BEYOND NOISY PROXY: MITIGATING THE GAP FOR CLIP BASED SOURCE-FREE DOMAIN ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Source-free domain adaptation (SFDA) aims to adapt a pretrained source model to an unlabeled target domain without access to source data, aligning with the increasing demand for data security and privacy. Recent advances in vision-language models (VLMs), such as CLIP, provide a promising alternative by exploiting cross-modal knowledge as proxy supervision to facilitate SFDA. However, even with pretrained CLIP to guide SFDA, performance remains sub-optimal. Beyond attributing this to noisy or task-agnostic proxy supervision, we explore a more comprehensive perspective: the modality gap between the image and text embeddings. On the one hand, this gap hinders the CLIP proxy from early and fully exhibiting task-specific distinctions for SFDA. On the other hand, under domain shift, it inevitably induces noisy pseudo-labels from the source model and proxy model. To mitigate these challenges, we propose a unified framework, termed Semantic Cross-modal reAlignment with seLective Pseudo-labeling (SCALP), which facilitates SFDA via improved CLIP proxy supervision and reliable pseudo-label learning. Specifically, we first propose a Class-centroids Distributional Modality Realignment (CDMR) to match textual embeddings with confident visual class-centroids via a training-free affine transformation, so as to mitigate the modality gap. We subsequently induce an Agreement-based Selective Pseudo-label Learning (ASPL) strategy that leverages dual-model agreement to reduce noisy pseudo-labels during adaptation. Extensive experiments demonstrate that SCALP consistently improves adaptation performance across multiple benchmarks, achieving overall state-of-the-art results for SFDA.

1 INTRODUCTION

Modern deep models often suffer substantial performance degradation under domain shift (Mitsuzumi et al., 2024; Li et al., 2025; Lee et al., 2025). While unsupervised domain adaptation (UDA) mitigates this issue through cross-domain alignment (Ben-David et al., 2010; Tian et al., 2022), it usually requires persistent access to source data. Source-free domain adaptation (SFDA) (Liang et al., 2020) removes this requirement by adapting only a pretrained source model with unlabeled target data. Existing SFDA methods either reconstruct source distributions (Li et al., 2020; Ding et al., 2022) or exploit pseudo-labeling and target structures (Yang et al., 2021; Huang et al., 2021), but remain largely constrained by the knowledge encoded in the source model. To overcome this limitation, recent studies leverage pretrained vision-language models (VLMs), such as CLIP (Radford et al., 2021), as external proxies that render generalizable semantic priors. Nevertheless, even recast pretrained CLIP as external supervision for source model adaptation, the overall performance remains sub-optimal due to two key challenges (Zhang et al., 2023; Tang et al., 2024; 2025; Lee et al., 2025) *Challenge 1: Unsatisfied proxy supervision for SFDA*. *Challenge 2: Noisy pseudo-label during the adaptation training*.

Regarding *Challenge 1*, prior works mainly attribute the limited effectiveness of VLM proxies to task-agnostic or noisy supervision, and accordingly focus on proxy denoising or refinement (Xiao et al., 2024; Tang et al., 2024; 2025). However, these efforts remain sorely investigating at the prediction level rather than a more comprehensive level of the embedding space: the *modality gap* (Liang et al., 2022b). The misalignment between visual and textual embeddings hinders class distinctions and recasts a task-specific CLIP proxy for SFDA, as illustrated in Fig. 1 (a).

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.

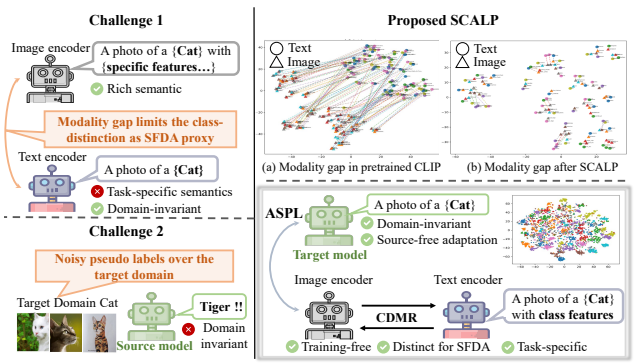


Figure 1: Overview of two challenges in SFDA with pretrained CLIP (*Left*), our proposed solutions (*Right*).

Prior studies (Hu et al., 2024; Schrodi et al., 2025) suggest that this gap arises from semantic imbalance across modalities. As image encoders are assumed to retrieve specific and domain-invariant semantics from input data of target domain (Tang et al., 2025), whereas customized textual prompt (e.g., *"A photo of a Cat"*) remains informatively sparse. As a result, this sparsity hinders early and full expression of task-oriented semantics retrieved by the image encoder, corresponding to *Challenge 1* in Fig. 1. Consequently, the proxy supervision exhibits task-agonist vague, puzzling inter-class distinctions among textual embeddings corresponding to different classes. On the other hand, despite the incorporation of proxy supervision, domain shift and residual proxy bias inevitably produce noisy pseudo-labels during adaptation training as *Challenge 2* in Fig. 1.

To mitigate both *Challenge 1* and *2*, we propose a unified framework, Semantic Cross-modal re-Alignment with seLective Pseudo-labeling (SCALP), which enhances the task-specific distinction of CLIP proxy before adaptation while reducing noisy pseudo-labels throughout training. Unlike existing solutions that mitigate the modality gap through post-hoc feature shifting (Liang et al., 2022b; Eslami & de Melo, 2025) or paired-semantic contrastive learning via captions (Grassucci et al., 2026; Yin et al., 2026), these methods are inapplicable to SFDA, where only unlabeled target images are available without paired captions. Accordingly, we propose Class-centroids Distributional Modality Realignment **(CDMR)** to mitigate the modality gap through a training-free affine transformation tailored to SFDA. Specifically, CDMR aligns semantic-sparse text embeddings with semantic-informative vision centroids estimated from target image embeddings regarding different classes, thereby realigning cross-modal semantics and reformulating task-specific proxy supervision before adaptation, as shown in Fig. 1 (b). Furthermore, we propose an Agreement-based Selective Pseudo-label Learning (**ASPL**) strategy, which assigns one-hot labels to dual-agreed samples for supervision, while applying neighbor-based consistency to disagreed samples to align local structure without introducing additional noise. Our main contributions are summarized:

- We reveal the modality gap, an overlooked yet essential factor that results in sub-optimal proxy performance toward SFDA, as the intrinsic image-text semantic misalignment in pretrained CLIP weakens task-specific proxy distinct patterns.
- We propose SCALP, a unified framework that refines the CLIP proxy to be task-specific while reduces noisy pseudo-labels during the adaptation training process, laying a foundation of developing the promising CLIP-based SFDA methods.
- In SCALP, we propose CDMR, which exploits a training-free affine transformation to realign the image–text embeddings of confident targe -domain samples, thereby recasting CLIP as a task-specific proxy without target labels or paired captions. Based on this refined proxy, ASPL is introduced to assign selective supervision for reliable SFDA.

2 Related Work

Source Free Domain Adaptation. SFDA is more challenging than conventional domain adaptation, as only a pretrained source model and unlabeled target data are available. Consequently, most domain alignment methods become inapplicable due to the absence of source data (Kim et al., 2021; Bourtoule et al., 2021; Liang et al., 2022a). Although recent methods attempt to address this limitation through self-training or proxy refinement, the available transferable knowledge remains inherently restricted, often resulting in performance saturation under severe domain shifts. This motivates the introduction of external knowledge to break such adaptation bottlenecks, where pretrained VLMs provide a promising solution owing to their rich and domain-invariant knowledge.

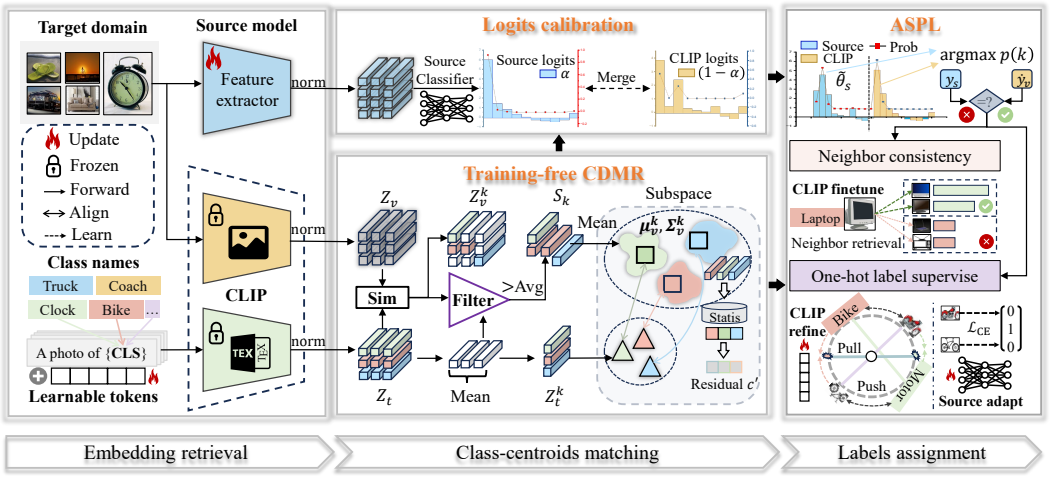


Figure 2: Overview of our SCALP. We mitigate two challenges via three actions in sequence.

Pretrained Vision-Language Models for SFDA. Pretrained VLMs, especially CLIP, provide rich prior semantics for SFDA by offering external guidance on unlabeled target data (Zhang et al., 2023; Tang et al., 2025). However, directly treating proxy predictions as ground truth leads to an unsatisfying proxy (Xiao et al., 2024; Zhang et al., 2025). Prior works mainly attribute this to the mismatch between task-generic CLIP priors and task-specific adaptation, thereby refining CLIP through cross-model alignment or pseudo supervision (Xiao et al., 2024; Tang et al., 2024; 2025; Lee et al., 2025). Beyond such noisy or task-agnostic supervision, we further identify the intrinsic image-text *modality gap* as another important factor limiting CLIP-based SFDA.

Modality Gap. The modality gap is pervasive in VLMs such as CLIP, where image and text embeddings are separated in the latent space (Liang et al., 2022b; An et al., 2025; Zhu et al., 2025b). In the classification task, this gap leads to overly similar class-oriented textual embeddings (Hu et al., 2024), weakening the discriminability and robustness for zero-shot (Chowers et al., 2026). However, existing realignment methods either depend on post-hoc feature shifting (Liang et al., 2022b) or group-semantic contrastive learning (Grassucci et al., 2026), requiring image-text pairs under the same class, that are unavailable in SFDA. Since SFDA sorely offers target images without any labels or paired captions. This motivates us to develop a new SFDA-compatible modality realignment strategy that leverages target visual semantics without requiring paired cross-modal supervision.

3 Methodology

3.1 Preliminary

Problem Setup. In SFDA, the labeled source domain $\mathcal{D}_S = \{\mathcal{X}_s, \mathcal{Y}_s\}$ is inaccessible, while only N unlabeled target images $\mathcal{X}_t = \{\mathbf{x}_t^i\}_{i=1}^N$ from $\mathcal{D}_T = \{\mathcal{X}_t, \mathcal{Y}_t\}$ are available. Adaptation relies on a pretrained source model and a pretrained CLIP proxy. For CLIP, the frozen image encoder produces $\mathrm{z}_v^i = f_v(\mathbf{x}_t^i)$, while the frozen text encoder generates class-specific embeddings $\mathrm{Z}_t = g_t(\mathcal{T}, v)$ from task-customized prompts $\mathcal{T}$ and learnable tokens v; the VLM logits are $\theta_v(\mathbf{x}_t^i) = \mathrm{sim}(\mathrm{Z}_v^i, \mathrm{z}_t)$. The source model predicts $\theta_s(\mathbf{x}_t) = h_s(f_s(\mathbf{x}_t))$, where h_s is frozen and f_s is adapted to the target domain, with soft predictions obtained by the softmax function $\phi(\cdot)$. The core problem of SFDA is to adapt a pretrained source model $\theta_s : \mathcal{X}_s \to \mathcal{Y}_s$ over unseen target domain, yielding an accurate target model $\theta_t : \mathcal{X}_t \to \mathcal{Y}_t$ against the *challenge 1* and *2*. Three terms are provided (1) pretrained source model θ_s, (2) unlabeled target domain $\mathcal{X}_t$, and (3) pretrained CLIP: θ_v as the external proxy.

3.2 Semantic cross-modal Realignment with Selective Pseudo-labeling

Overview. As illustrated in Fig. 2, **SCALP** follows sequential actions from the embedding retrieval, class-centroids matching, and pseudo-labels for each target $\mathbf{x}_t$. During the embedding re-

trieval, we construct visual class-centroids and a target neighbor bank from unlabeled target image embeddings. For class-level, CDMR aligns textual embeddings with retrieved visual class-centroids to mitigate modality gap while ensure a task-specific proxy in the beginning. Finally, ASPL exploits dual-agreement to selectively assign hard labels for each image or conducts the neighbor consistency to constrain each noisy prediction for adaptation.

