BEYOND PROTOTYPE: META-PROMPTABLE MAHALANOBIS MATCHING FOR CROSS-DOMAIN FEW-SHOT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-domain few-shot learning (CD-FSL) aims to transfer knowledge from a label-rich source domain to novel target domains with only a few labeled examples. Across segmentation, classification, and detection, most existing methods still rely on prototype matching, where each class is represented by a single vector and queries are assigned by Euclidean or cosine distance. This paradigm implicitly treats class-conditional features as isotropic, an assumption that is often violated under severe domain shift: target-domain features commonly exhibit anisotropic covariance structures, large intra-class variation, and class-dependent uncertainty. As a result, nearest-prototype decision rules can produce poorly calibrated boundaries and limited cross-domain generalization. We propose a meta-promptable Mahalanobis matching framework that replaces isotropic prototypes with prompt-conditioned anisotropic Gaussian class models. Given a few support examples and class-specific instructions, large foundation models generate semantic prompts that summarize category concepts, domain context, and variation cues. Conditioned on these prompts and support features, an episodically trained meta-learner estimates class-wise means and covariance matrices, using a low-rank-plus-diagonal factorization to remain stable in the few-shot regime. Query assignment is then performed by Mahalanobis distance, enabling task-adaptive matching that accounts for both feature location and direction-dependent uncertainty. Theoretically, we show that the resulting hypothesis space strictly contains prototype matching as a special case and yields lower Bayes risk when target features are anisotropic. Empirically, our framework achieves consistent state-of-the-art performance on fourteen cross-domain few-shot benchmarks, including four segmentation datasets, four classification datasets, and six detection datasets, demonstrating that distribution-aware matching provides a unified and effective alternative to prototype-based CD-FSL. Code will be made publicly available.

1 INTRODUCTION

Cross-domain few-shot learning (CD-FSL) aims to recognize novel target-domain categories with only a handful of labeled examples, after learning from a label-rich source domain. It is a practical and challenging setting for visual intelligence, since many real applications, such as medical imaging, remote sensing, underwater perception, and industrial inspection, differ substantially from the natural-image domains on which modern models are commonly trained. This problem has been studied across image classification Zou et al. (2025), semantic segmentation He et al. (2024), and object detection Fu et al. (2025). Despite their different prediction formats, these tasks are largely built upon the same prototype-based principle: each class is compressed into a single vector from the support set, and each query image, pixel, or region proposal is assigned to the nearest prototype under cosine or Euclidean distance.

The success of prototype matching has made it a default choice in few-shot recognition, but this convenience hides a restrictive geometric assumption. A prototype represents a class only by its center, implicitly treating the corresponding feature distribution as isotropic: all feature directions are assumed to have comparable uncertainty, correlations among dimensions are ignored, and the same distance geometry is used for all classes and domains. Such an assumption is especially fragile

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

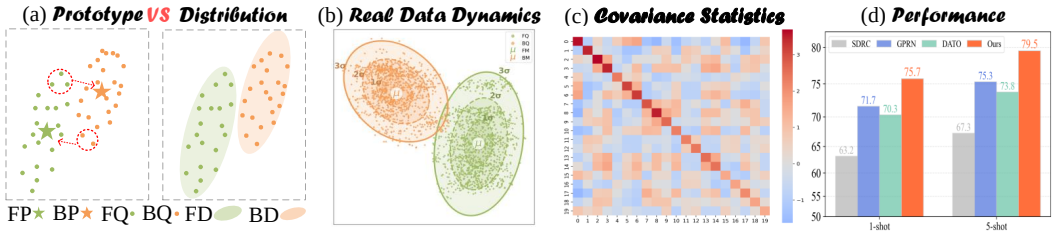


Figure 1: Motivation and empirical evidence on cross-domain few-shot segmentation (CD-FSS). (a) Prototype-based methods represent each class with a single prototype and implicitly assume isotropic Gaussian feature distributions, which may misclassify queries under domain shift. In contrast, our method models class-wise anisotropic Gaussian distributions for distribution-aware matching. P, F, B, Q, and D denote prototype, foreground, background, query, and feature distribution, respectively. (b) Feature visualization on a randomly sampled task shows clear anisotropic foreground/background distributions, while our estimated distributions better cover the corresponding query features. (c) Covariance heatmaps exhibit pronounced off-diagonal correlations, further confirming the anisotropic structure of target-domain features. (d) Average results on four CD-FSS benchmarks show consistent gains over prototype-based methods.

in CD-FSL. Under severe domain shift, target-domain features are often elongated, correlated, and class-dependent; the reliable evidence provided by a few support samples may reveal the class center, but not the full shape of its distribution. Thus, a query may be close to a prototype yet unlikely under the true class distribution, causing poorly calibrated boundaries and cross-domain errors.

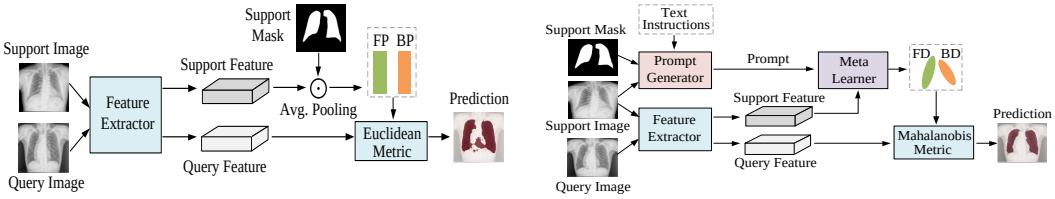
This limitation is not merely conceptual. As illustrated in Fig. 1, target-domain foreground and background features extracted by a pre-trained backbone on DeepGlobe Demir et al. (2018) exhibit strong anisotropic structures. Their covariance matrices contain clear off-diagonal correlations, indicating that the target feature geometry cannot be faithfully captured by spherical prototypes. Similar patterns are observed across classification and detection tasks. These observations suggest that the bottleneck of CD-FSL is not only how to obtain better features, but also how to match queries to support classes under a more faithful distributional geometry.

A natural remedy is to replace point prototypes with class-conditional distributions and use Mahalanobis distance for matching. However, directly estimating full covariance matrices from few-shot support examples is statistically unstable and computationally expensive in high-dimensional deep feature spaces. This creates a central dilemma: CD-FSL needs anisotropic class models to reduce the bias of prototype matching, yet the target support set is too small to estimate such models reliably from visual samples alone. Resolving this dilemma requires additional priors that are informative, transferable, and usable under scarce supervision.

We address this challenge with a meta-promptable Mahalanobis matching framework. The core idea is to use semantic knowledge from large foundation models (LFMs) Islam & Moushi (2025) as a distributional prior, while retaining visual grounding from the support examples. Given support images and class-specific textual instructions, we design a chain-of-thought (CoT) prompt generation strategy that elicits structured descriptions of category semantics, domain context, and intra-class variation. These descriptions are encoded into semantic prompts and supplied to an episodically trained meta-learner, which estimates class-wise anisotropic Gaussian distributions for target tasks.

The proposed meta-learner makes distribution estimation feasible in the few-shot regime. It first estimates class means through cross-modal fusion between support visual features and domain-aware semantic prompts, producing centers that are both visually grounded and semantically informed. It then estimates covariance matrices using a low-rank-plus-diagonal factorization, where prompt-guided variation cues help model direction-dependent uncertainty while the structured parameterization avoids the instability of unconstrained full covariance prediction. Query assignment is finally performed by Mahalanobis distance, allowing the decision rule to account for both feature location and the anisotropic uncertainty of each class. This formulation is task-agnostic: pixels in segmentation, image embeddings in classification, and region proposals in detection can all be matched to class distributions under the same principle.

We further support the proposed framework with theoretical and empirical evidence. Theoretically, we show that prototype matching is a special case of our hypothesis space, corresponding to isotropic covariance, and that under anisotropic target-domain feature distributions the Mahalanobis decision rule can achieve lower Bayes risk than prototype-based classifiers. Empirically, we evaluate our



(a) Existing prototype matching. (b) Our meta-promptable Mahalanobis matching.

Figure 2: Comparison between prototype matching and our distribution-aware framework, illustrated with CD-FSS. FP/BP denote foreground/background prototypes, and FD/BD denote foreground/background distributions. (a) Prototype-based methods compress each class into a single vector and perform nearest-prototype matching, implicitly assuming isotropic feature distributions. (b) Our method replaces point prototypes with prompt-conditioned anisotropic Gaussian distributions and performs Mahalanobis matching for cross-domain prediction.

method on fourteen cross-domain few-shot benchmarks covering four segmentation datasets, four classification datasets, and six detection datasets. The consistent improvements over state-of-the-art (SoTA) methods demonstrate that distribution-aware matching provides a unified and effective alternative to the conventional prototype paradigm.

In summary, we make three contributions. First, we identify the geometric limitation of prototype-based CD-FSL and replace isotropic point prototypes with anisotropic Gaussian class distributions, enabling more faithful modeling of target-domain feature geometry under domain shift. Second, we propose a meta-promptable Mahalanobis matching framework, where CoT-generated semantic prompts provide category, domain, and variation priors, and an episodically trained meta-learner estimates class-wise means and covariances through cross-modal fusion and low-rank factorization. Third, we show that this distribution-aware principle provides a unified solution for segmentation, classification, and detection, with theoretical guarantees over prototype matching and new SoTA results on fourteen CD-FSL benchmarks.

2 METHODOLOGY

Problem Definition. CD-FSL transfers knowledge from a source domain $\mathcal{D}_s$ to unseen target domains $\mathcal{D}_t$ under domain shift and disjoint label spaces, using only a few labeled support examples Guo et al. (2020). In a standard task, a support set $\mathcal{S} = \{(x_s^k, y_s^k)\}_{k=1}^{\mathcal{K}}$ contains $\mathcal{K}$ labeled target-domain examples, and the goal is to predict the label of a query sample x_q, which can be an image in classification, a pixel in segmentation, or a region proposal in detection. In the following, we present the method using CD-FSS as the main example, while the same formulation naturally extends to cross-domain few-shot classification (CD-FSC) and detection (CD-FSD). Standard CD-FSS benchmarks define a $\mathcal{C}$-way $\mathcal{K}$-shot task with $\mathcal{C} = 2$, corresponding to foreground and background Lei et al. (2022); Wang et al. (2022). Each support sample contains an image I_s and binary mask M_s, and the model predicts the query mask.

Overview. As shown in Fig. 2(b), our framework replaces prototype matching with prompt-conditioned anisotropic Gaussian modeling. It has two components: a CoT-based Prompt Generator and a Meta-Promptable Distribution Modeling module, termed the Meta-Learner. As illustrated in Fig. 3, the Prompt Generator derives semantic prompts from support images and class-specific instructions, capturing category, domain, and variation cues. The Meta-Learner fuses these prompts with support visual features to estimate class-wise means and covariances for Mahalanobis matching.

2.1 COT-BASED PROMPT GENERATOR

Estimating class-specific distributions from only a few support samples is challenging, since sparse visual evidence cannot fully capture category diversity, domain context, and intra-class variation. To provide complementary semantic priors, we introduce a CoT-based prompt generation strategy that combines support images with class-specific textual instructions. The generated descriptions are encoded as semantic prompts to guide foreground/background distribution modeling. Specifically, we generate three types of semantic cues. For category semantics, we use support masks to isolate foreground objects and background regions, which are then processed by a pre-trained vision-language model (e.g., DAM Lian et al. (2025)) to obtain fine-grained descriptions of the target class and its context. For domain semantics, we design a multi-turn CoT protocol for large language models (e.g., GPT Islam & Moushi (2025)) to decompose domain characteristics step by step. Since CD-FSL tasks

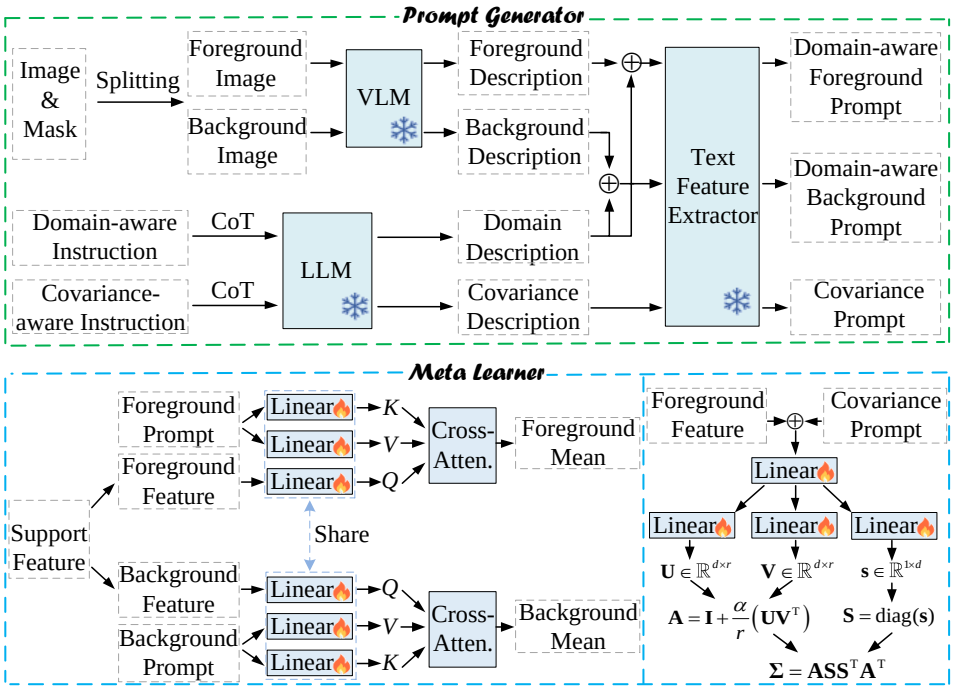


Figure 3: Overview of the proposed CoT-based Prompt Generator and Meta-Learner. The Prompt Generator integrates support images and class-specific instructions to produce semantic prompts describing category concepts, domain context, and variation cues. Conditioned on these prompts and support features, the Meta-Learner estimates the means and covariance matrices of foreground/background anisotropic Gaussian distributions, with low-rank covariance factorization for efficient few-shot modeling.

usually specify the target domain, such as medical imagery or remote sensing, this protocol elicits structured domain knowledge that helps contextualize visual features. For covariance semantics, we further use CoT prompting to describe class-specific variation patterns. For foreground classes, multi-turn templates guide the LLM to summarize likely appearance, shape, and distributional variations. For background covariance, whose semantics are unbounded and less coherent, we do not generate text prompts and instead rely on visual features for distribution estimation.

