WHY WEAK-TO-STRONG CONSISTENCY REGULARIZATION ENABLES FIXMATCH TO OUTPERFORM PSEUDO-LABELING: A FEATURE-LEARNING PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

FixMatch, a representative semi-supervised learning (SSL) method, has achieved strong performance and become a widely adopted baseline for subsequent SSL research. Its success is commonly attributed to the combination of pseudo-labeling and weak-to-strong consistency regularization. However, the interaction between these two components remains theoretically underexplored. To address this gap, we explain FixMatch's advantage over Pseudo-Labeling in theory by connecting feature learning with pseudo-label selection during training. With limited labeled data, a model initially relies on features that fit the labeled samples but generalize poorly, producing confident yet incorrect pseudo-labels. Pseudo-Labeling trains on the same weakly augmented images used to select these targets, leaving small prediction errors that can limit further feature learning. Incorrect targets can also reinforce poorly generalizing features and delay the selection of additional correct samples. FixMatch instead trains on strongly augmented versions of the selected images. Strong augmentation can weaken the features supporting current predictions and help the model learn shared features that improve predictions across more samples. More samples can then enter training with correct pseudo-labels, providing supervision for further feature learning. Under the conditions of our theory, this feedback can give FixMatch better representations, higher pseudo-label precision, and lower classification error than Pseudo-Labeling. However, weak-to-strong consistency regularization can also strengthen parameter updates that push the model toward predicting incorrect pseudo-labels. Experiments on CIFAR-10 and CIFAR-100 support key parts of this explanation: strong views increase learning signals toward both correct and incorrect pseudo-labels, FixMatch learns more discriminative features, and replacing the selected pseudo-labeled dataset changes subsequent feature learning and pseudo-label selection.

1 INTRODUCTION

By learning complex representations from data, deep neural networks have enabled substantial advances in fields such as computer vision and natural language processing (He et al., 2016; Vaswani et al., 2017; Radford et al., 2018; Dosovitskiy et al., 2020; Ho et al., 2020; Mildenhall et al., 2021; Ouyang et al., 2022; Schick et al., 2023). However, their strong performance often relies on large-scale labeled datasets, whose annotation is expensive, time-consuming, and dependent on human expertise (Ouali et al., 2020). Semi-supervised learning (SSL) addresses this limitation by leveraging abundant unlabeled data together with limited labeled data, thereby reducing annotation costs while improving model generalization (Van Engelen & Hoos, 2020).

Modern deep SSL methods are largely built upon two principles: pseudo-labeling and consistency regularization (Sohn et al., 2020). Pseudo-Labeling, a foundational SSL approach, converts confident model predictions on unlabeled samples into training targets, thereby extending supervision beyond the labeled set (Lee, 2013). Consistency regularization, in contrast, encourages the model to produce stable predictions under different perturbations or augmentations of the same input (Bachman et al., 2014; Sajjadi et al., 2016; Laine & Aila, 2016). FixMatch, a widely used SSL method,

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.

combines these two ideas by generating pseudo-labels from weakly augmented images and retaining predictions that exceed a confidence threshold. It then uses these pseudo-labels to supervise predictions on strongly augmented versions of the same images, introducing weak-to-strong consistency regularization into pseudo-labeling (Sohn et al., 2020). FixMatch achieved state-of-the-art performance at the time of its introduction (Sohn et al., 2020) and has since inspired influential FixMatch-based SSL methods, including Dash, FlexMatch, and FreeMatch (Xu et al., 2021; Zhang et al., 2021; Wang et al., 2023).

Despite this empirical success, the mechanism underlying FixMatch's advantage over Pseudo-Labeling remains unclear. The original FixMatch experiments demonstrated a substantial performance gap over Pseudo-Labeling and highlighted the importance of the strong training view (Sohn et al., 2020). However, these findings identify an effective design but do not explain the learning mechanism responsible for its success. Existing studies provide only partial explanations. Consistency-based analyses attribute the benefits to perturbation invariance, neighborhood expansion, and label propagation (Xie et al., 2020; Wei et al., 2021), whereas self-training studies emphasize confirmation bias and error accumulation arising from incorrect pseudo-labels (Arazo et al., 2020; Chen et al., 2022). More recently, Li et al. (2025) showed that strong augmentation can suppress learned semantic features and expose complementary ones. However, it remains unclear how weak-to-strong consistency regularization affects feature learning and pseudo-label selection in FixMatch compared with Pseudo-Labeling.

We develop a theory of how feature learning and pseudo-label selection interact during training, starting when models select unlabeled samples using features learned from limited labeled data. Features that fit this small labeled set may generalize poorly to other samples (Chen et al., 2022). Neural networks can also favor easy-to-learn patterns, including spurious features: patterns that predict a class in some samples but are unreliable across the broader population (Nam et al., 2020; Shah et al., 2020). We compare FixMatch and Pseudo-Labeling from the same model and initial selected samples, allowing each method to update its own predictions and selections as training proceeds.

Our central observation is that fitting confident, self-generated labels can preserve the features that produced them. In our comparison, Pseudo-Labeling trains on the same weakly augmented image used to generate its pseudo-label. Its target residual, defined as one minus the probability assigned to that pseudo-label, is therefore small (Mummadi et al., 2021). This leaves little prediction error to support learning beyond the current features. Incorrect pseudo-labels can further obstruct this process. A feature that produced an incorrect prediction can still fit the resulting training target, whereas a feature consistent with the true class may conflict with it. Our theory shows how newly selected incorrect labels reinforce reliance on spurious features and delay the selection of additional correct samples.

Weak-to-strong consistency regularization changes the information available for fitting these targets. Strong augmentation can weaken the features supporting a prediction while preserving other useful information (Shen et al., 2022; Li et al., 2025). The resulting prediction error can help FixMatch learn shared class-relevant features—patterns that support correct predictions across multiple samples. However, a larger error signal is not always beneficial: when the pseudo-label is wrong, it can also strengthen parameter updates toward the wrong class. Our analysis accounts for both effects and identifies sufficient conditions for useful learning to dominate.

To follow their consequences during training, we analyze a neural network in which different samples share feature-extraction parameters and can contain overlapping sets of features. Learning from selected samples therefore changes predictions on other samples. As shared features improve, additional correctly predicted samples become confident enough to enter training. These samples provide supervision for further feature learning, extending the model's ability to distinguish classes beyond the initially selected data. Meanwhile, incorrect selections can reinforce poorly generalizing features and slow this progression. By following both methods' evolving parameters and selections, we show how FixMatch can obtain more correct training targets and learn features that generalize to a broader population. In this model, its selections have higher pseudo-label precision, the fraction of selected pseudo-labels that are correct.

We test these mechanisms on CIFAR-10 with 40 labels and CIFAR-100 with 400 labels. We compare residuals for correct and incorrect pseudo-labels and replace selected datasets while holding the

starting model fixed. We then measure subsequent selection and assess feature quality using linear probes: new linear classifiers trained on frozen model features.

Our key contributions are summarized as follows:

First, we analyze how the limited error signal in Pseudo-Labeling affects subsequent feature learning and sample selection. Our theory shows how incorrect selections can reinforce spurious-feature dependence and delay further correct supervision.

Second, we establish the dual effect of weak-to-strong consistency regularization: it can promote useful feature learning but also strengthen updates toward incorrect targets. We derive sufficient conditions for the useful contribution to outweigh the error contribution and experimentally examine residual amplification for both target groups.

Finally, we develop a dynamic feature-learning theory connecting changes in learned features with changes in selected supervision. Under the conditions of our theory, we prove that FixMatch achieves higher pseudo-label precision and lower classification error throughout a finite interval of common training times. We also establish an advantage when both methods' features are evaluated using their best linear classifiers, showing that the difference extends to the learned representations.

2 Related Works

Modern Deep SSL Algorithms. Pseudo-Labeling (Lee, 2013) and consistency regularization (Bachman et al., 2014; Sajjadi et al., 2016) are fundamental components of many modern deep SSL algorithms. Classical consistency-based methods include the Π-model and Temporal Ensembling (Laine & Aila, 2016), Mean Teacher (Tarvainen & Valpola, 2017), and Virtual Adversarial Training (VAT) (Miyato et al., 2019). These approaches encourage stable predictions under input or model perturbations. FixMatch (Sohn et al., 2020) combines pseudo-labeling with weak-to-strong consistency regularization by training on strongly augmented images using confident weak-view targets. Following FixMatch, FlexMatch, Dash, and FreeMatch dynamically adjust pseudo-label selection criteria (Zhang et al., 2021; Xu et al., 2021; Wang et al., 2023), while SoftMatch replaces hard selection with confidence-based sample weighting (Chen et al., 2023).

Early Learning and Pseudo-Label Quality. Nam et al. (2020) show that neural networks may favor spurious features early in training when these features are easier to learn than the true target signal. Shah et al. (2020) show that this preference can persist even when more complex features are more predictive. In SSL settings, both the feature extractor and classifier can overfit a small labeled set, leading to biased predictions and unreliable pseudo-labels (Chen et al., 2022). When Pseudo-Labeling uses these predictions as training targets, errors arising from poorly generalizing features can affect later learning. We therefore study how weak-to-strong consistency regularization changes this interaction between feature learning and pseudo-label selection.

Explaining FixMatch's Advantage over Pseudo-Labeling. Several lines of research provide partial explanations for the benefits of weak-to-strong training. Wei et al. (2021) analyze consistency-regularized self-training under assumptions of expansion, class separation, and initial pseudo-labeler accuracy. Their theoretical guarantees concern population-objective minimizers rather than optimization trajectories. Mummadi et al. (2021) identify vanishing logit gradients caused by high-confidence hard pseudo-labels in test-time adaptation. Pezeshki et al. (2021) analyze gradient starvation, where learning one predictive feature suppresses the learning of others. In supervised learning, Shen et al. (2022) demonstrate that feature-balancing augmentations can expose informative views that may otherwise remain underutilized. Most closely related to our work, Li et al. (2025) show that strong augmentation can suppress learned semantic features and expose complementary ones, motivating SA-FixMatch. Their primary comparison is with supervised learning, in a regime where selected pseudo-labels are correct with high probability. Building on these studies, we follow feature updates and selection along both methods' training trajectories. Our model connects incorrect selections to spurious-feature reinforcement and connects new correct supervision to broader representation learning.

3 A THEORY OF FEATURE-LEARNING DYNAMICS IN FIXMATCH

Starting from a shared predictor that fits labeled samples but generalizes poorly, we study how training views and evolving pseudo-label selection shape feature learning.

3.1 A SHARED STARTING POINT FOR PSEUDO-LABEL TRAINING

FixMatch and Pseudo-Labeling start with the same architecture, parameters, labeled data, unlabeled population, and weak selection views. Their first selected samples and pseudo-labels are therefore identical. Afterwards, each method selects samples using its own current predictions.

Let x be an input, K the number of classes, and d the feature dimension. At training time $t \geq 0$, write the parameters as $\Theta_m(t) = (\psi_m(t), W_m(t))$, where $m = F$ denotes FixMatch and $m = P$ denotes Pseudo-Labeling. The feature map $h_\psi(x) \in \mathbb{R}^d$ has parameters ψ, and the classifier has weights $W \in \mathbb{R}^{K\times d}$. Define the representation Z_t^m, logits f_m, and class probabilities p_m by

$$Z_t^m(x) = h_{\psi_m(t)}(x), \qquad f_m(t,x) = W_m(t)Z_t^m(x), \qquad p_m(t,x) = \operatorname{softmax} f_m(t,x). \tag{1}$$

The common initialization gives $Z_0^F = Z_0^P$; it does not prescribe the order in which later features are learned.

Write x^w and x^s for weakly and strongly augmented versions of x. For class index $k \in \{1,\dots,K\}$, $p_{m,k}$ is the kth probability. The weak-view pseudo-label and selection indicator are

$$\widetilde{y}_m(t,x) = \arg\max_k p_{m,k}(t,x^w), \qquad \chi_m(t,x) = \mathbf{1}\{p_{m,\widetilde{y}_m(t,x)}(t,x^w) \geq \tau\}. \tag{2}$$

Here $\mathbf{1}\{\cdot\}$ is the indicator function; ties in the maximum are resolved by choosing the smallest class index. The confidence threshold τ is usually high; our experiments use 0.95. For target class y and probability vector p, cross-entropy is $\ell(y,p) = -\log p_y$. Suppressing the current time argument, both methods use

$$\mathcal{L}_m = \alpha\mathcal{L}_\ell + \lambda\mathbb{E}_u[\chi_m \ell(\widetilde{y}_m, p_m(x^{v_m}))], \qquad v_F = s, \quad v_P = w. \tag{3}$$

The supervised loss $\mathcal{L}_\ell$ averages over labeled weak views; $\alpha \geq 0$ and $\lambda > 0$ are common loss weights, and v_m denotes the training view. The unlabeled expectation $\mathbb{E}_u$ includes the whole population and augmentation randomness, with rejected samples contributing zero. Pseudo-labels and indicators are held fixed when computing a gradient and refreshed from the current model. Pseudo-Labeling fits the exact weak view used for selection; FixMatch fits a strong view x^s.

3.2 FROM PREDICTION RESIDUALS TO FEATURE UPDATES

Consider one selected sample at a shared model state, with fixed pseudo-label $\widetilde{y}$. For view $v \in \{w,s\}$, abbreviate its logits and probabilities as $f^v = Wh_\psi(x^v)$ and $p^v = \operatorname{softmax} f^v$, with components f_k^v and p_k^v. Its target residual is $r_v = 1 - p_{\widetilde{y}}^v$. Cross-entropy gives

$$\begin{aligned} r_w &\leq 1-\tau, & r_s - r_w &= p_{\widetilde{y}}^w - p_{\widetilde{y}}^s, \\ \frac{\partial \ell}{\partial f_k^v} &= p_k^v - \mathbf{1}\{k=\widetilde{y}\}, & \|\nabla_f \ell(\widetilde{y}, p^v)\|_1 &= 2r_v. \end{aligned} \tag{4}$$

Here $\|\cdot\|_1$ sums absolute component values. Thus, at $\tau = 0.95$, the same-view residual is at most 0.05. The predictions used for selection already fit their own targets closely. A strong view has a larger residual when it reduces the probability of the fixed target. This can happen when augmentation weakens the features supporting the weak-view prediction.

To connect this signal to feature learning, let $J_\psi(z) = \partial h_\psi(z)/\partial\psi$ be the feature Jacobian for input z at a differentiable state, and let $e_{\widetilde{y}}$ be the target's one-hot vector. The sample's negative gradients with respect to ψ and W, denoted by g_ψ^v and g_W^v, are

$$g_\psi^v = J_\psi(x^v)^\top W^\top (e_{\widetilde{y}} - p^v), \qquad g_W^v = (e_{\widetilde{y}} - p^v)h_\psi(x^v)^\top. \tag{5}$$

The residual supplies the prediction error; the feature response determines how that error changes the model. Retaining useful information can allow the increased residual to support feature learning, whereas deleting that information can remove its gradient. Appendices A.1 and A.2 derive these identities and the effect of an update on other samples.

3.3 CORRECT AND INCORRECT SUPERVISION

A larger residual also increases learning pressure toward an incorrect target. For a random sample (X, Y), Y is the true class, used only for analysis. Write $\widetilde{Y} = \widetilde{y}(X)$ and $\chi = \chi(X)$ for its pseudo-label and selection indicator at the shared state. Let $\mathcal{C} = \{\chi = 1, \widetilde{Y} = Y\}$ and $\mathcal{E} = \{\chi = 1, \widetilde{Y} \neq Y\}$ be correct and incorrect selection events. Their masses are $c = \Pr(\mathcal{C})$ and $e = \Pr(\mathcal{E})$, with precision $Q = c/(c + e)$ when $c + e > 0$. Probabilities and expectations include sample and paired-view randomness. Then

$$\mathbb{E}[\chi g_\psi^v] = c\,\mathbb{E}[g_\psi^v \mid \mathcal{C}] + e\,\mathbb{E}[g_\psi^v \mid \mathcal{E}]. \tag{6}$$

The two terms need not point in the same direction. For a specified differentiable representation score $\Phi(\psi)$, with larger values indicating better predictions under a fixed true-label readout, set $u = \nabla_\psi \Phi$ at the shared state. Define Δ_+ as the conditional mean of $\langle u, g_\psi^s - g_\psi^w \rangle$ on $\mathcal{C}$, and Δ_- as that of $\langle u, g_\psi^w - g_\psi^s \rangle$ on $\mathcal{E}$. They measure the additional useful contribution and the additional error cost along this direction; the latter may itself be negative. Under negative-gradient flow, a dot denotes a derivative with respect to training time, so $\dot{\psi}_m = d\psi_m/dt$.

Proposition 1 (A sufficient local advantage)**.** *Suppose $c + e > 0$. Let $A_+ > 0$ and $A_- \geq 0$ be constants bounding the conditional contributions: $\Delta_+ \geq A_+$ and $\Delta_- \leq A_-$. The two gradient-flow directions evaluated at this shared state satisfy*

$$\langle u, \dot{\psi}_F - \dot{\psi}_P \rangle \geq \lambda(c + e)\{QA_+ - (1 - Q)A_-\}. \tag{7}$$

The difference is positive if $Q > A_-/(A_+ + A_-)$.

Bounds on a conditional mean are required only when its event has positive probability; otherwise its weighted contribution is zero. The common supervised contribution cancels in this comparison. Appendix A.4 gives the proof and specifies how the direction and view-dependent gradients enter the bounds. Incorrect supervision can do more than slow learning: features that reproduce an erroneous target may be reinforced while class-relevant features receive opposing updates. The next model derives this effect and follows its consequences for later selection. Appendix A.5 also gives the contribution decomposition along each learner's own trajectory.

3.4 FEEDBACK BETWEEN FEATURE LEARNING AND PSEUDO-LABEL SELECTION

A network with shared features. We use a binary patch model to calculate how learning on selected samples affects other samples. Labels $Y \in \{-1, +1\}$ are balanced within each population group. A *feature pathway* contains class-specific filters and output weights, shared across samples containing the corresponding input pattern. Pathways $0, 1, 2, 3$ carry semantic features associated with the true class; pathway s carries a spurious feature that agrees with the labels on some groups and conflicts with them on others. Pathway 0 is the root of the cooccurrence structure because it appears in the initial selected samples. These indices name feature pathways, not training stages. Each pathway j has observable input channels $x_j^+, x_j^- \in \mathbb{R}^2$ for the positive and negative classes. Using a single logit for binary classification, the network is

$$f_\Theta(x) = \sum_{j \in \{0,1,2,3,s\}} \left(a_j^+ [\langle w_j^+, x_j^+ \rangle]_+ - a_j^- [\langle w_j^-, x_j^- \rangle]_+ \right). \tag{8}$$

Here $[z]_+ = \max(z, 0)$ is the ReLU activation; $w_j^\pm \in \mathbb{R}^2$ and $a_j^\pm \in \mathbb{R}$ are the network parameters. The representation is the vector of hidden activations in (8). The predicted positive-class probability is $\sigma(f_\Theta(x))$, where $\sigma(z) = (1 + e^{-z})^{-1}$; a zero logit predicts $+1$. Thus, the confidence gate is $|f_\Theta(x)| \geq \kappa$ with $\kappa = \log[\tau/(1 - \tau)]$ and $1/2 < \tau < 1$.

