SARA: SEMI-SUPERVISED MULTI-DOMAIN LEARNING OF VISION-LANGUAGE MODELS WITHOUT TEST-DOMAIN LABELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) such as CLIP provide a promising foundation for multi-domain learning (MDL) by introducing rich multimodal semantics and transferable pretrained knowledge. However, existing MDL methods that preserve domain-specific differences often rely on test-domain labels to select specialized parameters or computation paths. Without such labels, domain-specific information can provide useful cues for some samples but induce misleading prediction variation for others. We refer to changes in such information for a fixed image as *domain perturbations*: some predictions remain stable, whereas others become unstable and may even undergo prediction flips. Moreover, most MDL approaches rely heavily on labeled data, while existing semi-supervised learning methods do not explicitly account for such domain-specific prediction variation, which can aggravate confirmation bias during self-training. To address these challenges, we propose **S**tability-**A**ware **R**outing and **A**daptation (**SARA**), a CLIP-based semi-supervised MDL framework without test-domain labels. SARA assesses sample stability under domain perturbations through *predictive consistency* and *decision margin*, and consists of two coordinated components: *(i)* Stability-Guided Self-Training applies a conservative stability assessment to filter unreliable pseudo-labels; *(ii)* Stability-Based Prediction Routing exploits domain-specific information for stable samples while reducing its influence on unstable samples through sample-specific prediction mechanisms. Extensive experiments on four multi-domain benchmarks demonstrate consistent improvements over MDL and SSL baselines, including a 1.39% average gain over the state of the art.

1 INTRODUCTION

Multi-domain learning (MDL) aims to train a model with minimal average risk across multiple overlapping but non-identical domains (Wang et al., 2024). Over the years, MDL has been extensively studied in visual recognition, while related multi-domain learning paradigms have also been explored in areas such as natural language processing, recommendation, and remote sensing (Joshi et al., 2012; Liu et al., 2017; Ma et al., 2018; Tang et al., 2020; Kim et al., 2021). Recently, vision-language models (VLMs) such as CLIP (Radford et al., 2021), pretrained on large-scale image–text data, have demonstrated strong transferability and zero-shot recognition ability and have been widely adapted to downstream visual tasks (Zhou et al., 2022b;a; Khattak et al., 2023a;b; Zhang et al., 2024b). The generalizable vision-language representations learned by CLIP provide a promising foundation for MDL, allowing multi-domain recognition to leverage rich multimodal semantics and transferable pretrained knowledge.

Existing MDL approaches can be broadly viewed from two perspectives: seeking commonalities and preserving differences (Wang et al., 2024). While distribution alignment and domain-invariant representation learning reduce discrepancies across domains (Sun & Saenko, 2016; Ganin et al., 2016; Schoenauer Sebag et al., 2019), relying only on shared representations may overlook discriminative domain-specific information (Liu et al., 2019; Wang et al., 2024). Existing methods therefore preserve domain differences through specialized parameters or computation paths, such as domain-specific adapters (Rebuffi et al., 2017; 2018; Iakovleva et al., 2023) and domain-dependent experts (Ma et al., 2018; Tang et al., 2020). However, these components are often associated with predefined domain

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.

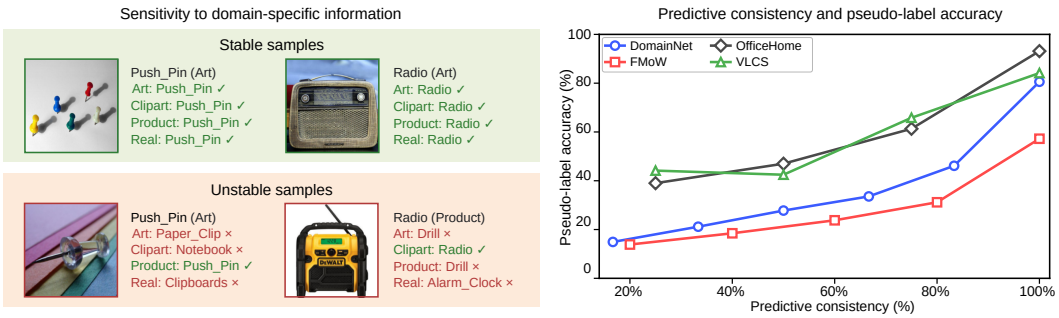


Figure 1: Sample stability under domain perturbations. Left: Stable and unstable samples from Office-Home. Right: Higher predictive consistency corresponds to higher pseudo-label accuracy across four multi-domain benchmarks.

identities, requiring the test-domain label to select the corresponding parameters or computation path. Such domain information cannot always be assumed available at inference. For example, visual conditions in autonomous systems may vary continuously with weather and illumination, while web images collected from arbitrary users often lack explicit domain annotations.

The absence of test-domain labels further raises the issue that domain-specific information may not be equally reliable for different samples. As shown in Figure 1 (left), even Office-Home samples from the same categories can respond differently when the domain-specific information involved in prediction changes: some preserve their categorical predictions, whereas others exhibit substantial changes or even prediction flips. We refer to such changes in domain-specific information for a fixed image as *domain perturbations*, and distinguish samples with stable prediction behavior under these perturbations as *stable samples* from those showing greater prediction variation as *unstable samples*. Without test-domain labels, the key question is therefore not simply which domain a test sample belongs to, but when domain-specific information can be reliably trusted for its prediction.

Beyond the dependence on test-domain labels, another practical challenge is that existing MDL methods are predominantly developed under fully supervised settings. With limited annotations, models may overfit the small labeled set or fail to adequately capture diverse domain-specific patterns, whereas unlabeled data are often much easier to collect (Yang et al., 2023). This motivates us to study MDL in a semi-supervised setting, i.e., Semi-Supervised Multi-Domain Learning (SSMDL) (Chavhan et al., 2023; Li et al., 2019; Meshgi & Mirzaei, 2020). The sensitivity to domain-specific information discussed above also affects pseudo-label reliability during self-training. As shown in Figure 1 (right), pseudo-label accuracy generally increases with predictive consistency across four multi-domain benchmarks, indicating that such consistency provides an informative signal for pseudo-label reliability. However, existing SSL methods mainly assess pseudo-label reliability through confidence-based selection (Sohn et al., 2020; Huang et al., 2022), adaptive thresholding (Zhang et al., 2021; Wang et al., 2023b), class-balanced pseudo-label construction (Menghini et al., 2023), candidate labels (Zhang et al., 2024a), or calibration (Wang et al., 2025). These strategies do not explicitly account for prediction variation caused by domain-specific information, limiting their ability to mitigate confirmation bias in MDL. This motivates incorporating such variation into pseudo-label reliability assessment for SSMDL.

To address these challenges, we propose **S**tability-**A**ware **R**outing and **A**daptation (**SARA**) , a CLIP-based SSMDL framework without test-domain labels. SARA constructs multiple domain-specific streams using shared and domain-specific multimodal prompts and assesses sample stability under domain perturbations using two criteria: *predictive consistency*, which reflects whether categorical predictions remain consistent across streams, and *decision margin*, which measures their separation from competing predictions. SARA consists of two coordinated components: *(i)* Stability-Guided Self-Training applies a conservative stability assessment to unlabeled samples to filter unreliable pseudo-labels. *(ii)* Stability-Based Prediction Routing applies sample-specific prediction mechanisms to stable and unstable samples. For stable samples, it exploits domain-sensitive information to weight and integrate knowledge from multiple streams; for unstable samples, it uses a shared classifier on complementary features to reduce the influence of domain-specific variations. Together, the two

components use sample stability as a common signal for selectively trusting pseudo-labels during training and domain-specific information during prediction, without requiring test-domain labels.

Our contributions are summarized as follows:

- We investigate VLM-based SSMDL without test-domain labels and identify sample-dependent stability under domain perturbations as a key issue for prediction and self-training.
- We propose SARA, which assesses sample stability through predictive consistency and decision margin and applies it to Stability-Guided Self-Training and Stability-Based Prediction Routing for reliable pseudo-label selection and prediction.
- Experiments on four multi-domain benchmarks demonstrate consistent improvements over representative MDL and SSL baselines, including a 1.39% average gain over the state of the art.

2 RELATED WORK

2.1 MULTI-DOMAIN LEARNING

Multi-domain learning (MDL) aims to learn a unified model over multiple related but non-identical domains (Wang et al., 2024). Existing methods generally balance knowledge sharing across domains with the preservation of domain-specific information. One line of work reduces domain discrepancy through distribution alignment or domain-invariant representation learning, including moment matching (Sun & Saenko, 2016), adversarial learning (Schoenauer Sebag et al., 2019), and shared-private representation decomposition (Liu et al., 2017). Another line explicitly models domain-specific knowledge. Adapter-based methods introduce lightweight domain-specific transformations into a shared backbone (Rebuffi et al., 2017; 2018; Iakovleva et al., 2023), while MMoE (Ma et al., 2018) and PLE (Tang et al., 2020) capture domain relationships through shared and domain-specific experts with gating mechanisms. Other studies improve MDL through specialized training or architecture design, such as decoupled training (Wang et al., 2024) and neural architecture search (Wang et al., 2023a). Methods with domain-specific components often rely on domain identity to select or combine specialized parameters at inference. Our method instead exploits domain-specific information without requiring test-domain labels and uses prediction variations across multiple domain-specific streams to assess sample stability under domain perturbations.

2.2 SEMI-SUPERVISED LEARNING

Semi-supervised learning (SSL) exploits abundant unlabeled data to improve learning when labeled data are scarce (Yang et al., 2023). Modern SSL methods commonly combine pseudo-labeling with consistency regularization. FixMatch (Sohn et al., 2020) uses confident predictions from weakly augmented views to supervise strongly augmented views, while subsequent methods improve pseudo-label selection through curriculum learning (Zhang et al., 2021), adaptive thresholds (Wang et al., 2023b), and soft sample weighting (Chen et al., 2023). For CLIP-based adaptation, UPL (Huang et al., 2022) selects high-confidence unlabeled samples for prompt learning, FPL and GRIP (Menghini et al., 2023) construct class-balanced pseudo-labeled sets, CPL (Zhang et al., 2024a) uses candidate label sets, and CAP (Wang et al., 2025) mitigates class-wise pseudo-label imbalance through calibration. Existing SSL methods mainly assess pseudo-label reliability through confidence, calibration, class balance, or consistency under data augmentation, without explicitly considering prediction variation caused by domain-specific information in MDL. Our self-training strategy instead evaluates predictive consistency across domain-specific streams to identify unstable samples and filter their pseudo-labels.

2.3 VISION-LANGUAGE MODELS AND PROMPT LEARNING

Vision-language models (VLMs) learn transferable cross-modal representations from large-scale image–text pairs and support zero-shot visual recognition (Zhang et al., 2024b). CLIP (Radford et al., 2021) is a representative VLM trained with image–text contrastive learning. Prompt learning adapts pretrained VLMs by optimizing a small set of learnable prompts while keeping most model parameters frozen. CoOp (Zhou et al., 2022b) learns continuous textual context vectors, CoCoOp (Zhou et al., 2022a) introduces conditional prompt learning to improve generalization, and VPT (Jia et al., 2022) introduces learnable visual prompts into the image encoder. MaPLe (Khattak et al., 2023a) extends

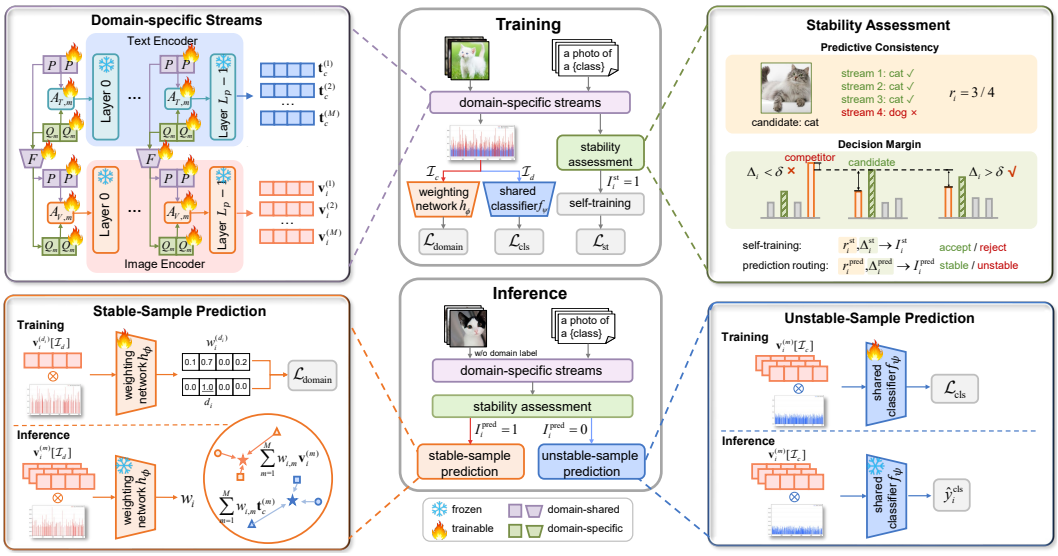


Figure 2: Overview of SARA. Sample Stability Assessment based on predictive consistency and decision margin guides both pseudo-label filtering during training and stable/unstable prediction routing at inference.

prompt learning to both visual and textual modalities, while PromptSRC (Khattak et al., 2023b) improves generalization through self-regulating constraints. Our method combines domain-shared and domain-specific multimodal prompts to construct multiple domain-specific streams. Their prediction variations are further exploited for sample stability assessment in both prediction routing and self-training.