The resulting textual descriptions are encoded by a pre-trained text encoder (e.g., BERT Devlin et al. (2019)). We combine domain semantics with foreground and background category semantics to obtain the domain-aware prompts $\mathbf{t}_{fg}$ and $\mathbf{t}_{bg}$, respectively. The foreground variation descriptions are separately encoded as the covariance prompt $\mathbf{t}_{cov}$. These prompts are then fed into the meta-learner to guide anisotropic Gaussian distribution estimation.

2.2 Meta-Promptable Distribution Modeling

Our framework employs a Meta-Promptable Distribution Modeling Module (i.e., Meta-learner) to generate the parameters of anisotropic Gaussian distributions for each class. This module takes visual features from support images and semantic prompts as inputs, and produces distribution parameters through two dedicated procedures: 1) Mean estimation via cross-modal fusion, where domain-aware prompts interact with visual features through cross-attention to generate refined mean vectors; 2) Covariance estimation via low-rank diagonal factorization, a strategy that directly generates the components for constructing full covariance matrices, thereby enabling stable and accurate estimation.

Mean Estimation via Cross-modal Fusion. Given a support image I_s, we first employ a visual encoder to extract dense feature maps $\mathbf{F}_s \in \mathbb{R}^{h\times w\times d}$. These spatial representations are then aggregated using the corresponding binary mask M_s to obtain foreground and background feature vectors $\mathbf{f}_{fg}$ and $\mathbf{f}_{bg}$, respectively. The resulting features serve as compact yet comprehensive visual summaries, encapsulating the essential characteristics of each class. Subsequently, the foreground mean vector $\boldsymbol{\mu}_{fg}$ is generated through cross-modal fusion between the visual feature $\mathbf{f}_{fg}$ and the domain-aware semantic prompt $\mathbf{t}_{fg}$ via a cross-attention operation:

$$\boldsymbol{\mu}_{fg} = \mathrm{CrossAttn}\left(\mathbf{f}_{fg}, \mathbf{t}_{fg}\right). \tag{1}$$

The background mean vector $\boldsymbol{\mu}_{bg}$ is produced following the same procedure using the shared cross-attention layer with inputs $\mathbf{f}_{bg}$ and $\mathbf{t}_{bg}$. This cross-modal fusion yields mean vectors that are both visually grounded and semantically informed, establishing a robust foundation for modeling class-specific distributions.

Covariance Estimation via Low-Rank Factorization. Accurate estimation of the covariance matrix is essential for capturing complex intra-class variations in real-world categories. However, directly estimating a full covariance matrix in a d-dimensional feature space poses fundamental challenges: it requires $\mathcal{O}(d^2)$ parameters, often leads to ill-conditioning, and introduces significant optimization difficulties. To address these issues, we introduce a computationally efficient factorization strategy that is applied uniformly to both foreground and background classes.

We describe the procedure for the foreground class; the background class follows identically with its own set of parameters. Specifically, we decompose the foreground covariance matrix $\boldsymbol{\Sigma}_{fg}$ into a full-rank transformation matrix $\mathbf{A}$ and a diagonal scaling matrix $\mathbf{S}_{fg}$, where $\mathbf{A}$ is further reparameterized via the product of two low-rank matrices $\mathbf{U}_{fg}$ and $\mathbf{V}_{fg}$. The parameters of these matrices are generated by lightweight subnetworks that integrate visual and semantic information. The network first concatenates the visual feature $\mathbf{f}_{fg}$ and the covariance prompt $\mathbf{t}_{cov}$, then projects them into a hidden representation:

$$\mathbf{h}_{fg} = \mathbf{W}_h[\mathbf{f}_{fg}; \mathbf{t}_{cov}] + \mathbf{b}_h, \tag{2}$$

where $\mathbf{W}_h$ and $\mathbf{b}_h$ are the weight and bias of the linear projection layer, respectively.

This hidden representation is subsequently mapped to the corresponding matrix parameters via separate linear layers:

$$\mathbf{U}_{fg} = \mathbf{W}_U\mathbf{h}_{fg}, \quad \mathbf{V}_{fg} = \mathbf{W}_V\mathbf{h}_{fg}, \quad \mathbf{s}'_{fg} = \mathbf{W}_s\mathbf{h}_{fg}, \tag{3}$$

where $\mathbf{W}_U$, $\mathbf{W}_V$, $\mathbf{W}_s$ are learnable weight matrices.

To ensure numerical stability, the scaling vector is passed through a softplus activation Glorot et al. (2011): $\mathbf{s}_{fg} = \mathrm{softplus}(\mathbf{s}'_{fg})$. The transformation matrix $\mathbf{A}$ is constructed as a controlled perturbation of the identity matrix:

$$\mathbf{A} = \mathbf{I} + \frac{\alpha}{r}\mathbf{U}_{fg}\mathbf{V}_{fg}^\top, \quad \mathbf{A} \in \mathbb{R}^{d\times d}, \tag{4}$$

where $\mathbf{U}_{fg}, \mathbf{V}_{fg} \in \mathbb{R}^{d\times r}$ are low-rank factors with rank $r = 30$ and feature dimension $d = 1024$ (thus $r \ll d$), and $\alpha = 1$ is a hyperparameter controlling the perturbation strength. The identity matrix $\mathbf{I}$ serves as a strong prior, ensuring that $\mathbf{A}$ is initialized near identity and remains well-conditioned throughout training. The diagonal scaling matrix is defined as $\mathbf{S}_{fg} = \mathrm{diag}(\mathbf{s}_{fg})$, which encodes feature-specific variances. Finally, the complete foreground covariance matrix is assembled as:

$$\boldsymbol{\Sigma}_{fg} = \mathbf{A}\mathbf{S}_{fg}\mathbf{S}_{fg}^\top\mathbf{A}^\top. \tag{5}$$

For $\boldsymbol{\Sigma}_{fg}$ to be positive definite, $\mathbf{A}$ must be full rank. In our construction, $\mathbf{A} = \mathbf{I} + \frac{\alpha}{r}\mathbf{U}_{fg}\mathbf{V}_{fg}^\top$ is initialized near identity because $\mathbf{U}_{fg}, \mathbf{V}_{fg}$ are small. During training, the cross-entropy loss naturally penalizes near-singular covariance matrices (as they produce extreme Mahalanobis distances and poor predictions), so the network avoids learning $\mathbf{U}_{fg}, \mathbf{V}_{fg}$ that would render $\mathbf{A}$ singular. Given $\mathbf{s}_{fg} > \mathbf{0}$ and $\mathbf{A}$ full rank, for any non-zero vector $\mathbf{x} \in \mathbb{R}^d$:

$$\mathbf{x}^\top\boldsymbol{\Sigma}_{fg}\mathbf{x} = \mathbf{x}^\top\mathbf{A}\mathbf{S}_{fg}\mathbf{S}_{fg}^\top\mathbf{A}^\top\mathbf{x} = \|\mathbf{S}_{fg}^\top\mathbf{A}^\top\mathbf{x}\|_2^2 > 0, \tag{6}$$

which ensures $\boldsymbol{\Sigma}_{fg}$ is positive definite and invertible, as required for computing Mahalanobis distances.

The background covariance matrix $\boldsymbol{\Sigma}_{bg}$ is computed from the background visual feature $\mathbf{f}_{bg}$ using the same low-rank parameterization as the foreground, but without a background-specific covariance prompt. As discussed earlier, background variations are too unstructured for reliable textual prompting, since background regions lack coherent semantic boundaries.

Residual Enhancement Strategy. To stabilize the optimization of the meta-learner, we introduce a residual enhancement strategy that leverages statistical priors from support image features for robust distribution initialization. Specifically, the visual features $\mathbf{f}_{fg}$ and $\mathbf{f}_{bg}$ serve as semantically

meaningful anchors, enabling the meta-learner to refine distribution parameters via residual learning. Mean vectors are updated as:

$$\boldsymbol{\mu}_{fg} \leftarrow \boldsymbol{\mu}_{fg} + \mathbf{f}_{fg}, \quad \boldsymbol{\mu}_{bg} \leftarrow \boldsymbol{\mu}_{bg} + \mathbf{f}_{bg}. \tag{7}$$

For covariance estimation, we further enhance both foreground and background covariance matrices via statistical regularization. Channel-wise variances $\sigma^2_{fg}(d)$ and $\sigma^2_{bg}(d)$ are computed as the sample variances across spatial dimensions, and the regularized covariance matrices are updated as:

$$\begin{aligned} \boldsymbol{\Sigma}_{fg} &\leftarrow \boldsymbol{\Sigma}_{fg} + \mathrm{diag}\big(\sigma^2_{fg}(1), \ldots, \sigma^2_{fg}(d)\big), \\ \boldsymbol{\Sigma}_{bg} &\leftarrow \boldsymbol{\Sigma}_{bg} + \mathrm{diag}\big(\sigma^2_{bg}(1), \ldots, \sigma^2_{bg}(d)\big). \end{aligned} \tag{8}$$

The resulting probabilistic model, represented by the anisotropic Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_{fg}, \boldsymbol{\Sigma}_{fg})$ and $\mathcal{N}(\boldsymbol{\mu}_{bg}, \boldsymbol{\Sigma}_{bg})$, establishes a stable and expressive foundation for few-shot prediction based on Mahalanobis matching.

2.3 Training and Inference Protocol

The segmentation of the query image is performed by computing the Mahalanobis distance between its pixel-level features and the estimated Gaussian distributions of the foreground and background. Specifically, for a query image I_q, we first extract a dense feature map $\mathbf{F}_q$ using the visual encoder. Each pixel feature $\mathbf{x}_q = \mathbf{F}_q(h, w)$ is then classified based on its Mahalanobis distance to the two class distributions:

$$\begin{aligned} d_{\text{fg}}(\mathbf{x}_q) &= \sqrt{(\mathbf{x}_q - \boldsymbol{\mu}_{\text{fg}})^\top \boldsymbol{\Sigma}_{\text{fg}}^{-1} (\mathbf{x}_q - \boldsymbol{\mu}_{\text{fg}})}, \\ d_{\text{bg}}(\mathbf{x}_q) &= \sqrt{(\mathbf{x}_q - \boldsymbol{\mu}_{\text{bg}})^\top \boldsymbol{\Sigma}_{\text{bg}}^{-1} (\mathbf{x}_q - \boldsymbol{\mu}_{\text{bg}})}. \end{aligned} \tag{9}$$

The mask $\hat{M}_q$ is generated by assigning each pixel to the class with the minimal Mahalanobis distance:$\hat{M}_q(h, w) = 1$ if $d_{\text{fg}}(\mathbf{x}_q) < d_{\text{bg}}(\mathbf{x}_q)$, and 0 otherwise. The cross-entropy loss between the predicted mask $\hat{M}_q$ and ground truth M_q is computed to enable end-to-end optimization. During meta-training, the model is exposed to a large number of episodic tasks, thereby acquiring a transferable inductive bias that enables it to dynamically infer distribution from limited visual-linguistic evidence.

After training, the model serves as a strong initialization for novel target domains. For a target task, we perform rapid adaptation—termed meta-finetuning—using only its support set. Following prior practice Lei et al. (2022); Hu et al. (2022), we construct pseudo-query samples via data augmentation on the support images and fine-tune the model on these samples. The adapted model then estimates the final anisotropic Gaussian distributions and predicts the query sample via Mahalanobis distance.

2.4 Generalizing to Classification and Detection

Our framework naturally extends to cross-domain few-shot classification and detection, which are also dominated by the prototype-based paradigm. For classification, our method replaces Euclidean distance with Mahalanobis distance to the estimated anisotropic Gaussian distribution. For detection, it substitutes prototype-based operations throughout the pipeline with their distribution-based counterparts. Importantly, this upgrade serves as a drop-in replacement for existing methods Hu et al. (2022); Zou et al. (2025); Fu et al. (2025), enhancing class semantics while preserving original network architectures.

2.5 Theoretical Analysis

We provide a theoretical analysis tailored to CD-FSL, where the learner is given only a few support examples from a target domain that suffers from severe domain shift. In this challenging setting, prototype matching—which implicitly assumes isotropic Gaussian distributions—suffers from high bias, while directly estimating a full anisotropic covariance matrix from the limited support set is ill-posed. Our framework resolves this dilemma by leveraging large foundation models to generate informative CoT prompts, which guide a low-rank factorization of the covariance matrix for each class. We prove that, under the same few-shot conditions, the resulting Mahalanobis decision rule achieves strictly lower Bayes risk than the optimal prototype-based classifier. Moreover, this advantage becomes more pronounced as the true anisotropy of the feature distributions increases—a characteristic that is particularly strong in CD-FSL tasks.

Table 1: Comparison with state-of-the-art methods on $\mathcal{K}$-shot CD-FSS.

