WHEN CONFIDENCE COLLAPSES: ROBUST PSEUDO-LABEL SELECTION FOR EXTREMELY SPARSE SEMI-SUPERVISED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Semi-supervised learning (SSL) reduces annotation costs by exploiting abundant unlabeled data, yet its behavior under extremely sparse supervision remains poorly understood. We systematically investigate this regime and identify a common failure mode in confidence-driven pseudo-label selection, where an over-fit classification head produces saturated confidence scores that become poorly discriminative. As a result, existing selection strategies can retain large numbers of unreliable pseudo-labels, causing training performance to peak prematurely and subsequently regress. Motivated by this observation, we propose a confidence-robust pseudo-label selection framework that combines a Confidence Calibration Prior with Weak-Strong Consistency. The Confidence Calibration Prior uses a fixed confidence floor and a per-class top-K constraint to control pseudo-label selection without relying solely on the confidence distribution, while Weak-Strong Consistency improves the reliability of the selected supervision. We further use MoCo pretraining as a complementary representation initialization and evaluate the resulting framework under a unified protocol across three image-classification benchmarks and multiple label densities. Experiments show that our method consistently improves over representative confidence-driven baselines in the extremely sparse regime, while its advantage becomes dataset-dependent as the labeling budget increases. These results reveal an operating boundary of confidence-driven pseudo-label selection and motivate sparse-label SSL methods that are less dependent on the confidence distribution of the classification head.

1 INTRODUCTION

Semi-supervised learning (SSL) reduces annotation costs by combining limited labeled data with abundant unlabeled data. Modern SSL methods based on consistency regularization and pseudo-labeling have achieved strong performance on standard benchmarks (Sohn et al., 2020; Wang et al., 2023). These advances have made SSL an important paradigm for learning when annotation is expensive or difficult to obtain (Cheng et al., 2025). However, their effectiveness under extremely limited supervision remains insufficiently understood.

In practical applications, only a handful of labeled examples may be available for each class. This sparse-label regime commonly arises in medical imaging, rare-class recognition, and cold-start deployment (Rebuffi et al., 2020). Yet existing SSL methods are predominantly evaluated with substantially larger labeling budgets (Oliver et al., 2018; Su et al., 2021). To examine this gap, we systematically compare eight SSL baselines using only 4 or 25 labels per class.

Our experiments reveal that existing SSL methods can degrade substantially when only 4 labels per class are available. Across CIFAR-10, SVHN, and STL-10, pseudo-label selection becomes increasingly unreliable despite rapidly increasing model confidence. In particular, confidence-driven methods may accept most unlabeled samples while their pseudo-label accuracy remains close to random chance. These observations suggest that poor performance is primarily caused by unreliable pseudo-label selection under extremely sparse supervision.

This failure stems from an implicit assumption underlying most existing selection strategies. They assume that confidence scores provide sufficient variation for distinguishing reliable from unreliable

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.

unlabeled samples. With very few labels, however, the classifier can become over-confident and produce nearly one-hot predictions for many incorrect samples. Consequently, threshold-based and threshold-free confidence selection mechanisms can both fail despite their different selection rules.

We therefore propose a selection strategy designed specifically for the sparse-label regime. Our Confidence Calibration Prior combines a fixed confidence floor with a per-class top-K constraint, while Weak-Strong Consistency provides more reliable representations for pseudo-label selection. Extensive experiments show consistent improvements over eight baselines at 4 labels per class across three benchmarks. We further demonstrate that the proposed approach remains competitive at higher label densities, clarifying when confidence-driven selection becomes reliable again.

- **Sparse-label failure analysis.** We systematically evaluate eight SSL baselines and uncover a common confidence-collapse phenomenon, where over-confident predictions undermine pseudo-label selection under extremely limited labels.
- **Confidence-robust pseudo-label selection.** We introduce a new framework that combines a Confidence Calibration Prior with Weak-Strong Consistency, reducing the reliance on unreliable confidence distributions while improving the quality of selected pseudo-labels.
- **Strong performance under sparse supervision.** Extensive experiments on CIFAR-10, SVHN, and STL-10 show that our method consistently achieves the best performance with only four labels per class, outperforming all compared baselines in this challenging regime.

2 Related Work

2.1 Consistency-based and confidence-driven SSL

Modern SSL predominantly combines consistency regularization with confidence-driven pseudo-label selection. Early approaches such as the Π-model (Laine & Aila, 2017) and Mean Teacher (Tarvainen & Valpola, 2017) established consistency training under input perturbations, while MixMatch (Berthelot et al., 2019) and ReMixMatch (Berthelot et al., 2020) further improved this paradigm through sharpening and MixUp. FixMatch (Sohn et al., 2020) subsequently popularized weak/strong augmentation with fixed confidence thresholding. Subsequent methods have focused on more adaptive or robust pseudo-label selection, including class-dependent thresholds in FlexMatch (Zhang et al., 2021), adaptive thresholds in FreeMatch (Wang et al., 2023), confidence weighting in SoftMatch (Chen et al., 2023), margin-based selection in MarginMatch (Sosea & Caragea, 2024), dual-mask selection in CGMatch (Cheng et al., 2025), adaptive thresholding with RegMixup in RegMixMatch (Han et al., 2025), and confidence-variance partitioning in CoVar (Liu et al., 2026). Other approaches, such as SemiReward (Li et al., 2024) and PabLO (Vishwakarma et al., 2025), explicitly address softmax miscalibration through learned reward or confidence models. Despite these advances, most methods continue to rely on confidence as a discriminative signal for pseudo-label selection, while its reliability under extremely sparse supervision remains insufficiently studied. Additional discussion of confidence-driven SSL methods is provided in Appendix A.

2.2 Sparse-label SSL and confirmation bias

Sparse-label SSL has received substantially less attention than conventional settings with moderate labeling budgets. Although FixMatch (Sohn et al., 2020) reports results with 4 labels per class on CIFAR-10, subsequent methods such as FreeMatch (Wang et al., 2023), CGMatch (Cheng et al., 2025), RegMixMatch (Han et al., 2025), and CoVar (Liu et al., 2026) are predominantly evaluated with larger labeling budgets. Complementary approaches attempt to extract additional information from low-confidence samples, such as reliable negative information in CCL (Deng et al., 2024) and feature-level perturbation with supervised contrastive learning in CLFP (Wang et al., 2026). Meanwhile, prior studies on confirmation bias show that erroneous pseudo-labels can reinforce themselves during training (Arazo et al., 2020; Rizve et al., 2021), while broader evaluations highlight the sensitivity of SSL to labeling budgets and experimental conditions (Oliver et al., 2018; Su et al., 2021). These studies primarily characterize pseudo-label pollution or labeling sensitivity, whereas our work focuses on confidence collapse as a potential mechanism underlying selection failure in the sparse-label regime. Self-supervised representation learning methods, including MoCo (He et al., 2020), SimCLR (Chen et al., 2020a;b), and BYOL (Grill et al., 2020), provide complementary approaches

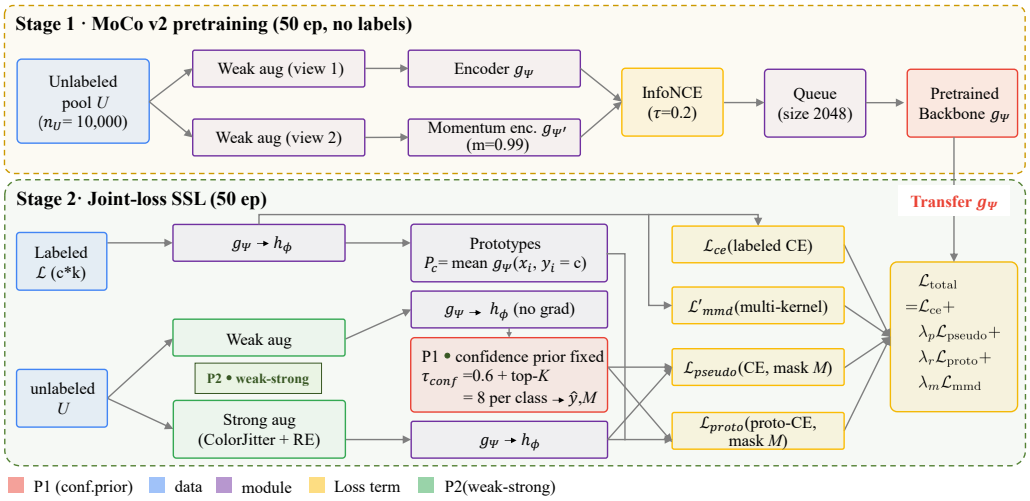


Figure 1: Overall pipeline. Stage 1 performs unlabeled MoCo pretraining. Stage 2 combines labeled cross-entropy, pseudo-label cross-entropy filtered by the Confidence Calibration Prior mask $\mathcal{M}$, prototype cross-entropy, and MMD alignment. Weak views generate pseudo-labels; strong views carry the consistency losses.

for learning representations from unlabeled data; we use MoCo only for representation initialization, while our main contribution targets pseudo-label selection. Further discussion of sparse-label SSL, confirmation bias, and self-supervised representation learning is provided in Appendix A.

3 Method

We propose a two-stage semi-supervised learning framework designed for the sparse-label regime, where only four labeled examples are available per class. Let $\mathcal{L} = \{(x_i, y_i)\}_{i=1}^{n_L}$ denote the labeled set and $\mathcal{U} = \{u_j\}_{j=1}^{n_U}$ the unlabeled pool, with $n_L = Ck$, $k = 4$, $C = 10$, and $n_U = 10{,}000$. The classifier is factorized as $f_\theta = h_\phi \circ g_\psi$, where g_ψ is a WRN-28-2 backbone with feature dimension $d = 128$, and h_ϕ is a linear classification head.

Stage 1 performs MoCo v2 pretraining (Chen et al., 2020b) on the unlabeled pool for 50 epochs, producing view-invariant representations without using class labels. Stage 2 initializes a new classification head and jointly optimizes the backbone and head for 50 epochs using:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ce}} + \lambda_p \mathcal{L}_{\text{pseudo}} + \lambda_r \mathcal{L}_{\text{proto}} + \lambda_m \mathcal{L}_{\text{mmd}}, \tag{1}$$

where $\lambda_p = 0.5$, $\lambda_r = 0.3$, and $\lambda_m = 0.3$. The proposed framework contains two method-defining components: a Confidence Calibration Prior that determines which unlabeled samples are trusted, and a Weak-Strong Consistency module that determines how their supervision is propagated. Figure 1 provides an overview of the complete pipeline.

