Prototypical Few-Shot Classification with Vision-Language Models: A Taxonomy and Systematic Study

João Pedro Perez Resmer[1,2], **Jorge Kysnney Santos Kamassury**[1], **Alexsandro Gehlen**[1]

[1]Centro de Convergência Digital e Mecatrônica (CDM), Fundação CERTI
Florianópolis – SC – Brasil

[2]Universidade Federal de Santa Catarina (UFSC)
Florianópolis – SC – Brasil

{jzr, jkk, agh}@certi.org.br

Abstract. *Prototypical few-shot methods on frozen vision-language models combine a feature transformation, a similarity metric, and a visual-text fusion strategy in ways that resist direct comparison. We organize the quadratic-distance methods as a single scoring family parametrized by a feature transformation and a geometry matrix, recovering Prototypical Networks, cosine similarity, and SimpleShot's L2N variant; Tip-Adapter sits in a separate kernel-based sibling family. Inside it we propose a diagonal Mahalanobis variant whose geometry is estimated from the support set in closed form, with no base-class data and no gradient training, gaining about 2 points over cosine on CUB-200-2011 and 4 on EuroSAT at 1-shot, with overlapping intervals on miniImageNet. All methods are released in one open-source library.*

1. Introduction

Few-shot classification asks a model to discriminate N unseen classes from only K labeled examples per class. The metric-learning approach projects instances into an embedding space where same-class examples sit close to each other. Prototypical Networks [Snell et al. 2017] set the dominant template: build a class prototype as the mean of its support embeddings, then classify a query by its Euclidean distance to those prototypes. SimpleShot [Wang et al. 2019] later showed that simple feature transformations (mean-centering, L2-normalization) plus a nearest-neighbor rule are often enough to match or beat more elaborate meta-learning algorithms in the low-shot regime.

Vision-language models such as CLIP [Radford et al. 2021] and Flamingo [Alayrac et al. 2022], trained on hundreds of millions of image-text pairs, opened a second route. Their textual encoder turns a class name into a zero-shot prototype combinable with visual support to adapt without fine-tuning. Tip-Adapter [Zhang et al. 2022] demonstrated this in a particularly direct form: treat the support set as a key-value cache and add its kernel-based predictions to the zero-shot text logits. Subsequent methods [Zhou et al. 2022, Gao et al. 2024, Zhu et al. 2023b] combine, in different ways, a feature transformation, a similarity metric, and a visual-text fusion strategy. Comparing them head-to-head is awkward in practice: each method ships its own training script, its own data loader, and its own definition of an episode, and the contribution of any single design choice ends up tangled with the rest of the implementation.

This paper recasts the quadratic-distance methods as instances of a single scoring family with two free choices, a feature transformation T and a geometry matrix $\boldsymbol{A}$, and treats Tip-Adapter as a structurally separate kernel-based sibling. Within the quadratic family we study a diagonal Mahalanobis variant whose $\boldsymbol{A}$ is read off the support set in closed form. All methods are evaluated under a shared protocol on frozen CLIP, so that the gap between, for instance, cosine and diagonal Mahalanobis traces to the geometry matrix and not to a difference in episode sampling or feature pre-processing.

The main contributions of this paper are the following. We propose a two-axis taxonomy (T, $\boldsymbol{A}$) that covers Prototypical Networks, cosine similarity, and SimpleShot's L2N variant under a common scoring function, with the kernel-based Tip-Adapter variants placed in a separate sibling family. Within this taxonomy, we introduce a diagonal Mahalanobis variant $\boldsymbol{A} = \hat{\boldsymbol{\Sigma}}_{\gamma}^{-1}$ in which $\hat{\boldsymbol{\Sigma}}$ is a pooled estimator built from the (augmented) support set, with no fine-tuning and no gradient training; the classifier reduces to cosine when intra-class variation vanishes, and at 1-shot it gains roughly 2pp on CUB-200-2011 and 4pp on EuroSAT, the benchmarks furthest from CLIP's pre-training, with overlapping intervals on miniImageNet. Finally, we release a library that exposes the methods compared here behind a common interface, so that swapping the metric is a single argument change rather than a per-method re-implementation.

2. Related Work

2.1. Prototypical Methods and Metric Learning

Snell et al. [Snell et al. 2017] introduced the prototypical formulation in which each class is represented by the centroid of its support embeddings. SimpleShot [Wang et al. 2019] later argued that, once the embeddings come from a sufficiently expressive feature extractor, a nearest-neighbor classifier combined with simple transformations (centering, L2-normalization) is enough to match or exceed meta-learning approaches in the low-shot regime.

Richer metric matrices have also been studied. Bateni et al. [Bateni et al. 2020] use a Mahalanobis distance whose covariance is meta-learned from a disjoint set of base classes and then applied to novel classes under an episodic protocol. The Mahalanobis form itself goes back to Fisher's Linear Discriminant Analysis [Fisher 1936], where the within-class inverse covariance reweights feature directions. We work in a different regime: the backbone is a frozen VLM that has already been pre-trained on hundreds of millions of image-text pairs, and the covariance is computed episode by episode from the support set alone, without a separate meta-training stage on base classes.

A different line of work modifies the prototype itself rather than the metric: self-attention fusion of global and local features [Zhang et al. 2025], transductive refinement via label propagation [Zhu and Koniusz 2023], among others. These are orthogonal to the axis of variation we study, which keeps both the feature extractor and the prototype construction fixed and only changes the scoring function (Section 3).

