# UNDERSTANDING THE LEARNING PHASES IN SELF-SUPERVISED LEARNING VIA CRITICAL PERIODS

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

## ABSTRACT

Self-supervised learning (SSL) has emerged as a powerful pretraining strategy to learn transferable representations from unlabeled data. Yet, it remains unclear how long SSL models should be pretrained for such representations to emerge. Contrary to the prevailing heuristic that longer pretraining translates to better downstream performance, we identify a *transferability trade-off*: across diverse SSL settings, intermediate checkpoints often yield stronger out-of-domain (OOD) generalization, whereas additional pretraining primarily benefits in-domain (ID) accuracy. From this observation, we hypothesize that SSL progresses through learning phases that can be characterized through the lens of *critical periods* (CP). Prior work on CP has shown that supervised learning models exhibit early phases of high plasticity, followed by a consolidation phase where adaptability declines but task-specific performance keeps increasing. Since traditional CP analysis depends on supervised labels, for SSL we rethink CP in two ways. First, we inject deficits to perturb the pretraining data and measure the quality of learned representations via downstream tasks. Second, to estimate network plasticity during pretraining we compute the Fisher Information matrix on pretext objectives, quantifying the sensitivity of model parameters to the supervisory signal defined by the pretext tasks. We conduct several experiments to demonstrate that SSL models do exhibit their own CP, with CP closure marking a *sweet spot* where representations are neither underdeveloped nor overfitted to the pretext task. Leveraging these insights, we propose *CP-guided checkpoint selection* as a mechanism for identifying intermediate checkpoints during SSL that improve OOD transferability. Finally, to balance the transferability trade-off, we propose *CP-guided self-distillation*, which selectively distills layer representations from the sweet spot (CP closure) checkpoint into their overspecialized counterparts in the final pretrained model.

## 1 INTRODUCTION

Self-supervised learning (SSL) leverages pretext tasks (e.g., contrasting views or predicting masked inputs) to learn representations from unlabeled data that transfer well to downstream tasks (Ozbulak et al., 2023; Gui et al., 2024). While prior work has studied *how well* SSL models transfer (Ericsson et al., 2021a), *why* they transfer (Ericsson et al., 2021b), and *under what conditions* they succeed (Tian et al., 2020; Zhao et al., 2020; Cole et al., 2022; Dubois et al., 2022; 2023), it remains unclear **how long to pretrain SSL models** for transferable representations to emerge.

Without knowing when the SSL model has learned enough from its pretext task, pretraining risks both under- and over-training. Stopping too early yields underdeveloped representations, such that a common practice is that *longer pretraining is beneficial* (Chen et al., 2020; He et al., 2022). However, pretraining for too long increases computational costs and risks overfitting to pretraining biases.

Determining the optimal pretraining duration is difficult because SSL objectives are only implicitly aligned with downstream transferability (Balestriero et al., 2023; Reizinger et al., 2025). Typically, the quality of SSL representations is assessed *after pretraining* via linear probing or finetuning (Chen et al., 2020; Kumar et al., 2022; Balestriero et al., 2023). Such post-hoc evaluation is costly to repeat across tasks and, more importantly, provides no guidance *during pretraining* about whether learned representations are underdeveloped or already overspecialized to the pretext task.

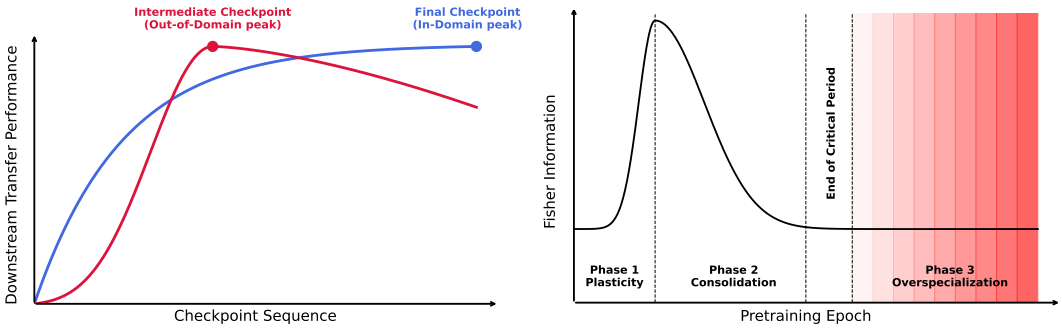



Figure 1: **(Left)** Conceptual schematic of downstream performance of SSL models across a sequence of pretrained checkpoints. In-domain (ID) downstream performance increases with pretraining. Out-of-domain (OOD) transferability, however, peaks at an intermediate checkpoint and declines thereafter, indicating that broadly transferable representations emerge early in pretraining. **(Right)** Conceptual schematic of Fisher Information (FI) dynamics during SSL pretraining. The curve shows three phases. *Phase 1 (Plasticity)* shows a rise in FI when representations are highly sensitive to changes. *Phase 2 (Consolidation)*, where FI declines and plateaus as representations stabilize. The *Critical Period (CP)* closes once FI levels off into a plateau, but before the phase enters *Phase 3 (Overspecialization)*, where FI remains stable but OOD transferability declines due to overtraining. Red shading highlights the loss of transferability beyond CP closure. **(Takeaway)** The end of CP marks a *sweet spot* where representations are neither underdeveloped nor overfitted to pretext biases.

By evaluating checkpoints across the SSL pretraining trajectory (Figure 1, left) we identify a **transferability trade-off**: intermediate checkpoints often achieve better out-of-domain (OOD) transfer than later checkpoints, whereas extended pretraining gives higher in-domain (ID) performance. Here, ID denotes fine-tuning on a labeled version of the pretraining dataset, while OOD denotes fine-tuning on datasets drawn from different distributions (Marks et al., 2025). Specifically, we evaluate two families of SSL methods: discriminative SSL (contrastive SimCLR (Chen et al., 2020) and non-contrastive VICReg (Bardes et al., 2021) and DINO (Caron et al., 2021)) and generative SSL (MAE (He et al., 2022), pretrained on ImageNet-1K (Deng et al., 2009) and fMoW-RGB (Christie et al., 2018). This transferability trade-off pattern implies that pretraining does not simply yield representations that improve uniformly across domains. Instead, we hypothesize that SSL progresses through distinct *learning phases* where the properties of the learned representations shift: early phases support OOD generalization, while later phases specialize toward the pretraining data distribution and improve ID accuracy.

To build intuition, we draw on the notion of *critical periods* (CP). Prior work (Achille et al., 2018) reports that, much like in biological systems, neural networks exhibit CP: they undergo an early window of high plasticity (when representations are highly sensitive to changes) followed by a consolidation phase (when representations stabilize and adaptability declines). In supervised learning, these phases were revealed through perturbation experiments, where temporary distortions of the training data (e.g., perturbing inputs mid-training) permanently impaired generalization if applied during early epochs but had little effect if applied later. This temporal sensitivity can be explained through Fisher Information (FI) (Fisher, 1925), which quantifies how strongly small parameter changes affect model predictions and serves as a proxy for *information plasticity* (Achille et al., 2018; Berariu et al., 2021). Early in training, FI rises and plasticity is high, so perturbations strongly reshape representations and leave lasting effects. As training continues, FI declines and representations consolidate, so later perturbations have little impact. Overall, CP analyses reveal *when* representations are adaptable or rigid, which may offer insight into transfer dynamics (Achille et al., 2018).

Yet critical periods (CP) have not been studied in SSL, where transferability between pretraining and downstream tasks is key. Unlike supervised learning, SSL derives its supervisory signal from the structure of the data rather than explicit labels, making prior CP analyses inapplicable. We therefore reformulate CP analyses to track information plasticity during SSL without downstream supervision. This is achieved in two ways: (1) applying perturbations during pretraining to test stage-wise effects on downstream transferability, and (2) redefining Fisher Information with respect to pretext tasks.

*We find that SSL models also undergo a structured progression* (Figure 1, right). During SSL pretraining, Fisher Information (FI) rises sharply, indicating a phase of high plasticity in which representations are highly sensitive to updates. FI then declines and stabilizes, marking a consolidation phase where task-irrelevant variability is discarded and representations lose sensitivity to new information. We identify CP closure at the end of consolidation: representations are sufficiently developed for transfer, but not yet overfitted to the pretext task. Beyond CP closure, FI further stabilizes but OOD generalization degrades, revealing a previously undefined stage of *overspecialization*.

This pattern helps explain our empirical findings that intermediate SSL checkpoints often transfer better OOD than later checkpoints: before CP closure, models retain plasticity that supports generalization beyond the pretraining distribution, while later checkpoints past CP closure have anchored to pretext-specific biases. Although the timing of these transitions varies across different SSL settings, *the presence of critical periods is consistent*.

Building on these observations, we show that critical periods provide a guide to steer SSL. *CP-guided checkpoint selection* uses CP closure as an unsupervised indicator, favoring OOD transfer, while pretraining beyond closure prioritizes ID performance. To balance this trade-off, we propose *CP-guided self-distillation*: during fine-tuning, we distill early-layer features from CP-selected checkpoints into the early layers of longer-pretrained models while leaving later layers intact. We target early layers because they are widely understood to encode more general information (Yosinski et al., 2014; Skean et al., 2025), which may lessen with prolonged pretraining, while later layers often encode more task-specific structure to satisfy the training objective (Bordes et al., 2022).

