Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.077556

ARTICLE

LiRA-CLIP: Training-Free Posterior-Predictive Uncertainty for Few-Shot CLIP Classification

Mustafa Qaid Khamisi[1], Zuping Zhang[1,*], Mohammed Al-Habib[1], Muhammad Asim[2] and Sajid Shah[2]

[1]School of Computer Science and Engineering, Central South University, Changsha, China
[2]EIAS Data Science Lab, College of Computer and Information Sciences, Prince Sultan University, Riyadh, Saudi Arabia

*Corresponding Author: Zuping Zhang. Email: zpzhang@csu.edu.cn

Received: 11 December 2025; Accepted: 13 February 2026; Published: 08 May 2026

ABSTRACT: Large Vision-Language models (VLMs) such as Contrastive Language-Image Pretraining (CLIP) have transformed open world image recognition. Nevertheless, few-shot classification, particularly in the extremely low-shot regime, requires not only high accuracy but also reliably calibrated uncertainty for decisions with high confidence. Existing training-free CLIP adapters are primarily designed to increase accuracy and efficiency; integrate the zero-shot text logits with the few-shot feature caches, but not definitely model predictive uncertainty and therefore often exhibit considerable miscalibration and weak selective performance. Bayesian adapters move in the direction of probabilistic modeling by placing priors over adapter parameters and employing task-specific variational training; however, this requires gradient-based optimization for every new task, increases computational costs, and becomes fragile when only one or two labeled examples per class are available. Starting from this observation, we introduce a training-free posterior-predictive Likelihood Ratio Adapter(LiRA-CLIP) for few-shot CLIP classification, which directly addresses probabilistic reliability under strict low-shot and deployment constraints. LiRA-CLIP extends the frozen CLIP head by a text-conditioned generative model in feature space that produces heavy-tailed posterior-predictive likelihood ratios, fused with the CLIP logits via a small, reliability-driven calibration layer. This layer is optimized in order to minimize the negative log-likelihood under an explicit accuracy side constraint, which leads to calibrated probabilities and dependable selective decisions without any gradient-based task-specific training. Extensive experiments show that LiRA-CLIP matches or slightly surpasses strong CLIP adapters in top-1 accuracy, while reducing calibration error by roughly 40%–50% and significantly increasing 95% and 99% reliable coverage in the low-shot regime, and thus establishes a new state of the art with respect to probabilistic reliability for training-free few-shot CLIP models.

KEYWORDS: Vision–language models; few-shot learning; CLIP; training-free; uncertainty calibration; selective classification; posterior predictive modeling

1 Introduction

Large pre-trained Vision–Language Models (VLMs) such as CLIP [1] have provided transferable representations and a prompt-based zero-shot encoder. Web-scale variants [2] trained with noisy text supervision increase transfer under distribution shift. Adaptation have been applied in few-shot image classification to improve over zero-shot prompting, but reported gains can strongly depend on task-specific model selection and may collapse under distribution shift in low-data settings [3]. Most adaptation methods are gradient-based, including prompt-learning methods that optimize continuous context tokens while keeping the CLIP backbone frozen [4,5], and lightweight tuning methods that optimize a minimal set of

task parameters on top of frozen CLIP encoders [6]. In addition, feature-adapter approaches introduce bottleneck modules downstream of the encoders, where these are the only parameters updated during training to ensure that the CLIP backbone remains frozen [7]. To bridge domain shifts and exploit limited labeled data, few-shot adapters refine or extend CLIP with limited supervision [8,9]. Training-free caching approaches such as Tip-Adapter [10] combine few-shot visual features with text prompts in order to obtain robust performance gains. Subsequent nonparametric and prototype-based adapters show that cache-like adaptation can rival or surpass traditional fine-tuning under strict time and compute budgets [11,12], and kernel-based analyses provide theoretical grounding and starting points for further improvements [13]. Recent methods have explored multimodal fusion, attention mechanisms, and prototype mechanisms in order to improve generalization and robustness [14,15].

Despite these advances, there remains, for current few-shot CLIP adapters, a gap between accuracy and probabilistic reliability, in particular in the extreme low-shot regime. Training-free cache and prototype adapters [8,13] are optimized for discriminative performance. They reweight zero-shot text logits or fuse them with support-set similarities of feature representations, but they treat these scores as deterministic functionals in the embedding space and do not model predictive uncertainty in a probabilistic manner. As a consequence, they often exhibit substantial miscalibration and weak performance under selective classification metrics, even when their top-1 accuracy is high. Few-shot OOD detectors generally target detection metrics instead of calibrated in-domain confidence [16]. Although meta-learned online adapters [9] indeed avoid task-specific fine-tuning, they still rely on extensive offline training and optimized for discriminative accuracy rather than for calibrated uncertainty.

By contrast, Bayesian adapters such as BayesAdapter [17] do introduce probabilistic structure; however, they place priors over adapter parameters and require gradient-based, task-specific training by means of variational inference. This improves the quality of uncertainty estimation in moderate data regimes, but increases adaptation time and, according to reports, shows particular brittleness or instability in the extreme low-shot regime, especially when only one or two labeled examples per class are available. To the best of our knowledge, there currently exists no training-free CLIP adapter that, under a strict few-shot protocol with only a handful of labeled examples per class, simultaneously performs adaptation in closed form without task-specific gradient-based optimization, explicitly targets calibrated probabilities and high-confidence selective metrics, and operates under strict latency and memory constraints that are suitable for deployment. Existing training-free adapters prioritize discriminative accuracy without a well-founded probabilistic uncertainty model, whereas Bayesian adapters trade training-free adaptation for task-specific variational optimization.

We address this gap by proposing LiRA-CLIP, a training-free posterior-predictive Likelihood Ratio Adapter for few-shot CLIP classification. LiRA-CLIP extends the frozen CLIP classifier by a text-conditioned Bayesian generative model over background-whitened image features, whose posterior-predictive likelihood ratios with respect to a pooled background define a single heavy-tailed generative evidence stream. This generative stream is fused with the CLIP logits through a unified, reliability-driven mechanism a margin-based confidence gate and a lightweight calibration layer which together act as a two-stream probabilistic adapter and, in a fully training-free, deterministic setting, produce calibrated probabilities and reliable high-confidence selective decisions. Across six standard benchmarks for few-shot adaptation and two CLIP backbones. Our training-free, text-conditioned Bayesian decision rule consistently improves calibration metrics (ECE and AECE) and high-confidence selective coverage at 95% and 99% target accuracy in the low-shot regime relative to plain CLIP, fine-tuned, and trained Bayesian adapters. As the budget of labeled data grows, LiRA-CLIP remains competitive in top-1 accuracy on all benchmarks while preserving its reliability advantages. These gains persist across a broad range of recent state-of-the-art adapters and

support posterior-predictive likelihood-ratio fusion as a well-founded and practical path toward trustworthy few-shot CLIP adaptation.

Our contributions can be summarized as follows:

- We introduce LiRA-CLIP a training-free CLIP adapter that extends CLIP by a text-conditioned, posterior-predictive generative model over whitened image features. Producing a reliable heavy-tailed Student-*t* likelihood–ratio (t-PLLR) stream specifically in the extreme low-shot regime and can be fused with CLIP without any gradient-based, task-specific optimization.
- We define a few-shot CLIP adaptation as a single reliability-driven calibration problem in a compact probabilistic adapter, where unifying both stream-specific temperatures and a global generative fusion coefficient. Leading to calibrated probabilities and reliable selective decisions at 95% and 99%, while retaining competitive top-1 accuracy.
- Through extensive experiments on six few-shot benchmarks with two CLIP backbones, Consistently LiRA-CLIP improves probabilistic reliability with clearly reduced calibration errors (ECE and AECE) and higher reliable coverage at 95% and 99% in the low-shot regime. It matches or closely trails the best existing adapters in top-1 accuracy. Additionally, we conduct targeted ablation studies on prior hyperparameters and a lightly fine-tuned variant (LiRA-CLIP-F) showing that these reliability gains are robust and that the fully training-free formulation is particularly advantageous under extreme data scarcity.

2 Related Work

2.1 Few-Shot CLIP Adaptation

Few-shot adaptation to new classes can be commonly categorized into three groups: prompt learning, which optimizes continuous tokens while the CLIP encoders remain frozen [4,5], adapters in the embedding space, connect zero-shot and few-shot evidence with small residual modules [3,6,7], and training-free and key-value cache models [10,13]. Although gradient-based prompt learning has demonstrated strong gains [5], its requirement constrains the use of prompt learning in scenarios in which CLIP is available as a frozen, purely forward-going (black-box) feature extractor. Adapter-based strategies instead operate in the CLIP embedding space using either fine-tuning a lightweight linear head or a shallow MLP on frozen features [3,17,18], or by relying on training-free key-value caches that are initialized from the support set [8,10]. LiRA-CLIP belongs to the family of adapter-based procedures, but is fully training-free and designed for a protocol with inaccessible weights and purely forward inference. Where both vision and text encoders of CLIP remain frozen, no gradients are computed, and the adaptation proceeds entirely via a posterior-predictive decision rule and reliability-driven calibration on frozen features.

2.2 Training-Free Caches

Cache-based CLIP adaptation combines support-set similarities with zero-shot text logits, as in Tip-Adapter and its variants [8,10,13]. APE [8] refines such cache priors by analyzing inter-class discrepancies and leveraging the three way interaction between image, cache, and text. It offers both a training-free variant (APE) and a lightweight trained variant (APE-T) for high accuracy with few parameters. ProKeR [13] formalizes Tip-Adapter as a Nadaraya-Watson local estimator and highlights the benefit of global information via a proximal RKHS regularizer that is solved in closed form, leads to significant improvements in performance. These cache-based methods; however, treat cache scores as deterministic similarity functionals in feature space, and do not directly target calibrated uncertainty or selective decisions in the extreme low-data regime. Building on the low-overhead cache logic, LiRA-CLIP replaces deterministic cache weightings

with text-anchored Bayesian posterior-predictive scoring in a background-whitened CLIP space. And uses a pooled background Student-t posterior-predictive likelihood-ratio measure (t-PLLR) as a global generative reference without the need for any gradient-based optimization.

2.3 Bayesian Uncertainty and Calibration in CLIP Adapters

Recent work investigates uncertainty and calibration in VLMs adaptation methods [19–21]. BayesAdapter [17] in particular shows that a strong adapter can be interpreted as a maximum a posteriori solution in a probabilistic framework, and that the transition from a point estimate to a Bayesian posterior over adapter parameters improves calibration and selective classification. These Bayesian approaches for few-shot learning employ priors to adjust parameters and refine them through task-specific training. Yet, they have some limitations, specifically in the extreme low shot regime where their advantages diminish in the 1-shot setting or drop, and also demand considerable optimization costs.

LiRA-CLIP is complementary to this Bayesian parameter-space category but instead of placing priors over adapter weights and updating them through task-specific training, it performs training-free Bayesian modeling directly in feature space. Using a text-conditioned generative model over background-whitened CLIP representations to provide a posterior-predictive likelihood-ratio score with respect to a pooled background. Then, fused it with CLIP logits via a small, accuracy-protected calibration layer. Leading to calibrated probabilities and reliable high-confidence selective decisions in a fully training-free, posterior-predictive setting that, to our knowledge, is not obviously addressed by existing CLIP adaptation methods.

Our study focuses on the gap between accuracy and probabilistic reliability in the extreme low-shot regime for few-shot CLIP classification, with frozen encoders and deployment constraints. Targeted calibrated point probabilities and high-confidence, accuracy-constrained selective classification using a training-free adapter that extends the frozen CLIP head without gradient-based, task-specific training.

We have studied training-free few-shot CLIP classification with frozen encoders. LiRA-CLIP performs closed-form posterior-predictive updates in feature space and applies a fixed two-stream fusion rule, producing class probabilities over fused logits, Algorithm 1, Eqs. (20) and (21). We selected the fusion parameters by reliability-driven global calibration on an auxiliary task pool. We reused them unchanged across tasks, Section 3.9, Eq. (19). We evaluate point-probability reliability using Expected Calibration Error (ECE) and Adaptive ECE (AECE) and and assess selective classification by measuring coverage under accuracy constraints at 95% (Sel@95) and 99% (Sel@99) target accuracy as defined in Section 4, Tables 1 and 2. Some of the baseline methods are restricted to CLIP adapters and evaluated under the same strict few-shot protocol with frozen CLIP encoders and logits-to-probabilities outputs, making them directly comparable, Section 4. Other uncertainty quantification approaches target different objectives, conformal prediction [22] outputs prediction sets with coverage guarantees rather than point-probability calibration. Related probabilistic extensions of frozen VLMs also study uncertainty under different downstream tasks or uncertainty representations including generalized few-shot semantic segmentation [23] or post-hoc probabilistic embeddings via GPLVM [24]. These methods are complementary, not directly comparable to our point-probability calibration (ECE) and selective classification (Sel@99) results under the evaluation protocol in Section 4.

2.4 LiRA-CLIP's Key Distinction from Prior Work

We distinguished LiRA-CLIP from prior probabilistic VLM adaptation not by the presence of uncertainty modeling but by where uncertainty is represented and how it is computed at deployment. BayesAdapter [17] models uncertainty in parameter space employing priors over adapter parameters and performing Bayesian posterior inference via task-time optimization such variational updates, instead to

employ closed-form prediction. ProbVLM [20] instead learns a probabilistic embedding adapter, trained offline via gradient-based optimization to produce output distributions over frozen VLM embeddings. By contrast, LiRA-CLIP is a training free method, performs closed-form posterior-predictive inference directly in frozen CLIP feature space at task time, with no gradient-based updates. Specifically, when given only support set sufficient statistics in a background-whitened representation Eq. (2), a text-conditioned bind NIG prior, Eqs. (4) and (5) leading to a Student-t posterior predictive, Eqs. (6)–(9) and a pooled-background likelihood-ratio evidence stream Eqs. (10) and (11). And then it fused with CLIP logits via a globally calibrated and frozen scalar rule, Eqs. (16)–(19). This approach keeps the deployment profile strictly forward only, no backpropagation and no per-sample test-time optimization as in C-TPT [19], and does not require any per-task calibration or tuning at test time.

3 Methodology

In this section, we present LiRA-CLIP, a training-free posterior-predictive likelihood–ratio adapter for CLIP. LiRA-CLIP couples a text-conditioned Student-t generative head with the standard CLIP discriminative head via a confidence gate, and formulates few-shot adaptation as a single, reliability-driven calibration problem. An overview of the LiRA-CLIP architecture is shown in Fig. 1.

