# Safeguarding Mutual Correction in Source-Free Domain Adaptation via Cut Statistics

**Seongjun Lee**
Department of Artificial Intelligence
Korea University
Seoul, Republic of Korea
pyoung7307@korea.ac.kr

**Changhee Lee***
Department of Artificial Intelligence
Korea University
Seoul, Republic of Korea
changheelee@korea.ac.kr

## Abstract

Source-Free Domain Adaptation (SFDA) aims to adapt a source-pretrained model to an unlabeled target domain without access to the original source domain. While early single-model approaches rely on self-refinement, they are inherently susceptible to confirmation bias and struggle to correct their own systematic errors. To overcome this limitation, recent methods introduce Vision-Language (ViL) models as external knowledge sources. However, these approaches operate in a largely unidirectional paradigm, using the ViL model primarily to supervise the source-pretrained model. This overlooks a key structural property: the two models exhibit distinct failure modes – where one produces an incorrect prediction, the other may produce a correct one, creating a natural opportunity for mutual correction within the target domain. Yet, without ground-truth labels, identifying which model is correct on any given sample is non-trivial, and naively exchanging predictions risks propagating errors across models. To address this challenge, we propose `SafeCut`, a novel approach that leverages the cut statistic as a label-free measure of prediction reliability to gate cross-model supervision. Our approach dynamically controls both the direction and strength of supervision based on relative reliability, selectively amplifying true corrections while suppressing miscorrections on a per-sample basis. We further provide theoretical justification showing that this reliability-gated mechanism guarantees a net-positive correction signal. Extensive experiments across diverse SFDA benchmarks demonstrate that `SafeCut` achieves state-of-the-art performance, highlighting the effectiveness of safeguarding mutual correction in SFDA via cut statistics. Codes are available here.

## 1 Introduction

One of the central challenges in deploying deep learning models is the presence of distribution shift between training and deployment environments. Unsupervised Domain Adaptation (UDA) addresses this issue by aligning source and target distributions without relying on target labels [1, 2]. However, it requires access to source data – an assumption often violated in privacy- and regulation-sensitive domains such as healthcare and finance. Source-Free Domain Adaptation (SFDA) has therefore emerged as a practical alternative, operating under the *stricter constraint* that only a source-pretrained model and unlabeled target data are accessible, while both source data and target labels are *completely inaccessible* [3].

SFDA assumes that, despite distribution shifts, the source and target domains share an underlying semantic structure, allowing knowledge learned from the source domain to remain partially transferable. Early approaches capitalize on this shared structure by leveraging the structural proximity

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

of predictions or latent features induced by the source-pretrained model to derive self-supervised adaptation signals on the target domain [3–8]. However, because these single-model approaches rely exclusively on self-refinement, they are highly susceptible to confirmation bias and inherently struggle to correct their own systematic errors, as incorrect predictions are often distributed throughout the target domain, making them difficult to correct through local proximity or consistency-based supervision [8, 9].

Recent work introduces Vision-Language (ViL) models as an external knowledge source, enabling cross-model supervision beyond self-training [10, 11]. Their rich, generalized representations provide corrective signals in regions where the source-pretrained model is uncertain or biased. Despite these advantages, existing approaches treat ViL as a universally more reliable knowledge source. In practice, however, ViL models may have distinct failure modes from the source-pretrained models – cases where one model errs while the other is correct – and their relative reliability varies across samples and target domains. Unconditionally accepting ViL predictions ignores this structural variability, which limits potential gains from adaptation. Moreover, such one-sided reliance can hinder generalization in domains where ViL models are inherently less reliable, such as medical imaging [12, 13]. More fundamentally, it overlooks the opportunity to leverage these distinct failure modes for mutual correction.

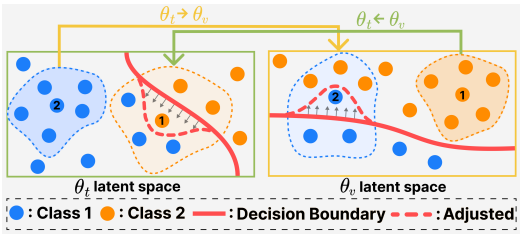



Figure 1: **Reliability-based Mutual Correction.** The model with lower cut corrects the other (Sample 1: $\theta_v \rightarrow \theta_t$, Sample 2: $\theta_t \rightarrow \theta_v$).

To address this, we propose a novel approach, **Safe**guarding mutual correction via **Cut** statistics (`SafeCut`), which establishes a positive feedback loop through *reliability-gated knowledge transfer* on the target domain. Using the *cut statistic* [14] as a label-free measure of local prediction consistency, our method dynamically controls both the direction and strength of cross-model supervision (Figure 1). This selectively amplifies corrective signals while suppressing miscorrections, enabling the two models to refine each other in a principled manner. As a result, `SafeCut` achieves state-of-the-art performance across diverse domain adaptation benchmarks.

**Our contributions are summarized as follows:**

- We provide extensive empirical evidence that the cut statistic is strongly and negatively correlated with prediction accuracy in domain adaptation settings, and further establish a theoretical foundation based on cut margins to exploit different failure modes.
- Building on this insight, we propose `SafeCut`, a reliability-gated method that leverages the cut statistic to dynamically control both the direction and strength of cross-model supervision.
- We demonstrate that `SafeCut` achieves state-of-the-art performance across diverse domain adaptation benchmarks, highlighting the effectiveness of exploiting complementary failure modes for SFDA.

## 2 Related Works

**Source-Free Domain Adaptation.** Early work aligns target representations via self-supervised pseudo-labeling [3], and subsequent methods exploit target neighborhood structure for prediction consistency, pseudo-label refinement, and contrastive adaptation [4–6]. Alternatively, several approaches estimate prediction reliability via GMM-based confidence or neighborhood uncertainty, and subsequently use these quantities for self-supervision [7, 8]. However, by relying solely on the source model's predictions and the resulting latent structures, these methods are inherently constrained by the model's own blind spots and confirmation bias. To address this, recent work introduces ViL as an auxiliary knowledge source, focusing on adapting ViL to the target domain or denoising its pseudo-labels [10, 11]. These methods are specifically tailored to the zero-shot capabilities of ViL and do not explore their extension to more general counterpart models; furthermore, they leave the critical question of how effective knowledge transfer can occur between the two models largely unaddressed.

**Cross-Model Supervision.** Cross-model supervision has been explored in mutual learning [15] and co-teaching [16] frameworks, where two models iteratively refine each other – either by exchanging predictions or by selecting reliable samples for the peer to learn from. However, these approaches fundamentally rely on labeled data or noisy label supervision, making them inapplicable to the fully unsupervised SFDA setting. More recently, weak-to-strong generalization [17, 18] demonstrates that a weaker model's imperfect supervision can nonetheless elicit strong generalization from a more capable peer. `SafeCut` builds on this insight and extends it to the label-free SFDA setting, leveraging the cut statistic to dynamically route supervision based on local prediction consistency — enabling reliable cross-model correction without any label information.

# 3 Mutual Correction in Source-free Domain Adaptation

## 3.1 Problem Setting

We consider a $C$-class classification task over the input space $\mathcal{X}$ and the label space $\mathcal{Y} = \{1, \ldots, C\}$. In the *source-free domain adaptation* (SFDA) setting, we are provided with a source model $\theta_s : \mathcal{X} \to \mathbb{R}^C$ pre-trained on a labeled source domain $\mathcal{D}_s$. Our objective is to adapt this model to an unlabeled target domain $\mathcal{D}_t = \{x_j\}_{j=1}^n$, whose underlying distribution differs from that of the source. Following the standard SFDA setting, the source data $\mathcal{D}_s$ is *strictly inaccessible* during adaptation, due to data privacy or storage limitations. We initialize our target model $\theta_t$ with the source parameters (i.e., $\theta_t = \theta_s$) and must bridge the domain gap using only the unlabeled samples in $\mathcal{D}_t$. The core challenge lies in recovering the target discriminative structure without the supervision of the source distribution or ground-truth target labels.

We further introduce a *peer (counterpart) model* $\theta_v$ designed to provide *complementary correctness patterns*. Specifically, the two models (i.e., $\theta_t$ and $\theta_v$) are assumed to make errors on different subsets of the target domain: there exist non-empty disagreement regions where only one of the two models is correct. This functional asymmetry enables *mutual correction*, whereby each model can provide informative supervision to its peer on samples where its own prediction is correct and the other's is not. When applied iteratively, such cross-model supervision forms a positive reinforcing feedback loop, in which progressively refined predictions from one model serve as higher-fidelity pseudo-labels for the other, ultimately improving both models beyond their initial performance [15, 17, 18].

**Dynamics of Mutual Correction.** To formalize the dynamics of mutual correction, we partition the target data based on the correctness of $\theta_t$ and $\theta_v$. For sample $x$, let $y(x) \in \mathcal{Y}$ denote the (unknown) ground-truth label and $\hat{y}_i(x)$ the predicted label from model $\theta_i$ with $i \in \{t, v\}$. We define the set of correct predictions for model $\theta_i$ as $\mathcal{C}_i = \{x \in \mathcal{D}_t : \hat{y}_i(x) = y(x)\}$ and the set of incorrect predictions as its complement, $\overline{\mathcal{C}}_i = \mathcal{D}_t \setminus \mathcal{C}_i$. The intersection of these sets induces a partition of the target domain into four mutually exclusive (unobservable) subsets:

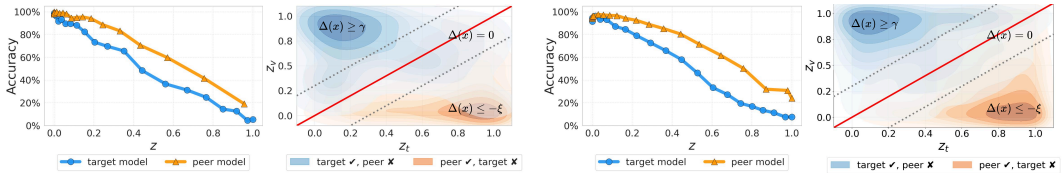
$$\mathcal{S}_{cc} = \mathcal{C}_t \cap \mathcal{C}_v, \quad \mathcal{S}_{c\bar{c}} = \mathcal{C}_t \cap \overline{\mathcal{C}}_v, \quad \mathcal{S}_{\bar{c}c} = \overline{\mathcal{C}}_t \cap \mathcal{C}_v, \quad \mathcal{S}_{\bar{c}\bar{c}} = \overline{\mathcal{C}}_t \cap \overline{\mathcal{C}}_v, \tag{1}$$

where each subset corresponds to a distinct joint correctness pattern of the two models.

Due to the absence of ground-truth labels in SFDA, mutual correction relies on each model using the other's predictions as pseudo-labels. This cross-model supervision induces two competing dynamics:

- **Correction:** This occurs when a model receives a correct pseudo-label from its peer that can rectify its own error. Specifically, for samples in $\mathcal{S}_{\bar{c}c}$, the peer model $\theta_v$ provides correct supervision to $\theta_t$; symmetrically, $\theta_t$ can correct $\theta_v$ for samples in $\mathcal{S}_{c\bar{c}}$. Incorporating such signals directly improves the receiving model's discriminative power.
- **Miscorrection:** This occurs when an incorrect prediction is used as supervision, potentially corrupting a model that is already correct. For instance, if $\theta_v$ supervises $\theta_t$ on $\mathcal{S}_{c\bar{c}}$ (or vice versa on $\mathcal{S}_{\bar{c}c}$), the peer model provides an erroneous pseudo-label. Accepting such signals propagates errors, leading to confirmation bias and a degradation of performance.

The ideal strategy would accept supervision exclusively on *correctly predicted* samples – i.e., employing $\theta_v$ as a supervisor only on $\mathcal{S}_{\bar{c}c}$ and $\theta_t$ only on $\mathcal{S}_{c\bar{c}}$ – thereby maximizing correction while avoiding miscorrection. However, this *oracle* selection is infeasible in practice: since ground-truth labels are *not available*, there is no direct way to determine which model is correct for a given sample. The central challenge is therefore to approximate this selection in a *fully unsupervised* manner, so that correction signals are amplified while miscorrection signals are suppressed.




(a) Cl → Pr (Office-Home) (b) S → C (DomainNet-126)

Figure 2: **Cut statistic as a reliability proxy.** For each adaptation task, the left panel shows that prediction accuracy decreases (approximately) monotonically with the cut statistic $z$ (computed via (9)) for both models. The right panel shows the sample density in the $(z_t, z_v)$ plane for $\mathcal{S}_{c\bar{c}}$ (blue) and $\mathcal{S}_{\bar{c}c}$ (orange), which are concentrated in $\Delta(x) \geq \gamma_0$ and $\Delta(x) \leq -\xi_0$, respectively.

### 3.2 When to Rely on Which Model: A Cut Statistic Approach

To address the challenge of identifying correctly predicted samples without access to ground-truth labels, we employ the *cut statistic* [14], a graph-based measure that quantifies consistency of model predictions within a sample's local neighborhood.

**Cut Statistic.** For each target sample $x \in \mathcal{D}_t$, let $\mathcal{N}_K(x)$ denote its $K$-nearest neighbors. Then, the cut statistic for model $\theta_i$, where $i \in \{t, v\}$, is defined as:

$$z_i(x) = \sum_{x' \in \mathcal{N}_K(x)} w(x, x') \cdot \mathbb{1}[\hat{y}_i(x) \neq \hat{y}_i(x')] \tag{2}$$

where $w(x, x')$ is a distance-based weighting function that assigns greater importance to closer neighbors. Intuitively, the cut statistic quantifies the *local smoothness* of model predictions over the data graph. A low value indicates that a prediction is consistent with those of its neighbors, whereas a high value indicates that such local consistency does not hold.

Under the cluster assumption [19], locally consistent predictions are generally more likely to be correct, suggesting that the cut statistic should be negatively correlated with predictive accuracy [14, 18]. However, in domain adaptation settings – where distribution shift may distort the underlying structure – this relationship is not guaranteed to hold. To examine whether this intuition remains valid, we empirically analyze the relationship between local consistency and prediction accuracy across diverse adaptation scenarios. Specifically, for each model $\theta_i$ with $i \in \{t, v\}$, we compute $z_i(x)$ for each target sample $x \in \mathcal{D}_t$, group samples into bins according to their cut values, and measure the average accuracy within each bin. The results (Figure 2 left) reveal a clear and consistent trend across all evaluated settings, motivating the following observation (see Appendix D.4 for additional results):

**Observation 3.1.** ***(Approximate Monotonicity)*** *For both $\theta_i$ with $i \in \{t, v\}$, prediction accuracy is approximately monotonically decreasing with respect to the cut statistic $z_i$.*

**Connecting Correction Dynamics with Cut Statistics.** Based on this observation, we assume that the conditional accuracy $\mathbb{P}\big(\hat{y}_i(x) = y(x) \mid z_i(x)\big)$ is monotonically decreasing in with respect to $z_i \in [0, 1]$.[2] Since this monotonicity holds for both $\theta_t$ and $\theta_v$, the relative difference in their cut statistics can offer a principled, *label-free* criterion for determining which model provides more reliable prediction on a per-sample basis. To formalize this, we define the *cut margin* $\Delta(x) := z_v(x) - z_t(x)$ to characterize the *relative expertise* of the two models on sample $x$. The sign of $\Delta(x)$ indicates the direction of reliability: a positive value indicates that $\theta_t$ is more likely to be correct, while a negative value suggests the peer model $\theta_v$ is more reliable. Crucially, the magnitude $|\Delta(x)|$ reflects the strength of this reliability gap, and thus measures the potential fidelity of the correction signal. Building on this intuition, we approximate the unobservable oracle partitions $\mathcal{S}_{c\bar{c}}$ and $\mathcal{S}_{\bar{c}c}$ in (1) by introducing margin thresholds $\gamma, \xi \in [0, 1]$, identifying regions of the target domain where one model provides a clear reliability gap over the other:

$$\hat{\mathcal{S}}_{c\bar{c}}(\gamma) = \{x \in \mathcal{D}_t : \Delta(x) \geq \gamma\}, \quad \hat{\mathcal{S}}_{\bar{c}c}(\xi) = \{x \in \mathcal{D}_t : \Delta(x) \leq -\xi\}. \tag{3}$$

Given different failure modes of $\theta_t$ and $\theta_v$, and recalling that the monotonic decrease of each model's accuracy with its own cut statistic (Observation 3.1), the model with higher local consistency is more

[2]For notational simplicity, we will omit the dependence on $x$ in $y(x)$ and $\hat{y}_i(x)$ when it is clear from context.

likely to be correct and can thus provide high-fidelity pseudo-labels to supervise its less reliable counterpart. Notably, this intuition remains valid under approximate monotonicity; we provide theoretical bounds under noisy conditions in Appendix A.1. We formalize this intuition through the following baseline condition for effective mutual correction:

**Proposition 3.1** (**Reliability-Separated Regions**). *There exist baseline margins $\gamma_0, \xi_0$ such that:*

$$\mathbb{P}\big(\hat{y}_t = y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma_0)\big) > \mathbb{P}\big(\hat{y}_v = y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma_0)\big), \tag{4}$$

$$\mathbb{P}\big(\hat{y}_v = y \mid x \in \hat{\mathcal{S}}_{\bar{c}c}(\xi_0)\big) > \mathbb{P}\big(\hat{y}_t = y \mid x \in \hat{\mathcal{S}}_{\bar{c}c}(\xi_0)\big). \tag{5}$$

*Please refer to Appendix A.1 for the full proof of Proposition 3.1.*

Here, $\gamma_0$ and $\xi_0$ denote the minimum reliability gaps required for one model's prediction to serve as a trustworthy supervisory signal for its peer.

