WHEN AND HOW DOES A REINFORCEMENT LEARNING OPTIMIZER HELP OUT-OF-DISTRIBUTION DETECTION? A THEORETICAL STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

Out-of-distribution (OOD) detection in dynamic open-world environments requires a model to continually adapt to evolving data distributions while generalizing to covariate-shifted inputs and rejecting semantic-shifted OOD examples. Most existing OOD detection methods optimize only the current-step objective and do not explicitly account for how post-deployment environment changes affect future OOD behavior. In this paper, we establish a theoretical framework that explains how and when reinforcement learning (RL)-guided update improves dynamic OOD detection. Our RL-guided optimizer explicitly favors updates that reduce the semantic OOD false positive rate over time. We analyze how the value function shapes the updates and affects the training trajectory, and derive a temporal error decomposition that separates the change in evaluation risk into a model-change term, controlled by the optimizer, and an environment-change term, derived by distribution shift. This decomposition yields measurable conditions under which the RL-guided optimizer lowers the semantic-OOD false positive rate relative to GD without degrading classification generalization error. Experimental evidence support better performance of the RL-guided optimizer over a GD optimizer.

1 INTRODUCTION

Consider an object-recognition system deployed on a mobile robot in a warehouse. During commissioning, the system is calibrated on summer imagery, but lighting, seasonal clutter, and equipment upgrades gradually shift the data distribution. An update that improves the current classification loss may simultaneously distort the detector's energy surface and weaken its response to genuinely anomalous inputs. This highlights a fundamental challenge in dynamic OOD detection: improving current-step performance may inadvertently degrade future-domain OOD robustness.

Most OOD detection methods are developed under a static deployment assumption, where the test distribution remains fixed after training (Hendrycks & Gimpel, 2016; Liang et al., 2018; Liu et al., 2020b). Even methods that adapt at test time, such as entropy minimization (Wang et al., 2020) and related test-time adaptation approaches (Niu et al., 2022b), optimize *instantaneous* objectives and do not explicitly account for how parameter updates affect the detector's behavior in subsequent environments. This paper addresses the question: *can we design an optimization framework for dynamic OOD detection that explicitly accounts for the future-domain consequences of each update?*

A natural approach is to incorporate information about future OOD performance into the current update. We formalize this idea through a value function perspective: a value function trained by temporal-difference (TD) learning estimates the cumulative return associated with future semantic-OOD detection performance and provides a differentiable signal for shaping the adaptation trajectory.

Building on this insight, we introduce an *RL-guided optimization framework* that augments standard gradient descent (GD) with a learned value correction. At each adaptation step, the gradient of the value function directly shapes the parameter update, without requiring a separate policy or discrete action space. Unlike general-purpose learned optimizers (Andrychowicz et al., 2016; Metz et al., 2019), the value function is explicitly tied to the semantic-OOD false positive rate (FPR), enabling a direct theoretical connection between optimization dynamics and OOD detection performance.

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.

Theoretical contributions. We provide a theoretical framework that explains how and when the RL-guided update improves dynamic OOD detection. The analysis proceeds in four steps (Sec. 4). *Temporal error decomposition*. We split the change in future-domain evaluation risk into an uncontrollable environment-change term and a model-change term set by the optimizer, isolating what optimization can influence. *Value-gradient decomposition*. We decompose the value correction into an explicit soft-FPR channel and the contributions of the remaining state coordinates, making clear how the state representation shapes the update. *Multi-step trajectory identity*. Over one adaptation window, we separate the accumulated value correction from the drift in subsequent base updates induced by earlier corrections. *Risk-specific effects* (Thm 4.2). Projecting this trajectory decomposition onto the semantic-OOD FPR and covariate-shifted classification-risk gradients yields signed identities, Eq. 3-Eq. 4, together with measurable conditions for reducing semantic-OOD FPR while controlling the corresponding change in classification risk. The analysis is instantiated for a frozen-feature linear head, where both evaluation-risk gradients and the corresponding smoothness constants admit closed-form expressions (Prop 4.3). In addition, we analyze the RL-GD optimization dynamics through loss-gap bounds (Thm 5.3), training-step gradient-energy states (Cor. 5.4), and environment-effective gradient flow (Thm 5.7 and Thm 5.8). *Our contributions are listed below:*

- We formulate dynamic OOD detection as a sequential adaptation problem and introduce an RL-guided update that augments a shared gradient-based objective with a TD-trained value correction.
- We develop a finite-window theory that characterizes the risk-specific effects of value guidance on future-domain semantic-OOD rejection and covariate-shifted classification.
- We analyze the RL-GD optimization dynamics through loss-gap bounds, training-step gradient-energy states, and environment-effective gradient flow.

2 Related Work

Many OOD detection methods assume *static* distributions, including post-hoc scoring approaches such as maximum softmax probability (Hendrycks & Gimpel, 2016), input perturbation (Liang et al., 2018), Mahalanobis distance (Lee et al., 2018), and energy-based scores (Liu et al., 2020b), as well as training-time strategies like outlier exposure (Hendrycks et al., 2018). Test-time adaptation methods Wang et al. (2020); Niu et al. (2022b) address distribution shift by updating models on unlabeled data (Wang et al., 2020; Niu et al., 2022b), but optimize *instantaneous objectives* without considering long-term OOD behavior. Recent work leverages wild or mixed ID/OOD data for detection and generalization (Hendrycks et al., 2018; Bai et al., 2023), yet typically uses batch-independent updates. Continual learning and online adaptation study stability under sequential shifts (Kirkpatrick et al., 2017; Lopez-Paz & Ranzato, 2017; Farajtabar et al., 2020), while RL and meta-learning approaches learn adaptive update rules (Andrychowicz et al., 2016; Finn et al., 2017), and energy-based theories provide geometric interpretations of OOD detection (Liu et al., 2020b). In contrast, our work provides a unified *temporal perspective*, explicitly modeling how optimization dynamics influence future-domain OOD robustness. Extended discussions are in Appx B.

3 Method

We formalize the dynamic OOD detection setting. At environment t, the detector receives data drawn from a non-stationary distribution $\mathbb{P}_t^{\text{wild}} = (1 - \alpha_t^{\text{cov}} - \alpha_t^{\text{sem}})\mathbb{P}^{\text{ID}} + \alpha_t^{\text{cov}}\mathbb{P}_t^{\text{Cov}} + \alpha_t^{\text{sem}}\mathbb{P}_t^{\text{Sem}}$, where $\mathbb{P}^{\text{ID}}$ is the ID data, $\mathbb{P}_t^{\text{Cov}}$ is covariate-shifted data that remain in the known label space, and $\mathbb{P}_t^{\text{Sem}}$ is semantic OOD data that should be rejected. Within each environment, the detector performs K adaptation updates indexed by k. The GD and RL-guided optimizers use the same base adaptation objective. Semantic-OOD performance is evaluated through the semantic-OOD FPR, while classification performance is evaluated on ID and covariate-shifted data.

Let $f_\theta : \mathcal{X} \to \mathbb{R}^N$ be the classifier and $E_\theta(x) = -\log \sum_{y=1}^{N} \exp\big(f_\theta^{(y)}(x)\big)$ its energy score. Lower energy indicates that a sample is more likely to belong to a known class. Optimizers share base objective $\mathcal{L}_{OW}(\theta) = \mathcal{L}_{ID}(\theta) + \lambda_{\text{Sem}}\mathcal{L}_{OOD}(\theta)$, where $\mathcal{L}_{ID}$ is the cross-entropy on ID data and $\mathcal{L}_{OOD}$ encourages high uncertainty on semantic-OOD.

Optimizing $\mathcal{L}_{OW}(\theta)$ alone does not explicitly account for how the update affects OOD behavior in subsequent environments. Two updates with similar current-step loss can therefore induce different future-domain energy separation. We aim to use temporal OOD information to shape this update.

RL-Guided optimizer. We now describe how the proposed approach is realized as an optimization rule that augments standard gradient descent with a learned, temporally-aware correction.

Pseudo-OOD subset. The RL-guided correction does not require semantic-OOD labels. At step t, we calibrate an energy threshold τ_t using ID reference energies and use this threshold to select a pseudo-semantic-OOD subset $B_{t,k}^{\text{sem}}$ from the available wild batch. This subset serves as a proxy for constructing the differentiable state and reward.

Soft-FPR surrogate. Since the hard FPR is non-differentiable, we use a surrogate $\widehat{\text{FPR}}_{t,k}^{\text{soft}} = \frac{1}{|B_{t,k}^{\text{sem}}|}\sum_{x\in B_{t,k}^{\text{sem}}}\sigma\Big(\frac{\tau_t - E_{\theta_{t,k}}(x)}{\beta}\Big)$, with $\beta > 0$ and $\sigma(\cdot)$ the logistic sigmoid. This provides a differentiable approximation of FPR used in the reward and in computing gradients through the state.

State. A state is defined as $s_{t,k} = \Big[\mu_E^{\text{id}},\ \mu_E^{\text{sem}},\ \text{Var}(E^{\text{id}}),\ \text{Var}(E^{\text{sem}}),\ \widehat{\text{FPR}}_{t,k}^{\text{soft}},\ \widehat{\text{Acc}}_{t,k}\Big]$. where μ_E^{id} and μ_E^{sem} denote the mean energy on ID and pseudo-semantic-OOD; the variance terms summarize the corresponding energy distributions; $\widehat{\text{FPR}}_{t,k}^{\text{soft}}$ measures the current pseudo-OOD rejection behavior; and $\widehat{\text{Acc}}_{t,k}$ measures classification performance on the labeled ID reference data. This low-dimensional state captures the key statistics for OOD separation.

Reward. The reward is $r_{t,k} = -\widehat{\text{FPR}}_{t,k}^{\text{soft}}$. Larger rewards correspond to lower pseudo-OOD FPR.

Value learning. A value function $V_\phi(s_{t,k})$ estimates the return of future rewards along the adaptation sequence. For an observed transition $(s_{t,k}, r_{t,k}, s_{t,k+1})$, we use TD learning, $\mathcal{L}_{\text{TD}}(\phi) = \frac{1}{2}\big[V_\phi(s_{t,k}) - (r_{t,k} + \gamma V_\phi(s_{t,k+1}))\big]^2$. During TD training, we detach the states, rewards, and bootstrap targets from the detector, while holding the critic fixed and backpropagating through the differentiable state coordinates only when computing the value-gradient correction.

Unlike a conventional actor-critic formulation, we do not learn a separate policy over a predefined action space. Instead, the detector update itself serves as the control mechanism, while the value gradient directly shapes the update direction.

Update rule. At environment t and update k with learning rate η, the GD baseline follows

$$\theta_{t,k+1}^{GD} = \theta_{t,k}^{GD} - \eta\nabla_\theta\mathcal{L}_{OW}(\theta_{t,k}^{GD}), \tag{1}$$

The RL-guided optimizer augments the same base update with a learned value correction:

$$\theta_{t,k+1}^{RL} = \theta_{t,k}^{RL} - \eta\nabla_\theta\mathcal{L}_{OW}(\theta_{t,k}^{RL}) + \eta_{RL}\nabla_\theta V_\phi(s_{t,k}). \tag{2}$$

For $\eta > 0$, Eq. 2 update is equivalent to one gradient step on the locally frozen objective $\mathcal{L}_{OW}(\theta_{t,k}^{RL}) - \frac{\eta_{RL}}{\eta}V_\phi(s_{t,k})$. This interpretation shows that the value function modifies the adaptation trajectory by favoring parameter directions associated with higher predicted future return. Sec. 4 analyzes the resulting trajectory through its effects on semantic-OOD rejection and covariate-shifted classification risk. At inference time, only the detector is used; the critic is required only during adaptation.

4 From Value Guidance to Risk-Specific Effects

The RL-guided optimizer modifies GD through correction $\nabla_\theta V_{\phi_{t,k}}(s_{t,k})$. We study how these corrections accumulate over adaptation and how the resulting parameter displacement affects semantic-OOD FPR and known-class classification. Our analysis proceeds in three steps. First, we decompose the value gradient through the state coordinates, separating the contribution of the soft-FPR statistic from the rest. Second, we derive an exact multi-step trajectory identity that separates the accumulated value correction from the base-update drift induced by earlier corrections. Third, we project this displacement onto the two evaluation risks. This yields separate, measurable conditions for reducing semantic-OOD FPR and known-class classification. All proofs of this section are in Appx C.

4.1 From Dynamic Evaluation to an Optimizer Comparison

We first compare the two optimizers over one adaptation window t. Both optimizers receive the same sequence of batches. Denote $d_{t,k} = \theta_{t,k}^{RL} - \theta_{t,k}^{GD}$, and $D_{t,k} = \|d_{t,k}\|$. When both optimizers start from the same model, $d_{t,0} = 0$. We retain $d_{t,0}$ in the analysis so that the analysis also covers settings in which the two trajectories enter an adaptation window with different parameters.

Evaluation risks. For environment t, a fixed τ_t and a held-out semantic-OOD distribution $\mathbb{P}_t^{\text{sem}}$, define soft FPR $\text{FPR}_t^{\text{soft}}(\theta) = \mathbb{E}_{x\sim\mathbb{P}_t^{\text{sem}}}\big[\sigma\big(\frac{\tau_t - E_\theta(x)}{\beta}\big)\big]$ and hard FPR $\text{FPR}(\theta) = \Pr_{x\sim\mathbb{P}_t^{\text{sem}}}[E_\theta(x) \le \tau_t]$. For a held-out distribution $\mathbb{P}_t^{\text{cov}}$ of shifted known-class samples, define the generalization error

$\mathrm{GErr}_t(\theta) = \mathbb{E}_{(x,y)\sim\mathbb{P}_t^{\mathrm{cov}}} \ell_{\mathrm{CE}}(f_\theta(x), y)$. The held-out semantic-OOD distributions are not used to construct the RL-guided update; adaptation uses the available data and the pseudo-OOD statistics.

At the final GD endpoint, write $h_{t,K} = \nabla \mathrm{FPR}_t^{\mathrm{soft}}(\theta_{t,K}^{\mathrm{GD}})$, $q_{t,K} = \nabla \mathrm{GErr}_t(\theta_{t,K}^{\mathrm{GD}})$. The two gradients define distinct local evaluation geometries for measuring the same RL-guided trajectory.

Connection to dynamic evaluation. Generalization error changes from t to $t+1$ is decomposed as

$$\mathrm{GErr}_{t+1}(\theta_{t,K}) - \mathrm{GErr}_t(\theta_{t,0}) = \underbrace{\mathrm{GErr}_{t+1}(\theta_{t,K}) - \mathrm{GErr}_{t+1}(\theta_{t,0})}_{\text{model-change}} + \underbrace{\mathrm{GErr}_{t+1}(\theta_{t,0}) - \mathrm{GErr}_t(\theta_{t,0})}_{\text{environment-change}}.$$

The first term captures the effect of the update, whereas the second term captures the effect of the environment shift with the model held fixed. When the two optimizers share the same $\theta_{t,0}$, the environment-change term is common to both trajectories. The finite-window analysis below therefore focuses on the optimizer-induced parameter displacement and its effect on the evaluation risks.

4.2 How the Value Function Shapes the Update

Value correction is obtained by differentiating the critic through the state. Because the state contains the soft FPR statistics, we can separate this contribution from those of the remaining coordinates.

At $\theta_{t,k}^{\mathrm{RL}}$, denote $p_{t,k} = \nabla_\theta \widehat{\mathrm{FPR}}_{t,k}^{\mathrm{soft}}$, and $c_{t,k} = -\frac{\partial V_\phi}{\partial \widehat{\mathrm{FPR}}^{\mathrm{soft}}}$, and let $b_{t,k} = \sum_{i \neq \mathrm{FPR}} \frac{\partial V_{\phi_{t,k}}}{\partial s^{(i)}} \nabla_\theta s_{t,k}^{(i)}$ collect the contributions of the remaining differentiable state coordinates. The critic parameters ϕ are held fixed when these derivatives are computed. By the chain rule, the correction can be decomposed as $\nabla_\theta V_\phi(s_{t,k}) = -c_{t,k} p_{t,k} + b_{t,k}$. The first term describes the path through the soft FPR coordinate. When $c_{t,k} > 0$, the critic assigns lower value to states with larger soft FPR, so $-c_{t,k}p_{t,k}$ points locally toward lower proxy FPR. The term $b_{t,k}$ collects directions induced by remaining state statistics.

To assess how this correction affects an evaluation risk, we project the correction onto the corresponding evaluation gradient. For $v \in \{h_{t,K}, q_{t,K}\}$, $-\left\langle v, \sum_{k=0}^{K-1} \nabla_\theta V(s_{t,k}) \right\rangle = \sum_{k=0}^{K-1} [c_{t,k}\langle v, p_{t,k}\rangle - \langle v, b_{t,k}\rangle]$. Thus, the effect of the explicit soft-FPR channel depends on both the critic sensitivity $c_{t,k}$ and the alignment between the proxy FPR gradient $p_{t,k}$ and the evaluation-risk gradient v. The remaining state coordinates contribute through the second term. Sec. 4.4 uses this projection to characterize the final risk change after the complete adaptation trajectory.

4.3 Multi-Step Effect on the Update Trajectory

A correction applied at one step also changes the parameters at which later base gradients are evaluated. We therefore compare the complete K-step trajectory. Denote the induced base-update drift as $e_{t,K} = -\eta \sum_{k=0}^{K-1} \left[\nabla \mathcal{L}_{\mathrm{OW}}(\theta_{t,k}^{\mathrm{RL}}) - \nabla \mathcal{L}_{\mathrm{OW}}(\theta_{t,k}^{\mathrm{GD}})\right]$. Subtracting the two update recursions and summing over the K steps gives the exact trajectory identity $d_{t,K} = d_{t,0} + \eta_{\mathrm{RL}} \sum_{k=0}^{K-1} \nabla_\theta V_\phi(s_{t,k}) + e_{t,K}$.