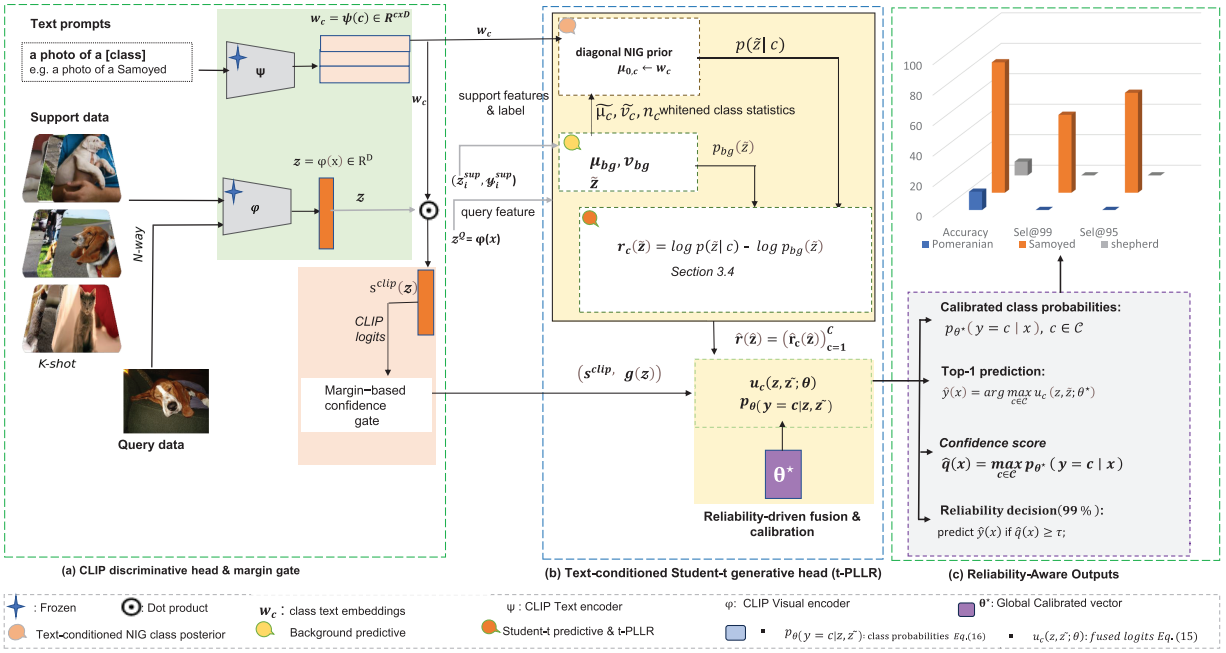

Figure 1: Overview of LiRA-CLIP framework, a training-free posterior-predictive likelihood-ratio adapter for few-shot CLIP classification. Fig. 1a, a frozen CLIP image encoder φ and text encoder ψ produce image features $z = \varphi(x)$ and class prompts $w_c = \psi(c)$. Their dot products yield zero-shot logits $s_c^{\text{clip}}(z)$ and a margin-based confidence gate $g(z)$ (discriminative stream). Fig. 1b, support-whitened features $\tilde{z}$ feed a text-conditioned diagonal normal-inverse-gamma (NIG) model, leading to Student-t posterior-predictive densities $p(\tilde{z} \mid c)$ and a pooled background density $p_{\text{bg}}(\tilde{z})$. Their standardized log-likelihood ratios $\hat{r}_c(\tilde{z})$ form the t-PLLR generative stream. Connected with a lightweight reliability-driven calibration layer fuses the discriminative and generative streams into final logits which are converted into calibrated class probabilities $p_{\theta^\star}(y = c \mid x)$ and high-confidence selective decisions as shown in Fig. 1c, with no gradient-based tuning or task-specific training.

3.1 Problem Setting

We consider a N-way few-shot classification task with a class set $\mathcal{N} = \{1, \ldots, N\}$. The labeled support set is $\mathcal{D}_{\text{sup}} = \{(x_i^{\text{sup}}, y_i^{\text{sup}})\}_{i=1}^{N_{\text{sup}}}$, with $y_i^{\text{sup}} \in \mathcal{N}$, and we evaluate on an unlabeled test set $\mathcal{D}_{\text{test}} = \{x_i^{\text{test}}\}_{i=1}^{N_{\text{test}}}$. We use a frozen CLIP image encoder $\varphi : \mathcal{X} \to \mathbb{R}^D$ and text encoder $\psi : \mathcal{N} \to \mathbb{R}^D$. For an image x and class $c \in \mathcal{N}$, we define:

$$z = \varphi(x), \qquad w_c = \psi(c) \in \mathbb{R}^D, \tag{1}$$

where w_c is obtained from a fixed prompt template, as in standard CLIP zero-shot classification [1]. On the support set we compute empirical per-class means $\bar{z}_c$, diagonal variances s_c^2, and counts n_c. We also compute pooled background statistics $(\mu_{\text{bg}}, \nu_{\text{bg}})$ across all support samples, with total count $n_{\text{bg}} = N_{\text{sup}}$. For one-shot classes ($n_c \leq 1$), we tie the class variance to the background variance for numerical stability.

3.2 Background Whitening

Whitening is a technique in signal and image processing, used to transform correlated background or clutter into approximately white noise so as to enhance target detectability in challenging environments such as in SAR speckle whitening [25]. Related ideas appear in deep networks through switchable whitening and normalization layers, which decorrelate features and improve optimization stability across tasks [26]. In LiRA-CLIP we adopt a lightweight variant at the representation level. We standardize all features using the pooled support statistics $(\mu_{\text{bg}}, \nu_{\text{bg}})$. Let $\sqrt{\nu_{\text{bg}}}$ denote the element-wise square root and $\oslash$ element-wise division. For any feature vector $z \in \mathbb{R}^D$ define as:

$$\tilde{z} = (z - \mu_{\text{bg}}) \oslash \sqrt{\nu_{\text{bg}}}. \tag{2}$$

Eq. (2) performs pooled-statistics, per-dimension standardization using μ_{bg} *and* ν_{bg} computed from the support set, Section 3.1. Since our generative head models $\tilde{z}$ with diagonal likelihood and uses closed-form text-conditioned NIG updates, Section 3.3, this standardization controls coordinate-wise scale when computing Student-t posterior predictives and the resulting posterior-predictive log-likelihood ratios, Eqs. (6)–(11). As with any pooled-statistic standardization, $(\mu_{\text{bg}}, \nu_{\text{bg}})$ may be higher-variance when the support set is extremely small or class frequencies are highly imbalanced. LiRA-CLIP incorporates stability safeguards already in the method: for one-shot classes we tie the class variance to the background variance, Section 3.1. We applied a variance floor $\varepsilon > 0$ in the per-sample standardization of LLRs, Eq. (11). And we enforced a minimum predictive degrees-of-freedom in the Student-t posterior predictive, Section 3.3. These choices are designed to mitigate degeneracy of likelihood-ratio scores and keep the posterior-predictive stream well-conditioned in the extreme low-shot setting. Empirically, LiRA-CLIP exhibits non-degenerate behavior in the low-shot regime across benchmarks, as reflected by improved ECE, AECE and selective coverage, Section 4.

To express the text-conditioned generative model in the same background-whitened coordinates as $\tilde{z}$. We also whiten the CLIP text embedding using the pooled support background statistics $(\mu_{\text{bg}}, \nu_{\text{bg}})$, Section 3.1. For each class c, we define the whitened text anchor as:

$$\tilde{w}_c = (w_c - \mu_{\text{bg}}) \oslash \sqrt{\nu_{\text{bg}}} \in \mathbb{R}^D. \tag{3}$$

We have used $\tilde{w}_c$ only in the posterior-predictive (generative) stream, while the standard CLIP discriminative stream remains defined on the original features z and text embeddings w_c, Section 3.6.

3.3 Text Conditioned Diagonal NIG Class Priors

The Normal Inverse Gamma (NIG) prior is a standard conjugate prior for the mean and variance of a normal likelihood and underpins a wide range of Bayesian regression and hierarchical models [27,28]. Integrating out the latent mean and variance yields Student-t posterior-predictive distributions. Where it naturally accommodate heavy tails and provide robustness to outliers and model misspecification [29]. In our method LiRA-CLIP, we model the class-conditional distribution of whitened image features $\tilde{z} \mid y = c$ as a diagonal Gaussian and place a NIG prior on its parameters (μ_c, σ_c^2). The prior is text-conditioned, its mean is anchored in the whitened text embedding $\tilde{w}_c$, Eq. (3), and therefore, it is defined in the same coordinate system as the likelihood for $\tilde{z}$. In the whitened space, we employ an independent NIG prior to each class-dimension pair (c, d) as:

$$\sigma_{cd}^2 \sim \text{Inv} - \text{Gamma}(\alpha_0, \beta_0), \quad \mu_{cd} \mid \sigma_{cd}^2 \sim \mathcal{N}\big(\mu_{0,cd}, \sigma_{cd}^2/\kappa_0\big), \tag{4}$$

with positive hyperparameters $(\kappa_0, \alpha_0, \beta_0)$ shared across classes and dimensions. Our key design choice is to set the class prior mean to this whitened text anchor as follows:

$$\mu_{0,c} = \tilde{w}_c \in \mathbb{R}^D, \qquad \tilde{w}_c \text{ defined in Eq. (3).} \tag{5}$$

Using the same class embedding that defines CLIP's frozen zero-shot linear head, Eq. (12). Since CLIP defines the image feature $z = \varphi(x)$ and text embedding of class $w_c = \psi(c)$ in the same D-dimensional representation space, Eq. (1) and compares them directly via dot products, applying the same background-whitening reparameterization to both, Eq. (2). IT simply, expresses this class anchor in the coordinate system of the generative likelihood for $\tilde{z}$. Where it also is consistent with our pooled background predictive construction, whose mean is defined as a whitened function of the text embeddings, Eq. (8).

3.4 Student-t Posterior Predictive and Log-Likelihood Ratios

Log-likelihood ratios are central to statistical decision theory, hypothesis testing, and modern likelihood-based machine learning [30,31]. LiRA-CLIP leverages this principle in a posterior-predictive setting. For each class, we form the log-ratio between its Student-t predictive density and a pooled background predictive in the background-whitened CLIP feature space and then apply per-sample standardization across classes. Resulting in Student-t posterior-predictive log-likelihood ratios (t-PLLRs), defining a single stream of generative evidence, which we later fused with CLIP logits by our reliability-driven calibration mechanism, Section 3.9. For a whitened feature $\tilde{z} \in \mathbb{R}^D$, the class-$c$ posterior predictive density factorizes across dimensions define as:

$$p(\tilde{z} \mid c) = \prod_{d=1}^{D} t_{\nu_c}\big(\tilde{z}_d \,\big|\, m_{cd}, s_{cd}^2\big), \tag{6}$$

where $t_\nu(\cdot \mid \mu, s^2)$ denotes the univariate Student-t distribution with degrees of freedom ν, location μ, and scale s. The predictive variance in dimension d has the usual NIG form:

$$s_{cd}^2 = \frac{\beta_{cd}}{\alpha_c} \cdot \frac{\kappa_c + 1}{\kappa_c}. \tag{7}$$

We denote the resulting generative log-likelihood by $\ell_c(\tilde{z}) = \log p(\tilde{z} \mid c)$.

3.5 Background Predictive Model

We construct a background Student-t predictive in the same whitened space using the global statistics $(\mu_{\mathrm{bg}}, v_{\mathrm{bg}}, n_{\mathrm{bg}})$ and the text embeddings, defined as follows:

$$\mu_{\mathrm{bg}}^{(\mathrm{t})} = \left(\frac{\frac{1}{C}\sum_{c=1}^{C} w_c - \mu_{\mathrm{bg}}}{\sqrt{v_{\mathrm{bg}}}} \right), \tag{8}$$

corresponding to the whitened mean of the class text embeddings, and take a unit background variance $v_{\mathrm{bg}}^{(\mathrm{t})} = \mathbf{1}$. Using n_{bg}, we obtain a diagonal Student-t posterior predictive from:

$$p_{\mathrm{bg}}(\tilde{z}) = \prod_{d=1}^{D} t_{\nu_{\mathrm{bg}}}\left(\tilde{z}_d \,\middle|\, \mu_{\mathrm{bg},d}^{(\mathrm{t})},\, s_{\mathrm{bg},d}^2 \right), \tag{9}$$

with degrees of freedom ν_{bg} and scale parameters $s_{\mathrm{bg},d}^2$ defined analogously to the class case. We denote the background log-likelihood by $\ell_{\mathrm{bg}}(\tilde{z}) = \log p_{\mathrm{bg}}(\tilde{z})$. For each class c, we form the log-likelihood ratio (LLR) as:

$$r_c(\tilde{z}) = \ell_c(\tilde{z}) - \ell_{\mathrm{bg}}(\tilde{z}). \tag{10}$$

To stabilize the scale of generative scores across images, we apply per-sample z-scoring across classes. Let $r(\tilde{z}) = (r_c(\tilde{z}))_{c=1}^{C}$ and denote its empirical mean and variance across classes by $\bar{r}(\tilde{z})$ and $v_r(\tilde{z})$, with a small variance floor $\varepsilon > 0$. We get the standardized generative score via:

$$\hat{r}_c(\tilde{z}) = \frac{r_c(\tilde{z}) - \bar{r}(\tilde{z})}{\sqrt{v_r(\tilde{z}) + \varepsilon}}, \qquad c \in \mathcal{C}. \tag{11}$$

These standardized posterior-predictive LLRs constitute the generative stream used by the adapter.

3.6 CLIP Discriminative Stream

The CLIP zero-shot classifier provides the discriminative stream via a linear head whose weights are given by the text embeddings w_c [1,4,10]. For a feature vector z, we define the CLIP logits as:

$$s_c^{\mathrm{clip}}(z) = \beta_{\mathrm{clip}} \langle z, w_c \rangle, \qquad c \in \mathcal{C}, \tag{12}$$

where $\beta_{\mathrm{clip}} > 0$ is the standard CLIP logit-scale (temperature) parameter [1]. These logits are converted to probabilities via a softmax, $p^{\mathrm{clip}}(y = c \mid z) \propto \exp\left(s_c^{\mathrm{clip}}(z)\right)$.

3.7 Margin Based Confidence Gate

To modulate the contribution of the generative stream, we construct a scalar confidence gate from the CLIP logits. Let $s_{(1)}(z)$ and $s_{(2)}(z)$ denote the largest and second-largest components of $s^{\mathrm{clip}}(z)$, and define the top-2 margin as follows:

$$m(z) = s_{(1)}(z) - s_{(2)}(z). \tag{13}$$

The margin is large and positive when CLIP is confident and small when it is ambiguous.

We map this margin to a raw gate via a squashing nonlinearity:

$$\tilde{g}(z) = \sigma\left(k_{\mathrm{gate}}\left(t_0 - \tanh m(z)\right)\right), \tag{14}$$

where $k_{\text{gate}} > 0$ and t_0 are hyperparameters and $\sigma(\cdot)$ is the logistic sigmoid. This parametrization yields gates close to one when CLIP is uncertain (small margin) and close to zero when CLIP is highly confident (large margin). We then normalize and recenter $\tilde{g}(z)$ using statistics from the defined set, optionally apply a power transform, and clamp the result to a fixed interval $[g_{\min}, g_{\max}]$, obtaining the final gate $g(z) \in [g_{\min}, g_{\max}]$. The gate is thus a deterministic function of the CLIP logits and introduces no additional learned parameters beyond those calibrated in the fusion stage.