3.2.1 CLASS-CENTROIDS DISTRIBUTIONAL MODALITY REALIGNMENT.

As noted in *Challenge 1*, customized text templates fail to capture essential task-specific semantics, exacerbating the modality gap and producing noisy pseudo-labels. To address this, we propose CDMR, a class-level affine transformation that realigns textual class-centroids toward visual ones, thereby mitigating the modality gap without specific labels or paired captions.

Class-centroids from Confidence-filtered. Given a target image $\mathbf{x}_t^i$, we extract its frozen CLIP visual embedding $\mathrm{z}_v^i = f_v(\mathbf{x}_t^i)$ and assign it to the class k related textual embedding z_t^k, $\hat{\mathrm{y}}_v^i = \arg\max_k \mathrm{sim}(\mathrm{z}_v^i, \mathrm{z}_t^k)$. Unlike prior works Liang et al. (2022b) directly apply post-hoc shifts without the awareness of grouped semantics across modalities, we first gather class-assigned embeddings $\mathcal{B}_k = \{\mathrm{z}_v^i \mid \hat{y}_v^i = k\}$, and then retain samples whose image-text similarity is higher than the class-wise average:

$$\mathcal{S}_k = \left\{ \mathrm{z}_v^i \;\middle|\; \mathrm{sim}(\mathrm{z}_v^i, \mathrm{z}_t^k) \geq \frac{1}{|B_k|} \sum_{\mathrm{z}_v^j \in \mathcal{B}_k} \mathrm{sim}(\mathrm{z}_v^j, \mathrm{z}_t^k) \right\}. \tag{1}$$

This class-wise filter is to select those candidates with initial higher confidence, thus only responsible for class-level coarse filter and statistical textual realignment, rather than the decisive criterion for reliable supervision. The k-th class centroid is computed by taking the expectation over each selected visual embeddings with respect to the class k, thus $\mu_v^k = \frac{1}{|\mathcal{S}_k|} \sum_{\mathrm{z}_v^i \in \mathcal{S}_k} \mathrm{z}_v^i$. Fine-grained reliability will be subsequently enhanced by ASPL at the instance level.

Distributional Modality Alignment. Inspired by Zhang et al. (2024), CDMR exploits class-level distributional alignment by matching the statistical properties of textual and visual embeddings under class semantics as Fig. 2. A statistical motivation of this realignment is provided in Appendix A. Let $\mathrm{z}_t^k \in \mathbb{R}^d$ denote the textual embedding of the class k, with mean μ_t^k and diagonal variance σ_t^k. μ_v^k and σ_v^k denote the corresponding statistics derived from visual class-centroids. Following the relationship observed in prior work (Liang et al., 2022b; Zhang et al., 2024), we assume after alignment, ideal class textual embedding satisfies with respect to a small random variant $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$,

$$\dot{\mathrm{z}}_t^k - \mu_v^k = \mathbf{c}' + \epsilon \quad \text{with } \mathbf{c}' = \mu_v^{\star,k} - \mu_v^k, \tag{2}$$

where $\dot{\mathrm{z}}_t^k$ denotes the aligned class embedding, and σ denotes a small variance, and $\mathbf{c}'$ refers to an estimation residual between ground truth class-centroids $\mu_v^{\star,k} = \mathbb{E}\left[\mathrm{z}_v^i | y^\star = k\right]$ on entire image embeddings and estimated class-centroids $\mu_v^k = \mathbb{E}_{\mathrm{z}_v^i \sim \mathcal{S}_k}\left[\mathrm{z}_v^i | \hat{\mathrm{y}}_v^i = k\right]$. ϵ captures the remaining stochastic discrepancy, viewed as Gaussian noise with small variance. Then, a linear affine,

$$\dot{\mathrm{z}}_t^k \approx \mathbf{A}_k \left(\mathrm{z}_t^k + \frac{\mathbf{c}'}{\mathbf{A}_k} - \mu_t^k \right) + \mu_v^k, \tag{3}$$

aligns the estimated μ_t^k and σ_t^k of the k-th class textual embeddings to relative visual class-centroids, we have $\mathbf{A}_k \approx \mathrm{diag}\left(\sqrt{\frac{\sigma_v^k}{\sigma_t^k + \varepsilon}}\right)$, where $\varepsilon > 0$ is a small constant for numerical stability and $\sigma_v^k = \mathbb{E}_{\mathrm{z}_v^i \sim \mathcal{S}_k}\left[\left(\mathrm{z}_v^i - \mu_v^k\right)^2\right]$. Here, μ_t^k and $\boldsymbol{\sigma}_t^k$ denote the element-wise mean and variance of the class-k embeddings through following prompt pool, respectively.

Specially, given a prompt pool $E_k = \{\mathcal{T}_e(k)\}_{e=1}^{180}$ (Lin et al., 2023), they are computed as $\mu_t^k = \mathbb{E}_{\mathcal{T}_e \sim E_k}\left[g_t(\mathcal{T}_e(k))\right]$ and $\sigma_t^k = \mathbb{E}\left[\left(g_t(\mathcal{T}_e(k)) - \mu_t^k\right)^2\right]$. $\mathcal{T}_e(k)$ can be any text prompts, e.g. "A photo of $\{cls\text{-}k\}$" or "A rendering of $\{cls\text{-}k\}$" in the pool, $g_t(\cdot)$ is the CLIP text encoder and $(\cdot)^2$ denotes element-wise square. The variance σ_t^k is used for diagonal variance calibration between the textual prompt distribution and the visual class distribution. Based on the residual bias $\mathbf{c}'$ in Eq. 2, we further introduce parametrize prompt to refine the textual embeddings toward more fine-grained

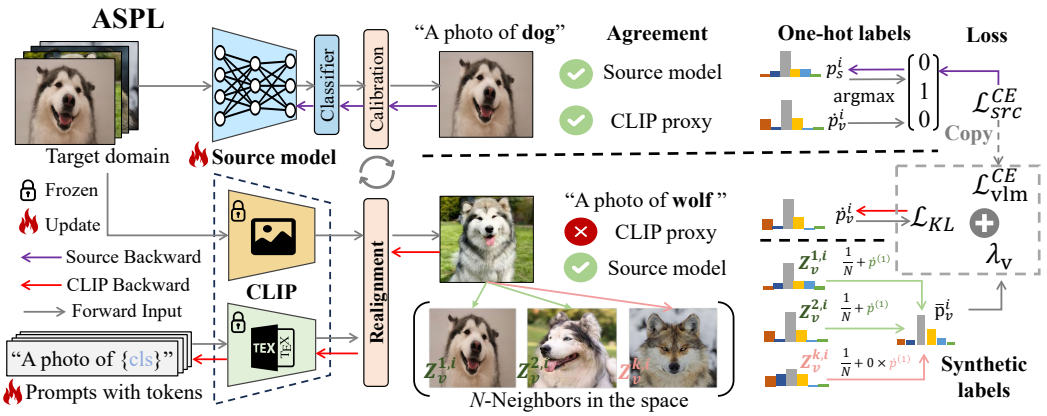


Figure 3: Overview of ASPL. Dual-agreed samples are assigned with one-hot supervision, while those disagreed are assigned with synthetic labels that weighted over the N nearest neighbors $\mathrm{z}_v^{i,j}$.

semantics for each individual z_v^i. Specifically, we aim to minimize the residual bias by inducing a learnable tokens that can be absorbed and optimized by parameterized textual embeddings, then reformulate the solution as:

$$\dot{\mathrm{z}}_t^k \leftarrow \mathbf{A}_k \left(g_t(\mathcal{T}(k), v) - \mu_t^k\right) + \mu_v^k, \tag{4}$$

where g_t is the CLIP text encoder, and $\mathcal{T}(k)$ denotes the class-k prompt with learnable tokens v, reducing the residual $\mathbf{c}'$. Since CDMR derives class-level semantics from confident samples in $\mathcal{S}_k$, its coverage remains limited for uncertain target instances. Thus, CDMR serves as a task-specific proxy initialization, while subsequent instance-level refinement further captures fine-grained target semantics, as illustrated in Fig. 2.

3.2.2 Agreement-based Selective Pseudo-label Learning.

Building upon the class-level aligned CLIP proxy and the pretrained source model, we propose Agreement-based Selective Pseudo-label Learning (ASPL) to further refine pseudo supervision at the instance level. In contrast to CDMR's class-level alignment, ASPL performs fine-grained adaptation by selectively assigning supervision according to prediction reliability of individuals. Specifically, reliable samples are supervised with one-hot pseudo-labels, while disagreed samples are regularized via neighbor-based synthetic targets, preventing noise propagation during adaptation.

Logits Calibration. Before dual-agreement selection, we calibrate the source logits using the CLIP predictions, following the logits merging principle in Tang et al. (2024):

$$\tilde{\theta}_s^i(1-\alpha)\,\mathrm{Norm}\left(\theta_s^i\right) + \alpha\,\mathrm{Norm}\left(\mathrm{sim}\left(\mathrm{z}_v^i, \dot{\mathrm{Z}}_t\right)\right), \tag{5}$$

where $\alpha \in (0,1)$ controls the calibration weight, $\dot{\mathrm{Z}}_t = \{\dot{\mathrm{z}}_t^k\}_{k=1}^K$ denotes the realigned textual class K embeddings, and $\mathrm{Norm}(\cdot)$ normalizes logits to a comparable scale before merging. The calibrated source logits are subsequently used for dual-agreement selection.

Selective Pseudo-label Assignment via Dual-agreement. To improve pseudo-label quality and stabilize adaptation, we introduce a selective pseudo-label assignment strategy based on dual-model agreement, as shown in Fig. 3. For the i-th target sample $\mathbf{x}_t^i$, the realigned CLIP proxy prediction and source-model prediction are defined as

$$\dot{p}_v^i = \phi\left(\mathrm{sim}(\mathrm{z}_v^i, \dot{\mathrm{Z}}_t)\right), \quad p_s^i = \phi(\tilde{\theta}_s^i), \tag{6}$$

where z_v^i is the i-th sample's visual embedding via CLIP, $\dot{\mathrm{Z}}_t$ denotes all realigned textual class-centroids with learnable tokens, and $\phi(\cdot)$ is the softmax function. We then define a binary mask:

$$m^i = \mathbf{1}\left[\arg\max_k p_s^i(k) = \arg\max_k \dot{p}_v^i(k)\right], \mathbf{y}_i = \delta(\dot{p}_v^i), \tag{7}$$

where $\mathbf{1}[\cdot]$ is the indicator function and $\delta(\cdot)$ denotes one-hot encoding. For samples with $m^i = 1$, the source model and CLIP proxy provide consistent predictions, and thus we assign one-hot pseudo-labels for supervised adaptation. For samples with $m^i = 0$, we instead exploit local neighborhood structure in the CLIP visual embedding space to construct synthetic soft labels as Fig. 3. The intuition is that synthesizing neighbor information exploits local smoothness to deal with shifts, due to local hypothesis Zhou et al. (2003). Distinct from uniform synthesis, we improve supervised contributions from dual-agreed neighbors, in case the propagation of noisy semantics. Specifically, for each embedding z_v^i, we retrieve its N-nearest neighbors in the f_v encoded space, $\mathcal{N}_i = \left\{\mathrm{z}_v^{i,j}\right\}_{j=1}^{N}$. The neighbor bank is constructed once during the initial forward pass and stored as a static feature bank for efficient retrieval, due to a frozen image encoder throughout adaptation. For each neighbor $\mathrm{z}_v^{i,j}$, we compute its realigned proxy prediction $\dot{p}_v^{i,j}$ following Eq. 6. Let $m^{i,j} \in \{0, 1\}$ denote the dual-agreed indicator. We assign each neighbor a variable weight:

$$w^{i,j} = \frac{1}{N} + \mathbf{1}[m^{i,j} = 1] \cdot \max_k \dot{p}_v^{i,j}(k), \tag{8}$$

where $\max_k \dot{p}_v^{i,j}(k)$ is the highest class probability of the j-th neighbor, denoted as $p^{(1)}$ in Fig. 3. This design assigns larger weights to confident dual-agreed neighbors, while the uniform term $\frac{1}{N}$ preserves contributions from all neighbors and prevents the synthetic target from being dominated by a single possibly noisy prediction. The synthetic soft label for the disagreed sample is then obtained by normalized neighbor aggregation:

$$\bar{\mathbf{p}}_v^i = \frac{\sum_{j=1}^{N} w^{i,j} \dot{p}_v^{i,j}}{\sum_{j=1}^{N} w^{i,j}}. \tag{9}$$

This normalization ensures a valid probability distribution and enables uncertain samples to be regularized by reliable local semantics. Compared with direct hard pseudo-labeling, the proposed neighbor-based synthetic target reduces prediction variance, alleviates noisy-label propagation, and improves training stability for dual-disagreed samples.

