BEYOND PSEUDO-LABELING: RETRIEVAL-AUGMENTED VISUAL TUNING FOR SEMI-SUPERVISED LEARNING IN THE WILD

Anonymous authors
Paper under double-blind review

ABSTRACT

Semi-supervised learning (SSL) commonly exploits unlabeled data via confidence-thresholded pseudo-labeling, assuming that high confidence implies correctness. This assumption often fails in *in-the-wild* regimes with strong distribution mismatch, where thresholding induces *two-sided* selection bias (admitting high-confidence errors while discarding informative low-confidence samples) and shift further aggravates confidence *miscalibration*. Their coupling makes single-threshold heuristics brittle, leading to confirmation bias and even the "unlabeled hurts" phenomenon. We propose **Retrieval-Augmented Visual Tuning (RAVT)**, which replaces pointwise pseudo-label acceptance with *evidence-backed neighborhoods* conditioned on labeled queries. RAVT constructs reliable neighborhoods through hybrid retrieval that combines global semantic recall with patch-level ranking, and strengthens semantic consistency using implicit textual cues; learning is driven by non-local neighborhood evidence and followed by supervised anchored tuning to stabilize representations. Across four progressively harder mismatch settings, RAVT is particularly effective in in-the-wild regimes, including cross-dataset and multi-source heterogeneous unlabeled pools, where it consistently mitigates negative transfer and delivers more stable cross-distribution generalization, achieving substantial gains over strong pseudo-labeling baselines; in standard near-i.i.d. SSL settings, RAVT can also serve as a plug-in component for existing pseudo-labeling methods and further improve their performance.

1 INTRODUCTION

Deep learning has achieved strong performance in computer vision, but this progress is largely enabled by massive labeled datasets whose collection is expensive and slow (Deng et al., 2009). Semi-supervised learning (SSL) alleviates this bottleneck by leveraging a small labeled set together with abundant unlabeled data. A dominant line of SSL methods, such as FixMatch (Sohn et al., 2020) and its variants (Zhang et al., 2021; Wang et al., 2022), relies on *pseudo-labeling*: the model predicts labels for unlabeled samples, retains predictions above a confidence threshold, and uses them as additional supervision. While effective on curated i.i.d.-like benchmarks, this paradigm often breaks down in real-world *in-the-wild* settings, where unlabeled data can exhibit extreme distribution shift.

We attribute this reliability gap to a core weakness of confidence-based pseudo-labeling selection under distribution shift. First, confidence thresholding induces **intrinsic selection bias**: it can both admit high-confidence yet incorrect samples, such as open-set or boundary cases, and discard low-confidence but correct samples, such as long-tailed or shifted-but-correct instances.

Second, mismatch aggravates **miscalibration**, where the model assigns spuriously high probabilities to incorrect predictions, so confidence ceases to reliably reflect correctness (Guo et al., 2017). Importantly, these two failure modes are coupled in thresholding: tuning the threshold both changes which unlabeled samples contribute to training and presumes that confidence faithfully tracks correctness. As a result, confidence-based heuristics cannot jointly mitigate two-sided selection bias and miscalibration under shift, so self-training tends to reinforce confidently wrong pseudo-labels while ignoring informative hard or rare samples, amplifying confirmation bias (Arazo et al., 2019) and harming cross-distribution generalization.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This motivates a more controllable alternative to global threshold filtering, especially for in-the-wild unlabeled distribution mismatch: *constructing neighborhood-level evidence conditioned on labeled queries.* Instead of treating unlabeled data as pointwise pseudo-label targets, we view the unlabeled pool as a source of supportive evidence. Concretely, we use each labeled sample as a query to retrieve semantically related neighbors from the unlabeled set, and then form learning signals through non-local evidence aggregation and neighborhood consistency modeling. This shifts supervision from isolated high-confidence predictions to *neighborhood evidence*.

Based on these insights, we propose **Retrieval-Augmented Visual Tuning (RAVT)**, which exploits unlabeled data through evidence-backed neighborhoods rather than pointwise pseudo-label acceptance. In Stage I, RAVT retrieves a semantically coherent neighborhood for each labeled query using hybrid retrieval (Yates et al., 2021), including global semantic recall and patch-level ranking, and enforces semantic consistency with implicit textual cues, producing a retrieval-augmented representation for adapter updates. This design reduces appearance-driven false neighbors and makes the training signal depend on neighborhood support instead of a single model confidence score. In Stage II, we perform supervised anchored tuning on labeled data to stabilize representations and suppress neighborhood-induced drift.

We evaluate RAVT under four progressively more challenging settings, spanning standard SSL to strong distribution shift and multi-source heterogeneous scenarios. The results reveal a clear contrast: when unlabeled data deviates substantially from the labeled distribution, either under single-source shift or multi-source mixtures, confidence-based pseudo-labeling approaches often fail to provide consistent gains and can even exhibit *"unlabeled hurts"*. In contrast, by constructing neighborhood-level evidence with semantic alignment constraints, RAVT substantially mitigates degradation under shift and delivers more stable and consistent cross-distribution generalization. In standard near-i.i.d. SSL settings, RAVT can also be used as a plug-in component for existing pseudo-labeling methods, improving their performance while preserving their original training recipes.

We summarize our contributions as follows:

- We analyze why confidence-based pseudo-labeling degrades under distribution shift via two coupled factors: intrinsic selection bias and miscalibration, whose interaction amplifies confirmation bias.
- We propose RAVT, a retrieval-augmented SSL framework that replaces pointwise confidence filtering with neighborhood evidence conditioned on labeled queries, enabled by hybrid retrieval and semantic alignment for robust unlabeled utilization.
- We devise a four-level evaluation suite with increasing mismatch difficulty, showing that RAVT is effective under *in-the-wild* shift and can be seamlessly combined with pseudo-labeling to further improve performance.

2 Related Works

Semi-Supervised Learning (SSL) is largely built on *pseudo-labeling* (Lee et al., 2013) and *consistency regularization* (Laine & Aila, 2016). A dominant paradigm uses confidence-based pseudo-labels to supervise strongly augmented unlabeled samples, as exemplified by FixMatch (Sohn et al., 2020). Subsequent methods mainly improve sample selection or weighting, including FlexMatch (Zhang et al., 2021), FreeMatch (Wang et al., 2022), SoftMatch (Chen et al., 2023), as well as recent matching strategies and adapter tuning approaches such as MultiMatch (Sirbu et al., 2025) and V-PET (Zhang et al., 2025). However, these methods still largely rely on confidence as a proxy for pseudo-label correctness, which becomes unreliable under distribution mismatch. In contrast, we treat unlabeled samples as *support evidence* for labeled queries, retrieving semantically related neighborhoods and constructing supervision from aggregated neighborhood evidence rather than global confidence filtering, enabling more robust use of shifted or heterogeneous unlabeled data.

Retrieval-Augmented Learning. Information retrieval (IR) provides a principled way to select relevant evidence from large corpora. BM25 (Robertson et al., 1995), a representative classical ranker, remains a strong retrieval baseline for estimating query–document relevance. Building on this idea, retrieval-augmented learning has become prevalent in NLP and multimodal systems, where models query external memories or knowledge bases for complementary context (Long et al., 2022; Chen et al., 2022; Marino et al., 2021; Yasunaga et al., 2022). Our work follows this paradigm but targets

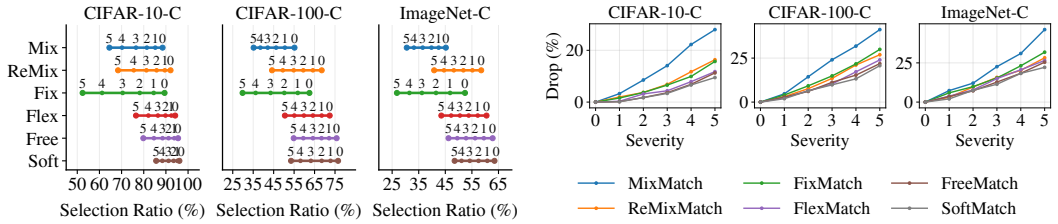


Figure 1: **Pseudo-label selection ratio (left).** Ratio of unlabeled samples selected by SSL methods, i.e., samples whose confidence exceeds the threshold, under severities 0–5. **Performance degradation (right).** Accuracy drop (%) of SSL baselines as corruption severity increases from 0 to 5.

semi-supervised learning. We treat the unlabeled pool as queryable evidence and construct neighborhood supervision via hybrid retrieval, using dense semantic retrieval, with BM25-style scoring for fine-grained cues, thereby improving SSL robustness under distribution mismatch.