3.8 Two-Stream Fusion

For each sample, LiRA-CLIP fuses the CLIP discriminative stream with the standardized generative LLR stream. Let $s_c^{\text{clip}}(z)$ denote the CLIP logits and $\hat{r}_c(\tilde{z})$ the standardized t-PLLR scores. Fusion parameters collected into a single vector as follows:

$$\theta = (\gamma_{\text{clip}}, \gamma_{\text{llr}}, \alpha), \tag{15}$$

where $\gamma_{\text{clip}} > 0$ and $\gamma_{\text{llr}} > 0$ are per-stream temperatures and $\alpha \geq 0$ controls the gated generative contribution. To calculate the fused logits, we used:

$$u_c(z, \tilde{z}; \theta) = \gamma_{\text{clip}}\, s_c^{\text{clip}}(z) + \alpha\, g(z)\, \gamma_{\text{llr}}\, \hat{r}_c(\tilde{z}), \qquad c \in \mathcal{C}, \tag{16}$$

and then we calculate class probabilities with:

$$p_\theta(y = c \mid z, \tilde{z}) = \frac{\exp\big(u_c(z, \tilde{z}; \theta)\big)}{\sum_{c' \in \mathcal{C}} \exp\big(u_{c'}(z, \tilde{z}; \theta)\big)}. \tag{17}$$

When $\alpha = 0$, Eqs. (16) and (17) reduce to temperature-scaled CLIP, setting additionally $\gamma_{\text{clip}} = 1$ recovers the original CLIP predictions. For $\alpha > 0$, the generative t-PLLR stream is adaptively up or down-weighted by the gate $g(z)$ depending on CLIP confidence. Once background whitening, priors, and the gate form are fixed, the entire adapter is parameterized only by θ: the CLIP stream, t-PLLR stream, and gate influence decisions entirely via the fused logits $u(z, \tilde{z}; \theta)$ and probabilities p_θ.

3.9 Reliability-Driven Global Calibration

We define LiRA-CLIP as the solution of a single reliability-driven calibration problem over θ, rather than tuning the gate and per-stream temperatures independently. Treated it as a unified probabilistic mechanism and choose a single global vector $\theta^\star = (\gamma^\star_{\text{clip}}, \gamma^\star_{\text{llr}}, \alpha^\star)$ by solving a constrained optimization problem on a small pool of auxiliary few-shot classification tasks whose label sets are disjoint from those used in our main evaluation.

Auxiliary Calibration Pool Composition

We built the auxiliary pool from five datasets that span different visual domains to reduce the risk of overfitting to specific task characteristics. These domains include Caltech101 (generic objects), DTD (textures), FGVC-Aircraft (fine-grained), OxfordPets (fine-grained), and UCF101 (actions). We performed a fixed, class-disjoint split into calibration classes and evaluation classes (80/20) for each dataset. Sampling auxiliary few-shot tasks exclusively from the calibration classes. $\theta^\star$ is selected using only the auxiliary calibration pool, no evaluation classes, benchmark datasets, or test episodes are used to choose $\theta^\star$. We excluded EuroSAT from the calibration pool because it contains only 10 classes, which makes class-disjoint calibration splits statistically small and unstable. EuroSAT is therefore used as an out-of-calibration transfer benchmark, evaluated using the same frozen $\theta^\star$ without any dataset-specific tuning. Each auxiliary task

follows the same few-shot protocol as in Section 4. for a chosen shot $K \in \{1, 2, 4, 8, 16, 32\}$, we sample K labeled support examples per class from the training split and form a labeled development/query set by sampling $Q = 16$ additional examples per class from the remaining training data (capped by availability). Generated a total of $T = 120$ auxiliary tasks, balanced across datasets and shot regimes (4 random episodes per dataset-shot pair). On this auxiliary pool, we minimized the mean negative log-likelihood by:

$$\mathcal{L}_{\mathrm{NLL}}(\theta) = -\frac{1}{N_{\mathrm{dev}}} \sum_{i=1}^{N_{\mathrm{dev}}} \log p_\theta^{(i)}\left(y_i^{\mathrm{dev}}\right), \tag{18}$$

where N_{dev} is the total number of development or query examples across all auxiliary tasks.

Formally, we select $\boldsymbol{\theta}^\star$ via:

$$\theta^\star \in \arg \min_{\substack{\theta \in \mathcal{G} \\ A(\theta) \geq A_{\mathrm{guard}}}} \mathcal{L}_{\mathrm{NLL}}(\theta), \tag{19}$$

where $\mathcal{G}$ is a fixed grid. Where A_{guard} specifies an accuracy slack on the auxiliary pool (i.e., we require $A(\theta) \geq A_{\max} - A_{\mathrm{guard}}$), protecting accuracy while selecting $\theta^\star$ by NLL. The resulting $\theta^\star$ is reused unchanged for all tasks, datasets, backbones, and shot regimes, with no test-time tuning. At deployment, LiRA-CLIP remains fully training-free: for each new few-shot task it performs only closed-form posterior-predictive updates and evaluates Eqs. (16) and (17). Although Eq. (19) is a constrained optimization problem, we solve it by grid search because the fusion vector $\theta = (\gamma_{\mathrm{clip}}, \gamma_{\mathrm{llr}}, \alpha)$ has only three scalar degrees of freedom. A fixed grid $\mathcal{G}$ yields a deterministic and stable selection of $\theta^\star$ and avoids introducing any gradient-based optimization into the calibration stage, which aligns with our deployment setting. At test time, the adapter remains fully training-free: for each new few-shot task it performs only closed-form posterior-predictive updates in the Student-t head and evaluates Eqs. (16) and (17) with the frozen $\theta^\star$, with no per-task tuning.

3.10 Test-Time Prediction with Frozen Fusion Parameters

Algorithm 1 describes how LiRA-CLIP adapts itself to a new few-shot task exclusively by means of closed-form computations and a globally calibrated, frozen fusion rule. Given the cached streams and the gate, the final phase applies the globally calibrated fusion parameters $\theta^\star$ to each test. For each $x^{\mathrm{test}} \in \mathcal{D}_{\mathrm{test}}$, the algorithm forms fused logits

$$u_c = \gamma_{\mathrm{clip}}^\star \, s_c^{\mathrm{clip}}(z^{\mathrm{test}}) + \alpha^\star \, g(z^{\mathrm{test}}) \, \gamma_{\mathrm{llr}}^\star \, \hat{r}_c(\tilde{z}^{\mathrm{test}}), \tag{20}$$

and then, converts these fused logits by a softmax into class probabilities $p_{\theta^\star}(y = c \mid x^{\mathrm{test}})$; we evaluate p_θ in Eq. (17) at $\theta = \theta^\star$ and at the features $(z^{\mathrm{test}}, \tilde{z}^{\mathrm{test}})$ derived from x^{test}, with:

$$p_{\theta^\star}(y = c \mid x^{\mathrm{test}}) = \frac{\exp\left(u_c(z^{\mathrm{test}}, \tilde{z}^{\mathrm{test}}; \theta^\star)\right)}{\sum_{c' \in \mathcal{C}} \exp\left(u_{c'}(z^{\mathrm{test}}, \tilde{z}^{\mathrm{test}}; \theta^\star)\right)}, \tag{21}$$

where the vector $\theta^\star$ is not updated on the new task, adaptation is completely training-free and reduces to closed-form posterior-predictive updates in the Student-t head together with evaluation of this frozen fusion rule.

Algorithm 1: LiRA-CLIP: training-free two-stream t-predictive adapter for a new few-shot task

Input: support set $\mathcal{D}_{\text{sup}}$, test set $\mathcal{D}_{\text{test}}$; frozen CLIP encoders φ, ψ; fixed hyperparameters; globally calibrated fusion parameters $\theta^\star = (\gamma^\star_{\text{clip}}, \gamma^\star_{\text{llr}}, \alpha^\star)$.
Output: $p_{\theta^\star}(y \mid x)$ for all $x \in \mathcal{D}_{\text{test}}$.

Feature extraction
for each image x in $\mathcal{D}_{\text{sup}} \cup \mathcal{D}_{\text{test}}$ **do**
    Compute CLIP image feature $z = \varphi(x)$.
end for
for each class $c \in \mathcal{N}$ **do**
    Compute CLIP text embedding $w_c = \psi(c)$ Eq. (1).
end for
Support statistics and whitening
From $\mathcal{D}_{\text{sup}}$, compute per-class statistics $(\bar{z}_c, s_c^2, n_c)$ and pooled background statistics $(\mu_{\text{bg}}, \nu_{\text{bg}}, n_{\text{bg}})$.
Whiten support and test features to obtain $\tilde{z}$ Eq. (2).
Text-conditioned NIG priors and *t*-predictives
For each class c, form text-conditioned NIG priors anchored at w_c Eq. (4) with $\mu_{0,c} = w_c$).
Calculate $p(\tilde{z} \mid c)$ and $p_{\text{bg}}(\tilde{z})$ Eqs. (6)–(9).
Fixed streams and gate (no tuning at task time)
for each image x in $\mathcal{D}_{\text{sup}} \cup \mathcal{D}_{\text{test}}$ **do**
    Compute generative log-likelihoods $\ell_c(\tilde{z}) = \log p(\tilde{z} \mid c)$ and $\ell_{\text{bg}}(\tilde{z})$; form LLRs $r_c(\tilde{z}) = \ell_c(\tilde{z}) - \ell_{\text{bg}}(\tilde{z})$ Eq. (10).
    Standardize across classes to obtain $\hat{r}_c(\tilde{z})$ Eq. (11) ▷ generative t-PLLR stream.
    Compute CLIP logits $s_c^{\text{clip}}(z)$ Eq. (12) ▷ discriminative stream.
    Compute margin $m(z)$ and confidence gate $g(z)$ Eqs. (13), (14).
end for
Test-time prediction with frozen fusion parameters
for each $x^{\text{test}} \in \mathcal{D}_{\text{test}}$ **do**
    Using cached $\hat{r}(\tilde{z}^{\text{test}})$, $s^{\text{clip}}(z^{\text{test}})$ and $g(z^{\text{test}})$, form fused logits Eq. (20)
    and convert them to class probabilities $p_{\theta^\star}(y = c \mid x^{\text{test}})$ Eq. (21).
end for

4 Experiments

4.1 Experimental Setting

4.1.1 Datasets and Protocol

In line with earlier work on CLIP adapters [6,7,10,17], we evaluate on six established vision benchmarks: Oxford Pets [32], Caltech101 [33], FGVC-Aircraft [34], DTD [35], EuroSAT [36], and UCF101 [37]. Taken together, these datasets cover generic object recognition, fine-granular categories, textures, and remote sensing. We adopt the strict few-shot adaptation protocol from [3,6,17], for each task and each class we draw uniformly K labeled support examples from the training split, with $K \in 1, 2, 4, 8, 16, 32$, and use the official test partition for evaluation. All reported LiRA-CLIP results are averaged over three random seeds. Unless stated otherwise, we use a single set of hyperparameters and a globally calibrated fusion vector $\theta^\star$ that is fixed for all tasks; at test time, there is no task-specific tuning or additional supervision.

4.1.2 Baselines

We compare LiRA-CLIP with eight recent CLIP adapters: Linear Probing (LP) [1], TIP-Adapter and TIP-Adapter-f [10], TaskRes [6], CrossModal [38], BayesAdapter [17], LP++ [18], and CLAP [3]. It is noteworthy that all baseline results reported in this work are taken from [17].

4.1.3 Implementation Details

We compute CLIP features with two common visual encoders, ResNet-50 [39] and ViT-B/16 [40], unless indicated otherwise, ablation studies are carried out on ResNet-50. During feature extraction we apply random zoom, crop, and horizontal flip augmentations, following [3,6,17], and we reuse the same text prompt templates. For the fine-tuning ablation LiRA-CLIP-F, in which only the fusion parameters are updated while CLIP and the generative head remain frozen, we adopt the training configuration from [17]: 300 epochs, batch size 256, as well as SGD with momentum 0.9 and a learning rate of 0.1. All experiments for LiRA-CLIP and LiRA-CLIP-F are averaged over three random seeds. We reported the mean performance in the main text, while standard errors and dataset-specific results are provided in the appendices.

4.2 Analysis of the Experimental Results

4.2.1 Calibration

We first investigate how well confidence values reflect actual correctness. In accordance with common practice, we report the Expected Calibration Error (ECE) [17,41] as well as its adaptive variant AECE, which reduces the bias that arises when standard ECE is influenced by bins with insufficient or zero sample size [17,42]. ECE partitions the predictions into B confidence bins and averages the absolute difference between empirical accuracy and mean confidence in each bin; we use $B = 10$. AECE keeps the same definition but chooses the bin boundaries such that each bin contains approximately the same number of samples, thereby reducing artefacts due to sparsely populated regions of the confidence range.

4.2.2 Overall Accuracy and Calibration

We have reported means over six datasets and six shot regimes ($K \in 1, 2, 4, 8, 16, 32$), using three random seeds per setting on two backbones. The results are shown in Table 1. On ResNet-50, LiRA-CLIP attains the highest average top-1 accuracy (68.44%), remaining essentially on a par with the strongest previous adapters such as CrossModal, CLAP, and BayesAdapter (67.85%–68.26%). At the same time, LiRA-CLIP exhibits a clearly recognizable gain in probabilistic reliability, its ECE and AECE (2.49% and 2.45%, respectively) correspond to a reduction of about 40%–45% compared to the best-calibrated baseline (BayesAdapter, 4.32% and 4.24%), whereas strongly accuracy-oriented methods such as LP++, CLAP, and Tip-Adapter-f show markedly higher calibration errors. For ViT-B/16, LiRA-CLIP remains competitive in accuracy (74.36%), with a performance that is statistically not distinguishable from the strongest baselines (CrossModal, CLAP, TaskRes; 74.16%–74.42% with overlapping standard deviations). In this setting as well, LiRA-CLIP attains the best calibration, reducing ECE from 3.46%–4.19% (CrossModal, BayesAdapter) to 2.28% and AECE from 3.38%–4.14% to 2.36%. Across both architectures, LiRA-CLIP thus offers a fully training-free adaptation with state-of-the-art calibration and competitive top-1 accuracy and consistently delivers more reliable probability estimates than all considered CLIP-adapter baselines. For more detailed results, see Appendix A.1, Table A1, and Appendix A.2, Table A2, and Appendix A.3, Table A3.

Table 1: Calibration and discrimination comparison across two backbones.

Method	ResNet50			ViT-16		
	Accuracy (%)	ECE	AECE	Accuracy (%)	ECE	AECE
LP [1]	54.136 ± 1.311	25.621 ± 1.330	25.614 ± 1.326	61.852 ± 1.047	22.014 ± 1.272	22.001 ± 1.273
TipA [10]	60.405 ± 0.364	8.068 ± 0.393	8.005 ± 0.398	62.009 ± 0.943	33.646 ± 0.993	33.624 ± 0.989
TipA-f [10]	65.754 ± 0.490	6.296 ± 0.438	6.206 ± 0.445	63.956 ± 0.985	31.512 ± 1.019	31.493 ± 1.014
CrossModal [38]	68.046 ± 0.576	4.383 ± 0.463	4.367 ± 0.468	**74.421** ± 0.417	3.459 ± 0.382	3.376 ± 0.397
TaskRes [6]	67.854 ± 0.521	5.676 ± 0.484	5.691 ± 0.484	74.238 ± 0.446	4.211 ± 0.420	4.190 ± 0.439
LP++ [18]	67.306 ± 0.647	11.378 ± 1.748	11.397 ± 1.733	74.164 ± 0.494	5.951 ± 1.523	5.959 ± 1.514
CLAP [3]	68.262 ± 0.469	7.169 ± 0.460	7.146 ± 0.454	74.379 ± 0.428	5.711 ± 0.403	5.675 ± 0.392
BayesAdapter [17]	67.937 ± 0.544	4.323 ± 0.436	4.236 ± 0.438	73.992 ± 0.509	4.185 ± 0.490	4.144 ± 0.497
LiRA-CLIP(ours)	**68.441** ± 0.514	**2.492** ± 0.379	**2.448** ± 0.394	74.361 ± 0.616	**2.284** ± 0.284	**2.359** ± 0.299

Note: Best result is **bold**; Second best underlined; Higher is better for Accuracy; Lower is better for ECE and AECE.