The final RL-GD parameter difference therefore consists of three parts: the inherited difference $d_{t,0}$, the accumulated value correction $\eta_{\mathrm{RL}} \sum_k \nabla_\theta V_\phi(s_{t,k})$, and the induced base-update drift $e_{t,K}$. The drift captures the indirect effect of RL guidance: earlier corrections move the RL-guided trajectory to different parameters, which changes the subsequent base gradients relative to those along GD.

4.4 Risk-Specific Effects of the Value-Guided Trajectory

We now connect the parameter-space trajectory to the two evaluation risks. Let the accumulated value correction be measured against the semantic-OOD and classification risk gradients by $A_{t,K}^{\mathrm{FPR}} = -\left\langle h_{t,K}, \sum_{k=0}^{K-1} \nabla_\theta V_\phi(s_{t,k}) \right\rangle$, and $A_{t,K}^{\mathrm{GErr}} = -\left\langle q_{t,K}, \sum_{k=0}^{K-1} \nabla_\theta V_\phi(s_{t,k}) \right\rangle$. Here, $h_{t,K}$ and $q_{t,K}$ are the gradients of the held-out environment-t evaluation risks defined in Sec. 4.1. The two quantities evaluate the same accumulated correction under different risk geometries. Using the decomposition from Sec. 4.2, $A_{t,K}^{\mathrm{FPR}} = \sum_{k=0}^{K-1} [c_{t,k}\langle h_{t,K}, p_{t,k}\rangle - \langle h_{t,K}, b_{t,k}\rangle]$, $A_{t,K}^{\mathrm{GErr}} = \sum_{k=0}^{K-1} [c_{t,k}\langle q_{t,K}, p_{t,k}\rangle - \langle q_{t,K}, b_{t,k}\rangle]$. For semantic-OOD rejection, the term $c_{t,k}\langle h_{t,K}, p_{t,k}\rangle$ combines two factors: how strongly the critic values a reduction in soft-FPR, and how well that soft-FPR direction aligns with the held-out semantic-OOD risk. The analogous decomposition with $q_{t,K}$ describes how the same correction acts on classification risk.

Assumption 4.1 (Local quadratic regularity)**.** *For each evaluation risk $R \in \{\mathrm{FPR}^{\mathrm{soft}}, \mathrm{GErr}\}$, the first-order expansion between the GD and value-guided endpoints satisfies $\left|R(\theta^{\mathrm{RL}}_{t,K}) - R(\theta^{\mathrm{GD}}_{t,K}) - \left\langle \nabla_\theta R(\theta^{\mathrm{GD}}_{t,K}), d_{t,K} \right\rangle\right| \le \frac{L_{R,t}}{2} D^2_{t,K}$.*

Theorem 4.2 (Finite-window effect of value guidance)**.** *Under Assumption 4.1, the endpoint differences satisfy*

$$\mathrm{FPR}^{\mathrm{soft}}(\theta^{\mathrm{RL}}_{t,K}) - \mathrm{FPR}^{\mathrm{soft}}(\theta^{\mathrm{GD}}_{t,K}) = -\eta_{\mathrm{RL}} A^{\mathrm{FPR}}_{t,K} + \langle h_{t,K}, d_{t,0} + e_{t,K} \rangle + \rho^{\mathrm{FPR}}_{t,K}, \tag{3}$$

$$\mathrm{GErr}(\theta^{\mathrm{RL}}_{t,K}) - \mathrm{GErr}(\theta^{\mathrm{GD}}_{t,K}) = -\eta_{\mathrm{RL}} A^{\mathrm{GErr}}_{t,K} + \langle q_{t,K}, d_{t,0} + e_{t,K} \rangle + \rho^{\mathrm{GErr}}_{t,K}, \tag{4}$$

where $|\rho^{\mathrm{FPR}}_{t,K}| \le \frac{L_{\mathrm{FPR},t}}{2} D^2_{t,K}$, and $|\rho^{\mathrm{GErr}}_{t,K}| \le \frac{L_{\mathrm{GErr},t}}{2} D^2_{t,K}$.

Thm. 4.2 decomposes the endpoint risk difference into three components: the first-order contribution of the accumulated value correction, the change in subsequent base updates caused by following a different trajectory, and a local nonlinear remainder.

For semantic-OOD FPR, the theorem provides the sufficient condition $\eta_{\mathrm{RL}} A^{\mathrm{FPR}}_{t,K} > \langle h_{t,K}, d_{t,0} + e_{t,K} \rangle + \frac{L_{\mathrm{FPR},t}}{2} D^2_{t,K}$, under which $\mathrm{FPR}^{\mathrm{soft}}(\theta^{\mathrm{RL}}_{t,K}) < \mathrm{FPR}^{\mathrm{soft}}(\theta^{\mathrm{GD}}_{t,K})$. Thus, value guidance improves semantic OOD rejection when its accumulated risk-aligned contribution is large enough to dominate the signed base-update drift and the local curvature term.

The classification equation quantifies the effect of the same trajectory on known-class prediction. For any tolerance $\varepsilon \ge 0$, $\eta_{\mathrm{RL}} A^{\mathrm{GErr}}_{t,K} \ge \langle q_{t,K}, d_{t,0} + e_{t,K} \rangle + \frac{L_{\mathrm{GErr},t}}{2} D^2_{t,K} - \varepsilon$ implies $\mathrm{GErr}_t(\theta^{\mathrm{RL}}_{t,K}) \le \mathrm{GErr}_t(\theta^{\mathrm{GD}}_{t,K}) + \varepsilon$. Hence, the decomposition provides a sufficient condition for improved semantic-OOD FPR and a tolerance-based bound on the corresponding change in classification risk.

4.5 Specialization to Frozen-Feature Linear Heads

The analysis in Sec. 4.4 applies to general differentiable detector parameterizations. We now specialize it to a frozen-feature linear-head setting. In this setting, the two evaluation-risk gradients and their local curvature admit explicit forms, making the quantities in Thm. 4.2 directly interpretable.

In this specialization, the adapted parameter is the classifier head itself, so we identify θ with W. Let $f_W(x) = Wh(x)$, $\|h(x)\|_2 \le H$, where the feature extractor $h(\cdot)$ is fixed and only the classifier matrix W is adapted. Let $\pi_W(x) = \mathrm{softmax}(Wh(x))$, and define the energy score as $E_W(x) = -\log \sum_{j=1}^{C} \exp([Wh(x)]_j)$. Its gradient with respect to the classifier head is $\nabla_W E_W(x) = -\pi_W(x) h(x)^\top$.

Semantic-OOD risk geometry. For $z_W(x) = \frac{\tau_t - E_W(x)}{\beta}$, the soft-FPR is $\mathrm{FPR}^{\mathrm{soft}}_t(W) = \mathbb{E}_{x \sim \mathbb{P}^{\mathrm{sem}}_t}[\sigma(z_W(x))]$. $\nabla_W \mathrm{FPR}^{\mathrm{soft}}_t(W) = \frac{1}{\beta} \mathbb{E}_{x \sim \mathbb{P}^{\mathrm{sem}}_t}\left[\sigma(z_W(x))\big(1 - \sigma(z_W(x))\big)\pi_W(x) h(x)^\top\right]$. The factor $\sigma(z_W(x))(1 - \sigma(z_W(x)))$ is largest when the energy of an OOD sample lies near the rejection threshold. Thus, the gradient $h_{t,K}$ appearing in Thm. 4.2 places the largest first-order weight on semantic-OOD examples close to that threshold.

Known-class risk geometry. For shifted known-class data, the classification risk is $\mathrm{GErr}(W) = \mathbb{E}_{(x,y) \sim \mathbb{P}^{\mathrm{cov}}}[\ell_{\mathrm{CE}}(Wh(x), y)]$. Its gradient is $\nabla_W \mathrm{GErr}_t(W) = \mathbb{E}_{(x,y) \sim \mathbb{P}^{\mathrm{cov}}_t}\left[\big(\pi_W(x) - e_y\big) h(x)^\top\right]$, where e_y denotes the one-hot vector of the true class.

These derivations make the two risk geometries in Sec. 4.4 explicit. The semantic OOD gradient is determined by OOD samples near the rejection threshold, whereas the classification gradient is determined by the prediction residuals of known-class samples. Consequently, the alignments $A^{\mathrm{FPR}}_{t,K}$ and $A^{\mathrm{GErr}}_{t,K}$ measure different properties of the same accumulated value correction.

Proposition 4.3 (Linear-head risk regularity)**.** *Under $\|h(x)\|_2 \le H$, the gradients of the two evaluation risks are Lipschitz in W. Valid constants are $L_{\mathrm{FPR}_t} \le \frac{H^2}{6\sqrt{3}\,\beta^2} + \frac{H^2}{4\beta}$, $L_{\mathrm{GErr}_t} \le H^2$. Therefore, the quadratic remainder condition in Assumption 4.1 holds for the frozen-feature linear-head model with finite constants.*

Prop. 4.3 gives an explicit instantiation of the curvature terms appearing in Thm. 4.2. In particular, the sufficient condition for lower semantic-OOD soft FPR becomes $\eta_{\mathrm{RL}} A^{\mathrm{FPR}}_{t,K} > \langle h_{t,K}, d_{t,0} + e_{t,K} \rangle +$

$\frac{1}{2}\left(\frac{H^2}{6\sqrt{3}\,\beta^2}+\frac{H^2}{4\beta}\right)D_{t,K}^2$. Similarly, the classification-risk change is controlled by $\mathrm{GErr}(W_{t,K}^{\mathrm{RL}}) - \mathrm{GErr}(W_{t,K}^{\mathrm{GD}}) \le -\eta_{\mathrm{RL}} A_{t,K}^{\mathrm{GErr}} + \langle q_{t,K}, d_{t,0}+e_{t,K}\rangle + \frac{H^2}{2}D_{t,K}^2$.

Thus, in the frozen-feature linear-head setting, the abstract terms in Thm. 4.2 acquire a concrete interpretation. The semantic OOD alignment is governed by the interaction between the value-guided correction and threshold-local OOD examples. The same parameter displacement also determines changes in generalization error through the prediction residuals of known-class samples.

5 Understanding the Role of Training Regimes

This section analyzes the behavior of the RL optimizer in comparison to the GD optimizer across training steps. A key assumption for this section is that the learning dynamics can be characterized using a second-order Taylor expansion of the loss function during model training. Here we introduce the effect of loss gap between GD vs RL training on a sequence of environments. Proofs in Appx D.

Definition 5.1 (Optimizer loss gap)**.** *For environment t, F_t^{OW} is the loss gap of the optimized parameter under both GD and RL optimizers (i.e. $\theta_{t,*}^{RL}$ and $\theta_{t,*}^{GD}$): $F_t^{OW} := \mathcal{L}_{OW}(\theta_{t,*}^{RL}) - \mathcal{L}_{OW}(\theta_{t,*}^{GD})$.*

Definition 5.2 (Training step-based gradient energy state)**.** *For the environment t, we define the training step-based gradient energy state at optimization step k, denoted by $S_{t,\nabla}^{(k)}$, as follows $S_{t,\nabla}^{(k)} := \mathbb{E}_{x\sim P_{t+1}^{sem}}\left[\nabla_\theta^{(k)} E_{\theta_t}(x)\right] - \mathbb{E}_{x\sim P_t^{sem}}\left[\nabla_\theta^{(k)} E_{\theta_t}(x)\right]$, where $\nabla_\theta^{(k)} E_{\theta_t}(x) \approx E_{\theta_{t,k}}(x) - E_{\theta_{t,k-1}}(x)$.*

The results in this section rely on regularity conditions ensuring smooth gradients, bounded iterates, and slow cross-environment variation; their formal statements and proofs are provided in Appx D.

5.1 RL-GD Loss Gap Bounds

We next analyze the training gap between GD and our RL optimizers. We recall Def. 5.1 and show that the loss gap is bounded by the average gap of RL agent state during training steps.

Theorem 5.3 (Loss gap bound)**.** *Suppose Assumptions D.1 and D.2 hold true. We bound F_t^{OW} by*

$$F_t^{OW} \le \frac{1}{2}\lambda_{max}\big[\eta_{K,r} + \eta_{RL}\sum_{k=1}^{K}\sum_{k'=1}^{k-1}\|s_{t,k'} - s_{t,k'-1}\|^2\big], \tag{5}$$

where η_{RL} is constant and sufficiently small, $s_{t,k}$ is the state of RL agent at time step t and training step k. $\eta_{K,r}$ is a constant in terms of total training steps K and radius r, and λ_{max} is the maximum eigenvalue of the Hessian matrix of loss at $\theta_{t,*}^{GD}$, $\nabla^2\mathcal{L}_{OW}(\theta_{t,*}^{GD})$.

Corollary 5.4 (Training step-based gradient energy state bounds loss gap)**.** *The F_t^{OW} bound in Thm. 5.3 is derived in terms of training step-based gradient energy state in Def. 5.2:*

$$F_t^{OW} \le \frac{1}{2}\lambda_{max}\big[\eta_{K,r} + \eta_{RL}\sum_{k=1}^{K}\sum_{k'=1}^{k-1}\|S_{t,\nabla}^{(k')}\|^2\big], \tag{6}$$

where η_{RL} is constant, $\eta_{K,r}$ is a constant in terms of total training steps K and radius r, and λ_{max} is the maximum eigenvalue of $\nabla^2\mathcal{L}_{OW}(\theta_{t,*}^{GD})$.

5.2 RL Improves the Effective Gradient Flow

Thm. 5.3 develops the loss gap for RL optimizer compared to GD optimizer. We next study environment and training effective gradient flow (EGF) and how our RL-guided optimizer improves the EGF across environment change and during training steps.

Definition 5.5 (Effective gradient flow)**.** *We define two versions of effective gradient flow (EGF):*

• *(Gradient flow (GF) over environment change) measures optimization dynamics and is typically approximated using the norm of the gradients of the loss over time. Define the matrix of $\mathcal{G}_T := [\nabla_\theta\mathcal{L}_G(\theta_1)\ \nabla_\theta\mathcal{L}_G(\theta_2)\ \dots\ \nabla_\theta\mathcal{L}_G(\theta_T)] = [\mathbf{G}_1\mathbf{G}_2\dots\mathbf{G}_T]$ for T environments $\mathcal{E} = \{\mathcal{E}_1,\dots,\mathcal{E}_T\}$. Note that $\mathcal{L}_G = \mathcal{L}_{OW}$ (for GD optimizer) and $\mathcal{L}_G = \mathcal{L}_{OW} + \eta_{RL}V(s)$ (for RL optimizer). The environment-effective gradient flow for $t = 1,\dots,T$ is defined as $EGF_p^{env} := \frac{1}{T}\sum_{t=1}^{T}\|\mathbf{G}_t\|_p$, where $\mathbf{G}_t = \nabla_\theta\mathcal{L}_G(\theta_t)$.*

• *(Gradient flow (GF) over training steps) measures optimization per training update and is typically approximated using the norm of the gradients of the loss over epochs at each time step. Define the matrix of* $\mathcal{G}_K := [\nabla_\theta \mathcal{L}_G(\theta_{t,1})\ \nabla_\theta \mathcal{L}_G(\theta_{t,2})\ \ldots\ \nabla_\theta \mathcal{L}_G(\theta_{t,K})] = [\mathbf{G}_1 \mathbf{G}_2 \ldots \mathbf{G}_K]$ *for* K *training updates. The training-effective gradient flow at time* t *is defined as* $EGF_{t,p}^{train} := \frac{1}{K}\sum_{k=1}^{K} \|\mathbf{G}_k\|_p$, *where* $\mathbf{G}_k = \nabla_\theta \mathcal{L}_G(\theta_{t,k})$.

Consider three consecutive environments $\mathcal{E}_{t-1}, \mathcal{E}_t$, and $\mathcal{E}_{t+1}$. The model f_θ has been trained using RL/GD optimizer, with energy values $E_{\theta_{t-1}}$, E_{θ_t}, and $E_{\theta_{t+1}}$ (resp.). To evaluate the RL and GD optimizers, we introduce two new query-based notions: *(1) Energy generalization transferability.* Is the gap in EGF_p^{env} between RL and GD optimizers indicative of energy generalization transfer across environments? Here, energy generalization transferability refers to evaluating $E_{\theta_{t-1}}$ and E_{θ_t} on semantically shifted environments $\mathcal{E}_t^{sem}$ and $\mathcal{E}_{t+1}^{sem}$, drawn from distributions P_t^{sem} and P_{t+1}^{sem} (resp.). *(2) Semantic-shifted energy gap.* Is the gap in EGF_p^{env} between RL and GD optimizers related to the energy gap induced by semantic shifts across environments? In this case, E_{θ_t} and $E_{\theta_{t+1}}$ are evaluated on the semantic-shifted environments $\mathcal{E}_t^{sem}$ and $\mathcal{E}_{t+1}^{sem}$ with distributions P_t^{sem} and P_{t+1}^{sem} (resp.). Before investigating these queries (Thms 5.8, 5.7) we introduce the definition below.

