uses a CPU generator seeded by a deterministic hash of the ordered class names, followed by reduced QR on a $K \times r$ Gaussian matrix. Its construction is independent of model weights and data.

For the cumulative local step index $t_i \geq 1$ of client i, the schedules below use $t = t_i$:

$$c_t = 0.005t^{-0.2}, \quad a_t = 0.01(t+1)^{-0.4}, \quad h_{b,t} = 0.9h_{b,t-1} + 0.1\widehat{g}_{b,t}, \quad \theta_{b,t+1} = \theta_{b,t} - a_t h_{b,t}. \tag{17}$$

Step counters and momentum persist across rounds. The configuration learning-rate field of 0.001 is not the actual SPSA gain. The local training batch size is 32, using full batches when possible and at least one batch for a small client. All clients train before the 18-client subset is chosen for aggregation; averaging renormalizes sample weights within the chosen subset. Validation chooses among 24 post-aggregation checkpoints, retaining later ones on ties.

Second-stage references use two independently augmented views per training image, with a separate random-number stream. Statistics are accumulated in float64. In addition to S_i, T_i, the implementation records residual energy, the sum of normalization factors, and observation count for diagnostics. These auxiliary scalars do not change Eq. (7). CG starts from zero, targets relative residual 10^{-10}, and runs for at most $4(r+1)$ iterations; a relative residual above 10^{-8} fails the numerical check. The intercept is regularized. Because the objective sums losses, λ/N is the equivalent regularization coefficient for an average-loss objective with N reference observations.

The validation set contains four examples per class and is disjoint from the training references. It is reused for checkpoint and strength selection. The implementation evaluates candidate strengths using one cached validation score pass, writes the selected strength and validation counts to a selection record, and then queries the test set. Test labels never fit B or choose γ. The original base, selected, and full-strength predictions can subsequently be compared on the same test cache. Sensitivity analyses reuse those scores for evaluation without changing the frozen main protocol.

A.2 Mean–contrast coordinates and finite-budget estimation

Latent family and inverse map. Let $p = 5 \cdot 64 = 320$, $z_T = \mathrm{vec}(Z_T)$, $z_V = \mathrm{vec}(Z_V)$, and $q = (\mathrm{vec}(C), \mathrm{vec}(D))$. The same vectorization order is used in both modalities. Equation (2) gives

$$\begin{bmatrix} z_T \\ z_V \end{bmatrix} = \mathcal{H}q, \qquad \mathcal{H} = \begin{bmatrix} I_p & I_p \\ I_p & -I_p \end{bmatrix}, \qquad \mathcal{H}^\top \mathcal{H} = 2I_{2p}. \tag{18}$$

Multiplying by $\frac{1}{2}\mathcal{H}^\top$ recovers q. In matrix form, $C = (Z_T + Z_V)/2$ and $D = (Z_T - Z_V)/2$. Every pair of latent text and visual kernels therefore has one mean–contrast representation. The equivalence concerns the prescribed lifted family, not arbitrary prompts outside those subspaces.

Isometry of each lift. For either embedding lift V and any latent displacement Z, orthonormal columns give $U^\top U = I_5$ and $V^\top V = I_{64}$. By the Frobenius trace identity and cyclic permutation

Algorithm 1 FedMUSE: server coordination and client computation

Require: Local data $\{\mathcal{D}_i\}$, sample counts $\{n_i\}$, validation set, public lifts and projection

function SERVER
  Initialize global kernels $\theta = (C, D)$
  for round $q = 1, \ldots, 24$ **do**
    Broadcast $\langle \text{SEARCH}, \theta \rangle$ to all clients
    After local updates complete, select $\mathcal{A}_q$ from all clients
    Request and receive $\{\theta_i : i \in \mathcal{A}_q\}$
    $\theta \leftarrow \sum_{i \in \mathcal{A}_q} \frac{n_i}{\sum_{j \in \mathcal{A}_q} n_j} \theta_i$
    Evaluate validation accuracy and retain the best $\theta_\star$
  end for
  Broadcast $\langle \text{REFINE}, \theta_\star \rangle$ to all clients
  Receive each client's sufficient statistics (S_i, T_i)
  Solve $(\sum_i S_i + I)B^\star = \sum_i T_i$
  Select $\widehat{\gamma}$ by Eq. (13); record before test
  Broadcast $(B^\star, \widehat{\gamma})$ for prediction via Eq. (12)
end function

function CLIENT($i, \langle \text{phase}, \theta \rangle$)
  if phase $=$ SEARCH **then**
    Initialize $\theta_i \leftarrow \theta$; restore local momentum and step counter
    for local minibatch $s = 1, \ldots, 5|\mathcal{L}_i|$ **do**
      Take the next minibatch; advance the persistent counter $t_i \leftarrow t_i + 1$
      Estimate both blocks at θ_i using c_{t_i}, then jointly update using a_{t_i}
    end for
    Retain optimizer state locally
    Send only $\theta_i = (C_i, D_i)$ when requested by the server
  else
    Freeze received $\theta_\star$; query two augmented views per local image
    Construct Φ_i, E_i using the public projection and Eq. (6)
    Send $S_i = \Phi_i^\top \Phi_i,\ T_i = \Phi_i^\top E_i$ to the server
  end if
end function

of the trace,

$$\|UZV^\top\|_F^2 = \operatorname{tr}(VZ^\top U^\top UZV^\top) \tag{19}$$

$$= \operatorname{tr}(Z^\top ZV^\top V) = \operatorname{tr}(Z^\top Z) = \|Z\|_F^2. \tag{20}$$

The inverse on the lifted subspace is $Z = U^\top (UZV^\top)V$. Thus differing encoder widths do not rescale the displacement of a latent kernel.

Gradient transformation. Write $G_T = \nabla_{P_T} \ell$ and $G_V = \nabla_{P_V} \ell$ for mathematical prompt derivatives; the algorithm never requests them from the VLM. Their latent-coordinate versions are $g_T = \operatorname{vec}(U^\top G_T V_T)$ and $g_V = \operatorname{vec}(U^\top G_V V_V)$. Using $\langle X, Y \rangle_F = \operatorname{tr}(X^\top Y)$, the differential of the loss is

$$d\ell = \langle G_T, U(dC + dD)V_T^\top \rangle_F + \langle G_V, U(dC - dD)V_V^\top \rangle_F \tag{21}$$

$$= \langle U^\top G_T V_T + U^\top G_V V_V, dC \rangle_F \tag{22}$$

$$+ \langle U^\top G_T V_T - U^\top G_V V_V, dD \rangle_F. \tag{23}$$

The coefficients of dC and dD yield $g_C = g_T + g_V$ and $g_D = g_T - g_V$. Expanding their squared norms gives

$$\|g_C\|^2 = \|g_T\|^2 + \|g_V\|^2 + 2\langle g_T, g_V \rangle, \tag{24}$$

$$\|g_D\|^2 = \|g_T\|^2 + \|g_V\|^2 - 2\langle g_T, g_V \rangle. \tag{25}$$

Adding and subtracting these lines proves Eq. (5).

Exact-gradient comparison. For $z = \mathcal{H}q$, the chain rule gives $\nabla_q \ell(\mathcal{H}q) = \mathcal{H}^\top \nabla_z \ell(z)$. A Euclidean gradient step of size η in q induces

$$z^+ = \mathcal{H}(q - \eta \mathcal{H}^\top \nabla_z \ell) = z - 2\eta \nabla_z \ell. \tag{26}$$

Its latent trajectory agrees with ordinary gradient descent after matching initializations and doubling the latent learning rate. Finite-budget perturbation estimators need not share this equivalence: transforming the coordinates also transforms the directions sampled by the optimizer.

Rademacher directional estimate. Condition on the current parameters and local minibatch, so the loss and its gradient are fixed. For block b, assume an affine loss in a neighborhood of the probed point. If $\Delta_b \in \{-1, 1\}^{d_b}$ has independent equiprobable entries, the two-point SPSA estimate (Spall, 1992) becomes

$$\frac{\ell(\theta + c\Delta_b) - \ell(\theta - c\Delta_b)}{2c} = g_b^\top \Delta_b, \qquad \widehat{g}_b = (g_b^\top \Delta_b)\Delta_b. \tag{27}$$

Here Δ_b is embedded only in block b; the other block is held fixed. The affine assumption makes the central difference exact for every nonzero c.

Expectation and squared error. For coordinate k, $\Delta_{b,k}^2 = 1$, so

$$(\widehat{g}_b)_k = (g_b)_k + \sum_{\ell \neq k} (g_b)_\ell \Delta_{b,\ell} \Delta_{b,k}. \tag{28}$$

Independence and zero means imply $\mathbb{E}[\Delta_{b,\ell}\Delta_{b,k}] = 0$ for $\ell \neq k$, proving $\mathbb{E}\widehat{g}_b = g_b$. For the error of coordinate k, expand the square before taking expectation:

$$\mathbb{E}[(\widehat{g}_b - g_b)_k^2] = \sum_{\ell \neq k} \sum_{s \neq k} (g_b)_\ell (g_b)_s \, \mathbb{E}[\Delta_{b,\ell}\Delta_{b,s}\Delta_{b,k}^2] \tag{29}$$

$$= \sum_{\ell \neq k} (g_b)_\ell^2. \tag{30}$$

The terms with $\ell \neq s$ vanish because those entries are independent with zero mean; the terms with $\ell = s$ equal one. Summing over k counts each $(g_b)_\ell^2$ exactly $d_b - 1$ times:

$$\mathbb{E}\|\widehat{g}_b - g_b\|^2 = \sum_k \sum_{\ell \neq k} (g_b)_\ell^2 = (d_b - 1)\|g_b\|^2. \tag{31}$$

Averaging independent probes. Let $e_j = \widehat{g}_b^{(j)} - g_b$ for m_b independent directions. Their means vanish, and independence gives $\mathbb{E}[e_j^\top e_s] = 0$ for $j \neq s$. Therefore

$$\mathbb{E}\left\| \frac{1}{m_b} \sum_j e_j \right\|^2 = \frac{1}{m_b^2} \sum_j \mathbb{E}\|e_j\|^2 + \frac{1}{m_b^2} \sum_{j \neq s} \mathbb{E}[e_j^\top e_s] \tag{32}$$

$$= \frac{d_b - 1}{m_b} \|g_b\|^2. \tag{33}$$

This proves Eq. (4). The full error norm is the sum of the two block error norms because their coordinates are disjoint.

Cross-block noise in joint perturbation. Joint perturbation yields

$$\widehat{g}_b^{\text{joint}} = (g_b^\top \Delta_b)\Delta_b + (g_{b'}^\top \Delta_{b'})\Delta_b. \tag{34}$$

Set $\xi = g_{b'}^\top \Delta_{b'}$. It is independent of Δ_b, with $\mathbb{E}\xi = 0$ and $\mathbb{E}\xi^2 = \|g_{b'}\|^2$. The cross term between the first block's error and $\xi\Delta_b$ consequently has zero expectation. Since $\|\Delta_b\|^2 = d_b$,

$$\mathbb{E}\|\widehat{g}_b^{\text{joint}} - g_b\|^2 = (d_b - 1)\|g_b\|^2 + \mathbb{E}\xi^2 \, \mathbb{E}\|\Delta_b\|^2 \tag{35}$$

$$= (d_b - 1)\|g_b\|^2 + d_b\|g_{b'}\|^2. \tag{36}$$

Averaging m independent joint directions divides both terms by m. In the affine model, block perturbation therefore prevents the other block's gradient from creating error in a stationary block.

