Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Training-Free Few-Shot Image Classification via Diagnostic Verification and Bounded Correction

**Lei Chen**
Naval University of Engineering
**Li Duan**
Naval University of Engineering
**Rui Fang**
Naval University of Engineering
**HongPing Zhang**
d25181002@nue.edu.cn

Naval University of Engineering

**Research Article**

**Keywords:** Few-shot image classification, Training-free adaptation, Ambiguity resolution, Bounded correction

**Posted Date:** November 20th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11133832/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Training-Free Few-Shot Image Classification via Diagnostic Verification and Bounded Correction

Lei Chen, Li Duan, Rui Fang, Hongping Zhang*

## Abstract

Training-free adaptation of CLIP has become a promising paradigm for few-shot image classification because it avoids gradient optimization and overfitting. Existing methods typically construct visual and textual classifiers independently and fuse them with globally fixed weights, implicitly assuming that all samples share the same optimal fusion weight. This assumption fails in few-shot settings: when the top-2 logits are close, fixed weights cannot resolve ambiguity, while unbounded z-score correction may over-correct ambiguous samples. We propose DiVe, a Diagnostic Verification and Bounded Correction framework. For each sample, DiVe verifies its top-2 candidates, constructs a discriminative direction from fine-grained textual prototypes, projects test features and visual prototypes onto this direction to quantify attribution, and converts verification confidence into a bounded correction signal via a tanh mapping. DiVe is fully training-free and can be plugged into existing training-free CLIP classifiers as post-processing. On 11 datasets with 4 CLIP backbones under the 16-shot setting, DiVe achieves the best average accuracy among training-free methods, reaching 76.43% with ResNet-50, and improves GDA-CLIP by 0.92% and TIMO by 0.51% on six representative datasets. Cross-dataset experiments on ImageNet variants demonstrate robustness under distribution shift, while also revealing boundary conditions under extreme domain gaps.

**Keywords:** Few-shot image classification; Training-free adaptation; Ambiguity resolution; Bounded correction

## 1 Introduction

Contrastive Language-Image Pre-training (CLIP) [1] and ALIGN [2] learn aligned visual-text representations through image-text contrastive learning, obtaining strongly generalizable cross-modal representations whose zero-shot transfer capability can adapt to various visual tasks without downstream annotations. However, in few-shot scenarios where only 1 to 16 annotated samples per class are available, directly using the zero-shot classifier fails to fully exploit the limited annotation information, leaving significant room for performance improvement. Few-shot learning aims to rapidly generalize to novel categories from extremely few annotated samples. Early research mostly adopted meta-learning as the fundamental framework, such as Matching Networks [3] and Model-Agnostic Meta-Learning [4], which learn transferable metrics or model initializations across tasks. In recent years, direct downstream adaptation based on large-scale pre-trained image-text models such as CLIP has gradually become the mainstream paradigm for few-shot image classification, as it eliminates the need for multi-round meta-training on base-class tasks.

CLIP-based few-shot adaptation methods can be divided into two categories: training-based and training-free. Training-based methods introduce learnable parameters and perform gradient optimization on few-shot data: CoOp [5] replaces handcrafted prompts with learnable continuous prompt vectors; CLIP-Adapter [6] introduces lightweight residual adapters on frozen features; CoCoOp [7] further enables prompt vectors to be dynamically generated conditioned on input images; Visual Prompt Tuning

(VPT) [8] introduces a small number of learnable tokens at the visual input; MaPLe [9] jointly learns multimodal prompts in both visual and language branches. These methods achieve higher accuracy but incur additional training overhead and carry overfitting risks when samples are extremely scarce. Training-free methods completely avoid gradient optimization: Tip-Adapter [10], GDA-CLIP [11], TIMO [12], etc., exploit support set information through non-parametric classifiers or statistical models; Test-time Prompt Tuning (TPT) [13] also does not update parameters and achieves prompt adaptation based solely on individual test samples. These methods combine efficiency and stability, and are the focus of enhancement in this paper.

Most training-free methods share a common assumption: separately construct classifiers based on visual features and text prompts, then integrate logits through globally fixed weights. Taking the performance of GDA-CLIP [11] on ImageNet 16-shot as an example (statistics shown in Fig. 1), samples with a top-2 logit margin less than 0.5 have an error rate of approximately 69.08%, while samples with a margin greater than 10.0 have an error rate of only 3.4%, indicating that ambiguous samples are the primary source of errors. Specifically, the limitations of the existing paradigm manifest at two levels: (1) when top-2 predicted logits are close, there lacks an explicit binary verification mechanism to determine which candidate is more reliable; (2) a few post-processing correction methods employ z-score normalization with a fixed scaling factor, whose unbounded nature may generate correction signals with excessive amplitude on ambiguous samples, causing already correct predictions to be erroneously flipped. In short, existing methods lack a sample-level decision mechanism for "where to correct and how much to correct."

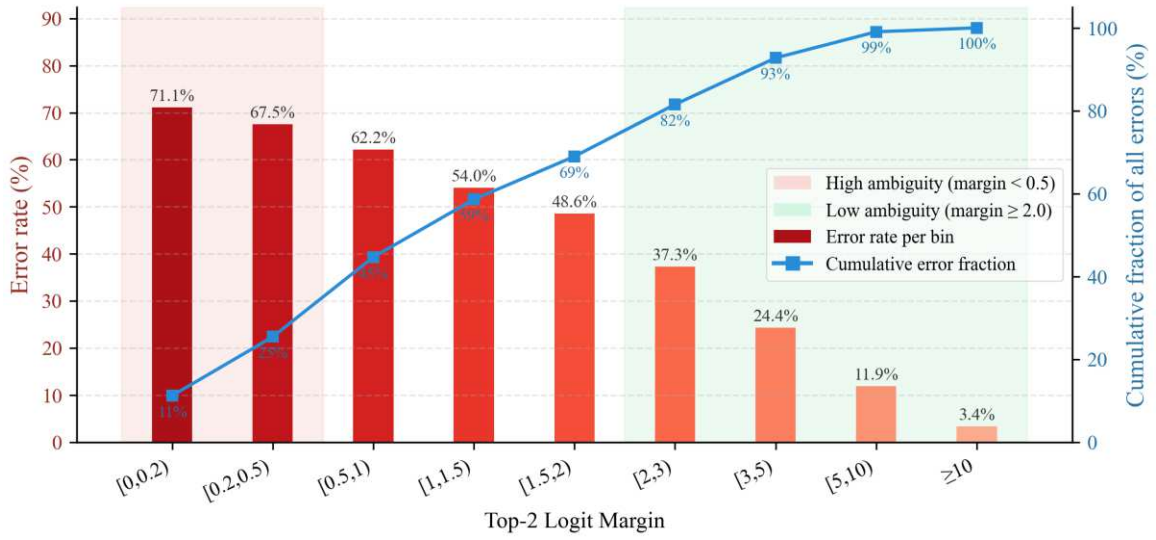



Fig. 1. Statistics of prediction ambiguity and error distribution for the GDA-CLIP model. Bars represent the error rate of samples within different top-2 logit margin intervals; the blue line shows the cumulative proportion of erroneous samples in each interval relative to all erroneous samples. The vast majority of errors are concentrated in highly ambiguous samples with margin<2.0, providing statistical evidence for the motivation of the proposed method.