Definition 5.6. *At time steps* t *and* $t-1$, *we define the environment-transfer generalization energy,* $S_{t\to t+1}$, *by* $\mathbb{E}_{x\sim P_{t+1}^{sem}}[E_{\theta_t}(x)] - \mathbb{E}_{x\sim P_t^{sem}}\left[E_{\theta_{t-1}}(x)\right]$. *And we define the environment-based semantic shifted energy bridge,* $E_{t\to t+1}$, *by* $\mathbb{E}_{x\sim P_{t+1}^{sem}}\left[E_{\theta_{t+1}}(x)\right] - \mathbb{E}_{x\sim P_t^{sem}}[E_{\theta_t}(x)]$.

Theorem 5.7 (Training effective gradient flow gap bound)**.** *Consider training-effective gradient flow* $EGF_{t,p}^{train}$ *under both GD and RL optimizers. Then under Assumptions D.1 andD.2, we have*

$$EGF_{t,2}^{train}(RL) - EGF_{t,2}^{train}(GD) \le L_r + L_{K-RL}\sum_{k=1}^{K}\sum_{k'=1}^{k} \|S_{t,\nabla}^{(k')}\|_2, \tag{7}$$

where L_r *and* L_{K-RL} *are constants and* $L_r = 2\sqrt{L}r$, $L_{K-RL} = \frac{\eta_{RL}\sqrt{L_V}}{K}$. *Here* $S_{t,\nabla}^{(k')}$ *is the training step-based gradient energy state at training step* k' *introduced in Def. 5.2.*

Theorem 5.8 (Environment effective gradient flow gap bound)**.** *Consider environment-effective gradient flow* EGF_p^{env} *under both GD and RL optimizers. Under the Assumptions D.1,D.2, D.4, and D.5, we bound environment-effective gradient flow gap between RL and GD optimizers as follows:*

$$EGF_p^{env}(RL) - EGF_p^{env}(GD) \le L_r + L_{T-RL}\sum_{t=1}^{T}\sum_{\tau=1}^{t}\left(\mathcal{O}(\xi_{\tau\to\tau+1}^{S}) + \mathcal{O}(\xi_{\tau-1\to\tau}^{E})\right), \tag{8}$$

for $p = 2$. *In (Eq. 8),* L_r *and* L_{T-RL} *are constants and* $L_{T-RL} = \frac{\eta_{RL}\sqrt{L_V}}{T}$, *for constant* L_V *and* η_{RL}. $\xi_{r\to r+1}^{S}$ *and* $\xi_{r-1\to r}^{E}$ *are environment-transfer generalization energy and environment-based semantic shifted energy rates in Def. 5.6.*

Discussion. Thm. 5.8 provides the central theoretical foundation for the gradient flow behavior of our RL-guided optimization method. Specifically, it shows that the RL-based correction enhances gradient flow effectiveness compared to standard gradient descent. This improvement arises from the aggregated contributions of environment-transfer generalization energy drift and environment-specific semantic shift energy drift across all environments.

6 Experiments

We build online OOD streams from six benchmarks, each environments with a fixed set of in-distribution classes and a disjoint held-out set that serves as semantic OOD throughout. Core50 (Lomonaco & Maltoni, 2017), iCubWorld (Pasquale et al., 2015), and OpenLORIS (She et al., 2020) drift gradually between environments, DomainNet (Peng et al., 2018) changes abruptly, and CLEAR-10 and CLEAR-100 (Lin et al., 2022) are nearly saturated under frozen features, and Terra Incognita (Beery et al., 2018) and Century-of-Portraits (Ginosar et al., 2019) are reported in the supplementary material. As in Sec. 4.5, only a linear head on frozen features is adapted, for $K = 30$ inner steps per environment. We compare the RL-guided update (Eq. 2) with GD on the same objective (Eq. 1) and with five test-time adaptation (TTA) methods on the same head: Tent (Wang et al., 2021), EATA (Niu et al., 2022a), energy fine-tuning (Liu et al., 2020a), SAR (Niu et al., 2023), and CoTTA (Wang et al., 2022). Features come from five frozen encoders, CLIP ViT-B/16 and

ViT-L/14 (Radford et al., 2021), SigLIP (Zhai et al., 2023), SigLIP2 (Tschannen et al., 2025), and ALIGN (Jia et al., 2021) optionally with a MiniCPM-V (Yao et al., 2024) caption embedding. We report the mean and standard deviation over 5 seeds with paired bootstrap tests. Stream construction and the protocol are in the supplementary material.

6.1 RL Lowers Semantic-OOD False Positives

The Eq. 3 of Thm. 4.2 writes the FPR difference between the two optimizers as the accumulated value correction $-\mathrm{RL}A^{\mathrm{FPR}}_{t,K}$ plus the base-update drift and a curvature remainder, and gives the condition under which RL attains the lower FPR. On the three gradual streams RL has the lowest FPR@95 of all seven methods (Tab. 1), 2.0, 1.9, and 0.3 points below GD on Core50, iCubWorld, and OpenLORIS. The TTA baselines optimize prediction confidence rather than the energy separation targeted by $\mathcal{L}_{OW}$, and all of them stay above GD. The gain holds for all five encoders, significantly ($p < 0.05$) for 5, 3, and 2 of them on the three gradual streams, and it costs no accuracy. Per-encoder results and accuracies are in the supplementary material.

Table 1: FPR@95 (%, lower is better) and RL AUROC, mean ± std over 5 seeds, multimodal CLIP-B/16. The best FPR in each row is in bold, with ties marked together.

Dataset	Tent	EATA	en-FT	SAR	CoTTA	GD	RL	RL AUROC
Gradual drift								
Core50	76.1 ± 1.0	70.0 ± 1.2	65.6 ± 0.9	70.8 ± 1.0	67.4 ± 1.0	62.7 ± 0.5	**60.7 ± 3.0**	0.912 ± 0.008
iCubWorld	88.6 ± 0.6	88.2 ± 0.6	87.8 ± 0.5	88.3 ± 0.6	88.1 ± 0.6	87.6 ± 0.6	**85.7 ± 1.3**	0.628 ± 0.025
OpenLORIS	57.7 ± 0.7	51.5 ± 0.8	50.9 ± 0.7	52.4 ± 0.8	49.6 ± 1.0	49.5 ± 0.7	**49.2 ± 2.9**	0.896 ± 0.011
Abrupt drift								
DomainNet	55.9 ± 1.0	55.6 ± 0.8	55.0 ± 1.0	55.6 ± 0.9	55.1 ± 0.8	**54.4 ± 1.2**	54.6 ± 1.0	0.916 ± 0.001
Saturated								
CLEAR-10	3.8 ± 0.4	3.6 ± 0.3	3.4 ± 0.3	3.6 ± 0.3	3.4 ± 0.3	**3.2 ± 0.3**	**3.2 ± 0.2**	0.988 ± 0.000
CLEAR-100	1.0 ± 0.1	1.0 ± 0.1	**0.7 ± 0.0**	1.0 ± 0.1	**0.7 ± 0.0**	**0.7 ± 0.0**	**0.7 ± 0.1**	0.993 ± 0.000

6.2 Gradual vs. Abrupt Drift

The condition in Thm. 4.2 requires the accumulated correction $\eta_{\mathrm{RL}}A^{\mathrm{FPR}}_{t,K}$ to dominate the inherited difference and base-update drift $\langle h_{t,K}, d_{t,0} + e_{t,K}\rangle$ together with the curvature term. Nothing forces this to hold when environments change abruptly, because corrections accumulated earlier are carried over through $d_{t,0}$ while the evaluation geometry $h_{t,K}$ changes. Fig. 1(a) matches this prediction. Averaged over encoders, RL improves on GD by 2.6, 0.9, and 0.7 points on the gradual streams, and not at all on DomainNet, which is not harder (AUROC 0.916 versus 0.912 on Core50) but changes in large steps. On the saturated CLEAR streams RL matches GD, so the correction does no harm.

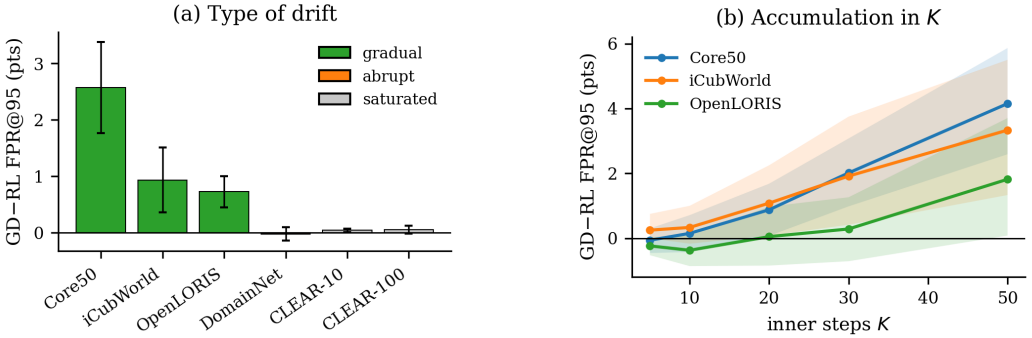


Figure 1: (a) FPR@95 improvement of RL over GD, mean ± std over five encoders. (b) GD−RL FPR@95 gap as a function of the number of inner steps K on the gradual streams (multimodal CLIP-B/16), with the 95% bootstrap interval shaded.

6.3 Covariate Shift

The wild distribution contains covariate-shifted samples $\mathbb{P}^{\mathrm{Cov}}_t$, on which Eq. 4 of Thm. 4.2 measures the effect of the same value-guided trajectory on GErr. For this, we replace half of the known-class samples of every environment, in both the adaptation and the test split, with copies corrupted by Gaussian noise at ImageNet-C severity s_t (Hendrycks & Dietterich, 2019), keeping their labels and leaving the semantic-OOD classes untouched. We run this on Core50, iCubWorld, OpenLORIS, and DomainNet. A gradual schedule raises s_t from 1 to 5 in near-equal steps, and an abrupt schedule jumps from 1 to 5 halfway through the stream. Clean and corrupted known samples both count as ID for FPR@95, and covariate accuracy is measured on the corrupted ones.

Results. The value correction improves generalization under covariate shift. On Core50, it raises accuracy on the corrupted samples by 2.4 points under both schedules and lowers the generalization error from 0.608 to 0.592, and maintains this trend across all five encoders. The gain is smaller on iCubWorld, OpenLORIS, and DomainNet, matching the FPR ordering gain in Tab. 1. The accuracy for the clean, known samples does not change, within 0.33 points in each run. Eq. 4 attributes this to an accumulated correction that is also aligned with the classification-risk gradient $q_{t,K}$ Counting the corrupted samples as in-distribution raise the FPR of both optimizers. GD FPR on Core50 increases from 62.7 to 88.7, and the difference between GD and RL FPR stays within 1.3 points in either direction. RL is slightly worse on Core50 and slightly better on iCubWorld. The two schedules give similar results, so the pace of the shift does not change the outcome.

Table 2: Covariate shift with visual CLIP-B/16, mean $\pm$ std over 5 seeds. FPR@95 counts clean and corrupted known samples as in-distribution, and covariate accuracy is measured on the corrupted samples. All values are percentages.

	Gradual schedule				Abrupt schedule			
Dataset	FPR GD	FPR RL	Cov. acc. GD	Cov. acc. RL	FPR GD	FPR RL	Cov. acc. GD	Cov. acc. RL
Core50	88.7 ± 1.2	89.8 ± 0.8	57.8 ± 0.9	60.2 ± 1.5	86.6 ± 1.2	86.8 ± 0.7	53.7 ± 0.7	56.1 ± 1.2
iCubWorld	96.9 ± 0.3	96.1 ± 1.6	48.7 ± 0.9	49.5 ± 1.5	96.2 ± 0.6	94.9 ± 1.7	52.1 ± 1.1	53.2 ± 1.0
OpenLORIS	94.3 ± 0.7	94.4 ± 0.8	57.9 ± 0.6	58.2 ± 0.4	87.1 ± 1.0	87.1 ± 1.4	53.8 ± 0.3	54.1 ± 0.3
DomainNet	67.4 ± 1.1	67.4 ± 1.0	88.0 ± 0.4	88.1 ± 0.4	67.6 ± 0.6	67.6 ± 0.7	86.5 ± 0.4	86.6 ± 0.3

6.4 Mechanism, Dynamics, and the Value Function

The alignment $A^{\mathrm{FPR}}_{t,K}$ in Eq. 3 sums one contribution per inner step, and the FPR gap grows with K (Fig. 1(b)), from -0.1 to 4.2 points on Core50 and from 0.3 to 3.5 on iCubWorld between $K = 5$ and $K = 50$. The intermediate quantities of the analysis behave as predicted on the gradual streams, reported in the supplementary material. RL raises the mean semantic-OOD energy above GD in 90% to 95% of environments, the direction implied by the soft-FPR channel of Sec. 4.2, and this gap nearly vanishes on DomainNet and CLEAR. The critic sensitivity $c_{t,k} = -\partial V_\phi / \partial \widehat{\mathrm{FPR}}^{\mathrm{soft}}$ of Sec. 4.2 is positive in 80% to 90% of environments, the sign under which that channel points toward lower proxy FPR. The effective gradient flow (Definition 5.5) stays within the bounds of Thms 5.7 and 5.8, and RL forgets significantly less than GD ($p \leq 0.044$). Given a sweep of λ_{Sem} (Fig. 2), RL stays below GD at every setting on Core50 and iCubWorld. Its best setting beats the best setting of GD by 1.2 and 1.7 points ($p = 0.008$ and $p = 0.018$). The supplementary material reports further ablations. Caption features are not needed, since the gain is larger than the features alone. A frozen head is worse than both optimizers on every stream. When the encoder is fine-tuned as well, outside the head-only setting of Sec. 4.5, RL no longer improves on GD.

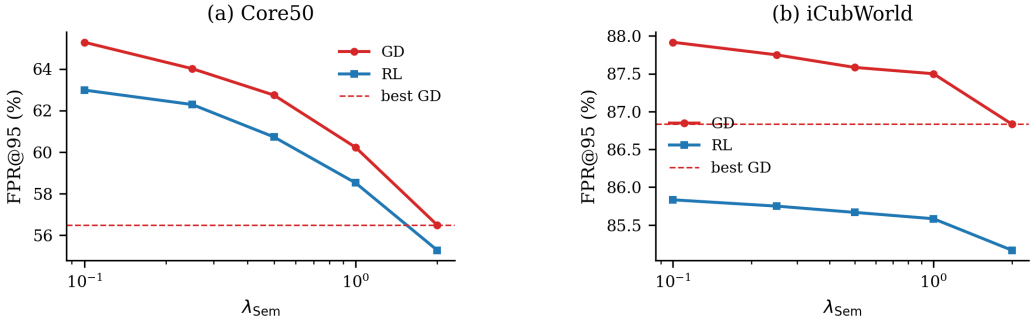


Figure 2: FPR@95 of GD (red) and RL (blue) as the weight λ_{Sem} is swept on (a) Core50 and (b) iCubWorld (multimodal CLIP-B/16), with the best setting of GD dashed.

7 Conclusion

We presented an RL-guided optimization framework for OOD detection under dynamic distribution shifts. By augmenting a shared gradient-based objective with a TD-trained value correction, the method shapes the adaptation trajectory toward long-term OOD rejection. Our finite-window theory decomposes the effect of value guidance into the accumulated correction, induced base-update drift, and local curvature. It provides measurable conditions for reducing semantic-OOD FPR while controlling the change in covariate-shifted classification risk. Experiments show consistent improvements under gradual drift, consistent with the proposed risk-alignment condition.

8 AI DISCLOSURE

In the multimodal experiments, MiniCPM-V was used to generate one caption for each image. These captions were used as synthetic textual annotations: they were encoded by the corresponding text encoder and concatenated with visual features. We have not used generative AI tools for help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, and interpret results are not applicable to this work. Additionally, we used generative AI tools for edit a research paper to improve readability. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. Learning to learn by gradient descent by gradient descent. *Advances in neural information processing systems*, 29, 2016.

Haoyue Bai, Gregory Canal, Xuefeng Du, Jeongyeol Kwon, Robert D Nowak, and Yixuan Li. Feed two birds with one scone: Exploiting wild data for both out-of-distribution generalization and detection. In *International Conference on Machine Learning*, pp. 1454–1471. PMLR, 2023.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

David Duvenaud, Jackson Wang, Jorn Jacobsen, Kevin Swersky, Mohammad Norouzi, and Will Grathwohl. Your classifier is secretly an energy based model and you should treat it like one. *ICLR 2020*, 2020.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *International conference on artificial intelligence and statistics*, pp. 3762–3773. PMLR, 2020.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Shiry Ginosar, Kate Rakelly, Sarah Sachs, Brian Yin, Crystal Lee, Philipp Krahenbuhl, and Alexei A. Efros. A century of portraits: A visual historical record of american high school yearbooks, 2019. URL https://arxiv.org/abs/1511.02575.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *ArXiv*, abs/1807.01697, 2019. URL https://api.semanticscholar.org/CorpusID:56657912.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jia21b.html.

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.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

Zhiqiu Lin, Jia Shi, Deepak Pathak, and Deva Ramanan. The clear benchmark: Continual learning on real-world imagery. *ArXiv*, abs/2201.06289, 2022. URL https://api.semanticscholar.org/CorpusID:244906195.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21464–21475. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f5496252609c43eb8a3d147ab9b9c006-Paper.pdf.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020b.