While Proposition 3.1 shows that the supervising model is more accurate in regions where it is locally more consistent than its peer, effective mutual correction requires a stronger guarantee: cross-model supervision must produce a *net-positive correction signal*, where corrections outnumber miscorrections. The following theorem proves that this dominance holds under the baseline margins:

**Theorem 3.1** (**Correction Dominance**). *For the baseline margins $\gamma_0, \xi_0$ satisfying Proposition 3.1:*

$$\mathbb{P}\big(\hat{y}_t = y \wedge \hat{y}_v \neq y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma_0)\big) > \mathbb{P}\big(\hat{y}_t \neq y \wedge \hat{y}_v = y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma_0)\big), \tag{6}$$

$$\mathbb{P}\big(\hat{y}_v = y \wedge \hat{y}_t \neq y \mid x \in \hat{\mathcal{S}}_{\bar{c}c}(\xi_0)\big) > \mathbb{P}\big(\hat{y}_v \neq y \wedge \hat{y}_t = y \mid x \in \hat{\mathcal{S}}_{\bar{c}c}(\xi_0)\big). \tag{7}$$

*Please refer to Appendix A.1 for the full proof of Theorem 3.1.*

Figure 2 (right) illustrates this net-positive correction signal: as samples move toward the center of each mode in the $(z_t, z_v)$ plane, the density of corrections increases while that of miscorrections diminishes. This suggests that choosing sufficiently strict margins $\gamma_{\text{strict}}, \xi_{\text{strict}}$ beyond the baseline can simultaneously maximize corrections and minimize miscorrections. The following corollary formalizes this margin-driven trade-off:

**Corollary 3.1** (**Margin-Driven Strict Dominance**). *For sufficiently strict margins $\gamma_{strict}, \xi_{strict}$, the ratio of true correction to miscorrection increases compared to the baseline margins $\gamma_0, \xi_0$:*

$$\frac{\mathbb{P}\big(\hat{y}_t = y \wedge \hat{y}_v \neq y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma_{strict})\big)}{\mathbb{P}\big(\hat{y}_t \neq y \wedge \hat{y}_v = y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma_{strict})\big)} \geq \frac{\mathbb{P}\big(\hat{y}_t = y \wedge \hat{y}_v \neq y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma_0)\big)}{\mathbb{P}\big(\hat{y}_t \neq y \wedge \hat{y}_v = y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma_0)\big)} > 1. \tag{8}$$

*An analogous result holds for $\hat{\mathcal{S}}_{\bar{c}c}(\cdot)$ with thresholds $\xi_0$ and $\xi_{strict}$.*

**From Theory to Practice.** Corollary 3.1 shows that sufficiently enlarging the cut margin amplifies the dominance of true correction over miscorrection. In practice, however, selecting optimal thresholds $\gamma$ and $\xi$ is infeasible in a fully unlabeled setting. To bridge this gap, we replace hard thresholding with a soft gating mechanism that assigns supervision weights monotonically increasing in the cut margin $|\Delta(x)|$, prioritizing samples with a larger reliability gap. For simplicity, we adopt a symmetric formulation with $\gamma = \xi$, reducing the gating mechanism to a single margin-based criterion.

# 4 Training with Reliability-Gated Supervision

## 4.1 Estimating Prediction Reliability via Cut Statistics

To quantify prediction reliability via the cut statistic defined in (2), we first characterize the local geometry of the target domain under both $\theta_t$ and $\theta_v$. Let $\phi_i \colon \mathcal{X} \to \mathbb{R}^{d_i}$ denote the feature extractor for model $i \in \{t, v\}$, mapping inputs into a latent space where pairwise distances are computed.[3] Using these representations, we construct a local neighborhood graph by identifying the $K$-nearest neighbors $\mathcal{N}_K(x)$ for each sample $x$ under a standard distance metric $d$. To mitigate spurious connections – particularly for samples located in low-density regions – we restrict the neighborhood to *mutual* neighbors, i.e., $\mathcal{N}_K^{\text{mut}}(x) = \{x' \mid x' \in \mathcal{N}_K(x) \wedge x \in \mathcal{N}_K(x')\}$. Then, we construct a

[3] In our experiments, $\phi_i(x)$ corresponds to the output of the penultimate layer of each model (i.e., pre-logits).

distance-weighted cut statistic for each model $i \in \{t, v\}$ as follows

$$\hat{z}_i(x) = \frac{\sum_{x' \in \mathcal{N}_K^{\text{mut}}(x)} \exp(-d(\phi_i(x), \phi_i(x'))) \cdot \mathbb{1}[\hat{y}_i(x) \neq \hat{y}_i(x')]}{\sum_{x' \in \mathcal{N}_K^{\text{mut}}(x)} \exp(-d(\phi_i(x), \phi_i(x')))}. \tag{9}$$

Here, $w(x, x') \propto \exp(-d(\phi_i(x), \phi_i(x')))$ and normalization ensures that cut statistics remain comparable across regions with different sample densities. Finally, to translate the cut statistic into an intuitive measure of prediction reliability, we apply an exponential mapping $r_i(x) = \exp(-\hat{z}_i(x)) \in (0, 1]$, ensuring that highly consistent samples receive scores near 1.

Naively computing these scores requires an additional forward pass over the entire target data at every epoch, which is prohibitively expensive in SFDA settings where target domains often contain hundreds of thousands of samples. To circumvent this overhead, we reuse the latent features cached during the optimization forward pass of the preceding epoch. To further stabilize the estimates, we apply an exponential moving average across epochs, i.e., $\bar{r}_i^{(e)}(x) = m \cdot \bar{r}_i^{(e-1)}(x) + (1 - m) \cdot r_i^{(e)}(x)$, where $e$ denotes the epoch index and $m \in [0, 1)$ is the momentum parameter. For notational simplicity, we omit the epoch index hereafter and use $\bar{r}_i(x)$ in place of $r_i(x)$; the complete epoch-indexed formulation and algorithm details are provided in Appendix C.1. These reliability scores $\bar{r}_t$ and $\bar{r}_v$ serve as the foundation for determining how the two models supervise one another, which we will describe in the next subsection.

### 4.2 Reliability-Gated Mutual Correction

Given the per-sample reliability estimated from Section 4.1, we now formulate a novel mutual correction objective where $\theta_t$ and $\theta_v$ iteratively refine each other by utilizing their predictions as pseudo-labels. The central challenge in this cross-model supervision is that we must dynamically determine the *direction* and *strength* of the supervision for each sample while *preventing* reliable predictions from being corrupted by erroneous supervision from the less reliable peer.

**Margin-based Reliability Gating.** Recall from Section 3.2 that there exists a margin threshold that identifies regions of the target domain where one model can reliably supervise the other. Rather than relying on such discrete thresholds, we adopt a continuous formulation in which the reliability gap directly governs both the *direction* and *strength* of supervision on a per-sample basis. Formally, we define the cross-model *margin-based reliability gating* mechanism:

$$T_{v \to t}(x) = \sigma(\beta(\bar{r}_v(x) - \bar{r}_t(x))), \quad T_{t \to v}(x) = \sigma(\beta(\bar{r}_t(x) - \bar{r}_v(x))) \tag{10}$$

where $\sigma$ denotes the sigmoid function. By the symmetry of the sigmoid, $T_{v \to t}(x) + T_{t \to v}(x) = 1$ holds for all $x$, ensuring that the total supervisory budget is conserved across the two directions. Here, $T_{v \to t}(x)$ represents the weight assigned to the supervisory signal from $\theta_v$ to $\theta_t$ (and vice versa). The temperature parameter $\beta > 0$ controls the sharpness of the gating mechanism. As $\beta$ increases, the weights become more sensitive to the reliability gap $|\bar{r}_v(x) - \bar{r}_t(x)|$, approaching the hard-threshold behavior implied by Proposition 3.1. Conversely, smaller values of $\beta$ produce a smoother transition, allowing the model to remain less committed when the reliability difference is small (e.g., when both models are similarly uncertain or similarly confident). By applying this gating directly to the mutual correction signal, we obtain the following cross-supervision loss:

$$\mathcal{L}_{\text{cross}}^t(x) = T_{v \to t}(x) \cdot D_{\text{KL}}\big(\text{sg}(p_v) \,\|\, p_t\big), \quad \mathcal{L}_{\text{cross}}^v(x) = T_{t \to v}(x) \cdot D_{\text{KL}}\big(\text{sg}(p_t) \,\|\, p_v\big), \tag{11}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator. As a result, when one model has a clear reliability advantage on a given sample, it naturally assumes a stronger supervisory role, while the influence of the less reliable model is correspondingly attenuated.

**Temporal Self-Anchoring.** While soft gating in (11) effectively mitigates miscorrection, it does not eliminate it entirely. Even small, non-zero gating weights from an unreliable (incorrect) peer can allow residual error signals that can accumulate over successive epochs, gradually degrading predictions that were initially correct. To prevent this, we further introduce a *temporal self-anchoring* mechanism that encourages each model to remain consistent with its own previous predictions, dynamically weighted by its own reliability score:

$$\mathcal{L}_{\text{anc}}^t(x) = T_{t \to v}(x) \cdot D_{\text{KL}}(\tilde{p}_t \,\|\, p_t), \quad \mathcal{L}_{\text{anc}}^v(x) = T_{v \to t}(x) \cdot D_{\text{KL}}(\tilde{p}_v \,\|\, p_v) \tag{12}$$

where $\tilde{p}_t$ and $\tilde{p}_v$ denote the predictions from the previous epoch. Since $T_{t \to v}(x)$ is large only when $\theta_t$ is highly reliable, and symmetrically for $T_{v \to t}(x)$, the anchoring term selectively preserves confident predictions while allowing less reliable ones to be flexibly updated through cross-model supervision.

**Training Objective.** Combining reliability-gated mutual correction in (11) with temporal self-anchoring in (12), we define the overall training objective for each model as:

$$\mathcal{L}_t = \underbrace{\mathbb{E}_x \left[ T_{v\to t} \cdot D_{\mathrm{KL}}(\mathrm{sg}(p_v) \| p_t) + T_{t\to v} \cdot D_{\mathrm{KL}}(\tilde{p}_t \| p_t) \right]}_{\text{sample-wise supervision}} + \underbrace{\lambda_{\mathrm{IIC}} \mathcal{L}_{\mathrm{IIC}} - \lambda_{\mathrm{div}} H(\bar{p}_t)}_{\text{batch-wise regularization}} \tag{13}$$

$$\mathcal{L}_v = \underbrace{\mathbb{E}_x \left[ T_{t\to v} \cdot D_{\mathrm{KL}}(\mathrm{sg}(p_t) \| p_v) + T_{v\to t} \cdot D_{\mathrm{KL}}(\tilde{p}_v \| p_v) \right]}_{\text{sample-wise supervision}} + \underbrace{\lambda_{\mathrm{IIC}} \mathcal{L}_{\mathrm{IIC}} - \lambda_{\mathrm{div}} H(\bar{p}_v)}_{\text{batch-wise regularization}} \tag{14}$$

where $\bar{p}_t$ and $\bar{p}_v$ denote the mean predictions over the current batch. While the sample-wise supervision drives local, reliability-aware updates, we complement it with two batch-level regularizers that impose global structure. The IIC loss $\mathcal{L}_{\mathrm{IIC}}$ [20] maximizes the mutual information between $p_t$ and $p_v$ aggregated over the batch, aligning the prediction distributions of the two different views. This is particularly valuable on samples where the reliability gap $|\bar{r}_t - \bar{r}_v|$ is small, where the cross-supervision signal becomes nearly directionless. In addition, the diversity regularizer [3] maximizes the entropy of $\bar{p}_i$, preventing collapse to a few dominant classes.

# 5 Experiments

## 5.1 Experimental Setup

**Settings.** We evaluate `SafeCut` against state-of-the-art baselines on four benchmarks: **Office-31** [21] and **Office-Home** [22] as small- and medium-scale datasets, and **DomainNet-126** [23] and **VisDA** [24] as large-scale datasets (details in Appendix B). We adopt ResNet-50 [25] as the backbone for the first three datasets and ResNet-101 for VisDA, under a closed-set SFDA protocol. Following recent works [10, 11], we use CLIP [26] with ViT-B/32 as the peer model in most experiments, applying CoOp [27] for prompt learning with a frozen image encoder. Beyond this default setup, we additionally evaluate `SafeCut` in two practical scenarios: (i) replacing CLIP with a non-CLIP peer to test peer generality, and (ii) privacy-constrained medical settings using Camelyon17 from the **WILDS** benchmark [28] and diabetic retinopathy adaptation from EyePACS [29] to APTOS 2019.

## 5.2 Performance Evaluation

Table 1: Closed-set SFDA results (Acc, %) on **Office-31**.

| Method | Venue | A→D | A→W | D→A | D→W | W→A | W→D | Avg. |
|---|---|---|---|---|---|---|---|---|
| Source | – | 79.1 | 76.5 | 59.9 | 95.5 | 61.4 | 98.8 | 78.5 |
| SHOT | ICML20 | 94.4 | 88.4 | 73.2 | 96.7 | 74.9 | 99.8 | 87.9/88.6 |
| NRC | NIPS21 | 92.4 | 94.1 | 75.6 | 97.2 | 74.6 | 98.6 | 88.7/89.4 |
| AaD | NIPS22 | 87.6 | 86.5 | 58.9 | 90.6 | 58.7 | 90.6 | 78.8/89.9 |
| AdaCon | CVPR22 | 87.8 | 86.3 | 71.3 | 96.7 | 67.2 | 97.6 | 84.5/– |
| CoWA | ICML22 | 88.0 | 90.9 | 74.6 | 98.1 | 75.8 | 99.4 | 87.8/90.3 |
| PLUE | CVPR23 | 89.4 | 89.2 | 70.6 | 97.1 | 67.8 | 96.6 | 85.1/– |
| ELR | ICLR23 | 92.2 | 92.2 | 76.5 | 97.4 | 76.1 | 99.6 | 89.0/89.6 |
| DIFO-V | CVPR24 | 96.2 | 95.0 | 83.2 | 96.1 | 83.0 | 99.2 | 92.1/92.5 |
| ProDe-V | ICLR25 | 96.8 | **96.6** | 82.9 | 97.2 | 82.0 | 99.0 | 92.4/92.6 |
| `SafeCut` | NIPS26 | **98.6** | 95.2 | **83.9** | **98.2** | **83.3** | **100.0** | **93.2** |

In all benchmark experiments, we report the *last-epoch performance*. For a fair comparison, the best results reported in the literature are highlighted in blue alongside our reproduced baseline numbers (detailed per-task results are provided in Appendix D.1). As shown in Table 1, `SafeCut` achieves an average accuracy of 93.2% on Office-31, outperforming all reproduced baselines in 5 out of 6 adaptation tasks (+0.8% over reproduced ProDe-V) and exceeding the best reported result by +0.6%. Tables 2 and 3 demonstrate larger gains on Office-Home and DomainNet-126, with improvements of +2.3% and +2.4% over reproduced ProDe-V, and +2.2% and +0.8% over the paper-reported best results, respectively. On VisDA, our method remains competitive with the state of the art (90.9% vs. 91.0% paper-reported; +0.3% over reproduced ProDe-V). We attribute these consistent improvements across benchmarks to the reliability-based gating mechanism, which filters out miscorrections and mitigates the confirmation bias inherent in naive cross-supervision.

**Peer Model Generality.** To verify that `SafeCut` does not rely on vision-language pretraining, we replace the CLIP peer model with a second ResNet-50 pretrained on a different source domain. Since the two models (denoted as $\theta_1$ and $\theta_2$) are trained on distinct source distributions, they naturally exhibit complementary failure modes on the target domain (i.e., the condition required by `SafeCut`). As shown in Table 4, both models improve consistently across all 12 source domain pairs, regardless of the initial performance gap between $\theta_1$ and $\theta_2$. Notably, the adapted performance surpasses non-CLIP SFDA baselines (e.g., SHOT, NRC, AaD, ELR) in most combinations, confirming that `SafeCut` serves as a general-purpose correction framework applicable to diverse knowledge sources as long as they exhibit distinct failure modes with the source-pretrained model.

Table 2: Comparison on the Office-Home and VisDA datasets (Acc, %).

| Method | Venue | Office-Home | | | | | | | | | | | | | VisDA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Ar→Cl | Ar→Pr | Ar→Rw | Cl→Ar | Cl→Pr | Cl→Rw | Pr→Ar | Pr→Cl | Pr→Rw | Rw→Ar | Rw→Cl | Rw→Pr | Avg. | Sy→Re |
| Source | – | 43.6 | 67.0 | 73.9 | 49.9 | 60.2 | 62.6 | 51.7 | 40.9 | 72.6 | 64.1 | 46.3 | 78.1 | 59.2 | 49.1 |
| SHOT | ICML20 | 55.2 | 78.5 | 81.3 | 68.9 | 79.0 | 78.8 | 68.2 | 54.3 | 81.7 | 73.5 | 59.1 | 83.6 | 71.8/71.8 | 82.3/82.9 |
| NRC | NIPS21 | 54.7 | 77.1 | 79.4 | 64.7 | 76.2 | 75.3 | 63.3 | 53.6 | 78.8 | 70.2 | 56.5 | 82.5 | 69.4/72.2 | 83.4/85.9 |
| AaD | NIPS22 | 57.7 | 78.5 | 80.8 | 68.0 | 79.6 | 78.9 | 66.5 | 56.6 | 81.6 | 70.9 | 57.3 | 83.7 | 71.7/72.7 | 78.8/88.0 |
| AdaCon | CVPR22 | 49.5 | 76.1 | 77.0 | 63.5 | 73.2 | 74.4 | 60.7 | 47.1 | 77.9 | 67.8 | 49.7 | 80.3 | 66.4/– | 67.2/78.7 |
| CoWA | ICML22 | 56.7 | 79.0 | 80.9 | 68.6 | 78.7 | 79.8 | 68.1 | 55.6 | 82.6 | 72.4 | 61.7 | 83.6 | 72.3/72.5 | 84.3/86.9 |
| PLUE | CVPR23 | 46.3 | 61.9 | 71.6 | 59.2 | 63.7 | 65.8 | 58.0 | 44.3 | 72.1 | 63.0 | 49.2 | 76.6 | 61.0/– | 86.6/90.0 |
| ELR | ICLR23 | 55.9 | 77.6 | 79.9 | 61.4 | 76.8 | 76.5 | 62.4 | 53.3 | 80.9 | 68.9 | 58.7 | 83.6 | 69.7/72.6 | 85.8/85.8 |
| DIFO-V | CVPR24 | 70.7 | 90.5 | 89.2 | 82.4 | 90.2 | 89.2 | 80.6 | 71.1 | 89.4 | 82.9 | 71.0 | 91.0 | 83.2/83.1 | 89.8/90.3 |
| ProDe-V | ICLR25 | 71.8 | 92.2 | 90.5 | 82.7 | 91.6 | 90.6 | 82.4 | 72.8 | 90.8 | 82.8 | 72.7 | 92.0 | 84.4/84.5 | 90.6/91.0 |
| `SafeCut` | NIPS26 | **76.9** | **93.0** | **91.9** | **85.9** | **92.5** | **91.6** | **85.0** | **76.7** | **92.0** | **85.7** | **76.1** | **93.2** | **86.7** | **90.9** |

Table 3: Comparison on the DomainNet dataset (Acc, %).