3 METHOD

Problem setup. We consider semi-supervised multi-domain learning with M training domains $\{\mathcal{D}_m\}_{m=1}^{M}$ sharing the same C-class label space. The training data consist of labeled samples (x_i^l, y_i, d_i) and unlabeled samples (x_i^u, d_i), where $d_i \in \{1, \ldots, M\}$ denotes the domain index. Domain annotations are available for both labeled and unlabeled samples during training but are unavailable at inference. Our goal is to exploit domain-specific knowledge and unlabeled data during training while producing predictions without requiring the domain identity of a test sample.

Overview. We propose Stability-Aware Routing and Adaptation (SARA), as illustrated in Figure 2. SARA constructs multiple domain-specific streams using shared and domain-specific multimodal prompts and assesses sample stability under domain perturbations using predictive consistency and decision margin. Based on this assessment, Stability-Guided Self-Training applies a conservative criterion to filter unreliable pseudo-labels, while Stability-Based Prediction Routing uses sample-specific prediction mechanisms, leveraging domain-specific information for stable samples while adopting a less domain-dependent predictor for unstable samples.

3.1 DOMAIN-SPECIFIC REPRESENTATION LEARNING

Domain-specific streams. Although CLIP (Radford et al., 2021) provides transferable image–text representations, fully shared adaptation may overlook domain-dependent patterns, whereas independent domain-specific adaptation can weaken knowledge sharing. We therefore construct M domain-specific streams using domain-shared and domain-specific multimodal prompts while keeping the CLIP image and text encoders frozen.

Let $E_I(\cdot)$ and $E_T(\cdot)$ denote the CLIP image and text encoders. At prompt layer ℓ, $P^{(\ell)}$ denotes the shared prompt and $Q_m^{(\ell)}$ denotes the prompt specific to domain m. The prompts used by stream m

are constructed as

$$T_m^{(\ell)} = A_{T,m}^{(\ell)}\left(P^{(\ell)}, Q_m^{(\ell)}\right), \qquad V_m^{(\ell)} = A_{V,m}^{(\ell)}\left(F^{(\ell)}(P^{(\ell)}), F^{(\ell)}(Q_m^{(\ell)})\right), \tag{1}$$

where $A_{T,m}^{(\ell)}$ and $A_{V,m}^{(\ell)}$ are lightweight adapters that learn fusion weights for combining the shared and domain-specific prompts, and $F^{(\ell)}$ denotes the shared text-to-vision prompt projection.

For image x_i and class c, the normalized visual and textual representations produced by stream m are

$$\mathbf{v}_i^{(m)} = E_I\left(x_i; \{V_m^{(\ell)}\}_{\ell=0}^{L_p-1}\right), \qquad \mathbf{t}_c^{(m)} = E_T\left(c; \{T_m^{(\ell)}\}_{\ell=0}^{L_p-1}\right). \tag{2}$$

Following CLIP, we compute the stream-wise predictive distribution $\mathbf{p}_i^{(m)}$ by normalizing the scaled image–text similarities $\tau \mathbf{v}_i^{(m)\top} \mathbf{t}_c^{(m)}$ over the C classes, where τ is the CLIP logit scale. We denote its class-c probability by $p_{i,c}^{(m)}$ and the corresponding top-1 prediction by $\hat{y}_i^{(m)} = \arg\max_c p_{i,c}^{(m)}$. Hence, the same image can be evaluated through multiple streams carrying different domain-specific information.

Domain-sensitive channel selection. Before adaptation, we estimate a fixed class-conditional domain-sensitivity ranking from frozen zero-shot CLIP visual features. For class c and feature channel j, $\eta_{c,j}^2$ measures the proportion of class-conditional feature variation associated with differences across training domains. Labeled samples use their ground-truth classes, whereas unlabeled samples are grouped according to their zero-shot CLIP predictions. The sensitivity of channel j is summarized as

$$s_j = \frac{1}{|\mathcal{C}_j|} \sum_{c \in \mathcal{C}_j} \eta_{c,j}^2, \tag{3}$$

where $\mathcal{C}_j$ contains the classes for which the statistic is well defined. We denote the indices of the K most domain-sensitive channels by $\mathcal{I}_d$ and their complementary set by $\mathcal{I}_c = [d_v] \setminus \mathcal{I}_d$, where d_v is the visual feature dimension.

3.2 STABILITY-GUIDED SELF-TRAINING

Direct self-training may reinforce domain-dependent errors through unreliable pseudo-labels. We therefore assess each pseudo-label candidate according to its stability under domain perturbations. Since self-training can discard unreliable candidates, we use the strongest competing prediction from the alternative streams to construct a conservative selection criterion.

Sample stability assessment. For an unlabeled image x_i, the corresponding-domain stream d_i generates the candidate pseudo-label $\tilde{y}_i = \arg\max_c p_{i,c}^{(d_i)}$. We assess its stability through predictive consistency and decision margin:

$$r_i^{\mathrm{st}} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}\{\hat{y}_i^{(m)} = \tilde{y}_i\}, \qquad \Delta_i^{\mathrm{st}} = p_{i,\tilde{y}_i}^{(d_i)} - \max_{\substack{m \neq d_i \\ c \neq \tilde{y}_i}} p_{i,c}^{(m)}. \tag{4}$$

Here, $\mathbf{1}\{\cdot\}$ denotes the indicator function. r_i^{st} measures predictive consistency as the fraction of streams agreeing with the candidate, while Δ_i^{st} measures its separation from the strongest competing prediction in the alternative streams.

The candidate is accepted when

$$I_i^{\mathrm{st}} = \mathbf{1}\left\{r_i^{\mathrm{st}} \geq \rho_s\right\} \mathbf{1}\left\{\Delta_i^{\mathrm{st}} \geq \delta_s\right\}, \tag{5}$$

where ρ_s and δ_s denote the predictive-consistency and decision-margin thresholds, respectively. Accepted pseudo-labels are used for self-training, while the remaining candidates are discarded.

Training objectives. For labeled samples, standard supervised classification is applied to the corresponding-domain stream. Together with the accepted unlabeled samples, the main prompt-learning objective is $\mathcal{L}_{\mathrm{main}} = \mathcal{L}_{\mathrm{sup}} + \lambda_u \mathcal{L}_{\mathrm{st}}$.

To support prediction routing at inference, we additionally train a lightweight weighting network h_ϕ and a classifier f_ψ shared across all streams. The weighting network predicts an M-dimensional affinity vector $\mathbf{w}_i^{(d_i)} = h_\phi(\mathbf{v}_i^{(d_i)}[\mathcal{I}_d])$ and is supervised by the known domain annotation:

$$\mathcal{L}_{\text{domain}} = \text{CE}\left(\mathbf{w}_i^{(d_i)}, d_i\right). \tag{6}$$

This supervision encourages h_ϕ to capture sample-specific affinities over the learned streams rather than performing a hard selection of a single domain at inference.

Let $f_\psi(\cdot)$ denote the predictive distribution of the shared classifier. For stream m, its prediction on the complementary channels is $\mathbf{p}_i^{(m,\text{cls})} = f_\psi(\mathbf{v}_i^{(m)}[\mathcal{I}_c])$. The classifier is supervised across all streams by

$$\mathcal{L}_{\text{cls}} = \frac{1}{M}\sum_{m=1}^{M} \text{CE}\left(\mathbf{p}_i^{(m,\text{cls})}, y_i^\star\right), \tag{7}$$

where $y_i^\star$ denotes the ground-truth label for a labeled sample or the pseudo-label for an unlabeled sample selected by the stability assessment for self-training. Both h_ϕ and f_ψ are optimized independently using their respective objectives on detached visual features.

3.3 STABILITY-BASED PREDICTION ROUTING

Without test-domain labels, domain-specific information need not be equally reliable for every test sample. We therefore use different prediction mechanisms according to sample stability: stable samples exploit domain-specific information through adaptive stream aggregation, whereas unstable samples are predicted with reduced reliance on such information.

Sample-specific prediction mechanisms. *Stable-sample prediction.* We first form a weighted multi-stream prediction for each test sample, which is retained as the final prediction when the sample is assessed as stable. The domain-sensitive channels are used to estimate soft affinities over the learned streams. Specifically, we average the stream-wise affinity vectors as $\mathbf{w}_i = M^{-1}\sum_{m=1}^{M}\mathbf{w}_i^{(m)}$, where $w_{i,m}$ denotes the m-th component of $\mathbf{w}_i$. The resulting weights are used to aggregate the visual and textual representations:

$$\bar{\mathbf{v}}_i = \sum_{m=1}^{M} w_{i,m}\mathbf{v}_i^{(m)}, \qquad \bar{\mathbf{t}}_{i,c} = \sum_{m=1}^{M} w_{i,m}\mathbf{t}_c^{(m)}. \tag{8}$$

The class probabilities $p_{i,c}^{\text{w}}$ are obtained by normalizing $\exp\left(\tau\bar{\mathbf{v}}_i^\top\bar{\mathbf{t}}_{i,c}\right)$ over the C classes, yielding the weighted predictive distribution $\mathbf{p}_i^{\text{w}}$ with top-1 prediction $\hat{y}_i^{\text{w}} = \arg\max_c p_{i,c}^{\text{w}}$.

Unstable-sample prediction. For unstable samples, we exclude the most domain-sensitive channels and use the shared classifier f_ψ. We average its stream-wise predictive distributions as

$$\mathbf{p}_i^{\text{cls}} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{p}_i^{(m,\text{cls})}, \tag{9}$$

and denote its top-1 prediction by $\hat{y}_i^{\text{cls}} = \arg\max_c p_{i,c}^{\text{cls}}$. Sharing the classifier across streams and operating on the complementary channels reduces dependence on stream-specific domain variations while retaining category-discriminative information.

Sample stability assessment. We use the weighted prediction $\hat{y}_i^{\text{w}}$ as the candidate and assess its stability under domain perturbations through

$$r_i^{\text{pred}} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{1}\{\hat{y}_i^{(m)} = \hat{y}_i^{\text{w}}\}, \qquad \Delta_i^{\text{pred}} = p_{i,\hat{y}_i^{\text{w}}}^{\text{w}} - \frac{1}{M}\sum_{m=1}^{M}\max_{c\neq\hat{y}_i^{\text{w}}} p_{i,c}^{(m)}. \tag{10}$$

Here, r_i^{pred} measures predictive consistency across streams, while Δ_i^{pred} measures the separation between the weighted candidate and the competing predictions averaged across streams.