The DiVe framework proposed in this paper (overall structure shown in Fig. 2) is designed to address the above issues. The Diagnostic Verification Module (DVM) is responsible for "where to correct": it identifies top-2 ambiguous candidates, constructs a stable discriminative direction using fine-grained text prototypes, projects test features and visual prototypes onto this direction, and computes verification scores to determine attribution. The Diagnostic Confidence-Driven Correction (DCDC) module is responsible for "how much to correct": it converts verification confidence into a correction signal through a bounded *tanh* mapping, applying strong correction at high confidence and conservative correction at low confidence, to avoid the over-correction risk of unbounded normalization. DVM and DCDC form an execution-level coupling: DCDC takes the verification score

output by DVM as its sole input and does not independently compute intermediate quantities. The main contributions of this paper are as follows:

- To address the deficiency of globally fixed-weight fusion in training-free CLIP, we propose DiVe, a sample-level hypothesis-verification post-processing pipeline: it integrates top-2 ambiguity identification, fine-grained text prototype discriminative projection, and bounded confidence-driven correction to achieve adaptive decision-making for "where to correct and how much to correct";
- We propose DCDC bounded confidence correction, and demonstrate the advantages of *tanh* mapping over z-score normalization in few-shot training-free scenarios, mitigating over-correction;
- As a general post-processing plugin, DiVe can seamlessly integrate with Tip-Adapter/GDA-CLIP/TIMO. Its generalization is validated on 11 datasets, multiple backbones, and distribution shift scenarios, and the failure boundaries of the method are explicitly revealed.

## 2 Related Work

### *2.1 CLIP-based Few-Shot Image Classification*

Since the proposal of CLIP [1], adapting its visual-language aligned representations to downstream few-shot classification has become an important research direction. Training-based adaptation methods exploit few-shot information through learnable parameters: CoOp [5] pioneered the replacement of handcrafted prompts with learnable continuous prompt vectors, significantly improving few-shot performance; CLIP-Adapter [6] introduced a lightweight residual adapter network on CLIP features; CoCoOp [7] extended CoOp to a conditional prompt generation mechanism, enabling prompt vectors to dynamically adjust with input images; CLAP [7]

proposed a training strategy based on feature calibration through class-adaptive linear probing. These methods achieve higher accuracy, but the computational overhead of gradient optimization and the overfitting risk under few-shot settings have prompted researchers to turn to training-free adaptation.

*2.2 Training-Free CLIP Adaptation*

Training-free adaptation methods avoid gradient optimization and exploit support set information through non-parametric or statistical methods. Tip-Adapter [10] is an early representative work in this direction, which constructs support set features as a key-value cache, performs non-parametric classification through similarity weighting, and integrates with the zero-shot classifier; Tip-X [15] removes unnecessary learnable parameters on this basis to further simplify the model; GDA-CLIP [11] revisits Gaussian Discriminant Analysis and proves that a GDA classifier based on empirical Bayes estimation in the CLIP feature space can serve as a highly competitive training-free baseline; APE [16] improves the quality of the zero-shot classifier through adaptive feature selection and prior refinement.

Recent work has begun to focus on cross-modal interaction: TIMO [12] addresses the issues of anomalous matching and prompt quality disparity in CLIP, proposing a text-image mutual guidance mechanism; ProKeR [17] provides a unified interpretation from the theoretical perspective of kernel ridge regression, explaining Tip-Adapter and GDA-CLIP as its special cases. Most of the above methods still follow the "independent construction—global fusion" paradigm, lacking adaptive processing of sample-level ambiguity. The DiVe proposed in this paper introduces explicit sample-level

diagnostic verification and bounded correction, which can serve as a general plugin to enhance existing training-free methods.

*2.3 Human-Inspired Visual Recognition and Ambiguity Resolution*

Incorporating human visual cognitive mechanisms into computer vision is a long-standing research direction. Cognitive science research shows that human visual recognition follows a "hypothesis generation—verification" paradigm, i.e., generating candidate hypotheses based on initial perception, then obtaining discriminative evidence through selective attention for verification [18,19]. In computer vision, the attention mechanism can be regarded as a simulation of selective attention [20] curriculum learning draws on the cognitive development law of progressing from easy to difficult [21]; prototype-based classification methods align with the human pattern of classifying based on conceptual prototypes [22]. However, in the training-free CLIP few-shot adaptation scenario, existing work has not yet explicitly introduced the complete cognitive flow of "hypothesis generation—targeted verification—confidence decision." This paper systematically introduces this mechanism into the training-free few-shot classification framework.

## 3 Method

*3.1 Problem Definition and Preliminaries*

Let the pre-trained CLIP visual encoder $f_v(\cdot)$ and the text encoder $f_t(\cdot)$. The normalized visual feature of an input test image $I$ is denoted as $x \in \mathbb{R}^D$, where $D$ is the feature dimension. Given a $K$-class few-shot task, the support set $S=\{(x_i,y_i)\}_{i=1}^{K\times N_s}$ contains $N_s$ annotated samples per class. Support image

features are extracted by the visual encoder and L2-normalized: $x_i = f_v(I_i)/\|f_v(I_i)\|_2$.

**Base classifier construction.** We use GDA-CLIP [23] as base classifier construction , for each category $k$, the class mean $\mu_k^{vis}$ is estimated from support sample features. An empirical Bayes ridge estimator is used for inversion:

$$\hat{\Sigma}^{-1} = D\cdot((N_sK\text{-}1)\hat{\Sigma} + tr(\hat{\Sigma})I_D)^{-1}, \quad (1)$$

where $N_sK$ is the total number of support samples, $D$ is the dimension of the features, and $\hat{\Sigma}$ represents the empirical covariance. The GDA classifier weights $\mathbf{W}_k=\hat{\Sigma}^{-1}\mu_k^{vis}$ and bias $b_k=-\frac{1}{2}(\mu_k^{vis})^T\hat{\Sigma}^{-1}\mu_k^{vis}+\log(1/K)$ are obtained.

The zero-shot text classifier weights $\mathbf{W}_c\in\mathbb{R}^{K\times D}$ are obtained by encoding category prompt texts (e.g., "a photo of a {class}" [1]) through $f_t$ encoding. The final logits of the base classifier are:

$$l = 100\cdot x\,\mathbf{W}_c^T + \alpha\,(x\mathbf{W}^T + \boldsymbol{b}), \quad (2)$$

where $\mathbf{W} = [\mathbf{W}_1,\ldots\mathbf{W}_K]^T$, $\boldsymbol{b} = [b_1,\ldots,b_K]^T$, and $\alpha> 0$ is the visual-text fusion weight, its search range $[10^{-3}, 10^2]$ is inherited from GDA-CLIP [11] and kept unchanged for all compared variants. The proposed method performs sample-level adaptive post-processing on the base classifier without modifying its parameters or introducing training.

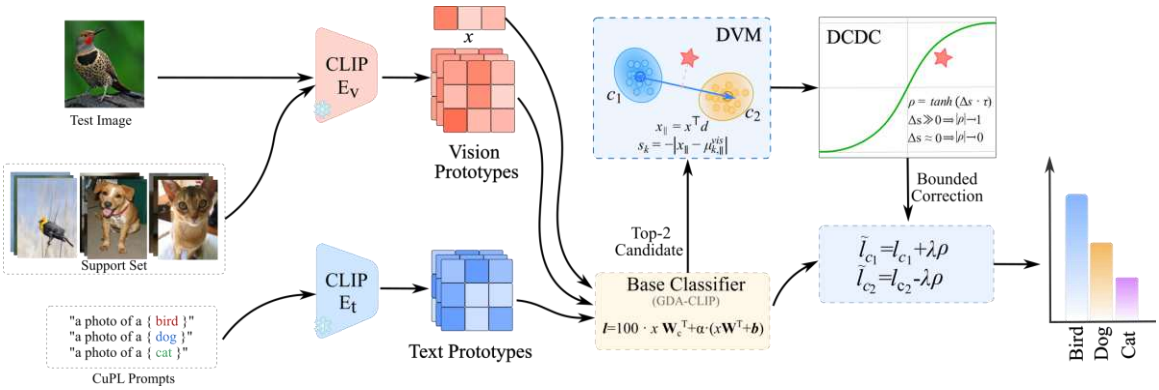



Fig. 2. Overview of the DiVe framework. Given a test image, the base GDA-CLIP classifier produces initial logits. DVM identifies the top-2 ambiguous categories, constructs a discriminative direction $d$ from fine-grained text prototypes, and projects the test feature and visual prototypes onto this direction to obtain verification scores. DCDC converts the score difference into a bounded correction signal $\rho$ through *tanh* mapping and applies it to the top-2 logits. The entire process is training-free and runs per sample.