2.2. Vision-Language Models for Few-Shot Adaptation

CLIP [Radford et al. 2021], Flamingo [Alayrac et al. 2022], and similar models pre-trained on large collections of image-text pairs ship with aligned visual and textual embedding spaces, in which a class name passed through the textual encoder becomes a

zero-shot prototype, matchable against a query embedding without any labeled image of the target class.

Several lines of work then adapt these models with minimal extra training. CoOp [Zhou et al. 2022] replaces hand-written text templates with learned continuous prompts. Tip-Adapter [Zhang et al. 2022] treats the support set as a key-value cache and adds its kernel-based affinity scores to the zero-shot text logits. CLIP-Adapter [Gao et al. 2024] inserts lightweight residual modules on top of the frozen embeddings. The Tip-Adapter score is exponential in the query-key distance, not quadratic in the query-prototype residual, which puts it in a separate family from the Mahalanobis one; we treat it as a sibling in Section 3 and make the relationship precise in Proposition 1. Other CLIP adaptation methods [Zhu et al. 2023b, Udandarao et al. 2023] operate outside this scoring-function structure and are not directly comparable within our framework.

Most of these methods share a further design choice: how to combine the support-driven visual prediction with the prompt-based textual prediction. Tip-Adapter uses a scalar weight on the visual logits; CLIP-Adapter uses a residual interpolation in feature space; CaFo [Zhang et al. 2023] and ProGrad [Zhu et al. 2023a] extend the same idea across multiple foundation models. We use an additive combination in probability space (Section 3.4).

2.3. Software Infrastructure for Few-Shot Evaluation

Tooling for few-shot evaluation has lagged behind the methods themselves. learn2learn [Arnold et al. 2020] and Torchmeta [Deleu et al. 2019] provide standardized episodic data loaders and classical meta-learning algorithms, but both predate the VLM era and offer no clean way to plug a frozen multimodal backbone into a training-free metric-based pipeline.

We therefore built our experiments on top of a library designed to fill this gap.[1] A single prototypical module attaches to any visual or multimodal backbone, picks out the visual and textual encoders when both are available, and combines visual prototypical logits with textual logits from class-describing prompts. The methods compared here plug into the same module through a single argument that selects the scoring function. This is the property we use to keep the empirical comparison paired across methods (Section 4.1).

3. Methodology

3.1. Problem Setup and Notation

Few-shot classification operates on a support set $\mathcal{S} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N \cdot K}$ with K examples (shots) per class for N classes (ways); the task is to classify queries $\boldsymbol{q}$ into one of the N classes. The frozen visual encoder is $f_\theta : \mathcal{X} \to \mathbb{R}^d$ and (for VLMs) the textual encoder is $g_\phi : \mathcal{T} \to \mathbb{R}^d$, both mapping into the same $\mathbb{R}^d$. Let $T : \mathbb{R}^d \to \mathbb{R}^d$ be a deterministic transformation (identity, L2-normalization, etc.) and $\mathcal{I}_k = \{i : y_i = k\}$ the support indices of class k. When test-time augmentation is used, every support image is passed R times through the augmentation pipeline and the resulting augmented features are appended to $\mathcal{I}_k$, so $|\mathcal{I}_k| = KR$ [Zhang et al. 2022]; $R = 1$ reduces to CLIP's standard

[1] https://github.com/fundacaocerti/Fewpy

preprocessing. The transformed prototype is the transform of the mean of the (augmented) support embeddings:

$$\boldsymbol{\mu}_k^{(T)} \;=\; T\!\left(\frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k} f_\theta(\boldsymbol{x}_i)\right). \tag{1}$$

The two orderings (mean-then-transform vs. transform-then-mean) coincide for linear T and disagree for non-linear T at $|\mathcal{I}_k| \geq 2$; we use mean-then-transform throughout, following [Wang et al. 2019].

The prototypical classifier scores the transformed query $\boldsymbol{z}_q = T(f_\theta(\boldsymbol{q}))$ against each prototype through a function s, with class probabilities given by a temperature-τ softmax:

$$p(y=k \mid \boldsymbol{q}) \;=\; \frac{\exp\!\Big(\tau\, s(\boldsymbol{z}_q, \boldsymbol{\mu}_k^{(T)})\Big)}{\sum_{j=1}^{N}\exp\!\Big(\tau\, s(\boldsymbol{z}_q, \boldsymbol{\mu}_j^{(T)})\Big)}. \tag{2}$$

For support points, $\boldsymbol{z}_i = T(f_\theta(\boldsymbol{x}_i))$. Throughout the paper f_θ stays fixed; only s (and the geometry it depends on) is adapted per episode, and τ is the CLIP default $\log(1/0.07)$ [Radford et al. 2021] for every method.

3.2. The Generalized Mahalanobis Family

The Mahalanobis family in this paper consists of scoring functions of the form

$$s_{T,\boldsymbol{A}}(\boldsymbol{z}_q, \boldsymbol{\mu}_k^{(T)}) \;=\; -\,\|\boldsymbol{z}_q - \boldsymbol{\mu}_k^{(T)}\|_{\boldsymbol{A}}^2, \tag{3}$$

with $\|\boldsymbol{v}\|_{\boldsymbol{A}}^2 \coloneqq \boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v}$ for a symmetric positive semi-definite $\boldsymbol{A} \in \mathbb{R}^{d\times d}$. A member is identified by the pair $(T, \boldsymbol{A})$; kernel-based scores, non-linear in the residual, are treated separately in Section 3.3.