3.1 Confidence Calibration Prior

Existing confidence-driven SSL methods select pseudo-labels using adaptive thresholds (Wang et al., 2023; Han et al., 2025) or confidence-variance partitions (Liu et al., 2026). These strategies implicitly assume that confidence values preserve meaningful differences between reliable and unreliable predictions. As shown in Section 4.4, this assumption breaks down under sparse supervision because the classification head rapidly becomes over-confident on unlabeled samples. Consequently, increasing the sophistication of the thresholding rule does not resolve the underlying selection problem.

We address this issue by introducing a Confidence Calibration Prior that limits the number of selected pseudo-labels independently of the global confidence distribution. For an unlabeled sample u_j, the weak-view prediction is:

$$p_j = \text{softmax}(h_\phi(g_\psi(u_j^{\text{w}}))), \tag{2}$$

with pseudo-label $\hat{y}_j = \arg\max_c p_{j,c}$ and confidence $\text{conf}_j = \max_c p_{j,c}$. We first apply a fixed confidence floor $\tau_{\text{conf}} = 0.6$ and retain only samples satisfying $\text{conf}_j \geq \tau_{\text{conf}}$. Unlike adaptive thresholds, this floor is not affected by the upward drift of confidence values caused by head overfitting.

More importantly, we impose a per-class top-K constraint. For each predicted class c, we define $\mathcal{S}_c$ and construct the selection mask ($\mathcal{M}$):

$$\mathcal{S}_c = \{j : \hat{y}_j = c,\ \text{conf}_j \geq \tau_{\text{conf}}\}, \tag{3}$$

$$\mathcal{M} = \bigcup_{c=1}^{C} \text{TopK}(\mathcal{S}_c, K), \tag{4}$$

where $K = 8$. Thus, at most K samples are selected for each predicted class within a mini-batch. This absolute class-wise constraint prevents confidence saturation from causing an uncontrolled increase in pseudo-label density. More importantly, it changes the selection principle from estimating a reliable global threshold to controlling the amount of trusted supervision. The proposed prior therefore remains well-defined even when the confidence distribution collapses.

3.2 Weak-Strong Consistency and distribution alignment

The Confidence Calibration Prior controls which pseudo-labels are used, while Weak-Strong Consistency determines how these selected samples contribute to training. For each unlabeled sample u_j, we generate a weak view u_j^{w} and a strongly augmented view u_j^{s}. The weak view produces pseudo-labels without gradient propagation, while the selected samples in $\mathcal{M}$ are supervised through their strong views. The standard pseudo-label consistency loss is:

$$\mathcal{L}_{\text{pseudo}} = \frac{1}{|\mathcal{M}|} \sum_{j \in \mathcal{M}} \text{CE}\left(h_\phi(g_\psi(u_j^{\text{s}})), \hat{y}_j\right). \tag{5}$$

To further stabilize learning from the extremely small labeled set, we construct class prototypes from labeled features and define prototype-based logits as:

$$P_c = \text{mean}\{g_\psi(x_i) : y_i = c\}, \tag{6}$$

$$\ell_{j,c} = -\frac{\|g_\psi(u_j^{\text{s}}) - P_c\|_2^2}{T}. \tag{7}$$

The corresponding prototype loss is:

$$\mathcal{L}_{\text{proto}} = \frac{1}{|\mathcal{M}|} \sum_{j \in \mathcal{M}} \text{CE}(\ell_j, \hat{y}_j). \tag{8}$$

Both losses use the same mask $\mathcal{M}$, ensuring that unreliable pseudo-labels do not directly enter either supervision branch. We additionally apply a multi-kernel MMD regularizer (Gretton et al., 2012) to reduce the distribution discrepancy between labeled and unlabeled representations:

$$\mathcal{L}_{\text{mmd}} = \text{MMD}^2\left(\{g_\psi(x_i)\}, \{g_\psi(u_j^{\text{w}})\}; \Sigma\right), \tag{9}$$

where $\Sigma = \{0.5, 1.0, 2.0, 4.0\}$. Unlike the pseudo-label losses, this term does not depend on pseudo-label correctness and therefore provides an additional distribution-level regularization signal.

The two proposed components are mutually reinforcing rather than independent. Weak-strong consistency encourages representations to remain stable across perturbations, improving the reliability of samples considered by the calibration prior. Conversely, the class-wise top-K constraint prevents confirmation bias from propagating through unreliable pseudo-labels. This interaction allows the framework to exploit consistency learning without relying on a well-calibrated confidence distribution, which is precisely the assumption violated in the sparse-label regime.

Table 1: Peak test accuracy (%) on CIFAR-10 / SVHN / STL-10 across three label densities. Labels/cls is abbreviated as L/c. Mean±std over $N = 4$ seeds; **bold** = best per row. WRN-28-2 backbone, 50 epochs, $|\mathcal{U}| = 10{,}000$. Per-seed values are provided in Appendix E.

Dataset	L/c	Sup-only	Mean	FixMatch	FreeMatch	CGMatch	RegMix	CLFP	CoVar	**Ours**
CIFAR-10	4	$22.05_{\pm 1.55}$	$22.73_{\pm 1.48}$	$24.97_{\pm 1.85}$	$25.33_{\pm 2.26}$	$26.56_{\pm 2.63}$	$29.82_{\pm 5.21}$	$23.32_{\pm 1.42}$	$25.53_{\pm 0.23}$	**32.79**$_{\pm 3.09}$
	25	$35.40_{\pm 2.28}$	$37.34_{\pm 1.82}$	$47.08_{\pm 5.26}$	$54.86_{\pm 2.95}$	$52.95_{\pm 1.63}$	**57.73**$_{\pm 2.50}$	$47.47_{\pm 2.29}$	$53.51_{\pm 2.84}$	$53.82_{\pm 2.11}$
	40	$40.20_{\pm 0.88}$	$42.91_{\pm 1.91}$	$52.17_{\pm 3.44}$	$59.04_{\pm 0.91}$	$55.41_{\pm 1.48}$	**61.78**$_{\pm 1.39}$	$51.36_{\pm 4.94}$	$58.27_{\pm 0.77}$	$57.18_{\pm 1.25}$
SVHN	4	$11.93_{\pm 0.31}$	$16.23_{\pm 1.23}$	$14.00_{\pm 0.68}$	$17.42_{\pm 1.72}$	$14.91_{\pm 0.87}$	$16.66_{\pm 2.86}$	$13.76_{\pm 1.60}$	$15.28_{\pm 3.09}$	**18.15**$_{\pm 3.35}$
	25	$27.66_{\pm 7.59}$	$45.12_{\pm 9.43}$	$70.86_{\pm 11.74}$	$80.30_{\pm 1.18}$	$79.86_{\pm 1.42}$	**84.67**$_{\pm 2.04}$	$76.64_{\pm 5.94}$	$83.82_{\pm 1.99}$	$69.75_{\pm 3.40}$
	40	$50.87_{\pm 8.22}$	$63.31_{\pm 3.76}$	$82.35_{\pm 1.52}$	$81.71_{\pm 2.09}$	$81.95_{\pm 1.75}$	$85.37_{\pm 4.32}$	$83.55_{\pm 1.89}$	**85.80**$_{\pm 1.48}$	$77.94_{\pm 1.39}$
STL-10	4	$22.69_{\pm 1.80}$	$23.85_{\pm 1.96}$	$27.47_{\pm 2.23}$	$29.40_{\pm 3.61}$	$28.37_{\pm 1.78}$	$30.36_{\pm 2.43}$	$26.90_{\pm 2.89}$	$28.32_{\pm 2.14}$	**32.47**$_{\pm 1.35}$
	25	$35.85_{\pm 1.97}$	$37.66_{\pm 1.29}$	$41.25_{\pm 1.10}$	$44.80_{\pm 1.98}$	$43.54_{\pm 1.80}$	**45.92**$_{\pm 1.39}$	$42.55_{\pm 1.01}$	$43.94_{\pm 1.72}$	$44.84_{\pm 1.49}$
	40	$39.65_{\pm 1.11}$	$41.68_{\pm 0.44}$	$46.33_{\pm 0.53}$	$47.80_{\pm 1.47}$	$47.57_{\pm 1.52}$	$48.90_{\pm 1.61}$	$45.51_{\pm 0.55}$	$47.75_{\pm 1.23}$	**49.34**$_{\pm 1.35}$

FixMatch seed 3 diverged on SVHN 25/class (peak 53.36%); result retained without cherry-picking. CGMatch and RegMixMatch each had one divergent seed on SVHN 40/class; the reported value is the $N = 3$ mean±std, following the same treatment as FixMatch. CoVar has no divergent seeds. Full per-seed data are provided in Appendix E.

4 Experiments

4.1 Datasets and Baselines

Datasets and label densities. We evaluate on three standard image-classification benchmarks: CIFAR-10 (Krizhevsky, 2009), SVHN (Netzer et al., 2011), and STL-10 (Coates et al., 2011). All images are converted to 32×32 RGB resolution, with STL-10 downsampled accordingly. We consider three label densities of 4, 25, and 40 labels per class, corresponding to $|\mathcal{L}| = 40$, 250, and 400 labeled samples, respectively. The unlabeled pool is fixed to $|\mathcal{U}| = 10{,}000$ across all datasets, while the test sets contain 10,000, 26,032, and 8,000 samples for CIFAR-10, SVHN, and STL-10, respectively. This setup explicitly emphasizes the sparse-label regime while also evaluating whether the observed behavior persists as the labeling budget increases.

Baselines. We compare against eight baselines covering several representative SSL paradigms. These include two classical baselines, Sup-only and Mean Teacher (shortened to Mean) (Tarvainen & Valpola, 2017); the fixed-threshold FixMatch (Sohn et al., 2020); three self-adaptive threshold methods, FreeMatch (Wang et al., 2023), CGMatch (Cheng et al., 2025), and RegMixMatch (shortened to RegMix) (Han et al., 2025); the complementary-label and feature-perturbation method CLFP (Wang et al., 2026); and the threshold-free confidence-variance method CoVar (Liu et al., 2026). Together, these baselines cover fixed-threshold, adaptive-threshold, complementary-label, and threshold-free selection strategies.

For a controlled comparison, all methods use the same WRN-28-2 backbone (Zagoruyko & Komodakis, 2016) with feature dimension $d = 128$ and approximately 1.47M parameters. We use a shared 50-epoch AdamW training protocol with learning rate 10^{-3}, weight decay 5×10^{-4}, and labeled/unlabeled batch sizes of $32/128$. Each method is evaluated with four random seeds for every dataset-density combination, without dataset-specific hyperparameter tuning. For CGMatch, RegMixMatch, CLFP, and CoVar, we implement the core components that characterize each method and verify each implementation against FreeMatch under the shared training protocol. Detailed implementation choices and verification results are provided in Appendix C.