### *3.2 Failure Analysis of the Global Fusion Paradigm and Diagnostic Motivation*

The base classifier applies the same $\alpha$ to all test samples, implicitly assuming that all samples have the same classification difficulty. This assumption severely deviates from reality in few-shot scenarios. To quantitatively characterize this problem, we analyze the prediction behavior of the base GDA-CLIP classifier on the ImageNet dataset (ResNet-50, 16-shot):

Among samples with a top-2 logit margin $\delta = l_{c_1} - l_{c_2} < 1.0$ (accounting for approximately 45% of the test set), the error rate is 60.02%; among samples with $\delta > 2.0$ (accounting for approximately 31% of the test set), the error rate is only 26.71%.

This means that the base classifier can already make stable and correct predictions for high-confidence samples, and its performance bottleneck is

concentrated on boundary samples with small top-2 margins, which account for a non-negligible proportion (approximately 45%) of the test set and constitute the main source of errors. However, the global fusion paradigm cannot distinguish boundary samples from simple samples: applying the same correction strength to both wastes the high-confidence information of simple samples and lacks sufficient discriminative power on boundary samples. The core design principle of this paper is: perform targeted explicit verification and correction only on ambiguous samples, while maintaining minimal intervention on high-confidence samples. Specifically, DVM first locates ambiguous candidates based on the top-2 margin, then measures the direction and strength of evidence using the verification confidence $\Delta s$, while DCDC converts it into correction strength through bounded mapping, thereby achieving "correction on demand."

*3.3 Diagnostic Verification Module (DVM)*

As shown in Fig. 3(a), the goal of DVM is: for each test sample, construct a discriminative axis in the feature space, project the test feature and the visual prototypes of candidate categories onto this axis, and determine attribution by comparing projection positions.

*3.3.1 Candidate Hypothesis Generation*

Inspired by the "hypothesis generation" mechanism in human visual cognition, DVM first identifies the two most likely candidate categories for each test sample, forming a verification hypothesis set. The category indices corresponding to the maximum and second-largest values are selected from the base classifier logits $l$:

$$(c_1, c_2) = \left( \operatorname*{argmax}_{k} l_k , \operatorname*{argmax}_{k \neq c_1} l_k \right) \tag{3}$$

Where $c_1$ and $c_2$ denote the top-1 and top-2 candidate categories, respectively. Selecting only top-2 rather than top-k (k>2) is based on the following considerations: as shown in Fig. 1, the vast majority of errors are concentrated in confusion between top-2 categories; selecting more candidates not only increases computational overhead but also introduces noise from low-ranked categories, while top-2 verification already covers the main ambiguity cases.

*3.3.2 Discriminative Direction Construction Based on Fine-Grained Text Prototypes*

After determining candidate categories, DVM needs to find a discriminative direction in the feature space that can maximally distinguish$c_1$ and $c_2$. An intuitive idea is to use the difference vector of visual class prototypes, i.e., $d_{vis} = \mu_{c_1}^{vis} - \mu_{c_2}^{vis}$. However, in the few-shot setting, visual class prototypes are estimated from very few support samples, resulting in high estimation variance and poor stability. In particular, when support samples are not representative of the category, the visual prototype difference may fail to stably characterize the true inter-category difference.

To this end, this paper constructs the discriminative direction using fine-grained text prototypes. We adopt category descriptive prompts generated by a large language model (CuPL [23]) and generate $P$ descriptive texts for each category $k$ (e.g., "A tench looks like a green freshwater fish with a brownish hue." "Gnocchi are small, doughy dumplings."), from which a set of text embeddings$T_k = \{t_{k,p}\}_{p=1}^{P}$ is obtained through the CLIP text encoder. The fine-grained text prototype of category $k$ is defined as the mean of descriptive text embeddings, L2-normalized:

$$\mu_k^{txt} = \text{normalize}\left(\frac{1}{P}\sum_{p=1}^{P} t_{k,p}\right) \quad (4)$$

Compared to simple prompts containing only category names, fine-grained descriptive prompts contain rich semantic attributes (shape, color, texture, function, etc.), and text prototypes can more accurately characterize inter-category semantic differences. The discriminative direction between $c_1$ and $c_2$ is defined as the normalized vector of their text prototype difference:

$$d = \frac{\mu_{c_1}^{txt} - \mu_{c_2}^{txt}}{\|\mu_{c_1}^{txt} - \mu_{c_2}^{txt}\|_2} \quad (5)$$

Since CLIP visual and text features are aligned to the same embedding space through contrastive learning, the discriminative direction constructed from text prototypes can be directly used to project visual features. This discriminative direction *d* has three properties: (1) text prototypes are aggregated from dozens of descriptive prompts, independent of scarce support images, and are more stable than visual prototypes estimated from *Ns* support samples under few-shot conditions; (2) *d* lies along the connecting direction of the two text prototypes, characterizing their discriminative direction in semantic space; projecting visual features onto this direction is equivalent to evaluating test image attribution on a "semantic discriminative axis," consistent with the human cognitive pattern of distinguishing similar objects using semantic attributes; (3) the construction of *d* is entirely text-based, providing a stable discriminative axis even under extreme few-shot conditions.

*3.3.3 Projection Verification and Scoring*

After constructing the discriminative direction, DVM projects the test image feature and the candidate category visual prototypes onto this direction, evaluating which category the test image is closer to on the

discriminative axis. The visual class prototype $\mu_k^{vis}$ is defined as the mean of support sample features for category $k$. The projection is computed as follows:

$$x_{\parallel} = x^T d, \quad \mu_{k,\parallel}^{vis} = \left(\mu_k^{vis}\right)^T d, \ k \in \{c_1, c_2\} \tag{6}$$

Where $x_{\parallel}$ is the scalar projection of the test feature onto the discriminative direction, and $\mu_{k,\parallel}^{vis}$ is the projection of the visual prototype of category $k$. The verification score for each candidate category is defined as the negative Euclidean distance between the test projection and the prototype projection of that category:

$$s_k = -\left| x_{\parallel} - \mu_{k,\parallel}^{vis} \right|, \ k \in \{c_1, c_2\} \tag{7}$$

The physical meaning of the verification score $s_k$: a larger value (closer to 0) indicates that the test image is closer to the visual prototype of category $k$ on the discriminative direction, providing stronger supporting evidence for that category; a smaller value (more negative) indicates a greater distance and weaker supporting evidence. The verification process of DVM is conducted in the discriminative subspace defined by text prototypes, combining the stability of text semantics with the discriminative power of visual features.

As shown in Fig. 3(a), the above projection verification process can be intuitively understood on a one-dimensional discriminative axis: the fine-grained text prototypes $\mu_{c_1}^{txt}$ and $\mu_{c_2}^{txt}$ define the two endpoints of the discriminative axis, and their connecting direction is the normalized discriminative direction $d$; the visual prototypes $\mu_{c_1}^{vis}$, $\mu_{c_2}^{vis}$ of candidate categories and the test feature $x$ are projected onto this axis, forming three scalar positions. The verification score $s_k$ is equivalent to the negative distance between the test projection $x_{\parallel}$ and the $\mu_{k,\parallel}^{vis}$ prototype projection of

category $k$: if the test projection is closer to $\mu_{c_1,\|}^{vis}$, then $s_{c_1} > s_{c_2}$ ), and verification supports the original top-1 prediction; otherwise, it supports the top-2 candidate. This design simplifies category attribution judgment in the high-dimensional feature space to position comparison on a one-dimensional semantic discriminative axis, leveraging both the semantic stability of text prototypes and the instance-level discriminative information of visual features.

The verification confidence is defined as the difference between the two scores:

$$\Delta s = s_{c_1} - s_{c_2} \tag{8}$$

$\Delta s \gg 0$ indicates that verification strongly supports the original top-1 prediction $c_1$; $\Delta s \approx 0$ indicates insufficient evidence and high classification uncertainty; $\Delta s \ll 0$ indicates that verification leans toward the top-2 candidate $c_2$.

