STAR: A SPECTRAL RESONANCE FRAMEWORK FOR TRAINING-FREE MULTI-DOMAIN LEARNING WITH CLIP

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-domain learning (MDL) aims to learn a model that performs reliably across multiple domains. As a powerful vision-language foundation model, CLIP is a promising basis for MDL due to its large-scale pretraining. To better exploit CLIP's pretrained priors, MDL calls for a different design instead of existing heavy architectures and complex training strategies. In this paper, we study multi-domain CLIP features from both spectral and Gaussian geometric perspectives. Our analyses first reveal that dominant class geometry remains domain-dependent, while nuisance-sensitive variations concentrate in spectral-tail directions, providing a reliable basis for self-domain recognition. More importantly, we find that self-domain failures can be complemented by evidence from other domains, and their recoverability is strongly associated with self-domain Gaussian geometry. Based on these findings, we propose *SpecTral Adaptive Resonance* (STAR), a training-free framework for CLIP-based MDL. STAR first establishes a reliable self-domain base through *Subspace Anchor Guidance* (SAG), which preserves dominant class-domain spectral structure while suppressing nuisance-sensitive directions. On top of this base, *CrOss-domain REctification* (CORE) exploits complementary cross-domain spectral evidence under self-domain Gaussian compatibility to enhance self-domain predictions. Extensive experiments show that STAR consistently outperforms previous MDL baselines, and further analyses validate the effectiveness of our design choices. By building reliable self-domain anchors and selectively resonating with complementary cross-domain structures, STAR offers a new geometric perspective on CLIP-based MDL.

1 INTRODUCTION

Multi-domain learning (MDL) aims to learn a model that maintains reliable performance across domains while achieving strong average performance over the full set (Joshi et al., 2012; Schoenauer Sebag et al., 2019; Liu et al., 2017), which is of great importance that real-world systems often face domain shifts that can degrade robustness and generalization across environments (Gulrajani & Lopez-Paz, 2021; Koh et al., 2021). As a powerful vision-language foundation model, CLIP (Radford et al., 2021) serves as the basis for a broad range of subsequent methods across visual tasks (Zhang et al., 2024; Chen et al., 2023; Awais et al., 2025), and its rich semantic knowledge acquired during pretraining makes it a promising foundation for MDL. In particular, MDL poses a challenge for CLIP: the model must retain its text-aligned semantic discrimination while accounting for domain-specific visual structures. It highlights the need for a CLIP-based solution that can reconcile semantic consistency with domain-specific structure in MDL.

Before the emergence of CLIP, MDL had already been extensively studied in visual learning, which can be grouped as: (1) parameter-allocation (Ma et al., 2018; Tang et al., 2020; Rebuffi et al., 2017; 2018; Iakovleva et al., 2023; Li & Vasconcelos, 2019; Wang et al., 2023), which balance commonality and domain specificity through shared and domain-specific parameters; (2) feature-decomposition (Liu et al., 2019), which aligns shared and separating domain-specific information in the feature space; and (3) optimization-level methods (Schoenauer Sebag et al., 2019; Wang et al., 2024a), which address MDL through dedicated optimization strategies. Despite their effectiveness,

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.

these methods often rely on heavy architectures or complex optimization, making them less suited to CLIP, whose pretrained priors can be better exploited through lightweight feature-space refinement.

To gain a deeper understanding of geometric structure in multi-domain CLIP features, we conduct preliminary analyses from spectral and Gaussian geometry perspectives. We first examine how spectral structure supports reliable self-domain recognition, and then investigate whether remaining self-domain failures can be complemented by cross-domain evidence. This leads to two key findings: (1) nuisance-sensitive variations are more concentrated in spectral-tail directions, while dominant components preserve stable semantic structures but remain domain-dependent; (2) for the remaining self-domain failures, cross-domain evidence provides substantial complementary information, while recoverable and unrecoverable cases exhibit distinct patterns in self-domain Gaussian geometry. These findings lead to an effective training-free framework for CLIP-based MDL, termed Spectral Adaptive Resonance (STAR). Specifically, STAR constructs adaptive class-domain subspaces by retaining only the dominant singular directions, forming self-domain Subspace Anchor Guidance (SAG). Building on this reliable self-domain base, Gaussian geometry-guided CrOss-domain REctification (CORE) adaptively exploits complementary evidence from other domains. Together, they integrate reliable self-domain structure with cross-domain complementarity for downstream recognition. Extensive experiments in Sec. 5 demonstrate STAR as an efficient training-free solution for CLIP-based MDL. Across diverse MDL scenarios, STAR substantially improves zero-shot CLIP, matching or surpassing training-based baselines while consistently boosting their performance when combined.

Our main contributions can be summarized as follows:

- We uncover that in spectral geometry, multi-domain nuisance-sensitive variations concentrate in spectral-tail directions, while dominant class geometry remains domain-dependent.
- We reveal that cross-domain evidence complements remaining self-domain failures, while recoverable and unrecoverable cases exhibit distinct self-domain Gaussian geometry.
- Following these findings, we propose Spectral Adaptive Resonance (STAR), a training-free approach for CLIP-based MDL that models self-domain subspace anchor guidance and enhances it through Gaussian geometry-guided cross-domain rectification.
- Extensive experiments demonstrate the effectiveness and general applicability of STAR, with comprehensive analyses validating the proposed design principles.

2 Related Works

2.1 Multi-Domain Learning

Multi-domain learning (MDL) aims to learn a unified model that performs effectively across multiple domains. In visual MDL, such domains may exhibit distributional shifts arising from variations in style, imaging conditions, or environments. Methods such as MMoE (Ma et al., 2018) and PLE (Tang et al., 2020) model domain relationships through mixtures of shared and domain-specific experts together with gating mechanisms. Another representative direction (Rebuffi et al., 2017; 2018; Iakovleva et al., 2023) studies heterogeneous MDL settings by inserting domain-specific adapters into a shared backbone to achieve specialization. Some works address MDL from the perspective of optimization strategy or overall framework construction, for example through adversarial learning (Schoenauer Sebag et al., 2019), decoupled training (Wang et al., 2024a), or neural architecture search (Wang et al., 2023). JOARN(Liu et al., 2019) learns domain-aware representations by structuring shared and private feature spaces, while CovNorm improves efficiency through covariance-normalized lightweight adaptation layers (Li & Vasconcelos, 2019). Different from the above methods that rely on complex architectures or optimization procedures, we construct a training-free MDL framework from the perspective of spectral and Gaussian geometry to better exploit CLIP's pretrained knowledge.

2.2 CLIP-based Training-Free Adaptation

CLIP-based training-free adaptation aims to adapt pretrained CLIP to downstream tasks by directly exploiting its pretrained knowledge on support-set statistics without back-propagation. Cache-based methods constitute a major line of research in this area. Tip-Adapter (Zhang et al., 2022) introduces a key-value cache model built from training samples to inject support-set information into predictions,

Tip-X (Udandarao et al., 2023) replaces direct image-to-image matching with a comparison of the class-response distributions induced by the text classifier, APE (Zhu et al., 2023) refines the prior through the affinities between the test image and the cache-text representations, ProKeR (Bendou et al., 2025) incorporates a proximal global regularization induced by the zero-shot predictor in a reproducing kernel Hilbert space. Apart from those, GDA-CLIP (Wang et al., 2024b) builds a closed-form linear classifier under Gaussian discriminant analysis, while TIMO (Li et al., 2025) integrates bidirectional image-text interaction into the classifier construction process. Different from those that overlook domain relationships, we construct the framework from an MDL perspective by establishing reliable class-domain spectral anchors and then exploiting cross-domain complementarity.

3 Exploring Visual Evidence in Multi-Domain Features

3.1 Exploring Spectral Sensitivity and Domain Dependence

Since CLIP visual features often exhibit structured spectral concentration, we conduct a preliminary analysis from a spectral-geometry perspective on FMoW and DomainNet, which exhibit markedly different forms of inter-domain discrepancy: FMoW reflects real-world appearance variations across *geographic regions*, whereas DomainNet captures stylistic variations induced by different *artistic rendering domains*. Let x^{base} denote the original image, we define the two augmented views as

$$\tilde{x}^{\text{str}}(u,v) = (1-\rho)\, x^{\text{base}}(u,v) + \rho\, \mathcal{T}_{\text{AdaIN}}\big(x^{\text{base}},\, x^{\text{ref}}_{d'}\big)(u,v), \tag{1}$$

$$\tilde{x}^{\text{inst}}(u,v) = \begin{cases} \lambda\, x^{\text{base}}(u,v), & (u,v) \in \Omega_c, \\ x^{\text{base}}(u,v), & (u,v) \notin \Omega_c, \end{cases} \tag{2}$$

where $x^{\text{ref}}_{d'}$ is sampled from a different domain $d' \neq d$, $\mathcal{T}_{\text{AdaIN}}(\cdot,\cdot)$ denotes the AdaIN style transfer (Huang & Belongie, 2017), ρ controls the blending ratio which is set to 0.15 in our experiments, Ω_c is a centered rectangular region, and $\lambda > 1$ is a brightness scaling factor. For each class-domain pair (c,d), we collect its corresponding visual features into a matrix $\mathbf{X}_{c,d} \in \mathbb{R}^{N_{c,d} \times D}$ and compute

$$\mathbf{X}_{c,d} = \mathbf{U}_{c,d}\mathbf{\Sigma}_{c,d}\mathbf{V}_{c,d}^{\top}, \quad \mathbf{B}^{\text{head}}_{c,d} = \mathbf{V}_{c,d}[:, 1 : K^{\text{head}}_{c,d}], \quad \mathbf{B}^{\text{tail}}_{c,d} = \mathbf{V}_{c,d}[:, K^{\text{head}}_{c,d} + 1 : R_{c,d}], \tag{3}$$

where $R_{c,d}$ is the rank of $\mathbf{X}_{c,d}$, the singular values are sorted in descending order, and we split the spectrum into two halves, i.e., $K^{\text{head}}_{c,d} = \lceil R_{c,d}/2 \rceil$ and $K^{\text{tail}}_{c,d} = R_{c,d} - K^{\text{head}}_{c,d}$.

We then define the metrics by the *reference-normalized projection score* (RNPS) as follows:

$$\text{RNPS}_r(\tilde{x}; c, d) = \frac{E_r(\tilde{x}; c, d) - \bar{E}^{\text{ref}}_r(c,d)}{\max\big(\bar{E}^{\text{ref}}_r(c,d),\, \epsilon\big)}, \qquad E_r(\tilde{x}; c, d) = \Big\| \big(\mathbf{B}^{r}_{c,d}\big)^{\top} f(\tilde{x}) \Big\|_2^2, \tag{4}$$

where $f(\cdot)$ is the normalized zero-shot CLIP visual encoder and $\bar{E}^{\text{ref}}_r(c,d)$ is the mean original projection energy of the reference class-domain pair (c,d) on subspace $r \in \{\text{head}, \text{tail}\}$.

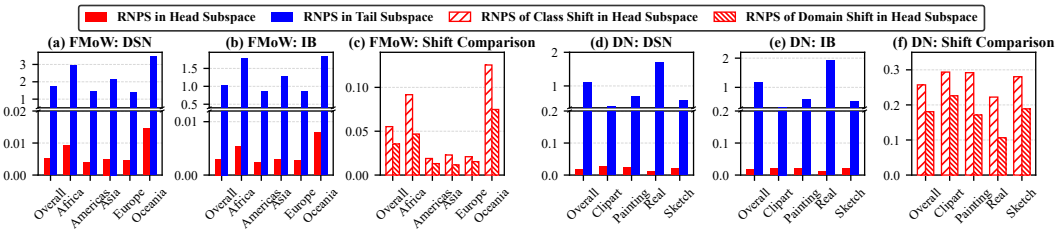

Figure 1: Analysis of spectral selectivity and domain dependence. DSP denotes Domain-related Structured Nuisance, and IB denotes Instance-specific Bias. The vertical axis reports the RNPS value.