4.2 Main Results

Table 1 reports peak test accuracy across the $3 \times 3 \times 9$ experimental grid, with $N = 4$ seeds for each setting. Figure 2 highlights the sparse-label regime, while Figure 3 shows performance trends across label densities. We summarize the results through three aggregate observations rather than enumerating individual table entries.

Sparse-label regime. With only 4 labels per class, the proposed method consistently outperforms all five recent SOTA baselines across the three datasets. It achieves higher accuracy in all 15 pairwise dataset-level comparisons, with a mean improvement of 4.34 pp and a standard deviation of 2.43 pp. The largest gains occur on CIFAR-10 and STL-10, whereas the margins on SVHN are

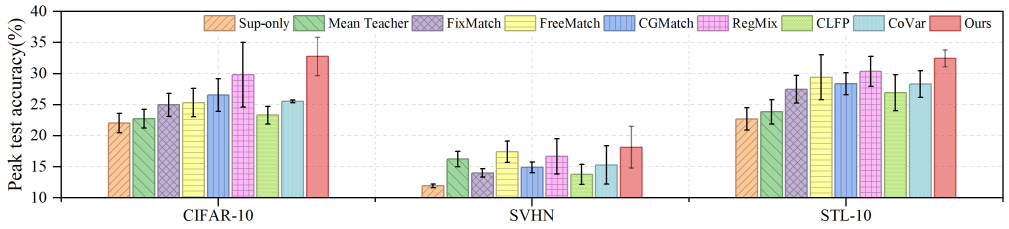


Figure 2: Sparse-label (4 labels/class) benchmark. The proposed method ranks first on all three datasets against eight baselines. Error bars: $\pm 1\sigma$ over $N = 4$ seeds.

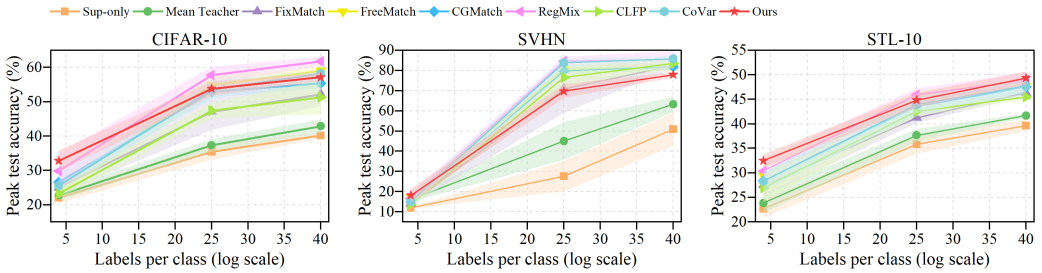


Figure 3: Peak test accuracy versus label density (log-scale x-axis). The proposed method (blue $\star$) dominates the sparse regime on CIFAR-10 and STL-10; at 25 and 40 labels per class, RegMixMatch and CoVar take the lead on several cells, and the SVHN mid-density dip is analyzed in Section 4.5. Shading: $\pm 1\sigma$ over 4 seeds; the two 2026 baselines are drawn as dashed lines.

smaller but remain positive across all five comparisons. The two smallest margins are 0.73 pp against FreeMatch and 1.49 pp against RegMixMatch on SVHN, both within one pooled standard deviation and therefore better viewed as comparable rather than decisive improvements. At this label density, the proposed method also ranks first among all nine methods on every dataset.

Mid- and dense-label regimes. As the labeling budget increases from 4 to 25 and 40 labels per class, the relative performance changes substantially. RegMixMatch achieves the highest accuracy on CIFAR-10 at 25 and 40 labels per class, reaching 57.73% and 61.78%, respectively, while CoVar achieves the highest accuracy on SVHN at 40 labels per class with 85.80%. On SVHN with 25 labels per class, four confidence-driven baselines outperform the proposed method by 10.11–14.92 pp. Specifically, FreeMatch, CGMatch, RegMixMatch, and CoVar exceed the proposed method by 10.55, 10.11, 14.92, and 14.07 pp, respectively. Per-epoch analysis in Sections 4.4 and 4.5 indicates that this gap is associated with a signal-volume bottleneck rather than a degradation in pseudo-label quality. Despite these mid-density differences, the proposed method remains competitive at higher label densities and achieves the highest accuracy on STL-10 with 40 labels per class.

Common failure mode of confidence-driven baselines. Under sparse supervision, several confidence-driven baselines exhibit substantial gaps between their peak and final test accuracy, indicating unstable training dynamics. For example, on CIFAR-10 with 4 labels per class, FreeMatch on seed 0 regresses by 15.6 pp between its mid-training peak and epoch 50, ending at only 10.7% test accuracy, and comparable single-seed regressions of 14.61 and 17.15 pp are observed for RegMixMatch on CIFAR-10 seed 3 and CoVar on STL-10 seed 1, respectively, with the latter being the largest single-seed drop among the baselines. Averaged over four seeds on CIFAR-10 with 4 labels per class, the peak-to-final drops are 8.32 pp for FreeMatch, 6.37 pp for RegMixMatch, and 9.02 pp for CoVar, compared with 1.67 pp for the proposed method. These observations are consistent with a common failure mode of confidence-driven pseudo-label selection when the classifier becomes over-confident under sparse supervision. Further analysis of pseudo-label quality, confidence distributions, and training dynamics is provided in Section 4.4.

Table 2: Ablation on the three datasets at 4 labels per class (WRN-28-2, $N = 4$ seeds). Δ vs Sup measures the gain over the supervised baseline. The full method combines the Confidence Calibration Prior and Weak-Strong Consistency.

Dataset	Configuration	Peak test acc. (%)	Δ vs Sup (pp)	Interpretation
CIFAR-10	Sup only	22.05±1.55	–	Labeled CE only
	Confidence Calibration Prior only	24.06±0.95	$+2.01$	Fixed floor + top-K cap, weak view
	Weak-Strong Consistency only	27.46±2.37	$+5.41$	Weak-strong aug + MoCo, no calibration prior
	Full (Ours)	**32.79±3.09**	$\mathbf{+10.74}$	Interaction: $+3.32$ over sum
SVHN	Sup only	11.93±0.31	–	Labeled CE only
	Confidence Calibration Prior only	12.17±1.06	$+0.24$	Fixed floor + top-K cap, weak view
	Weak-Strong Consistency only	14.84±1.02	$+2.91$	Weak-strong aug + MoCo, no calibration prior
	Full (Ours)	**18.15±3.35**	$\mathbf{+6.22}$	Interaction: $+3.07$ over sum
STL-10	Sup only	22.69±1.80	–	Labeled CE only
	Confidence Calibration Prior only	26.57±1.71	$+3.88$	Fixed floor + top-K cap, weak view
	Weak-Strong Consistency only	29.24±1.05	$+6.55$	Weak-strong aug + MoCo, no calibration prior
	Full (Ours)	**32.47±1.35**	$\mathbf{+9.78}$	Interaction: -0.65 vs sum

4.3 Ablation Study

Table 2 evaluates the two proposed components on CIFAR-10, SVHN, and STL-10 with 4 labels per class. We compare the supervised baseline, each component individually, and their full combination under the same WRN-28-2 architecture and training protocol. On CIFAR-10, the Confidence Calibration Prior and Weak-Strong Consistency improve accuracy by $+2.01$ pp and $+5.41$ pp, respectively, while their combination yields a $+10.74$ pp improvement. This corresponds to a $+3.32$ pp gain beyond the sum of the individual improvements, indicating a positive interaction between the two components.

A similar pattern appears on SVHN. The Confidence Calibration Prior alone provides a modest $+0.24$ pp improvement, while Weak-Strong Consistency contributes $+2.91$ pp. Combining both components increases the improvement to $+6.22$ pp, corresponding to a $+3.07$ pp positive interaction. These results suggest that the two components can reinforce each other under extremely sparse supervision, although their relative contributions vary across datasets.

STL-10 exhibits a different interaction pattern. The Confidence Calibration Prior alone produces the largest single-component improvement across the three datasets, increasing accuracy by $+3.88$ pp, while Weak-Strong Consistency contributes $+6.55$ pp. The full method achieves a $+9.78$ pp improvement, which is 0.65 pp below the additive expectation of $+10.43$ pp. Thus, the two components remain complementary, but their combination does not exhibit the positive interaction observed on CIFAR-10 and SVHN.

These results establish that both components contribute independently to sparse-label performance, while their combination provides additional benefits on two of the three datasets. We further analyze the underlying pseudo-label dynamics and the role of confidence calibration in Section 4.4. Additional results with a SmallCNN backbone are provided in Appendix G.

An additional ablation with a SmallCNN backbone is reported in Appendix G. It shows the same qualitative contribution pattern, while the Confidence Calibration Prior provides a relatively larger marginal gain than with WRN-28-2. This result suggests that the effectiveness of the calibration prior can depend on representation capacity, which we further examine in Appendix G.

4.4 Mechanism Analysis

To understand why confidence-driven baselines degrade under sparse supervision, we record per-epoch diagnostics for FreeMatch, RegMixMatch, CoVar, and our method on CIFAR-10 with 4 labels per class. Complementary diagnostics on SVHN at 4 and 25 labels per class are reported in Appendix F and support the boundary analysis in Section 4.5. We track four metrics: pseudo-label accuracy against the hidden ground truth (pl_acc), the fraction of unlabeled samples retained by the selection rule (keep ratio), mean maximum-softmax confidence, and test accuracy. Figure 4 shows the mean±std trajectories over four seeds, while Table 3 summarizes peak-to-final test-accuracy degradation and pseudo-label class balance at epoch 50.

Table 3: CIFAR-10 with 4 labels per class: mechanism-analysis summary over 4 seeds. Peak-to-final drop measures the regression from peak to final test accuracy. Class-kept $\min$ / $\max$ measures the balance of retained pseudo-labels at epoch 50, where 1.0 indicates perfect balance. Peak values are recorded from the mechanism-analysis runs, which enable per-epoch diagnostic logging; small deviations from the corresponding main-table cells (Table 1) are expected because per-epoch evaluation and logging perturb the stochastic training state.