Allocation under a fixed direction budget. Let $a_b = (d_b - 1)\|g_b\|^2$ and initially allow positive real allocations with $\sum_b m_b = m$. The total error is $\sum_b a_b/m_b$. For positive a_b, the Lagrangian is

$$\mathcal{L} = \sum_b \frac{a_b}{m_b} + \nu \left(\sum_b m_b - m \right).$$

Its stationarity equations give $-a_b/m_b^2 + \nu = 0$, hence $m_b = \sqrt{a_b/\nu}$. Enforcing the budget yields

$$\sqrt{\nu} = \frac{\sum_b \sqrt{a_b}}{m}, \qquad m_b = \frac{m\sqrt{a_b}}{\sum_s \sqrt{a_s}}, \qquad \min \sum_b \frac{a_b}{m_b} = \frac{(\sum_b \sqrt{a_b})^2}{m}. \tag{37}$$

This proves the allocation statement in Proposition 1. A zero-gradient block contributes zero affine-model error and can be omitted in the continuous limit. Actual numbers of probes are integers, so feasible splits must be compared after rounding.

Balanced energies and the 3/2 split. Let $\mathcal{G} = \|g_T\|^2 + \|g_V\|^2 > 0$. If $|\langle g_T, g_V \rangle| \le \epsilon \mathcal{G}$ for $\epsilon < 1/2$, the two energy expansions give

$$(1 - 2\epsilon)\mathcal{G} \le \|g_b\|^2 \le (1 + 2\epsilon)\mathcal{G} \quad (b = C, D).$$

Dividing the lower bound of one block by the upper bound of the other gives

$$\frac{1 - 2\epsilon}{1 + 2\epsilon} \le \frac{\|g_C\|^2}{\|g_D\|^2} \le \frac{1 + 2\epsilon}{1 - 2\epsilon}. \tag{38}$$

With equal block dimensions, the optimal allocation ratio is the square root of this energy ratio. Exact balance makes 3/2 and 2/3 the two optimal positive integer splits of five directions. If one modality dominates, Cauchy–Schwarz gives $|\langle g_T, g_V \rangle|/\mathcal{G} \le \rho/(1 + \rho^2)$ for $\rho = \|g_V\|/\|g_T\|$, so the balance condition also follows as ρ tends to zero. Random lifts alone do not establish this condition along a learned trajectory. The fixed 3/2 split avoids estimating unknown gradient norms with additional queries.

Total-error comparison at equal budget. For equal dimensions $d_b = p$, let $u = \|g_C\|$ and $w = \|g_D\|$. The continuous optimal block error is $(p - 1)(u + w)^2/m$, while joint perturbation has total error $(2p - 1)(u^2 + w^2)/m$. Their difference is

$$\frac{p(u - w)^2 + 2uw}{m} \ge 0. \tag{39}$$

This comparison uses the gradient-dependent continuous optimum. An integer, fixed split can lose that advantage after rounding. Proposition 1 therefore identifies noise containment and a conditional allocation rule, rather than claiming uniformly smaller total error for the fixed 3/2 split.

Nonlinear finite differences. For a nonlinear loss, define $\varphi(t) = \ell(\theta + t\Delta_b)$ with $\varphi'(0) = g_b^\top \Delta_b$. Assume the third derivative tensor of ℓ has operator norm at most L_3 along every possible probed segment. Taylor's theorem at c and $-c$ gives

$$\frac{\varphi(c) - \varphi(-c)}{2c} = \varphi'(0) + \frac{c^2}{12}\big(\varphi'''(\xi_+) + \varphi'''(\xi_-)\big), \tag{40}$$

$$\left| \frac{\varphi(c) - \varphi(-c)}{2c} - g_b^\top \Delta_b \right| \le \frac{L_3 c^2}{6} \|\Delta_b\|^3 = \frac{L_3 c^2 d_b^{3/2}}{6}. \tag{41}$$

Multiplying by Δ_b bounds the estimator remainder norm by $L_3 c^2 d_b^2/6$. The affine identities omit this finite-step term and any additional oracle noise. This distinguishes the exact local-linear calculation from general zeroth-order smooth-loss analysis (Spall, 1992; Nesterov & Spokoiny, 2017).

Exchange-symmetric curvature. For a conditional quadratic illustration, suppose the latent Hessian is $\mathcal{K} = \begin{bmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{B} & \mathcal{A} \end{bmatrix}$ with symmetric $\mathcal{A}, \mathcal{B}$. Set $Q = \mathcal{H}/\sqrt{2}$. Direct block multiplication gives

$$\mathcal{K}Q = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathcal{A} + \mathcal{B} & \mathcal{A} - \mathcal{B} \\ \mathcal{A} + \mathcal{B} & -(\mathcal{A} - \mathcal{B}) \end{bmatrix}, \tag{42}$$

$$Q^\top \mathcal{K} Q = \operatorname{diag}(\mathcal{A} + \mathcal{B}, \mathcal{A} - \mathcal{B}). \tag{43}$$

The unnormalized kernel coordinates have Hessian $\mathcal{H}^\top \mathcal{K} \mathcal{H} = 2\operatorname{diag}(\mathcal{A} + \mathcal{B}, \mathcal{A} - \mathcal{B})$. Common and differential directions separate under this symmetry assumption; the actual VLM towers need not satisfy it.

A.3 Score coordinates and reference preservation

Centering and normalization. For a raw score vector z, let $u = z - \bar{z}\mathbf{1}$ with $\bar{z} = \mathbf{1}^\top z/K$, and let $s = \sqrt{\|u\|^2 + \eta_0}$ with $\eta_0 = 10^{-12}$. Then $s > 0$ and, for every pair of classes,

$$v_j - v_k = \frac{u_j - u_k}{s} = \frac{z_j - z_k}{s}. \tag{44}$$

Thus class ordering and ties are unchanged by Eq. (6). Furthermore, $\|v\|^2 = \|u\|^2/(\|u\|^2+\eta_0) \le 1$. The bound gives a common scale for fitting residual labels while preserving the reference decisions.

Shared projection and intercept. Since $RR^\top = I_r$, R has spectral norm one and $\|Rv\| \le \|v\|$. For $\psi = [Rv; 1]$, it follows that $\|\psi\|^2 = \|Rv\|^2 + 1 \le 2$. Projection keeps the linear system at order $d = r + 1 \le 33$. The appended constant allows a class-dependent offset, and regularizing its row makes the stated penalty apply to every learned coefficient. A shared checkpoint, class order, and R make all uploaded statistics components of one pooled objective.

Residual targets and deployment. For one-hot label y, set $e = y - v$. Fitting $\psi^\top B$ to $e^\top$ gives the corrected scores $v + B^\top\psi$. The deployed scores can also be written

$$f_\gamma = (1-\gamma)v + \gamma(v + B^\top\psi). \tag{45}$$

For $\gamma \in [0, 1]$, this interpolates between the reference and full correction. These are decision scores, not assumed calibrated probabilities. At $\gamma = 0$, Eq. (44) recovers the raw reference prediction under the same tie-breaking rule.

A sufficient decision-margin condition. Suppose the reference has unique winning class $k_\star$ and margin $m = v_{k_\star} - \max_{j\neq k_\star} v_j > 0$. Write $a = B^\top\psi$. For each other class j,

$$(v_{k_\star} + \gamma a_{k_\star}) - (v_j + \gamma a_j) \ge m - \gamma(|a_{k_\star}| + |a_j|) \tag{46}$$

$$\ge m - 2\gamma\|a\|_\infty. \tag{47}$$

Hence $m > 2\gamma\|a\|_\infty$ preserves that prediction. The condition explains how strength limits boundary disruption; it preserves a reference decision whether that decision was correct or wrong. The population effect of validation-based strength selection is analyzed separately in Theorem 2.

A.4 Proof of Theorem 1

Dimensions and additive objective statistics. Let $d = r+1$, let $N_i > 0$ be the number of reference observations on client i, and set $N = \sum_i N_i$. Two views per training image give $N_i = 2|\mathcal{D}_i|$, so $\rho_i = N_i/N$ equals the training-image weight. Each client has $\Phi_i \in \mathbb{R}^{N_i\times d}$, $E_i \in \mathbb{R}^{N_i\times K}$, and a candidate head $B \in \mathbb{R}^{d\times K}$. If observation j supplies column vectors ψ_{ij} and e_{ij}, matrix multiplication gives

$$S_i = \sum_{j=1}^{N_i} \psi_{ij}\psi_{ij}^\top, \qquad T_i = \sum_{j=1}^{N_i} \psi_{ij}e_{ij}^\top. \tag{48}$$

These statistics retain the quadratic and linear terms of the pooled residual objective. The constant residual energy is needed to evaluate the objective value, but not to determine its minimizer. This uses the classical ridge construction (Hoerl & Kennard, 1970) and the distributed-statistic principle of Fed3R (Fanì et al., 2024) in the shared residual-score coordinates.

Quadratic expansion and stationary point. For each client, expand the squared Frobenius norm:

$$\|\Phi_i B - E_i\|_F^2 = \operatorname{tr}\big((\Phi_i B - E_i)^\top(\Phi_i B - E_i)\big) \tag{49}$$

$$= \operatorname{tr}(B^\top S_i B) - 2\operatorname{tr}(B^\top T_i) + \|E_i\|_F^2. \tag{50}$$

Writing $S = \sum_i S_i$, $T = \sum_i T_i$, $A = S + \lambda I_d$, and $c = \frac{1}{2}\sum_i \|E_i\|_F^2$ gives

$$J(B) = c - \operatorname{tr}(B^\top T) + \tfrac{1}{2}\operatorname{tr}(B^\top AB), \qquad \nabla_B J = AB - T. \tag{51}$$

The derivative follows by differentiating each column's symmetric quadratic form. For any $x \in \mathbb{R}^d$,

$$x^\top Ax = \sum_i \|\Phi_i x\|^2 + \lambda\|x\|^2 \ge \lambda\|x\|^2. \tag{52}$$

Thus A is positive definite when $\lambda > 0$, including when the score features are rank deficient. Along any nonzero matrix direction H, the second derivative is $\operatorname{tr}(H^\top AH) \geq \lambda \|H\|_F^2 > 0$. Strict convexity and the zero-gradient equation therefore give the unique minimizer $B^\star = A^{-1}T$.

Partition invariance and sample weighting. Equation (48) sums the same outer products regardless of their client assignments. Keeping observations, score map, labels, and projection fixed leaves S, T, and hence $B^\star$, unchanged after repartition. No matrix inverse is taken until after summation. Averaging losses instead of summing would divide the data terms by N; the equivalent averaged objective must therefore use penalty λ/N. Summing unnormalized statistics implements the stated pooled objective without giving a small client the weight of a large one. This invariance concerns the fixed-reference fit; changing a training partition can also change the learned prompt predictor.

Excess objective around the optimum. Let $H = \widehat{B} - B^\star$ for any candidate head. Substitute $B^\star + H$ into Eq. (51) and subtract $J(B^\star)$:

$$J(B^\star + H) - J(B^\star) = \operatorname{tr}\big(H^\top (AB^\star - T)\big) + \tfrac{1}{2}\operatorname{tr}(H^\top AH) \tag{53}$$

$$= \tfrac{1}{2}\operatorname{tr}(H^\top AH) = \tfrac{1}{2}\|A^{1/2}H\|_F^2. \tag{54}$$

The first term vanishes because $AB^\star = T$. The last equality uses $\|A^{1/2}H\|_F^2 = \operatorname{tr}(H^\top AH)$. Positive definiteness makes this excess nonnegative, with equality only at the pooled solution.