Setting $\boldsymbol{A} = \boldsymbol{I}$ recovers two familiar methods: $T = \mathrm{id}$ gives Prototypical Networks [Snell et al. 2017], while $T = T_{L2}(\boldsymbol{x}) = \boldsymbol{x}/\|\boldsymbol{x}\|_2$ gives, for unit-norm $\boldsymbol{z}_q$ and $\boldsymbol{\mu}_k^{(T_{L2})}$, the score $-\|\boldsymbol{z}_q - \boldsymbol{\mu}_k^{(T_{L2})}\|^2 = 2\,\boldsymbol{z}_q^\top \boldsymbol{\mu}_k^{(T_{L2})} - 2$, which is monotone in cosine similarity once the additive constant and factor of 2 are absorbed into τ [Wang et al. 2017]. SimpleShot [Wang et al. 2019]'s L2N variant coincides with our Cosine; its CL2N variant assumes base-class statistics and falls outside our protocol. Established quadratic-distance methods that operate without base classes have all kept $\boldsymbol{A} = \boldsymbol{I}$ and varied only T; the next subsection introduces a choice of $\boldsymbol{A}$ that also requires no base-class data.

Support-estimated geometry. We take $\boldsymbol{A} = \hat{\boldsymbol{\Sigma}}_\gamma^{-1}$ with $\hat{\boldsymbol{\Sigma}}_\gamma = \hat{\boldsymbol{\Sigma}} + \gamma \boldsymbol{I}$ and $\gamma > 0$, where $\hat{\boldsymbol{\Sigma}}$ is computed from the (augmented) support set. The form is motivated by Fisher's Linear Discriminant Analysis [Fisher 1936], in which an inverse covariance reweights the embedding space to favor directions of low within-class variance; Bateni et al. [Bateni et al. 2020] apply this idea with a covariance meta-learned from base classes, whereas ours is computed from the episode support alone. Writing $\boldsymbol{r}_{k,i} := \boldsymbol{z}_i - \boldsymbol{\mu}_k^{(T)}$ for the residual between transformed support embedding i and prototype k, with $k \in \{1,\ldots,N\}$, $i \in \{1,\ldots,M\}$, and $M = N \cdot KR$, the full pooled estimator is the second moment of these residuals:

$$\hat{\boldsymbol{\Sigma}}^{\text{full}} \;=\; \frac{1}{N\cdot M}\sum_{k=1}^{N}\sum_{i=1}^{M} \boldsymbol{r}_{k,i}\,\boldsymbol{r}_{k,i}^\top. \tag{4}$$

The diagonal estimator $\hat{\boldsymbol{\Sigma}}^{\text{diag}} = \text{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_d^2)$ keeps only the diagonal:

$$\hat{\sigma}_j^2 \;=\; \frac{1}{N \cdot M} \sum_{k=1}^{N} \sum_{i=1}^{M} (z_{i,j} - \mu_{k,j}^{(T)})^2. \tag{5}$$

The full form has $\mathcal{O}(d^2)$ parameters and rank at most $\min(N \cdot M, d)$, which at $d = 512$, $N = 5$, $R = 12$ is 300 for $K = 1$ (rank-deficient) and 1500 for $K = 5$, though the realized rank is usually lower since the residuals are not linearly independent. The diagonal form has $\mathcal{O}(d)$ parameters and is full-rank by construction. The $+\gamma \boldsymbol{I}$ term makes both forms positive definite. Expanding Equation 5 yields a decomposition into three non-negative terms:

$$\hat{\sigma}_j^2 \;=\; \underbrace{\tfrac{1}{M} \textstyle\sum_i (z_{i,j} - \bar{z}_j)^2}_{V_z \text{ (support spread)}} \;+\; \underbrace{\tfrac{1}{N} \textstyle\sum_k (\mu_{k,j}^{(T)} - \bar{\mu}_j)^2}_{V_\mu \text{ (prototype spread)}} \;+\; \underbrace{(\bar{z}_j - \bar{\mu}_j)^2}_{\Delta_j^2 \text{ (centroid offset)}}, \tag{6}$$

with $\bar{z}_j = M^{-1} \sum_i z_{i,j}$ and $\bar{\mu}_j = N^{-1} \sum_k \mu_{k,j}^{(T)}$. This is not Fisher's within-class estimator [Fisher 1936], which keeps only the within-class portion of V_z and is undefined at $K = 1$ where our pooled form remains well-defined. We sweep γ in Section 4.6.