3.3 Overall Training Objectives

ASPL selectively employs a cross-entropy optimization according to the dual-agreement mask m^i. For dual-agreed samples, we use one-hot pseudo-labels to refine both the CLIP proxy and source model:

$$\mathcal{L}_{\mathrm{vlm}}^{\mathrm{CE}} = \mathbb{E}_{m^i=1}\left[\mathrm{CE}(\dot{p}_v^i, \mathbf{y}_i)\right], \quad \mathcal{L}_{\mathrm{src}}^{\mathrm{CE}} = \mathbb{E}_{m^i=1}\left[\mathrm{CE}(p_s^i, \mathbf{y}_i)\right], \tag{10}$$

where $\dot{p}_v^i$ and p_s^i denote the realigned proxy prediction and source prediction, respectively, and $\mathbf{y}_i$ is the one-hot pseudo-label derived from Eq. 7. For those samples with $m^i = 0$, hard pseudo-labels are prevented, we instead regularize proxy predictions by the synthetic soft $\bar{\mathbf{p}}_v^i$ in Eq. 9:

$$\mathcal{L}_{\mathrm{vlm}}^{\mathrm{KL}} = \mathbb{E}_{m^i=0}\left[\mathrm{KL}\left(\mathrm{sg}(\bar{\mathbf{p}}_v^i) \parallel \dot{p}_v^i\right)\right], \tag{11}$$

where KL refers to the KL-divergence and sg indicates the stop-gradient action. We impose a λ_{v} to balance two proxy refinements, thus the overall training objective of two models are then defined as

$$\mathcal{L}_{\mathrm{vlm}} = \lambda_{\mathrm{v}} \mathcal{L}_{\mathrm{vlm}}^{\mathrm{CE}} + \mathcal{L}_{\mathrm{vlm}}^{\mathrm{KL}}, \quad \mathcal{L}_{\mathrm{src}} = \mathcal{L}_{\mathrm{src}}^{\mathrm{CE}}. \tag{12}$$

Here, $\mathcal{L}_{\mathrm{vlm}}^{\mathrm{CE}}$ helps to recast the proxy task-specific discriminability via reliable hard supervision, while $\mathcal{L}_{\mathrm{vlm}}^{\mathrm{KL}}$ regularizes noisy supervision through soft neighborhood consistency. Finally, the source model is lead to achieve reliable source-free adaptation by learning with dual-agreed supervision with respect to $\mathcal{L}_{\mathrm{src}}^{\mathrm{CE}}$. The pseudo-code of SCALP is provided in Appendix E.

4 Experiments

4.1 Setups

Datasets. We evaluate on four standard domain adaptation benchmarks: **Office-31** (Saenko et al., 2010) contains three domains: Amazon (A), DSLR (D), and Webcam (W), with 31 classes. **Office-Home** (Venkateswara et al., 2017) includes 65 classes across four domains: Art (A), Clipart (C), Product (P), and RealWorld (R). **VisDA-2017** (Peng et al., 2017) consists of synthetic and real domains. **DomainNet-126** (Saito et al., 2019) is a large-scale dataset with four domains: Clipart (C), Sketch (S), Painting (P), and Real (R).

Table 1: Comparison on Office-Home in SFDA setting, where V denotes whether a frozen CLIP ViT-L/14@336 model is used as the proxy. Full results are provided in Appendix B.1.

			Office-Home												
Method	Venue	V	A→C	A→P	A→R	C→A	C→P	C→R	P→A	P→C	P→R	R→A	R→C	R→P	Avg
Source	ICML20	×	43.7	67.0	73.9	49.9	60.1	62.5	51.7	40.9	72.6	64.2	46.3	78.1	59.2
CLIP-L/14@336	Zero-shot	✓	72.6	86.5	85.2	81.8	86.5	85.2	81.8	72.6	85.2	81.8	72.6	86.5	81.5
SHOT	ICML20	×	56.7	77.9	80.6	68.0	78.0	79.4	67.9	54.5	82.3	74.2	58.6	84.5	71.9
AaD	NIPS22	×	59.3	79.3	82.1	68.9	79.8	79.5	67.2	57.4	83.1	72.1	58.5	85.4	72.7
CoWA	ICML22	×	56.9	78.4	81.0	69.1	80.0	79.9	67.7	57.2	82.4	72.8	60.5	84.5	72.5
ELR	ICLR23	×	58.4	78.7	81.5	69.2	79.5	79.3	66.3	58.0	82.6	73.4	59.8	85.1	72.6
C-SFDA	CVPR23	×	60.3	80.2	82.9	69.3	80.1	78.8	67.3	58.1	83.4	73.6	61.3	86.3	73.5
Co-learn	ICCV23	×	69.6	89.5	91.2	82.7	88.4	91.3	82.5	68.5	91.5	82.8	71.3	92.1	83.5
I-SFDA	CVPR24	×	60.7	78.9	82.0	69.9	79.5	79.7	67.1	58.8	82.3	74.2	61.3	86.4	73.4
R-SFDA	CVPR25	×	61.2	80.9	82.7	69.3	81.2	81.4	68.1	58.8	83.4	74.6	62.4	85.7	74.1
UC-SFDA	ICLR25	×	–	–	–	–	–	–	–	–	–	–	–	–	73.8
DIFO	CVPR24	✓	80.2	93.3	92.0	84.5	93.1	91.9	86.7	79.4	93.1	87.5	78.7	93.6	87.8
Co-learn++	IJCV25	✓	80.0	91.2	91.8	83.4	92.7	91.3	83.4	78.9	92.0	85.5	80.6	94.7	87.1
ProDe	ICLR25	✓	81.4	92.5	91.8	85.5	93.2	91.0	90.2	82.2	93.8	86.1	81.9	94.0	88.6
DUET	NIPS25	✓	80.2	92.6	92.5	85.2	93.0	92.1	85.0	80.3	92.1	86.5	79.7	92.9	87.7
SCALP	–	✓	**82.2**	**96.1**	**94.9**	**90.6**	**96.4**	**94.6**	**90.5**	**82.7**	**95.1**	**90.6**	**82.2**	**95.6**	**91.0**

Table 2: Comparison on DomainNet-126 in SFDA setting, where V denotes whether a frozen CLIP ViT-L/14@336 model is used as the proxy. Full results are provided in Appendix B.1.

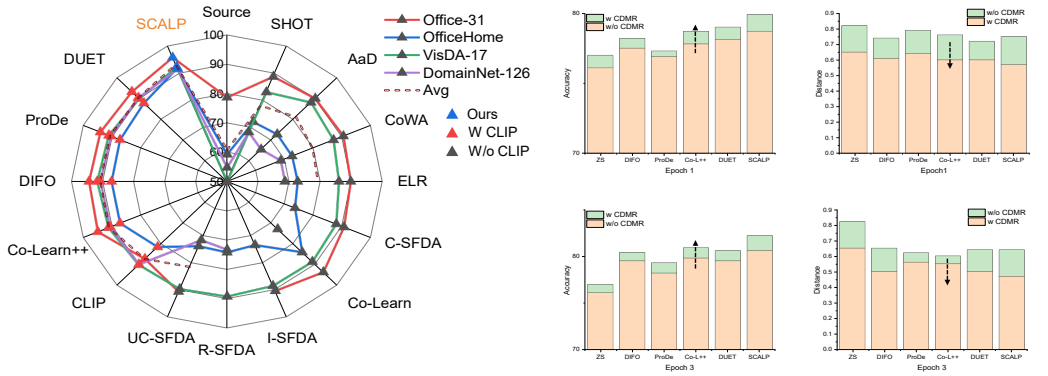
			DomainNet-126												
Method	Venue	V	C→P	C→R	C→S	P→C	P→R	P→S	R→C	R→P	R→S	S→C	S→P	S→R	Avg
Source	ICML20	×	44.6	59.8	47.5	53.3	75.3	46.2	55.3	62.7	46.4	55.1	50.7	59.5	54.7
CLIP-L/14@336	Zero-shot	✓	88.8	93.6	88.3	89.4	93.6	88.3	89.4	88.8	88.3	89.4	88.8	93.6	90.0
SHOT	ICML20	×	63.5	78.2	59.5	67.9	81.3	61.7	67.7	67.6	57.8	70.2	64.0	78.0	68.1
AaD	NIPS22	×	60.8	74.8	55.9	62.2	78.3	58.2	63.1	68.1	55.6	67.1	66.0	75.4	65.5
CoWA	ICML22	×	64.6	80.6	60.6	66.2	79.8	60.8	69.0	67.2	60.0	69.0	65.8	79.9	68.6
Co-learn	ICCV23	×	65.9	82.7	65.3	71.0	84.0	64.9	73.5	71.7	64.6	75.7	70.6	84.5	72.9
R-SFDA	ICCV25	×	–	–	68.0	71.9	83.2	–	77.6	72.9	67.3	–	71.2	–	73.2
UC-SFDA	ICLR25	×	–	–	66.5	69.3	81.0	–	75.2	71.1	64.3	–	68.1	–	71.6
DIFO	CVPR24	✓	89.2	93.6	88.3	90.0	94.3	88.4	90.1	89.2	88.6	89.8	89.4	94.5	90.5
Co-learn++	IJCV25	✓	89.5	93.9	88.6	90.0	93.8	88.7	90.3	89.4	88.5	90.1	89.5	93.9	90.5
ProDe	ICLR25	✓	88.6	94.8	88.1	89.3	94.6	88.3	89.5	88.6	88.4	89.3	89.4	94.9	90.3
DUET	NIPS25	✓	89.1	93.2	88.7	90.0	93.3	89.0	89.9	89.5	89.0	90.0	89.1	93.6	90.4
SCALP	–	✓	**90.5**	**95.3**	**89.1**	**90.3**	**95.1**	**89.4**	**90.7**	**90.4**	**89.2**	**90.5**	**90.2**	**95.2**	**91.4**

Competitors. We compare our method with diverse baselines into four categories. (1) **Baselines.** A ResNet-based source model and CLIP zero-shot. (2) **SFDA with pretrained image encoders.** Representative methods include SHOT (Liang et al., 2020), AaD (Yang et al., 2022), CoWA (Lee et al., 2022), C-SFDA (Karim et al., 2023), Co-Learn (Zhang et al., 2023), ELR (Yi et al., 2023), I-SFDA (Mitsuzumi et al., 2024), R-SFDA (Zhu et al., 2025a), and UC-SFDA (Xu et al., 2025). (3) **SFDA with pretrained CLIP.** We further compare with recent CLIP-based methods, including DIFO (Tang et al., 2024), ProDe (Tang et al., 2025), Co-Learn++ (Zhang et al., 2025), and DUET (Lee et al., 2025), where CLIP serves as the external proxy. (4) **Only CLIP adaptation.** We additionally compare with ADCLIP (Singha et al., 2023), SigCLIP (Zhai et al., 2023), DAMP (Du et al., 2024), and DAPL (Ge et al., 2025), that directly refine CLIP for unsupervised adaptations.

Implementation. Following prior SFDA settings (Liang et al., 2020), we adapt the pretrained source model using only unlabeled target data and a pretrained CLIP proxy. The CLIP text encoder g_t is frozen, with only the prompt tokens v optimized, while the source encoder is updated separately using Adam and SGD, respectively. We evaluate multiple CLIP backbones from ViT-B/16 (Radford et al., 2021) to ViT-L/14@336, with comparisons to ADCLIP (Singha et al., 2023). All experiments are implemented in PyTorch on NVIDIA H100 GPUs and reported with the averaged performance over three independent runs. More details are provided in Appendix C.

4.2 Results and comparisons

Main results. Compared to the best method without CLIP proxy, we obtain gains of 7.5% and 18.2%, respectively on Office-Home (Table 1) and DomainNet-126 (Table 2). Fig 4 reports overall results across four datasets, where our method with the same CLIP backbone achieves the best average accuracy among SFDA methods. Against CLIP-based competitors, we further improve around 1.4% across four benchmarks, where our SCALP achieves the best averaged 92.4%, and



(a) Spider map of overall comparisons on four benchmarks. (b) Early evolution of accuracy and modality gap.

Figure 4: Overall comparison and modality gap analysis. (a) Performance across four SFDA benchmarks. (b) Evolution of accuracy and modality gap in early epochs.

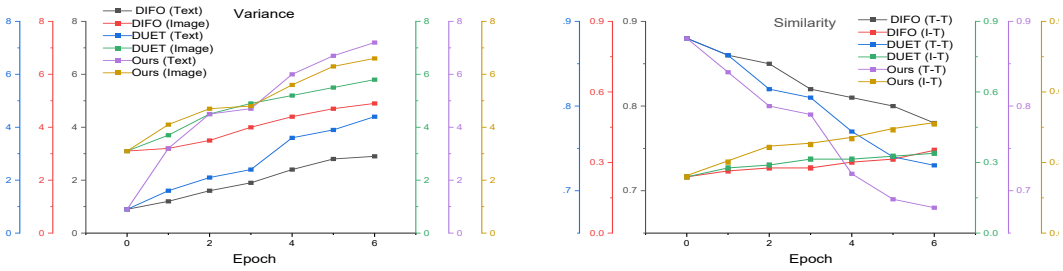
see full results in Appendix B.1, respectively outperforming the second best competitor on each benchmark, by 0.7%, 2.4%, 1.1%, 0.9%. Table 3 further isolates our refined CLIP proxy to compete with unsupervised adaptation of CLIP only.