Method	Peak (%) (4-seed mean)	Final (%) (4-seed mean)	Peak→final drop (pp) mean / max (pp)	Class-kept $\min$ / $\max$ (4-seed mean)
FreeMatch (SAT, 2023)	26.51±2.77	18.19±8.76	+8.32 / +15.58	0.029
RegMixMatch (SAT+SRM, 2025)	28.64±3.28	22.27±8.87	+6.37 / +16.13	0.048
CoVar (spectral, 2026)	25.82±1.21	16.80±5.83	+9.02 / +14.91	**0.000**
Ours	**32.02±3.16**	**30.35±2.56**	**+1.67 / +2.23**	**0.976**

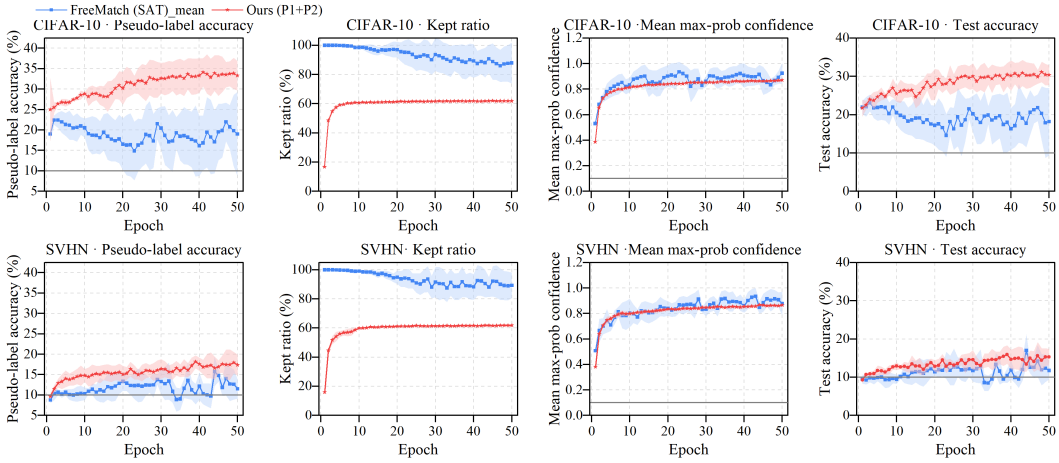


Figure 4: Mechanism analysis on CIFAR-10 (top) and SVHN (bottom) at 4 labels per class over 4 seeds. Shading denotes mean$\pm 1\sigma$. Columns show (1) pseudo-label accuracy against hidden ground truth, (2) keep ratio, (3) mean maximum-softmax confidence, and (4) test accuracy. Confidence-driven baselines exhibit confidence saturation and subsequent test-accuracy regression, whereas our method maintains a bounded keep ratio and more stable training dynamics.

First, FreeMatch exhibits a clear confidence-collapse pattern: mean confidence approaches one within the first ten epochs, while pseudo-label accuracy remains close to the 10% random baseline. Consequently, its SAT rule retains 73–100% of the unlabeled batch despite predominantly incorrect pseudo-labels, followed by substantial test-accuracy regression after the training peak. Second, our method avoids this pattern: the keep ratio stabilizes around 61–62% due to the per-class top-K constraint, while mean confidence increases to 0.86–0.88 without saturating near one. Meanwhile, pseudo-label accuracy improves progressively to 30–38%, and none of the four seeds exhibits the pronounced confidence-collapse behavior observed for FreeMatch.

Third, similar instability appears in RegMixMatch and CoVar despite their different selection mechanisms. RegMixMatch inherits the SAT-based selection mechanism and additionally applies a high-confidence gate ($\max_c p_c \geq 0.99$) to its RegMixup branch, which selects very few clean pseudo-labels under sparse supervision. CoVar instead performs threshold-free spectral partitioning on confidence-variance features, but this signal becomes less discriminative when maximum-softmax confidence approaches one across samples. Thus, both threshold-based and threshold-free strategies can lose effective discrimination when confidence becomes poorly calibrated.

Fourth, the final pseudo-label accuracy gap between our method and the confidence-driven baselines is accompanied by a corresponding test-accuracy gap. Together with the observed peak-to-final regressions, these results are consistent with pseudo-label quality and selection stability being important factors underlying sparse-label performance. The lower pseudo-label accuracy ceiling observed on SVHN further indicates that dataset difficulty affects the attainable quality of selected pseudo-labels. Overall, the diagnostics support our central hypothesis that confidence-driven selection becomes unreliable when sparse supervision produces an over-confident classification head.

4.5 DISCUSSION

The results clarify both the strength and the boundary of the proposed method. First, performance does not improve monotonically relative to existing methods as the labeling budget increases. On CIFAR-10, the gap against FreeMatch changes from $+7.46$ pp at 4 labels per class to -1.05 pp and -1.86 pp at 25 and 40 labels per class, respectively. On STL-10, the corresponding gaps remain positive at $+3.07$, $+0.05$, and $+1.54$ pp. In contrast, SVHN exhibits a pronounced mid-density reversal, with the gap changing from $+0.73$ pp to -10.55 pp and then to -3.77 pp.

The SVHN reversal at 25 labels per class appears to reflect a limitation of the fixed per-class selection budget rather than a degradation in pseudo-label quality. All four confidence-driven baselines outperform our method by 10–15 pp in this setting. Per-epoch diagnostics in Appendix F show that our method and FreeMatch both reach approximately 82% pseudo-label accuracy, while their keep ratios differ substantially, at approximately 62% and 92%, respectively. This suggests that once pseudo-label confidence becomes sufficiently informative, the fixed top-K constraint can become a bottleneck by limiting the volume of usable unlabeled supervision.

This observation also reveals the operating boundary of the Confidence Calibration Prior. Its main advantage is most pronounced when confidence scores are unreliable and pseudo-label selection would otherwise admit large numbers of incorrect samples. As the labeling budget increases and confidence becomes more informative, confidence-driven methods can exploit a larger fraction of the unlabeled pool without suffering the same degree of selection instability. Thus, the proposed method should be viewed as a sparse-label-oriented selection strategy rather than a universally superior replacement for confidence-driven SSL.

The mechanism analysis further suggests that both threshold-based and threshold-free confidence-driven methods can be affected by confidence collapse, although the precise failure dynamics differ across methods. When maximum-softmax confidence becomes nearly saturated, confidence-based discrimination provides less useful information for separating reliable from unreliable samples. Our Confidence Calibration Prior addresses this issue by imposing an absolute per-class selection budget rather than relying exclusively on the shape of the confidence distribution. This design trades some potential signal volume for greater control over pseudo-label selection, which explains both its robustness in the sparse regime and its reduced advantage when larger labeling budgets make confidence more informative.

Overall, the experiments identify a practical regime in which confidence calibration is particularly important: extremely sparse supervision combined with an over-confident classification head. Further analyses of label density, backbone capacity, and alternative explanations for the SVHN behavior are provided in Appendix D.

5 CONCLUSION

This work identifies a common failure mode of confidence-driven pseudo-label selection under extremely sparse supervision. As the classification head over-fits the small labeled set, maximum-softmax confidence can rapidly saturate, reducing its ability to distinguish reliable from unreliable pseudo-labels and leading to substantial test-accuracy regression. We observe this behavior across four representative baselines spanning threshold-based and threshold-free selection paradigms, indicating that the issue is not specific to a single selection rule. To address this problem, we propose a two-component framework that combines a Confidence Calibration Prior, consisting of a fixed confidence floor and per-class top-K constraint, with Weak-Strong Consistency built on MoCo initialization.

At 4 labels per class, the two components exhibit positive interaction on CIFAR-10 and SVHN, providing an additional $+3.32$ and $+3.07$ pp beyond the sum of their individual improvements, respectively, while their combination remains approximately additive on STL-10. Across CIFAR-10, SVHN, and STL-10, our method achieves the highest accuracy at this sparse-label density with $15/15$ favorable dataset-level comparisons and a mean margin of $+4.34$ pp; at higher label densities this advantage becomes dataset-dependent, revealing an operating boundary that motivates further investigation of confidence-robust pseudo-label selection strategies.

AI USE STATEMENT

In accordance with the ICLR 2027 AI Policy for Authors, the use of generative AI tools in the preparation of this manuscript is disclosed as follows. **Claude Sonnet 4.5** (Anthropic) was used for two tasks in the recommended-disclosure category: (i) *translation* of an initial Chinese draft into English, and (ii) *editing for readability*, including grammar correction, sentence-level rephrasing, and improving the flow of technical exposition.

Generative AI tools were *not* used for any research subtask in the required-disclosure category. Specifically, they were not used to design the methodology or experimental protocol, to formulate mathematical claims, to generate synthetic data, or to implement the proposed method or the baseline reproductions; formal proofs and theoretical derivations are not applicable to this work. Every number reported in Tables 1–14 and every per-epoch trajectory in Figure 4 was produced by an actual training run on the hardware described in Appendix H, and was cross-checked against the retained raw experiment logs; no result, citation, or numerical value in this paper was generated, hallucinated, or paraphrased by an AI tool without such verification.

All AI-assisted text was reviewed by the author(s), with technical claims re-verified against the underlying data or code and stylistic suggestions accepted only when they preserved the intended meaning. The author(s) take full responsibility for the final content of this paper, including all text, claims, figures, tables, and citations.

REFERENCES

E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *International Joint Conference on Neural Networks*, 2020.

D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel. MixMatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, 2019.

D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel. ReMixMatch: Semi-supervised learning with distribution alignment and augmentation anchoring. In *International Conference on Learning Representations*, 2020.

H. Chen, R. Tao, Y. Fan, Y. Wang, J. Wang, B. Schiele, X. Xie, B. Raj, and M. Savvides. SoftMatch: Addressing the quantity-quality trade-off in semi-supervised learning. In *International Conference on Learning Representations*, 2023.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020a.

X. Chen, H. Fan, R. Girshick, and K. He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

B. Cheng, J. Lu, Y. Tian, H. Zhao, Y. Chang, and L. Du. CGMatch: A different perspective of semi-supervised learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15381–15391, 2025.

A. Coates, A. Ng, and H. Lee. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics*, 2011.

Q. Deng, Y. Guo, Z. Yang, H. Pan, and J. Chen. Boosting semi-supervised learning with contrastive complementary labeling. *Neural Networks*, 170:417–426, 2024.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 2012.

J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, 2020.

H. Han, J. Yuan, C. Wei, and Z. Yu. RegMixMatch: Optimizing Mixup utilization in semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17032–17040, 2025.

K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

S. Laine and T. Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2017.

A. Le Coz, S. Herbin, and F. Adjed. Confidence calibration of classifiers with many classes. In *Advances in Neural Information Processing Systems*, 2024.