Our contributions can be summarized as follows:

- We reveal a *transferability trade-off* in SSL pretraining. Across the diverse SSL settings we evaluate, intermediate checkpoints often yield stronger out-of-domain (OOD) transferability, while models pretrained longer tend to improve in-domain (ID) accuracy. This calls for rethinking the standard practice in SSL that longer pretraining translates to better representations for downstream tasks (§2).
- We connect this phenomenon to the notion of critical periods (CP), providing the *first study of CP in SSL* and their impact on transferability. Since SSL objectives differ from supervised learning, we reformulate CP analyses for SSL by introducing perturbations into pretraining and redefining Fisher Information in terms of pretext tasks rather than downstream labels. These analyses reveal that SSL models also exhibit their own CP (§3).
- We identify a previously uncharacterized *overspecialization phase*, where prolonged pretraining anchors models to pretext-specific biases and reduces OOD generalization. Building on this insight, we propose two interventions: *CP-guided checkpoint selection*, which uses CP closure to identify intermediate checkpoints with stronger OOD robustness, and *CP-guided self-distillation*, which restores early-layer features from CP checkpoints into later checkpoints to recover OOD performance while retaining ID strength (§4).

## 2 Does longer self-supervised pretraining always improve downstream transferability?

Prior work in self-supervised learning (SSL) reported that longer pretraining improves downstream performance (Goyal et al., 2019; Chen et al., 2020; He et al., 2022). This has led to the de facto practice of pretraining SSL models for as long as compute budgets allow. We show that this improvement does not universally hold. Instead, we observe a **transferability trade-off**: *while extended pretraining improves in-domain (ID) performance, it often diminishes out-of-domain (OOD) transferability.*

### 2.1 Experimental Setup

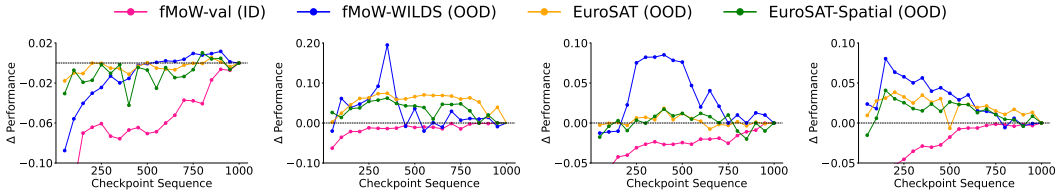
To study how pretraining duration affects downstream transferability, we evaluate two families of SSL using the methods, architectures, and datasets described below. For discriminative SSL, we include both contrastive (SimCLR (Chen et al., 2020)) and non-contrastive methods (VICReg (Bardes et al., 2021), DINO (Caron et al., 2021)). For generative SSL, we use MAE (He et al., 2022). For architectures, we use ResNet-50 (SimCLR, VICReg) and ViT-B16 (DINO, MAE), covering

convolutional and transformer-based backbones. Across these schemes, optimizers used vary between SGD, LARS, and AdamW. For datasets, we pretrain on ImageNet-1K (Deng et al., 2009) and fMoW-RGB (Christie et al., 2018), both large-scale datasets with over a million images that span complementary visual domains. This combination allows us to study the effect of pretraining duration in both an object-centric natural image setting (Beyer et al., 2020; Fang et al., 2023) and a real-world satellite domain with well-defined distribution shifts (Koh et al., 2021; Rolf et al., 2024).

We pretrain each model from scratch for 1000 epochs, saving checkpoints every 50 epochs. Downstream transfer is evaluated along two dimensions (Marks et al., 2025). In-domain (ID) performance is measured by fine-tuning on labeled versions of the pretraining data and reporting accuracy on its held-out test set. Out-of-domain (OOD) performance is measured by fine-tuning on datasets outside the pretraining distribution and evaluating on their test sets. For each checkpoint, we fine-tune and compare against the model pretrained for 1000 epochs, which we refer to as *final checkpoints*. Details on pretraining, downstream settings and datasets are provided in Appendix A.

### 2.2 Results

Figure 2 shows downstream transfer performance of SSL models across pretraining durations.




(a) SimCLR-RN50-fMoW (b) VICReg-RN50-fMoW (c) DINO-ViT-B16-fMoW (d) MAE-ViT-B16-fMoW

Figure 2: Transferability trade-off in SSL. The x-axis shows a sequence of checkpoints (every 50 epochs), and the y-axis shows downstream performance relative to the final checkpoints.

**Extended pretraining induces a transferability trade-off along the *pretraining data distribution*** We find that OOD transfer peaks at intermediate checkpoints and declines thereafter, while ID performance continues to rise (Figure 2). For example, VICReg-RN50 reaches its highest OOD accuracy around 350 epochs before dropping. A similar trend appears for MAE and DINO with ViT-B16: OOD transfer rises early, peaks, and then declines, while ID performance on fMoW-val continues to increase. SimCLR-RN50 follows the same trade-off but peaks later, around 850 epochs. This divergence indicates that intermediate checkpoints yield broadly transferable representations, though the exact timing varies by method, with later checkpoints increasingly specializing to the pretraining distribution. *ImageNet-based* results show a similar trend, as reported in Appendix B. Specifically, across SSL methods, intermediate checkpoints yield the strongest OOD transfer before declining without returning to earlier levels, while ID accuracy on ImageNet-val continues to rise throughout pretraining.

## 3 Critical Periods in Self-Supervised Learning

Insights from Section 2 raise the question: *why do different stages of pretraining yield such different transfer properties?* We hypothesize that SSL pretraining progresses through structured learning phases. To examine this, we draw on the notion of *critical periods* (CP). Prior work shows that neural networks undergo phases of early plasticity, when representations are highly sensitive to change, followed by reduced plasticity and consolidation (Achille et al., 2018; Kim et al., 2023). Yet whether such phases exist in SSL, and how they relate to transferability, remains unexplored. If SSL models pass through periods of heightened plasticity followed by stabilization, these transitions may underlie the observed transferability trade-off. Probing when plasticity is present or lost during pretraining offers a way to map the learning phases and examine their link to transferability.

To investigate possible explanations, in the next section we revisit critical period analyses in supervised learning (§3.1), followed by its reformulation for SSL via two approaches: perturbation experiments on pretraining data (§3.2) and Fisher Information on pretext objectives (§3.3).

### 3.1 PRIOR CRITICAL PERIOD ANALYSES REQUIRE RETHINKING FOR SSL

**How critical periods have been studied in supervised learning (SL).** Prior work identifies critical periods in SL in two ways (Achille et al., 2018). First, perturbation experiments probe whether the *timing* of perturbations matters. If altering the input distribution early in training degrades final accuracy, while the same change later has little effect, this marks a critical early phase. Second, Fisher Information (FI) (Fisher, 1925) analysis provides a continuous marker of plasticity. Computed with respect to class-label likelihoods, FI quantifies the sensitivity of model predictions to small parameter changes. Intuitively, a rise in FI reflects heightened plasticity, when the network is responsive to updates and can reorganize its representations. As FI declines, plasticity decreases and the network consolidates what it has learned, becoming less adaptable to new information (Achille et al., 2018).

**Why critical periods analyses must be rethought for SSL.** Both approaches assume labeled data, but SSL pretraining is decoupled from labels and optimizes proxy objectives on unlabeled data. One could study critical periods during fine-tuning, when downstream labels are available, but this only reveals how a *fixed representation* adapts to one task, not how transferable representations emerge *during pretraining*. Our focus is SSL pretraining itself, since this stage defines a generic prior aimed at broad downstream applicability. Formally, pretraining on unlabeled data $D_A$ produces a posterior $p(\theta \mid D_A)$, which serves as the prior for downstream data $D_B$. To capture how this prior evolves, critical periods must be analyzed *during pretraining*, not after.

**Probing critical periods in SSL.** To study critical periods in SSL, we introduce two probes during pretraining. (1) *Deficit injection on the unlabeled pretraining data* perturbs the input distribution. The pretext task remains unchanged, but the self-supervision signal is degraded (e.g., input perturbations remove fine-grained cues, making data pairs harder to align or reconstructions less informative). By varying when deficits are introduced and measuring their impact on downstream transfer, we can identify phases when representations are more or less sensitive to change. (2) *Fisher Information on pretext objectives* quantifies the sensitivity of model parameters to the supervisory signal defined by the pretext tasks. Tracking FI over pretraining reveals when parameters remain adaptable and when they consolidate, which is crucial in SSL since the value of pretraining lies in producing transferable representations. Identifying when representations are still malleable versus when they have committed helps explain when they are effective for downstream transfer.

### 3.2 PROBE 1: DEFICIT INJECTION ON UNLABELED PRETRAINING DATA

The central question is: *does the impact of input perturbations during SSL pretraining depend on **when** they occur?* If perturbations early in pretraining change the final representations, as reflected in downstream performance, while the same perturbations later in pretraining have little effect, then the SSL model exhibits a critical period.

Let $\mathcal{D} = \{x_i\}_{i=1}^N$ denote samples from a clean distribution $p(x)$. A model learns a representation function $f_\theta : \mathcal{X} \to \mathbb{R}^d$ with parameters $\theta$, trained with a self-supervised loss $\ell_{\text{SSL}}(f_\theta(x))$, such as contrastive loss or reconstruction error.

To inject a deficit, we replace clean training samples with data drawn from a perturbed distribution $p'(x)$ starting at onset epoch $t_0$ and lasting for a duration of $\Delta t$ epochs. After this window, training resumes on clean data until epoch $T$, where $T > t_0 + \Delta t$.

We denote the encoder trained entirely on clean data as $f_{\theta^*}$ (baseline) and the encoder trained with a deficit window as $f_{\theta'}$. To quantify the effect of the intervention, we compare downstream transfer performance between these models. Let $\Phi(\cdot)$ denote a downstream evaluation metric (e.g., classification accuracy). The sensitivity score is defined as

$$S(t_0) = \Phi(f_{\theta^*}) - \Phi(f_{\theta'}). \tag{1}$$

This score reflects the relative degradation in downstream performance caused by the intervention. A critical period exists if early interventions consistently yield higher sensitivity than later ones.