Dominant Subspaces are Less Nuisance-Sensitive but Remain Domain-Dependent. We compare the RNPS of the head and tail subspaces under *domain-related structured nuisance* and *instance-specific bias*, using the cross-domain style transfer and local-intensity perturbation in equation 1 and equation 2, respectively. Across FMoW and DomainNet, the head subspace exhibits substantially smaller RNPS than the tail subspace, indicating that nuisance sensitivity is more pronounced in the

spectral tail. We further assess the domain dependence of the dominant geometry by comparing head-subspace RNPS under *domain shift* and *class shift*. The former uses original samples from *the same class but different domains*, whereas the latter uses samples from *different classes within the same domain*. Domain shift induces slightly smaller RNPS than class shift, but both are of the same order of magnitude, indicating that the dominant class geometry remains domain-dependent. *Together, these findings motivate retaining dominant spectral directions within class-domain-specific subspaces, preserving relatively stable visual structure while accounting for domain-dependent geometry.*

3.2 Exploring Cross-Domain Complementarity

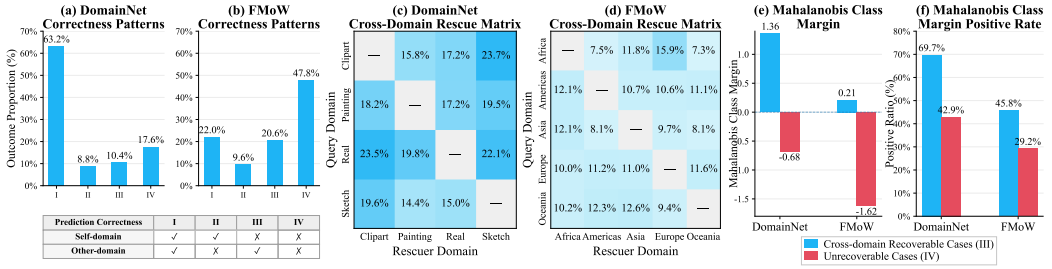

Figure 2: Analysis of cross-domain complementarity on DomainNet and FMoW.

Based on the class-domain subspaces in Sec. 3.1, we further use frozen CLIP features of 16-shot support sets to analyze how domain-specific head subspaces interact on the validation sets. Firstly, we categorize samples according to the prediction patterns of the self-domain and other-domain subspaces' projection energy, as illustrated in Fig. 2(a,b). Specifically, pattern III corresponds to cross-domain recoverable cases where the self-domain subspace fails, whereas pattern IV represents cases where all subspaces make incorrect predictions. To further characterize the distribution of cross-domain compensation, we compute the rescue matrix in Fig. 2(c,d). Each entry records the proportion of self-domain errors from a query domain that can be corrected by a rescuer domain subspace. Lastly, we investigate the discrepancy of patterns III and IV based on their Mahalanobis distances. Specifically, we compute the Mahalanobis Class Margin (MCM):

$$\mathrm{MCM}(x) = Q(z, w_x, d_x) - Q(z, y, d_x)\,, \tag{5}$$

where d_x denotes the query domain of sample x, y is the ground-truth class, and w_x denotes the incorrectly predicted class of the self-domain branch. The function $Q(z, c, d_x)$ denotes the squared Mahalanobis distance between normalized feature z and the normalized prototype of class c in the query domain d_x, using a shared covariance estimated from within-class residuals across all domains.

Self-domain Gaussian geometry guides cross-domain compensation. As shown in Fig. 2(a,b), among the self-domain failure cases that form the focus of our study, a substantial fraction can be corrected by at least one other-domain subspace (pattern III), which motivates distinguishing cross-domain recoverable errors from fully unrecoverable ones before incorporating cross-domain evidence. Fig. 2(c,d) further show that, although some domain pairs exhibit slightly stronger rescue relationships, the rescue rates are broadly distributed across different domains. Since complementary information can emerge from different domains for different samples, cross-domain compensation should account for the potential contributions of all other domains rather than pre-selecting a single auxiliary one. Finally, Fig. 2(e,f) compares the mean and positive ratio of the self-domain MCM. Although both patterns are misclassified by the self-domain subspace, they exhibit a clear discrepancy in the query-domain Gaussian geometry. This suggests that self-domain Gaussian discrepancy can serve as a reliability signal for modulating aggregated other-domain energy, preserving complementary cross-domain evidence while suppressing unreliable transfer. *Together, these findings motivate a reliability-aware cross-domain compensation strategy that aggregates complementary evidence across domains while suppressing compensation from geometrically unreliable transfers.*

4 Method

This paper investigates Multi-Domain Learning (MDL), which aims to adapt the model to multiple domains under a shared label space despite domain-specific distribution shifts. Given multiple

domains $\mathcal{D} = \{d_t\}_{t=1}^{T}$ with $d_t = \{(x_i^t, y_i^t)\}_{i=1}^{n_t}$, where T is the number of domains and n_t is the sample size of domain t. For CLIP, the goal of MDL is to improve CLIP's semantic discriminability under domain-specific visual variations, thereby enhancing performance on each domain d_t.

Different from existing MDL methods (Tang et al., 2020; Ma et al., 2018) relying on task-adaptive components or dedicated training strategies, we formulate a subspace-based training-free framework. It is architecture-agnostic and can be seamlessly plugged into any CLIP-based MDL framework.

4.1 Method Overview

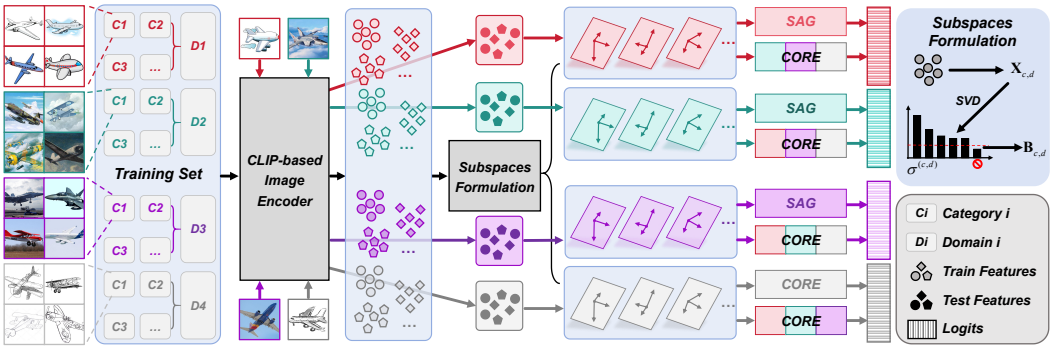

Figure 3: Overall pipeline of STAR. Along the build route, class-domain features are extracted and used to construct truncated subspaces via SVD. Along the test route, the test feature is matched with the constructed subspaces to derive SAG and CORE scores, which are combined through adaptive weighting to obtain the final classification logits.

This paper proposes Spectral Adaptive Resonance (STAR), a training-free framework for CLIP-based MDL, motivated by the findings in Sec. 3. The overall pipeline is illustrated in Fig. 3. First, we construct class-domain subspaces via singular value decomposition (SVD) to characterize domain-dependent geometric structures in CLIP representations. Based on these subspaces, STAR derives Subspace Anchor Guidance (SAG) from self-domain spectral resonance, establishing a reliable recognition base. Building on this base, CrOss-domain REctification (CORE) adaptively exploits complementary resonance from non-target domains under self-domain Gaussian geometry. Finally, SAG and CORE are integrated with CLIP predictions to obtain the final classification logits.

4.2 Subspace Anchor Guidance

For each labeled image x_i with class label $y_i \in \mathcal{C}$ and domain label $d_i \in \mathcal{D}$, we extract a CLIP-based normalized visual feature $\mathbf{z}_i = f_{d_i}(x_i)$ with the corresponding domain-specific expert or zero-shot CLIP, and group the features by each class-domain pair (c, d) to obtain $\mathcal{V}_{c,d} = \{\mathbf{z}_i \mid y_i = c,\ d_i = d\}$. Then we stack the features in $\mathcal{V}_{c,d}$ row-wise into a matrix $\mathbf{X}_{c,d} \in \mathbb{R}^{N_{c,d} \times D}$ and perform thin SVD:

$$\mathbf{X}_{c,d} = \mathbf{U}_{c,d}\mathbf{\Sigma}_{c,d}\mathbf{V}_{c,d}^{\top} = [\mathbf{u}_1\ \mathbf{u}_2\ \cdots\ \mathbf{u}_r]\,\mathrm{diag}(\sigma_1^{(c,d)}, \sigma_2^{(c,d)}, \ldots, \sigma_r^{(c,d)})\,[\mathbf{v}_1\ \mathbf{v}_2\ \cdots\ \mathbf{v}_r]^{\top}. \quad (6)$$

where $N_{c,d} = |\mathcal{V}_{c,d}|$, $r = \mathrm{rank}(\mathbf{X}_{c,d})$, and D denotes the dimension of visual feature.

Instead of fixing the subspace dimension, we choose it adaptively from the singular value spectrum of each class-domain feature matrix. As shown in Sec.3, the leading singular directions capture the dominant class-domain structure, and their large singular values indicate that class-related variation is concentrated along these directions. In contrast, the trailing directions associated with small singular values are more strongly influenced by background noise and other unstable variations. We therefore retain only the smallest number of components whose cumulative spectral energy exceeds a threshold τ. We then define $K_{c,d}$ as the smallest integer satisfying:

$$\left(\sum_{j=1}^{K_{c,d}} \left(\sigma_j^{(c,d)}\right)^2\right) \Big/ \left(\sum_{j} \left(\sigma_j^{(c,d)}\right)^2\right) \geq \tau. \quad (7)$$

This criterion yields an adaptive estimate of the effective rank, preserving the dominant class-domain variations without introducing a manually fixed dimension. The domain subspace basis is defined as:

$$\mathbf{B}_{c,d} = \mathbf{V}_{c,d}[:, 1:K_{c,d}] = \left[\mathbf{v}_1^{(c,d)} \;\; \mathbf{v}_2^{(c,d)} \;\; \cdots \;\; \mathbf{v}_{K_{c,d}}^{(c,d)}\right] \in \mathbb{R}^{D\times K_{c,d}}, \quad \mathbf{B}_{c,d}^\top \mathbf{B}_{c,d} = \mathbf{I}. \tag{8}$$

Given a test sample x with known domain label d, we extract its normalized in-domain visual feature $\mathbf{z} = f_d(x)$. We then compute the class-wise projection energies onto all class-domain subspaces in domain d, and form the SAG logits as

$$\boldsymbol{s}^{\mathrm{SAG}}(\mathbf{z}, d) = [s_1(\mathbf{z}, d),\, s_2(\mathbf{z}, d),\, \cdots,\, s_C(\mathbf{z}, d)] = \left[\|\mathbf{B}_{1,d}^\top \mathbf{z}\|_2^2,\, \|\mathbf{B}_{2,d}^\top \mathbf{z}\|_2^2,\, \cdots,\, \|\mathbf{B}_{C,d}^\top \mathbf{z}\|_2^2\right]. \tag{9}$$

Since $\mathbf{B}_{c,d}$ is orthonormal, $s_c(\mathbf{z}, d)$ is exactly the squared norm of the orthogonal projection of $\mathbf{z}$ onto the subspace spanned by the dominant directions of class c in domain d. A larger value indicates stronger spectral resonance between $\mathbf{z}$ and the corresponding dominant geometry.