Lemma 1 (Graceful degeneration)**.** *Let $\boldsymbol{A} = \hat{\boldsymbol{\Sigma}}_\gamma^{-1}$ with $\hat{\boldsymbol{\Sigma}}_\gamma = \hat{\boldsymbol{\Sigma}} + \gamma \boldsymbol{I}$ and $\gamma > 0$, where $\hat{\boldsymbol{\Sigma}}$ is either the full estimator of Equation 4 or its diagonal restriction (Equation 5). Under $T = T_{L2}$, as $\hat{\boldsymbol{\Sigma}} \to \mathbf{0}$, the classifier defined by Equation 3 converges to the cosine classifier with temperature $\tau' = \tau/\gamma$.*

Proof. As $\hat{\boldsymbol{\Sigma}} \to \mathbf{0}$ (the zero matrix), $\boldsymbol{A} = (\hat{\boldsymbol{\Sigma}} + \gamma \boldsymbol{I})^{-1} \to \gamma^{-1} \boldsymbol{I}$, so by Equation 3 $s_{T_{L2}, \boldsymbol{A}} \to \gamma^{-1} s_{T_{L2}, \boldsymbol{I}}$. Scaling every score by $\gamma^{-1} > 0$ inside Equation 2 is equivalent to $\tau \mapsto \tau/\gamma$, and $s_{T_{L2}, \boldsymbol{I}}$ shares the cosine softmax (Section 3.2). For the diagonal estimator, Equation 6 expresses $\hat{\boldsymbol{\Sigma}} \to \mathbf{0}$ as the simultaneous vanishing of V_z, V_μ, and Δ_j^2 in every coordinate j. □

Remark 1 (Empirical signature of the limit)**.** *The limit of Lemma 1 is approached in our setup when there is nothing left for the support to disagree about: at $K = 1$ and $R = 1$, each class is represented by a single embedding that is also its own prototype ($\boldsymbol{z}_i = \boldsymbol{\mu}_{y_i}^{(T)}$), so the only residuals feeding $\hat{\boldsymbol{\Sigma}}$ are between-class. In this regime, Cosine, both Mahalanobis variants, and Tip-Adapter (KV) return the same ranking by construction. Augmenting the support set is what pushes $\hat{\boldsymbol{\Sigma}}$ away from $\mathbf{0}$ and gives the metric something to do.*

3.3. The RBF-Kernel Sibling Family

Tip-Adapter [Zhang et al. 2022] sits in a different family. Instead of a quadratic score, it applies a Gaussian (RBF) kernel to L2-normalized embeddings, and there are two natural ways to aggregate the kernel over the support set. The *prototype-based* variant applies the kernel directly to the class prototype,

$$s_\beta^{\text{RBF}}(\boldsymbol{z}_q, \boldsymbol{\mu}_k^{(T_{L2})}) \;=\; \exp\Big(\beta\, \boldsymbol{z}_q^\top \boldsymbol{\mu}_k^{(T_{L2})}\Big) \;=\; e^\beta \cdot \exp\Big(-\tfrac{\beta}{2} \|\boldsymbol{z}_q - \boldsymbol{\mu}_k^{(T_{L2})}\|^2\Big), \tag{7}$$

where the second equality uses $\|\boldsymbol{z}_q\| = \|\boldsymbol{\mu}_k^{(T_{L2})}\| = 1$. The *cache-based* variant, which is the original Tip-Adapter formulation, sums kernels over individual support points instead:

$$s_\beta^{\text{KV}}(\boldsymbol{z}_q, k) \;=\; \sum_{i : y_i = k} \exp\big(-\beta \|\boldsymbol{z}_q - \boldsymbol{z}_i\|^2\big). \tag{8}$$

The two variants give the same score when each class has a single effective support point ($|\mathcal{I}_k| = 1$) and otherwise diverge. With $K = 1$ and no augmentation ($R = 1$) the cache has a single point per class and the two scores coincide; once $K \geq 2$ or augmentation populates the cache, the two scores can give different rankings. The following proposition makes the divergence precise in terms of the prototype.

Proposition 1 (Argmax equivalence and divergence)**.** *Let $\beta > 0$ and assume L2-normalized inputs.* (i) *For any $K \geq 1$, the prototype-based RBF classifier (Equation 7) and the cosine classifier produce the same argmax over k.* (ii) *Whenever any class has more than one effective support point ($|\mathcal{I}_k| \geq 2$ for some k), the cache-based RBF classifier (Equation 8) is not a function of $\boldsymbol{\mu}_k^{(T_{L2})}$ alone and therefore is not argmax-equivalent to cosine in general.*

Proof. (i) The exponential and the constant e^{β} in Equation 7 are strictly monotone in $\boldsymbol{z}_q^\top \boldsymbol{\mu}_k^{(T_{L2})}$, so $\arg\max_k s_\beta^{\text{RBF}} = \arg\max_k \boldsymbol{z}_q^\top \boldsymbol{\mu}_k^{(T_{L2})}$, which is the cosine argmax. *(ii)* Two support sets that share the same prototype $\boldsymbol{\mu}_k^{(T_{L2})}$ but differ in their individual embeddings $\{\boldsymbol{z}_i\}_{i \in \mathcal{I}_k}$ yield identical cosine scores yet, in general, different values of s_β^{KV} at any query $\boldsymbol{z}_q$, and therefore different argmaxes. □

When we extend a Tip-Adapter variant to the multimodal setting (Section 3.4), its visual score s^{vis} is either s_β^{RBF} or s_β^{KV}, optionally passed through a softmax. We report both variants.