Self-training uses the strongest competing prediction to conservatively reject a pseudo-label whenever any alternative stream provides strong contradictory evidence. At inference, however, the goal is

Table 1: Main results on four benchmarks under the SSMDL setting. We report classification accuracy (mean±std, %) over three random seeds. † indicates that test-domain labels are required at inference. The best result in each column is highlighted in bold.

Benchmark	Method	Domain-wise Accuracy						
		Art	Clipart	Product	Real			Average
Office-Home	CLIP-ZS	82.64±0.00	67.98±0.00	89.15±0.00	88.17±0.00			81.99±0.00
	Joint	82.39±1.52	74.71±0.27	93.33±0.44	91.61±0.35			85.51±0.45
	Separate†	83.65±0.71	73.54±0.21	94.40±0.66	92.26±0.30			85.96±0.29
	MMoE†	84.78±0.89	75.15±1.02	93.48±0.66	92.04±0.18			86.36±0.38
	PLE†	86.29±1.25	75.44±0.65	94.18±0.27	91.76±0.20			86.92±0.41
	CPL	82.39±1.57	72.59±1.22	93.62±0.74	90.47±0.66			84.77±0.79
	CAP	84.65±0.79	72.66±0.46	93.69±0.44	**92.54**±0.33			85.89±0.14
	SARA(Ours)	**87.30**±0.22	**78.65**±0.46	**95.18**±0.33	**92.54**±0.33			**88.42**±0.05
		Caltech	LabelMe	SUN	VOC			Average
VLCS	CLIP-ZS	**100.00**±0.00	65.54±0.00	72.42±0.00	88.76±0.00			81.68±0.00
	Joint	98.83±0.66	66.54±4.43	67.88±2.27	82.54±3.89			78.95±2.27
	Separate†	**100.00**±0.00	50.44±0.64	60.10±3.61	87.28±1.98			74.45±1.06
	MMoE†	99.53±0.33	**72.66**±0.53	76.77±3.54	87.57±1.47			84.13±1.18
	PLE†	**100.00**±0.00	51.19±0.64	64.44±4.68	89.15±1.75			76.20±0.71
	CPL	99.30±0.00	44.19±2.09	55.45±3.87	64.69±2.55			65.91±2.10
	CAP	99.30±0.00	46.19±11.49	59.80±0.00	63.32±0.00			67.15±3.88
	SARA(Ours)	**100.00**±0.00	69.29±0.37	**83.74**±1.22	**89.64**±1.36			**85.67**±0.22
		Asia	Europe	Africa	Americas	Oceania		Average
FMoW	CLIP-ZS	14.15±0.00	19.79±0.00	17.16±0.00	18.38±0.00	28.50±0.00		19.60±0.00
	Joint	25.56±0.06	35.70±0.69	25.35±1.03	32.39±1.86	39.70±0.26		31.74±0.54
	Separate†	23.78±0.67	30.51±0.77	20.50±0.96	22.78±0.61	32.72±1.28		26.06±0.11
	MMoE†	**27.13**±0.85	36.18±0.30	26.32±1.20	32.15±1.94	41.64±1.18		32.68±0.46
	PLE†	25.84±2.48	33.92±1.63	23.69±0.85	28.02±2.67	34.74±2.78		29.24±1.33
	CPL	23.41±0.60	27.52±0.37	21.41±0.95	26.81±0.18	32.12±0.76		26.26±0.18
	CAP	24.23±0.00	29.21±0.00	23.05±0.00	**34.62**±0.24	40.73±0.00		30.37±0.25
	SARA(Ours)	25.87±0.65	**37.24**±0.75	**26.76**±4.75	33.80±2.04	**46.85**±3.32		**34.10**±0.51
		Clipart	Painting	Real	Sketch	Infograph	Quickdraw	Average
DomainNet	CLIP-ZS	70.34±0.00	66.28±0.00	83.15±0.00	62.69±0.00	46.66±0.00	13.65±0.00	57.13±0.00
	Joint	74.41±0.09	69.33±0.06	85.29±0.17	66.91±0.30	54.25±0.24	21.96±0.14	62.02±0.02
	Separate†	74.35±0.45	70.58±0.15	85.32±0.12	67.00±0.11	54.30±0.39	22.25±0.73	62.30±0.14
	MMoE†	73.93±0.27	69.46±0.23	85.18±0.20	66.81±0.21	54.28±0.15	21.52±0.46	61.86±0.09
	PLE†	74.88±0.24	70.55±0.33	**85.67**±0.06	67.08±0.14	**54.58**±0.08	22.67±0.56	62.57±0.13
	CPL	73.76±0.32	70.38±0.23	84.71±0.15	66.82±0.01	52.72±0.26	19.73±0.35	61.35±0.05
	CAP	75.40±0.01	70.48±0.20	85.05±0.09	67.29±0.28	54.22±0.13	23.06±0.28	62.58±0.06
	SARA(Ours)	**75.42**±0.25	**70.62**±0.09	85.43±0.33	**67.66**±0.15	54.42±0.19	**28.44**±0.09	**63.66**±0.04

to balance the use of domain-specific information rather than reject samples. We therefore average the competing evidence across streams, allowing sufficiently stable samples to retain the weighted predictor and exploit domain-specific information, while routing less stable samples to the less domain-dependent predictor.

The sample is assessed as stable when

$$I_i^{\text{pred}} = \mathbf{1}\left\{r_i^{\text{pred}} \geq \rho_p\right\}\mathbf{1}\left\{\Delta_i^{\text{pred}} \geq \delta_p\right\}, \tag{11}$$

where ρ_p and δ_p denote the corresponding thresholds. The final prediction is then

$$\hat{y}_i = \begin{cases} \hat{y}_i^{\text{w}}, & I_i^{\text{pred}} = 1, \\ \hat{y}_i^{\text{cls}}, & I_i^{\text{pred}} = 0, \end{cases} \tag{12}$$

where $I_i^{\text{pred}} = 1$ and $I_i^{\text{pred}} = 0$ correspond to stable and unstable samples, respectively.

Table 2: Ablation of Stability-Guided Self-Training (ST) and Stability-Based Prediction Routing (PR) (mean±std, %).

ST	PR	Office-Home	VLCS	FMoW	DomainNet	Average
×	×	85.69±0.64	83.20±0.59	32.22±0.60	62.06±0.18	65.79±0.28
✓	×	87.52±0.28	85.45±0.52	32.11±0.35	62.21±0.20	66.82±0.20
×	✓	85.73±0.81	82.14±0.30	34.07±0.46	62.33±0.16	66.07±0.29
✓	✓	**88.42**±0.05	**85.67**±0.22	**34.10**±0.51	**63.66**±0.04	**67.96**±0.17

Table 3: Ablation of predictive consistency and decision margin (mean±std, %).

Consistency	Margin	Office-Home	VLCS	FMoW	DomainNet	Average
×	×	86.90±0.29	80.17±1.31	33.71±2.59	60.96±0.13	65.44±0.24
✓	×	87.26±0.27	84.53±1.46	32.79±1.59	63.26±0.21	66.96±0.08
✓	✓	**88.42**±0.05	**85.67**±0.22	**34.10**±0.51	**63.66**±0.04	**67.96**±0.17

Table 4: Pseudo-label filtering ablation (mean±std, %). † uses the corresponding-domain stream.

Dataset	Method	Acc.†	PL Acc.	Sel.
Office-Home	Random	86.84±0.34	88.12	83.73
	FixMatch	87.01±0.49	95.45	79.17
	Consistency-only	86.76±0.39	92.53	85.90
	SARA (Ours)	**87.77**±0.68	92.88	85.45
VLCS	Random	77.92±3.15	81.35	56.87
	FixMatch	80.59±1.88	79.52	91.48
	Consistency-only	84.56±1.83	85.78	92.62
	SARA (Ours)	**85.59**±0.38	85.36	92.73
FMoW	Random	21.06±17.93	47.15	33.58
	FixMatch	32.07±1.34	65.84	27.87
	Consistency-only	32.43±1.73	42.89	51.71
	SARA (Ours)	**32.74**±0.27	48.22	44.41

4 EXPERIMENTS

We evaluate the proposed framework on four multi-domain classification benchmarks under the semi-supervised multi-domain setting. We compare our method with zero-shot CLIP, CLIP-based semi-supervised learning baselines, and representative multi-domain learning baselines. We further conduct ablation studies and additional analyses to examine the contribution of each component and better understand the behavior of the proposed method.

4.1 MAIN RESULTS

Table 1 summarizes the main results. SARA achieves the highest average accuracy on all four benchmarks. Compared with the strongest baseline on each benchmark, SARA improves the average accuracy by 1.50%, 1.54%, 1.42%, and 1.08% on Office-Home, VLCS, FMoW, and DomainNet, respectively, yielding an average improvement of 1.39%. Notably, the strongest baselines on Office-Home, VLCS, and FMoW require test-domain labels at inference, whereas SARA achieves higher accuracy without requiring such information. Averaging the benchmark-level accuracies across the four datasets, SARA achieves an overall average accuracy of 67.96%, compared with 66.26% for MMoE, the strongest baseline under this overall comparison, corresponding to an improvement of 1.71%.

The improvements are also consistent across individual domains. SARA obtains the best or tied-best accuracy on 14 of the 19 domains, showing that its gains are not concentrated in only a few domains. Among the remaining five domains, SARA is surpassed on four only by baselines that require test-domain labels. On the only other domain, FMoW-Americas, SARA is 0.82% below CAP. These results indicate that the improvements of SARA are broadly distributed across diverse domains.

4.2 ABLATION STUDY

Effect of ST and PR. Table 2 evaluates the two main components of SARA. When ST is disabled, all trainable parameters are optimized using labeled samples only; when PR is disabled, predictions are obtained by uniformly averaging the class probabilities from all domain-specific streams since test-domain labels are unavailable. ST alone improves the average accuracy by 1.03%, while PR alone yields a smaller gain of 0.28% under labeled-only training. With ST enabled, adding PR further improves the average by 1.14%, and the complete model improves the baseline by 2.17%. These results show that ST and PR are mutually reinforcing. Full domain-wise results are provided in Appendix B.1.

Effect of stability assessment. Table 3 jointly ablates the stability assessment used in both self-training and prediction routing. The $\times/\times$ setting replaces predictive consistency and decision margin with FixMatch-style confidence using a threshold of 0.85. Predictive consistency alone improves the overall average from 65.44% to 66.96%, while adding decision margin further improves all four benchmarks and raises the average to 67.96%. This shows that combining cross-stream

Table 6: Weighting and classifier accuracy on stable and unstable subsets (mean±std, %).

Subset	Prediction	Office-Home	VLCS	FMoW	DomainNet
Stable	Weighting	**91.41**±0.40	**85.91**±0.45	**42.67**±0.12	**73.11**±0.26
	Classifier	89.82±0.51	84.46±1.22	42.64±0.43	69.41±0.24
Unstable	Weighting	48.12±2.82	46.67±21.86	17.09±0.52	23.21±0.34
	Classifier	**56.47**±2.86	**63.33**±32.15	**20.51**±0.73	**23.93**±0.71

Table 7: Routing on exclusive-correct samples (mean±std).

Subset	Metric	Office-Home	FMoW	DomainNet
W-only	# Samples	53.33±3.79	3015.00±83.22	4634.33±48.52
	Correct routing (%)	**70.48**±10.53	43.46±0.46	**67.82**±0.64
C-only	# Samples	41.33±6.66	3603.67±253.72	3550.33±113.05
	Correct routing (%)	**66.58**±4.51	**62.47**±0.30	**57.95**±0.88

predictive consistency with separation from competing predictions provides a more effective stability assessment.

Analysis of pseudo-label filtering. To isolate the self-training side, all variants use the corresponding-domain stream prediction and differ only in pseudo-label filtering. In Table 4, PL Acc. and Sel. denote pseudo-label accuracy and selection rate, respectively. SARA achieves the highest corresponding-stream accuracy on all three benchmarks. Compared with FixMatch, SARA accepts more pseudo-labels and thus exploits more unlabeled supervision, while the comparison with Consistency-only shows that the decision-margin criterion improves the quality–coverage trade-off rather than simply increasing selection. Random performs substantially worse, especially on VLCS and FMoW, since low-quality pseudo-labels can degrade the model and subsequently reduce the stability mask and later supervision.