Locally fitted heads and the heterogeneity residual. For a matched distribution of the global penalty, set $A_i = S_i + \lambda \rho_i I_d$ and $B_i = A_i^{-1} T_i$. Each A_i is positive definite because $\rho_i > 0$, and $\sum_i A_i = A$. The averaged local head $\bar{B} = \sum_i \rho_i B_i$ satisfies $\sum_i \rho_i (B_i - \bar{B}) = 0$. Using these identities, expand the residual in Eq. (9):

$$\Xi = \sum_i (A_i - \rho_i A)(B_i - \bar{B}) \tag{55}$$

$$= \sum_i A_i (B_i - \bar{B}) - A \sum_i \rho_i (B_i - \bar{B}) \tag{56}$$

$$= \sum_i A_i B_i - \left(\sum_i A_i \right) \bar{B} = T - A\bar{B} \tag{57}$$

$$= A(B^\star - \bar{B}). \tag{58}$$

Consequently $\bar{B} - B^\star = -A^{-1}\Xi$. Inserting this into Eq. (54) gives

$$J(\bar{B}) - J(B^\star) = \tfrac{1}{2}\operatorname{tr}(\Xi^\top A^{-1}\Xi) = \tfrac{1}{2}\|A^{-1/2}\Xi\|_F^2, \tag{59}$$

which proves Eq. (10). Moreover,

$$A_i - \rho_i A = S_i - \rho_i \sum_j S_j = N_i \left(\frac{S_i}{N_i} - \frac{S}{N} \right). \tag{60}$$

The regularizers cancel; the remaining factor is a difference in empirical score second moments. Its product with $B_i - \bar{B}$ explains how local geometry and local predictor disagreement combine. Either identical second moments per observation or identical local heads makes $\Xi = 0$; cancellation can also make it zero when neither factor is identical.

Scalar illustration of local inversion. Consider two equal-weight scalar quadratic problems with $S_1 = 1$, $S_2 = 9$, $T_1 = 1$, $T_2 = 0$, and $\lambda = 1$. Then $A_1 = 3/2$, $A_2 = 19/2$, $B_1 = 2/3$, and $B_2 = 0$, so $\bar{B} = 1/3$. The pooled system instead has $A = 11$, $T = 1$, and $B^\star = 1/11$. Equation (54) gives excess $\frac{1}{2} \cdot 11(1/3 - 1/11)^2 = 32/99$. The penalty fractions match exactly; inversion before aggregation causes the mismatch.

Residual-strength ray. The normal equation implies

$$\operatorname{tr}((B^\star)^\top AB^\star) = \operatorname{tr}((B^\star)^\top T) = \operatorname{tr}(T^\top A^{-1}T) = \mathcal{E} \geq 0. \tag{61}$$

Applying Eq. (51) to $\gamma B^\star$ yields

$$J(\gamma B^\star) = c - \gamma \mathcal{E} + \tfrac{1}{2}\gamma^2 \mathcal{E} \tag{62}$$

$$= J(0) - \gamma(1 - \gamma/2)\mathcal{E}. \tag{63}$$

This proves Eq. (11). For $\gamma \in [0,1]$, the factor $\gamma(1-\gamma/2)$ is nonnegative, and the derivative is $-(1-\gamma)\mathcal{E} \le 0$. Thus full strength minimizes this training-objective ray. Classification accuracy depends on score ordering, so its best deployment strength need not be the ray's training optimum.

CG solution and numerical error. Conjugate gradients solves the positive-definite system without explicitly forming A^{-1} (Hestenes & Stiefel, 1952). For one column with right-hand side t, initialize $b_0 = 0$, $r_0 = t$, $p_0 = r_0$. Each iteration uses

$$\alpha_s = \frac{r_s^\top r_s}{p_s^\top A p_s}, \qquad b_{s+1} = b_s + \alpha_s p_s, \qquad r_{s+1} = r_s - \alpha_s A p_s, \tag{64}$$

$$\beta_s = \frac{r_{s+1}^\top r_{s+1}}{r_s^\top r_s}, \qquad p_{s+1} = r_{s+1} + \beta_s p_s. \tag{65}$$

The columns share A but have their own steps and residuals. A zero right-hand side has solution zero and needs no iteration. Stopping uses the full matrix Frobenius residual ratio defined in Appendix A.

For any computed matrix $\widehat{B}$, write $\mathcal{Q} = A\widehat{B} - T$. Subtracting the exact normal equation gives $A(\widehat{B} - B^\star) = \mathcal{Q}$. Equation (52) implies $\|A^{-1}\|_2 \le 1/\lambda$, and therefore

$$\|\widehat{B} - B^\star\|_F = \|A^{-1}\mathcal{Q}\|_F \le \|A^{-1}\|_2 \|\mathcal{Q}\|_F \le \frac{\|\mathcal{Q}\|_F}{\lambda}, \tag{66}$$

$$J(\widehat{B}) - J(B^\star) = \tfrac{1}{2}\operatorname{tr}(\mathcal{Q}^\top A^{-1}\mathcal{Q}) \le \frac{\|\mathcal{Q}\|_F^2}{2\lambda}. \tag{67}$$

For a relative residual η, substitute $\|\mathcal{Q}\|_F \le \eta\|T\|_F$ to obtain explicit parameter and objective tolerances. A direct solve and converged CG therefore target the same global predictor; averaging local solutions instead introduces the distinct heterogeneity residual Ξ.

A.5 Score-coordinate drift

Fixed observations and changing score maps. Fix the N reference views and labels, the projection, and λ. Changing only the prompts changes v_j to v_j', the design vector ψ_j to ψ_j', and residual label $e_j = y_j - v_j$ to $e_j' = y_j - v_j'$. Set $\Delta S = S' - S$, $\Delta T = T' - T$, and $A' = S' + \lambda I$. Both the design matrix and response move because both depend on the prompt scores.

Exact change in the fitted head. The two normal equations are $AB^\star = T$ and $A'B' = T'$. Using $A' = A + \Delta S$,

$$A'(B' - B^\star) = T' - A'B^\star \tag{68}$$

$$= (T + \Delta T) - (AB^\star + \Delta S B^\star) = \Delta T - \Delta S B^\star. \tag{69}$$

Since S' remains a Gram matrix, $A' \succeq \lambda I$ and is invertible. Thus

$$B' - B^\star = (A')^{-1}(\Delta T - \Delta S B^\star). \tag{70}$$

The spectral bound $\|(A')^{-1}\|_2 \le 1/\lambda$, followed by the triangle inequality and $\|XY\|_F \le \|X\|_2\|Y\|_F$, gives

$$\|B' - B^\star\|_F \le \|(A')^{-1}\|_2\big(\|\Delta T\|_F + \|\Delta S B^\star\|_F\big) \tag{71}$$

$$\le \frac{\|\Delta T\|_F + \|\Delta S\|_2\|B^\star\|_F}{\lambda}. \tag{72}$$

From score displacement to feature and target displacement. Assume $\|v_j' - v_j\|_2 \le \epsilon$ for every reference view. Define $\delta\psi_j = \psi_j' - \psi_j = [R(v_j' - v_j); 0]$ and $\delta e_j = e_j' - e_j = -(v_j' - v_j)$. Projection contraction gives $\|\delta\psi_j\| \le \epsilon$, and $\|\delta e_j\| \le \epsilon$ follows directly. The bounds in Appendix A.3 give $\|\psi_j\| \le \sqrt{2}$. Since a one-hot label has norm one and $\|v_j\| \le 1$, the triangle inequality also gives $\|e_j\| \le 2$.

Change in second moments. Expanding each new outer product and canceling its old counterpart gives

$$\Delta S = \sum_{j=1}^{N} (\psi_j \delta\psi_j^\top + \delta\psi_j \psi_j^\top + \delta\psi_j \delta\psi_j^\top). \tag{73}$$

For vectors a, b, $\|ab^\top\|_2 = \|a\|\|b\|$, because this rank-one operator has its single nonzero singular value equal to that product. Applying this identity term by term yields

$$\|\Delta S\|_2 \le \sum_j (2\|\psi_j\|\|\delta\psi_j\| + \|\delta\psi_j\|^2) \tag{74}$$

$$\le N(2\sqrt{2}\,\epsilon + \epsilon^2). \tag{75}$$

Change in feature–target moments. The same expansion gives

$$\Delta T = \sum_{j=1}^{N} (\delta\psi_j e_j^\top + \psi_j \delta e_j^\top + \delta\psi_j \delta e_j^\top). \tag{76}$$

Here $\|ab^\top\|_F^2 = \sum_{k,\ell} a_k^2 b_\ell^2 = \|a\|^2\|b\|^2$. Taking square roots and using the triangle inequality gives

$$\|\Delta T\|_F \le \sum_j (\|\delta\psi_j\|\|e_j\| + \|\psi_j\|\|\delta e_j\| + \|\delta\psi_j\|\|\delta e_j\|) \tag{77}$$

$$\le N((2+\sqrt{2})\epsilon + \epsilon^2). \tag{78}$$

Substitution into Eq. (72) explicitly bounds the head displacement in terms of normalized-score displacement and the regularization floor.

Effect on decision scores. For a query with the same displacement bound, hold $\gamma \in [0, 1]$ fixed. Add and subtract $(B^\star)^\top \psi'$ to expand the difference between corrected predictions:

$$f'_\gamma - f_\gamma = (v' - v) + \gamma(B' - B^\star)^\top \psi' + \gamma(B^\star)^\top(\psi' - \psi).$$

Since $\|\psi'\| \le \sqrt{2}$, this gives

$$\|f'_\gamma - f_\gamma\|_2 \le \epsilon + \gamma\big(\sqrt{2}\|B' - B^\star\|_F + \epsilon\|B^\star\|_F\big). \tag{79}$$

Prompt changes affect both the identity path and the fitted correction. Freezing the reference sets this displacement to zero for fixed views. Fresh augmentation can still change sampled statistics; the bound isolates prompt-induced movement rather than augmentation variance.

A.6 Proof of Theorem 2

Candidate family and target risk. Let $h_{\ell,\gamma}$ be the classifier associated with checkpoint $\ell \in \{1, \dots, L\}$ and strength $\gamma \in \Gamma$. Condition on training data and all randomness constructing those checkpoints and training-fitted readouts. The theorem assumes that the resulting finite family is independent of the validation observations $Z_j = (X_j, Y_j)$. The observations are independent but may have different class-conditional distributions. Define

$$\mathcal{R}_{\ell,\gamma} = \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\mathbf{1}[h_{\ell,\gamma}(X_j) \ne Y_j]. \tag{80}$$

This is the average expected risk for the validation sampling design.

Paired gains and their bounded range. On the same observation, compare correction and reference through

$$D_{\ell,\gamma,j} = \mathbf{1}[h_{\ell,0}(X_j) \ne Y_j] - \mathbf{1}[h_{\ell,\gamma}(X_j) \ne Y_j]. \tag{81}$$

It equals 1 for an error corrected, -1 for a correct decision disrupted, and 0 otherwise. Thus $D_{\ell,\gamma,j} \in [-1, 1]$. Its empirical and expected averages are

$$\widehat{G}_{\ell,\gamma} = \frac{1}{n}\sum_j D_{\ell,\gamma,j}, \qquad G_{\ell,\gamma} = \mathbb{E}\widehat{G}_{\ell,\gamma} = \mathcal{R}_{\ell,0} - \mathcal{R}_{\ell,\gamma}. \tag{82}$$

At $\gamma = 0$, both gains are identically zero. Pairing allows the bound to target improvement over the same trained reference.

Uniform concentration over checkpoint–strength pairs. Hoeffding's inequality (Hoeffding, 1963) bounds the deviation of a mean of independent variables with ranges $b_j - a_j$ by

$2\exp[-2n^2\epsilon^2/\sum_j (b_j - a_j)^2]$. Here every range has width two, so a fixed nonzero-strength pair satisfies

$$\Pr\{|\widehat{G}_{\ell,\gamma} - G_{\ell,\gamma}| > \epsilon\} \le 2\exp(-n\epsilon^2/2). \tag{83}$$

There are $L(|\Gamma| - 1)$ nonzero-strength pairs. A union bound gives

$$\Pr\{\exists \ell, \gamma \neq 0 : |\widehat{G}_{\ell,\gamma} - G_{\ell,\gamma}| > \epsilon\} \le 2L(|\Gamma| - 1)\exp(-n\epsilon^2/2). \tag{84}$$

Setting the right-hand side to δ and solving for ϵ gives Eq. (14). With probability at least $1 - \delta$,

$$|\widehat{G}_{\ell,\gamma} - G_{\ell,\gamma}| \le \varepsilon_n \quad \text{simultaneously for all } \ell, \gamma. \tag{85}$$

The zero-strength pair satisfies the event exactly. Since it holds simultaneously, it also holds at a checkpoint selected with those same validation observations. This is the finite-family hold-out principle (Maillard et al., 2021) specialized to paired residual gains.

