# CONDITIONAL ROLE ALLOCATION FOR HETEROGENEOUS SUPERVISION IN SOURCE-FREE DOMAIN ADAPTATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Source-Free Domain Adaptation (SFDA) adapts a pre-trained source model to an unlabeled target domain without access to source data. Foundation models provide rich and heterogeneous prior knowledge for generating target-domain supervision, yet existing approaches often assume that additional supervision is uniformly beneficial across samples. However, we find instead that its incremental value strongly depends on where it is applied: supervision from a newly introduced model, such as a multimodal large language model (MLLM), is often redundant or even disruptive for samples already supported by reliable pseudo-labels, but exhibits substantially greater corrective value on unresolved samples. Motivated by this supervision-position effect, we propose Conditional Role Allocation (CRA), which coordinates complementary foundation models according to existing reliable coverage rather than globally fusing their predictions. CRA progressively adapts the student through three stages. It first establishes reliable anchors by combining CLIP semantic confidence with DINOv2 target-domain structural support. It then expands supervision through class-wise budgets derived from structural evidence and confidence-based selection, forming a broader reliable supervision base. Finally, CRA reserves candidate-constrained MLLM adjudication for residual samples beyond reliable coverage and jointly adapts the student using both established and complementary supervision. In this way, CRA preserves reliable supervision while allocating additional model capabilities to samples with greater correction potential. Extensive experiments on three SFDA benchmarks covering closed-set, partial-set, and open-set settings demonstrate the effectiveness of CRA. CRA achieves state-of-the-art performance on most evaluated settings and consistently improves over strong baselines, without accessing source data or fine-tuning any foundation model.

## 1 INTRODUCTION

Unsupervised domain adaptation (UDA) aims to improve model performance in a target domain by jointly using labeled source data and unlabeled target data. However, access to source data can be restricted by privacy, security, and data-sharing requirements. Source-free domain adaptation (SFDA) addresses this constraint by adapting a source-trained model using only unlabeled target data. Most prior approaches derive supervision from source predictions and target data through pseudo-label self-training and neighborhood consistency (Liang et al., 2020; Yang et al., 2021). Confidence weighting and uncertainty modeling are also used to reduce the influence of noisy supervision (Lee et al., 2022; Litrico et al., 2023; Xu et al., 2025). Although these strategies improve pseudo-label quality and mitigate error accumulation, their effectiveness remains constrained by source-model bias and the quality of target representations.

Foundation models such as CLIP (Radford et al., 2021) and DINOv2 (Oquab et al., 2024) provide rich semantic and visual priors learned from large-scale pre-training, offering an external source of supervision beyond the source model itself. Recent studies have improved the reliability of foundation-model supervision through prompt learning, prediction denoising, and knowledge distillation (Tang et al., 2024b; Bai et al., 2024a; Tang et al., 2025). Beyond relying on individual model predictions, subsequent works have incorporated additional vision and vision-language



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.

models to provide complementary supervision for target adaptation (Zhang et al., 2025). More recently, multimodal large language models (MLLMs), with strong visual-semantic understanding and instruction-following capabilities, have further expanded the available sources of external supervision (Chen et al., 2026). While these approaches demonstrate that additional model supervision can improve adaptation performance on average, such aggregate gains do not necessarily imply that the added supervision is uniformly beneficial across target samples. This raises a fundamental question: *is the incremental value of additional model supervision uniform across target samples, or does it depend on where that supervision is applied?*

To examine this question, we divide target samples according to existing reliable CLIP-based supervision and evaluate five representative MLLMs under the same supervision budget. Despite their higher overall target accuracy than CLIP, Table 7 reveals a clear position-dependent effect: on reliably covered samples, four of the five MLLMs damage more correct pseudo-labels than they repair, whereas on samples outside this coverage, all five produce more corrections than errors. Figure 1 visualizes one representative case from Table 7: InternVL2.5-8B yields a net loss of 41 correct labels within the reliable coverage but a net gain of 221 outside it, with 2,033 samples evaluated in each case. We term this phenomenon the *supervision-position effect*: the value of additional supervision depends on where it is introduced relative to existing reliable coverage. This observation motivates allocating complementary supervision selectively rather than applying it uniformly across target samples.

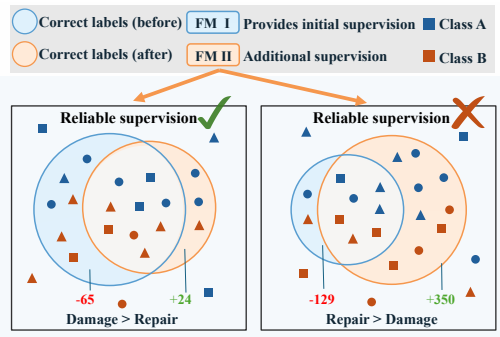


Figure 1: Conceptual illustration of the *supervision-position effect*: *(Left)* For target samples already covered by reliable supervision from FM I (e.g., CLIP), additional supervision from FM II (e.g., InternVL) can damage more correct pseudo-labels than it repairs. *(Right)* For samples beyond this reliable coverage, the same additional supervision can correct more erroneous pseudo-labels than it damages. These contrasting effects motivate allocating complementary supervision according to existing reliable coverage.

Building on this finding, we propose **Conditional Role Allocation (CRA)**. Since the value of additional supervision depends on existing reliable coverage, CRA progressively adapts the student through three stages. It first establishes reliable anchors by combining CLIP semantic confidence with DINOv2 structural support, and then expands reliable supervision through class-wise budgets derived from structural evidence and confidence-based selection. For residual samples beyond the established coverage, CRA constructs CLIP-derived candidate sets and introduces candidate-constrained MLLM adjudication. Each stage inherits the previous checkpoint while retaining established supervision, allowing complementary models to contribute where their knowledge is most valuable. Our main contributions are as follows:

- We reveal a *supervision-position effect* in foundation-model collaboration: additional supervision can be redundant or even harmful for samples already supported by reliable pseudo-labels, while providing substantial error correction for samples beyond existing reliable coverage.
- We propose CRA, a conditional role allocation framework that progressively establishes, expands, and complements target supervision by assigning semantic, structural, and MLLM-based supervision according to sample-level coverage.
- Extensive experiments across closed-set, partial-set, and open-set SFDA benchmarks demonstrate the effectiveness of CRA, achieving strong performance without access to source data or foundation-model fine-tuning.

## 2 Related Work

**Source-Free Domain Adaptation.** Existing SFDA methods construct target supervision through various strategies, including synthesizing samples or class prototypes to approximate unavailable source-domain information (Kurmi et al., 2021; Qiu et al., 2021), and learning discriminative target

representations through pseudo-label self-training, information maximization, and neighborhood consistency (Liang et al., 2020; Yang et al., 2021; 2022). To mitigate noisy supervision, subsequent studies employ confidence weighting, uncertainty modeling, and contrastive regularization to improve pseudo-label reliability and target representation quality (Lee et al., 2022; Litrico et al., 2023; Xu et al., 2025; Rai et al., 2025). However, most conventional SFDA methods still derive adaptation signals primarily from the source-trained model and unlabeled target data, making them susceptible to source-model bias and unreliable target supervision. In contrast, our work introduces heterogeneous foundation-model knowledge and focuses on how such external supervision should be selectively allocated across target samples.

**Foundation Models for Domain Adaptation.** Foundation models such as CLIP (Radford et al., 2021) and DINOv2 (Oquab et al., 2024) provide transferable semantic and visual representations that benefit learning under domain shifts. Recent studies exploit foundation-model knowledge through prompt learning, prediction denoising, uncertainty-aware optimization and knowledge distillation across domain generalization and adaptation settings (Bai et al., 2024b; Zhang et al., 2026; Tang et al., 2024b; 2025; Lee et al., 2025). In SFDA, more recent works further incorporate additional pre-trained models into target adaptation by leveraging their semantic priors to guide source knowledge transfer (Han et al., 2026) and refine predictions (Zhang et al., 2025), or by introducing multiple MLLMs through reliability-based curricula (Chen et al., 2026). These approaches primarily improve the reliability or diversity of external supervision, but largely treat its usefulness as independent of where it is introduced. Our work instead studies the sample-dependent incremental value of additional supervision and coordinates heterogeneous foundation models according to existing reliable coverage through conditional role allocation.

# 3 METHOD

**Overview.** Motivated by the supervision-position effect, we propose **Conditional Role Allocation (CRA)**, which assigns complementary supervisory roles according to reliable target-domain coverage. As illustrated in Figure 2, CRA follows three cumulative stages: establishing reliable anchors through semantic confidence and structural support, expanding supervision with class-wise budgets, and introducing MLLM guidance for residual samples. Each stage inherits the preceding checkpoint and retains previously established supervision.

**Problem Setup.** Source-free domain adaptation adapts a source-trained model $f_s$ to an unlabeled target dataset $\mathcal{D}_T = \{x_i\}_{i=1}^N$ without source data. Let $\mathcal{Y}_s$ and $\mathcal{Y}_t$ denote the source and target label spaces. We consider closed-set ($\mathcal{Y}_s = \mathcal{Y}_t$), partial-set ($\mathcal{Y}_t \subsetneq \mathcal{Y}_s$), and open-set ($\mathcal{Y}_s \subsetneq \mathcal{Y}_t$) adaptation, with target-private classes grouped as unknown. The student $f_\theta$ is initialized from $f_s$, while CLIP, DINOv2, and the MLLM remain frozen. We denote target indices by $\mathcal{I}_T = \{1, \dots, N\}$ and known classes by $\mathcal{C} = \mathcal{Y}_s$. The following procedure applies to closed-set and partial-set adaptation without access to the true target class subset. For open-set adaptation, we additionally allow explicit unknown decisions during MLLM adjudication, filter classification supervision using a dynamic known-sample mask, and incorporate an auxiliary information-maximization objective with entropy-based rejection at inference. These settings share the same three-stage adaptation framework, with setting-specific adjustments detailed in Appendix A.4

## 3.1 STAGE I: SEMANTIC–STRUCTURAL ANCHORING

Stage I selects anchors whose CLIP predictions exhibit both high semantic confidence and local support in DINOv2 space. Let $s_{ic}^C$ denote the CLIP classification score for image $x_i$ and class $c$, and let $p_{ic}^C = \mathrm{softmax}_c(s_i^C)$. The pseudo-label and confidence margin are

$$\hat{y}_i^C = \arg\max_{c \in \mathcal{C}} p_{ic}^C, \qquad m_i^C = p_{i,(1)}^C - p_{i,(2)}^C, \tag{1}$$

where $(1)$ and $(2)$ indicate the highest and second-highest scores. We use a class-specific threshold $t_c^C = Q_{\tau_C}(\{m_i^C : \hat{y}_i^C = c\})$, where $Q_{\tau_C}$ is the $\tau_C$-quantile within each nonempty predicted class.