| Method | Venue | DomainNet | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 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 | – | 44.7 | 59.9 | 47.5 | 53.3 | 75.3 | 46.2 | 55.1 | 62.6 | 46.4 | 55.2 | 50.8 | 59.6 | 54.7 |
| SHOT | ICML20 | 63.1 | 78.6 | 60.1 | 68.4 | 80.7 | 62.0 | 68.0 | 68.3 | 58.7 | 70.8 | 64.9 | 77.7 | 68.4/– |
| NRC | NIPS21 | 60.0 | 75.7 | 55.4 | 59.6 | 80.8 | 58.0 | 60.0 | 67.3 | 54.3 | 66.9 | 65.0 | 77.7 | 65.1/– |
| AaD | NIPS22 | 62.4 | 79.1 | 59.3 | 67.1 | 81.6 | 62.6 | 69.7 | 66.7 | 58.7 | 71.1 | 63.6 | 79.7 | 68.5/– |
| AdaCon | CVPR22 | 56.9 | 64.1 | 51.9 | 57.5 | 64.9 | 52.9 | 60.7 | 64.4 | 51.5 | 62.7 | 60.1 | 61.1 | 59.0/– |
| CoWA | ICML22 | 64.7 | 80.6 | 63.0 | 70.1 | 80.6 | 64.3 | 70.5 | 69.5 | 63.1 | 71.7 | 67.3 | 80.0 | 70.5/– |
| PLUE | CVPR23 | 55.5 | 62.0 | 53.9 | 57.2 | 68.4 | 53.0 | 58.5 | 61.8 | 52.3 | 60.8 | 58.5 | 65.6 | 58.9/– |
| ELR | ICLR23 | 64.8 | 76.1 | 61.3 | 66.7 | 81.8 | 63.7 | 69.1 | 69.2 | 62.0 | 72.3 | 66.8 | 77.5 | 69.3/79.4 |
| DIFO-V | CVPR24 | 76.5 | 87.8 | 75.4 | 78.8 | 87.7 | 75.7 | 80.0 | 77.3 | 75.6 | 79.2 | 76.4 | 87.8 | 79.9/80.0 |
| ProDe-V | ICLR25 | 80.9 | 91.5 | 77.4 | 84.1 | 91.3 | 76.9 | 84.2 | 81.3 | 76.6 | 84.0 | 81.2 | 91.2 | 83.4/85.0 |
| `SafeCut` | NIPS26 | **82.8** | **92.3** | **81.8** | **86.2** | **92.2** | **81.0** | **86.9** | **82.9** | **81.7** | **86.6** | **83.0** | **92.3** | **85.8** |

Table 4: Mutual correction between two ResNet-50 models from different source domains (Office-Home, Acc, %). Bef./Aft.: accuracy before/after adaptation.

| Target: **Ar** | | | | | | Target: **Cl** | | | | | | Target: **Pr** | | | | | | Target: **Rw** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\theta_1$ | | $\theta_2$ | | | | $\theta_1$ | | $\theta_2$ | | | | $\theta_1$ | | $\theta_2$ | | | | $\theta_1$ | | $\theta_2$ | |
| $\theta_1$ | $\theta_2$ | Bef. | Aft. | Bef. | Aft. | $\theta_1$ | $\theta_2$ | Bef. | Aft. | Bef. | Aft. | $\theta_1$ | $\theta_2$ | Bef. | Aft. | Bef. | Aft. | $\theta_1$ | $\theta_2$ | Bef. | Aft. | Bef. | Aft. |
| Cl | Pr | 49.9 | 72.2 | 51.7 | 72.2 | Ar | Pr | 43.6 | 59.8 | 40.9 | 59.9 | Ar | Cl | 67.0 | 81.3 | 60.2 | 81.2 | Ar | Cl | 73.9 | 83.1 | 62.6 | 83.1 |
| Cl | Rw | 49.9 | **75.8** | 64.2 | **75.8** | Ar | Rw | 43.6 | **61.7** | 46.3 | **61.6** | Cl | Rw | 60.2 | **86.4** | 78.1 | **86.4** | Ar | Pr | 73.9 | **84.3** | 72.6 | **84.2** |
| Pr | Rw | 51.7 | 74.4 | 64.2 | 74.4 | Pr | Rw | 40.9 | 60.4 | 46.3 | 60.4 | Ar | Rw | 67.0 | 83.0 | 78.1 | 83.0 | Cl | Pr | 62.6 | 83.9 | 72.6 | 83.9 |

### 5.3 Additional Analysis

**Ablation Studies.** We analyze the contribution of each component through ablation and sensitivity studies on Office-31, Office-Home, and DomainNet-126. Table 5(a) ablates each loss term. Here, $0/1$ denote variants of `SafeCut` in which each term is disabled or enabled with uniform (constant) weighting, replacing our reliability-based gating. $\mathcal{L}_{\mathrm{c}}$ provides the primary performance gain, while $\mathcal{L}_{\mathrm{a}}$ improves stability by mitigating noise from pseudo-labels by anchoring to earlier predictions. Combining both with adaptive gating in `SafeCut` achieves the best performance, highlighting their complementary roles and the importance of reliability-aware weighting. Table 5(b) compares key design variants: freezing $\theta_v$ degrades correction quality, and entropy-based gating underperforms cut-based gating, supporting the use of the cut statistic as a more reliable sample-level signal due to

Table 5: Ablation and sensitivity analysis (Avg. Acc, %). $\mathcal{L}_{\mathrm{c}}$: cross-supervision KL, $\mathcal{L}_{\mathrm{a}}$: temporal anchoring.

**(a) Component ablation**

| | O31 | OH | DN |
|---|---|---|---|
| $\mathcal{L}_{\mathrm{c}}=0, \mathcal{L}_{\mathrm{a}}=0$ | 91.2 | 81.8 | 72.1 |
| $\mathcal{L}_{\mathrm{c}}=0, \mathcal{L}_{\mathrm{a}}=1$ | 91.2 | 83.3 | 78.7 |
| $\mathcal{L}_{\mathrm{c}}=1, \mathcal{L}_{\mathrm{a}}=0$ | 92.4 | 86.0 | 83.6 |
| $\mathcal{L}_{\mathrm{c}}=1, \mathcal{L}_{\mathrm{a}}=1$ | 92.1 | 86.2 | 84.3 |
| `SafeCut` | **93.2** | **86.7** | **85.8** |

**(b) Variant**

| Variant | O31 | OH | DN |
|---|---|---|---|
| Frozen $\theta_v$ | 89.0 | 84.8 | 84.6 |
| Entropy gating | 92.5 | 86.3 | 85.0 |
| `SafeCut` | **93.2** | **86.7** | **85.8** |
| Oracle partition $S$ | 96.0 | 93.3 | 92.7 |

**(c) CLIP adaptation**

| | O31 | OH | DN |
|---|---|---|---|
| Zero-shot | 79.4 | 79.0 | 80.2 |
| **Adapted** | **92.8** | **85.8** | **84.0** |

**(d) $\beta$ sensitivity (OH)**

| $\beta$ | OH |
|---|---|
| 2 | 86.7 |
| 4 | 86.6 |
| 6 | **86.7** |

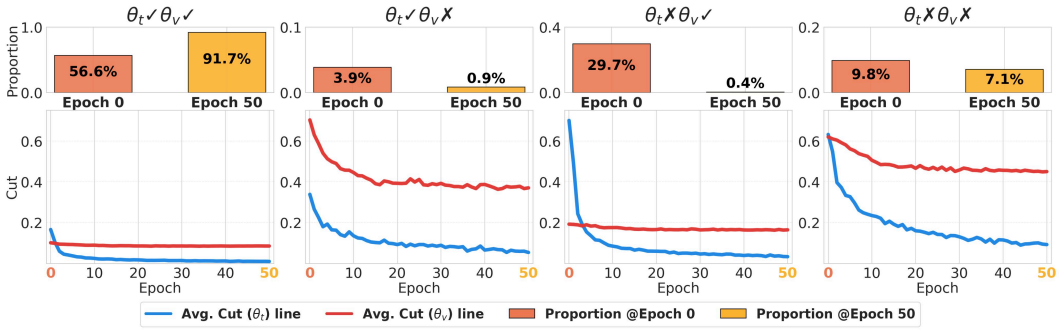



Figure 3: **Correctness pattern (top) and average cut trajectories (bottom) on Office-Home Cl→Pr.** Samples are grouped by their initial state and tracked throughout adaptation.

its incorporation of local neighborhood structure. The oracle partition provides an upper bound, indicating that further improvements could be achieved by better identifying relative expertise. Table 5(c) confirms that ViL adaptation is essential, as zero-shot CLIP significantly underperforms the adapted peer. Table 5(d) shows robustness to the gating sharpness parameter $\beta$, requiring minimum tuning (full ablation details are provided in Appendix E.3).

**Training Dynamics.** As a diagnostic, we track ground-truth state transitions and per-group cut dynamics throughout adaptation (Figure 3). The proportion of jointly correct samples $\mathcal{S}_{\mathrm{cc}}$ increases sharply from 56.6% to 91.7%, while the one-sided correct sets $\mathcal{S}_{\mathrm{c\bar{c}}}$ and $\mathcal{S}_{\mathrm{\bar{c}c}}$ nearly disappear, indicating effective bidirectional knowledge transfer. The jointly incorrect set $\mathcal{S}_{\mathrm{\bar{c}\bar{c}}}$ also decreases modestly (9.8%→7.1%), suggesting gradual improvement through neighborhood generalization. To further assess the reliability signal, we group samples by their initial ground-truth state and track the per-group average cut value over training epochs: samples in $\mathcal{S}_{\mathrm{cc}}$ maintain consistently low cut values, reflecting stable neighborhoods, samples in the one-sided sets show a steady decrease as the models become aligned, and samples in $\mathcal{S}_{\mathrm{\bar{c}\bar{c}}}$ maintain consistent high cut values. This confirms that `SafeCut` assigns low reliability to samples that are difficult to correct. See Appendix D.5 for additional results.

**Effect of Peer Complementarity.** To examine how peer complementarity affects mutual correction, we use the 12 dual-ResNet pairs in Table 4. For each pair, we measure the *pre-adaptation prediction agreement*, defined as the fraction of target samples on which the two models predict the same class, and compare it with the *adaptation gain*, defined as the average accuracy improvement of the two models after `SafeCut` adaptation. As shown in Figure 4, higher prediction agreement is strongly associated with smaller adaptation gains (Pearson $r = -0.81$), supporting our motivation that complementary predictions provide more opportunities for cross-model correction. Nevertheless, the gain decreases gracefully as agreement increases: even at the highest observed agreement (69.6%), SafeCut still yields a +10.4%p improvement, with no degradation across the evaluated pairs. See Appendix D.6 for detailed results.

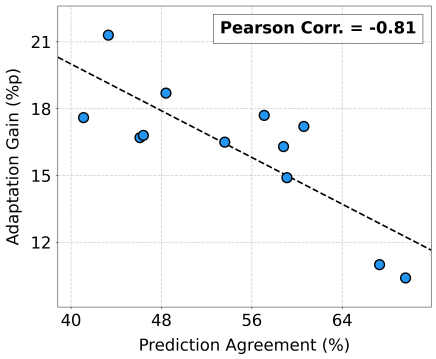



Figure 4: **Peer complementarity.** Prediction agreement vs. adaptation gain in the dual-ResNet setting of Table 4.

**Medical Domain Setup.** Medical imaging is a canonical SFDA setting, where sharing source data is often restricted due to privacy concerns. We evaluate `SafeCut` on Camelyon17 from WILDS benchmark [28], training the source model on centers $\{0, 3, 4\}$, validating on $\{1\}$, and adapting to $\{2\}$. Unlike natural-image benchmarks, zero-shot CLIP transfers poorly to histopathology, even underperforming the source-pretrained target model (50.01% vs. 64.14%). In this setting, ProDe achieves only modest improvements on the target model and leaves (updated) CLIP similar to its zero-shot performance (50.00%, with hyperparameters tuned via grid search;

Table 6: SFDA on Camelyon17 (Acc, %).

| Method | Target | CLIP |
|---|---|---|
| Source | 64.14 | 50.01 |
| ProDe | 76.27 | 50.00 |
| $\mathcal{L}_{\mathrm{c}}=1, \mathcal{L}_{\mathrm{a}}=0$ | 77.29 | 77.20 |
| $\mathcal{L}_{\mathrm{c}}=1, \mathcal{L}_{\mathrm{a}}=1$ | 87.96 | 87.78 |
| `SafeCut` | **94.96** | **94.20** |

see Appendix E.2). This is because ProDe updates CLIP only through $\mathcal{L}_{\text{IIC}}$, providing no direct source-driven supervision; as a result, it fails to improve and continues to provide weak guidance. In contrast, `SafeCut` trains both models through mutual, reliability-gated correction, enabling joint improvement. Using $\mathcal{L}_c$ alone matches ProDe, while adding $\mathcal{L}_a$ further stabilizes training. The complete model (`SafeCut`) achieves $94.96\%$, significantly outperforming ProDe. These results highlight the importance of reliability-aware mutual correction when the peer model is initially unreliable.

To further evaluate this weak-peer regime, we additionally consider diabetic retinopathy severity classification, pretraining the source model on EyePACS [29] and adapting to APTOS 2019. As shown in Table 7, zero-shot CLIP achieves only $25.4\%$ accuracy on this 5-class task, indicating a particularly weak initial peer. ProDe reduces the target-model accuracy to $35.2\%$, whereas `SafeCut` improves the target and peer models to $46.0\%$ and $45.9\%$, respectively. Together with the Camelyon17 results, these findings show that `SafeCut` remains effective in specialized medical domains where pretrained ViL models transfer poorly, supporting the benefit of reliability-gated mutual correction beyond natural-image benchmarks.

Table 7: EyePACS→APTOS2019 (Acc, %).

| Method | Target | CLIP |
|---|---|---|
| Source | 39.0 | 25.4 |
| ProDe | 35.2 | 35.3 |
| `SafeCut` | **46.0** | **45.9** |

## 6 Conclusion

In this paper, we presented `SafeCut`, a novel SFDA framework driven by reliability-gated mutual correction. Unlike existing methods that unconditionally trust external models, we exploit the different failure modes of the target and peer models. Backed by solid theoretical grounding, `SafeCut` uses the cut statistic as a label-free reliability proxy to dynamically control the direction and strength of cross-model supervision. Coupled with temporal self-anchoring, this mechanism amplifies valid corrective signals while mitigating miscorrections and suppressing error accumulation. Extensive experiments confirm that `SafeCut` achieves state-of-the-art performance across multiple benchmarks. Crucially, `SafeCut` not only generalizes across diverse peer models but also excels in challenging domains like medical imaging, demonstrating its versatility as a general-purpose correction framework.


## Acknowledgments and Disclosure of Funding

We thank the NeurIPS reviewers for their valuable comments and suggestions. This work is supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00358602) and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT): Artificial Intelligence Graduate School Program (No. RS-2019-II190079, Korea University), Development of Unified Reasoning Technology Mimicking Human Cognition for Hierarchical Understanding and Unbounded Problem Solving (No. RS-2026-25522672), and Development of AI Co-Scientist based on Scientific Causal World Model (No.RS-2026-25507543).


## References


[1] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[2] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.

[3] 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*, pages 6028–6039. PMLR, 2020.

[4] Shiqi Yang, Joost Van de Weijer, Luis Herranz, Shangling Jui, et al. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. *Advances in neural information processing systems*, 34:29393–29405, 2021.

[5] Shiqi Yang, Shangling Jui, Joost Van De Weijer, et al. Attracting and dispersing: A simple approach for source-free domain adaptation. *Advances in Neural Information Processing Systems*, 35:5802–5815, 2022.

[6] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 295–305, 2022.

[7] Mattia Litrico, Alessio Del Bue, and Pietro Morerio. Guiding pseudo-labels with uncertainty estimation for source-free unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7640–7650, 2023.

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

[9] Li Yi, Gezheng Xu, Pengcheng Xu, Jiaqi Li, Ruizhi Pu, Charles Ling, A Ian McLeod, and Boyu Wang. When source-free domain adaptation meets learning with noisy labels. *arXiv preprint arXiv:2301.13381*, 2023.

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

[11] Song Tang, Wenxin Su, Yan Gan, Mao Ye, Jianwei Zhang, and Xiatian Zhu. Proxy denoising for source-free domain adaptation. *arXiv preprint arXiv:2406.01658*, 2024.

[12] 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*, pages 2994–3003, 2024.

[13] Xiaowei Yu, Zihao Wu, Lu Zhang, Jing Zhang, Yanjun Lyu, and Dajiang Zhu. Cp-clip: Core-periphery feature alignment clip for zero-shot medical image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 88–97. Springer, 2024.

[14] Fabrice Muhlenbach, Stéphane Lallich, and Djamel A Zighed. Identifying and handling mislabelled instances. *Journal of Intelligent Information Systems*, 22(1):89–109, 2004.

[15] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4320–4328, 2018.

[16] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.

[17] Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023.

[18] Hunter Lang, David Sontag, and Aravindan Vijayaraghavan. Theoretical analysis of weak-to-strong generalization. *Advances in neural information processing systems*, 37:46837–46880, 2024.

[19] O Chapelle. Semi-supervised learning, 2006.

[20] Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9865–9874, 2019.

[21] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010.

[22] 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*, pages 5018–5027, 2017.

[23] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8050–8058, 2019.

[24] 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.

[25] 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*, pages 770–778, 2016.

[26] 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*, pages 8748–8763. PmLR, 2021.

[27] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348, 2022.

[28] Shiori Sagawa, Pang Wei Koh, Tony Lee, Irena Gao, Sang Michael Xie, Kendrick Shen, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, Henrik Marklund, et al. Extending the wilds benchmark for unsupervised adaptation. *arXiv preprint arXiv:2112.05090*, 2021.

[29] Jorge Cuadros and George Bresnick. Eyepacs: an adaptable telemedicine system for diabetic retinopathy screening. *Journal of diabetes science and technology*, 3(3):509–516, 2009.

[30] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019.

[31] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *International Conference on Learning Representations*, 2020.

[32] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in neural information processing systems*, 34:18408–18419, 2021.

# Appendix

## Table of Contents


**A Proofs** **14**

A.1 Theoretical Justification under the Noisy Scenario 15

**B Evaluation Datasets** **16**

**C Implementation Details** **17**

C.1 Epoch-Indexed Formulation 17

**D Additional Results** **18**

D.1 Comparison Against Paper-Reported Best Results 18

D.2 VisDA Full Results 20

D.3 Computational Analysis 20

D.4 Extended Empirical Validation of Cut Statistic Properties 20

D.5 Extended Correctness Pattern Transition Results 26

D.6 Effect of Peer Complementarity 27

**E Experimental Details** **29**

E.1 Implementation and Hyperparameter Details 29

E.2 ProDe Hyperparameter Search on Camelyon17 29

E.3 Full Per-Task Ablation Results 30

**F Limitation and Future Directions** **33**

**G Broader Impacts** **33**

**H Asset Licenses and Terms of Use** **33**

# A Proofs

**Proposition A.1 (Reliability-Separated Regions).** *There exist baseline margins $\gamma_0, \xi_0$ such that:*

$$\mathbb{P}\big(\hat{y}_t = y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma_0)\big) > \mathbb{P}\big(\hat{y}_v = y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma_0)\big), \tag{15}$$

$$\mathbb{P}\big(\hat{y}_v = y \mid x \in \hat{\mathcal{S}}_{\bar{c}c}(\xi_0)\big) > \mathbb{P}\big(\hat{y}_t = y \mid x \in \hat{\mathcal{S}}_{\bar{c}c}(\xi_0)\big). \tag{16}$$

*Proof.* Motivated by the approximate monotonicity observed in Observation 3.1, we consider a sufficiently separated positive-margin region and let $C := \{x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma_0)\}$. By definition, any sample in this region satisfies $\Delta(x) = z_v(x) - z_t(x) \geq \gamma_0 > 0$.