3.4. Multimodal Extension via Textual Prototypes

A VLM gives us a second prototype for every class, derived from a class-describing prompt t_k: $\boldsymbol{t}_k = T_{L2}(g_\phi(t_k))$. We can therefore form two probability vectors, $\boldsymbol{p}^{\text{vis}}$ from the visual score and $\boldsymbol{p}^{\text{txt}}$ from the textual cosine score, and add them. Writing $\tilde{\boldsymbol{z}}_q = T_{L2}(f_\theta(\boldsymbol{q}))$ for the L2-normalized query,

$$p_k^{\text{vis}} \;=\; \text{softmax}_k\big(\tau\, s^{\text{vis}}(\boldsymbol{z}_q, k)\big), \qquad p_k^{\text{txt}} \;=\; \text{softmax}_k\big(\tilde{\boldsymbol{z}}_q^\top \boldsymbol{t}_k\big), \tag{9}$$

where s^{vis} is the visual score of whichever classifier we are using (Mahalanobis or RBF) and $\tilde{\boldsymbol{z}}_q^\top \boldsymbol{t}_k$ is cosine similarity in the VLM-aligned space. The fused prediction is

$$\tilde{p}_k \;=\; p_k^{\text{vis}} \;+\; \lambda_{\text{txt}} \cdot p_k^{\text{txt}}, \qquad \lambda_{\text{txt}} \geq 0, \tag{10}$$

with $\arg\max_k \tilde{p}_k$ the predicted class. Because $\tilde{p}$ is no longer a normalized distribution, λ_{txt} is only required to be non-negative. The textual score lives in $[-1, 1]$ and does not need temperature scaling, so τ applies only to the visual side. The combination is in the same spirit as the ensembling weight in WiSE-FT [Wortsman et al. 2022], but interpolates in probability space at inference time rather than in weight space.

4. Experiments and Discussion

4.1. Experimental Setup

Datasets and backbone. We run our experiments on three benchmarks chosen to expose different failure modes: *miniImageNet* [Vinyals et al. 2016], natural images, where CLIP is already very strong (saturation regime); *CUB-200-2011* [Wah et al. 2011], bird

species (fine-grained); and *EuroSAT* [Helber et al. 2019], Sentinel-2 satellite imagery, far from CLIP's pre-training distribution (domain shift). For miniImageNet we take the test split of Ravi and Larochelle [Ravi and Larochelle 2017] (20 classes); for CUB-200-2011, the test split of Hilliard et al. [Hilliard et al. 2018] (50 classes). EuroSAT has no standard few-shot split, so we follow [Zhang et al. 2022] and sample episodes from its full 10-class test set. The backbone is a frozen CLIP ViT-B/32 [Radford et al. 2021] with $d = 512$, identical across methods, so that any accuracy difference in the tables can be traced back to the choice of $(T, \boldsymbol{A})$ and (when relevant) the fusion weight λ_{txt}.

Protocol. We follow the standard episodic protocol [Vinyals et al. 2016]: $N = 5$ ways, $K \in \{1, 5\}$ shots, 15 queries per class, 600 episodes per configuration, with deterministic seeds fixed once per configuration so that every method sees the same episode sequence (paired comparison). We report mean accuracy and 95% confidence intervals over the 600 episode accuracies. The support set is augmented by passing each image $R = 12$ times through CLIP's training-time pipeline (random resized crop and random horizontal flip on top of standard preprocessing) [Zhang et al. 2022]; the augmented features are computed once and shared across all methods. The role of augmentation is examined in Section 4.4.

Methods. Eight methods are compared. Two have $\boldsymbol{A} = \boldsymbol{I}$: ProtoNet ($T$ = id) and Cosine ($T = T_{L2}$). Three have a support-estimated or learned $\boldsymbol{A}$: Mahalanobis-diag ($\boldsymbol{A} = (\hat{\boldsymbol{\Sigma}}_\gamma^{\text{diag}})^{-1}$, our main proposal), Mahalanobis-full ($\boldsymbol{A} = (\hat{\boldsymbol{\Sigma}}_\gamma^{\text{full}})^{-1}$), and Learnable Mahalanobis, a parameterized variant with a learnable diagonal $\boldsymbol{s} \in \mathbb{R}^d$ in $\boldsymbol{A}$ that recovers ProtoNet at $\boldsymbol{s} = \boldsymbol{0}$. Three are RBF-kernel methods: Tip-Adapter (KV) and Tip-Adapter (KV+softmax), both applied to class prototypes; and TipAdapter, the original cache-based formulation [Zhang et al. 2022] that aggregates kernels over individual support points (Equation 8). SimpleShot's L2N variant coincides with our Cosine under the mean-then-transform convention and is reported under that name; CL2N assumes base-class statistics and falls outside our protocol.

Hyperparameters. The temperature is fixed at the CLIP initial value $\tau = \log(1/0.07) \approx 2.66$ [Radford et al. 2021] for every method, isolating the effect of $\boldsymbol{A}$ from temperature calibration; argmax predictions are invariant to τ, which only affects the softmax probabilities entering the multimodal fusion. The regularizer is $\gamma = 10^{-6}$ throughout, swept in Section 4.6 on a configuration disjoint from the main results and found to have a plateau spanning $\gamma \in [10^{-8}, 10^{-5}]$; we did not tune it per dataset or per shot count. λ_{txt} is profiled across six orders of magnitude in Table 4.

4.2. When Does a Data-Adaptive Geometry Help?

Table 1 compares all eight methods at $K \in \{1, 5\}$. At 1-shot, the diagonal Mahalanobis is ahead of Cosine on all three benchmarks, decisively on CUB-200-2011 ($+2.09$pp) and EuroSAT ($+4.20$pp) and by a margin within the intervals on miniImageNet ($+0.98$pp). The pattern holds in three regimes that stress the backbone in very different ways (saturation, fine-grained discrimination, domain shift), suggesting that the benefit of an episode-local geometry is not confined to a single type of dataset. At 5-shot, the diagonal form remains in front on miniImageNet (98.21%) and CUB (96.59%) but is overtaken by the full estimator on EuroSAT (84.76% versus 81.24%); Section 4.3 discusses this crossover.