3 METHODS

3.1 WHY PSEUDO-LABELING FAILS IN THE WILD

Despite strong results on curated, near-distribution benchmarks (e.g., CIFAR-10 (Krizhevsky et al., 2009), SVHN (Netzer et al., 2011)) where cluster/low-density assumptions plausibly hold and confidence is relatively well calibrated, pseudo-labeling (PL) is brittle *in the wild*. Real deployments routinely face covariate and appearance shift, class-prior shift (Sugiyama & Kawanabe, 2012), long tails (Kang et al., 2019), and open-set (Bendale & Boult, 2016) contamination, all of which degrade calibration and violate sample-selection assumptions. Consequently, thresholding by confidence no longer "purifies" the unlabeled pool, and self-training can amplify systematic errors. We therefore make explicit the two proximate causes that govern PL's success or failure.

Let $C(u) = \max_y \tilde{\mathbf{y}}(u)_y$ be the weak-view confidence for $u \in \mathcal{D}_u$, and define the selected pool $\mathcal{U}_\tau = \{u \in \mathcal{D}_u : C(u) \geq \tau\}$ with mass $p_\tau = \Pr(C \geq \tau)$. Write $\hat{y}(u) = \arg\max_y \tilde{\mathbf{y}}(u)_y$ and let $y^*(u)$ denote the (unknown) true label. The pseudo-label error rate on the confidence-selected pool is

$$\varepsilon_{\text{pl}|\tau} = \Pr\big[\hat{y}(u) \neq y^*(u) \,\big|\, u \in \mathcal{U}_\tau\big].$$

Define the reliability curve $A(c) = \Pr(\text{correct} \mid C = c)$ and the calibration gap $\delta_{\text{cal}}(c) = c - A(c)$ (Guo et al., 2017). A simple conditioning argument then yields the decomposition

$$\varepsilon_{\text{pl}|\tau} = 1 - \mathbb{E}[C \mid C \geq \tau] \;+\; \mathbb{E}[\delta_{\text{cal}}(C) \mid C \geq \tau]$$

which separates an *intrinsic selection* term $1 - \mathbb{E}[C \mid C \geq \tau]$ (from biased inclusion of boundary/rare/open-set samples even at high C) from a *miscalibration* term $\mathbb{E}[\delta_{\text{cal}}(C) \mid C \geq \tau]$ (when confidence overstates correctness under shift).

Remark. We validate these failure modes under a controlled clean-labeled / corrupted-unlabeled shift on CIFAR and ImageNet (Fig. 1), where unlabeled images are synthesized using standard corruption operators with increasing severity (Hendrycks & Dietterich, 2019). Fig. 2 shows a clear calibration breakdown under shift: as severity increases, accuracy drops while the model remains overconfident on misclassified samples, indicating a pronounced decoupling between confidence and correctness in the high-confidence region. This overconfidence directly undermines confidence based selection (Fig. 1 (left)): although the retention rate decreases as the shift strengthens, the selected subset is not reliably cleaner—the conditional error among retained pseudo-labels remains high because confidence no longer predicts correctness. Moreover, the shift alters the difficulty structure of the unlabeled pool (more hard/ambiguous instances), so thresholding can simultaneously admit high-confidence errors (FP-type) and discard correct-but-underconfident samples (FN-type), reducing both label quality and coverage. As a consequence, self-training continues to ingest a non-trivial amount of erroneous supervision, reinforcing confirmation bias and leading to

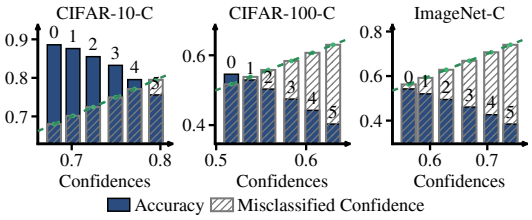


Figure 2: **Calibration Failure under Distribution Shift.** Solid bars denote accuracy and hatched bars denote confidence on misclassified samples across severities 0–5. Despite declining accuracy, confidence remains high on misclassified samples ("confidently wrong").

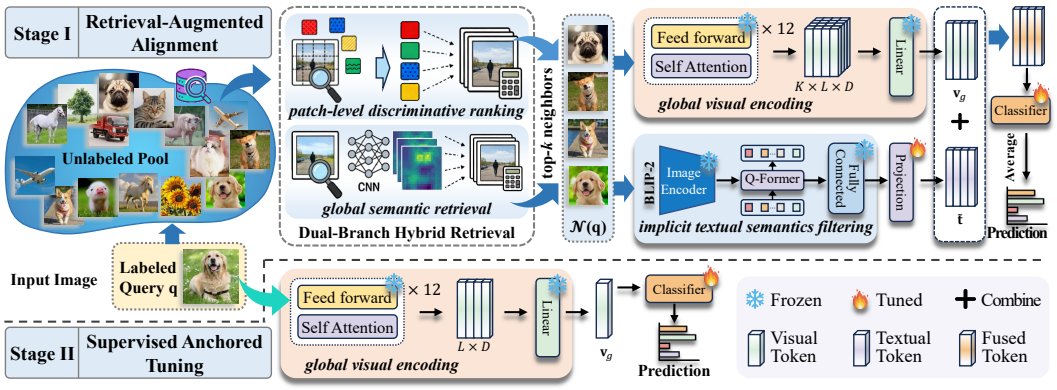


Figure 3: **RAVT framework overview**. For each labeled query, **Stage I** (Retrieval-Augmented Alignment) retrieves unlabeled neighbors via hybrid retrieval, and aggregates visual–language evidence to optimize the lightweight projection and classifier. **Stage II** (Supervised Anchored Tuning) fine-tunes the classifier on labeled data to re-anchor predictions and stabilize training.

the pronounced performance degradation of confidence-threshold pseudo-labeling under distribution shift (Fig. 1 (right)).

3.2 Retrieval-Augmented Visual Tuning

These observations suggest that pseudo-labeling under distribution shift is brittle not merely due to threshold selection, but to its inherently *pointwise* mechanism: unlabeled samples are accepted or rejected by confidence scores that can be severely miscalibrated under shift. RAVT therefore replaces global confidence filtering with *neighborhood-level evidence construction conditioned on labeled queries* (cf. Fig. 3). Stage I retrieves coherent unlabeled neighborhoods around clean labeled anchors and aggregates their evidence, reducing reliance on individual confidence scores and mitigating false-positive/false-negative errors. Hybrid retrieval and implicit textual semantics improve neighborhood quality, while Stage II re-anchors training on clean labeled data to suppress drift.

Problem Formulation and Foundations. We consider a semi-supervised classification setting with a labeled set $\mathcal{D}_\ell = \{(x_i, y_i)\}_{i=1}^{n_\ell}$, where $y_i \in \{1, \dots, C\}$, and an unlabeled set $\mathcal{D}_u = \{u_j\}_{j=1}^{n_u}$. We denote by $a_w(\cdot)$ and $a_s(\cdot)$ the weak and strong views, respectively. For visual representation, we adopt a frozen image encoder $\phi_v(\cdot)$ (e.g., ViT (Dosovitskiy, 2020) or ALBEF (Li et al., 2021)). Given an input image x, the encoder outputs local visual tokens $\mathbf{V}_{\text{loc}}(x) \in \mathbb{R}^{L \times D}$ and a global visual embedding $\mathbf{v}_g(x) \in \mathbb{R}^D$. For implicit textual semantics, we use a frozen BLIP-2 (Li et al., 2023) model $\phi_t(\cdot)$ to produce a text feature $\mathbf{t}(x) \in \mathbb{R}^{D_t}$, which is projected into the visual embedding space via a learnable linear adapter $\tilde{\mathbf{t}}(x) = \mathbf{W}\mathbf{t}(x) \in \mathbb{R}^D$ with $\mathbf{W} \in \mathbb{R}^{D \times D_t}$. To enable efficient global retrieval over unlabeled data, we precompute and ℓ_2-normalize $\{\mathbf{v}_g(u) \mid u \in \mathcal{D}_u\}$, and build an approximate k-NN index $\mathcal{I}$ using FAISS-HNSW (Johnson et al., 2019; Malkov & Yashunin, 2018) with cosine similarity.