4.2.3 Selective Classification at High Confidence

We evaluate selective classification under high-confidence conditions, a central requirement in safety-critical deployment scenarios [17,43]. Given a confidence threshold τ, the classifier only predicts for those points whose maximum class probability exceeds τ, and abstains on all others. Following [17,44,45], we call a method reliable at level X% if its accuracy on the selected subset is at least X%. Under this side constraint, the objective is to maximize coverage, that is, the fraction of test examples for which the system issues a prediction. Table 2 reports overall reliable coverage at reliability levels of 95% and 99% on the test set, where LiRA-CLIP attains the highest coverage consistently on both backbones. On ResNet-50, it improves 99% reliable coverage from 10.8% to 17.6% and 95% reliable coverage from 22.2% to 28.3% relative to BayesAdapter, the strongest previous uncertainty baseline. On ViT-B/16, LiRA-CLIP increases 95% and 99% reliable coverage from 31.1% and 16.6% to 36.0% and 21.9%, respectively. These aggregated results support our central claim that LiRA-CLIP provides a reliability-focused, training-free adaptation with superior coverage at high confidence levels.

Table 2: Average reliable coverage (Sel@95 and Sel@99, %), computed over six few-shot benchmarks with $K \in 1, 2, 4$ and three random seeds. For more details results see Appendix A.4, Table A4.

Method	ResNet-50		ViT-B/16	
	Sel@99	Sel@95	Sel@95	Sel@99
CrossModal [38]	8.85 ± 0.50	20.07 ± 0.52	30.53 ± 0.65	14.13 ± 0.53
TaskRes [6]	7.63 ± 0.41	18.46 ± 0.48	27.40 ± 0.60	12.73 ± 0.53
LP++ [18]	0.22 ± 0.10	2.27 ± 0.59	14.87 ± 2.19	2.90 ± 0.70
CLAP [3]	8.04 ± 0.37	18.75 ± 0.36	27.77 ± 0.62	12.65 ± 0.46
BayesAdapter [17]	10.83 ± 0.51	22.23 ± 0.44	31.12 ± 0.74	16.61 ± 0.79
LiRA-CLIP (ours)	**17.63 ± 0.23**	**28.25 ± 0.70**	**36.04 ± 0.92**	**21.89 ± 0.75**

4.3 Per-Dataset Selective Classification

In the low-shot regime ($K \in 1, 2, 4$), LiRA-CLIP systematically enlarges the set of test points for which high-confidence, accuracy-controlled decisions can be made. For instance, at a target accuracy of 95% and 99% on Caltech101, it increases coverage on ResNet-50 from about 59%–70% (BayesAdapter) to 65%–72%

and on ViT-B/16 from 69%–77% to 78%–83%; on OxfordPets, LiRA-CLIP roughly doubles reliable coverage compared to the strongest baselines. By contrast, many prompt- and cache-based adapters either fail to satisfy the 95%–99% accuracy side constraints (which leads to ✗) or attain only marginal coverage in the extremely low-shot regime, whereas LiRA-CLIP remains calibrated and non-degenerate across datasets. As the number of shots increases ($K \geq 8$), LiRA-CLIP continues to perform strongly, but parameter-based methods such as BayesAdapter and CrossModal can match or slightly surpass its high-confidence coverage on some texture and remote-sensing tasks (DescribableTextures, EuroSAT). This pattern is consistent with our design of a training-free adapter that deliberately trades a small amount of high-shot coverage in favor of conservative, well-calibrated selective decisions in the low-data regimes where training-based uncertainty estimates are particularly fragile. To illustrate this behavior, Fig. 2 shows 99%-reliable coverage as a function of the number of shots ($K \in 1, 2, 4$) for the strongest competing baselines and LiRA-CLIP on six representative benchmarks with the ResNet-50 backbone. Across all datasets, the LiRA-CLIP curves in the low-shot regime lie on or above those of the baselines and translate a strict 99% reliability target into a substantial increase in coverage, while the method remains fully training-free. Full numerical results are given in Appendix A.4, Table A4.

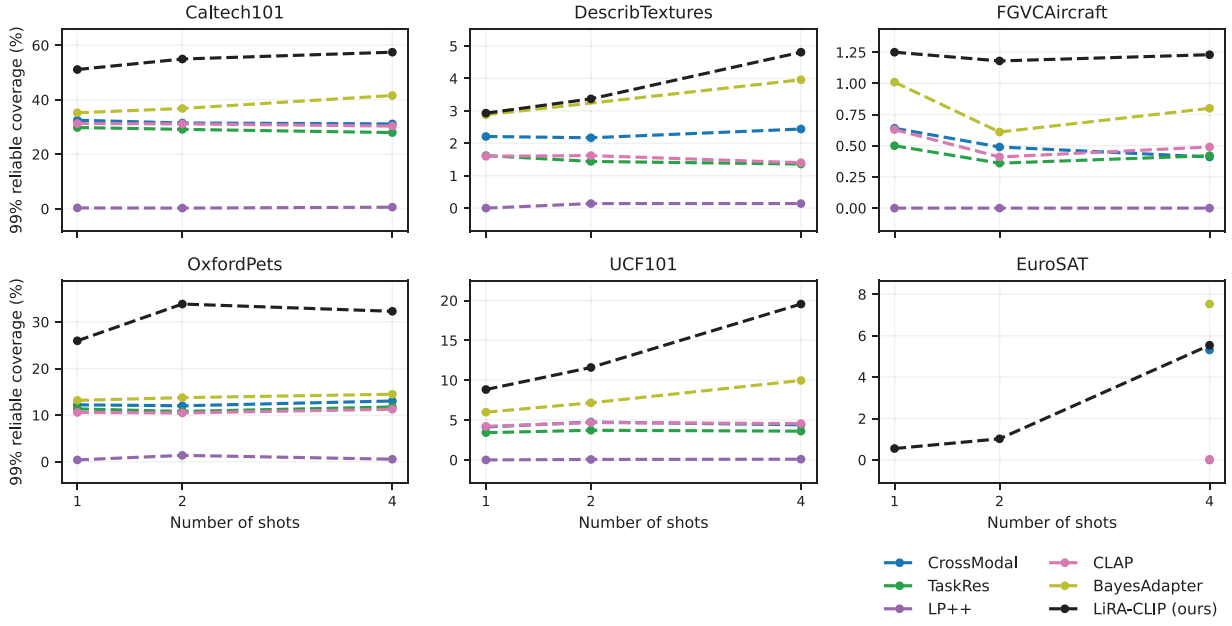

Figure 2: 99%-reliable coverage (%) on ResNet-50 across low-shot regimes ($K \in \{1, 2, 4\}$), broken down per dataset. Each curve shows how many test examples a method can classify while satisfying a 99% accuracy constraint.

4.4 Ablation Study

Our main method, LiRA-CLIP, is completely training-free. To assess whether the same posterior-predictive fusion architecture remains effective also under limited supervised fine-tuning, we additionally consider LiRA-CLIP-F, which fine-tunes exclusively the three scalar fusion parameters by gradient descent, while all CLIP and generative-model parameters remain frozen. Table 3 reports ablation results on EuroSAT. The training-free LiRA-CLIP already attains accuracy comparable to or slightly higher than recent adapters (64.9% vs. 64.7% for TaskRes and 64.6% for BayesAdapter) and at the same time reduces ECE from 4.8–10.6 to 2.9. The permitted mild fine-tuning of the fusion parameters (LiRA-CLIP-F) further increases accuracy to 69.1%, while maintaining competitive calibration, ECE 4.5%. This suggests that posterior-predictive likelihood-ratio fusion remains effective even when it is deployed as a small trainable adapter. In the extreme low-shot regime (1–2 shots), however, the labeled support set provides only very limited information to

reliably re-estimate even a small number of fusion parameters. In this setting, task-specific fine-tuning of $(\gamma_{\text{clip}}, \gamma_{\text{llr}}, \alpha)$ tends to improve accuracy, but at the price of overconfident, less stable probability estimates, whereas the fully training-free variant preserves the once-on-auxiliary-tasks calibrated, cross-task posterior-predictive structure. In line with this bias-variance intuition, our per-shot ablations (1, 2, 4 shots on EuroSAT) show that LiRA-CLIP-F indeed achieves higher accuracy, but systematically worse calibration and high-confidence coverage than LiRA-CLIP, which underscores that the fully training-free, posterior-predictive formulation is particularly advantageous for reliability under extreme data scarcity. For full numerical results see Appendix B.1, Table A5.

Table 3: Fine-tuning ablation on EuroSAT with a ResNet-50 backbone. Results are averaged over $K \in \{1, 2, 4\}$ shots and three random seeds.

Method	Accuracy	ECE	AECE
TipA [10]	49.167 ± 0.953	10.057 ± 0.957	9.983 ± 0.967
TipA-f [10]	59.203 ± 1.087	16.410 ± 1.187	16.287 ± 1.267
CrossModal [38]	64.167 ± 1.913	7.623 ± 1.410	7.617 ± 1.380
TaskRes [6]	64.663 ± 1.737	4.790 ± 1.740	4.863 ± 1.647
LP++ [18]	59.433 ± 1.670	23.550 ± 3.267	23.507 ± 3.277
CLAP [3]	66.030 ± 1.467	9.640 ± 1.213	9.573 ± 1.240
BayesAdapter [17]	64.627 ± 1.833	10.550 ± 1.590	10.530 ± 1.597
LiRA-CLIP (ours)	**64.890** ± 2.713	**2.927** ± 0.913	**3.093** ± 0.990
LiRA-CLIP-F (ours)	**69.133** ± 1.537	4.487 ± 0.890	3.423 ± 1.107

Note: Best result is **bold**; Higher is better for Accuracy; Lower is better for ECE and AECE.

Another ablation we conduct to assess the sensitivity of LiRA-CLIP to the normal-inverse-gamma prior hyperparameters $(\kappa_0, \alpha_0, \beta_0)$. Table 4 illustrates the sensitivity of LiRA-CLIP, where sweeping the NIG hyperparameters has only a minor effect on accuracy and reliability, confirming that performance is driven by the structure of the posterior-predictive model rather than fine-tuning of prior scales. We report top-1 accuracy (Accuracy %), ECE (%), AECE (%) and 99% reliable selective coverage (Sel@99%).

Table 4: Sensitivity of LiRA-CLIP to the NIG hyperparameters on Caltech101 (1-shot averaged on three random seeds.) on ResNet50.

$(\kappa_0, \alpha_0, \beta_0)$	Accuracy	ECE	AECE	Sel@99
(0.05, 0.5, 0.05)*	88.56 ± 0.04	1.63 ± 0.76	1.61 ± 0.50	51.09 ± 1.52
(0.04, 0.4, 0.04)	88.56 ± 0.04	1.93 ± 0.73	1.84 ± 0.42	50.90 ± 1.52
(0.03, 0.3, 0.03)	88.54 ± 0.06	2.36 ± 0.34	1.92 ± 0.46	51.39 ± 0.96
(0.02, 0.2, 0.02)	88.54 ± 0.02	2.35 ± 0.40	1.88 ± 0.52	51.14 ± 0.94
(0.01, 0.1, 0.01)	88.52 ± 0.00	2.26 ± 0.24	1.84 ± 0.52	50.97 ± 0.94

Note: *Indicate to the default settings used in our experiment.

4.4.1 Computational Complexity and Runtime

To calculate Computational complexity let N be the number of classes (ways), K the number of shots per class, D dimension the CLIP feature, and N_{test} the number of test images in an episode. LiRA-CLI training-free at task time, conditioned on frozen CLIP features. It computes only closed-form sufficient statistics

and t-predictive parameters and excluding CLIP feature extraction, the one-off episode setup processes NK support features in a single pass and costs $\mathcal{O}(NKD)$ time where it stores class-wise parameters with $\mathcal{O}(ND)$ memory and $\mathcal{O}(NKD)$ only when explicitly caching all support features). At inference, each test image evaluates N class-wise diagonal t-predictive scores and fuses them with CLIP logits, which are $\mathcal{O}(ND)$ time per image and fully vectorized (no backpropagation).

4.4.2 Paired Non-Parametric Significance Analysis

The main objectives of LiRA-CLIP is to improve probabilistic reliability under extreme low-shot and deployment constraints while preserving few-shot accuracy. We evaluated its statistical stability at the level of dataset × shot operating points. Specifically, we treated each dataset and shot regime as one paired observation and applied paired Wilcoxon signed-rank and sign tests, complemented by a paired permutation (sign-flip) test on the mean improvement. Following standard multi-dataset comparison practice [46], we controlled family-wise error with Holm correction. We conducted the Paired non-parametric test on ResNet50. LiRA-CLIP maintains accuracy parity with the strongest baseline adapters; however, per-setting accuracy differences are small $p_{\text{Holm}} = 0.194$. In contrast, the reliability gains are large and consistent across settings. LiRA-CLIP significantly improved calibration and selective decision-making. ECE improved in 25/36 dataset × shot settings, p_{Holm} *is* 1.9×10^{-3} and selective coverage improved in 17/18 settings for both Selective coverage sel@95% and Sel@99%, p_{Holm} *is* 5.0×10^{-4} and 5.7×10^{-4}, respectively. Further details reported in Appendix C.1, Table A6.

4.4.3 Implementation and Measured Cost

LiRA-CLIP performs no task-time optimization (training time is 0). Thus we reported in Table 5 its test-time inference cost as average latency per image (ms/img) includes adapter-only cost, which excludes CLIP image encoding and captures only the closed-form scoring and fusion stage. We also reported end-to-end latency, includes CLIP encoding plus evaluation stage. BayesAdapter reported the average adaptation time averaged over three independent seeds (in seconds). All LiRA-CLIP experiments were run on NVIDIA Tesla T4 GPU (16 GB VRAM), PyTorch 2.9.0, CUDA 12.6, and cuDNN 9.1.

Table 5: Adaptation cost vs. inference cost on Caltech101 (ResNet50). LiRA-CLIP performs no task-time optimization and therefore we reported test-time inference latency normalized per image (ms/img) including adapter-only overhead and end-to-end latency (CLIP encoding & evaluation). LiRA-CLIP reported values are averaged over three independent random seeds.