Comparison of prediction mechanisms. All variants use the same fully trained model and differ only in the prediction mechanism. As shown in Table 5, stability-based routing achieves the highest accuracy on all four benchmarks and an overall average of 67.96%. It also outperforms the corresponding stream despite the latter requiring test-domain labels, supporting sample-specific prediction routing. Detailed domain-wise results are provided in Appendix B.4.

Table 5: Comparison of prediction mechanisms (mean±std, %). † requires test-domain labels.

Prediction	Office-Home	VLCS	FMoW	DomainNet	Average
Corresponding stream†	87.77±0.68	85.59±0.38	32.74±0.27	63.35±0.01	67.37±0.04
Weighting	87.75±0.31	85.64±0.34	32.84±0.13	63.15±0.05	67.34±0.17
Classifier	86.97±0.49	84.22±1.00	34.08±0.71	60.49±0.11	66.44±0.48
Stability-based routing	**88.42**±0.05	**85.67**±0.22	**34.10**±0.51	**63.66**±0.04	**67.96**±0.17

Analysis of stability-based routing. Table 6 shows that weighting performs better on stable samples, whereas the classifier performs better on unstable samples. Table 7 further shows that on Office-Home and DomainNet, the stability assessment selects the correct mechanism for more than half of both W-only and C-only samples, consistent with Figure 3; on FMoW, the corresponding rates are 43.46% and 62.47%, respectively.This asymmetry likely reflects the more ambiguous domain structure and stronger domain interference in FMoW, which tends to assign more samples as unstable and is consistent with the classifier being relatively stronger on this benchmark. Together, these results suggest that sample stability provides a useful signal for selecting between the two prediction mechanisms.

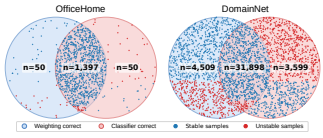


Figure 3: Exclusive-correct routing.

5 Conclusion

We study semi-supervised multi-domain learning without test-domain labels. Domain-specific information can provide useful discriminative cues but may also introduce misleading variation, making it important to trust and exploit such information selectively according to sample stability. Based on this insight, SARA assesses stability through predictive consistency and decision margin to guide both self-training and prediction routing: it filters unreliable pseudo-labels during training and adapts the use of domain-specific information at inference. Experiments on four multi-domain benchmarks show a 1.39% average improvement over the strongest baseline on each benchmark. These results suggest that sample stability under domain perturbations provides a useful principle for deciding both when to trust pseudo-labels and how to exploit domain-specific information without test-domain labels.

6 LIMITATIONS

While SARA removes the need for test-domain labels at inference, it assumes reliable and discrete source-domain annotations during training. In practice, domain labels may be noisy or ambiguous, particularly when domain boundaries change gradually. Moreover, a labeled source domain may itself contain multiple latent sub-domains, which may not be fully captured by the current one-hot domain supervision used to learn stream affinities. Extending SARA to uncertain domain annotations and latent or hierarchical domain structures is therefore an interesting direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools only to aid and polish the writing of the manuscript, including improving grammar, clarity, and presentation. All AI-assisted edits were reviewed and revised by the authors. We take full responsibility for the final content of this work, including its text, claims, experiments, and reported results.

ETHICS STATEMENT

This work uses existing benchmark datasets and does not involve new human-subject data collection or the release of new personal data. As with other methods based on pretrained vision-language models, the learned representations may inherit biases from the pretraining data and benchmark datasets, and performance may vary across domains and deployment conditions. We therefore encourage careful evaluation of dataset bias and domain mismatch before deployment in sensitive applications.

REPRODUCIBILITY STATEMENT

The model formulation, training objectives, and evaluation protocol are described in the main paper. Appendix A provides implementation and training details, including optimization settings, preprocessing, threshold selection, and model components, while Appendix A.4 describes the dataset splits and few-shot construction. Additional ablations, sensitivity analyses, and diagnostic experiments are reported in Appendix B. All experimental comparisons use fixed data partitions and labeled splits, and results are reported over three independent training runs where applicable.

REFERENCES

Ruchika Chavhan, Biplab Banerjee, and Nibaran Das. Semi-supervised multi-domain learning for medical image classification. In *Recent Trends in Image Processing and Pattern Recognition*, volume 1704 of *Communications in Computer and Information Science*, pp. 22–33. Springer, 2023.

Hao Chen, Ran Tao, Yue Fan, Yidong Wang, Jindong Wang, Bernt Schiele, Xing Xie, Bhiksha Raj, and Marios Savvides. Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. In *International Conference on Learning Representations*, 2023.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6172–6180, 2018.

Chen Fang, Ye Xu, and Daniel N. Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1657–1664, 2013.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022.

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

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Proceedings of the European Conference on Computer Vision*, pp. 709–727, 2022.

Mahesh Joshi, Mark Dredze, William W. Cohen, and Carolyn P. Rosé. Multi-domain learning: When do domains matter? In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1302–1312. Association for Computational Linguistics, 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*, pp. 19113–19122, 2023a.

Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15190–15200, 2023b.

Heejae Kim, Kyung-Chae Lee, Changha Lee, Sanghyun Hwang, and Chan-Hyun Youn. An alternating training method of attention-based adapters for visual explanation of multi-domain satellite images. *IEEE Access*, 9:62332–62346, 2021. doi: 10.1109/ACCESS.2021.3074640.

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 Earnshaw, Imran 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 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 2021.

Yitong Li, Timothy Baldwin, and Trevor Cohn. Semi-supervised stochastic multi-domain learning using variational inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1923–1934. Association for Computational Linguistics, 2019.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1–10. Association for Computational Linguistics, 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*, pp. 7193–7201, 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 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1930–1939. ACM, 2018.

Cristina Menghini, Andrew Delworth, and Stephen H. Bach. Enhancing CLIP with CLIP: Exploring pseudolabeling for limited-label prompt tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 23234–23254, 2023.

Kourosh Meshgi and Maryam Sadat Mirzaei. Adversarial semi-supervised multi-domain tracking. In *Computer Vision – ACCV 2020*, volume 12623 of *Lecture Notes in Computer Science*, pp. 612–630. Springer, 2020.

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*, pp. 1406–1415, 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 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems*, volume 30, pp. 506–516, 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*, pp. 8119–8127, 2018.

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

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alexey Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, volume 33, pp. 596–608, 2020.

Baochen Sun and Kate Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. In *Proceedings of the European Conference on Computer Vision Workshops*, pp. 443–450, 2016.

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 14th ACM Conference on Recommender Systems*, pp. 269–278. ACM, 2020.

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*, pp. 5018–5027, 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*, pp. 20094–20104, 2023a.

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*, volume 38, pp. 15644–15652, 2024.

Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, Bernt Schiele, and Xing Xie. Freematch: Self-adaptive thresholding for semi-supervised learning. In *International Conference on Learning Representations*, 2023b.

Yuchen Wang, Xuefeng Bai, Xiucheng Li, Weili Guan, Liqiang Nie, and Xinyang Chen. Handling imbalanced pseudolabels for vision-language models with concept alignment and confusion-aware calibrated margin. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 62357–62372. PMLR, 2025.

Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):8934–8954, 2023.

Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18408–18419, 2021.

Jiahan Zhang, Qi Wei, Feng Liu, and Lei Feng. Candidate pseudolabel learning: Enhancing vision-language models by prompt tuning with unlabeled data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 60004–60020. PMLR, 2024a.

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*, 46(8):5625–5644, 2024b.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.

A Implementation Details

A.1 Training Settings

We use a fixed two-shot labeled split with two labeled training images per class in each domain, while the remaining training data provide unlabeled samples according to each method's training procedure. For FMoW, both labeled and unlabeled samples are drawn from the official training split (Koh et al., 2021). We report domain-averaged accuracy, defined as the unweighted mean of accuracies independently computed on all test domains. All methods use the same image preprocessing and CLIP ViT-B/16 backbone. For trainable methods, the CLIP backbone is frozen and only lightweight adaptation modules are optimized. Results are reported as mean accuracy and standard deviation over three independent training runs.

Table 8 summarizes the training and dataset-specific hyperparameters of SARA. We use CLIP ViT-B/16 (Radford et al., 2021) as the frozen backbone and follow the fixed two-shot split described in the main paper. The self-training thresholds (ρ_s, δ_s) and prediction-routing thresholds (ρ_p, δ_p) are both selected by grid search on the validation set. For both components, the consistency threshold is searched over $\{0.25, 0.50, 0.75, 1.00\}$, and the margin threshold is searched over $\{0.00, 0.05, 0.10, 0.15, 0.20, 0.25\}$. The model is trained for a fixed 10 epochs without epoch-wise checkpoint selection, and the test set is used only for final evaluation.

Self-training implementation. For each unlabeled image, pseudo-label generation and stability assessment use the standard CLIP preprocessing, consisting of bicubic resizing to 224, center cropping, and CLIP normalization. An independently augmented view is used for self-training supervision, with `RandomResizedCrop`$(224, \text{scale} = (0.08, 1.0))$, bicubic interpolation, `RandomHorizontalFlip`$(p = 1.0)$, and CLIP normalization. Let $\mathbf{z}_i^{(d_i)}(x_i^a)$ denote the class logits produced by the corresponding-domain stream on the augmented view. For the accepted pseudo-labels, the self-training loss is

$$\mathcal{L}_{\text{st}} = \frac{\sum_{i=1}^{B} I_i^{\text{st}} \ \text{CE}\Big(\mathbf{z}_i^{(d_i)}(x_i^a), \tilde{y}_i\Big)}{\max\Big(1, \sum_{i=1}^{B} I_i^{\text{st}}\Big)}, \tag{A.1}$$

where $\tilde{y}_i$ is generated from the standard view as described in Section **??**. Thus, only unlabeled samples accepted by the stability assessment contribute to the self-training loss, while their supervision is applied to the augmented view.

A.2 Architecture and Feature-Selection Details

A.2.1 Fusion modules.

The fusion modules $A_{T,m}^{(\ell)}$ and $A_{V,m}^{(\ell)}$ in Eq. equation 1 are implemented as small learnable lookup tables, separately for each prompt layer and modality. For each learned stream m, the corresponding entry contains two logits that are normalized by softmax to weight the shared and domain-specific prompts, with no hidden layers or image-conditioned input.

A.2.2 Domain-sensitive channel scoring.

For completeness, we detail the class-conditional domain-sensitivity statistic used to rank visual channels. We first extract L2-normalized visual features $\mathbf{z}_i \in \mathbb{R}^{512}$ using the frozen zero-shot CLIP image encoder on the training pool. Labeled samples are grouped by their ground-truth classes, whereas unlabeled samples are grouped by their CLIP-ZS top-1 predictions. For class c, let $\mathcal{D}_c$ denote the domains containing samples assigned to c, and consider only classes with $|\mathcal{D}_c| \geq 2$. For channel j, the between-domain and within-domain sums of squares are

$$SS_{\text{B}}^{(c,j)} = \sum_{d \in \mathcal{D}_c} n_{c,d} \left(\mu_{c,d,j} - \mu_{c,j}\right)^2, \qquad SS_{\text{W}}^{(c,j)} = \sum_{d \in \mathcal{D}_c} \sum_{i \in \mathcal{G}_{c,d}} \left(z_{i,j} - \mu_{c,d,j}\right)^2, \tag{A.2}$$

Table 8: Training settings used in our experiments. Most hyperparameters are shared across datasets; dataset-specific settings are listed in the corresponding columns.