Stage I: Retrieval-Augmented Alignment. Given a labeled query $\mathbf{q} \in \mathcal{D}_\ell$, we first retrieve a top-k neighborhood $\mathcal{N}(\mathbf{q}) \subset \mathcal{D}_u$ from the unlabeled pool (details of retrieval are deferred to the following section). For each neighbor $r \in \mathcal{N}(\mathbf{q})$, we extract a global visual embedding $\mathbf{v}_g(r) \in \mathbb{R}^D$ using the frozen visual encoder and map it into the original label-discriminative space. Rather than relying on a single unlabeled instance and its potentially overconfident prediction, we aggregate predictions across the non-local neighborhood, forming a *cluster-level prediction* for the labeled query that is substantially more stable under distribution shift. Under substantial appearance variation, the top-k list may be dominated by superficial correlations (e.g., background, illumination, material, or style) rather than semantic agreement, resulting in false neighbors that inject spurious evidence and induce negative transfer. To address this issue, we incorporate implicit textual semantics through cross-modal feature fusion, encouraging the neighborhood evidence to rely less on superficial *visual resemblance* and more on *semantic consistency*.

Specifically, for every $r \in \mathcal{N}(\mathbf{q})$, we extract a global visual embedding $\mathbf{v}_g(r) \in \mathbb{R}^D$ using the frozen visual encoder, and obtain an implicit textual semantics $\mathbf{t}(r) \in \mathbb{R}^{D_t}$ using the frozen BLIP-2 model. The implicit textual semantics is projected into the visual embedding space via a linear mapping $\tilde{\mathbf{t}}(r) = \mathbf{W}\mathbf{t}(r) \in \mathbb{R}^D$. We then construct a retrieval-augmented representation by fusing

the ℓ_2-normalized visual and textual embeddings:

$$\mathbf{f}_r = \frac{1}{2}\left(\frac{\mathbf{v}_g(r)}{\|\mathbf{v}_g(r)\|_2} + \frac{\tilde{\mathbf{t}}(r)}{\|\tilde{\mathbf{t}}(r)\|_2}\right) \in \mathbb{R}^D. \tag{1}$$

This normalization-aware fusion prevents either modality from dominating due to scale mismatch and yields a more semantically consistent neighbor feature. Rather than deriving supervision from a single unlabeled instance, we fuse evidence *non-locally* across the retrieved neighborhood. We first compute per-neighbor logits with a lightweight classifier $h : \mathbb{R}^D \to \mathbb{R}^C$, and then aggregate them to obtain a neighborhood-level prediction for the query:

$$\bar{\mathbf{h}}(\mathbf{q}) = \frac{1}{K}\sum_{r\in\mathcal{N}(\mathbf{q})} h(\mathbf{f}_r), \qquad K = |\mathcal{N}(\mathbf{q})|. \tag{2}$$

We train the retrieval branch with a standard cross-entropy loss between the neighborhood-aggregated logits $\bar{\mathbf{h}}(\mathbf{q})$ and the ground-truth label $y_{\mathbf{q}}$. Thus, off-class neighbors only affect the aggregated evidence. By fusing aligned visual–language evidence across the retrieved neighborhood, Stage I yields a more stable supervisory signal than pointwise confidence-based pseudo-labeling, particularly under distribution shift.

Stage II: Supervised Anchored Tuning. In the second stage, we switch to the base branch and continue optimizing the same lightweight classifier $h(\cdot)$ using *only* the labeled data. For each labeled query image $\mathbf{q} \in \mathcal{D}_\ell$, we extract its global visual embedding $\mathbf{v}_g(\mathbf{q}) \in \mathbb{R}^D$ from the frozen visual encoder and feed it into the classifier, which preserves dimensional consistency with the retrieval-augmented features $\mathbf{f}_r \in \mathbb{R}^D$ used in Stage I. The base-branch objective is the standard supervised cross-entropy:

$$\mathcal{L}_{\text{anchor}} = \frac{1}{n_\ell}\sum_{(\mathbf{q},y_{\mathbf{q}})\in\mathcal{D}_\ell} \mathrm{CE}\big(y_{\mathbf{q}},\, \mathrm{softmax}(h(\mathbf{v}_g(\mathbf{q})))\big).$$

By refocusing on explicit supervision from labeled data, Stage II anchors the learned representations to the labeled distribution, suppresses neighborhood-induced drift, and sharpens class decision boundaries.

Dual-Branch Hybrid Retrieval. To construct a high-quality neighborhood for each labeled query, RAVT integrates *global semantic retrieval* with *patch-level discriminative ranking*, so that both coarse semantics and fine-grained correspondences are jointly captured. Given a labeled query image $\mathbf{q} \in \mathcal{D}_\ell$, the global branch performs efficient approximate k-NN lookup in a pre-built index over ℓ_2-normalized global embeddings $\mathbf{v}_g(\cdot) \in \mathbb{R}^D$, yielding a coarse candidate set $\mathcal{N}_{\text{global}}(\mathbf{q})$. In parallel, the local branch applies our PatchLex Local Ranker to recover neighbors whose patch-token patterns align with the query beyond what pooled global features can reliably capture, producing $\mathcal{N}_{\text{local}}(\mathbf{q})$. The final neighborhood is formed by $\mathcal{N}(\mathbf{q}) = [\mathcal{N}_{\text{global}}(\mathbf{q}), \mathcal{N}_{\text{local}}(\mathbf{q})]$, with $K = |\mathcal{N}(\mathbf{q})|$. This complementary design provides a more reliable evidence pool for subsequent alignment and neighborhood aggregation. $\mathcal{N}\mathrm{global}(\mathbf{q})$ and $\mathcal{N}\mathrm{local}(\mathbf{q})$ can naturally overlap, this overlap effectively upweights such "consensus" neighbors in the merged neighborhood.

PatchLex adapts BM25-style lexical ranking to patch tokens. For any image $\mathbf{x}$, a frozen encoder extracts local patch tokens $\mathbf{T}(\mathbf{x}) \in \mathbb{R}^{L\times D}$. We discretize the unlabeled token space by stacking $\{\mathbf{T}(\mathbf{x}) \mid \mathbf{x} \in \mathcal{D}_u\}$ into $\mathbf{T}(\mathcal{D}_u) \in \mathbb{R}^{(N_u L)\times D}$ and running MiniBatchKMeans (Sculley, 2010) to obtain M prototypes $\mathcal{V} = \{\mathbf{c}_m\}_{m=1}^M$, which serve as *visual words* (Sivic & Zisserman, 2003).