			ISIC		Chest X-Ray		FSS-1000		Deepglobe		Average	
Method	**Backbone**	**Mark**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**
ABCDFSS Herzog (2024)	ResNet-50	CVPR-24	45.7	53.3	79.8	81.4	74.6	76.2	42.6	49.0	60.7	65.0
DMTNet Chen et al. (2024b)	ResNet-50	IJCAI-24	43.6	52.3	73.8	77.0	81.5	83.3	40.1	51.2	59.7	66.0
APM-M Tong et al. (2024)	ResNet-50	NeurIPS-24	41.71	51.16	78.25	82.81	79.29	81.83	40.86	44.92	60.03	65.18
DR-Adapter Su et al. (2024)	ResNet-50	CVPR-24	40.77	48.87	82.35	82.31	79.05	80.40	41.29	50.12	60.86	65.42
IFA Nie et al. (2024)	ResNet-50	CVPR-24	66.3	69.8	74.0	74.6	80.1	82.4	50.6	58.8	67.8	71.4
TVGTANet Liu et al. (2025a)	ResNet-50	CVPR-24	47.21	58.75	84.58	87.27	78.32	81.44	42.04	50.67	63.04	69.53
DATO Li et al. (2025)	ResNet-50	CVPR-25	68.76	70.30	79.58	81.07	81.79	84.61	51.10	59.30	70.31	73.82
GPRN Peng et al. (2025)	ResNet-50	AAAI-25	66.8	72.2	87.0	87.1	81.1	82.6	51.7	59.3	71.7	75.3
Ours	ResNet-50	-	**76.7**	**79.2**	**89.3**	**89.9**	**85.4**	**87.6**	**54.9**	**65.1**	**76.6**	**80.5**

Table 2: Comparison with state-of-the-art methods on 5-way $\mathcal{K}$-shot CD-FSC.

			Chest		ISIC		EuroSAT		CropDisease		Average	
Method	**Backbone**	**Mark**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**
FLoR Zou et al. (2024a)	ViT-Small	CVPR-24	23.26	27.02	35.49	53.06	73.09	90.75	83.55	96.47	53.85	66.83
AttnTemp Zou et al. (2024b)	ViT-Small	NeurIPS-24	23.63	28.03	38.05	54.91	75.09	90.82	84.78	96.66	55.39	67.61
ReCIT Yi et al. (2025b)	ViT-Small	ICML-25	23.84	28.88	38.48	54.91	75.23	91.58	85.92	96.85	55.87	68.06
REAP Yi et al. (2025a)	ViT-Small	ICML-25	24.17	28.34	38.67	55.28	75.97	91.79	85.33	96.71	56.04	68.03
PMF Hu et al. (2022)	ViT-Small	CVPR-22	21.73	27.27	30.36	50.12	70.74	85.98	80.79	92.96	50.91	64.08
PMF+Ours	ViT-Small	-	**24.62**	**28.75**	**36.84**	**54.33**	**79.58**	**92.47**	**83.31**	**97.13**	**56.09**	**68.17**
CD-CLS Zou et al. (2025)	ViT-Small	NeurIPS-24	23.39	27.66	35.56	54.69	74.97	91.53	84.54	96.27	54.62	67.54
CD-CLS+Ours	ViT-Small	-	**26.46**	**30.42**	**39.74**	**58.91**	**82.69**	**96.13**	**91.26**	**98.95**	**60.04**	**71.10**

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

For CD-FSS, we follow the standard benchmark Lei et al. (2022): models are pre-trained on PASCAL VOC Everingham et al. (2010) and evaluated on DeepGlobe, ISIC2018, FSS-1000, and Chest X-ray. We report mIoU over 600 randomly sampled tasks for both 1-shot and 5-shot settings. Following common practice Chen et al. (2024a); Peng et al. (2025), we use ResNet-50 He et al. (2016) as the visual backbone. The prompt generator uses frozen DAM Lian et al. (2025), GPT-4 Islam & Moushi (2025), and BERT Devlin et al. (2019), while the meta-learner consists of a shared cross-attention layer for mean estimation and four-layer linear networks for covariance factorization. Images are resized to 400×400, and models are optimized with SGD at a learning rate of 5×10^{-4}. For CD-FSC, we follow Guo et al. Guo et al. (2020), using 64 mini-ImageNet Vinyals et al. (2016) base classes as source and evaluating on EuroSAT, ISIC, ChestX, and CropDisease. For CD-FSD, we follow Fu et al. Fu et al. (2025), using COCO Lin et al. (2014) as source and evaluating on ArTaxOr, Clipart1k, DIOR, DeepFish, NEU-DET, and UODD. We integrate our Mahalanobis matching into representative prototype-based baselines: PMF Hu et al. (2022) and CD-CLS Zou et al. (2025) for classification, and CD-ViTO Fu et al. (2025) for detection. Only the matching module is replaced; original architectures, data splits, and hyperparameters are kept unchanged.

3.2 COMPARISON WITH SOTA METHODS

Tables 1, 2, and 3 present comparisons across segmentation, classification, and detection. Across all three tasks, our method consistently improves over strong SoTA baselines. On CD-FSS, our framework achieves the best results on all four benchmarks under both 1-shot and 5-shot settings, improving the average mIoU over GPRN Peng et al. (2025) by 4.9 and 5.2 points. On CD-FSC and CD-FSD, replacing prototype-style matching in representative baselines, including PMF Hu et al. (2022), CD-CLS Zou et al. (2025), and CD-ViTO Fu et al. (2025), also yields consistent gains across domains and shot settings. These results support a unified conclusion: the isotropic prototype assumption is a shared bottleneck in CD-FSL, regardless of whether the query unit is a pixel, an image, or a region proposal. Replacing point prototypes with prompt-conditioned anisotropic Gaussians lets our framework capture class-specific covariance and domain uncertainty, positioning Mahalanobis matching as a general alternative to prototype-based CD-FSL.

3.3 ABLATION STUDY

Effect of the Prompt Generator. The Prompt Generator provides semantic priors from support images and class-specific instructions. In "Ours w/o Prompt Generator," these semantic prompts are removed, and class distributions are estimated by the meta-learner from visual features alone. As shown in Table 4, removing this component consistently reduces performance across all benchmarks. This confirms that semantic prompts provide complementary and transferable cues for distribution estimation under domain shift, and that they cannot be fully replaced by visual features alone.

Effect of Meta-Promptable Distribution Modeling. The Meta-Promptable Distribution Modeling module learns to estimate class-wise anisotropic Gaussian distributions from visual features and semantic prompts. In "Ours w/o Meta-Promptable Distribution Modeling," the entire learnable

Table 3: Comparison with state-of-the-art methods on $\mathcal{A}ll$-way $\mathcal{K}$-shot CD-FSD.

Method	Backbone	Mark	Shot	ArTaxOr	Clipart1k	DIOR	DeepFish	NEU-DET	UODD	Average
DE-ViT Zhang et al. (2024)	ViT-Large	CoRL-24	1-shot	0.4	0.5	2.7	0.4	0.4	1.5	1.0
DE-ViT-FT Zhang et al. (2024)	ViT-Large	CoRL-24	1-shot	10.5	13.0	14.7	19.3	0.6	2.4	10.1
CD-ViTO Fu et al. (2025)	ViT-Large	ECCV-24	1-shot	21.0	17.7	17.8	20.3	3.6	3.1	13.9
CD-ViTO+Ours	ViT-Large	-	1-shot	**26.2**	**25.8**	**21.3**	**24.1**	**12.6**	**12.5**	**20.4**
DE-ViT Zhang et al. (2024)	ViT-Large	CoRL-24	5-shot	10.1	5.5	7.8	2.5	1.5	3.1	5.1
DE-ViT-FT Zhang et al. (2024)	ViT-Large	CoRL-24	5-shot	38.0	38.1	23.4	21.2	7.8	5.0	22.3
CD-ViTO Fu et al. (2025)	ViT-Large	ECCV-24	5-shot	47.9	41.1	26.9	22.3	11.4	6.8	26.1
CD-ViTO+Ours	ViT-Large	-	5-shot	**56.1**	**48.5**	**34.7**	**27.5**	**18.9**	**16.8**	**33.8**
DE-ViT Zhang et al. (2024)	ViT-Large	CoRL-24	10-shot	9.2	11.0	8.4	2.1	1.8	3.1	5.9
DE-ViT-FT Zhang et al. (2024)	ViT-Large	CoRL-24	10-shot	49.2	40.8	25.6	21.3	8.8	5.4	25.2
CD-ViTO Fu et al. (2025)	ViT-Large	ECCV-24	10-shot	60.5	44.3	30.8	22.3	12.8	7.0	29.6
CD-ViTO+Ours	ViT-Large	-	10-shot	**68.9**	**52.6**	**37.1**	**33.8**	**19.5**	**22.8**	**39.1**

Table 4: Ablation study on $\mathcal{K}$-shot CD-FSS.

	ISIC		Chest X-Ray		FSS-1000		Deepglobe		Average	
Method	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**
Baseline	41.2	53.6	66.6	70.2	78.6	81.2	37.9	43.0	56.1	62.0
Ours	**76.7**	**79.2**	**89.3**	**89.9**	**85.4**	**87.6**	**54.9**	**65.1**	**76.6**	**80.5**
Ours w/o Prompt Generator	64.4	69.3	74.6	75.1	74.4	82.1	45.1	57.6	64.6	71.0
Ours w/o Meta-Promptable Distribution Modeling	61.7	66.3	73.3	80.1	79.9	80.8	44.7	55.0	64.9	70.6
Ours w/o Covariance	72.9	74.9	85.1	85.9	78.4	84.1	49.3	60.2	71.4	76.3
Ours w/o CoT Strategy	73.4	76.1	86.4	86.9	82.8	84.2	51.7	61.4	73.6	77.2
Ours w/o Covariance Factorization	74.8	77.1	88.1	87.3	82.6	85.3	52.1	61.8	74.4	77.9
Ours w/o Residual Strategy	73.5	75.7	85.8	86.3	81.1	83.4	49.1	61.5	72.3	76.7

module is replaced by direct empirical estimation: class means and covariances are computed from visual feature statistics, without using either the meta-learner or the generated prompts. This variant therefore measures the joint contribution of meta-learned estimation and prompt-guided modeling, relative to a purely visual, non-learned baseline. As shown in Table 4, it suffers a substantial performance drop, indicating that direct empirical estimation is unreliable under limited support samples. The meta-learned estimator, in contrast, leverages transferable knowledge to infer stable and task-adaptive class distributions.

Effect of Covariance Modeling. We further isolate the contribution of covariance modeling by removing the covariance term entirely, yielding the variant "Ours w/o Covariance." In this setting, each class is represented only by its mean, and the model degenerates to prototype matching, while the semantic prompts are still used to refine the prototype. This variant therefore measures how much of our gain comes from anisotropic covariance modeling beyond semantic prototype refinement. As shown in Table 4, "Ours w/o Covariance" drops from 76.6/80.5 to 71.4/76.3 on average, i.e., a 5.2/4.2 point decrease, despite still benefiting from prompt-based refinement. This confirms that the improvement of our full model cannot be attributed to semantic prompts alone: explicitly modeling class-wise covariance and matching with Mahalanobis distance provides a substantial additional gain, and the isotropic prototype assumption remains a key bottleneck in CD-FSS.

Effect of Design Choices. We further ablate the design choices within the Prompt Generator and the distribution modeling module. We compare the full model with three variants: "Ours w/o CoT Strategy," which replaces multi-turn CoT prompting with a single-turn question–answering template; "Ours w/o Covariance Factorization," which predicts covariance matrices directly with MLPs instead of low-rank factorization; and "Ours w/o Residual Strategy," which removes the residual connection between visual statistics and meta-learned parameters. As shown in Table 4, removing any of these designs consistently degrades performance. Among them, the CoT strategy contributes the most, confirming that structured multi-turn reasoning yields richer semantic descriptions than single-turn prompting. Covariance factorization and residual enhancement also both improve performance, with residual enhancement causing the larger drop when removed, confirming its role in stabilizing few-shot distribution modeling.

3.4 Further Analysis

Robustness Analysis. Our framework improves robustness through two complementary mechanisms. First, text-image cross-attention acts as an adaptive semantic filter: instead of treating LFM outputs as fixed instructions, it measures token-level relevance between textual descriptions and visual features. Hallucinated or ambiguous tokens tend to receive lower attention because they are weakly aligned with the support evidence, reducing the impact of noisy prompts. Second, the residual enhancement strategy provides a visual-statistical fallback. By injecting mean

Table 5: Evaluating the robustness of Prompt Generator.

	ISIC		Chest X-Ray		FSS-1000		Deepglobe		Average	
Method	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**
Baseline	41.2	53.6	66.6	70.2	78.6	81.2	37.9	43.0	56.1	62.0
Ours w/ BLIP	74.8	77.7	87.4	87.8	84.1	85.9	53.9	64.1	75.1	78.9
Ours w/ CoCa	74.2	76.9	87.2	87.4	83.5	85.3	53.2	63.9	74.5	78.4
Ours w/ Llama	74.9	77.8	87.9	88.2	84.1	86.1	54.1	64.3	75.3	79.1
Ours w/ Qwen-VL	75.2	78.3	88.1	88.3	84.4	86.3	54.2	64.5	75.5	79.4
Ours	**76.7**	**79.2**	**89.3**	**89.9**	**85.4**	**87.6**	**54.9**	**65.1**	**76.6**	**80.5**

and covariance priors computed from support features, the meta-learner refines rather than replaces visual statistics, preventing unreliable semantic prompts from dominating distribution estimation.

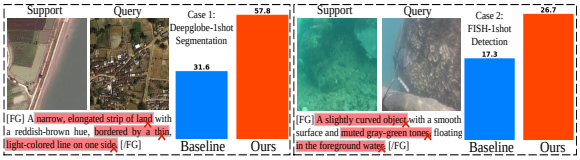


Figure 4: Case study on model robustness.

We empirically examine robustness from three aspects. First, Fig. 4 shows that our method remains stable under noisy or ambiguous captions, as visual grounding helps suppress misleading semantic cues. Second, Table 5 evaluates different foundation models, including BLIP Li et al. (2022), CoCa Yu et al. (2022), Llama Touvron et al. (2023), and Qwen-VL Wang et al. (2024). The small performance variation indicates that our method does not depend on a specific VLM or LLM. Third, results on fourteen segmentation, classification, and detection benchmarks show consistent gains across medical, remote-sensing, underwater, industrial, and artistic domains, demonstrating that our method remains robust to imperfect semantic descriptions.

Table 6: Efficiency comparison on 1-shot CD-FSS.