Using the law of total probability, the reliability gap between the two models over region $C$ can be expressed as an expectation (See detail proof in noisy version A.2):

$$\mathbb{P}\big(\hat{y}_t = y \mid C\big) - \mathbb{P}\big(\hat{y}_v = y \mid C\big) = \mathbb{E}_{x|C}\Big[\mathbb{P}\big(\hat{y}_t = y \mid z_t(x)\big) - \mathbb{P}\big(\hat{y}_v = y \mid z_v(x)\big)\Big].$$

For any $x \in C$, we have $\Delta(x) \geq \gamma_0 > 0$, and hence $z_t(x) < z_v(x)$. Under the cross-model reliability ordering induced by the cut statistic in this reliability-separated region, a lower cut statistic corresponds to a higher probability of correctness; thus, $\mathbb{P}(\hat{y}_t = y \mid z_t(x)) > \mathbb{P}(\hat{y}_v = y \mid z_v(x))$, implying that the integrand is strictly positive for all $x \in C$. Therefore, it follows that:

$$\mathbb{P}\big(\hat{y}_t = y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma_0)\big) - \mathbb{P}\big(\hat{y}_v = y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma_0)\big) > 0.$$

An analogous derivation confirms the symmetric result for $\hat{\mathcal{S}}_{\bar{c}c}(\xi_0)$. □

**Theorem A.1 (Correction Dominance).** *For the baseline margins $\gamma_0, \xi_0$ satisfying Proposition A.1:*

$$\mathbb{P}\big(\hat{y}_t = y \wedge \hat{y}_v \neq y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma_0)\big) > \mathbb{P}\big(\hat{y}_t \neq y \wedge \hat{y}_v = y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma_0)\big), \tag{17}$$

$$\mathbb{P}\big(\hat{y}_v = y \wedge \hat{y}_t \neq y \mid x \in \hat{\mathcal{S}}_{\bar{c}c}(\xi_0)\big) > \mathbb{P}\big(\hat{y}_v \neq y \wedge \hat{y}_t = y \mid x \in \hat{\mathcal{S}}_{\bar{c}c}(\xi_0)\big). \tag{18}$$

*Proof.* Let $E_1$ be the event $\{\hat{y}_t = y\}$ and $E_2$ be the event $\{\hat{y}_v = y\}$. For any conditional event $C$, the law of total probability yields:

$$\mathbb{P}(E_1 \mid C) = \mathbb{P}(E_1 \wedge E_2 \mid C) + \mathbb{P}(E_1 \wedge E_2^c \mid C),$$
$$\mathbb{P}(E_2 \mid C) = \mathbb{P}(E_1 \wedge E_2 \mid C) + \mathbb{P}(E_1^c \wedge E_2 \mid C).$$

Subtracting the second equation from the first, we establish the following fundamental identity:

$$\mathbb{P}(E_1 \mid C) - \mathbb{P}(E_2 \mid C) = \mathbb{P}(E_1 \wedge E_2^c \mid C) - \mathbb{P}(E_1^c \wedge E_2 \mid C). \tag{19}$$

For the first inequality, let the condition $C$ be $x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma_0)$. By Proposition A.1, we have $\mathbb{P}(E_1 \mid C) > \mathbb{P}(E_2 \mid C)$. Applying this to Equation (19), it directly follows that:

$$\mathbb{P}(E_1 \wedge E_2^c \mid C) - \mathbb{P}(E_1^c \wedge E_2 \mid C) > 0.$$

Substituting the events back to their original notation, we obtain:

$$\mathbb{P}\big(\hat{y}_t = y \wedge \hat{y}_v \neq y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma_0)\big) > \mathbb{P}\big(\hat{y}_t \neq y \wedge \hat{y}_v = y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma_0)\big),$$

which satisfies the first claim of the theorem.

Similarly, for the second inequality, let the condition $C'$ be $x \in \hat{\mathcal{S}}_{\bar{c}c}(\xi_0)$. Proposition A.1 states that $\mathbb{P}(E_2 \mid C') > \mathbb{P}(E_1 \mid C')$. Applying the same probability identity for condition $C'$, we get:

$$\mathbb{P}(E_2 \mid C') - \mathbb{P}(E_1 \mid C') = \mathbb{P}(E_1^c \wedge E_2 \mid C') - \mathbb{P}(E_1 \wedge E_2^c \mid C') > 0.$$

Therefore, we deduce:

$$\mathbb{P}\big(\hat{y}_v = y \wedge \hat{y}_t \neq y \mid x \in \hat{\mathcal{S}}_{\bar{c}c}(\xi_0)\big) > \mathbb{P}\big(\hat{y}_v \neq y \wedge \hat{y}_t = y \mid x \in \hat{\mathcal{S}}_{\bar{c}c}(\xi_0)\big),$$

□

### A.1 Theoretical Justification under the Noisy Scenario

While Section 3.2 assumes a strict monotonic relationship between the cut statistic and prediction accuracy, real-world distribution shifts in the target domain inevitably introduce noise. In an ideal, noise-free scenario, the difference in predictive accuracy between the two models would scale linearly with the cut margin on a per-sample basis. That is, within the valid bounds of probability, the accuracy difference would be proportional to the cut margin at a rate of $\eta$. In practice, however, individual samples exhibit empirical deviations from this ideal relationship. To account for such sample-wise variance, we introduce a noise tolerance term  to establish a rigorous lower bound. Here, we formalize this relaxed condition as *noisy monotonicity* and prove that correction dominance is still theoretically guaranteed provided the margin is sufficiently large.

**Assumption A.1 (Noisy Monotonicity).** For any target sample $x \in \mathcal{D}_t$ with cut statistics $z_t(x)$ and $z_v(x)$, there exist a decay rate $\eta > 0$ and noise level $\epsilon \geq 0$ such that:

$$\mathbb{P}\big(\hat{y}_t = y \mid z_t(x)\big) - \mathbb{P}\big(\hat{y}_v = y \mid z_v(x)\big) \geq \eta\Delta(x) - \epsilon, \tag{20}$$

where $\Delta(x) = z_v(x) - z_t(x)$ is the cut margin.

Based on this assumption, we can establish the lower bound for the reliability gap in the margin-separated regions.

**Proposition A.2 (Reliability Gap under Noise).** *Under Assumption A.1, for the margin-separated regions $\hat{\mathcal{S}}_{c\bar{c}}(\gamma)$ and $\hat{\mathcal{S}}_{\bar{c}c}(\xi)$, the reliability gap between $\theta_t$ and $\theta_v$ is bounded by:*

$$\mathbb{P}\big(\hat{y}_t = y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma)\big) - \mathbb{P}\big(\hat{y}_v = y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma)\big) \geq \eta\gamma - \epsilon, \tag{21}$$

$$\mathbb{P}\big(\hat{y}_v = y \mid x \in \hat{\mathcal{S}}_{\bar{c}c}(\xi)\big) - \mathbb{P}\big(\hat{y}_t = y \mid x \in \hat{\mathcal{S}}_{\bar{c}c}(\xi)\big) \geq \eta\xi - \epsilon. \tag{22}$$

*Proof.* Let $C$ denote the event $\{x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma)\}$. By the definition of this region in Equation (3), any sample conditioned on $C$ strictly satisfies $\Delta(x) \geq \gamma$.

Using the definition of conditional probability, the probability that model $\theta_t$ is correct given $C$ is:

$$\mathbb{P}(\hat{y}_t = y \mid C) = \int_{x \in \mathcal{D}_t} \mathbb{P}(\hat{y}_t = y \mid x, C) \cdot \mathbb{P}(x \mid C) dx.$$

Since $\mathbb{P}(x \mid C) = 0$ for any $x \notin C$, we can restrict the integration domain to $C$. Furthermore, assuming the prediction correctness primarily depends on the cut statistic $z_t(x)$, we can express this as an expectation over the region $C$:

$$\mathbb{P}(\hat{y}_t = y \mid C) = \int_{x \in C} \mathbb{P}\big(\hat{y}_t = y \mid z_t(x)\big) \cdot \mathbb{P}(x \mid C) dx = \mathbb{E}_{x|C}\big[\mathbb{P}\big(\hat{y}_t = y \mid z_t(x)\big)\big].$$

Applying the same logic for $\hat{y}_v = y$ and subtracting the two yields:

$$\mathbb{P}(\hat{y}_t = y \mid C) - \mathbb{P}(\hat{y}_v = y \mid C) = \mathbb{E}_{x|C}\Big[\mathbb{P}\big(\hat{y}_t = y \mid z_t(x)\big) - \mathbb{P}\big(\hat{y}_v = y \mid z_v(x)\big)\Big].$$

By Assumption A.1, the integrand is lower-bounded by $\eta\Delta(x) - \epsilon$. Since $\Delta(x) \geq \gamma$ for all $x \in C$, we have:

$$\mathbb{E}_{x|C}\Big[\mathbb{P}\big(\hat{y}_t = y \mid z_t(x)\big) - \mathbb{P}\big(\hat{y}_v = y \mid z_v(x)\big)\Big] \geq \mathbb{E}_{x|C}\big[\eta\gamma - \epsilon\big] = \eta\gamma - \epsilon.$$

Therefore, $\mathbb{P}\big(\hat{y}_t = y \mid C\big) - \mathbb{P}\big(\hat{y}_v = y \mid C\big) \geq \eta\gamma - \epsilon$. □

For $\hat{\mathcal{S}}_{\bar{c}c}(\xi)$, applying the symmetric noisy-monotonicity relation $\mathbb{P}(\hat{y}_v = y \mid z_v(x)) - \mathbb{P}(\hat{y}_t = y \mid z_t(x)) \geq \eta(-\Delta(x)) - \epsilon$ together with $\Delta(x) \leq -\xi$ yields the second inequality analogously. Building upon this reliability gap, we now establish the condition under which true correction strictly dominates miscorrection in the presence of noise.

**Theorem A.2 (Correction Dominance under Noise).** *For the margin-separated region $\hat{\mathcal{S}}_{c\bar{c}}(\gamma)$ satisfying the condition $\gamma > \frac{\epsilon}{\eta}$, the probability of true correction strictly dominates miscorrection by at least the margin $\eta\gamma - \epsilon$:*

$$\mathbb{P}\big(\hat{y}_t = y \wedge \hat{y}_v \neq y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma)\big) - \mathbb{P}\big(\hat{y}_t \neq y \wedge \hat{y}_v = y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma)\big) \geq \eta\gamma - \epsilon > 0. \tag{23}$$

*Symmetrically, for $\hat{\mathcal{S}}_{\bar{c}c}(\xi)$ satisfying $\xi > \frac{\epsilon}{\eta}$:*

$$\mathbb{P}\big(\hat{y}_v = y \wedge \hat{y}_t \neq y \mid x \in \hat{\mathcal{S}}_{\bar{c}c}(\xi)\big) - \mathbb{P}\big(\hat{y}_v \neq y \wedge \hat{y}_t = y \mid x \in \hat{\mathcal{S}}_{\bar{c}c}(\xi)\big) \geq \eta\xi - \epsilon > 0. \tag{24}$$

*Proof.* Let $E_1$ be the event $\{\hat{y}_t = y\}$ and $E_2$ be the event $\{\hat{y}_v = y\}$. For any conditional event $C$, the law of total probability yields:

$$\mathbb{P}(E_1 \mid C) = \mathbb{P}(E_1 \wedge E_2 \mid C) + \mathbb{P}(E_1 \wedge E_2^c \mid C),$$
$$\mathbb{P}(E_2 \mid C) = \mathbb{P}(E_1 \wedge E_2 \mid C) + \mathbb{P}(E_1^c \wedge E_2 \mid C).$$

Subtracting the second equation from the first, we establish the following fundamental identity:

$$\mathbb{P}(E_1 \mid C) - \mathbb{P}(E_2 \mid C) = \mathbb{P}(E_1 \wedge E_2^c \mid C) - \mathbb{P}(E_1^c \wedge E_2 \mid C). \tag{25}$$

For the first inequality, let the condition $C$ be $x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma)$. By Proposition A.2, we have:

$$\mathbb{P}(E_1 \mid C) - \mathbb{P}(E_2 \mid C) \geq \eta\gamma - \epsilon.$$

Applying this to Equation (25), it directly follows that:

$$\mathbb{P}(E_1 \wedge E_2^c \mid C) - \mathbb{P}(E_1^c \wedge E_2 \mid C) \geq \eta\gamma - \epsilon.$$

Given the condition $\gamma > \frac{\epsilon}{\eta}$, we know that $\eta\gamma - \epsilon > 0$. Substituting the events back to their original notation, we obtain:

$$\mathbb{P}\big(\hat{y}_t = y \wedge \hat{y}_v \neq y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma)\big) - \mathbb{P}\big(\hat{y}_t \neq y \wedge \hat{y}_v = y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma)\big) \geq \eta\gamma - \epsilon > 0,$$

which satisfies the first claim of the theorem.

Similarly, for the second inequality, let the condition $C'$ be $x \in \hat{\mathcal{S}}_{\bar{c}c}(\xi)$. Proposition A.2 states that $\mathbb{P}(E_2 \mid C') - \mathbb{P}(E_1 \mid C') \geq \eta\xi - \epsilon$. Applying the same probability identity for condition $C'$, we get:

$$\mathbb{P}(E_2 \mid C') - \mathbb{P}(E_1 \mid C') = \mathbb{P}(E_1^c \wedge E_2 \mid C') - \mathbb{P}(E_1 \wedge E_2^c \mid C') \geq \eta\xi - \epsilon.$$

Given $\xi > \frac{\epsilon}{\eta}$, we deduce:

$$\mathbb{P}\big(\hat{y}_v = y \wedge \hat{y}_t \neq y \mid x \in \hat{\mathcal{S}}_{\bar{c}c}(\xi)\big) - \mathbb{P}\big(\hat{y}_v \neq y \wedge \hat{y}_t = y \mid x \in \hat{\mathcal{S}}_{\bar{c}c}(\xi)\big) \geq \eta\xi - \epsilon > 0.$$

□

# B Evaluation Datasets

In this paper, our method is evaluated on five widely used benchmarks for domain adaptation as follows.

- **Office-31** [21] is a small-scale dataset consisting of three domains: Amazon (A), DSLR (D), and Webcam (W), all featuring real-world objects in office environments. The Amazon domain contains images downloaded from `amazon.com` under studio lighting conditions, whilst the DSLR and Webcam domains consist of images captured with a high-resolution SLR camera and a low-resolution webcam, respectively. The dataset contains 4,652 images across 31 categories in total.
- **Office-Home** [22] is a medium-scale dataset containing approximately 15,500 images of 65 everyday object categories across four visually distinct domains: Artistic images (Ar), Clip Art (Cl), Product images (Pr), and Real-World images (Rw). The dataset has on average 70 images per category and is notably more challenging than Office-31 due to its larger number of categories and greater domain diversity.

- **VisDA-C** [24] is a large-scale dataset designed for the simulation-to-reality domain shift in object classification. The source domain comprises 152,397 synthetic images rendered from 3D CAD models under varying viewpoints and lighting conditions, whilst the target validation domain consists of 55,388 real object images cropped from the Microsoft COCO dataset. The dataset spans 12 object categories and poses a substantial distribution gap due to the synthetic-to-real nature of the shift.
- **DomainNet-126** [23] is a large-scale and challenging benchmark derived from the original DomainNet dataset [30], which contains approximately 600k images of 345 classes across 6 domains. Following [30], we use a curated subset of 4 domains—Clipart (C), Painting (P), Real (R), and Sketch (S)—and 126 classes with reduced label noise, resulting in roughly 145k images in total.
- **Camelyon17-WILDS** [28] is a histopathology benchmark for evaluating robustness to domain shift across hospitals. Each input is a $96 \times 96$ tissue patch, and the task is binary classification of whether the patch contains tumor tissue. The 5 hospitals serve as distinct domains, with models trained on 3 hospitals and evaluated on a held-out hospital, yielding 455,954 labeled examples in total.
- **EyePACS** [29] is a retinal fundus image dataset for diabetic retinopathy screening and severity assessment. It contains fundus photographs acquired under diverse imaging conditions, resulting in substantial variation in image appearance and quality. We use EyePACS as the source domain.
- **APTOS 2019**[4] is a retinal fundus image dataset for diabetic retinopathy severity classification. The task consists of five classes corresponding to increasing levels of diabetic retinopathy severity. We use APTOS 2019 as the target domain.

# C Implementation Details

## C.1 Epoch-Indexed Formulation

The main text omits epoch indices for notational brevity; here we make the full temporal structure explicit and describe the efficient caching scheme that underlies the implementation.

**Feature Caching.** Computing cut statistics requires the complete feature set $\{\phi_i(x_j)\}_{j=1}^n$ over the entire target domain, as nearest-neighbor relationships cannot be resolved within a single mini-batch. A naïve implementation would therefore require a dedicated full forward pass at every epoch solely for cut statistic computation, incurring significant overhead when the target domain contains hundreds of thousands of samples.

To eliminate this cost, we maintain a feature bank $\mathcal{B}_i^{\phi}$ and a prediction bank $\mathcal{B}_i^{p}$ for each model $i \in \{t, v\}$, populated *in-place* during the standard optimization forward pass:

$$\mathcal{B}_i^{\phi,(e)}[j] \leftarrow \phi_i^{(e)}(x_j), \qquad \mathcal{B}_i^{p,(e)}[j] \leftarrow p_i^{(e)}(x_j), \qquad j \in \mathcal{B}, \tag{26}$$

so that by the end of epoch $e$ the banks hold the most recent representations without any additional forward pass. Because the CLIP image encoder $\phi_v$ is entirely frozen, $\mathcal{B}_v^{\phi}$ is computed once during warmup and never modified thereafter. At the end of each epoch, the epoch-indexed cut statistic is computed from the accumulated banks:

$$\hat{z}_i^{(e)}(x) = \frac{\displaystyle\sum_{x' \in \mathcal{N}_K^{\mathrm{mut},(e)}(x)} \exp\big(-d(\mathcal{B}_i^{\phi,(e)}(x),\, \mathcal{B}_i^{\phi,(e)}(x'))\big) \cdot \mathbf{1}\Big[\mathcal{B}_i^{p,(e)}(x) \neq \mathcal{B}_i^{p,(e)}(x')\Big]}{\displaystyle\sum_{x' \in \mathcal{N}_K^{\mathrm{mut},(e)}(x)} \exp\big(-d(\mathcal{B}_i^{\phi,(e)}(x),\, \mathcal{B}_i^{\phi,(e)}(x'))\big)}, \tag{27}$$

yielding the raw reliability score $r_i^{(e)}(x) = \exp(-\hat{z}_i^{(e)}(x))$.