Pseudo-label quality and modality gap for SFDA. Pseudo-label quality in early epochs is critical for reliable adaptation. Appendix B.9 reports early pseudo-label accuracy on dual-agreed samples and agreement ratios on Office-Home Art. DIFO keeps a ratio of 1 without sample selection, while SCALP achieves higher accuracy and agreement ratios than DUET after CDMR and logits calibration. It indicates SCALP's higher pseudo-label quality over recent SFDA methods. Fig 4 (b) further analyzes CDMR, compared without realignment for SFDA, and dynamical gap on Office-Home Clipart domain, where $\text{Dist.} = \frac{1}{K}\sum_{i=1}^{K}(1-\cos\langle z_t^k, \mu_v^k\rangle)$ following An et al. (2025). In Appendix B.3, CDMR reduces the centroid distance by 0.16, while SCALP improves pseudo-label accuracy over 2% in the first three epochs, indicating a more reliable initialization.

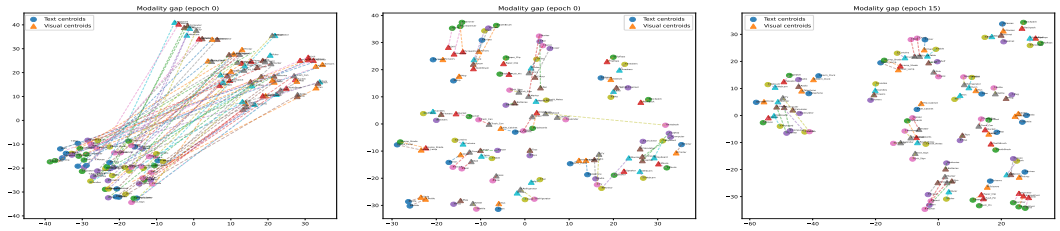
4.3 Ablation Analysis

Analysis of SCALP. (1) **CDMR improves proxy initialization.** In statistics, Fig. 5(a) shows how CDMR closes the gap and enhances class distinctions by larger intra-class variance and lower inter-class textual similarity. Appendix B.3 shows that CDMR improves four recent SFDA methods by averaged 1.0% while reducing the cross-modal distance by approximated 0.13. Furthermore, the confidence filter in Eq. 1 contributes around 0.5% gain by restricting realignment to confident samples, providing a distinct task-specific initialization toward adaptation. (2) **ASPL enables complementary learning of reliable semantics and noise constraints.** In addition to CDMR, ASPL introduces selective supervision and neighbor consistency, yielding an extra 1.5% improvement. Disabling neighbor consistency or replacing one-hot supervision with consistency-only training causes clear performance drops, confirming the complementary roles of the selective optimization. In addition, we also evaluate over multiple ViT backbone in Table 4 and more details in Appendix B.5. (3) **Additional components further contribute adaptation.** We further employ logits calibration to suppress unreliable source-biased predictions before ASPL selection, which actually increases the ratio and quality of dual-agreed samples during early training in Appendix B.9. Meanwhile, the consistently positive gradient cosine similarity indicates compatible optimization between supervised and neighbor objectives. Detailed component-wise, filtering, calibration, and overall analyses are provided in Table 4 and the Appendix B respectively. (4) **Further validations on implementations.** We additionally examine prompt customization and different fine-tuning strategies for ASPL, indicating our method remains relatively template-agonist. Detailed component-wise and design analyses are provided in Table 4 and Appendix B.

Hyperparameter Evaluation. SCALP involves four key hyperparameters: the number of neighbors N, the selective training trade-off λ_v, calibration weight α and temperature τ applied to CLIP



(a) Class-centroid variance (↑). (b) Different textual class similarity (↓).



(c) Gap of CLIP. (d) Gap after CDMR. (e) Gap after SCALP.

Figure 5: (a)-(b): Class-centroids variance and cross-modal similarity. (c)-(e): Modality gaps by different methods, full images are in Appendix B.8.

Table 3: Comparison with only proxy adaptation. See Appendix B.9 for details.

Proxy	Backbone	SF	O31	OH	Vis17	DN126
CLIP	B/16	✓	76.9	82.4	88.9	82.3
CLIP+CDMR	B/16	✓	77.4	83.2	89.2	83.2
SigCLIP	B/16	✓	86.0	84.5	89.1	84.6
SigCLIP+CDMR	B/16	✓	87.3	85.0	90.1	85.0
Method	Backbone	SF	O31	OH	Vis17	DN126
DAPL	B/16	×	88.9	84.0	89.5	85.8
ADCLIP	B/16	×	–	86.1	90.7	86.9
DAMP	B/16	×	–	87.0	90.9	87.6
SCALP	B/16	✓	93.5	**87.0**	**91.6**	**87.7**

Table 4: Ablation of SCALP across CLIP backbones. See Appendix B.4 for more ablation details and Appendix B.5 for multiple CLIP backbones.

Method	CDMR	Dual.	Neigh.	B/16	B/32	L/14	336
Zero-shot	×	×	×	82.6	78.7	87.6	87.3
Post-hoc	×	×	×	82.8	78.4	87.4	87.0
CDMR w/o Filter	✓	×	×	83.0	78.7	87.9	87.0
CDMR full	✓	×	×	83.9	81.0	88.9	89.1
CDMR + Dual.	✓	✓	×	86.2	84.6	90.8	90.4
CDMR + Neigh.	✓	×	✓	85.0	83.9	89.1	88.8
ASPL w/o CDMR	×	✓	✓	89.0	88.5	91.0	91.1
w/o calibration	✓	✓	✓	88.4	87.7	91.4	91.7
SCALP full	✓	✓	✓	**89.3**	**88.9**	**91.9**	**92.4**

proxy logits. Empirically, the performance is relatively insensitive to these hyperparameters within reasonable ranges in Appendix B.12, specifically $N \in [3, 11]$, $\lambda_v \in [0.1, 2]$, and $\tau \in [0.6, 1.4]$, where the results imply insensitive effects across a broad range.

Visualization of cross-modal semantics from images. Beyond the quantitative distance in Fig. 4 (b), we further provide heatmap visualizations as complementary diagnostic evidence. As denoted in Appendix B.11, CDMR tends to present more localized responses on object-oriented regions, whereas the pretrained CLIP proxy only shows less activation on the region related to each class. This observation further indicates that CDMR mitigates the image-text gap by synchronizing with class-level semantics, thereby enhancing the task-specific distinction of CLIP proxy toward SFDA.

5 Conclusion

We facilitate SFDA by mitigating the modality gap of CLIP proxies before adaptation while reducing noisy pseudo-labels during adaptation. Specifically, we propose a unified framework SCALP, which integrates distributional realignment and selective pseudo-labeling to close the proxy gap while enhance task-specific distinctions. Extensive experiments across four benchmarks validate the effectiveness of our approach, offering new insights for future CLIP-based SFDA development.

AI USE STATEMENT

In this work, we used generative AI tools for language refinement, coding and mathematical calibration. We have not used generative AI tools for methodological discussion, idea generation, mathematical derivation, experimental design feedback, and interpretation of experimental results, and using AI for survey and reference is not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, personally identifiable information, sensitive data, or high-risk applications. All experiments are conducted on publicly available benchmark datasets and standard pretrained models. We are not aware of any specific ethical concerns arising from the proposed method or its experimental evaluation.

REPRODUCIBILITY STATEMENT

Implementation details, training settings, and experimental protocols are provided in the main paper and Appendix. The source code has been partially released at `https://anonymous.4open.science/r/5872-SCALP-D081`.

REFERENCES

Na Min An, Eunki Kim, James Thorne, and Hyunjung Shim. I0T: Embedding standardization method towards zero modality gap. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27182–27199, Vienna, Austria, July 2025. Association for Computational Linguistics.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1–2):151–175, 2010.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159, 2021.

Rhea Chowers, Oshri Naparstek, Udi Barzelay, and Yair Weiss. Is the modality gap a bug or a feature? a robustness perspective, 2026.

Ning Ding, Yixing Xu, Yehui Tang, Chao Xu, Yunhe Wang, and Dacheng Tao. Source-free domain adaptation via distribution estimation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7202–7212, 2022.

Zhekai Du, Xinyao Li, Fengling Li, Ke Lu, Lei Zhu, and Jingjing Li. Domain-agnostic mutual prompting for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23375–23384, June 2024.

Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Improving cross-modal alignment in CLIP. In *The Thirteenth International Conference on Learning Representations*, 2025.

Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1):1160–1170, 2025.

Eleonora Grassucci, Giordano Cicchetti, Emanuele Frasca, Aurelio Uncini, and Danilo Comminiello. Closing the modality gap aligns group-wise semantics. In *The Fourteenth International Conference on Learning Representations*, 2026.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Xuefeng Hu, Ke Zhang, Lu Xia, Albert Chen, Jiajia Luo, Yuyin Sun, Ken Wang, Nan Qiao, Xiao Zeng, Min Sun, et al. Reclip: Refine contrastive language image pre-training with source free domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2994–3003, 2024.

Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 3635–3649, 2021.

Nazmul Karim, Niluthpol Chowdhury Mithun, Abhinav Rajvanshi, Han-pang Chiu, Supun Samarasekera, and Nazanin Rahnavard. C-sfda: A curriculum learning aided self-training framework for efficient source free domain adaptation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24120–24131, 2023.

Youngeun Kim, Donghyeon Cho, Kyeongtak Han, Priyadarshini Panda, and Sungeun Hong. Domain adaptation without source data. *IEEE Transactions on Artificial Intelligence*, 2(6):508–518, 2021.

Jae Yun Lee, Jae Hyeon Park, Gyoomin Lee, Bogyeong Kim, Min Hee Cha, Hyeok Nam, Joo Hyeon Jeon, Hyunse Lee, and Sung In Cho. DUET: Dual-perspective pseudo labeling and uncertainty-aware exploration & exploitation training for source-free domain adaptation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Jonghyun Lee, Dahuin Jung, Junho Yim, and Sungroh Yoon. Confidence score for source-free unsupervised domain adaptation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 12365–12377. PMLR, 2022.

Jiao Li, Houcheng Su, Bingli Wang, Yuandong Min, Mengzhu Wang, Nan Yin, Shanshan Wang, and Jingcai Guo. Esbn: Estimation shift of batch normalization for source-free universal domain adaptation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 1386–1394. International Joint Conferences on Artificial Intelligence Organization, 2025.

Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 6028–6039. PMLR, 2020.

Jian Liang, Dapeng Hu, Yunbo Wang, Ran He, and Jiashi Feng. Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8602–8617, 2022a.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022b.

Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramanan. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19325–19337, 2023.

Yu Mitsuzumi, Akisato Kimura, and Hisashi Kashima. Understanding and improving source-free domain adaptation from a theoretical perspective. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28515–28524, 2024.

Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge, 2017.

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, pp. 8748–8763. PMLR, 18–24 Jul 2021.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Computer Vision – ECCV 2010*, pp. 213–226. Springer Berlin Heidelberg, 2010.

Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *2019 IEEE/CVF International Conference on Computer Vision*, pp. 8049–8057, 2019.

Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Mainak Singha, Harsh Pal, Ankit Jha, and Biplab Banerjee. Ad-clip: Adapting domains in prompt space using clip. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 4357–4366, 2023.

Song Tang, Wenxin Su, Mao Ye, and Xiatian Zhu. Source-free domain adaptation with frozen multimodal foundation model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23711–23720, 2024.

Song Tang, Wenxin Su, Yan Gan, Mao Ye, Jianwei Dr. Zhang, and Xiatian Zhu. Proxy denoising for source-free domain adaptation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jiayi Tian, Jing Zhang, Wen Li, and Dong Xu. Vdm-da: Virtual domain modeling for source data-free domain adaptation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(6): 3749–3760, 2022.

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*, 2017.

Siying Xiao, Mao Ye, Qichen He, Shuaifeng Li, Song Tang, and Xiatian Zhu. Adversarial experts model for black-box domain adaptation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8982–8991. Association for Computing Machinery, 2024.

Gezheng Xu, Hui Guo, Li Yi, Charles Ling, Boyu Wang, and Grace Yi. Revisiting source-free domain adaptation: a new perspective via uncertainty control. In *The Thirteenth International Conference on Learning Representations*, 2025.

Shiqi Yang, yaxing wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29393–29405, 2021.

Shiqi Yang, yaxing wang, kai wang, Shangling Jui, and Joost van de Weijer. Attracting and dispersing: A simple approach for source-free domain adaptation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 5802–5815. Curran Associates, Inc., 2022.

Li Yi, Gezheng Xu, Pengcheng Xu, Jiaqi Li, Ruizhi Pu, Charles Ling, Ian McLeod, and Boyu Wang. When source-free domain adaptation meets learning with noisy labels. In *The Eleventh International Conference on Learning Representations*, 2023.

Wenzhe Yin, Zehao Xiao, Pan Zhou, Shujian Yu, Jiayi Shen, Jan-Jakob Sonke, and Stratis Gavves. Distributional vision-language alignment by cauchy-schwarz divergence. In *The Fourteenth International Conference on Learning Representations*, 2026.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952, 2023.

Wenyu Zhang, Li Shen, and Chuan-Sheng Foo. Rethinking the role of pre-trained networks in source-free domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18841–18851, 2023.