	Office-Home	VLCS	FMoW	DomainNet
Backbone	Frozen CLIP ViT-B/16			
Few-shot	2-shot			
Prompt length	2			
Prompt depth L_p	8			
Epochs	10			
Iterations per epoch	200			
Batch size	8			
Optimizer	SGD, momentum 0.9			
Learning rate	0.01			
Weight decay	0			
LR scheduler	200-step linear warm-up + cosine decay			
Unlabeled loss weight λ_u	1.0			
Self-training consistency threshold ρ_s	1.00		0.75	
Self-training margin threshold δ_s	0.00			
Domain-sensitive channels K	32	64	16	128
Prediction consistency threshold ρ_p	0.75		0.50	
Prediction margin threshold δ_p	0.10	0.00	0.20	0.15

where $\mathcal{G}_{c,d}$ is the corresponding class–domain group, $\mu_{c,d,j}$ is its channel mean, and $\mu_{c,j}$ is the class-wise mean. We define

$$\eta^2_{c,j} = \frac{SS_{\mathrm{B}}^{(c,j)}}{SS_{\mathrm{B}}^{(c,j)} + SS_{\mathrm{W}}^{(c,j)} + \epsilon}, \qquad \epsilon = 10^{-12}, \tag{A.3}$$

and obtain the channel score by macro-averaging over valid classes,

$$s_j = \frac{1}{|\mathcal{C}_{\mathrm{valid}}|} \sum_{c \in \mathcal{C}_{\mathrm{valid}}} \eta^2_{c,j}. \tag{A.4}$$

Channels are ranked by s_j in descending order, yielding the fixed ranking used to construct $\mathcal{I}_d$ and its complementary set $\mathcal{I}_c$.

A.2.3 Weighting network.

We compute a fixed ranking of domain-sensitive visual channels using frozen CLIP-ZS features from the training data only. Ground-truth classes are used for labeled samples, while CLIP-ZS top-1 predictions are used for unlabeled samples when computing the class-conditional channel sensitivity. The top-K channels form the domain-sensitive set $\mathcal{I}_d$, where $K = 32$, 64, 16, and 128 for Office-Home, VLCS, FMoW, and DomainNet, respectively.

Given a stream-wise visual representation $\mathbf{v}_i^{(m)} \in \mathbb{R}^{512}$, the weighting network operates on $\mathbf{v}_i^{(m)}[\mathcal{I}_d]$ and predicts a distribution over the M training domains:

$$h_\phi(\mathbf{z}) = \mathrm{Softmax}(W_2\,\mathrm{Dropout}_{0.3}(\mathrm{GELU}(\mathrm{LN}(W_1\mathbf{z} + \mathbf{b}_1))) + \mathbf{b}_2)\,, \tag{A.5}$$

where $W_1 \in \mathbb{R}^{3072\times K}$ and $W_2 \in \mathbb{R}^{M\times 3072}$. The network therefore consists of a linear layer with hidden width 3072, followed by LayerNorm, GELU, dropout, and a linear output layer.

During training, h_ϕ receives the corresponding-domain stream feature of each labeled and unlabeled batch sample, $\mathbf{v}_i^{(d_i)}[\mathcal{I}_d]$. The known training-domain label d_i provides the supervision for $\mathcal{L}_{\mathrm{domain}}$, with domain-label smoothing of 0.005. We optimize h_ϕ using SGD with learning rate 10^{-3}, momentum 0.9, and weight decay 0.1; its learning rate uses a 200-step linear warm-up followed by cosine decay. In implementation, $\mathcal{L}_{\mathrm{domain}}$ is scaled by 0.5. The visual features are detached, so this loss updates only h_ϕ.

For prediction, every domain-specific stream produces a weight vector, which is averaged as

$$\mathbf{w}_i = \frac{1}{M}\sum_{m=1}^{M} h_\phi\Big(\mathbf{v}_i^{(m)}[\mathcal{I}_d]\Big)\,. \tag{A.6}$$

The resulting weights are used to aggregate both visual and textual representations:

$$\bar{\mathbf{v}}_i = \sum_{m=1}^{M} w_{i,m}\mathbf{v}_i^{(m)}, \qquad \bar{\mathbf{t}}_{i,c} = \sum_{m=1}^{M} w_{i,m}\mathbf{t}_c^{(m)}. \tag{A.7}$$

A.2.4 SHARED CLASSIFIER.

The shared classifier operates on the complementary visual channels

$$\mathcal{I}_c = [d_v] \setminus \mathcal{I}_d, \tag{A.8}$$

where $d_v = 512$ for CLIP ViT-B/16. Its input dimension is therefore $512 - K$. We implement f_ψ as

$$f_\psi(\mathbf{z}) = \mathrm{Softmax}(W_2\,\mathrm{Dropout}_{0.3}(\mathrm{ReLU}(\mathrm{BN}(W_1\mathbf{z} + \mathbf{b}_1))) + \mathbf{b}_2)\,, \tag{A.9}$$

where $W_1 \in \mathbb{R}^{1024\times(512-K)}$ and $W_2 \in \mathbb{R}^{C\times 1024}$.

For each labeled image, all M stream representations $\mathbf{v}_i^{(m)}[\mathcal{I}_c]$ are used to train the classifier with the ground-truth class label. For an unlabeled image accepted by Stability-Guided Self-Training ($I_i^{\mathrm{st}} = 1$), all M stream representations are likewise used with its accepted pseudo-label; rejected unlabeled samples do not update f_ψ. We optimize f_ψ independently using AdamW with learning rate 10^{-3}, weight decay 10^{-4}, class-label smoothing 0.05, and a cosine-annealing schedule with $T_{\max} = 10$. Its visual inputs are detached, so the classifier loss does not back-propagate to the prompts or frozen CLIP backbone.

For prediction, we average the stream-wise class probabilities:

$$\mathbf{p}_i^{\mathrm{cls}} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{p}_i^{(m,\mathrm{cls})}, \qquad \hat{y}_i^{\mathrm{cls}} = \arg\max_c p_{i,c}^{\mathrm{cls}}. \tag{A.10}$$

Thus, the shared-classifier prediction averages probabilities rather than logits or visual features.

A.3 BASELINE ADAPTATION DETAILS

All in-house baselines use the same underlying dataset partitions, fixed two-shot labeled split, and frozen CLIP ViT-B/16 backbone as SARA. The test set is used only for final evaluation.

CLIP-based SSL baselines. For **CAP** (Wang et al., 2025), we retain its original pseudo-label selection rule and SSL objective on the corresponding SSMDL split. CAP uses learnable prompts with prompt depth 8 and two context tokens. It is trained for 10 epochs with batch size 8, SGD, learning rate 0.01, weight decay 0.1, a cosine learning-rate schedule, and one warm-up epoch. The number of iterations per epoch is 100 for Office-Home and VLCS and 200 for FMoW and DomainNet.

For **CPL** (Zhang et al., 2024a), we follow the released textual prompt-learning implementation and retain its original pseudo-label construction and loss. CPL is trained for 50 epochs with batch size 8, SGD, learning rate 0.02, weight decay 0.05, temperature 1.0, and a cosine learning-rate schedule with two warm-up epochs. We otherwise follow its released method-specific settings.

Multi-domain learning baselines. We implement **Joint**, **Separate**, **MMoE** (Ma et al., 2018), and **PLE** (Tang et al., 2020) as frozen-CLIP prompt-based MDL baselines and equip them with FixMatch (Sohn et al., 2020) to exploit unlabeled data under the SSMDL setting. All four methods use prompt length 2 and prompt depth 8.

Joint uses a single prompt learner shared by all domains, consisting of one shallow prompt and seven deep prompts. The corresponding visual prompts are obtained from the textual prompts through the associated visual projections, and no domain-conditioned gate is used.

Table 9: Dataset statistics for Office-Home (Venkateswara et al., 2017), VLCS (Fang et al., 2013), FMoW (Christie et al., 2018; Koh et al., 2021), and DomainNet (Peng et al., 2019). Office-Home, VLCS, and DomainNet use the generated training, validation, and test splits, while FMoW follows the official WILDS split.

Dataset	Domains	# Domain	# Class	Train Size	Val Size	Test Size
Office-Home	Art, Clipart, Product, Real-World	4	65	12375	1557	1656
VLCS	Caltech, LabelMe, SUN, VOC	4	5	8574	1073	1078
FMoW	Asia, Europe, Africa, Americas, Oceania	5	62	363350	53025	53460
DomainNet	Clipart, Infograph, Painting, Quickdraw, Real, Sketch	6	345	476109	59524	60373

Separate maintains an independent Joint-style prompt learner for each domain. The M prompt learners do not share prompt parameters or gating modules, and each sample uses the learner associated with its domain.

MMoE uses M shared prompt experts at each prompt layer, where M is the number of training domains. For each domain, prompt layer, and modality, a learnable domain-conditioned softmax gate produces mixing weights over the M shared experts. Textual and visual prompts use separate gates.

PLE maintains one shared prompt and one private prompt for each domain at every prompt layer. For a sample from domain d, separate textual and visual domain-conditioned binary softmax coefficients combine the shared prompt with the private prompt associated with domain d. Thus, MMoE mixes multiple shared prompt experts, whereas PLE combines a shared prompt with a domain-specific prompt.

All four MDL baselines are trained for 10 epochs with batch size 8 and 200 iterations per epoch. They use SGD with learning rate 0.01, momentum 0.9, a warm-up stage followed by cosine learning-rate decay, while keeping the CLIP backbone frozen.

A.4 Dataset Details

For Office-Home, VLCS, and DomainNet, we construct fixed train/validation/test splits independently within each domain–class pair using an 80/10/10 ratio with a fixed split seed. Within the resulting training split, two images are sampled for each domain–class pair to form the labeled set, while the remaining training images form the unlabeled pool. For FMoW, we retain the official WILDS train/validation/test split and sample two labeled images per region–class pair from the official training set, with the remaining training images used as unlabeled data. The dataset partitions and two-shot labeled sets are fixed across experimental runs; different random seeds affect only model training rather than the data splits. Ground-truth labels associated with unlabeled samples are not used for self-training.

We provide additional details on the four benchmarks used in our experiments. The data split and evaluation protocols are described in the main paper and Appendix A.1.

Office-Home (Venkateswara et al., 2017) contains 15,588 images from 65 object categories across four domains: Art, Clipart, Product, and Real-World. The domains cover artistic depictions, clipart images, product images, and natural photographs.

VLCS (Fang et al., 2013) combines images from Caltech101, LabelMe, SUN09, and VOC2007 and defines a five-way classification task over the shared classes bird, car, chair, dog, and person.

FMoW (Christie et al., 2018; Koh et al., 2021) is a satellite image classification benchmark with 62 building or land-use categories. Following the WILDS protocol (Koh et al., 2021), we use five geographical domains: Asia, Europe, Africa, Americas, and Oceania. FMoW follows the official training, validation, and test split.

DomainNet (Peng et al., 2019) is a large-scale multi-domain object recognition benchmark containing approximately 0.6 million images from 345 categories across six domains: Clipart, Infograph, Painting, Quickdraw, Real, and Sketch. We use all six domains throughout our experiments.

Table 10: Detailed ablation of Stability-Guided Self-Training (ST) and Stability-Based Prediction Routing (PR). We report classification accuracy (mean±std, %) over three random seeds. The best result in each column is highlighted in bold.