Shots	Adaptation/Training Time (s)		LiRA-CLIP Inference Latency (ms/img)	
K	BayesAdapter [17]	LiRA-CLIP*	Adapter-Only	End-to-End
1	6.47 ± 0.74	0	0.900 ± 0.078	6.865 ± 0.086
2	7.34 ± 0.72		0.896 ± 0.044	6.846 ± 0.041
4	8.04 ± 0.79		0.884 ± 0.050	6.912 ± 0.040
8	14.93 ± 1.53		0.731 ± 0.016	6.729 ± 0.071
16	26.62 ± 2.68		0.767 ± 0.032	7.305 ± 0.615
32	42.40 ± 4.39		1.206 ± 0.214	7.392 ± 0.097

Note: *LiRA-CLIP performs no task-time optimization so the training time is 0 s for all shots K.

4.4.4 LiRA-CLIP Seed-Sensitivity Analysis

We verified the sensitivity of LiRA-CLIP results to the choice of a three-seed protocol by conducting a robustness check on Caltech101 (ResNet50) using five random seeds. The five-seed estimates (mean ± SE) align closely with LiRA-CLIP original findings, showing negligible deviations with a maximum of 0.30 pp in accuracy, 0.14–0.18 in calibration (ECE/AECE), and under 0.90 pp in selective coverage. These results confirm the stability of LiRA-CLIP core trends, details in Table 6.

Table 6: LiRA-CLIP Seed-sensitivity check on Caltech101 and ResNet50.

	Accuracy and Calibration						Selective Coverage			
K	Acc_3	Acc_5	ECE_3	ECE_5	$AECE_3$	$AECE_5$	$Sel@95_3$	$Sel@95_5$	$Sel@99_3$	$Sel@99_5$
1	88.56 ± 0.04	88.86 ± 0.18	1.93 ± 0.76	1.99 ± 0.31	1.61 ± 0.50	1.79 ± 0.21	65.45 ± 1.70	66.22 ± 1.11	51.09 ± 1.52	51.42 ± 0.98
2	89.67 ± 0.30	89.76 ± 0.23	1.75 ± 0.23	1.74 ± 0.16	1.58 ± 0.35	1.60 ± 0.25	70.05 ± 0.63	70.44 ± 0.45	54.94 ± 0.66	55.29 ± 0.45
4	90.53 ± 0.08	90.55 ± 0.04	1.43 ± 0.12	1.46 ± 0.07	1.42 ± 0.13	1.42 ± 0.09	72.16 ± 0.18	72.52 ± 0.17	57.44 ± 0.15	58.33 ± 0.38
8	91.08 ± 0.16	91.15 ± 0.11	0.96 ± 0.08	1.09 ± 0.08	1.04 ± 0.04	1.13 ± 0.06	73.06 ± 0.49	73.08 ± 0.44	60.42 ± 0.65	60.71 ± 0.64
16	92.01 ± 0.04	91.89 ± 0.12	0.75 ± 0.23	0.89 ± 0.13	0.82 ± 0.16	0.92 ± 0.13	74.36 ± 0.32	73.80 ± 0.46	62.39 ± 0.38	61.87 ± 0.43
32	92.09 ± 0.12	92.22 ± 0.10	1.08 ± 0.15	0.96 ± 0.10	1.11 ± 0.31	0.93 ± 0.18	75.27 ± 0.40	75.03 ± 0.27	63.83 ± 0.72	63.45 ± 0.50

4.4.5 Sensitivity to the Accuracy Guard

The reliability-driven calibration in Eq. (19) includes an accuracy-guard constraint $A(\theta) \geq A_{\text{guard}}$ to protect discriminative performance while optimizing NLL. To assess how sensitive the resulting global fusion parameters are to this choice, we vary $A_{\text{guard}} \in \{0.005, 0.01, 0.02\}$ on a representative calibration setting (Caltech101, 1-shot, ResNet-50; three random seeds) and report the resulting top-1 accuracy, calibration errors (ECE and AECE), and reliable selective coverage at 95% and 99% target accuracy. Table 7 summarizes the results.

Table 7: Sensitivity of LiRA-CLIP to the accuracy-guard parameter A_{guard} in Eq. (19) on Caltech101 (1-shot, ResNet-50; averaged over three random seeds). We report top-1 accuracy (Acc.), calibration errors (ECE/AECE), and reliable selective coverage at 95% and 99% target accuracy (Cov@95/Cov@99).

A_{guard}	Shots	Acc. (%)	ECE (%)	AECE (%)	Cov@95 (%)	Cov@99 (%)
0.005	1	88.56 ± 0.04	1.93 ± 0.76	1.73 ± 0.50	65.45 ± 1.70	51.09 ± 1.52
0.010	1	88.56 ± 0.04	1.93 ± 0.76	1.73 ± 0.50	65.45 ± 1.70	51.09 ± 1.52
0.020	1	88.88 ± 0.37	2.32 ± 0.36	2.04 ± 0.19	65.45 ± 1.70	51.09 ± 1.52

5 Conclusion

LiRA-CLIP, a training-free text-conditioned, posterior-predictive likelihood-ratio adapter for few-shot CLIP classification. We introduced it in this article to improve probabilistic reliability in the extremely low-shot regime. LiRA-CLIP operations take place in a background-whitened CLIP feature space, and placing diagonal Normal Inverse Gamma priors over both the class-conditional distributions and a pooled background; leading to a Student-t posterior-predictive likelihood-ratio stream (t-PLLR) capturing heavy distributional outliers and data scarcity. We build a two-stream fusion mechanism to combine the generative t-PLLR scores with zero-shot CLIP logits; using a lightweight global calibration layer without gradient based task specific tuning or accessing to the CLIP weights. For a new few-shot task, LiRA-CLIP performs training-free adaptation via closed-form posterior-predictive updates and evaluation of a frozen fusion rule. Experimental results on six standard benchmarks confirm the robust adaptability and

reliability of LiRA-CLIP; notably, ablation studies validate these reliability gains stem from the proposed posterior-predictive likelihood-ratio architecture rather than from brittle hyperparameter tuning. Establishing LiRA-CLIP as a simple efficient route to training-free, reliably calibrated few-shot CLIP adaptation. LiRA-CLIP by design trades a small amount of high-shot accuracy for substantially improved low-shot reliability, making it preferable to choose when labels are scarce and complementary to fully fine-tune adapters in data-rich regimes. For future work, we will explore lightweight task-aware refinements and richer generative components to narrow the remaining high-shot gap, at the same time preserving the method's training-free.

Acknowledgement: The authors wish to express their gratitude to Prince Sultan University for their support.

Funding Statement: This research was funded by the National Nature Science of China, grant numbers U23A20321 and 62272490. Also, the authors would like to thank Prince Sultan University for paying the APC of this article.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Mustafa Qaid Khamisi, Zuping Zhang and Mohammed Al-Habib; methodology, Mustafa Qaid Khamisi; software, Mustafa Qaid Khamisi and Mohammed Al-Habib; validation, Mustafa Qaid Khamisi, Zuping Zhang and Mohammed Al-Habib; formal analysis, Mustafa Qaid Khamisi and Zuping Zhang; investigation, Mustafa Qaid Khamisi, Zuping Zhang and Mohammed Al-Habib; resources, Mustafa Qaid Khamisi and Zuping Zhang; data curation, Mustafa Qaid Khamisi and Zuping Zhang; writing—original draft preparation, Mustafa Qaid Khamisi; writing—review and editing, Mustafa Qaid Khamisi, Zuping Zhang, Mohammed Al-Habib, Muhammad Asim and Sajid Shah; visualization, Mustafa Qaid Khamisi and Mohammed Al-Habib; supervision, Zuping Zhang; project administration, Mustafa Qaid Khamisi and Zuping Zhang; funding acquisition, Muhammad Asim and Sajid Shah. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the Corresponding Author, Zuping Zhang, upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CLIP	Contrastive Language–Image Pre-Training
LiRA	Likelihood Ratio Adapter
t-PLLR	Student-t Posterior-Predictive Likelihood Ratio
NLL	Negative Log-Likelihood
ECE	Expected Calibration Error
AECE	Adaptive Expected Calibration Error
NIG	Normal Inverse Gamma

Appendix A Detailed Numerical Values for the Reported Metrics in Manuscript

Appendix A.1 Per-Dataset AECE (%, Lower Is Better) on ResNet50. Results Are Reported over Three Random Seeds

Table A1: Per-dataset AECE (%, lower is better) on ResNet50. Results are reported over three random seeds.

No.shots	Method	Caltech101	DescribTextures	FGVCAircraft	OxfordPets	UCF101	EuroSAT
1	LP [1]	30.70 ± 1.50	56.37 ± 0.76	59.82 ± 1.13	51.32 ± 1.26	48.99 ± 2.24	43.73 ± 1.33
	TipA [10]	2.10 ± 0.24	4.59 ± 0.25	4.11 ± 0.14	5.58 ± 0.18	2.19 ± 0.43	6.42 ± 0.27
	TipA-f [10]	2.55 ± 0.18	5.35 ± 0.24	3.56 ± 0.37	6.65 ± 0.50	3.34 ± 0.51	15.84 ± 0.98
	CrossModal [38]	2.31 ± 0.30	5.52 ± 0.37	9.32 ± 0.19	2.42 ± 0.53	3.31 ± 0.65	11.54 ± 2.78
	TaskRes [6]	3.37 ± 0.27	3.87 ± 0.63	4.70 ± 0.39	4.23 ± 0.76	5.72 ± 0.29	5.90 ± 2.42
	LP++ [18]	14.47 ± 0.39	17.38 ± 1.55	5.33 ± 1.50	14.08 ± 3.58	28.17 ± 1.63	17.33 ± 1.95
	CLAP [3]	2.36 ± 0.37	4.49 ± 0.74	13.03 ± 0.27	5.36 ± 0.53	2.92 ± 0.77	7.14 ± 1.02
	BayesAdapter [17]	1.62 ± 0.30	8.58 ± 0.35	15.06 ± 0.72	1.83 ± 0.45	1.94 ± 0.20	17.06 ± 1.21
	LiRA-CLIP (ours)	1.61 ± 0.50	3.68 ± 0.24	4.07 ± 0.42	1.44 ± 0.26	2.44 ± 0.06	3.77 ± 0.99
2	LP [1]	21.81 ± 1.72	47.11 ± 1.97	50.17 ± 0.96	42.25 ± 3.18	37.19 ± 0.32	34.36 ± 4.17
	TipA [10]	2.58 ± 0.17	4.07 ± 0.14	5.14 ± 0.29	4.87 ± 0.50	1.93 ± 0.23	11.60 ± 1.31
	TipA-f [10]	2.08 ± 0.11	5.98 ± 0.69	3.78 ± 0.25	6.28 ± 0.58	3.88 ± 0.16	19.21 ± 1.29
	CrossModal [38]	2.99 ± 0.32	3.20 ± 0.55	5.47 ± 0.52	2.68 ± 0.13	5.70 ± 0.38	8.91 ± 0.41
	TaskRes [6]	4.33 ± 0.22	5.62 ± 1.13	2.25 ± 0.15	4.99 ± 0.05	8.06 ± 0.42	3.64 ± 0.76
	LP++ [18]	20.26 ± 0.47	14.94 ± 6.49	4.51 ± 0.92	10.59 ± 3.11	17.91 ± 4.56	23.43 ± 4.56
	CLAP [3]	3.28 ± 0.28	2.43 ± 0.09	9.29 ± 0.79	6.10 ± 0.36	4.29 ± 0.32	7.61 ± 0.73
	BayesAdapter [17]	1.58 ± 0.10	5.11 ± 1.18	10.17 ± 0.50	1.63 ± 0.26	2.21 ± 0.03	11.29 ± 2.07
	LiRA-CLIP (ours)	1.58 ± 0.35	2.74 ± 0.69	2.53 ± 0.08	1.45 ± 0.25	2.02 ± 0.20	2.90 ± 0.75
4	LP [1]	13.47 ± 0.42	37.17 ± 0.89	40.10 ± 1.48	29.30 ± 1.97	26.82 ± 0.54	23.60 ± 1.96
	TipA [10]	1.35 ± 0.09	5.84 ± 0.31	7.12 ± 0.29	3.88 ± 0.11	3.38 ± 0.21	11.93 ± 1.32
	TipA-f [10]	1.32 ± 0.22	6.41 ± 0.16	5.47 ± 0.37	5.00 ± 0.19	1.49 ± 0.09	13.81 ± 1.53
	CrossModal [38]	3.11 ± 0.47	3.52 ± 0.47	2.67 ± 0.09	3.89 ± 0.61	6.39 ± 0.35	2.40 ± 0.95
	TaskRes [6]	4.47 ± 0.42	8.69 ± 0.04	1.87 ± 0.31	5.94 ± 0.25	8.50 ± 0.46	5.05 ± 1.76
	LP++ [18]	16.06 ± 1.02	10.29 ± 4.38	6.40 ± 0.19	17.71 ± 0.50	15.10 ± 2.32	29.76 ± 3.32
	CLAP [3]	3.55 ± 0.49	4.49 ± 0.13	4.77 ± 0.37	6.67 ± 0.16	5.16 ± 0.25	13.97 ± 1.97
	BayesAdapter [17]	1.30 ± 0.30	3.42 ± 0.30	10.35 ± 0.55	2.37 ± 0.66	1.69 ± 0.16	3.24 ± 1.51
	LiRA-CLIP (ours)	1.42 ± 0.13	2.82 ± 0.81	2.79 ± 0.65	1.13 ± 0.39	2.51 ± 0.52	2.61 ± 1.23

Appendix A.2 Detailed ECE Results of LiRA-CLIP in Comparison with Baseline Methods

Table A2: Per-dataset ECE (%, lower is better) on ResNet50. Results are reported over three random seeds.