Let $u_0 = (1, 1)^\top$ and $q_0 = (0, 1)^\top$. An initial correct group of mass π_0 has inputs u_0 at the root and spurious pathway in its true-class channel. Three further groups, each of mass π_e, have semantic inputs u_0 in the true-class channel on pairs $(0, 1)$, $(0, 2)$, and $(1, 3)$. There is one evaluation group per semantic pathway, each of mass $\pi_T/4$, with only the input q_0 in that pathway's true-class channel. A group of mass ν has the root input in its true-class channel and the spurious input in the opposite channel. A further group of mass ω has the same root input but the scaled input ρu_0 in the opposite spurious channel, where $\rho > 0$. All other inputs are zero. All group masses are positive population fractions and satisfy $\pi_0 + 3\pi_e + \pi_T + \nu + \omega = 1$. The channel structure describes observed data; labels are not supplied to the pseudo-labeling rule. The labeled set contains one initial correct sample per class. All groups belong to the same unlabeled population.

Training and representation change. Weak views leave inputs unchanged. Strong views independently delete the spurious input with probability p_d and retain each semantic input with probability p_r, where $p_d, p_r \in (0, 1]$. This distribution stays fixed during training. We analyze population gradient flow for the binary-logit specialization of (3), using the logistic loss $\ell(y, f) = \log(1+e^{-yf})$ for $y \in \{-1, +1\}$ and scalar logit f. The two class channels have equal initial parameters; population symmetry preserves this equality.

We omit method and time indices when discussing either trajectory. For either channel, a_j, w_j denote its parameters. Define the hidden preactivation $s_j = \langle w_j, u_0 \rangle$, logit contribution $\beta_j = a_j s_j$, and spurious-feature contribution $S = \beta_s$. The level $h = \kappa/2$ will mark the onset of an evaluation response. Choose initial semantic strengths $0 < \varepsilon < \delta < h$ and set $b = \kappa - \varepsilon$, sufficient for a pair to pass the gate when its other pathway has strength at least ε. Both methods start from $\beta_0 = \delta$, $\beta_{1,2,3} = \varepsilon$, and $S = M$, where $M > 0$ is the initial spurious-feature contribution, with $s_j = \sqrt{2}a_j > 0$ on every pathway. Semantic filters satisfy $w_{j,1} - w_{j,2} = \sqrt{\kappa/\sqrt{2}}$; the spurious-feature filter has equal coordinates. Appendix B.1 gives the full setup and the parameter conditions. Choose ρ so that $0 < \rho M - \delta < \kappa$. Initially, the groups of mass π_0 and ν pass the gate; the additional group of mass ω is predicted incorrectly but rejected. The pair groups are correctly predicted but not yet confident, while both classes in each evaluation group have zero hidden response.

Until evaluation samples enter the loss, direct differentiation gives

$$\dot{a}_j = \tfrac{1}{2} s_j g_j, \qquad \dot{s}_j = a_j g_j, \qquad \dot{\beta}_j = \sqrt{2}\beta_j g_j, \qquad \gamma_j = \tfrac{1}{2}[\beta_j - \sqrt{h\beta_j}]_+. \tag{9}$$

Here g_j is the loss-weighted signed residual signal on pathway j, with the class probability $1/2$ factored out, and $\gamma_j = Y f_\Theta(x)$ is the true-label margin for an evaluation sample on semantic pathway $j \in \{0, 1, 2, 3\}$. Appendix B.2 gives the formulas for g_j and derives these identities. Updates first move the filter's preactivation. When $\beta_j > h$, the previously zero evaluation responses activate in different class channels, making those samples linearly distinguishable. This is a change in representation, not just a larger output weight.

Incorrect selections reinforce spurious features. The additional group's wrong-label margin is $q_m = \rho S^m - \beta_0^m$. It enters with an incorrect pseudo-label when $\chi_\omega^m := \mathbf{1}\{q_m \geq \kappa\} = 1$. Writing $r(z) = \sigma(-z)$, its added Pseudo-Labeling coefficients are

$$\Delta g_{0,P} = -\lambda\omega\chi_\omega^P r(q_P), \qquad \Delta g_{s,P} = \lambda\omega\rho\chi_\omega^P r(q_P). \tag{10}$$

They oppose the semantic feature while reinforcing the spurious feature. Before the first correct pair enters, their relative contributions satisfy

$$\frac{d}{dt}\log\frac{S^P}{\beta_0^P} = 2\sqrt{2}\lambda\nu r(S^P - \beta_0^P) + \sqrt{2}\lambda\omega(1+\rho)\chi_\omega^P r(q_P) > 0. \tag{11}$$

The hidden preactivation ratio s_s^P/s_0^P has the same direction of change. Thus, erroneous supervision increases dependence on the feature that generated the labels, even while the semantic feature continues to grow. Under the conditions below, q_P crosses the threshold, whereas q_F decreases and the group remains rejected. Comparing Pseudo-Labeling with an otherwise identical continuation that omits only this group's loss shows a strict delay in its first correct pair selection. These are consequences of the gradient and gate equations (Appendices B.5–B.6).

Correct selections support broader feature learning. A pair (i, j) enters when $\beta_i + \beta_j \geq \kappa$, supplying gradients to both pathways. Learning on pair $(0, 1)$ therefore updates pathway 1, which also appears in pair $(1, 3)$. Once the latter passes the threshold, it supplies a gradient to pathway 3. The shared filters also change the representations of evaluation samples that remain unselected. If the initial selected identities and targets are frozen, pathways $1, 2, 3$ receive no gradient. Their improvement therefore requires the new supervision (Appendix B.8).

Let $c_m(t)$ and $e_m(t)$ be the correctly and incorrectly selected population masses, $Q_m(t) = c_m(t)/(c_m(t) + e_m(t))$ their precision, and $\mathcal{R}_m(t)$ the trained classifier's error. For a representation Z, let $\mathcal{R}_{\text{lin}}(Z)$ be the smallest population error of an affine linear classifier on Z.

Comparing the two trajectories. Choose a horizon $H > 0$ and bounds $U > b$, $V > M$. Appendix B.3 gives parameter inequalities that keep both trajectories within $\varepsilon \le \beta_j \le U$ for $j \in \{0, 1, 2, 3\}$, $\beta_0 \ge \delta$, and $M \le S \le V$ through time H. The additional inequalities $M - U > \kappa$ and $(U - \sqrt{hU})/2 < \kappa$ keep the original wrong group confidently wrong and the evaluation groups unselected. Within these bounds, the FixMatch root has

$$g_0^F \ge \lambda p_r[p_d\pi_0 - \nu e^U]r(U) > 0. \tag{12}$$

The positive term comes from retaining the root when the spurious input is deleted; the negative term bounds the incorrect-target contribution. Thus $p_d\pi_0 > \nu e^U$ supplies a concrete sufficient condition for net useful progress in this model.

Integrating each learner's rate bounds gives an upper time bound $\overline{T}_F$ for FixMatch to learn all four evaluation representations, an upper bound $\overline{T}_P$ for Pseudo-Labeling to learn three, and a lower bound $\underline{T}_{P,\mathrm{far}}$ for Pseudo-Labeling to select the remaining pair. Let T_a bound its new erroneous admission time from above. Appendix B.7 derives the representation time bounds; Appendix B.5 derives T_a and shows that the new errors remain selected through H.

Theorem 2 (Representation and selection at a common training time)**.** *Suppose the parameters satisfy the explicit inequalities in Appendix B.3. These bound both trajectories through time H, keep evaluation groups unselected, and ensure*

$$t_- := \max(\overline{T}_F, \overline{T}_P, T_a) < t_+ := \min(H, \underline{T}_{P,\mathrm{far}}). \tag{13}$$

For every $t \in (t_-, t_+)$, both methods have selected new samples and learned new evaluation representations, but

$$\begin{aligned} c_F(t) &= \pi_0 + 3\pi_e > \pi_0 + 2\pi_e = c_P(t), & e_F(t) &= \nu < \nu + \omega = e_P(t), \\ Q_F(t) &> Q_P(t), & \mathcal{R}_F(t) &= \nu + \omega < \nu + \omega + \pi_T/8 = \mathcal{R}_P(t), \\ \mathcal{R}_{\mathrm{lin}}(Z_t^F) &= 0, & \mathcal{R}_{\mathrm{lin}}(Z_t^P) &= \pi_T/8. \end{aligned} \tag{14}$$

Why the comparison holds. FixMatch activates all four evaluation pathways by $\overline{T}_F$. Pseudo-Labeling has the root and two further pathways active after $\overline{T}_P$, but cannot select the last pair before $\underline{T}_{P,\mathrm{far}}$. Its remaining evaluation group still has an identical zero representation for both classes. This gives the error gap in (14). FixMatch also selects an additional correct pair and avoids the new erroneous admission, giving higher precision. Both classifiers still predict the new group incorrectly, so its mass ω is included in both risks. Appendices B.4–B.9 prove the bounds. A verified parameter choice gives strict differences throughout $[0.45, 0.48]$ in gradient-flow time (Appendix B.10).

4 Experiments

We test how the training view changes learning pressure, how selected supervision affects later features and selections, and how representation quality and pseudo-label precision evolve. We compare Pseudo-Labeling and FixMatch on CIFAR-10 with 40 labeled samples and CIFAR-100 with 400 labeled samples (Krizhevsky, 2009). Our implementation builds on the Unified Semi-supervised Learning Benchmark (USB; Wang et al., 2022).[1] Following Section 3, only the unlabeled training view differs: Pseudo-Labeling reuses the exact weak-view tensor used for selection. Both methods use WRN-28-2 (Zagoruyko & Komodakis, 2016), a confidence threshold of $\tau = 0.95$, and three paired training seeds. Paired 1024-epoch runs share initialization, labeled splits, and data-sampling controls. We follow USB's default FixMatch settings for batch sizes, the unlabeled loss weight, and augmentation operations. Tables report results as mean $\pm$ sample standard deviation (SD) across three training seeds. Appendix C.1 provides full implementation details.

4.1 Strong views increase the learning pressure

At each fixed checkpoint in evaluation mode, we compare selected samples' target residuals on the exact weak selection view and four random strong augmentations per image. Pseudo-labels stay fixed; true labels separate correct and incorrect targets.

[1] Official codebase: `https://github.com/microsoft/Semi-supervised-learning`.

Table 1: Pseudo-label residuals for FixMatch at epoch 64. Correct/Wrong indicate pseudo-label correctness. Same weak uses the selection view; Strong averages four augmentations per image.

Dataset	Targets	Same weak	Strong
CIFAR-10	Correct	0.0061 ± 0.0024	0.2016 ± 0.0644
CIFAR-10	Wrong	0.0132 ± 0.0021	0.2849 ± 0.0079
CIFAR-100	Correct	0.0081 ± 0.0009	0.2946 ± 0.0085
CIFAR-100	Wrong	0.0144 ± 0.0007	0.4081 ± 0.0088

Table 2: Effect of pseudo-labeled dataset source on continuations from epoch 64. Recipient denotes the model being continued. Entries are paired differences after 50 epochs: FixMatch-generated data minus Pseudo-Labeling-generated data. Test, probe, and Q_{new} differences are in percentage points (pp); C_{new} differences are sample counts (mean $\pm$ sample SD across three training seeds).

Dataset	Recipient	Δ test	Δ probe	ΔC_{new}	ΔQ_{new}
CIFAR-10	FixMatch	43.76 ± 26.51	22.27 ± 10.99	172.0 ± 51.0	33.72 ± 25.00
	Pseudo-Labeling	47.23 ± 29.73	31.69 ± 13.42	158.3 ± 17.7	44.56 ± 27.53
CIFAR-100	FixMatch	13.47 ± 1.42	9.20 ± 0.59	911.7 ± 206.0	18.84 ± 1.81
	Pseudo-Labeling	14.89 ± 1.40	17.16 ± 1.63	945.0 ± 89.6	17.31 ± 2.40

Table 1 reports the results for FixMatch at epoch 64. On CIFAR-10, strong augmentation increases the mean residual from 0.0061 to 0.2016 for correct pseudo-labels and from 0.0132 to 0.2849 for incorrect pseudo-labels, with similar increases on CIFAR-100. The mean strong-view residual exceeds the weak-view residual for both methods and both pseudo-label groups at all nine shared checkpoints from epoch 64 to 1024 (Appendix C.3). Since the residual is the target-logit gradient magnitude, these paired measurements show how strong views relax the small-residual constraint and increase learning pressure toward both correct and incorrect targets (Sections 3.2–3.3).

4.2 Intervening on subsequent supervision

To test how selected supervision affects later features and selections (Section 3.4), we copy each method's epoch-64 checkpoint into two branches, supplying one with FixMatch's full selected dataset and the other with Pseudo-Labeling's. Both datasets retain every selected sample and its original pseudo-label. FixMatch selects fewer samples overall but more correct targets: mean precision is 74.57% versus 17.34% on CIFAR-10 and 46.08% versus 13.99% on CIFAR-100. Appendix C.4 gives the complete counts.

All branches continue training for 50 epochs using fixed identities and labels. For pseudo-labeled samples, FixMatch recipients use strong views and Pseudo-Labeling recipients use weak views. Within each starting checkpoint, the two branches share model and optimizer states, learning-rate schedule, and sampling controls; only the supplied supervision differs. The loss combines labeled and pseudo-labeled cross-entropy. Before evaluation, all branches undergo the same clean-image batch-normalization (BN) recalibration procedure (Appendix C.4).

Test accuracy evaluates the existing classifier. For *linear-probe accuracy*, we freeze the feature extractor and fit a new linear classifier for 100 epochs on all 50,000 training images with true labels. Its accuracy on the 10,000 test images measures the features' class separability (Appendix C.2).

We also evaluate subsequent selection on a common pool excluding both supplied datasets and the labeled samples. Among images rejected by the starting recipient, let C_{new} count those selected correctly after continuation and Q_{new} be the precision of all newly selected images. Baseline and endpoint selections use the same BN evaluation procedure and threshold 0.95.

All four outcomes favor FixMatch-generated over Pseudo-Labeling-generated supervision for every recipient and seed (Table 2). Probe gains show a representation effect; new selections show that learning changes predictions on samples excluded from the continuation loss. This supports the supervision–learning–selection feedback. New wrong selections also increase in most pairs, and previous correct selections can be lost. Appendix C.5 reports these transitions and both BN evaluation policies.

4.3 Representation trajectories of FixMatch and Pseudo-Labeling

Figure 1 shows probe accuracy (Section 4.2) and offline selected-set precision: the fraction of high-confidence predictions matching true labels. Probes are shown at epochs 4, 8, 16, 32, 64, 128, 256, 512, 768, and 1024; precision is measured every eight epochs on fixed weak views of genuine unlabeled samples in evaluation mode.

FixMatch gains a probe advantage early and retains it at the measured checkpoints. Pseudo-Labeling's probe accuracy changes little, consistent with limited feature learning. Individual runs can nevertheless deteriorate. On CIFAR-100, all three FixMatch runs have small late probe changes, yet precision falls by about 10–19 percentage points from epoch 256 to 1024. On CIFAR-10, seed 1's probe accuracy falls from 88.31% at epoch 128 to 65.32% at epoch 256. Appendix C.2 gives individual runs and selection changes.

At epoch 1024, exponential moving average (EMA) test accuracy is $67.91 \pm 28.02\%$ for FixMatch versus $22.32 \pm 2.10\%$ for Pseudo-Labeling on CIFAR-10, and $39.35 \pm 1.32\%$ versus $3.45 \pm 4.00\%$ on CIFAR-100, using the trained classifiers rather than probes. The early probe gains and controlled comparisons support the proposed mechanism; the declines show that a relative advantage need not mean continuous improvement. Fixed-checkpoint diagnostics reveal substantial training/evaluation differences in selection, without identifying the cause of the decline (Appendix C.8).

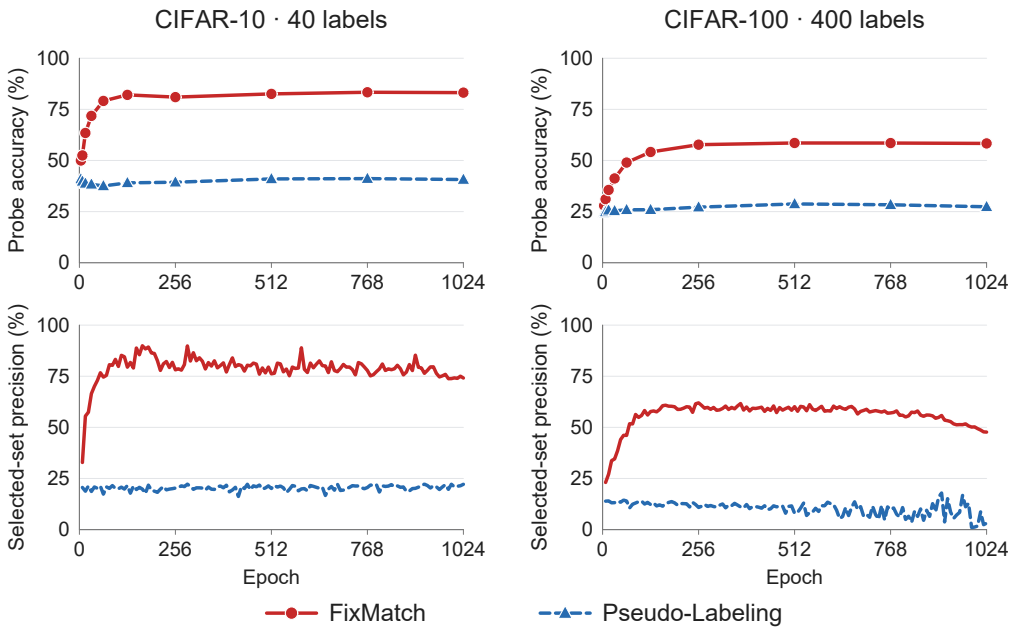

Figure 1: Three-seed mean probe accuracy (top) and offline selected-set precision (bottom) throughout training. FixMatch: red solid lines and circles; Pseudo-Labeling: blue dashed lines and triangles. Markers indicate evaluated probe checkpoints.

5 Conclusion

We explain FixMatch's advantage through feedback between feature learning and pseudo-label selection. In our neural model, incorrect selections reinforce spurious features and delay correct supervision. Weak-to-strong consistency regularization can instead support shared features and further correct selections, yielding higher precision and lower classification error under the stated conditions. It can also amplify updates toward wrong targets. CIFAR-10 and CIFAR-100 experiments show increased learning pressure and effects of selected supervision on both later representations and subsequent selection. Future work will extend this framework to other FixMatch training phenomena.

AI USE STATEMENT

We used ChatGPT and Codex to assist with developing the theoretical model, formulating mathematical claims, developing, drafting, and checking proofs, and refining hypotheses and experimental designs. These tools also assisted with implementing and debugging experimental and verification code, processing experimental outputs, interpreting results, searching the literature, and drafting, translating, and editing the manuscript. They also helped prepare tables and figures. The reported empirical results are based on executed experiments and saved outputs.