**Deficit Settings.** Following prior work (Achille et al., 2018), we simulate sensory deprivation by replacing inputs with Gaussian noise. For SSL methods, the pretext objectives (e.g., contrastive alignment or masked reconstruction) continue updating during the deficit window, but the supervisory signal comes from noise rather than meaningful images. As a result, the model learns nuisances that are not useful for downstream transfer. Each deficit is applied for a fixed window (5,

30, 50 epochs) at varying onset times $t_0$: early (epoch 0), middle (epoch 450), and late (epoch 750), following (Kleinman et al., 2024). After the deficit window, training resumes on clean inputs (until epoch $T = 1000$). We use the same evaluation settings as in Section 2.1.

**Early SSL pretraining phases are more sensitive to deficits.** Figure 3 shows the sensitivity $S(t_0)$ of learned representations to Gaussian noise deficits introduced at different times during pretraining.

Across all evaluated SSL methods, we find that deficits applied at the start of pretraining cause larger degradation than when the same deficits are introduced later. On average across methods and deficit durations, early deficits reduce accuracy by about 14 points, compared to 8 points for middle deficits and only 3 points for late deficits. SimCLR is the most vulnerable overall, followed by VICReg and DINO, while MAE is comparatively more robust. While the absolute magnitude of sensitivity varies by method, the trend is consistent: the beginning of pretraining is a critical window where perturbations to the data distribution leave long-lasting effects on learned representations. A similar trend is observed for *ImageNet-pretrained SSL models*, with results provided in Appendix B.

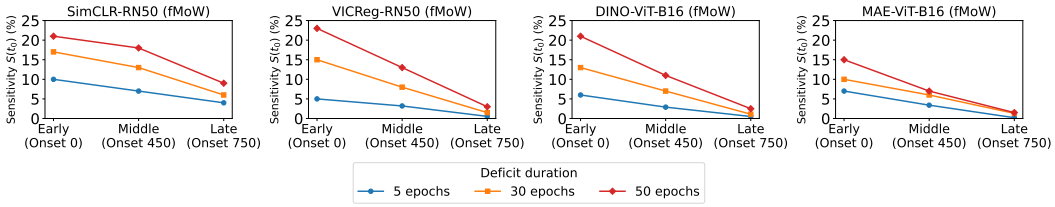



Figure 3: Sensitivity $S(t_0)$ to input perturbations introduced at different stages of SSL pretraining on fMoW. Each curve shows the effect of applying noise of varying duration at different phases (early, middle, late). Higher values indicate stronger lasting degradation in downstream accuracy relative to a clean baseline. Models are fine-tuned on fMoW-train and evaluated on held-out fMoW-val.

### 3.3 Probe 2: Tracking Fisher Information on Pretext Objectives

Perturbation experiments reveal whether temporary interventions have lasting effects, but they do not explain *why* sensitivity varies across pretraining. To provide an analytical perspective, we study the evolution of Fisher Information (FI) (Fisher, 1925) during SSL pretraining. FI measures how strongly parameters influence the predictive distribution and has been used to quantify parameter importance (Amari, 1998; Kirkpatrick et al., 2017; Achille et al., 2018). FI is also a positive semi-definite approximation of the Hessian, capturing local curvature of the loss landscape (Martens, 2020). Unlike the full Hessian, the trace of FI can be estimated efficiently during pretraining.

In SSL, supervision is provided by pretext tasks that define targets $y$ based directly on the input (Balestriero & LeCun, 2024). For instance, in contrastive learning, $y$ specifies positive and negative pairs from augmentations of $x$, while in masked image modeling, $y$ denotes masked input regions to be reconstructed. More generally, the model's approximation of these supervisory signals can be represented as a conditional distribution $p_\theta(y|x)$ defined by the model parameters $\theta$, which governs training (Alshammari et al., 2025). From this perspective, FI computed on $p_\theta(y|x)$ quantifies parameter sensitivity to the supervisory signal from pretext tasks. This follows prior work linking FI to network plasticity (Kirkpatrick et al., 2017; Achille et al., 2018; Lewandowski et al., 2023): increasing FI corresponds to phases of heightened plasticity, while stabilization of FI reflects consolidation. These phases indicate when representations are most malleable and when they begin to resist change, with implications for transferability (Jastrzebski et al., 2021; Berariu et al., 2021).

Consider a model with parameters $\theta \in \mathbb{R}^d$, trained on inputs $x \sim \hat{p}(x)$ where $\hat{p}(x)$ is the empirical distribution of $\mathcal{D}$. To quantify local sensitivity, we consider an infinitesimal perturbation of the parameters, $\theta' = \theta + \delta\theta$. The effect of this perturbation is measured by the Kullback–Leibler (KL) divergence between $p_{\theta'}(y|x)$ and $p_\theta(y|x)$. A second-order Taylor expansion gives

$$\mathbb{E}_{x\sim\hat{p}(x)}\,\mathrm{KL}(p_\theta(y|x)\,\|\,p_{\theta'}(y|x)) = \tfrac{1}{2}\,\delta\theta^\top F\,\delta\theta \;+\; o(\|\delta\theta\|^2), \tag{2}$$

where the Fisher Information Matrix (FIM) is

$$F := \mathbb{E}_{x\sim\hat{p}(x)}\,\mathbb{E}_{y\sim p_\theta(y|x)}\left[\nabla_\theta \log p_\theta(y|x)\,\nabla_\theta \log p_\theta(y|x)^\top\right]. \tag{3}$$

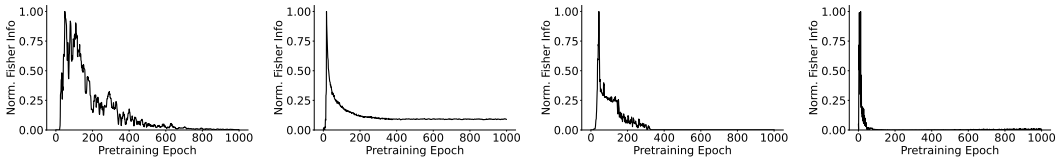
The matrix $F$ characterizes how perturbations to the parameters $\theta$ influence the model's predictive distribution. Parameter-space directions with large eigenvalues of $F$ correspond to high sensitivity, whereas directions with small eigenvalues can be altered with minimal impact on model behavior.

Since computing the full FIM is intractable, we use its trace as a scalar measure of total sensitivity:

$$\mathrm{tr}(F) = \mathbb{E}_{x\sim\hat{p}(x)}\,\mathbb{E}_{y\sim p_\theta(y|x)}\left[\|\nabla_\theta \log p_\theta(y|x)\|^2\right]. \tag{4}$$

The trace of $F$ is the expected squared norm of the score function. In practice, we approximate $\mathrm{tr}(F)$ using gradients of the self-supervised loss with respect to $\theta$, which correspond to gradients of $\log p_\theta(y|x)$ under the pretext task.

**Plasticity rises, peaks, and stabilizes in SSL pretraining.** Figure 4 shows Fisher Information (FI) trajectories during SSL pretraining, providing a quantitative view of how plasticity evolves over time. Across methods, FI rises early, peaks, and then declines before stabilizing. For example, VICReg exhibits an FI peak around epoch 50 followed by stabilization around epoch 350. For MAE, FI rises sharply until about epoch 50, then declines and stabilizes around epoch 150.




(a) SimCLR-RN50-fMoW (b) VICReg-RN50-fMoW (c) DINO-ViT-B16-fMoW (d) MAE-ViT-B16-fMoW

Figure 4: Fisher Information dynamics during SSL pretraining on fMoW.

We define the *critical period* (CP) as the sequence of epochs before FI stabilizes. During this phase, the model is highly plastic and representations remain malleable. Once FI stabilizes, the CP is considered closed: representations commit to existing knowledge and become less sensitive to new information as task-irrelevant variability is discarded. These dynamics align with the perturbation experiments in Figure 3. Deficits introduced during the early phase, while the network was still in its CP, had lasting effects on representation quality. In contrast, deficits introduced after the CP produced only minor effects because the model had already consolidated and become less responsive to change. The decline of FI therefore captures a temporal asymmetry in SSL pretraining and provides an indicator of when the CP is open or closed. *ImageNet-based* FI results are also provided in Appendix B, showing a similar rise-peak-stabilization trend. This consistency indicates that the observed FI behavior is not a dataset-specific artifact.

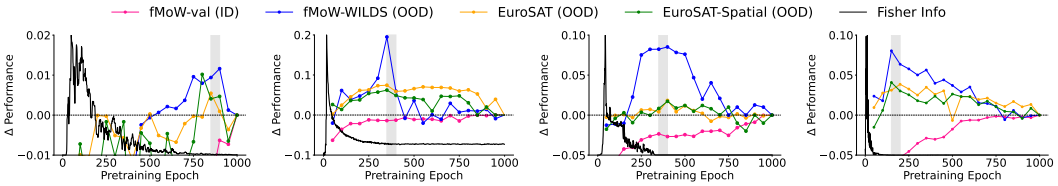
## 4 Critical Periods as a Guide for Efficient and Transferable SSL

In the previous section, our analyses revealed that self-supervised learning (SSL) also exhibits critical periods (CP). Here, we investigate how CP dynamics relate to downstream transferability and propose two simple yet effective CP-guided interventions for efficient and transferable SSL.

### 4.1 Connecting Critical Periods with Downstream Transferability

To test whether critical periods (CP) relate to transferability trade-offs (§2), we align Fisher Information (FI) trajectories with downstream performance across pretraining epochs. Figure 5 shows a consistent pattern across SSL methods: out-of-domain (OOD) transferability peaks near the point where FI stabilizes (grey shading marks CP closure), then declines and does not recover, even as in-domain (ID) accuracy continues to rise. *ImageNet-based* connections are reported in Appendix B, which exhibit a similar alignment between FI stabilization and the peak in OOD transfer.