Accepted correction relative to the reference. On Eq. (85), a selected nonzero strength has empirical gain at least τ. Therefore

$$G_{\widehat{\ell},\widehat{\gamma}} \ge \widehat{G}_{\widehat{\ell},\widehat{\gamma}} - \varepsilon_n \ge \tau - \varepsilon_n, \tag{86}$$

$$\mathcal{R}_{\widehat{\ell},\widehat{\gamma}} - \mathcal{R}_{\widehat{\ell},0} = -G_{\widehat{\ell},\widehat{\gamma}} \le \varepsilon_n - \tau. \tag{87}$$

If selection returns zero, the two classifiers have identical decisions, so their risk difference is exactly zero. This proves Eq. (2). The fixed $\tau = 2/n$ rule requires two additional correct validation examples; it is not automatically larger than the concentration radius.

Approximate empirical maximization. For the chosen checkpoint, write $M = \max_{\gamma\in\Gamma} \widehat{G}_\gamma \ge 0$. If $M \ge \tau$, the rule selects an empirical maximizer and attains M. If $M < \tau$, it returns zero and attains $0 \ge M - \tau$. In both cases,

$$\widehat{G}_{\widehat{\gamma}} \ge \max_{\gamma\in\Gamma} \widehat{G}_\gamma - \tau. \tag{88}$$

Choosing the smallest tied strength does not change the attained empirical gain.

Comparison with the best strength at the selected checkpoint. Let γ° maximize expected gain for this checkpoint. Using the uniform event twice and Eq. (88) once,

$$G_{\widehat{\gamma}} \ge \widehat{G}_{\widehat{\gamma}} - \varepsilon_n \tag{89}$$

$$\ge \widehat{G}_{\gamma^\circ} - \tau - \varepsilon_n \tag{90}$$

$$\ge G_{\gamma^\circ} - \tau - 2\varepsilon_n. \tag{91}$$

Replace G_γ with $\mathcal{R}_0 - \mathcal{R}_\gamma$ and cancel the reference risk. This gives

$$\mathcal{R}_{\widehat{\ell},\widehat{\gamma}} \le \min_{\gamma\in\Gamma} \mathcal{R}_{\widehat{\ell},\gamma} + \tau + 2\varepsilon_n, \tag{92}$$

proving Eq. (15). The comparator varies strength at the selected checkpoint; it does not replace that checkpoint with the best member of the entire training trajectory.

Checkpoint reuse. The family can include all checkpoints from a predetermined training trajectory, with a training-only readout specified for each. The implementation fits only the selected checkpoint's readout; the other training-defined candidates provide the finite family for analysis. Validation may select among this family, but the assumption does not allow its observations to alter the family itself through unlimited method or hyperparameter redesign. The factor L accounts for checkpoint selection within that fixed construction.

Class-balanced validation and shifted test priors. Let $\mathcal{R}_k(h)$ be class-k conditional risk. Equal validation counts per class make the target $\mathcal{R}_{\mathrm{bal}}(h) = K^{-1}\sum_k \mathcal{R}_k(h)$. For test class prior π, holding class-conditionals fixed gives

$$\mathcal{R}_\pi(h) - \mathcal{R}_{\mathrm{bal}}(h) = \sum_k (\pi_k - 1/K)\mathcal{R}_k(h). \tag{93}$$

For a paired gain, write $d_k = \mathcal{R}_k(h_0) - \mathcal{R}_k(h_\gamma) \in [-1, 1]$. Then

$$|G_{\pi,\gamma} - G_{\mathrm{bal},\gamma}| = \left|\sum_k (\pi_k - 1/K)d_k\right| \tag{94}$$

$$\le \sum_k |\pi_k - 1/K|\,|d_k| \le \|\pi - \pi_{\mathrm{bal}}\|_1. \tag{95}$$

An accepted correction therefore has shifted-prior risk difference at most $\varepsilon_n - \tau + \|\pi - \pi_{\mathrm{bal}}\|_1$ under these assumptions. Zero strength remains the same classifier under either prior. Conditional-distribution changes require a further assumption; observed test gains are not a consequence of the fixed two-example threshold alone.

A.7 Readout search complexity

The directional-estimation calculation above applies to any vector of searched parameters, including an explicit readout (Spall, 1992; Nesterov & Spokoiny, 2017). With d coordinates and m directions, its affine-model mean-square error is $(d-1)\|g\|^2/m$. Adding the head increases the total dimension from 640 to $640 + (r+1)K$; in the joint-SPSA control, its coefficients join the common block. The corresponding block dimension and cross-coordinate error terms grow even though the returned scores already make the head objective explicitly solvable. Solving that objective removes head coordinates from stochastic search rather than spending additional VLM probes on their gradients. This is a dimension-dependent local estimate, not a convergence theorem for the nonlinear federated update with momentum (Ling et al., 2024).

A.8 The Interleaved adaptation variant

Interleaved tests coupling prompt adaptation to an explicitly fitted readout. At the start of round q, all clients form residual statistics using the current prompts θ^q. The server solves

$$B^q = \left(\sum_i S_i(\theta^q) + \lambda I \right)^{-1} \sum_i T_i(\theta^q), \qquad \lambda = 1. \tag{96}$$

Clients then hold B^q fixed during that round's prompt search and minimize

$$F_q(\theta; B^q) = \frac{\alpha_q}{2} \sum_i w_i \mathbb{E}_{(x,y)\sim\mathcal{D}_i} \left\| v(z_\theta(x)) + (B^q)^\top \psi(z_\theta(x)) - e_y \right\|_2^2, \tag{97}$$

where e_y is the one-hot label. If $s_j = \sqrt{\|u_j\|^2 + 10^{-12}}$ is the raw-score normalization factor for reference j, the scale is $\alpha_q = N^{-1} \sum_j s_j$. It remains fixed during that round's prompt updates. Prompt updates are aggregated, and the next round refits the head in the new score coordinates. There is no differentiation through the server solve during these local updates: only the prompts are searched while the broadcast head is fixed. Predictions use full strength, $\gamma = 1$. Unlike the final method, Interleaved does not preserve a separately selected prompt reference or choose a residual strength. Its mechanism tests in Table 8 modify the prompt objective, head tracking, aggregation, or solution within this organization.

A.9 Query and communication formulas

A query is one batched VLM forward invocation. For client i with n_i training examples and batch size 32, write $b_i = \max(1, \lfloor n_i/32 \rfloor)$ and $a_i = \lceil n_i/32 \rceil$. Each client has b_i optimization minibatches per epoch, with at least one for a small dataset. A prompt step costs $2(m_C + m_D) = 10$ calls. Multiplying epochs, rounds, and client batch counts gives

$$Q_{\mathrm{prompt}} = 24 \cdot 5 \cdot 10 \sum_i b_i, \qquad Q_{\mathrm{ref}} = 2 \sum_i a_i, \qquad Q_{\mathrm{train}} = Q_{\mathrm{prompt}} + Q_{\mathrm{ref}}. \tag{98}$$

Two reference views each require a_i calls, producing the reference term above. The complete predictor has $640 + (r+1)K$ learned coefficients; γ is a selected hyperparameter. Validation and test calls are separate.

For Interleaved, client i makes $5b_i$ prompt steps and pays $2a_i$ reference calls every round. Its first a_i prompt steps use four perturbation pairs (eight calls), and the remaining steps use five pairs (ten calls). Provided $a_i \le 5b_i$, the roundwise total is

$$2a_i + 8a_i + 10(5b_i - a_i) = 10 \cdot 5b_i.$$

Thus the alternative pays for repeated fits by reallocating its prompt probes; it has the same total training-call budget as prompt search.

For $d = r + 1$, S_i has d^2 entries and T_i has dK entries. The implemented full-matrix serialization also includes three diagnostic scalars, giving $8[d^2 + dK + 3]$ bytes in float64 per client. A returned head and scalar strength require $4[dK + 1]$ bytes in float32. Although S_i is symmetric, this count charges all its entries rather than assuming packed-triangle compression. The selected reference kernels require one additional $20 \cdot 640 \cdot 4$-byte broadcast. Under 24 rounds, 20 client broadcasts, and 18 uploads per round, the complete parameter/statistic payload in MiB is

$$\mathcal{C}_K = \frac{4 \cdot 24 \cdot (20 + 18) \cdot 640 + 20\big(4 \cdot 640 + 8[d^2 + dK + 3] + 4[dK + 1]\big)}{2^{20}}. \tag{99}$$

This payload convention excludes control traffic and backbone distribution.

B Extended experimental analysis

B.1 Dataset coverage

Table 3 summarizes the recognition tasks, split sizes, and class-name templates. The benchmarks span generic objects, fine-grained categories, scenes, textures, remote sensing, action frames, and synthetic counting. This diversity tests whether the same adaptation rule can exploit different kinds of class evidence. The main experiments use 16 training examples and four validation examples per class; the 4- and 8-shot subsets use fewer training examples. UCF101 is evaluated on image frames, and CLEVR is an eight-class counting task. Unless stated otherwise, results are means $\pm$ sample standard deviations over three seeds.

Table 3: **Statistics of datasets.** Pool sizes are the retained training/validation splits before few-shot sampling. Used columns report the 16-shot training and four-example-per-class validation sets. Test counts match the final-method score caches; UCF101 counts image frames, not videos. Manual prompts are the class templates used by the implementation.

Dataset	Classes	Train pool	Val. pool	Test	Train used	Val. used
Caltech101	100	4,128	1,649	2,465	1,600	400
Oxford Pets	37	2,944	736	3,669	592	148
CLEVR	8	55,999	14,001	15,000	128	32
Flowers102	102	4,093	1,633	2,463	1,632	408
Food101	101	50,500	20,200	30,299	1,616	404
FGVC Aircraft	100	3,334	3,333	3,333	1,600	400
SUN397	397	15,880	3,970	19,850	6,352	1,588
DTD	47	2,820	1,128	1,692	752	188
EuroSAT	10	13,500	5,400	8,100	160	40
Stanford Cars	196	6,509	1,635	8,041	3,136	784
RESISC45	45	15,750	6,300	9,450	720	180
UCF101	101	1,430,742	357,683	697,865	1,616	404
Mini-ImageNet	100	30,000	12,000	18,000	1,600	400

Dataset	Task type	Manual prompt
Caltech101	Generic objects	“a photo of a [CLASS].”
Oxford Pets	Fine-grained	“a photo of a [CLASS].”
CLEVR	Object counting	“This is a photo of [CLASS] objects”
Flowers102	Fine-grained	“a photo of a [CLASS].”
Food101	Fine-grained	“a photo of [CLASS].”
FGVC Aircraft	Fine-grained	“a photo of a [CLASS].”
SUN397	Scenes	“a photo of a [CLASS].”
DTD	Textures	“[CLASS] texture.”
EuroSAT	Remote sensing	“a centered satellite photo of [CLASS].”
Stanford Cars	Fine-grained	“a photo of a [CLASS].”
RESISC45	Remote sensing	“This is a photo of a [CLASS]”
UCF101	Action frames	“a photo of a person doing [CLASS].”
Mini-ImageNet	Generic objects	“a photo of a [CLASS].”