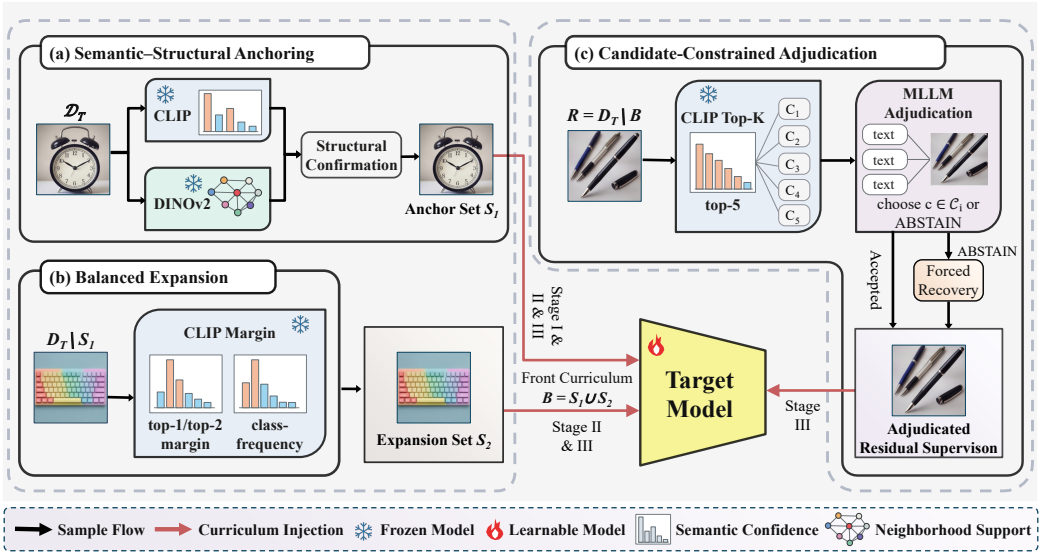


Figure 2: Overview of CRA. (a) Semantic–structural anchoring combines CLIP and DINOv2 evidence to establish reliable anchors. (b) Class-balanced expansion broadens their coverage to form the base supervision set. (c) Candidate-constrained adjudication introduces complementary MLLM supervision for the remaining residual samples. *Forced recovery* re-queries samples without valid first-pass labels, requiring a choice from the same candidate set. *Front Curriculum* denotes the base supervision set $\mathcal{B} = \mathcal{S}_1 \cup \mathcal{S}_2$ formed by anchoring and expansion.

Using normalized DINOv2 features $z_i^D$, we construct a cosine-similarity neighborhood $\mathcal{N}_{k_1}(i)$ excluding $i$. Support for the CLIP prediction is

$$\rho_i^D = \frac{1}{k_1} \sum_{j \in \mathcal{N}_{k_1}(i)} \mathbf{1}\big(\hat{y}_j^C = \hat{y}_i^C\big), \tag{2}$$

where $\mathbf{1}(\cdot)$ is the indicator function. Let $v_i$ be the most frequent CLIP pseudo-label in this neighborhood. The anchor set is

$$\mathcal{S}_1 = \Big\{ i \in \mathcal{I}_T : m_i^C \geq t_{\hat{y}_i^C}^C,\ v_i = \hat{y}_i^C,\ \rho_i^D \geq \tau_D \Big\}, \tag{3}$$

where $\tau_D$ controls neighborhood support.

We refer to this neighborhood-based verification as *structural confirmation*: DINOv2 provides visual neighborhoods to verify local structural support for CLIP predictions, rather than independently assigning semantic labels.

The student is initially trained only on $\mathcal{S}_1$ using the fixed pseudo-labels $\hat{y}_i^C$ and label-smoothed cross-entropy. Sampling uses inverse pseudo-label class frequencies with replacement to mitigate class imbalance.

### 3.2 Stage II: Class-Balanced Expansion

Stage II broadens supervision while retaining the anchors. For each non-anchor sample, neighboring anchors vote in DINOv2 space among its top-three CLIP classes, with similarity-based weights. A sample enters a reference set $\mathcal{V} \subseteq \mathcal{I}_T \backslash \mathcal{S}_1$ when it has sufficient anchor neighbors and candidate-class voting support, and the winning class agrees with $\hat{y}_i^C$. Voting criteria are detailed in Appendix A.2.

The reference set determines the expansion budget $b_c = \sum_{i \in \mathcal{V}} \mathbf{1}(\hat{y}_i^C = c)$. For the non-anchor pool $\mathcal{U}_c = \{i \in \mathcal{I}_T \setminus \mathcal{S}_1 : \hat{y}_i^C = c\}$, we select

$$\mathcal{S}_{2,c} = \mathrm{Top}_{b_c}(\mathcal{U}_c; m^C), \qquad \mathcal{S}_2 = \bigcup_{c \in \mathcal{C}} \mathcal{S}_{2,c}, \qquad \mathcal{B} = \mathcal{S}_1 \cup \mathcal{S}_2, \tag{4}$$

where $\mathrm{Top}_{b_c}$ returns the $b_c$ highest-margin indices, with an empty selection when $b_c = 0$. Structural support determines how many samples to add per class, whereas CLIP confidence determines which samples are selected. Expansion samples need not belong to $\mathcal{V}$ and retain their CLIP pseudo-labels. Thereby broadening target-domain coverage while preserving anchor supervision.

Training continues from Stage I by sampling both anchors and expansion samples, using inverse-class-frequency sampling within each set. Anchors receive label-smoothed cross-entropy, while expansion samples receive ordinary cross-entropy. We subsequently consolidate the complete base set $\mathcal{B}$ using class-balanced sampling and label-smoothed cross-entropy, denoted by $\mathcal{L}_{\mathrm{base}}$. This consolidation changes neither set membership nor pseudo-labels, preserving the established supervision for joint training in Stage III. Sampling ratios, loss weights, and checkpoint smoothing are specified in Appendix A.1.

### 3.3 Stage III: Candidate-Constrained Residual Adjudication

Stage III introduces complementary supervision for the residual set $\mathcal{R} = \mathcal{I}_T \setminus \mathcal{B}$ while retaining the complete base set established in the previous stages.

**MLLM Adjudication and Force Recovery.** For each $i \in \mathcal{R}$, CLIP supplies the candidate set $\mathcal{C}_i = \mathrm{Top}_q(\mathcal{C}; s_i^C)$. The MLLM receives $x_i$ and $\mathcal{C}_i$ and is instructed to select a candidate or return `ABSTAIN`. Parseable candidate labels are accepted directly. Other responses, including abstentions, trigger *forced recovery*: a second query that retains the same candidate set but requires the MLLM to select one class rather than abstain. Valid recovered labels are accepted; samples still unresolved remain available for student-based supervision. `ABSTAIN` allows the MLLM to defer a decision when visual evidence is insufficient, whereas *forced recovery* explicitly requires a choice among the same candidates. Here, `ABSTAIN` denotes an unassigned first-pass label, not an unknown-class prediction. Accepted MLLM labels $y_i^M$ and adjudication states are fixed before training.

**Complementary Target Selection.** Let $f_{\bar{\theta}}$ denote the parameter-smoothed student, whose weak-view prediction is $\pi_i = f_{\bar{\theta}}(a_w(x_i))$. Define its predicted class $\hat{y}_i^S = \arg\max_{c \in \mathcal{C}} \pi_{ic}$ and confidence $u_i^S = \max_{c \in \mathcal{C}} \pi_{ic}$. The residual training target is

$$\tilde{y}_i = \begin{cases} \hat{y}_i^S, & u_i^S \geq \tau_S, \\ y_i^M, & u_i^S < \tau_S, \text{ valid MLLM label}, \\ \arg\max_{c \in \mathcal{C}_i} \pi_{ic}, & u_i^S < \tau_S, \text{ otherwise}, \end{cases} \tag{5}$$

where $\tau_S$ is the student-confidence threshold. Confident student predictions can select any known class; otherwise, MLLM guidance or a candidate-constrained student prediction supplies the target. Targets are detached during each update but may change during training.

**Joint Optimization.** Following the weak/strong augmentation scheme in RCL (Chen et al., 2026), we generate a weakly augmented view $a_w(x_i)$ with mild appearance changes and a strongly augmented view $a_s(x_i)$ with more substantial perturbations, both intended to preserve class identity. Both views are supervised by the same target $\tilde{y}_i$. The weak-view loss $\mathcal{L}_{\mathrm{guided}}$ is a weighted cross-entropy that learns the selected target, while the strong-view loss $\mathcal{L}_{\mathrm{cons}}$ applies the same supervision under stronger perturbations to encourage augmentation consistency. Each loss is normalized by the sum of sample weights. The weight $w_i$ depends on the fixed adjudication state: 1 for valid first-pass labels, $\eta$ for recovered labels, and $\gamma$ for unresolved samples, regardless of the current target source.

Each iteration samples from both $\mathcal{B}$ and $\mathcal{R}$ and optimizes

$$\mathcal{L}_{\mathrm{III}} = \lambda_b \alpha \mathcal{L}_{\mathrm{base}} + (1 - \alpha)\left(\lambda_g \mathcal{L}_{\mathrm{guided}} + \lambda_c \mathcal{L}_{\mathrm{cons}}\right), \tag{6}$$

where $\alpha$ balances base and residual supervision, and $\lambda_b, \lambda_g, \lambda_c$ weight the loss terms. Both sets use inverse-class-frequency sampling. Sampling labels are fixed: CLIP pseudo-labels for $\mathcal{B}$, and accepted MLLM labels or, when unavailable, CLIP pseudo-labels for $\mathcal{R}$. These sampling labels are distinct from the dynamic targets in Equation 5. All losses update $f_\theta$, while $f_{\bar{\theta}}$ is updated by a moving average of student parameters without backpropagation. Detailed adjudication and training settings appear in Appendices A.2 and A.1. Only $f_{\bar{\theta}}$ is retained for deployment.