Each token $\mathbf{T}(\mathbf{x})_l$ is assigned to its nearest prototype $a_l \in \{1,\dots,M\}$, yielding a Bag-of-Words (BoW) vector $\mathbf{b}(\mathbf{x}) \in \mathbb{N}^M$ (Gabriela, 2004):

$$b_m(\mathbf{x}) = \sum_{l=1}^{L} \mathbf{1}[a_l = m], \qquad |\mathbf{x}| = \sum_{m=1}^{M} b_m(\mathbf{x}) = L.$$

We compute frequency and IDF (Robertson et al., 1995) over $\mathcal{D}_u$ as $\mathrm{df}_m = \sum_{\mathbf{x}\in\mathcal{D}_u} \mathbf{1}[b_m(\mathbf{x}) > 0]$ and $\mathrm{IDF}_m = \log\Big(\frac{N_u - \mathrm{df}_m + 0.5}{\mathrm{df}_m + 0.5} + 1\Big)$. For a query $\mathbf{q}$ and an unlabeled candidate $\mathbf{d} \in \mathcal{D}_u$, PatchLex ranks candidates by $S_{\text{lex}}(\mathbf{q},\mathbf{d})$ and retrieves the top-k_{local} neighbors:

$$S_{\text{lex}}(\mathbf{q},\mathbf{d}) = \sum_{m:b_m(\mathbf{q})>0} \mathrm{IDF}_m \frac{b_m(\mathbf{d})(k_{\text{lex}}+1)}{b_m(\mathbf{d}) + k_{\text{lex}}}, \quad \mathcal{N}_{\text{local}}(\mathbf{q}) = \mathrm{TopK}_{k_{\text{local}}}\{S_{\text{lex}}(\mathbf{q},\mathbf{d}) : \mathbf{d} \in \mathcal{D}_u\}.$$

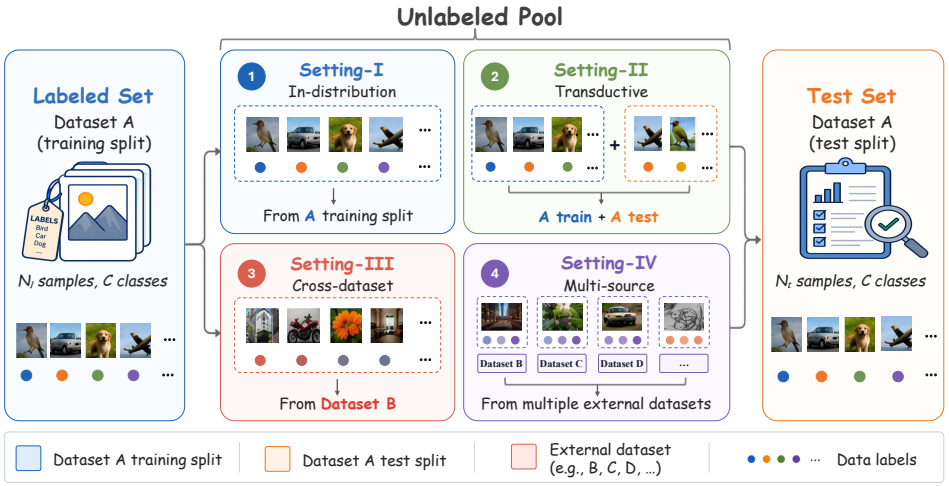


Figure 4: **Four Evaluation Protocols.**

Table 1: **Comparison with state-of-the-art under increasing distribution shift.** The best and second-best results in each setting are highlighted in red and blue, respectively. Setting-I: in-distribution, Setting-II: transductive, Setting-III: cross-dataset, Setting-IV: multi-source.

Method	Backbone	CIFAR-10				CIFAR-100			
		Setting-I	Setting-II	Setting-III	Setting-IV	Setting-I	Setting-II	Setting-III	Setting-IV
Labeled Only	ViT	75.78				49.90			
MixMatch Berthelot et al. (2019b)	ViT	63.81	68.84	60.31	51.63	32.41	30.85	22.36	27.60
ReMixMatch Berthelot et al. (2019a)	ViT	90.12	94.30	58.63	57.41	57.25	56.83	41.77	46.60
FixMatch Sohn et al. (2020)	ViT	92.53	85.19	75.29	63.65	53.58	54.58	51.75	53.25
FlexMatch Zhang et al. (2021)	ViT	95.03	91.44	69.56	65.72	60.06	59.90	47.28	50.70
FreeMatch Wang et al. (2022)	ViT	95.10	95.95	74.02	77.33	62.02	61.14	45.64	50.66
SoftMatch Chen et al. (2023)	ViT	94.90	95.05	70.90	73.49	62.13	61.61	46.91	54.29
RAVT	ViT	92.25	92.73	87.71	83.63	61.79	61.36	58.72	57.47
Labeled Only	ALBEF	79.72				50.04			
FixMatch Sohn et al. (2020)	ALBEF	93.96	88.47	75.81	73.43	54.60	54.47	53.69	54.93
FlexMatch Zhang et al. (2021)	ALBEF	95.11	93.27	73.69	72.41	60.54	60.28	48.96	50.74
FreeMatch Wang et al. (2022)	ALBEF	95.19	96.07	79.91	80.15	62.33	62.26	50.84	50.57
SoftMatch Chen et al. (2023)	ALBEF	95.04	95.13	79.84	77.05	61.41	61.31	50.60	53.78
MultiMatch (Sirbu et al., 2025)	ALBEF	95.52	95.61	77.43	73.95	61.68	61.54	48.32	51.11
V-PET (Zhang et al., 2025)	ALBEF	95.59	95.67	82.14	79.85	62.06	61.98	54.27	55.45
RAVT	ALBEF	93.02	93.90	90.19	87.17	62.50	62.43	61.60	59.24
RAVT + FixMatch	ALBEF	93.71	94.46	90.96	86.64	66.35	64.96	62.37	60.60
RAVT + FlexMatch	ALBEF	95.25	94.69	89.44	85.26	67.04	66.42	61.45	59.94
RAVT + FreeMatch	ALBEF	95.33	95.47	91.52	87.99	69.61	67.93	61.83	60.94
RAVT + SoftMatch	ALBEF	94.97	95.06	91.28	87.49	69.33	68.04	62.15	60.84
RAVT + MultiMatch	ALBEF	95.18	95.25	90.88	87.36	69.42	68.28	61.92	60.12
RAVT + V-PET	ALBEF	95.03	95.21	90.46	87.24	69.15	68.22	61.71	59.65

4 Experiments

4.1 Setup and Details

As illustrated in Fig. 4, we evaluate RAVT under four progressively challenging semi-supervised settings for robustness and generalization. **Setting-I (in-distribution)** follows the standard protocol, using labeled(with N_{labeled} samples across C classes) and unlabeled data from the same training split, providing an i.i.d. baseline. **Setting-II (transductive (Chapelle et al., 2006))** adds test images to the unlabeled pool while keeping the labeled set unchanged, allowing access to unlabeled test images during training. **Setting-III (cross-dataset)** introduces single-source distribution shift by sourcing unlabeled data from an independent external dataset, creating substantial statistical and semantic mismatch. Finally, **Setting-IV (multi-source)** aggregates multiple external datasets into a diverse unlabeled pool, yielding heterogeneous distributions that better reflect real-world in-the-wild deployments. All settings use a test set from the same dataset as the labeled set.

We optimize learnable modules with AdamW (lr 1×10^{-3}, wd 1×10^{-4}) under a cosine schedule. PatchLex uses 196 ViT patch tokens, excluding `[CLS]`. We build a MiniBatchKMeans visual-word codebook with M=1,000 prototypes and use k_{lex}=1.5. For neighborhoods, we use $K = 26$ by default, split evenly between global and PatchLex retrieval ($K/2$) without deduplication.

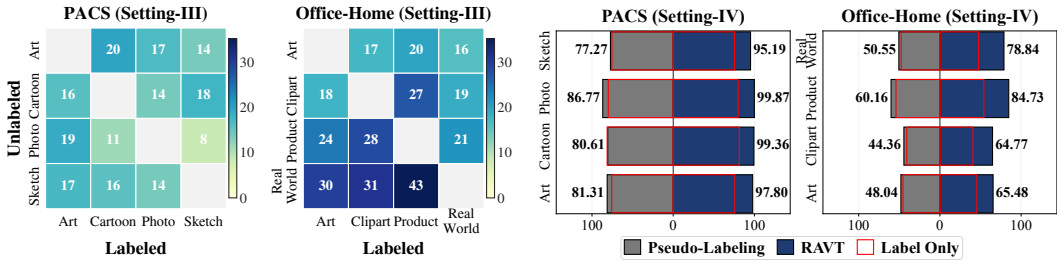


Figure 5: **Cross-domain evaluation on PACS and Office-Home. Closed-loop validation under Setting-III (left)**, where RAVT gains (%) over pseudo-labeling are reported for each labeled/unlabeled domain pair. **Multi-source under Setting-IV (right)**, where one domain is the labeled anchor with 4 shots/class and the remaining domains form a heterogeneous unlabeled pool.