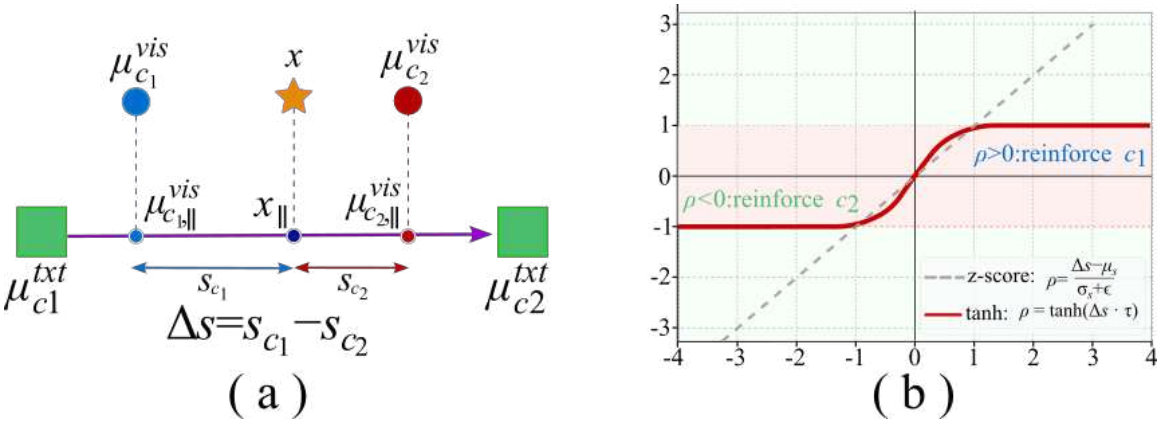



Fig. 3. Schematic illustration of the DVM principle. (a) DVM discriminative direction projection principle: constructing a discriminative axis based on fine-grained text prototypes, projecting test sample features and visual prototypes onto this direction to obtain verification scores; (b) comparing z-score unbounded mapping and *tanh* bounded mapping for generating correction signals $\rho$.

Large score differences (high confidence) trigger strong correction, while small score differences (low confidence) trigger weak correction.

*3.4 Diagnostic Confidence-Driven Correction (DCDC)*

The verification scores output by DVM provide diagnostic signals for ambiguity resolution, but how to safely and stably apply these signals to logit correction remains to be studied. This section proposes the Diagnostic Confidence-Driven Correction module DCDC, which takes the verification confidence $\Delta s$ output by DVM as its sole input, aiming to convert verification confidence into a bounded and controllable correction signal.

*3.4.1 From Unbounded z-score to Bounded tanh Mapping*

When converting verification scores to correction signals, traditional methods often employ z-score normalization:

$$\rho = \frac{\Delta s - \mu_s}{\sigma_s + \epsilon}, \tag{9}$$

where $\mu_s$ and $\sigma_s$ are the mean and standard deviation of the verification confidence $\Delta s$ across samples in the current test batch, respectively. This approach has two potential issues: (1) its value range is $(-\infty, +\infty)$, and when $\Delta s$ is large, it may produce correction signals with amplitude exceeding 1, leading to over-correction; (2) it depends on the stability of within-batch statistics, while in few-shot scenarios the batch sample size is limited, and the estimation variance of $\sigma_s$ is non-negligible, potentially amplifying the influence of outliers.

This paper proposes correction signal generation based on bounded *tanh* mapping:

$$\rho = \tanh(\Delta s \cdot \tau) \tag{10}$$

where $\tau > 0$ is a scaling hyperparameter controlling the steepness of the *tanh* mapping. Fig. 3(b) intuitively compares the behavioral differences between

z-score and *tanh* mappings in converting verification confidence $\Delta s$ to correction signal $\rho$. The z-score mapping (dashed line) has a value range of $(-\infty,+\infty)$; when $|\Delta s|$ is large, the correction signal amplitude can far exceed 1, and its slope depends on the standard deviation of $\Delta s$ within the batch $\sigma_s$, in few-shot scenarios with limited batch sample size, $\sigma_s$ has large estimation variance, potentially causing the same $\Delta s$ to produce significantly different correction strengths across batches. In contrast, the *tanh* mapping (solid line) strictly constrains the output to (-1, 1), when $|\Delta s|$ is small, *tanh* is approximately linear, and the correction signal is proportional to confidence; when $|\Delta s|$ increases, *tanh* gradually saturates, and the correction signal approaches $\pm 1$ rather than growing indefinitely. This saturation property fundamentally eliminates the risk of over-correction, and the *tanh* mapping depends only on the hyperparameter $\tau$, without involving any batch statistics, enabling DCDC to run completely independently per sample, suitable for practical deployment scenarios of single-image inference.

The correction signal $\rho$ has the following properties:

**1. Boundedness:** $\rho \in [-1, 1]$, the correction signal amplitude is strictly limited within a bounded interval, effectively suppressing the risk of over-correction.

**2. Confidence-awareness:** when $|\Delta s|$ is large (high verification confidence), $|\rho| \to 1$, and the correction signal reaches maximum strength; when $\Delta s \approx 0$ (low verification confidence), $\rho \to 0$, and the correction signal tends to zero. The nonlinear nature of the mapping creates a smooth adaptive relationship between correction strength and confidence.

**3. Direction preservation**: the sign of $\rho$ is consistent with $\Delta s$: $\rho > 0$

indicates reinforcing the top-1 prediction $c_1$, while $\rho < 0$ indicates correcting toward the top-2 prediction $c_2$ · *tanh* is a monotonically increasing function, ensuring consistency between the correction direction and the verification evidence direction.

**4. Smooth differentiability:** *tanh* is a smooth "S"-shaped curve, continuously differentiable over the entire domain, avoiding the gradient discontinuity problem caused by hard thresholds (such as the sign function), and facilitating future end-to-end training extensions.

*3.4.2 Logit Correction and Final Prediction*

After obtaining the correction signal $\rho$, symmetric correction is applied to the top-2 logits of the base classifier:

$$\tilde{l}_{c_1}=l_{c_1}+\lambda\rho,\ \tilde{l}_{c_2}=l_{c_2}-\lambda\rho \tag{11}$$

where $\lambda > 0$ is the global correction strength scaling factor, determined by grid search on the validation set; logits of other categories remain unchanged. Symmetric correction ensures the invariance of the sum of top-2 category logits ( $\tilde{l}_{c_1} + \tilde{l}_{c_2} = l_{c_1} + l_{c_2}$ ), maintaining the overall energy distribution of the classifier output. The final prediction is given by the argmax of the corrected logits:

$$\hat{y} = \arg\max_k \tilde{l}_k \tag{12}$$

When $\lambda = 0$, no correction is performed, and DiVe is equivalent to the base classifier. Since λ=0 is always a candidate value, model selection on the validation set will not underperform the base classifier; however, this property does not constitute a strict lower bound on the test set, and test performance depends on the distribution consistency between the validation and test sets.

*3.5 Algorithm Flow and Complexity Analysis*

The complete inference flow of DiVe is summarized in pseudocode form in Algorithm 1.

**Computational complexity analysis.** The additional operations introduced by DiVe are all per-sample vector operations: top-2 identification has complexity $O(K)$, discriminative direction construction has complexity $O(D)$, feature projection and verification scoring have complexity $O(D)$, and *tanh* mapping and logit correction have complexity $O(1)$. Therefore, the total additional computational complexity of DiVe per sample is $O(K+D)$, where $K$ is the number of categories. When the number of categories $K$ is large (ImageNet K=1000), the $O(K)$ overhead comes from finding the top-2.