The text branch and the SAG branch produce class-wise scores in different numeric spaces. While the former follows CLIP's similarity-based logit geometry with learned temperature scaling, SAG serves as a self-domain spectral anchor through projection-based resonance. Direct fusion is therefore suboptimal due to scale mismatch. To make the spectral anchor compatible with the text branch while preserving its relative class ranking, we perform sample-wise moment matching on the class score vectors. For a test sample with text similarity scores $\boldsymbol{s}^{\text{text}}$, let

$$\mu_{\text{text}} = \mathrm{Mean}(\boldsymbol{s}^{\text{text}}), \quad \varsigma_{\text{text}} = \mathrm{Std}(\boldsymbol{s}^{\text{text}}), \quad \mu_{\mathrm{SAG}} = \mathrm{Mean}(\boldsymbol{s}^{\mathrm{SAG}}), \quad \varsigma_{\mathrm{SAG}} = \mathrm{Std}(\boldsymbol{s}^{\mathrm{SAG}}). \tag{10}$$

We then map the SAG logits to the text-score distribution by Z-score Calibration:

$$\boldsymbol{\ell}^{\mathrm{SAG}} = \mathrm{ZCal}\left(\boldsymbol{s}^{\mathrm{SAG}}, \boldsymbol{s}^{\text{text}}\right) = T_{\mathrm{CLIP}} \left(\frac{\boldsymbol{s}^{\mathrm{SAG}} - \mu_{\mathrm{SAG}}}{\varsigma_{\mathrm{SAG}} + \varepsilon}\varsigma_{\text{text}} + \mu_{\text{text}}\right), \tag{11}$$

where ε is a small constant for numerical stability, and T_{CLIP} denotes the learned logit scale in CLIP. This calibration aligns the SAG branch with the per-sample score geometry of the text branch and places it under the same learned temperature, yielding a numerically compatible fusion space.

4.3 CROSS-DOMAIN RECTIFICATION

Although the SAG branch provides domain-specific geometric evidence, it still leaves a subset of samples incorrectly classified, some of which can be recovered by complementary signals from other domains. Building upon SAG, CORE therefore further exploits cross-domain subspaces to recover these residual self-domain failures.

As discussed in Sec. 3.2, complementary signals are not dominated by a single auxiliary domain. Therefore, we aggregate the responses from all non-query domains instead of selecting a specific auxiliary domain. Given a query sample x from domain q, we compute the projection energy of class c on every other-domain subspace and obtain the aggregated cross-domain evidence (CDE):

$$\bar{E}_{\mathrm{CDE}}(x, c) = \frac{1}{|\mathcal{D}| - 1} \sum_{\substack{d\in\mathcal{D} \\ d\neq q}} E_d(x, c), \tag{12}$$

where $E_d(x, c)$ denotes the projection energy onto the class-c subspace constructed from domain d.

However, directly applying CDE may introduce inaccurate compensation, as CDE contributes unequally across samples. As observed in Sec. 3.2, recoverable and unrecoverable self-domain failures exhibit distinct self-domain Gaussian geometries. This indicates a positive correlation between self-domain Gaussian compatibility and CORE effectiveness, making Gaussian geometry a natural reliability signal for modulating CDE.

Accordingly, we quantify the self-domain Gaussian compatibility (GC) using Mahalanobis distance:

$$g_q(x, c) = -Q(\mathbf{z}_x, c, q), \tag{13}$$

where a larger $g_q(x, c)$ indicates stronger compatibility between the query feature and the corresponding self-domain class distribution.

Table 1: Full-supervised accuracy (%) on all evaluated datasets. Dataset-level averages are underlined; best results within each method block are shown in **bold**; rows with **+ours** are shaded gray.

Model			DomainNet					Office-Home					FMoW					
	Method	+ours	Clipart	Painting	Real	Sketch	Avg.	Art	Clipart	Product	Real	Avg.	Asia	Europe	Africa	America	Oceania	Avg.
ZS			74.25	68.81	83.46	65.91	73.11	84.15	70.39	90.00	87.10	82.91	17.84	22.26	15.66	20.09	24.00	19.97
		✓	**81.84**	**77.88**	**87.79**	**74.50**	**80.50**	**87.17**	**83.22**	**96.60**	**92.15**	**89.78**	**48.26**	**47.28**	**54.32**	**47.17**	**67.13**	**52.83**
CoOp	jointly		76.91	73.45	86.74	69.47	76.64	87.55	77.52	94.36	89.78	87.31	41.78	42.63	42.02	42.56	45.51	42.90
	jointly	✓	**81.69**	**77.79**	**87.86**	**74.57**	**80.48**	**89.06**	**83.22**	**95.74**	**92.90**	**90.23**	**49.87**	**47.83**	**53.78**	**47.83**	**62.89**	**52.44**
	separately		79.49	75.92	87.52	71.19	78.53	87.73	76.75	93.94	90.00	87.11	50.70	45.30	46.98	45.71	56.32	49.01
	separately	✓	**82.07**	**77.96**	**87.87**	**74.26**	**80.54**	**89.81**	**82.46**	**95.85**	**92.47**	**90.15**	**51.53**	**48.63**	**56.13**	**48.46**	**62.20**	**53.39**
	MMoE		78.19	74.85	87.33	70.75	77.78	87.13	76.21	93.51	91.08	86.98	49.69	44.90	46.17	45.55	51.42	47.55
	MMoE	✓	**82.09**	**77.88**	**87.84**	**74.13**	**80.48**	**88.87**	**82.89**	**96.38**	**93.01**	**90.29**	**51.49**	**48.79**	**55.98**	**48.37**	**59.42**	**52.81**
	PLE		79.22	75.22	87.59	70.89	78.23	88.68	76.64	93.94	90.53	87.45	51.55	45.13	47.16	46.02	56.32	49.23
	PLE	✓	**82.02**	**77.72**	**87.80**	**74.15**	**80.42**	**88.68**	**82.68**	**95.80**	**92.47**	**89.91**	**51.48**	**48.93**	**56.36**	**48.51**	**61.50**	**53.36**
MaPLe	jointly		79.67	75.78	87.58	72.29	78.83	**87.55**	77.74	93.94	90.64	87.47	58.40	61.38	61.27	60.99	64.79	61.37
	jointly	✓	**83.51**	**78.51**	**88.11**	**76.04**	**81.54**	**90.38**	**83.66**	**95.96**	**92.90**	**90.73**	**66.02**	**63.63**	**67.17**	**65.05**	**78.71**	**68.11**
	separately		81.34	77.86	88.29	73.53	80.25	88.30	77.08	94.58	91.08	87.75	64.31	63.72	63.38	63.45	70.02	64.97
	separately	✓	**83.03**	**78.85**	**88.26**	**75.76**	**81.47**	**88.68**	**83.22**	**96.38**	**92.80**	**90.27**	**65.55**	**64.65**	**68.06**	**65.51**	**75.02**	**67.76**
	MMoE		79.81	76.19	87.74	72.45	79.06	86.61	75.33	93.51	89.78	86.31	61.77	61.37	61.39	60.78	67.09	62.48
	MMoE	✓	**83.13**	**78.55**	**88.14**	**75.62**	**81.36**	**88.68**	**83.22**	**95.85**	**93.01**	**90.19**	**65.04**	**62.82**	**67.64**	**63.67**	**74.69**	**66.77**
	PLE		81.59	77.94	88.32	74.16	80.50	87.92	79.61	94.47	91.50	88.38	66.88	63.85	63.60	63.46	73.85	66.33
	PLE	✓	**83.30**	**78.86**	**88.30**	**76.00**	**81.61**	**88.68**	**83.04**	**95.82**	**93.12**	**90.16**	**66.45**	**64.68**	**70.33**	**65.68**	**80.50**	**69.53**

We then couple this Gaussian compatibility with the aggregated cross-domain response:

$$s^{\mathrm{CORE}}(x,c) = \left(\bar{E}_{\mathrm{CDE}}(x,c)\right)^{\alpha} \exp\left(\beta g_q(x,c)\right), \tag{14}$$

where α controls the contribution of the aggregated CDE and β controls the strength of GC. The resulting modulation preferentially emphasizes CDE that exhibit stronger compatibility with the self-domain Gaussian geometry, forming an adaptive cross-domain resonance for rectification.

Finally, the CORE scores are calibrated into the CLIP logit space before fusion:

$$\boldsymbol{\ell}^{\mathrm{CORE}} = \mathrm{ZCal}(\boldsymbol{s}^{\mathrm{CORE}}, \boldsymbol{s}^{\mathrm{text}}). \tag{15}$$

4.4 Spectral Adaptive Resonance

STAR organizes these spectral responses through an anchor–rectification mechanism, where SAG establishes self-domain resonance as an anchor and CORE further rectifies it with Gaussian-compatible cross-domain resonance. We finally combine all branches through a convex combination:

$$\boldsymbol{\ell}^{\mathrm{fuse}} = (1 - \omega_{\mathrm{SAG}} - \omega_{\mathrm{CORE}})\,\boldsymbol{\ell}^{\mathrm{text}} + \omega_{\mathrm{SAG}}\boldsymbol{\ell}^{\mathrm{SAG}} + \omega_{\mathrm{CORE}}\boldsymbol{\ell}^{\mathrm{CORE}}, \tag{16}$$

where ω_{SAG} and ω_{CORE} are fusion weights with $0 < \omega_{\mathrm{SAG}}, \omega_{\mathrm{CORE}} < 1$ and $\omega_{\mathrm{SAG}} + \omega_{\mathrm{CORE}} < 1$, all of which are selected from a predefined validation grid with all model parameters fixed.

5 Experiments

5.1 Setup.

Datasets. Following previous work (Wang et al., 2024a), we evaluate CLIP-based multi-domain learning on three public image classification benchmarks: Office-Home (Venkateswara et al., 2017), DomainNet (Peng et al., 2019), and FMoW-WILDS (Christie et al., 2018; Koh et al., 2021). Office-Home and DomainNet mainly involve style-driven domain shifts, while FMoW defines domains according to geographic regions. We focus on the full-supervised and few-shot settings in the main paper, with additional evaluations under broader learning scenarios provided in the Appendix.

Baselines and protocol. All baselines are built upon OpenAI CLIP (Radford et al., 2021). We evaluate our method under the full-supervised and few-shot settings, comparing with zero-shot CLIP, two representative CLIP-based prompt-learning methods, CoOp (Zhou et al., 2022) and MaPLe (Khattak et al., 2023), and their multi-domain extensions with MMoE (Ma et al., 2018) and

Table 2: Few-shot accuracy (%) with 16 shots per class-domain. Dataset-level averages are underlined; best results within each method block are shown in **bold**; rows with **+ours** are shaded gray.