4.2 Compared with State-of-The-Art Methods

RAVT Excels in Extreme Distribution Shift Regimes. As summarized in Table 1, we evaluate RAVT under progressively increasing domain mismatch using CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009) as labeled anchors, with 4 labeled images per class. Setting-III introduces 100k cross-dataset unlabeled images from STL-10 (Coates et al., 2011), while Setting-IV uses a heterogeneous 100k-image pool from Caltech-101 (Fei-Fei et al., 2004), ImageNet (Deng et al., 2009), iNaturalist (Van Horn et al., 2018), STL-10, and CIFAR-10/100. We compare RAVT with representative and recent SOTA SSL methods. RAVT shows clear advantages under severe distribution shift: on CIFAR-10, it achieves 90.19% vs. 82.14% in Setting-III and 87.17% vs. 80.15% in Setting-IV, with similar trends on CIFAR-100. In near-i.i.d. regimes (Setting-I/II), pseudo-labeling methods remain highly competitive, while combining pseudo-labeling with RAVT achieves the strongest overall performance, together demonstrating the complementarity between RAVT and standard SSL techniques. These gains are also consistent across different backbones (e.g., ViT and ALBEF), indicating that RAVT generalizes beyond a specific architecture.

Closed-Loop Validation under Distribution Shift. To evaluate robustness to unlabeled-domain shift, we perform *closed-loop* evaluations on PACS (Li et al., 2017) and Office-Home (Venkateswara et al., 2017), where the test domain matches the labeled domain.In Setting-III, each domain serves as the labeled anchor and pairs with each remaining domain as unlabeled data, yielding 12 cycles (Fig. 5 (left)). RAVT consistently outperforms pseudo-labeling (PL), with average gains of **+15.33%/+24.5%** on PACS/Office-Home. In **Setting-IV**, the remaining domains are merged into a heterogeneous unlabeled pool for each labeled anchor (Fig. 5(right)), where RAVT still improves over PL by **+16.57%/+22.68%**. We further compare RAVT with PL on DTD, SUN397, RESISC45, Retinopathy, CLEVR-C, and KITTI, where RAVT performs better in most Setting-III pairs and across Setting-IV labeled targets (Fig. 6). These results show that retrieval-backed, neighborhood-level supervision remains robust under diverse domain mismatch.

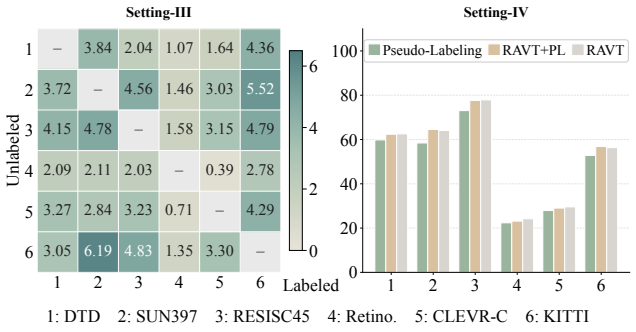


Figure 6: **Evaluation on Diverse Datasets.** RAVT gains over pseudo-labeling (PL) in Setting-III **(left)**, where columns/rows indicate labeled/unlabeled datasets; and PL, RAVT, and RAVT+PL accuracy in Setting-IV **(right)** using a mixed unlabeled pool from the other five datasets.

4.3 Ablation Study

The Impact of Different Components. We explore RAVT via ablations on (i) the neighborhood retrieval and (ii) the two-stage protocol, isolating the effect of each design choice as follows.

❶ Stage I *vs.* Stage II. Table 2 shows that Stage I alone already achieves 50.82% on Tiny-ImageNet-200 (Deng et al., 2009) (4-shot/class), while Stage II alone drops to 39.95%, indicating that robustness to shifted unlabeled data mainly comes from retrieval-augmented alignment. Combining both stages yields the best accuracy (52.64%), improving over Stage I by +1.82%, confirming that Stage II complements Stage I by re-anchoring to clean supervision and sharpening the decision boundary.

Table 2: **Ablation on training protocol.** We study the effect of **Stage I** and **Stage II** on Tiny-ImageNet-200.

Ablations \ Components	Stage I	Stage II	Acc. (%) ↑
exp-I (Stage II only)	✗	✓	39.95
exp-II (Stage I only)	✓	✗	50.82
exp-III (Stage I + II)	✓	✓	**52.64**

Table 3: **Ablation on feature modalities.** Effect of **GVE** and **ITS** on Tiny-ImageNet-200.

Ablations \ Components	GVE	ITS	Acc. (%) ↑
exp-I (Baseline)	✗	✗	39.95
exp-II (GVE only)	✓	✗	46.79 (+6.84%)
exp-III (ITS only)	✗	✓	51.18 (+11.23%)
exp-IV (GVE + ITS)	✓	✓	**52.64** (+12.69%)

Table 4: **Effect of BLIP-2 support.** Accuracies after equipping SSL baselines with BLIP-2; parentheses show gains over ViT baselines. RAVT remains stronger under distribution mismatch.

Method	CIFAR-10				CIFAR-100			
	Setting-I	Setting-II	Setting-III	Setting-IV	Setting-I	Setting-II	Setting-III	Setting-IV
SSL baselines equipped with BLIP-2								
FixMatch	92.81 (+0.28)	85.64 (+0.45)	77.31 (+2.02)	64.22 (+0.57)	54.00 (+0.42)	54.92 (+0.34)	52.01 (+0.26)	54.02 (+0.77)
FlexMatch	95.18 (+0.15)	91.76 (+0.32)	71.42 (+1.86)	65.98 (+0.26)	60.32 (+0.26)	60.27 (+0.37)	48.82 (+1.54)	50.73 (+0.03)
FreeMatch	95.22 (+0.12)	95.98 (+0.03)	74.69 (+0.67)	78.46 (+1.13)	62.24 (+0.22)	61.45 (+0.31)	47.08 (+1.44)	51.44 (+0.78)
SoftMatch	94.98 (+0.08)	95.24 (+0.19)	71.41 (+0.51)	73.85 (+0.36)	62.31 (+0.18)	61.88 (+0.27)	47.03 (+0.12)	54.87 (+0.58)
RAVT with BLIP-2 support								
RAVT	92.25	92.73	87.71	83.63	61.79	61.36	58.72	57.47
RAVT + FixMatch	93.17	92.84	86.53	82.26	62.71	62.08	57.34	56.19
RAVT + FlexMatch	94.22	93.51	85.86	81.33	63.85	62.74	56.62	56.48
RAVT + FreeMatch	94.73	95.14	88.07	83.82	64.67	63.92	59.13	57.81
RAVT + SoftMatch	**95.28**	**95.43**	**89.26**	**85.17**	**66.24**	**65.73**	**60.45**	**58.82**

❷ Global Visual Embeddings (**GVE**) *vs.* Implicit Textual Semantics (**ITS**). Table 3 shows that both modalities improve Tiny-ImageNet-200, with **ITS** contributing more than **GVE**. From the no-feature baseline (39.95%), **GVE**-only reaches 46.79%, **ITS**-only reaches 51.18%, and their combination achieves the best result of 52.64%. This indicates that textual semantics provides stronger robustness under mismatch, while visual embeddings offer complementary cues. Since neither modality alone matches the full model, the improvement is not simply due to an auxiliary encoder. To further isolate the effect of BLIP-2, Table 4 equips the compared SSL baselines with the same BLIP-2 feature support; RAVT still outperforms them under in-the-wild mismatch, confirming that the main gain comes from evidence-backed neighborhood construction rather than merely using BLIP-2.