[4] https://www.kaggle.com/competitions/aptos2019-blindness-detection/data

**Algorithm 1** Training of `SafeCut`

**Input**: Unlabeled target data $\{x_j\}_{j=1}^n$, pretrained student $\theta_t$, frozen CLIP encoder $\phi_v$, learnable prompt $\theta_v$, total epochs $E$.
**Procedure**:

**Initialisation**: Set $\theta_t \leftarrow \theta_s$; set prompt context $\boldsymbol{v}$ = "a photo of a".
**Warmup**: Forward all $\{x_j\}$ through $\theta_t$ and $\phi_v$; populate banks $\mathcal{B}_i^{\phi}$, $\mathcal{B}_i^{p}$ for $i \in \{t, v\}$.
Compute $r_i^{(0)}$ via Eq. (9); set $\bar{r}_i^{(0)} \leftarrow r_i^{(0)}$ for $i \in \{t, v\}$.
**for** $e = 1, 2, \ldots, E$ **do**
    Compute gating weights $T_{v\to t}^{(e-1)}$, $T_{t\to v}^{(e-1)}$ via Eq. (10) using $\bar{r}_i^{(e-1)}$.
    **For each** mini-batch $\mathcal{B}$: forward through $\theta_t$ and $\phi_v$ to obtain $p_t^{(e)}$, $p_v^{(e)}$.
        Update banks: $\mathcal{B}_t^{\phi}[j] \leftarrow \phi_t^{(e)}(x_j)$, $\mathcal{B}_i^{p}[j] \leftarrow p_i^{(e)}(x_j)$ for $j \in \mathcal{B}$.
        Compute $\mathcal{L}_t^{(e)}$ and $\mathcal{L}_v^{(e)}$ via Eq. (30); update $\theta_t$ and $\boldsymbol{v}$.
    Compute $r_i^{(e)}$ from cached banks via Eq. (9) (no extra forward pass).
    Update EMA: $\bar{r}_i^{(e)} \leftarrow m^{(e)} \cdot \bar{r}_i^{(e-1)} + (1 - m^{(e)}) \cdot r_i^{(e)}$ for $i \in \{t, v\}$.
**end for**
**return** Adapted student model $\theta_t$.

**EMA Smoothing with Cosine Momentum Schedule.** To stabilize reliability estimates across epochs, we apply an exponential moving average:

$$\bar{r}_i^{(e)}(x) = m^{(e)} \cdot \bar{r}_i^{(e-1)}(x) + \big(1 - m^{(e)}\big) \cdot r_i^{(e)}(x), \tag{28}$$

where the momentum $m^{(e)}$ follows a cosine schedule that increases monotonically:

$$m^{(e)} = m_{\text{start}} + (m_{\text{end}} - m_{\text{start}}) \cdot \frac{1 - \cos\Big(\pi \cdot \frac{e}{E-1}\Big)}{2}. \tag{29}$$

This assigns low momentum in early epochs to allow rapid adaptation, and high momentum in later epochs to suppress noise once training stabilizes. At initialization, we set $\bar{r}_i^{(0)} = r_i^{(0)}$ directly from the warmup pass.

**Epoch-Indexed Training Objective.** Substituting $\bar{r}_i^{(e-1)}$ into the gating functions of Eq. (10), the full epoch-indexed objective becomes:

$$\mathcal{L}_t^{(e)} = \mathbb{E}_x\Big[T_{v\to t}^{(e-1)} \cdot D_{\text{KL}}\Big(\text{sg}(p_v^{(e)}) \,\|\, p_t^{(e)}\Big) + T_{t\to v}^{(e-1)} \cdot D_{\text{KL}}\Big(p_t^{(e-1)} \,\|\, p_t^{(e)}\Big)\Big] + \mathcal{R}(p_t^{(e)}), \tag{30}$$

$$\mathcal{L}_v^{(e)} = \mathbb{E}_x\Big[T_{t\to v}^{(e-1)} \cdot D_{\text{KL}}\Big(\text{sg}(p_t^{(e)}) \,\|\, p_v^{(e)}\Big) + T_{v\to t}^{(e-1)} \cdot D_{\text{KL}}\Big(p_v^{(e-1)} \,\|\, p_v^{(e)}\Big)\Big] + \mathcal{R}(p_v^{(e)}), \tag{31}$$

where $\mathcal{R}(p) = \lambda_{\text{IIC}}\mathcal{L}_{\text{IIC}} - \lambda_{\text{div}} H(\bar{p})$ denotes the batch-level regularization term. Crucially, the gating weights $T^{(e-1)}$ are derived from the *previous* epoch's reliability scores, preventing a circular dependency between the current supervision signal and the reliability estimate used to weight it. The complete procedure is summarized in Algorithm 1.

# D Additional Results

## D.1 Comparison Against Paper-Reported Best Results

Tables 8–11 compare `SafeCut` against the best results reported in each paper, rather than our reproduced numbers. Methods for which per-task results were not reported in the original paper are excluded from the respective tables. `SafeCut` achieves state-of-the-art performance on Office-31 (+0.6% over ProDe-V), Office-Home (+2.2% over ProDe-V), and DomainNet-126 (+0.8% over ProDe-V), winning the majority of per-task comparisons on each benchmark. On VisDA-C, `SafeCut` remains competitive (90.9% vs. 91.0% for ProDe-V), with per-class results detailed in Table 11. These results confirm that the improvements reported in the main text are not artifacts of favorable reproduction conditions, but reflect genuine gains over the published state of the art.

Table 8: Comparison against paper-reported best results on **Office-31** (Acc, %).

| Method | Venue | A→D | A→W | D→A | D→W | W→A | W→D | Avg. |
|---|---|---|---|---|---|---|---|---|
| Source | – | 79.1 | 76.5 | 59.9 | 95.5 | 61.4 | 98.8 | 78.5 |
| SHOT | ICML20 | 94.0 | 90.1 | 74.7 | 98.4 | 74.3 | 99.9 | 88.6 |
| NRC | NIPS21 | 96.0 | 90.8 | 75.3 | **99.0** | 75.0 | **100.0** | 89.4 |
| AaD | NIPS22 | 96.4 | 92.1 | 75.0 | **99.1** | 76.5 | **100.0** | 89.9 |
| CoWA | ICML22 | 94.4 | 95.2 | 76.2 | 98.5 | 77.6 | 99.8 | 90.3 |
| ELR | ICLR23 | 93.8 | 93.3 | 76.2 | 98.0 | 76.9 | **100.0** | 89.6 |
| DIFO-V | CVPR24 | 97.2 | 95.5 | 83.0 | 97.2 | 83.2 | 98.8 | 92.5 |
| ProDe-V | ICLR25 | 96.8 | **96.4** | 83.1 | 97.0 | 82.5 | 99.8 | 92.6 |
| Ours | NIPS26 | **98.6** | 95.2 | **83.9** | 98.2 | **83.3** | **100.0** | **93.2** |

Table 9: Comparison against paper-reported best results on **Office-Home** (Acc, %).

| Method | Venue | Ar→Cl | Ar→Pr | Ar→Rw | Cl→Ar | Cl→Pr | Cl→Rw | Pr→Ar | Pr→Cl | Pr→Rw | Rw→Ar | Rw→Cl | Rw→Pr | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Source | – | 43.6 | 67.0 | 73.9 | 49.9 | 60.2 | 62.6 | 51.7 | 40.9 | 72.6 | 64.1 | 46.3 | 78.1 | 59.2 |
| SHOT | ICML20 | 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 |
| NRC | NIPS21 | 57.7 | 80.3 | 82.0 | 68.1 | 79.8 | 78.6 | 65.3 | 56.4 | 83.0 | 71.0 | 58.6 | 85.6 | 72.2 |
| 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 |
| DIFO-V | CVPR24 | 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-V | ICLR25 | 72.7 | 92.3 | 90.5 | 82.5 | 91.5 | 90.7 | 82.5 | 72.5 | 90.8 | 83.0 | 72.6 | 92.2 | 84.5 |
| Ours | NIPS26 | **76.9** | **93.0** | **91.9** | **85.9** | **92.5** | **91.6** | **85.0** | **76.7** | **92.0** | **85.7** | **76.1** | **93.2** | **86.7** |

Table 10: Comparison against paper-reported best results on **DomainNet-126** (Acc, %).

| Method | Venue | 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 | – | 44.7 | 59.9 | 47.5 | 53.3 | 75.3 | 46.2 | 55.1 | 62.6 | 46.4 | 55.2 | 50.8 | 59.6 | 54.7 |
| ELR | ICLR23 | 77.2 | 91.2 | 68.5 | 79.8 | 92.7 | 67.5 | 75.6 | 82.2 | 65.7 | 85.1 | 77.6 | 89.3 | 79.4 |
| DIFO-V | CVPR24 | 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 |
| ProDe-V | ICLR25 | **83.2** | **92.4** | 79.0 | 85.0 | **92.3** | 79.3 | 85.5 | **83.1** | 79.1 | 85.5 | **83.4** | **92.4** | 85.0 |
| Ours | NIPS26 | 82.8 | 92.3 | **81.8** | **86.2** | 92.2 | **81.0** | **86.9** | 82.9 | **81.7** | **86.6** | 83.0 | 92.3 | **85.8** |

Table 11: Comparison against paper-reported best results on **VisDA-C** (per-class Acc, %).

| Method | Venue | plane | bcycl | bus | car | horse | knife | mcycl | person | plant | sktbrd | train | truck | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Source | – | 60.7 | 21.7 | 50.9 | 68.5 | 71.8 | 5.3 | 86.3 | 20.2 | 67.1 | 43.2 | 83.3 | 10.6 | 49.1 |
| SHOT | ICML20 | 94.3 | 88.5 | 80.1 | 57.3 | 93.1 | 94.9 | 80.7 | 80.3 | 91.5 | 89.1 | 86.3 | 58.2 | 82.9 |
| NRC | NIPS21 | 96.8 | 91.3 | 82.4 | 62.4 | 96.2 | 95.9 | 86.1 | 80.6 | 94.8 | 94.1 | 90.4 | 59.7 | 85.9 |
| AaD | NIPS22 | 97.4 | 90.5 | 80.8 | 76.2 | 97.3 | 96.1 | 89.8 | 82.9 | 95.5 | 93.0 | 92.0 | 64.7 | 88.0 |
| AdaCon | CVPR22 | 97.0 | 84.7 | 84.0 | 77.3 | 96.7 | 93.8 | 91.9 | 84.8 | 94.3 | 93.1 | 94.1 | 49.7 | 86.8 |
| 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 |
| PLUE | CVPR23 | 97.3 | **96.2** | **90.5** | **91.8** | 90.0 | 94.2 | 87.4 | **87.7** | **97.0** | 84.3 | 93.0 | **81.0** | 90.0 |
| DIFO-V | CVPR24 | 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-V | ICLR25 | 98.5 | 92.4 | 86.6 | 80.5 | **98.1** | 98.0 | 92.3 | 84.3 | 94.7 | 97.0 | 94.1 | 75.6 | **91.0** |
| Ours | NIPS26 | **98.7** | 91.9 | 86.8 | 82.3 | 97.6 | **98.1** | **94.1** | 86.8 | 94.6 | 94.8 | **94.5** | 70.9 | 90.9 |

### D.2 VisDA Full Results

Table 12: Full per-class results on VisDA-C (Synthesis → Real, acc, %).

| Method | Venue | plane | bcycl | bus | car | horse | knife | mcycl | person | plant | sktbrd | train | truck | Per-class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Source | – | 60.7 | 21.7 | 50.9 | 68.5 | 71.8 | 5.3 | 86.3 | 20.2 | 67.1 | 43.2 | 83.3 | 10.6 | 49.1 |
| SHOT | ICML20 | 94.9 | 86.9 | 79.7 | 54.8 | 93.6 | 94.1 | 78.5 | 81.4 | 90.6 | 90.0 | 84.5 | 58.3 | 82.3 |
| NRC | NIPS21 | 97.2 | 91.5 | 83.5 | 39.3 | 96.4 | 96.2 | 82.8 | 79.7 | 93.8 | 93.3 | 91.2 | 55.8 | 83.4 |
| AaD | NIPS22 | 97.5 | 91.8 | 70.6 | 45.5 | 92.3 | 90.0 | 65.6 | 65.6 | 95.0 | 88.4 | 87.5 | 56.0 | 78.8 |
| AdaCon | CVPR22 | 88.4 | 69.1 | 78.5 | 55.2 | 85.4 | 91.6 | 79.7 | 55.8 | 80.3 | 48.4 | 64.7 | 8.9 | 67.2 |
| CoWA | ICML22 | 95.5 | 84.2 | 83.0 | 68.9 | 94.3 | 96.9 | 88.8 | 83.1 | 91.3 | 89.0 | 85.5 | 51.5 | 84.3 |
| ELR | ICLR23 | 97.2 | 89.7 | 84.0 | 60.1 | 95.9 | 96.7 | 87.1 | 81.1 | 93.3 | 92.9 | 91.0 | 57.3 | 85.5 |
| PLUE | CVPR23 | 97.1 | 88.5 | 84.7 | 59.0 | 96.1 | 94.5 | 89.2 | **86.6** | 93.7 | 86.9 | 92.3 | 70.5 | 86.6 |
| DIFO-V | CVPR24 | 97.7 | 87.7 | 86.4 | 74.7 | 97.3 | **98.1** | 92.9 | 85.5 | 92.2 | **96.2** | 93.0 | 75.7 | 89.8 |
| ProDe-V | ICLR25 | 98.5 | **93.4** | 83.9 | 78.3 | **98.1** | 97.8 | 90.6 | 84.3 | 94.6 | 96.4 | 93.9 | **77.1** | 90.6 |
| Ours | NIPS26 | **98.6** | 92.1 | **87.3** | **83.3** | 97.8 | **98.1** | **94.1** | 86.0 | **94.8** | 94.6 | **94.5** | 69.5 | **90.9** |

Table 12 presents the full per-class breakdown on VisDA-C. `SafeCut` achieves the highest per-class average (90.9%), outperforming ProDe-V by 0.3%p despite falling behind on individual categories such as *bcycl*, *horse*, and *truck*. The largest gap occurs on *truck* (−7.6%p vs. ProDe-V), which we attribute to the visually ambiguous nature of the category under domain shift. Notably, `SafeCut` achieves strong gains on *car* (+5.0%p) and *bus* (+3.4%p) over ProDe-V, categories that are particularly sensitive to appearance-level domain shift, suggesting that reliability-gated mutual correction is especially effective where visual statistics differ significantly between domains.

### D.3 Computational Analysis

Table 13 reports the per-component runtime of `SafeCut` alongside the total wall-clock time of ProDe.[5] `SafeCut` maintains two models and computes per-sample reliability estimates via a mutual $k$-NN graph at each epoch, which introduces additional cost over a single-model baseline. However, two design choices keep this overhead manageable. First, the cut statistic is computed *once per epoch* from feature representations cached during the preceding optimization forward pass, avoiding any separate full forward pass over the target domain; the resulting per-epoch KNN cost remains below 0.5 s across all benchmarks. Second, target-domain visual features for the CLIP peer are extracted and cached prior to adaptation and reused throughout training via an EMA-updated reliability bank, eliminating repeated invocations of the CLIP backbone during training. As a result, despite the additional complexity of bidirectional mutual correction, `SafeCut` completes adaptation faster than ProDe across all four benchmarks.

Table 13: **Computational cost analysis between `SafeCut` (ours) and ProDe.**

| | `SafeCut` (Ours) | | | | **ProDe** |
|---|---|---|---|---|---|
| Dataset | Avg. KNN | Avg. Train | Avg. Infer. | Total | Total |
| Office-31 | 0.05 s | 8.92 s | 6.57 s | **0.78 h** | 1.11 h |
| Office-Home | <0.01 s | 17.85 s | 15.83 s | **5.62 h** | 5.87 h |
| DomainNet-126 | 0.25 s | 72.21 s | 44.86 s | **19.55 h** | 29.58 h |
| VisDA | 0.40 s | 160.80 s | 80.70 s | **2.02 h** | 3.49 h |

### D.4 Extended Empirical Validation of Cut Statistic Properties

We provide extended empirical validation of the two key properties underlying our framework: (i) approximate monotonicity of predictive accuracy with respect to the cut statistic (Observation 3.1), and (ii) spatial separation of complementary correction regions in the joint cut plane. We examine these properties across three settings: Office-Home with a CLIP peer (Figures 5, 8), DomainNet-126 with a CLIP peer (Figures 6, 9), and Office-Home with a heterogeneous dual-ResNet peer (Figures 7, 10).

**Approximate Monotonicity.** Across all three settings and all 12 adaptation tasks per setting, both models consistently exhibit approximately monotonically decreasing accuracy with respect to their own cut statistic $z$, confirming that Observation 3.1 holds broadly regardless of benchmark, peer

[5] All experiments are conducted on a single NVIDIA RTX PRO 6000 Blackwell Server Edition GPU (97 GB VRAM).