Model			DomainNet					Office-Home					FMoW					
	Method	+ours	Clipart	Painting	Real	Sketch	Avg.	Art	Clipart	Product	Real	Avg.	Asia	Europe	Africa	America	Oceania	Avg.
ZS			74.25	68.81	83.46	65.91	73.11	84.15	70.39	90.00	87.10	82.91	17.84	22.26	15.66	20.09	24.00	19.97
		✓	**78.81**	**74.79**	**86.14**	**70.40**	**77.54**	**88.30**	**82.46**	**95.53**	**90.75**	**89.26**	**35.64**	**37.09**	**43.93**	**35.87**	**51.53**	**40.81**
CoOp	jointly		76.38	71.85	85.77	67.42	75.36	86.61	76.20	93.61	90.43	86.71	27.45	32.19	24.09	31.86	37.33	30.59
	jointly	✓	**79.32**	**75.39**	**86.46**	**70.57**	**77.93**	**89.81**	**82.02**	**95.43**	**91.29**	**89.64**	**37.50**	**37.98**	**41.38**	**37.46**	**47.19**	**40.30**
	separately		77.06	72.56	86.12	67.66	75.85	87.73	75.22	93.19	90.32	86.61	27.73	27.93	26.29	29.61	38.45	30.00
	separately	✓	**79.17**	**75.32**	**86.50**	**70.53**	**77.88**	**89.25**	**81.58**	**95.32**	**91.29**	**89.36**	**36.89**	**37.56**	**42.34**	**36.28**	**51.53**	**40.92**
	MMoE		75.93	71.23	85.53	67.45	75.03	86.04	74.56	92.66	90.64	85.97	26.32	30.77	22.91	31.87	38.72	30.11
	MMoE	✓	**79.23**	**75.22**	**86.39**	**70.61**	**77.86**	**89.25**	**81.25**	**94.68**	**91.61**	**89.20**	**36.94**	**37.98**	**41.99**	**37.01**	**50.07**	**40.80**
	PLE		76.88	72.14	86.24	68.14	75.86	87.92	75.22	92.87	90.54	86.64	27.46	28.38	25.29	31.49	37.33	29.99
	PLE	✓	**79.24**	**75.58**	**86.52**	**70.87**	**78.05**	**89.81**	**82.13**	**95.21**	**90.86**	**89.50**	**37.20**	**37.67**	**41.54**	**36.19**	**49.16**	**40.35**
MaPLe	jointly		77.89	72.37	85.97	68.78	76.25	86.80	76.64	92.34	91.18	86.74	36.93	38.81	31.29	41.02	45.91	38.79
	jointly	✓	**80.08**	**75.57**	**86.51**	**71.26**	**78.36**	**89.25**	**82.35**	**95.00**	**91.72**	**89.58**	**43.25**	**46.04**	**47.20**	**43.84**	**53.76**	**46.82**
	separately		77.34	72.96	86.26	68.33	76.22	87.36	74.67	93.40	90.43	86.47	29.39	31.21	28.82	31.64	39.48	32.11
	separately	✓	**79.29**	**75.43**	**86.66**	**70.41**	**77.95**	**89.06**	**81.47**	**94.57**	**91.61**	**89.18**	**38.08**	**38.48**	**41.41**	**37.18**	**49.38**	**40.91**
	MMoE		78.00	72.42	86.17	68.92	76.37	86.23	75.44	93.08	90.64	86.35	35.51	36.81	29.05	39.16	44.34	36.98
	MMoE	✓	**80.24**	**75.55**	**86.62**	**71.23**	**78.41**	**87.55**	**81.36**	**94.79**	**90.32**	**88.50**	**42.97**	**44.99**	**47.43**	**43.45**	**54.82**	**46.73**
	PLE		78.47	73.47	86.53	69.64	77.03	86.79	76.86	93.41	91.29	87.08	37.23	36.83	34.56	39.70	48.50	39.37
	PLE	✓	**80.36**	**75.90**	**86.74**	**71.59**	**78.65**	**88.49**	**81.58**	**95.74**	**91.94**	**89.44**	**44.15**	**46.21**	**46.87**	**42.81**	**57.27**	**47.46**

Table 3: Ablation of STAR components on zero-shot ViT-B/16 CLIP. ✓/× denotes inclusion/exclusion, the best results are in **bold**, and rows of complete STAR are shaded gray.

Full-supervised						Few-shot					
SAG	CORE-GC	CORE-CDE	Office-Home	DomainNet	FMoW	SAG	CORE-GC	CORE-CDE	Office-Home	DomainNet	FMoW
×	×	×	82.91±0.00	73.11±0.19	19.97±0.30	×	×	×	82.91±0.00	73.11±0.19	19.97±0.30
✓	×	×	88.68±0.21	79.87±0.06	45.50±0.85	✓	×	×	87.65±0.39	76.99±0.06	36.71±0.55
✓	×	✓	88.64±0.09	79.91±0.09	45.65±0.70	✓	×	✓	88.58±0.65	77.09±0.03	36.09±1.39
✓	✓	×	88.75±1.21	79.88±0.27	50.90±0.08	✓	✓	×	88.06±1.44	77.06±0.40	40.09±0.54
×	✓	✓	89.23±0.19	78.34±0.07	49.52±0.23	×	✓	✓	88.38±0.53	77.06±0.11	40.11±0.82
✓	✓	✓	**89.78±0.06**	**80.50±0.16**	**52.83±0.06**	✓	✓	✓	**89.26±0.53**	**77.54±0.01**	**40.81±0.31**

PLE (Tang et al., 2020). For all trainable methods, we use the same optimization setting and select the best checkpoint according to validation domain-average accuracy. All results are reported over three runs. Evaluations under additional learning settings are provided in Appendix A.2.

5.2 Main results

Full-supervised learning. As shown in Tab. 1, separately trained models generally outperform jointly trained ones, which suggests that preserving domain-specific parameters is more beneficial. Among the two MDL architectures, PLE is overall more effective and stable than MMoE, indicating that hierarchical shared-private modeling is better suited to CLIP-based MDL than mixture-style expert routing. MaPLe-based methods generally outperform their CoOp-based counterparts, likely because MaPLe also adapts the image encoder while CoOp keeps it frozen. Since the shift in FMoW is less style-dominated than in DomainNet and Office-Home, MaPLe is further favored over CoOp. Our method consistently improves all baselines, including the strongest MaPLe+PLE combination, showing that it provides complementary gains beyond existing prompt-learning and MDL designs.

Few-shot learning. As shown in Tab. 2, the trends of few-shot differ from those in the full-supervised case. For MaPLe, jointly training generally outperforms separately training, whereas for CoOp the two are comparable, suggesting that the advantage of domain-specific modeling becomes less pronounced when labeled data are limited, which is likely because splitting the data by domain further weakens the already limited supervision. Both PLE and MMoE remain effective in the few-shot regime, with PLE showing slightly better overall performance. Finally, our method consistently improves all baselines, further verifying its effectiveness in the few-shot setting.

Ablation study. We perform ablation studies on zero-shot CLIP to examine each component of STAR. As shown in Table 3, both SAG and CORE improve the text-only baseline, while their combination

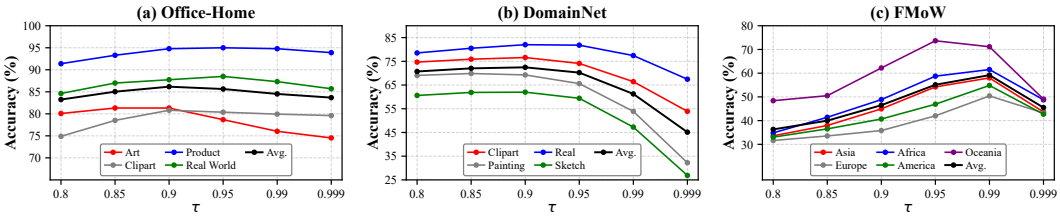

Figure 4: Influence of the truncation threshold τ under the full-supervised setting on zero-shot CLIP.

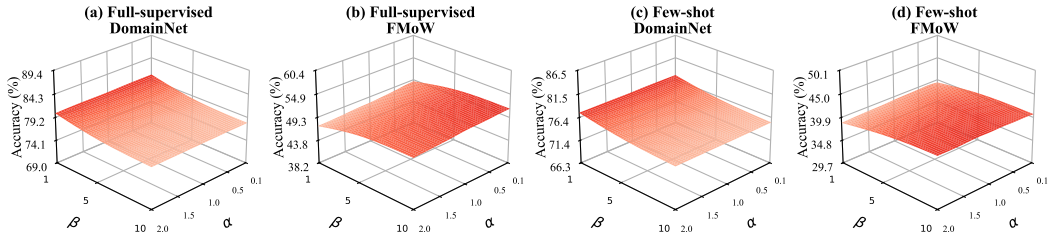

Figure 5: Sensitivity analysis of α and β on DomainNet and FMoW.

consistently performs best, confirming their complementarity. With SAG already applied, CDE alone brings only marginal gains, indicating that cross-domain evidence requires reliability guidance. Incorporating self-domain GC improves their effectiveness, with a stronger impact on FMoW likely due to its pronounced geographic distribution shifts. Overall, full STAR consistently achieves the best performance, showing that GC complements rather than replaces subspace-based evidence.

5.3 Further Analysis

We further analyze the design of STAR, where all hyperparameters are selected on the validation set through lightweight post-hoc search that takes only seconds and preserves the training-free nature.

Influence of τ. We analyze the influence of τ on zero-shot CLIP under the full-supervised setting using the SAG branch alone, as shown in Fig. 4. Office-Home and DomainNet consistently achieve their best performance around $\tau = 0.9$–0.95, suggesting that τ is transferable across datasets with similar domain shifts. FMoW favors a larger threshold around $\tau = 0.99$, likely due to its more complex geographic variations requiring broader spectral retention. When τ approaches 1, performance consistently drops on all datasets, as truncation vanishes and the projection energy degenerates toward the squared feature norm. This further supports the necessity of spectral truncation.

Sensitivity to α and β. We analyze the calibration parameters α and β. As shown in Fig. 5, performance varies smoothly across a broad range of α, suggesting that STAR is not sensitive to narrow calibration of CDE. In contrast, β, which controls the sharpness of GC for CORE, shows stronger dataset-dependence. DomainNet favors mild Gaussian modulation, whereas FMoW benefits from stronger modulation. These results are consistent with the design of STAR: CDE and GC play distinct roles and thus require separate calibration, while the broad high-performance regions indicate that STAR does not rely on a single hyperparameter configuration.

6 Conclusion

In this paper, we reveal two geometric insights in multi-domain CLIP features: (1) nuisance-sensitive variations concentrate in spectral-tail directions, while dominant class geometry remains domain-dependent; and (2) complementary cross-domain evidence can recover residual self-domain failures, with their recoverability reflected in self-domain Gaussian geometry. Building on these findings, we propose STAR, a training-free framework that establishes self-domain spectral resonance as an anchor and further rectifies it with Gaussian-compatible cross-domain resonance. Extensive experiments validate the effectiveness of STAR, offering a new geometric perspective for CLIP-based MDL.

AI USE STATEMENT

Generative AI tools were used solely as assistive tools for improving the clarity, grammar, and readability of author-written text, and for assisting with the implementation and debugging of research code under the authors' supervision. They were not delegated responsibility for scientific decisions or the validation of research outcomes. All AI-assisted text and code were manually reviewed, revised where necessary, and, for code, tested by the authors to ensure correctness and consistency with the intended methodology. The authors have further verified the final manuscript for factual accuracy, proper attribution, and originality, and confirmed that it contains no AI-generated fabricated or misleading statements, plagiarism, or other misrepresentations. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide the implementation details, hyperparameter settings, and experimental protocols in the main paper and appendix. STAR is training-free and keeps all model parameters fixed, with its subspace construction, calibration, and fusion procedures fully specified. Additional ablations and sensitivity analyses are provided to facilitate reproducibility.

REFERENCES

Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundation models defining a new era in vision: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2025.

Yassir Bendou, Amine Ouasfi, Vincent Gripon, and Adnane Boukhayma. Proker: A kernel perspective on few-shot adaptation of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 2023.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Ekaterina Iakovleva, Karteek Alahari, and Jakob Verbeek. Multi-domain learning with modulation adapters. *arXiv preprint arXiv:2307.08528*, 2023.

Mahesh Joshi, Mark Dredze, William W. Cohen, and Carolyn Rosé. Multi-domain learning: When do domains matter? In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2012.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Yayuan Li, Jintao Guo, Lei Qi, Wenbin Li, and Yinghuan Shi. Text and image are mutually beneficial: Enhancing training-free few-shot classification with clip. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

Yunsheng Li and Nuno Vasconcelos. Efficient multi-domain learning by covariance normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Adversarial multi-task learning for text classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.

Yajing Liu, Xinmei Tian, Ya Li, Zhiwei Xiong, and Feng Wu. Compact feature learning for multi-domain image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H. Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (SIGKDD)*, 2018.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Efficient parametrization of multi-domain deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018.