Author review included critical discussion of the experimental interpretations, refinement of the theoretical narrative, and detailed review and revision of the manuscript. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study analyzes semi-supervised learning using existing CIFAR-10 and CIFAR-100 benchmarks. It involves no recruitment of human participants or collection of new personal data. The analysis highlights how incorrect pseudo-labels can influence later learning. These benchmark results do not establish the reliability of pseudo-labeling in deployment settings where misclassification has significant consequences.

REPRODUCIBILITY STATEMENT

Appendices A and B give the assumptions, derivations, and proofs. Appendix B.10 specifies a parameter instance and its interval-arithmetic checks. Appendix C documents the implementation, training and evaluation protocols, random seeds, and additional results. It includes per-seed selection transitions and distinguishes native BN statistics from recalibrated and training-mode evaluations.

REFERENCES

Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks*, pp. 1–8. IEEE, 2020.

Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Baixu Chen, Junguang Jiang, Ximei Wang, Pengfei Wan, Jianmin Wang, and Mingsheng Long. Debiased self-training for semi-supervised learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

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. In *International Conference on Learning Representations*, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

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

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

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on Challenges in Representation Learning*, 2013.

Jingyang Li, Jiachun Pan, Vincent Y. F. Tan, Kim-Chuan Toh, and Pan Zhou. Towards understanding why FixMatch generalizes better than supervised learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2019.

Chaithanya Kumar Mummadi, Robin Hutmacher, Kilian Rambach, Evgeny Levinkov, Thomas Brox, and Jan Hendrik Metzen. Test-time adaptation to distribution shift by confidence maximization and input transformation. *arXiv preprint arXiv:2106.14999*, 2021.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Debiasing classifier from biased classifier. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20673–20684, 2020.

Yassine Ouali, Céline Hudelot, and Myriam Tami. An overview of deep semi-supervised learning. *arXiv preprint arXiv:2006.05278*, 2020.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C. Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1256–1272, 2021.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Ruoqi Shen, Sebastien Bubeck, and Suriya Gunasekar. Data augmentation as feature manipulation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 19773–19808. PMLR, 2022.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, volume 33, pp. 596–608, 2020.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jesper E. Van Engelen and Holger H. Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yidong Wang, Hao Chen, Yue Fan, Wang Sun, Ran Tao, Wenxin Hou, Renjie Wang, Linyi Yang, Zhi Zhou, Lan-Zhe Guo, Heli Qi, Zhen Wu, Yu-Feng Li, Satoshi Nakamura, Wei Ye, Marios Savvides, Bhiksha Raj, Takahiro Shinozaki, Bernt Schiele, Jindong Wang, Xing Xie, and Yue Zhang. USB: A unified semi-supervised learning benchmark for classification. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, Bernt Schiele, and Xing Xie. FreeMatch: Self-adaptive thresholding for semi-supervised learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. In *International Conference on Learning Representations*, 2021.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6256–6268, 2020.

Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. Dash: Semi-supervised learning with dynamic thresholding. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11525–11536, 2021.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference*, pp. 87.1–87.12. BMVA Press, 2016.

Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. FlexMatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18408–18419, 2021.

A Training-View Effects on Representation Updates

Let (X, Y) be a random input and its true class; lowercase x, y denote realized inputs and labels. We analyze a shared current state $\Theta = (\psi, W)$, where ψ contains the feature-map parameters and $W \in \mathbb{R}^{K\times d}$ is the classification head for $K \geq 2$ classes and a d-dimensional representation. The representation, logits, and probabilities are $Z(x) = h_\psi(x) \in \mathbb{R}^d$, $f_\Theta(x) = Wh_\psi(x) \in \mathbb{R}^K$, and $p_\Theta(x) = \mathrm{softmax}(f_\Theta(x))$. Here $m = F$ denotes FixMatch and $m = P$ denotes Pseudo-Labeling; $v = w$ and $v = s$ denote weak and strong views, with $v_F = s$ and $v_P = w$. Let X^w, X^s be paired views of X. For a confidence threshold $\tau \in (0, 1)$, define the pseudo-label $\widetilde{Y}$ and selection indicator χ by

$$\widetilde{Y} = \arg\max_k p_{\Theta,k}(X^w), \qquad \chi = \mathbf{1}\{p_{\Theta,\widetilde{Y}}(X^w) \geq \tau\}.$$

The class index k ranges over $\{1, \ldots, K\}$. If several classes share the maximum probability, choose the smallest class index. The weak training view is the exact view used for selection. Targets, selection, and view realizations are held fixed when differentiating

$$\mathcal{L}_m = \alpha\mathcal{L}_\ell + \lambda\mathbb{E}[\chi\ell(\widetilde{Y}, p_\Theta(X^{v_m}))], \qquad \ell(a, p) = -\log p_a.$$

Here $\mathcal{L}_\ell$ is the supervised loss averaged over labeled weak views, $\alpha \geq 0$ and $\lambda > 0$ are loss weights, and $\ell(a, p)$ is cross-entropy for class index a and probability vector p. Unindexed expectations and probabilities refer to the full unlabeled population and augmentation randomness. The true label Y is used only for analysis. We assume the displayed gradients are integrable and that differentiation under the expectation is valid, for example under an integrable local bound on their norms.

A.1 Residuals and parameter gradients

Write $f^v = f_\Theta(X^v)$ and $p^v = p_\Theta(X^v)$, let $e_k \in \mathbb{R}^K$ be the one-hot vector for class k, and define the target residual r_v and feature Jacobian J_ψ by

$$r_v = 1 - p^v_{\widetilde{Y}}, \qquad J_\psi(z) = D_\psi h_\psi(z).$$

The Jacobian maps a parameter increment to its first-order representation change at input z. Unsubscripted vector norms are Euclidean; $\|\cdot\|_1$ denotes the sum of absolute coordinates. Direct differentiation of $-f^v_{\widetilde{Y}} + \log \sum_k \exp(f^v_k)$ gives

$$\begin{aligned} \nabla_f \ell(\widetilde{Y}, p^v) &= p^v - e_{\widetilde{Y}}, \\ g^v_\psi := -\nabla_\psi \ell(\widetilde{Y}, p^v) &= J_\psi(X^v)^\top W^\top (e_{\widetilde{Y}} - p^v), \\ g^v_W := -\nabla_W \ell(\widetilde{Y}, p^v) &= (e_{\widetilde{Y}} - p^v) h_\psi(X^v)^\top. \end{aligned} \tag{15}$$

Because the non-target probabilities sum to r_v,

$$\|\nabla_f \ell(\widetilde{Y}, p^v)\|_1 = 2r_v, \qquad r_w \leq 1 - \tau \quad \text{on } \{\chi = 1\}. \tag{16}$$

Moreover,

$$r_s - r_w = p^w_{\widetilde{Y}} - p^s_{\widetilde{Y}}. \tag{17}$$

Thus, for a selected sample, a strong-view target-probability bound $p^s_{\widetilde{Y}} \leq \epsilon_s < \tau$ implies $r_s - r_w \geq \tau - \epsilon_s > 0$, irrespective of target correctness. Equivalently, $r_v = (1+\exp(M_v))^{-1}$ for the target logit margin $M_v = f^v_{\widetilde{Y}} - \log \sum_{k \neq \widetilde{Y}} \exp(f^v_k)$. A reduction of the target logit alone does not determine the residual when other logits also change.

The weak residual cap concerns the current selector and the same weak view. Replaying a frozen selected set at later states does not preserve this cap without a separate confidence bound. For finite logits, define the residual-normalized negative logit gradient

$$d_v = \frac{e_{\widetilde{Y}} - p^v}{r_v}.$$

Then $\|d_v\|_1 = 2$ and

$$\|g^v_\psi\| = r_v \|J_\psi(X^v)^\top W^\top d_v\|. \tag{18}$$

Residual size alone therefore does not order parameter-gradient norms. In the multiclass case, d_v can also change between views.

A.2 Transfer to another input

Let $V_{\psi,m}$ and $V_{W,m}$ denote the negative gradients of $\mathcal{L}_m$ at the shared state. In particular,

$$V_{\psi,m} = -\alpha \nabla_\psi \mathcal{L}_\ell + \lambda \mathbb{E}[\chi g^{v_m}_\psi]. \tag{19}$$

For step size $\eta > 0$, after a gradient step $(\psi^+_m, W^+_m) = (\psi, W) + \eta(V_{\psi,m}, V_{W,m})$, the representation at any fixed input z satisfies

$$\begin{aligned} h_{\psi^+_m}(z) - h_\psi(z) = &- \eta\alpha J_\psi(z) \nabla_\psi \mathcal{L}_\ell \\ &+ \eta\lambda \mathbb{E}\big[\chi J_\psi(z) J_\psi(X^{v_m})^\top W^\top (e_{\widetilde{Y}} - p^{v_m})\big] + \mathcal{E}_m(z). \end{aligned} \tag{20}$$

Here $\mathcal{E}_m(z)$ is the representation's Taylor remainder. If the map $\psi \mapsto J_\psi(z)$ is Lipschitz with constant $L_z \geq 0$ in operator norm along the parameter segment, Taylor's theorem gives

$$\|\mathcal{E}_m(z)\| \leq \tfrac{1}{2} L_z \eta^2 \|V_{\psi,m}\|^2. \tag{21}$$

Differentiability at the initial state alone gives an $o(\eta)$ remainder. Subtracting the two versions of (20) cancels the supervised first-order term; the remaining term is the difference of the two displayed unlabeled expectations, plus $\mathcal{E}_F(z) - \mathcal{E}_P(z)$.

For a ReLU representation, the expansion is local to a differentiable activation region. A segment staying away from the relevant activation boundaries supplies the usual smooth Taylor bound. At a boundary one must instead use the appropriate directional derivative or an explicit piecewise calculation. In particular, a dormant evaluation ReLU has zero local output derivative although updates from other inputs can move its preactivation toward zero. A subsequent activation is not justified by extrapolating its initial Taylor expansion across that boundary.

For any fixed direction u in the parameter space of ψ, define

$$\mathcal{T}_v(u) = \left\langle u, J_\psi(X^v)^\top W^\top d_v \right\rangle. \tag{22}$$

Taking $u = J_\psi(z)^\top \xi$ for a representation direction $\xi \in \mathbb{R}^d$, the unlabeled first-order change of $\xi^\top h_\psi(z)$ is $\eta\lambda\mathbb{E}[\chi r_v \mathcal{T}_v(u)]$. Here $\mathcal{T}_v(u)$ is a signed inner product of parameter derivatives at two inputs. It can be negative or zero even for a correct target. A larger residual does not determine its sign.

A.3 Correct and incorrect contributions

Let

$$\begin{gathered} \mathcal{C} = \{\chi = 1, \widetilde{Y} = Y\}, \qquad \mathcal{E} = \{\chi = 1, \widetilde{Y} \neq Y\}, \\ c = \Pr(\mathcal{C}), \qquad e = \Pr(\mathcal{E}), \qquad \mu = c + e, \qquad Q = c/\mu \quad (\mu > 0). \end{gathered} \tag{23}$$

Here c and e are the correctly and incorrectly selected masses, μ is their total mass, and Q is selected-set precision. Since $\mathcal{C}$ and $\mathcal{E}$ partition the selected population,

$$\mathbb{E}[\chi g_\psi^v] = c\,\mathbb{E}[g_\psi^v \mid \mathcal{C}] + e\,\mathbb{E}[g_\psi^v \mid \mathcal{E}]. \tag{24}$$

This proves the selected-label decomposition. A conditional mean on a zero-probability event is set to zero; only its mass-weighted contribution is used.

Fix a differentiable, larger-is-better representation utility $\Phi(\psi)$ specified independently of the pseudo-label update, and evaluate $u = \nabla_\psi \Phi(\psi)$ at the shared state. For example, Φ may be negative true-label population loss under a fixed probe, provided the required derivatives exist. Hold u fixed in the following comparison. This definition does not assume that either update improves Φ. Set

$$\begin{aligned} \Delta_+ &= \mathbb{E}[\langle u, g_\psi^s - g_\psi^w \rangle \mid \mathcal{C}], \\ \Delta_- &= \mathbb{E}[\langle u, g_\psi^w - g_\psi^s \rangle \mid \mathcal{E}]. \end{aligned} \tag{25}$$

The signed quantity Δ_- is the additional cost from incorrect targets in this direction; it need not be nonnegative.

A.4 A sufficient projected comparison

Combining (19) and (24) at the common state gives

$$\langle u, V_{\psi,F} - V_{\psi,P} \rangle = \lambda(c\Delta_+ - e\Delta_-). \tag{26}$$

Let $A_+ > 0$ bound the useful contribution from below and $A_- \geq 0$ bound the additional error cost from above. If $\mu > 0$, $\Delta_+ \geq A_+$ whenever $c > 0$, and $\Delta_- \leq A_-$ whenever $e > 0$, then

$$\begin{gathered} \langle u, V_{\psi,F} - V_{\psi,P} \rangle \geq \lambda\mu[QA_+ - (1-Q)A_-], \\ Q > \frac{A_-}{A_+ + A_-} \quad \Longrightarrow \quad \langle u, V_{\psi,F} - V_{\psi,P} \rangle > 0. \end{gathered} \tag{27}$$

The first inequality follows by substituting the two contribution bounds; rearranging its strict positivity gives the second. It compares directional derivatives of the specified utility. Positive relative progress does not imply that the F update itself has positive utility derivative. For finite steps, a strict first-order gap gives a utility advantage for sufficiently small η, with an $o(\eta)$ remainder under differentiability of Φ. When $\mu = 0$, both unlabeled contributions vanish and Q is undefined.

These statements compare a common current state and selector. Appendix A.5 accounts for the contributions after the two trajectories separate.

Residual gain with controlled transfer. The contribution bounds above can be checked through residuals and transfer factors, but require both. Write $\mathcal{T}_v = \mathcal{T}_v(u)$ for the factor in (22). The exact identity is

$$r_s\mathcal{T}_s - r_w\mathcal{T}_w = (r_s - r_w)\mathcal{T}_w + r_s(\mathcal{T}_s - \mathcal{T}_w). \tag{28}$$

For example, suppose $c > 0$, $r_s \geq r_w$ and $\mathcal{T}_w \geq 0$ on $\mathcal{C}$. On an event $\mathcal{A} \subseteq \mathcal{C}$ of conditional probability $\pi_{\mathcal{A}} = \Pr(\mathcal{A} \mid \mathcal{C})$, suppose $r_s - r_w \geq \delta_r > 0$ and $\mathcal{T}_w \geq \gamma_u > 0$. Here δ_r and γ_u are lower bounds on the residual gain and transfer factor. If $\epsilon_T \geq 0$ bounds the mean transfer change as $\mathbb{E}[r_s|\mathcal{T}_s - \mathcal{T}_w| \mid \mathcal{C}] \leq \epsilon_T$, then

$$\Delta_+ \geq \pi_{\mathcal{A}}\delta_r\gamma_u - \epsilon_T. \tag{29}$$

Indeed, the first term in (28) is nonnegative on $\mathcal{C}$ and at least $\delta_r\gamma_u$ on $\mathcal{A}$; the second has conditional mean at least $-\epsilon_T$. The confidence rule alone supplies neither the transfer sign nor its stability between views.

A.5 Contributions along the training trajectories

Let $\Theta_m(t) = (\psi_m(t), W_m(t))$ be the training trajectory of method $m \in \{F, P\}$. Evaluate its pseudo-labels, selected events $\mathcal{C}_m(t), \mathcal{E}_m(t)$, and their masses $c_m(t), e_m(t)$ at its own current state. Write $g_{\psi,m}(t)$ for the negative feature-parameter gradient of one pseudo-labeled sample at this state, using view v_m. For a fixed continuously differentiable representation utility Φ, define $u_m(t) = \nabla_\psi\Phi(\psi_m(t))$ and

$$\begin{aligned} L_m(t) &= -\alpha\langle u_m(t), \nabla_\psi\mathcal{L}_\ell(\Theta_m(t))\rangle, \\ U_m(t) &= \mathbb{E}[\langle u_m(t), g_{\psi,m}(t)\rangle \mid \mathcal{C}_m(t)], \\ H_m(t) &= -\mathbb{E}[\langle u_m(t), g_{\psi,m}(t)\rangle \mid \mathcal{E}_m(t)]. \end{aligned} \tag{30}$$

Here L_m is the supervised contribution to utility change, U_m is the mean contribution from correctly selected samples, and H_m is the negative mean contribution from incorrectly selected samples. Conditional means with zero mass are set to zero. These quantities are signed: incorrect supervision need not reduce every representation utility, nor does correct supervision necessarily improve it.

Along an absolutely continuous trajectory with integrable displayed terms, the chain rule and the selected-label decomposition give, almost everywhere,

$$\frac{d}{dt}\Phi(\psi_m(t)) = L_m(t) + \lambda[c_m(t)U_m(t) - e_m(t)H_m(t)]. \tag{31}$$

For the same initialization, integration yields

$$\begin{aligned} \Phi(\psi_F(t)) - \Phi(\psi_P(t)) = &\int_0^t [L_F(s) - L_P(s)]\,ds \\ &+ \lambda\int_0^t [c_F(s)U_F(s) - c_P(s)U_P(s)]\,ds \\ &- \lambda\int_0^t [e_F(s)H_F(s) - e_P(s)H_P(s)]\,ds. \end{aligned} \tag{32}$$

The supervised terms no longer cancel once the parameter states differ. The identity follows the actual stop-gradient training flow; it does not differentiate through the selection rule. At switching times, the parameters remain continuous and the identity holds through integration. It records accumulated contributions along each trajectory. Comparing training with and without an error cohort instead requires comparing the resulting trajectories, as done in Appendix B.

A.6 Target errors and binary logistic specialization

For multiclass cross-entropy, the identity

$$e_{\widetilde{Y}} - p^v = (e_Y - p^v) + (e_{\widetilde{Y}} - e_Y)$$

gives the exact separation

$$\begin{aligned} \mathbb{E}[\chi g_\psi^v] = &\mathbb{E}\big[\chi J_\psi(X^v)^\top W^\top(e_Y - p^v)\big] \\ &+ \mathbb{E}\big[\mathbf{1}_{\mathcal{E}} J_\psi(X^v)^\top W^\top(e_{\widetilde{Y}} - e_Y)\big]. \end{aligned} \tag{33}$$

The first term uses true targets on the same selected views; the second is the change caused by target errors. Their projections on a separate evaluation utility need not have prescribed signs.

For the single-logit model, take $W \in \mathbb{R}^{1\times d}$ and $f_\Theta(x) = Wh_\psi(x) \in \mathbb{R}$. With binary pseudo-label $\widetilde{y} \in \{-1, +1\}$, define the logistic loss and residual function

$$\ell_{\mathrm{bin}}(\widetilde{y}, f) = \log(1 + \exp(-\widetilde{y}f)), \qquad r(z) = \frac{1}{1+\exp(z)}.$$

Direct differentiation yields

$$\begin{aligned} g_\psi^v &= \widetilde{y}\, r(\widetilde{y} f_\Theta(X^v)) J_\psi(X^v)^\top W^\top, \\ g_W^v &= \widetilde{y}\, r(\widetilde{y} f_\Theta(X^v)) h_\psi(X^v)^\top. \end{aligned} \tag{34}$$

These formulas, averaged over sample types and augmentation masks, give the binary parameter dynamics used in the feedback model. They use one scalar logit; introducing two independently parameterized softmax rows would require its own parameter-gradient calculation.