B.2 Robustness to client heterogeneity and label scarcity

Changing client heterogeneity. Table 4 evaluates the full pipeline under IID and two Dirichlet partitions, using the main 16-shot pathological partition as the common comparison. Mean accuracy reaches 70.70%, 70.51%, and 70.87% for IID, $\alpha = 0.1$, and $\alpha = 0.5$, respectively: differences of $+0.56$, $+0.37$, and $+0.72$ pp from the main setting. The modest average variation indicates that the method accommodates different distributions of client class evidence. DTD improves under all three partitions, while EuroSAT falls by 1.37 pp at $\alpha = 0.1$, showing greater sensitivity to strongly skewed local supervision. Each partition also changes the prompt-search trajectory, so the results reflect adaptation of both the learned score map and the pooled residual predictor.

Table 4: **Full-method accuracy under different client partitions.** All settings use 16 shots. Accuracy is in %; Δ_{16} is the accuracy difference (pp) from the full method under the main pathological partition in Table 5 (16-shot column). Entries are mean $\pm$ sample SD over three seeds; differences are computed within each seed.

Dataset	IID	Δ_{16}	Dir. $\alpha = .1$	Δ_{16}	Dir. $\alpha = .5$	Δ_{16}
Caltech	93.09 ± 1.22	$+0.42 \pm 1.43$	93.50 ± 0.80	$+0.82 \pm 2.05$	93.40 ± 0.83	$+0.73 \pm 1.32$
Pets	92.72 ± 0.15	$+0.15 \pm 1.40$	92.77 ± 0.91	$+0.19 \pm 0.43$	93.00 ± 0.24	$+0.42 \pm 1.41$
CLEVR	28.23 ± 3.60	$+2.20 \pm 8.34$	28.46 ± 6.20	$+2.44 \pm 7.92$	29.55 ± 2.65	$+3.52 \pm 3.95$
Flowers	85.41 ± 0.89	-0.03 ± 0.55	85.61 ± 0.87	$+0.18 \pm 1.77$	85.97 ± 0.20	$+0.53 \pm 1.20$
Food	86.72 ± 0.20	$+0.31 \pm 0.47$	86.45 ± 0.51	$+0.04 \pm 0.66$	86.62 ± 0.20	$+0.21 \pm 0.31$
Aircraft	32.01 ± 0.21	$+0.77 \pm 0.72$	32.26 ± 0.30	$+1.02 \pm 0.89$	32.24 ± 0.69	$+1.00 \pm 0.22$
SUN397	68.96 ± 0.23	$+0.00 \pm 0.44$	68.65 ± 0.18	-0.31 ± 0.38	69.12 ± 0.24	$+0.16 \pm 0.45$
DTD	59.77 ± 0.69	$+1.28 \pm 1.06$	59.65 ± 1.09	$+1.16 \pm 0.22$	60.50 ± 0.36	$+2.01 \pm 0.89$
EuroSAT	69.47 ± 2.75	$+1.09 \pm 2.57$	67.00 ± 2.56	-1.37 ± 3.33	67.73 ± 1.45	-0.64 ± 2.05
Cars	69.51 ± 0.83	$+0.61 \pm 0.73$	69.34 ± 0.41	$+0.45 \pm 0.90$	69.69 ± 0.40	$+0.80 \pm 0.79$
RESISC	75.52 ± 1.28	$+0.40 \pm 0.62$	75.79 ± 1.14	$+0.67 \pm 0.91$	75.80 ± 2.29	$+0.68 \pm 1.63$
UCF	68.90 ± 0.38	$+0.02 \pm 0.47$	68.69 ± 1.61	-0.19 ± 1.77	68.49 ± 0.16	-0.39 ± 0.15
Mini-IN	88.84 ± 0.79	$+0.01 \pm 0.68$	88.49 ± 0.80	-0.33 ± 0.66	89.17 ± 0.78	$+0.35 \pm 0.65$

Reducing labeled supervision. Table 5 measures the effect of reducing training examples while retaining the main client-partition rule. The full method reaches 67.53% at four shots and 68.84% at eight shots, compared with 70.15% at 16 shots: reductions of 2.62 and 1.31 pp. The losses concentrate on Flowers, DTD, EuroSAT, and RESISC45, whereas Caltech101, Pets, and Food101 remain within 0.6 pp of their 16-shot accuracies even at four shots. Tasks with stronger pretrained class discrimination therefore require less additional supervision; tasks that depend more heavily on adaptation benefit from the extra examples available to both prompt learning and residual fitting. Eight shots recover about half of the four-shot accuracy deficit.

Table 5: **Full-method accuracy with fewer training examples.** All settings use the main pathological partition. Accuracy is in %; Δ_{16} is the accuracy difference (pp) from the 16-shot full method. Entries are mean $\pm$ sample SD over three seeds; differences are computed within each seed.

Dataset	16-shot	4-shot	Δ_{16}	8-shot	Δ_{16}
Caltech	92.67 ± 1.26	92.12 ± 0.88	-0.55 ± 0.90	92.72 ± 1.31	$+0.05 \pm 2.55$
Pets	92.58 ± 1.24	92.48 ± 0.58	-0.10 ± 1.44	92.77 ± 0.27	$+0.19 \pm 1.46$
CLEVR	26.03 ± 4.74	25.68 ± 5.05	-0.35 ± 9.63	25.40 ± 1.56	-0.63 ± 6.28
Flowers	85.44 ± 1.08	79.12 ± 1.26	-6.32 ± 1.47	82.15 ± 0.92	-3.29 ± 1.60
Food	86.41 ± 0.40	86.09 ± 0.35	-0.32 ± 0.59	86.19 ± 0.30	-0.22 ± 0.58
Aircraft	31.24 ± 0.60	28.60 ± 0.93	-2.64 ± 0.38	30.06 ± 0.22	-1.18 ± 0.51
SUN397	68.96 ± 0.21	67.57 ± 0.38	-1.38 ± 0.39	68.16 ± 0.41	-0.80 ± 0.62
DTD	58.49 ± 1.09	51.95 ± 0.82	-6.54 ± 1.40	56.05 ± 1.89	-2.44 ± 1.01
EuroSAT	68.37 ± 0.77	62.14 ± 3.86	-6.23 ± 4.63	63.33 ± 4.61	-5.05 ± 4.99
Cars	68.89 ± 0.50	67.13 ± 0.43	-1.76 ± 0.87	68.20 ± 0.82	-0.69 ± 1.27
RESISC	75.11 ± 1.07	70.85 ± 1.50	-4.26 ± 0.71	74.12 ± 0.94	-0.99 ± 0.97
UCF	68.88 ± 0.31	67.30 ± 0.34	-1.58 ± 0.61	67.60 ± 0.38	-1.28 ± 0.46
Mini-IN	88.82 ± 0.14	86.84 ± 0.11	-1.99 ± 0.17	88.14 ± 1.07	-0.68 ± 0.96

B.3 Global refinement across prompt predictors

FedMUSE refinement on FedBPT scores. To isolate the contribution of our global residual refinement, we freeze each accepted FedBPT checkpoint and apply FedMUSE's score normalization, client statistic pooling, ridge solve, and validation strength rule. FedBPT's prompt search remains unchanged. The intervention therefore tests whether our globally fitted decision stage can use class evidence that another prompt predictor has already exposed, without putting its readout coefficients into further black-box search.

Across 13 tasks, adding our refinement raises FedBPT from 62.56% to 68.28%, a paired gain of 5.72 pp. The task-level results in Table 6 show large gains on DTD, RESISC45, and Flowers; on CLEVR, validation instead preserves FedBPT's original decisions. This transfer across prompt predictors supports the global residual stage as a substantive contribution in its own right. Full FedMUSE reaches 70.15%, another 1.86 pp above refined FedBPT on the dataset average. The gap shows that the prompt-search stage also matters: its learned shared score map determines the evidence available to the same global solver. The zero-shot counterpart appears as "w/o prompt search" in Table 2.

Table 6: **Transferring FedMUSE global residual refinement to FedBPT.** Test accuracy (%) before and after our pooled residual readout and validation strength selection. FedBPT prompts remain fixed. The FedBPT column repeats the main-table mean; uncertainty on the newly refined accuracy and paired changes uses three rerun checkpoint seeds. Gain is refined FedBPT minus its original predictor; Gap is full FedMUSE minus refined FedBPT. Differences use unrounded paired accuracies; macro averages weight datasets equally.

Dataset	FedBPT	+ global residual	Gain (pp)	Gap to FedMUSE (pp)
Caltech	90.99	91.52 ± 0.76	$+0.53 \pm 0.91$	$+1.15 \pm 0.52$
Pets	87.37	90.03 ± 1.01	$+2.66 \pm 2.31$	$+2.54 \pm 0.41$
CLEVR	25.52	25.52 ± 2.20	$+0.00 \pm 0.00$	$+0.51 \pm 2.87$
Flowers	70.19	84.76 ± 1.18	$+14.58 \pm 1.59$	$+0.68 \pm 0.19$
Food	85.48	85.60 ± 0.41	$+0.12 \pm 0.21$	$+0.81 \pm 0.59$
Aircraft	23.45	28.51 ± 0.81	$+5.07 \pm 1.25$	$+2.73 \pm 1.38$
SUN397	64.70	66.87 ± 1.21	$+2.17 \pm 0.66$	$+2.08 \pm 1.20$
DTD	39.20	57.78 ± 0.63	$+18.58 \pm 1.16$	$+0.71 \pm 0.61$
EuroSAT	53.77	59.22 ± 3.95	$+5.44 \pm 4.76$	$+9.15 \pm 3.52$
Cars	63.80	67.78 ± 0.78	$+3.98 \pm 0.67$	$+1.11 \pm 0.39$
RESISC	58.59	75.11 ± 1.83	$+16.52 \pm 1.18$	$+0.01 \pm 1.87$
UCF	64.59	67.93 ± 0.13	$+3.34 \pm 0.47$	$+0.95 \pm 0.35$
Mini-IN	85.67	87.04 ± 0.71	$+1.36 \pm 1.18$	$+1.79 \pm 0.63$
Macro avg.	62.56	68.28 ± 0.31	$+5.72 \pm 0.02$	$+1.86 \pm 0.57$

B.4 Prompt coordinates and finite-budget estimation

A matched coordinate–estimator comparison. The mean–contrast map is invertible within the chosen latent prompt family, so its effect must come from search behavior under finite queries. Table 7 crosses mean–contrast versus independent text/visual coordinates with joint versus block-wise SPSA on CLEVR, DTD, EuroSAT, and UCF101. All four settings start from identical physical text and image prompts, use the same cross-entropy objective and 10 batched calls per local step, and receive the same residual fitting procedure. Coordinate-specific perturbation and update scales match the physical prompt displacement induced by the two parameterizations.

After refinement, mean–contrast with block estimation averages 55.44% across the four tasks, compared with 54.51% for independent coordinates with block estimation. Under joint estimation, the ordering reverses (54.38% versus 54.70%). Prompt-only mean–contrast/block search also trails independent/block search (48.97% versus 49.78%). The mixed effects identify a conditional benefit of the selected coordinate–estimator pairing; they do not imply that mean–contrast coordinates or block estimation uniformly dominate their alternatives.

Table 7: **Prompt coordinates and zeroth-order estimators.** Test accuracy (%) with prompt search alone (top) and after the same global residual refinement (bottom). MC denotes mean–contrast coordinates; Ind. denotes independent text/visual coordinates. Joint perturbs both parameter blocks together; block estimates them separately. Entries are mean ± sample SD over three seeds. The four settings share the physical prompt initialization, search loss, and 10 batched calls per local step.