Alice Schoenauer Sebag, Louise Heinrich, Marc Schoenauer, Michele Sebag, Lani F. Wu, and Steven J. Altschuler. Multi-domain adversarial learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Hongyan Tang, Junning Liu, Ming Zhao, and Xudong Gong. Progressive layered extraction (ple): A novel multi-task learning (mtl) model for personalized recommendations. In *Proceedings of the ACM Conference on Recommender Systems (RecSys)*, 2020.

Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. Sus-x: Training-free name-only transfer of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Shiguang Wang, Tao Xie, Jian Cheng, Xingcheng Zhang, and Haijun Liu. Mdl-nas: A joint multi-domain learning framework for vision transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2023.

Ximei Wang, Junwei Pan, Xingzhuo Guo, Dapeng Liu, and Jie Jiang. Decoupled training: Return of frustratingly easy multi-domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2024a.

Zhengbo Wang, Jian Liang, Lijun Sheng, Ran He, Zilei Wang, and Tieniu Tan. A hard-to-beat baseline for training-free CLIP-based adaptation. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2024b.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2024.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2022.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022.

Xiangyang Zhu, Renrui Zhang, Bowei He, Aojun Zhou, Dong Wang, Bin Zhao, and Peng Gao. Not all features matter: Enhancing few-shot clip with adaptive prior refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

A Appendix

A.1 Details of Experimental Setup

A.1.1 Dataset Construction

Table 4: Number of samples in the train, validation, and test splits for DomainNet, Office-Home, and FMoW under our data construction protocol.

Split	DomainNet				Office-Home				FMoW				
	Clipart	Painting	Real	Sketch	Art	Clipart	Product	Real	Asia	Europe	Africa	America	Oceania
Train	38,361	55,164	138,221	57,683	1,696	3,099	3,148	3,087	82,613	129,841	26,569	126,145	10,580
Val	4,660	6,763	17,137	7,067	466	810	821	805	10,302	16,206	3,300	15,743	1,303
Test	5,108	7,201	17,589	7,516	265	456	470	465	10,384	16,286	3,370	15,823	1,369

We evaluate all methods on three public multi-domain image classification datasets: DomainNet (Peng et al., 2019), Office-Home (Venkateswara et al., 2017), and FMoW-WILDS (Christie et al., 2018; Koh et al., 2021). Table 4 reports the numbers of training, validation, and test samples for all domains under our data construction protocol. Although all three datasets are used in the multi-domain setting, their split construction follows dataset-specific procedures. To the best of our knowledge, DomainNet and FMoW-WILDS are among *the largest publicly available datasets* that have been adopted for multi-domain learning (MDL), providing large-scale and diverse benchmarks for evaluating MDL methods.

DomainNet. We use four domains: *clipart*, *painting*, *real*, and *sketch*. Before splitting, we remove duplicate samples according to the official `domain_net_duplicates.txt` file released with DomainBed (Gulrajani & Lopez-Paz, 2021), then split each class-domain subset into Train, Val, and Test sets with an 8:1:1 ratio.

Office-Home. We use four domains, including *art*, *clipart*, *product*, and *real_world*. For each class in each domain, we split the original `train` subset into training and validation sets with an 8:2 ratio, while directly using the original `val` subset as the test set.

FMoW. We follow the WILDS (Koh et al., 2021) setup and group samples into five geographic regions: *Asia*, *Europe*, *Africa*, *Americas*, and *Oceania*. We then split each class-region subset into training, validation, and test sets with an 8:1:1 ratio. Together, these datasets cover both style-driven domain shifts, as in DomainNet and Office-Home, and geographically induced domain shifts, as in FMoW.

A.1.2 Learning Settings

We evaluate all methods under five representative MDL settings: full-supervised learning, few-shot learning, domain-imbalanced learning, semi-supervised learning, and unsupervised learning.

Full-supervised learning. All available labeled training samples are used for model training or for constructing the proposed class-domain subspaces.

Few-shot learning. We use 16 labeled samples for each class-domain pair. This setting evaluates whether different methods remain effective when only a limited number of labeled examples are available in each domain.

Domain-imbalanced learning. The number of labeled samples varies across domains according to the native data scale. The domain-specific shot numbers are reported in Tab. 5. This setting is used to evaluate robustness under imbalanced domain supervision.

Semi-supervised learning. We use 2 labeled samples for each class-domain pair, while the remaining training samples are treated as unlabeled data. For trainable baselines, optimization is performed with FixMatch (Sohn et al., 2020).

Unsupervised learning. No ground-truth labels are used for constructing the training supervision. For trainable baselines, pseudo-label based optimization is also performed with FixMatch (Sohn et al., 2020).

A.1.3 BASELINES

All baselines are built upon OpenAI CLIP (Radford et al., 2021). We include zero-shot CLIP as the basic training-free baseline. We further evaluate two representative CLIP-based parameter-efficient fine-tuning methods, CoOp (Zhou et al., 2022) and MaPLe (Khattak et al., 2023). To construct stronger multi-domain baselines, we integrate two classical multi-domain learning architectures, MMoE (Ma et al., 2018) and PLE (Tang et al., 2020), into CoOp and MaPLe. The corresponding architecture designs are described in the following sections of the Appendix.

A.1.4 OPTIMIZATION AND EVALUATION

For all trainable methods, we use the same optimization setting, with a learning rate of 0.001, 10 epochs, and a batch size of 32. Models are trained on the training set with the cross-entropy loss, and the best checkpoint is selected according to domain-average accuracy on the validation set. Final performance is reported on the test set. All experiments are conducted on a single NVIDIA L20 GPU with 48GB memory and two Intel Xeon Gold 6132 CPUs. We report results over three runs.

A.2 RESULTS OF MORE SETTINGS

Table 5: Domain-imbalanced accuracy (%). Dataset-level averages are underlined; best results within each method block are shown in **bold**; rows with **+ours** are shaded gray.

Model			DomainNet					Office-Home					FMoW					
	Method	+ours	Clipart	Painting	Real	Sketch	Avg.	Art	Clipart	Product	Real	Avg.	Asia	Europe	Africa	America	Oceania	Avg.
	Shots		8	64	128	16		4	4	16	32		32	64	16	128	8	
ZS			74.25	68.81	83.46	65.91	73.11	84.15	70.39	90.00	87.10	82.91	17.84	22.26	15.66	20.09	24.00	19.97
		✓	**77.43**	**76.54**	**87.49**	**70.82**	**78.07**	**85.66**	**74.67**	**95.21**	**90.86**	**86.60**	**39.95**	**42.55**	**44.15**	**43.17**	**49.60**	**43.88**
CoOp	jointly		75.75	72.32	86.60	67.65	75.58	86.03	72.48	92.56	90.97	85.51	28.20	35.74	24.68	34.48	39.01	32.42
	jointly	✓	**77.82**	**77.00**	**88.36**	**70.58**	**78.44**	**86.79**	**76.81**	**94.95**	**92.10**	**87.66**	**40.15**	**42.73**	**44.57**	**43.75**	**50.47**	**44.34**
	separately		76.17	74.03	87.20	68.36	76.44	86.17	71.93	92.44	91.08	85.41	21.94	34.11	26.14	33.73	30.93	29.37
	separately	✓	**77.96**	**77.38**	**88.50**	**70.58**	**78.60**	**87.08**	**75.77**	**94.84**	**92.26**	**87.49**	**40.25**	**42.68**	**43.55**	**43.60**	**49.45**	**43.91**
	MMoE		75.58	73.11	86.87	67.67	75.81	85.85	72.81	92.66	90.64	85.49	22.39	33.07	22.89	33.26	40.62	30.45
	MMoE	✓	**77.67**	**77.25**	**88.38**	**70.69**	**78.50**	**88.11**	**75.99**	**95.85**	**91.72**	**87.92**	**40.24**	**42.80**	**44.17**	**43.53**	**51.28**	**44.40**
	PLE		76.41	74.01	87.14	68.58	76.53	85.28	71.49	92.77	90.97	85.13	25.24	32.83	25.27	33.74	33.31	30.07
	PLE	✓	**78.06**	**77.61**	**88.34**	**71.21**	**78.81**	**87.08**	**76.43**	**95.00**	**91.56**	**87.52**	**40.23**	**42.66**	**43.56**	**43.36**	**49.05**	**43.77**
MaPLe	jointly		76.70	74.39	87.19	69.02	76.83	85.28	75.11	93.08	90.97	86.11	37.88	46.38	35.46	46.22	48.69	42.92
	jointly	✓	**79.20**	**78.11**	**88.58**	**71.80**	**79.42**	**86.70**	**76.15**	**94.79**	**92.31**	**87.49**	**50.43**	**53.79**	**50.37**	**53.32**	**60.56**	**53.69**
	separately		76.14	75.53	87.60	69.05	77.08	86.04	72.47	92.88	91.19	85.64	24.09	42.17	29.45	40.10	32.10	33.59
	separately	✓	**78.25**	**78.18**	**88.69**	**71.24**	**79.08**	**87.55**	**75.98**	**94.79**	**91.61**	**87.48**	**41.04**	**46.88**	**43.28**	**51.13**	**50.29**	**46.53**
	MMoE		76.75	73.91	87.22	68.36	76.56	85.85	73.58	92.77	90.53	85.68	33.63	42.43	31.70	42.96	46.38	39.42
	MMoE	✓	**78.94**	**78.16**	**88.67**	**71.66**	**79.36**	**86.22**	**76.32**	**94.78**	**92.05**	**87.34**	**48.40**	**51.83**	**49.84**	**51.56**	**60.15**	**52.36**
	PLE		77.45	75.56	87.64	69.47	77.53	85.85	73.58	92.97	91.29	85.92	38.81	44.50	37.59	46.08	47.55	42.91
	PLE	✓	**79.60**	**78.13**	**88.76**	**71.81**	**79.58**	**87.55**	**76.37**	**94.95**	**92.10**	**87.74**	**49.42**	**53.27**	**50.30**	**52.42**	**58.80**	**52.84**

Domain-imbalanced learning. The results of domain-imbalanced learning are shown in Tab. 5. Jointly training generally outperforms separately training under domain imbalance, suggesting that shared representations provide more stable supervision when some domains are under-sampled. MaPLe also consistently outperforms CoOp in most cases, indicating the benefit of adapting both visual and textual branches in this challenging setting. The behavior of PLE and MMoE depends on the backbone: PLE is usually stronger with MaPLe, while MMoE is more competitive with CoOp, implying that structured expert decomposition is more beneficial when the backbone has stronger adaptation capacity. On top of these baselines, our method brings consistent improvements in all settings, even on lower-resource domains, verifying its effectiveness under imbalanced supervision.

Semi-supervised and unsupervised learning. As shown in Tab. 6, our method consistently improves both ZS-CLIP and prompt-learning variants under semi-supervised and unsupervised settings

Table 6: Semi-supervised and unsupervised accuracy (%). Dataset-level averages are underlined; best results within each method block are shown in **bold**; rows with **+ours** are shaded gray.