Table 1: **Closed-set** SFDA accuracy (%) on Office-Home. SF means source-free.

| Method | SF | CLIP | MLLM | 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. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SHOT | ✓ | ✗ | ✗ | 57.1 | 78.1 | 81.5 | 68.0 | 78.2 | 78.1 | 67.4 | 54.9 | 82.2 | 73.3 | 58.8 | 84.3 | 71.8 |
| TPDS | ✓ | ✗ | ✗ | 59.3 | 80.3 | 82.1 | 70.6 | 79.4 | 80.9 | 69.8 | 56.8 | 82.1 | 74.5 | 61.2 | 85.3 | 73.5 |
| UCon-SFDA | ✓ | ✗ | ✗ | 65.6 | 87.8 | 91.0 | 78.6 | 79.3 | 87.6 | 80.2 | 65.9 | 87.3 | 83.2 | 69.1 | 88.7 | 80.3 |
| DIFO | ✓ | ✓ | ✗ | 70.6 | 90.6 | 88.8 | 82.5 | 90.6 | 88.8 | 80.9 | 70.1 | 88.9 | 83.4 | 70.5 | 91.2 | 83.1 |
| ProDe | ✓ | ✓ | ✗ | 64.0 | 90.0 | 88.3 | 81.1 | 90.1 | 88.6 | 79.8 | 65.4 | 89.0 | 80.9 | 65.5 | 90.2 | 81.1 |
| DIFO++ | ✓ | ✓ | ✗ | 73.2 | 91.4 | 89.7 | 83.6 | 91.0 | 89.8 | 80.8 | 73.8 | 90.0 | 84.3 | 74.2 | 92.3 | 84.5 |
| RCL | ✓ | ✗ | ✓ | 82.5 | 95.3 | 93.3 | 89.1 | 95.3 | 92.7 | 89.3 | 82.4 | 92.8 | 89.4 | 82.1 | 95.4 | 90.0 |
| **CRA (Ours)** | ✓ | ✓ | ✓ | **84.3** | **96.3** | **94.2** | **89.7** | **96.3** | **94.2** | **89.6** | **84.2** | **94.2** | **89.7** | **84.3** | **96.3** | **91.1** |

Table 2: **Closed-set** SFDA results (%) on DomainNet-126 and VisDA-C. SF denotes source-free.

| | | | | **DomainNet-126** | | | | | | | | | | | | | **VisDA-C** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | SF | CLIP | MLLM | 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. | S→R |
| SHOT | ✓ | ✗ | ✗ | 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 | 82.7 |
| TPDS | ✓ | ✗ | ✗ | 62.9 | 77.1 | 59.8 | 65.6 | 79.0 | 61.5 | 66.4 | 67.0 | 58.2 | 68.6 | 64.3 | 75.3 | 67.1 | 87.6 |
| DIFO | ✓ | ✓ | ✗ | 76.6 | 87.2 | 74.9 | 80.0 | 87.4 | 75.6 | 80.8 | 77.3 | 75.5 | 80.5 | 76.7 | 87.3 | 80.0 | 90.3 |
| ProDe | ✓ | ✓ | ✗ | 79.3 | 91.0 | 75.3 | 80.0 | 90.9 | 75.6 | 80.4 | 78.9 | 75.4 | 80.4 | 79.2 | 91.0 | 81.5 | 88.7 |
| DIFO++ | ✓ | ✓ | ✗ | 78.1 | 89.7 | 76.5 | 81.7 | 89.6 | 77.2 | 82.3 | 80.2 | 75.9 | 82.3 | 79.4 | 89.7 | 81.9 | 90.5 |
| RCL | ✓ | ✗ | ✓ | 87.6 | 92.8 | **87.9** | 89.2 | 92.7 | **87.8** | 89.6 | 87.7 | **87.6** | 89.4 | 87.5 | 92.7 | 89.4 | **93.2** |
| **CRA (Ours)** | ✓ | ✓ | ✓ | **87.8** | **93.5** | 87.1 | **90.1** | **93.5** | 87.0 | **90.1** | **87.9** | 87.1 | **90.1** | **87.8** | **93.5** | **89.6** | 91.9 |

# 4 EXPERIMENTS

## 4.1 EXPERIMENTAL SETUP

**Benchmarks.** We evaluate our method on three widely used domain adaptation benchmarks: Office-Home (Venkateswara et al., 2017), DomainNet-126 (Peng et al., 2019), and VisDA-C (Peng et al., 2017). Office-Home evaluates cross-domain object recognition, DomainNet-126 is a 126-class subset of DomainNet for large-scale multi-domain adaptation, and VisDA-C focuses on synthetic-to-real adaptation. Together covering diverse visual styles and varying degrees of domain shift Dataset statistics and domain splits are provided in Appendix A.3.

**Implementation Details.** We use ResNet-50 (He et al., 2016) as the student backbone for Office-Home and DomainNet-126, and ResNet-101 for VisDA-C. The student is initialized from the source checkpoints provided by SHOT (Liang et al., 2020) for Office-Home and VisDA-C, and by PLUE (Litrico et al., 2023) for DomainNet-126. During adaptation, CRA uses three frozen foundation models: OpenCLIP ViT-B/32 (Radford et al., 2021) for semantic scoring and pseudo-labeling, DINOv2 ViT-S/14 (Oquab et al., 2024) for target-domain neighborhood construction, and Qwen2-VL-7B-Instruct (Wang et al., 2024) for residual-sample adjudication. Only the student is optimized during adaptation and deployed at inference.

**Baselines.** We evaluate CRA under three widely studied SFDA scenarios: closed-set, partial-set, and open-set adaptation. We organize the baselines according to the supervision and external knowledge used during target adaptation. Traditional SFDA methods rely primarily on source-model predictions and target-domain structure, including SHOT (Liang et al., 2020), TPDS (Tang et al., 2024a), UCon-SFDA (Xu et al., 2025), HCL (Huang et al., 2021), CoWA (Lee et al., 2022), and CRS (Zhang et al., 2023). Foundation-model-based methods further exploit vision-language knowledge, including DIFO (Tang et al., 2024b), DIFO++ (Tang et al., 2026), and ProDe (Tang et al., 2025). We additionally compare with the MLLM-guided RCL (Chen et al., 2026), which constructs reliability-based curriculum supervision using multiple MLLMs.

## 4.2 MAIN RESULTS

**Comparison on Closed-set SFDA.** Tables 1 and 2 report the closed-set SFDA results. CRA achieves the best average performance on both Office-Home and DomainNet-126. On Office-Home, CRA reaches 91.1% average accuracy, outperforming RCL by 1.1 percentage points and ranking first on all 12 transfer tasks, highlighting its advantages across diverse and challenging domain shifts. On

DomainNet-126, CRA achieves 89.6%, improving over RCL by 0.2 points and obtaining the best results on 9 of the 12 transfers, demonstrating strong scalability to larger label spaces. Notably, the only three transfers where CRA does not outperform RCL all use Sketch as the target domain, indicating that the remaining variation is concentrated in this challenging domain.

On VisDA-C, CRA obtains 91.9% mean-class accuracy, compared with 93.2% for RCL. One relevant difference is that RCL employs multiple large MLLMs as zero-shot teachers, whose individual predictions are already particularly strong on VisDA-C, whereas CRA uses a single 7B MLLM and invokes it selectively only for residual samples. Thus, VisDA-C represents a setting in which broad direct MLLM supervision is especially competitive. Nevertheless, CRA remains competitive while following a more selective supervision strategy designed to preserve reliable existing knowledge. Detailed class-wise results are provided in Appendix B.1.

**Comparison on Partial-set and Open-set SFDA.** For a more comprehensive evaluation, we further evaluate CRA under partial-set and open-set SFDA settings on Office-Home. CRA achieves substantial improvements in both settings. As shown in Table 3, CRA obtains an average accuracy of 94.0% under partial-set SFDA and an average OS of 89.2% under open-set SFDA, surpassing the strongest compared baselines, DIFO++ and ProDe, by 7.3 and 6.6 percentage points, respectively. These results demonstrate that CRA remains effective under both missing and unseen target classes. Full per-transfer results are provided in Appendix B.2.

Table 3: **Partial-set and open-set** SFDA results (%) on Office-Home. FM indicates external foundation models used for adaptation.

| **Partial-set** | FM | Avg. | **Open-set** | FM | Avg. |
|---|---|---|---|---|---|
| Source | ✗ | 62.8 | Source | ✗ | 46.6 |
| SHOT | ✗ | 79.3 | SHOT | ✗ | 72.8 |
| HCL | ✗ | 79.6 | HCL | ✗ | 72.6 |
| CoWA | ✗ | 83.2 | CoWA | ✗ | 73.2 |
| CRS | ✗ | 80.6 | CRS | ✗ | 73.2 |
| DIFO | ✓ | 85.6 | DIFO | ✓ | 75.9 |
| ProDe | ✓ | 84.2 | ProDe | ✓ | <u>82.6</u> |
| DIFO++ | ✓ | <u>86.7</u> | DIFO++ | ✓ | 77.0 |
| **CRA (Ours)** | ✓ | **94.0** | **CRA (Ours)** | ✓ | **89.2** |

### 4.3 Ablation and Analysis Study

**Effect of MLLM Residual Supervision.** Table 4 examines whether MLLM judgments provide complementary supervision for residual samples on Office-Home A→C and P→C. We keep the residual set, CLIP candidate sets, adjudication-state partition, sample weights, student initialization, and training protocol fixed, and replace only the accepted Qwen2-VL-7B-Instruct labels with labels derived from CLIP or DINOv2-assisted CLIP. The latter uses similarity-weighted voting over anchor pseudo-labels in DINOv2 neighborhoods to provide structural evidence for reranking the CLIP candidate classes. Compared with CLIP supervision, MLLM supervision improves accuracy by 7.5 and 7.0 percentage points on A→C and P→C, respectively; compared with DINOv2-assisted CLIP, the gains increase to 9.9 and 9.4 points. These results show that, under matched residual samples and training conditions, MLLM adjudication provides complementary supervision that cannot be recovered by CLIP semantic confidence or DINOv2-assisted structural evidence alone.