Vincenzo Lomonaco and Davide Maltoni. Core50: a new dataset and benchmark for continuous object recognition. In Sergey Levine, Vincent Vanhoucke, and Ken Goldberg (eds.), *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pp. 17–26. PMLR, 13–15 Nov 2017. URL https://proceedings.mlr.press/v78/lomonaco17a.html.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

Luke Metz, Niru Maheswaranathan, Jeremy Nixon, Daniel Freeman, and Jascha Sohl-Dickstein. Understanding and correcting pathologies in the training of learned optimizers. In *International Conference on Machine Learning*, pp. 4556–4565. PMLR, 2019.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16888–16905. PMLR, 17–23 Jul 2022a. URL https://proceedings.mlr.press/v162/niu22a.html.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pp. 16888–16905. PMLR, 2022b.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world, 2023. URL https://arxiv.org/abs/2302.12400.

Giulia Pasquale, Carlo Ciliberto, Francesca Odone, Lorenzo Rosasco, and Lorenzo Natale. Real-world object recognition with off-the-shelf deep conv nets: How many objects can icub learn?, 2015. URL https://arxiv.org/abs/1504.03154.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1406–1415, 2018. URL https://api.semanticscholar.org/CorpusID:54458071.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Qi She, Fan Feng, Xinyue Hao, Qihan Yang, Chuanlin Lan, Vincenzo Lomonaco, Xuesong Shi, Zhengwei Wang, Yao Guo, Yimin Zhang, Fei Qiao, and Rosa H. M. Chan. OpenLORIS-Object: A robotic vision dataset and benchmark for lifelong deep learning. In *2020 International Conference on Robotics and Automation (ICRA)*, pp. 4767–4773, 2020.

Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in neural information processing systems*, 34:144–157, 2021.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pp. 9229–9248. PMLR, 2020.

Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in neural information processing systems*, 33:11839–11852, 2020.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim M. Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier H'enaff, Jeremiah Harmsen, Andreas Steiner, and Xiao-Qi Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *ArXiv*, abs/2502.14786, 2025. URL https://api.semanticscholar.org/CorpusID:276482166.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization, 2021. URL https://arxiv.org/abs/2006.10726.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation, 2022. URL https://arxiv.org/abs/2203.13591.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Jun-Bo Cui, Hongji Zhu, Tianchi Cai, Hao-Yu Li, Weilin Zhao, Zhihui He, Qi-An Chen, Hua Zhou, Zhensheng Zou, Hao-Ye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A gpt-4v level mllm on your phone. *ArXiv*, abs/2408.01800, 2024. URL https://api.semanticscholar.org/CorpusID:271709626.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, October 2023.

APPENDIX

A ALGORITHM

Algorithm 1 summarizes the overall training procedure. The outer loop indexes the evolving environment, while the inner loop performs detector adaptation within each environment. At each outer step t, we first calibrate the threshold τ_t from the current ID batch. We then sample labeled ID data and unlabeled wild data, construct the pseudo semantic-OOD partition using the current detector, and compute the OOD-aware objective. The state is estimated, the value function is updated, and finally the detector parameters are updated using the RL-guided rule.

Algorithm 1 RL-guided dynamic OOD adaptation

Initialize detector parameters θ_0 and value parameters ϕ
for $t = 0, \ldots, T-1$ **do**
  Receive data stream from current environment P_t^{wild}
  Sample a reference ID batch and calibrate τ_t from a fixed percentile of its energy distribution
  Set $\theta_{t,0} \leftarrow \theta_t$
  for $k = 0, \ldots, K-1$ **do**
    Sample a batch of labeled ID data and unlabeled wild data from the current environment
    Use the current energy detector $E_{\theta_{t,k}}(\cdot)$ and threshold τ_t to form a pseudo semantic-OOD subset and a residual wild subset
    Compute the OOD-aware loss $\mathcal{L}_{OW}(\theta_{t,k})$
    Estimate the state $s_{t,k}$ from current batch statistics
    Compute the reward $r_{t,k}$
    Update value parameters ϕ by minimizing the squared TD error.
    Update detector parameters using Eq. 2
  end for
  Set $\theta_{t+1} \leftarrow \theta_{t,K}$
end for

We update the value network before the detector within each inner step so that the detector update uses the most recent estimate of the value function while avoiding a same-step circular dependency between ϕ and θ. Relative to standard gradient descent, the additional computation in our method comes from evaluating $\nabla_\theta V_\phi(s_{t,k})$, which requires one forward-backward pass through the lightweight value network and differentiation through the state statistics. Since $s_{t,k}$ is low-dimensional and the value network is shallow, this overhead is modest compared with the detector update itself.

B EXTENDED RELATED WORK

Static OOD detection. A large body of work studies OOD detection under a fixed deployment distribution. Post-hoc scoring methods assign anomaly scores using maximum softmax probability (Hendrycks & Gimpel, 2016), input perturbation (Liang et al., 2018), Mahalanobis distance (Lee et al., 2018), and energy-based scores (Liu et al., 2020a). Additional improvements exploit feature clipping (Sun et al., 2021) and contrastive learning (Tack et al., 2020). Training-time approaches further improve separation between ID and OOD samples through outlier exposure (Hendrycks et al., 2018). These methods are designed for *static* settings and do not address how OOD detectors should evolve under non-stationary environments. In contrast, our work studies how *optimization dynamics* affect OOD robustness over time.

Test-time adaptation. Test-time adaptation (TTA) methods update model parameters using unlabeled test data at deployment. Tent (Wang et al., 2021) minimizes prediction entropy through batch normalization, while EATA (Niu et al., 2022a) improves stability via sample selection and regularization. Test-time training adds a self-supervised objective at deployment (Sun et al., 2020), SAR (Niu et al., 2023) selects reliable samples and flattens the entropy landscape, and CoTTA (Wang et al., 2022) targets continually changing target domains. We compare against all five of these methods in our experiments. While these methods improve robustness to non-stationarity, they optimize *instantaneous objectives* and do not explicitly model how current updates affect future-domain OOD

performance. Our framework instead introduces a temporally-aware objective that explicitly accounts for the long-term impact of each update.

Wild-data and open-world OOD detection. Recent work explores OOD detection using unlabeled data that mixes ID and OOD samples. Outlier Exposure (Hendrycks et al., 2018) demonstrates that auxiliary outlier data can improve detection performance, while SCONE (Bai et al., 2023) shows that wild data can simultaneously benefit OOD detection and generalization. These approaches assume a single wild distribution that is fixed during training. In contrast, we consider a *temporal setting* in which data distributions evolve over time, and we analyze how the *trajectory of updates* influences long-run OOD robustness. A related line of work studies model adaptation under non-stationary environments, including test-time training with self supervision (Sun et al., 2020), entropy-based adaptation (Wang et al., 2021), and stability-aware updates (Niu et al., 2022a). These methods address online distribution shift by improving predictive performance at each step, but they do not explicitly analyze how optimization updates influence *future-domain* OOD detection behavior. Our work complements this literature by providing a theoretical analysis of optimization dynamics and their role in temporal robustness.

Continual learning and forgetting. Continual learning methods (Kirkpatrick et al., 2017; Lopez-Paz & Ranzato, 2017; Farajtabar et al., 2020) study how models retain knowledge across sequential tasks. The model-change and environment-change terms we separate in Section 4 play a role similar to the stability-plasticity trade-off in this literature, and we measure forgetting directly in our streams. However, our focus differs in two key aspects: *i)* we study *OOD detection* rather than classification accuracy, and *ii)* we analyze how *optimization dynamics* shape the evaluation risks under evolving distributions.

RL for optimization and meta-learning. RL and meta-learning have been used to design learned optimizers and adaptive update rules (Andrychowicz et al., 2016; Metz et al., 2019). Unlike these general-purpose approaches, our framework introduces a *task-structured value function* tied specifically to semantic-OOD false positive rate. This structure enables a direct connection between optimization dynamics and OOD detection performance, allowing us to derive explicit theoretical guarantees on future-domain robustness.

Energy-based OOD scores. Energy-based formulations interpret classifiers as defining energy functions that separate ID and OOD samples (Liu et al., 2020a; Duvenaud et al., 2020). We adopt the energy score as the detector statistic, and its smoothed false positive rate is the state coordinate through which the value function acts on the update. Our analysis therefore connects update dynamics to OOD robustness through this soft-FPR channel and the two evaluation-risk gradients rather than through a separation bound on energies.

Optimization dynamics and training regimes. Our Section 5 studies the optimizer itself, bounding the loss gap between the RL-guided and GD trajectories and comparing their effective gradient flow. This connects to analyses of optimization dynamics and learned update rules (Andrychowicz et al., 2016; Metz et al., 2019), which characterize how an update rule shapes the trajectory rather than a single step.

C Section 4 Proofs

C.0.1 Exact Multi-Step Trajectory Identity

Let $\widehat{\mathcal{L}}_{\mathrm{OW},t,k}$ denote the realized batch-specific base objective at environment t and update step k. Both optimizers use the same sequence of batches and the same base objective, but evaluate its gradient at their respective parameters. Define

$$u^o_{t,k} := \nabla_\theta \widehat{\mathcal{L}}_{\mathrm{OW},t,k}\left(\theta^o_{t,k}\right), \qquad o \in \{\mathrm{RL}, \mathrm{GD}\}, \tag{9}$$

and let

$$g_{t,k} := \nabla_\theta V_{\phi_{t,k}}(s_{t,k}) \tag{10}$$

denote the value-gradient correction used by the RL-guided optimizer.

Lemma C.1 (Exact trajectory identity)**.** *Suppose the two optimizers follow*

$$\theta^{\mathrm{GD}}_{t,k+1} = \theta^{\mathrm{GD}}_{t,k} - \eta u^{\mathrm{GD}}_{t,k}, \tag{11}$$

$$\theta^{\mathrm{RL}}_{t,k+1} = \theta^{\mathrm{RL}}_{t,k} - \eta u^{\mathrm{RL}}_{t,k} + \eta_{\mathrm{RL}} g_{t,k}. \tag{12}$$

Define the parameter difference and the induced base-update drift by

$$d_{t,k} := \theta^{\mathrm{RL}}_{t,k} - \theta^{\mathrm{GD}}_{t,k}, \qquad e_{t,K} := -\eta \sum_{k=0}^{K-1} \left(u^{\mathrm{RL}}_{t,k} - u^{\mathrm{GD}}_{t,k}\right). \tag{13}$$

Then

$$d_{t,K} = d_{t,0} + \eta_{\mathrm{RL}} \sum_{k=0}^{K-1} g_{t,k} + e_{t,K}. \tag{14}$$

Proof. Subtracting Eq. Eq. 11 from Eq. Eq. 12 gives

$$d_{t,k+1} - d_{t,k} = -\eta \left(u^{\mathrm{RL}}_{t,k} - u^{\mathrm{GD}}_{t,k}\right) + \eta_{\mathrm{RL}} g_{t,k}. \tag{15}$$

Summing over $k = 0, \ldots, K-1$ telescopes the left-hand side:

$$d_{t,K} - d_{t,0} = -\eta \sum_{k=0}^{K-1} \left(u^{\mathrm{RL}}_{t,k} - u^{\mathrm{GD}}_{t,k}\right) + \eta_{\mathrm{RL}} \sum_{k=0}^{K-1} g_{t,k}. \tag{16}$$

Using the definition of $e_{t,K}$ yields Eq. Eq. 14. □

The identity is exact and requires neither convexity nor smoothness of the base objective. It also does not require a shared initialization. The term $d_{t,0}$ is retained because the two optimizers may enter the adaptation window with different parameters as a result of their previous trajectories.

C.0.2 Evaluation-Risk Theorem

Let

$$R^{\mathrm{FPR}}_t(\theta) := \mathrm{FPR}^{\mathrm{soft}}_{t,\beta}(\theta), \qquad R^{\mathrm{GErr}}_t(\theta) := \mathrm{GErr}_t(\theta), \tag{17}$$

denote the held-out evaluation risks for environment t. Define

$$h_{t,K} := \nabla R^{\mathrm{FPR}}_t \left(\theta^{\mathrm{GD}}_{t,K}\right), \qquad q_{t,K} := \nabla R^{\mathrm{GErr}}_t \left(\theta^{\mathrm{GD}}_{t,K}\right), \tag{18}$$

and

$$A^{\mathrm{FPR}}_{t,K} := -\left\langle h_{t,K}, \sum_{k=0}^{K-1} g_{t,k} \right\rangle, \tag{19}$$

$$A^{\mathrm{GErr}}_{t,K} := -\left\langle q_{t,K}, \sum_{k=0}^{K-1} g_{t,k} \right\rangle. \tag{20}$$

We assume that, for each $R \in \{R^{\mathrm{FPR}}_t, R^{\mathrm{GErr}}_t\}$, ∇R is $L_{R,t}$-Lipschitz on a convex neighborhood containing the segment between $\theta^{\mathrm{GD}}_{t,K}$ and $\theta^{\mathrm{RL}}_{t,K}$.

Proof of Theorem 4.2. Fix one of the two evaluation risks R, and abbreviate

$$\theta_0 := \theta^{\mathrm{GD}}_{t,K}, \qquad d := d_{t,K}, \qquad \theta_1 := \theta^{\mathrm{RL}}_{t,K} = \theta_0 + d. \tag{21}$$

By the fundamental theorem of calculus,

$$\begin{aligned} R(\theta_1) - R(\theta_0) &= \int_0^1 \langle \nabla R(\theta_0 + \lambda d), d \rangle \, d\lambda \\ &= \langle \nabla R(\theta_0), d \rangle + \rho_R, \end{aligned} \tag{22}$$

where

$$\rho_R := \int_0^1 \langle \nabla R(\theta_0 + \lambda d) - \nabla R(\theta_0), d \rangle \, d\lambda. \tag{23}$$

By Cauchy–Schwarz and the Lipschitz assumption,

$$|\rho_R| \le \int_0^1 \|\nabla R(\theta_0 + \lambda d) - \nabla R(\theta_0)\|_2 \, \|d\|_2 \, d\lambda \le \int_0^1 L_{R,t} \lambda \|d\|_2^2 \, d\lambda = \frac{L_{R,t}}{2} \|d\|_2^2. \tag{24}$$

For $R = R_t^{\mathrm{FPR}}$, using Eq. Eq. 14,

$$\langle h_{t,K}, d_{t,K} \rangle = \langle h_{t,K}, d_{t,0} + e_{t,K} \rangle + \eta_{\mathrm{RL}} \left\langle h_{t,K}, \sum_{k=0}^{K-1} g_{t,k} \right\rangle = \langle h_{t,K}, d_{t,0} + e_{t,K} \rangle - \eta_{\mathrm{RL}} A_{t,K}^{\mathrm{FPR}}. \tag{25}$$

Combining this equality with Eqs. Eq. 22–Eq. 24 gives

$$R_t^{\mathrm{FPR}}\left(\theta_{t,K}^{\mathrm{RL}}\right) - R_t^{\mathrm{FPR}}\left(\theta_{t,K}^{\mathrm{GD}}\right) = -\eta_{\mathrm{RL}} A_{t,K}^{\mathrm{FPR}} + \langle h_{t,K}, d_{t,0} + e_{t,K} \rangle + \rho_{t,K}^{\mathrm{FPR}}, \tag{26}$$

where

$$\left|\rho_{t,K}^{\mathrm{FPR}}\right| \le \frac{L_{\mathrm{FPR},t}}{2} \|d_{t,K}\|_2^2. \tag{27}$$

The proof for $R = R_t^{\mathrm{GErr}}$ is identical. Therefore,

$$R_t^{\mathrm{GErr}}\left(\theta_{t,K}^{\mathrm{RL}}\right) - R_t^{\mathrm{GErr}}\left(\theta_{t,K}^{\mathrm{GD}}\right) = -\eta_{\mathrm{RL}} A_{t,K}^{\mathrm{GErr}} + \langle q_{t,K}, d_{t,0} + e_{t,K} \rangle + \rho_{t,K}^{\mathrm{GErr}}, \tag{28}$$

where

$$\left|\rho_{t,K}^{\mathrm{GErr}}\right| \le \frac{L_{\mathrm{GErr},t}}{2} \|d_{t,K}\|_2^2. \tag{29}$$

□

Sufficient conditions. Because

$$\rho_{t,K}^{\mathrm{FPR}} \le \frac{L_{\mathrm{FPR},t}}{2} \|d_{t,K}\|_2^2,$$

Eq. Eq. 26 is strictly negative whenever

$$\eta_{\mathrm{RL}} A_{t,K}^{\mathrm{FPR}} > \langle h_{t,K}, d_{t,0} + e_{t,K} \rangle + \frac{L_{\mathrm{FPR},t}}{2} \|d_{t,K}\|_2^2. \tag{30}$$

Similarly, for any $\varepsilon \ge 0$,

$$\eta_{\mathrm{RL}} A_{t,K}^{\mathrm{GErr}} \ge \langle q_{t,K}, d_{t,0} + e_{t,K} \rangle + \frac{L_{\mathrm{GErr},t}}{2} \|d_{t,K}\|_2^2 - \varepsilon \tag{31}$$

implies

$$R_t^{\mathrm{GErr}}\left(\theta_{t,K}^{\mathrm{RL}}\right) \le R_t^{\mathrm{GErr}}\left(\theta_{t,K}^{\mathrm{GD}}\right) + \varepsilon. \tag{32}$$

These two conditions are separate; neither implies the other.

C.0.3 Frozen-Feature Linear-Head Regularity

Let $f_W(x) = Wh(x)$ with $\|h(x)\|_2 \le H$ and $\pi_W(x) = \mathrm{softmax}(Wh(x))$. All matrix gradients are equipped with the Frobenius norm, and Hessians use the induced operator norm.