Dataset	MC / block	MC / joint	Ind. / block	Ind. / joint
Prompt search only				
CLEVR	26.03 ± 4.74	25.03 ± 5.03	28.60 ± 0.82	25.00 ± 1.14
DTD	44.46 ± 2.34	43.24 ± 0.22	42.89 ± 1.09	44.19 ± 0.60
EuroSAT	58.42 ± 5.55	60.51 ± 2.17	61.98 ± 2.99	57.57 ± 2.62
UCF	66.96 ± 0.34	66.08 ± 0.94	65.65 ± 0.75	66.15 ± 0.81
Macro avg.	48.97 ± 2.79	48.72 ± 0.95	49.78 ± 0.52	48.23 ± 0.35
With identical global residual refinement				
CLEVR	26.03 ± 4.74	25.03 ± 5.03	29.48 ± 1.60	25.94 ± 2.47
DTD	58.49 ± 1.09	58.69 ± 0.82	55.58 ± 2.15	58.27 ± 1.62
EuroSAT	68.37 ± 0.77	65.52 ± 10.14	65.27 ± 2.35	65.87 ± 7.33
UCF	68.88 ± 0.31	68.28 ± 0.15	67.73 ± 0.53	68.72 ± 0.01
Macro avg.	55.44 ± 1.46	54.38 ± 2.18	54.51 ± 0.28	54.70 ± 2.82

B.5 Mechanism analysis of adaptation organization

Mechanism tests of interleaved fitting. The INTERLEAVED strategy introduced in Section 3 alternates prompt updates with pooled head fitting. Table 8 groups its mechanism tests by the question they answer. The first two rows compare separated and interleaved fitting; the others test specific training mechanisms.

Contribution of prompt learning. Freezing initial prompts and fitting a head tests how much can be gained without supervised prompt search. Searching prompts with square loss but no head tests whether the changed objective alone explains INTERLEAVED's behavior. CE search under the INTERLEAVED query schedule retains cross-entropy while matching the reference-query allocation. On DTD, the frozen-initial-prompt head reaches 52.62%, but the final method reaches 58.49%; useful residual fitting does not remove the need to learn the score map. Square-loss prompt search alone reaches only 39.66%, suggesting that exchanging the prompt objective is insufficient.

Tracking the score map. Keeping the first-round head tests whether a head fitted early remains suitable as prompts evolve. Its DTD accuracy falls from INTERLEAVED's 54.67% to 47.28%, and EuroSAT falls from 63.27% to 56.52%. These results are consistent with the design matrix and residual targets changing during prompt training. A frozen reference removes prompt-induced coordinate drift.

Solving versus searching the head. Averaging local heads tests the aggregation mechanism; replacing the pooled solver with CG tests numerical solution; searching prompts and head together tests putting the explicit readout into zeroth-order optimization. Local averaging substantially weakens DTD and UCF101, while CG stays close to the pooled-solve result. Joint SPSA reaches only 34.83% on DTD and 55.29% on UCF101, compared with 54.67% and 64.07% for INTERLEAVED. The head increases the searched dimension to 2,191 and 3,973, respectively, while also introducing local parameter aggregation. These controls favor solving the accessible head objective explicitly.

Task-dependent behavior of interleaved fitting. The final method improves over INTERLEAVED on DTD, EuroSAT, and UCF101, but INTERLEAVED is stronger on CLEVR (31.98% versus 26.03%). Interleaved updates can reshape the score map specifically for the fitted head, which may help a task with a weak prompt-only predictor. The final method instead preserves a separately learned reference and allows conservative strength selection. This accounts for a useful trade-off: stronger task-specific coupling can help individual tasks, whereas separating the two stages improves the 13-dataset average from 66.88% to 70.15%.

Table 8: **Mechanism tests of adaptation organization.** The first two rows compare the final method with INTERLEAVED. Other rows test prompt learning, score-map tracking, or head aggregation and optimization, as explained in the text. CE is cross-entropy; CG is conjugate gradients. Accuracy is mean ± sample SD over three seeds.

Predictor / mechanism test	CLEVR	DTD	EuroSAT	UCF
FedMUSE	26.03 ± 4.74	58.49 ± 1.09	68.37 ± 0.77	68.88 ± 0.31
INTERLEAVED	31.98 ± 1.32	54.67 ± 2.17	63.27 ± 2.08	64.07 ± 0.70
Contribution of prompt learning and its objective				
Freeze initial prompts; fit head	26.58 ± 0.90	52.62 ± 1.11	60.94 ± 2.21	63.25 ± 0.66
Square-loss search; no head	29.99 ± 1.98	39.66 ± 1.66	58.72 ± 4.68	63.07 ± 1.60
CE search; INTERLEAVED query schedule	27.90 ± 2.21	44.09 ± 2.46	60.85 ± 6.88	65.94 ± 0.02
Tracking the changing prompt-score map				
Keep first-round head	30.35 ± 3.02	47.28 ± 1.66	56.52 ± 1.36	62.97 ± 1.12
Global head aggregation and explicit solution				
Average local heads	30.06 ± 4.54	40.74 ± 0.76	55.87 ± 1.91	58.47 ± 1.12
Solve pooled head with CG	30.26 ± 2.05	54.65 ± 2.14	63.60 ± 1.64	63.80 ± 1.17
SPSA search of prompts + head	29.77 ± 1.88	34.83 ± 1.93	63.16 ± 3.55	55.29 ± 0.93

B.6 REFERENCE PRESERVATION UNDER HEAD INTERVENTIONS

Separating reference preservation from head removal. Table 9 compares the independent prompt-search reference with three predictions using the same trained INTERLEAVED prompts. INTERLEAVED*:* $\gamma = 0$ removes the head from those prompts; *saved head* uses their learned head; *refitted head* recomputes a pooled head on training references. Thus the zero-strength column belongs to INTERLEAVED, whose prompt checkpoint differs from $\theta_\star$ in the final method.

Removing INTERLEAVED's head reduces DTD from 54.67% to 22.71% and UCF101 from 64.07% to 41.16%. Their independent prompt-search references reach 44.46% and 66.96%. The head and prompts have adapted together, so removing the former does not undo changes to the latter. A freshly fitted head restores much of the accuracy but does not uniformly exceed the saved head. This dependence motivates the final method's fixed reference: its $\gamma = 0$ path preserves the exact prompt predictor selected before residual learning.

Table 9: **Head dependence of jointly trained prompts.** The prompt-search reference is trained separately. All three INTERLEAVED columns retain the same jointly trained prompts: $\gamma = 0$ removes their head; saved uses their checkpoint head; refitted solves a new head on training references. The reference checkpoint $\theta_\star$ is not used in the INTERLEAVED columns. Three-seed accuracy is mean ± sample SD.

Dataset	Prompt-search reference	INTERLEAVED $\gamma = 0$	INTERLEAVED saved head	INTERLEAVED refitted head
CLEVR	26.03 ± 4.74	17.49 ± 4.17	31.98 ± 1.32	28.21 ± 1.59
DTD	44.46 ± 2.34	22.71 ± 5.19	54.67 ± 2.17	54.37 ± 2.07
EuroSAT	58.42 ± 5.55	31.26 ± 19.12	63.27 ± 2.08	62.29 ± 1.37
UCF	66.96 ± 0.34	41.16 ± 13.00	64.07 ± 0.70	63.06 ± 1.00
RESISC	61.22 ± 0.75	49.54 ± 5.57	75.65 ± 1.40	74.53 ± 2.17
Cars	64.85 ± 0.59	54.32 ± 2.53	58.06 ± 1.51	58.85 ± 0.77

B.7 Global pooling and local-head aggregation

Optimization mismatch and its classification consequence.

Independent local solves need not remain optimal after parameter averaging. To quantify this mismatch, Figure 4 relates their excess residual objective to their test-accuracy deficit. The horizontal axis is

$$100\,\frac{J(\bar{B}) - J(B^\star)}{J(0) - J(B^\star)},$$

where $\bar{B}$ is the sample-weighted local-head average. It measures the fraction of the pooled solution's objective improvement lost by averaging local solutions. The vertical axis is the test-accuracy advantage of pooled fitting, with each readout using validation-selected strength. Each point represents a dataset mean.

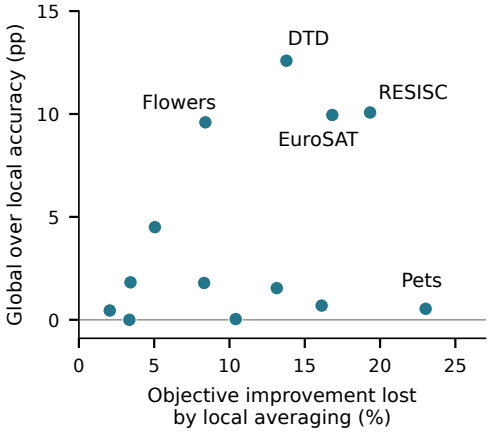


Figure 4: **Pooled fitting versus local-head averaging.** Fraction of the pooled objective improvement lost by averaging local heads versus the resulting test-accuracy deficit. Each point is a dataset mean; the horizontal zero line marks equal accuracy.

The local-head objective gap is positive on all tasks, while its classification consequence is especially large on DTD, Flowers, EuroSAT, and RESISC45. DTD loses 12.59 pp and EuroSAT loses 9.95 pp relative to pooled fitting. These results connect the heterogeneity gap in Theorem 1 to the benefit of exchanging sufficient statistics. Food101 has a positive objective gap but little accuracy change: improving a quadratic fitting objective need not move an already correct classification boundary. The largest classification deficits occur where the reference leaves substantial correction opportunities.

B.8 Component effects across datasets

Dataset-level component effects.

Figure 5 expands the main component ablation into paired accuracy changes for each task. Columns remove the residual, replace global pooling with local-head averaging, fix γ to one or one-half, remove the two-example selection threshold, or replace CG with a direct solve. Each value is the intervention's accuracy minus the final method's accuracy.

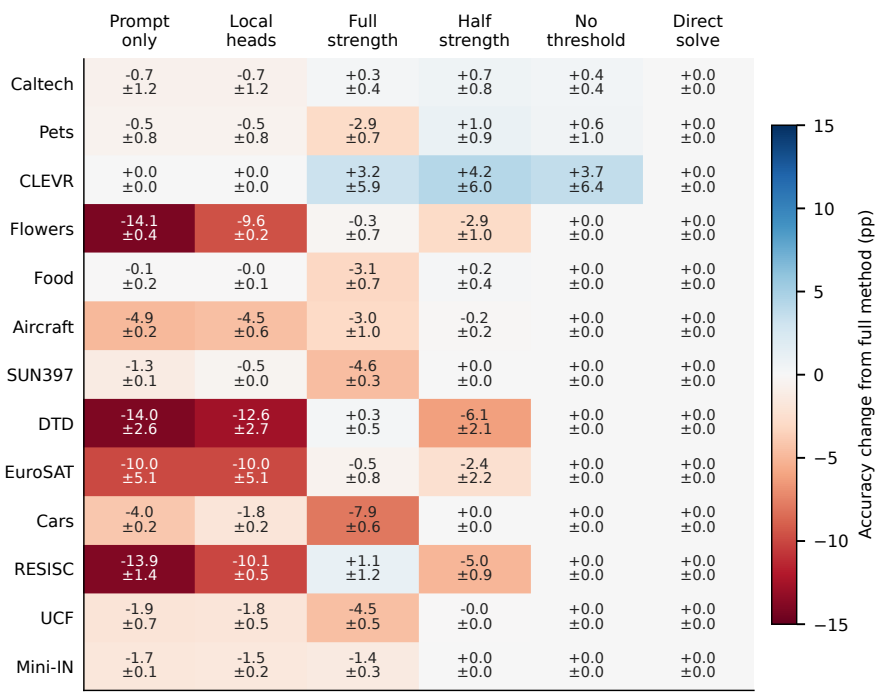


Figure 5: **Task-level effects of refinement choices.** Rows are datasets; columns are the interventions named above. Cells show accuracy change from the final method (pp), with sample SD beneath; negative values favor the final method. "No gain threshold" retains validation strength selection but removes its two-example requirement.