Benchmark	ST	PR	Domain-wise Accuracy				
			Art	Clipart	Product	Real	Average
Office-Home	×	×	85.28±0.65	72.66±1.13	93.05±0.44	91.76±0.50	85.69±0.64
	✓	×	86.67±0.79	77.41±0.58	94.04±0.21	91.97±0.54	87.52±0.28
	×	✓	84.78±1.21	73.98±1.70	93.12±0.49	91.04±0.69	85.73±0.81
	✓	✓	**87.30**±0.22	**78.65**±0.46	**95.18**±0.33	**92.54**±0.33	**88.42**±0.05
			Caltech	LabelMe	SUN	VOC	Average
VLCS	×	×	**100.00**±0.00	68.66±3.97	75.66±0.97	88.46±0.78	83.20±0.59
	✓	×	99.53±0.40	**69.79**±0.94	83.03±1.89	89.45±1.20	85.45±0.52
	×	✓	**100.00**±0.00	69.16±1.77	71.72±1.22	87.67±0.95	82.14±0.30
	✓	✓	**100.00**±0.00	69.29±0.37	**83.74**±1.22	**89.64**±1.36	**85.67**±0.22
			Asia	Europe	Africa	Americas	Oceania	Average
FMoW	×	×	25.38±0.77	35.37±0.36	25.13±0.59	34.13±0.76	41.07±2.26	32.22±0.60
	✓	×	24.35±0.96	35.68±0.97	24.00±2.90	32.66±1.88	43.84±1.30	32.11±0.35
	×	✓	**26.32**±0.67	35.22±0.31	**31.48**±1.68	**34.66**±0.66	42.66±0.51	34.07±0.46
	✓	✓	25.87±0.65	**37.24**±0.75	26.76±4.75	33.80±2.04	**46.85**±3.32	**34.10**±0.51
			Clipart	Painting	Real	Sketch	Infograph	Quickdraw	Average
DomainNet	×	×	74.94±0.21	69.73±0.45	85.52±0.05	66.70±0.19	53.66±0.25	21.79±0.75	62.06±0.18
	✓	×	74.61±0.17	69.75±0.20	**85.55**±0.27	67.14±0.07	54.03±0.58	22.16±0.23	62.21±0.20
	×	✓	74.80±0.28	69.83±0.52	85.33±0.11	66.49±0.11	52.54±0.17	25.00±0.18	62.33±0.16
	✓	✓	**75.42**±0.25	**70.62**±0.09	85.43±0.33	**67.66**±0.15	**54.42**±0.19	**28.44**±0.09	**63.66**±0.04

B Additional Experimental Results

B.1 Detailed Component Ablation

Table 10 provides the complete domain-wise results of the component ablation. The two components exhibit different relative contributions across benchmarks: ST alone brings substantial gains of 1.83% and 2.25% on Office-Home and VLCS, respectively, while PR alone is particularly effective on FMoW, improving the average accuracy by 1.85%. More importantly, their benefits become consistently positive when used together. Adding PR on top of ST further improves Office-Home, VLCS, FMoW, and DomainNet by 0.90%, 0.22%, 1.99%, and 1.45%, respectively; conversely, adding ST to the PR setting also improves all four benchmarks by 2.69%, 3.53%, 0.03%, and 1.33%. This explains why PR alone provides limited gains on some datasets: without ST, the weighting network and shared classifier are learned only from the limited labeled data, whereas self-training allows these prediction components to further exploit unlabeled samples. The dataset-dependent gains also indicate different emphases of the two components: Office-Home and VLCS benefit more directly from self-training, while sample-specific routing plays a more prominent role on FMoW and DomainNet. Overall, the complete model achieves the highest benchmark-level accuracy in all four cases and improves or matches the setting without ST and PR on 17 of the 19 individual domains, with only minor decreases on FMoW-Americas (0.33%) and DomainNet-Real (0.09%). These results support that ST and PR mutually reinforce each other while contributing differently across datasets.

B.2 Analysis of Domain Perturbations and Domain-Specific Representations

Control study for domain perturbations. To examine whether cross-stream prediction variation reflects meaningful domain-conditioned perturbations rather than generic disagreement among multiple model paths, we perturb only the training-domain assignments while keeping class labels and the model architecture unchanged. Shuffle randomly permutes domain labels within the labeled and unlabeled training pools while preserving their domain counts, whereas Pseudo-equal randomly assigns samples to approximately balanced pseudo-domains. The final routed accuracy drops from

Table 11: Control study for domain perturbations (mean±std, %).

	Office-Home					VLCS				
Setting	Art	Clipart	Product	Real	Avg.	Caltech	LabelMe	SUN	VOC	Avg.
Pseudo-equal	86.16±1.21	76.61±1.43	94.82±0.54	92.40±0.25	87.50±0.42	99.53±0.81	64.92±12.76	76.87±10.41	82.15±10.19	80.87±8.49
Shuffle	86.29±0.95	77.63±0.96	95.04±0.44	92.33±0.33	87.82±0.43	100.00±0.00	64.54±11.21	79.70±6.36	88.17±1.80	83.10±4.64
Original	**87.30**±0.22	**78.65**±0.46	**95.18**±0.33	**92.54**±0.33	**88.42**±0.05	**100.00**±0.00	**69.29**±0.37	**83.74**±1.22	**89.64**±1.36	**85.67**±0.22

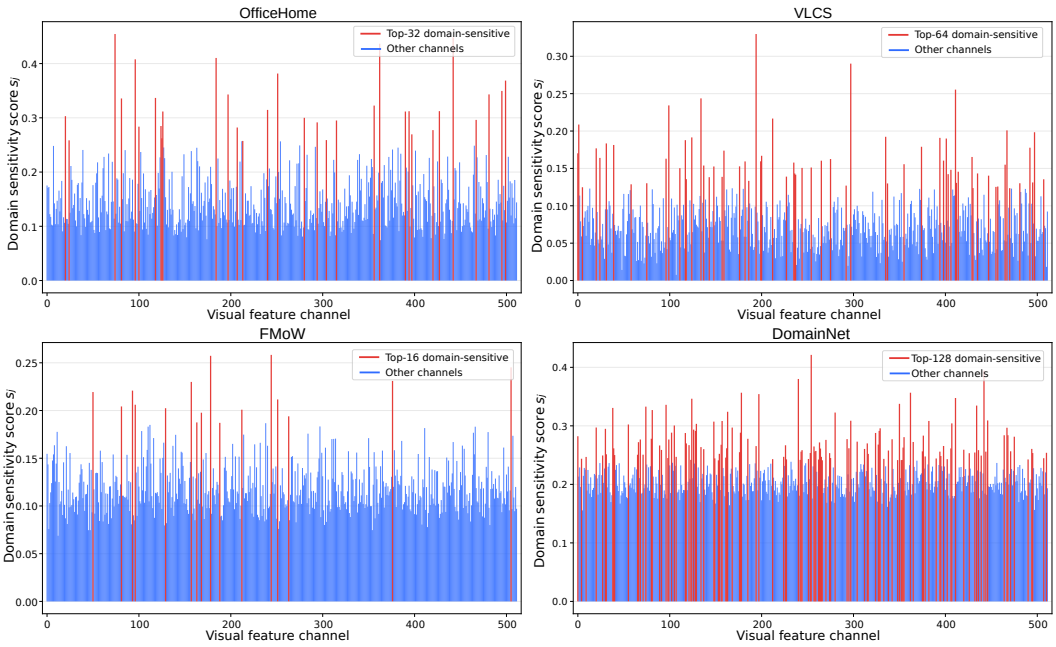


Figure 4: Channel-wise domain sensitivity of frozen CLIP visual features on the four benchmarks. Each bar shows the domain sensitivity score s_j of one visual channel, with the Top-K channels used by SARA highlighted for each dataset.

88.42% to 87.82% and 87.50% on Office-Home, and more substantially from 85.67% to 83.10% and 80.87% on VLCS, with markedly larger variance under the perturbed assignments. These results support the interpretation that useful cross-stream variation depends on meaningful domain-specific specialization rather than merely on independent multi-stream disagreement.

Domain-sensitive channel analysis. To better understand the domain-sensitive channels used by SARA, we analyze the channel-wise domain sensitivity scores s_j of the 512 frozen CLIP visual feature channels. Figure 4 visualizes the scores on the four benchmarks, with the Top-K channels actually used by SARA highlighted for each dataset. Table 12 further summarizes their distributions. Specifically, Mean measures the overall level of domain sensitivity, Std. reflects its variation across channels, P90 and P99 characterize the upper tail of the distribution, and Max denotes the most domain-sensitive channel.

The four benchmarks exhibit noticeably different domain-sensitivity profiles. DomainNet shows the strongest and broadest domain effect, with the highest Mean and P90, consistent with the relatively large $K = 128$. In contrast, Office-Home has the most pronounced high-sensitivity tail, as indicated by its highest Std., P99, and Max values, while the selected Top-32 channels focus on this strongly sensitive subset. VLCS has the lowest overall sensitivity, although a distinct subset of channels remains substantially more sensitive than the majority. FMoW exhibits the most uniform profile, with the smallest Std.; under this relatively flat distribution, $K = 16$ retains only the highest-scoring channels.

These observations show that domain sensitivity differs substantially across benchmarks in both magnitude and distribution, supporting dataset-specific selection of domain-sensitive channels rather

Table 12: Statistics of the domain sensitivity scores s_j over the 512 frozen CLIP visual feature channels. K denotes the number of domain-sensitive channels used by SARA for each benchmark.

Dataset	K	Mean	Std.	P90	P99	Max
Office-Home	32	0.153	0.062	0.232	0.380	0.455
VLCS	64	0.082	0.042	0.133	0.216	0.330
FMoW	16	0.121	0.029	0.158	0.221	0.258
DomainNet	128	0.221	0.040	0.272	0.354	0.421

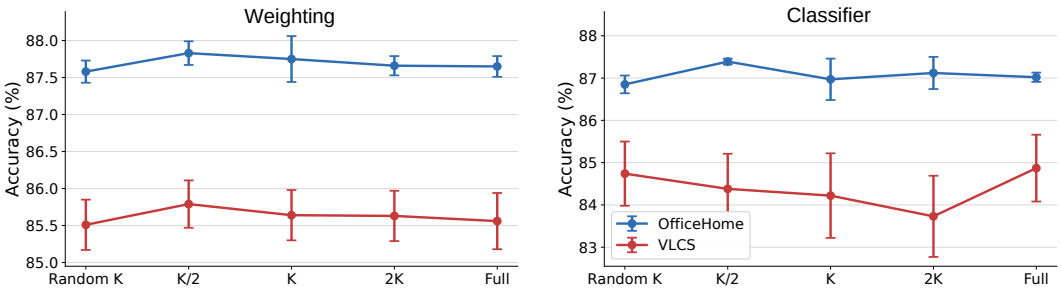


Figure 5: Sensitivity to the number of selected channels on Office-Home and VLCS. K denotes the dataset-specific default setting; Random-K randomly selects K channels, while $K/2$, K, and $2K$ follow the domain-sensitivity ranking. Full uses the complete feature representation. Error bars show the standard deviation over three runs.

than a universal K. In SARA, the selected set $\mathcal{I}_d$ provides domain-sensitive information for estimating stream affinities, while its complementary set $\mathcal{I}_c$ is used by the shared classifier to reduce reliance on domain-specific variations.

We choose the number of domain-sensitive channels K from the coarse candidate set $\{16, 32, 64, 128, 256\}$ according to the dataset-specific distribution of channel sensitivity scores. The selection considers both the magnitude of domain sensitivity and how strongly the high-sensitivity channels are concentrated. Accordingly, we use $K = 32$, 64, 16, and 128 for Office-Home, VLCS, FMoW, and DomainNet, respectively. For example, Office-Home exhibits a pronounced high-sensitivity tail, whereas FMoW has a relatively flat sensitivity distribution, for which a smaller K retains only the most sensitive channels. DomainNet exhibits stronger and broader domain sensitivity and therefore uses a larger K.

We further examine the sensitivity of SARA to the number of selected channels, using the original dataset-specific setting as K and comparing Random-K, $K/2$, K, $2K$, and Full. As shown in Figure 5, performance remains generally stable over a broad range of channel numbers, indicating that SARA does not depend on a narrowly tuned value of K. In particular, $K/2$ provides a favorable balance for both weighting-based and classifier predictions, suggesting that a relatively compact subset already captures sufficient useful information. Random-K exhibits less consistent behavior across datasets and prediction mechanisms, supporting the use of the domain-sensitivity ranking rather than arbitrary channel subsets. Using all channels provides little additional benefit for weighting-based prediction, whereas the classifier remains competitive and can benefit from the broader representation, especially on VLCS. Overall, these results show that the channel-selection design is robust to the exact choice of K while the two prediction mechanisms exhibit different preferences for feature breadth.