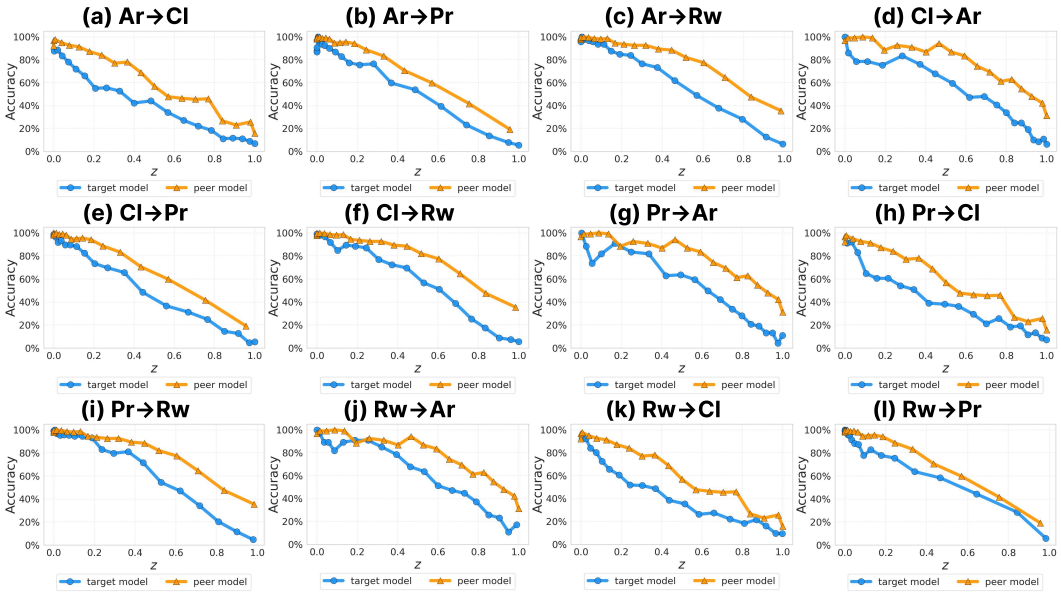



Figure 5: **Accuracy vs. cut statistic $z$ for all 12 Office-Home tasks (CLIP peer).**

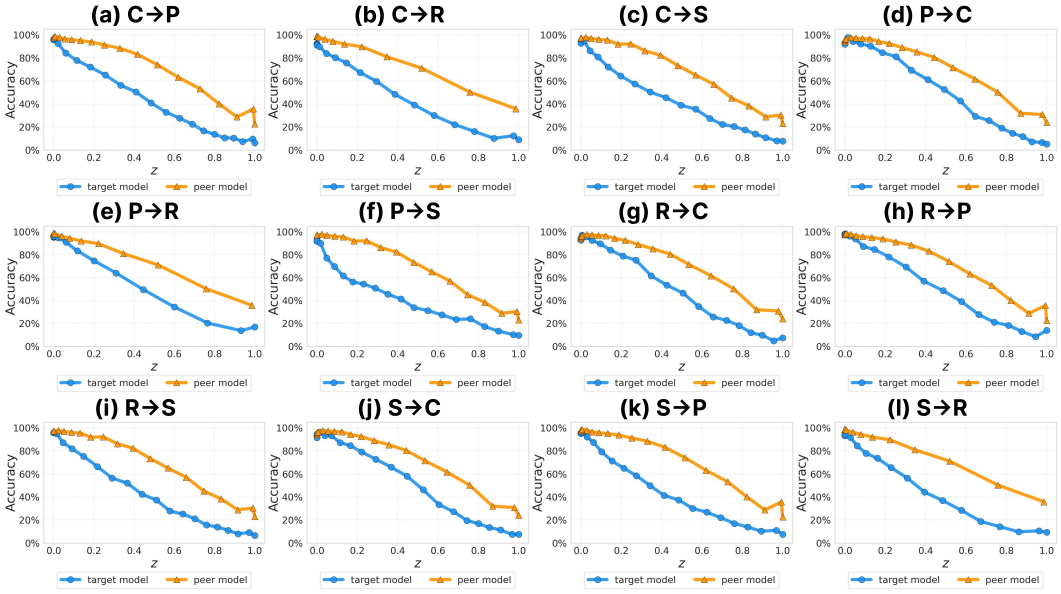



Figure 6: **Accuracy vs. cut statistic $z$ for all 12 DomainNet-126 tasks (CLIP peer).**

architecture, and inter-model accuracy gap. This robustness is most directly evidenced by contrasting the two peer configurations: in the CLIP-based settings (Figures 5, 6), $\theta_v$ maintains a large and persistent accuracy advantage over $\theta_t$ across the full range of $z$, whereas in the heterogeneous dual-ResNet setting (Figure 7), the curves of $\theta_1$ and $\theta_2$ run in close proximity and frequently cross at intermediate cut values. Despite these contrasting regimes – one with strong directional asymmetry and one with near-symmetric accuracy – the monotonic relationship between cut value and predictive accuracy is preserved in both cases, confirming that the cut statistic functions as a reliable per-model reliability proxy independent of the degree of inter-model asymmetry.

**Spatial Separation of Correction Regions.** Figures 8, 9, and 10 display KDE density plots of $\mathcal{S}_{c\bar{c}}$ (blue) and $\mathcal{S}_{\bar{c}c}$ (orange) in the joint cut plane. The red diagonal line marks $\Delta(x) = 0$, i.e., equal cut

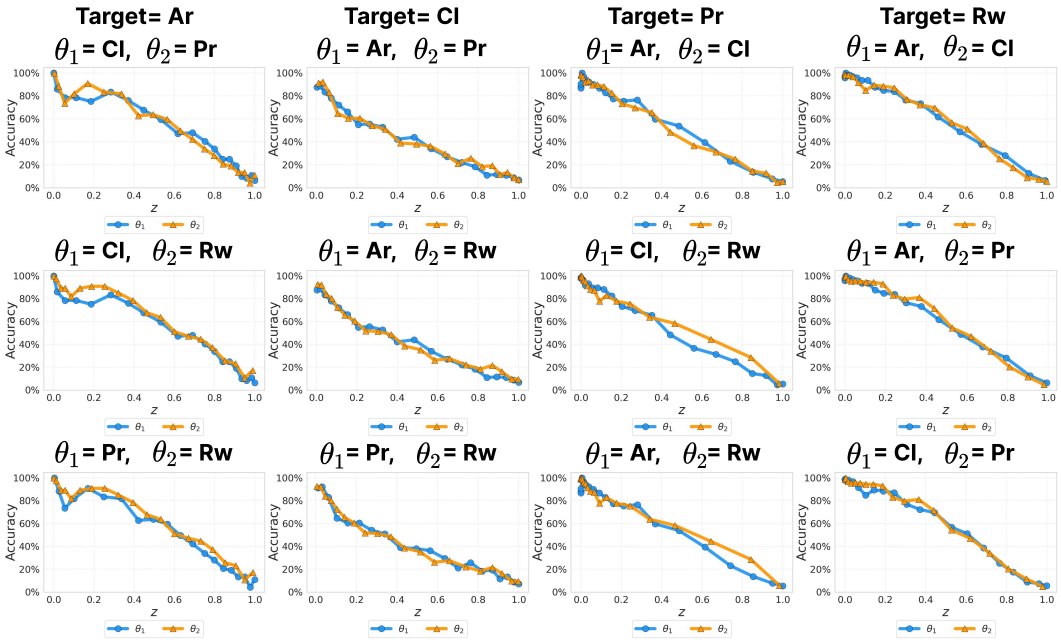



Figure 7: **Accuracy vs. cut statistic $z$ for all 12 Office-Home source-pair combinations (dual-ResNet peer).**

values for both models; the baseline thresholds $\gamma_0$ and $\xi_0$ from Proposition 3.1 reside in the upper and lower half-spaces relative to this line, respectively, and need not coincide with any fixed offset.

Across all three settings, the blue mode concentrates in the upper half-space ($\Delta(x) > 0$) and the orange mode in the lower half-space ($\Delta(x) < 0$), confirming that the two complementary correction regions are separated by the $\Delta(x) = 0$ boundary as predicted. The degree of this separation, however, varies with the choice of target domain. In most tasks, the two modes are pushed toward opposite extremes of the joint cut plane, yielding a wide effective margin and making the identification of $\gamma_0$ and $\xi_0$ straightforward. When **Art (Ar)** serves as the target domain, both distributions shift toward the red line with smaller $|\Delta(x)|$, compressing the separation. Nevertheless, the relative ordering of the two modes – blue above and orange below the $\Delta(x) = 0$ boundary – is preserved even in this compressed regime, ensuring that valid baseline thresholds $\gamma_0 > 0$ and $\xi_0 > 0$ satisfying Proposition 3.1 can still be identified within their respective half-spaces.

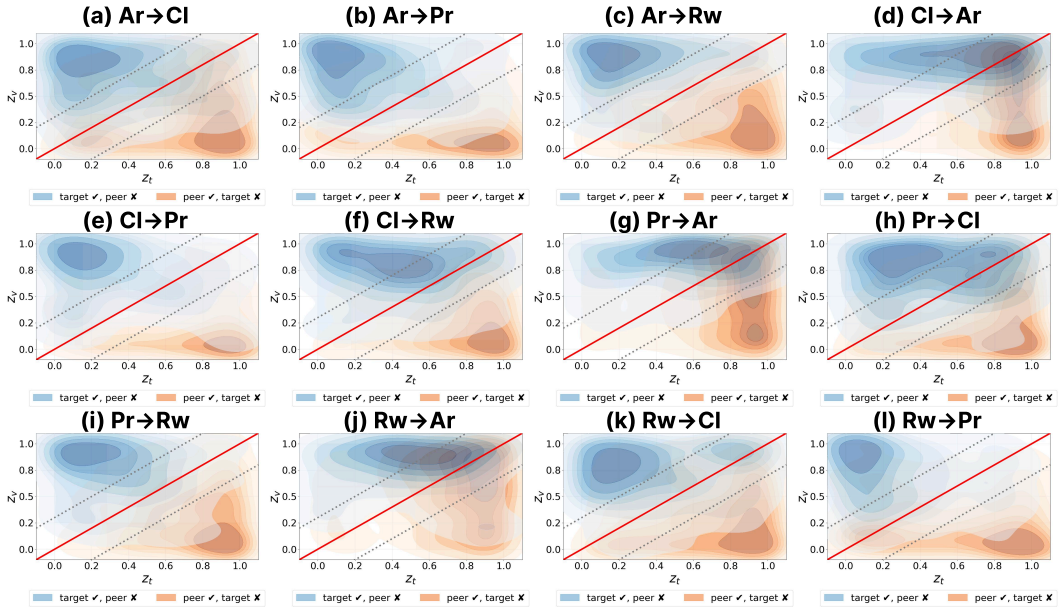



Figure 8: **Joint cut density of $\mathcal{S}_{c\bar{c}}$ and $\mathcal{S}_{\bar{c}c}$ for all 12 Office-Home tasks (CLIP peer).**

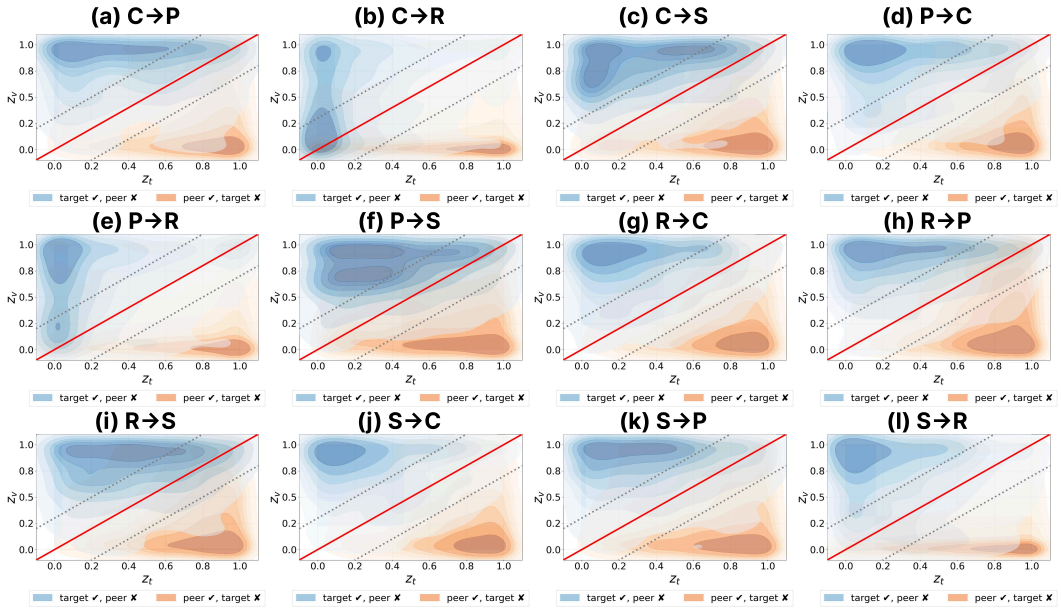



Figure 9: **Joint cut density of $\mathcal{S}_{c\bar{c}}$ and $\mathcal{S}_{\bar{c}c}$ for all 12 DomainNet-126 tasks (CLIP peer).**

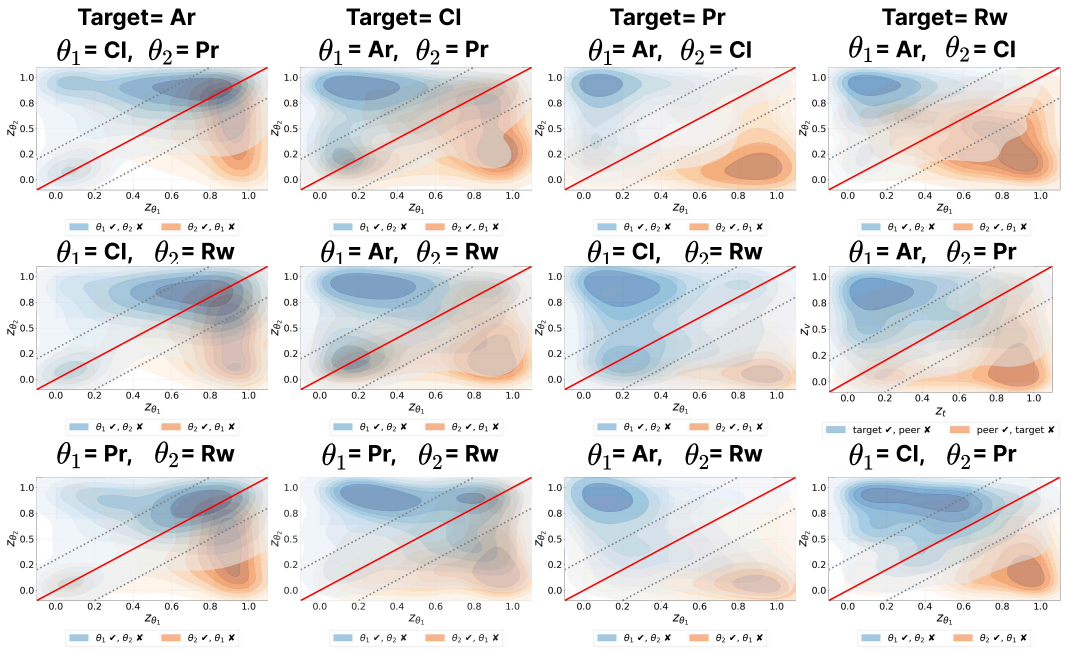



Figure 10: **Joint cut density of $\mathcal{S}_{c\bar{c}}$ and $\mathcal{S}_{\bar{c}c}$ for all 12 Office-Home source-pair combinations (dual-ResNet peer).**

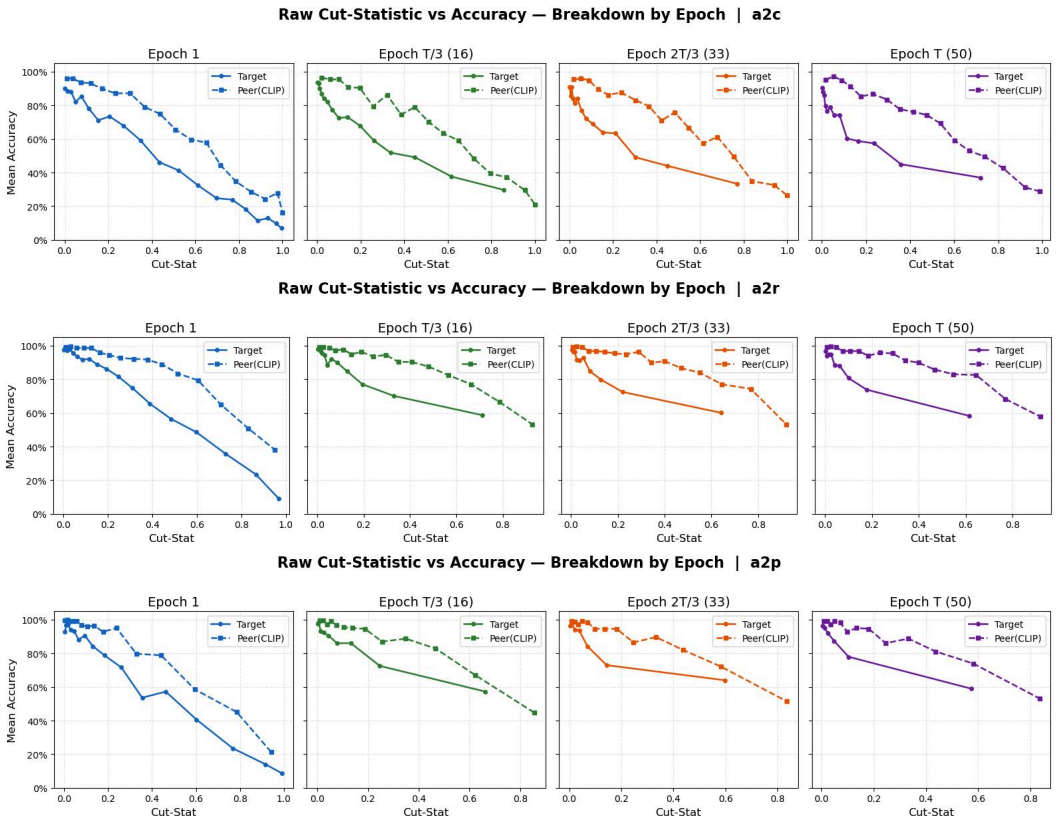



Figure 11: **Temporal stability of the cut–accuracy monotonicity.** Mean accuracy versus cut statistic $z$ for the target model (solid) and CLIP peer (dashed) at four checkpoints (Epoch 1, $T/3$, $2T/3$, $T$) on three Office-Home tasks. The monotonically decreasing trend and the relative ordering between the two curves are preserved throughout adaptation, while the target curve progressively converges toward the peer.

**Temporal Stability of Monotonicity.** Beyond the snapshot validation in Observation 3.1, a stronger property would be that the cut–accuracy monotonicity holds *throughout* adaptation—ensuring that the reliability scores $\bar{r}_i$ accumulated through EMA remain well-calibrated as both models evolve, and that reliability-gated supervision retains its fidelity precisely when mutual correction is most active. Figure 11 verifies this stronger property by tracking the cut–accuracy relationship across four representative checkpoints (Epoch 1, $T/3$, $2T/3$, and $T$) on three Office-Home tasks (Ar→Cl, Ar→Rw, Ar→Pr). Across every task and every checkpoint, both the target model and the CLIP peer maintain a clear monotonically decreasing trend, and the relative ordering of the two curves is consistently preserved. Even more encouragingly, as training progresses the target curve steadily shifts upward and tightens toward the peer curve, directly reflecting the joint improvement driven by reliability-gated mutual correction—while the qualitative cut–accuracy shape remains intact. These results establish the cut statistic as a temporally robust reliability proxy: its validity extends seamlessly from initialization through the entire training trajectory, providing a stable foundation on which reliability-aware supervision continues to deliver well-calibrated guidance throughout adaptation.