Wenyu Zhang, Li Shen, and Chuan-Sheng Foo. Source-free domain adaptation guided by vision and vision-language pre-training. *International Journal of Computer Vision*, 133(2):844–866, 2025.

Yuhui Zhang, Elaine Sui, and Serena Yeung. Connect, collapse, corrupt: Learning cross-modal tasks with uni-modal data. In *The Twelfth International Conference on Learning Representations*, 2024.

Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In S. Thrun, L. Saul, and B. Schölkopf (eds.), *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16795–16804, 2022.

Ronghang Zhu, Mengxuan Hu, Weiming Zhuang, Lingjuan Lyu, Xiang Yu, and Sheng Li. Revisiting source-free domain adaptation: Insights into representativeness, generalization, and variety. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25688–25697, 06 2025a.

Xingyu Zhu, Beier Zhu, Shuo Wang, Kesen Zhao, and Hanwang Zhang. Enhancing CLIP robustness via cross-modality alignment. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

APPENDIX

A SKETCH DERIVATION OF REALIGNMENT

For clarity, we omit the class index k. Let the aligned textual embedding be defined as

$$\dot{z}_t = \mathbf{A} z_t + \boldsymbol{\beta}. \tag{13}$$

We assume that the ideal aligned textual embedding still satisfies a reduced residual modality gap:

$$\dot{z}_t - \mu_v = \mathbf{c}' + \epsilon, \qquad \text{with } \epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}), \tag{14}$$

where μ_v denotes the estimated class-centroids based on $\{z_v^i | y = k\}$. $\mathbf{c}'$ is the remaining centroid estimation residual as $\mathbf{c}' = \mu_v^\star - \mu_v$, where $\mu_v^\star = \mathbb{E}\left[z_v^i | y^\star = k\right]$ and σ is small. Then, we take expectation of Eq. 14 and equation holds,

$$\mathbb{E}[\dot{z}_t] - \mathbb{E}[z_v] = \mathbf{c}', \quad \mathbb{E}[z_t] = \mu_t. \tag{15}$$

Thus we have

$$\mathbf{A}\mu_t + \boldsymbol{\beta} = \mu_v + \mathbf{c}' \rightarrow \boldsymbol{\beta} = \mu_v + \mathbf{c}' - \mathbf{A}\mu_t. \tag{16}$$

Substituting it back yields

$$\dot{z}_t = \mathbf{A}(z_t - \mu_t) + \mu_v + \mathbf{c}'. \tag{17}$$

Next, we match the variance. Since $\mathbf{c}'$ is deterministic and noise variable is independent from image class-centroids μ_v. Therefore,

$$\mathrm{Var}[\dot{z}_t] = \mathrm{Var}[\mu_v + \epsilon] = \Sigma_v + \sigma^2 \mathbf{I}. \tag{18}$$

On the other side,

$$\mathrm{Var}[\dot{z}_t] = \mathbf{A}\Sigma_t \mathbf{A}^\top = \Sigma_v + \sigma^2 \mathbf{I} \approx \Sigma_v. \tag{19}$$

Under the diagonal covariance assumption, $\Sigma_t = \mathrm{diag}(\sigma_t)$ and $\Sigma_v = \mathrm{diag}(\sigma_v)$, we obtain the element-wise approximated solution due to a small σ,

$$\mathbf{A} \approx \mathrm{diag}\left(\sqrt{\frac{\sigma_v}{\sigma_t + \varepsilon}}\right), \tag{20}$$

where $\varepsilon > 0$ is used for numerical stability.

Since the residual centroid bias $\mathbf{c}'$ is not explicitly estimated, we absorb them into the residual term and reformulate the practical form

$$\dot{z}_t \approx \mathrm{diag}\left(\sqrt{\frac{\sigma_v}{\sigma_t + \varepsilon}}\right)(z_t - \mu_t) + \mu_v + \mathbf{c}'. \tag{21}$$

Thus, the affine transformation aligns textual embeddings to visual class statistics by matching both mean and diagonal variance. Further, we treat the residual bias $\mathbf{c}'$ as a learnable component that can be absorbed and optimized through parameterized textual representations, then reformulate the solution as:

$$\dot{z}_t \approx \mathbf{A}(z_t + \frac{\mathbf{c}'}{\mathbf{A}} - \mu_t) + \mu_v \rightarrow \mathbf{A}(g(\mathcal{T}, v) - \mu_t) + \mu_v, \tag{22}$$

where $g(\cdot, v)$ denotes the text embedding by learnable prompt tokens v, which compensates for the approximation residual that should be refined with instance-level fine-tune, as similar as Eq. 4 in the main text.

B ADDITIONAL EXPERIMENTAL RESULTS

B.1 FULL RESULTS ON FOUR SFDA BENCHMARKS.

We present full results on SFDA benchmarks: Office-Home, VisDA-2017, and DomainNet-126. Our method achieves consistent improvements over prior methods across all datasets, generally larger gains on more challenging scenarios. Detailed comparisons are shown in Tables 6–8. In short, we summarize overall results in Table 9.

Table 5: Comparison on SFDA Office-31, where V denotes whether a frozen CLIP ViT-L/14@336 model is used as the proxy.

Method	Venue	V	A→D	A→W	D→W	D→A	W→A	W→D	Avg
Source	ICML20	×	81.9	78.0	93.6	59.4	61.4	98.8	78.9
CLIP-L/14@336	Zero-shot	✓	87.8	89.2	89.2	85.5	85.5	87.8	87.5
SHOT	ICML20	×	93.7	91.1	98.2	74.2	74.6	**100.0**	88.6
AaD	NIPS22	×	96.4	92.1	**99.1**	75.0	76.5	**100.0**	89.9
CoWA	ICML22	×	94.4	95.2	98.5	76.2	77.6	99.8	90.3
ELR	ICLR23	×	93.8	93.3	98.0	76.2	76.9	100.0	89.7
Co-learn	ICCV23	×	97.4	**98.0**	**99.1**	84.5	82.2	**100.0**	93.5
C-SFDA	CVPR23	×	96.2	93.9	98.8	77.3	77.9	99.7	90.6
I-SFDA	CVPR24	×	95.3	94.2	98.3	76.4	77.5	99.9	90.3
UC-SFDA	ICLR25	×	94.8	95.4	98.9	77.1	77.1	100.0	90.6
DIFO	CVPR24	✓	98.0	97.7	98.2	86.7	87.0	98.2	94.3
ProDe	ICLR25	✓	96.6	97.0	97.9	86.9	87.3	98.4	94.0
DUET	NIPS25	✓	98.4	96.0	96.6	85.0	85.6	99.0	93.4
SCALP	–	✓	**99.0**	97.2	98.5	**88.5**	**88.4**	98.2	**95.0**

Table 6: Comparison on Office-Home in SFDA setting. V refers to using different CLIPs as proxy supervision. L means CLIP ViT-L/14 and XL means L/14@336.

Method	Venue	V	A→C	A→P	A→R	C→A	C→P	C→R	P→A	P→C	P→R	R→A	R→C	R→P	Avg
Source	ICML20	×	43.7	67.0	73.9	49.9	60.1	62.5	51.7	40.9	72.6	64.2	46.3	78.1	59.2
CLIP-L/14	Zero-shot	✓	76.1	93.5	92.3	83.8	93.5	92.3	83.8	76.1	92.3	83.8	76.1	93.5	86.4
CLIP-L/14@336	Zero-shot	✓	72.6	86.5	85.2	81.8	86.5	85.2	81.8	72.6	85.2	81.8	72.6	86.5	81.5
SHOT	ICML20	×	56.7	77.9	80.6	68.0	78.0	79.4	67.9	54.5	82.3	74.2	58.6	84.5	71.9
AaD	NIPS22	×	59.3	79.3	82.1	68.9	79.8	79.5	67.2	57.4	83.1	72.1	58.5	85.4	72.7
CoWA	ICML22	×	56.9	78.4	81.0	69.1	80.0	79.9	67.7	57.2	82.4	72.8	60.5	84.5	72.5
ELR	ICLR23	×	58.4	78.7	81.5	69.2	79.5	79.3	66.3	58.0	82.6	73.4	59.8	85.1	72.6
C-SFDA	CVPR23	×	60.3	80.2	82.9	69.3	80.1	78.8	67.3	58.1	83.4	73.6	61.3	86.3	73.5
Co-learn	ICCV23	×	69.6	89.5	91.2	82.7	88.4	91.3	82.5	68.5	91.5	82.8	71.3	92.1	83.5
I-SFDA	CVPR24	×	60.7	78.9	82.0	69.9	79.5	79.7	67.1	58.8	82.3	74.2	61.3	86.4	73.4
R-SFDA	CVPR25	×	61.2	80.9	82.7	69.3	81.2	81.4	68.1	58.8	83.4	74.6	62.4	85.7	74.1
UC-SFDA	ICLR25	×	–	–	–	–	–	–	–	–	–	–	–	–	73.8
Co-learn++(L)	IJCV25	✓	81.2	93.6	92.3	88.3	94.5	92.8	88.7	80.5	92.4	88.7	81.0	93.9	89.0
Co-learn++(XL)	IJCV25	✓	80.0	91.2	91.8	83.4	92.7	91.3	83.4	78.9	92.0	85.5	80.6	94.7	87.1
DIFO(L)	CVPR24	✓	80.2	94.0	91.7	88.4	93.7	91.6	88.9	80.3	93.0	90.0	80.4	93.8	88.8
DIFO(XL)	CVPR24	✓	80.2	93.3	92.0	84.5	93.1	91.9	86.7	79.4	93.1	87.5	78.7	93.6	87.8
ProDe(L)	ICLR25	✓	81.4	92.5	91.8	85.5	93.2	91.0	90.2	82.2	93.8	86.1	81.9	94.0	88.6
ProDe(XL)	ICLR25	✓	81.4	92.5	91.8	85.5	93.2	91.0	90.2	82.2	93.8	86.1	81.9	94.0	88.6
DUET(L)	NIPS25	✓	79.2	92.3	91.0	85.2	91.5	91.3	84.7	79.2	91.7	86.0	78.6	92.3	86.9
DUET(XL)	NIPS25	✓	80.2	92.6	92.5	85.2	93.0	92.1	85.0	80.3	92.1	86.5	79.7	92.9	87.7
SCALP(L)	–	✓	**82.4**	**96.3**	**95.3**	90.0	96.1	**95.2**	90.2	81.8	94.9	90.5	82.0	**96.1**	90.9
SCALP(XL)	–	✓	82.2	96.1	94.9	**90.6**	**96.4**	94.6	**90.5**	**82.7**	**95.1**	**90.6**	**82.2**	95.6	**91.0**

B.2 Overall Analysis of SCALP.

We perform a systematic ablation study to evaluate the individual contribution of each component in Table 4. The introduction of **ASPL** results in an averaged gain of +1.0%, suggesting its dominant role among others. Building upon this, **CDMR** further improves the performance by +0.7%, highlighting its complementary in CLIP proxy to be more task-specific distinctions and better adaptation results. When three components engage at the same time, the average 91.9% across four datasets with CLIP L/14, 1.1% and 1.4% higher than the second highest ProDe and the third highest Co-Learn++ in Table 9, respectively.

B.3 Relationship between CDMR, Modality Gap and Early Adaptation.

Following An et al. (2025), we quantify the modality gap by cosine similarity of two embeddings, a lower value indicates better alignment between textual class embeddings and visual class centroids. As shown in Table 10, CDMR consistently reduces the modality gap across different SFDA methods, while simultaneously improving pseudo-label accuracy during the early adaptation stage. For SCALP, the centroid distance decreases from $0.75/0.69/0.64$ to $0.57/0.50/0.47$ over the first three epochs, accompanied by pseudo-label accuracy gains around 1.2%, 1.5%, and 1.6%, respectively. This consistent trend suggests a close relationship among modality realignment, proxy discriminability, and early adaptation: by narrowing the image–text gap, CDMR provides a more task-

Table 7: Comparison on DomainNet-126 in SFDA setting. V refers to using different CLIPs as proxy supervision. L means CLIP ViT-L/14 and XL means L/14@336.