**The Overspecialization Phase.** We define the divergence between rising ID and declining OOD performance as the onset of an *overspecialization phase*. After CP closure, representations continue to specialize on the pretext distribution by discarding variability deemed irrelevant for the pretext task. While this pruning benefits ID performance, it also discards information that is useful for OOD transfer, leading to a divergence between the two (Figure 1). This indicates that CP closure provides a *sweet spot* where representations are sufficiently learned but not yet overfitted to the pretext task.




(a) SimCLR-RN50-fMoW (b) VICReg-RN50-fMoW (c) DINO-ViT-B16-fMoW (d) MAE-ViT-B16-fMoW

Figure 5: Relation between Fisher Information (FI) dynamics and downstream transferability. FI trajectories (black) are aligned with downstream performance (colored lines) across checkpoints.

**Fisher Information (FI) dynamics help explain delayed trade-offs.** As noted in Section 2, SimCLR's transferability trade-off emerges later than in other SSL methods. FI trajectories show that SimCLR's critical period (CP) closes much later, delaying overspecialization. This aligns with prior findings that contrastive losses converge slowly (Shah et al., 2022; Tong et al., 2023). A key distinction is that SimCLR (Chen et al., 2020) is the only method among those we study whose objective depends on both positive and negative pairs. Since the gradients of the contrastive loss depend on each sample's position relative to all other negatives, the optimization objective shifts as the minibatch changes (Chen et al., 2022). This forces the model to repeatedly reorganize the global structure of the representation space to keep positives aligned while pushing all other samples apart. We conjecture that SimCLR's delayed CP-closure reflects this repeated global reshaping.

In contrast, VICReg (Bardes et al., 2021), DINO (Caron et al., 2021), and MAE (He et al., 2022) do not rely on negatives. VICReg regulates per-batch variance and covariance, DINO enforces alignment with a slowly updated teacher, and MAE reconstructs masked parts of a single image. These objectives do not require maintaining relationships to all other samples in the batch, unlike SimCLR (Chen et al., 2022), which may partly account for the faster CP closure we observe.

### 4.2 Critical Period-guided Checkpoint Selection (CPCS)

Selecting the right SSL checkpoint is non-trivial, as earlier checkpoints risk underdeveloped representations while later ones overspecialize to the pretext task. The finding that OOD transferability peaks near the end of the critical period (CP) suggests a practical strategy. Rather than defaulting to the conventional final checkpoint, we propose *Critical Period-guided Checkpoint Selection (CPCS)*, which leverages Fisher Information (FI) dynamics to identify CP closure checkpoints.

CPCS requires no extra cost post-pretraining and provides a label-free signal for selecting a checkpoint at CP closure, which our results show coincides with peak OOD transferability, just before overspecialization. In practice, one can (i) monitor FI trace across epochs, (ii) identify the point where the FI curve first enters a stable plateau, and (iii) select the nearest saved checkpoint for downstream transfer. This rule-of-thumb narrows the search space: CP closure offers a safe choice when OOD transfer is important, while continuing pretraining beyond CP closure remains beneficial when ID accuracy is the priority. Additional results are provided in Appendix C.

### 4.3 Critical Period-guided Self-Distillation (CPSD)

While intermediate checkpoints exhibit stronger out-of-domain (OOD) generalization, later checkpoints continue to achieve higher in-domain (ID) accuracy. This trade-off reflects complementary properties: *CP checkpoints* capture broadly transferable features, while *post-CP checkpoints* specialize toward pretext-specific signals, increasing alignment with the pretraining distribution.

To mitigate this loss of OOD transferability, we propose *CP-guided self-distillation* (CPSD), a light post-pretraining strategy that reuses existing checkpoints. The idea is simple: use the CP checkpoint as a teacher for the intermediate layers of the post-CP checkpoint (student). During downstream fine-tuning, we optimize the task loss $L_{\text{task}}$ (e.g., cross-entropy for classification) together with a distillation loss applied only to intermediate layers $\mathcal{L}$. The overall objective is

$$L = L_{\text{task}} + \lambda \sum_{l \in \mathcal{L}} \| f_l^{\text{student}} - f_l^{\text{teacher}} \|_2^2, \tag{5}$$

where $\lambda$ is a hyperparameter and the last layers are optimized only with $L_{\text{task}}$.

**The intuition** comes from our layer-wise probing analysis (Figure 6): CP checkpoints achieve consistently stronger OOD performance across the network, with the gap largest in the early layers. In contrast, post-CP checkpoints provide higher ID accuracy in the later layers, reflecting the benefits of extended pretraining when downstream ID tasks are aligned with the pretraining data distribution. Crucially, the ID gains in the later layers build on early layers specialized to the pretraining distribution, which helps explain why stronger ID performance comes at the cost of reduced OOD generalization. CPSD addresses this trade-off by restoring the early layers of the final checkpoint toward their CP state to recover OOD robustness, while preserving the late layers of the post-CP checkpoint to maintain the ID strength gained through extended pretraining.

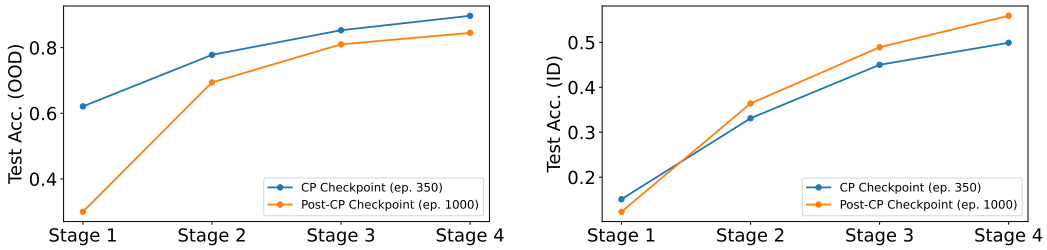



Figure 6: Layer-wise probing for OOD transfer (left: VICReg pretrained on fMoW, evaluated on EuroSAT-Spatial) and for ID performance (right: VICReg pretrained and evaluated on fMoW).

**Results.** Table 1 reports top-1 reports downstream classification accuracy. The final pretrained checkpoint achieves the strongest ID performance but suffers OOD degradation. The CP-guided checkpoint reverses this pattern: it trades a small amount of ID accuracy for a large OOD gain. CP-guided self-distillation combines these strengths: distilling early-layer features from the CP checkpoint into the final checkpoint yields a balanced overall performance. Distilling all layers, however, performs worse than early-layer distillation. We suspect this is because pulling the entire network toward the CP checkpoint restores generality from the intermediate model but also overwrites the useful ID-specific refinements learned in later layers. Distillation settings and additional results are provided in Appendix D.

Table 1: Downstream classification results after pretraining with VICReg-RN50 on fMoW. Results are averaged over 3 runs. Results style: **best**, second best.

| Model | fMoW-val (ID) | fMoW-WILDS (OOD) | EuroSAT (OOD) | EuroSAT-Spatial (OOD) |
|---|---|---|---|---|
| Final ckpt (ep. 1000) | **0.621** ± 0.021 | 0.241 ± 0.034 | 0.864 ± 0.017 | 0.851 ± 0.028 |
| CP-guided ckpt (ep. 350) | 0.610 ± 0.025 | 0.430 ± 0.031 | 0.931 ± 0.013 | 0.912 ± 0.022 |
| CP-guided self-distill | 0.617 ± 0.018 | **0.445** ± 0.029 | **0.954** ± 0.011 | **0.925** ± 0.019 |
| CP-guided self-distill (all layers) | 0.611 ± 0.019 | 0.421 ± 0.023 | 0.929 ± 0.049 | 0.908 ± 0.012 |

## 5 Discussion & Related Work

In this work, we studied a simple yet underexplored question: *how long should we pretrain self-supervised learning (SSL) models?* Contrary to the prevailing heuristic that longer pretraining translates to better downstream performance (Chen et al., 2020; He et al., 2022), we find that the answer is more nuanced. Surprisingly, earlier checkpoints achieve stronger out-of-domain (OOD) transfer than later ones, while the latter improve in-domain (ID) performance. The transferability trade-off across pretraining duration indicates that SSL undergoes a phase transition, akin to *critical periods*.

**Critical early learning phases.** Originating in biology, critical periods refer to windows of heightened plasticity during which neural circuits are particularly sensitive to early experience (Kandel et al., 2000; Hensch, 2004; Knudsen, 2004). A similar effect has been reported in artificial neural networks: changes in the early training phase shape the final representation, whereas changes later have limited impact. Input perturbations applied early permanently reduce generalization, while the same perturbations applied later are recoverable (Achille et al., 2018; Kleinman et al., 2024;

Altıntaş et al., 2025). Moreover, regularization methods (weight decay or data augmentation) only have large effects when applied early in training (Golatkar et al., 2019; Liu et al., 2020; Kalra & Barkeshli, 2023). Conceptually, critical periods mark a transition from a high-plasticity stage, where representations are rapidly formed, to a consolidation stage, where representations stabilize and task-irrelevant information is discarded (Shwartz-Ziv & Tishby, 2017; Achille et al., 2018).

**Exploring critical periods in SSL.** Whether SSL exhibits critical periods (CP) similar to supervised learning, and how these phases affect downstream transfer, remains unexplored. Building on recent calls for a temporal perspective on SSL (Simon et al., 2023; Reizinger et al., 2025), we investigate the emergence of CP in SSL and their impact on transferability. Our results reveal that SSL pretraining undergoes structured phases: early epochs exhibit high plasticity, while later epochs consolidate the model into patterns dictated by the pretraining setup. Beyond plasticity and consolidation, we identify a subsequent phase of *overspecialization* that has not been characterized before. During overspecialization, OOD generalization declines, indicating that representations become increasingly bound to pretraining source data and pretext task. This phased learning dynamics elucidate when representations are broadly transferable, complementing prior work that investigated SSL transferability only after full pretraining (Ericsson et al., 2021a;b).