Proposition C.2 (Linear-head risk regularity)**.** *Suppose that the evaluation threshold τ_t is fixed with respect to W. Then the held-out semantic-OOD soft-FPR and covariate-shifted classification risk have locally Lipschitz gradients. Valid constants are*

$$L_{\mathrm{FPR},t} \le \frac{H^2}{6\sqrt{3}\,\beta^2} + \frac{H^2}{4\beta}, \qquad L_{\mathrm{GErr},t} \le H^2. \tag{33}$$

Proof. The energy is

$$E_W(x) = -\log \sum_{j=1}^{N} \exp([Wh(x)]_j). \tag{34}$$

Its gradient is

$$\nabla_W E_W(x) = -\pi_W(x) h(x)^\top. \tag{35}$$

Since $\|\pi_W(x)\|_2 \le 1$,

$$\|\nabla_W E_W(x)\|_F \le H. \tag{36}$$

Moreover,

$$\nabla_W^2 E_W(x) = -\big(\mathrm{Diag}(\pi_W(x)) - \pi_W(x)\pi_W(x)^\top\big) \otimes h(x)h(x)^\top. \tag{37}$$

The softmax covariance has operator norm at most one. Therefore,

$$\|\nabla_W^2 E_W(x)\|_{\mathrm{op}} \le H^2. \tag{38}$$

For the fixed evaluation threshold τ_t, define

$$a_W(x) := \tau_t - E_W(x), \qquad z_W(x) := \frac{a_W(x)}{\beta}. \tag{39}$$

Let $\mathbb{E}_x$ denote expectation over $x \sim \mathbb{P}_t^{\mathrm{sem}}$. The soft-FPR gradient is

$$\begin{aligned} \nabla_W \mathrm{FPR}_{t,\beta}^{\mathrm{soft}}(W) &= \frac{1}{\beta}\mathbb{E}_x\left[\sigma'(z_W(x))\nabla_W a_W(x)\right] \\ &= \frac{1}{\beta}\mathbb{E}_x\left[\sigma(z_W(x))\big(1-\sigma(z_W(x))\big)\nabla_W a_W(x)\right]. \end{aligned} \tag{40}$$

Because τ_t is fixed with respect to W,

$$\nabla_W a_W(x) = -\nabla_W E_W(x), \qquad \nabla_W^2 a_W(x) = -\nabla_W^2 E_W(x). \tag{41}$$

Consequently,

$$\|\nabla_W a_W(x)\|_F \le H, \qquad \|\nabla_W^2 a_W(x)\|_{\mathrm{op}} \le H^2. \tag{42}$$

Differentiating Eq. Eq. 40 once more gives

$$\begin{aligned} \nabla_W^2 \mathrm{FPR}_{t,\beta}^{\mathrm{soft}}(W) = \mathbb{E}_x\Bigg[&\frac{\sigma''(z_W(x))}{\beta^2}\nabla_W a_W(x) \otimes \nabla_W a_W(x) \\ &+ \frac{\sigma'(z_W(x))}{\beta}\nabla_W^2 a_W(x)\Bigg]. \end{aligned} \tag{43}$$

The logistic sigmoid satisfies

$$0 \le \sigma'(z) \le \frac{1}{4}, \qquad |\sigma''(z)| \le \frac{1}{6\sqrt{3}}. \tag{44}$$

Taking the operator norm in Eq. Eq. 43 yields

$$\left\|\nabla_W^2 \mathrm{FPR}_{t,\beta}^{\mathrm{soft}}(W)\right\|_{\mathrm{op}} \le \frac{H^2}{6\sqrt{3}\,\beta^2} + \frac{H^2}{4\beta}. \tag{45}$$

Thus, the soft-FPR gradient is locally Lipschitz with constant $L_{\mathrm{FPR},t}$.

For a known-class example (x, y),

$$\nabla_W \ell_{\mathrm{CE}}(Wh(x), y) = \big(\pi_W(x) - e_y\big)h(x)^\top, \tag{46}$$

and its Hessian is

$$\big(\mathrm{Diag}(\pi_W(x)) - \pi_W(x)\pi_W(x)^\top\big) \otimes h(x)h(x)^\top. \tag{47}$$

Its operator norm is at most H^2. Therefore, taking expectations over $(x, y) \sim \mathbb{P}_t^{\mathrm{cov}}$ shows that $\mathrm{GErr}_t(W)$ has an H^2-Lipschitz gradient. □

D PROOFS FOR SECTION 5

For the purpose of analysis we use the semantic-drift summary as a reduced state representation, $z_{t,k} := \mathbb{E}_{x\sim P^{\text{sem}}_{t+1}}[E_{\theta_{t,k}}(x)] - \mathbb{E}_{x\sim P^{\text{sem}}_{t}}[E_{\theta_{t,k}}(x)]$. In the proofs below, we use $z_{t,k}$ and $s_{t,k}$ interchangeably.

Assumption D.1 (Lipschitz continuous gradient)**.** *We assume that the function $V(s_t)$ (and $\mathcal{L}_{OW}$) is L_V-Lipschitz (and L-Lipschitz) continuous gradient.*

Assumption D.2 (Compactness)**.** *At each training update k both $\theta^{GD}_{t,k}, \theta^{RL}_{t,k} \in \mathcal{B}(\theta, r)$ with radius r.*

Assumption D.3 (Slow generalization differ)**.** *We assume that the semantic-shifted environment does not change arbitrarily the gradient of energy at each training step: $\mathbb{E}_{x\sim P^{sem}_{t+1}}[\nabla_\theta E_{\theta_{t,k}}(x)] - \mathbb{E}_{x\sim P^{sem}_{t}}[\nabla_\theta E_{\theta_{t,k}}(x)] = \mathcal{O}(\xi_{t,k})$.*

Assumption D.4 (Slow environment-transfer generalization energy)**.** *We assume that the environment does not transfer energy arbitrarily fast that is $S_{t\to t+1} = \mathcal{O}(\xi^S_{t\to t+1})$.*

Assumption D.5 (Slow environment-based semantic shifted energy)**.** *We assume that the environment does not change semantic shifted energy arbitrarily fast that is $E_{t\to t+1} = \mathcal{O}(\xi^E_{t\to t+1})$.*

D.1 PROOF OF THEOREM 5.3

Proof. Recall the Def. 5.1: $F^{OW}_t := \mathcal{L}_{OW}(\theta^{\text{RL}}_{t,*}) - \mathcal{L}_{OW}(\theta^{\text{GD}}_{t,*})$. Applying second-order Taylor expansion, we approximate $\mathcal{L}_{OW}(\theta^{\text{RL}}_{t,*})$ around $\theta^{\text{GD}}_{t,*}$ by

$$\mathcal{L}_{OW}(\theta^{\text{GD}}_{t,*}) + (\theta^{\text{RL}}_{t,*} - \theta^{\text{GD}}_{t,*})^\top \nabla \mathcal{L}_{OW}(\theta^{\text{GD}}_{t,*}) + \frac{1}{2}(\theta^{\text{RL}}_{t,*} - \theta^{\text{GD}}_{t,*})^\top \nabla^2 \mathcal{L}_{OW}(\theta^{\text{GD}}_{t,*})(\theta^{\text{RL}}_{t,*} - \theta^{\text{GD}}_{t,*}), \quad (48)$$

where, $\nabla^2 \mathcal{L}_{OW}(\theta^{\text{GD}}_{t,*})$ is the Hessian matrix for loss $\mathcal{L}_{OW}$ at point $\theta^{\text{GD}}_{t,*}$. Note that since $\theta^{\text{GD}}_{t,*}$ in an optimizer therefore $\nabla \mathcal{L}_{OW}(\theta^{\text{GD}}_{t,*}) = 0$. This simplifies the approximation of $\mathcal{L}_{OW}(\theta^{\text{RL}}_{t,*})$ in (48) as

$$\mathcal{L}_{OW}(\theta^{\text{RL}}_{t,*}) \approx \mathcal{L}_{OW}(\theta^{\text{GD}}_{t,*}) + \frac{1}{2}(\theta^{\text{RL}}_{t,*} - \theta^{\text{GD}}_{t,*})^\top \nabla^2 \mathcal{L}_{OW}(\theta^{\text{GD}}_{t,*})(\theta^{\text{RL}}_{t,*} - \theta^{\text{GD}}_{t,*}).$$

Equivalently

$$F^{OW}_t := \mathcal{L}_{OW}(\theta^{\text{RL}}_{t,*}) - \mathcal{L}_{OW}(\theta^{\text{GD}}_{t,*}) \approx \frac{1}{2}(\theta^{\text{RL}}_{t,*} - \theta^{\text{GD}}_{t,*})^\top \nabla^2 \mathcal{L}_{OW}(\theta^{\text{GD}}_{t,*})(\theta^{\text{RL}}_{t,*} - \theta^{\text{GD}}_{t,*}). \quad (49)$$

Let λ_{max} be the maximum eigenvalue of $\nabla^2 \mathcal{L}_{OW}(\theta^{\text{GD}}_{t,*})$. Using (49), we upper-bound F^{OW}_t by

$$F^{OW}_t \le \frac{1}{2}\lambda_{max} \|\theta^{\text{RL}}_{t,*} - \theta^{\text{GD}}_{t,*}\|^2. \quad (50)$$

Let $\theta^{\text{GD}}_{t,K}$ and $\theta^{\text{RL}}_{t,K}$ denote the parameters after K training steps of standard GD and RL-augmented optimizers, respectively. Going back to the GD update training update in (1) and RL-guided training update in (2):

$$\theta^{\text{GD}}_{t,K} = \theta^{\text{GD}}_{t,K-1} - \eta \nabla_\theta L_{OW}(\theta^{\text{GD}}_{t,K-1}), \quad (51)$$

$$\theta^{\text{RL}}_{t,K} = \theta^{\text{RL}}_{t,K-1} - \eta \nabla_\theta L_{OW}(\theta^{\text{RL}}_{t,K-1}) + \eta_{\text{RL}} \nabla_\theta V(s_{t,K-1}). \quad (52)$$

We suppose that after K training steps , GD and RL-guided updates return the optimizers, that is $\theta^{\text{RL}}_{t,*} = \theta^{\text{RL}}_{t,K}$ and $\theta^{\text{GD}}_{t,*} = \theta^{\text{GD}}_{t,K}$, when the network is trained K updates (epochs). By utilizing Eq. (51) and Eq. (52) and applying triangle inequality, we have

$$\begin{aligned}\|\theta^{\text{RL}}_{t,*} - \theta^{\text{GD}}_{t,*}\|^2 \le & \|\theta^{\text{RL}}_{t,K-1} - \theta^{\text{GD}}_{t,K-1}\|^2 + \eta^2 \|\nabla_\theta L_{OW}(\theta^{\text{GD}}_{t,K-1}) - \nabla_\theta L_{OW}(\theta^{\text{RL}}_{t,K-1})\|^2 \\ & + \eta^2_{\text{RL}} \|\nabla_\theta V(s_{t,K-1})\|^2. \end{aligned} \quad (53)$$

Recursively for training updates $k = 0, \dots K$, we extend (53) by

$$\begin{aligned}\|\theta^{\text{RL}}_{t,*} - \theta^{\text{GD}}_{t,*}\|^2 \le & \|\theta^{\text{RL}}_{t,0} - \theta^{\text{GD}}_{t,0}\|^2 + \eta^2 \sum_{k=1}^{K} \|\nabla_\theta L_{OW}(\theta^{\text{GD}}_{t,k-1}) - \nabla_\theta L_{OW}(\theta^{\text{RL}}_{t,k-1})\|^2 \\ & + \eta^2_{\text{RL}} \sum_{k=1}^{K} \|\nabla_\theta V(s_{t,k-1})\|^2 \end{aligned} \quad (54)$$

Without lose of generality, we assume that the initialization points for GD and RL-guided optimizers are the same meaning that $\|\theta^{\text{RL}}_{t,0} - \theta^{\text{GD}}_{t,0}\|^2 = 0$. This simplifies (54) as follows:

$$\|\theta^{\text{RL}}_{t,*} - \theta^{\text{GD}}_{t,*}\|^2 \leq \eta^2 \sum_{k=1}^{K} \|\nabla_\theta L_{OW}(\theta^{\text{GD}}_{t,k-1}) - \nabla_\theta L_{OW}(\theta^{\text{RL}}_{t,k-1})\|^2 + \eta^2_{\text{RL}} \sum_{k=1}^{K} \|\nabla_\theta V(s_{t,k-1})\|^2. \tag{55}$$

Remark: Note that for environment t, the state is a function of parameter θ_t, therefore during the training $s_{t,k} \neq s_{t,k'}$ for $k, k' \in \{0, 1, \dots K\}$, $i \neq j$.

Going back to the proof, next, we extend the term $\|\nabla_\theta V(s_{t,k-1})\|^2$ in (55): first we add and subtract the term $\nabla_\theta V(s_{t,k-2})$, then use triangle inequality,

$$\begin{aligned}\|\nabla_\theta V(s_{t,k-1})\|^2 &= \|\nabla_\theta V(s_{t,k-1}) - \nabla_\theta V(s_{t,k-2}) + \nabla_\theta V(s_{t,k-2})\|^2 \\ &\leq \|\nabla_\theta V(s_{t,k-1}) - \nabla_\theta V(s_{t,k-2})\|^2 + \|\nabla_\theta V(s_{t,k-2})\|^2,\end{aligned}$$

recursively we extend the above inequality for training update $k' = 1, \dots k-1$ as follows

$$\|\nabla_\theta V(s_{t,k-1})\|^2 \leq \sum_{k'=1}^{k-1} \|\nabla_\theta V(s_{t,k'}) - \nabla_\theta V(s_{t,k'-1})\|^2. \tag{56}$$

Note that the gradient of value function is zero at the initialization, i.e. $\nabla_\theta V(s_{t,0}) = 0$.

Next we first apply Assumption D.1 on value function V_θ and $\mathcal{L}_{OW}$. Hence the bound in (55) is upper-bounded by

$$\|\theta^{\text{RL}}_{t,*} - \theta^{\text{GD}}_{t,*}\|^2 \leq \eta^2 L \sum_{k=1}^{K} \|\theta^{\text{GD}}_{t,k-1} - \theta^{\text{RL}}_{t,k-1}\|^2 + \eta^2_{\text{RL}} L_V \sum_{k=1}^{K} \sum_{k'=1}^{k-1} \|s_{t,k'} - s_{t,k'-1}\|^2. \tag{57}$$

Secondly apply Assumption D.2, meaning that at each training update $\|\theta^{\text{GD}}_{t,k} - \theta^{\text{RL}}_{t,k}$ is within a parameter ball with radius r, $\mathcal{B}(\theta, r)$. This yields to the bound below for (57),

$$\|\theta^{\text{RL}}_{t,*} - \theta^{\text{GD}}_{t,*}\|^2 \leq 4\eta^2 L\ K\ r^2 + \eta^2_{\text{RL}} L_V \sum_{k=1}^{K} \sum_{k'=1}^{k-1} \|s_{t,k'} - s_{t,k'-1}\|^2 \tag{58}$$

Finally, combining (50) and (58), we derive the our main bound in (5):

$$F^{OW}_t \leq \frac{1}{2} \lambda_{max} \left[\eta_{K,r} + \eta_{RL} \sum_{k=1}^{K} \sum_{k'=1}^{k-1} \|s_{t,k'} - s_{t,k'-1}\|^2 \right], \tag{59}$$

where $\eta_{RL} = \eta^2_{RL} L_V$ and $\eta_{K,r} = 4\eta^2 L\ K\ r^2$. □

D.2 Proof of Corollary 5.4

Proof. Let us recall

$$z_{t,k} := \mathbb{E}_{x \sim P^{\text{sem}}_{t+1}}[E_{\theta_{t,k}}(x)] - \mathbb{E}_{x \sim P^{\text{sem}}_{t}}[E_{\theta_{t,k}}(x)]. \tag{60}$$

Hence for consecutive training updates k and $k-1$, we have

$$\begin{aligned}&z_{t,k} - z_{t,k-1} \\ &= \mathbb{E}_{x \sim P^{sem}_{t+1}}\left[E_{\theta_{t,k}}(x)\right] - \mathbb{E}_{x \sim P^{sem}_{t}}\left[E_{\theta_{t,k}}(x)\right] - \mathbb{E}_{x \sim P^{sem}_{t+1}}\left[E_{\theta_{t,k-1}}(x)\right] + \mathbb{E}_{x \sim P^{sem}_{t}}\left[E_{\theta_{t,k-1}}(x)\right] \\ &= \left(\mathbb{E}_{x \sim P^{sem}_{t+1}}\left[E_{\theta_{t,k}}(x)\right] - \mathbb{E}_{x \sim P^{sem}_{t+1}}\left[E_{\theta_{t,k-1}}(x)\right)\right) - \left(\mathbb{E}_{x \sim P^{sem}_{t}}\left[E_{\theta_{t,k}}(x)\right] - \mathbb{E}_{x \sim P^{sem}_{t}}\left[E_{\theta_{t,k-1}}(x)\right]\right) \\ &= \mathbb{E}_{x \sim P^{sem}_{t+1}}\left[E_{\theta_{t,k}}(x) - E_{\theta_{t,k-1}}(x)\right] - \mathbb{E}_{x \sim P^{sem}_{t}}\left[E_{\theta_{t,k}}(x) - E_{\theta_{t,k-1}}(x)\right].\end{aligned} \tag{61}$$

Approximate epoch-based energy gradient as $\nabla^{(k)}_\theta E_{\theta_t}(x) \approx E_{\theta_{t,k}}(x) - E_{\theta_{t,k-1}}(x)$, therefore the Eq. 61 turns into

$$z_{t,k} - z_{t,k-1} = \mathbb{E}_{x \sim P^{sem}_{t+1}}\left[\nabla^{(k)}_\theta E_{\theta_t}(x)\right] - \mathbb{E}_{x \sim P^{sem}_{t}}\left[\nabla^{(k)}_\theta E_{\theta_t}(x)\right] \tag{62}$$