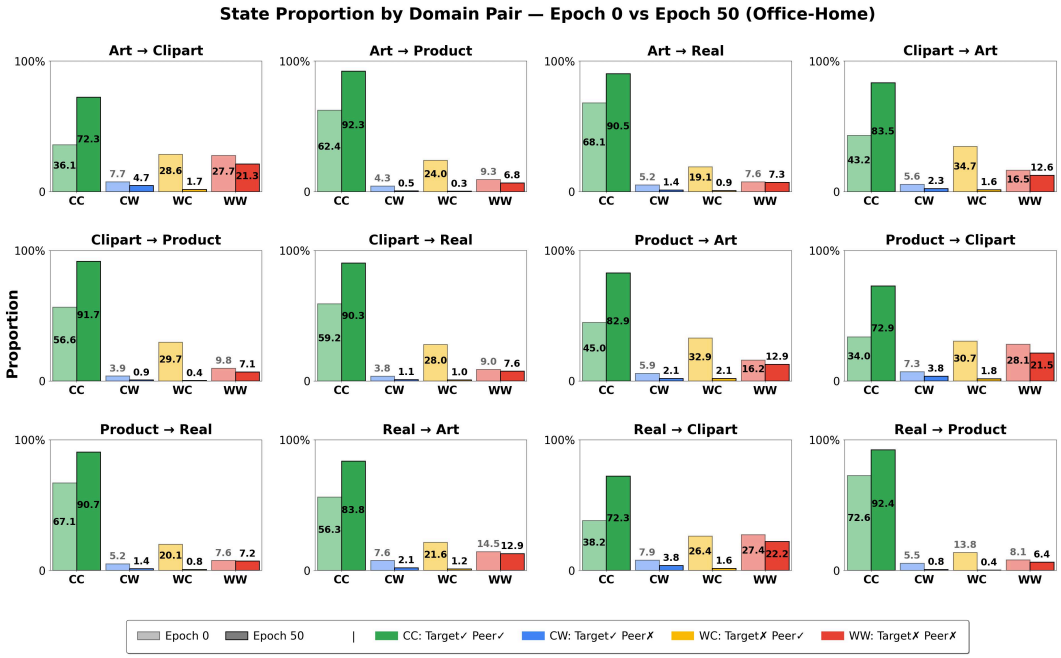



Figure 12: **Correctness pattern transition (Epoch 0 vs. 50) for all 12 Office-Home tasks (CLIP peer).**

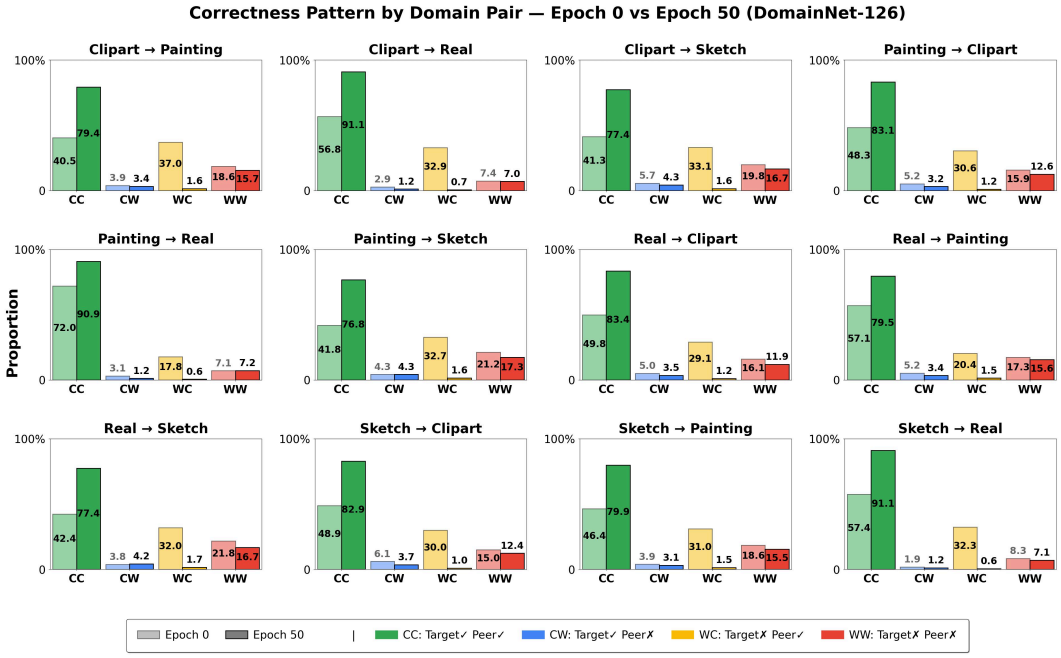



Figure 13: **Correctness pattern transition (Epoch 0 vs. 50) for all 12 DomainNet-126 tasks (CLIP peer).**

## D.5 Extended Correctness Pattern Transition Results

Figures 12, 13, and 14 report the correctness pattern transitions from Epoch 0 to Epoch 50 across all 12 adaptation tasks for Office-Home (CLIP peer), DomainNet-126 (CLIP peer), and Office-Home (dual-ResNet peer), respectively. Each bar group decomposes the target domain into the four oracle subsets – $\mathcal{S}_{cc}$, $\mathcal{S}_{c\bar{c}}$, $\mathcal{S}_{\bar{c}c}$, and $\mathcal{S}_{\bar{c}\bar{c}}$ – before and after adaptation.

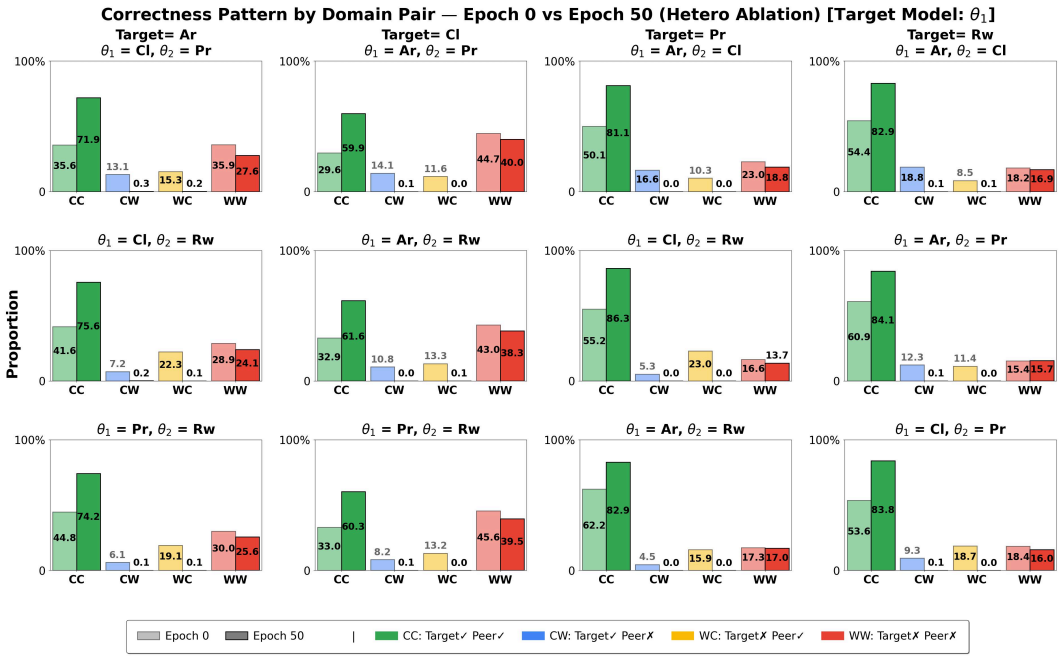



Figure 14: **Correctness pattern transition (Epoch 0 vs. 50) for all 12 Office-Home source-pair combinations (dual-ResNet peer, $\theta_1$).**

**Consistent Pattern Across All Settings.** Three transitions are observed consistently across all benchmarks and peer configurations. First, the jointly correct proportion $\mathcal{S}_{cc}$ increases substantially in every task, confirming that bidirectional knowledge transfer effectively drives both models toward agreement on correct predictions. Second, the one-sided correct sets $\mathcal{S}_{c\bar{c}}$ and $\mathcal{S}_{\bar{c}c}$ nearly vanish after adaptation – the populations that initially provided the source of mutual correction are progressively absorbed into $\mathcal{S}_{cc}$, indicating that cross-model supervision successfully resolves asymmetric errors in both directions. Third, the jointly incorrect set $\mathcal{S}_{\bar{c}\bar{c}}$ also exhibits a modest but consistent decrease across most tasks, suggesting that neighborhood-level generalization from corrected samples propagates improvement even to samples that neither model could directly supervise.

**Amplified Effect in the Dual-ResNet Setting.** While the above pattern holds in both peer configurations, two notable differences emerge in the dual-ResNet setting (Figure 14). First, the one-sided sets $\mathcal{S}_{c\bar{c}}$ and $\mathcal{S}_{\bar{c}c}$ collapse more sharply toward zero compared to the CLIP-based settings – in many source-pair combinations, both sets reduce to below 0.1% after adaptation, a more complete elimination than observed in the CLIP case. This more thorough resolution of asymmetric errors reflects the structural symmetry of the two ResNet models: since neither model maintains a persistent directional advantage, correction signals flow bidirectionally with equal effectiveness, leaving minimal residual one-sided errors. Second, despite starting from lower initial $\mathcal{S}_{cc}$ proportions – a consequence of both models being source-pretrained without the semantic priors of a vision-language model – the absolute gain in $\mathcal{S}_{cc}$ is comparably large, demonstrating that `SafeCut` can leverage complementary failure modes effectively even when the peer model is architecturally identical and initially less accurate than in the CLIP setting.

### D.6 Effect of Peer Complementarity

SafeCut relies on complementary predictions between the two peer models: when their failure modes differ, one model can provide corrective supervision on samples where the other model is unreliable. To quantitatively examine how this complementarity affects adaptation, we use the 12 dual-ResNet settings reported in Table 4. For each pair, we compute the prediction agreement before adaptation, defined as the proportion of target samples for which the two models predict the same class. We then measure the adaptation gain as the average accuracy improvement of the two models after `SafeCut` adaptation.

Table 14: Effect of peer complementarity in the dual-ResNet setting of Table 4. Agreement denotes the prediction agreement between the two models before adaptation, and Gain denotes their average accuracy improvement after SafeCut adaptation.

| Target | Source 1 | Source 2 | Agreement (%) | Gain (%p) |
|---|---|---|---|---|
| Cl | Ar | Pr | 41.1 | 17.6 |
| Ar | Cl | Pr | 43.3 | 21.3 |
| Cl | Ar | Rw | 46.1 | 16.7 |
| Cl | Pr | Rw | 46.4 | 16.8 |
| Ar | Cl | Rw | 48.4 | 18.7 |
| Ar | Pr | Rw | 53.6 | 16.5 |
| Pr | Ar | Cl | 57.1 | 17.7 |
| Rw | Cl | Pr | 58.8 | 16.3 |
| Rw | Ar | Cl | 59.1 | 14.9 |
| Pr | Cl | Rw | 60.6 | 17.2 |
| Rw | Ar | Pr | 67.3 | 11.0 |
| Pr | Ar | Rw | 69.6 | 10.4 |

We observe a strong negative relationship between pre-adaptation prediction agreement and the gain achieved by SafeCut (Pearson $r = -0.81$). Model pairs with lower agreement generally obtain larger improvements, supporting the premise that complementary predictions create more opportunities for cross-model correction. As the two models become more similar, the achievable benefit decreases, consistent with the fact that the one-sided correctness regions become less prevalent. Importantly, this reduction is gradual rather than harmful over the evaluated range. Even for the pair with the highest observed agreement (69.6%), `SafeCut` yields a +10.4 percentage-point improvement, and no performance degradation is observed across the 12 evaluated pairs. These results empirically support the role of peer complementarity in `SafeCut` while also indicating that the potential gain naturally diminishes as the two peers become increasingly similar.

# E Experimental Details

## E.1 Implementation and Hyperparameter Details

**Hyperparameters** Table 15 summarizes the hyperparameters used across all benchmarks. A single hyperparameter configuration is applied uniformly to all adaptation tasks within each dataset; no per-task search or adjustment is performed. The EMA schedule for cut statistic estimation is held strictly constant across all benchmarks, with $m_{\text{start}} = 0.7$ and $m_{\text{end}} = 0.95$ following the cosine annealing schedule of Equation (29), and the CLIP prompt is initialized uniformly as `a_photo_of_a` throughout. The gating sharpness $\beta$, the IIC weight $\lambda_{\text{iic}}$, and the neighborhood size $K$ are consistent across most benchmarks, with minor adjustments that follow intuitive dataset-scale considerations: smaller datasets (Office-31, Camelyon17) use a smaller $K$ due to reduced sample counts, and $\beta$ is slightly lower for Office-31 and Camelyon17 given their smaller scale and reduced domain complexity. The diversity weight $\lambda_{\text{div}}$ is the only parameter that varies more substantially across benchmarks, reflecting differences in class balance and dataset scale rather than task-specific tuning.

Table 15: Implementation details. Red: identical across all applicable benchmarks; underline: identical in most benchmarks.

| | O31 | OH | DN | VisDA | Cam17 |
|---|---|---|---|---|---|
| *Optimization* | | | | | |
| Epochs | 30 | 50 | 50 | 30 | 20 |
| Batch size | 64 | 192 | 192 | 192 | 128 |
| LR | 5e-3 | 1e-2 | 1e-2 | 5e-3 | 5e-3 |
| Prompt LR | 5e-3 | 5e-4 | 5e-4 | 1e-3 | 5e-4 |
| *Cut statistic* | | | | | |
| $K$ | 20 | 50 | 50 | 50 | 20 |
| EMA $m_{\text{start}}$ | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 |
| EMA $m_{\text{end}}$ | 0.95 | 0.95 | 0.95 | 0.95 | 0.95 |
| *Loss weights* | | | | | |
| $\beta$ | 3.0 | 6.0 | 6.0 | 6.0 | 5.0 |
| $\lambda_{\text{iic}}$ | 1.3 | 1.3 | 1.3 | 1.3 | 1.0 |
| $\lambda_{\text{div}}$ | 0.8 | 0.5 | 0.05 | 0.05 | 0.1 |
| *CLIP* | | | | | |
| Prompt init | `a_photo_of_a` | | | | |

**Baseline Reproduction.** All baselines are reproduced under a unified experimental setup following the protocol of ProDe [11], using the codebase and pretrained source models provided in the ProDe repository.[6] For methods not included in the ProDe repository, namely ELR [9] and AaD [5], we refer to their respective official implementations.[7] [8] Pretrained source models are taken directly from the ProDe repository without modification, ensuring that all methods are evaluated from identical initializations. All experiments are conducted using PyTorch with a single NVIDIA GPU [9].

## E.2 ProDe Hyperparameter Search on Camelyon17

As noted in Section 5, ProDe achieves only modest improvement on Camelyon17 despite extensive hyperparameter tuning. To ensure a fair comparison, we conduct a thorough grid search over all hyperparameters that could plausibly affect ProDe's performance in this setting, with the full search space summarized in Table 16. The CLIP prompt is initialized as `a_photo_of_a`, identical to the setting used in ProDe's original natural-image experiments. The best result obtained across all configurations is $76.27\%$ on the target model, with CLIP remaining near its zero-shot performance ($50.00\%$) regardless of the configuration. We attribute this performance ceiling to a fundamental limitation of ProDe's update structure in this setting, rather than suboptimal hyperparameter selection. ProDe supervises CLIP exclusively through $\mathcal{L}_{\text{IIC}}$ [20], a self-supervised clustering objective that operates without any source-driven signal. When the peer model starts from a poor initialization – as is the case for zero-shot CLIP on histopathology, where it underperforms even the source-pretrained model ($50.01\%$ vs. $64.14\%$) – $\mathcal{L}_{\text{IIC}}$ alone is insufficient to recover meaningful representations, leaving CLIP unable to provide reliable guidance throughout

Table 16: Hyperparameter search space for ProDe on Camelyon17.

| Hyperparameter | Search Space |
|---|---|
| Epochs $E$ | $\{1, 2, 3, 4, 5, 10\}$ |
| $\lambda_{\text{div}}$ | $\{0.05, 0.1\}$ |
| $\lambda_{\text{iic}}$ | $\{0.3, 0.5, 0.6, 0.7, 1.0\}$ |
| LR | $\{10^{-3}, 3{\times}10^{-3}, 5{\times}10^{-3}, 10^{-2}, 0.05, 0.06, \ldots, 0.25, 0.30, 0.35, 0.40\}$ |
| Prompt LR | $\{10^{-5}, 5{\times}10^{-5}, 10^{-4}, 5{\times}10^{-4}, 10^{-3}, 5{\times}10^{-3}, 0.01, 0.02, \ldots, 0.10\}$ |
| Prompt init | `a_photo_of_a` (fixed) |

[6] https://github.com/tntek/source-free-domain-adaptation
[7] https://github.com/liyi01827/SFDA_LLN
[8] https://github.com/Albert0147/AaD_SFDA
[9] NVIDIA RTX PRO 6000 Blackwell Server Edition GPU (97 GB VRAM).

training. This one-directional update structure lacks the mechanism to bootstrap a weak peer model, making the performance gap robust to hyperparameter variation.

### E.3 Full Per-Task Ablation Results

**Component Ablation (Table 17).** Table 17 reports the full per-task results for each loss configuration. For clarity, we present the explicit training objectives for each case.

$\mathcal{L}_{\mathrm{c}} = 0, \mathcal{L}_{\mathrm{a}} = 0$**:** Cross-model correction and temporal anchoring are both disabled, reducing the objective to batch-level regularization only:

$$\mathcal{L}_t = \lambda_{\mathrm{IIC}}\mathcal{L}_{\mathrm{IIC}} - \lambda_{\mathrm{div}}H(\bar{p}_t), \quad \mathcal{L}_v = \lambda_{\mathrm{IIC}}\mathcal{L}_{\mathrm{IIC}} - \lambda_{\mathrm{div}}H(\bar{p}_v). \tag{32}$$

$\mathcal{L}_{\mathrm{c}} = 0, \mathcal{L}_{\mathrm{a}} = 1$**:** Cross-model correction is disabled, but temporal anchoring is retained with uniform weighting (gate $\equiv 1$):

$$\mathcal{L}_t = \mathbb{E}_x[D_{\mathrm{KL}}(\tilde{p}_t \| p_t)] + \lambda_{\mathrm{IIC}}\mathcal{L}_{\mathrm{IIC}} - \lambda_{\mathrm{div}}H(\bar{p}_t), \tag{33}$$

$$\mathcal{L}_v = \mathbb{E}_x[D_{\mathrm{KL}}(\tilde{p}_v \| p_v)] + \lambda_{\mathrm{IIC}}\mathcal{L}_{\mathrm{IIC}} - \lambda_{\mathrm{div}}H(\bar{p}_v). \tag{34}$$

$\mathcal{L}_{\mathrm{c}} = 1, \mathcal{L}_{\mathrm{a}} = 0$**:** Cross-model correction is enabled with uniform weighting, while temporal anchoring is disabled:

$$\mathcal{L}_t = \mathbb{E}_x[D_{\mathrm{KL}}(\mathrm{sg}(p_v) \| p_t)] + \lambda_{\mathrm{IIC}}\mathcal{L}_{\mathrm{IIC}} - \lambda_{\mathrm{div}}H(\bar{p}_t), \tag{35}$$

$$\mathcal{L}_v = \mathbb{E}_x[D_{\mathrm{KL}}(\mathrm{sg}(p_t) \| p_v)] + \lambda_{\mathrm{IIC}}\mathcal{L}_{\mathrm{IIC}} - \lambda_{\mathrm{div}}H(\bar{p}_v). \tag{36}$$

$\mathcal{L}_{\mathrm{c}} = 1, \mathcal{L}_{\mathrm{a}} = 1$**:** Both cross-model correction and temporal anchoring are enabled with uniform weighting:

$$\mathcal{L}_t = \mathbb{E}_x[D_{\mathrm{KL}}(\mathrm{sg}(p_v) \| p_t) + D_{\mathrm{KL}}(\tilde{p}_t \| p_t)] + \lambda_{\mathrm{IIC}}\mathcal{L}_{\mathrm{IIC}} - \lambda_{\mathrm{div}}H(\bar{p}_t), \tag{37}$$

$$\mathcal{L}_v = \mathbb{E}_x[D_{\mathrm{KL}}(\mathrm{sg}(p_t) \| p_v) + D_{\mathrm{KL}}(\tilde{p}_v \| p_v)] + \lambda_{\mathrm{IIC}}\mathcal{L}_{\mathrm{IIC}} - \lambda_{\mathrm{div}}H(\bar{p}_v). \tag{38}$$

`SafeCut` **(Ours):** Both terms are enabled with adaptive reliability gating $T_{v\to t}$ and $T_{t\to v}$, replacing the uniform weights above:

$$\mathcal{L}_t = \mathbb{E}_x[T_{v\to t} \cdot D_{\mathrm{KL}}(\mathrm{sg}(p_v) \| p_t) + T_{t\to v} \cdot D_{\mathrm{KL}}(\tilde{p}_t \| p_t)] + \lambda_{\mathrm{IIC}}\mathcal{L}_{\mathrm{IIC}} - \lambda_{\mathrm{div}}H(\bar{p}_t), \tag{39}$$

$$\mathcal{L}_v = \mathbb{E}_x[T_{t\to v} \cdot D_{\mathrm{KL}}(\mathrm{sg}(p_t) \| p_v) + T_{v\to t} \cdot D_{\mathrm{KL}}(\tilde{p}_v \| p_v)] + \lambda_{\mathrm{IIC}}\mathcal{L}_{\mathrm{IIC}} - \lambda_{\mathrm{div}}H(\bar{p}_v). \tag{40}$$

`SafeCut` consistently outperforms all ablated variants across all three benchmarks, confirming the complementary roles of reliability-aware cross-model correction, temporal anchoring.