**Ablation on CRA Components.** Table 5 evaluates the contributions of the three stages of CRA. Stage I alone achieves an average accuracy of 73.1% using a limited set of reliable anchors, providing a stable starting point for adaptation. Adding Stage II expands reliable supervision across classes and improves the average accuracy to 80.9%. Alternatively, introducing Stage III directly after Stage I raises the accuracy to 85.7%, demonstrating the complementary value of MLLM supervision on samples outside the reliable anchor set. Combining all three stages further improves the average accuracy to 91.2% and achieves the best performance on all four transfers. These results validate the complementary roles of the three stages: Stage I establishes reliable supervision, Stage II broadens its coverage, and Stage III provides additional correction for the remaining residual samples.

**Parameter Sensitivity.** Figure 3 examines the sensitivity of CRA to the CLIP within-class margin quantile $\tau_C$, the DINOv2 neighborhood-support threshold $\tau_D$, and the recovery-supervision weight $\eta$ on Office-Home A→C and C→A. Performance remains stable across all tested ranges, with accuracy varying by at most 0.7 percentage points within each sweep. This indicates that CRA is relatively insensitive to these key hyperparameters. Additional experimental settings and the candidate-count analysis are provided in Appendix B.6.

Table 4: Ablation of residual supervision sources on Office-Home A→C and P→C.

| Residual supervision | A→C | P→C |
|---|---|---|
| **CRA (Ours)** | **84.3** | **84.2** |
| CRA w/ CLIP supervision | 76.8 | 77.2 |
| CRA w/ DINOv2-assisted CLIP | 74.4 | 74.8 |

Table 5: Ablation of the three stages of CRA.

| Stage I Anchoring | Stage II Expansion | Stage III Residual | Office-Home C→A | A→C | R→P | P→R |
|---|---|---|---|---|---|---|
| ✓ | ✗ | ✗ | 66.1 | 60.1 | 82.7 | 83.7 |
| ✓ | ✓ | ✗ | 72.8 | 69.7 | 90.9 | 90.4 |
| ✓ | ✗ | ✓ | 83.2 | 80.2 | 91.3 | 88.1 |
| ✓ | ✓ | ✓ | **89.7** | **84.4** | **96.3** | **94.2** |

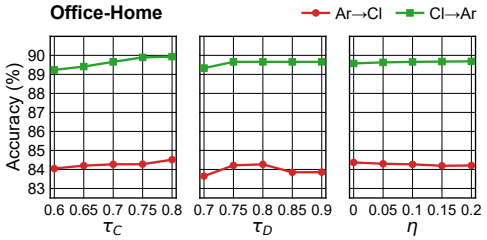


Figure 3: Parameter sensitivity of CRA on A→C and C→A. The panels vary the CLIP within-class margin quantile level $\tau_C$, DINOv2 neighborhood-support threshold $\tau_D$, and forced-recovery weight $\eta$, respectively. Here $\eta$ is distinct from label smoothing $\epsilon_{\text{ls}}$. In the image legends, Ar and Cl denote Art and Clipart.

**Conditional vs. Global Supervision.** Table 6 compares CRA with students trained using global CLIP, DINOv2, or MLLM supervision on three Office-Home transfers. Among the global strategies, MLLM supervision performs best, achieving an average student accuracy of 87.8%. CRA further improves the average to 89.5%, a gain of 1.7 percentage points, and consistently outperforms global MLLM supervision on all three transfers. These results show that CRA's gains do not simply arise from using a stronger supervision source; rather, conditionally assigning complementary supervision according to existing reliable coverage is more effective than applying a single supervisor globally. Detailed full-target prediction statistics are provided in Appendix B.3.

Table 6: Student accuracy (%) under global single-model supervision and CRA on three Office-Home transfers.

| Method | A→C | C→A | P→R | Avg. |
|---|---|---|---|---|
| CLIP | 75.1 | 83.0 | 90.8 | 83.0 |
| DINOv2 | 71.5 | 79.5 | 89.8 | 80.3 |
| MLLM | 81.1 | 89.5 | 92.8 | 87.8 |
| **CRA (Ours)** | **84.4** | **89.7** | **94.2** | **89.5** |

**Effect of Supervision Position.** Table 7 examines how the value of MLLM supervision changes with its position relative to existing reliable coverage. We compare adding MLLM labels within the reliably supervised region and beyond this region on Office-Home A→C, while matching the effective label budget and training protocol. Although the MLLMs generally achieve higher absolute label accuracy within the reliable region, their incremental value is substantially lower there. Specifically, four of the five MLLMs damage more correct CLIP labels than they repair, resulting in negative net correction. In contrast, beyond reliable coverage, all five MLLMs correct more erroneous CLIP labels than they corrupt and therefore provide positive net correction. This shows that high prediction accuracy does not necessarily imply high supervisory value when reliable supervision already exists.

Table 7: Effect of additional MLLM supervision within and beyond existing reliable coverage on Office-Home A→C under matched effective label budgets.

| MLLM | Within reliable coverage | | | Beyond reliable coverage | | |
|---|---|---|---|---|---|---|
| | Acc. | Repair / Damage / Net | Student | Acc. | Repair / Damage / Net | Student |
| Qwen2-VL-7B-Instruct | 94.0 | 40 / 15 / +25 | 77.0 | 75.6 | 461 / 61 / +400 | 84.3 |
| InternVL2.5-8B | 90.5 | 24 / 65 / −41 | 76.0 | 65.6 | 350 / 129 / +221 | 80.9 |
| Phi-3.5-Vision-Instruct | 90.6 | 26 / 65 / −39 | 75.6 | 65.4 | 352 / 135 / +217 | 79.8 |
| LLaVA-1.6-34B-AWQ | 90.1 | 34 / 83 / −49 | 75.9 | 63.8 | 348 / 162 / +186 | 80.4 |
| ShareGPT4V-7B | 89.2 | 15 / 82 / −67 | 75.1 | 58.7 | 208 / 127 / +81 | 78.6 |

Figure 4 further verifies that this position-dependent effect translates into downstream student performance. Under the same MLLM supervision budget, allocating additional judgments beyond reliable coverage consistently outperforms allocating them within reliable coverage across five transfers, improving student accuracy by 2.4–7.3 percentage points. Together, these results reveal the *supervision-position effect*: additional supervision is most useful when it complements missing or unreliable supervision rather than replacing already reliable knowledge. This directly motivates

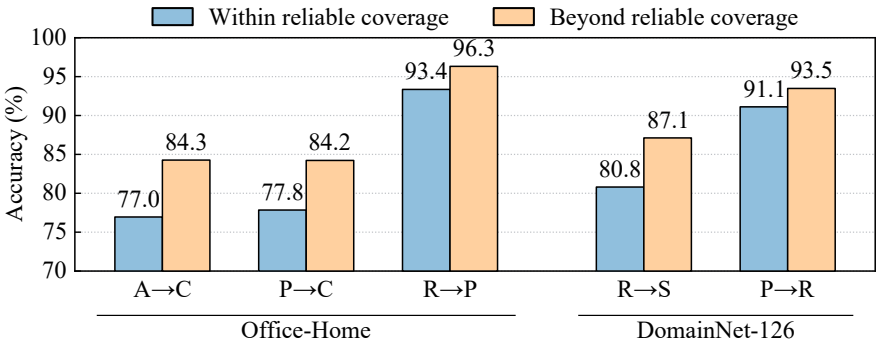


Figure 4: Effect of MLLM supervision position on student accuracy. Paired bars compare allocation within and beyond existing reliable coverage, with the effective label budget and student training protocol fixed within each transfer.

CRA to preserve reliable supervision and reserve MLLM adjudication for residual samples. Detailed cross-transfer results and additional allocation controls are provided in Appendix B.4.

### 4.4 Visualization

**T-SNE Visualization.** Figure 5 visualizes the feature distributions of the same target samples across different adaptation stages, with DIFO and the supervised Oracle included as references. Colors indicate ground-truth classes for visualization only. Compared with the source model and DIFO, CRA produces more compact and better-separated class clusters with reduced inter-class overlap, while progressively approaching the structure exhibited by the Oracle. These results provide qualitative evidence that the three-stage adaptation progressively improves the discriminability of target-domain representations. Additional stage-wise Grad-CAM visualizations and their selection settings are provided in Appendix B.8.

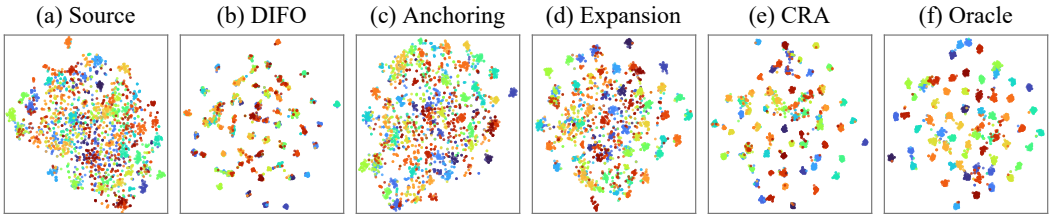


Figure 5: Stage-wise refinement of target features: (a) Source, (b) DIFO, (c) Anchoring, (d) Expansion, (e) CRA, and (f) Oracle. Oracle is a separate model trained with target-domain ground-truth labels solely as a supervised reference.

## 5 Conclusion

In this work, we identify a *supervision-position effect* in foundation-model collaboration: additional supervision can be redundant or even disruptive for samples already supported by reliable supervision, while providing substantial complementary correction for samples outside existing reliable coverage. Motivated by this finding, we propose Conditional Role Allocation (CRA), which first establishes and expands reliable target supervision using complementary semantic and structural evidence, and then reserves candidate-constrained MLLM adjudication for the remaining residual samples. Extensive experiments across three benchmarks and closed-set, partial-set, and open-set SFDA settings demonstrate the effectiveness of CRA without access to source data or foundation-model fine-tuning. Our results suggest that effective foundation-model collaboration depends not simply on introducing stronger or more diverse supervision, but on allocating supervisory roles according to where additional knowledge provides incremental value.

### AI USE STATEMENT

The authors used generative AI tools solely for language editing and polishing of the manuscript, including improving grammar, clarity, readability, and presentation. All research ideas, methodological design, experimental procedures, results, analyses, and conclusions were independently developed and verified by the authors. The authors carefully reviewed and approved all AI-assisted edits and take full responsibility for the accuracy, integrity, and final content of the paper.