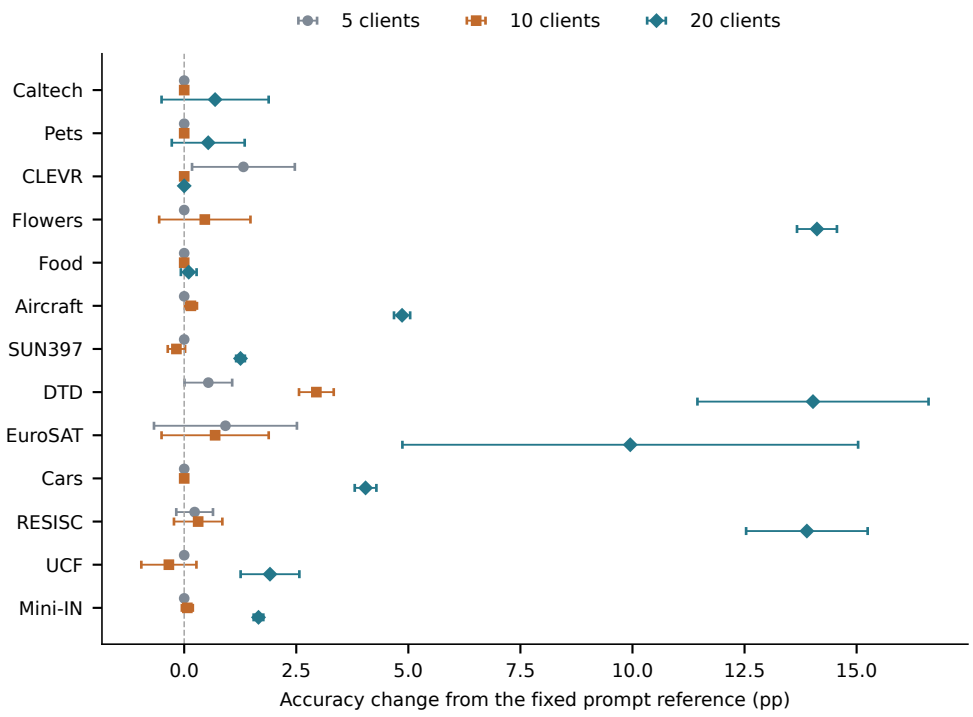
Removing residual refinement has the largest effects on Flowers, DTD, and RESISC45. Local averaging recovers only part of these gains, reinforcing the need to combine heterogeneous evidence before solving. Fixed $\gamma = 1$ particularly harms strong-reference tasks, while $\gamma = 0.5$ is competitive on some tasks but cannot adapt to their different correction needs. Removing the two-example threshold raises the overall mean by 0.36 pp: the threshold can reject useful small validation gains. Its role is a more conservative adoption rule, whereas selecting the strength accounts for the larger improvement over full-strength deployment. Replacing CG with a direct solve leaves accuracy unchanged on every task, confirming that the numerical solver does not account for these component effects.

B.9 Client coverage in residual refinement

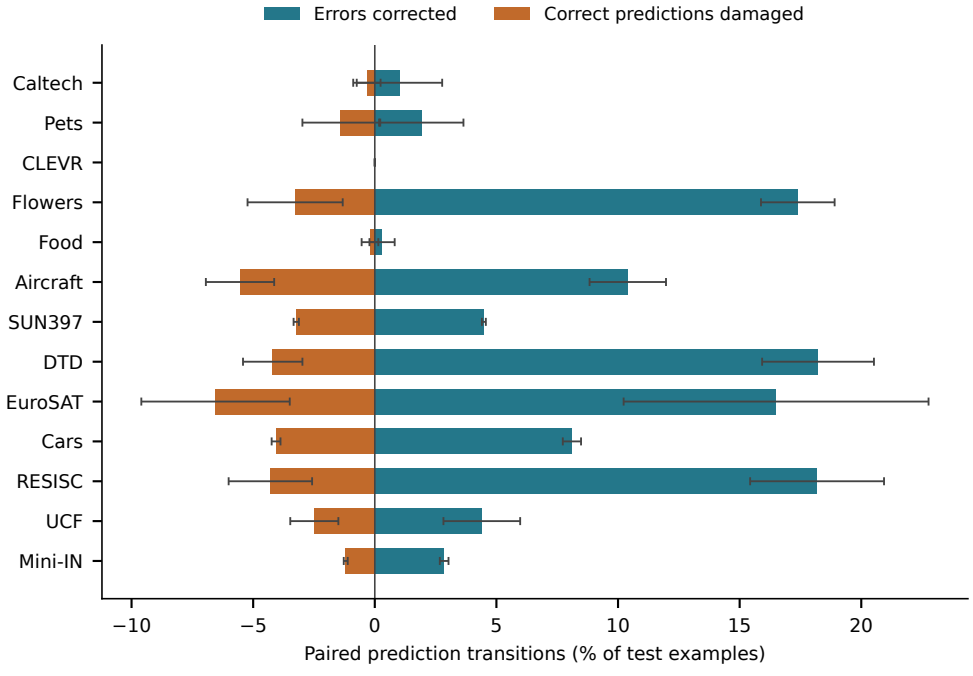
Coverage of the global evidence pool. The coverage experiment holds the learned prompt predictor fixed and changes only which clients contribute residual statistics. Figure 6a reports every task rather than only a macro-average. Five and ten contributing clients improve on the reference by just 0.23 and 0.32 pp on average, compared with 5.16 pp when all 20 clients contribute. Flowers, DTD, EuroSAT, and RESISC45 account for much of this gap: under label skew, small client subsets omit class evidence that the global solve needs. The result supports broad statistic collection in the second stage; it does not represent retraining the prompt predictor with fewer clients.

B.10 Prediction corrections and disruptions

Correction and disruption across tasks. Figure 6b decomposes each task's accuracy change into wrong-to-correct and correct-to-wrong transitions. The right bar counts corrected errors; the left bar counts newly introduced errors, both relative to all test examples. The correction surplus determines the net gain.



(a) Gain over the fixed prompt reference when 5, 10, or 20 clients supply residual statistics.



(b) Wrong-to-correct transitions (right) and correct-to-wrong transitions (left), as percentages of all test examples.

Figure 6: **Coverage and prediction changes across tasks.** Points and bars are three-seed means with sample SD; panel (a) also averages three client subsets within each seed.

Flowers, DTD, and RESISC45 show large correction surpluses, explaining their substantial average improvements. Strong-reference tasks such as Food101 have little remaining headroom. CLEVR selects the reference-only path in the main setting, so both transition rates are zero. This task dependence supports retaining the reference as a deployment option rather than applying the same correction to every task. Flowers also improves macro-F1 by 17.26 pp, indicating benefits beyond the aggregate correct count. SUN397's lower-quartile class recall instead falls by 0.30 pp: selecting overall validation accuracy can overlook a small group of difficult classes. Class-aware strength selection is a possible remedy for this remaining imbalance.

B.11 Parameter and query costs

Search dimension versus complete adaptation size. Table 10 separates prompt-training calls from reference calls and reports the full learned parameter count. "Extra (%)" is $100Q_{\mathrm{ref}}/Q_{\mathrm{prompt}}$; "Upload (KiB/client)" is one client's additional statistics payload. The readout adds $(r+1)K$ coefficients to the complete predictor while keeping search dimension fixed at 640.

Table 10: Learned parameter counts and end-to-end training-call totals. Extra is the reference-stage overhead (%); upload is the additional per-client statistics payload in KiB.

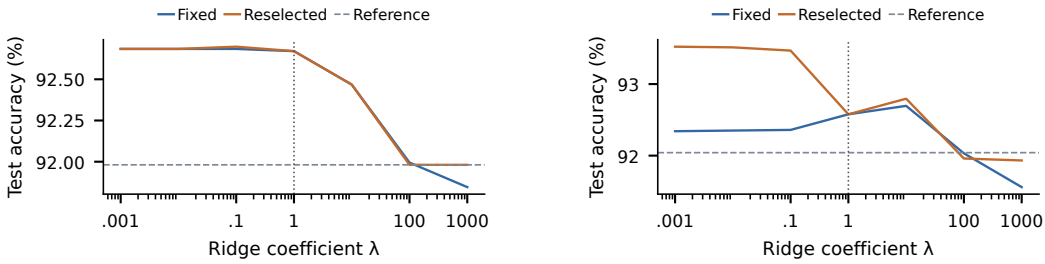
Dataset	Parameters	Prompt Q	Ref. Q	Total Q	Extra (%)	Upload (KiB/client)
Caltech	3940	48,000	120	48,120	0.250	34.31
Pets	1861	24,000	40	24,040	0.167	18.07
CLEVR	712	24,000	40	24,040	0.167	1.22
Flowers	4006	48,000	120	48,120	0.250	34.83
Food	3973	48,000	120	48,120	0.250	34.57
Aircraft	3940	48,000	120	48,120	0.250	34.31
SUN397	13741	216,000	400	216,400	0.185	110.88
DTD	2191	24,000	80	24,080	0.333	20.65
EuroSAT	750	24,000	40	24,040	0.167	1.83
Cars	7108	96,000	200	96,200	0.208	59.06
RESISC	2125	24,000	80	24,080	0.333	20.13
UCF	3973	48,000	120	48,120	0.250	34.57
Mini-IN	3940	48,000	120	48,120	0.250	34.31

Refinement uses only 40–400 extra batched calls, adding 0.17–0.33% to prompt-training calls. For DTD, 80 reference calls yield a 14.03 pp gain over its prompt reference; SUN397 has a larger head because it has more classes, yet its query overhead is only 0.19%. This separation lets the method combine distributed supervision without paying zeroth-order search costs for every head coefficient. Appendix A.9 derives the complete communication payload.

B.12 Regularization sensitivity across tasks

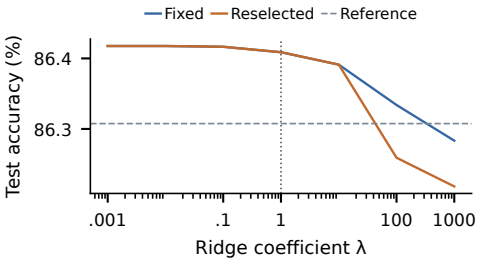
Regularization and task-dependent correction. Figures 7 and 8 show the seven-coefficient sweep for all 13 tasks. Blue curves keep the strength originally selected at $\lambda = 1$; orange curves reselect strength on validation. Horizontal dashed lines are the prompt-reference accuracies, and vertical lines mark the common default. Each task has its own accuracy scale.

Across small coefficients, the macro-average varies modestly: validation-reselected accuracy ranges from 70.15% to 70.64% for $\lambda \in [0.001, 1]$. Larger penalties shrink the residual enough to remove useful class corrections, approaching the reference-only predictor. The unit default supplies a common positive diagonal floor in the normalized-score system while keeping most of the refinement gain. Using $\lambda = 0.1$ improves the mean by 0.49 pp over the unit default, indicating that a weaker penalty can retain additional useful corrections. Fixed-strength and reselected curves further reveal the interaction of head magnitude with deployment strength.

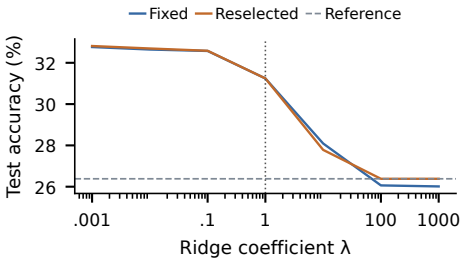


(a) Caltech101.

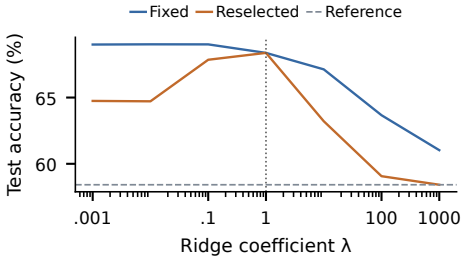
(b) Oxford Pets.