Stream affinity weights and feature fusion. To further examine the sample-specific prediction mechanism, we visualize the stream affinity weights and the resulting weighted visual representations. For each benchmark, we randomly select 20 test samples from each domain. As shown in Figure 6, each row corresponds to a test sample and each column to a domain-specific stream, with brighter values indicating larger affinity weights. Source-domain labels are used only to organize the visualization and are not available to the model at inference. For Figure 7, each test sample is passed through all M streams, producing M stream-wise visual representations $\{\mathbf{v}_i^{(m)}\}_{m=1}^{M}$. We apply LDA

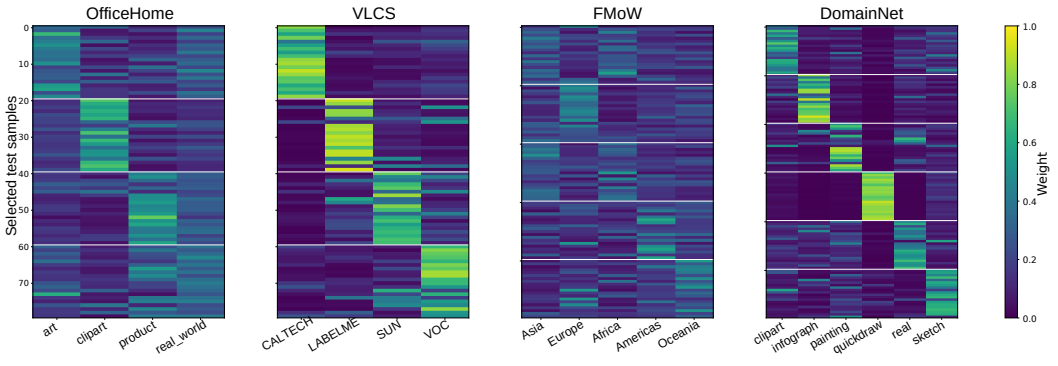


Figure 6: Sample-specific stream affinity weights on the four benchmarks. Rows denote test samples, columns denote domain-specific streams, and brighter values indicate larger affinity weights.

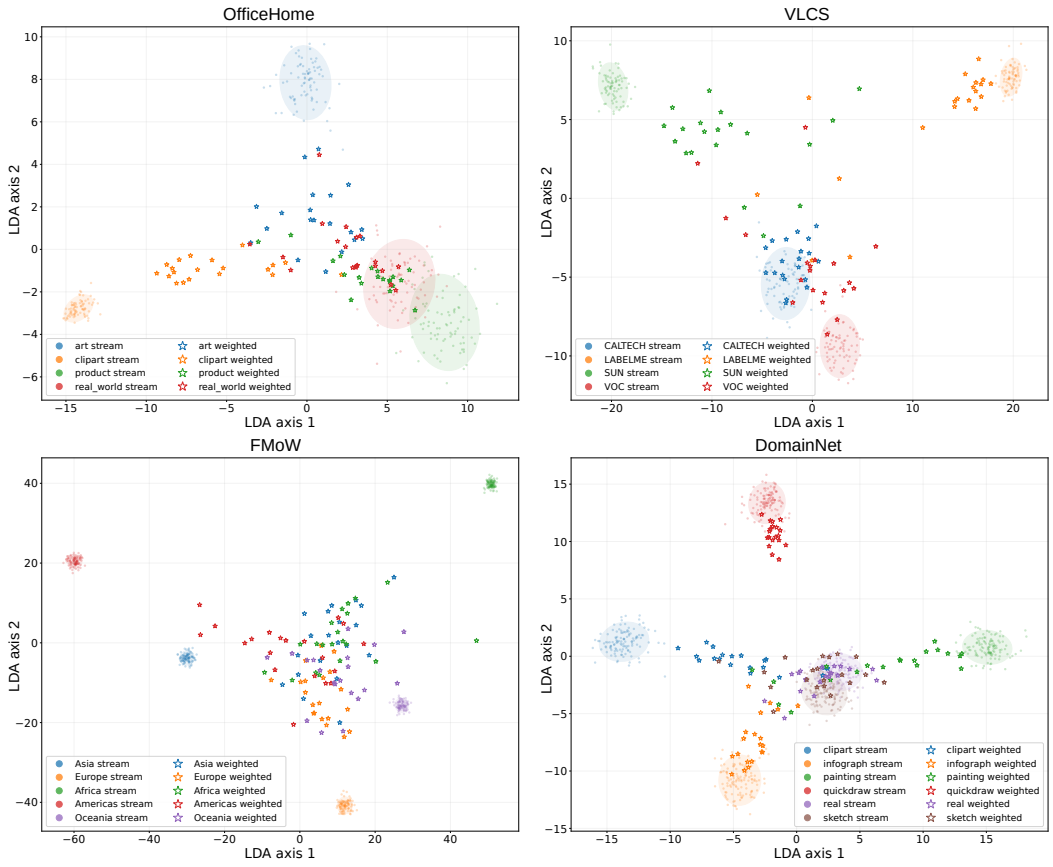


Figure 7: LDA visualization of stream-wise and weighted visual representations. Filled points and ellipses denote stream-wise features and their distributions, while stars denote weighted representations.

using stream identities and project the weighted representation $\bar{\mathbf{v}}_i$ into the same space; thus, each sample contributes M stream-wise points and one weighted point.

The learned affinity weights exhibit clear domain-related structure while remaining soft and sample-specific. VLCS and DomainNet show the most pronounced block structures, whereas Office-Home exhibits a softer but still structured allocation. FMoW shows more diffuse affinities, indicating stronger cross-stream mixing. Moreover, samples within the same source domain can receive

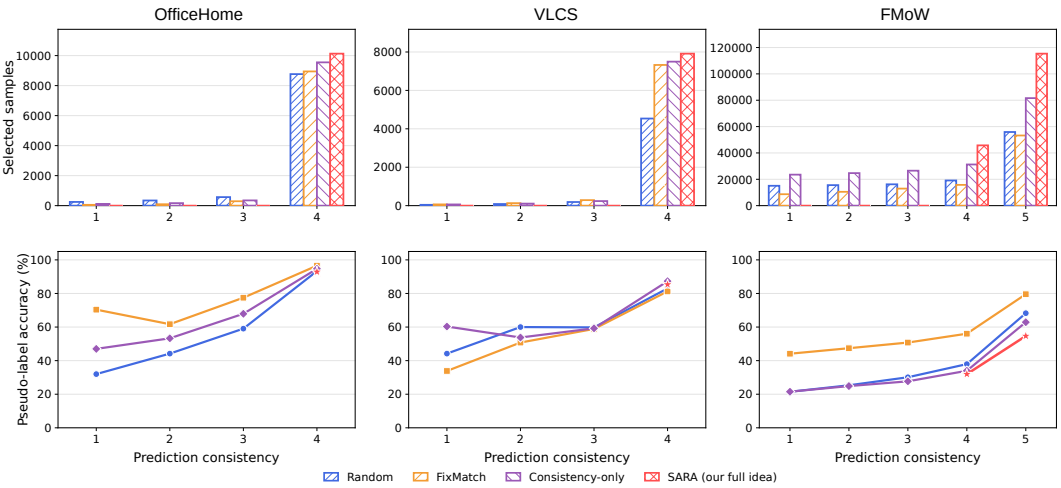


Figure 8: Pseudo-label selection across prediction-consistency levels on Office-Home, VLCS, and FMoW. The top row shows the number of selected samples, and the bottom row reports their pseudo-label accuracy.

noticeably different affinity patterns, showing that the weighting mechanism does not reduce to a fixed domain-level assignment or hard one-hot routing.

Figure 7 further shows how these affinities affect the representation geometry. Stream-wise representations form distinct stream-dependent regions, while weighted representations often lie between or near multiple stream clusters rather than coinciding with a single one. This pattern is particularly evident on FMoW, consistent with its more diffuse affinity weights, while VLCS and DomainNet show that a clear stream preference can still coexist with multi-stream fusion. Together, the two visualizations show that SARA learns sample-specific soft stream affinities and uses them to form fused representations rather than performing hard routing to a presumed test-domain stream.

B.3 Detailed Analysis of Stability-Guided Self-Training

Pseudo-label selection across consistency levels. Figure 8 further analyzes pseudo-label selection at different prediction-consistency levels, where the consistency level denotes the number of domain-specific streams agreeing with the candidate pseudo-label. Across the three benchmarks, pseudo-label accuracy generally increases with consistency, and this trend is also evident for Random and FixMatch, indicating that cross-stream consistency provides an informative reliability signal beyond the filtering rule itself. SARA concentrates its selected pseudo-labels in the high-consistency region, whereas Random and FixMatch retain more low-consistency samples that generally have lower pseudo-label accuracy. On FMoW, the additional decision-margin criterion further removes part of the less reliable consistency-4 samples while retaining a large number of consistency-5 samples. These results complement the overall statistics in Table 4 by showing how the different filtering strategies distribute their supervision across consistency levels.

Sensitivity to self-training thresholds. Figure 9 evaluates the effect of the self-training thresholds using the corresponding-domain stream for each test sample. This diagnostic requires test-domain labels and is used only to directly measure how different threshold settings affect the training quality of the domain-specific streams, without the influence of prediction routing. Overall, the corresponding-stream accuracy remains competitive over a reasonably broad range of threshold values, indicating that self-training does not rely on a narrowly tuned configuration. Office-Home and DomainNet are relatively stable under both threshold sweeps, whereas VLCS and FMoW exhibit larger variations, suggesting greater sensitivity to the quality–coverage trade-off of pseudo-label selection. FMoW in particular benefits from a more conservative consistency threshold, consistent with its more challenging pseudo-label filtering behavior observed above. The validation-selected default settings lie in competitive regions of the curves but are not consistently the test-optimal points, further indicating that the reported behavior is not due to fine-grained tuning on the test set.

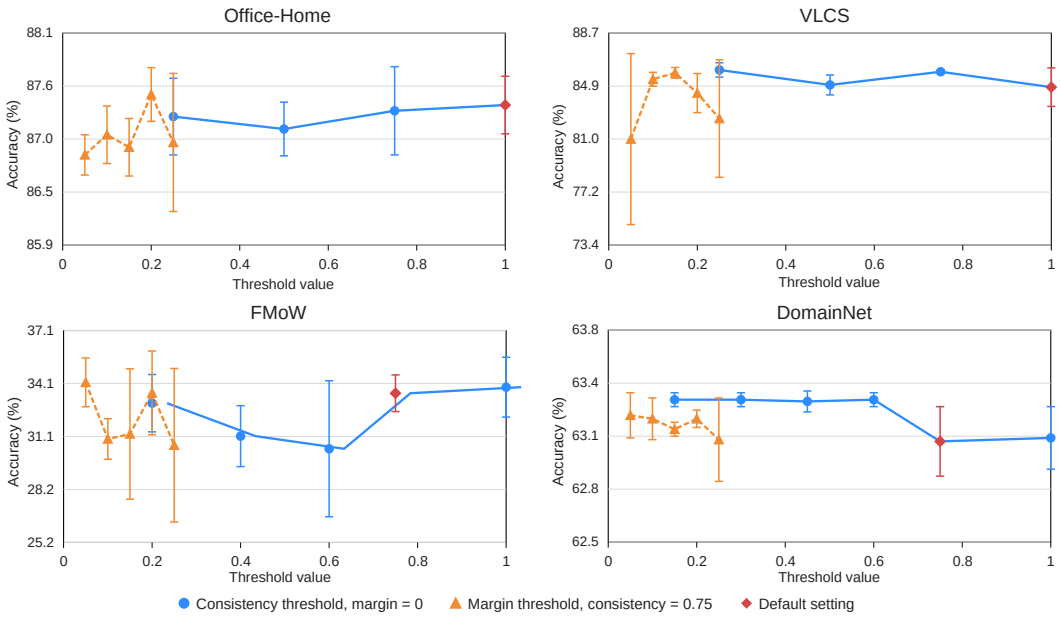


Figure 9: Sensitivity to the self-training thresholds on the four benchmarks. We report the accuracy of the corresponding-domain stream, which uses test-domain labels only for this diagnostic analysis. We vary the consistency threshold ρ_s with $\delta_s = 0$, and the margin threshold δ_s with $\rho_s = 0.75$. Red diamonds indicate the default settings selected on the validation set.

Table 13: Detailed domain-wise comparison of different prediction mechanisms on Office-Home and FMoW. All variants use the same fully trained model. † indicates that test-domain labels are required. We report classification accuracy (mean±std, %) over three random seeds.