Method	Backbone	Time (S)↓	Params (M)↓	Acc. (%)↑
DR-Adapter Su et al. (2024)	ResNet-50	0.81	50.6	60.68
GPRN Peng et al. (2025)	ResNet-50	1.08	**10.0**	71.7
Ours	ResNet-50	**0.45**	17.7	**76.6**

Efficiency Analysis. Our framework is efficient by design: the Prompt Generator is inference-only and introduces no learnable parameters, while all additional parameters are confined to a lightweight meta-learner. We evaluate efficiency on 1-shot CD-FSS using an NVIDIA RTX 3090 GPU and compare with representative methods Su et al. (2024); Peng et al. (2025). Table 6 reports online processing time, extra parameters beyond the backbone, and average mIoU over four benchmarks. Our method achieves 76.6% mIoU with 0.45s per task and 17.7M extra parameters, outperforming GPRN Peng et al. (2025) and DR-Adapter Su et al. (2024) while remaining computationally lightweight. The reported time includes only online computation, i.e., text encoding, few-shot adaptation, and query inference. Text prompt generation is performed offline, including VLM captioning (0.81s/task) and LLM CoT reasoning (1.42s/task), for a total of 2.23s per task. Since these operations depend only on support samples, not query images, and domain-aware prompts are generated once per target domain, their amortized deployment cost is negligible.

4 Related Work

CD-FSL has been studied in image classification Zou et al. (2025), semantic segmentation He et al. (2024), and object detection Fu et al. (2025). Despite different output forms, most methods adopt a prototype-based formulation, where support examples are compressed into class prototypes and queries are assigned by Euclidean or cosine nearest-prototype matching Snell et al. (2017). This paradigm has been extended from classification to segmentation and detection with task-specific adaptations. Recent works mainly improve it through prototype refinement, such as informative support selection Wang et al. (2023), support-query consistency He et al. (2024), and query self-similarity Chen et al. (2024b), or through domain adaptation, including frequency filtering Tong et al. (2024), style suppression Su et al. (2024); Kong et al. (2024), test-time adaptation Herzog (2024), correspondence refinement Nie et al. (2024), feature regularization Liu et al. (2025b), adapter-based modeling Tong et al. (2025), adversarial training Li et al. (2025), and SAM-assisted masks Peng et al. (2025). In detection, representative methods such as DeViT Zhang et al. (2024) and CD-ViTO Fu et al. (2025) also rely on prototype-style class representations. Although effective, these methods largely preserve the isotropic geometry of prototype matching, representing each class as a single point while ignoring class-specific covariance. This limitation is critical under domain shift, where target features are often anisotropic and correlated. In contrast, our method replaces point prototypes with prompt-conditioned anisotropic Gaussian distributions and performs Mahalanobis matching, providing a distribution-aware alternative across segmentation, classification, and detection.

5 Conclusion

We presented a meta-promptable Mahalanobis matching framework for CD-FSL. By replacing point prototypes with prompt-conditioned anisotropic Gaussian distributions, our method better captures class-specific covariance and target-domain uncertainty. Experiments across fourteen segmentation, classification, and detection benchmarks demonstrate consistent SoTA performance and validate distribution-aware matching as a unified alternative to prototype-based CD-FSL.

AI USE STATEMENT

In this work, we used generative AI tools for two purposes. First, they assisted with manuscript preparation, including translation, language polishing, and refinement of wording, sentence structure, and presentation. Second, large foundation models were used as components of our proposed framework: a pretrained vision-language model (e.g., DAM) generated category-level descriptions from support images, a large language model (e.g., GPT-4) produced chain-of-thought prompts for domain and variation semantics, and a pretrained text encoder (e.g., BERT) encoded the resulting descriptions into semantic prompts. These models are used as fixed, off-the-shelf feature extractors and prompt generators within our pipeline; they are not used to formulate mathematical claims or proofs, and all theoretical results in this work were derived by the authors.

All AI-assisted text was reviewed and revised by the authors. The semantic prompts produced by the foundation models were used only as inputs to the meta-learner, and all experimental results, comparisons, and conclusions were obtained from our own implementation and evaluation. We take full responsibility for the final content of this work, including its text, claims, results, and artifacts.

ETHICS STATEMENT

This work uses only publicly available datasets collected under their original licenses; no human-subject experiments were conducted. Our framework relies on pretrained foundation models to generate semantic prompts, which may inherit biases from their training data and could produce inaccurate or stereotyped descriptions for some domains or categories. Our cross-attention filtering and residual enhancement partially mitigate this risk, but we do not claim full immunity to such biases. Predictions on medical and remote-sensing data are for research use only and must not be treated as clinical or operational advice. We declare no conflicts of interest or external sponsorship.

REPRODUCIBILITY STATEMENT

The main text specifies the backbone, the CoT-based Prompt Generator, the Meta-Promptable Distribution Modeling module, the low-rank covariance factorization, and the Mahalanobis matching rule, together with all training and adaptation procedures. Implementation details, including the rank r, the prompt design, the optimizers, and the evaluation protocols for CD-FSS, CD-FSC, and CD-FSD, are provided in the Experimental Setup and Appendix, along with ablation studies, hyperparameter validation, robustness and efficiency analyses, and qualitative results. Source code and prompts will be publicly released.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Hecheng Cai, Yang Liu, Shudong Huang, and Jiancheng Lv. With a little help from language: semantic enhanced visual prototype framework for few-shot learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 3751–3759, 2024.

Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P Musco, Rahul K Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Antani, George Thoma, and Clement J McDonald. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE transactions on medical imaging*, 33(2):577–590, 2013.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

Hao Chen, Yonghan Dong, Zheming Lu, Yunlong Yu, and Jungong Han. Pixel matching network for cross-domain few-shot segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 978–987, 2024a.

Jiayi Chen, Rong Quan, and Jie Qin. Cross-domain few-shot semantic segmentation via doubly matching transformation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 641–649, 2024b.

Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 172–181, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

G. Drange. Arthropod taxonomy orders object detection dataset. 2019.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010.

Yuqian Fu, Yu Wang, Yixuan Pan, Lian Huai, Xingyu Qi, Zeyu Shangguan, Tong Liu, Yanwei Fu, Luc Van Gool, and Xingqun Jiang. Cross-domain few-shot object detection via enhanced open-set object detector. *COMPUTER VISION-ECCV 2024, PT LVIII*, 15116:247–264, 2025.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 315–323. JMLR Workshop and Conference Proceedings, 2011.

Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. In *European conference on computer vision*, pp. 124–141. Springer, 2020.

John C Harsanyi. Bayesian decision theory and utilitarian ethics. *The American Economic Review*, 68(2):223–228, 1978.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Weizhao He, Yang Zhang, Wei Zhuo, Linlin Shen, Jiaqi Yang, Songhe Deng, and Liang Sun. Apseg: auto-prompt network for cross-domain few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23762–23772, 2024.

Jonas Herzog. Adapt before comparison: A new perspective on cross-domain few-shot segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23605–23615, 2024.

Shell Xu Hu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy M Hospedales. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9068–9077, 2022.

Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5001–5009, 2018.

Raisa Islam and Owana Marzia Moushi. Gpt-4o: The cutting-edge advancement in multimodal llm. In *Intelligent Computing-Proceedings of the Computing Conference*, pp. 47–60. Springer, 2025.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pp. 709–727. Springer, 2022.

Lihao Jiang, Yi Wang, Qi Jia, Shengwei Xu, Yu Liu, Xin Fan, Haojie Li, Risheng Liu, Xinwei Xue, and Ruili Wang. Underwater species detection using channel sharpening attention. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4259–4267, 2021.

Qiuyu Kong, Jiangming Chen, Jie Jiang, Zanxi Ruan, and Lai Kang. Dual-branch fusion with style modulation for cross-domain few-shot semantic segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 2166–2174, 2024.

Shuo Lei, Xuchao Zhang, Jianfeng He, Fanglan Chen, Bowen Du, and Chang-Tien Lu. Cross-domain few-shot semantic segmentation. In *European conference on computer vision*, pp. 73–90. Springer, 2022.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020a.

Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. Fss-1000: A 1000-class dataset for few-shot segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2869–2878, 2020b.

Zhaoyang Li, Yuan Wang, Wangkai Li, Tianzhu Zhang, and Xiang Liu. Dual-agent optimization framework for cross-domain few-shot segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9849–9859, 2025.

Long Lian, Yifan Ding, Yunhao Ge, Sifei Liu, Hanzi Mao, Boyi Li, Marco Pavone, Ming-Yu Liu, Trevor Darrell, Adam Yala, et al. Describe anything: Detailed localized image and video captioning. *arXiv e-prints*, pp. arXiv–2504, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pp. 740–755. Springer, 2014.

Jianming Liu, Wenlong Qiu, and Haitao Wei. Textual and visual guided task adaptation for source-free cross-domain few-shot segmentation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5150–5159, 2025a.

Yuhan Liu, Yixiong Zou, Yuhua Li, and Ruixuan Li. The devil is in low-level features for cross-domain few-shot segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4618–4627, 2025b.

Jiahao Nie, Yun Xing, Gongjie Zhang, Pei Yan, Aoran Xiao, Yap-Peng Tan, Alex C Kot, and Shijian Lu. Cross-domain few-shot segmentation via iterative support-query correspondence mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3380–3390, 2024.

Shi-Feng Peng, Guolei Sun, Yong Li, Hongsong Wang, and Guo-Sen Xie. Sam-aware graph prompt reasoning network for cross-domain few-shot segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6488–6496, 2025.

Alzayat Saleh, Issam H Laradji, Dmitry A Konovalov, Michael Bradley, David Vazquez, and Marcus Sheaves. A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Scientific reports*, 10(1):14671, 2020.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pp. 4077–4087, 2017.

Kechen Song and Yunhui Yan. A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. *Applied Surface Science*, 285:858–864, 2013. ISSN 0169-4332. doi: https://doi.org/10.1016/j.apsusc.2013.09.002. URL https://www.sciencedirect.com/science/article/pii/S0169433213016437.

Jiapeng Su, Qi Fan, Wenjie Pei, Guangming Lu, and Fanglin Chen. Domain-rectifying adapter for cross-domain few-shot segmentation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 24036–24045, 2024.

Jintao Tong, Yixiong Zou, Yuhua Li, and Ruixuan Li. Lightweight frequency masker for cross-domain few-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 37: 96728–96749, 2024.

Jintao Tong, Ran Ma, Yixiong Zou, Guangyao Chen, Yuhua Li, and Ruixuan Li. Adapter naturally serves as decoupler for cross-domain few-shot semantic segmentation. In *Forty-second International Conference on Machine Learning*, 2025.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pp. 3630–3638, 2016.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Wenjian Wang, Lijuan Duan, Yuxi Wang, Qing En, Junsong Fan, and Zhaoxiang Zhang. Remember the difference: Cross-domain few-shot semantic segmentation via meta-memory transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7065–7074, 2022.

Yuxi Wang, Jian Liang, Jun Xiao, Shuqi Mei, Yuran Yang, and Zhaoxiang Zhang. Informative data mining for one-shot cross-domain semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1064–1074, 2023.

Jiamin Wu, Xin Liu, Xiaotian Yin, Tianzhu Zhang, and Yongdong Zhang. Task-adaptive prompted transformer for cross-domain few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6012–6020, 2024.

Shuai Yi, Yixiong Zou, Yuhua Li, and Ruixuan Li. Random registers for cross-domain few-shot learning. *arXiv preprint arXiv:2506.02843*, 2025a.

Shuai Yi, Yixiong Zou, Yuhua Li, and Ruixuan Li. Revisiting continuity of image tokens for cross-domain few-shot learning. *arXiv preprint arXiv:2506.03110*, 2025b.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022.

Xinyu Zhang, Yuhan Liu, Yuting Wang, and Abdeslam Boularias. Detect everything with few examples. *Proceedings of Machine Learning Research*, 270:3986–4004, 2024.

Yixiong Zou, Yicong Liu, Yiman Hu, Yuhua Li, and Ruixuan Li. Flatten long-range loss landscapes for cross-domain few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23575–23584, 2024a.

Yixiong Zou, Ran Ma, Yuhua Li, and Ruixuan Li. Attention temperature matters in vit-based cross-domain few-shot learning. *Advances in Neural Information Processing Systems*, 37:116332–116354, 2024b.

Yixiong Zou, Shuai Yi, Yuhua Li, and Ruixuan Li. A closer look at the cls token for cross-domain few-shot learning. *Advances in Neural Information Processing Systems*, 37:85523–85545, 2025.

A Appendix Overview

The supplementary material presents additional analyses and experiments supporting our contributions. Specifically, it includes:

1. a theoretical re-examination of prototype matching, revealing its restrictive assumptions and establishing the advantages of Mahalanobis matching in hypothesis space and Bayes risk (Section B and Section C);
2. a comparison with zero-shot methods, including SAM-3 on CD-FSS and Qwen3-VL on CD-FSD (Section D.1);
3. ablation studies on chain-of-thought prompting (Section D), individual prompt types, the low-rank hyperparameter (Section D.6), visual prompt tuning (Section 11), and multimodal fusion (Section 12);
4. covariance visualizations confirming the anisotropy of real-world features and the effectiveness of the estimated distributions (Section D.7);
5. qualitative segmentation and detection results demonstrating improved localization, fine-grained perception, and robustness under domain shift (Section D.8 and Section D.9); and
6. a discussion of limitations and promising future directions (Section E).