❸ Global Semantic Retrieval (**GSR**) *vs.* PatchLex Local Ranker (**PLR**). Table 5 shows that retrieval quality is crucial on CIFAR-10/100. From the random-indexing baseline (77.38%/47.53%), **GSR** gains +10.06/+14.07% and **PLR** +13.98/+14.44%, showing patch-level ranking better suppresses appearance-driven false neighbors. The hybrid **GSR+PLR** performs best, improving over the baseline by +15.64/+14.97% and the stronger single branch by +1.66/+0.53%, confirming that global recall and local discrimination are complementary for building reliable neighborhoods under shift.

RAVT under Varying Unlabeled Capacity and Diversity. We study RAVT against the strongest pseudo-labeling baseline along unlabeled volume (N_u) and category diversity (C_u). Caltech-101 and CIFAR-100 serve as 4-shot/class labeled anchors, while the unlabeled pool is uniformly sampled from DTD (Cimpoi et al., 2014), Flowers102 (Nilsback & Zisserman, 2008), Pets (Parkhi et al., 2012), SVHN (Netzer et al., 2011), and SUN397 (Xiao et al., 2010), with categories and per-category sample counts balanced across sources.

❶ Varying Number of Unlabeled Samples. To isolate unlabeled *volume*, we fix C_u=200 and vary $N_u \in \{0, 2\text{k}, 4\text{k}, 8\text{k}, 12\text{k}, 16\text{k}, 20\text{k}\}$. Fig. 7(a) reports accuracy (left) and gain over the N_u=0 baseline (right) on Caltech101 and CIFAR-100. Across budgets, **RAVT scales more favorably with** N_u: on Caltech101 it sustains ∼**5–7%** gains, while pseudo-labeling peaks at ∼**2–3%** and saturates; on CIFAR-100, RAVT maintains ∼**3–5%** gains whereas pseudo-labeling offers only ≤**2%** and diminishes as N_u grows. This reflects a mechanistic difference: pseudo-labeling accumulates calibration-sensitive noisy labels under mismatch, whereas higher N_u gives RAVT denser retrieval neighborhoods and stronger neighborhood-level evidence aggregation, yielding stable improvements.

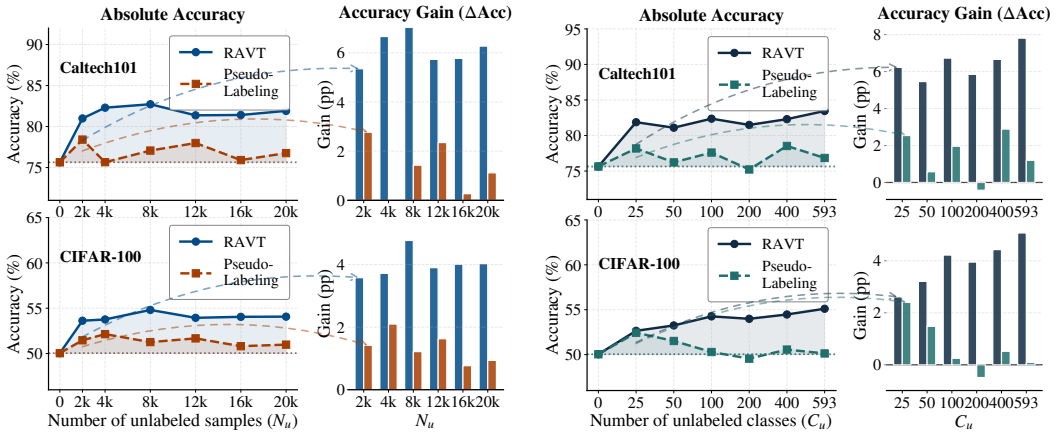
❷ Varying Categories of Unlabeled Samples. To study unlabeled *diversity*, we fix N_u=8k and vary $C_u \in \{25, 50, 100, 200, 400, 593\}$ (593: full category set). Fig. 7(b) shows a clear divergence as C_u grows: pseudo-labeling quickly plateaus, reflecting the brittleness of confidence-based supervision as the unlabeled space becomes increasingly open and disjoint from the labeled classes, where open-set noise and miscalibration accumulate. In contrast, **RAVT improves steadily with increasing** C_u,

Table 5: **Ablation on hybrid retrieval.** Effect of **GSR** and **PLR** on CIFAR-10/100.

Components / Ablations	GSR	PLR	CIFAR-10	CIFAR-100
exp-I (Baseline)	✗	✗	77.38	47.53
exp-II (GSR only)	✓	✗	87.44	61.60
exp-III (PLR only)	✗	✓	91.36	61.97
exp-IV (GSR + PLR)	✓	✓	**93.02**	**62.50**

Table 6: **Long-tailed unlabeled data.** Accuracy on CIFAR-10/100-LT and Tiny-ImageNet-LT under imbalance factor 100.

Methods	CIFAR-10-LT	CIFAR-100-LT	Tiny-ImageNet-LT
Labeled Only	79.72	50.04	39.95
FreeMatch	88.75	58.46	46.77
SoftMatch	87.93	56.79	46.42
RAVT	**92.27**	**59.65**	**50.73**



(a) (b)

Figure 7: **Ablation on unlabeled scale and diversity.** Accuracy and baseline gains on Caltech101/CIFAR-100 (**Top/Bottom**) under varying (a) N_u and (b) C_u.

indicating stronger tolerance to highly diverse mixtures. By grounding learning in *retrieval neighborhoods* and aggregating evidence non-locally, RAVT exploits additional categories as transferable primitives while suppressing spurious signals from heterogeneous unlabeled data.

Robustness to Long-Tailed Unlabeled Data. We test robustness to unlabeled class imbalance with balanced labeled/test splits and a long-tailed unlabeled pool (IF=100). As shown in Table 6, RAVT is best across all benchmarks: outperforming SoftMatch by **+4.34/+2.86/+4.31%** and FreeMatch by **+3.52/+1.19/+3.96%** on CIFAR-10/100-LT and Tiny-ImageNet-LT, respectively. These results show that RAVT is more resilient to unlabeled long-tail shifts, where confidence-based pseudo-labeling tends to be biased toward head classes.

Training Overhead with Zero Inference Cost. Table 7 reports RAVT's per-image runtime across unlabeled pool sizes. RAVT incurs a one-time *offline* database construction cost that grows with N_u but is reduced by product quantization (PQ). During *training*, PQ and asynchronous updates reduce RAVT from 10.44–24.08 ms to 6.68–10.17 ms per image, comparable to standard SSL baselines (8.29–8.40 ms) even at 100k samples. At *inference* (test-time cost), retrieval is disabled and RAVT remains at 4.68 ms per image across all N_u, confirming **no extra inference-time overhead**.

Table 7: **Per-image runtime (ms) with varying unlabeled samples N_u.**

Stage	Method	2k	4k	8k	100k
Offline	RAVT	41.71	52.14	81.07	149.50
	RAVT + PQ	32.14	38.65	56.82	99.34
Inference	RAVT	4.68	4.68	4.68	4.68
Training	RAVT	10.44	12.36	15.75	24.08
	RAVT + PQ	9.63	11.24	14.12	21.19
	RAVT + Async	7.12	7.81	8.64	11.39
	RAVT + PQ + Async	**6.68**	**7.24**	**7.96**	10.17
	FlexMatch	8.29	8.29	8.29	**8.29**
	FreeMatch	8.34	8.34	8.34	8.34
	SoftMatch	8.40	8.40	8.40	8.40

5 Conclusion and Limitations

We present **RAVT**, a retrieval-augmented SSL pipeline that remains effective in challenging *in-the-wild* distribution shifts. By grounding unlabeled utilization in neighborhood-level evidence, RAVT is *complementary* to existing SSL methods and can be seamlessly combined with pseudo-labeling to further improve performance. Our current study focuses on image classification; extending RAVT to *dense prediction* tasks (e.g., detection and segmentation) is a promising direction for future work.