On the other hand, by Def. 5.2, we have

$$\|z_{t,k} - z_{t,k-1}\|^2 = \|S^{(k)}_{t,\nabla}\|^2,$$

where $S^{(k)}_{t,\nabla}$ is the epoch-based energy gradient state. This implies an alternative upper bound for F^{OW}_t as expressed in (6). □

D.3 Proof of Theorem 5.7

Proof. We start with the gradient of loss function $\mathcal{L}_G$ under GD and RL-guided optimizers for environment t and training step k:

$$\mathbf{G}_{t,k}^{GD} := \nabla_\theta \mathcal{L}_{OW}(\theta_{t,k}^{GD}) \quad \text{and} \quad \mathbf{G}_{t,k}^{RL} := \nabla_\theta \mathcal{L}_{OW}(\theta_{t,k}^{RL}) + \eta_{RL} \nabla_\theta V(s_{t,k}). \tag{63}$$

Using the triangle inequality, we bound the length of the gradient difference vector between $\mathbf{G}_{t,k}^{GD}$ and $\mathbf{G}_{t,k}^{RL}$ by

$$\begin{aligned} \|\mathbf{G}_{t,k}^{RL} - \mathbf{G}_{t,k}^{GD}\|_2 &= \|\nabla_\theta \mathcal{L}_{OW}(\theta_{t,k}^{RL}) + \eta_{RL} \nabla_\theta V(s_{t,k}) - \nabla_\theta \mathcal{L}_{OW}(\theta_{t,k}^{GD})\|_2 \\ &\leq \|\nabla_\theta \mathcal{L}_{OW}(\theta_{t,k}^{RL}) - \nabla_\theta \mathcal{L}_{OW}(\theta_{t,k}^{GD})\|_2 + \eta_{RL} \|\nabla_\theta V(s_{t,k})\|_2. \end{aligned} \tag{64}$$

Next, let us focus on $\|\nabla_\theta V(s_{t,k})\|_2$ and subtract and add the term $\nabla_\theta V(s_{t,k-1})$:

$$\begin{aligned} \|\nabla_\theta V(s_{t,k})\|_2 &= \|\nabla_\theta V(s_{t,k}) - \nabla_\theta V(s_{t,k-1}) + \nabla_\theta V(s_{t,k-1})\|_2 \\ &\leq \|\nabla_\theta V(s_{t,k}) - \nabla_\theta V(s_{t,k-1})\|_2 + \|\nabla_\theta V(s_{t,k-1})\|_2. \end{aligned}$$

Recursively, the above arguments lead to the following bound:

$$\|\nabla_\theta V(s_{t,k})\|_2 \leq \sum_{k'=1}^{k} \|\nabla_\theta V(s_{t,k'}) - \nabla_\theta V(s_{t,k'-1})\|_2 \leq \sqrt{L_V} \sum_{k'=1}^{k} \|s_{t,k'} - s_{t,k'-1}\|_2, \tag{65}$$

where the second inequality is concluded from Assumption D.1. Note that $\|\nabla_\theta V(s_{t,0})\|_2 = 0$.

Now, we recall

$$z_{t,k} := \mathbb{E}_{x \sim P_{t+1}^{\text{sem}}}[E_{\theta_{t,k}}(x)] - \mathbb{E}_{x \sim P_t^{\text{sem}}}[E_{\theta_{t,k}}(x)]. \tag{66}$$

By Def. 5.2, and similar arguments in Corollary 5.4, we redeem the epoch-based gradient energy state

$$\|z_{t,k'} - z_{t,k'-1}\|_2 = \|S_{t,\nabla}^{(k')}\|_2. \tag{67}$$

Plugging (67) in the upper bound (65), we have

$$\|\nabla_\theta V(s_{t,k})\|_2 \leq \sqrt{L_V} \sum_{k'=1}^{k} \|S_{t,\nabla}^{(k')}\|_2. \tag{68}$$

Finally, we focus on $\|\nabla_\theta \mathcal{L}_{OW}(\theta_{t,k}^{RL}) - \nabla_\theta \mathcal{L}_{OW}(\theta_{t,k}^{GD})\|_2$: Using Assumptions D.1 and D.2 for $\mathcal{L}_{OW}$, we bound the difference of the gradient of loss for GD and RL-guided optimizers by

$$\|\nabla_\theta \mathcal{L}_{OW}(\theta_{t,k}^{RL}) - \nabla_\theta \mathcal{L}_{OW}(\theta_{t,k}^{GD})\|_2 \leq 2\sqrt{L}\, r, \tag{69}$$

where r is the radius of ball $\mathcal{B}(\theta, r)$. We refer Def. 5.5, apply inverse triangle inequality on $EGF_{t,p}^{train}$ difference between GD and RL-guided optimizers:

$$\frac{1}{K} \sum_{k=1}^{K} \|\mathbf{G}_{t,k}^{RL}\|_2 - \frac{1}{K} \sum_{k=1}^{K} \|\mathbf{G}_{t,k}^{GD}\|_2 \leq \frac{1}{K} \sum_{k=1}^{K} \|\mathbf{G}_{t,k}^{RL} - \mathbf{G}_{t,k}^{GD}\|_2. \tag{70}$$

Sequentially Plugging 68) and (69) in (64) and plugging (64) in (70), we conclude

$$\begin{aligned} \frac{1}{K} \sum_{k=1}^{K} \|\mathbf{G}_{t,k}^{RL}\|_2 - \frac{1}{K} \sum_{k=1}^{K} \|\mathbf{G}_{t,k}^{GD}\|_2 &\leq \frac{1}{K} \sum_{k=1}^{K} \left(2\sqrt{L} r + \eta_{RL} \sqrt{L_V} \sum_{k'=1}^{k} \|S_{t,\nabla}^{(k')}\|_2 \right) \\ &\leq 2\sqrt{L} r + \frac{\eta_{RL} \sqrt{L_V}}{K} \sum_{k=1}^{K} \sum_{k'=1}^{k} \|S_{t,\nabla}^{(k')}\|_2, \end{aligned} \tag{71}$$

which proves our claim in (7). □

D.4 Proof of Theorem 5.8

Proof. We start with the gradient of loss function $\mathcal{L}_G$ under GD and RL-guided optimizers for environment t:

$$\mathbf{G}_t^{GD} := \nabla_\theta \mathcal{L}_{OW}(\theta_t^{GD}) \quad \text{and} \quad \mathbf{G}_t^{RL} := \nabla_\theta \mathcal{L}_{OW}(\theta_t^{RL}) + \eta_{RL} \nabla_\theta V(s_t). \tag{72}$$

In addition, recall Def. 5.5, and apply inverse triangle inequality on $EGF_{t,p}^{env}$ difference between GD and RL-guided optimizers:

$$\frac{1}{T}\sum_{t=1}^{T} \|\mathbf{G}_t^{RL}\|_2 - \frac{1}{T}\sum_{t=1}^{T} \|\mathbf{G}_t^{GD}\|_2 \leq \frac{1}{T}\sum_{t=1}^{T} \|\mathbf{G}_t^{RL} - \mathbf{G}_t^{GD}\|_2. \tag{73}$$

Using $\mathbf{G}_t^{GD}$ and $\mathbf{G}_t^{RL}$ in (72), we bound the length of $\mathbf{G}_t^{RL} - \mathbf{G}_t^{GD}$ by triangle inequality as follows:

$$\begin{aligned} \|\mathbf{G}_t^{RL} - \mathbf{G}_t^{GD}\|_2 &= \|\nabla_\theta \mathcal{L}_{OW}(\theta_t^{RL}) + \eta_{RL}\nabla_\theta V(s_t) - \nabla_\theta \mathcal{L}_{OW}(\theta_t^{GD})\|_2 \\ &\leq \|\nabla_\theta \mathcal{L}_{OW}(\theta_t^{RL}) - \nabla_\theta \mathcal{L}_{OW}(\theta_t^{GD})\|_2 + \eta_{RL}\|\nabla_\theta V(s_t)\|_2. \end{aligned} \tag{74}$$

Using Assumptions D.1 and D.2 for $\mathcal{L}_{OW}$ in environment t, we have

$$\|\nabla_\theta \mathcal{L}_{OW}(\theta_t^{RL}) - \nabla_\theta \mathcal{L}_{OW}(\theta_t^{GD})\|_2 \leq 2\sqrt{L}\, r, \tag{75}$$

where r in the radius of $\mathcal{B}(\theta_t, r)$ in Assumption D.2. Next, similar to arguments in Theorem 5.7, we have

$$\begin{aligned} &\|\nabla_\theta V(s_t)\|_2 = \|\nabla_\theta V(s_t) - \nabla_\theta V(s_{t-1}) + \nabla_\theta V(s_{t-1})\|_2 \\ &\leq \|\nabla_\theta V(s_t) - \nabla_\theta V(s_{t-1})\|_2 + \|\nabla_\theta V(s_{t-1})\|_2 = \sum_{\tau=1}^{t} \|\nabla_\theta V(s_\tau) - \nabla_\theta V(s_{\tau-1})\|_2 \\ &\leq \sum_{\tau=1}^{t} \sqrt{L_V}\|s_\tau - s_{\tau-1}\|_2 \ \text{ using Assumption D.2.} \end{aligned} \tag{76}$$

Note that $\|\nabla_\theta V(s_0)\|_2 = 0$. Recall

$$z_\tau := \mathbb{E}_{x\sim P_{\tau+1}^{\text{sem}}}[E_{\theta_\tau}(x)] - \mathbb{E}_{x\sim P_{\tau}^{\text{sem}}}[E_{\theta_\tau}(x)]. \tag{77}$$

Hence for two environments $\tau - 1$ and τ:

$$\begin{aligned} &z_\tau - z_{\tau-1} \\ &= \mathbb{E}_{x\sim P_{\tau+1}^{sem}}\left[E_{\theta_\tau}(x)\right] - \mathbb{E}_{x\sim P_{\tau}^{sem}}\left[E_{\theta_\tau}(x)\right] - \mathbb{E}_{x\sim P_{\tau}^{sem}}\left[E_{\theta_{\tau-1}}(x)\right] + \mathbb{E}_{x\sim P_{\tau-1}^{sem}}\left[E_{\theta_{\tau-1}}(x)\right] \\ &= \left(\mathbb{E}_{x\sim P_{\tau+1}^{sem}}\left[E_{\theta_\tau}(x)\right] - \mathbb{E}_{x\sim P_{\tau}^{sem}}\left[E_{\theta_{\tau-1}}(x)\right]\right) - \left(\mathbb{E}_{x\sim P_{\tau}^{sem}}\left[E_{\theta_\tau}(x)\right] - \mathbb{E}_{x\sim P_{\tau-1}^{sem}}\left[E_{\theta_{\tau-1}}(x)\right]\right) \\ &= S_{\tau\to\tau+1} - E_{\tau-1\to\tau}, \end{aligned} \tag{78}$$

The last equality of (78) is based on Def. 5.6. Applying Assumptions D.4 and D.5

$$\|z_\tau - z_{\tau-1}\|_2 \leq \mathcal{O}(\xi_{\tau\to\tau+1}^{S}) + \mathcal{O}(\xi_{\tau-1\to\tau}^{E}) \tag{79}$$

Going back to (76), substituting (81) in (76) implies

$$\|\nabla_\theta V(s_t)\|_2 \leq \sqrt{L_V}\sum_{\tau=1}^{t}\left(\mathcal{O}(\xi_{\tau\to\tau+1}^{s}) + \mathcal{O}(\xi_{\tau-1\to\tau}^{E})\right). \tag{80}$$

Combining (75) and (80) and use them in (74), we have

$$\|\mathbf{G}_t^{RL} - \mathbf{G}_t^{GD}\|_2 \leq 2\sqrt{L}\, r + \eta_{RL}\sqrt{L_V}\sum_{\tau=1}^{t}\left(\mathcal{O}(\xi_{\tau\to\tau+1}^{s}) + \mathcal{O}(\xi_{\tau-1\to\tau}^{E})\right). \tag{81}$$

Hence by using (81) in (73), we derive the upper bound below:

$$\frac{1}{T}\sum_{t=1}^{T} \|\mathbf{G}_t^{RL}\|_2 - \frac{1}{T}\sum_{t=1}^{T} \|\mathbf{G}_t^{GD}\|_2 \leq 2\sqrt{L}\, r + \frac{\eta_{RL}\sqrt{L_V}}{T}\sum_{t=1}^{T}\sum_{\tau=1}^{t}\left(\mathcal{O}(\xi_{\tau\to\tau+1}^{s}) + \mathcal{O}(\xi_{\tau-1\to\tau}^{E})\right), \tag{82}$$

which proves the theorem and yield our claim in (8). □

D.5 Datasets and Preprocessing

Each benchmark is turned into an ordered stream of environments. The environment axis is the ordering the dataset provides. For Core 50 this is acquisition sessions, days for iCubWorld, task difficulty for OpenLORIS, rendition domains for DomainNet, camera locations for Terra Incognita, and temporal periods for Century-of-Portraits, and time bins of two consecutive buckets for CLEAR. In each stream a fixed set of classes are held out as semantic OOD and the remaining classes are considered in-distribution. The held-out classes appear in every environment.

Every dataset is split into training, validation, and test sets separately for each class and environment using a fixed seed. When a dataset provides an official train and test partition (CLEAR-10, CLEAR-100, DomainNet, iCubWorld, and OpenLORIS), the test images come from the official test partition and the training and validation images from the official training partition. So, no official test image is used for adaptation or validation. CLEAR-10 keeps its full official test set, CLEAR-100 caps it at 15% of the total, such that the train/val/test split is 70/15/15. Core50 keeps 50 evenly spaced frames per object and session. Terra Incognita caps the result at 150, 40, and 40 images, and Century-of-Portraits caps each class at 1250 images before splitting. For the video frames of OpenLORIS we sample evenly spaced frames, rather than consecutive ones. In OpenLORIS, Terra Incognita, iCubWorld, and Century-of-Portraits, any class and environment with too few images is dropped. In Century-of-Portraits the in-distribution labels is gender (male (M) or female (F)) and the held-out class consists of faces from decades not used for training.

Each image is passed once through the frozen backbone. The feature is ℓ_2-normalized. For the multi-modal variant, captions are generated once per image with Mini-CPM-V at temperature 0 with a fixed seed. It is then embedded by the text encoder of the same backbone, and concatenated with the visual feature before normalization. Splitting, captioning, and feature extraction use seed 42.

D.6 Optimizer and Protocol

Both Rl and GD minimize the objective $\mathcal{L}_{OW}$, where $\mathcal{L}_{OOD}$ is the soft FPR surrogate computed on the pseudo semantic-OOD subset. The subset is formed by thresholding the energy at a percentile of the in-distribution energies of the current batch, and the surrogate uses the sharpness β. The value network V_ϕ is a small MLP on the state $s_{t,k}$, the head at every inner step. RL applies the update of Eq. 2, and GD applies Eq. 1, which is the same update without the value correction. The five TTA baselines adapt the same linear head with their own objectives. In each environment the head takes K inner steps, starting from the parameters reached at the end of the previous environment, and is then evaluated on the test split of that environment, so FPR@95, AUROC, generalization error, and accuracy are averaged over the environments of the stream. All methods start each seed from the same initial head θ_0, fit on the first environment, and GD and RL see the same batches. One set of hyperparameters is used for every dataset, backbone, and method (Table 3), with no per-dataset tuning. Results are averaged over seeds 0 to 4, and significance is paired bootstrap with 10^4 resamples over per-(seed, environment) differences. The two hyperparameters that control the strength of the correction were ablated, the number of inner steps $k \in \{5, 10, 20, 30, 50\}$ (supplementary material D.12) and the weight $\lambda_{\text{Sem}} \in \{0.1, 0.25, 0.5, 1, 2\}$ (supplementary material D.11). We also ablated the feature modality (supplementary material D.8), the encoder (supplementary material D.7), and the adaptation itself through a frozen head (supplementary material D.9).

Table 3: Hyperparameters shared by all datasets, backbones, and methods, as used in every run configuration.

Hyperparameter	Value	Hyperparameter	Value
inner steps K	30	GD step η	0.05
RL step η_{RL}	0.05	OOD weight λ_{Sem}	0.5
soft-FPR sharpness β	1.0	threshold percentile	0.95
TD discount γ	0.9	value-network learning rate	0.001
batch size	256	θ_0 warm-up epochs	20
warm-up learning rate	0.05		

D.7 Per-backbone Results

Table 4 reports FPR@95 of GD and RL on the gradual streams for every backbone. RL is lower than GD in all fifteen configurations, including the non-CLIP encoder ALIGN, which indicates that the effect comes from the optimizer and not from a particular pretrained feature space.

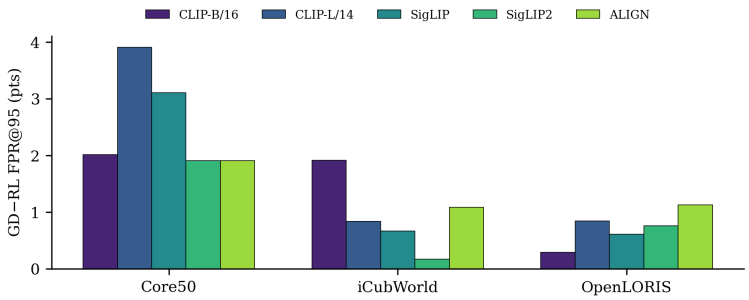


Figure 3: FPR@95 improvement of RL over GD for each encoder on the gradual streams (multimodal). The gain is positive for all five encoders.