For true binary label $y \in \{-1, +1\}$ and wrong target $\widetilde{y} = -y$, the direction on the same input is opposite to the true-margin derivative:

$$\left\langle \nabla_\psi [y f_\Theta(X^v)], g_\psi^v \right\rangle = -r(-y f_\Theta(X^v)) \| J_\psi(X^v)^\top W^\top \|^2 \le 0. \tag{35}$$

Its magnitude depends on both the residual and the parameter derivative. This same-input sign does not establish a sign for transfer to another input or for every parameter coordinate.

A.7 A RESIDUAL INCREASE CAN HARM A HELDOUT REPRESENTATION

Consider a two-class softmax network

$$h_\psi(x) = [\psi^\top x]_+, \qquad \psi = (0, 0, 1)^\top, \qquad W = (1, 0)^\top.$$

Here $[a]_+ = \max(a, 0)$ is the ReLU activation and $\sigma(a) = (1 + e^{-a})^{-1}$ is the sigmoid function. Let $x^w = (1, 0, 3)^\top$, $x^s = (1, -5, 1)^\top$, and $z = (1, 1, 1)^\top$. A ground-truth rule assigning class 1 when $x_1 > 0$ makes both views and z correctly labeled as class 1. The weak target is accepted at $\tau = 0.95$, since $p_1^w = \sigma(3) > 0.95$. Nevertheless, $r_s = \sigma(-1) > \sigma(-3) = r_w$.

For one unlabeled gradient step with weight one and no supervised term, (15) gives

$$\psi_m^+ = \psi + \eta r_{v_m} x^{v_m}, \qquad W_m^+ = (1 + \eta r_{v_m} h_\psi(x^{v_m}), -\eta r_{v_m} h_\psi(x^{v_m}))^\top.$$

For $0 < \eta r_s < 1/3$, the heldout unit remains active and

$$h_{\psi_F^+}(z) = 1 - 3\eta r_s, \qquad h_{\psi_P^+}(z) = 1 + 4\eta r_w. \tag{36}$$

The corresponding true-class logit margins are

$$(1 + 2\eta r_s)(1 - 3\eta r_s) < 1, \qquad (1 + 6\eta r_w)(1 + 4\eta r_w) > 1,$$

respectively. The F step increases true-label loss at z, whereas the P step decreases it. Thus a larger residual with a correct, label-preserving training view does not by itself imply a more useful representation update.

B REPRESENTATION–SELECTION FEEDBACK IN A TRAINABLE NETWORK

This appendix proves Theorem 2 by following each learner's parameters and confidence gates. It links the initial feature dependence to new incorrect selections, shows how their supervision delays later correct selections, and then compares the learned representations. Appendix B.10 verifies the conditions for an explicit parameter instance.

B.1 POPULATION, NETWORK, AND COMMON INITIALIZATION

Let (X, Y) denote a random input and its true label, with $Y \in \{-1, +1\}$ balanced within each population group. An input x has two-dimensional patches $x_j^+, x_j^- \in \mathbb{R}^2$ for $j \in \{0, 1, 2, 3, s\}$. Indices $0, 1, 2, 3$ denote semantic feature pathways: samples with a patch at the same index use the same pair of class-channel filters and output coefficients. They are not training stages or separate models. Pathway 0 is the root because it is present in the initially selected samples and provides the starting point for later learning. Index s denotes the pathway for a spurious feature, whose association with the true label differs across population groups. The superscripts name observable input channels. Their association with the label is a property of the data distribution; the training algorithm does not read Y on unlabeled examples. Define the seed patch u_0, evaluation patch q_0, and set $\mathcal{G}$ of cooccurring semantic pathway pairs by

$$u_0 = (1, 1)^\top, \qquad q_0 = (0, 1)^\top, \qquad \mathcal{G} = \{(0, 1), (0, 2), (1, 3)\}. \tag{37}$$

For a sample of class y, its true-class channel means $+$ when $y = +1$ and $-$ when $y = -1$. Unmentioned patches are zero. The population is:

Group	Nonzero patches	Mass
Correct seed	u_0 at root 0 and spurious pathway s, both in the true-class channel	π_0
Cooccurrence edge $(i, j) \in \mathcal{G}$	u_0 at i and j, both in the true-class channel	π_e each
Target singleton $j \in \{0, 1, 2, 3\}$	q_0 at j in the true-class channel	$\pi_T/4$ each
Initial error group	u_0 at root 0 in the true-class channel and at s in the opposite channel	ν
Candidate error group	u_0 at root 0 in the true-class channel and ρu_0 at s in the opposite channel	ω

All masses are positive and

$$\pi_0 + 3\pi_e + \pi_T + \nu + \omega = 1. \tag{38}$$

The positive scalar ρ controls the candidate group's spurious input. An edge (i, j) records feature cooccurrence in a sample, not a connection between neural layers. The labeled set consists of the two correct seed samples, one per class. The target singletons belong to the same unlabeled and evaluation population as the other groups.

With ReLU activation $[z]_+ = \max(z, 0)$, the network has scalar binary logit

$$f_\Theta(x) = \sum_{j \in \{0,1,2,3,s\}} \left\{ a_j^+ [\langle w_j^+, x_j^+ \rangle]_+ - a_j^- [\langle w_j^-, x_j^- \rangle]_+ \right\}. \tag{39}$$

All $a_j^\pm \in \mathbb{R}$ and $w_j^\pm \in \mathbb{R}^2$ are independent trainable parameters in Θ. At continuous training time t, the hidden representation $Z_t^m(x)$ for method $m \in \{F, P\}$, with F denoting FixMatch and P Pseudo-Labeling, is the vector of the ten ReLU activations in (39); it does not include the output coefficients $a_j^\pm$. Thus the representation and the trained readout can be evaluated separately.

We use the sigmoid $\sigma(z) = (1 + e^{-z})^{-1}$, binary logistic loss $\ell(y, f) = \log(1 + e^{-yf})$, and residual at margin z

$$r(z) = \sigma(-z) = \frac{1}{1 + e^z}. \tag{40}$$

A prediction is $\mathrm{sign}_+(f)$, where $\mathrm{sign}_+(z) = +1$ for $z \geq 0$ and -1 otherwise; the same rule assigns pseudo-labels. With probability threshold $1/2 < \tau < 1$, the corresponding logit-magnitude threshold is $\kappa = \log(\tau/(1 - \tau))$, so acceptance means $|f_\Theta(x)| \geq \kappa$. Each method computes its own current weak-view pseudo-label $\widetilde{y}_m(t, x)$ and indicator $\chi_m(t, x)$. They are held fixed in each derivative and recomputed from the updated model.

Weak augmentation is the identity. A strong view deletes a nonzero spurious patch with probability $p_d \in (0, 1]$ and independently retains each nonzero semantic patch with probability $p_r \in (0, 1]$. The mask law is the same in both label classes and is fixed throughout training. FixMatch trains on the

strong view and Pseudo-Labeling on the weak view. For either method, all parameters follow the negative Euclidean gradient of

$$\alpha\mathcal{L}_\ell(\Theta) + \lambda\mathbb{E}_u\big[\chi_m(t,X)\ell\big(\widetilde{y}_m(t,X), f_\Theta(X^{v_m})\big)\big], \qquad v_F = s, \quad v_P = w, \tag{41}$$

where $\alpha \geq 0$ and $\lambda > 0$ weight the labeled and unlabeled losses, respectively; $\mathcal{L}_\ell$ is the mean loss of the two labeled samples, and $\mathbb{E}_u$ averages over the full unlabeled population and the view randomness. Here X^{v_m} is the training view, with w denoting weak and s strong augmentation. In particular, the unlabeled term is not divided by the currently selected mass.

The paired channels have equal initial parameters. Because the population, mask law, and loss are invariant under exchanging the two channels and reversing the label, this equality is preserved along the flow. It is a symmetry of independently trained parameters, not a parameter-tying rule. For their common values, suppress the channel superscripts and write

$$s_j = \langle w_j, u_0\rangle, \qquad d_j = w_{j,1} - w_{j,2}, \qquad \beta_j = a_j s_j, \qquad S = \beta_s. \tag{42}$$

Here s_j is the filter's response to u_0, d_j is the difference between the two coordinates of w_j, β_j is the resulting logit contribution, and S is the spurious-feature contribution. The coordinate β_j is not the hidden representation.

Let $M > 0$ be the initial spurious-feature contribution. Choose the target activation level h, initial nonroot semantic and root strengths ε and δ, and the level b at which a root paired with an unchanged neighbor passes the gate:

$$h = \kappa/2, \qquad 0 < \varepsilon < \delta < h, \qquad b = \kappa - \varepsilon, \qquad \beta_0(0) = \delta, \quad \beta_{1,2,3}(0) = \varepsilon, \quad S(0) = M. \tag{43}$$

Every pathway, including s, initially satisfies $s_j = \sqrt{2}a_j > 0$. Set

$$d_j = \sqrt{\kappa/\sqrt{2}} \quad (j = 0,1,2,3), \qquad d_s = 0. \tag{44}$$

These requirements specify actual parameters: for one pathway with initial strength $\beta > 0$ and prescribed filter-coordinate difference d, take $a = \sqrt{\beta/\sqrt{2}}$, $s = \sqrt{2}a$, and $w = ((s+d)/2, (s-d)/2)^\top$. Both methods start from the same parameters. These parameters define the common warm start.

Choose $0 < \zeta < \kappa$ and set

$$\rho = \frac{\kappa + \delta - \zeta}{M}, \qquad q_m(t) = \rho S^m(t) - \beta_0^m(t). \tag{45}$$

Here q_m is the candidate group's wrong-label margin. Initially $q_m(0) = \kappa - \zeta$, so this group is predicted incorrectly but rejected. The actual confidence rule accepts it with an incorrect label if $q_m \geq \kappa$, accepts it with a correct label if $q_m \leq -\kappa$, and otherwise rejects it. Both boundaries are part of the rule; the bounds below establish which can be reached.

The conditions below imply $M - \delta > \kappa$. Initially the seed is confidently correct and the initial error group confidently incorrect. All edges are correctly classified but rejected because their margins are at most $\delta + \varepsilon < \kappa$. Every target singleton has zero hidden representation and is also rejected. Thus both methods initially select exactly the same two group identities and pseudo-labels.

B.2 Exact parameter dynamics

We derive the equations while the seed and initial error group retain their initial acceptance and label states and the target singletons remain rejected. Appendix B.4 proves these properties from explicit parameter conditions, thereby closing the use of these equations.

Dots below denote derivatives with respect to training time t; a superscript m identifies the method and is omitted when an identity holds for either one. Every nonzero semantic training patch is then u_0. For one class channel, the probability of that class supplies a factor $1/2$. If g_j denotes its signed training coefficient after this class probability has been factored out, direct differentiation gives

$$\dot{a}_j = \frac{s_j}{2} g_j, \qquad \dot{w}_j = \frac{a_j}{2} u_0 g_j, \qquad \dot{s}_j = a_j g_j, \qquad \dot{d}_j = 0. \tag{46}$$

The same equations hold for the spurious pathway, with the scale ρ included in g_s for the candidate group. A negative g_j, such as the root contribution of an incorrect target, reverses both parameter updates. Since $d(s_j - \sqrt{2}a_j)/dt = -(g_j/\sqrt{2})(s_j - \sqrt{2}a_j)$, the initial balance is preserved. Hence

$$s_j = \sqrt{2}a_j, \qquad \dot{a}_j = \frac{g_j}{\sqrt{2}}a_j, \qquad \dot{\beta}_j = \sqrt{2}\beta_j g_j. \tag{47}$$

In particular, heads and u_0 preactivations remain positive on every finite interval where the coefficients are bounded. Both filter coordinates and the head are trained; the constant d_j follows from the direction of the filter gradient.

For an edge $(i, j) \in \mathcal{G}$, let $\chi_{ij}^m(t)$ be its selection indicator:

$$\chi_{ij}^m(t) = \mathbf{1}\{\beta_i^m(t) + \beta_j^m(t) \geq \kappa\}. \tag{48}$$

For endpoint strength $u = \beta_j$ and other-endpoint strength $v = \beta_i$, an accepted edge contributes to endpoint j the coefficient

$$\begin{aligned} \mathcal{D}_F(u, v) &= \lambda\pi_e p_r\{(1 - p_r)r(u) + p_r r(u + v)\}, \\ \mathcal{D}_P(u, v) &= \lambda\pi_e r(u + v). \end{aligned} \tag{49}$$

The outer p_r is the probability that endpoint j survives; the two terms condition on deletion or retention of the other endpoint. If j is deleted its derivative is zero.

With $u = \beta_0$ now denoting the root strength, write $\mathcal{G}_{0,m}(u, S)$ for the combined seed and initial-error contribution to the root coefficient:

$$\begin{aligned} \mathcal{G}_{0,F}(u, S) &= \alpha r(S + u) + \lambda\pi_0 p_r\{p_d r(u) + (1 - p_d)r(S + u)\} \\ &\quad - \lambda\nu p_r\{p_d r(-u) + (1 - p_d)r(S - u)\}, \\ \mathcal{G}_{0,P}(u, S) &= (\alpha + \lambda\pi_0)r(S + u) - \lambda\nu r(S - u). \end{aligned} \tag{50}$$

The negative terms are the actual opposing semantic gradients from accepted incorrect pseudo-labels. In a strong view with the root retained, the wrong group's pseudo-label margin is $-u$ when the spurious patch is deleted and $S - u$ when it is retained. This gives precisely the last bracket in $\mathcal{G}_{0,F}$.

At a common semantic and spurious state (u, S), denote the magnitudes of these opposing coefficients by

$$\begin{aligned} \mathcal{H}_F(u, S) &= \lambda\nu p_r\{p_d r(-u) + (1 - p_d)r(S - u)\}, \\ \mathcal{H}_P(u, S) &= \lambda\nu r(S - u). \end{aligned} \tag{51}$$

Their difference is positive exactly when

$$p_r p_d\{r(-u) - r(S - u)\} > (1 - p_r)r(S - u). \tag{52}$$

For proposed upper bounds $U > b$ and $V > M$, consider the root box $u \in [\delta, U]$, $S \in [M, V]$. A sufficient condition is the same inequality with $r(-u)$ replaced by $r(-\delta)$ and $r(S-u)$ by $r(M-U)$. In the instance of Appendix B.10, the resulting left side minus right side exceeds 0.369. Thus strong-view training can increase the adverse semantic coefficient; its effect on spurious-feature reinforcement is different, as shown below.

For the candidate group, write $\chi_\omega^m = \mathbf{1}\{q_m \geq \kappa\}$ on the domain $q_m > 0$. Its additional root coefficients are

$$\begin{aligned} \Delta g_{0,F}(u, S) &= -\lambda\omega\chi_\omega^F p_r\{p_d r(-u) + (1 - p_d)r(\rho S - u)\}, \\ \Delta g_{0,P}(u, S) &= -\lambda\omega\chi_\omega^P r(\rho S - u). \end{aligned} \tag{53}$$

The negative signs follow from fitting an incorrect target while the root patch lies in the true-class channel. Writing $i \sim j$ for the neighbors of j in $\mathcal{G}$, the complete semantic coefficients are

$$\begin{aligned} g_j^m = \mathbf{1}\{j = 0\}\{\mathcal{G}_{0,m}(\beta_0^m, S^m) + \Delta g_{0,m}(\beta_0^m, S^m)\} \\ + \sum_{i \sim j} \chi_{ij}^m \mathcal{D}_m(\beta_j^m, \beta_i^m). \end{aligned} \tag{54}$$

Each expression is evaluated at that method's current parameters and gate, with no synchronization after initialization.

Define $C = \alpha + \lambda\pi_0$ and $N = \lambda\nu$. The spurious-feature coefficients are

$$\begin{aligned} g_s^F &= \alpha r(S+\beta_0) \\ &\quad + \lambda(1-p_d)\Big[\pi_0\{(1-p_r)r(S) + p_r r(S+\beta_0)\} \\ &\qquad\qquad + \nu\{(1-p_r)r(S) + p_r r(S-\beta_0)\}\Big] \\ &\quad + \lambda\omega\rho\chi_\omega^F(1-p_d)\{(1-p_r)r(\rho S) + p_r r(\rho S - \beta_0)\}, \\ g_s^P &= Cr(S+\beta_0) + Nr(S-\beta_0) + \lambda\omega\rho\chi_\omega^P r(\rho S - \beta_0), \\ \dot{S} &= \sqrt{2} S g_s^m. \end{aligned} \tag{55}$$

Every spurious-feature term is nonnegative. Both error groups reinforce this feature when its patch is retained and oppose the semantic root. The factor ρ in the candidate term is the derivative of its scaled input. Deleting the spurious patch removes its own derivative but can increase the opposing semantic derivative.

For a target singleton at semantic pathway $j \in \{0,1,2,3\}$, the true-class preactivation is $(s_j - d_j)/2$. From $s_j^2 = \sqrt{2}\beta_j$ and $d_j^2 = \kappa/\sqrt{2}$, its true margin equals

$$\gamma(\beta_j) = a_j \left[\frac{s_j - d_j}{2}\right]_+ = \frac{1}{2}\left[\beta_j - \sqrt{h\beta_j}\right]_+. \tag{56}$$

For $\beta_j \le h$, both labels of this target group have hidden vector zero. For $\beta_j > h$, they have positive activations in distinct class channels. Thus crossing h changes the separability of Z_t^m. The function γ is nondecreasing, and strictly increasing above h. It is separate from the acceptance threshold κ.

B.3 Explicit sufficient conditions

Choose a horizon $H > 0$ and proposed bounds $U > b$ and $V > M$. The constants below depend only on these bounds and the data, mask, loss, and initialization parameters. Let E_m bound one admitted edge's coefficient and A_m bound the total root coefficient from above:

$$\begin{aligned} E_F &= \lambda\pi_e p_r\{(1-p_r)r(\varepsilon) + p_r r(2\varepsilon)\}, \\ E_P &= \lambda\pi_e r(2\varepsilon), \\ A_F &= \alpha r(M+\varepsilon) + \lambda\pi_0 p_r\{p_d r(\varepsilon) + (1-p_d)r(M+\varepsilon)\} + 2E_F, \\ A_P &= (\alpha + \lambda\pi_0)r(M+\varepsilon) + 2E_P. \end{aligned} \tag{57}$$

Let B_m bound the spurious-feature coefficient from above, with FixMatch's candidate gate closed:

$$\begin{aligned} B_F &= \alpha r(M+\varepsilon) + \lambda(1-p_d)\Big[\pi_0\{(1-p_r)r(M) + p_r r(M+\varepsilon)\} \\ &\qquad\qquad + \nu\{(1-p_r)r(M) + p_r r(M-U)\}\Big], \\ B_P &= Cr(M+\varepsilon) + Nr(M-U) + \lambda\omega\rho r(\kappa). \end{aligned} \tag{58}$$

For Pseudo-Labeling, a selected candidate satisfies $r(q_P) \le r(\kappa)$. Its adverse root term can therefore be bounded without assuming when its gate is on. Define the root lower coefficients

$$\begin{aligned} L_F &= \lambda p_r[p_d\pi_0 - \nu e^U] r(U), \\ L_P &= Cr(V+U) - Nr(M-U) - \lambda\omega r(\kappa). \end{aligned} \tag{59}$$

The required range and sign conditions are

$$\delta + \sqrt{2}UA_m H < U, \qquad \varepsilon + 2\sqrt{2}UE_m H < U \qquad (m = F, P), \tag{60}$$

$$M + \sqrt{2}VB_m H < V \qquad (m = F, P), \tag{61}$$

$$L_F > 0, \qquad L_P > 0, \tag{62}$$

$$M - U > \kappa, \qquad \gamma(U) < \kappa. \tag{63}$$

To keep the candidate rejected by FixMatch, require

$$\rho M - U > 0, \qquad \rho V B_F < \delta L_F. \tag{64}$$

These inequalities will close the state bounds and gate conditions together in Appendix B.4.