**Algorithm 1 DiVe Reasoning Process**

**Input**: test image feature $x$, base logits $\boldsymbol{l}$, $\{\boldsymbol{\mu}_k^{txt}\}_{k=1}^{K}$, $\{\boldsymbol{\mu}_k^{vis}\}_{k=1}^{K}$, $\tau$, $\lambda$

$c_1 = \underset{k}{\operatorname{argmax}}\, l_k$, $c_2 = \underset{k \neq c_1}{\operatorname{argmax}}\, l_k$

$d \leftarrow \dfrac{\mu_{c_1}^{txt} - \mu_{c_2}^{txt}}{\|\mu_{c_1}^{txt} - \mu_{c_2}^{txt}\|_2}$

$x_{\|} \leftarrow x^T d$

$\mu_{k,\|}^{vis} \leftarrow \left(\mu_k^{vis}\right)^T d,\ k \in \{c_1, c_2\}$

$s_k \leftarrow -\left|x_{\|} - \mu_{k,\|}^{vis}\right|,\ k \in \{c_1, c_2\}$

$\rho \leftarrow \tanh\left((s_{c_1} - s_{c_2}) \cdot \tau\right)$

$\tilde{l}_{c_1} \leftarrow l_{c_1} + \lambda\rho$, $\tilde{l}_{c_2} \leftarrow l_{c_2} - \lambda\rho$

$\tilde{l}_k \leftarrow l_k$ for $k \notin \{c_1, c_2\}$

$\hat{y} \leftarrow \underset{k}{\operatorname{argmax}}\, \tilde{l}_k$

return $\tilde{\boldsymbol{l}}, \hat{y}$

**Storage and preprocessing overhead.** Additional storage is required for fine-grained text prototypes of each category $\{\mu_k^{txt}\}_{k=1}^{K}$ ($K\times D$ dimensions), which can be computed once and cached during the dataset preparation phase; visual prototypes $\{\mu_k^{vis}\}_{k=1}^{K}$ ($K\times D$ dimensions) are a necessary component of the GDA base classifier and require no additional storage.

# 4 Experiments

## *4.1 Experimental Setup*

### *4.1.1Datasets*

Following the standard evaluation protocols of GDA-CLIP [11] and TIMO [12], this paper conducts few-shot classification experiments on 11 image classification datasets, covering general object recognition ImageNet [24], fine-grained classification Stanford Cars [25], Oxford Flowers [26], FGVC Aircraft [27], and Oxford Pets [28], scene recognition SUN397 [29], texture classification DTD [30], food classification Food101 [31], remote sensing image classification EuroSAT [32], action recognition UCF101 [33], and general object recognition Caltech101 [34], among other task types. Cross-dataset generalization evaluation is conducted on 4 variants of ImageNet: ImageNet-V2 [35] (newly collected test set), ImageNet-Sketch [36] (hand-drawn sketches), ImageNet-A [37] (natural adversarial examples), and ImageNet-R [38] (rendered/artistic versions), to evaluate robustness under distribution shift.

### *4.1.2 Implementation Details*

This paper defaults to using ResNet-50 as the CLIP visual backbone, results for ResNet-101 , ViT-B/32, and ViT-B/16 are also reported to verify backbone-agnosticism. The main experiments adopt the 16-shot setting with 16 annotated samples per class, and each setting is run independently with 3

different random seeds, reporting average accuracy. The base GDA classifier uses an empirical Bayes ridge estimator for inversion, with regularization strength searched on the validation set. Fine-grained text prototypes are computed from CuPL [23] descriptive prompts, with 20-50 descriptive texts generated per class. Correction strength $\lambda \in [0, 0.1, 0.2, 0.4, 0.6, 0.8, 1.0, 2.0, 4.0, 8.0]$, and *tanh* scale $\tau \in [1.0, 2.0, 4.0, 8.0]$. All experiments are completed on a single NVIDIA RTX3090 GPU.

### *4.1.3 Comparison Methods*

This paper compares with two categories of methods: (1) Training-free methods: Zero-Shot CLIP [1], Linear-probe [1], CALIP [39], Tip-Adapter [10], Tip-X [15], APE [16], GDA-CLIP [11], TIMO [12]; (2) Training-required methods: Linear-probe, CoOp [5], CLIP-Adapter [6], Tip-Adapter-F [7], APE-T [16], and CLAP [14]. Results for all comparison methods are from their original papers or reproduced by this paper under a unified protocol, to ensure fair comparison.

## *4.2 Main Results*

Table 1 and Fig. 4 present the performance of DiVe relative to training-free and training-required baseline methods on 11 datasets under the ResNet-50 backbone, 16-shot setting. The following observations can be made:

Table 1. Performance (%) of few-shot classification on 11 datasets, showing Top-1 accuracy averaged over 3 random seeds, √ denotes training-required,× denotes training-free.

| **Methods** | **Train** | **Pets** | **Flowers** | **FGVC** | **DTD** | **EuroSAT** | **Cars** | **Food** | **SUN** | **Caltech** | **UCF** | **ImageNet** | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Linear-probe | √ | 76.42 | 94.95 | 36.39 | 63.97 | 82.76 | 70.08 | 70.17 | 67.15 | 90.63 | 73.72 | 55.87 | 71.10 |
| CoOp | √ | 87.01 | 94.51 | 31.26 | 63.58 | 83.53 | 73.36 | 74.67 | 69.26 | 91.83 | 75.71 | 62.95 | 73.42 |
| CLIP-Adapter | √ | 87.84 | 93.90 | 32.10 | 65.96 | 84.43 | 74.01 | 75.25 | 69.55 | 92.49 | 76.76 | 63.59 | 74.44 |
| Tip-Adapter-F | √ | 89.70 | 94.80 | 35.55 | 66.55 | 84.54 | **75.74** | <u>79.43</u> | 71.47 | 92.86 | <u>78.03</u> | **65.51** | 75.83 |
| APE-T | √ | 88.48 | 94.82 | 31.27 | 66.84 | 81.87 | 67.87 | 79.41 | 71.52 | 92.56 | 76.05 | <u>65.16</u> | 74.17 |
| CLAP | √ | 87.43 | 86.10 | 36.53 | **68.50** | 85.25 | 65.09 | 77.22 | **71.78** | <u>93.39</u> | **78.14** | 65.09 | 75.97 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Zero-Shot CLIP | × | 85.69 | 66.02 | 17.04 | 42.85 | 36.27 | 55.79 | 77.39 | 58.84 | 85.72 | 60.38 | 60.32 | 58.76 |
| CALIP | × | 86.21 | 66.38 | 17.76 | 42.39 | 38.90 | 56.27 | 77.42 | 58.95 | 87.71 | 61.72 | 60.57 | 59.48 |
| Tip-Adapter | × | 87.60 | 89.47 | 29.24 | 58.85 | 72.27 | 66.26 | 77.88 | 66.94 | 90.40 | 68.02 | 61.41 | 69.94 |
| Tip-X | × | 88.91 | 90.16 | 29.67 | 61.35 | 76.89 | 67.67 | 77.91 | 67.66 | 90.67 | 69.85 | 62.61 | 71.21 |
| APE | × | 88.01 | 89.95 | 28.18 | 64.62 | 76.30 | 66.94 | 78.54 | 67.66 | 91.39 | 70.99 | 62.18 | 71.34 |
| GDA-CLIP | × | 88.89 | 95.66 | 40.24 | 66.02 | 86.03 | 75.15 | 79.04 | 70.77 | 92.63 | 76.02 | 63.65 | 75.83 |
| TIMO | × | 89.79 | 95.85 | 40.01 | 67.67 | 86.05 | 74.07 | 77.88 | 71.61 | **93.48** | 76.84 | 64.14 | 76.13 |
| DiVe(ours) | × | **90.19** | **95.93** | **40.49** | 66.77 | **86.07** | 75.65 | **79.58** | 71.60 | 92.66 | 77.00 | 64.82 | **76.43** |

DiVe achieves an average accuracy of 76.43% on 11 datasets, the best average among training-free methods. It should be noted that DiVe does not achieve the best performance on all datasets: on DTD it is 66.77%, lower than TIMO's 67.67%; on SUN397 it is 71.60%, essentially tied with TIMO's 71.61%; on Caltech101 it is 92.66%, lower than TIMO's 93.48%. This indicates that the top-2 text discriminative axis is not yet sufficient to fully replace the global fusion information of the base classifier on categories with large intra-class variation and limited text semantic discriminative power, such as textures and scenes, and also points out directions for future improvement.