B Revisiting the Prototype Matching Paradigm in CD-FSL

B.1 Prototype Matching Pipeline

The prototype matching paradigm, illustrated in Fig. 2, has emerged as the dominant framework in CD-FSL, with numerous studies Wang et al. (2023); He et al. (2024); Tong et al. (2024); Su et al. (2024); Kong et al. (2024); Herzog (2024); Nie et al. (2024); Liu et al. (2025b); Tong et al. (2025); Li et al. (2025); Peng et al. (2025) built upon this foundation. To formalize its operational pipeline, we consider a standard $\mathcal{C}$-way $\mathcal{K}$-shot cross-domain few-shot segmentation (CD-FSS) task comprising a support set $\mathcal{S} = \{(I_s^k, M_s^k)\}_{k=1}^{\mathcal{K}}$ with annotations and a query image I_q to be segmented. The process begins with the shared feature extractor f_θ—pre-trained on the source domain—jointly embedding both support and query images into a unified feature space to generate dense support features $\{\mathbf{F}_s^k\}_{k=1}^{\mathcal{K}}$ and query features $\mathbf{F}_q$. Subsequently, for each class $c \in \mathcal{C}$ in the current task, its corresponding prototype vector $\mathbf{p}_c \in \mathbb{R}^d$, where d denotes the feature dimensionality, is computed via masked average pooling over the support features:

$$\mathbf{p}_c = \frac{1}{N_c} \sum_{k=1}^{\mathcal{K}} \sum_{(h,w)\in\mathcal{R}_c^k} \mathbf{F}_s^k(h,w), \tag{10}$$

where $\mathcal{R}_c^k = \{(h,w) : M_s^k(h,w) = c\}$ denotes the set of support pixel positions belonging to class c, and $N_c = \sum_{k=1}^{\mathcal{K}} |\mathcal{R}_c^k|$ represents the total cardinality of such pixels across the support set. In the final segmentation stage, each query pixel feature $\mathbf{x}_q = \mathbf{F}_q(h,w)$ is classified by identifying its nearest prototype in the embedding space via :

$$\hat{M}_q(h,w) = \arg\min_{c\in\mathcal{C}} \|\mathbf{x}_q - \mathbf{p}_c\|^2, \tag{11}$$

where $\hat{M}_q$ denotes the predicted segmentation mask.

B.2 Bayesian Decision Analysis

To establish a rigorous theoretical foundation for analyzing prototype matching, we first revisit the fundamental principles of Bayesian decision theory Harsanyi (1978) for CD-FSS. The optimal classification rule derives from maximizing the posterior probability of class assignment given the observed features and support set. For a query pixel with feature vector $\mathbf{x}_q \in \mathbb{R}^d$, the posterior probability conditioned on the support set $\mathcal{S}$ is given by:

$$P(c|\mathbf{x}_q, \mathcal{S}) = \frac{p(\mathbf{x}_q|c, \mathcal{S})P(c|\mathcal{S})}{p(\mathbf{x}_q|\mathcal{S})}, \tag{12}$$

where the term $p(\mathbf{x}_q|\mathcal{S}) = \sum_{c' \in \mathcal{C}} p(\mathbf{x}_q|c', \mathcal{S})P(c'|\mathcal{S})$ serves as a class-independent normalization constant. Maximizing the posterior probability is therefore equivalent to:

$$c_q^* = \arg\max_{c \in \mathcal{C}} p(\mathbf{x}_q|c, \mathcal{S})P(c|\mathcal{S}). \tag{13}$$

Applying the negative logarithm transformation yields the fundamental Bayesian decision rule:

$$c_q^* = \arg\min_{c \in \mathcal{C}} \left[-\log p(\mathbf{x}_q|c, \mathcal{S}) - \log P(c|\mathcal{S}) \right]. \tag{14}$$

B.3 Underlying Assumptions within Prototype Matching

Within this general Bayesian framework, we now rigorously examine how prototype matching emerges as a specific instance under restrictive assumptions. Specifically, the prototype matching decision rule:

$$c_{q,\text{proto}}^* = \arg\min_{c \in \mathcal{C}} \|\mathbf{x}_q - \mathbf{p}_c\|^2 \tag{15}$$

is equivalent to the Bayesian optimal decision rule equation 14 if and only if the following three conditions hold simultaneously: 1) Isotropic Gaussian Likelihood: $p(\mathbf{x}_q|c, \mathcal{S}) = \mathcal{N}(\mathbf{x}_q|\mu_c, \sigma^2\mathbf{I}_d)$; 2) Uniform Class Prior: $P(c|\mathcal{S}) = \frac{1}{|\mathcal{C}|}$; 3) Accurate Mean Estimation: $\mathbf{p}_c = \mu_c$.

This derivation is straightforward. Under assumption (1), the class-conditional likelihood follows an isotropic Gaussian distribution:

$$p(\mathbf{x}_q|c, \mathcal{S}) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{x}_q - \mu_c\|^2 \right). \tag{16}$$

Computing the negative log-likelihood yields:

$$-\log p(\mathbf{x}_q|c, \mathcal{S}) = \frac{1}{2\sigma^2} \|\mathbf{x}_q - \mu_c\|^2 + \frac{d}{2} \log(2\pi\sigma^2). \tag{17}$$

Substituting into the Bayesian decision rule equation 14 with assumption (2):

$$c_q^* = \arg\min_{c \in \mathcal{C}} \left[\frac{1}{2\sigma^2} \|\mathbf{x}_q - \mu_c\|^2 + \frac{d}{2} \log(2\pi\sigma^2) - \log P(c|\mathcal{S}) \right]. \tag{18}$$

Incorporating the uniform prior assumption:

$$c_q^* = \arg\min_{c \in \mathcal{C}} \left[\frac{1}{2\sigma^2} \|\mathbf{x}_q - \mu_c\|^2 + \underbrace{\frac{d}{2} \log(2\pi\sigma^2) + \log |\mathcal{C}|}_{\text{class-independent constant}} \right]. \tag{19}$$

Removing class-independent constants and applying assumption (3) with $\mathbf{p}_c = \mu_c$ yields the prototype matching rule.

This reveals that prototype matching represents a special case of the general Bayesian optimal classifier, one that requires multiple strong assumptions to hold exactly.

B.4 Pitfalls of Prototype Matching

Above analysis reveals that prototype matching relies on idealized assumptions that prove untenable in CD-FSS. The method presupposes isotropic feature distributions, and precise mean estimation—conditions systematically violated in practice. Real-world features exhibit complex anisotropic covariance structures poorly approximated by $\sigma^2\mathbf{I}_d$, while limited support samples under domain shift induce substantial estimation variance and bias. By collapsing rich class distributions into singular prototypes, the method discards crucial intra-class variability and inherently limits its capacity to model prediction uncertainty across diverse object regions and instances. These fundamental constraints explain prototype matching's diminished performance when confronted with the complex distribution shifts characteristic of cross-domain scenarios.

Although our analysis is presented using the example of few-shot segmentation, the underlying conclusions are task-agnostic. In cross-domain few-shot classification Hu et al. (2022); Zou et al.

(2025), each image is represented as a single feature vector, and the same prototype matching formulation applies directly. In cross-domain few-shot detection Zhang et al. (2024); Fu et al. (2025), region proposals serve as the basic units for feature aggregation and prototype assignment. Across all these tasks, the core limitation remains identical: the implicit isotropic Gaussian assumption and the reliance on single-point prototypes fail to capture the true anisotropic nature of feature distributions under domain shift. Therefore, the theoretical insights and the proposed Mahalanobis matching framework generalize straightforwardly to classification and detection, as validated in our experiments (see Tables 1–3 in the main text).

C Theoretical Comparison: Prototype vs. Mahalanobis Matching

Without loss of generality, we illustrate our theoretical analysis using cross-domain few-shot segmentation (CD-FSS). The same analysis and conclusions apply directly to other CD-FSL tasks (e.g., classification and detection) by treating each entire image (for classification) or each region proposal (for detection) as a basic sample unit.

C.1 Theoretical Foundation

Our framework models the feature distribution of each class conditioned on the support set $\mathcal{S}$ as an anisotropic Gaussian:

$$p(\mathbf{x}_q \mid c, \mathcal{S}) = \mathcal{N}(\mathbf{x}_q \mid \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) = \frac{\exp\left(-\frac{1}{2}\Delta_c(\mathbf{x}_q)\right)}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_c|^{1/2}}, \tag{20}$$

where $\Delta_c(\mathbf{x}_q) = (\mathbf{x}_q - \boldsymbol{\mu}_c)^\top \boldsymbol{\Sigma}_c^{-1}(\mathbf{x}_q - \boldsymbol{\mu}_c)$. This formulation captures both the central tendency and the covariance structure, thereby going beyond point-estimate prototypes.

The Bayes decision rule minimizes the expected risk:

$$c^* = \arg\min_{c \in \mathcal{C}} \left[-\log p(\mathbf{x}_q \mid c, \mathcal{S}) - \log P(c \mid \mathcal{S}) \right]. \tag{21}$$

From equation 20, the negative log-likelihood is:

$$-\log p(\mathbf{x}_q \mid c, \mathcal{S}) = \frac{1}{2}\Delta_c(\mathbf{x}_q) + \frac{1}{2}\log|\boldsymbol{\Sigma}_c| + \frac{d}{2}\log(2\pi). \tag{22}$$

Substituting this into the decision rule and eliminating constants independent of c yields:

$$c^* = \arg\min_{c \in \mathcal{C}} \left\{ \Delta_c(\mathbf{x}_q) + \log|\boldsymbol{\Sigma}_c| - 2\log P(c \mid \mathcal{S}) \right\}. \tag{23}$$

Using the estimated parameters $\hat{\boldsymbol{\mu}}_c, \hat{\boldsymbol{\Sigma}}_c$ and the empirical prior $\hat{P}(c \mid \mathcal{S})$, we obtain the full decision rule:

$$c_q^* = \arg\min_{c \in \mathcal{C}} \left\{ \underbrace{\Delta_c(\mathbf{x}_q)}_{\text{Mahalanobis}} + \underbrace{\log|\hat{\boldsymbol{\Sigma}}_c|}_{\text{Volume penalty}} - \underbrace{2\log \hat{P}(c \mid \mathcal{S})}_{\text{Prior term}} \right\}, \tag{24}$$

where $\Delta_c(\mathbf{x}_q) = (\mathbf{x}_q - \hat{\boldsymbol{\mu}}_c)^\top \hat{\boldsymbol{\Sigma}}_c^{-1}(\mathbf{x}_q - \hat{\boldsymbol{\mu}}_c)$.

C.2 Hypothesis Space Comparison

We now formally compare the hypothesis spaces of prototype matching and the proposed Mahalanobis matching.

Prototype Matching Hypothesis Space $\mathcal{H}_{\text{proto}}$. Classifiers in this space partition the feature space $\mathbb{R}^d$ via Euclidean distance comparisons to learned prototype vectors. The decision region for class c is:

$$\mathcal{H}_c^{\text{proto}} = \left\{ \mathbf{x}_q \in \mathbb{R}^d : \|\mathbf{x}_q - \mathbf{p}_c\|^2 \leq \|\mathbf{x}_q - \mathbf{p}_{c'}\|^2,\ \forall c' \neq c \right\}. \tag{25}$$

This space is parameterized by C prototype vectors $\{\mathbf{p}_c\}_{c=1}^C$ and induces piecewise-linear (hyperplane) decision boundaries.

Mahalanobis Matching Hypothesis Space $\mathcal{H}_{\text{dist}}$. Classifiers in this space leverage Mahalanobis distance to capture second-order statistics. The decision region for class c is a hyper-ellipsoid:

$$\mathcal{H}_c^{\text{dist}} = \left\{ \mathbf{x}_q \in \mathbb{R}^d : \Delta_c(\mathbf{x}_q) \leq \Delta_{c'}(\mathbf{x}_q),\ \forall c' \neq c \right\}, \tag{26}$$

where $\Delta_c(\mathbf{x}_q) = (\mathbf{x}_q - \hat{\boldsymbol{\mu}}_c)^\top \hat{\boldsymbol{\Sigma}}_c^{-1} (\mathbf{x}_q - \hat{\boldsymbol{\mu}}_c)$. This space is parameterized by C mean-covariance pairs $\{(\hat{\boldsymbol{\mu}}_c, \hat{\boldsymbol{\Sigma}}_c)\}_{c=1}^C$ and yields quadratic decision boundaries that account for feature correlations and scales.

Lemma C.1 (Hypothesis Space Inclusion Relation)**.** *The prototype matching hypothesis space $\mathcal{H}_{proto}$ is a proper subset of the distribution learning hypothesis space $\mathcal{H}_{dist}$:*

$$\mathcal{H}_{proto} \subset \mathcal{H}_{dist}. \tag{27}$$

Moreover, there exist infinitely many classifiers that belong to $\mathcal{H}_{dist}$ but not to $\mathcal{H}_{proto}$.

Proof. For any prototype classifier $h \in \mathcal{H}_{\text{proto}}$ with prototypes $\{\mathbf{p}_c\}$, construct a distribution learning classifier $h' \in \mathcal{H}_{\text{dist}}$ by setting $\hat{\boldsymbol{\mu}}_c = \mathbf{p}_c$ and $\hat{\boldsymbol{\Sigma}}_c = \sigma^2 \mathbf{I}_d$ (same isotropic covariance for all c). Then the Mahalanobis distance reduces to the Euclidean distance up to a constant factor, so h' defines exactly the same decision regions as h. Hence $h \in \mathcal{H}_{\text{dist}}$, i.e., $\mathcal{H}_{\text{proto}} \subseteq \mathcal{H}_{\text{dist}}$.

To see that the inclusion is proper, consider a binary classification problem ($C = 2$). Let a distribution learning classifier have parameters satisfying $\hat{\boldsymbol{\mu}}_1 = \hat{\boldsymbol{\mu}}_2$ but $\hat{\boldsymbol{\Sigma}}_1 \neq \hat{\boldsymbol{\Sigma}}_2$. Its decision boundary is a quadratic surface, which cannot be realized by any Euclidean-distance-based prototype classifier (whose decision boundary is always a hyperplane). Therefore this classifier belongs to $\mathcal{H}_{\text{dist}} \setminus \mathcal{H}_{\text{proto}}$, proving $\mathcal{H}_{\text{proto}} \subset \mathcal{H}_{\text{dist}}$. □

C.3 Bayes Risk Advantage

Theorem C.2 (Bayes Risk Advantage)**.** *Let h^* be the Bayes optimal classifier that knows the true data distributions $P(\mathbf{x}_q \mid c)$ and $P(c)$. Let h^*_{proto} be the optimal classifier within $\mathcal{H}_{proto}$ and h^*_{dist} the optimal classifier within $\mathcal{H}_{dist}$. Their Bayes risks satisfy:*

$$R(h^*) \leq R(h^*_{dist}) \leq R(h^*_{proto}). \tag{28}$$

Furthermore, if the true data distribution $P(\mathbf{x}_q \mid c)$ is such that the Bayes optimal decision boundary h^* does not belong to $\mathcal{H}_{proto}$ (i.e., the assumptions of prototype matching are violated), then the second inequality becomes strict:

$$R(h^*) \leq R(h^*_{dist}) < R(h^*_{proto}). \tag{29}$$

Proof. 1) $R(h^*) \leq R(h^*_{\text{dist}})$ follows directly from the definition of Bayes optimality: h^* minimizes the risk over all measurable classifiers, so its risk is a lower bound for any restricted class.