No.shots	Method	Caltech101	DescribTextures	FGVCAircraft	OxfordPets	UCF101	EuroSAT
1	LP [1]	30.71 ± 1.50	56.38 ± 0.76	59.82 ± 1.13	51.33 ± 1.26	48.99 ± 2.24	43.73 ± 1.33
	TipA [10]	2.44 ± 0.15	4.42 ± 0.26	4.11 ± 0.14	5.71 ± 0.19	2.54 ± 0.46	6.53 ± 0.27
	TipA-f [10]	2.74 ± 0.14	5.54 ± 0.36	3.55 ± 0.43	6.66 ± 0.51	3.41 ± 0.27	15.96 ± 0.91
	CrossModal [38]	2.22 ± 0.33	5.16 ± 0.48	9.32 ± 0.19	2.37 ± 0.56	3.33 ± 0.74	11.54 ± 2.78
	TaskRes [6]	3.47 ± 0.31	3.49 ± 0.51	4.80 ± 0.40	4.44 ± 0.62	5.73 ± 0.30	5.76 ± 2.50
	LP++ [18]	14.47 ± 0.39	17.37 ± 1.55	5.23 ± 1.75	14.08 ± 3.57	28.17 ± 1.64	17.42 ± 1.89
	CLAP [3]	2.53 ± 0.49	4.46 ± 0.96	13.04 ± 0.27	5.33 ± 0.54	2.88 ± 0.86	7.19 ± 0.97
	BayesAdapter [17]	1.65 ± 0.13	8.59 ± 0.37	15.06 ± 0.72	1.86 ± 0.51	2.04 ± 0.08	17.07 ± 1.21
	LiRA-CLIP (ours)	1.63 ± 0.76	3.66 ± 0.06	4.52 ± 0.55	1.26 ± 0.18	2.44 ± 0.06	3.71 ± 0.80

(Continued)

Table A2 (continued)

No.shots	Method	Caltech101	DescribTextures	FGVCAircraft	OxfordPets	UCF101	EuroSAT
2	LP [1]	21.81 ± 1.72	47.12 ± 1.97	50.17 ± 0.96	42.26 ± 3.18	37.19 ± 0.32	34.36 ± 4.17
	TipA [10]	2.44 ± 0.23	4.25 ± 0.35	5.14 ± 0.31	4.97 ± 0.45	1.59 ± 0.26	11.71 ± 1.27
	TipA-f [10]	2.16 ± 0.24	6.09 ± 0.62	4.00 ± 0.16	6.29 ± 0.58	3.72 ± 0.19	19.33 ± 1.23
	CrossModal [38]	3.13 ± 0.30	3.04 ± 0.33	5.47 ± 0.53	2.69 ± 0.25	5.56 ± 0.28	8.95 ± 0.45
	TaskRes [6]	4.35 ± 0.21	5.65 ± 1.12	2.37 ± 0.11	5.05 ± 0.08	8.06 ± 0.42	3.71 ± 0.80
	LP++ [18]	20.26 ± 0.47	14.83 ± 6.57	4.36 ± 1.04	10.61 ± 3.11	17.91 ± 4.56	23.43 ± 4.56
	CLAP [3]	3.30 ± 0.29	2.48 ± 0.12	9.32 ± 0.80	6.21 ± 0.28	4.32 ± 0.29	7.75 ± 0.70
	BayesAdapter [17]	1.81 ± 0.08	5.39 ± 1.16	10.17 ± 0.50	1.74 ± 0.16	2.20 ± 0.06	11.32 ± 2.05
	LiRA-CLIP (ours)	1.75 ± 0.23	2.37 ± 0.88	2.87 ± 0.38	1.27 ± 0.28	2.02 ± 0.20	2.40 ± 0.65
4	LP [1]	13.48 ± 0.41	37.19 ± 0.89	40.10 ± 1.48	29.31 ± 1.97	26.83 ± 0.54	23.60 ± 1.96
	TipA [10]	1.71 ± 0.12	5.88 ± 0.09	7.12 ± 0.29	3.88 ± 0.07	3.35 ± 0.21	11.93 ± 1.33
	TipA-f [10]	1.31 ± 0.20	6.71 ± 0.14	5.39 ± 0.31	5.02 ± 0.19	1.72 ± 0.11	13.94 ± 1.42
	CrossModal [38]	3.15 ± 0.46	3.44 ± 0.46	2.54 ± 0.10	3.98 ± 0.56	6.42 ± 0.37	2.38 ± 1.00
	TaskRes [6]	4.51 ± 0.39	8.70 ± 0.04	1.39 ± 0.09	5.96 ± 0.28	8.54 ± 0.48	4.90 ± 1.92
	LP++ [18]	16.07 ± 1.01	10.39 ± 4.33	6.40 ± 0.19	17.76 ± 0.47	15.11 ± 2.32	29.80 ± 3.35
	CLAP [3]	3.60 ± 0.44	4.47 ± 0.28	4.88 ± 0.35	6.68 ± 0.17	5.17 ± 0.24	13.98 ± 1.97
	BayesAdapter [17]	1.80 ± 0.22	3.19 ± 0.05	10.38 ± 0.53	2.37 ± 0.62	1.92 ± 0.33	3.26 ± 1.51
	LiRA-CLIP (ours)	1.43 ± 0.12	3.40 ± 0.49	3.32 ± 0.31	1.25 ± 0.50	2.51 ± 0.52	2.67 ± 1.29
8	LP [1]	7.45 ± 0.57	29.83 ± 0.99	25.02 ± 0.64	21.71 ± 1.42	16.72 ± 1.25	16.84 ± 0.92
	TipA [10]	2.54 ± 0.23	7.40 ± 0.17	12.01 ± 0.43	2.28 ± 0.35	6.85 ± 0.37	12.42 ± 0.66
	TipA-f [10]	1.17 ± 0.23	4.98 ± 0.36	9.97 ± 0.95	3.73 ± 0.23	2.69 ± 0.37	4.90 ± 2.00
	CrossModal [38]	2.23 ± 0.13	6.20 ± 0.57	3.73 ± 0.30	5.90 ± 0.25	4.63 ± 0.42	3.89 ± 1.18
	TaskRes [6]	3.34 ± 0.12	11.52 ± 0.58	1.67 ± 0.32	7.18 ± 0.27	7.43 ± 0.60	8.28 ± 1.58
	LP++ [18]	9.23 ± 0.92	11.47 ± 4.88	3.76 ± 0.88	15.75 ± 0.26	10.83 ± 2.65	5.26 ± 0.47
	CLAP [3]	3.80 ± 0.08	6.50 ± 0.48	4.60 ± 0.39	7.44 ± 0.23	6.82 ± 0.31	16.14 ± 1.81
	BayesAdapter [17]	1.12 ± 0.20	3.60 ± 0.44	8.53 ± 0.82	2.93 ± 0.09	1.92 ± 0.22	1.96 ± 0.60
	LiRA-CLIP (ours)	0.96 ± 0.08	2.73 ± 0.30	1.81 ± 0.05	1.44 ± 0.30	1.66 ± 0.62	2.22 ± 0.76
16	LP [1]	4.30 ± 0.65	20.30 ± 1.41	12.62 ± 1.08	9.81 ± 1.42	9.40 ± 0.52	14.05 ± 2.58
	TipA [10]	3.76 ± 0.11	14.16 ± 0.61	19.13 ± 0.52	2.00 ± 0.43	12.47 ± 0.29	4.97 ± 0.50
	TipA-f [10]	1.50 ± 0.12	3.38 ± 0.63	16.12 ± 0.51	0.98 ± 0.34	7.21 ± 0.18	1.85 ± 0.53
	CrossModal [38]	2.00 ± 0.23	5.50 ± 0.52	2.54 ± 0.11	5.05 ± 0.12	3.23 ± 0.30	6.94 ± 1.17
	TaskRes [6]	2.72 ± 0.07	10.07 ± 0.42	2.23 ± 0.21	6.67 ± 0.11	6.09 ± 0.23	11.64 ± 1.13
	LP++ [18]	5.42 ± 0.87	7.10 ± 2.69	3.31 ± 0.43	13.07 ± 0.72	7.22 ± 0.34	2.34 ± 0.58
	CLAP [3]	4.01 ± 0.19	9.38 ± 0.51	2.47 ± 0.16	7.80 ± 0.08	8.04 ± 0.30	19.40 ± 1.09
	BayesAdapter [17]	1.38 ± 0.20	2.11 ± 0.30	5.56 ± 0.25	2.42 ± 0.20	1.48 ± 0.05	4.68 ± 0.48
	LiRA-CLIP (ours)	0.75 ± 0.23	3.06 ± 0.32	2.17 ± 0.41	1.67 ± 0.14	1.92 ± 0.13	5.99 ± 0.84
32	LP [1]	2.14 ± 0.38	14.78 ± 1.58	3.87 ± 0.95	4.06 ± 0.82	3.58 ± 1.49	11.48 ± 1.41
	TipA [10]	15.27 ± 0.11	23.77 ± 0.27	29.39 ± 0.31	6.93 ± 0.51	18.32 ± 0.50	7.05 ± 1.34
	TipA-f [10]	3.05 ± 0.06	8.37 ± 0.17	22.28 ± 0.11	3.42 ± 0.22	11.44 ± 0.47	6.07 ± 0.26
	CrossModal [38]	1.61 ± 0.14	2.76 ± 0.05	1.91 ± 0.39	4.55 ± 0.24	2.74 ± 0.26	7.69 ± 0.10
	TaskRes [6]	2.48 ± 0.08	5.37 ± 0.20	3.03 ± 0.21	5.95 ± 0.12	5.47 ± 0.40	12.33 ± 0.18
	LP++ [18]	2.67 ± 0.21	4.07 ± 1.76	5.23 ± 0.38	10.24 ± 1.33	6.08 ± 0.45	2.89 ± 0.34
	CLAP [3]	3.68 ± 0.06	9.72 ± 0.17	3.34 ± 0.08	8.20 ± 0.06	9.48 ± 0.17	19.68 ± 0.38
	BayesAdapter [17]	0.98 ± 0.05	2.78 ± 0.30	2.54 ± 0.34	3.20 ± 0.18	1.42 ± 0.42	5.20 ± 0.28
	LiRA-CLIP (ours)	1.08 ± 0.15	2.21 ± 0.27	3.97 ± 0.15	1.85 ± 0.06	1.90 ± 0.17	7.41 ± 0.39

Appendix A.3 Detailed Accuracy Results of LiRA-CLIP in Comparison with Baseline Methods

Table A3: Accuracy (%) on ResNet50. Results are reported over three random seeds.

No.shots	Method	Caltech101	DescribTextures	FGVCAircraft	OxfordPets	UCF101	EuroSAT
1	LP [1]	57.96 ± 2.12	28.37 ± 0.62	12.06 ± 0.17	28.19 ± 1.69	33.53 ± 2.04	42.93 ± 3.93
	TipA [10]	84.95 ± 0.09	44.27 ± 0.21	17.76 ± 0.12	83.57 ± 0.15	59.58 ± 0.28	42.77 ± 0.65
	TipA-f [10]	86.13 ± 0.42	45.29 ± 0.28	18.61 ± 0.34	85.05 ± 0.58	61.34 ± 0.30	52.00 ± 0.52
	CrossModal [38]	88.07 ± 0.37	47.85 ± 1.06	20.44 ± 0.32	81.43 ± 1.16	64.45 ± 0.77	57.88 ± 3.03
	TaskRes [6]	88.52 ± 0.29	49.00 ± 1.16	20.46 ± 0.12	83.16 ± 1.10	65.02 ± 0.43	58.98 ± 2.72
	LP++ [18]	88.21 ± 0.48	46.43 ± 0.67	20.13 ± 0.21	83.51 ± 2.04	63.94 ± 0.57	50.50 ± 1.11
	CLAP [3]	88.26 ± 0.28	47.60 ± 0.45	20.46 ± 0.29	84.41 ± 0.72	64.09 ± 1.20	60.77 ± 2.17
	BayesAdapter [17]	87.94 ± 0.08	46.61 ± 0.70	19.85 ± 0.59	80.97 ± 0.67	63.05 ± 0.59	57.47 ± 1.29
	LiRA-CLIP (ours)	88.56 ± 0.04	47.22 ± 0.35	20.27 ± 0.59	85.60 ± 0.70	64.42 ± 0.26	57.53 ± 4.66
2	LP [1]	68.05 ± 2.29	36.43 ± 1.09	17.04 ± 0.57	38.28 ± 2.23	44.40 ± 0.22	50.58 ± 5.48
	TipA [10]	86.22 ± 0.31	45.15 ± 0.22	18.85 ± 0.44	83.44 ± 0.51	61.18 ± 0.23	50.16 ± 1.02
	TipA-f [10]	87.61 ± 0.20	48.66 ± 0.51	20.77 ± 0.11	85.48 ± 0.51	65.05 ± 0.14	60.07 ± 0.88
	CrossModal [38]	89.53 ± 0.25	53.07 ± 1.05	22.46 ± 0.18	81.74 ± 0.48	68.13 ± 0.50	62.91 ± 1.00
	TaskRes [6]	89.74 ± 0.15	53.23 ± 0.92	22.16 ± 0.20	83.71 ± 0.35	68.20 ± 0.37	63.38 ± 0.63
	LP++ [18]	89.02 ± 0.25	51.20 ± 1.91	22.06 ± 0.43	82.95 ± 0.99	67.95 ± 0.24	59.65 ± 2.61
	CLAP [3]	89.68 ± 0.21	53.17 ± 0.39	22.82 ± 0.29	84.74 ± 0.61	67.93 ± 0.47	65.06 ± 0.63
	BayesAdapter [17]	89.06 ± 0.05	51.87 ± 1.45	22.68 ± 0.12	80.17 ± 0.75	67.35 ± 0.29	63.83 ± 2.32
	LiRA-CLIP (ours)	89.67 ± 0.30	52.22 ± 0.95	22.93 ± 0.09	86.78 ± 0.25	69.03 ± 0.38	66.35 ± 1.44
4	LP [1]	77.54 ± 1.34	47.01 ± 0.76	21.14 ± 0.82	49.88 ± 1.66	55.19 ± 0.48	61.28 ± 2.07
	TipA [10]	87.68 ± 0.17	48.31 ± 0.57	20.26 ± 0.19	83.70 ± 0.16	62.78 ± 0.17	54.57 ± 1.19
	TipA-f [10]	89.41 ± 0.20	54.61 ± 0.17	22.84 ± 0.28	86.26 ± 0.27	66.59 ± 0.67	65.54 ± 1.86
	CrossModal [38]	90.37 ± 0.18	58.37 ± 0.14	24.45 ± 0.27	84.56 ± 0.68	69.71 ± 0.39	71.71 ± 1.71
	TaskRes [6]	90.51 ± 0.17	58.02 ± 0.19	23.09 ± 0.19	85.89 ± 0.29	68.64 ± 0.49	71.63 ± 1.86
	LP++ [18]	90.80 ± 0.19	57.51 ± 0.50	24.03 ± 0.67	87.47 ± 0.15	70.40 ± 0.49	68.15 ± 1.29
	CLAP [3]	90.78 ± 0.20	58.87 ± 0.34	25.08 ± 0.40	86.48 ± 0.23	69.70 ± 0.38	72.26 ± 1.60
	BayesAdapter [17]	90.01 ± 0.34	57.58 ± 0.23	25.57 ± 0.44	83.27 ± 0.86	68.62 ± 0.25	72.58 ± 1.89
	LiRA-CLIP (ours)	90.53 ± 0.08	58.67 ± 0.41	26.03 ± 0.23	86.94 ± 0.08	72.34 ± 0.65	70.79 ± 2.04
8	LP [1]	83.46 ± 1.74	54.06 ± 0.55	27.78 ± 0.45	58.86 ± 0.55	64.27 ± 0.83	68.16 ± 3.08
	TipA [10]	88.51 ± 0.21	52.32 ± 0.17	20.91 ± 0.12	84.12 ± 0.32	64.95 ± 0.25	63.34 ± 0.57
	TipA-f [10]	90.68 ± 0.30	61.15 ± 0.40	26.38 ± 0.79	87.07 ± 0.24	70.19 ± 0.47	71.64 ± 2.36
	CrossModal [38]	91.79 ± 0.06	62.61 ± 0.58	28.29 ± 0.53	86.75 ± 0.34	73.55 ± 0.39	77.74 ± 1.91
	TaskRes [6]	91.59 ± 0.09	61.35 ± 0.33	26.53 ± 0.57	87.23 ± 0.27	72.87 ± 0.51	76.84 ± 2.24
	LP++ [18]	91.62 ± 0.19	62.63 ± 0.09	26.24 ± 0.66	88.06 ± 0.20	73.69 ± 0.20	72.98 ± 2.69
	CLAP [3]	91.49 ± 0.07	62.88 ± 0.30	29.25 ± 0.53	87.48 ± 0.16	73.61 ± 0.19	76.39 ± 2.14
	BayesAdapter [17]	91.68 ± 0.15	62.67 ± 0.67	29.40 ± 0.71	84.56 ± 0.24	73.26 ± 0.31	78.34 ± 1.83
	LiRA-CLIP (ours)	91.08 ± 0.16	63.24 ± 0.36	29.10 ± 0.75	87.95 ± 0.11	75.32 ± 0.01	77.03 ± 1.03
16	LP [1]	87.34 ± 1.71	60.86 ± 0.42	34.66 ± 0.63	71.75 ± 0.98	70.96 ± 0.53	72.42 ± 0.31
	TipA [10]	89.36 ± 0.09	54.91 ± 0.34	23.03 ± 0.22	83.55 ± 0.45	66.35 ± 0.21	69.51 ± 0.54
	TipA-f [10]	91.94 ± 0.24	65.58 ± 0.14	31.32 ± 0.56	86.42 ± 0.14	72.30 ± 0.10	77.21 ± 1.61
	CrossModal [38]	92.75 ± 0.04	66.90 ± 0.37	33.32 ± 0.16	87.47 ± 0.08	76.39 ± 0.22	82.05 ± 0.70
	TaskRes [6]	92.56 ± 0.16	65.64 ± 0.40	31.37 ± 0.27	88.18 ± 0.12	75.73 ± 0.12	81.15 ± 0.51
	LP++ [18]	92.08 ± 0.18	66.65 ± 0.51	32.19 ± 0.43	88.81 ± 0.12	76.39 ± 0.32	79.84 ± 0.51
	CLAP [3]	92.14 ± 0.04	66.69 ± 0.48	34.20 ± 0.25	88.38 ± 0.07	76.12 ± 0.15	80.35 ± 0.48
	BayesAdapter [17]	92.87 ± 0.03	66.57 ± 0.12	34.88 ± 0.28	85.95 ± 0.19	76.84 ± 0.40	83.28 ± 0.16
	LiRA-CLIP (ours)	92.01 ± 0.04	66.67 ± 0.35	33.47 ± 0.11	89.00 ± 0.26	77.29 ± 0.40	80.40 ± 0.45