**Several implications follow.** SSL is often targeted as a pathway to task-agnostic representations (Qiang et al., 2024; Reizinger et al., 2025). This has fueled the rise of foundation models, whose general-purpose representations transfer across tasks and domains (Bommasani, 2021). From this perspective, SSL pretraining defines a distribution over parameters that serves as a prior for all possible downstream tasks. This prior is only useful to the extent that it supports adaptation, yet SSL is not devoid of specialization. Even without labels, every pretext objective imposes implicit supervisory signals (Balestriero & LeCun, 2024; Wang et al., 2024), shaping the invariances and biases the model encodes. Since downstream tasks are unknown at pretraining time, SSL has no guidance for distinguishing task-relevant from task-irrelevant variation (Kleinman et al., 2021), so models may capture nuisances alongside useful features (Xiao et al., 2020; Robinson et al., 2021; Wang et al., 2022; Bandara et al., 2023; Rabin et al., 2024; Qiang et al., 2025). With extended pretraining, the prior increasingly aligns with the pretext task, reducing network plasticity. This tension is salient for foundation models, whose utility depends on SSL producing broadly adaptable priors.

**Limitations & Future Work.** Our analysis centers on vision models and covers two SSL families: discriminative SSL (contrastive: SimCLR (Chen et al., 2020); non-contrastive: VICReg (Bardes et al., 2021), DINO (Caron et al., 2021)) and generative SSL (MAE (He et al., 2022)). Whether similar phase-like behavior occur in language or multimodal settings, or beyond our SSL settings (e.g., JEPA (Assran et al., 2023)), is an open question.

Another promising direction is the integration of label-free representation quality metrics such as RankMe (Garrido et al., 2023) and LiDAR (Thilak et al., 2023) into CP analysis. RankMe assesses the effective rank of the feature covariance, and LiDAR assesses discriminative directions that distinguish one image's features from another's. This information helps identify when structured representations begin to form during pretraining, and combining these perspectives with CP analysis may reveal a more complete picture of the temporal evolution of SSL representations.

While our exploration is by no means exhaustive, our findings suggest that transferability in self-supervised pretraining may follow a non-monotonic trajectory. Analyses that rely solely on the final asymptotic checkpoint therefore risk missing the stages at which transferability is acquired, altered, or lost. Understanding these transient phases could ultimately be as important as understanding the asymptotic properties of pretrained models.

## REFERENCES

Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical learning periods in deep networks. In *International Conference on Learning Representations*, 2018.

Shaden Alshammari, John Hershey, Axel Feldmann, William T Freeman, and Mark Hamilton. I-con: A unifying framework for representation learning. *arXiv preprint arXiv:2504.16929*, 2025.

Gül Sena Altıntaş, Devin Kwok, Colin Raffel, and David Rolnick. The butterfly effect: Neural network training trajectories are highly sensitive to initial conditions. In *Forty-second International Conference on Machine Learning*, 2025.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Randall Balestriero and Yann LeCun. The birth of self supervised learning: A supervised theory. In *NeurIPS 2024 Workshop: Self-Supervised Learning-Theory and Practice*, 2024.

Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, et al. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023.

Wele Gedara Chaminda Bandara, Celso M De Melo, and Vishal M Patel. Guarding barlow twins against overfitting with mixed samples. *arXiv preprint arXiv:2312.02151*, 2023.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Tudor Berariu, Wojciech Czarnecki, Soham De, Jorg Bornschein, Samuel Smith, Razvan Pascanu, and Claudia Clopath. A study on the plasticity of neural networks. *arXiv preprint arXiv:2106.00042*, 2021.

Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

Rishi Bommasani. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Florian Bordes, Randall Balestriero, Quentin Garrido, Adrien Bardes, and Pascal Vincent. Guillotine regularization: Why removing layers is needed to improve generalization in self-supervised learning. *arXiv preprint arXiv:2206.13378*, 2022.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Changyou Chen, Jianyi Zhang, Yi Xu, Liqun Chen, Jiali Duan, Yiran Chen, Son Tran, Belinda Zeng, and Trishul Chilimbi. Why do we need large batchsizes in contrastive learning? a gradient-bias perspective. *Advances in Neural Information Processing Systems*, 35:33860–33875, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6172–6180, 2018.

Elijah Cole, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge Belongie. When does contrastive visual representation learning work? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14755–14764, 2022.

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

Yann Dubois, Stefano Ermon, Tatsunori B Hashimoto, and Percy S Liang. Improving self-supervised learning by characterizing idealized representations. *Advances in Neural Information Processing Systems*, 35:11279–11296, 2022.

Yann Dubois, Tatsunori Hashimoto, and Percy Liang. Evaluating self-supervised learning via risk decomposition. In *International Conference on Machine Learning*, pp. 8779–8820. PMLR, 2023.

Linus Ericsson, Henry Gouk, and Timothy M Hospedales. Why do self-supervised models transfer? investigating the impact of invariance on downstream tasks. *arXiv preprint arXiv:2111.11398*, 2021a.

Linus Ericsson, Henry Gouk, and Timothy M Hospedales. How well do self-supervised models transfer? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5414–5423, 2021b.

Alex Fang, Simon Kornblith, and Ludwig Schmidt. Does progress on imagenet transfer to real-world datasets? *Advances in Neural Information Processing Systems*, 36:25050–25080, 2023.

Ronald Aylmer Fisher. Theory of statistical estimation. In *Mathematical proceedings of the Cambridge philosophical society*, pp. 700–725. Cambridge University Press, 1925.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International conference on machine learning*, pp. 10929–10974. PMLR, 2023.

Aditya Sharad Golatkar, Alessandro Achille, and Stefano Soatto. Time matters in regularizing deep networks: Weight decay and data augmentation affect early learning dynamics, matter little near convergence. *Advances in Neural Information Processing Systems*, 32, 2019.

Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *Proceedings of the ieee/cvf International Conference on computer vision*, pp. 6391–6400, 2019.

Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9052–9071, 2024.

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.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Takao K Hensch. Critical period regulation. *Annu. Rev. Neurosci.*, 27(1):549–579, 2004.

Stanislaw Jastrzebski, Devansh Arpit, Oliver Astrand, Giancarlo B Kerg, Huan Wang, Caiming Xiong, Richard Socher, Kyunghyun Cho, and Krzysztof J Geras. Catastrophic fisher explosion: Early phase fisher matrix impacts generalization. In *International Conference on Machine Learning*, pp. 4772–4784. PMLR, 2021.

Dayal Singh Kalra and Maissam Barkeshli. Phase diagram of early training dynamics in deep neural networks: effect of the learning rate, depth, and width. *Advances in Neural Information Processing Systems*, 36:51621–51662, 2023.

Eric R Kandel, James H Schwartz, Thomas M Jessell, Steven Siegelbaum, A James Hudspeth, Sarah Mack, et al. *Principles of neural science*, volume 4. McGraw-hill New York, 2000.

Youngeun Kim, Yuhang Li, Hyoungseob Park, Yeshwanth Venkatesha, Anna Hambitzer, and Priyadarshini Panda. Exploring temporal information dynamics in spiking neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 8308–8316, 2023.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Michael Kleinman, Alessandro Achille, Daksh Idnani, and Jonathan Kao. Usable information and evolution of optimal representations during training. In *International Conference on Learning Representations*, 2021.

Michael Kleinman, Alessandro Achille, and Stefano Soatto. Critical learning periods emerge even in deep linear networks. In *The Twelfth International Conference on Learning Representations*, 2024.

Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4347–4355, 2025.

Eric I Knudsen. Sensitive periods in the development of the brain and behavior. *Journal of cognitive neuroscience*, 16(8):1412–1425, 2004.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

Alex Lewandowski, Haruto Tanaka, Dale Schuurmans, and Marlos C Machado. Directions of curvature as an explanation for loss of plasticity. *arXiv preprint arXiv:2312.00246*, 2023.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342, 2020.

Markus Marks, Manuel Knott, Neehar Kondapaneni, Elijah Cole, Thijs Defraeye, Fernando Perez-Cruz, and Pietro Perona. A closer look at benchmarking self-supervised pre-training with image classification. *International Journal of Computer Vision*, pp. 1–13, 2025.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Zhuo Ouyang, Kaiwen Hu, Qi Zhang, Yifei Wang, and Yisen Wang. Projection head is secretly an information bottleneck. *arXiv preprint arXiv:2503.00507*, 2025.

Utku Ozbulak, Hyun Jung Lee, Beril Boga, Esla Timothy Anzaku, Homin Park, Arnout Van Messem, Wesley De Neve, and Joris Vankerschaver. Know your self-supervised learning: A survey on image-based generative and discriminative training. *arXiv preprint arXiv:2305.13689*, 2023.

Wenwen Qiang, Jingyao Wang, Changwen Zheng, Hui Xiong, and Gang Hua. On the universality of self-supervised learning. *arXiv preprint arXiv:2405.01053*, 2024.

Wenwen Qiang, Jingyao Wang, Zeen Song, Jiangmeng Li, and Changwen Zheng. On the out-of-distribution generalization of self-supervised learning. *arXiv preprint arXiv:2505.16675*, 2025.

Zachary Rabin, Jim Davis, Benjamin Lewis, and Matthew Scherreik. Overfitting in contrastive learning? *arXiv preprint arXiv:2407.15863*, 2024.

Patrik Reizinger, Randall Balestriero, David Klindt, and Wieland Brendel. Position: An empirically grounded identifiability theory will accelerate self supervised learning research. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. Can contrastive learning avoid shortcut solutions? *Advances in neural information processing systems*, 34:4974–4986, 2021.

Esther Rolf, Konstantin Klemmer, Caleb Robinson, and Hannah Kerner. Position: Mission critical–satellite data is a distinct modality in machine learning. In *Forty-first International Conference on Machine Learning*, 2024.