The two prototype-based Tip-Adapter variants post exactly the same accuracies as Cosine, confirming Proposition 1; the cache-based TipAdapter departs from them as soon as the cache holds more than one point per class (already at $K = 1$ once augmentation is on, and more visibly at $K = 5$).

Table 1. Main results: 5-way accuracy (%) with 95% CIs over 600 episodes, $R = 12$. Tip-Adapter (KV) variants belong to the RBF-kernel sibling family; the others are Mahalanobis instances. Bold marks the best per column.

	miniImageNet		CUB-200-2011		EuroSAT	
Method	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
ProtoNet	91.42±0.56	97.86±0.18	83.53±0.81	95.80±0.31	60.46±0.86	74.15±0.66
Cosine	90.72±0.59	97.89±0.18	83.98±0.80	95.84±0.30	61.14±0.87	75.14±0.65
Tip-Adapter (KV)	90.72±0.59	97.89±0.18	83.98±0.80	95.84±0.30	61.14±0.87	75.14±0.65
Tip-Adapter (KV+softmax)	90.72±0.59	97.89±0.18	83.98±0.80	95.84±0.30	61.14±0.87	75.14±0.65
TipAdapter	90.73±0.59	97.50±0.21	84.02±0.80	94.98±0.37	62.59±0.87	68.13±0.75
Learnable Mahalanobis	89.04±0.66	97.86±0.18	83.53±0.81	95.80±0.31	60.46±0.86	74.15±0.66
Mahalanobis-full	83.77±0.85	96.58±0.24	81.03±0.84	92.18±0.41	**68.15±0.81**	**84.76±0.51**
Mahalanobis-diag (ours)	**91.70±0.56**	**98.21±0.16**	**86.07±0.77**	**96.59±0.27**	65.34±0.83	81.24±0.54

The Learnable Mahalanobis row in Table 1 shows the behavior of this parameterized variant. After training on the support set with the standard cross-entropy loss, it matches ProtoNet to the second decimal on five of the six configurations, in both mean accuracy and confidence interval. The scale vector $\boldsymbol{s}$ stays at its initialization, a stationary point gradient training does not escape at this data scale.

4.3. Diagonal versus Full Covariance Estimation

Comparing the diagonal and full forms of $\boldsymbol{A} = \hat{\boldsymbol{\Sigma}}_\gamma^{-1}$ at fixed $T = T_{L2}$ in Table 1 shows that one form does not dominate the other. On miniImageNet and CUB the diagonal form is in front at both shot counts (e.g., 91.70% vs 83.77% at mini 1-shot, 86.07% vs 81.03% at CUB 1-shot). Here a per-dimension scaling captures most of the signal, while the full estimator must fit $\mathcal{O}(d^2)$ entries from $N \cdot M$ rank-one outer products, estimating many small eigenvalues from very little data.

On EuroSAT, where the visual domain is unlike anything in CLIP's pre-training, the full estimator overtakes at both shot counts (68.15% vs 65.34% at 1-shot, 84.76% vs 81.24% at 5-shot). The most likely reason is that off-diagonal covariance structure carries useful signal under domain shift, and the diagonal restriction discards it. Cosine accuracy itself is a usable proxy for this distance, and across the three benchmarks it falls (90.72%, 83.98%, 61.14%) as the gain from an episode-local geometry rises ($+0.98$, $+2.09$, $+4.20$pp).

We adopt the diagonal form as our main proposal: it only needs $\mathcal{O}(d)$ parameters and is stable across several orders of magnitude of γ (Section 4.6). We still report the full form, since for sufficiently shifted target domains it is the better choice.

4.4. Behavior When Intra-Class Variation Vanishes

Lemma 1 says that the classifier reduces to cosine in the limit $\hat{\boldsymbol{\Sigma}} \to \mathbf{0}$. The closest regime we can reach experimentally is $K = 1$ with no augmentation: each class contributes a

single embedding which is also its prototype, so the intra-class portion of the residuals is empty (Remark 1). Table 2 shows what happens. At $R = 1$ on EuroSAT 1-shot, both support-estimated Mahalanobis variants and Tip-Adapter (KV) collapse to 61.90%, matching the Cosine baseline. ProtoNet and Learnable Mahalanobis land 0.04pp away (61.86%); the gap is entirely the absence of L2-normalization. At $R = 12$ the gap over Cosine returns, 4 percentage points (diagonal) and 7 (full), while both $\boldsymbol{A} = \boldsymbol{I}$ baselines lose accuracy, so augmentation alone does not explain the gain. The geometry needs intra-class variation in the support to do useful work; augmentation is what supplies it.

Table 2. Effect of support augmentation (EuroSAT 5-way 1-shot, 600 episodes, 95% CIs). Cosine, Tip-Adapter (KV), and both Mahalanobis variants coincide at $R = 1$ (the empirical signature of Lemma 1); bold marks the best method at $R = 12$.