S. Li, W. Jin, Z. Wang, F. Wu, Z. Liu, C. Tan, and S. Z. Li. SemiReward: A general reward model for semi-supervised learning. In *International Conference on Learning Representations*, 2024.

J. Liu, P. Liu, and L. He. A confidence-variance theory for pseudo-label selection in semi-supervised learning. *arXiv preprint arXiv:2601.11670*, 2026.

Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

A. Oliver, A. Odena, C. A. Raffel, E. D. Cubuk, and I. Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*, 2018.

S.-A. Rebuffi, S. Ehrhardt, K. Han, A. Vedaldi, and A. Zisserman. Semi-supervised learning with scarce annotations. In *CVPR Workshops*, 2020.

M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *International Conference on Learning Representations*, 2021.

K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, 2020.

M. Sosea and C. Caragea. MarginMatch: Improving semi-supervised learning with pseudo-margins. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

J. C. Su, Z. Cheng, and S. Maji. A realistic evaluation of semi-supervised learning for fine-grained classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, 2017.

H. Vishwakarma, Y. Chen, S. S. S. Namburi GNVV, S. J. Tay, R. K. Vinayak, and F. Sala. Rethinking confidence scores and thresholds in pseudolabeling-based SSL. In *International Conference on Machine Learning*, 2025.

K. Wang, J. Yang, Y. Guo, and A. Xue. Unified complementary learning with feature perturbation for semi-supervised learning. *Algorithms*, 19(1):56, 2026.

Y. Wang, H. Chen, Q. Heng, W. Hou, Y. Fan, Z. Wu, J. Wang, M. Savvides, T. Shinozaki, B. Raj, et al. FreeMatch: Self-adaptive thresholding for semi-supervised learning. In *International Conference on Learning Representations*, 2023.

S. Zagoruyko and N. Komodakis. Wide residual networks. In *British Machine Vision Conference*, 2016.

B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, and T. Shinozaki. FlexMatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *Advances in Neural Information Processing Systems*, 2021.

A Related Work

A.1 Consistency-based and confidence-driven SSL

Modern SSL predominantly combines consistency regularization with confidence-driven pseudo-label selection. Early approaches, including the Π-model (Laine & Aila, 2017) and Mean Teacher (Tarvainen & Valpola, 2017), established consistency training under input perturbations. MixMatch (Berthelot et al., 2019) and ReMixMatch (Berthelot et al., 2020) further improved this paradigm through temperature sharpening and MixUp augmentation. FixMatch (Sohn et al., 2020) subsequently simplified SSL with weak/strong augmentation and a fixed confidence threshold, becoming a standard baseline for later methods.

A substantial body of work has since focused on improving pseudo-label selection. FlexMatch (Zhang et al., 2021) introduces curriculum pseudo-labeling with class-dependent thresholds, while FreeMatch (Wang et al., 2023) adapts thresholds using the exponential moving average of maximum probabilities and a per-class fairness regularizer. SoftMatch (Chen et al., 2023) assigns confidence-dependent weights through a truncated Gaussian to balance pseudo-label quantity and quality. MarginMatch (Sosea & Caragea, 2024) exploits confidence margins across training epochs, while CGMatch (Cheng et al., 2025) uses dual masks to separately handle easy and ambiguous samples. RegMixMatch (Han et al., 2025) further combines adaptive thresholding with a semi-supervised RegMixup branch. More recently, CoVar (Liu et al., 2026) replaces scalar thresholding with an SVD-based spectral partition over confidence and variance. Other studies, such as SemiReward (Li et al., 2024) and PabLO (Vishwakarma et al., 2025), explicitly address softmax miscalibration by learning reward models or confidence estimators. Despite these advances, existing approaches still rely on confidence as a meaningful signal for distinguishing reliable from unreliable unlabeled samples. Whether this assumption remains valid under extremely sparse supervision has received little systematic investigation, leaving confidence collapse largely unaddressed.

A.2 Sparse-label SSL and confirmation bias

Sparse-label SSL has received considerably less attention than conventional SSL settings with moderate labeling budgets. FixMatch (Sohn et al., 2020) reports results with 4 labels per class on CIFAR-10, but does not systematically investigate the corresponding failure modes. More recent methods, including FreeMatch, CGMatch, RegMixMatch, and CoVar, are typically evaluated with at least 25 labels per class, leaving their behavior under more severe label scarcity unclear. A complementary line of research attempts to exploit low-confidence samples rather than discarding them. CCL (Deng et al., 2024) extracts reliable negative information from the least-probable classes, while CLFP (Wang et al., 2026) incorporates feature-level dropout perturbations and supervised contrastive learning. These methods expand the information available from unlabeled data, but do not directly address the possibility that the classification head becomes systematically over-confident under sparse supervision.

Confirmation bias provides another perspective on pseudo-labeling failures. Prior analyses (Arazo et al., 2020; Rizve et al., 2021) show that erroneous pseudo-labels can reinforce themselves during training, while broader studies (Oliver et al., 2018; Su et al., 2021) highlight the sensitivity of SSL to realistic labeling budgets. However, these studies primarily characterize pseudo-label pollution rather than identifying confidence collapse as a specific mechanism underlying selection failure. Our work instead focuses on this mechanism and develops a selection strategy that does not depend solely on the shape of the confidence distribution. Finally, self-supervised representation learning methods such as MoCo (He et al., 2020), SimCLR (Chen et al., 2020a;b), and BYOL (Grill et al., 2020) provide complementary approaches for learning robust visual representations. We employ MoCo as a Stage-1 initialization strategy, while the proposed calibration prior directly addresses the pseudo-label selection problem.

Algorithm 1 Stage-2 training step: Confidence Calibration Prior plus Weak-Strong Consistency.

Require: Labeled batch $B_L = \{(x_i, y_i)\}$; unlabeled batch $B_U = \{u_j\}$; backbone g_ψ; head h_ϕ; fixed confidence floor τ_{conf}; per-class cap K; loss weights $\lambda_p, \lambda_r, \lambda_m$; MMD bandwidths Σ.

$z_i^L \leftarrow g_\psi(x_i)$; $\mathcal{L}_{\text{ce}} \leftarrow \text{mean}\{\text{CE}(h_\phi(z_i^L), y_i)\}$. ▷ Labeled loss.

$z_j^{\text{w}} \leftarrow g_\psi(u_j^{\text{w}})$; $p_j \leftarrow \text{softmax}(h_\phi(z_j^{\text{w}}))$; $\hat{y}_j \leftarrow \arg\max_c p_{j,c}$; $\text{conf}_j \leftarrow \max_c p_{j,c}$. ▷ Weak-view head predictions.

$\mathcal{S}_c \leftarrow \{j : \hat{y}_j = c \wedge \text{conf}_j \geq \tau_{\text{conf}}\}$ for $c = 1, \ldots, C$.

$\mathcal{M} \leftarrow \bigcup_c \text{TopK}(\mathcal{S}_c, K)$. ▷ Confidence Calibration Prior mask.

$z_j^{\text{s}} \leftarrow g_\psi(u_j^{\text{s}})$ for $j \in \mathcal{M}$. ▷ Weak-Strong Consistency.

$\mathcal{L}_{\text{pseudo}} \leftarrow |\mathcal{M}|^{-1} \sum_{j \in \mathcal{M}} \text{CE}(h_\phi(z_j^{\text{s}}), \hat{y}_j)$.

$P_c \leftarrow \text{mean}\{z_i^L : y_i = c\}$; $\ell_{j,c} \leftarrow -\|z_j^{\text{s}} - P_c\|_2^2/T$; $\mathcal{L}_{\text{proto}} \leftarrow |\mathcal{M}|^{-1} \sum_{j \in \mathcal{M}} \text{CE}(\ell_j, \hat{y}_j)$.

$\mathcal{L}_{\text{mmd}} \leftarrow \text{MMD}^2(\{z_i^L\}, \{z_j^{\text{w}}\}; \Sigma)$.

$\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{ce}} + \lambda_p \mathcal{L}_{\text{pseudo}} + \lambda_r \mathcal{L}_{\text{proto}} + \lambda_m \mathcal{L}_{\text{mmd}}$.

Update (ψ, ϕ) by one AdamW step on $\mathcal{L}_{\text{total}}$.

B Algorithm and Hyperparameters

This appendix collects the implementation-level specification of the proposed Stage-2 training step. Algorithm 1 details the per-batch computation of the Confidence Calibration Prior mask and the Weak-Strong Consistency losses introduced in Section 3, and Table 4 lists the five method-defining hyperparameters held fixed across all datasets and label densities. Together they provide a self-contained reference for reproducing the proposed method.

Table 4: Method-defining hyperparameters, held fixed across the three datasets and three label densities of the main evaluation. Shared training details common to all baselines (WRN-28-2 backbone, 50-epoch AdamW at lr $= 10^{-3}$, weight decay 5×10^{-4}, labeled/unlabeled batch sizes $32/128$) are given in Section 4.1; Stage-1 MoCo v2 pretraining settings (50 epochs, $d = 128$, $|Q| = 2048$, $m = 0.99$, $\tau = 0.2$), the weak/strong augmentation recipes, and the RBF bandwidths $\Sigma = \{0.5, 1, 2, 4\}$ are standard defaults and stated inline in the preceding paragraphs of Section 3.

Component	Symbol	Value
Fixed confidence floor	τ_{conf}	0.6
Per-class top-K cap	K	8
Pseudo-label loss weight	λ_p	0.5
Prototype loss weight	λ_r	0.3
MMD loss weight	λ_m	0.3

C Baseline Implementations

The main text (Section 4.1) compresses baseline descriptions to a single paragraph. This appendix collects the detailed per-baseline implementation notes for all eight comparison baselines evaluated in Table 1. Sections C.1–C.4 describe the four standard baselines (Sup-only, Mean Teacher, FixMatch, FreeMatch), which are implemented as-published under the shared 50-epoch AdamW protocol without simplification. Sections C.5–C.8 describe the four recent baselines (CGMatch, RegMixMatch, CLFP, CoVar) that are simplified in-house reproductions, each identifying the components retained and omitted relative to the published method and each accompanied by a single-point verification against FreeMatch under the shared protocol to confirm that the simplified version occupies the same baseline stratum as the original. Throughout this appendix, verification runs use a standalone single-seed sanity-check script whose seeding and data-loader initialization differ from the main-table pipeline; the goal is only to confirm each simplified reproduction reaches a baseline stratum comparable to FreeMatch under the sanity-check protocol, not to reproduce the corresponding main-table seed 0 values reported in Appendix E.