2) $R(h^*_{\text{dist}}) \leq R(h^*_{\text{proto}})$ follows from the hypothesis space inclusion. Lemma C.1 gives $\mathcal{H}_{\text{proto}} \subset \mathcal{H}_{\text{dist}}$, and minimizing a function over a superset cannot increase the minimum:

$$\min_{h \in \mathcal{H}_{\text{dist}}} R(h) \leq \min_{h \in \mathcal{H}_{\text{proto}}} R(h),$$

which is exactly $R(h^*_{\text{dist}}) \leq R(h^*_{\text{proto}})$.

3) Strict inequality. Assume for contradiction that $R(h^*_{\text{dist}}) = R(h^*_{\text{proto}})$. Since $\mathcal{H}_{\text{proto}} \subset \mathcal{H}_{\text{dist}}$, this equality would imply that h^*_{proto} attains the minimal risk achievable in the larger space $\mathcal{H}_{\text{dist}}$. However, the condition $h^* \notin \mathcal{H}_{\text{proto}}$ implies that there exists a classifier in $\mathcal{H}_{\text{dist}} \setminus \mathcal{H}_{\text{proto}}$ that better approximates the Bayes optimal decision boundary, thereby achieving a lower risk. This contradicts the assumption that h^*_{proto} is optimal within $\mathcal{H}_{\text{dist}}$. Hence the assumption is false and the inequality must be strict: $R(h^*_{\text{dist}}) < R(h^*_{\text{proto}})$. □

Corollary C.3. *In cross-domain few-shot learning tasks, the true data distribution $P(\mathbf{x}_q \mid c)$ almost always makes the Bayes optimal classifier h^* not belong to $\mathcal{H}_{proto}$. Consequently, the strict inequality $R(h^*_{dist}) < R(h^*_{proto})$ holds almost necessarily.*

The practical validity of this corollary stems from the inherent characteristics of cross-domain scenarios. Domain shift induces complex transformations in feature distributions $P(\mathbf{x}_q \mid c)$, where pixel-wise features of both foreground objects and background environments exhibit pronounced anisotropy. This renders the Bayes optimal decision boundary h^* a highly complex, non-linear manifold. While the prototype matching space $\mathcal{H}_{\text{proto}}$ is fundamentally limited to linear decision boundaries, the distribution learning space $\mathcal{H}_{\text{dist}}$ produces quadratic boundaries that provide substantially superior approximation capability. Consequently, under domain shift, h^* cannot be accurately represented within $\mathcal{H}_{\text{proto}}$, and by Theorem C.2, Mahalanobis matching guarantees strictly lower Bayes risk than prototype matching.

C.4 Practical Simplification

In our implementation, we adopt a simplified decision rule that relies exclusively on the Mahalanobis distance:

$$c_q^* = \arg\min_{c \in \mathcal{C}} \left[(\mathbf{x}_q - \hat{\boldsymbol{\mu}}_c)^\top \hat{\boldsymbol{\Sigma}}_c^{-1} (\mathbf{x}_q - \hat{\boldsymbol{\mu}}_c) \right]. \tag{30}$$

This rule omits the class-dependent log-determinant term $\log |\hat{\boldsymbol{\Sigma}}_c|$ and the prior term $2 \log \hat{P}(c \mid \mathcal{S})$ that appear in the full Bayesian decision rule (Eq. equation 24). The omission is justified as follows.

(1) Uniform prior: The standard few-shot evaluation protocol ensures that each episode contains a balanced set of classes with equal support sizes. Consequently, the empirical prior $\hat{P}(c \mid \mathcal{S})$ is uniform across classes and can be dropped without affecting the decision.

(2) Log-determinant term: The remaining term $\log |\hat{\boldsymbol{\Sigma}}_c|$ is class-dependent. Omitting it is not an approximation to the Bayes classifier; rather, it is a deliberate design choice rooted in two theoretical observations.

First, high estimation variance under few-shot conditions. In high-dimensional spaces ($d = 1024$), the sample covariance matrix estimated from only a handful of support examples (e.g., $\mathcal{K} = 1$ or 5) is inherently noisy. Its determinant, $\log |\hat{\boldsymbol{\Sigma}}_c|$, suffers from large sampling variability. Including a term with high variance in the decision rule can degrade generalization, as the classifier becomes sensitive to spurious fluctuations in the volume estimate. By omitting it, we obtain a more stable decision function that relies primarily on the Mahalanobis distance, which already captures the shape (correlation structure) of the class-conditional distribution.

Second, implicit regularization via the training objective. Our meta-learner is trained end-to-end to minimize the cross-entropy loss using only the Mahalanobis distance as the scoring function. The loss does not directly penalize $\log |\hat{\boldsymbol{\Sigma}}_c|$. However, if the estimated covariances differed substantially in volume across classes, the scale of Mahalanobis distances would vary systematically across classes: a class with an overly large covariance volume would tend to yield smaller squared Mahalanobis distances (because its inverse covariance matrix is scaled down in a typical sense), while a class with a very small volume would produce larger distances. Such imbalance would skew the classifier's predictions and harm classification accuracy. Therefore, the training process implicitly regularizes $\log |\hat{\boldsymbol{\Sigma}}_c|$ to stay within a moderate range and to be comparable across classes. Under this implicit regularization, omitting the log-determinant term merely discards an approximately constant additive offset, which does not affect the relative ordering of the distances.

Thus, the Mahalanobis distance alone serves as a sufficient statistic for discrimination in our framework, preserving theoretical soundness while avoiding unnecessary computational overhead.

C.5 Generalization to Classification and Detection

Although the preceding theoretical analysis is presented using cross-domain few-shot segmentation as an illustrative example, all conclusions are task-agnostic. In cross-domain few-shot classification, each entire image is treated as a single sample with its feature vector; the same anisotropic Gaussian modeling and Mahalanobis distance apply directly. In cross-domain few-shot detection, each region proposal serves as the basic unit, and the formulation remains identical. Consequently, the established results—strictly larger hypothesis space and strictly lower Bayes risk under domain shift—hold uniformly across all CD-FSL tasks. Thus, our theoretical framework provides a unified foundation for the proposed method's advantages in segmentation, classification, and detection.

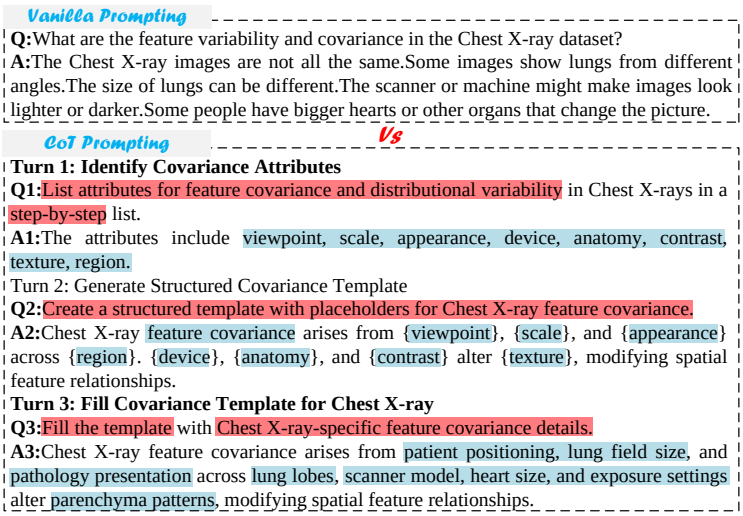


Figure 5: Illustration of the chain-of-thought (CoT) prompting strategy for foreground semantic covariance generation. Compared with vanilla prompting, multi-turn CoT decomposes semantic inference into structured steps, producing more precise descriptions of domain context and distributional variation.

D More Experiments

D.1 Comparison with Zero-Shot Methods

Large vision-language models (VLMs) provide semantic priors but often degrade under domain shift, since zero-shot prediction relies mainly on class names without target-domain visual grounding. Conversely, purely visual few-shot models may overfit limited support examples. Our framework bridges this gap by grounding semantic descriptions with few-shot visual evidence and using the resulting prompts to estimate class-wise anisotropic Gaussian distributions.

Table 7: Comparison with SAM-3 on CD-FSS.

Method	ISIC	Chest X-Ray	FSS-1000	Deepglobe	Average
SAM-3	28.1	51.8	73.0	29.2	45.5
Ours (1-shot)	**76.7**	**89.3**	**85.4**	**54.9**	**76.6**

Table 8: Comparison with Qwen3-VL on CD-FSD.

Method	DIOR	NEU-DET	UODD	Average
Qwen3-VL (zero-shot)	18.3	2.0	9.2	9.8
Ours (1-shot)	**21.3**	**12.6**	**12.5**	**15.5**
Ours (5-shot)	**34.7**	**18.9**	**16.8**	**23.5**

As shown in Table 7, our 1-shot results substantially outperform SAM-3 Carion et al. (2025) on four CD-FSS benchmarks, while our method surpasses zero-shot Qwen3-VL Bai et al. (2025) on CD-FSD even with a single support example, with further gains in the 5-shot setting. These results suggest that zero-shot generalization to unseen domains remains fundamentally limited when relying solely on textual semantics. Visual grounding from few-shot examples is therefore not merely supplementary but essential for robust cross-domain adaptation. Our framework shows that even one support example per class can dramatically improve domain generalization by anchoring abstract semantics to concrete visual evidence—a capability that zero-shot methods intrinsically lack

D.2 Further Analysis of Chain-of-Thought Prompt Strategy

The CoT prompting strategy for foreground semantic covariance generation is illustrated in Fig. 5, and a corresponding example for domain semantics generation is provided in Fig. 6.

In addition, we systematically evaluate how different CoT strategies affect performance. Table 9 compares four variants with increasing reasoning depth: (1) Ours w/o CoT: A simple question–answering baseline without explicit reasoning steps (e.g., “Describe this domain”). The LLM responds directly,

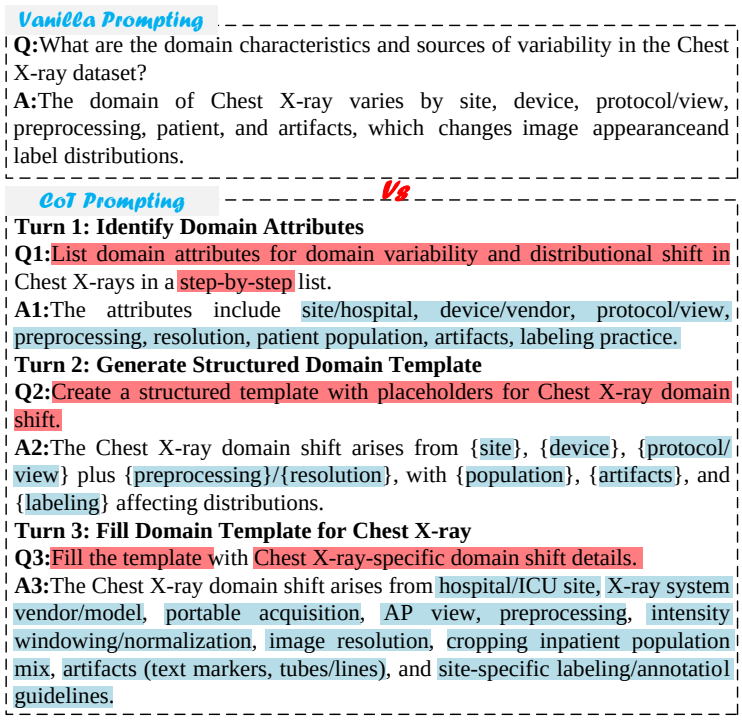


Figure 6: Illustration of the chain-of-thought (CoT) prompting strategy for domain semantics generation. Vanilla prompting yields coarse-grained descriptions that fail to capture complex domain dynamics. In contrast, our multi-turn CoT approach scaffolds LLM reasoning through step-by-step decomposition, generating precise and nuanced semantic descriptions.

Table 9: Analysis of chain-of-thought prompt strategy on $\mathcal{K}$-shot cross-domain few-shot segmentation.

	ISIC		Chest X-Ray		FSS-1000		Deepglobe		Average	
Method	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**
Baseline	41.2	53.6	66.6	70.2	78.6	81.2	37.9	43.0	56.1	62.0
Ours w/o CoT	73.4	76.1	86.4	86.9	82.8	84.2	51.7	61.4	73.6	77.2
Ours w/ Single-turn CoT	73.1	76.4	86.2	86.8	83.4	84.8	54.1	63.6	74.2	77.9
Ours w/ Multi-turn CoT with Simplified Prompts	74.3	77.6	86.8	87.3	83.7	85.4	54.2	63.9	74.8	78.6
Ours w/ Multi-turn CoT	**76.7**	**79.2**	**89.3**	**89.9**	**85.4**	**87.6**	**54.9**	**65.1**	**76.6**	**80.5**

without structured inference. (2) Ours w/ Single-turn CoT: All reasoning instructions are consolidated into a single prompt. The LLM performs domain analysis, semantic understanding, and covariance inference in one pass, offering structured reasoning but no multi-turn interaction. (3) Ours w/ Multi-turn CoT with Simplified Prompts: Reasoning is decomposed into sequential steps, each using concise prompts that retain only core instructions without detailed guidance. (4) Ours w/ Multi-turn CoT (final): Our full strategy employs comprehensive multi-turn reasoning with detailed prompts. Each turn builds on previous outputs, forming a coherent reasoning chain guided by task-specific instructions.