Method	Venue	V	C→P	C→R	C→S	P→C	P→R	P→S	R→C	R→P	R→S	S→C	S→P	S→R	Avg
Source	ICML20	×	44.6	59.8	47.5	53.3	75.3	46.2	55.3	62.7	46.4	55.1	50.7	59.5	54.7
CLIP-L/14	Zero-shot	✓	87.3	93.8	86.6	87.8	93.8	86.6	87.8	87.3	86.6	87.8	87.3	93.8	88.9
CLIP-L/14@336	Zero-shot	✓	88.8	93.6	88.3	89.4	93.6	88.3	89.4	88.8	88.3	89.4	88.8	93.6	90.0
SHOT	ICML20	×	63.5	78.2	59.5	67.9	81.3	61.7	67.7	67.6	57.8	70.2	64.0	78.0	68.1
AaD	NIPS22	×	60.8	74.8	55.9	62.2	78.3	58.2	63.1	68.1	55.6	67.1	66.0	75.4	65.5
CoWA	ICML22	×	64.6	80.6	60.6	66.2	79.8	60.8	69.0	67.2	60.0	69.0	65.8	79.9	68.6
Co-learn	ICCV23	×	65.9	82.7	65.3	71.0	84.0	64.9	73.5	71.7	64.6	75.7	70.6	84.5	72.9
R-SFDA	ICCV25	×	–	–	68.0	71.9	83.2	–	77.6	72.9	67.3	–	71.2	–	73.2
UC-SFDA	ICLR25	×	–	–	66.5	69.3	81.0	–	75.2	71.1	64.3	–	68.1	–	71.6
DIFO(L)	CVPR24	✓	88.7	94.1	88.0	89.2	93.6	87.2	89.4	88.7	88.1	88.9	88.9	94.2	89.9
DIFO(XL)	CVPR24	✓	89.2	93.6	88.3	90.0	94.3	88.4	90.1	89.2	88.6	89.8	89.4	94.5	90.5
Co-learn++(L)	IJCV25	✓	89.6	93.5	87.2	89.5	93.0	87.6	89.7	89.0	87.9	89.5	88.5	94.3	89.9
Co-learn++(XL)	IJCV25	✓	89.5	93.9	88.6	90.0	93.8	88.7	90.3	89.4	88.5	90.1	89.5	93.9	90.5
ProDe(L)	ICLR25	✓	88.1	94.7	87.4	88.7	94.2	87.5	89.1	88.2	87.7	89.2	89.0	94.7	89.9
ProDe(XL)	ICLR25	✓	88.6	94.8	88.1	89.3	94.6	88.3	89.5	88.6	88.4	89.3	89.4	94.9	90.3
DUET(L)	NIPS25	✓	88.0	92.0	87.1	88.7	92.5	87.5	88.5	89.0	87.5	88.9	88.8	92.5	89.2
DUET(XL)	NIPS25	✓	89.1	93.2	88.7	90.0	93.3	89.0	89.9	89.5	89.0	90.0	89.1	93.6	90.4
SCALP(L)	–	✓	89.7	95.0	88.3	89.7	**95.4**	88.5	90.0	89.4	88.5	89.8	89.7	**95.3**	90.8
SCALP(XL)	–	✓	**90.5**	**95.3**	**89.1**	**90.3**	95.1	**89.4**	**90.7**	**90.4**	**89.2**	**90.5**	**90.2**	95.2	**91.4**

Table 8: Comparison on VisDA-2017 in SFDA setting. V refers to using different CLIPs as proxy supervision. L means CLIP ViT-L/14 and XL means L/14@336.

Method	Venue	V	plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	Avg
Source	ICML20	×	51.5	15.3	43.4	75.4	71.2	6.8	85.5	18.8	49.4	46.4	82.1	5.4	45.9
CLIP-L/14	Zero-shot	✓	99.6	91.4	92.6	69.3	99.5	92.2	97.1	79.8	83.9	99.2	97.0	71.6	89.4
CLIP-L/14@336	Zero-shot	✓	99.6	92.1	92.7	70.5	99.4	92.4	97.4	79.4	85.2	98.9	97.1	71.6	89.7
SHOT	ICML20	×	95.3	87.1	79.1	55.1	93.2	95.5	79.5	79.6	91.6	89.5	87.9	56.0	82.4
AaD	NIPS22	×	96.9	90.2	85.7	82.8	97.4	96.0	89.7	83.2	96.8	94.4	90.8	49.0	87.7
CoWA	ICML22	×	96.2	89.7	83.9	73.8	96.4	97.4	89.3	86.8	94.6	92.1	88.7	53.8	86.9
ELR	ICLR23	×	97.1	89.7	82.7	62.0	96.2	97.0	87.6	81.2	93.7	94.1	90.2	58.6	85.8
Co-learn	ICCV23	×	98.9	93.2	81.0	83.0	98.6	98.8	95.7	84.8	94.8	97.3	95.1	41.6	88.6
C-SFDA	CVPR23	×	97.6	88.8	86.1	72.2	97.2	94.4	92.1	84.7	93.0	90.7	93.1	63.5	87.8
I-SFDA	CVPR24	×	97.5	91.4	87.9	79.4	97.2	97.2	92.2	83.0	96.4	94.2	91.1	53.0	88.4
R-SFDA	ICCV25	×	94.6	86.4	85.4	96.8	96.7	92.2	96.1	82.6	88.2	88.4	89.8	72.4	89.1
UC-SFDA	ICLR25	×	98.4	90.7	88.6	80.7	97.9	96.9	93.1	83.8	97.6	95.9	92.6	59.1	89.6
Co-learn++(L)	IJCV25	✓	98.7	91.0	90.0	79.4	97.6	97.4	92.0	87.2	92.3	95.6	93.1	66.8	90.3
Co-learn++(XL)	IJCV25	✓	**99.6**	94.6	90.9	77.8	**99.6**	99.0	96.4	80.1	90.0	**99.2**	96.3	70.1	91.1
DIFO(L)	CVPR24	✓	98.3	91.6	90.8	81.7	97.9	98.3	92.4	87.5	92.1	95.8	93.6	68.4	90.7
DIFO(XL)	CVPR24	✓	99.0	92.4	**91.6**	83.2	98.7	99.0	93.5	88.8	93.4	95.7	94.3	70.7	91.5
ProDe(L)	ICLR25	✓	97.9	90.5	86.7	88.5	97.8	96.4	94.2	88.4	95.7	96.2	93.3	71.5	91.4
ProDe(XL)	ICLR25	✓	97.5	90.8	87.2	88.6	97.7	96.5	94.3	88.6	94.2	93.1	95.0	68.0	91.1
DUET(L)	NIPS25	✓	97.5	90.2	88.6	79.8	97.2	97.8	91.0	86.3	90.8	94.2	90.7	66.3	89.2
DUET(XL)	NIPS25	✓	97.9	91.3	89.7	82.1	97.8	96.0	92.4	87.6	91.8	95.1	85.3	70.6	89.8
SCALP(L)	–	✓	99.2	94.6	91.0	75.2	99.2	98.7	**96.4**	82.5	90.6	99.0	**97.3**	**73.2**	91.4
SCALP(XL)	–	✓	**99.6**	**96.1**	89.2	**89.1**	98.0	**99.6**	95.2	**90.9**	**95.7**	95.6	95.8	71.8	**92.2**

aligned proxy initialization, which yields more reliable initialization and facilitates subsequent proxy supervision for SFDA.

B.4 Quality of Confidence-Filter $\mathcal{S}_k$.

We further examine whether the confidence-filtered visual sets $\mathcal{S}_k$ provide reliable and sufficiently balanced statistics for class-centroid construction. As shown in Table 11, our class-wise mean threshold achieves the best trade-off between pseudo-label quality and class coverage: it substantially improves the overall accuracy over using the full target set while preserving all 65 classes. In contrast, fixed confidence thresholds provide only marginal accuracy gains at high values of γ, but discard samples from several classes, resulting in incomplete class coverage and potentially enlarging the residual semantic gap. Therefore, our adaptive class-wise selection not only maintains reliable samples within each $\mathcal{S}_k$, but also avoids the class imbalance induced by a globally fixed threshold. Table 11 further analyzes the relationship between the quality of $\mathcal{S}_k$ and the resulting class-wise accuracy and class coverages.

Table 9: Overall accuracy (%) on four benchmarks under SFDA settings. Methods are identified by CLIP usage. Pretrained backbones (PB) like ResNet (RN), SwinTransformer (Swin)ViT-L/14 (L/14) and ViT-L/14@336 (336) are adopted.

Method	Venue	CLIP	PB	O31	OH	Vis17	DN126	Avg
Source	ICML20	×	RN	78.9	59.2	45.9	54.7	59.7
SHOT	ICML20	×	RN	88.6	71.9	82.4	68.1	77.8
AaD	NeurIPS22	×	RN	89.9	72.7	87.7	65.5	79.0
CoWA	ICML22	×	RN	90.3	72.5	86.9	68.6	79.6
ELR	ICLR23	×	RN	89.7	72.6	85.8	–	–
C-SFDA	CVPR23	×	RN	90.6	73.5	87.8	–	–
Co-Learn	ICCV23	×	Swin	93.6	84.0	88.6	72.9	84.8
I-SFDA	CVPR24	×	RN	90.3	73.4	88.4	–	–
R-SFDA	CVPR25	×	RN	–	74.1	89.1	73.2	–
UC-SFDA	ICLR25	×	RN	90.6	73.8	89.6	71.6	81.4
CLIP	ZS	✓	L/14	86.4	86.4	89.4	88.9	87.8
CLIP	ZS	✓	336	87.5	81.5	89.7	90.0	87.2
Co-Learn++	IJCV25	✓	L/14	93.7	89.0	90.3	89.9	90.5
Co-Learn++	IJCV25	✓	336	94.8	87.1	91.1	90.5	90.9
DIFO	CVPR24	✓	L/14	92.5	88.8	90.7	89.9	90.5
DIFO	CVPR24	✓	336	94.3	87.8	91.5	90.5	91.0
ProDe	ICLR25	✓	L/14	93.1	88.6	91.4	89.9	90.8
ProDe	ICLR25	✓	336	94.0	88.6	91.1	90.3	91.0
DUET	NeurIPS25	✓	L/14	93.0	86.9	89.2	89.2	89.6
DUET	NeurIPS25	✓	336	93.4	87.7	89.8	90.4	90.3
SCALP	–	✓	L/14	94.2	90.9	91.4	90.7	91.8
SCALP	–	✓	336	**95.0**	**91.0**	**92.2**	**91.4**	**92.4**

Table 10: Early pseudo-label accuracy and modality gap on Clipart. Each entry is training-free post-hoc / CDMR. Lower centroid distance (Dist.) indicates closer modality gap as epoch (Ep.).

Ep.	Metric	ZS	DIFO	ProDe	Co-L++	DUET	SCALP
1	Acc. ↑	76.1/77.0	77.5/78.2	76.9/77.3	77.8/78.7	78.1/79.0	**78.7/79.9**
	Dist. ↓	0.82/0.65	0.74/0.61	0.79/0.64	0.76/0.60	0.72/0.60	**0.75/0.57**
2	Acc. ↑	76.1/77.0	78.3/79.5	77.5/78.9	78.6/80.1	79.2/80.9	**80.2/81.7**
	Dist. ↓	0.82/0.65	0.70/0.55	0.69/0.58	0.65/0.58	0.64/0.53	**0.69/0.50**
3	Acc. ↑	76.1/77.0	79.5/80.4	78.2/79.3	79.8/80.9	79.5/80.6	**80.6/82.2**
	Dist. ↓	0.82/0.65	0.65/0.50	0.62/0.56	0.60/0.55	0.64/0.50	**0.64/0.47**

Table 11: Comparison between our class-wise mean confidence filter and globally fixed confidence thresholds on Office-Home Clipart with CLIP ViT-L/14.

Selection strategy	Accuracy(%)	Classes (#)
Class-wise mean $\mathcal{S}_k$	88.1	65
No filter over $\mathcal{X}_t$	76.1	65
Fixed threshold $\gamma = 0.9$	88.8	63
Fixed threshold $\gamma = 0.8$	88.7	63
Fixed threshold $\gamma = 0.7$	88.4	63
Fixed threshold $\gamma = 0.6$	87.6	64
Fixed threshold $\gamma = 0.5$	86.7	65

Table 12: Ablation study on different fine-tuning strategies on Office-Home. (-1) means append a learnable adapter at the end of transformer bottleneck.

Method	Office-Home
Adapter (-1)	80.3
LoRA	83.2
Prompt tuning	86.0
CoOp	85.8

B.5 Ablation on Different CLIP Backbones.

We evaluate the performance of our method under different CLIP backbones, including ViT-B/16 and ViT-L/14 (with and without higher-resolution fine-tuning). As shown in Table 13, our method consistently achieves competitive or superior performance across all backbones. Notably, stronger backbones lead to improved overall accuracy, while our approach maintains stable gains over baselines under identical settings, demonstrating its robustness and general applicability.

Table 13: Comparison across different CLIP backbones (ViT-B/16 and ViT-B/32) for SFDA.

Method	Backbone	Office-31				Office-Home					VisDA	DomainNet-126				
		→A	→D	→W	Avg.	→Ar	→Cl	→Pr	→Rw	Avg.	Sy→Re	→C	→P	→R	→S	Avg.
DIFO	ViT-B/16	–	–	–	92.2	–	–	–	–	82.5	90.0	–	–	–	–	84.8
ProDe	ViT-B/16	–	–	–	92.2	–	–	–	–	82.9	90.7	–	–	–	–	85.6
DUET	ViT-B/16	80.6	96.0	96.6	91.1	83.2	69.4	87.7	88.0	82.1	89.3	83.1	80.3	86.8	76.5	81.7
SCALP	ViT-B/16	**82.5**	99.0	98.4	**93.3**	**84.2**	**75.5**	**92.5**	**91.8**	**86.0**	**91.0**	**85.0**	**85.8**	**93.5**	**82.9**	**86.8**
DIFO	ViT-B/32	**83.1**	98.0	96.4	92.5	82.3	70.4	90.8	88.8	83.1	90.3	80.4	76.9	87.3	75.3	80.0
ProDe	ViT-B/32	82.8	**98.3**	96.7	**92.6**	82.7	72.6	92.0	90.7	84.5	91.0	85.3	83.2	92.4	79.1	85.0
DUET	ViT-B/32	81.2	96.0	95.5	90.9	83.3	**73.8**	90.8	91.0	84.7	**91.4**	82.7	80.3	89.7	79.0	82.9
SCALP	ViT-B/32	82.5	97.2	**97.9**	92.5	**83.8**	**74.0**	**92.9**	**92.0**	**85.7**	91.2	**85.8**	**84.7**	**93.6**	**81.2**	**86.3**

Table 14: Early-stage pseudo-label quality and agreement-ratio dynamics on Office-Home using CLIP ViT-B/16 and ViT-B/32 within SCALP.