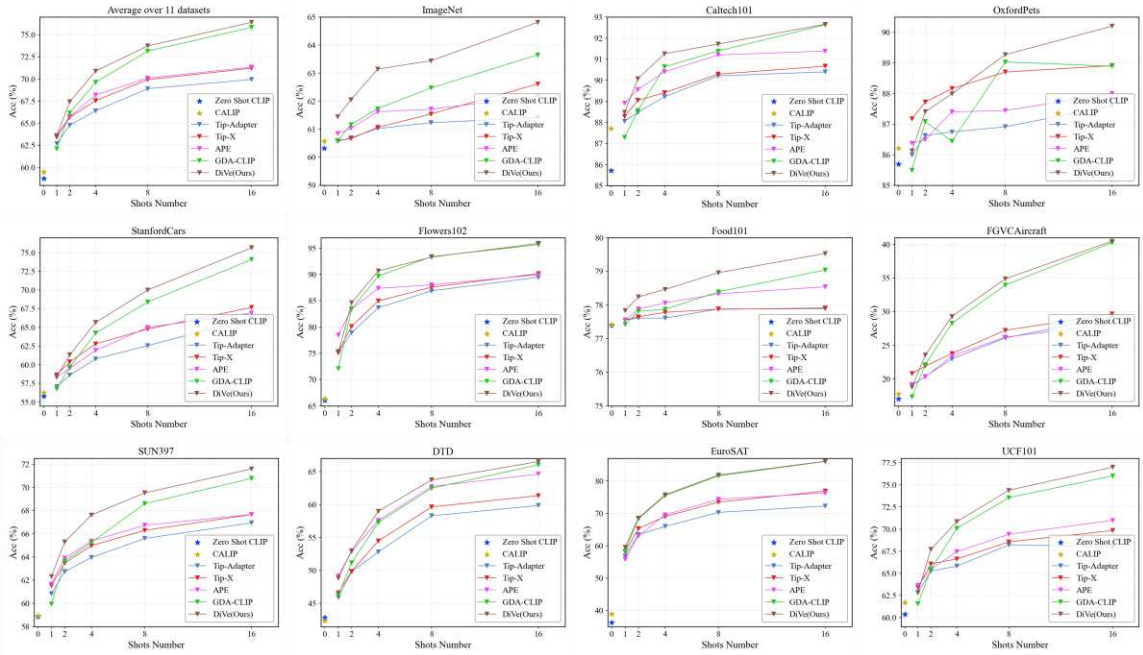


Fig. 4. Results of few-shot classification experiments on 11 datasets. We compare the proposed method with 6 training-free methods under the ResNet-50 CLIP model across sample sizes of 1, 2, 4, 8, and 16.

### *4.3 Results with Different Backbone Networks*

To verify the backbone-agnosticism of DiVe, Table 2 reports results for ResNet-101, ViT-B/32, and ViT-B/16 backbones under 16-shot. On all three backbones, DiVe consistently improves the performance of the base GDA-CLIP and surpasses all competing training-free methods: on ResNet-101 it reaches 78.27%, surpassing GDA-CLIP (78.04%) and TIMO (77.67%); on ViT-B/32 it reaches 78.29%, surpassing the best competing method TIMO (78.12%); on ViT-B/16 it reaches 82.01%, achieving the best among training-free methods and surpassing TIMO (81.96%). These results demonstrate that the diagnostic verification and bounded correction mechanism of DiVe does not depend on a specific visual feature extractor and can stably function in feature spaces of different architectures and dimensions.

Table 2. Average accuracy (%) of few-shot classification using different visual encoders on 11 datasets, averaged over 3 random seeds.

| Methods | RN50 | RN101 | ViT-B/32 | ViT-B/16 | Average |
|---|---|---|---|---|---|
| Tip-Adapter | 69.94 | 71.73 | 72.25 | 77.09 | 72.75 |
| CoOp | 73.42 | 75.96 | 75.7 | 79.71 | 76.2 |
| APE | 71.34 | 72.31 | 73.91 | 78.13 | 73.92 |
| GDA-CLIP | 75.83 | <u>78.04</u> | 78.01 | 81.8 | 78.42 |
| TIMO | <u>76.13</u> | 77.67 | <u>78.12</u> | <u>81.96</u> | <u>78.47</u> |
| DiVe(ours) | **76.43** | **78.27** | **78.29** | **82.01** | **78.74** |

### *4.4 Cross-Dataset Generalization*

To evaluate the robustness and generalization ability of DiVe under distribution shift, this paper conducts cross-dataset generalization

experiments: the model constructs a classifier at 16-shot on the ImageNet support set, without any adaptation or fine-tuning, and is directly tested on four variants: ImageNet-V2, ImageNet-Sketch, ImageNet-A, and ImageNet-R. Results are shown in Table 3.

DiVe achieves an average accuracy of 44.36% on the four target variants, surpassing TIMO by 0.27% and GDA-CLIP by 0.66%, respectively. By variant, DiVe achieves the best results among training-free methods on ImageNet-V2, ImageNet-Sketch, and ImageNet-R (56.36%, 36.21%, and 61.04%, respectively); among them, ImageNet-R contains a large number of rendered, cartoon, and artistic style images with significant domain shift from natural images, and the stable gain of DiVe on it indicates that the discriminative direction verification mechanism based on text prototypes can better cope with visual domain shift — text semantics are relatively stable across different visual styles, and thus the constructed discriminative direction has a certain degree of domain invariance. It should be objectively noted that on ImageNet-A, DiVe achieves 23.82%, slightly lower than APE (24.48%) and TIMO (24.08%); on ImageNet-Sketch, the gain relative to TIMO is only 0.05%. This indicates that under extreme domain shifts such as natural adversarial examples and hand-drawn sketches, the alignment between text prototypes and visual features may still fail, and the domain invariance of the discriminative direction is conditional.

Table 3. Cross-dataset generalization results (%) on ImageNet variants under ResNet-50 16-shot, averaged over 3 random seeds. The classifier is constructed from the ImageNet 16-shot support set without training, then directly tested on target datasets.

| Methods | Source | Target | | | | |
|---|---|---|---|---|---|---|
| | ImageNet | -V2 | -Sketch | -A | -R | Average |
| CLIP | 60.32 | 53.41 | 35.48 | 23.76 | 60.39 | 43.26 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Tip-Adapter | 61.42 | 54.12 | 35.73 | 23.78 | 60.38 | 43.5 |
| APE | 62.18 | 54.49 | 36.14 | **24.48** | 60.71 | 43.96 |
| GDA-CLIP | 63.65 | 54.8 | 35.82 | 23.75 | 60.41 | 43.7 |
| TIMO | 64.14 | 55.97 | 36.16 | 24.08 | 60.14 | 44.09 |
| DiVe(ours) | **64.81** | **56.36** | **36.21** | 23.82 | **61.04** | **44.36** |

### *4.5 Ablation Study*

#### *4.5.1 Component Analysis*

Table 4 compares three configurations: (1) base classifier; (2) +DVM (using z-score normalization correction); (3) +DVM+DCDC (full method, *tanh* mapping). Since DCDC takes the verification score output by DVM as its sole input, using DCDC alone does not generate correction signals, so the core comparison is the incremental difference between z-score and *tanh* mappings. Experiments are conducted on six representative datasets: ImageNet, Food101, Oxford Pets, DTD, SUN397, and UCF101.