AI USE STATEMENT

We used generative AI tools only for language editing and improving the clarity and readability of the manuscript. Generative AI was not used to formulate the research problem, develop the proposed method, derive technical formulations, design experiments, or generate experimental results. All technical content, analyses, and reported results were produced and verified by the authors, who take full responsibility for the content of this paper.

ETHICS STATEMENT

This work studies semi-supervised learning under distribution mismatch using publicly available research benchmarks and pretrained models. No new human-subject data or personally identifiable information are collected. As with other learning systems built on existing datasets and pretrained models, RAVT may inherit biases or limitations from these resources. The method is intended for research purposes, and its robustness under distribution shift should not be interpreted as guaranteeing reliability or fairness in safety-critical applications.

REPRODUCIBILITY STATEMENT

We provide the full formulation of RAVT, including Retrieval-Augmented Alignment, Supervised Anchored Tuning, and the dual-branch hybrid retrieval mechanism. The experimental section specifies the four evaluation settings, training configuration, retrieval setup, and key hyperparameters. We further report extensive ablations and sensitivity analyses on the training stages, feature modalities, retrieval components, unlabeled-data scale and diversity, and robustness to long-tailed unlabeled data. We will release the source code and configuration files upon publication.

REFERENCES

Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *International conference on machine learning*, pp. 312–321. PMLR, 2019.

Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1563–1572, 2016.

David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019a.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019b.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. A discussion of semi-supervised learning and transduction. In *Semi-supervised learning*, pp. 473–478. MIT Press, 2006.

Hao Chen, Ran Tao, Yue Fan, Yidong Wang, Jindong Wang, Bernt Schiele, Xing Xie, Bhiksha Raj, and Marios Savvides. Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. *arXiv preprint arXiv:2301.10921*, 2023.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5558–5570, 2022.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178. IEEE, 2004.

Csurka Gabriela. Visual categorization with bags of keypoints. 2004.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.(2009), 2009.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, pp. 896. Atlanta, 2013.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, and Anton Van den Hengel. Retrieval augmented classification for long-tail visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6959–6969, 2022.

Yu A Malkov and Dmitry A Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836, 2018.

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14111–14121, 2021.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 7. Granada, 2011.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. *Okapi at TREC-3*. British Library Research and Development Department, 1995.

David Sculley. Web-scale k-means clustering. In *Proceedings of the 19th international conference on World wide web*, pp. 1177–1178, 2010.

Iustin Sirbu, Robert-Adrian Popovici, Cornelia Caragea, Stefan Trausan-Matu, and Traian Rebedea. MultiMatch: Multihead consistency regularization matching for semi-supervised text classification. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2792–2808, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.139. URL https://aclanthology.org/2025.emnlp-main.139/.

Sivic and Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proceedings ninth IEEE international conference on computer vision*, pp. 1470–1477. IEEE, 2003.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Masashi Sugiyama and Motoaki Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, 2012.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5018–5027, 2017.

Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, et al. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*, 2022.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561*, 2022.

Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. Pretrained transformers for text ranking: Bert and beyond. In *Proceedings of the 14th ACM International Conference on web search and data mining*, pp. 1154–1156, 2021.

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.

Ping Zhang, Zheda Mai, Quang-Huy (Percy) Nguyen, and Wei-Lun (Harry) Chao. Revisiting semi-supervised learning in the era of foundation models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 59295–59324. Curran Associates, Inc., 2025. doi: 10.52202/085713-1984. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/558159930b584e7f137a1dcbd380d9fd-Paper-Conference.pdf.

A Additional Experimental Results

This section evaluates whether the proposed retrieval-backed neighborhood supervision generalizes beyond visual classification.

A.1 Generalization Beyond Vision: Results on AG News

The main paper focuses on visual semi-supervised learning under distribution mismatch, with progressively more challenging regimes involving cross-dataset, multi-source, and domain-shifted unlabeled pools. To examine whether the proposed idea is limited to vision, we additionally adapt RAVT to a text classification setting and evaluate it on AG News. As shown in Table 8, RAVT remains competitive in the near-i.i.d. settings and provides substantially stronger robustness in the more challenging mismatch settings. Moreover, combining RAVT with conventional SSL methods consistently improves their performance, suggesting that retrieval-augmented neighborhood evidence is also effective beyond the visual domain.

B Analysis of Potential Pretraining Leakage

Because RAVT uses a pretrained cross-modal model to derive implicit textual semantics, we conduct two controlled experiments to examine whether its gains can be attributed to hidden memorization or leakage from cross-modal pretraining. First, we construct a deliberately unfavorable worst-case control in which BLIP-2 is explicitly exposed to the original unlabeled image–label pairs. Second, we replace BLIP-2 with CLIP and verify whether the advantage of RAVT persists with a different pretrained cross-modal model.

B.1 Worst-Case Joint Leakage of (X_u, Y_u)

Although there is no direct evidence that BLIP-2 or CLIP-style pretraining memorizes the unlabeled image–label pairs used in our evaluation, we test a strict worst-case scenario. Specifically, we fine-tune BLIP-2 on the original unlabeled pairs (X_u, Y_u), thereby forcing explicit exposure to the unlabeled data space, and then use its predictions to replace the pseudo-labels of the baseline SSL methods. Table 9 reports three values for each baseline: performance using BLIP-2 zero-shot predictions, performance using predictions from the explicitly fine-tuned BLIP-2, and performance using the baseline's original pseudo-labels. Fine-tuning BLIP-2 improves several baselines over raw zero-shot labeling, but all variants remain far below RAVT. Thus, even explicit access to (X_u, Y_u) cannot explain the performance margin of the proposed method.

B.2 Replacing BLIP-2 with CLIP

We further replace BLIP-2 with CLIP to determine whether the improvements are specific to a particular cross-modal foundation model. As shown in Table 10, BLIP-2 generally provides stronger

Table 8: Results on **AG News** under four settings with increasing unlabeled-data mismatch difficulty. Conventional SSL methods perform competitively in the near-i.i.d. regimes (Settings I and II), whereas **RAVT** provides clear gains in the more challenging mismatch regimes (Settings III and IV). Combining RAVT with existing SSL methods further improves performance, indicating that retrieval-augmented neighborhood supervision is complementary to confidence-based pseudo-labeling and is not restricted to visual data.

Method	Setting-I	Setting-II	Setting-III	Setting-IV
FixMatch	82.90	81.65	71.80	67.50
FlexMatch	84.51	83.82	74.35	69.85
SoftMatch	87.32	86.85	76.64	73.37
RAVT	85.48	85.70	82.46	81.32
RAVT + FixMatch	85.57	85.42	82.95	81.45
RAVT + FlexMatch	86.24	86.15	83.18	81.70
RAVT + SoftMatch	**87.49**	**87.25**	**83.65**	**82.15**

Table 9: Worst-case joint-leakage analysis under Settings III and IV. Each baseline entry reports **BLIP-2 zero-shot / BLIP-2 fine-tuned / original baseline pseudo-label** accuracy (%). BLIP-2 is deliberately fine-tuned on the original unlabeled image–label pairs (X_u, Y_u) before its predictions are used as pseudo-labels. Although this explicit exposure improves some baselines, their results remain substantially below **RAVT**, showing that potential pretraining leakage cannot account for the observed gains.

Method	CIFAR-10 (Setting-III)	CIFAR-10 (Setting-IV)	CIFAR-100 (Setting-III)	CIFAR-100 (Setting-IV)
Labeled Only	75.78	75.78	49.90	49.90
FixMatch	71.93 / 79.12 / 75.29	63.29 / 63.38 / 63.65	46.91 / 52.46 / 51.75	45.64 / 47.88 / 53.25
FlexMatch	68.74 / 75.48 / 69.56	63.82 / 66.14 / 65.72	45.27 / 49.83 / 47.28	44.89 / 45.92 / 50.70
FreeMatch	73.11 / 80.05 / 74.02	72.84 / 78.41 / 77.33	41.83 / 46.55 / 45.64	43.71 / 51.23 / 50.66
SoftMatch	69.56 / 78.31 / 70.90	71.97 / 72.18 / 73.49	43.58 / 50.12 / 46.91	47.67 / 49.34 / 54.29
RAVT (Ours)	**87.71**	**83.63**	**58.72**	**57.47**

Table 10: Comparison between CLIP and BLIP-2 under Settings III and IV. Each numerical entry reports **CLIP / BLIP-2** accuracy (%). Although BLIP-2 is generally stronger, RAVT maintains a large advantage with either cross-modal model. The consistent improvements of RAVT-equipped variants over pure SSL baselines support the conclusion that the gains originate from retrieval-augmented neighborhood evidence rather than hidden pretraining leakage.