Domain	Epoch			
	Epoch 1	Epoch 2	Epoch 3	Epoch 4
	CLIP ViT-B/32			
A	89.6% (1451/2464)	87.0% (2168/2464)	85.1% (2344/2464)	84.4% (2417/2464)
C	80.2% (3231/4365)	77.1% (3860/4365)	74.0% (4146/4365)	74.2% (4250/4365)
P	95.7% (3877/4484)	94.2% (4048/4484)	94.3% (4176/4484)	93.5% (4250/4484)
R	95.0% (3569/4362)	93.4% (4011/4362)	93.3% (4138/4362)	92.9% (4215/4362)
	CLIP ViT-B/16			
A	90.8% (1924/2464)	87.4% (2270/2464)	85.7% (2367/2464)	85.0% (2401/2464)
C	81.8% (2829/4365)	77.9% (3829/4365)	76.5% (4076/4365)	74.5% (4301/4365)
P	94.6% (3266/4484)	93.2% (3894/4484)	93.3% (4073/4484)	92.8% (4188/4484)
R	93.3% (3518/4362)	92.9% (4222/4362)	92.6% (4285/4362)	92.4% (4299/4362)

We further investigate if smaller CLIP backbones still facilitate SFDA in early adaptation. Results in Table 14 show that SCALP promptly improves pseudo-label quality and dual-agreement sampling ratios from the beginning, yielding more reliable proxy supervision throughout training.

B.6 Ablation on Prompt Customization.

Overall, the performance remains stable across a broad range as Table 16, indicating the insensitivity of initial prompt templates with respect to CLIP L/14@336 as proxy.

B.7 Ablation on Post-Fine-tuning Methods.

To evaluate SCALP under different fine-tuning strategies, we compare prompt learning with LoRA (Hu et al., 2022), CoOp (Zhou et al., 2022) and the post-adapter strategy (Lee et al., 2025). As shown in Table 12, prompt learning achieves the best performance among the three alternatives. This suggests that directly refining textual prompts toward task-specific semantics with respect to class distinctions is particularly effective for improving the CLIP proxy, compared with other methods. More importantly, this result provides complementary evidence that the image–text modality gap is an important factor limiting proxy performance: by learning more detailed semantics with respect to SFDA class vocabulary, prompt tuning partially alleviates the cross-modal semantic mismatch and consequently yields a better proxy for developing SFDA.

B.8 Dynamic Visualization of modality gap.

The evolution of the modality gap across training epochs shows in Fig. 6. We observe that the gap is gradually reduced as training progresses, indicating improved cross-modal alignment. Notably, our method achieves a smaller and more stable modality gap compared to baselines, especially in early stages.

Table 15: Performance on Office-Home across epochs using CLIP ViT-L/14 proxy. Each entry reports accuracy and the dual-agreed sampling ratio.

Method	Domain	Epoch			
		Ep. 1	Ep. 2	Ep. 3	Ep. 4
DUET^-	A	92.4% (1447/2464)	91.0% (1862/2464)	89.3% (2089/2464)	89.1% (2173/2464)
	C	88.5% (1872/4365)	82.6% (3125/4365)	79.5% (3800/4365)	78.5% (3969/4365)
	P	95.0% (2913/4484)	94.3% (3938/4484)	94.1% (4006/4484)	94.0% (4128/4484)
	R	95.7% (3466/4362)	95.1% (3866/4362)	94.4% (3979/4362)	93.2% (3916/4362)
SCALP^-	A	93.4% (1534/2464)	92.9% (1689/2464)	91.7% (2186/2464)	91.9% (2281/2464)
	C	87.9% (2284/4365)	87.4% (2806/4365)	86.5% (3524/4365)	85.3% (3926/4365)
	P	96.2% (3977/4484)	95.2% (4071/4484)	95.2% (4155/4484)	95.3% (4206/4484)
	R	95.2% (3695/4362)	93.7% (4130/4362)	94.2% (4119/4362)	94.8% (4024/4362)
DUET^+	A	92.9% (1765/2464)	91.4% (2029/2464)	89.6% (2357/2464)	90.0% (2353/2464)
	C	86.5% (2059/4365)	85.8% (3052/4365)	83.2% (3718/4365)	82.0% (4006/4365)
	P	95.7% (3771/4484)	95.2% (4101/4484)	95.3% (4124/4484)	95.0% (4112/4484)
	R	95.1% (3466/4362)	94.1% (3845/4362)	94.0% (4068/4362)	94.3% (4050/4362)
SCALP^+	A	93.9% (1858/2464)	92.6% (2111/2464)	92.3% (2208/2464)	92.5% (2394/2464)
	C	88.9% (2422/4365)	88.6% (3141/4365)	87.6% (3863/4365)	86.0% (4074/4365)
	P	96.0% (4028/4484)	95.6% (4141/4484)	95.7% (4416/4484)	95.5% (4368/4484)
	R	95.8% (3749/4362)	95.9% (4211/4362)	95.7% (4197/4362)	95.3% (4203/4362)

$+$: CDMR with calibration; $-$: without calibration.

Table 16: Results of Different Prompt Initialization Templates.

#	Initialization Template	Office-Home	VisDA-17	DN-126
1	`'There is a [CLS].'`	90.4	91.6	90.7
2	`'This is a photo of a [CLS].'`	90.5	91.1	91.0
3	`'This maybe a photo of a [CLS].'`	90.2	**92.2**	90.7
4	`'This is almost a photo of a [CLS].'`	90.5	91.0	91.1
5	`'This is a picture of a [CLS].'`	90.7	92.0	91.2
6	`'A picture of [CLS].'`	**91.0**	91.0	**91.4**
7	`'A photo of a [CLS].'`	90.8	92.0	91.0
8	`'A [CLS].'`	90.0	91.5	91.0
9	`'This is definitely a [CLS].'`	90.5	90.9	90.8

B.9 Pseudo-label Quality and dual-agreed Ratios in Early Adaptation.

We analyze pseudo-label accuracy (on dual-agreed samples) and agreement ratios within early four epochs on Office-Home Art domain (Fig. 7). Specifically, DIFO maintains a constant ratio of 1 due to the absence of selection process, while our SCALP achieves higher accuracy and agreement ratios than DUET, after CDMR and logits calibration. Although the gain in final accuracy is relatively modest, calibration consistently improves the agreement ratio during early epochs, supporting more comprehensive and correcter adaptation in subsequent, as shown in Table 15.

B.10 Analysis of Gradient Conflict in ASPL.

Unlike DUET, where supervised learning and consistency training via augmentation may introduce gradient conflicts during joint optimization, ASPL separates samples according to dual-agreement reliability. Specifically, $\mathcal{L}_{CE}$ is applied to confident dual-agreed samples, while $\mathcal{L}_{KL}$ regularizes uncertain samples through neighborhood consistency. As shown in Fig. 8, the cosine similarity between gradients of $\mathcal{L}_{CE}$ and $\mathcal{L}_{KL}$ remains consistently positive during training, indicating compatible optimization directions without any gradient conflicts.

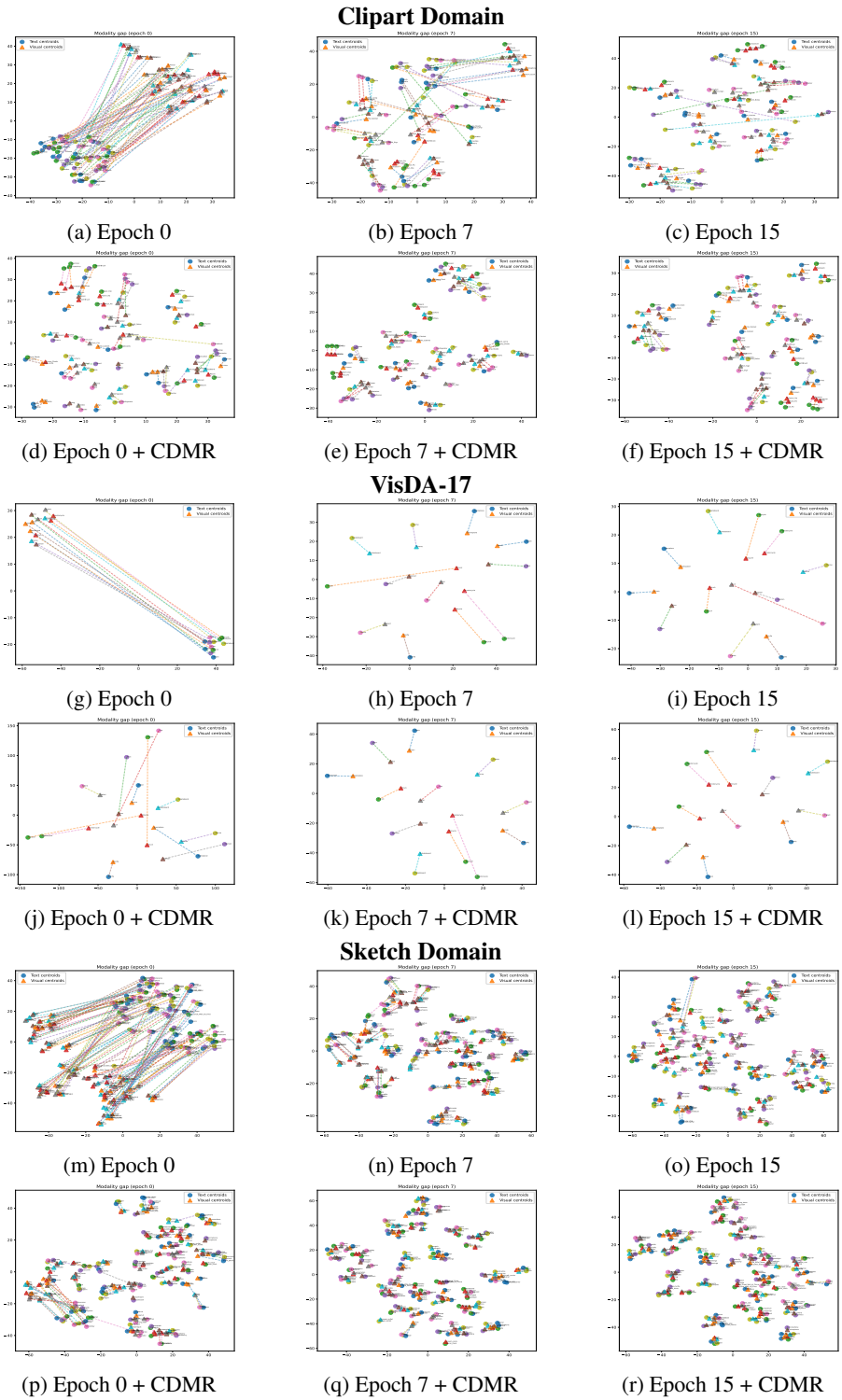
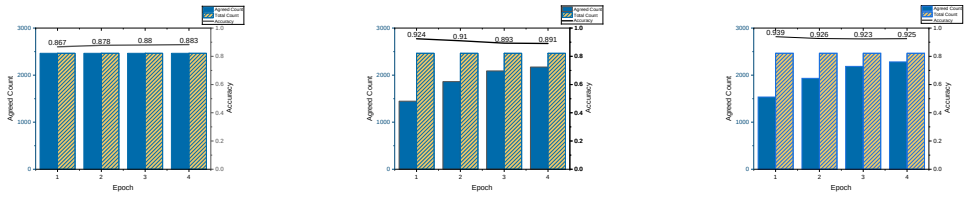


Figure 6: Visualization of cross-modal gaps on Clipart, VisDA-17, and Sketch domains as epoch.

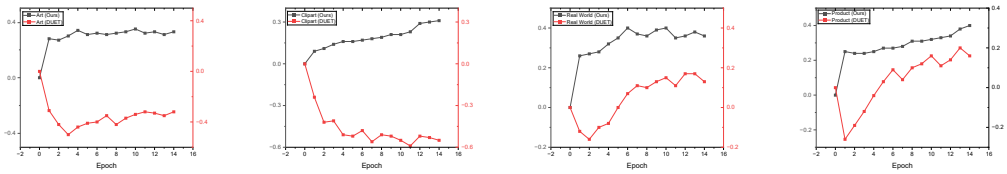
B.11 Heatmaps of Image-text Shared Attention.