### REPRODUCIBILITY STATEMENT

We provide detailed implementation and evaluation information to facilitate reproducibility. The main paper specifies the datasets, source-model initialization, supervision construction, optimization objective, and complete CRA pipeline, while the appendix provides additional implementation details, training configurations, and supplementary analyses. We will release the codebase after publication, including training and evaluation code, configuration files, source-checkpoint information, and scripts for reproducing the main experimental results.

## REFERENCES

Shuanghao Bai, Min Zhang, Wanqi Zhou, Siteng Huang, Zhirong Luan, Donglin Wang, and Badong Chen. Prompt-based distribution alignment for unsupervised domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 729–737, 2024a.

Shuanghao Bai, Yuedi Zhang, Wanqi Zhou, Zhirong Luan, and Badong Chen. Soft prompt generation for domain generalization. In *European Conference on Computer Vision*, pp. 434–450. Springer, 2024b.

Dongjie Chen, Kartik Patwari, Zhengfeng Lai, Xiaoguang Zhu, Sen-ching Cheung, and Chen-Nee Chuah. Empowering source-free domain adaptation via mllm-guided reliability-based curriculum learning. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4262–4272. IEEE, 2026.

Shuo Han, Xu Tang, Jingjing Ma, and Xiangrong Zhang. Vision-language model guided source-free domain adaptation via optimal transport. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36989–36998, June 2026.

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

Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. In *Advances in neural information processing systems*, volume 34, pp. 3635–3649, 2021.

Vinod K Kurmi, Venkatesh K Subramanian, and Vinay P Namboodiri. Domain impression: A source data free domain adaptation method. In *2021 IEEE Winter conference on applications of computer vision (WACV)*, pp. 615–625. IEEE, 2021.

JaeYun Lee, Jae Hyeon Park, Gyoomin Lee, Bogyeong Kim, Min Hee Cha, Hyeok Nam, Joo Jeon, Hyunse Lee, and Sung In Cho. Duet: Dual-perspective pseudo labeling and uncertainty-aware exploration & exploitation training for source-free domain adaptation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 74927–74954. Curran Associates, Inc., 2025. doi: 10.52202/085713-2516. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/6c8dfbbd1ba3e22339e58a336cbed52b-Paper-Conference.pdf.

Jonghyun Lee, Dahuin Jung, Junho Yim, and Sungroh Yoon. Confidence score for source-free unsupervised domain adaptation. In *International conference on machine learning*, pp. 12365–12377. PMLR, 2022.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International conference on machine learning*, pp. 6028–6039. PMLR, 2020.

Mattia Litrico, Alessio Del Bue, and Pietro Morerio. Guiding pseudo-labels with uncertainty estimation for source-free unsupervised domain adaptation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7640–7650. IEEE, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1406–1415. IEEE, 2019.

Zhen Qiu, Yifan Zhang, Hongbin Lin, Shuaicheng Niu, Yanxia Liu, Qing Du, and Mingkui Tan. Source-free domain adaptation via avatar prototype generation and adaptation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 2921–2927, 2021. doi: 10.24963/ijcai.2021/402.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Shivangi Rai, Rini Smita Thakur, Kunal Jangid, and Vinod K Kurmi. Label calibration in source free domain adaptation. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pp. 6446–6455, February 2025.

Song Tang, An Chang, Fabian Zhang, Xiatian Zhu, Mao Ye, and Changshui Zhang. Source-free domain adaptation via target prediction distribution searching. *International journal of computer vision*, 132(3):654–672, 2024a.

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 (CVPR)*, pp. 23711–23720. IEEE, 2024b.

Song Tang, Wenxin Su, Yan Gan, Mao Ye, Jianwei Zhang, and Xiatian Zhu. Proxy denoising for source-free domain adaptation. In *International Conference on Learning Representations*, volume 2025, pp. 82548–82569, 2025.

Song Tang, Yunxiang Bai, Wenxin Su, Mao Ye, Jianwei Zhang, and Xiatian Zhu. Source-free domain adaptation with vision-language prior. *International Journal of Computer Vision*, 134 (10):433, 2026.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5385–5394. IEEE, 2017.

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

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 *International Conference on Learning Representations*, volume 2025, pp. 92900–92939, 2025.

Shiqi Yang, Joost Van de Weijer, Luis Herranz, Shangling Jui, et al. 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, Shangling Jui, Joost Van De Weijer, et al. Attracting and dispersing: A simple approach for source-free domain adaptation. In *Advances in neural information processing systems*, volume 35, pp. 5802–5815, 2022.

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.

Yixin Zhang, Zilei Wang, and Weinan He. Class relationship embedded learning for source-free unsupervised domain adaptation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7619–7629. IEEE, 2023.

Yuedi Zhang, Shuanghao Bai, Wanqi Zhou, Zhirong Luan, and Badong Chen. Dual-path stable soft prompt generation for domain generalization. In *2026 IEEE International Conference on Multimedia and Expo*, 2026.

This supplementary material provides implementation details and additional results for CRA. Appendix A describes the training configurations, supervision construction, MLLM adjudication pipeline, and evaluation datasets. Appendix B presents additional class-wise results, full-target prediction statistics, experimental controls, candidate-coverage analyses, supervision-request statistics, and visualization details. Appendix C discusses the implications, limitations, and future directions of CRA.

# A Training and Implementation Details

## A.1 Training Details

CRA is implemented in PyTorch, and all experiments are conducted on a single NVIDIA RTX 4090 GPU. The student is optimized with SGD during Stage I and Adam during Stages II and III, using a batch size of 64 throughout. All foundation models remain frozen, and their outputs are pre-computed and cached for student training. We use cosine learning-rate annealing and a weight decay of $10^{-4}$ for all training stages. SGD additionally uses a momentum of 0.9 with Nesterov acceleration. Dataset-specific training configurations are summarized in Table S1.

Table S1: Dataset-specific student-training configurations.

| Dataset | Stage | Optimizer | Initial learning rate | Batch size | Training duration |
|---|---|---|---|---|---|
| Office-Home | Stage I: Anchoring | SGD | $10^{-3} \rightarrow 5 \times 10^{-4}$ | 64 | 10 + 15 epochs |
| Office-Home | Stage II: Expansion | Adam | $10^{-5}$ | 64 | 1,000 + 600 training steps |
| Office-Home | Stage III: Residual | Adam | $10^{-5}$ | 64 | 3,000 training steps |
| DomainNet-126 | Stage I: Anchoring | SGD | $10^{-3} \rightarrow 5 \times 10^{-4}$ | 64 | 10 + 15 epochs |
| DomainNet-126 | Stage II: Expansion | Adam | $10^{-5}$ | 64 | 1,000 + 600 training steps |
| DomainNet-126 | Stage III: Residual | Adam | $10^{-5}$ | 64 | 15,000 training steps |
| VisDA-C | Stage I: Anchoring | SGD | $10^{-3} \rightarrow 5 \times 10^{-4}$ | 64 | 10 + 15 epochs |
| VisDA-C | Stage II: Expansion | Adam | $10^{-5}$ | 64 | 1,000 + 600 training steps |
| VisDA-C | Stage III: Residual | Adam | $10^{-5}$ | 64 | 15,000 training steps |

Stage I consists of two consecutive training segments over the same anchor set: 10 epochs with an initial learning rate of $10^{-3}$, followed by 15 epochs with an initial learning rate of $5 \times 10^{-4}$. Stage II first performs 1,000 training steps using equal proportions of anchor and expansion samples, followed by 600 additional training steps on the resulting base supervision set $\mathcal{B} = \mathcal{S}_1 \cup \mathcal{S}_2$. Stage III jointly trains on base and residual supervision using a 1:1 sampling ratio. The optimizer and learning-rate schedule are initialized separately for each training segment.

The key hyperparameters are summarized in Table S2.

Table S2: Key hyperparameters of CRA.

| Parameter | Default value |
|---|---|
| Semantic-margin quantile $\tau_C$ | 0.70 |
| Structural-support threshold $\tau_D$ | 0.80 |
| Candidate count $q$ | 5 |
| Recovery weight $\eta$ | 0.10 |
| Base/residual mixing ratio $\alpha$ | 0.50 |
| Loss coefficients $(\lambda_b, \lambda_g, \lambda_c)$ | $(1.0, 1.0, 0.5)$ |

The quantile $\tau_C$ is applied independently within each CLIP-predicted class to determine the class-specific semantic threshold $\tau_c^C$. Label smoothing with $\epsilon_{\text{ls}} = 0.10$ is applied to anchor supervision in Stage I and to the base-supervision loss over $\mathcal{B}$ in Stage III. During Stage II, newly selected expansion samples are trained with ordinary cross-entropy, while anchor samples are replayed with label-smoothed cross-entropy. Residual supervision in Stage III uses weighted cross-entropy. The anchor/expansion sampling ratio in Stage II and the base/residual sampling ratio in Stage III are both 1:1. Three-run experiments use student-training seeds 2026, 2027, and 2028; single-seed results are explicitly identified in the corresponding tables and figures.

### A.2 Supervision Construction and Adjudication Details

**Stages I–II: Supervision Construction.** For Stage I, the semantic threshold is computed independently within each CLIP-predicted class using the empirical quantile specified by $\tau_C$. Linear interpolation is used when the desired quantile falls between samples. Ties may cause the retained proportion to deviate slightly from the nominal quantile, and classes receiving no CLIP predictions contribute no anchors. Structural support is computed using the $k = 10$ nearest neighbors under cosine similarity in the $\ell_2$-normalized DINOv2 feature space, excluding the query itself.

For Stage II, CRA constructs a reference set $\mathcal{V}$ used only to determine the class-wise expansion budgets. For each non-anchor sample, we retrieve its 20 nearest target neighbors in DINOv2 feature space and convert their similarities into weights using a softmax temperature of 0.07. Only anchor-neighbor votes are retained, and voting is restricted to the query's top-three CLIP candidate classes. A sample enters $\mathcal{V}$ when it has at least three anchor neighbors, receives positive restricted voting evidence, achieves a normalized winning-class support of at least 0.80, and the winning class agrees with its CLIP top-1 prediction.

The class budget $b_c$ is the number of reference samples assigned to class $c$. CRA then selects the $b_c$ highest-margin samples among all non-anchor samples predicted as class $c$ by CLIP. Thus, $\mathcal{V}$ determines how many samples are added for each class, while CLIP margin ranking determines which samples are selected. The reference and expansion sets therefore need not contain the same sample indices, and no expansion sample is selected when $b_c = 0$.