**Design Variant Comparison (Table 18).** Table 18 compares three design variants against `SafeCut` and an oracle upper bound. **Frozen** $\theta_v$ holds the peer model fixed throughout adaptation (i.e., $\mathcal{L}_v \equiv 0$), so that $\theta_v$ remains at its zero-shot initialization. The consistent performance drop confirms that joint peer adaptation is essential: a static peer provides increasingly stale supervision as the target model evolves. **Entropy Gating** replaces the cut-based reliability weights with entropy-derived gates:

$$T_i^{\mathrm{ent}}(x) = \exp(-H(p_i(x))), \quad i \in \{t, v\}, \tag{41}$$

where $H(p_i(x)) = -\sum_c p_i^c(x) \log p_i^c(x)$. While entropy captures per-sample confidence from the marginal output distribution, it does not incorporate local neighborhood structure. The consistent underperformance relative to `SafeCut` supports the cut statistic as a more reliable reliability proxy. **Oracle Partition** $\mathcal{S}$ provides an upper bound by assuming access to ground-truth correctness labels, applying cross-model supervision exclusively via cross-entropy on the subset where the supervising model is known to be correct:

$$\mathcal{L}_t^{\mathrm{oracle}} = \mathbb{E}_{x\in\mathcal{S}_{\bar{c}c}}\left[-\sum_c \mathbb{1}[c = \hat{y}_v(x)] \log p_t^c(x)\right], \tag{42}$$

$$\mathcal{L}_v^{\mathrm{oracle}} = \mathbb{E}_{x\in\mathcal{S}_{c\bar{c}}}\left[-\sum_c \mathbb{1}[c = \hat{y}_t(x)] \log p_v^c(x)\right]. \tag{43}$$

The gap between `SafeCut` and Oracle $\mathcal{S}$ (e.g., $86.7\%$ vs. $93.3\%$ on Office-Home) establishes a concrete upper bound for reliability-gated mutual correction, and suggests that more precise approximations of the oracle partition remain a promising direction for future work.

Table 17: Full per-task component ablation results (acc, %). $\mathcal{L}_c$: cross-supervision KL, $\mathcal{L}_a$: temporal anchoring.

**(a) Office-31**

| Method | A→D | A→W | D→A | D→W | W→A | W→D | Avg. |
|---|---|---|---|---|---|---|---|
| $\mathcal{L}_c=0, \mathcal{L}_a=0$ | 97.2 | 91.7 | 81.2 | 97.4 | 80.6 | 99.0 | 91.2 |
| $\mathcal{L}_c=0, \mathcal{L}_a=1$ | 97.2 | 92.2 | 80.9 | 97.5 | 79.6 | 99.8 | 91.2 |
| $\mathcal{L}_c=1, \mathcal{L}_a=0$ | 97.0 | 94.2 | 83.7 | 96.7 | **84.2** | 98.8 | 92.4 |
| $\mathcal{L}_c=1, \mathcal{L}_a=1$ | 96.2 | 94.0 | 83.6 | 96.5 | 83.5 | 99.0 | 92.1 |
| SafeCut | **98.6** | **95.2** | **83.9** | **98.2** | 83.3 | **100.0** | **93.2** |

**(b) Office-Home**

| Method | Ar→Cl | Ar→Pr | Ar→Rw | Cl→Ar | Cl→Pr | Cl→Rw | Pr→Ar | Pr→Cl | Pr→Rw | Rw→Ar | Rw→Cl | Rw→Pr | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{L}_c=0, \mathcal{L}_a=0$ | 69.6 | 84.5 | 90.3 | 80.8 | 85.6 | 89.7 | 78.2 | 68.9 | 89.5 | 83.3 | 70.4 | 90.6 | 81.8 |
| $\mathcal{L}_c=0, \mathcal{L}_a=1$ | 71.2 | 88.2 | 90.7 | 81.8 | 90.3 | 90.0 | 79.6 | 71.1 | 89.5 | 83.4 | 71.8 | 92.2 | 83.3 |
| $\mathcal{L}_c=1, \mathcal{L}_a=0$ | 74.7 | 92.7 | 91.7 | 85.2 | **92.7** | 91.4 | 84.6 | 75.8 | 91.4 | 84.5 | 74.4 | 92.7 | 86.0 |
| $\mathcal{L}_c=1, \mathcal{L}_a=1$ | 76.2 | 92.5 | 91.5 | 85.7 | 92.7 | 91.3 | **85.0** | 76.1 | 91.3 | 84.8 | 75.1 | 92.5 | 86.2 |
| SafeCut | **76.9** | **93.0** | **91.9** | **85.9** | 92.5 | **91.6** | **85.0** | **76.7** | **92.0** | **85.7** | **76.1** | **93.2** | **86.7** |

**(c) DomainNet-126**

| Method | 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. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{L}_c=0, \mathcal{L}_a=0$ | 63.1 | 78.2 | 62.5 | 71.2 | 86.5 | 69.0 | 71.1 | 72.0 | 64.9 | 73.2 | 72.0 | 81.8 | 72.1 |
| $\mathcal{L}_c=0, \mathcal{L}_a=1$ | 71.3 | 85.9 | 70.9 | 78.9 | 89.0 | 73.8 | 81.1 | 78.8 | 73.8 | 79.6 | 74.9 | 86.1 | 78.7 |
| $\mathcal{L}_c=1, \mathcal{L}_a=0$ | 80.6 | 91.7 | 78.9 | 83.3 | 91.5 | 78.0 | 83.9 | 80.8 | 78.2 | 83.9 | 80.9 | 91.8 | 83.6 |
| $\mathcal{L}_c=1, \mathcal{L}_a=1$ | 81.7 | 91.9 | 79.3 | 84.2 | 92.0 | 79.1 | 84.2 | 81.6 | 79.1 | 84.3 | 81.9 | 92.1 | 84.3 |
| SafeCut | **82.8** | **92.3** | **81.8** | **86.2** | **92.2** | **81.0** | **86.9** | **82.9** | **81.7** | **86.6** | **83.0** | **92.3** | **85.8** |

**CLIP Adaptation.** Table 19 compares zero-shot CLIP against the peer model adapted within SafeCut across all tasks. Despite zero-shot CLIP's strong vision-language pretraining, prompt learning within SafeCut consistently and substantially improves peer accuracy across all benchmarks: $+13.4\%$ on Office-31 ($79.4\% \rightarrow 92.8\%$), $+7.7\%$ on Office-Home ($79.0\% \rightarrow 86.7\%$), and $+3.8\%$ on DomainNet-126 ($80.2\% \rightarrow 84.0\%$). These gains confirm that target-domain specialization of the peer model is essential: an adapted peer provides higher-fidelity supervision signals throughout mutual correction, and the improvement in peer accuracy directly translates to stronger correction of the target model.

Table 18: Full per-task variant comparison results (acc, %).

**(a) Office-31**

| Variant | A→D | A→W | D→A | D→W | W→A | W→D | Avg. |
|---|---|---|---|---|---|---|---|
| Frozen $\theta_v$ | 90.8 | 90.8 | 81.5 | 94.3 | 81.3 | 95.4 | 89.0 |
| Entropy Gating | 97.8 | 94.2 | **83.9** | 97.5 | **83.6** | 98.4 | 92.5 |
| `SafeCut` | **98.6** | **95.2** | **83.9** | **98.2** | 83.3 | **100.0** | **93.2** |
| Oracle partition $\mathcal{S}$ | 99.4 | 99.9 | 88.8 | 99.3 | 88.8 | 100.0 | 96.0 |

**(b) Office-Home**

| Variant | Ar→Cl | Ar→Pr | Ar→Rw | Cl→Ar | Cl→Pr | Cl→Rw | Pr→Ar | Pr→Cl | Pr→Rw | Rw→Ar | Rw→Cl | Rw→Pr | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Frozen $\theta_v$ | 73.5 | 90.4 | 90.9 | 84.2 | 90.7 | 90.6 | 83.6 | 73.4 | 91.1 | 84.1 | 74.4 | 90.8 | 84.8 |
| Entropy Gating | 75.7 | 92.7 | **92.0** | 85.5 | 92.3 | **91.7** | **85.0** | 76.1 | 91.8 | 85.1 | 75.7 | 92.3 | 86.3 |
| `SafeCut` | **76.9** | **93.0** | **92.0** | **85.9** | **92.5** | 91.6 | **85.0** | **76.7** | **92.0** | **85.7** | **76.1** | **93.2** | **86.7** |
| Oracle partition $\mathcal{S}$ | 88.0 | 97.0 | 95.8 | 91.9 | 97.1 | 95.7 | 92.0 | 87.8 | 95.8 | 93.0 | 88.2 | 97.3 | 93.3 |

**(c) DomainNet-126**

| Variant | 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. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Frozen $\theta_v$ | 81.8 | 91.7 | 80.0 | 84.7 | 91.7 | 80.0 | 84.6 | 81.9 | 79.5 | 85.0 | 81.8 | 91.8 | 84.6 |
| Entropy Gating | 82.4 | 92.2 | 80.2 | 85.1 | **92.2** | 80.2 | 85.2 | 82.5 | 80.2 | 84.9 | 82.4 | 92.2 | 85.0 |
| `SafeCut` | **82.8** | **92.3** | **81.8** | **86.2** | **92.2** | **81.0** | **86.9** | **82.9** | **81.7** | **86.6** | **83.0** | **92.3** | **85.8** |
| Oracle partition $\mathcal{S}$ | 91.2 | 96.1 | 90.2 | 93.4 | 96.0 | 90.2 | 93.3 | 91.2 | 90.1 | 93.5 | 91.2 | 96.1 | 92.7 |

Table 19: Full per-task CLIP adaptation results (acc, %). Zero-shot: CLIP without adaptation. Adaptation: CLIP adapted as peer model in `SafeCut`.

**(a) Office-31**

| | A→D | A→W | D→A | D→W | W→A | W→D | Avg. |
|---|---|---|---|---|---|---|---|
| Zero-shot | 81.1 | 80.0 | 77.0 | 80.0 | 77.0 | 81.1 | 79.4 |
| Adaptation | **98.6** | **94.8** | **83.5** | **98.0** | **82.0** | **99.8** | **92.8** |

**(b) Office-Home**

| | Ar→Cl | Ar→Pr | Ar→Rw | Cl→Ar | Cl→Pr | Cl→Rw | Pr→Ar | Pr→Cl | Pr→Rw | Rw→Ar | Rw→Cl | Rw→Pr | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Zero-shot | 64.6 | 86.4 | 87.2 | 77.9 | 86.4 | 87.2 | 77.9 | 64.6 | 87.2 | 77.9 | 64.6 | 86.4 | 79.0 |
| Adaptation | **76.9** | **93.0** | **91.9** | **85.9** | **92.5** | **91.6** | **85.0** | **76.7** | **92.0** | **85.7** | **76.1** | **93.2** | **86.7** |

**(c) DomainNet-126**

| | 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. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Zero-shot | 77.5 | 89.8 | 74.5 | 78.9 | 89.8 | 74.5 | 78.9 | 77.5 | 74.5 | 78.9 | 77.5 | 89.8 | 80.2 |
| Adaptation | **81.1** | **91.8** | **79.1** | **84.3** | **91.6** | **78.5** | **84.6** | **80.9** | **79.2** | **84.2** | **81.5** | **91.8** | **84.0** |

## F Limitations and Future Directions

SafeCut establishes a principled foundation for reliability-gated mutual correction, while several directions remain open for further exploration. The framework benefits from complementary prediction patterns between the two peers, while increasingly similar predictions provide fewer opportunities for cross-model correction. Samples on which both models are incorrect also remain comparatively challenging, as neither peer provides an immediately corrective signal, although some of these samples are gradually improved through changes in the surrounding neighborhood structure. These observations suggest several promising extensions. In particular, samples in low-margin regions could benefit from complementary reliability cues, such as prediction entropy or temporal consistency, to further sharpen directional discrimination. Incorporating additional pseudo-label-based strategies [31, 32] could also provide richer supervisory signals during adaptation and potentially move performance closer to the oracle correction bound identified in Table 5(b). We leave these directions as future work.

## G Broader Impacts

SafeCut contributes to the broader goal of deploying machine learning models reliably across diverse real-world environments without requiring labeled target data or access to sensitive source data. This has direct positive implications in privacy-sensitive domains such as healthcare and finance, where sharing training data is often legally or ethically restricted. By enabling effective adaptation under such constraints, SafeCut lowers the barrier to deploying high-quality models in underserved or data-scarce settings. Beyond privacy, the source-free constraint also reduces the computational and logistical costs associated with retaining and transferring large-scale source datasets, making model deployment more accessible to practitioners with limited infrastructure. Furthermore, the mutual correction framework introduced in SafeCut is model-agnostic and does not assume a specific type of peer model, broadening its applicability to heterogeneous deployment scenarios where diverse pretrained models are increasingly available. Finally, by establishing a theoretically grounded reliability criterion for cross-model supervision, SafeCut contributes to the broader agenda of trustworthy machine learning, providing principled mechanisms to mitigate confirmation bias and error propagation in self-supervised adaptation pipelines. We do not foresee significant negative societal impacts from this work. However, as with any adaptation method that operates without human supervision, there remains a risk that errors shared across both peer models go undetected, potentially leading to incorrect predictions in high-stakes applications such as medical diagnosis. Practitioners deploying SafeCut in such settings should complement adaptation with appropriate human oversight and validation protocols.

## H Asset Licenses and Terms of Use

**Datasets.** Office-31 [21] does not specify an explicit license and is used in this work solely for academic research purposes, with proper attribution to the original authors. Office-Home [22] is distributed under a fair-use notice for non-commercial research and educational purposes. DomainNet-126 [23] contains copyrighted material distributed under a fair-use notice for non-commercial research and educational purposes; uses beyond this scope may require permission from the respective copyright holders. VisDA [24] is used for non-commercial research and educational purposes in accordance with its usage terms. Camelyon17 from the WILDS benchmark [28] is distributed under the Creative Commons CC0 1.0 Universal license (public domain). EyePACS [29] is used for academic research with attribution to EyePACS, as requested by the dataset provider for research and publication use. APTOS 2019 is used under the competition rules permitting non-commercial use of the competition data for academic research and education.

**Models.** CLIP [26] is released by OpenAI under the MIT license. The ResNet-50 and ResNet-101 implementations used in our experiments are provided via torchvision, which is released under the BSD 3-Clause license. CoOp [27] is released under the MIT license.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly present the proposed concept of reliability-gated mutual correction between heterogeneous peer models, and all claims are substantiated by the theoretical analysis and experimental results provided in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in a dedicated appendix section (Section F), along with future directions for addressing them.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All assumptions and proofs are provided in full in Appendix A. Specifically, Proposition 3.1 and Theorem 3.1 are proved under strict monotonicity, and Appendix A.1 provides extended versions under the noisy monotonicity assumption(Assumption A.1), establishing quantitative lower bounds on the correction dominance margin even in the presence of sample-wise noise.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Full implementation details and hyperparameter configurations are provided in Appendix E.1, and all baseline reproductions follow the unified protocol described therein. To further support reproducibility, we provide an anonymous code repository linked in the abstract.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All datasets used in this paper are publicly available and referenced in Appendix B. The full experimental protocol, including hyperparameter configurations and baseline reproduction details, is documented in Appendix E.1. Code is made available via an anonymous repository linked in the abstract.(updated to a public repository for the camera-ready version).

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All training and evaluation details, including hyperparameter configurations, and hardware specifications, are documented in Appendix E.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Statistical robustness is demonstrated through consistent improvements across a large number of adaptation tasks (6, 12, 12, and 12 tasks on Office-31, Office-Home, DomainNet-126, and VisDA-C, respectively), confirming that the reported gains are not artifacts of specific domain pairs or benchmarks.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments are conducted on a single NVIDIA RTX PRO 6000 Blackwell Server Edition GPU (97 GB VRAM), with per-dataset wall-clock runtime reported in Appendix D.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work uses only publicly available datasets for non-commercial research purposes, involves no human subjects, and poses no foreseeable risks of harm, discrimination, or misuse.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Both positive and negative societal impacts are discussed in Appendix G.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper proposes a domain adaptation method and does not release any pretrained generative models, scraped datasets, or other artifacts that pose a high risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Licenses and terms of use for all datasets and models are documented in Appendix H, and codebases used for baseline reproduction are credited via footnotes in Appendix E.1.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: This paper does not introduce new datasets, model weights, or other standalone assets beyond the proposed method itself.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This paper involves no crowdsourcing experiments or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper involves no research with human subjects and therefore requires no IRB approval or equivalent review.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: This paper does not use LLMs as a core methodological component. LLMs were used solely for writing and editing assistance, which does not impact the scientific methodology or originality of the research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.