(Continued)

Table A3 (continued)

No.shots	Method	Caltech101	DescribTextures	FGVCAircraft	OxfordPets	UCF101	EuroSAT
32	LP [1]	89.83 ± 1.37	64.95 ± 1.15	40.22 ± 0.57	78.05 ± 0.58	76.39 ± 0.47	75.00 ± 1.70
	TipA [10]	77.90 ± 0.11	56.50 ± 0.12	24.93 ± 0.11	81.20 ± 0.58	67.15 ± 0.39	70.83 ± 1.61
	TipA-f [10]	91.32 ± 0.12	68.07 ± 0.27	36.23 ± 0.16	86.61 ± 0.25	73.92 ± 0.36	77.81 ± 0.90
	CrossModal [38]	93.71 ± 0.15	69.66 ± 0.48	38.34 ± 0.28	88.34 ± 0.33	79.32 ± 0.31	83.54 ± 0.25
	TaskRes [6]	93.64 ± 0.08	68.89 ± 0.48	35.91 ± 0.11	88.71 ± 0.14	79.08 ± 0.39	82.13 ± 0.33
	LP++ [18]	92.85 ± 0.26	68.91 ± 0.80	34.98 ± 0.27	88.95 ± 0.19	78.48 ± 0.49	83.77 ± 0.38
	CLAP [3]	92.17 ± 0.12	68.44 ± 0.18	36.97 ± 0.23	88.90 ± 0.08	78.26 ± 0.15	81.56 ± 0.41
	BayesAdapter [17]	93.54 ± 0.03	69.58 ± 0.34	40.66 ± 0.24	88.08 ± 0.26	79.53 ± 0.39	85.55 ± 0.31
	LiRA-CLIP (ours)	92.09 ± 0.12	67.34 ± 0.49	36.11 ± 0.14	89.00 ± 0.01	79.20 ± 0.03	81.70 ± 0.18

Appendix A.4 99%-Reliable Prediction of LiRA-CLIP in Comparison with Baseline Methods

Table A4: Selective classification at 99% confidence using a ResNet-50 backbone.

No.shots	Method	Caltech101	DescribTextures	FGVCAircraft	OxfordPets	UCF101	EuroSAT
1	LP [1]	✗	✗	✗	✗	✗	✗
	TipA [10]	24.71 ± 0.18	0.85 ± 0.05	0.23 ± 0.04	6.05 ± 0.12	2.58 ± 0.09	0.00 ± 0.00
	TipA-f [10]	25.46 ± 0.13	0.73 ± 0.05	0.29 ± 0.09	6.82 ± 0.12	2.74 ± 0.11	0.00 ± 0.00
	CrossModal [38]	32.49 ± 0.26	2.21 ± 0.15	0.64 ± 0.15	12.22 ± 0.11	4.13 ± 0.29	✗
	TaskRes [6]	29.80 ± 0.43	1.62 ± 0.14	0.50 ± 0.10	11.35 ± 0.09	3.42 ± 0.23	✗
	LP++ [18]	0.31 ± 0.14	0.00 ± 0.00	0.00 ± 0.00	0.40 ± 0.28	0.00 ± 0.00	✗
	CLAP [3]	31.28 ± 0.26	1.60 ± 0.17	0.63 ± 0.11	10.58 ± 0.24	4.21 ± 0.38	✗
	BayesAdapter [17]	35.21 ± 0.56	2.88 ± 0.36	1.01 ± 0.15	13.17 ± 0.75	5.99 ± 0.35	✗
	LiRA-CLIP (ours)	51.09 ± 1.52	2.93 ± 0.12	1.25 ± 0.14	25.95 ± 0.05	8.83 ± 0.05	0.55 ± 0.17
2	LP [1]	✗	✗	✗	✗	✗	✗
	TipA [10]	32.85 ± 0.48	1.50 ± 0.15	0.44 ± 0.02	7.08 ± 0.27	4.07 ± 0.00	0.00 ± 0.00
	TipA-f [10]	33.66 ± 0.33	1.26 ± 0.08	0.46 ± 0.04	7.56 ± 0.51	4.71 ± 0.23	0.00 ± 0.00
	CrossModal [38]	31.48 ± 2.66	2.17 ± 0.24	0.49 ± 0.10	12.01 ± 0.65	4.75 ± 0.28	✗
	TaskRes [6]	29.13 ± 2.74	1.44 ± 0.25	0.36 ± 0.06	10.80 ± 0.77	3.72 ± 0.29	✗
	LP++ [18]	0.24 ± 0.06	0.14 ± 0.14	0.00 ± 0.00	1.38 ± 0.54	0.06 ± 0.05	✗
	CLAP [3]	31.14 ± 2.25	1.62 ± 0.25	0.41 ± 0.06	10.47 ± 0.58	4.69 ± 0.20	✗
	BayesAdapter [17]	36.80 ± 0.95	✗	0.61 ± 0.11	13.76 ± 0.72	7.16 ± 0.15	✗
	LiRA-CLIP (ours)	54.94 ± 0.66	3.37 ± 0.41	1.18 ± 0.03	33.84 ± 0.35	11.60 ± 0.00	1.02 ± 0.09
4	LP [1]	✗	✗	✗	✗	✗	✗
	TipA [10]	43.34 ± 0.23	✗	0.79 ± 0.02	9.45 ± 0.20	10.53 ± 0.21	0.00 ± 0.00
	TipA-f [10]	44.16 ± 0.30	2.36 ± 0.03	0.84 ± 0.05	10.20 ± 0.16	11.01 ± 0.28	0.02 ± 0.01
	CrossModal [38]	31.13 ± 2.35	2.44 ± 0.04	0.41 ± 0.08	13.04 ± 0.43	4.41 ± 0.18	5.31 ± 1.06
	TaskRes [6]	27.96 ± 1.55	1.36 ± 0.15	0.42 ± 0.05	11.76 ± 0.39	3.61 ± 0.11	✗
	LP++ [18]	0.58 ± 0.15	0.14 ± 0.14	0.00 ± 0.00	0.55 ± 0.36	0.09 ± 0.02	0.00 ± 0.00
	CLAP [3]	30.33 ± 1.43	1.40 ± 0.19	0.49 ± 0.06	11.27 ± 0.25	4.56 ± 0.18	0.03 ± 0.02
	BayesAdapter [17]	41.56 ± 2.51	3.96 ± 0.09	0.80 ± 0.13	14.49 ± 0.43	9.97 ± 0.07	7.53 ± 1.76
	LiRA-CLIP (ours)	57.44 ± 0.15	4.81 ± 0.13	1.23 ± 0.13	32.28 ± 0.06	19.56 ± 0.00	5.54 ± 0.12
8	LP [1]	✗	✗	✗	✗	✗	✗
	TipA [10]	52.44 ± 0.17	✗	✗	13.09 ± 0.07	✗	0.00 ± 0.00
	TipA-f [10]	52.82 ± 0.21	✗	✗	13.85 ± 0.39	19.68 ± 0.11	0.00 ± 0.00
	CrossModal [38]	42.56 ± 1.01	2.70 ± 0.28	0.96 ± 0.05	12.67 ± 0.27	10.06 ± 0.38	4.57 ± 0.84
	TaskRes [6]	34.25 ± 1.01	1.73 ± 0.14	0.61 ± 0.05	11.44 ± 0.23	6.21 ± 0.02	0.51 ± 0.16
	LP++ [18]	2.80 ± 0.78	0.75 ± 0.75	0.00 ± 0.00	0.64 ± 0.12	1.22 ± 0.60	1.84 ± 1.04
	CLAP [3]	31.49 ± 0.48	1.71 ± 0.12	0.58 ± 0.07	11.09 ± 0.21	5.75 ± 0.09	0.02 ± 0.01
	BayesAdapter [17]	51.12 ± 0.81	7.05 ± 0.63	1.46 ± 0.03	15.90 ± 0.36	15.22 ± 0.37	9.33 ± 1.52
	LiRA-CLIP (ours)	60.42 ± 0.65	✗	1.36 ± 0.04	37.41 ± 0.80	22.01 ± 0.67	6.13 ± 0.09
16	LP [1]	63.83 ± 2.24	✗	2.72 ± 0.16	✗	✗	✗
	TipA [10]	65.07 ± 0.16	✗	✗	20.54 ± 0.21	✗	0.11 ± 0.04
	TipA-f [10]	64.46 ± 0.24	✗	2.96 ± 0.11	22.34 ± 0.71	✗	8.74 ± 0.64
	CrossModal [38]	52.33 ± 0.86	4.10 ± 0.21	1.10 ± 0.10	14.40 ± 0.12	15.15 ± 0.36	5.23 ± 1.02
	TaskRes [6]	44.83 ± 0.86	2.29 ± 0.22	0.75 ± 0.10	12.66 ± 0.39	9.89 ± 0.37	0.56 ± 0.16
	LP++ [18]	18.82 ± 3.29	0.93 ± 0.46	0.05 ± 0.01	2.23 ± 0.17	4.60 ± 0.32	4.70 ± 1.64

(Continued)

Table A4 (continued)

No.shots	Method	Caltech101	DescribTextures	FGVCAircraft	OxfordPets	UCF101	EuroSAT
	CLAP [3]	32.20 ± 0.44	1.77 ± 0.12	0.59 ± 0.05	11.03 ± 0.17	5.80 ± 0.08	0.02 ± 0.01
	BayesAdapter [17]	57.59 ± 0.88	✗	1.77 ± 0.18	17.62 ± 0.53	19.16 ± 0.32	9.53 ± 1.14
	LiRA-CLIP (ours)	62.39 ± 0.38	6.46 ± 0.66	1.83 ± 0.15	41.70 ± 0.70	✗	7.08 ± 0.03
32	LP [1]	64.29 ± 2.87	✗	1.92 ± 0.16	✗	✗	✗
	TipA [10]	✗	✗	✗	✗	✗	10.44 ± 0.30
	TipA-f [10]	70.85 ± 0.23	✗	✗	35.25 ± 0.30	✗	✗
	CrossModal [38]	58.39 ± 0.63	7.64 ± 0.26	1.63 ± 0.05	16.31 ± 0.14	20.68 ± 0.30	7.02 ± 0.32
	TaskRes [6]	53.50 ± 0.71	4.57 ± 0.29	1.16 ± 0.07	14.00 ± 0.39	16.05 ± 0.32	0.49 ± 0.06
	LP++ [18]	47.88 ± 0.63	3.61 ± 0.71	0.23 ± 0.06	4.51 ± 0.83	12.58 ± 0.34	7.70 ± 0.87
	CLAP [3]	34.13 ± 0.07	1.64 ± 0.14	0.47 ± 0.04	10.97 ± 0.20	5.98 ± 0.09	0.02 ± 0.02
	BayesAdapter [17]	61.93 ± 0.22	10.95 ± 0.20	2.24 ± 0.07	19.84 ± 0.12	24.82 ± 0.23	12.75 ± 0.21
	LiRA-CLIP (ours)	63.83 ± 0.72	5.67 ± 0.00	3.03 ± 0.03	43.05 ± 0.60	✗	9.14 ± 0.05

Appendix B Detailed Numerical Values for Ablation Study

Appendix B.1 Ablation Results for LiRA-CLIP-F in Comparison with Baseline Methods

Table A5: Results on EuroSAT using ResNet50 averaged over three random seeds where Sel@99% denotes 99%-reliable coverage (%); a value of ✗ indicates that the 99% target accuracy could not be satisfied.