C.1 Sup-only Baseline

The Sup-only baseline is the labeled-set lower bound: WRN-28-2 trained by cross-entropy on the labeled subset alone, with no access to the unlabeled pool. The shared 50-epoch AdamW protocol applies (learning rate 10^{-3}, weight decay 5×10^{-4}, labeled batch size 32, cosine annealing schedule matched to the SSL baselines). Only weak augmentation (random crop with 4-pixel padding,

horizontal flip, normalize) is applied; no strong augmentation and no consistency loss. This baseline is the reference for computing Δ vs Sup in Table 2 and for verifying that each SSL method extracts non-trivial signal from the unlabeled pool: the sparse-label numbers in the first column of Table 1 (22.05, 11.93, 22.69 pp for CIFAR-10, SVHN, STL-10 respectively) are consistent with the $\sim n_L/C$ scaling expected when the network memorizes 40 examples across 10 classes.

C.2 Mean Teacher Baseline

The Mean Teacher baseline follows Tarvainen & Valpola (2017): a student network updated by gradient descent, and a teacher network whose weights are the exponential moving average of the student's (EMA decay 0.999). The consistency loss is mean-squared error between student and teacher predictions on the unlabeled batch under independent weak-augmentation draws, and no confidence-based masking is applied. A linear ramp-up over the first 10 epochs scales the consistency weight from 0 to $\lambda_c = 1.0$; both networks share the WRN-28-2 backbone and the shared 50-epoch AdamW protocol. Because Mean Teacher predates the confidence-driven family, it serves as an important reference point in the sparse-label comparison: its performance (22.73, 16.23, 23.85 pp at 4 labels per class on the three datasets) tracks Sup-only more closely than the confidence-driven methods do, which is consistent with the interpretation that consistency without confidence gating is neither victim nor beneficiary of the head-over-confidence pathology diagnosed in Section 4.4.

C.3 FixMatch Baseline

The FixMatch baseline follows Sohn et al. (2020). Weak augmentation (random crop, horizontal flip, normalize) generates pseudo-labels without gradient; strong augmentation (weak + RandAugment at magnitude 10 + Cutout of size 16) is used for the unsupervised cross-entropy loss on samples whose maximum-softmax probability exceeds the fixed threshold $\tau = 0.95$, with $\lambda_u = 1.0$. No class-balancing, no EMA teacher, and no adaptive threshold are used; the labeled loss is the standard cross-entropy under weak augmentation. FixMatch is the direct predecessor of the SAT family compared in this paper and provides the fixed-threshold reference point against which the confidence-collapse pattern of Section 4.4 is diagnosed: at 4 labels per class the $\tau = 0.95$ gate lets through very few samples early on but is quickly overwhelmed as the head over-fits, yielding the confirmation-bias trajectory that motivates the adaptive-threshold family and, ultimately, the calibration-prior alternative proposed here.

C.4 FreeMatch Baseline

The FreeMatch baseline reproduces the three core components of Wang et al. (2023): (i) a global self-adaptive threshold τ_t that tracks the exponential moving average of the batch-wide maximum-softmax mean with EMA decay 0.999; (ii) local per-class thresholds derived from the EMA of per-class softmax probabilities, combined with the global threshold as $\tau_t^c = (\tilde{p}_t^c / \max_{c'} \tilde{p}_t^{c'}) \cdot \tau_t$; and (iii) the self-adaptive fairness (SAF) regularizer that penalizes deviations of the model's marginal class distribution from uniform on the accepted high-confidence pseudo-labels. Weak and strong augmentation follow the FixMatch recipe (Appendix C.3), and the shared WRN-28-2 backbone and 50-epoch AdamW protocol apply. FreeMatch itself serves as the calibration anchor against which the simplified reproductions of Appendices C.5–C.8 are single-point verified, so no separate verification run is reported for this baseline; its per-seed values on the sparse cell (Table 5, $\{24.46, 25.41, 28.40, 23.06\}$ pp) are the reference distribution used to check whether the simplified 2025–2026 baselines occupy the same stratum under the shared protocol.

C.5 CGMatch Simplified Implementation

The CGMatch baseline is a simplified in-house reproduction retaining the three mechanisms most identifying to the published method (Cheng et al., 2025): EMA(max-probability) threshold, easy/ambiguous partition ($n_\sigma = 2.0$), and generalized cross-entropy loss ($q = 0.7$) on the ambiguous branch. Omitted for parity with the shared training protocol: memory bank, count-gap statistics, warm-up phase, distribution alignment hook, and non-linear ambiguous-weight schedule (fixed $(t/T)^2$ used). Single-point verification on CIFAR-10 at 40 labels per class (seed 0) yields

58.31% peak test accuracy, within 1σ of FreeMatch under identical conditions ($59.04 \pm 0.91\%$), confirming that the simplified version occupies the same baseline stratum.

C.6 RegMixMatch Simplified Implementation

The RegMixMatch baseline is a simplified reproduction of Han et al. (2025) retaining the full FreeMatch SAT backbone (SAT rule plus entropy fairness) and the semi-supervised RegMixup (SRM) branch gated on max-prob ≥ 0.99 with ResizeMix (tao $\in [0.1, 0.8]$, $\alpha = 1.0$), and omitting the class-aware mixup branch (per the official README's `disab_cam=True` recommendation) and the EMA teacher. Single-point verification on CIFAR-10 at 40 labels per class (seed 0) yields 62.22%, comfortably above FreeMatch under identical conditions and consistent with the paper's claim of a $+3$ pp improvement in dense-label regimes. The SRM branch's ≥ 0.99 gate admits essentially no clean pseudo-labels under sparse labels, so under 4 labels per class the method reverts to FreeMatch behavior — a design property (rather than a simplification) that drives the shared collapse pattern reported in Section 4.4.

C.7 CLFP Simplified Implementation

The CLFP baseline is a simplified reproduction of Wang et al. (2026) retaining two of the four original components: complementary-label negative learning on the $k = 7$ least-probable classes for low-confidence samples ($\mathcal{L}_{\text{neg}} = -\mathbb{E}\log(1 - p_c)$ over complementary labels), and feature-level dropout ($p = 0.5$) MSE consistency on high-confidence samples; the supervised contrastive branch and the dual strong-augmentation view branch are omitted to preserve the shared data-loader interface used by all baselines. Single-point verification on CIFAR-10 at 40 labels per class (seed 0) yields 54.83%, reaching a baseline stratum comparable to FreeMatch under the sanity-check protocol. The gap to the published 94.91% (SGD + 1024-epoch cosine schedule) is a protocol difference and is not comparable; the meaningful comparison is the within-protocol ranking on the main table, where CLFP sits above the Sup-only lower bound but below FixMatch and FreeMatch on dense labels and fails to close the sparse-label gap (Section 4.2).

C.8 CoVar Simplified Implementation

The CoVar baseline is a simplified reproduction of Liu et al. (2026), grafting the threshold-free confidence-variance mechanism onto a shared FixMatch weak/strong data loader. Retained: the $(\log C_j, g_j v_j)$ 2D embedding with $g_j = (K-1)^2/(2(1-C_j))$, SVD-based spectral partitioning into reliable and unreliable clusters, higher-mean-$\log C_j$ cluster taken as reliable, and Gaussian per-sample weighting ($w = \exp(-\text{conf_z}^2/\alpha) \cdot \exp(-\text{res_z}^2/\alpha)$, $\alpha = 2.0$, $w = 1$ inside the reliable cluster). Omitted: the SimPLE-specific pair loss, EMA teacher, and ramp-up scheduler, replaced by the shared FixMatch weak/strong cross-entropy with constant $\lambda_u = 1$. Single-point verification on CIFAR-10 at 40 labels per class (seed 0) yields 59.62%, within 0.5σ of FreeMatch under identical conditions (59.19% seed 0), consistent with the original paper's claim of near-parity with FreeMatch in well-calibrated dense-label regimes.

D Extended Discussion

This appendix collects the extended discussion material moved out of the main text (Section 4.5) to keep the main narrative focused.

D.1 Label-Density Dependence, Per-Dataset Breakdown

Three distinct label-density-versus-gain patterns are visible across the three datasets. On CIFAR-10 the gap against FreeMatch decreases monotonically from $+7.46$ (4/class) to -1.05 (25/class) to -1.86 pp (40/class); the 25/40 gaps are within 2σ and are interpreted as *on par*, whereas the 4/class gap has non-overlapping seed distributions. On STL-10 the gap is non-monotonic but always non-negative ($+3.07 \rightarrow +0.05 \rightarrow +1.54$ pp); STL-10 is the only dataset on which the proposed method dominates at every label density tested, an outcome consistent with prior findings that contrastive pretraining is more effective on higher-resolution or texture-rich imagery (Chen et al., 2020b). On

SVHN the gap varies non-monotonically as $+0.73 \to -10.55 \to -3.77$ pp; the middle point is not only negative but nearly an order of magnitude larger than either boundary, warranting the boundary analysis of Appendix D.2.

D.2 The SVHN Mid-Density Regime as a Method Boundary

Two candidate explanations for the SVHN mid-density loss are considered, and only the second is supported. *Alternative 1 (rejected): baseline-specific artifact.* If the loss were driven by an idiosyncratic property of FreeMatch, replacing FreeMatch with an independently designed baseline should change the sign. However, all four confidence-driven baselines yield gaps in the -10 to -15 pp range on SVHN 25/class ($\Delta_{\text{FM}} = -10.55$, $\Delta_{\text{CGM}} = -10.11$, $\Delta_{\text{RMM}} = -14.92$, $\Delta_{\text{CoVar}} = -14.07$ pp), spanning both threshold-based and threshold-free selection rules and four different publication years, so the gap is a general property of confidence-driven selection at that density rather than a defect of any one baseline. *Alternative 2 (supported): task-specific ceiling on the calibration prior's contribution.* On SVHN 25/class, FreeMatch reaches pseudo-label accuracy of 82.6% (4-seed mean) with a keep ratio of 91.7%, indicating that SAT is no longer collapsing but operating on high-quality pseudo-labels; the proposed method attains a comparable pseudo-label accuracy (81.7%) but retains only 61.7% of the batch, so effective training signal per batch drops from $\sim 77\%$ (FreeMatch) to $\sim 51\%$ (proposed), and the deficit accumulates over 50 epochs into the observed -10.55 pp gap. At 40 labels per class the effect diminishes as the baselines saturate near 82% and the proposed method catches up to 78%, giving the smaller -3.77 pp gap. The method's applicability boundary is thus characterized by two coupled conditions: *low task complexity* (few classes with visually distinct patterns) combined with *mid-density labels* where confidence-driven selection is neither collapsing nor saturated.