Prediction	Office-Home					FMoW					
	Art	Clipart	Product	Real	Avg.	Asia	Europe	Africa	Americas	Oceania	Avg.
Corresponding stream†	86.54±1.70	76.75±1.44	94.89±0.21	**92.90**±0.99	87.77±0.68	25.30±1.76	35.58±1.61	26.48±3.81	32.37±1.22	43.97±3.10	32.74±0.27
Weighting	86.79±1.13	77.27±0.46	94.61±0.44	92.33±0.25	87.75±0.31	24.95±0.41	36.10±0.87	25.89±4.15	32.93±1.50	44.33±2.68	32.84±0.13
Classifier	85.53±1.57	77.27±0.70	94.47±0.21	90.61±0.33	86.97±0.49	25.86±0.92	**37.54**±0.73	26.63±4.79	33.79±2.43	46.57±3.63	34.08±0.71
Prediction Routing	**87.30**±0.22	**78.65**±0.46	**95.18**±0.33	92.54±0.33	**88.42**±0.05	**25.87**±0.65	37.24±0.75	**26.76**±4.75	**33.80**±2.04	**46.85**±3.32	**34.10**±0.51

B.4 Detailed Analysis of Stability-Based Prediction Routing

Domain-wise comparison of prediction mechanisms. Table 13 further shows that the benchmark-level gains of PR are consistent across individual domains. On Office-Home, PR outperforms both fixed prediction mechanisms on all four domains and achieves the best result on Art, Clipart, and Product, while also surpassing the corresponding stream on three of the four domains without requiring test-domain labels. On FMoW, the classifier is the stronger fixed mechanism, yet PR matches or improves it on four of the five regions and outperforms both weighting and the corresponding stream on every region. Consequently, PR achieves the highest average accuracy on both benchmarks, showing that its gains are broadly distributed across domains rather than being driven by only a few cases.

Comparison with alternative routing strategies. Equal probability averages the two predictive distributions as $\arg\max_c(p_c^{\mathrm{w}} + p_c^{\mathrm{cls}})/2$, while Max-confidence selects the prediction mechanism with the larger top-1 probability. Table 14 shows that stability-based routing achieves the highest benchmark-level accuracy on both Office-Home and VLCS, improving over both simple alternatives by about 0.4% and 0.2%, respectively. This indicates that the benefit does not arise merely from combining the two predictors or choosing the one with higher raw confidence; instead, cross-stream stability provides a more informative sample-specific signal for deciding which prediction mechanism to trust.

Table 14: Comparison with alternative routing strategies (mean±std, %). All strategies use the same trained model and differ only in how the weighting-based and classifier predictions are combined.

	Office-Home					VLCS				
Method	Art	Clipart	Product	Real	Avg.	Caltech	LabelMe	SUN	VOC	Avg.
Equal probability	86.42±0.00	78.22±0.13	95.39±0.12	92.04±0.43	88.02±0.11	100.00±0.00	68.16±0.99	83.74±1.22	89.94±0.89	85.46±0.72
Max-confidence	86.67±0.22	77.85±0.22	95.46±0.33	92.11±0.33	88.02±0.09	100.00±0.00	68.16±0.99	83.74±1.22	89.84±0.90	85.44±0.70
Stability-based routing	**87.30**±0.22	**78.65**±0.46	95.18±0.33	**92.54**±0.33	**88.42**±0.05	100.00±0.00	**69.29**±0.37	83.74±1.22	89.64±1.36	**85.67**±0.22

Table 15: Sensitivity to the prediction-routing thresholds ρ_p and δ_p. Each entry reports accuracy (%) using the same trained model. Bold values indicate the default settings selected on the validation set.

$\rho_p \backslash \delta_p$	0.00	0.05	0.10	0.15	0.20
Office-Home					
1.00	87.90	87.90	87.90	87.90	87.90
0.75	88.39	88.32	**88.42**	88.36	88.35
0.50	88.10	88.10	88.24	88.22	88.19
0.25	87.99	88.00	88.15	88.15	88.14
VLCS					
1.00	85.13	85.13	85.13	85.13	85.13
0.75	85.47	85.44	85.39	85.44	85.47
0.50	**85.67**	85.64	85.54	85.54	85.49
0.25	85.61	85.59	85.51	85.54	85.49
FMoW					
1.00	34.09	34.09	34.09	34.11	34.11
0.75	33.97	33.99	34.01	34.04	34.08
0.50	33.55	33.68	33.79	33.93	**34.04**
0.25	33.50	33.67	33.82	33.96	34.06

Sensitivity to routing thresholds. Table 15 shows that prediction routing is robust to a broad range of threshold combinations: across all 20 settings, the accuracy ranges are only 0.52%, 0.54%, and 0.61% on Office-Home, VLCS, and FMoW, respectively. The validation-selected defaults therefore do not rely on a narrow operating point. The results also reveal a sensible interaction between the two criteria: when the consistency requirement is relaxed, a larger decision margin can partially compensate by requiring stronger separation from competing predictions, which is particularly visible on FMoW.

B.5 Efficiency Analysis

Table 16 evaluates efficiency on six-domain DomainNet using a single NVIDIA L20 GPU with 224×224 inputs, AMP FP16, and batch size 1. SARA uses only 4.478M trainable parameters (2.99% of CLIP) and maintains peak inference memory comparable to single-stream PLE†. This efficiency in parameters and memory comes with higher inference computation: since test-domain labels are unavailable, SARA evaluates all M domain-specific streams, giving approximately $O(M)$ inference cost. On DomainNet, this results in 204.50 GFLOPs, close to six single-stream PLE† evaluations ($6 \times 34.08 = 204.48$ GFLOPs). Thus, SARA trades additional multi-stream computation for prediction without test-domain labels, and this cost may become more substantial as the number of domains grows. Sparse stream selection or lightweight pre-routing is a promising direction for improving scalability to larger domain sets.

B.6 Qualitative Analysis of Prediction Behaviors

Figure 10 presents representative test samples selected according to the correctness of the two prediction mechanisms. W and C denote the weighting-based prediction and the shared-classifier prediction, respectively. The examples are organized into four cases: both predictions are correct, only weighting is correct, only the classifier is correct, and both predictions are incorrect. These groups are defined by prediction correctness for qualitative analysis rather than by the stability assessment used for prediction routing.

For *Both correct* samples, the two prediction mechanisms generally agree with the ground-truth class, and the domain-specific streams also often exhibit highly consistent predictions. These examples

Table 16: Efficiency–accuracy trade-off on six-domain DomainNet. Class-conditioned text features are cached when measuring steady-state inference FLOPs. † denotes methods that require test-domain labels at inference.

Method	Train. Params (M) ↓	GFLOPs/img ↓	Peak Mem. (GB) ↓	Acc. (%) ↑
Joint (Wang et al., 2024)	3.160	34.07	0.640	62.02
Separate† (Wang et al., 2024)	18.960	34.07	0.794	62.30
PLE† (Tang et al., 2020)	3.209	34.08	0.655	62.57
CAP (Wang et al., 2025)	3.555	34.07	0.924	62.58
SARA	4.478	204.50	0.657	**63.66**

typically contain clear and distinctive category evidence, making them reliably classified by different prediction mechanisms.

For *Weighting only correct* samples, the weighting-based prediction succeeds while the shared classifier fails. In some cases, the corresponding-domain stream is also incorrect, and several examples are misclassified by all individual streams while the weighted prediction still recovers the correct class. This shows that weighting is not equivalent to selecting or copying a single stream prediction; soft aggregation can exploit class evidence distributed across multiple stream representations.

For *Classifier only correct* samples, the shared classifier recovers the correct class when the weighting-based prediction fails. This occurs both when the domain-specific streams disagree and when multiple streams share the same incorrect prediction. Such cases are consistent with the classifier operating on complementary channels and reducing dependence on domain-sensitive variations.

Finally, *Both incorrect* samples illustrate common failure cases of the two prediction mechanisms. These examples include atypical object appearances, multiple salient objects or strong contextual cues, fine-grained scene ambiguity in FMoW, and highly abstract visual content in DomainNet. Such intrinsic visual or semantic ambiguity can remain challenging for both mechanisms, suggesting object- or region-level reasoning as a useful direction for future work.

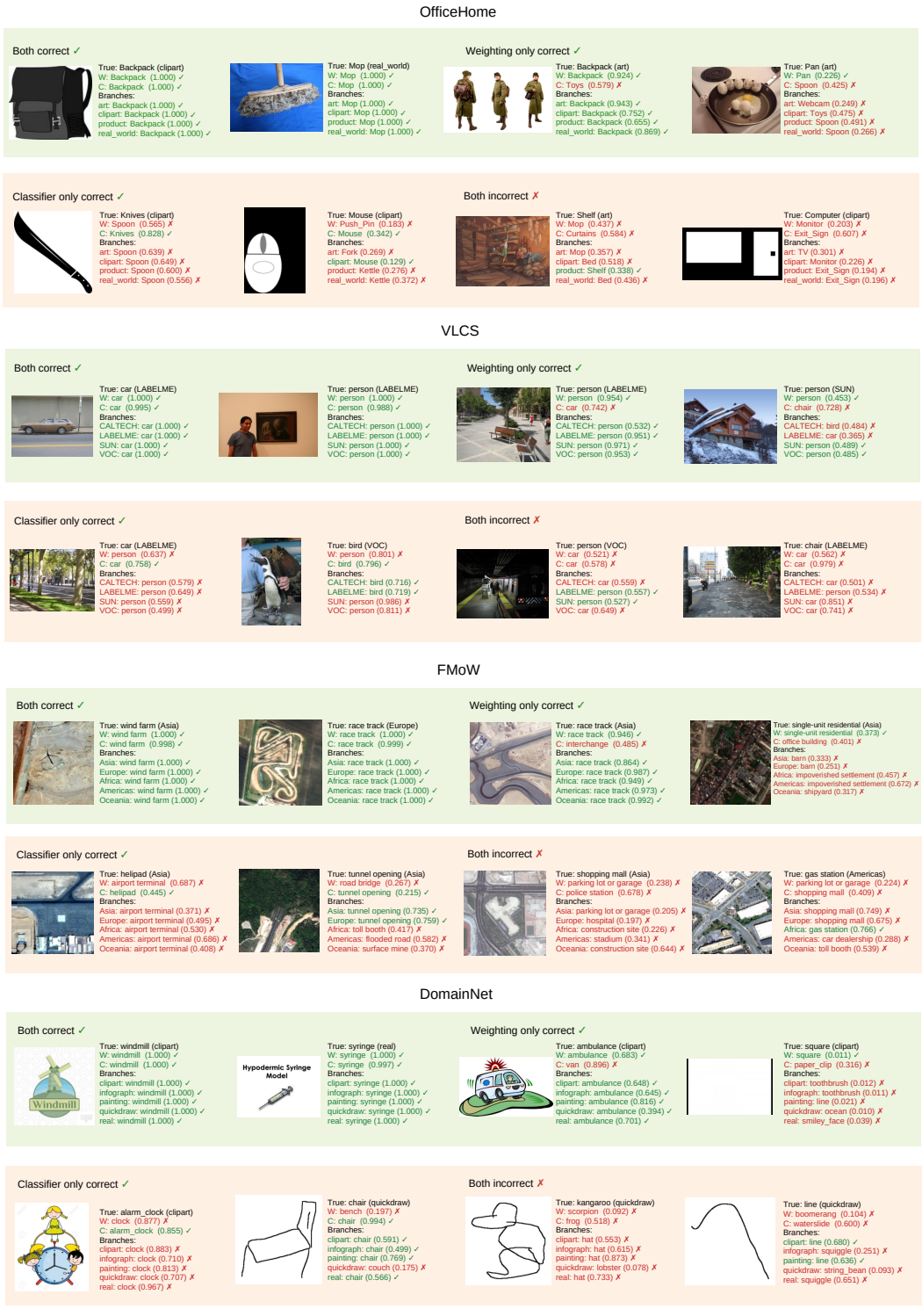


Figure 10: Representative examples grouped by the correctness of weighting and classifier predictions.