The results reveal key insights. First, all CoT variants outperform the baseline, confirming that semantic guidance helps compensate for limited visual evidence. Second, performance improves monotonically with CoT sophistication, validating that structured reasoning yields richer semantic descriptions. Notably, single-turn CoT suffers from context fragmentation, limiting its ability to capture nuanced relationships among domain characteristics, categories, and covariance structures. In contrast, our multi-turn approach decomposes reasoning into logically connected steps, transforming fragmented information into coherent semantic knowledge and enabling more accurate distribution estimation under domain shift.

D.3 Effect of Individual Prompt Types

We further examine the contribution of individual prompt types. Table 10 compares the full model with two variants: "Ours w/o Domain-aware Prompt," which removes the domain-specific context from the semantic prompts, and "Ours w/o Covariance Prompt," which removes the variation cues used to guide covariance estimation. As shown in the table, removing either prompt type consistently degrades performance across all four benchmarks. The drop is slightly larger for the covariance prompt on FSS-1000 and Deepglobe, where intra-class variation is more pronounced, while the domain-aware prompt contributes more on Chest X-Ray and ISIC, where domain-specific context is critical for separating foreground from background. These results confirm that domain context and variation cues play complementary roles, and that both are necessary for accurate distribution estimation under domain shift.

Table 10: Additional prompt design ablation on $\mathcal{K}$-shot CD-FSS.

	ISIC		Chest X-Ray		FSS-1000		Deepglobe		Average	
Method	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**
Ours	**76.7**	**79.2**	**89.3**	**89.9**	**85.4**	**87.6**	**54.9**	**65.1**	**76.6**	**80.5**
Ours w/o Domain-aware Prompt	72.9	75.6	85.8	86.3	81.6	83.2	50.7	60.4	72.8	76.4
Ours w/o Covariance Prompt	72.2	75.1	85.6	86.2	81.3	82.9	50.2	60.1	72.3	76.1

D.4 Comparison with Traditional Visual Prompt Tuning

We compare our meta-promptable framework against representative visual prompt tuning methods, VPT Jia et al. (2022) and MetaPrompt Wu et al. (2024). These methods introduce learnable visual prompts to adapt pre-trained models, but they still operate under the isotropic prototype assumption, lacking the ability to model complex feature distributions. In contrast, our approach adopts a promptable meta-learning paradigm: semantic prompts derived from language models guide a meta-learner to dynamically estimate class-specific anisotropic Gaussian distributions, replacing the fixed prototype with a more expressive distribution. As shown in Table 11, our method consistently outperforms both baselines. This demonstrates that a promptable, distribution-aware meta-learning strategy is more effective than traditional visual prompt tuning.

D.5 Comparison with Traditional Multimodal Fusion

We compare our framework against conventional multimodal fusion strategies, including SEVPro Cai et al. (2024) and a baseline we construct named SimpleFusion. SimpleFusion is built upon the standard prototype matching paradigm, using the same semantic prompts as our method but with simple visual-textual fusion. In contrast, our framework does not merely fuse modalities; it uses semantic prompts to guide a meta-learner in estimating class-specific anisotropic Gaussian distributions. This promptable distribution modeling effectively embeds high-level semantics into the decision process, enabling robust cross-domain generalization. As shown in Table 12, our method consistently outperforms both SimpleFusion and SEVPro. These results confirm that semantic-guided distribution modeling — not naive feature fusion — is key to bridging modality gaps in CD-FSL.

D.6 Validation of the Low-Rank Hyperparameter r

The proposed Meta-Learner module employs a low-rank factorization, reparameterizing the high-dimensional covariance matrix via two factor matrices of rank r. This hyperparameter critically governs the trade-off between the representational capacity of the covariance model and the overall parameter complexity. To determine its optimal value, we systematically evaluate the impact of r on 1-shot segmentation performance across four benchmark datasets: Deepglobe Demir et al. (2018), ISIC2018 Tschandl et al. (2018), FSS-1000 Li et al. (2020b), and Chest X-ray Candemir et al. (2013). We vary r from 5 to 500 and report the average performance and parameter count in Fig. 7. The results reveal a consistent non-monotonic relationship: performance initially improves with increasing

Table 11: Comparison with visual prompt tuning methods on 5-way K-shot CD-FSC.

				Chest		ISIC		EuroSAT		CropDisease		Average	
Method	**Backbone**	**Type**	**Mark**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**
VPT Jia et al. (2022)	ViT-Small	Prompt Tuning	ECCV-22	22.11	27.39	31.45	50.39	72.31	86.77	80.83	93.23	51.68	64.45
MetaPrompt Wu et al. (2024)	ViT-Small	Prompt Tuning	AAAI-24	22.72	27.65	32.29	50.88	74.14	87.14	81.17	93.76	52.58	64.86
Ours	ViT-Small	Proposed	-	**26.46**	**30.42**	**39.74**	**58.91**	**82.69**	**96.13**	**91.26**	**98.95**	**60.04**	**71.10**

Table 12: Comparison with multimodal fusion methods on 5-way K-shot CD-FSC.

				Chest		ISIC		EuroSAT		CropDisease		Average	
Method	**Backbone**	**Type**	**Mark**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**
SimpleFusion	ViT-Small	Multimodal Fusion	-	20.31	22.69	29.69	48.11	70.03	84.29	80.09	92.31	50.03	61.85
SEVPro Cai et al. (2024)	ViT-Small	Multimodal Fusion	IJCAI-24	23.01	26.46	33.08	47.82	71.63	89.90	80.86	94.99	52.14	64.79
Ours	ViT-Small	Proposed	-	**26.46**	**30.42**	**39.74**	**58.91**	**82.69**	**96.13**	**91.26**	**98.95**	**60.04**	**71.10**

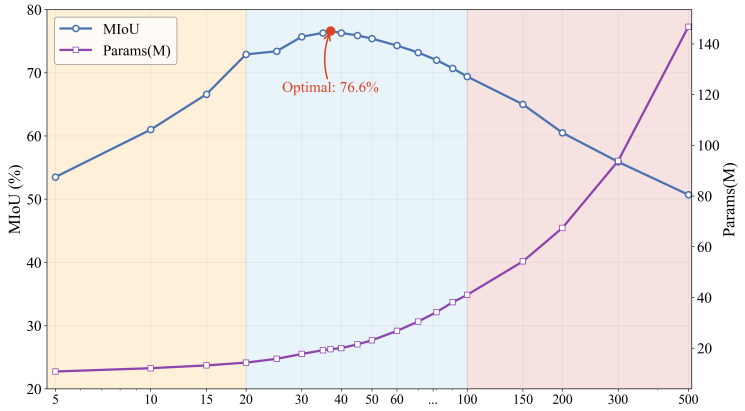


Figure 7: Analysis of LoRA Rank Selection.

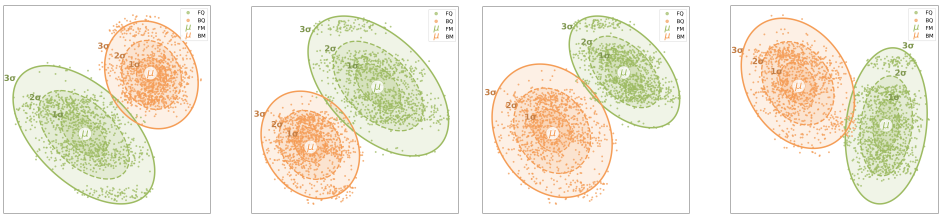
r, peaks in the range $r = 30$–40, and then gradually declines. This trend reflects a fundamental trade-off. For small values of r (e.g., $r = 5$), the limited dimensionality of the factor matrices creates a representational bottleneck, constraining the model's ability to capture the full covariance structure and leading to suboptimal performance. As r increases, the representational capacity improves. However, beyond the optimum, an excessively large r unnecessarily inflates the number of learnable parameters. This increase in model complexity heightens the risk of overfitting, particularly under the data-scarce regime of few-shot learning, which explains the subsequent performance degradation. Balancing performance and parameter efficiency, we identify $r = 30$ as the optimal configuration. This value provides sufficient representational power while maintaining a parsimonious parameterization. Consequently, we adopt $r = 30$ for all datasets and task settings in our experiments.

D.7 Further Analysis of Distribution Dynamics

This section presents additional experimental analysis investigating the distribution characteristics of real-world visual features. We randomly sample one few-shot task from each of the four cross-domain segmentation datasets, with each task consisting of a support image, a query image, and their corresponding masks. Dense visual features are extracted from both images using a visual backbone network. For the query image, we directly project the foreground and background pixel features to a two-dimensional space using t-SNE Van der Maaten & Hinton (2008) for visualization. Concurrently, we estimate the foreground and background distributions from the support image using our method, characterizing each by its mean vector and covariance matrix. These estimated distributions are subsequently visualized in the same embedding space. The distribution visualization employs confidence ellipses derived from the distribution parameters. For each category (foreground and background), we compute a two-dimensional mean and a 2×2 covariance matrix. Eigenvalue decomposition of the covariance matrix yields the principal directions and scales of the distribution. The semi-axes of the confidence ellipse are determined by the square roots of the eigenvalues: the major semi-axis equals $\sqrt{\lambda_{\max}}$ and the minor semi-axis equals $\sqrt{\lambda_{\min}}$. The orientation is defined by the eigenvector corresponding to the largest eigenvalue. Confidence levels of 1σ, 2σ, and 3σ are represented by scaling the semi-axes by factors of 1, 2, and 3 respectively, while maintaining the same center at the mean. As shown in Fig. 8, two key observations emerge. First, the query features exhibit clear anisotropic Gaussian characteristics in the embedded space. Second, and more importantly, the distributions learned from the support set effectively encompass the query feature distributions. This demonstrates our model's capability to accurately capture and generalize the underlying feature structure from limited support examples.

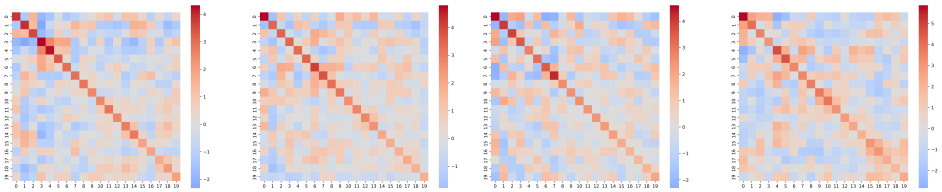
In addition, we provide additional quantitative results to elucidate the covariance structures inherent in the real data distribution. Specifically, for each of the four datasets, we randomly sample one category and compute the empirical covariance matrix from its foreground features. To facilitate visualization of these high-dimensional features, we employ PCA for dimensionality reduction, projecting them

into a 20-dimensional subspace. The resulting covariance matrices are visualized as heatmaps in Fig. 9. A key observation is the consistent presence of pronounced off-diagonal values across all datasets. This non-diagonal structure provides clear evidence of anisotropy in the feature distributions, demonstrating that our model successfully captures the complex, correlated feature relationships that a simple diagonal covariance matrix would fail to represent.



(a) ISIC (b) Chest X-ray (c) FSS-1000 (d) Deepglobe

Figure 8: Visualization of feature statistics across dimensions.



(a) ISIC (b) Chest X-ray (c) FSS-1000 (d) Deepglobe

Figure 9: Visualization of feature statistics across dimensions.

D.8 Additional Qualitative Results on Segmentation

We present comprehensive visualization results of our method across four cross-domain few-shot segmentation benchmarks, with comparative visualizations provided in Fig. 10, 11, 12, and 13. A consistent and notable improvement is observed over the prototype matching baseline across all datasets, with each domain highlighting distinct advantages of our approach. Specifically, on the ISIC2018 and Chest X-ray datasets (Figs. 11 and 13), our method achieves more precise localization of pathological regions, effectively reducing both false positives and undersegmentation. For the FSS-1000 dataset (Fig. 12), our model demonstrates enhanced capability in identifying the most semantically relevant targets, yielding predictions with sharper boundaries and more structurally complete objects. On the Deepglobe dataset (Fig. 10), our approach provides superior pixel-level classification in remote sensing imagery, more accurately capturing intricate spatial structures and land-cover details. The consistent qualitative improvements observed across a range of domains and tasks demonstrate the robustness and generalizability of our method in tackling the underlying challenges of cross-domain few-shot segmentation.

D.9 Additional Qualitative Results on Object Detection

We present qualitative detection results across multiple datasets with substantial domain diversity, including ArTaxOr Drange (2019), Clipart1k Inoue et al. (2018), DIOR Li et al. (2020a), DeepFish Saleh et al. (2020), NEU-DET Song & Yan (2013), and UODD Jiang et al. (2021). For each dataset, several images are randomly selected, and we display the detection results of our method (in blue), the baseline method CD-ViTO Fu et al. (2025) (in red), and the ground-truth annotations (in green).

The visualization results demonstrate that our proposed method achieves superior detection performance compared to the baseline, producing results that more closely align with the ground truth. Specifically, as illustrated in Fig. 14, Fig. 15, and Fig. 19, in complex scenarios characterized by a large number of objects with varying distances and scales, our method identifies more object

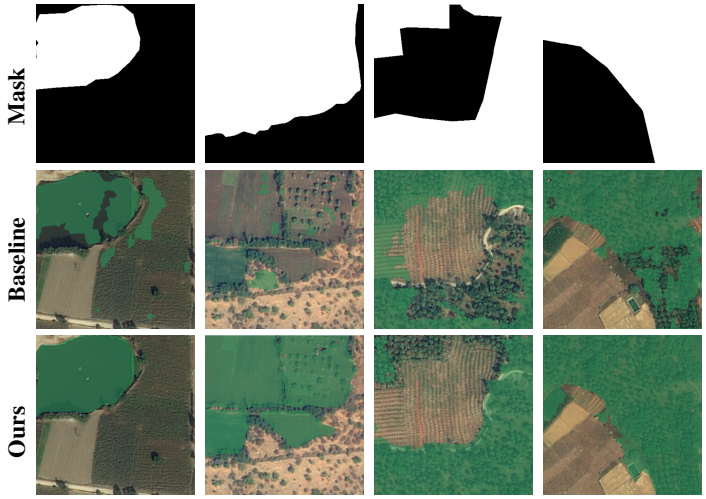


Figure 10: Segmentation results on Deepglobe dataset.