**Stage III: MLLM Adjudication and Residual Targets.** CRA uses Qwen2-VL-7B-Instruct with deterministic decoding for residual adjudication. For each residual sample, the first-pass prompt provides the target image together with the CLIP-derived candidate class names:

```
Classify the main object in this target-domain image.
Choose only one label from the candidates below, or ABSTAIN
when the image does not visibly distinguish them.
Candidates:
{candidate list}
Return exactly one JSON object: {"label":"<candidate or
ABSTAIN>"}. Do not add any other text.
```

A valid first-pass prediction is accepted directly. Samples returning `ABSTAIN` or an invalid response enter a second constrained recovery query using the same candidate set:

```
Make the final closed-set decision for a difficult
target-domain image. Earlier systems could not make a
reliable decision, so inspect the image again and compare
every candidate using visible object shape, parts, material,
markings, and plausible function. You must choose the
single best-supported candidate even when the evidence is
imperfect; ABSTAIN is not allowed. Do not assume that the
first candidate is correct.
Candidates:
{candidate list}
Return exactly one JSON object: {"label":"<one
candidate>","evidence":"one short contrastive
visible-evidence phrase","certainty":"high|medium|low"}. Do
not add any other text.
```

For both passes, the returned label is normalized and checked for membership in the candidate set. Missing or malformed JSON, out-of-candidate labels, and abstentions are treated as undecided. Outputs that remain invalid after recovery retain this state. The generated `certainty` field is not used for supervision weighting.

Residual targets follow a confidence-aware hierarchy. A student prediction with maximum class probability at least 0.90 on the weakly augmented view takes priority. Otherwise, CRA uses a valid first-pass MLLM prediction, followed by a valid recovered prediction. Samples that remain

undecided use the student prediction restricted to the fallback candidate set. High-confidence student predictions are not subject to this restriction.

Supervision weights are determined by the adjudication state: valid first-pass decisions receive a weight of $1.0$, recovered decisions receive $\eta = 0.10$, and still-undecided samples receive $0.30$. These weights are retained even when a high-confidence student prediction supplies the final target.

For undecided samples, DINOv2 neighborhood voting may additionally provide a structural proposal using the same minimum-neighbor and support criteria described above. A valid proposal must differ from the CLIP top-1 prediction and lie within the query's CLIP top-three classes. Since the default candidate set contains the CLIP top-five classes, this proposal does not enlarge the candidate label space, but instead provides auxiliary structural evidence within the existing CLIP-constrained decision space.

**Algorithm 1** Execution pipeline of CRA

1: Initialize the target student from the source model and extract frozen CLIP and DINOv2 features for all target samples.
2: Construct the anchor set $\mathcal{S}_1$ using class-wise CLIP confidence and DINOv2 neighborhood support.
3: Train the student on anchor supervision.
4: Construct the reference set $\mathcal{V}$, derive class-wise budgets $b_c$, and select expansion samples $\mathcal{S}_2$; set $\mathcal{B} = \mathcal{S}_1 \cup \mathcal{S}_2$.
5: Train the student on the expanded base supervision $\mathcal{B}$ with anchor replay.
6: Form the residual set $\mathcal{R} = \mathcal{I}_T \setminus \mathcal{B}$ and construct CLIP top-$q$ candidate sets for residual samples.
7: Apply candidate-constrained MLLM adjudication and recovery to obtain weighted residual training targets.
8: Jointly optimize the student using base supervision from $\mathcal{B}$ and residual supervision from $\mathcal{R}$; retain only the adapted student for inference.

### A.3 Evaluation Datasets

We provide the dataset statistics and evaluation protocols for the three benchmarks used in our experiments.

**Office-Home.** Office-Home (Venkateswara et al., 2017) contains 15,588 images from 65 object categories across four domains: Art (A), Clipart (C), Product (P), and Real-World (R). We evaluate all twelve directed transfer tasks between distinct domains.

**DomainNet-126.** DomainNet-126 is a 126-class subset of DomainNet (Peng et al., 2019), containing 142,334 images across four domains: Clipart (C), Painting (P), Real (R), and Sketch (S). Following PLUE (Litrico et al., 2023), we use this four-domain subset and evaluate all twelve directed transfer tasks.

**VisDA-C.** VisDA-C (Peng et al., 2017) is a synthetic-to-real classification benchmark with 12 object categories. The source training split contains 152,397 synthetic images rendered from 3D models, while the target validation split contains 55,388 real images cropped from MS COCO. We evaluate the single synthetic-to-real transfer using the validation split as the target domain.

**Evaluation Metrics.** We report classification accuracy for Office-Home and DomainNet-126, with the benchmark average computed over the twelve transfer tasks. For VisDA-C, we report mean-class accuracy, defined as the unweighted mean of the 12 per-class accuracies. Detailed class-wise VisDA-C results are provided in Appendix B.1. All results are reported as percentages.

### A.4 Task Settings and Additional Details

**Task Settings.** Source-free domain adaptation adapts a source-trained model to an unlabeled target domain without access to source data. Let $\mathcal{Y}_s$ and $\mathcal{Y}_t$ denote the source and target label spaces, respectively. Closed-set SFDA assumes $\mathcal{Y}_s = \mathcal{Y}_t$, so all target samples belong to source-known classes. Partial-set SFDA assumes $\mathcal{Y}_t \subsetneq \mathcal{Y}_s$: some source classes are absent from the target domain,

and the actual target-class subset is not provided during adaptation. Open-set SFDA assumes $\mathcal{Y}_s \subsetneq \mathcal{Y}_t$: the target domain additionally contains target-private classes, which are grouped into a single unknown class for evaluation. Thus, partial-set adaptation must avoid assigning target samples to source-private classes, whereas open-set adaptation must recognize known classes while rejecting target-private samples as unknown.

**Partial-set Training Details.** For Office-Home partial-set experiments, the source domain contains all 65 classes, while the target domain contains the first 25 classes in alphabetical order. This split is used only to construct the benchmark; the target-class list is not supplied to the adaptation procedure. The target model retains the complete 65-class classifier, and CLIP predictions and MLLM candidate construction use the full source vocabulary. Predictions are not masked using the target-class subset at evaluation. CRA follows the same cumulative three-stage training procedure as in the closed-set setting: expansion samples are introduced alongside the original anchors, followed by joint training with base and residual supervision. Class-wise expansion budgets are estimated from anchor-supported target-domain structure, without an additional partial-set-specific class-filtering rule or loss. Residual adjudication retains the CLIP top-5 candidate constraint and the closed-set forced-recovery procedure.

**Open-set Training Adjustments.** In the Office-Home open-set setting, the source model is trained on 25 known classes, while the target domain contains all 65 classes, with the remaining 40 classes grouped as unknown during evaluation. The target model retains a 25-class classifier, and CLIP and the MLLM receive only the known-class vocabulary, without unknown-class names. The first two stages follow the closed-set anchoring and balanced expansion procedures. In Stage III, MLLM adjudication additionally distinguishes candidate omission, unknown objects (`UNKNOWN`), and insufficient evidence (`ABSTAIN`). Only candidate omission triggers a second verification query, which expands the top-5 shortlist to the complete known-class vocabulary. This query still permits unknown or abstained responses rather than imposing the forced choice used in the closed-set setting. Only accepted known-class labels provide residual supervision; other responses do not generate known-class pseudo-labels.

During final training, the current student's predictive entropy over all target samples is partitioned into two clusters and the resulting selection is periodically refreshed. Only samples in the low-entropy cluster contribute to known-class supervision from the base and residual sets, and previously excluded samples may re-enter after a later refresh. An additional unlabeled batch is sampled uniformly from the entire target domain for information maximization: conditional entropy minimization applies only to the currently selected known samples, whereas prediction diversity uses all samples in this batch. At test time, entropy clustering is recomputed using the final model; the high-entropy cluster is assigned to the unknown class, while the remaining samples receive the highest-probability known-class prediction. Neither training-time selection nor unknown-class rejection uses target ground-truth labels.

# B Supplementary Experiments and Diagnostics

## B.1 Additional Class-wise Results

Table S3 reports the class-wise accuracy on VisDA-C, where Avg. denotes the mean class accuracy over the 12 categories. CRA results are averaged over three independent training runs with seeds 2026, 2027, and 2028, following the same reporting protocol as in the main text. Results for competing methods are reported according to their original papers. CRA achieves the best performance on the bus category and the second-best overall mean-class accuracy, while remaining competitive across most categories.

## B.2 Full Partial-set and Open-set Results

Table S4 complements the average results in Table 3 with results on all twelve Office-Home transfer tasks. The upper part reports classification accuracy for partial-set SFDA, while the lower part reports OS for open-set SFDA. CRA achieves the best performance on all twelve transfers in both settings, with average scores of 94.0% and 89.2%, outperforming the strongest baselines, DIFO++ and ProDe, by 7.3 and 6.6 percentage points, respectively. The improvements are particularly pronounced when Clipart is the target domain: across A→C, P→C, and R→C, CRA exceeds

Table S3: Class-wise accuracy (%) on VisDA-C Synthetic→Real. CRA results are averaged over three independent training seeds.

| Method | plane | bcycl | bus | car | horse | knife | mcycl | person | plant | sktbrd | train | truck | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SHOT (Liang et al., 2020) | 95.0 | 87.4 | 80.9 | 57.6 | 93.9 | 94.1 | 79.4 | 80.4 | 90.9 | 89.8 | 85.8 | 57.5 | 82.7 |
| TPDS (Tang et al., 2024a) | 97.6 | 91.5 | 89.7 | 83.4 | 97.5 | 96.3 | 92.2 | 82.4 | 96.0 | 94.1 | 90.9 | 40.4 | 87.6 |
| UCon-SFDA (Xu et al., 2025) | 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 |
| DIFO (Tang et al., 2024b) | 97.5 | 89.0 | 90.8 | 83.5 | 97.8 | 97.3 | 93.2 | 83.5 | 95.2 | 96.8 | 93.7 | 65.9 | 90.3 |
| ProDe (Tang et al., 2025) | 96.6 | 90.3 | 83.9 | 80.2 | 96.1 | 96.9 | 90.3 | 86.4 | 90.8 | 94.0 | 91.3 | 67.0 | 88.7 |
| DIFO++ (Tang et al., 2026) | 98.0 | 91.3 | 89.2 | 79.1 | 96.4 | **98.6** | 92.6 | **88.3** | 93.2 | 92.9 | 94.2 | **71.7** | 90.5 |
| RCL (Chen et al., 2026) | **99.5** | **96.1** | 92.6 | **89.4** | **99.1** | 97.1 | **97.0** | 85.8 | 96.6 | **98.1** | **97.3** | 70.0 | **93.2** |
| **CRA (Ours)** | 99.3 | 93.4 | **93.7** | 83.6 | 98.9 | 97.6 | 96.0 | 88.2 | 90.4 | 96.1 | 95.6 | 69.7 | 91.9 |

the strongest baseline for each transfer by 13.4–15.6 points under partial-set SFDA and 9.0–11.1 points under open-set SFDA. These results demonstrate that CRA remains effective across different source–target label-space mismatches and challenging domain shifts.