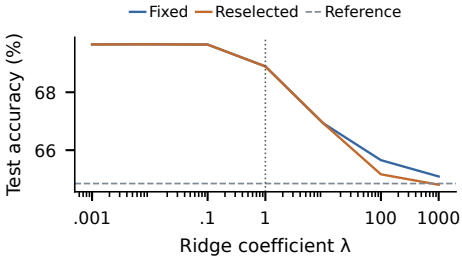
(c) Food101.



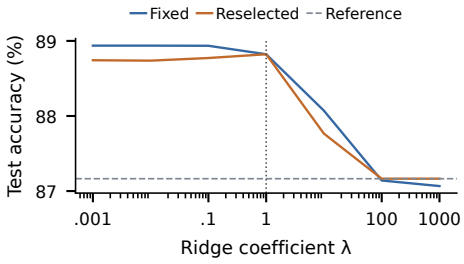
(d) FGVC Aircraft.



(e) EuroSAT.

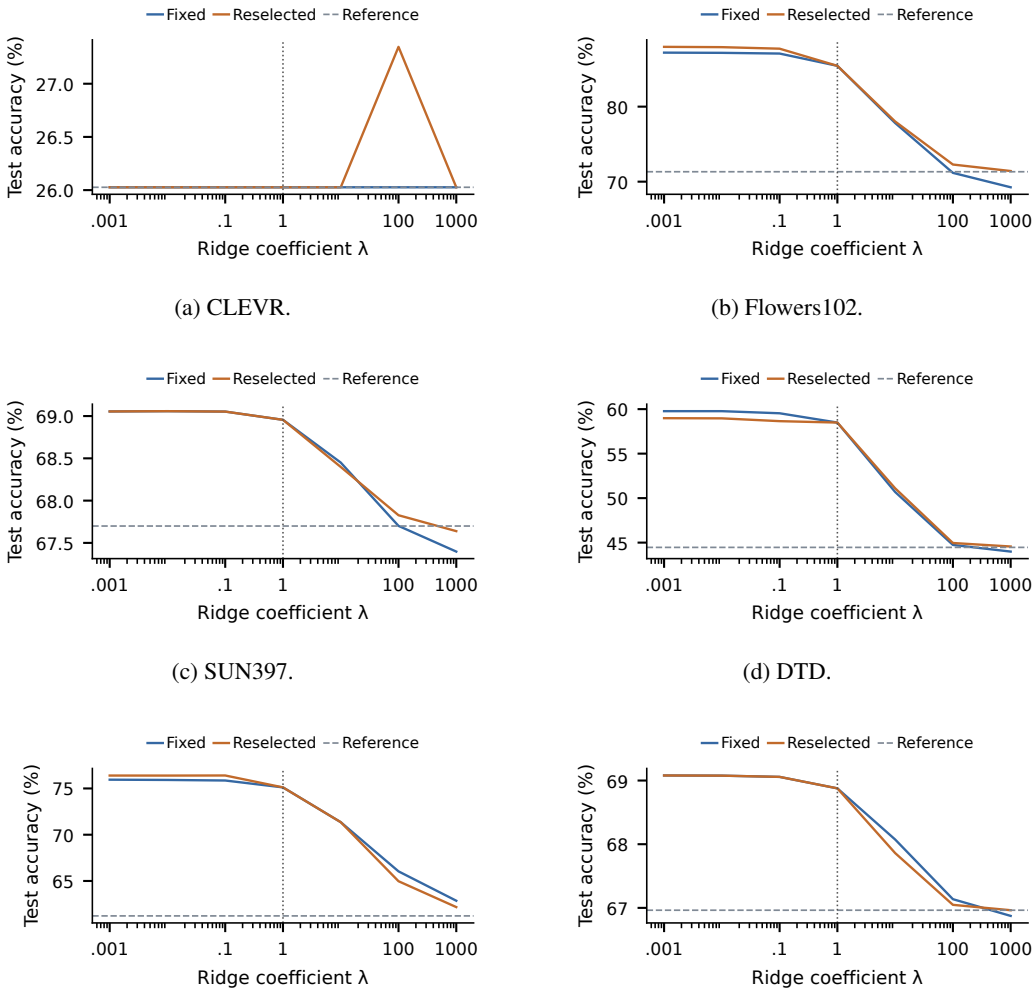


(f) Stanford Cars.



(g) Mini-ImageNet.

Figure 7: **Per-dataset ridge sensitivity, first task group.** Each panel shows three-seed mean test accuracy versus λ on a log scale. Blue fixes strength to its $\lambda = 1$ selection; orange reselects it on validation. Horizontal dashed lines are prompt-reference accuracies; vertical lines mark $\lambda = 1$.



(a) CLEVR.

(b) Flowers102.

(c) SUN397.

(d) DTD.

(e) RESISC45.

(f) UCF101.

Figure 8: **Per-dataset ridge sensitivity, second task group.** Colors, axes, and strength rules match Figure 7. Each task uses its own vertical scale; panels separate improvements over the reference from suppression under strong regularization.

B.13 SEED VARIABILITY IN THE MAIN COMPARISON

Table 11 expands Table 1 with per-dataset standard deviations. Rows identify datasets, and columns identify the evaluated predictors; each entry reports mean test accuracy and its variation across three seeds. The largest adaptation gaps occur on Flowers, DTD, EuroSAT, and RESISC45. Relative to FedBPT, the full method improves their means by 15.25, 19.29, 14.60, and 16.52 pp, respectively, exceeding the seed variability of either predictor on these tasks. CLEVR is less stable: the full method's 26.03% accuracy has a 4.74 pp standard deviation, and its 0.51 pp mean advantage over FedBPT is small in comparison. The evidence is therefore strongest on tasks where pooling heterogeneous class evidence yields substantial improvements, while the ranking on synthetic counting remains sensitive to the training seed.

Table 11: **Main comparison with seed variability.** Test accuracy (%) as mean ± standard deviation across three seeds. Means match Table 1. The two blocks cover all ten predictors; Ours denotes FedMUSE. CLIP zero-shot has zero seed variation.

Dataset	CLIP (ZS)	PromptFL (WB)	Fed BlackVIP	FedEffi BlackVIP	FedBBT
Caltech	92.37 ± 0.00	94.32 ± 0.67	92.24 ± 0.06	92.32 ± 0.12	88.72 ± 1.15
Pets	88.17 ± 0.00	92.52 ± 0.45	88.85 ± 0.21	87.14 ± 0.85	87.31 ± 2.06
CLEVR	14.69 ± 0.00	25.01 ± 3.33	17.60 ± 0.55	15.48 ± 0.55	13.87 ± 0.82
Flowers	67.03 ± 0.00	87.95 ± 1.60	67.33 ± 0.08	66.61 ± 0.58	70.02 ± 0.65
Food	83.76 ± 0.00	86.58 ± 0.26	83.95 ± 0.16	83.34 ± 0.27	85.03 ± 1.14
Aircraft	23.91 ± 0.00	29.61 ± 1.43	23.38 ± 0.24	22.87 ± 0.60	22.90 ± 2.06
SUN397	62.57 ± 0.00	69.13 ± 0.52	63.66 ± 0.21	61.98 ± 0.64	62.74 ± 1.60
DTD	43.09 ± 0.00	61.80 ± 0.28	42.59 ± 0.71	41.74 ± 0.18	40.07 ± 3.33
EuroSAT	48.65 ± 0.00	76.21 ± 2.67	52.48 ± 1.85	44.60 ± 1.86	51.26 ± 6.10
Cars	65.51 ± 0.00	69.60 ± 0.87	65.46 ± 0.18	64.51 ± 0.26	64.17 ± 0.14
RESISC	57.94 ± 0.00	76.44 ± 1.68	59.46 ± 0.30	57.07 ± 1.43	57.90 ± 0.92
UCF	63.66 ± 0.00	69.85 ± 2.62	63.96 ± 0.05	63.64 ± 0.28	60.99 ± 1.21
Mini-IN	85.76 ± 0.00	91.01 ± 0.65	85.85 ± 0.06	85.40 ± 0.26	83.26 ± 1.14

Dataset	Fed-DISC	FedZIP	FedBPT	FedBPTVLM	**Ours**
Caltech	88.78 ± 3.36	89.32 ± 1.82	90.99 ± 1.59	88.61 ± 2.10	92.67 ± 1.26
Pets	87.83 ± 1.22	84.23 ± 3.27	87.37 ± 1.42	85.50 ± 2.14	92.58 ± 1.24
CLEVR	18.94 ± 3.79	16.58 ± 6.54	25.52 ± 2.58	14.66 ± 1.34	26.03 ± 4.74
Flowers	68.43 ± 1.22	59.78 ± 2.33	70.19 ± 0.52	68.28 ± 0.27	85.44 ± 1.08
Food	85.88 ± 0.03	83.62 ± 1.39	85.48 ± 0.44	84.63 ± 0.76	86.41 ± 0.40
Aircraft	22.28 ± 0.46	14.17 ± 9.05	23.45 ± 2.31	23.29 ± 0.59	31.24 ± 0.60
SUN397	61.62 ± 0.32	62.67 ± 0.77	64.70 ± 0.89	58.79 ± 1.36	68.96 ± 0.21
DTD	40.74 ± 1.55	38.34 ± 1.34	39.20 ± 0.39	38.22 ± 2.75	58.49 ± 1.09
EuroSAT	54.00 ± 5.28	50.04 ± 3.88	53.77 ± 5.42	38.29 ± 4.20	68.37 ± 0.77
Cars	65.55 ± 0.42	62.55 ± 2.28	63.80 ± 0.24	63.26 ± 1.36	68.89 ± 0.50
RESISC	56.64 ± 3.10	53.06 ± 1.04	58.59 ± 0.87	52.95 ± 1.53	75.11 ± 1.07
UCF	62.50 ± 0.47	57.42 ± 0.83	64.59 ± 0.91	61.21 ± 1.15	68.88 ± 0.31
Mini-IN	82.68 ± 0.58	82.95 ± 0.73	85.67 ± 1.78	81.86 ± 1.75	88.82 ± 0.14

C BASELINE IMPLEMENTATION DETAILS

Federated prompting controls. FedBBT, FedBlackVIP, FedEffiBlackVIP, FedZIP, and FedBPTVLM provide text, visual, and multimodal prompting controls under distributed supervision. They test whether changing the local adaptation mechanism is enough to exploit heterogeneous clients. FedBPT and Fed-DISC additionally provide native federated derivative-free alternatives adapted to classification with a VLM.

Personalized versus global deployment. ZooPFL's private embeddings and local remappers define client-specific predictors (Lu et al., 2023b). A numerical comparison of its native method would require a personalized evaluation, whereas this study deploys the same predictor on every client. Collapsing the private modules would change the adaptation rule. We discuss ZooPFL as a personalized approach outside the global-predictor comparison.

Output interfaces. FedOT applies external orthogonal transformations to returned embeddings (Kong et al., 2026); these embeddings are absent from our score-returning output contract. Its encoder can remain black-box despite the external trainable classifier. Our own interface additionally assumes soft-prompt insertion, so neither interface is categorically weaker along every dimension. FDPT uses discrete prompts, textual responses, and a language agent (Wu et al., 2025); FedOne and FedPOB use discrete prompt learning or selection (Wang et al., 2025; Lu et al., 2025). Comparisons require harmonizing these input capabilities, auxiliary models, and all query costs.

Remaining comparisons. CraFT provides a centralized score-refinement design for opaque VLMs (Wang et al., 2024), while Fed3R develops global ridge-statistic aggregation for returned features (Fanì et al., 2024). FedMUSE draws on both ideas but evaluates a different combination: continuous multimodal prompt search followed by one residual solve in a frozen, shared class-score space. Direct numerical comparisons require harmonizing CraFT with distributed supervision and feature-based optimizers with the score-only interface; the current table therefore reports methods implemented under the shared global-predictor protocol.