D.3 Confidence-Collapse Mechanism Across Selection Paradigms

The theoretical basis of confidence-driven selection — whether FreeMatch's moment-matched threshold, CGMatch's easy/ambiguous partition, RegMixMatch's SAT-plus-strict-gate hybrid, or CoVar's spectral partition on $(\log C_j, g_j v_j)$ — implicitly assumes that the maximum-softmax confidence or its residual-variance summary carries discriminative variance across the unlabeled pool. This assumption is violated when the classification head over-fits a small labeled set to the point of producing one-hot outputs on essentially every unlabeled sample: $\max_c p_c \to 1$ and $v_j \to 0$ uniformly, so any confidence-based selection accepts nearly all pseudo-labels regardless of correctness. The over-confidence of modern classification heads with many classes is itself a well-documented calibration failure (Le Coz et al., 2024), providing an independent anchor for the mechanism observed here. Neither a stricter high-confidence gate (RegMixMatch's ≥ 0.99) nor a spectral partition (CoVar's SVD-based split) restores discriminative power in this regime; the former disables its augmentation, and the latter commits to an incorrect reliable cluster. The Confidence Calibration Prior avoids this failure not by choosing a better threshold formula but by bypassing the threshold-selection problem: the per-class top-K cap accepts a fixed absolute number of pseudo-labels per class regardless of the shape of the confidence distribution, so it is well-defined even when confidence has collapsed. The fixed floor $\tau_{\text{conf}} = 0.6$ is a secondary safeguard that primarily matters in the very early training epochs before the head has over-fit.

D.4 The Calibration Prior as a Capacity-Conditional Regularizer

Comparing the WRN-28-2 ablation on CIFAR-10 at 4 labels per class (Table 2) with the earlier SmallCNN ablation on SVHN at 40 labels per class (Appendix G) reveals a systematic pattern: the marginal contribution of the Confidence Calibration Prior is larger on smaller backbones. On SmallCNN plus SVHN, the calibration prior is worth $+17.91$ pp on top of the Weak-Strong Consistency baseline; on WRN-28-2 plus CIFAR-10 the corresponding gap is $+5.33$ pp. Both signs are positive; only the magnitude changes. This scaling is consistent with the interpretation of the calibration prior as a regularizer against confidence over-fitting whose effect depends on how prone the model is to over-fit. Smaller backbones over-fit the labeled set more severely and therefore benefit more from a hard filter on confidence-based selection; larger backbones can afford some over-confidence and still learn discriminative features, so the marginal benefit shrinks — but does not vanish. Absolute peak test accuracy is higher on the larger backbone in every configuration, as expected.

E PER-SEED RESULTS

Tables 5–13 report the per-seed peak test accuracy underlying each cell of the main table. All cells are complete at $N = 4$ seeds; the mean±std of each row matches the corresponding main-table cell up to rounding. Three anomalies flagged in Table 1 are visible in the raw data: FixMatch seed 3 on SVHN 25/class (peak 53.36%), CGMatch seed 2 on SVHN 40/class (peak 11.07%), and RegMixMatch seed 2 on SVHN 40/class (peak 19.59%). All are retained without cherry-picking.

Table 5: Per-seed peak test accuracy (%) on CIFAR-10 at 4 labels per class.

Method	seed 0	seed 1	seed 2	seed 3	mean±std
Sup-only	20.39	24.13	21.91	21.79	22.05±1.55
Mean Teacher	20.99	24.59	22.83	22.52	22.73±1.48
FixMatch	22.77	24.34	25.64	27.12	24.97±1.85
FreeMatch	24.46	25.41	28.40	23.06	25.33±2.26
CGMatch	26.07	25.53	30.35	24.29	26.56±2.63
RegMixMatch	32.05	26.49	36.05	24.68	29.82±5.21
CLFP	22.67	23.47	25.24	21.92	23.32±1.42
CoVar	25.73	25.46	25.70	25.23	25.53±0.23
Ours	28.88	33.88	32.18	36.23	32.79±3.09

Table 6: Per-seed peak test accuracy (%) on SVHN at 4 labels per class.

Method	seed 0	seed 1	seed 2	seed 3	mean±std
Sup-only	11.80	11.63	11.92	12.35	11.93±0.31
Mean Teacher	17.76	15.87	14.82	16.49	16.23±1.23
FixMatch	13.13	14.74	13.86	14.26	14.00±0.68
FreeMatch	15.64	17.26	17.01	19.77	17.42±1.72
CGMatch	14.07	16.00	15.22	14.35	14.91±0.88
RegMixMatch	12.76	17.08	17.14	19.66	16.66±2.86
CLFP	14.54	12.75	12.14	15.62	13.76±1.60
CoVar	14.26	19.62	14.87	12.37	15.28±3.09
Ours	13.20	20.29	20.11	18.98	18.14±3.35

Table 7: Per-seed peak test accuracy (%) on STL-10 at 4 labels per class.

Method	seed 0	seed 1	seed 2	seed 3	mean±std
Sup-only	21.96	22.04	25.36	21.40	22.69±1.80
Mean Teacher	24.00	25.46	24.89	21.05	23.85±1.96
FixMatch	26.70	30.61	27.19	25.37	27.47±2.23
FreeMatch	30.11	28.46	33.86	25.17	29.40±3.61
CGMatch	28.56	28.56	30.35	26.01	28.37±1.79
RegMixMatch	31.76	30.53	32.26	26.89	30.36±2.43
CLFP	26.74	26.94	30.50	23.44	26.90±2.89
CoVar	29.28	29.30	29.60	25.11	28.32±2.14
Ours	34.15	32.75	32.02	30.94	32.47±1.35

Table 8: Per-seed peak test accuracy (%) on CIFAR-10 at 25 labels per class.

Method	seed 0	seed 1	seed 2	seed 3	mean±std
Sup-only	33.51	34.04	35.46	38.59	35.40±2.28
Mean Teacher	36.38	35.52	37.78	39.68	37.34±1.82
FixMatch	49.27	44.29	41.44	53.30	47.08±5.26
FreeMatch	54.77	55.63	50.98	58.08	54.86±2.95
CGMatch	53.43	51.72	51.60	55.04	52.95±1.63
RegMixMatch	58.97	56.86	54.69	60.41	57.73±2.50
CLFP	45.04	48.02	46.45	50.38	47.47±2.29
CoVar	51.04	53.51	52.01	57.48	53.51±2.84
Ours	52.95	53.45	52.01	56.86	53.82±2.11

F ADDITIONAL MECHANISM CURVES

Figure 5 extends the per-epoch mechanism analysis of Section 4.4 to SVHN at 25 labels per class, and Figure 6 visualizes the epoch-50 per-class distribution of retained pseudo-labels across all three diagnosed configurations.

Table 9: Per-seed peak test accuracy (%) on SVHN at 25 labels per class.

Method	seed 0	seed 1	seed 2	seed 3	mean±std
Sup-only	23.51	28.48	20.68	37.97	27.66±7.59
Mean Teacher	41.84	51.75	33.27	53.60	45.12±9.43
FixMatch	77.62	77.62	74.83	53.36	70.86±11.74
FreeMatch	81.83	80.12	78.95	80.32	80.30±1.18
CGMatch	80.08	81.20	77.87	80.30	79.86±1.41
RegMixMatch	85.49	85.86	81.61	85.73	84.67±2.04
CLFP	68.87	83.14	78.38	76.19	76.64±5.94
CoVar	84.19	85.78	81.05	84.26	83.82±1.99
Ours	66.00	70.95	68.22	73.85	69.75±3.40

Table 10: Per-seed peak test accuracy (%) on STL-10 at 25 labels per class.

Method	seed 0	seed 1	seed 2	seed 3	mean±std
Sup-only	38.21	36.64	34.79	33.77	35.85±1.97
Mean Teacher	38.91	37.12	36.10	38.51	37.66±1.29
FixMatch	40.09	40.58	42.41	41.93	41.25±1.10
FreeMatch	44.24	46.75	42.27	45.92	44.80±1.98
CGMatch	44.90	44.54	40.92	43.80	43.54±1.81
RegMixMatch	45.55	47.20	44.12	46.83	45.92±1.39
CLFP	41.49	41.90	43.55	43.25	42.55±1.01
CoVar	43.43	45.73	41.79	44.84	43.94±1.72
Ours	46.19	43.36	43.77	46.06	44.84±1.49

Table 11: Per-seed peak test accuracy (%) on CIFAR-10 at 40 labels per class.

Method	seed 0	seed 1	seed 2	seed 3	mean±std
Sup-only	40.47	39.41	39.61	41.32	40.20±0.88
Mean Teacher	42.37	40.81	43.05	45.41	42.91±1.91
FixMatch	57.30	49.90	50.62	50.88	52.17±3.44
FreeMatch	59.19	59.79	57.72	59.45	59.04±0.91
CGMatch	53.80	56.60	54.49	56.73	55.41±1.48
RegMixMatch	60.79	62.94	60.37	63.01	61.78±1.39
CLFP	45.26	55.68	49.43	55.05	51.36±4.94
CoVar	57.76	59.14	57.50	58.69	58.27±0.77
Ours	56.26	57.23	56.31	58.93	57.18±1.25

G SmallCNN Ablation on SVHN

The earlier ablation cited in Appendix D.4 that motivates interpreting the Confidence Calibration Prior as a capacity-conditional regularizer differs from the main ablation in three respects: SmallCNN backbone ($\sim$ 0.3M parameters) instead of WRN-28-2 ($\sim$ 1.47M); SVHN at 40 labels per class instead of CIFAR-10 at 4 labels per class; and a 15+15 epoch schedule instead of 50+50. Table 14 reports the two configurations that isolate the calibration prior's contribution on top of the Weak-Strong Consistency baseline.

Two features of this earlier study are worth noting. First, the seed-level distributions do not overlap: the maximum peak accuracy across four seeds of the Weak-Strong Consistency-only configuration is 33.83%, whereas the minimum peak of the full method is 39.19%, a ~ 5 pp gap between the two distributions; the calibration prior's contribution is therefore a seed-stable effect rather than a mean shift with high variance overlap. Second, the class balance improves from $\min / \max = 0.238$ to 0.493; the WRN-28-2 setting achieves higher balance (≥ 0.97, Figure 6) because the larger backbone learns discriminative features earlier, letting the per-class top-K cap saturate on high-confidence samples across all classes.