	Model		DomainNet					Office-Home					FMoW					
	Method	+ours	Clipart	Painting	Real	Sketch	Avg.	Art	Clipart	Product	Real	Avg.	Asia	Europe	Africa	America	Oceania	Avg.
Semi-supervised	ZS		74.26	68.81	83.46	65.91	73.11	84.15	70.39	90.00	87.10	82.91	17.85	22.26	15.66	20.09	24.00	19.97
		✓	**75.83**	**70.34**	**84.94**	**66.98**	**74.52**	**86.42**	**72.37**	**91.91**	**89.35**	**85.01**	**21.32**	**25.99**	**27.54**	**27.15**	**35.94**	**27.59**
	MaPLe	jointly	76.23	70.53	85.02	68.36	75.03	86.93	74.13	92.34	89.03	85.61	23.64	23.23	20.44	24.26	33.16	24.95
		jointly ✓	**77.57**	**71.79**	**85.87**	**69.06**	**76.07**	**86.98**	**76.65**	**93.09**	**90.86**	**86.90**	**26.80**	**31.82**	**30.31**	**29.78**	**41.49**	**32.04**
		separately	75.03	70.85	85.38	67.23	74.62	85.29	71.05	92.77	89.14	84.56	22.48	18.94	20.95	20.80	31.63	22.96
		separately ✓	**77.02**	**72.26**	**85.91**	**68.03**	**75.81**	**86.23**	**75.00**	**93.72**	**90.43**	**86.35**	**26.86**	**28.63**	**29.53**	**26.10**	**37.95**	**29.81**
		MMoE	76.58	71.25	84.85	68.65	75.33	86.04	74.67	91.81	89.57	85.53	26.56	24.25	19.60	26.21	35.47	26.42
		MMoE ✓	**77.76**	**72.23**	**85.59**	**69.58**	**76.29**	**86.61**	**75.99**	**92.77**	**89.79**	**86.29**	**27.65**	**32.68**	**32.83**	**29.22**	**41.12**	**32.70**
		PLE	76.04	70.82	85.45	67.87	75.04	85.47	75.00	92.66	**89.90**	85.76	21.71	19.04	19.11	20.43	33.68	22.79
		PLE ✓	**77.37**	**71.93**	**86.20**	**68.67**	**76.04**	**86.42**	**77.41**	**92.88**	**89.68**	**86.60**	**27.13**	**29.83**	**30.36**	**26.58**	**39.92**	**30.76**
Unsupervised	ZS		74.25	68.81	83.46	65.91	73.11	84.15	70.39	90.00	87.10	82.91	17.84	22.26	15.66	20.09	24.00	19.97
		✓	**75.76**	**70.90**	**84.83**	**67.21**	**74.68**	**86.79**	**74.12**	**92.55**	**89.46**	**85.73**	**18.01**	**21.34**	**21.08**	**23.93**	**29.18**	**22.71**
	MaPLe	jointly	74.07	70.30	84.86	68.38	74.40	84.72	71.39	89.58	87.96	83.41	12.79	11.64	13.12	16.08	18.63	14.45
		jointly ✓	**76.35**	**71.88**	**85.46**	**69.39**	**75.77**	**86.42**	**75.11**	**92.55**	**89.78**	**85.97**	**19.52**	**23.88**	**20.21**	**23.74**	**33.75**	**24.22**
		separately	75.58	70.59	85.08	68.35	74.90	83.78	70.61	90.75	89.03	83.54	21.09	15.49	13.89	17.39	23.08	18.18
		separately ✓	**76.79**	**72.54**	**85.81**	**69.41**	**76.14**	**87.17**	**74.23**	**92.45**	**89.35**	**85.80**	**20.23**	**24.46**	**22.20**	**23.88**	**30.61**	**24.28**
		MMoE	74.05	70.19	84.48	67.72	74.11	84.34	71.06	91.17	87.96	83.63	21.16	15.54	11.87	17.07	22.35	17.60
		MMoE ✓	**76.60**	**71.67**	**85.51**	**68.70**	**75.62**	**86.23**	**74.67**	**91.91**	**89.25**	**85.51**	**19.24**	**23.16**	**23.35**	**22.63**	**29.36**	**23.55**
		PLE	75.31	70.19	84.80	68.56	74.72	85.80	71.82	89.68	89.36	84.16	24.65	17.35	13.11	21.98	23.05	19.99
		PLE ✓	**77.03**	**72.30**	**85.45**	**69.55**	**76.08**	**87.55**	**74.67**	**92.87**	**90.32**	**86.35**	**27.02**	**22.40**	**20.26**	**25.82**	**29.73**	**25.04**

Table 7: Standard deviations (std, ↓) of full-supervised accuracies (%) on DomainNet, Office-Home, and FMoW. The average std for each dataset is underlined, and rows with **+ours** are shaded gray.

	Model		DomainNet					Office-Home					FMoW					
	Method	+ours	Clipart	Painting	Real	Sketch	Avg.	Art	Clipart	Product	Real	Avg.	Asia	Europe	Africa	America	Oceania	Avg.
ZS			0.17	0.21	0.11	1.08	0.15	0.03	0.04	0.02	0.01	0.03	0.10	0.14	0.69	0.25	0.72	0.34
		✓	0.98	0.25	0.04	0.61	0.20	0.49	0.89	0.03	0.41	0.25	0.91	0.52	0.32	0.59	0.62	0.14
CoOp	jointly		0.56	0.11	0.12	1.05	0.09	0.99	0.51	0.41	0.19	0.27	0.43	0.39	1.08	0.72	2.49	0.31
	jointly	✓	0.72	0.15	0.09	0.48	0.17	2.31	0.04	0.80	1.13	0.19	0.31	0.85	0.55	0.50	0.40	0.24
	separately		0.55	0.16	0.24	0.14	0.23	0.84	2.07	0.81	1.45	0.83	0.52	0.51	0.93	0.12	1.66	0.57
	separately	✓	0.85	0.19	0.16	0.23	0.22	0.49	1.62	0.11	0.34	0.33	0.13	0.84	1.09	0.25	1.48	0.35
	MMoE		0.29	0.16	0.12	0.72	0.24	0.76	0.20	0.12	1.44	0.43	0.23	0.46	0.72	0.08	1.84	0.37
	MMoE	✓	1.11	0.33	0.18	0.76	0.23	0.58	0.42	0.34	0.98	0.40	0.63	0.70	0.78	0.10	1.11	0.33
	PLE		0.12	0.52	0.10	0.59	0.05	0.50	0.19	0.41	0.34	0.24	0.33	0.67	1.05	0.16	0.57	0.16
	PLE	✓	0.42	0.39	0.17	0.27	0.20	1.13	1.05	0.23	0.17	0.37	0.62	0.89	0.79	0.14	1.41	0.12
MaPLe	jointly		0.52	0.31	0.32	0.64	0.14	1.14	0.12	0.50	0.71	0.04	1.01	0.33	0.11	0.34	0.37	0.29
	jointly	✓	0.75	0.46	0.15	0.41	0.22	1.12	0.58	0.20	0.72	0.16	0.31	0.56	0.21	0.33	1.16	0.21
	separately		0.06	0.12	0.21	0.29	0.05	0.58	0.12	0.49	0.19	0.21	0.98	0.15	0.19	0.12	0.31	0.24
	separately	✓	0.32	0.06	0.30	0.63	0.04	0.76	0.19	0.26	0.04	0.28	0.82	0.09	0.11	0.58	0.17	0.24
	MMoE		0.37	0.05	0.12	0.90	0.24	0.30	0.74	0.20	0.42	0.22	0.49	0.10	0.07	0.12	1.21	0.17
	MMoE	✓	0.65	0.34	0.24	0.64	0.20	0.04	1.17	0.04	0.26	0.27	0.41	0.49	0.09	0.27	1.04	0.10
	PLE		0.43	0.47	0.11	0.49	0.15	1.69	0.26	1.13	0.50	0.61	0.08	0.06	0.32	0.14	0.96	0.16
	PLE	✓	0.48	0.46	0.24	0.54	0.14	0.04	0.58	0.19	0.57	0.08	0.25	0.08	0.46	0.27	0.91	0.05

across all datasets, demonstrating its complementarity to different CLIP-based backbones. The improvements indicate that the proposed subspace formulation remains effective when only limited labels or pseudo labels are available. The gains are particularly clear on FMoW, where complex geographic shifts make domain-related nuisance more prominent and several FixMatch-trained baselines even underperform their zero-shot counterparts due to noisy pseudo labels. In contrast, our method still brings clear improvements, suggesting that suppressing less informative domain-related components is beneficial under both scarce-label and noisy-pseudo-label scenarios.

A.3 Standard Deviations of Main Results

For each dataset, the standard deviation reported in the *Avg.* column is computed from the final averaged performance across all domains, rather than the mean of the per-domain standard deviations.

Table 8: Standard deviations (std, ↓) of few-shot accuracies (%) on DomainNet, Office-Home, and FMoW. The average std for each dataset is underlined, and rows with **+ours** are shaded gray.

Model			DomainNet					Office-Home					FMoW					
	Method	+ours	Clipart	Painting	Real	Sketch	Avg.	Art	Clipart	Product	Real	Avg.	Asia	Europe	Africa	America	Oceania	Avg.
ZS			0.17	0.21	0.11	1.08	0.15	0.03	0.02	0.04	0.01	0.03	0.10	0.14	0.69	0.25	0.72	0.34
		✓	0.62	0.49	0.35	0.19	0.18	0.04	0.58	0.34	0.57	0.34	0.56	0.10	2.08	0.41	0.25	0.40
CoOp	jointly		0.61	0.18	0.21	0.97	0.17	0.84	1.79	0.56	0.20	0.30	3.75	1.34	0.77	1.50	0.38	0.13
	jointly	✓	0.73	0.46	0.20	0.61	0.12	0.04	0.67	0.40	1.15	0.33	1.77	0.49	0.61	0.32	0.81	0.34
	separately		0.30	0.31	0.20	0.39	0.14	1.27	0.35	0.56	1.30	0.15	3.71	1.65	0.50	0.84	1.03	0.88
	separately	✓	0.12	0.19	0.29	1.15	0.09	0.49	0.20	0.26	0.34	0.10	0.28	0.24	1.49	0.48	2.91	0.39
	MMoE		0.20	0.58	0.12	1.15	0.19	2.25	0.58	0.79	0.11	0.64	2.72	0.67	0.39	0.29	1.53	0.62
	MMoE	✓	0.62	0.67	0.05	0.29	0.21	0.30	0.26	0.94	0.41	0.24	1.06	1.12	1.35	0.68	0.21	0.53
	PLE		0.34	0.25	0.07	0.94	0.13	0.04	0.89	0.79	0.57	0.15	4.92	0.20	1.78	0.53	0.58	1.18
	PLE	✓	0.22	0.33	0.18	0.31	0.09	1.25	0.35	0.41	0.04	0.33	0.17	0.35	0.44	1.34	1.53	0.08
MaPLe	jointly		0.44	0.13	0.25	1.14	0.14	0.49	0.50	0.87	1.30	0.74	0.73	0.55	0.52	0.15	2.31	0.42
	jointly	✓	1.15	0.17	0.09	0.66	0.20	0.76	0.20	0.56	0.34	0.05	0.15	0.11	1.58	0.59	0.20	0.29
	separately		0.12	0.76	0.10	0.29	0.19	0.31	1.31	0.04	0.12	0.33	1.16	0.19	0.90	0.99	1.83	0.63
	separately	✓	0.90	0.84	0.06	0.37	0.29	1.15	0.74	1.12	0.87	0.08	1.49	0.28	0.31	0.49	0.53	0.25
	MMoE		0.29	0.09	0.05	0.37	0.07	0.23	0.35	0.72	1.15	0.55	0.19	1.04	0.05	0.41	5.11	1.09
	MMoE	✓	0.17	0.63	0.16	0.42	0.28	1.25	0.20	0.26	0.65	0.25	0.37	1.28	0.81	0.40	1.53	0.30
	PLE		0.42	0.21	0.11	0.87	0.14	1.69	1.61	0.64	0.41	0.33	3.58	0.10	0.06	0.17	0.68	0.89
	PLE	✓	0.24	0.35	0.10	0.29	0.12	0.50	0.97	0.11	0.34	0.21	1.38	0.82	0.99	0.31	0.29	0.41

Full-supervised learning. The standard-deviation results in Tab. 7 show that our method remains overall stable across different random seeds. Although a few domain-specific cases exhibit relatively larger fluctuations, the average Std are consistently small in most settings, indicating that these outliers do not dominate the overall behavior. This suggests that, beyond improving accuracy, our method also provides an important practical advantage in terms of stability and robustness.