To state the time condition, define six positive derivative bounds. D_F bounds FixMatch's root growth from below, and D_E is a lower bound on growth supplied by one admitted edge. For Pseudo-Labeling, D_P^- and D_P^+ are lower and upper root-growth bounds before its first edge opens; D_L^- and D_L^+ are corresponding bounds for a root neighbor between strengths ε and b after that edge opens:

$$\begin{aligned} D_F &= \sqrt{2}\delta L_F, \\ D_E &= \sqrt{2}\varepsilon\lambda\pi_e p_r r(2U), \\ D_P^- &= \sqrt{2}\delta\{Cr(V+b) - Nr(M-b) - \lambda\omega r(\kappa)\}, \\ D_P^+ &= \sqrt{2}bCr(M+\delta), \\ D_L^- &= \sqrt{2}\varepsilon\lambda\pi_e r(U+b), \qquad D_L^+ = \sqrt{2}b\lambda\pi_e r(\kappa). \end{aligned} \tag{65}$$

In particular, $D_P^- > 0$ follows from (62) and $b < U$. Define an upper time bound $\overline{T}_F$ for all four FixMatch target representations to activate, an upper bound $\overline{T}_P$ for the first three Pseudo-Labeling targets, and a lower bound $\underline{T}_{P,\mathrm{far}}$ for Pseudo-Labeling to select edge $(1,3)$:

$$\begin{aligned} \overline{T}_F &= \frac{b-\delta}{D_F} + \frac{2(b-\varepsilon)}{D_E}, \\ \overline{T}_P &= \frac{b-\delta}{D_P^-} + \frac{h-\varepsilon}{D_L^-}, \\ \underline{T}_{P,\mathrm{far}} &= \frac{b-\delta}{D_P^+} + \frac{b-\varepsilon}{D_L^+}. \end{aligned} \tag{66}$$

The representation comparison requires

$$\max\{\overline{T}_F, \overline{T}_P\} < \min\{H, \underline{T}_{P,\mathrm{far}}\}. \tag{67}$$

All quantities on the two sides are explicit functions of the parameters. The symbols denote bounds, not exact event times.

For the candidate's admission by Pseudo-Labeling, define tighter bounds U_P and V_P, a lower time bound T_0^- for its first correct edge, and constants L_q, E_q for the candidate's margin derivative:

$$\begin{aligned} U_P &= \delta + \sqrt{2}UA_PH, \qquad V_P = M + \sqrt{2}VB_PH, \\ T_0^- &= (b-\delta)/D_P^+, \\ L_q &= \sqrt{2}\{(\rho M - U_P)Cr(V_P + U_P) \\ &\qquad + (\rho M + \delta)Nr(V_P - \delta)\}, \\ E_q &= 2\sqrt{2}U_P\lambda\pi_e r(\kappa), \qquad T_a = \zeta/L_q. \end{aligned} \tag{68}$$

We require

$$\begin{aligned} 0 < L_q < E_q, \qquad T_a < T_0^- < H, \\ -\zeta + L_qH - E_q(H - T_0^-) > 0. \end{aligned} \tag{69}$$

These conditions ensure that the new error group enters before the first correct edge and stays selected. Finally, the causal comparison in Appendix B.6 uses

$$C \geq Ne^{2b}. \tag{70}$$

Proposition 3 (A parameter certificate for the common-time comparison)**.** *Under the population, network, and initialization of Appendix B.1, the stated loss and mask rules, and the state, gate, time, and comparison conditions in this subsection, the conclusions of Theorem 2 hold on the nonempty interval*

$$I = \left(\max\{\overline{T}_F, \overline{T}_P\}, \min\{H, \underline{T}_{P,\mathrm{far}}\}\right). \tag{71}$$

Both methods have selected new cooccurrence groups and learned target representations through nonroot semantic pathways there. FixMatch has learned all four target groups; Pseudo-Labeling has learned groups $0, 1, 2$*, while group* 3 *still has zero hidden representation.*

The following subsections prove the proposition, including the additional incorrect admission, its causal delay, and the resulting selection and risk comparisons.

B.4 Closing the common parameter and gate bounds

For semantic indices $j \in \{0,1,2,3\}$ and methods $m \in \{F,P\}$, work up to a first possible failure on $[0,H]$ of

$$\varepsilon \le \beta_j^m \le U, \qquad \beta_0^m \ge \delta, \qquad M \le S^m \le V, \tag{72}$$

together with seed acceptance, initial-error acceptance with the incorrect label, target rejection, and candidate rejection for FixMatch. They hold initially: $M-\delta > M-U > \kappa$, $\delta+\varepsilon < \kappa$, and $\gamma(\delta)=\gamma(\varepsilon)=0$.

Each semantic pathway occurs in at most two cooccurrence pairs. On the candidate box, monotonicity of r bounds each admitted edge coefficient by E_m. Ignoring the adverse root terms in (50) bounds the root coefficient by A_m. Consequently, before any possible failure,

$$\beta_0^m(t) \le \delta + \sqrt{2}UA_m t, \qquad \beta_j^m(t) \le \varepsilon + 2\sqrt{2}UE_m t \quad (j=1,2,3). \tag{73}$$

The candidate contribution to the root is nonpositive, so it does not increase these upper bounds. They are strictly below U through time H by (60). Similarly, (55) gives

$$S^m(t) \le M + \sqrt{2}VB_m t < V, \tag{74}$$

using (61).

For the lower faces, edge coefficients are nonnegative and spurious-feature coefficients are nonnegative. For the FixMatch root, dropping nonnegative terms and using $r(S-u) \le r(-u)$ yields

$$\begin{aligned} g_0^F &\ge \lambda p_r\{p_d\pi_0 r(\beta_0) - \nu r(-\beta_0)\} \\ &= \lambda p_r[p_d\pi_0 - \nu e^{\beta_0}]r(\beta_0) \\ &\ge \lambda p_r[p_d\pi_0 - \nu e^{U}]r(U) = L_F > 0. \end{aligned} \tag{75}$$

For Pseudo-Labeling,

$$g_0^P \ge Cr(V+U) - Nr(M-U) - \lambda\omega r(\kappa) = L_P > 0. \tag{76}$$

Thus the root cannot decrease below δ, other semantic strengths cannot decrease below ε, and S cannot decrease below M. Equation (47) also preserves positive heads and u_0 preactivations.

The seed's weak true margin is $S+\beta_0 > \kappa$, and the initial error group's is $\beta_0 - S \le U - M < -\kappa$. Their labels and gates therefore persist. Every target margin is at most $\gamma(U) < \kappa$, so targets remain rejected. For the candidate group, $q_m \ge \rho M - U > 0$ keeps its prediction incorrect in both methods. Before a hypothetical first FixMatch admission, its gate is closed and

$$\dot{q}_F = \rho\dot{S}^F - \dot{\beta}_0^F \le \sqrt{2}(\rho VB_F - \delta L_F) < 0. \tag{77}$$

Starting at $\kappa - \zeta$, this margin cannot first cross upward through κ; the positive lower bound also prevents correct admission through $-\kappa$. Thus the candidate remains rejected by FixMatch. This closes its gate condition jointly with the state bounds, establishing (72) and all stated gate properties on the closed interval $[0,H]$.

All semantic strengths are now nondecreasing. Hence each correct edge gate can switch from rejected to accepted at most once. Between such switches the reduced vector field is smooth, and the positive heads keep training preactivations away from zero. At a reached edge boundary, the accepted edge adds a nonnegative coefficient, with positive endpoint derivatives. Continuing with the inclusive gate convention gives a continuous, piecewise smooth trajectory satisfying the equations almost everywhere. Appendix B.5 proves that the candidate has one transversal admission under Pseudo-Labeling and never leaves. There are therefore at most four selection events. A target ReLU crossing does not change this training vector field because its loss coefficient is zero throughout $[0,H]$.

B.5 Incorrect selection reinforces the existing feature dependence

Relative dependence on the spurious feature. Before Pseudo-Labeling selects its first correct edge, define $R_P = \log(S^P/\beta_0^P)$. With $u=\beta_0^P$ and $q_P = \rho S^P - u$, the exact equations give

$$\dot{R}_P = 2\sqrt{2}Nr(S^P-u) + \sqrt{2}\lambda\omega\chi_\omega^P(1+\rho)r(q_P) > 0. \tag{78}$$

The seed contribution cancels in this ratio. Initial errors increase the relative spurious contribution, and newly admitted errors add a second positive term. Since $s_j^2 = \sqrt{2}\beta_j$, the filter-response ratio obeys $\log(s_s/s_0) = R_P/2$: the same bias occurs in the feature extractor. After correct edges enter, their semantic gradients also enter this derivative. For FixMatch, the closed candidate gate and the state bounds give

$$\frac{d}{dt}\log(S^F/\beta_0^F) = \sqrt{2}(g_s^F - g_0^F) \le \sqrt{2}(B_F - L_F) < 0. \tag{79}$$

Indeed, $\rho V > \delta$ and (64) imply $B_F < L_F$.

Admission of an additional error group. Let $t_{P,0}$ denote Pseudo-Labeling's first correct root-edge admission. Until that event, all nonroot strengths stay at ε, so the event occurs when $u = b$. Ignoring adverse terms gives $\dot{u} \le D_P^+$ and hence $t_{P,0} \ge T_0^-$. The exact derivative of the candidate's wrong-label margin is

$$\begin{aligned}\dot{q}_P/\sqrt{2} = {} & q_P Cr(S^P + u) + (\rho S^P + u)Nr(S^P - u) \\ & + \lambda\omega\chi_\omega^P(\rho^2 S^P + u)r(q_P) \\ & - u\sum_{i=1,2}\chi_{0i}^P\lambda\pi_e r(u + \beta_i^P).\end{aligned} \tag{80}$$

The candidate's own contribution is nonnegative. Each admitted correct edge has residual at most $r(\kappa)$. Using the tighter state bounds $u \le U_P$ and $S^P \le V_P$ therefore yields

$$\begin{aligned}\dot{q}_P &\ge L_q - E_q\mathbf{1}\{t \ge t_{P,0}\}, \\ q_P(t) &\ge \kappa - \zeta + L_q t - E_q(t - T_0^-)_+.\end{aligned} \tag{81}$$

Here $(z)_+ = \max(z, 0)$. Before the first correct edge, the margin increases at rate at least $L_q > 0$. It reaches κ at some $t_a \le T_a < T_0^-$ and the candidate is selected incorrectly. It remains selected until T_0^- because its derivative is positive before the first correct edge. On $[T_0^-, H]$, the lower bound in (81) is decreasing, but its value at H exceeds κ by (69). Thus the gate stays open through H. In particular, the incorrectly selected mass is

$$e_P(t) = \nu + \omega\mathbf{1}\{t \ge t_a\}, \qquad e_F(t) = \nu \quad (0 \le t \le H). \tag{82}$$

The population has not changed; the current predictor has admitted previously rejected errors, whose loss further favors the spurious feature over the semantic root.

B.6 A CAUSAL DELAY CAUSED BY THE NEW ERRORS

Compare the actual Pseudo-Labeling trajectory with a continuation from the same initialization that excludes only the candidate group's loss. All other groups, confidence rules, and loss weights remain unchanged; the full-population objective is not renormalized. Subscripts err and ex denote the actual and excluded trajectories. Before their first correct-edge admissions, the excluded root and spurious dynamics have the field

$$\begin{aligned}F_0(u, S) &= \sqrt{2}u\{Cr(S + u) - Nr(S - u)\}, \\ G_s(u, S) &= \sqrt{2}S\{Cr(S + u) + Nr(S - u)\}.\end{aligned} \tag{83}$$

The actual field adds $-\sqrt{2}u\lambda\omega\chi_\omega^P r(q_P)$ to $\dot{u}$ and $\sqrt{2}S\lambda\omega\rho\chi_\omega^P r(q_P)$ to $\dot{S}$. The excluded continuation satisfies the same state bounds: removing this group improves the root lower bound and reduces the spurious upper bound.

Let $a(z) = r(z)(1 - r(z))$, so $r'(z) = -a(z)$. In coordinates $(u, -S)$, both off-diagonal derivatives of the excluded field are nonnegative if $Ca(S + u) \ge Na(S - u)$. For $0 < u \le b$,

$$\frac{a(S - u)}{a(S + u)} = e^{-2u}\left(\frac{1 + e^{S+u}}{1 + e^{S-u}}\right)^2 \le e^{2u} \le e^{2b}. \tag{84}$$

Thus (70) makes this field cooperative: increasing either coordinate cannot decrease the other's derivative. The added error loss subtracts from both coordinates $(u, -S)$. A first-contact comparison starting from equal states gives

$$u_{\mathrm{err}} \le u_{\mathrm{ex}}, \qquad S_{\mathrm{err}} \ge S_{\mathrm{ex}} \tag{85}$$

up to the excluded trajectory's first correct-edge time T_{ex}. At a hypothetical first contact with $u_{\mathrm{err}} = u_{\mathrm{ex}}$, the ordered spurious coordinate and additional negative root term prevent crossing; the other boundary follows in the same way. A small positive perturbation of the upper trajectory's field makes this comparison strict, and taking the perturbation to zero gives (85). An earlier correct edge on the actual trajectory is impossible because its root cannot reach b while the excluded root remains below b.

For a quantitative bound, set $w = u_{\mathrm{ex}} - u_{\mathrm{err}}$ and define

$$\begin{aligned} L_d &= \sqrt{2}\{N + b(C+N)/4\}, \\ \eta_d &= \sqrt{2}\delta\lambda\omega r(\rho V - \delta), \qquad d = T_0^- - T_a > 0. \end{aligned} \tag{86}$$

For $u \in [\delta, b]$, $a(z) \le 1/4$ gives $\partial_u F_0 \ge -L_d$, and the same comparison condition gives $\partial_S F_0 \le 0$. Once the candidate is admitted, its adverse root rate is at least $\eta_d > 0$. Hence, almost everywhere before T_{ex},

$$\dot{w} \ge -L_d w + \eta_d \mathbf{1}\{t \ge t_a\}. \tag{87}$$

Both trajectories are identical before t_a. Also $T_{\mathrm{ex}} \ge T_0^-$ and $t_a \le T_a$, so integrating gives

$$w(T_{\mathrm{ex}}) \ge \frac{\eta_d}{L_d}(1 - e^{-L_d d}) > 0. \tag{88}$$

At this time the excluded root has reached b, but the actual root has not. Until its own first correct edge, the actual root rate remains at most D_P^+. Therefore

$$t_{P,0} - T_{\mathrm{ex}} \ge \frac{\eta_d(1 - e^{-L_d d})}{L_d D_P^+} > 0. \tag{89}$$

After T_{ex} the excluded learner receives new correct supervision, so only the actual root-rate bound is used for this last step. The positive lower root bound ensures that the actual event occurs by $(b-\delta)/D_P^- < H$. Thus the added errors cause a delay between complete training continuations, not just a negative instantaneous update at a shared state.

B.7 Separate time bounds for the two learners

FixMatch. Equation (75) and $\beta_0^F \ge \delta$ imply $\dot{\beta}_0^F \ge D_F$. Let $t_{F,0}$ be the first time its root reaches b. Then

$$t_{F,0} \le \frac{b-\delta}{D_F}. \tag{90}$$

Both its neighbors are at least ε, so this opens edges $(0,1)$ and $(0,2)$ because $b + \varepsilon = \kappa$. Each admitted edge supplies either endpoint with derivative at least D_E. Indeed, conditional on retaining that endpoint, deleting the other endpoint can only increase the residual, and therefore

$$\sqrt{2}\beta_j \mathcal{D}_F(\beta_j, \beta_i) \ge \sqrt{2}\varepsilon\lambda\pi_e p_r r(2U) = D_E. \tag{91}$$

Pathways 1 and 2 reach strength b within another $(b-\varepsilon)/D_E$ units of time. Once pathway 1 reaches b, edge $(1,3)$ is accepted and pathway 3 reaches b within at most the same additional duration. Consequently, by $\overline{T}_F$ every semantic strength is at least $b > h$. All four target representations are active, with margin at least $\gamma(b) > 0$, and every graph edge is selected. The confidence gates and gradient bounds therefore determine when this stage of feature learning has completed.

Pseudo-Labeling before its first edge. Until any edge is selected, only the root semantic pathway changes. The other strengths stay equal to ε, so the two root edges first open together when $\beta_0^P = b$. Edge $(1,3)$ cannot open earlier, since both its endpoints would still equal ε. Let $t_{P,0}$ be this first root-edge time. On the segment $\delta \le \beta_0^P \le b$, the seed and both error-group terms give

$$D_P^- \le \dot{\beta}_0^P \le D_P^+. \tag{92}$$

For the lower bound, use $S \le V$, $S - \beta_0 \ge M - b$, the bound $\chi_\omega^P r(q_P) \le r(\kappa)$, and $\beta_0 \ge \delta$; the bracket is positive. For the upper bound, discard the adverse terms and use $S + \beta_0 \ge M + \delta$ and $\beta_0 \le b$. It follows that

$$\frac{b-\delta}{D_P^+} \le t_{P,0} \le \frac{b-\delta}{D_P^-}. \tag{93}$$

The upper bound is below H under (67), so the event occurs within the certified box.

Pseudo-Labeling after its root edges. Before edge $(1,3)$ opens, pathways 1 and 2 receive only their respective root-edge gradients. For either pathway while its strength lies in $[\varepsilon, b]$,

$$D_L^- \le \dot{\beta}_j^P \le D_L^+. \tag{94}$$

The lower bound uses $\beta_0 \le U$ and $\beta_j \le b$. For the upper bound, the admitted root edge has $\beta_0 + \beta_j \ge \kappa$ and $\beta_j \le b$. Thus each pathway reaches strength h no later than $\overline{T}_P$, provided the last edge has not already opened. If that edge has opened, pathway 1 has already reached $b > h$, and its additional edge gradient is nonnegative; the same upper bound on reaching h remains valid. Pathway 2 has no additional edge. In particular, both target representations are strictly active at times greater than $\overline{T}_P$ while the last edge is still absent.