Table 4. Ablation study (%) under ResNet-50 16-shot. Evaluating the contributions of DVM and DCDC on three base classifiers.

| Base Classifier | Setting | ImageNet | Food | Pets | DTD | SUN | UCF | Avg. | Δ |
|---|---|---|---|---|---|---|---|---|---|
| TIP-Adapter | — | 61.41 | 77.88 | 87.60 | 59.85 | 66.94 | 68.02 | 70.28 | - |
| | +DVM(z-score) | 61.69 | 78.61 | 88.37 | 60.87 | 68.29 | 69.09 | 71.15 | +0.87 |
| | +DVM+DCDC | 61.83 | 78.74 | 88.73 | 61.33 | 68.52 | 69.92 | 71.51 | +1.23 |
| GDA-CLIP | — | 63.65 | 79.04 | 88.89 | 66.02 | 70.77 | 76.02 | 74.07 | - |
| | +DVM(z-score) | 64.49 | 79.47 | 88.92 | 66.45 | 71.26 | 76.82 | 74.57 | +0.5 |
| | +DVM+DCDC | 64.82 | 79.58 | 90.19 | 66.77 | 71.60 | 77.00 | 74.99 | +0.92 |
| TIMO | — | 64.14 | 77.88 | 89.79 | 67.67 | 71.61 | 76.84 | 74.66 | - |
| | +DVM(z-score) | 64.22 | 77.95 | 90.18 | 67.81 | 71.82 | 77.38 | 74.89 | +0.23 |
| | +DVM+DCDC | 64.52 | 78.47 | 90.33 | 68.32 | 71.95 | 77.42 | 75.17 | +0.51 |

#### *4.5.2 General Plugin Capability Verification*

One finding of the ablation study is that DVM+DCDC, as a general post-processing plugin, provides stable gains on all three base classifiers, and does not degrade base classifier performance under any configuration: when

applied to Tip-Adapter, the average improvement is +1.23%; when applied to GDA-CLIP, it is +0.92%; when applied to TIMO, it is +0.51%. The difference in improvement magnitude reflects the difference in "room for improvement" across base classifiers: Tip-Adapter uses the simplest non-parametric key-value cache classifier, lacking statistical modeling, with the weakest ambiguity resolution ability and the largest post-processing correction room; GDA-CLIP is based on Gaussian discriminant analysis, with stronger statistical modeling ability and moderate room for improvement; TIMO already incorporates a text-image mutual guidance mechanism, with the highest classifier quality and relatively smaller exploitable gains. Further comparing the two mappings shows that bounded *tanh* brings additional stable increments of +0.36% (Tip-Adapter: 71.15→71.51), +0.42% (GDA-CLIP: 74.57→74.99), and +0.28% (TIMO: 74.89→75.17) over unbounded z-score for the three base classifiers, respectively, with no performance regression on any dataset, verifying the necessity of replacing z-score normalization with bounded mapping. This result validates the general effectiveness of DVM+DCDC: the performance gain from post-processing is negatively correlated with the ambiguity resolution ability of the base classifier itself.

*4.5.3 Impact of Discriminative Direction Selection*

One of the core designs of DVM is to construct the discriminative direction using the fine-grained text prototype difference, rather than directly using the visual class prototype difference. This section quantitatively verifies the rationality of this choice through ablation experiments. We compare three discriminative direction construction methods:

(1) Visual prototype difference ( $d_{vis}$ ): $d_{vis} = normalize(\mu_{c_1}^{vis} - \mu_{c_2}^{vis})$ , directly constructed from the difference of support sample feature means of candidate categories;

(2) Text-visual fusion difference ( $d_{fuse}$ ): $d_{fuse} = normalize(\mu_{c_1}^{txt} - \mu_{c_2}^{txt} + \mu_{c_1}^{vis} - \mu_{c_2}^{vis})$, summing the two directions with equal weights and then normalizing;

(3) Text prototype difference ( $d_{text}$ ): $d_{text} = normalize(\mu_{c_1}^{txt} - \mu_{c_2}^{txt})$, constructed from the difference of CuPL fine-grained text prototypes.

All three configurations share exactly the same DCDC correction module and hyperparameters, with only the discriminative direction d replaced. Experiments are conducted on six datasets: ImageNet, Food101, Oxford Pets, DTD, SUN397, and UCF101 (ResNet-50, 16-shot), and results are shown in Table 5.

Table 5. Impact of discriminative direction selection on classification performance (%). Comparing accuracy of visual prototype difference, text prototype difference (ours), and text-visual fusion under ResNet-50 16-shot.

| Discrimination Direction | ImageNet | Food | Pets | DTD | SUN | UCF | Avg. |
|---|---|---|---|---|---|---|---|
| $d_{vis}$ | 63.66 | 79.11 | 88.8 | 66.27 | 70.76 | 76.02 | 74.10 |
| $d_{fuse}$ | 63.76 | 79.18 | 89.16 | 66.65 | 70.90 | 76.13 | 74.30 |
| $d_{text}$ | **64.82** | **79.58** | **90.19** | **66.77** | **71.60** | **77.00** | **74.99** |

The following observations can be made from Table 5:

(1) Text prototype difference outperforms visual prototype difference. In terms of overall average accuracy, $d_{text}$ improves by +0.89 over $d_{vis}$ , confirming the core motivation of this paper: under few-shot conditions, visual prototypes are estimated from only *Ns*=16 support samples, with high estimation variance; when support samples are not representative of the

category, the visual prototype difference cannot stably characterize the true inter-category decision boundary; in contrast, text prototypes are aggregated from dozens of descriptive prompts, independent of scarce support images, and are more stably estimated under few-shot conditions.

(2) Text-visual fusion does not bring further gains, but is slightly lower than pure text prototype difference. The average accuracy of $d_{fuse}$ is 74.30%, lower than 74.99% of $d_{text}$. This indicates that under the few-shot setting, the noise contained in the visual prototype difference (introduced by support sample estimation variance) has a negative impact on the fusion direction, offsetting the supplementary information it might provide. This result supports the design choice of "using text prototype difference as the sole discriminative direction" in this paper, and also shows that simple equal-weight fusion is not an effective way to utilize visual prototype information.

### *4.6 Qualitative Analysis and Discussion*

To understand the confidence-aware correction behavior of DCDC, this paper conducts a statistical analysis of the distribution of the correction signal $\rho$ on the ImageNet dataset (ResNet-50, 16-shot), and the results are shown in Fig. 5(a). For erroneous samples corrected by DCDC, the distribution of the correction signal $\rho$ is mainly concentrated in the negative region (mean approximately $-0.23$), indicating that DCDC can identify erroneous predictions and correct toward the correct category (top-2); for still-erroneous samples, the distribution of $\rho$ leans toward positive values (mean approximately +0.13), indicating that the model reinforces the erroneous top-1 prediction. Whether for correct or erroneous predictions, the distribution of the correction signal $\rho$ is concentrated within the interval $[-0.75,+0.75]$, with extreme values close to $\pm 1$ rarely occurring, which

intuitively verifies the effectiveness of the *tanh* bounded mapping in suppressing over-correction.

Fig. 5(b) further analyzes the performance gain of DiVe on samples with different degrees of ambiguity. The horizontal axis is the base classifier top-2 logit margin (smaller margin indicates higher ambiguity), and the vertical axis is classification accuracy. In the region with the highest ambiguity (logit margin close to 0), DiVe has the largest gain relative to the base GDA-CLIP (exceeding 5%); as ambiguity decreases (logit margin increases), the gain gradually diminishes. This result confirms the core motivation of this paper: the DVM+DCDC framework is designed for ambiguous samples, exerting the greatest effect on the most ambiguous samples while adopting a conservative correction strategy for easily classifiable samples to avoid unnecessary interference. This adaptive on-demand correction property is the core advantage of DiVe over the global fusion paradigm. In addition, the applicable boundaries of this method should be noted: first, DVM only verifies and corrects between top-2 candidates, and when the true category falls outside top-2, it cannot be recovered through this mechanism, which is also the direct motivation for future extension to top-k verification; second, the effectiveness of the discriminative direction depends on the alignment quality between CLIP text and visual spaces, with limited gains on categories that are difficult to distinguish by text semantics.