Anshul Shah, Suvrit Sra, Rama Chellappa, and Anoop Cherian. Max-margin contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 8220–8230, 2022.

Ravid Shwartz Ziv and Yann LeCun. To compress or not to compress—self-supervised learning and information theory: A review. *Entropy*, 26(3):252, 2024.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

James B Simon, Maksis Knutins, Liu Ziyin, Daniel Geisz, Abraham J Fetterman, and Joshua Albrecht. On the stepwise nature of self-supervised learning. In *International Conference on Machine Learning*, pp. 31852–31876. PMLR, 2023.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. *arXiv preprint arXiv:2502.02013*, 2025.

Adam J Stewart, Caleb Robinson, Isaac A Corley, Anthony Ortiz, Juan M Lavista Ferres, and Arindam Banerjee. Torchgeo: deep learning with geospatial data. In *Proceedings of the 30th international conference on advances in geographic information systems*, pp. 1–12, 2022.

Vimal Thilak, Chen Huang, Omid Saremi, Laurent Dinh, Hanlin Goh, Preetum Nakkiran, Joshua M Susskind, and Etai Littwin. Lidar: Sensing linear probing performance in joint embedding ssl architectures. *arXiv preprint arXiv:2312.04000*, 2023.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Shengbang Tong, Yubei Chen, Yi Ma, and Yann Lecun. Emp-ssl: Towards self-supervised learning in one training epoch. *arXiv preprint arXiv:2304.03977*, 2023.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Caltech-ucsd birds-200-2011 (cub-200-2011). Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Haoqing Wang, Xun Guo, Zhi-Hong Deng, and Yan Lu. Rethinking minimal sufficient representation in contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16041–16050, 2022.

Wenhao Wang, Muhammad Ahmad Kaleem, Adam Dziedzic, Michael Backes, Nicolas Papernot, and Franziska Boenisch. Memorization in self-supervised learning improves downstream generalization. *arXiv preprint arXiv:2401.12233*, 2024.

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

Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. *arXiv preprint arXiv:2008.05659*, 2020.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

Nanxuan Zhao, Zhirong Wu, Rynson WH Lau, and Stephen Lin. What makes instance discrimination good for transfer learning? *arXiv preprint arXiv:2006.06606*, 2020.

# A Experimental Details from Section 2

## A.1 Pretraining Setup

**Datasets.** We use two large-scale datasets for SSL pretraining, chosen as the standard benchmarks for natural and satellite imagery. **ImageNet-1K** (Deng et al., 2009) contains 1.28M training images and 50K validation images across 1,000 object categories of natural images. **fMoW-RGB** (Christie et al., 2018) contains over 1M satellite images spanning 63 building and land-use categories.

**Hyperparameters.** We follow the official setups from SimCLR (Chen et al., 2020), VICReg (Bardes et al., 2021), DINO (Caron et al., 2021), and MAE (He et al., 2022). All experiments use a batch size of 1024 with cosine learning rate schedules. For SimCLR, we use SGD with momentum 0.9, learning rate 0.3, and weight decay $1\times10^{-4}$ (replacing the LARS optimizer used in the original large-batch setup). For VICReg, we use LARS with momentum 0.9, learning rate 0.2, and weight decay $1\times10^{-6}$. For DINO, we use AdamW with learning rate $5\times10^{-4}$ and weight decay scheduled from 0.04 to 0.4, with student temperature $\tau_s = 0.1$, teacher temperature $\tau_t = 0.07$, and teacher momentum increasing from 0.996 to 1.0. For MAE, we use AdamW with learning rate $1.5\times10^{-4}$, weight decay 0.05, and masking ratio 0.75. All experiments were run on 8× NVIDIA A100-SXM4-80GB GPUs with mixed-precision training.

**Augmentations.** For all SSL methods, we use the official augmentation pipelines. SimCLR (Chen et al., 2020): contrastive loss with temperature $\tau = 0.1$; random resized cropping (scale 0.08–1.0), random horizontal flipping (p=0.5), color jitter (brightness=0.8, contrast=0.8, saturation=0.8, hue=0.2; p=0.8), random grayscale (p=0.2), and Gaussian blur (p=0.5). VICReg (Bardes et al., 2021): invariance, variance, and covariance loss weights (25, 25, 1) with the same augmentation pipeline as SimCLR. DINO (Caron et al., 2021): multi-crop with two global crops (224px, scale 0.4–1.0) and eight local crops (96px, scale 0.05–0.4). Augmentations include color jitter (0.8,0.8,0.8,0.2; p=0.8), Gaussian blur (p=1.0 on global crops, p=0.5 on local crops), and solarization (p=0.2 on one global crop). Student and teacher temperatures are $\tau_s = 0.1$, $\tau_t = 0.07$, and teacher momentum increases from 0.996 to 1.0. MAE (He et al., 2022): masking ratio 0.75 with random cropping (224px) and random horizontal flipping (p=0.5).

## A.2 Downstream Setup

**Tasks.** Our primary downstream task is *image classification*, and we follow the definition of *in-domain (ID)* and *out-of-domain (OOD)* transfer from (Marks et al., 2025). After pre-training, models are fine-tuned on a labeled downstream dataset and evaluated on its held-out split. ID transfer is where the downstream dataset matches the pre-training distribution. OOD transfer is where the downstream dataset differs from the pre-training distribution.

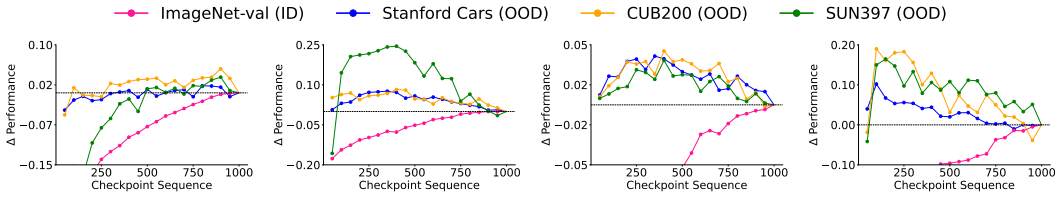
**Datasets.** For models pretrained on *ImageNet-1K*, ID transfer is measured by fine-tuning on the ImageNet-1K training set and evaluating on the validation set. OOD transfer reflects shifts away from this source. We consider three OOD datasets. *Stanford Cars* (Krause et al., 2013) for fine-grained object recognition, *CUB-200* (Wah et al., 2011) for fine-grained natural categories, and *SUN397* (Xiao et al., 2010) for scene recognition.

For models pretrained on *fMoW-RGB*, ID transfer is measured on the held-out fMoW-RGB validation split. We consider three OOD datasets. *fMoW-WILDS* (Koh et al., 2021) partitions the same dataset by geographic region, inducing spatial domain shifts. *EuroSAT* (Helber et al., 2019) uses Sentinel-2 imagery with a different sensing modality. *EuroSAT-Spatial* (Stewart et al., 2022) uses the same EuroSAT data but splits data along longitude to induce spatial distribution shifts.

**Evaluation.** We follow standard evaluation protocols in SSL (Balestriero et al., 2023). Models are evaluated using top-1 classification accuracy on the held-out validation or test split of each downstream dataset. We use official splits when available and adopt a standard $80/20$ split otherwise. For ResNet backbones, we fine-tune with SGD (momentum 0.9) for 50 epochs using batch size 256, a cosine learning rate schedule with base LR 0.05, and weight decay $10^{-4}$. For ViT backbones, we fine-tune with AdamW for 50 epochs using batch size 256, a cosine learning rate schedule with base LR $5\times10^{-4}$, and weight decay 0.05.

## B IMAGENET-BASED RESULTS

**Transferability Trade-Off in SSL (§2).** Figure 7 shows that the trade-off between in-domain (ID) and out-of-domain (OOD) performance is present in ImageNet-pretrained models.




(a) SimCLR-RN50-IN (b) VICReg-RN50-IN (c) DINO-ViT-B16-IN (d) MAE-ViT-B16-IN

Figure 7: The x-axis shows a sequence of checkpoints (every 50 epochs), and the y-axis shows downstream performance relative to the final checkpoints.

**Representation Sensitivity to Perturbations (§3.2).** Figure 8 shows that in ImageNet-pretrained models, sensitivity to input perturbations is highest during the early phase of pretraining.

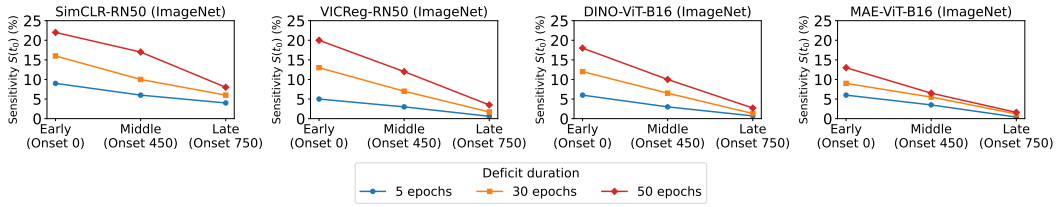
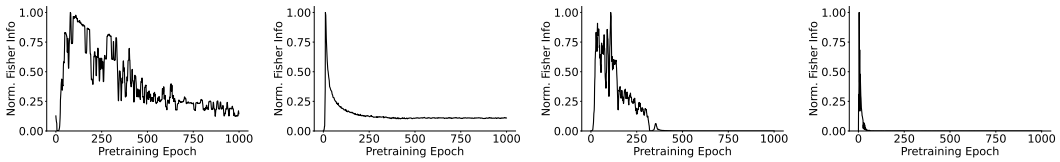



Figure 8: Sensitivity to input perturbations during pretraining reflected in downstream performance.