Table S4: Full partial-set (top) and open-set (bottom) SFDA results (%) on Office-Home. Partial-set results report classification accuracy, whereas open-set results report OS. Avg. denotes the reported average over 12 transfers. FM indicates external foundation models used for adaptation. The best results are shown in bold and the second-best results are underlined.

| **Partial-set** | FM | 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 | ✗ | 45.2 | 70.4 | 81.0 | 56.2 | 60.8 | 66.2 | 60.9 | 40.1 | 76.2 | 70.8 | 48.5 | 77.3 | 62.8 |
| SHOT (Liang et al., 2020) | ✗ | 64.8 | 85.2 | 92.7 | 76.3 | 77.6 | 88.8 | 79.7 | 64.3 | 89.5 | 80.6 | 66.4 | 85.8 | 79.3 |
| HCL (Huang et al., 2021) | ✗ | 65.6 | 85.2 | 92.7 | 77.3 | 76.2 | 87.2 | 78.2 | 66.0 | 89.1 | 81.5 | 68.4 | 87.3 | 79.6 |
| CoWA (Lee et al., 2022) | ✗ | 69.6 | 93.2 | 92.3 | 78.9 | 81.3 | 92.1 | 79.8 | 71.7 | 90.0 | 83.8 | 72.2 | 93.7 | 83.2 |
| CRS (Zhang et al., 2023) | ✗ | 68.6 | 85.1 | 90.9 | 80.1 | 79.4 | 86.3 | 79.2 | 66.1 | 90.5 | 82.2 | 69.5 | 89.3 | 80.6 |
| DIFO (Tang et al., 2024b) | ✓ | 70.2 | 91.7 | 91.5 | 87.8 | 92.6 | 92.9 | 87.3 | 70.7 | 92.9 | 88.5 | 69.6 | 91.5 | 85.6 |
| ProDe (Tang et al., 2025) | ✓ | 70.2 | 89.7 | 90.4 | 84.1 | 90.7 | 91.4 | 85.5 | 69.9 | 92.9 | 87.8 | 68.5 | 89.7 | 84.2 |
| DIFO++ (Tang et al., 2026) | ✓ | 75.9 | 90.2 | 90.9 | 88.4 | 91.8 | 92.5 | 87.5 | 75.6 | 93.3 | 89.5 | 74.3 | 90.8 | 86.7 |
| **CRA (Ours)** | ✓ | **89.3** | **96.4** | **96.5** | **93.4** | **96.6** | **96.5** | **93.8** | **89.8** | **96.5** | **93.6** | **89.9** | **96.4** | **94.0** |
| **Open-set** | FM | 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 | ✗ | 36.3 | 54.8 | 69.1 | 33.8 | 44.4 | 49.2 | 36.8 | 29.2 | 56.8 | 51.4 | 35.1 | 62.3 | 46.6 |
| SHOT (Liang et al., 2020) | ✗ | 64.5 | 80.4 | 84.7 | 63.1 | 75.4 | 81.2 | 65.3 | 59.3 | 83.3 | 69.6 | 64.6 | 82.3 | 72.8 |
| HCL (Huang et al., 2021) | ✗ | 64.0 | 78.6 | 82.4 | 64.5 | 73.1 | 80.1 | 64.8 | 59.8 | 75.3 | 78.1 | 69.3 | 81.5 | 72.6 |
| CoWA (Lee et al., 2022) | ✗ | 63.3 | 79.2 | 85.4 | 67.6 | 83.6 | 82.0 | 66.9 | 56.9 | 81.1 | 68.5 | 57.9 | 85.9 | 73.2 |
| CRS (Zhang et al., 2023) | ✗ | 65.2 | 76.6 | 80.2 | 66.2 | 75.3 | 77.8 | 70.4 | 61.8 | 79.3 | 71.1 | 61.1 | 78.3 | 73.2 |
| DIFO (Tang et al., 2024b) | ✓ | 64.5 | 86.2 | 87.9 | 68.2 | 79.3 | 86.1 | 67.2 | 62.1 | 88.3 | 71.9 | 65.3 | 84.4 | 75.9 |
| ProDe (Tang et al., 2025) | ✓ | 75.9 | 85.6 | 87.9 | 81.3 | 86.8 | 87.2 | 81.1 | 74.3 | 86.3 | 83.0 | 75.7 | 86.1 | 82.6 |
| DIFO++ (Tang et al., 2026) | ✓ | 63.0 | 87.5 | 87.0 | 69.6 | 81.0 | 87.1 | 68.0 | 64.0 | 89.1 | 72.9 | 66.4 | 88.0 | 77.0 |
| **CRA (Ours)** | ✓ | **84.9** | **92.2** | **94.2** | **84.9** | **92.5** | **93.8** | **84.9** | **85.4** | **94.2** | **86.1** | **84.8** | **92.7** | **89.2** |

### B.3 Additional Analysis of Global Supervision

**Global Supervision Quality and Student Adaptation.** Table S5 further analyzes the relationship between global supervision quality and downstream student adaptation on Office-Home A→C. We use Qwen2-VL-7B-Instruct as the MLLM and compare its valid first-pass predictions with supervision augmented by forced recovery. We also include global CLIP supervision and DINOv2-assisted CLIP supervision, where DINOv2 neighborhood structure is combined with CLIP semantic evidence. Coverage denotes the fraction of target samples receiving valid supervision, Selective Acc. measures prediction accuracy among valid outputs, and Correct Yield denotes the fraction of all target samples receiving a valid and correct label. Student Acc. is evaluated on the full target set. All metrics are reported as percentages.

Forced recovery increases MLLM coverage from 96.8% to 99.1% and slightly improves the correct-label yield from 84.6% to 84.8%, but does not improve the adapted student. More importantly, although global MLLM supervision achieves substantially higher output accuracy than CLIP-based

Table S5: Supervision output quality and student adaptation under global-supervision strategies on Office-Home A→C.

| Predictor / policy | Coverage | Selective Acc. | Correct Yield | Student Acc. |
|---|---|---|---|---|
| CLIP | 100.0 | 75.1 | 75.1 | 75.1 |
| DINOv2-assisted CLIP | 100.0 | 72.0 | 72.0 | 71.5 |
| Global MLLM (first pass) | 96.8 | **87.4** | 84.6 | 81.2 |
| Global MLLM (+ recovery) | 99.1 | 85.5 | **84.8** | 81.1 |
| **CRA (Ours)** | — | — | — | **84.3** |

alternatives, the resulting student still underperforms CRA. These results indicate that aggregate supervision quality and coverage alone do not determine adaptation performance, supporting our central finding that the value of additional supervision also depends on where it is applied.

**Global Prediction Statistics across MLLMs.** Table S6 reports the global prediction quality of five MLLMs on the same 4,365 target samples from Office-Home A→C. First-pass adjudication allows abstention, whereas recovery re-queries only unresolved samples using the same candidate set while preserving all valid first-pass decisions. We report coverage, selective accuracy over valid outputs, and correct yield over the full target set.

Recovery consistently increases valid-output coverage, but the newly recovered predictions are generally less reliable, leading to lower selective accuracy and only limited gains in correct yield. These results characterize the global predictive quality of different MLLMs, but do not determine where their supervision is most useful. Together with the supervision-position analysis, they show that strong global prediction performance should not be equated with high incremental supervisory value on every target sample.

Table S6: Prediction statistics of five MLLMs on Office-Home A→C before and after recovery. CLIP is included as a reference. All results are reported as percentages.

| | First pass | | | After recovery | | |
|---|---|---|---|---|---|---|
| Model | Coverage | Selective Acc. | Correct Yield | Coverage | Selective Acc. | Correct Yield |
| CLIP | 100.0 | 75.1 | 75.1 | 100.0 | 75.1 | 75.1 |
| Qwen2-VL-7B-Instruct | 96.8 | 87.4 | 84.6 | 99.1 | 85.5 | 84.8 |
| InternVL2.5-8B | 94.1 | 83.6 | 78.7 | 100.0 | 79.2 | 79.2 |
| Phi-3.5-Vision-Instruct | 98.6 | 80.2 | 79.0 | 100.0 | 79.2 | 79.2 |
| LLaVA-1.6-34B-AWQ | 89.2 | 84.5 | 75.3 | 100.0 | 78.2 | 78.2 |
| ShareGPT4V-7B | 99.2 | 75.8 | 75.2 | 99.8 | 75.3 | 75.2 |

### B.4 Supervision Position and Allocation Controls

**Cross-Transfer Supervision-Position Comparison.** Table S7 extends the main-text supervision-position analysis with per-transfer results and effective label budgets. For each transfer, supervision within and beyond existing reliable coverage uses the same number of valid MLLM labels and follows the same student-training protocol. Different transfers use their respective matched budgets, where $M$ denotes the number of accepted effective labels. Across all five transfers, allocating additional MLLM supervision beyond existing reliable coverage consistently yields higher student performance than allocating the same effective supervision budget within the reliable region. This result further supports the supervision-position effect across different domain shifts.

**Additional-Supervision Allocation Strategies.** Table S8 compares CRA with two alternative allocation strategies: random allocation and frozen-source uncertainty. The latter ranks target samples by the predictive entropy of the frozen source model and prioritizes samples with higher uncertainty. All strategies use the same full-target MLLM output pool, matched effective label budgets, and identical student-training protocols. CRA outperforms both random and source-uncertainty-based allocation on the two evaluated transfers. These results indicate that allocating additional supervision according to existing reliable coverage is more effective than either unguided allocation or uncertainty ranking based solely on the frozen source model.