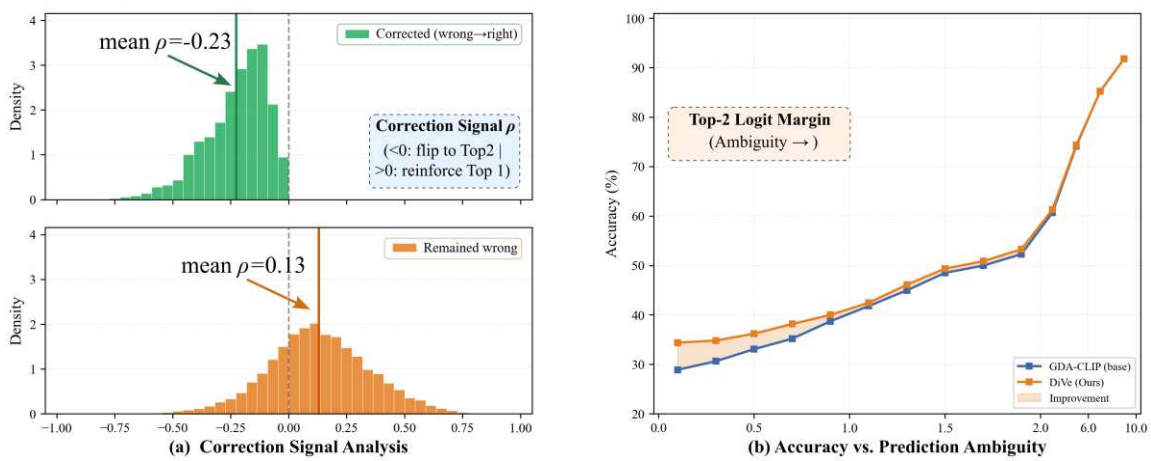



Fig. 5. Qualitative analysis of correction behavior. (a) Distribution of correction signals $\rho$ for corrected erroneous samples and remaining erroneous samples on ImageNet. Corrections for correct predictions are mainly positive (reinforcement), while corrections for erroneous predictions are more likely negative (correction), and the bounded *tanh* mapping prevents extreme corrections. (b) Accuracy changes with prediction ambiguity (top-2 logit margin), showing that DiVe gains more on ambiguous samples.

## 5 Conclusion and Future Work

This paper proposes DiVe, a training-free few-shot image classification framework that alleviates the problem in existing training-free methods where the global fusion paradigm struggles to handle sample-level ambiguity, through the Diagnostic Verification Module (DVM) and the Diagnostic Confidence-Driven Correction (DCDC) module. As the core of the framework, DVM achieves targeted ambiguity resolution by explicitly identifying top-2 ambiguous candidates, constructing a discriminative direction using fine-grained text prototypes, and performing feature projection verification. As the enabling mechanism of DVM, DCDC constrains verification confidence to a bounded interval through bounded *tanh* mapping, achieving strong correction at high confidence and conservative correction at low confidence, reducing the risk of over-

correction. DVM and DCDC form a serially dependent system that works in concert to achieve full effectiveness.

The proposed method is completely training-free, with an additional per-sample overhead of O($K$+$D$), and can be integrated as a lightweight post-processing plugin into various mainstream training-free classifiers. Experiments on 11 standard datasets with 4 CLIP backbones under 16-shot show that DiVe achieves the best average performance among training-free methods, with an average accuracy of 76.43% on ResNet-50 16-shot; cross-dataset generalization experiments verify its robustness under distribution shift; ablation experiments and qualitative analysis reveal the contributions of each component and the working principle of the correction mechanism.

Future work will proceed in the following directions: first, extending the current top-2 candidate verification mechanism to top-k multi-candidate verification to handle complex samples where multiple categories are simultaneously highly confused; second, exploring the combination of the DVM+DCDC framework with training-based methods (such as CoOp [5], CLIP-Adapter [6]) to verify whether end-to-end training can further improve performance; third, extending the diagnostic verification mechanism to other visual tasks such as object detection and semantic segmentation, verifying the universality of cognitively inspired ambiguity resolution in broader visual problems; fourth, researching adaptive strategies based on sample ambiguity degree, dynamically adjusting the verification candidate range and correction strength to further balance inference efficiency and classification performance.

**Funding Declaration:** This research was conducted without external funding.

## References


1. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning, 8748–8763 (2021)
2. Jia, C., Yang, Y., Xia, Y., et al.: Scaling up visual and vision-language representation learning with noisy text supervision. In: Proceedings of the 38th International Conference on Machine Learning, 4904–4916 (2021)
3. Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., Wierstra, D.: Matching networks for one shot learning. In: Advances in Neural Information Processing Systems, 3630–3638 (2016)
4. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning, 1126–1135 (2017)
5. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. Int. J. Comput. Vis. 130(9): 2337–2348 (2022)
6. Gao, P., Geng, S., Zhang, R., et al.: CLIP-Adapter: Better vision-language models with feature adapters. Int. J. Comput. Vis. 131(2): 419–436 (2023)
7. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 16816–16825 (2022)
8. Jia, M., Tang, L., Chen, B.-C., et al.: Visual prompt tuning. In: Proceedings of the European Conference on Computer Vision, 709–727 (2022)
9. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: MaPLe: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 19113–19122 (2023)
10. Zhang, R., Fang, R., Gao, P., et al.: Tip-Adapter: Training-free adaption of CLIP for few-shot classification. In: Proceedings of the European Conference on Computer Vision, 493–510 (2022)
11. Wang, Z., Liang, J., Sheng, L., He, R., Wang, Z., Tan, T.: A hard-to-beat baseline for training-free CLIP-based adaptation. In: Proceedings of the International Conference on Learning Representations (2024)
12. Li, Y., Guo, J., Qi, L., Li, W., Shi, Y.: Text and image are mutually beneficial: Enhancing training-free few-shot classification with CLIP. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 39(5): 5039–5047 (2025)
13. Shu, M., Nie, W., Huang, D.-A., et al.: Test-time prompt tuning for zero-shot generalization in vision-language models. In: Advances in Neural Information Processing Systems, 14274–14289 (2022)
14. Silva-Rodríguez, 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, 23681–23690 (2024)

15. Udandarao, V., Gupta, A., Albanie, S.: SuS-X: Training-free name-only transfer of vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2725–2736 (2023)
16. Zhu, X., Zhang, R., He, B., Zhou, A., Wang, D., Zhao, B., Gao, P.: Not all features matter: Enhancing few-shot CLIP with adaptive prior refinement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2605–2615 (2023)
17. 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 Conference on Computer Vision and Pattern Recognition, 25092–25102 (2025)
18. Palmer, S.E.: Vision Science: Photons to Phenomenology. MIT Press, Cambridge, MA (1999)
19. Kourtzi, Z., Kanwisher, N.: The representation of object perception in human occipitotemporal cortex. J. Cogn. Neurosci. 12(3): 465–475 (2000)
20. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, 5998–6008 (2017)
21. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th International Conference on Machine Learning, 41–48 (2009)
22. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems, 4077–4087 (2017)
23. Pratt, S., Covert, I., Liu, R., Farhadi, A.: What does a platypus look like? Generating customized prompts for zero-shot image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 15691–15701 (2023)
24. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 248–255 (2009)
25. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3D object representations for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, 554–561 (2013)
26. Nilsback, M.-E., Zisserman, A.: Automated flower classification over a large number of classes. In: Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing, 722–729 (2008)
27. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
28. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.V.: Cats and dogs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3498–3505 (2012)
29. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: SUN database: Large-scale scene recognition from abbey to yard. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3485–3492 (2010)

30. 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, 3606–3613 (2014)
31. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 – mining discriminative components with random forests. In: Proceedings of the European Conference on Computer Vision, 446–461 (2014)
32. 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. 12(7): 2217–2226 (2019)
33. Soomro, K., Zamir, A.R., Shah, M.: UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
34. 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 178–186 (2004)
35. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do ImageNet classifiers generalize to ImageNet? In: Proceedings of the 36th International Conference on Machine Learning, 5389–5400 (2019)
36. Wang, H., Ge, S., Xing, E.P., Lipton, Z.C.: Learning robust global representations by penalizing local predictive power. In: Advances in Neural Information Processing Systems (2019)
37. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 15262–15271 (2021)
38. Hendrycks, D., Basart, S., Mu, N., et al.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 8340–8349 (2021)
39. Guo, Z., Zhang, R., Qiu, L., Ma, X., Miao, X., He, X., Cui, B.: CALIP: Zero-shot enhancement of CLIP with parameter-free attention. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 37(1): 746–754 (2023)