Table 4: FPR@95 (%, lower is better) on the gradual streams for every backbone, multimodal, mean $\pm$ std over 5 seeds. * marks RL below GD at $p < 0.05$ (paired bootstrap).

Dataset	Backbone	GD	RL	GD−RL
Core50	CLIP-B/16	62.7 ± 0.5	$60.7 \pm 3.0^*$	+2.0
Core50	CLIP-L/14	50.2 ± 0.4	$46.3 \pm 5.9^*$	+3.9
Core50	SigLIP	41.0 ± 0.9	$37.9 \pm 4.3^*$	+3.1
Core50	SigLIP2	30.1 ± 0.5	$28.2 \pm 3.9^*$	+1.9
Core50	ALIGN	21.2 ± 0.7	$19.3 \pm 2.3^*$	+1.9
iCubWorld	CLIP-B/16	87.6 ± 0.6	$85.7 \pm 1.3^*$	+1.9
iCubWorld	CLIP-L/14	83.3 ± 2.0	82.5 ± 2.1	+0.8
iCubWorld	SigLIP	83.3 ± 1.1	$82.7 \pm 1.4^*$	+0.7
iCubWorld	SigLIP2	82.8 ± 0.5	82.7 ± 1.2	+0.2
iCubWorld	ALIGN	81.4 ± 1.3	$80.3 \pm 2.3^*$	+1.1
OpenLORIS	CLIP-B/16	49.5 ± 0.7	49.2 ± 2.9	+0.3
OpenLORIS	CLIP-L/14	41.2 ± 0.4	$40.4 \pm 2.0^*$	+0.8
OpenLORIS	SigLIP	26.8 ± 0.9	26.1 ± 2.0	+0.6
OpenLORIS	SigLIP2	36.7 ± 0.4	35.9 ± 2.6	+0.8
OpenLORIS	ALIGN	38.2 ± 0.7	$37.1 \pm 1.6^*$	+1.1

D.8 Feature-modality Ablation

Table 5 compares visual features alone with visual features concatenated with a MiniCPM-V caption embedding, for the same encoder. Caption embeddings are not needed for the gain. With visual features alone, RL improves on GD by 2.9, 4.1, and 0.9 points on the three gradual streams, more than with captions (2.0, 1.9, and 0.3), and the improvement is significant for 5, 5, and 2 of the five encoders. Under abrupt drift and on the saturated streams the gain stays negligible in both modalities. The absolute FPR depends on the modality because the caption embedding changes the feature space. it is lower with captions on OpenLORIS, DomainNet, and CLEAR and higher on Core50, and iCubWorld.

D.9 Generalization Error and the Frozen Control

Table 6 compares generalization error (cross-entropy on the known classes) and accuracy for RL and GD without covariate shift, together with the FPR of a head that is never adapted. GD and RL have the same generalization error within 0.002 and the same accuracy to within 0.1 points. So, the lower FPR of RL does not lead to worse classification. The frozen head has a higher FPR than both GD and

Table 5: Feature-modality ablation. FPR@95 (%) of GD and RL, mean ± std over 5 seeds, CLIP-B/16, with visual features alone and with a caption embedding appended. The columns marked sig. count the encoders (out of five) for which RL is below GD at $p < 0.05$.

Dataset	Visual			Visual + caption		
	GD	RL	sig.	GD	RL	sig.
Core50	53.3 ± 0.5	50.4 ± 2.5	5/5	62.7 ± 0.5	60.7 ± 3.0	5/5
iCubWorld	82.3 ± 2.1	78.2 ± 5.6	5/5	87.6 ± 0.6	85.7 ± 1.3	3/5
OpenLORIS	55.4 ± 0.9	54.5 ± 3.8	2/5	49.5 ± 0.7	49.2 ± 2.9	2/5
DomainNet	58.3 ± 0.7	58.2 ± 0.5	2/5	54.4 ± 1.2	54.6 ± 1.0	0/5
CLEAR-10	3.9 ± 0.2	3.8 ± 0.1	1/5	3.2 ± 0.3	3.2 ± 0.2	1/5
CLEAR-100	1.2 ± 0.1	1.2 ± 0.0	1/5	0.7 ± 0.0	0.7 ± 0.1	0/5

RL on every unsaturated stream, which confirms that online adaptation of the head is beneficial in these streams.

Table 6: Generalization error (GErr), accuracy, and the FPR@95 of a frozen head, multimodal CLIP-B/16, mean ± std over 5 seeds for GErr and accuracy.

Dataset	GErr GD	GErr RL	Acc. GD (%)	Acc. RL (%)	Frozen FPR (%)
Core50	0.1165 ± 0.0005	0.1171 ± 0.0005	96.4 ± 0.1	96.3 ± 0.1	67.4
iCubWorld	0.2237 ± 0.0019	0.2255 ± 0.0024	92.3 ± 0.1	92.2 ± 0.3	88.1
OpenLORIS	0.4450 ± 0.0025	0.4455 ± 0.0050	85.9 ± 0.1	85.9 ± 0.2	49.6
DomainNet	0.2275 ± 0.0015	0.2270 ± 0.0013	94.4 ± 0.1	94.4 ± 0.1	55.1
CLEAR-10	0.0329 ± 0.0004	0.0330 ± 0.0007	99.0 ± 0.0	99.0 ± 0.0	3.4
CLEAR-100	0.0257 ± 0.0001	0.0257 ± 0.0001	99.4 ± 0.0	99.4 ± 0.0	0.7

D.10 Mechanism and Dynamics

Table 7 reports two quantities used in the analysis. The energy bridge is the difference in mean semantic-OOD energy between GD and RL at the end of each environment, the direction implied by the soft-FPR channel of Section 4.2. We report its mean and the fraction of environments in which it is positive. Theorem 4.2 compares the two optimizers within one window from a shared initialization $\theta_{t,0}$ where we measure along the two full trajectories. So, the optimizers also differ in their starting point from the second environment going forward. The derivative $\partial V / \partial \mathrm{FPR}$ of the value network with respect to its FPR input is the negative of the critic sensitivity $c_{t,k}$ of Section 4.2, which the soft-FPR channel requires to be positive. We report its mean and the fraction of environments in which the derivative is negative. Both behave as the analysis requires on the gradual streams and are much smaller in magnitude on DomainNet and CLEAR, where RL brings no gain.

Table 8 reports the difference in effective gradient flow (Definition 5.5) between GD and RL, with $\mathcal{L}_G = \mathcal{L}_{OW}$ for GD and $\mathcal{L}_G = \mathcal{L}_{OW} + \eta_{\mathrm{RL}} V$ for RL and the norm $p = 2$. The training version averages $\|\nabla_\theta \mathcal{L}_G\|$ over the K inner steps of an environment, and the environment version uses the gradient at the start of each environment. Both differences are small and negative on every dataset. So, they lie within the upper bounds of Theorem 5.7, which assumes Assumptions D.1 and D.2, and of Theorem 5.8, which also assumes Assumptions D.4 and D.5. The table also reports forgetting, defined as the average change in FPR@95 on an environment between the time it is seen and the end of the stream. Negative values mean that detection on earlier environments improved. RL forgets less than GD on the gradual streams and on DomainNet. Figure 4 shows FPR@95 in every environment of the three gradual streams. On Core50, RL is at or below GD in all 11 environments.

D.11 Value-function Ablation Across Datasets

To test whether the learned value function can be replaced by a better-tuned penalty, we sweep the OOD weight λ_{Sem} of Eq. **??** for both optimizers. Table 9 gives the full sweep on the gradual streams, and Table 10 compares the best setting of each optimizer on every dataset and modality. The learned value beats the best fixed penalty on Core50 and iCubWorld in both modalities. On OpenLORIS, DomainNet, and CLEAR the best settings are within about one point of each other, and the difference is significant only in a few visual-feature cases with small margins.

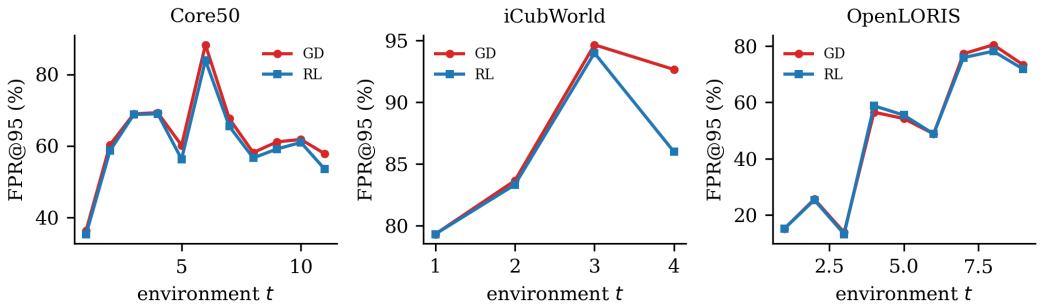


Figure 4: FPR@95 of GD (red) and RL (blue) in every environment of the gradual streams, multimodal CLIP-B/16, averaged over 5 seeds.

Table 7: Energy bridge (RL−GD mean semantic-OOD energy) and value monotonicity ($\partial V/\partial \mathrm{FPR}$), multimodal CLIP-B/16, mean ± std over environments and seeds.

Dataset	Energy bridge	% env. > 0	$\partial V/\partial \mathrm{FPR}$	% env. < 0
Core50	0.355 ± 0.270	95	-0.084 ± 0.067	85
iCubWorld	0.313 ± 0.269	90	-0.100 ± 0.077	90
OpenLORIS	0.108 ± 0.098	93	-0.067 ± 0.063	80
DomainNet	0.011 ± 0.013	80	-0.031 ± 0.044	70
CLEAR-10	0.015 ± 0.021	84	-0.015 ± 0.044	60
CLEAR-100	0.002 ± 0.009	68	-0.000 ± 0.040	60

Table 8: Effective-gradient-flow gap (RL−GD) over inner steps and over environments, and forgetting of GD and RL in FPR@95 points, multimodal CLIP-B/16, mean ± std over 5 seeds. p is the paired bootstrap probability that RL does not forget less than GD.

Dataset	$\mathrm{EGF}^{\mathrm{train}}$ gap	$\mathrm{EGF}^{\mathrm{env}}$ gap	Forgetting GD	Forgetting RL	p
Core50	-0.0039 ± 0.0018	-0.0039 ± 0.0022	-3.9 ± 0.5	-4.5 ± 1.0	0.044
iCubWorld	-0.0065 ± 0.0024	-0.0059 ± 0.0027	0.3 ± 0.6	-1.9 ± 2.0	0.006
OpenLORIS	-0.0015 ± 0.0012	-0.0013 ± 0.0012	0.4 ± 0.4	-1.1 ± 1.0	0.001
DomainNet	-0.0007 ± 0.0003	-0.0006 ± 0.0003	-0.2 ± 0.2	-0.6 ± 0.3	<0.001
CLEAR-10	-0.0001 ± 0.0003	-0.0002 ± 0.0002	-0.2 ± 0.1	-0.1 ± 0.1	0.990
CLEAR-100	-0.0002 ± 0.0008	-0.0001 ± 0.0009	-0.1 ± 0.1	-0.0 ± 0.1	1.000

D.12 Inner-step Ablation

Table 11 reports the FPR@95 gap between RL and GD as the number of inner steps per environment varies. On Core50 and iCubWorld the gap rises steadily with K, and on OpenLORIS it becomes significant at $K = 50$, as expected from $A^{\mathrm{FPR}}_{t,K}$ in Eq. 3, which sums one contribution per inner step. On DomainNet and CLEAR the gap stays near zero for every K.

D.13 Loss Gap

Definition 5.1 compares the objective $\mathcal{L}_{OW}$ at the parameters $\theta^{\mathrm{RL}}_{t,*}$ and $\theta^{\mathrm{GD}}_{t,*}$ reached by the two optimizers, and Theorem 5.3 bounds this gap F^{OW}_t under Assumptions D.1 and D.2. For every environment we run RL and GD for K steps from the same head, with a value network trained within that environment, and measure F^{OW}_t at the two end points. It is at most of order 10^{-3} on every dataset (Table 12). So, the value correction leaves the training objective essentially unchanged. The right-hand side of the theorem contains the contestant $\eta_{K,r}$, which cannot be measured. So, we evaluate the expansion on which the proof rests instead. The proof expands $\mathcal{L}_{OW}$ to second order around $\theta^{\mathrm{GD}}_{t,*}$, drops the first-order term because $\theta^{\mathrm{GD}}_{t,*}$ is the optimized parameter, and bounds the Hessian by its largest eigenvalue $\lambda_{\max}$. After $K = 30$ steps the gradient at $\theta^{\mathrm{GD}}_{t,*}$ does not vanish. So, we keep the first-order term and compare F^{OW}_t with $\langle \nabla \mathcal{L}_{OW}(\theta^{\mathrm{GD}}_{t,*}), \Delta\theta \rangle + \frac{1}{2}\lambda_{\max}\|\Delta\theta\|^2$, where

Table 9: FPR@95 (%) of GD and RL as the weight λ_{Sem} is swept, multimodal CLIP-B/16, mean $\pm$ std over environments and seeds.

Dataset	λ_{Sem}	GD	RL
Core50	0.1	65.3 ± 12.4	63.0 ± 12.5
Core50	0.25	64.0 ± 12.1	62.3 ± 12.5
Core50	0.5	62.7 ± 11.9	60.7 ± 12.1
Core50	1.0	60.2 ± 11.5	58.5 ± 11.9
Core50	2.0	56.5 ± 11.2	55.3 ± 11.3
iCubWorld	0.1	87.9 ± 6.7	85.8 ± 5.8
iCubWorld	0.25	87.8 ± 6.6	85.8 ± 5.8
iCubWorld	0.5	87.6 ± 6.4	85.7 ± 5.8
iCubWorld	1.0	87.5 ± 6.3	85.6 ± 6.0
iCubWorld	2.0	86.8 ± 6.2	85.2 ± 6.2
OpenLORIS	0.1	50.8 ± 25.3	50.4 ± 24.9
OpenLORIS	0.25	50.3 ± 24.9	50.1 ± 24.8
OpenLORIS	0.5	49.5 ± 24.5	49.2 ± 24.4
OpenLORIS	1.0	48.1 ± 24.2	47.5 ± 23.8
OpenLORIS	2.0	44.3 ± 22.6	44.6 ± 22.9

Table 10: Best setting of λ_{Sem} for each optimizer. FPR@95 (%), mean $\pm$ std over environments and seeds, the difference GD$-$RL in points, and the paired bootstrap probability that RL does not beat the best fixed penalty.

Dataset	Modality	best GD	best RL	GD$-$RL	p
Core50	multimodal	56.5 ± 11.2	55.3 ± 11.3	+1.21	0.008
Core50	visual	44.9 ± 11.6	43.7 ± 11.5	+1.21	0.003
iCubWorld	multimodal	86.8 ± 6.2	85.2 ± 6.2	+1.67	0.018
iCubWorld	visual	77.2 ± 7.5	71.6 ± 7.6	+5.67	<0.001
OpenLORIS	multimodal	44.3 ± 22.6	44.6 ± 22.9	-0.22	0.629
OpenLORIS	visual	55.4 ± 31.2	54.5 ± 29.8	+0.88	0.088
DomainNet	multimodal	53.1 ± 27.5	53.1 ± 27.5	+0.01	0.458
DomainNet	visual	57.5 ± 25.5	57.3 ± 25.4	+0.14	0.012
CLEAR-10	multimodal	2.8 ± 1.3	2.8 ± 1.2	+0.01	0.348
CLEAR-10	visual	3.2 ± 1.3	3.1 ± 1.2	+0.12	0.026
CLEAR-100	multimodal	0.6 ± 0.3	0.6 ± 0.3	+0.04	0.265
CLEAR-100	visual	0.9 ± 0.8	0.8 ± 0.8	+0.04	0.200

Table 11: FPR@95 gap GD$-$RL in points for each number of inner steps K, multimodal CLIP-B/16. Entries are the mean over per-(seed, environment) gaps, with the 95% bootstrap interval in brackets.

Dataset	$K = 5$	$K = 10$	$K = 20$	$K = 30$	$K = 50$
Core50	-0.06 [-0.45, +0.30]	+0.15 [-0.42, +0.71]	+0.88 [+0.08, +1.68]	+2.02 [+0.98, +3.08]	+4.15 [+2.59, +5.86]
iCubWorld	+0.25 [+0.00, +0.75]	+0.33 [-0.17, +1.00]	+1.08 [+0.08, +2.25]	+1.92 [+0.42, +3.75]	+3.33 [+1.33, +5.50]
OpenLORIS	-0.24 [-0.52, +0.03]	-0.37 [-0.86, +0.08]	+0.05 [-0.84, +0.98]	+0.29 [-0.71, +1.26]	+1.82 [+0.08, +3.70]
DomainNet	-0.03 [-0.08, +0.00]	-0.03 [-0.08, +0.00]	+0.00 [-0.08, +0.09]	-0.18 [-0.34, -0.04]	-0.03 [-0.36, +0.33]
CLEAR-10	+0.03 [+0.00, +0.07]	+0.04 [+0.00, +0.12]	+0.03 [-0.03, +0.08]	+0.03 [-0.05, +0.09]	-0.11 [-0.20, -0.03]
CLEAR-100	+0.00 [+0.00, +0.00]	+0.02 [+0.00, +0.06]	+0.02 [+0.00, +0.06]	+0.04 [+0.00, +0.10]	-0.02 [-0.08, +0.04]

$\Delta\theta = \theta^{\text{RL}}_{t,*} - \theta^{\text{GD}}_{t,*}$ and $\lambda_{\max}$ is estimated by power iteration. This quadratic estimate holds in 15% to 48% of environments. So, at the small scale of F_t^{OW} the curvature between the two end points is not captured by $\lambda_{\max}$ at $\theta^{\text{GD}}_{t,*}$. Checking the theorem itself would require running GD to convergence inn each environment, which lies outside the online protocol.