Here we give a few comparisons of image-text shared semantic before and after CDMR. In theory, high joint attention on the same part means the balance and alignment of semantics across modali-



(a) DIFO (b) DUET (c) SCALP

Figure 7: Comparison of pseudo-label accuracy and labeled ratios on Art (Dual-agreed).



(a) Art (b) Clipart. (c) Real World. (d) Product.

Figure 8: Cosine similarity between optimization gradients induced by supervised training and consistency training in ASPL, which consistently demonstrates our alignment (positive), rather than gradient conflicts (negative).

Table 17: Ablation on ASPL joint optimization with CLIP ViT-L/14.

Method	Simplify	O-31	Office-Home	VisDA-17	DomainNet-126	TEP ↓ (s)
w/ alternative optimization	×	**94.6**	91.0	**91.7**	**91.0**	302.6
w/ joint optimization	✓	94.2	**91.1**	91.4	90.7	**231.2**

Table 18: Hyperparameter sensitivity analysis with CLIP ViT-L/14@336 on DomainNet-126.

Parameter	**Value**	**Range**	**Accuracy (%)**	**Sensitivity**
N (neighbors)	5	$\{3, 5, 7, 9, 11\}$	$\{91.2, 91.4, 91.3, 91.4, 91.1\}$	↓
λ (trade-off)	0.2	$[0.1, 2]$	$[91.0, 91.4, 91.2, 90.4, 90.4]$	↓
τ (temperature)	1.2	$[0.6, 1.4]$	$[89.6, 91.4]$	→
α (calibration)	0.6	$[0.1, 0.9]$	$[90.5, 91.4]$	↓

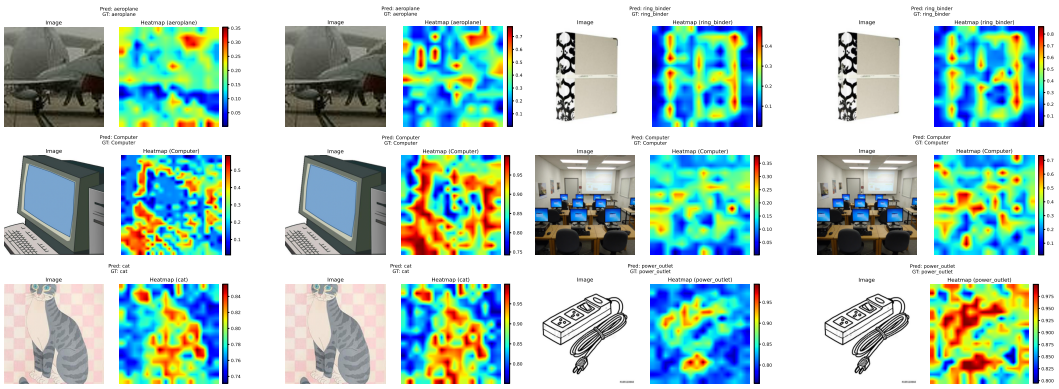
ties, since both image and text encoders are watching on that specific part. On the other hand, more task-specific proxy is supposed to offer darker red attention within following heatmaps Fig. 9.

B.12 Hyperparameter Evaluation.

In our proposed method, four key hyperparameters are considered: the number of neighbors N, the selective training trade-off coefficient λ, calibration weight α and the temperature τ applied to predictive logits. Empirically, the performance is relatively insensitive to these hyperparameters within reasonable ranges, specifically $N \in \{3, 5, 7, 9, 11\}$, $\alpha \in [0.1, 0.9]$, $\lambda \in [0.1, 2]$, and $\tau \in [0.6, 1.4]$, as illustrated in Table 18. The best performance is achieved with the configuration $\{N = 5, \lambda = 0.2, \tau = 1.2\}$.

B.13 Computational Overhead.

Although the proposed method inherently requires more resources at the first epoch due to the memory of visual features and neighbor calculation, the GPU memory and time per epoch (TPE) consumption exhibit acceptable difference when set up batchsize as 32. As Table 19, the GPU memory explodes when a larger pretrained CLIP ViT-L/14 is implemented, but Inference Time (IT) is iden-



(a) Before (b) After (c) Before (d) After

Figure 9: Comparison of image–text joint similarity on images before and after CDMR.

Table 19: Computational overhead comparison on Office-Home Art domain with CLIP ViT-L/14.

Method	TPE ↓ (s)	IT ↓ (ms)	GPU Memory ↓ (GB)
DIFO	277.6	0.421	32.0
ProDe	207.0	0.421	33.2
DUET	285.5	0.421	35.0
Colearn++	230.1	0.421	35.6
SCALP	226.4	0.421	33.8

tical across methods. We further evaluate the overhead by using an alternative training approach, as Table 17 shows minor improvements at the cost of running more time and complexity.

C Implementation and Setup

C.1 Pretraining on Source Domain.

For all transfer tasks on the four evaluation datasets, the source model θ_s is trained on the source domain using the standard supervised cross-entropy loss, following the same training protocol Eq. 23 as prior SFDA methods for fair comparison.

$$\mathcal{L}_{\text{src}} = \mathbb{E}_{(\mathbf{x}_s, y_s)\sim\mathcal{D}_s} \left[d(\phi(\theta_s(\mathbf{x}_s)), \mathbf{y}_s) \right], \tag{23}$$

where ϕ and $d(,)$ are softmax function and distance measures. The source model is trained for 30 epochs with batch size 64 and learning rate $1e^{-3}$ on Office-31, Office-Home, and DomainNet-126, for 20 epochs with batch size 32 and learning rate $3e^{-3}$ on VisDA-2017. The source dataset is divided into the training set and validation set in a 0.9:0.1 ratio as DIFO (Tang et al., 2024).

C.2 Image Preprocessing.

For all images in source domain, we follow the prior works. For target-domain images, since a pretrained CLIP is leveraged to supervise source-model adaptation, we apply the CLIP default image preprocessor by *from transformers import CLIPImageProcessor*. Specifically, given an input image $\mathbf{x}_t$, CLIP processor conducts resizing, center cropping, and normalization with defaults. Specifically, resizing the image to 224×224 or 336×336, converting it to a tensor in dependence of selected CLIP backbones. For SigLIP, we employ the corresponding `SiglipProcessor` from Hugging Face, where both image and text inputs follow the preprocessing configuration associated with the pretrained SigLIP checkpoint.

C.3 Adaptation Training.

Training with ASPL involves two objectives described in the main text. On Office-31 and Office-Home, we adopt Adam with a learning rate of 1×10^{-5} and batch size 32 for source model adaptation, while the proxy prompt v is optimized by SGD with learning rate 1×10^{-6} and momentum 0.9. On VisDA-2017 and DomainNet-126, the batch size is reduced to 16, with learning rates 1×10^{-6} and 3×10^{-7} for the source model and CLIP proxy, respectively. Training is conducted for 15 epochs on all datasets. During adaptation, the same mini-batch is jointly forwarded through both models to compute $\mathcal{L}_v^{CE}$ and $\mathcal{L}^{CE}$src, while two independent optimizers are used for simultaneous updates, as shown in Alg. 1.

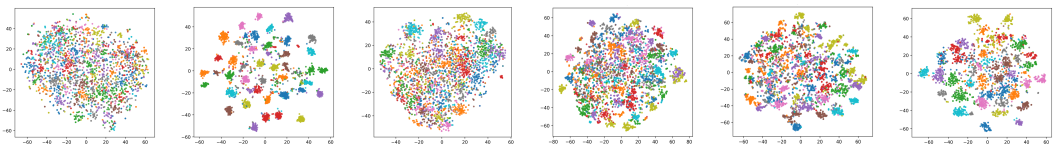
C.4 Training Implementation.

In practice, two optimization strategies can be adopted for adaptation training. The first strategy, detailed in Alg. 1, jointly updates the source and proxy models within a single forward pass for each mini-batch. Specifically, the predictions of both models are computed simultaneously to identify reliable and uncertain samples, after which the corresponding selective objectives are optimized in a unified training step. This implementation enables efficient information exchange between the two models while avoiding repeated feature extraction, and is therefore adopted for all main experiments.

Alternatively, ASPL can be implemented through an alternating optimization scheme, as described in Alg. 2. Under this setting, the source and proxy models are updated sequentially while keeping the other model fixed during each optimization stage. Such a decoupled procedure provides each model with a relatively stable supervisory signal and reduces the potential interference caused by simultaneous parameter updates. It may therefore lead to more stable optimization and better exploit the complementary predictions produced by the two models. However, alternating optimization requires separate forward and backward passes for the two update stages, resulting in higher computational overhead. As shown in Table 17, this strategy achieves slightly higher adaptation accuracy, whereas the joint optimization scheme offers a more favorable trade-off between performance and computational efficiency.

D T-SNE Map

Fig. 10 visualizes the feature representations produced by the SCALP-refined CLIP using t-SNE. Compared with the source model, adapted target model and other CLIP-based baselines, SCALP forms more compact intra-class clusters and clearer inter-class separation. This indicates that the refined proxy better captures task-specific semantics for class distinctions and reduces agonist ambiguity across classes. The visualization therefore provides qualitative evidence that SCALP improves the discriminability of target-domain representations during adaptation.



(a) Source (b) Target (c) CLIP-ZSC (d) CLIP-DUET (e) CLIP-ProDe (f) SCALP

Figure 10: t-SNE visualization of target-domain representations. (a)–(b) compare the pretrained source model and adapted target model; (c)–(f) show ZSC, DUET, ProDe, and SCALP.

E PSEUDO CODE

Algorithm 1 SCALP: Standard Training Procedure

Require: Source model θ_s; pretrained VLM θ_v with prompts v; unlabeled target dataset $\mathcal{X}_t$; number of iterations M; neighbor consistency weight λ.
Ensure: Adapted target model $\theta_t \leftarrow \theta_s^M$; refined prompts $v \leftarrow v^M$

Initialization: Set $\theta_s^{t=0} \leftarrow \theta_s$; initialize prompts $v^{t=0}$ = "A photo of a {cls}" and text embeddings $\mathrm{Z}_t^{m=0}$ from frozen θ_v; construct image embeddings bank $\{\mathrm{z}_v^i\}$ and corresponding neighbor set $\{\mathcal{N}_i\}$ from frozen CLIP.
Estimate class-centroids μ_v^k and z_t^k ▷ **CDMR**
Affine transform for realigned $\dot{\mathrm{Z}}_t^{m=0}$ via Eq.4 ▷ **CDMR**
for $m = 1$ to M **do**
    Sample a mini-batch $\mathcal{B}_T$ from $\{\mathbf{x}_t^i\}$
    Forward $\mathcal{B}_T$ through $\theta_s^{t=m-1}$ and θ_v with $\dot{\mathrm{Z}}_t^{m-1}$
    Calibration on θ_s via Eq.5
    Assign hard pseudo-labels on dual-agreed predictions via Eq.7 ▷ **ASPL**
    Synthesize predictions in $\{\mathcal{N}_i\}$ via Eq. 9 ▷ **ASPL**
    Independently update θ_s^m via Eq.10; update v^m via Eq.12.
end for
return $\theta_t \leftarrow \theta_s^M$, $v \leftarrow v^M$

Algorithm 2 CDMR-ASPL: (Optional Alternative Training)

Require: Source model θ_s; pretrained VLM θ_v with prompts v; unlabeled target dataset $\mathcal{X}_t$; number of epochs M; neighbor consistency weight λ.
Ensure: Adapted target model $\theta_t \leftarrow \theta_s^M$; refined prompts $v \leftarrow v^M$

Initialization: Set $\theta_s^0 \leftarrow \theta_s$; initialize prompts v^0 = "A photo of a {cls}"; extract initial text embeddings $\mathrm{Z}_t^{m=0}$ from frozen CLIP.
for $m = 1$ to M **do**
Source model update with fixed CLIP
    Freeze θ_v and prompts v^{m-1}
    for each mini-batch $\mathcal{B}_T \subset \{\mathbf{x}_t^i\}$ **do**
        Forward $\mathcal{B}_T$ through θ_s^{m-1} and frozen θ_v
        Construct image embedding bank $\{\mathrm{z}_v^i\}$ and neighbor sets $\{\mathcal{N}_i\}$
        Logits calibration on θ_s via Eq. 5
        Assign hard pseudo-labels on dual-agreed predictions via Eq. 7 ▷ **ASPL**
        Update θ_s^m via Eq. 10
    end for
CDMR and prompt update with fixed source model
    Freeze θ_s^m
    Estimate class-centroids μ_v^k and z_t^k
    Affine transform text embeddings $\dot{\mathrm{Z}}_t^m$ via Eq. 4 for alignment ▷ **CDMR**
    for each mini-batch $\mathcal{B}_T \subset \{\mathbf{x}_t^i\}$ **do**
        Forward $\mathcal{B}_T$ through frozen θ_s^m and θ_v with $\dot{\mathrm{Z}}_t^m$
        Assign selective supervision via Eq. 7 ▷ **ASPL**
        Update prompts v^m via Eq. 12
    end for
end for
return $\theta_t \leftarrow \theta_s^M$, $v \leftarrow v^M$