Until its first incident edge is selected, pathway 3 stays at strength ε. The first possible selection of $(1,3)$ therefore requires $\beta_1 = b$. Reaching this level from ε takes at least $(b-\varepsilon)/D_L^+$ after the first root-edge event. Let $t_{P,\mathrm{far}}$ be the first selection time of edge $(1,3)$. Together with (93), this gives

$$t_{P,\mathrm{far}} \ge \frac{b-\delta}{D_P^+} + \frac{b-\varepsilon}{D_L^+} = \underline{T}_{P,\mathrm{far}}. \tag{95}$$

If no such event occurs by H, the assertion means that it cannot occur in $[0, H]$ before this lower bound.

For $t \in I$, these estimates place each learner at its own current state: FixMatch has all four target representations active, whereas Pseudo-Labeling has pathways $0, 1, 2$ active and pathway 3 unchanged. Both have selected initially rejected cooccurrence samples. The interval comparison uses no coordinate-wise ordering between the two complete parameter vectors.

B.8 Why new selections provide necessary subsequent supervision

The hidden response learned on an unselected target and the confidence of a cooccurrence sample share trainable filter parameters. An admitted edge supplies the positive endpoint derivatives in (49); these move a previously untouched filter and its head, eventually changing the corresponding target representation according to (56). Pathway 1 also participates in edge $(1,3)$, so its changing logit contribution can admit a new source of gradient for pathway 3.

This mechanism distinguishes two populations of examples. Cooccurrence edges are correctly classified initially but have insufficient confidence for admission. Target singletons change from collapsed to class-separable representations and remain unselected throughout the proof. The result does not require that the same newly corrected examples are subsequently admitted.

For a direct counterfactual, freeze the identities and pseudo-labels of the initially selected seed and initial error groups, leaving the loss, views, and parameter updates otherwise unchanged. Neither those groups nor the labeled samples contain a nonzero patch at pathways $1, 2, 3$. Every partial derivative with respect to their heads and filters is therefore zero for either learner. Those parameters stay at their initial values and their target representations remain zero. Allowing the candidate group alone would not change this conclusion because it also has no nonroot semantic patches. New admissions are consequently necessary for the nonroot semantic representation improvements proved above. The fixed graph specifies which patches cooccur; the admission events themselves are generated by each learner's current model.

B.9 Selection precision, classifier risk, and representation risk

For $t \in I$, FixMatch selects all three edge groups and Pseudo-Labeling selects exactly the two root-edge groups. The seed remains selected and correct, the initial error group remains selected and incorrect, and no target group is selected. The candidate is selected only by Pseudo-Labeling. Let $c_m(t)$ and $e_m(t)$ denote correctly and incorrectly selected population masses, and let $Q_m(t)$ denote selected-set precision. Then

$$\begin{aligned} c_F(t) &= \pi_0 + 3\pi_e, & c_P(t) &= \pi_0 + 2\pi_e, \\ e_F(t) &= \nu, & e_P(t) &= \nu + \omega, \\ Q_m(t) &= \frac{c_m(t)}{c_m(t) + e_m(t)}. \end{aligned} \tag{96}$$

Since $\pi_e, \nu, \omega > 0$,

$$Q_F(t) - Q_P(t) = \frac{\nu\pi_e + \omega(\pi_0 + 3\pi_e)}{(\pi_0 + 3\pi_e + \nu)(\pi_0 + 2\pi_e + \nu + \omega)} > 0. \tag{97}$$

FixMatch's precision advantage combines more correct selections with avoidance of the new error group.

All seed and edge margins are positive. Both error groups have negative true margins under both classifiers, since $S - \beta_0 > \kappa$ and $q_m > 0$. Every active target group has positive true margin. For Pseudo-Labeling, only target group 3 has zero score, and predicting $+1$ incurs half of its balanced mass. Therefore the trained classifiers' population error rates $\mathcal{R}_m(t)$ satisfy

$$\mathcal{R}_F(t) = \nu + \omega, \qquad \mathcal{R}_P(t) = \nu + \omega + \frac{\pi_T}{8}. \tag{98}$$

For a fixed hidden map $Z : \mathcal{X} \to \mathbb{R}^{10}$ on input space $\mathcal{X}$, define the optimal affine linear-probe risk by minimizing over a weight vector $\xi \in \mathbb{R}^{10}$ and intercept $\eta \in \mathbb{R}$:

$$\mathcal{R}_{\text{lin}}(Z) = \inf_{\xi,\eta} \Pr\{\text{sign}_+(\xi^\top Z(X) + \eta) \neq Y\}. \tag{99}$$

The same conclusions hold if the intercept is fixed to zero. Under Pseudo-Labeling, both labels in target group 3 have exactly the same hidden vector zero, so every such classifier errs on at least half that group's mass. This lower bound is attained by assigning positive weights to all positive-class semantic hidden coordinates, negative weights to the negative-class coordinates, zero weights to the spurious-feature coordinates, and zero intercept. It correctly classifies every other population group, including the two error groups, whose root patches lie in the true-class channel. Under FixMatch the same semantic-only readout is correct on every group. Hence

$$\mathcal{R}_{\text{lin}}(Z_t^F) = 0, \qquad \mathcal{R}_{\text{lin}}(Z_t^P) = \frac{\pi_T}{8}. \tag{100}$$

This proves Proposition 3 and the claimed common-time selection and risk conclusions. The representation gap cannot be attributed solely to a different trained head, because the lower bound for Z_t^P holds for every linear readout.

Both methods improve their representations relative to initialization: initially all four target groups collapse to zero, giving optimal linear-probe risk $\pi_T/2$. The comparison interval therefore contains productive learning in both methods. It describes transfer within the specified population to never-selected target examples; it does not identify optimal population probe risk with a probe fitted for finitely many steps in an experiment.

B.10 A VERIFIED PARAMETER INSTANCE

The sufficient conditions are jointly satisfiable. Take

$$\begin{gathered}
\tau = 0.95, \quad \kappa = \log 19, \quad h = \kappa/2, \quad \alpha = \lambda = 1, \\
\pi_0 = 0.3, \quad \pi_e = 0.15, \quad \pi_T = 0.248, \quad \nu = \omega = 0.001, \\
p_d = 0.6, \quad p_r = 0.8, \quad M = 5, \\
\varepsilon = h - 0.001, \quad \delta = h - 0.0005, \quad b = h + 0.001, \\
U = 1.525, \quad V = 5.01, \quad H = 0.48, \\
\zeta = 0.0001, \qquad \rho = (\kappa + \delta - \zeta)/M.
\end{gathered} \tag{101}$$

The target input is q_0 itself; no small target scaling is used.

For clarity, the following strict enclosures report evaluations of the analytic bounds rather than simulated event times:

Quantity	Verified enclosure
$\delta + \sqrt{2}UA_F H$	$(1.52054, 1.52056)$
$\delta + \sqrt{2}UA_P H$	$(1.48935, 1.48937)$
$\varepsilon + 2\sqrt{2}UE_F H$	$(1.49045, 1.49047)$
$\varepsilon + 2\sqrt{2}UE_P H$	$(1.48677, 1.48679)$
$M + \sqrt{2}VB_F H$	$(5.00633, 5.00635)$
$M + \sqrt{2}VB_P H$	$(5.00708, 5.00709)$
L_F	$(0.02507, 0.02509)$
L_P	$(0.0018044, 0.0018045)$
$\gamma(U)$	$(0.01331, 0.01332)$
$\rho M - U$	$(2.89105, 2.89106)$
$\sqrt{2}(\rho VB_F - \delta L_F)$	$(-0.040535, -0.040534)$

All semantic and spurious-feature upper bounds are strictly below U and V. Both root lower coefficients are positive. The remaining confidence condition holds since $M - U = 3.475 > \log 19$; also $\gamma(U) < 0.014 < \kappa$. The candidate remains incorrectly predicted in both methods and rejected by FixMatch, as established by the last two rows.

For Pseudo-Labeling's candidate admission and retention, the bounds are

$$\begin{aligned} L_q &\in (0.0083419, 0.0083420), \qquad E_q \in (0.0315940, 0.0315941), \\ T_a &\in (0.0119875, 0.0119876), \qquad T_0^- \in (0.3586482, 0.3586483), \\ -\zeta + L_q H - E_q(H - T_0^-) &\in (0.00007015, 0.00007016). \end{aligned} \tag{102}$$

Thus the extra error group enters before time 0.012 and remains selected through H. The comparison condition also holds, with $C - Ne^{2b} > 1.28096$. Substituting into (89) gives a strictly positive delay:

$$\frac{\eta_d(1 - e^{-L_d d})}{L_d D_P^+} \in (0.0076220, 0.0076221). \tag{103}$$

The candidate-excluded root reaches its first correct edge by 0.368463, whereas the actual root reaches it by 0.378130; these are upper bounds, and the causal difference is bounded below by (103).

The separate time estimates give

$$\begin{aligned} \frac{b-\delta}{D_F} &\in (0.02873, 0.02874), \qquad \overline{T}_F \in (0.38304, 0.38305), \\ \frac{b-\delta}{D_P^+} &\in (0.35864, 0.35866), \qquad \frac{b-\delta}{D_P^-} \in (0.37812, 0.37814), \\ \overline{T}_P &\in (0.44557, 0.44558), \qquad \underline{T}_{P,\mathrm{far}} \in (0.48664, 0.48665). \end{aligned} \tag{104}$$

These bounds establish the separation on $[0.45, 0.48)$. The same conclusions hold at $H = 0.48$: the invariant box was proved on the closed horizon, both activation upper bounds are strictly below H, and the last Pseudo-Labeling edge has lower admission time strictly above H. Thus the certified comparison interval includes $[0.45, 0.48]$. The quantities are continuous training times under (41), not experimental epochs.

On this interval the exact population quantities are

Quantity	FixMatch	Pseudo-Labeling
Correct selected mass	0.75	0.60
Incorrect selected mass	0.001	0.002
Selected-set precision	$0.75/0.751$	$0.60/0.602$
Trained-classifier risk	0.002	0.033
Optimal linear-probe risk	0	0.031

Furthermore, $\gamma(b) \in (0.00025004, 0.00025005)$, so the learned FixMatch targets have strictly positive margin. The small margin reflects an initialization close to the target ReLU activation boundary, despite unit target input scale.

The numerical enclosures were checked using 60-digit outward interval evaluation of the displayed inequalities. The accompanying `verification/` directory provides a self-contained Python checker and its certificate. All conditions are strict and continuous in the listed parameters, so nearby choices within this symmetric network and population family also satisfy them.

C Experimental Details and Additional Results

We compare classification accuracy, linear-probe accuracy, and selected pseudo-label populations along paired SSL training trajectories. Holding each checkpoint fixed, we measure how the training view changes the target residual and its associated learning pressure. We then use controlled continuations to examine how selected datasets affect representation learning and the next pseudo-label selection. All checkpoint diagnostics use the same three training seeds per dataset.

C.1 Main training protocol and statistical units

We compare Pseudo-Labeling and FixMatch on CIFAR-10 with 40 labeled samples and CIFAR-100 with 400 labeled samples, corresponding to four labels per class. Both methods use WRN-28-2 with dropout disabled and BN momentum $.001$. For each dataset, seeds 0, 1, and 2 define three paired training runs. In the main SSL runs and controlled continuations, one reported epoch corresponds to 1024 optimizer updates rather than a full pass through the dataset. Main training therefore comprises 1,048,576 updates over 1024 epochs. Our implementation builds on USB (Wang et al., 2022), using the upstream commit `1ef4cbebcc0b368158315aeb425053858cf6c845` of `https://github.com/microsoft/Semi-supervised-learning`. We describe the paired training protocol and configuration below. The USB defaults used in the main text specify the batch sizes, unlabeled loss weight, and augmentation operations. Labeled and unlabeled batch sizes are 64 and 448, respectively. The unlabeled loss weight is one, and the confidence threshold is $.95$. We average the masked unlabeled loss over the full unlabeled batch, including the zero contributions of rejected examples.

We train in FP32 using Nesterov SGD with momentum $.9$, initial learning rate $.03$, and weight decay 5×10^{-4}. BN parameters and biases are excluded from weight decay. We use no warmup, gradient clipping, or mixed precision. At update u, the learning rate is $.03\cos((7\pi/16)u/1{,}048{,}576)$; it therefore remains nonzero at the endpoint. The EMA decay is $.999$. We refer to the non-averaged model as the student and report its accuracy separately. EMA evaluation uses averaged weights together with live student buffers, including BN running statistics.

Weak augmentation consists of horizontal flipping and random cropping. For strong augmentation, we use the pinned USB implementation with `RandAugment(3,5)`. It samples three operations with replacement, draws a parameter uniformly from each operation's prescribed range, and applies its built-in random Cutout. In this implementation, the constructor argument `m=5` is deprecated and does not determine the augmentation magnitude. Both datasets use normalization with mean $(.485, .456, .406)$ and standard deviation $(.229, .224, .225)$. Each update uses three separate forward passes in the following order: the labeled batch, the weak view for pseudo-label selection without gradients, and the unlabeled training view. The inputs are not concatenated. To implement the comparison between Pseudo-Labeling and FixMatch, we keep the entire training recipe fixed and vary only the unlabeled training view. Pseudo-Labeling reuses the exact weak-view tensor used during pseudo-label selection for the final forward pass. The model, data, optimizer, selection rule, and loss normalization are shared.

The unlabeled training loader includes all 50,000 training identities, including labeled examples. For offline population diagnostics, we remove the labeled identities, leaving 49,960 and 49,600 genuine unlabeled examples. Their true labels are reserved for diagnostics and the oracle intervention.

The main figure displays linear probes at epochs $4, 8, 16, 32, 64, 128, 256, 512, 768, 1024$. Additional measurements at epochs $100, 300, 500$ enter the trajectory audit below. Population audits follow the denser checkpoint schedule specified in Appendix C.2. All plotted means use three seeds. We report the prescribed main-training endpoint at epoch 1024.

Unless otherwise stated, standard deviations are sample SDs across three training seeds. We average the four strong-augmentation measurements per image within each seed before aggregating across

seeds. Checkpoints and augmentation draws are repeated measurements from the same runs and should not be interpreted as independent replications. Additional residual tables report means only.

C.2 Selected populations and representation probes

We measure selected-set precision every eight epochs, giving 128 checkpoints for each dataset, method, and seed. At each checkpoint, we evaluate a canonical weak view of each genuine unlabeled identity. Precision is the number of correctly pseudo-labeled accepted samples divided by the total number of accepted samples; it is undefined when the selected set is empty. These offline measurements count unique identities in evaluation mode. During online selection, the model uses stochastic weak views and batch-dependent BN, and it may encounter each identity repeatedly. Table 3 reports correct and incorrect pseudo-label counts together with precision.

Table 3: Final offline selected populations. Values are reported as mean $\pm$ sample SD across three training seeds. The total selected count is the sum of correct and wrong counts. Precision is computed separately for each seed and then averaged across seeds.

Dataset	Method	Correct count	Wrong count	Precision (%)
CIFAR-10	Pseudo-Labeling	9848.7 ± 547.1	34586.3 ± 426.9	22.16 ± 1.01
CIFAR-10	FixMatch	28951.0 ± 16515.5	7469.3 ± 4995.9	74.13 ± 25.18
CIFAR-100	Pseudo-Labeling	714.3 ± 922.0	23551.0 ± 20467.6	3.14 ± 2.04
CIFAR-100	FixMatch	14869.3 ± 617.5	16304.3 ± 529.8	47.69 ± 0.86

We use linear probing to assess learned features under a common classifier-training protocol, separating the representation from its co-trained classification head. We freeze the non-averaged model's feature extractor and compute representations of all 50,000 clean training images, standardizing features using training means and standard deviations only. We then fit a new linear head using all training labels for 100 epochs, with batch size 512, SGD learning rate .1, momentum .9, no weight decay, and epoch-wise cosine decay. Head initialization and data order are fixed with seed 712345. We evaluate the head on the 10,000 test images, without test-based tuning or early stopping. For this diagnostic classifier, one epoch corresponds to a full pass through the 50,000 training representations. The fixed update budget used to define an SSL epoch does not apply to probe fitting. Only the diagnostic head is fitted with these true labels; the SSL representation remains fixed. The probes in Figure 1 use epochs $4, 8, 16, 32, 64, 128, 256, 512, 768, 1024$.

Table 4 reports Student, EMA, and probe accuracies at the prescribed endpoint. FixMatch achieves higher mean accuracy under all three measures. Its probe advantage shows that the common linear readout can recover class labels more accurately from its learned representation; the Student and EMA results evaluate the classifiers produced by SSL training.

Figures 2–4 show the individual paired runs underlying the means in Figure 1. Lines connect recorded checkpoints throughout training. The probe and precision curves follow different parts of the same process: the class information available in the representation and the correctness of the model's offline high-confidence predictions.

The final main-training EMA accuracies for seeds 0/1/2 are $21.93/24.59/20.44\%$ (CIFAR-10 Pseudo-Labeling), $91.54/36.95/75.23\%$ (CIFAR-10 FixMatch), $8.06/1.04/1.24\%$ (CIFAR-100 Pseudo-Labeling), and $40.75/38.14/39.16\%$ (CIFAR-100 FixMatch). The corresponding mean best-observed test accuracies across available scheduled population-audit checkpoints are $24.31 \pm 1.78\%$, $89.75 \pm 2.82\%$, $12.32 \pm .85\%$, and $45.03 \pm 3.55\%$, respectively. These values are descriptive maxima based on test accuracy; no validation set is used to select them.

On CIFAR-10, FixMatch seed 1 reaches 86.52% EMA accuracy at epoch 112 and ends at 36.95%. Its probe also falls from 88.31% at epoch 128 to 65.32% at epoch 256, while student accuracy falls from 79.85% to 36.30%. This drop affects both the original classifier and measured linear separability. Seed 2 has a different late pattern: from epoch 256 to 1024, EMA accuracy falls from 90.63% to 75.23%, but probe accuracy changes from 89.90% to 89.40%. Thus, original-classifier and probe changes need not have the same size. All probes here use native student BN buffers. On CIFAR-100, approximately 99.995% and 100% of the final-window predictions from Pseudo-Labeling seeds 1 and 2, respectively, belong to a single class, with online selected-set precision near 1%. Online and offline selection also differ substantially: seed 2 accepts approximately 99.997% of

Table 4: Full accuracy results at the prescribed final checkpoint, epoch 1024. Values are mean ± sample SD over three training seeds. "Student" denotes the non-averaged model. "EMA" uses averaged weights with the non-averaged model's live BN buffers; probes use its frozen feature extractor and a newly fitted linear head.