	EuroSAT 1-shot	
Method	$R = 1$	$R = 12$
Cosine / Tip-Adapter (KV)	61.90±0.89	61.14±0.87
Mahalanobis-diag	61.90±0.89	65.34±0.83
Mahalanobis-full	61.90±0.89	**68.15±0.81**
ProtoNet	61.86±0.88	60.46±0.86
Learnable Mahalanobis	61.86±0.88	60.46±0.86

4.5. Multimodal Fusion

Table 3 compares visual-only with multimodal accuracy at $R = 12$, with λ_{txt} at each method's own optimum on each benchmark; Table 4 shows the corresponding profile for EuroSAT at 5-shot without augmentation.

Table 3. Multimodal fusion vs visual-only (5-way 1-shot, $R = 12$); λ_{txt} at each method's own optimum, an oracle choice applied equally to all rows. On miniImageNet the optimum sits at $\lambda_{\text{txt}} = 0$, so the multimodal column matches the visual one. Bold marks the best per column.

	miniImageNet		CUB-200-2011		EuroSAT	
Method	Visual	+ Text	Visual	+ Text	Visual	+ Text
Cosine	90.72±0.59	90.72±0.59	83.98±0.80	**96.71±0.35**	61.14±0.87	**71.95±0.73**
Mahalanobis-diag	**91.70±0.56**	**91.70±0.56**	**86.07±0.77**	96.46±0.34	**65.34±0.83**	71.27±0.71
Tip-Adapter (KV+softmax)	90.72±0.59	90.72±0.59	83.98±0.80	96.53±0.38	61.14±0.87	**71.95±0.72**

Two patterns stand out in Table 3. First, the optimum λ_{txt} varies across methods and benchmarks, from 0 on miniImageNet (where the textual signal does not help, regardless of metric) to non-zero values elsewhere, but the Mahalanobis variants are strikingly insensitive to it: their logits are scaled by the largest eigenvalues of $\hat{\boldsymbol{\Sigma}}_{\gamma}^{-1}$, of order $1/\gamma = 10^6$, so $\boldsymbol{p}^{\text{vis}}$ is sharply peaked and resists being displaced. Across four decades of λ_{txt} (Table 4) both Mahalanobis forms move by under 1pp, against 4 for Cosine and 23 for Tip-Adapter.

Second, once the textual prediction is fused, the visual-side choice is largely absorbed: Cosine, Mahalanobis-diag, and Tip-Adapter (KV+softmax) finish within statistical error of each other on CUB (96.71 / 96.46 / 96.53) and on EuroSAT (71.95 / 71.27

/ 71.95). The textual signal is the same for everyone and ends up dominating the combined prediction wherever it carries useful information at all; on miniImageNet, where the textual prototype adds nothing, the multimodal accuracy reduces to the visual-only one.

Table 4. Sensitivity of multimodal accuracy (%) to λ_{txt} (EuroSAT 5-way 5-shot, no augmentation). Bold marks each method's own peak across λ_{txt}.

λ_{txt}	10^{-2}	10^{-1}	1	10	10^{2}	10^{3}
Cosine	78.24	78.27	78.61	**80.34**	75.90	57.60
Mahalanobis-diag	**83.45**	82.75	82.50	82.61	82.92	65.30
Mahalanobis-full	**84.76**	84.26	83.97	84.34	84.42	65.61
Tip-Adapter (KV+softmax)	78.25	78.46	**80.49**	75.24	57.72	54.10
TipAdapter	**71.37**	71.36	71.36	71.22	66.27	19.78

4.6. Sensitivity to the Regularizer γ

Beyond the CLIP-default temperature, the support-estimated Mahalanobis has one hyperparameter: the regularizer γ. Table 5 sweeps γ over nine orders of magnitude on EuroSAT 5-way 5-shot with multimodal fusion at $\lambda_{\text{txt}} = 1$.

Table 5. Sensitivity of Mahalanobis estimators to γ (EuroSAT 5-way 5-shot, no augmentation, $\lambda_{\text{txt}} = 1$). Bold marks each estimator's own peak across γ.

γ	10^{-8}	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1
Mahalanobis-diag	82.89	82.81	**83.01**	82.68	82.08	80.94	79.29	79.48	78.46
Mahalanobis-full	84.10	83.80	84.46	**84.48**	84.06	84.02	82.80	80.36	78.55

Both estimators have a stable plateau across several orders of magnitude of γ: the diagonal form peaks at $\gamma \approx 10^{-6}$ and stays within a narrow band for $\gamma \in [10^{-8}, 10^{-5}]$ before degrading, consistent with the analytical limit $\boldsymbol{A} \to (1/\gamma)\boldsymbol{I}$; the full form peaks at $\gamma \approx 10^{-5}$ with an even broader plateau across $\gamma \in [10^{-6}, 10^{-3}]$.

5. Conclusion

A handful of prototypical methods on frozen vision-language models turn out to be the same scoring rule with different choices of two ingredients: the feature transformation T and the geometry matrix $\boldsymbol{A}$. Tip-Adapter does not fit this template; its exponential kernel puts it in a separate family, although the prototype-based version of its score collapses, in argmax, onto cosine. What is left to vary is essentially the choice of $\boldsymbol{A}$, and the contribution of this paper is the simplest non-identity choice that needs no extra data: $\boldsymbol{A} = \hat{\boldsymbol{\Sigma}}_{\gamma}^{-1}$ with $\hat{\boldsymbol{\Sigma}}$ a diagonal pooled estimator over the augmented support.