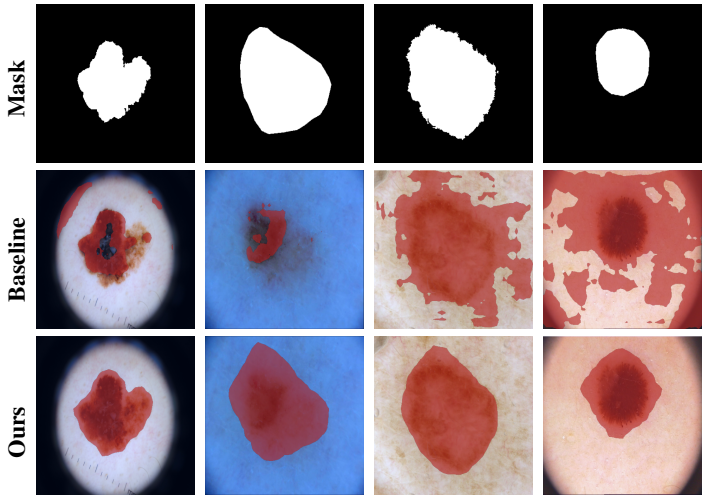


Figure 11: Segmentation results on ISIC2018 dataset.

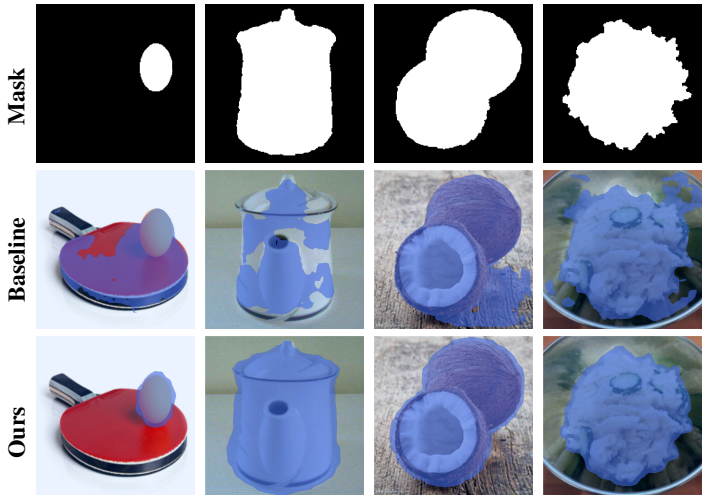


Figure 12: Segmentation results on FSS1000 dataset.

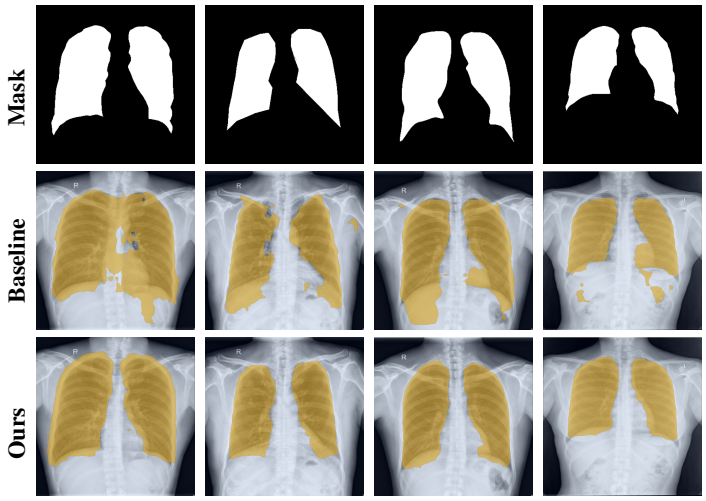


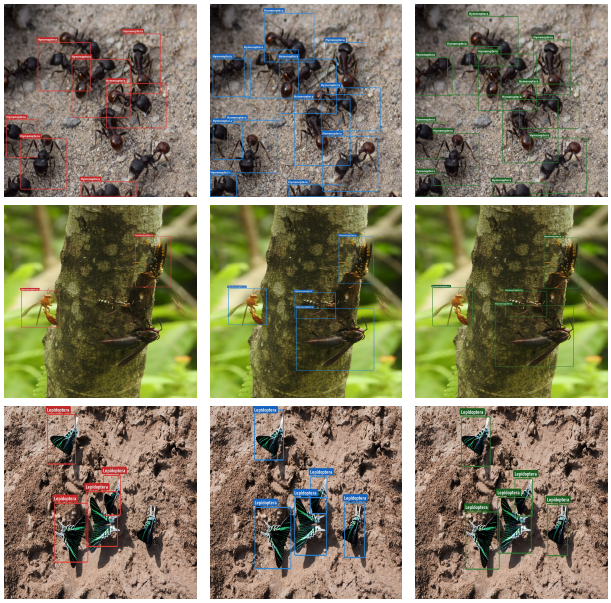
Figure 13: Segmentation results on Chest X-ray dataset.

instances than the baseline, exhibiting a particular advantage in small object detection. This suggests that, in contrast to the traditional prototype-based modeling approach employed by CD-ViTO, our distribution-based modeling paradigm possesses stronger semantic representation capabilities, thereby facilitating more discriminative object recognition. Besides, as shown in Fig. 16 and Fig. 17, in underwater environments with low visibility, complex backgrounds, and numerous distractors, our method continues to achieve more accurate object detection compared to the baseline, demonstrating enhanced environmental adaptability and robustness. And, as exemplified in Fig. 18, in remote sensing scenarios characterized by small object scales and dense distributions, our method successfully detects more true objects even when some targets are difficult to discern with the naked eye, further validating the advantages of distribution-based modeling in small object perception and fine-grained semantic representation.

Notably, the datasets employed in this study encompass significantly diverse domains, including cartoon, underwater, and aerial remote sensing, which collectively validate the cross-domain generalization capability of our proposed method. In contrast to the baseline method based on traditional prototype modeling that relies on a single centroid representation, our anisotropic Gaussian distribution-based modeling approach more comprehensively captures the variation in category semantics, thereby exhibiting stronger adaptability when confronted with domain shifts across different scenarios.

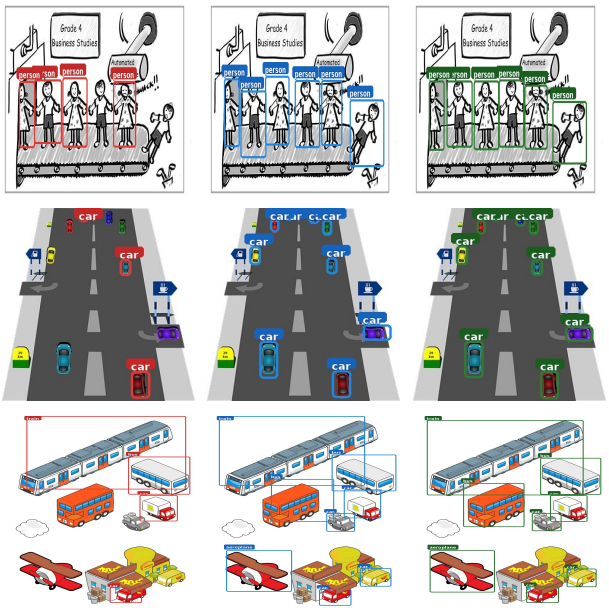
E Limitations and Future Work

Despite the online efficiency of our framework (0.45s per task), the offline generation of semantic prompts incurs a non-trivial computational overhead (2.23s per task), stemming from vision-language captioning and chain-of-thought reasoning. While this cost can be amortized when domain-aware prompts are reused, per-task prompt generation remains a bottleneck for large-scale evaluation or real-time deployment with many distinct classes. Furthermore, even with rich semantic descriptions, a single support example provides only a sparse visual anchor; the estimated anisotropic Gaussian distribution inevitably under-represents the true intra-class variability, particularly in domains with high diversity. This mismatch may lead to overconfident predictions and reduced generalization for query samples that deviate substantially from the observed support instance. Future work will investigate lightweight, task-specific adapters for efficient on-device prompt generation and semantic-aware augmentation that synthesizes plausible support-image variations from prompt descriptions, enriching visual evidence without additional labels.



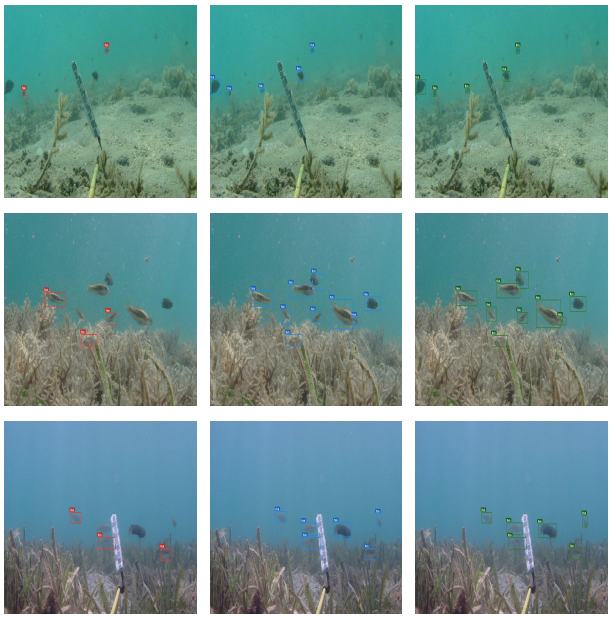
Baseline Baseline+Ours Ground Truth

Figure 14: Visualization of detection effects on ArTaxOr dataset.



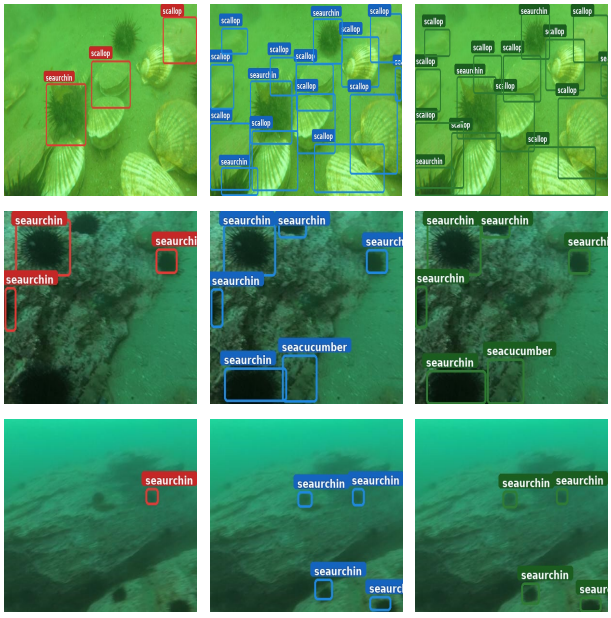
Baseline Ours Ground Truth

Figure 15: Visualization of detection effects on Clipart1k dataset.



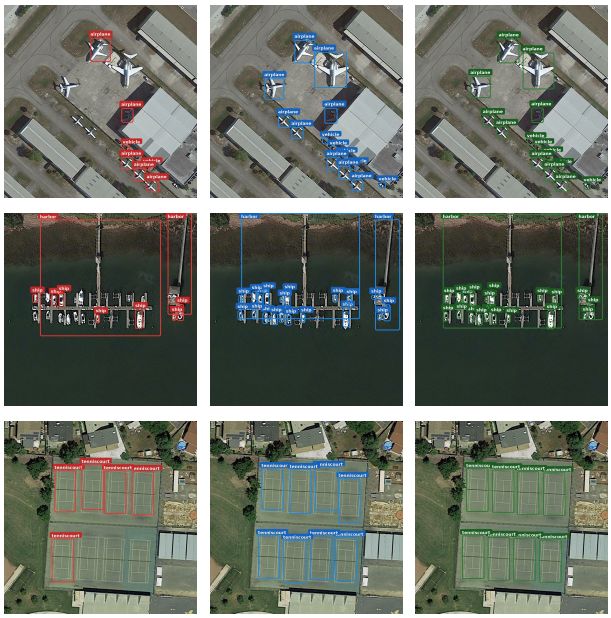
Baseline Ours Ground Truth

Figure 16: Visualization of detection effects on DeepFish dataset.



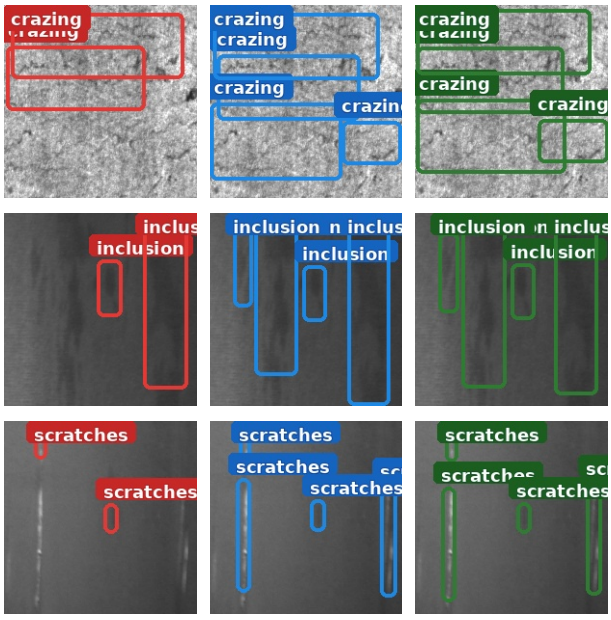
Baseline Ours Ground Truth

Figure 17: Visualization of detection effects on UODD dataset.



Baseline Ours Ground Truth

Figure 18: Visualization of detection effects on DIOR dataset.



Baseline Ours Ground Truth

Figure 19: Visualization of detection effects on NEU-DET dataset.