Dataset	Method	EMA test (%)	Student test (%)	Probe (%)
CIFAR-10	Pseudo-Labeling	22.32 ± 2.10	22.41 ± 1.64	40.69 ± 0.25
CIFAR-10	FixMatch	67.91 ± 28.02	66.40 ± 28.13	83.19 ± 13.40
CIFAR-100	Pseudo-Labeling	3.45 ± 4.00	2.44 ± 2.20	27.37 ± 0.36
CIFAR-100	FixMatch	39.35 ± 1.32	36.95 ± 0.64	58.35 ± 1.05

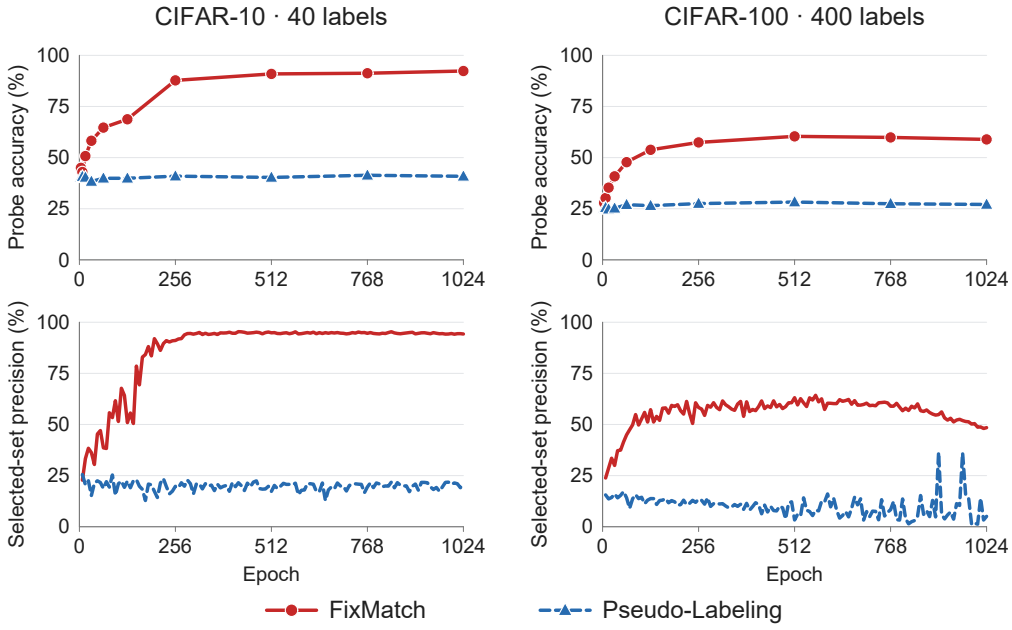

Figure 2: Seed 0: linear-probe accuracy (top) and offline selected-set precision (bottom), for CIFAR-10 (left) and CIFAR-100 (right). Red solid lines show FixMatch; blue dashed lines show Pseudo-Labeling. Probe markers denote evaluated checkpoints: red circles for FixMatch and blue triangles for Pseudo-Labeling. Precision uses lines only. Both datasets are shown throughout training.

online presentations but only 90 identities in its final offline snapshot. BN recalibration alone does not restore useful final classification in these models: every final student/EMA accuracy changes by at most 0.92 percentage points. Earlier low-accuracy episodes can include temporary recoveries; these measurements do not establish the cause of deterioration.

Selection changes after epoch 256. For CIFAR-100 FixMatch seeds 0/1/2, correct selected counts change from $11{,}482/10{,}593/13{,}603$ at epoch 256 to $14{,}875/14{,}249/15{,}484$ at epoch 1024. Wrong counts grow from $8{,}217/6{,}750/6{,}825$ to $15{,}823/16{,}218/16{,}872$. Precision therefore falls from $58.29/61.08/66.59\%$ to $48.46/46.77/47.86\%$, although the corresponding probe changes are only $+1.47/+0.34/+0.04$ percentage points. More correct selections do not necessarily mean higher precision.

Nor does a stable count imply a fixed selected population. For CIFAR-10 Pseudo-Labeling seed 0, epochs 288 and 296 select 41,106 and 41,055 identities, but 6,520 enter and 6,571 leave. Direct comparison of the stored identity arrays gives these counts. For CIFAR-100 FixMatch seed 1, epochs 944 and 952 both select 28,044 identities, with 6,635 entering and 6,635 leaving according to the within-run identity tracker. These comparisons use the same fixed weak view for each identity. They describe offline selection changes, rather than online minibatch exposures.

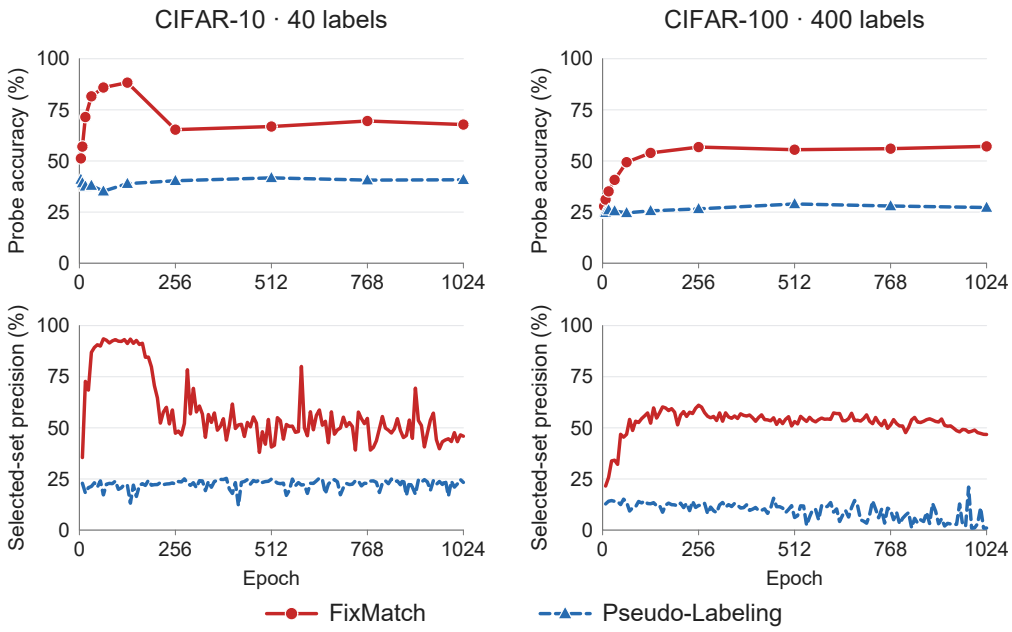

Figure 3: Seed 1: linear-probe accuracy (top) and offline selected-set precision (bottom), for CIFAR-10 (left) and CIFAR-100 (right). Red solid lines show FixMatch; blue dashed lines show Pseudo-Labeling. Probe markers denote evaluated checkpoints: red circles for FixMatch and blue triangles for Pseudo-Labeling. Precision uses lines only. Each curve shows this seed's recorded measurements throughout training.

C.3 Residuals across training checkpoints

At each checkpoint, we evaluate the full genuine unlabeled pool under a fixed weak view and select all images whose maximum predicted probability is at least .95, with the model in evaluation mode. We hold the selected images and pseudo-labels fixed throughout the comparison. This set contains only the images selected at the current checkpoint, excluding images selected only at earlier stages of training. We evaluate $r_v = 1 - p_{\tilde{y}}^v$ on the exact selection view and on four randomly generated strong augmentations per image, without updating the model. Ground-truth labels partition the selected samples into correct and incorrect pseudo-label groups and are not used as training targets.

Tables 5 and 6 report the nine checkpoints available for all three seeds on both datasets: $64, 100, 128, 256, 300, 500, 512, 768, 1024$. For each seed and pseudo-label group, we average the strong-view residual over the four augmentations and all images in that group. The weak-view residual is averaged over the same images. Each table entry averages the three seed-level means, giving every seed equal weight. The mean strong-view residual exceeds the weak-view residual for both methods and both groups at every reported checkpoint. Because the parameters, selected identities, and targets are fixed, this difference is caused by the training view. The residual equals the magnitude of the target-logit gradient, and twice the residual equals the full logit-gradient ℓ_1 norm (Appendix A.1). The measurements therefore show increased pressure toward both correct and incorrect pseudo-labels.

Additional one-step diagnostics. We conduct additional one-step diagnostics at epochs 100, 300, and 500 for both source methods. The correctly and incorrectly pseudo-labeled cohorts each contain 512 identities, matched by predicted class and weak-confidence bins with boundaries $.95, .96, .97, .98, .99, 1.00$. Incorrect pseudo-labels are the models' own predictions. A disjoint 1024-identity clean audit population is excluded from diagnostic updates, although it was available during preceding SSL training.

Each paired virtual update begins with identical parameters and buffers. BN statistics are frozen, dropout is disabled, and the update uses only the mean pseudo-label cross-entropy over the cohort.

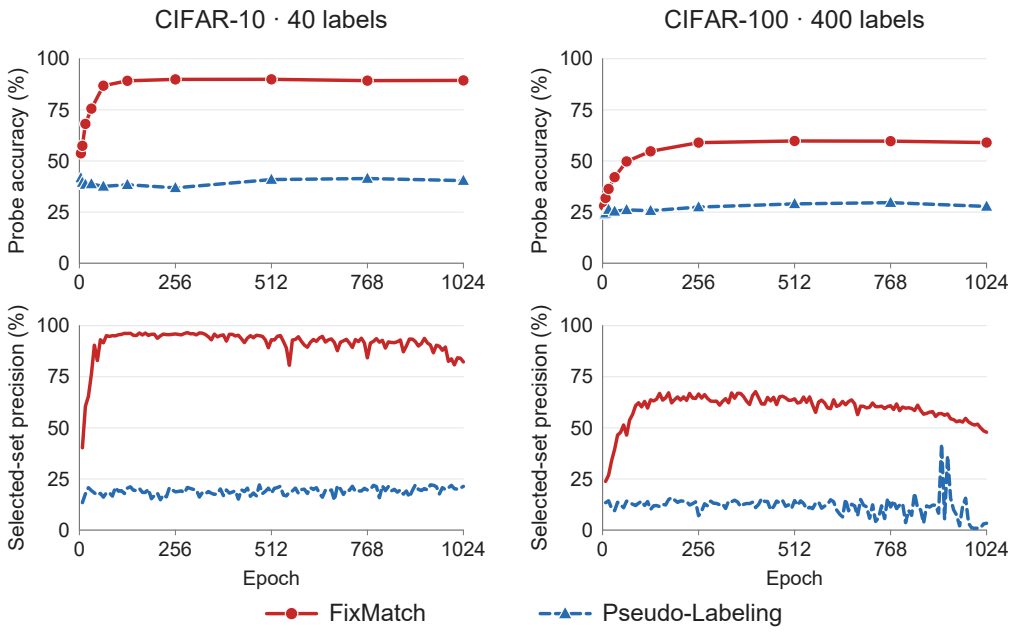

Figure 4: Seed 2: linear-probe accuracy (top) and offline selected-set precision (bottom), for CIFAR-10 (left) and CIFAR-100 (right). Red solid lines show FixMatch; blue dashed lines show Pseudo-Labeling. Probe markers denote evaluated checkpoints: red circles for FixMatch and blue triangles for Pseudo-Labeling. Precision uses lines only. Each curve shows this seed's recorded measurements throughout training.

Table 5: CIFAR-10 pseudo-label residuals at the nine shared checkpoints. Each model selects its high-confidence samples using a fixed weak view. Correct and Wrong partition them by pseudo-label correctness. Weak reuses the selection view; Strong averages four random strong augmentations per image. Entries are means across three training seeds.

Epoch	Method	Correct targets		Wrong targets	
		Same weak	Strong	Same weak	Strong
64	Pseudo-Labeling	0.0025	0.5751	0.0030	0.5788
64	FixMatch	0.0061	0.2016	0.0132	0.2849
100	Pseudo-Labeling	0.0030	0.6325	0.0037	0.6355
100	FixMatch	0.0054	0.1788	0.0134	0.3215
128	Pseudo-Labeling	0.0027	0.6225	0.0034	0.6341
128	FixMatch	0.0059	0.2045	0.0140	0.3097
256	Pseudo-Labeling	0.0027	0.6359	0.0032	0.6409
256	FixMatch	0.0064	0.2313	0.0142	0.3360
300	Pseudo-Labeling	0.0020	0.5984	0.0021	0.5853
300	FixMatch	0.0046	0.1700	0.0119	0.2638
500	Pseudo-Labeling	0.0018	0.6044	0.0023	0.6171
500	FixMatch	0.0062	0.2083	0.0129	0.3054
512	Pseudo-Labeling	0.0022	0.6115	0.0027	0.6123
512	FixMatch	0.0059	0.2008	0.0127	0.2830
768	Pseudo-Labeling	0.0019	0.6487	0.0024	0.6688
768	FixMatch	0.0053	0.2114	0.0107	0.3193
1024	Pseudo-Labeling	0.0015	0.6413	0.0020	0.6705
1024	FixMatch	0.0044	0.1693	0.0095	0.2666

There is no supervised term, momentum, weight decay, or reselection. We evaluate strong-view updates over four random augmentation realizations while keeping the original source model and optimizer unchanged.

Table 6: CIFAR-100 pseudo-label residuals at the nine shared checkpoints, using the same fixed-model evaluation and three-seed averaging as Table 5.

Epoch	Method	Correct targets		Wrong targets	
		Same weak	Strong	Same weak	Strong
64	Pseudo-Labeling	0.0075	0.6980	0.0114	0.7322
64	FixMatch	0.0081	0.2946	0.0144	0.4081
100	Pseudo-Labeling	0.0079	0.7369	0.0108	0.7622
100	FixMatch	0.0074	0.2441	0.0142	0.3621
128	Pseudo-Labeling	0.0073	0.7372	0.0109	0.7486
128	FixMatch	0.0070	0.2625	0.0135	0.3806
256	Pseudo-Labeling	0.0066	0.7013	0.0085	0.6883
256	FixMatch	0.0061	0.2504	0.0128	0.3567
300	Pseudo-Labeling	0.0066	0.6779	0.0099	0.7127
300	FixMatch	0.0063	0.2607	0.0127	0.3662
500	Pseudo-Labeling	0.0058	0.6876	0.0086	0.6864
500	FixMatch	0.0064	0.2680	0.0121	0.3773
512	Pseudo-Labeling	0.0055	0.6229	0.0071	0.4813
512	FixMatch	0.0060	0.2559	0.0118	0.3744
768	Pseudo-Labeling	0.0061	0.6513	0.0092	0.5764
768	FixMatch	0.0054	0.2326	0.0103	0.3432
1024	Pseudo-Labeling	0.0178	0.3191	0.0134	0.2263
1024	FixMatch	0.0044	0.1903	0.0073	0.2944

Table 7: Full pseudo-labeled datasets selected at epoch 64. Counts and precision are reported as mean $\pm$ sample SD across three training seeds. Precision is computed within each seed before averaging.

Dataset	Selector	Correct	Wrong	Total	Precision (%)
CIFAR-10	FixMatch	24856 ± 13466	6692 ± 7051	31548 ± 6461	74.57 ± 31.18
	Pseudo-Labeling	7163 ± 731	34149 ± 2467	41312 ± 2923	17.34 ± 1.32
CIFAR-100	FixMatch	6390 ± 904	7467 ± 927	13857 ± 1817	46.08 ± 1.00
	Pseudo-Labeling	2352 ± 439	14446 ± 2560	16798 ± 2996	13.99 ± 0.29

In the follow-up analysis of step size, 57 of 72 source/cohort conditions have a mean strong-update direction that favors lower clean-audit cross-entropy. However, only 25 reduce this loss at the full checkpoint learning rate. In the remaining 32 conditions with favorable directions, descent is recovered at all tested smaller multipliers $10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}$. The other 15 conditions already have unfavorable update directions. The mean strong-view gradient norm exceeds its weak-view counterpart in all 72 conditions, with ratios ranging from 7.48 to 917.44. Thus both the update direction and the step size matter when translating greater learning pressure into lower loss on other samples.

C.4 Training with fixed pseudo-labeled datasets

At epoch 64, each FixMatch and Pseudo-Labeling student selects all genuine unlabeled images whose maximum predicted probability is at least .95, using a fixed weak view in evaluation mode. We retain each selected image with its predicted label. Table 7 reports the resulting datasets, including their correct and incorrect pseudo-label counts. On both datasets, Pseudo-Labeling selects more images in total, but FixMatch selects more correct targets and fewer wrong targets. For CIFAR-10, the mean wrong counts are 6,692 for FixMatch and 34,149 for Pseudo-Labeling; for CIFAR-100, they are 7,467 and 14,446. These populations differ in the supervision they provide, not just in size.

For each dataset and seed, we copy each model checkpoint into two branches and train one branch on each selected dataset. This gives 24 continuations: two datasets, three seeds, two starting models, and two data sources. All selected identities and pseudo-labels remain fixed throughout each continuation. For each presented pseudo-labeled image, FixMatch branches generate a strong view and Pseudo-Labeling branches generate a weak view. Both methods use weak views for labeled images. Within each pair of branches, the training-view rule is fixed.

Each branch continues training for 50 epochs (51,200 optimizer updates), from update 65,536 to 116,736, retaining the checkpoint's SGD momentum and learning-rate schedule. Labeled and

Table 8: Starting and final accuracies for the epoch-64 dataset intervention. Initial denotes the model before continuation; FixMatch data and Pseudo-Labeling data identify the dataset used for the 50-epoch continuation. All entries use the same BN recalibration protocol and report percentages (mean $\pm$ sample SD across three training seeds).

Dataset	Recipient	Evaluation	Test (%)	Probe (%)
CIFAR-10	FixMatch	Initial	68.77 ± 24.50	79.22 ± 12.53
		FixMatch data	65.33 ± 27.36	79.44 ± 10.71
		Pseudo-Labeling data	21.57 ± 1.80	57.17 ± 2.08
	Pseudo-Labeling	Initial	20.68 ± 1.95	38.18 ± 1.41
		FixMatch data	63.31 ± 28.87	76.85 ± 12.71
		Pseudo-Labeling data	16.07 ± 1.42	45.16 ± 1.48
CIFAR-100	FixMatch	Initial	25.70 ± 1.65	48.14 ± 1.35
		FixMatch data	27.95 ± 0.99	49.07 ± 0.62
		Pseudo-Labeling data	14.48 ± 0.77	39.87 ± 0.29
	Pseudo-Labeling	Initial	10.04 ± 0.67	25.88 ± 0.75
		FixMatch data	23.94 ± 1.22	45.70 ± 1.08
		Pseudo-Labeling data	9.05 ± 0.55	28.54 ± 1.13

pseudo-labeled batch sizes are 64 and 128, respectively. Both loaders sample with replacement. The loss is the equally weighted sum of mean labeled cross-entropy and mean pseudo-labeled cross-entropy. Paired branches start from identical model, optimizer, and scheduler states. They also share labeled draws, augmentation base seeds, and uniform sampling draws for the pseudo-labeled datasets. True labels from the unlabeled pool are reserved for diagnostic measurements and linear-probe fitting.

During continuation, BN modules use current-batch statistics and update their running statistics with momentum .001. Non-BN modules and dropout remain in evaluation mode. For the primary evaluation, we recalibrate each branch's BN statistics on all 50,000 clean training images, presented in ascending identity order as 200 batches of 250 images. Before recalibration, we reset BN running means, variances, and batch counters to zero, one, and zero, respectively, and use cumulative averaging (`momentum=None`). Recalibration updates only BN buffers and uses neither labels nor parameter gradients. We apply the same procedure before continuation to evaluate the starting model. Test accuracy uses the existing classification head. Probe accuracy uses the frozen feature extractor and a new linear head fitted as described in Appendix C.2.