Table S7: Cross-transfer supervision-position comparison under matched effective label budgets. $M$ denotes the number of accepted labels at each position. Student accuracy is reported as a percentage.

| Dataset / transfer | $M$ | Within reliable coverage | Beyond reliable coverage |
|---|---|---|---|
| Office-Home A→C | 2,033 | 77.0 | **84.3** |
| Office-Home P→C | 2,033 | 77.8 | **84.2** |
| Office-Home R→P | 900 | 93.4 | **96.3** |
| DomainNet R→S | 9,308 | 80.8 | **87.1** |
| DomainNet P→R | 14,615 | 91.1 | **93.5** |

Table S8: Comparison of additional-supervision allocation strategies on Office-Home under matched effective label budgets and training protocols. Student accuracy is reported as a percentage.

| Allocation strategy | A→C | R→P |
|---|---|---|
| Random allocation | 80.5 | 94.0 |
| Frozen-source uncertainty | 81.7 | 94.9 |
| **CRA residual allocation** | **84.3** | **96.3** |

### B.5 Additional Ablations on Reliable Supervision Construction

**Label Sources and Anchor Selection.** Table S9 further examines the construction of reliable supervision through two groups of controls. The first varies the label source for the base supervision set by replacing CLIP labels with Qwen2-VL labels. The second evaluates anchor selection by comparing CRA's semantic–structural criterion with CLIP-margin selection, low-entropy selection based on the student model, and random class-matched selection. For each comparison, the supervision-set size and student-training protocol are kept fixed, while the anchor-selection controls additionally match the number of selected samples per class.

Replacing CLIP labels in the base supervision set with Qwen2-VL labels yields only a limited improvement, suggesting that extending MLLM supervision throughout the established reliable region provides little additional benefit. Meanwhile, CRA's semantic–structural anchor selection performs comparably to CLIP-margin selection while providing a principled combination of semantic confidence and target-domain structural support. Overall, these results support using lightweight foundation-model supervision to construct the reliable base and reserving additional MLLM supervision for samples outside this established coverage.

### B.6 Parameter Sensitivity and Candidate-Space Analysis

**Effect of Candidate-Space Size.** We next study the effect of the candidate-set size $q$ on residual supervision. Table S10 compares candidate recall, MLLM label accuracy, and downstream student accuracy for different values of $q$. We fix the upstream supervision construction and the resulting 2,067 residual samples on Office-Home A→C, while varying the candidate space used for MLLM adjudication. Each setting uses 3,000 student-training steps and three training seeds. Candidate recall measures the fraction of residual samples whose ground-truth class is contained in the candidate set. It is used only as a post-hoc evaluation metric; ground-truth labels are never used to construct candidate sets or generate supervision.

Within the tested range, increasing the candidate-set size improves candidate recall, MLLM label accuracy, and downstream student performance. This result shows that candidate-space coverage directly affects the corrective potential of residual MLLM supervision. We use $q = 5$ and $\eta = 0.10$ in the main experiments, while alternative settings are reported here for completeness.

### B.7 Supervision-Request Statistics

We further analyze the amount of MLLM supervision required by different strategies on Office-Home A→C. Global selective supervision queries the MLLM for all target samples while allowing abstention, whereas global forced supervision additionally applies recovery to unresolved predictions. In contrast, CRA requests MLLM supervision only for residual samples outside the estab-

Table S9: Ablations on label sources and anchor selection for reliable supervision construction. The two groups vary the base-set label source and the Stage I anchor-selection strategy, respectively. Student accuracy is reported as a percentage.

| Variant | Student Acc. |
|---|---|
| **(a) Base-set label source** | |
| CRA (Ours): CLIP labels | 84.3 |
| Qwen2-VL-7B replaces CLIP labels | **84.6** |
| **(b) Stage I anchor selector** | |
| CRA (Ours): DINOv2 confirmation | **84.3** |
| CLIP-margin selector | **84.3** |
| Student low-entropy selector | 83.7 |
| Random class-matched selector | 83.0 |

Table S10: Effect of candidate-set size on residual supervision for Office-Home A→C. Candidate recall, MLLM label accuracy, and student accuracy are reported as percentages. Student results are averaged over three training seeds.

| Top-$q$ | Candidate Recall | MLLM Label Acc. | Student Acc. |
|---|---|---|---|
| 3 | 75.8 | 71.8 | 82.7 |
| 5 | 81.6 | 75.6 | 84.3 |
| 7 | 85.3 | 78.5 | 85.6 |
| 9 | **87.5** | **80.0** | **85.9** |

lished reliable coverage, with recovery applied only when necessary. Figure S1 compares the resulting supervision requests and student accuracy.

We count the number of MLLM judgments required by each strategy, regardless of whether the corresponding outputs are newly computed or retrieved from cache. This metric therefore reflects supervision demand rather than actual model forward passes, wall-clock time, or monetary cost. CRA achieves higher student accuracy with fewer MLLM supervision requests than the global alternatives, showing that restricting additional supervision to residual samples improves its effectiveness while reducing supervision demand.

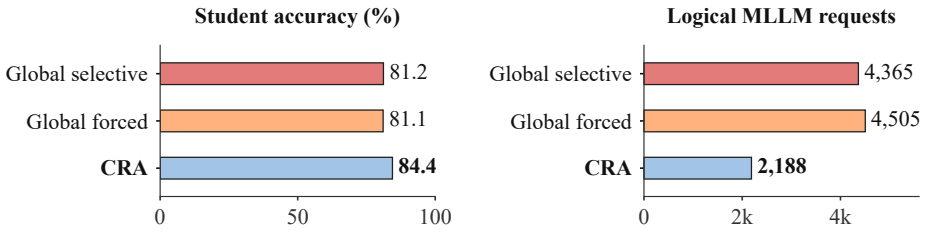


Figure S1: Logical MLLM requests versus student performance on A→C.

### B.8 Class Coverage and Grad-CAM Settings

**Class Coverage.** We additionally examine whether CRA's constructed supervision is concentrated on only a subset of target classes. On Office-Home with Art as the target domain, the supervision set covers 64 of the 65 classes, with only *Marker* uncovered. This statistic is computed post hoc for analysis and is not used to modify supervision construction or training.

**Grad-CAM Settings.** Grad-CAM examples are selected post hoc from the fixed residual set, requiring both the Source model and DIFO++ to make incorrect predictions while CRA predicts correctly. For a fair comparison, all methods use the same input image and crop, the final convolutional layer of ResNet-50, and the same target-class logit to compute the activation maps. The figure presents seven samples, with rows corresponding to the input images, Source, DIFO++, the CRA model after anchoring and expansion, and the final CRA model. In these examples, CRA places

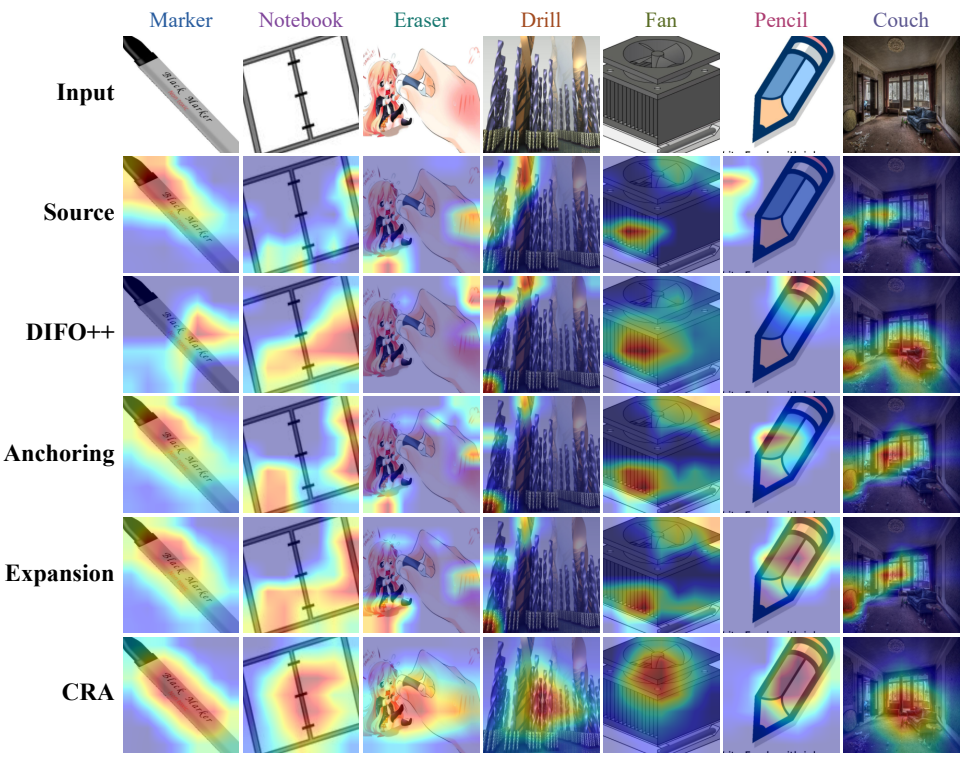


Figure S2: Grad-CAM comparisons and stage-wise visualizations on seven selected residual samples. Columns correspond to the indicated object categories. Rows show the input images, Source, DIFO++, the CRA model after anchoring and expansion, and the final CRA model, respectively. Each activation map is normalized independently, so response locations rather than absolute intensities should be compared.

greater emphasis on regions associated with the target object, whereas the baseline models often respond more strongly to less relevant regions.

## C LIMITATIONS

CRA has two main limitations. First, its effectiveness depends on the quality of supervision provided by the underlying foundation models. Errors in semantic confidence or structural support may lead to imperfect estimates of reliable coverage, while the CLIP-derived candidate space may exclude the correct class and thereby limit subsequent MLLM correction. These issues can become more pronounced when the target domain differs substantially from the pretraining distributions of the foundation models.

Second, CRA introduces additional computation and memory overhead during adaptation because it relies on multiple frozen foundation models and caches their predictions or features. Although CRA reduces unnecessary MLLM requests through selective residual adjudication and deploys only the adapted student at inference time, its adaptation-time cost remains higher than that of SFDA methods relying solely on the source model.