Few-shot learning. The few-shot Std results in Tab. 8 further show that our method is overall stable across different random seeds. In general, our method yields lower or at least comparable average standard deviations to the corresponding baselines, indicating that the performance improvements are not obtained at the cost of increased instability. Even when the standard deviation is not lower than the base model, it remains within a normal range and does not change the overall trend. These results further highlight stability as an important advantage of our method in the few-shot setting.

Domain-imbalanced learning. The Std results in Tab. 9 further demonstrate the overall stability of our method. Across datasets and methods, our method generally leads to lower average standard deviations, indicating that the performance gains are achieved in a stable and reliable manner. Although the variance is not reduced in every individual case, such fluctuations remain within a reasonable range and do not affect the overall trend. These results further suggest that stability is one of the practical advantages of our method under domain imbalance.

Semi-supervised learning. The Std results on Tab. 10 in the semi-supervised setting further demonstrate the strong stability of our method. In most cases, especially in terms of the average standard deviation, our method achieves lower or comparable variance to the corresponding baselines, indicating that the performance gains are obtained in a stable and reliable manner. Although a few individual cases do not show reduced variance, they remain within a normal range and do not affect the overall trend. Overall, these results further highlight stability as an important practical advantage of our method in the semi-supervised setting.

Unsupervised learning. The unsupervised Std results on Tab. 10 also verify the overall stability of our method. In the vast majority of cases, and particularly in terms of the average standard deviation, our method remains more stable than the corresponding baselines. We note that some standard deviations on FMoW are still relatively large in the unsupervised setting. This is mainly because the base models themselves already exhibit substantial fluctuations on this challenging dataset, and although our method can further improve stability, such large underlying variance makes it difficult to

Table 9: Standard deviations (std, ↓) of domain-imbalanced accuracies (%) on DomainNet, Office-Home, and FMoW across different random seeds. The average std for each dataset is underlined, and rows with **+ours** are shaded gray.

Model			DomainNet					Office-Home					FMoW					
	Method	+ours	Clipart	Painting	Real	Sketch	Avg.	Art	Clipart	Product	Real	Avg.	Asia	Europe	Africa	America	Oceania	Avg.
ZS			0.17	0.21	0.11	1.08	0.15	0.03	0.02	0.04	0.01	0.03	0.10	0.14	0.69	0.25	0.72	0.34
		✓	0.94	0.42	0.45	0.76	0.20	2.25	1.32	1.12	0.42	0.46	0.74	0.97	1.06	0.44	0.84	0.51
CoOp	jointly		0.43	0.39	0.06	0.78	0.05	1.14	0.43	0.64	1.15	0.04	0.61	1.67	0.22	0.17	0.58	0.27
	jointly	✓	0.74	0.43	0.11	1.19	0.06	1.27	1.16	1.13	0.19	0.17	0.25	0.82	0.51	0.17	0.71	0.36
	separately		0.19	0.70	0.15	0.38	0.05	0.58	0.27	0.80	0.11	0.13	3.47	0.82	0.20	0.44	2.21	1.34
	separately	✓	0.60	0.29	0.08	0.93	0.07	2.56	0.35	0.71	0.04	0.36	0.17	0.39	0.84	0.11	2.39	0.28
	MMoE		0.34	0.08	0.20	0.82	0.19	0.23	0.97	0.41	0.19	0.27	2.98	1.21	1.96	0.84	0.25	0.87
	MMoE	✓	0.96	0.14	0.08	0.94	0.05	0.41	0.12	0.79	0.41	0.25	1.28	0.68	0.75	0.33	0.26	0.53
	PLE		0.29	0.16	0.06	0.89	0.20	0.49	1.31	0.04	0.57	0.55	0.39	1.25	0.99	1.58	1.21	0.32
	PLE	✓	0.19	0.22	0.07	1.44	0.29	0.23	0.27	0.24	0.26	0.13	0.96	0.15	0.64	0.95	0.48	0.24
MaPLe	jointly		0.46	0.21	0.12	0.62	0.06	0.49	0.20	1.28	0.34	0.28	2.32	0.10	1.88	0.27	1.47	0.22
	jointly	✓	0.90	0.16	0.09	0.74	0.06	3.58	0.26	0.04	0.26	0.67	1.56	0.14	0.88	0.44	0.79	0.45
	separately		0.09	0.54	0.14	0.84	0.10	0.50	1.47	0.11	0.95	0.44	6.78	0.09	0.58	2.52	3.71	2.34
	separately	✓	0.60	0.49	0.08	1.09	0.09	0.63	0.43	0.41	0.34	0.17	1.05	0.85	0.27	1.59	2.22	0.72
	MMoE		0.11	0.25	0.08	0.84	0.16	1.26	0.50	1.14	0.95	0.27	0.85	0.57	0.05	1.14	2.47	0.48
	MMoE	✓	0.67	0.31	0.11	0.90	0.11	0.76	0.35	1.28	0.34	0.25	0.36	0.12	0.82	0.84	0.09	0.38
	PLE		0.45	0.39	0.04	0.36	0.15	0.77	0.50	1.43	0.49	0.36	0.29	0.07	2.98	0.88	0.56	0.82
	PLE	✓	0.77	0.24	0.12	0.70	0.08	0.84	0.12	1.13	0.65	0.33	0.65	0.46	1.52	0.34	0.68	0.70

Table 10: Standard deviations (std, ↓) of semi-supervised and unsupervised accuracies (%) on DomainNet, Office-Home, and FMoW. Dataset-level averages are underlined; with **+ours** are shaded gray.

Model				DomainNet					Office-Home					FMoW					
		Method	+ours	Clipart	Painting	Real	Sketch	Avg.	Art	Clipart	Product	Real	Avg.	Asia	Europe	Africa	America	Oceania	Avg.
Semi-supervised	ZS			0.17	0.21	0.11	1.08	0.15	0.03	0.02	0.04	0.01	0.03	0.10	0.14	0.69	0.25	0.72	0.34
			✓	0.82	0.05	0.06	0.79	0.05	0.77	0.66	0.12	0.49	0.39	2.46	0.27	0.41	0.87	1.27	0.43
	MaPLe	jointly		0.33	0.53	0.36	0.87	0.14	0.76	1.31	0.26	0.65	0.22	2.83	2.47	1.16	1.17	4.91	2.18
		jointly	✓	0.73	0.98	0.19	0.68	0.20	0.31	0.11	0.98	0.20	0.36	2.58	1.36	1.73	0.58	2.21	0.84
		separately		1.13	0.79	0.15	0.08	0.07	0.49	0.66	0.56	0.49	0.24	3.31	2.86	1.12	0.92	0.79	1.21
		separately	✓	0.46	0.70	0.15	0.42	0.10	0.76	0.35	0.41	0.49	0.24	2.87	0.07	0.41	1.49	2.96	0.73
		MMoE		0.05	0.24	0.34	0.30	0.14	2.25	0.74	0.49	0.72	0.46	2.36	4.42	1.11	5.52	2.23	2.05
		MMoE	✓	0.10	0.05	0.22	0.34	0.15	1.27	1.47	0.86	0.50	0.13	0.61	1.77	0.37	3.23	0.84	1.08
		PLE		0.43	0.80	0.13	0.38	0.27	0.23	2.29	1.86	0.34	0.84	9.12	3.71	0.85	2.46	3.68	0.72
		PLE	✓	0.82	0.29	0.06	0.56	0.13	0.49	0.04	2.07	0.57	0.73	4.78	1.07	2.46	0.37	1.78	0.27
Unsupervised	ZS			0.17	0.21	0.11	1.08	0.15	0.03	0.02	0.04	0.01	0.03	0.10	0.14	0.69	0.25	0.72	0.34
			✓	0.53	0.15	0.07	0.90	0.09	0.49	0.50	0.26	0.65	0.14	0.71	0.18	0.27	0.07	0.63	0.40
	MaPLe	jointly		0.29	0.49	0.15	0.80	0.23	0.31	1.62	0.64	1.15	0.91	0.70	0.85	2.10	0.93	1.34	0.81
		jointly	✓	0.43	0.08	0.04	1.33	0.30	0.23	0.35	0.71	0.04	0.09	1.15	1.41	0.80	0.20	1.56	0.53
		separately		0.63	0.04	0.93	0.14	0.34	1.14	1.47	0.80	0.87	0.28	6.21	1.72	0.73	0.10	2.08	1.28
		separately	✓	0.22	0.38	0.41	0.32	0.15	0.57	0.58	0.80	0.57	0.10	2.18	1.41	0.14	2.08	1.16	0.58
		MMoE		0.46	0.49	0.29	0.70	0.24	0.77	1.63	0.71	1.15	0.34	2.74	3.60	0.41	2.16	2.17	1.02
		MMoE	✓	0.34	0.44	0.28	1.34	0.36	0.50	0.82	0.56	0.95	0.10	1.29	1.86	0.31	0.99	1.27	0.12
		PLE		1.15	0.06	0.10	0.65	0.16	0.49	0.20	0.71	0.19	0.10	2.76	2.18	1.31	2.10	0.63	1.21
		PLE	✓	0.72	0.41	0.09	0.39	0.19	0.30	0.43	0.49	0.41	0.09	0.78	0.51	0.22	1.08	0.58	0.14

reduce the standard deviation to a very small range. Nevertheless, the overall results still consistently show that our method provides strong stability across different random seeds.

A.4 Efficiency Analysis of Subspace Formulation

Time consuming. As shown in Fig. 6, the computational cost of our subspace formulation remains modest across the few-shot, full-supervised, and domain-imbalanced settings. Across both datasets, our formulation step consistently requires significantly less time than model training. Specifically, it takes only about 1.2 to 4.9 minutes in the few-shot setting, 20.6 to 27.1 minutes in the full-supervised

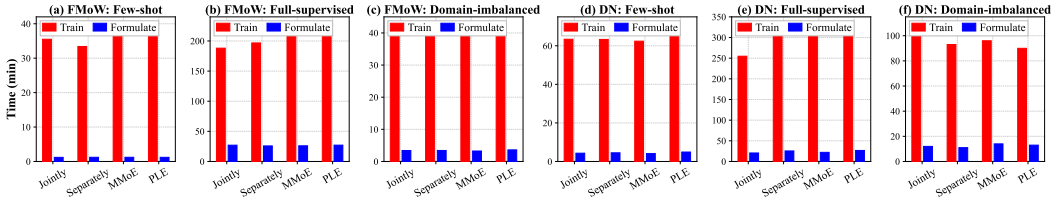

Figure 6: Computational cost comparison between subspace formulation and model training for CoOp on FMoW and DomainNet under few-shot, full-supervised, and domain-imbalanced settings.

setting, and 3.3 to 14 minutes in the domain-imbalanced setting. This efficiency comes from the training-free nature of our method: it only involves forward feature extraction followed by offline subspace construction, without iterative parameter updates or backpropagation.

More broadly, this efficiency advantage is not specific to CoOp or the settings visualized in Fig. 6: across all evaluated methods and learning settings, our subspace formulation consistently requires less than 15% of the corresponding training time. Meanwhile, it enables zero-shot CLIP to achieve comparable or even better performance than trained baselines, and when applied to trainable methods, it introduces only a small additional cost while further improving their accuracy.

Computational complexity. We further analyze the computational complexity of subspace formulation. For each class-domain pair (c, d), let $\mathbf{X}_{c,d} \in \mathbb{R}^{N_{c,d} \times D}$ denote the corresponding support feature matrix, where $N_{c,d}$ is the number of support samples and D is the feature dimension. Computing the thin singular value decomposition (SVD) of $\mathbf{X}_{c,d}$ requires $\mathcal{O}(N_{c,d} D \min(N_{c,d}, D))$ operations. Therefore, constructing all class-domain subspaces has an overall complexity of

$$\mathcal{O}\left(\sum_{c=1}^{C}\sum_{d=1}^{M} N_{c,d} D \min(N_{c,d}, D)\right) = \mathcal{O}\left(D^2 \sum_{c=1}^{C}\sum_{d=1}^{M} N_{c,d}\right), \quad \text{when } N_{c,d} > D, \tag{17}$$

where C and M denote the numbers of classes and domains, respectively. Since subspace construction is performed only once before inference and does not involve any iterative optimization or model parameter updates, its computational overhead remains limited compared with the training cost of conventional multi-domain learning methods.