The empirical picture is consistent across three very different domains. When intra-class variation is large enough for the estimator to see (under support augmentation), the data-adaptive geometry is ahead of cosine on the visual side at 1-shot, and the gain grows with the distance between the target domain and CLIP's pre-training. When that variation is suppressed, the diagonal Mahalanobis falls back to cosine by construction, so the method adds something only where there is something to add. The full covariance counterpart pays off in the most domain-shifted regime, where off-diagonal structure

carries useful signal. Once the textual prototype enters the prediction, the visual metric matters much less: methods that differ by 4 pp on visual EuroSAT collapse to within statistical noise after fusion. The most useful continuations of this work will likely come from running the same comparison on larger backbones (ViT-L/14, DINOv2, SigLIP) and on domains further from CLIP's training distribution, where the diagonal-versus-full crossover we already see on EuroSAT should become the rule rather than the exception.

Limitations and scope. Our taxonomy covers quadratic-distance and RBF-kernel scoring on frozen embeddings; the convergence to cosine is a limit statement, not a quantitative bound at finite $\hat{\Sigma}$. We report confidence intervals but no paired test; since the protocol is paired, these are conservative, but the miniImageNet 1-shot gap should be read as a direction rather than an established difference. The augmentation ablation covers EuroSAT 1-shot only, and we do not sweep R. We use a single backbone (CLIP ViT-B/32); our two axes predict that larger d starves the full estimator further, favoring the diagonal, and that a better-aligned backbone shrinks the gain by the saturation already visible on miniImageNet, but neither is verified here. EuroSAT has no standard class-disjoint few-shot split, so we evaluate on its full 10-class test set: cross-method comparisons remain valid in our paired protocol, but the absolute numbers are not directly comparable with FSL studies that use a different split.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. (2022). Flamingo: a visual language model for few-shot learning. *NeurIPS*, 35:23716–23736.

Arnold, S. M. R., Mahajan, P., Datta, D., Bunner, I., and Zarkadas, K. S. (2020). learn2learn: A library for meta-learning research. In *ICML AutoML Workshop*.

Bateni, P., Goyal, R., Masrani, V., Wood, F., and Sigal, L. (2020). Improved few-shot visual classification. In *CVPR*.

Deleu, T., Würfl, T., Samiei, M., Cohen, J. P., and Bengio, Y. (2019). Torchmeta: A meta-learning library for PyTorch. arXiv:1909.06576.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188.

Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., and Qiao, Y. (2024). Clip-adapter: Better vision-language models with feature adapters. *IJCV*, 132(2):581–595.

Helber, P., Bischke, B., Dengel, A., and Borth, D. (2019). Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J. Sel. Top. Appl. Earth Obs.*, 12(7):2217–2226.

Hilliard, N., Phillips, L., Howland, S., Yankov, A., Corley, C. D., and Hodas, N. O. (2018). Few-shot learning with metric-agnostic conditional embeddings. arXiv:1802.04376.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR.

Ravi, S. and Larochelle, H. (2017). Optimization as a model for few-shot learning. In *ICLR*.

Snell, J., Swersky, K., and Zemel, R. (2017). Prototypical networks for few-shot learning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *NeurIPS*, volume 30. Curran Associates, Inc.

Udandarao, V., Gupta, A., and Albanie, S. (2023). SuS-X: Training-Free Name-Only Transfer of Vision-Language Models. In *ICCV*, pages 2725–2736.

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., and Wierstra, D. (2016). Matching networks for one shot learning. In *NeurIPS*, volume 29.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. (2011). The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Wang, F., Xiang, X., Cheng, J., and Yuille, A. L. (2017). Normface: L_2 hypersphere embedding for face verification. In *ACM Multimedia*, MM '17, pages 1041–1049. ACM.

Wang, Y., Chao, W.-L., Weinberger, K. Q., and Maaten, L. v. d. (2019). SimpleShot: Revisiting nearest-neighbor classification for few-shot learning. arXiv:1911.04623.

Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith, S., Roelofs, R., Gontijo-Lopes, R., Hajishirzi, H., Farhadi, A., Namkoong, H., and Schmidt, L. (2022). Robust fine-tuning of zero-shot models. arXiv:2109.01903.

Zhang, L., Yang, X., Cheng, X., Cheng, W., and Lin, Y. (2025). Few-Shot Image Classification Algorithm Based on Global–Local Feature Fusion. *AI*, 6(10):265.

Zhang, R., Hu, X., Li, B., Huang, S., Deng, H., Qiao, Y., Gao, P., and Li, H. (2023). Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *CVPR*, pages 15211–15222.

Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., and Li, H. (2022). Tip-Adapter: Training-Free Adaption of CLIP for Few-Shot Classification. In Avidan, S., Brostow, G., Cissé, M., Farinella, G. M., and Hassner, T., editors, *ECCV*, volume 13695, pages 493–510. Springer Nature Switzerland, Cham. Series Title: Lecture Notes in Computer Science.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. (2022). Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348.

Zhu, B., Niu, Y., Han, Y., Wu, Y., and Zhang, H. (2023a). Prompt-aligned gradient for prompt tuning. In *ICCV*, pages 15659–15669.

Zhu, H. and Koniusz, P. (2023). Transductive few-shot learning with prototype-based label propagation by iterative graph refinement. In *CVPR*.

Zhu, X., Zhang, R., He, B., Zhou, A., Wang, D., Zhao, B., and Gao, P. (2023b). Not All Features Matter: Enhancing Few-shot CLIP with Adaptive Prior Refinement. In *ICCV*, pages 2605–2615.