No.shots	Method	Accuracy	ECE	AECE	Sel@99 %
1	TipA-f [10]	52.00 ± 0.52	15.96 ± 0.91	15.84 ± 0.98	0.00 ± 0.00
	CrossModal [38]	57.88 ± 3.03	11.54 ± 2.78	11.54 ± 2.78	✗
	TaskRes [6]	58.98 ± 2.72	5.76 ± 2.50	5.90 ± 2.42	✗
	LP++ [18]	50.50 ± 1.11	17.42 ± 1.89	17.33 ± 1.95	✗
	CLAP [3]	60.77 ± 2.17	7.19 ± 0.97	7.14 ± 1.02	✗
	BayesAdapter [17]	57.47 ± 1.29	17.07 ± 1.21	17.06 ± 1.21	✗
	LiRA-CLIP (ours)	57.53 ± 4.66	3.71 ± 0.80	3.77 ± 0.99	0.55 ± 0.17
	LiRA-CLIP-F (ours)	61.69 ± 0.79	5.87 ± 1.58	5.71 ± 1.90	✗
2	TipA-f [10]	60.07 ± 0.88	19.33 ± 1.23	19.21 ± 1.29	0.00 ± 0.00
	CrossModal [38]	62.91 ± 1.00	8.95 ± 0.45	8.91 ± 0.41	✗
	TaskRes [6]	63.38 ± 0.63	3.71 ± 0.80	3.64 ± 0.76	✗
	LP++ [18]	59.65 ± 2.61	23.43 ± 4.56	23.43 ± 4.56	✗
	CLAP [3]	65.06 ± 0.63	7.75 ± 0.70	7.61 ± 0.73	✗
	BayesAdapter [17]	63.83 ± 2.32	11.32 ± 2.05	11.29 ± 2.07	✗
	LiRA-CLIP (ours)	66.35 ± 1.44	2.40 ± 0.65	2.90 ± 0.75	1.02 ± 0.09
	LiRA-CLIP-F (ours)	68.75 ± 1.45	4.25 ± 0.93	4.38 ± 1.07	0.02 ± 0.00
4	TipA-f [10]	65.54 ± 1.86	13.94 ± 1.42	13.81 ± 1.53	0.02 ± 0.01
	CrossModal [38]	71.71 ± 1.71	2.38 ± 1.00	2.40 ± 0.95	5.31 ± 1.06
	TaskRes [6]	71.63 ± 1.86	4.90 ± 1.92	5.05 ± 1.76	✗
	LP++ [18]	68.15 ± 1.29	29.80 ± 3.35	29.76 ± 3.32	0.00 ± 0.00
	CLAP [3]	72.26 ± 1.60	13.98 ± 1.97	13.97 ± 1.97	0.03 ± 0.02
	BayesAdapter [17]	72.58 ± 1.89	3.26 ± 1.51	3.24 ± 1.51	7.53 ± 1.76
	LiRA-CLIP (ours)	70.79 ± 2.04	2.67 ± 1.29	2.61 ± 1.23	5.54 ± 0.12

(Continued)

Table A5 (continued)

No.shots	Method	Accuracy	ECE	AECE	Sel@99 %
	LiRA-CLIP-F (ours)	76.96 ± 2.37	3.34 ± 0.16	3.18 ± 0.35	2.54 ± 2.37
8	TipA-f [10]	71.64 ± 2.36	4.90 ± 2.00	5.04 ± 1.86	0.00 ± 0.00
	CrossModal [38]	77.74 ± 1.91	3.89 ± 1.18	3.88 ± 1.19	4.57 ± 0.84
	TaskRes [6]	76.84 ± 2.24	8.28 ± 1.58	8.27 ± 1.58	0.51 ± 0.16
	LP++ [18]	72.98 ± 2.69	5.26 ± 0.47	5.26 ± 0.46	1.84 ± 1.04
	CLAP [3]	76.39 ± 2.14	16.14 ± 1.81	16.14 ± 1.82	0.02 ± 0.01
	BayesAdapter [17]	78.34 ± 1.83	1.96 ± 0.60	1.73 ± 0.51	9.33 ± 1.52
	LiRA-CLIP (ours)	77.03 ± 1.03	2.22 ± 0.76	4.19 ± 0.67	6.13 ± 0.09
	LiRA-CLIP-F (ours)	80.94 ± 0.07	1.66 ± 0.27	1.68 ± 0.23	7.87 ± 0.19
16	CrossModal [38]	82.05 ± 0.70	6.94 ± 1.17	6.89 ± 1.22	5.23 ± 1.02
	TaskRes [6]	81.15 ± 0.51	11.64 ± 1.13	11.64 ± 1.13	0.56 ± 0.16
	LP++ [18]	79.84 ± 0.51	2.34 ± 0.58	2.29 ± 0.59	4.70 ± 1.64
	CLAP [3]	80.35 ± 0.48	19.40 ± 1.09	19.40 ± 1.09	0.02 ± 0.01
	BayesAdapter [17]	83.28 ± 0.16	4.68 ± 0.48	4.68 ± 0.42	9.53 ± 1.14
	LiRA-CLIP (ours)	80.40 ± 0.45	5.99 ± 0.84	5.99 ± 0.84	6.08 ± 0.03
	LiRA-CLIP-F (ours)	84.56 ± 0.02	3.85 ± 0.59	3.82 ± 0.56	6.77 ± 1.04
32	CrossModal [38]	83.54 ± 0.25	7.69 ± 0.10	7.68 ± 0.10	7.02 ± 0.32
	TaskRes [6]	82.13 ± 0.33	12.33 ± 0.18	12.32 ± 0.18	0.49 ± 0.06
	LP++ [18]	83.77 ± 0.38	2.89 ± 0.34	2.76 ± 0.36	7.70 ± 0.87
	CLAP [3]	81.56 ± 0.41	19.68 ± 0.38	19.68 ± 0.38	0.02 ± 0.02
	BayesAdapter [17]	85.55 ± 0.31	5.20 ± 0.28	5.25 ± 0.21	12.75 ± 0.21
	LiRA-CLIP (ours)	81.70 ± 0.18	7.41 ± 0.39	7.41 ± 0.39	7.94 ± 0.05
	LiRA-CLIP-F (ours)	88.14 ± 0.90	6.11 ± 0.11	6.08 ± 0.09	8.23 ± 0.58

Appendix C Statistical Significance and Robustness Analyses

Appendix C.1 Paired Non-Parametric Significance Test

Table A6: Paired non-parametric significance tests with BayesAdapter [17] on RN50. We computed the test over paired dataset × shot settings include Accuracy and ECE tested over 6 datasets×6 shots where the Selective coverage tested over 6 datasets×3 shots with $K \in \{1, 2, 4\}$. The improvements defined as $\Delta\text{Acc} = \text{Acc}_{\text{LiRA}} - \text{Acc}_{\text{BayesAdapter}}$, $\Delta\text{ECE} = \text{ECE}_{\text{BayesAdapter}} - \text{ECE}_{\text{LiRA}}$, and $\Delta\text{Sel} = \text{Sel}_{\text{LiRA}} - \text{Sel}_{\text{BayesAdapter}}$ (positive is better) where N is sample size for the paired comparison. Reported Wins(W), Losses(L) and Ties(T) across settings, bootstrap 95% CIs for the mean improvement. And then paired Wilcoxon, sign, and permutation (sign-flip) p-values. Holm-adjusted permutation p-values control family-wise error across all four metrics.

Metric	Alt.	N	W/L/T	Mean Δ (95% CI)	Median Δ	p_{perm}	$p_{\text{perm,Holm}}$
Accuracy	Two-sided	36	24/12/0	0.5044 [−0.2297, 1.2495]	0.4550	0.1939	0.1939
ECE	Greater	36	25/11/0	1.8431 [0.7522, 3.0664]	0.5200	0.00095	0.0019
Sel@95	Greater	18	17/1/0	6.0256 [3.2983, 9.0006]	5.6100	0.000125	0.00050
Sel@99	Greater	18	17/1/0	6.8061 [3.4300, 10.3662]	3.1050	0.00019	0.00057

Note: Alt is the alternative hypothesis used in the paired tests, where two-sided tests whether the median improvement differs from zero, while greater tests whether the median improvement is strictly positive.

References

1. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. London, UK: PMLR; 2021. p. 8748–63.
2. Jia C, Yang Y, Xia Y, Chen YT, Parekh Z, Pham H, et al. Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning. London, UK: PMLR; 2021. p. 4904–16.
3. Silva-Rodriguez J, Hajimiri S, Ben Ayed I, Dolz J. A closer look at the few-shot adaptation of large vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2024. p. 23681–90.
4. Zhou K, Yang J, Loy CC, Liu Z. Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2022. p. 16816–25.
5. Zhu B, Niu Y, Han Y, Wu Y, Zhang H. Prompt-aligned gradient for prompt tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ, USA: IEEE; 2023. p. 15659–69.
6. Yu T, Lu Z, Jin X, Chen Z, Wang X. Task residual for tuning vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2023. p. 10899–909.
7. Gao P, Geng S, Zhang R, Ma T, Fang R, Zhang Y, et al. Clip-adapter: better vision-language models with feature adapters. Int J Comput Vis. 2024;132(2):581–95.
8. Zhu X, Zhang R, He B, Zhou A, Wang D, Zhao B, et al. Not all features matter: enhancing few-shot clip with adaptive prior refinement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ, USA: IEEE; 2023. p. 2605–15.
9. Song L, Xue R, Wang H, Sun H, Ge Y, Shan Y, et al. Meta-adapter: an online few-shot learner for vision-language model. Adv Neural Inf Process Syst. 2023;36:55361–74. doi:10.52202/075280-2416.
10. Zhang R, Zhang W, Fang R, Gao P, Li K, Dai J, et al. Tip-adapter: training-free adaption of CLIP for few-shot classification. In: Computer vision—ECCV 2022. Cham, Switzerland: Springer Nature; 2022. p. 493–510. doi:10.1007/978-3-031-19833-5_29.

11. Kato N, Nota Y, Aoki Y. Proto-adapter: efficient training-free CLIP-adapter for few-shot image classification. Sensors. 2024;24(11):3624.
12. Wang Z, Liang J, Sheng L, He R, Wang Z, Tan T. A hard-to-beat baseline for training-free CLIP-based adaptation. arXiv:2402.04087. 2024.
13. Bendou Y, Ouasfi A, Gripon V, Boukhayma A. ProKeR: a kernel perspective on few-shot adaptation of large vision-language models. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference. Piscataway, NJ, USA: IEEE; 2025. p. 25092–102.
14. Li D, Wang R. Text-guided dual feature enhancement: a training-free paradigm for few-shot learning with CLIP. In: 2025 IEEE 6th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT). Piscataway, NJ, USA: IEEE; 2025. p. 1937–40. doi:10.1109/ainit65432.2025.11035349.
15. Guo Z, Zhang R, Qiu L, Ma X, Miao X, He X, et al. Calip: zero-shot enhancement of clip with parameter-free attention. In: Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, CA, USA: AAAI Press; 2023. p. 746–54.
16. Chen X, Li Y, Chen H. Dual-adapter: training-free dual adaptation for few-shot out-of-distribution detection. arXiv:2405.16146. 2024.
17. Morales-Álvarez P, Christodoulidis S, Vakalopoulou M, Piantanida P, Dolz J. BayesAdapter: enhanced uncertainty estimation in CLIP few-shot adaptation. arXiv:2412.09718. 2024.
18. Huang Y, Shakeri F, Dolz J, Boudiaf M, Bahig H, Ben Ayed I. Lp++: a surprisingly strong linear probe for few-shot clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2024. p. 23773–82.
19. Yoon HS, Yoon E, Tee JTJ, Hasegawa-Johnson M, Li Y, Yoo CD. C-TPT: calibrated test-time prompt tuning for vision-language models via text feature dispersion. arXiv:2403.14119. 2024.
20. Upadhyay U, Karthik S, Mancini M, Akata Z. Probvlm: probabilistic adapter for frozen vison-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ, USA: IEEE; 2023. p. 1899–910.
21. Oh C, Lim H, Kim M, Han D, Yun S, Choo J, et al. Towards calibrated robust fine-tuning of vision-language models. Adv Neural Inf Process Syst. 2024;37:12677–707. doi:10.52202/079017-0403.
22. Silva-Rodríguez J, Ben Ayed I, Dolz J. Conformal prediction for zero-shot models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2025. p. 19931–41.
23. Liu J, Shen J, Zhou P, Sonke JJ, Gavves E. Probabilistic prototype calibration of vision-language models for generalized few-shot semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE; 2025. p. 21155–65.
24. Venkataramanan A, Bodesheim P, Denzler J. Probabilistic embeddings for frozen vision-language models: uncertainty quantification with gaussian process latent variable models. In: Chiappa S, Magliacane S, editors. Proceedings of the Forty-First Conference on Uncertainty in Artificial Intelligence. Vol. 286 of Proceedings of Machine Learning Research. London, UK: PMLR; 2025. p. 4309–28.
25. Alparone L, Arienzo A, Lombardini F. Improved coherent processing of synthetic aperture radar data through speckle whitening of single-look complex images. Remote Sens. 2024;16(16):2955. doi:10.3390/rs16162955.
26. Pan X, Zhan X, Shi J, Tang X, Luo P. Switchable whitening for deep representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ, USA: IEEE; 2019. p. 1863–71.
27. Cai M, van Buuren S, Vink G. Joint distribution properties of fully conditional specification under the normal linear model with normal inverse-gamma priors. Sci Rep. 2023;13(1):644. doi:10.1038/s41598-023-27786-y.
28. Griffin JE, Brown PJ. Inference with normal-gamma prior distributions in regression problems. Bayesian Anal. 2010;5(1):171–88. doi:10.1214/10-ba507.
29. Geweke J. Bayesian treatment of the independent student-t linear model. J Appl Econom. 1993;8(S1):S19–40. doi:10.1002/jae.3950080504.
30. Dunn R, Ramdas A, Balakrishnan S, Wasserman L. Gaussian universal likelihood ratio testing. Biometrika. 2023;110(2):319–37. doi:10.1093/biomet/asac064.

31. Yodnual S, Chumnaul J. Signed log-likelihood ratio test for the scale parameter of Poisson Inverse Weibull distribution with the development of PIW4LIFETIME web application. PLoS One. 2025;20(8):e0329293. doi:10.1371/journal.pone.0329293.
32. Parkhi OM, Vedaldi A, Zisserman A, Jawahar C. Cats and dogs. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2012. p. 3498–505.
33. Fei-Fei L, Fergus R, Perona P. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In: 2004 Conference on Computer Vision and Pattern Recognition Workshop. Piscataway, NJ, USA: IEEE; 2004.
34. Maji S, Rahtu E, Kannala J, Blaschko M, Vedaldi A. Fine-grained visual classification of aircraft. arXiv:1306.5151. 2013.
35. Cimpoi M, Maji S, Kokkinos I, Mohamed S, Vedaldi A. Describing textures in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2014. p. 3606–13.
36. Helber P, Bischke B, Dengel A, Borth D. Eurosat: a novel dataset and deep learning benchmark for land use and land cover classification. IEEE J Sel Top Appl Earth Obs Remote Sens. 2019;12(7):2217–26.
37. Soomro K, Zamir AR, Shah M. Ucf101: a dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402. 2012.
38. Lin Z, Yu S, Kuang Z, Pathak D, Ramanan D. Multimodality helps unimodality: cross-modal few-shot learning with multimodal models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2023. p. 19325–37.
39. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2016. p. 770–8.
40. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv:2010.11929. 2021.
41. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: International Conference on Machine Learning. London, UK: PMLR; 2017. p. 1321–30.
42. Nixon J, Dusenberry MW, Zhang L, Jerfel G, Tran D. Measuring calibration in deep learning. arXiv:1904.01685. 2019.
43. Dadalto Câmara Gomes E, Romanelli M, Pichler G, Piantanida P. A data-driven measure of relative uncertainty for misclassification detection. In: Kim B, Yue Y, Chaudhuri S, Fragkiadaki K, Khan M, Sun Y, editors. International Conference on Learning Representations. Red Hook, NY, USA: Curran Associates, Inc.; 2024. p. 21826–48.
44. Geifman Y, El-Yaniv R. Selective classification for deep neural networks. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates, Inc.; 2017. p. 4885–94.
45. Wu YC, Lyu SH, Shang H, Wang X, Qian C. Confidence-aware contrastive learning for selective classification. In: Salakhutdinov R, Kolter Z, Heller K, Weller A, Oliver N, Scarlett J, editors. Proceedings of the 41st International Conference on Machine Learning. Vol. 235 of Proceedings of Machine Learning Research. London, UK: PMLR; 2024. p. 53706–29.
46. Demšar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res. 2006;7:1–30.