The practical measurements are consistent with this complexity analysis. In particular, full-supervised DomainNet contains 345 classes and 289,429 support samples, from which STAR constructs 1,380 class-domain subspaces. Even at this scale, constructing all subspaces requires only 20.6–27.1 minutes across the evaluated methods. Together with the results under other supervision regimes, this further demonstrates that subspace formulation does not introduce a substantial computational bottleneck at the evaluated scales.

Memory overhead. We further examine the additional memory overhead introduced by STAR. For each class-domain pair, STAR only stores the retained dominant basis $\mathbf{B}_{c,d} \in \mathbb{R}^{D \times K_{c,d}}$. Therefore, the total storage complexity is

$$\mathcal{O}\left(D \sum_{c=1}^{C}\sum_{d=1}^{M} K_{c,d}\right) = \mathcal{O}\big(CMD\bar{K}\big), \tag{18}$$

where $\bar{K}$ denotes the average retained rank across all class-domain pairs. Since $\bar{K} \ll D$ in practice, STAR introduces only modest additional memory overhead.

To quantify the memory footprint, we conduct additional measurements on full-supervised DomainNet. For reference, MaPLe-PLE training is measured on a single NVIDIA L20 GPU using two warm-up iterations followed by three measured iterations. Each iteration executes a complete training step, including forward propagation, loss computation, backward propagation, and optimizer update. We then construct all subspaces using the full DomainNet support set and measure the GPU memory footprint of the retained subspaces after loading. The results are summarized in Table 11.

Despite constructing 1,380 class-domain subspaces from 289,429 support samples, all retained STAR bases occupy only 1.02 GiB of GPU peak memory during formulation. This small footprint is

Table 11: Scale and memory statistics of STAR on full-supervised DomainNet. The MaPLe-PLE result reports peak allocated GPU memory during training, while the STAR result reports the memory footprint of all retained class-domain subspaces after loading.

Statistic	Value
GPU	NVIDIA L20
Support samples	289,429
Number of classes	345
Number of domains	4
Constructed class-domain subspaces	1,380
MaPLe-PLE peak training memory	9.40 GiB
STAR formulation peak memory	1.02 GiB

consistent with the low adaptive ranks produced by the energy-based truncation, where only a limited number of dominant spectral directions are retained. Overall, the theoretical analysis and empirical measurements jointly show that STAR introduces only modest additional computational and memory overhead.

A.5 Details of Baselines

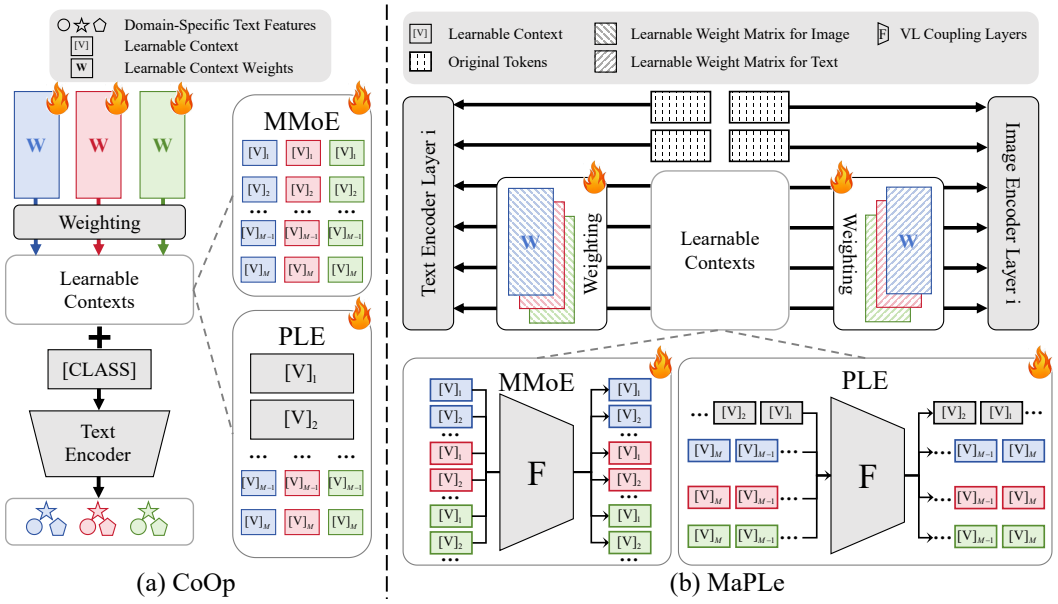

Figure 7: Overview of CoOp- and MaPLe-based multi-domain prompt learning with MMoE and PLE. In CoOp, domain-specific text features are obtained by adaptively weighting the learnable contexts. In MaPLe, the same mechanism is extended to both text and vision branches through vision-language coupling layers. MMoE uses a shared expert pool with domain-specific weights, whereas PLE combines shared and domain-specific prompts via adaptive gating.

We evaluate zero-shot CLIP and four trainable CLIP-based multi-domain baselines. For all trainable baselines, we use a unified optimization setup, including the cosine scheduler, SGD optimizer, and weight decay of 0.1. All trainable variants use shared context tokens across classes rather than class-specific context tokens. For prompt length, the CoOp-based variants use $N_{\text{CTX}} = 2$, while the MaPLe-based variants use $N_{\text{CTX}} = 4$. For the MaPLe-based variants, we further set the prompt depth to 8. Below, let T denote the number of domains.

- **Zero-shot CLIP.** We use the official OpenAI CLIP model without any trainable prompt parameters. Since domain labels are available in the multi-domain setting, we construct domain-specific textual classifiers using manually designed domain-conditioned templates. For each domain d, the text classifier is built by encoding prompts of the form $g_d(c)$ for every class name c.

  Specifically, for Office-Home, we use

$$\begin{aligned} g_{\text{art}}(c) &= \text{“a painting or artistic rendering of a } c\text{”}, \\ g_{\text{clipart}}(c) &= \text{“a clipart illustration of a } c\text{”}, \\ g_{\text{product}}(c) &= \text{“a product photo of a } c \text{ on a clean background”}, \\ g_{\text{real_world}}(c) &= \text{“a real-world photo of a } c \text{ in a natural environment”}. \end{aligned}$$

  For DomainNet, we use

$$\begin{aligned} g_{\text{clipart}}(c) &= \text{“a clipart image of a } c\text{”}, \\ g_{\text{painting}}(c) &= \text{“a painting of a } c\text{”}, \\ g_{\text{real}}(c) &= \text{“a real photo of a } c\text{”}, \\ g_{\text{sketch}}(c) &= \text{“a black-and-white sketch of a } c\text{”}. \end{aligned}$$

  For FMoW, where domains correspond to geographic regions, we use

$$g_d(c) = \text{“a satellite image in } d \text{ of a } c\text{”},$$

  where $d \in \{\text{Africa}, \text{Americas}, \text{Asia}, \text{Europe}, \text{Oceania}\}$. During inference, each image is classified using the textual classifier corresponding to its domain label.
- **CoOp-MMoE.** This variant uses CoOp-style text prompting only, without visual prompts or deep prompts. We set the number of prompt experts to T, and each expert contains N_{CTX} learnable context tokens. Therefore, the total number of learnable context tokens is $T\,N_{\text{CTX}}$. Each domain then learns a soft gate to combine these experts.
- **CoOp-PLE.** This variant also uses CoOp-style text prompting only. It contains one shared prompt branch and T domain-specific prompt branches, where each branch contains N_{CTX} learnable context tokens. Therefore, the total number of learnable context tokens is $(T + 1)N_{\text{CTX}}$. A learnable two-way soft gate is used to combine the shared and domain-specific branches for each domain.
- **MaPLe-MMoE.** This variant extends prompting to both the text and vision branches. For each prompt layer, we maintain T shared prompt experts, and each expert contains N_{CTX} learnable context tokens. The number of learnable context tokens in each layer is $T\,N_{\text{CTX}}$. Domain-specific gates are used to combine these experts for the text and vision branches.
- **MaPLe-PLE.** This variant also uses both text and vision prompts. For each encoder layer, it contains one shared branch and T domain-specific branches, and each branch contains N_{CTX} learnable context tokens. The number of learnable context tokens in each layer is $(T + 1)N_{\text{CTX}}$. A two-way soft gate is used to combine the shared and domain-specific branches, while the projection layers are shared across domains.

More detailed implementation can be found in the code, and we will make the full codebase publicly available upon acceptance.

A.6 Analysis of Design Choices and Hyperparameters

A.6.1 Selection of K

We additionally compare our *adaptive energy (AE)* truncation strategy with *fixed-rank (FR)* truncation under full-supervised setting, where FR uses a fixed rank K shared across all class-domain pairs. All rank-related hyperparameters are selected on the validation set. As shown in Table 12, we report the accuracy obtained using the *SAG branch alone*, without fusion with the text or CORE branches. To further examine the behavior of adaptive truncation, we report the mean adaptive rank $K_{c,d}$ within each domain in Table 13. The adaptive ranks vary noticeably across domains, indicating that the dominant spectral geometry is inherently domain-dependent.

Table 12: Comparison of rank selection strategies for the SAG branch under full-supervised zero-shot evaluation. AE (ours) adaptively determines rank $K_{c,d}$ using a validation-selected energy threshold, whereas FR exhaustively searches a globally fixed rank K. Results (%) averaged over 3 runs.

	Office-Home					DomainNet					FMoW					
Strategy	Art	Clipart	Product	Real-world	Avg.	Clipart	Painting	Real	Sketch	Avg.	Africa	Americas	Asia	Europe	Oceania	Avg.
AE (ours)	**83.00**	**88.62**	94.68	92.04	**89.59**	**80.84**	**76.62**	**85.65**	**72.46**	**78.89**	51.68	**47.40**	**48.63**	**47.78**	**66.59**	**52.42**
FR	82.46	87.92	**94.79**	**92.15**	89.33	79.53	75.25	85.27	71.02	77.77	**52.74**	46.90	48.06	47.25	64.39	51.87

Table 13: Mean adaptive rank $K_{c,d}$ retained by the threshold strategy for each domain under full-supervised zero-shot evaluation. The selected thresholds are $\tau = 0.90$ for Office-Home, $\tau = 0.90$ for DomainNet, and $\tau = 0.99$ for FMoW across 3 runs.

	Office-Home				DomainNet				FMoW				
Statistic	Art	Clipart	Product	Real-world	Clipart	Painting	Real	Sketch	Africa	Americas	Asia	Europe	Oceania
Mean $K_{c,d}$	9.72	6.81	6.68	11.30	22.49	37.88	48.64	31.75	61.44	123.20	108.99	129.02	43.10

Overall, AE outperforms fixed-rank truncation in the vast majority of cases and, more importantly, achieves consistent improvements in average accuracy across all three datasets, demonstrating the effectiveness of adaptive truncation. In addition, under the full-supervised setting, where the singular-value spectra can be substantially larger, fixed-rank truncation requires a considerably wider search over candidate ranks. In contrast, AE only requires a coarse-grained search over the energy threshold to reach its performance peak, as shown in Fig. 4, further demonstrating its efficiency. Moreover, the domain-wise variation of the resulting $K_{c,d}$ values suggests that different domains exhibit distinct spectral geometries and therefore require different numbers of dominant directions to adequately characterize their class-domain subspaces.