Table 8 reports the starting and final accuracies. FixMatch-generated data yield higher final test and probe accuracy in all 12 paired comparisons. Test accuracy evaluated with the native tracked BN statistics also favors FixMatch-generated data in all 12 comparisons. The main table reports mean paired differences across the three training seeds. Each seed contributes one difference for each dataset and recipient method. Within a pair, the starting parameters and training rule are identical, so the endpoint difference measures the effect of replacing the selected supervision. Higher probe accuracy locates part of this effect in the learned representation, rather than only in the existing classification head. The replacement changes image coverage, label assignments, and precision together. Because the loss is averaged over sampled pseudo-labeled images, the dataset size changes sample exposure rather than the unlabeled loss weight.

C.5 Subsequent selection after the dataset intervention

We evaluate the same epoch-64 starting models and the endpoints of the 50-epoch continuations in Section C.4, without further training. For each dataset and seed, let $\mathcal{U}$ be the genuine unlabeled pool and $\mathcal{S}_F, \mathcal{S}_P$ the two supplied pseudo-labeled identity sets. The common candidate pool is $\mathcal{H} = \mathcal{U} \setminus (\mathcal{S}_F \cup \mathcal{S}_P)$. Its images contribute to neither continuation's supervised nor pseudo-labeled loss. They may have appeared during earlier SSL training and are included in the all-image BN recalibration.

For each recipient and BN policy, let $\mathcal{A}_0$ be the identities in $\mathcal{H}$ selected at the starting checkpoint and $\mathcal{A}_D$ those selected after continuation on source $D \in \{F, P\}$. We use the same fixed weak views and threshold .95 at both times. Newly selected identities form $\mathcal{A}_D \setminus \mathcal{A}_0$. Their correct and wrong

Table 9: New selections on the common candidate pool after fixed-supervision continuation, with the primary BN recalibration. Dataset source identifies the frozen supervision used for continuation. Counts and new-selection precision are mean ± sample SD over three seeds.

Dataset source	C_{new}	W_{new}	Q_{new} (%)
CIFAR-10 / FixMatch recipient			
Pseudo-Labeling	13.7 ± 2.5	37.7 ± 15.5	27.95 ± 6.39
FixMatch	185.7 ± 50.9	132.0 ± 105.2	61.67 ± 19.60
CIFAR-10 / Pseudo-Labeling recipient			
Pseudo-Labeling	4.7 ± 3.1	43.3 ± 27.1	9.57 ± 0.42
FixMatch	163.0 ± 14.9	215.3 ± 244.8	54.14 ± 27.26
CIFAR-100 / FixMatch recipient			
Pseudo-Labeling	313.7 ± 24.5	1417.0 ± 71.9	18.15 ± 1.77
FixMatch	1225.3 ± 187.1	2104.7 ± 428.0	36.99 ± 3.24
CIFAR-100 / Pseudo-Labeling recipient			
Pseudo-Labeling	285.0 ± 19.5	4173.0 ± 617.1	6.45 ± 0.66
FixMatch	1230.0 ± 104.9	3986.3 ± 689.4	23.76 ± 2.18

Table 10: New selections with native tracked BN statistics, using the same candidate identities and continuation models as Table 9.

Dataset source	C_{new}	W_{new}	Q_{new} (%)
CIFAR-10 / FixMatch recipient			
Pseudo-Labeling	6.7 ± 3.5	22.0 ± 14.4	26.06 ± 8.26
FixMatch	411.7 ± 86.4	312.7 ± 164.9	58.53 ± 9.13
CIFAR-10 / Pseudo-Labeling recipient			
Pseudo-Labeling	40.0 ± 24.5	220.0 ± 133.9	15.50 ± 0.73
FixMatch	628.7 ± 127.0	695.7 ± 698.6	55.09 ± 19.49
CIFAR-100 / FixMatch recipient			
Pseudo-Labeling	233.3 ± 55.4	1135.3 ± 235.5	17.00 ± 1.25
FixMatch	1254.3 ± 161.0	2522.3 ± 462.5	33.39 ± 2.50
CIFAR-100 / Pseudo-Labeling recipient			
Pseudo-Labeling	332.0 ± 85.7	4437.0 ± 1440.7	7.15 ± 1.34
FixMatch	1413.3 ± 113.5	5310.3 ± 1063.1	21.31 ± 2.76

counts are C_{new} and W_{new}; their precision is $Q_{\text{new}} = C_{\text{new}}/(C_{\text{new}} + W_{\text{new}})$ when the denominator is positive. These targets are generated by the continued recipient, not supplied by the source model.

Evaluation policies and averaging. The primary evaluation uses the common clean-image BN recalibration procedure in Section C.4, applied separately to each model. We also evaluate each model with its native tracked BN statistics. Native baseline selections on $\mathcal{H}$ are empty by its definition, but recalibration can make baseline predictions cross the threshold. We therefore compute newly selected identities relative to the baseline under the same policy, rather than treating all endpoint selections as new. The two continuation branches for a given recipient share the same candidate pool and baseline. Tables 9 and 10 report three-seed means and sample SDs. Precision is computed within each seed before averaging. Table 2 uses the mean and sample SD of paired source differences.

New entries and changes in the full selected set. FixMatch-generated supervision gives more new correct selections and higher new-selection precision in all 12 recipient-level paired comparisons under both BN policies. New wrong selections also increase in 10 of the 12 primary comparisons and 11 of the 12 native comparisons. Tables 11 and 12 retain the per-seed counts, including losses of previous correct selections. Here $N = |\mathcal{H}|$, C_0 is the number of baseline correct selections, L_C counts baseline correct selections no longer selected correctly at the endpoint, and D_C is the endpoint correct count minus C_0. This net change also includes corrections of baseline incorrect selections. The source abbreviations F and P mean FixMatch-generated and Pseudo-Labeling-generated data, respectively.

New correct selections need not produce net growth. For CIFAR-10 FixMatch recipients trained on FixMatch-generated data, the primary baseline correct counts are $1{,}075/327/614$, while endpoint counts are $405/313/520$. These same branches produce $163/150/244$ newly correct selections. The

Table 11: Per-seed selection transitions with primary BN recalibration. Each seed appears twice, once per supervision source (P or F). C_0, L_C, D_C describe the full candidate pool; $C_{\text{new}}, W_{\text{new}}, Q_{\text{new}}$ describe its baseline-rejected members newly selected at the endpoint.

Seed/source	N	C_0	C_{new}	W_{new}	L_C	D_C	Q_{new} (%)
CIFAR-10 / FixMatch recipient							
0/P	5626	1075	11	22	1069	-1057	33.33
0/F	5626	1075	163	253	846	-670	39.18
1/P	1726	327	16	38	321	-305	29.63
1/F	1726	327	150	62	168	-14	70.75
2/P	2820	614	14	53	604	-589	20.90
2/F	2820	614	244	81	339	-94	75.08
CIFAR-10 / Pseudo-Labeling recipient							
0/P	5626	873	8	73	865	-825	9.88
0/F	5626	873	146	498	729	-298	22.67
1/P	1726	203	2	20	202	-192	9.09
1/F	1726	203	169	71	134	288	70.42
2/P	2820	334	4	37	327	-316	9.76
2/F	2820	334	174	77	235	387	69.32
CIFAR-100 / FixMatch recipient							
0/P	22654	691	312	1484	594	-273	17.37
0/F	22654	691	1091	2098	343	765	34.21
1/P	25844	912	290	1426	790	-478	16.90
1/F	25844	912	1439	2536	420	1070	36.20
2/P	24149	844	339	1341	742	-389	20.18
2/F	24149	844	1146	1680	450	719	40.55
CIFAR-100 / Pseudo-Labeling recipient							
0/P	22654	386	266	3468	283	1	7.12
0/F	22654	386	1109	3415	287	1123	24.51
1/P	25844	404	305	4436	315	28	6.43
1/F	25844	404	1286	4752	316	1238	21.30
2/P	24149	419	284	4615	308	19	5.80
2/F	24149	419	1295	3792	341	1261	25.46

comparison with the other data source improves even though each branch ends with fewer correct selections than it started with.

Initially correct but low-confidence candidates. We also examine the pre-specified subset of $\mathcal{H}$ that the starting recipient predicts correctly but rejects. Membership depends on the recipient and BN policy, and is fixed for its two continuations. All per-seed outcomes appear in Tables 13 and 14, where N now denotes this subset's size. For CIFAR-100 Pseudo-Labeling recipients, FixMatch-generated supervision gives lower new-selection precision in all three seeds under both BN policies. With native BN, its new correct counts are also lower in seeds 1 and 2, by 75 and 4 samples. Thus, the advantage on the full candidate pool does not extend uniformly to every subset.

The dataset replacement changes sample identities and labels together. It affects both the representation and subsequent predictions, as shown by the probe and selection measurements. It does not isolate how much of the selection effect passes through the feature extractor rather than the classifier. The continuation fixes supervision for 50 epochs and measures the next selection; the natural runs instead refresh targets throughout training. The accompanying file `data/next_selection_metrics.csv` contains the per-seed measurements.

Table 12: Per-seed transitions with native BN statistics. Symbols follow Table 11. Baseline selections are empty under this policy, so $C_0 = L_C = 0$ and $D_C = C_{\text{new}}$.

Seed/source	N	C_0	C_{new}	W_{new}	L_C	D_C	Q_{new} (%)
CIFAR-10 / FixMatch recipient							
0/P	5626	0	3	6	0	3	33.33
0/F	5626	0	457	495	0	457	48.00
1/P	1726	0	10	26	0	10	27.78
1/F	1726	0	312	174	0	312	64.20
2/P	2820	0	7	34	0	7	17.07
2/F	2820	0	466	269	0	466	63.40
CIFAR-10 / Pseudo-Labeling recipient							
0/P	5626	0	64	344	0	64	15.69
0/F	5626	0	726	1501	0	726	32.60
1/P	1726	0	15	78	0	15	16.13
1/F	1726	0	485	253	0	485	65.72
2/P	2820	0	41	238	0	41	14.70
2/F	2820	0	675	333	0	675	66.96
CIFAR-100 / FixMatch recipient							
0/P	22654	0	171	868	0	171	16.46
0/F	22654	0	1112	1991	0	1112	35.84
1/P	25844	0	252	1312	0	252	16.11
1/F	25844	0	1222	2741	0	1222	30.84
2/P	24149	0	277	1226	0	277	18.43
2/F	24149	0	1429	2835	0	1429	33.51
CIFAR-100 / Pseudo-Labeling recipient							
0/P	22654	0	281	2983	0	281	8.61
0/F	22654	0	1321	4088	0	1321	24.42
1/P	25844	0	431	5864	0	431	6.85
1/F	25844	0	1379	5824	0	1379	19.14
2/P	24149	0	284	4464	0	284	5.98
2/F	24149	0	1540	6019	0	1540	20.37

Table 13: Initially correct but rejected candidates: per-seed outcomes after primary BN recalibration. F and P denote the supervision sources; N is the baseline subset size.

Seed/source	N	C_{new}	W_{new}	Q_{new} (%)
CIFAR-10 / FixMatch recipient				
0/P	1050	7	4	63.64
0/F	1050	103	49	67.76
1/P	645	10	24	29.41
1/F	645	131	11	92.25
2/P	1128	8	29	21.62
2/F	1128	208	14	93.69
CIFAR-10 / Pseudo-Labeling recipient				
0/P	237	2	10	16.67
0/F	237	31	73	29.81
1/P	109	1	4	20.00
1/F	109	34	9	79.07
2/P	83	0	3	0.00
2/F	83	25	11	69.44
CIFAR-100 / FixMatch recipient				
0/P	2871	156	206	43.09
0/F	2871	687	148	82.28
1/P	3547	164	189	46.46
1/F	3547	905	173	83.95
2/P	3513	186	213	46.62
2/F	3513	727	114	86.44
CIFAR-100 / Pseudo-Labeling recipient				
0/P	874	142	92	60.68
0/F	874	168	142	54.19
1/P	1139	129	163	44.18
1/F	1139	165	219	42.97
2/P	936	104	150	40.94
2/F	936	126	183	40.78

Table 14: Initially correct but rejected candidates under native BN. Definitions follow Table 13; subset membership is determined using native baseline predictions.

Seed/source	N	C_{new}	W_{new}	Q_{new} (%)
CIFAR-10 / FixMatch recipient				
0/P	1452	2	3	40.00
0/F	1452	361	23	94.01
1/P	917	2	12	14.29
1/F	917	262	50	83.97
2/P	1408	5	16	23.81
2/F	1408	393	57	87.33
CIFAR-10 / Pseudo-Labeling recipient				
0/P	762	36	22	62.07
0/F	762	137	171	44.48
1/P	205	7	3	70.00
1/F	205	70	24	74.47
2/P	395	22	27	44.90
2/F	395	115	35	76.67
CIFAR-100 / FixMatch recipient				
0/P	3193	74	113	39.57
0/F	3193	760	126	85.78
1/P	3882	115	186	38.21
1/F	3882	891	139	86.50
2/P	3817	132	186	41.51
2/F	3817	1019	192	84.15
CIFAR-100 / Pseudo-Labeling recipient				
0/P	1167	182	60	75.21
0/F	1167	218	191	53.30
1/P	1344	251	132	65.54
1/F	1344	176	294	37.45
2/P	1136	164	95	63.32
2/F	1136	160	221	41.99

C.6 ORACLE PRECISION INTERVENTION

For the oracle-precision intervention, we use epoch-300 checkpoints and fix the image identities in the full dataset selected by Pseudo-Labeling at that epoch. We first assign each image its true label, then swap labels within disjoint cross-class pairs to construct precision levels $q \in \{.7, .8, .9, 1\}$. Nested prefixes of a common pair list determine the corruption levels. This procedure preserves the identity count and the oracle-corrected true-class histogram, while allowing the original dataset's pseudo-class histogram to change. The correct count and precision vary together at a fixed total count. At each precision level, we continue training the FixMatch starting models for 50 epochs using either weak or strong views. Batch sizes, loss, and BN evaluation procedures follow Section C.4. Oracle labels are used solely for this diagnostic intervention. Table 15 shows how the relative advantage of strong training views changes as the fraction of correct targets increases.

Table 15: Oracle-precision intervention from epoch-300 FixMatch checkpoints: strong-view minus weak-view endpoint test accuracy in percentage points, mean $\pm$ sample SD across three training seeds. Primary endpoints use branch-wise BN recalibration.

Dataset	$q = .7$	$q = .8$	$q = .9$	$q = 1$
CIFAR-10	19.31 ± 0.87	9.03 ± 1.85	1.70 ± 1.56	-5.88 ± 4.02
CIFAR-100	5.36 ± 1.82	1.84 ± 0.97	-1.28 ± 1.04	-4.47 ± 2.26

On both datasets, the mean strong-minus-weak difference in endpoint accuracy decreases as target precision increases. This comparison measures the relative effect of the training view; it does not measure the absolute benefit of correcting labels. The result also depends on the evaluation policy. For CIFAR-100 FixMatch starting models, the effects at $q = .9$ and $q = 1$ change from $-1.28/-4.47$ pp after recalibration to $+4.52/+1.96$ pp with native BN statistics. On CIFAR-10, all three strong-view continuations from FixMatch checkpoints at $q = .7$ yield empty selected sets at the primary, BN-recalibrated endpoint. Their selected-set precision is therefore undefined, while endpoint test accuracy remains defined.

C.7 RELATION TO THE THEORY

The residual measurements test the target-logit signal comparison in Section 3.2, including the increased pressure toward incorrect targets discussed in Section 3.3. The fixed dataset intervention tests the next links: selected supervision changes subsequent representation learning and selection from the same starting model. The training trajectories show how probe accuracy and selected-set precision develop under each method's own evolving model. Together, the controlled comparisons support the learning-pressure and supervision–learning–selection links in the theory. Natural trajectories show that relative advantages can coexist with deterioration within individual runs.

The newly admitted error cohort, its reinforcement of spurious features, and the resulting delay of correct selection are established for the model in Appendix B. The experiments measure residual amplification and the effect of complete selected datasets, rather than isolating that error-feedback mechanism or the individual effect of pseudo-label precision.

C.8 FIXED-CHECKPOINT TRAINING AND EVALUATION DIAGNOSTICS

To distinguish evaluation-mode selection from training-mode behavior, we examine both methods at epochs 64, 256, and 1024, using all three seeds on each dataset. Each checkpoint is evaluated on 3,584 genuine unlabeled identities, divided into eight batches of 448. The identities and augmented images are shared across methods and checkpoints within each dataset and seed. We use one fixed weak view and one fixed strong view per image, and one fixed batch of 64 samples drawn with replacement from the labeled set. Pseudo-Labeling uses its exact weak-view tensor for both selection and training-view prediction.

We compare four forward policies: evaluation mode with the checkpoint's native BN statistics; separate training-mode forwards in the recorded order labeled–weak–training; separate forwards in USB's labeled–training–weak order; and one training-mode forward with the three batches concatenated. For each policy, pseudo-labels are selected from its weak-view predictions at threshold .95. Model parameters and buffers are restored before each batch and policy. No gradient or optimizer

update is taken. We repeat the measurements with two partitions of the same identities. These are paired diagnostics, not additional training seeds or a replay of the original online batches.

Residuals under different forward policies. For FixMatch, the mean strong-view target residual exceeds the weak-view residual on both correct and incorrect pseudo-labels in all $18 \times 4 \times 2 = 144$ checkpoint/policy/partition combinations. At epoch 64 on CIFAR-100, concatenated forwards give three-seed mean residuals of $.00738$ versus $.26189$ for correct targets and $.01438$ versus $.37869$ for incorrect targets (weak versus strong). These averages give equal weight to the two partitions within each seed and then to the three seeds. Pseudo-Labeling's weak-view and training-view residuals are identical in every combination. The residual distinction therefore persists under both evaluation-mode and training-mode BN.

Selection depends on the evaluation policy. Table 16 illustrates the difference on the same checkpoint and images. CIFAR-100 Pseudo-Labeling seed 2 at epoch 1024 selects only four samples in evaluation mode, but all 3,584 in training mode. Both training-mode policies in Table 16 predict a single class for all images, with only 29 correct predictions. The native evaluation also predicts that class on 98.19% of images, but with lower confidence. Thus, a small offline selected set need not imply little exposure to incorrect pseudo-labels during training.

Table 16: Selection on 3,584 fixed images using the epoch-1024 CIFAR-100 Pseudo-Labeling checkpoint from seed 2. Counts agree under both batch partitions. Separate uses the recorded training order; Concatenated uses one combined forward. No parameters are updated.

Forward policy	Selected	Correct	Wrong	Precision (%)
Native evaluation	4	0	4	0.00
Separate training	3,584	29	3,555	0.81
Concatenated training	3,584	29	3,555	0.81

Concatenation does not consistently improve selection at these saved weights. At epoch 1024 on CIFAR-100, its mean precision is 1.85% versus 1.65% for separate forwards in Pseudo-Labeling, and 47.79% versus 48.12% in FixMatch. Swapping the order of the separate forwards leaves current probabilities and selections unchanged in all paired checks, although the resulting BN running statistics can differ. These measurements identify forward-policy sensitivity, not the cause of the training decline or the outcome of training with concatenation from initialization. The accompanying file `data/forward_policy_metrics.csv` contains all 288 measurements; `fmws` and `fmww` denote FixMatch and Pseudo-Labeling.