D.14 Covariate-shift Protocol and Full Results

We run the covariate experiments on the four benchmarks that are not saturated under frozen features, namely Core50, iCubWorld, OpenLORIS, and DomainNet. CLEAR-10 and CLEAR-100 are left out because every method is already at the FPR floor there.

Table 12: Loss gap F_t^{OW} between RL and GD, the second-order estimate evaluated with the measured first-order term, the fraction of environments where F_t^{OW} is below the estimate, and the top Hessian eigenvalue, multimodal CLIP-B/16, mean $\pm$ std over environments and seeds.

Dataset	F_t^{OW} ($\times 10^{-3}$)	second-order estimate ($\times 10^{-3}$)	% env. bound holds	$\lambda_{\max}$
Core50	-0.17 $\pm$ 2.32	-2.61 $\pm$ 3.34	25	0.063 $\pm$ 0.016
iCubWorld	-0.52 $\pm$ 1.59	-4.03 $\pm$ 4.14	15	0.158 $\pm$ 0.033
OpenLORIS	-0.49 $\pm$ 3.67	-1.33 $\pm$ 5.26	36	0.108 $\pm$ 0.060
DomainNet	-0.02 $\pm$ 0.47	-0.53 $\pm$ 0.91	30	0.035 $\pm$ 0.016
CLEAR-10	0.00 $\pm$ 1.45	-0.62 $\pm$ 1.37	36	0.031 $\pm$ 0.006
CLEAR-100	-0.16 $\pm$ 0.92	-0.37 $\pm$ 0.70	48	0.013 $\pm$ 0.003

Corruption and embedding. For every known-class image of a dataset, a corrupted copy is written at each severity with the same relative path, file name, and file format as the original. The corruption is additive Gaussian noise on pixel values in $[0, 1]$ followed by clipping, with the standard deviations of Table 13. Each image is corrupted with its own random generator, seeded from the global seed, the severity, and a checksum of the relative path of the image. The corruption of an image therefore depends only on the image, the seed, and the severity, and not on processing order or on the other files in the dataset. The copies are embedded with the same pipeline and backbone as the clean images and matched to their clean counterparts by image path. The embedding script verifies that at least 99% of corrupted images match a clean image and stops otherwise. The seeds, library versions, and match statistics are saved with the embeddings.

Mixing. Each environment realizes the three components of the wild distribution. In environment t, half of the known-class samples of both the adaptation split and the test split are replaced by the corrupted copy of the same image at severity s_t, drawn only among samples that have a corrupted copy. Labels, the remaining clean samples, and all held-out samples are unchanged. The optimizer is never told which samples are corrupted, and the pseudo semantic-OOD subset is formed from energies exactly as in the other experiments. For a multimodal feature only the visual block would be replaced and the row re-normalized, and we use visual features in all covariate experiments, with CLIP-B/16 in the main text and all five encoders in Table 14. The mixture of every environment is drawn once per seed and shared by GD and RL, so both optimizers adapt on identical data and forgetting is always measured on the same mixture. The run records the schedule and the realized fraction of corrupted samples, and it stops if the corrupted embeddings are missing or cannot be matched.

Metrics. FPR@95 treats every known-class test sample, clean or corrupted, as in-distribution and the held-out classes as OOD. Covariate accuracy is the classification accuracy on the corrupted known-class test samples and measures generalization under covariate shift. Clean accuracy is measured on the remaining known-class test samples. All metrics are computed on the test split of each environment after adaptation and averaged over the environments of the stream.

Schedules. Table 13 lists the noise level of each severity and the schedule of every dataset. The gradual schedule is $s_t = \lfloor 1 + 4t/(T-1) + \frac{1}{2} \rfloor$ for $t = 0, \ldots, T-1$. The abrupt schedule is $s_t = 1$ for $t < \lfloor T/2 \rfloor$ and $s_t = 5$ afterwards. The two schedules share their endpoints and differ in the size of the steps between consecutive environments.

Table 13: Covariate-shift protocol. Top: noise standard deviation at each severity. Bottom: per-environment severities of the two schedules, where T is the number of environments.

Severity	1	2	3	4	5
Gaussian noise std. σ	0.08	0.12	0.18	0.26	0.38

Dataset	T	Gradual schedule	Abrupt schedule
Core50	11	(1, 1, 2, 2, 3, 3, 3, 4, 4, 5, 5)	(1, 1, 1, 1, 1, 5, 5, 5, 5, 5, 5)
iCubWorld	4	(1, 2, 4, 5)	(1, 1, 5, 5)
OpenLORIS	9	(1, 2, 2, 3, 3, 3, 4, 4, 5, 5)	(1, 1, 1, 1, 1, 5, 5, 5, 5, 5)
DomainNet	4	(1, 2, 4, 5)	(1, 1, 5, 5)

Results. Table 14 reports every encoder and schedule, and Figure 5 summarizes the two differences between the optimizers. The value correction lowers the generalization error on the mixed known-class data in all forty runs, that is on every dataset, encoder, and schedule, and it raises accuracy on the corrupted samples in 36 of them. The margins follow the same ordering as the FPR gains of the clean streams, with Core50 largest (2.4 points of covariate accuracy and 0.014 of generalization error with CLIP-B/16) and DomainNet smallest (0.05 points). Accuracy on the clean known samples differs by at most 0.33 points between the two optimizers in any run, so the improvement comes from the corrupted samples and not from a different trade-off on the clean ones.

Semantic-OOD detection does not improve in this mixture. The FPR difference between the two optimizers ranges from -2.44 to 1.33 points across the forty runs, with RL behind on Core50 and ahead on iCubWorld, and the absolute FPR of both rises well above the clean streams of the main paper. Two things contribute to this. The corrupted samples are harder to separate from semantic OOD, and, because they are counted as in-distribution, they also move the threshold at which FPR@95 is read. This protocol therefore measures rejection under a harder in-distribution set than the clean streams of the main paper, and the two FPR numbers should not be compared directly. The gradual and abrupt schedules give similar results throughout, so these experiments do not separate the pace of the shift at this noise level.

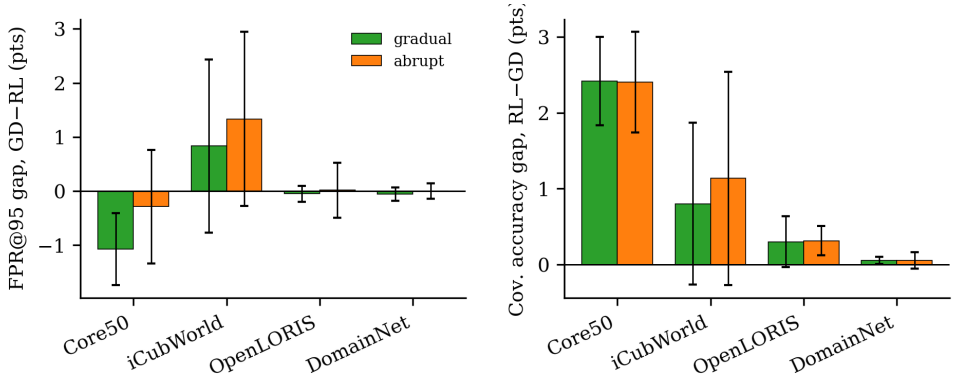


Figure 5: Covariate shift under the gradual and abrupt schedules (visual CLIP-B/16). Left: FPR@95 improvement of RL over GD. Right: covariate-accuracy difference between RL and GD. Error bars show the standard deviation over 5 seeds.

D.15 Additional Benchmarks

Table 15 reports Terra Incognita and Century-of-Portraits. On both, every method, RL and all baselines included, has an FPR@95 above 90%, and RL has an AUROC of at most 0.6. The frozen features do not encode the difference between in-distribution and held-out classes in these streams, so no adaptation of the head can separate them. Recovering this separation would require representation learning, which lies outside the head-only setting analyzed in Section 4.5, and these two benchmarks are therefore not informative about the optimizer.

Table 14: Covariate shift for every encoder and schedule (visual features, half of the known-class samples corrupted), mean ± std over 5 seeds. FPR@95 and covariate accuracy are percentages, and GErr is the cross-entropy on the mixed known-class test data.

Dataset	Encoder	Schedule	FPR GD	FPR RL	GErr GD	GErr RL	Cov. acc. GD	Cov. acc. RL
Core50	CLIP-B/16	gradual	88.7 ± 1.2	89.8 ± 0.8	0.608 ± 0.008	0.592 ± 0.008	57.8 ± 0.9	60.2 ± 1.5
Core50	CLIP-B/16	abrupt	86.6 ± 1.2	86.8 ± 0.7	0.644 ± 0.008	0.627 ± 0.008	53.7 ± 0.7	56.1 ± 1.2
Core50	CLIP-L/14	gradual	85.8 ± 1.6	86.8 ± 1.9	0.487 ± 0.007	0.467 ± 0.003	67.1 ± 0.4	69.2 ± 0.4
Core50	CLIP-L/14	abrupt	81.3 ± 2.5	81.2 ± 3.4	0.568 ± 0.005	0.546 ± 0.001	59.9 ± 0.3	61.7 ± 0.6
Core50	SigLIP	gradual	93.1 ± 1.4	93.2 ± 1.2	0.573 ± 0.008	0.562 ± 0.008	58.9 ± 0.7	61.9 ± 0.5
Core50	SigLIP	abrupt	90.9 ± 1.5	90.9 ± 1.1	0.598 ± 0.004	0.588 ± 0.004	55.2 ± 0.6	57.5 ± 0.2
Core50	SigLIP2	gradual	63.5 ± 0.9	65.3 ± 2.9	0.587 ± 0.005	0.575 ± 0.005	57.9 ± 0.6	59.7 ± 1.0
Core50	SigLIP2	abrupt	62.6 ± 1.0	65.1 ± 3.6	0.613 ± 0.002	0.602 ± 0.002	53.4 ± 0.5	55.1 ± 0.4
Core50	ALIGN	gradual	86.2 ± 0.8	87.0 ± 1.0	0.635 ± 0.005	0.625 ± 0.006	55.4 ± 0.5	56.4 ± 0.6
Core50	ALIGN	abrupt	80.8 ± 1.3	81.1 ± 1.2	0.653 ± 0.004	0.642 ± 0.006	52.5 ± 0.2	53.3 ± 0.4
iCubWorld	CLIP-B/16	gradual	96.9 ± 0.3	96.1 ± 1.6	0.732 ± 0.010	0.732 ± 0.009	48.7 ± 0.9	49.5 ± 1.5
iCubWorld	CLIP-B/16	abrupt	96.2 ± 0.6	94.9 ± 1.7	0.698 ± 0.009	0.697 ± 0.010	52.1 ± 1.1	53.2 ± 1.0
iCubWorld	CLIP-L/14	gradual	95.7 ± 1.2	95.7 ± 1.5	0.619 ± 0.013	0.610 ± 0.011	58.5 ± 1.6	59.1 ± 2.0
iCubWorld	CLIP-L/14	abrupt	92.6 ± 1.0	92.7 ± 1.5	0.593 ± 0.012	0.582 ± 0.010	58.6 ± 1.2	58.9 ± 1.6
iCubWorld	SigLIP	gradual	100.0 ± 0.0	100.0 ± 0.0	0.664 ± 0.007	0.654 ± 0.007	53.5 ± 0.6	54.5 ± 1.4
iCubWorld	SigLIP	abrupt	100.0 ± 0.0	100.0 ± 0.0	0.636 ± 0.010	0.624 ± 0.010	53.2 ± 1.4	54.9 ± 0.8
iCubWorld	SigLIP2	gradual	97.3 ± 0.6	96.8 ± 0.7	0.632 ± 0.010	0.630 ± 0.008	54.6 ± 2.3	55.8 ± 2.3
iCubWorld	SigLIP2	abrupt	97.2 ± 0.5	96.8 ± 0.7	0.609 ± 0.010	0.607 ± 0.008	56.3 ± 1.9	57.2 ± 1.7
iCubWorld	ALIGN	gradual	97.8 ± 1.2	97.8 ± 1.1	0.667 ± 0.007	0.663 ± 0.006	56.3 ± 1.7	57.1 ± 1.8
iCubWorld	ALIGN	abrupt	96.5 ± 0.9	96.2 ± 1.0	0.641 ± 0.007	0.636 ± 0.007	58.9 ± 1.5	60.5 ± 1.7
OpenLORIS	CLIP-B/16	gradual	94.3 ± 0.7	94.4 ± 0.8	0.909 ± 0.006	0.906 ± 0.006	57.9 ± 0.6	58.2 ± 0.4
OpenLORIS	CLIP-B/16	abrupt	87.1 ± 1.0	87.1 ± 1.4	0.935 ± 0.006	0.931 ± 0.006	53.8 ± 0.3	54.1 ± 0.3
OpenLORIS	CLIP-L/14	gradual	87.4 ± 0.9	87.4 ± 1.0	0.802 ± 0.003	0.801 ± 0.003	64.2 ± 0.6	64.2 ± 0.6
OpenLORIS	CLIP-L/14	abrupt	79.0 ± 0.6	78.9 ± 1.0	0.859 ± 0.004	0.858 ± 0.004	57.6 ± 0.7	57.9 ± 0.2
OpenLORIS	SigLIP	gradual	78.8 ± 1.9	79.2 ± 2.3	0.896 ± 0.007	0.891 ± 0.005	55.0 ± 0.3	55.5 ± 0.5
OpenLORIS	SigLIP	abrupt	63.4 ± 1.8	63.7 ± 2.0	0.933 ± 0.003	0.926 ± 0.003	51.0 ± 0.8	51.4 ± 0.8
OpenLORIS	SigLIP2	gradual	75.8 ± 2.1	75.9 ± 2.3	0.735 ± 0.006	0.735 ± 0.006	63.7 ± 0.9	63.6 ± 0.5
OpenLORIS	SigLIP2	abrupt	62.9 ± 1.1	62.8 ± 1.5	0.796 ± 0.004	0.796 ± 0.004	56.7 ± 0.5	56.8 ± 0.7
OpenLORIS	ALIGN	gradual	71.0 ± 1.9	71.4 ± 3.3	0.916 ± 0.007	0.913 ± 0.005	52.1 ± 0.7	52.4 ± 0.6
OpenLORIS	ALIGN	abrupt	60.7 ± 2.1	62.2 ± 3.0	0.962 ± 0.007	0.957 ± 0.008	47.7 ± 0.9	47.8 ± 0.9
DomainNet	CLIP-B/16	gradual	67.4 ± 1.1	67.4 ± 1.0	0.373 ± 0.006	0.372 ± 0.005	88.0 ± 0.4	88.1 ± 0.4
DomainNet	CLIP-B/16	abrupt	67.6 ± 0.6	67.6 ± 0.7	0.409 ± 0.006	0.408 ± 0.006	86.5 ± 0.4	86.6 ± 0.3
DomainNet	CLIP-L/14	gradual	49.9 ± 1.4	49.9 ± 1.4	0.246 ± 0.003	0.246 ± 0.003	92.6 ± 0.4	92.7 ± 0.4
DomainNet	CLIP-L/14	abrupt	48.8 ± 1.1	48.8 ± 1.1	0.273 ± 0.002	0.272 ± 0.002	91.5 ± 0.3	91.5 ± 0.3
DomainNet	SigLIP	gradual	40.2 ± 1.3	40.3 ± 1.5	0.214 ± 0.004	0.214 ± 0.004	94.2 ± 0.3	94.2 ± 0.3
DomainNet	SigLIP	abrupt	45.4 ± 1.2	45.4 ± 1.4	0.238 ± 0.004	0.238 ± 0.004	92.7 ± 0.3	92.7 ± 0.3
DomainNet	SigLIP2	gradual	35.9 ± 1.5	36.0 ± 1.9	0.199 ± 0.005	0.199 ± 0.006	94.1 ± 0.5	94.1 ± 0.5
DomainNet	SigLIP2	abrupt	38.3 ± 1.4	38.4 ± 1.5	0.220 ± 0.004	0.220 ± 0.004	93.1 ± 0.4	93.1 ± 0.4
DomainNet	ALIGN	gradual	46.4 ± 1.8	46.3 ± 1.7	0.324 ± 0.004	0.324 ± 0.004	90.4 ± 0.5	90.4 ± 0.5
DomainNet	ALIGN	abrupt	48.2 ± 2.1	48.2 ± 2.1	0.368 ± 0.006	0.367 ± 0.006	88.0 ± 0.5	88.0 ± 0.5

Table 15: Terra Incognita and Century-of-Portraits. FPR@95 (%) and RL AUROC, mean ± std over 5 seeds, multimodal CLIP-B/16.

Dataset	Tent	EATA	en-FT	SAR	CoTTA	GD	RL	RL AUROC
TerraIncognita	93.2 ± 1.2	92.3 ± 1.0	91.6 ± 1.0	92.7 ± 0.8	91.7 ± 1.2	**91.4 ± 1.2**	92.5 ± 0.8	0.597 ± 0.006
Portraits	95.5 ± 0.3	95.5 ± 0.3	95.5 ± 0.2	**95.4 ± 0.3**	95.5 ± 0.3	95.7 ± 0.3	95.5 ± 0.2	0.540 ± 0.001