Method	Backbone	CIFAR-10 (Setting-III)	CIFAR-10 (Setting-IV)	CIFAR-100 (Setting-III)	CIFAR-100 (Setting-IV)
FixMatch	CLIP / BLIP-2	76.12 / 77.31	61.54 / 64.22	50.48 / 52.01	51.85 / 54.02
FlexMatch	CLIP / BLIP-2	70.25 / 71.42	63.41 / 65.98	47.33 / 48.82	48.26 / 50.73
FreeMatch	CLIP / BLIP-2	73.44 / 74.69	75.62 / 78.46	45.52 / 47.08	49.08 / 51.44
SoftMatch	CLIP / BLIP-2	70.18 / 71.41	70.93 / 73.85	45.61 / 47.03	52.24 / 54.87
RAVT	CLIP / BLIP-2	87.17 / 87.71	81.04 / 83.63	57.62 / 58.72	55.38 / 57.47
RAVT + FixMatch	CLIP / BLIP-2	85.59 / 86.53	79.61 / 82.26	55.73 / 57.34	55.03 / 56.19
RAVT + FlexMatch	CLIP / BLIP-2	84.84 / 85.86	79.27 / 81.33	55.48 / 56.62	55.21 / 56.48
RAVT + FreeMatch	CLIP / BLIP-2	86.31 / 88.07	80.79 / 83.82	57.82 / 59.13	55.94 / 57.81
RAVT + SoftMatch	CLIP / BLIP-2	**89.06 / 89.26**	**83.12 / 85.17**	**59.15 / 60.45**	**56.07 / 58.82**

features than CLIP, but the same overall conclusion holds for both backbones: RAVT and all RAVT-equipped variants substantially outperform the corresponding pure SSL baselines in the challenging mismatch settings. This result indicates that the main gains arise from the retrieval-augmented neighborhood-evidence mechanism rather than from hidden data leakage or a specific pretrained model.

C Additional Ablation Studies

This section provides additional ablations on the auxiliary feature backbone, neighborhood size, global–local retrieval allocation, and PatchLex hyperparameters. All default hyperparameters are selected using a cross-validation strategy rather than by optimizing each individual test setting.

C.1 Effect of the Auxiliary Feature Backbone

The global visual embedding $v_g(r)$ and the implicit textual semantics $t(r)$ are not both produced by BLIP-2. Specifically, $v_g(r)$ is extracted from the visual backbone (ViT or ALBEF), whereas $t(r)$ is derived from the pretrained cross-modal model BLIP-2. To separate the contribution of cross-modal semantic grounding from that of stronger patch-level visual features, we replace BLIP-2 with DINO-v3 as a controlled visual-feature alternative. The comparison is reported in Table 11.

The DINO-v3 control confirms that high-quality local visual representations do improve robustness over the ALBEF-only baseline. Nevertheless, the full model using BLIP-2 is consistently stronger, especially under severe mismatch. This suggests that the improvement is driven primarily by cross-modal semantic grounding, which helps isolate OOD noise, rather than by local patch-level discriminative ranking alone.

C.2 Effect of Neighborhood Size

We select the neighborhood size K through cross-validation and report its sensitivity under the challenging mismatch settings in Table 12.

Table 11: Effect of different auxiliary feature backbones in **RAVT** on CIFAR-10 and CIFAR-100 under four settings with increasing unlabeled-data mismatch difficulty. DINO-v3 supplies strong patch-level visual features and improves upon the ALBEF baseline. However, incorporating **BLIP-2** yields substantially larger gains, particularly in the challenging mismatch settings (Settings III and IV), demonstrating the importance of cross-modal semantic grounding rather than merely stronger local patch discrimination.

Method	Backbone	CIFAR-10				CIFAR-100			
		Setting-I	Setting-II	Setting-III	Setting-IV	Setting-I	Setting-II	Setting-III	Setting-IV
RAVT	ALBEF	89.31	90.27	81.45	78.62	60.84	61.02	55.19	54.87
RAVT	ALBEF + DINO-v3	90.95	91.42	86.15	83.71	61.12	61.85	58.65	56.98
RAVT	ALBEF + BLIP-2	**93.02**	**93.90**	**90.19**	**87.17**	**62.50**	**62.43**	**61.60**	**59.24**

Table 12: Effect of neighborhood size K under the challenging mismatch settings. A small neighborhood provides insufficient consensus for suppressing severe OOD noise, whereas an overly large neighborhood may introduce semantically irrelevant samples and cause semantic drift. We use $K = 26$ as the default because it provides the strongest overall trade-off across datasets and settings. Bold entries denote the best result in each column.

Neighborhood Size (K)	CIFAR-10 (Setting-III)	CIFAR-10 (Setting-IV)	CIFAR-100 (Setting-III)	CIFAR-100 (Setting-IV)
$K = 8$	87.21	82.34	57.06	52.37
$K = 16$	88.94	85.51	59.43	56.74
$K = 22$	89.02	85.54	60.47	58.61
$K = 24$	**90.38**	86.79	60.82	58.83
$K = 26$	90.19	**87.17**	61.60	**59.24**
$K = 28$	89.62	87.03	**61.94**	59.08
$K = 32$	88.71	86.82	60.77	58.69
$K = 48$	85.93	84.41	58.61	56.51

A small K lacks sufficient neighborhood consensus to suppress severe OOD noise, while a large K increases the probability of retrieving irrelevant samples and therefore causes semantic drift. Although $K = 24$ and $K = 28$ attain isolated per-column maxima, $K = 26$ achieves the highest average accuracy across the four evaluated cases and is consequently adopted as the cross-validated default.

C.3 Global–Local Retrieval Allocation

Table 13 studies different allocations between the global retrieval branch and the PatchLex retrieval branch.

Table 13: Effect of different allocations between global retrieval and PatchLex retrieval under Setting-IV. Neither branch alone is optimal: global retrieval provides coarse semantic recall, whereas PatchLex provides fine-grained local discrimination. A balanced allocation yields the best performance.

Global	PatchLex	CIFAR-10 (Setting-IV)	CIFAR-100 (Setting-IV)
26	0	84.12	55.34
20	6	85.45	56.81
13	13	**87.17**	**59.24**
6	20	86.63	58.45
0	26	85.91	57.62

Global retrieval captures macro-level semantic context but is less robust to extreme corruptions when used alone. In contrast, PatchLex recovers robust, fine-grained local concepts, but lacks the broader semantic coverage provided by the global branch. The balanced 13:13 allocation therefore combines the complementary strengths of both branches and performs best on both datasets.

C.4 Sensitivity to the PatchLex Parameter

Table 14: Sensitivity analysis of the BM25-style PatchLex parameter k_{lex} on CIFAR-100 under **Setting-IV**. We adopt the classical default $k_{\text{lex}} = 1.5$. Performance remains stable over a broad parameter range, indicating that PatchLex is not overly sensitive to this parameter.

k_{lex}	0.5	1.2	**1.5**	2.0	3.0
CIFAR-100 Setting-IV	58.97	59.21	**59.24**	59.18	58.85

Table 14 reports the sensitivity of PatchLex to the BM25-style parameter k_{lex}. We use the classical default $k_{\mathrm{lex}} = 1.5$, selected through cross-validation.

The results vary only slightly around the selected default, showing that PatchLex remains stable across a broad range of k_{lex} values and that its performance does not depend on narrowly tuned parameter choices.