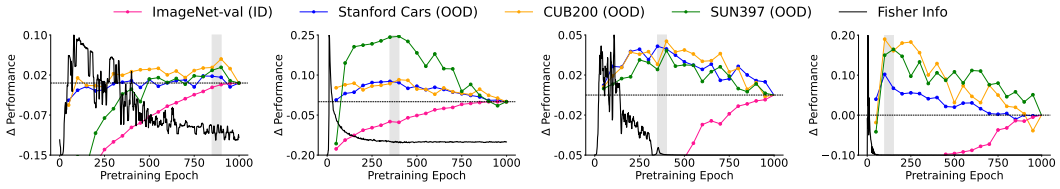
**Fisher Information Dynamics in SSL (§3.3).** Figure 9 shows that ImageNet-pretrained models follow a pattern where Fisher Information rises early, peaks, and then declines before stabilizing.




(a) SimCLR-RN50-IN (b) VICReg-RN50-IN (c) DINO-ViT-B16-IN (d) MAE-ViT-B16-IN

Figure 9: Fisher Information dynamics during SSL pretraining on ImageNet-1K.

**Critical Periods and Transferability (§4.1).** Figure 10 shows that in ImageNet-based models, OOD peaks near CP closure before declining, while ID continues to rise (overspecialization).




(a) SimCLR-RN50-IN (b) VICReg-RN50-IN (c) DINO-ViT-B16-IN (d) MAE-ViT-B16-IN

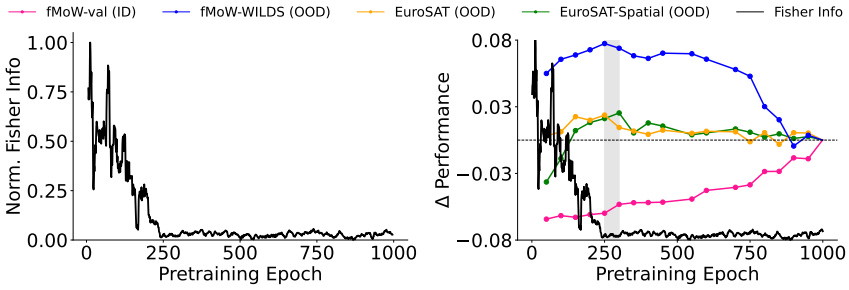
Figure 10: FI trajectories (black) are aligned with downstream performance (colored lines).

## C Additional: Critical Period-guided Checkpoint Selection

**Additional SSL Method (DINOv2).** We also use DINOv2 (Oquab et al., 2023), a state-of-the-art SSL method that combines image-level and patch-level objectives. Incorporating DINOv2 allows us to test whether critical periods also emerge in this setting and whether our findings still hold.

**Setting.** We pretrain DINOv2 with an EMA teacher and a student trained using AdamW with a cosine warmup schedule. Weight decay is annealed from $0.04$ to $0.4$ with a cosine schedule. The loss combines an image-level DINO objective (class tokens), a patch-level iBOT objective (masked tokens, applied only on the student), and a KoLeo regularizer ($\lambda = 0.1$). Teacher momentum is scheduled from $0.992$ to $1.0$. We use a ViT-S/16 backbone with DropPath $0.1$, LayerScale $10^{-5}$, and standard DINO multi-crop augmentations. For evaluation, we follow Appendix A.2.

**Results.** Figure 11 shows the Fisher Information (FI) dynamics during pretraining and their relation to downstream transfer for DINOv2. In Figure 11 (left), FI rises sharply at the start of pretraining but rapidly decays and stabilizes by epoch 200-250, indicating closure of the critical period. As shown in Figure 11 (right), using checkpoints around this closure yields peak OOD performance, while later training leads to overspecialization: OOD transfer declines and does not recover.




(a) Fisher Info during pretraining (b) Critical Period and Transferability

Figure 11: Analysis of DINOv2-ViT-S16 pretrained on fMoW-RGB.

## D Additional: Critical Period-guided Self-Distillation

**Setting.** For distillation, we use the *first residual stage* (`layer1`) of ResNet-50 (He et al., 2016) and the first three Transformer blocks of ViT-B16 (aligning [CLS] tokens). Models are trained with AdamW for $100$ epochs (batch $256$, base LR $10^{-4}$ with cosine schedule, weight decay $0.05$), using $\lambda = 0.5$ (Eq. 5).

Table 2: Classification results across methods pretrained on fMoW. Results style: **best**, second best.

| Method | fMoW-val (ID) | fMoW-WILDS (OOD) | EuroSAT (OOD) | EuroSAT-Spatial (OOD) |
|---|---|---|---|---|
| **SimCLR-RN50** | | | | |
| Final ckpt (ep. 1000) | 0.614 ± 0.019 | 0.401 ± 0.027 | 0.958 ± 0.012 | 0.879 ± 0.021 |
| CP-guided ckpt (ep. 850) | 0.593 ± 0.022 | 0.418 ± 0.025 | 0.960 ± 0.011 | 0.887 ± 0.018 |
| CP-guided self-distill | **0.616** ± 0.017 | **0.425** ± 0.023 | **0.971** ± 0.010 | **0.914** ± 0.016 |
| **VICReg-RN50** | | | | |
| Final ckpt (ep. 1000) | **0.621** ± 0.021 | 0.241 ± 0.034 | 0.864 ± 0.017 | 0.851 ± 0.028 |
| CP-guided ckpt (ep. 350) | 0.610 ± 0.025 | 0.430 ± 0.031 | 0.931 ± 0.013 | 0.912 ± 0.022 |
| CP-guided self-distill | 0.617 ± 0.018 | **0.445** ± 0.029 | **0.954** ± 0.011 | **0.925** ± 0.019 |
| **DINO-ViT-B16** | | | | |
| Final ckpt (ep. 1000) | **0.707** ± 0.018 | 0.364 ± 0.027 | 0.957 ± 0.013 | 0.887 ± 0.021 |
| CP-guided ckpt (ep. 400) | 0.684 ± 0.020 | 0.434 ± 0.026 | 0.975 ± 0.012 | 0.908 ± 0.019 |
| CP-guided self-distill | 0.692 ± 0.019 | **0.440** ± 0.024 | **0.979** ± 0.011 | **0.915** ± 0.018 |
| **MAE-ViT-B16** | | | | |
| Final ckpt (ep. 1000) | **0.679** ± 0.020 | 0.307 ± 0.028 | 0.915 ± 0.014 | 0.825 ± 0.023 |
| CP-guided ckpt (ep. 150) | 0.609 ± 0.018 | 0.387 ± 0.027 | 0.945 ± 0.012 | 0.858 ± 0.020 |
| CP-guided self-distill | 0.638 ± 0.019 | **0.388** ± 0.026 | **0.947** ± 0.011 | **0.861** ± 0.018 |

## E Ablation Study

Table 3: **Effect of distillation weight $\lambda$ on downstream transfer.** We use VICReg-RN50-fMoW and distill from the conv2_x block group (He et al., 2016). Results are averaged over 3 runs. Performance is stable across $\lambda$ values, with no single setting dominating across all OOD tasks.

| $\lambda$ | fMoW-val (ID) | fMoW-WILDS (OOD) | EuroSAT (OOD) | EuroSAT-Spatial (OOD) |
|---|---|---|---|---|
| 0.25 | 0.613 $\pm$ 0.017 | 0.447 $\pm$ 0.028 | 0.948 $\pm$ 0.014 | 0.928 $\pm$ 0.020 |
| 0.50 | 0.617 $\pm$ 0.018 | 0.445 $\pm$ 0.029 | 0.954 $\pm$ 0.011 | 0.929 $\pm$ 0.019 |
| 0.75 | 0.615 $\pm$ 0.022 | 0.438 $\pm$ 0.025 | 0.956 $\pm$ 0.018 | 0.924 $\pm$ 0.027 |
| 1.00 | 0.609 $\pm$ 0.020 | 0.431 $\pm$ 0.032 | 0.942 $\pm$ 0.020 | 0.912 $\pm$ 0.021 |

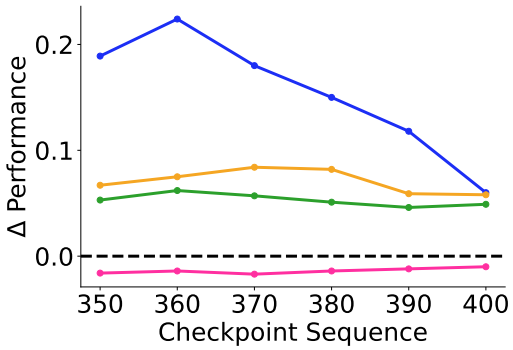



Figure 12: **Finer-resolution checkpoint evaluation near critical period (CP) closure.** Previously, the Fisher Information (FI) trajectory for VICReg-RN50-fMoW indicated a critical period (CP) closure region around epochs 350 to 400 (Fig. 4b). Since FI stabilization does not necessarily coincide with checkpoints saved every 50 epochs, checkpoints within this interval are further evaluated every 10 epochs. The plot shows $\Delta$ performance relative to the final 1000-epoch checkpoint, and OOD accuracy peaks near 360 to 370. **(Takeaway)** This indicates that qualitatively using FI stabilization is a way to narrow the search region for OOD transfer, and that selecting the nearest saved checkpoint within this interval is a practical strategy.