H Training Compute

All training runs were performed on a single laptop-grade RTX 4070 (8 GB VRAM), Intel Core i9-13900H, 32 GB system RAM, Windows 11, PyTorch 2.4 with CUDA 12. No cluster or cloud resources were used. Table 15 lists the training-run counts by experiment stage; total compute is approximately 175 GPU-hours across $\sim$376 WRN-28-2 training runs. The proposed method

Table 12: Per-seed peak test accuracy (%) on SVHN at 40 labels per class.

Method	seed 0	seed 1	seed 2	seed 3	mean±std
Sup-only	53.38	55.22	38.68	56.22	50.88±8.21
Mean Teacher	65.12	67.73	60.58	59.81	63.31±3.76
FixMatch	83.32	83.92	81.50	80.68	82.36±1.52
FreeMatch	78.98	82.31	81.54	84.01	81.71±2.09
CGMatch	80.24	83.42	11.07	82.20	64.23±35.47
RegMixMatch	80.03	88.40	19.59	87.67	68.92±33.10
CLFP	84.99	84.71	80.85	83.64	83.55±1.89
CoVar	84.11	87.40	85.08	86.60	85.80±1.48
Ours	78.15	76.25	77.72	79.63	77.94±1.39

Table 13: Per-seed peak test accuracy (%) on STL-10 at 40 labels per class.

Method	seed 0	seed 1	seed 2	seed 3	mean±std
Sup-only	39.69	40.83	39.92	38.16	39.65±1.11
Mean Teacher	41.75	42.21	41.14	41.64	41.69±0.44
FixMatch	46.45	46.11	46.99	45.75	46.33±0.53
FreeMatch	47.01	49.33	46.16	48.71	47.80±1.47
CGMatch	45.49	48.34	47.48	48.98	47.57±1.52
RegMixMatch	47.69	50.99	47.58	49.36	48.90±1.61
CLFP	44.92	45.17	46.10	45.84	45.51±0.55
CoVar	46.71	49.44	46.96	47.89	47.75±1.23
Ours	48.60	47.83	50.59	50.36	49.34±1.35

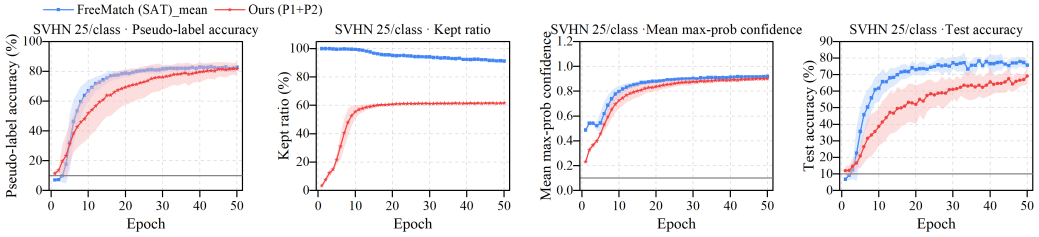


Figure 5: Mechanism analysis on SVHN at 25 labels per class (WRN-28-2, 50 epochs, 4-seed mean±std). FreeMatch and the proposed method reach comparable pseudo-label accuracy ($\sim$ 82% at epoch 50), but the proposed method retains only $\sim$ 62% of the batch versus FreeMatch's $\sim$ 92%. Neither method exhibits confidence collapse in this regime. The -10.55 pp test-accuracy gap is attributable to signal-volume starvation from the per-class top-K cap (Section 4.5, Appendix D.2).

additionally requires a 50-epoch Stage-1 MoCo pretraining before Stage 2, roughly doubling its per-run wall-clock relative to the confidence-driven baselines.

Training-loop-only compute (excluding data loading, checkpointing, and evaluation-time inference on the test set every epoch) is approximately 70% of the total wall-clock. The remaining 30% is dominated by per-epoch test-set evaluation on CIFAR-10 (10k) and SVHN (26k), retained because it is required by the per-epoch diagnostics of Section 4.

I LIMITATIONS

Two limitations bound the scope of the current study. First, the experiments were conducted under a limited compute budget of a single laptop-grade RTX 4070. The available budget was largely consumed by the 9 methods $\times$ 3 datasets $\times$ 3 label densities $\times$ 4 seeds evaluation, together with ablation, mechanism-analysis, and implementation-verification runs. Consequently, we could not perform a denser label-density sweep, such as 250 labels per class, or evaluate substantially larger backbones such as WRN-28-8 and ViT variants. Therefore, our conclusions primarily characterize the sparse-label regime rather than the complete performance curve across labeling budgets.

Second, the mechanism analysis in Section 4.4 could not be directly extended to STL-10 because its unlabeled split does not provide ground-truth labels for evaluating pseudo-label accuracy. Thus, the confidence-collapse behavior on STL-10 is supported indirectly by the observed peak-to-final

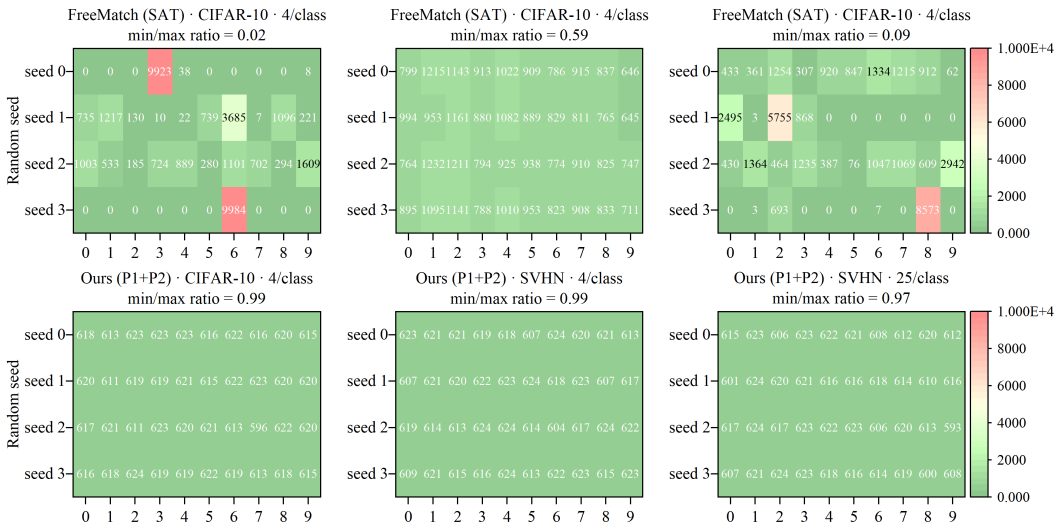


Figure 6: Per-class retained pseudo-labels at epoch 50 across three diagnosed configurations (CIFAR-10 4/class, SVHN 4/class, SVHN 25/class). FreeMatch (top row) is long-tailed under sparse labels (min/max ratio 0.02 on CIFAR-10 4/class, 0.09 on SVHN 4/class) and moderately imbalanced at mid-density (0.59 on SVHN 25/class). The proposed method (bottom row) is uniformly balanced (min/max ratio ≥ 0.97) as a direct consequence of the per-class top-K cap.

Table 14: SmallCNN ablation on SVHN at 40 labels per class (15+15 epochs, $N = 4$ seeds). The Confidence Calibration Prior's marginal contribution ($+17.91$ pp) is substantially larger than on WRN-28-2 CIFAR-10 4/class ($+5.33$ pp, Table 2), consistent with the capacity-conditional interpretation.

Configuration	Peak test acc. (%)	pl_acc (%)	Keep ratio (%)	Class min/max	Notes
Weak-Strong Consistency only	27.97±4.36	28.38±4.99	14.20±0.65	0.238	Confirmation-bias spiral
Full (Ours)	**45.88±5.42**	**66.40±5.58**	49.05±3.28	0.493	Stable
Δ (calibration prior)	$+17.91$	$+38.02$	$+34.85$	$+0.255$	Beyond noise

accuracy regressions rather than by direct pseudo-label diagnostics. An unlabeled benchmark with an accessible label oracle, or an equivalent evaluation protocol, would enable more direct verification of whether the same mechanism occurs on STL-10.

Table 15: Training-run inventory. Approximate wall-clock is per-run on the target hardware. Total compute is ∼175 GPU-hours across approximately 376 WRN-28-2 training runs. CCP-only and WSC-only in the ablation block refer to the Confidence-Calibration-Prior-only and Weak-Strong-Consistency-only configurations of Table 2, respectively.

Experiment stage	# runs	Per-run (min)	Notes
Main table (Table 1)			
Sup-only	36	∼8	3 ds × 3 dens × 4 seeds
Mean Teacher	36	∼12	3 ds × 3 dens × 4 seeds
FixMatch	36	∼15	3 ds × 3 dens × 4 seeds
FreeMatch	36	∼16	3 ds × 3 dens × 4 seeds
CGMatch	36	∼16	3 ds × 3 dens × 4 seeds
RegMixMatch	36	∼16	3 ds × 3 dens × 4 seeds
CLFP	36	∼15	3 ds × 3 dens × 4 seeds
CoVar	36	∼14	3 ds × 3 dens × 4 seeds
Ours	36	∼35	3 ds × 3 dens × 4 seeds; incl. Stage 1 MoCo
Ablation (Table 2)			
CCP-only, CIFAR-10 4/class	4	∼18	vs. Sup-only baseline (reused from main table)
WSC-only, CIFAR-10 4/class	4	∼30	incl. Stage 1 MoCo, no calibration prior
CCP-only, SVHN 4/class	4	∼18	–
WSC-only, SVHN 4/class	4	∼24	–
CCP-only, STL-10 4/class	4	∼91	STL-10 unlabeled load-heavy
WSC-only, STL-10 4/class	4	∼20–90	seeds 1–3 run in isolated subprocesses
Mechanism analysis (Figure 4, Appendix F)			
CIFAR-10 4/class	8	∼25–35	FreeMatch × 4 + Ours × 4
SVHN 4/class	8	∼25–35	FreeMatch × 4 + Ours × 4
SVHN 25/class	8	∼14–23	FreeMatch × 4 + Ours × 4
Verification runs (single-point smoke tests)			
CGMatch single-point	1	∼16	CIFAR-10 40/class seed 0 (Appendix C.5)
RegMixMatch single-point	1	∼16	CIFAR-10 40/class seed 0 (Appendix C.6)
CLFP single-point	1	∼15	CIFAR-10 40/class seed 0 (Appendix C.7)
CoVar single-point	1	∼14	CIFAR-10 40/class seed 0 (Appendix C.8)
Smoke tests (1-epoch)	–	<1 each	Not counted in totals
Total	**∼376**	–	∼175 GPU-hours end-to-end