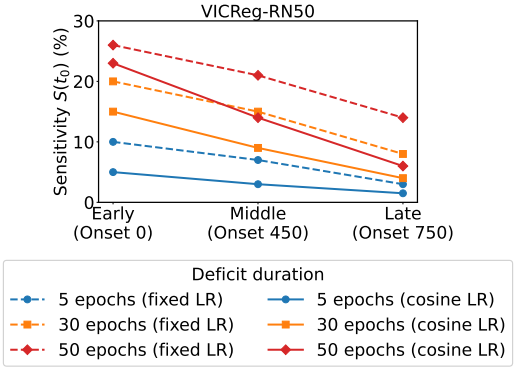
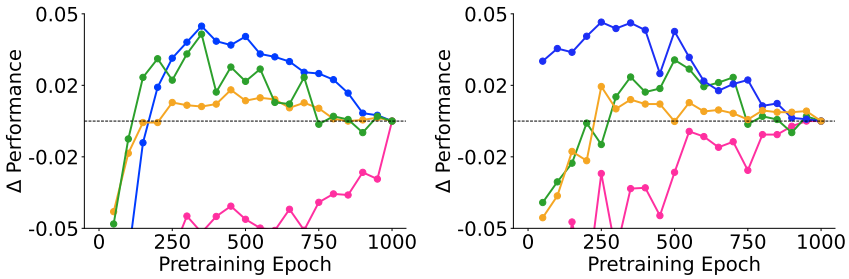



Figure 13: **Effect of learning rate schedule on sensitivity.** Using the same deficit setting as in Sec. 3.2, we measure the sensitivity of VICReg-RN50 pretrained on fMoW-RGB under different learning rate (LR) schedules (fixed vs. cosine). Across all deficit durations, both schedules exhibit similar qualitative behavior: early deficits cause the largest degradation in downstream performance, whereas later deficits have smaller impact. Unlike cosine LR, which anneals, the fixed LR maintains the same step size throughout, resulting in higher sensitivity overall. **(Takeaway)** This shows that the observed temporal sensitivity is not merely an artifact of the learning rate schedule.




(a) Pretrained on fMoW-RGB (b) Pretrained on ImageNet

Figure 14: **Linear probing results for VICReg-RN50.** After pretraining on either fMoW-RGB (left) or ImageNet (right), we freeze the backbone and attach a single linear classifier. We then evaluate checkpoints sampled throughout pretraining on one in-domain (ID) dataset (pink) and three out-of-domain (OOD) datasets (all other colors), and report each checkpoint's change in accuracy relative to the final checkpoint. **(Takeaway)** For both pretraining sources, intermediate checkpoints consistently outperform the final checkpoint on OOD tasks. In contrast, the ID pink curve steadily improves with longer pretraining. This divergence mirrors the transferability trade-off: continued pretraining refines features for the pretraining domain but can reduce their generality for other tasks.

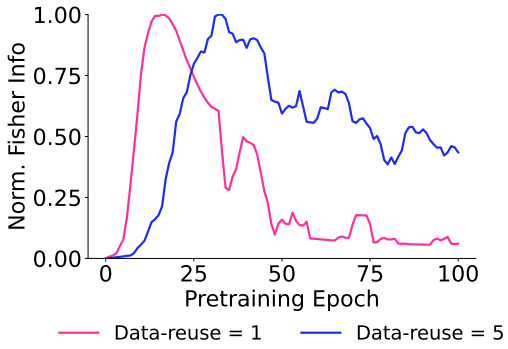
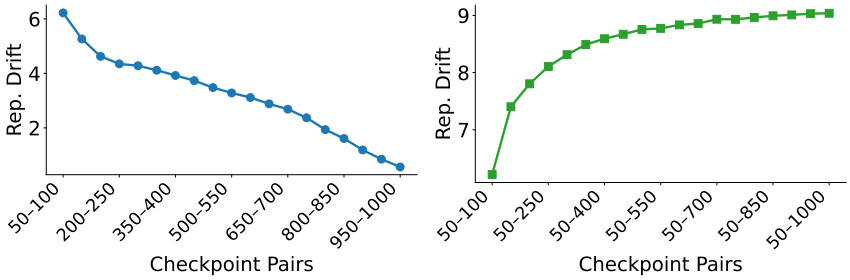
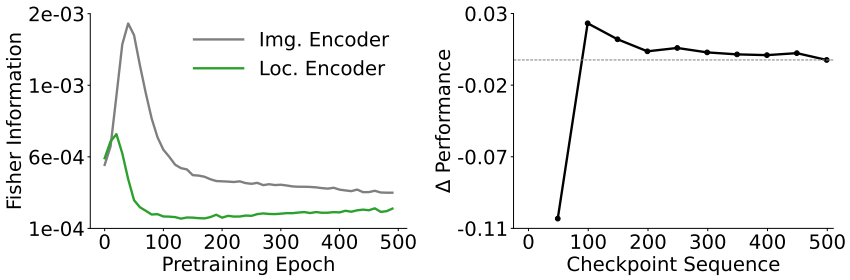



Figure 15: **Toy example on data overfitting via data reuse.** We pretrain a DINO ViT-B/16 encoder on the fMoW dataset while varying how each minibatch is reused during optimization. For a reuse value of $k$, the model processes the same minibatch $k$ times in succession and accumulates the resulting gradients before a single optimizer update. This increases the influence of each minibatch and reduces the diversity of gradient directions encountered per update. Increasing reuse from $k = 1$ to $k = 5$ causes the Fisher Information to rise more rapidly and peak earlier. Using each minibatch only once preserves greater gradient variability and leads to a later critical period closure.




(a) Drift between consecutive checkpoints (b) Drift relative to a fixed checkpoint

Figure 16: **Representation drift analyses for VICReg-RN50-IN.** To compute representation drift, we extract backbone features for a fixed set of held-out validation images (e.g., 2000) at each checkpoint and measure the mean feature-space distance between representations. The left plot reports drift between consecutive checkpoints (epoch t vs. epoch t+50), which is large early in pretraining and steadily decreases as pretraining continues. The right plot measures drift relative to an anchored early checkpoint (epoch 50), showing a rapid rise that later plateaus. **(Takeaway)** Together, these trends partially reinforce our observation that early phases exhibit representational plasticity, which may relate to improved downstream adaptation.




(a) Multimodal Fisher Information (b) Relative Downstream Task Performance

Figure 17: **Critical period (CP) analyses for multimodal settings.** To probe whether CP-like behavior extends to multimodal pretraining methods, we analyze a CLIP-style geo-foundation model (SatCLIP (Klemmer et al., 2025)). In all experiments, we follow the original SatCLIP pretraining scheme, using its contrastive objective between an MoCo-based image encoder and a spherical harmonics based location encoder, and we augment this setup by computing Fisher Information (FI) for both encoders throughout pretraining. **(Takeaway)** SatCLIP also exhibits CP-like behavior, with FI peaking early and stabilizing near epoch 100. When linear probing across seven downstream tasks (three classification and four regression tasks, following the original SatCLIP evaluation setup), the mean performance at the intermediate checkpoint at epoch 100 outperforms the final 500 epoch model. These findings indicate that CP effects may not be limited to unimodal vision settings.

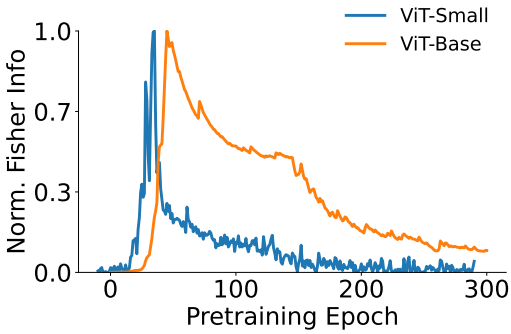



Figure 18: **Fisher Information (FI) trends for different model sizes.** As an initial check, we compare ViT-Small and ViT-Base under identical DINO pretraining. Both models exhibit a CP-like rise and stabilization in FI, but the larger ViT-Base shows a slower CP closure. This aligns with the expectation that larger models, due to their greater representational capacity, explore a broader feature space before consolidating their representations. **(Takeaway)** These results suggest that model size can shift the timing of CP dynamics due to the increased capacity of larger models to sustain exploration before settling into a stable representation.

# F Information Bottleneck Perspective on SSL Learning Phases

In the Information Bottleneck (IB) formulation (Tishby et al., 2000; Shwartz-Ziv & Tishby, 2017), a representation $Z$ balances two quantities: the information retained about the input $X$ and the information provided about a task variable $Y$. This trade-off is captured by the classical objective

$$\min_{p(z|x)} I(X;Z) \; - \; \beta\, I(Z;Y). \tag{6}$$

Prior work on the information dynamics of deep networks (Shwartz-Ziv & Tishby, 2017; Shwartz Ziv & LeCun, 2024; Ouyang et al., 2025) observes that training often proceeds in two phases: a *fitting phase* in which $I(Z;Y)$ increases, followed by a *compression phase* in which $I(X;Z)$ decreases.

Although self-supervised learning (SSL) does not explicitly optimize this IB objective, the IB perspective offers an intuition that helps explain the learning dynamics we observe. *Our goal is not to claim a formal equivalence, but to use the IB framework as an interpretive lens.*

In SSL, the only task variable $Y$ derives from the pretext objective, which we denote by $Y_{\text{pretext}}$. Examples include the identity of an augmented view, a teacher prediction, or a masked region to reconstruct. Under the IB perspective, optimization increases the information that the representation carries about the pretext task (Tian et al., 2020),

$$I(Z;\, Y_{\text{pretext}}), \tag{7}$$

which corresponds to learning invariances or reconstruction patterns needed for the SSL objective. As pretraining continues, the representation discards input variability that does not help predict $Y_{\text{pretext}}$. This corresponds to a reduction in

$$I(X;\, Z), \tag{8}$$

which can be interpreted as compression in the IB sense (Shwartz-Ziv & Tishby, 2017; Shwartz Ziv & LeCun, 2024).

**(Takeaway)** Since this compression is guided solely by the pretext task, the resulting representation may discard information that is unnecessary for the pretext objective but *useful for downstream tasks*. This pretext-driven narrowing of $Z$ can help explain, at least in part, the overspecialization effects observed during extended SSL pretraining.