How Representations Shape Label Requirements

Anonymous authors
Paper under double-blind review

Abstract

Pretraining can reduce the need for labels, but what makes a representation label-efficient? Labels reveal both what to predict and which inputs should share a prediction. We isolate these roles in a regression family whose candidate groupings have equal response complexity. Choosing a grouping before labels arrive can make a target risk unattainable, regardless of label count. A frozen representation retaining both candidates allows a supervised readout to overcome this structural barrier, at a higher estimation cost. We give upper and lower bounds on optimal risk within the committed class that match within a factor of three, and computable label budgets for known Gaussian tasks. Experiments on electrocardiograms, ImageNet, and robot actions vary input coverage and encoder update permissions. Under fixed recipes and on the measured grids, adaptation reaches specified mean-risk targets with one-tenth as many labels on ImageNet and robot actions.

1 Introduction

Learning representations from unlabeled inputs can make a small labeled set useful for downstream prediction (He et al., 2022; Lee et al., 2021). A learner that falls short of its target then faces several investments: more labels, more relevant inputs, or more computation for representation learning and supervised adaptation. Choosing among them requires knowing what remains unresolved. Poorly estimated responses and an uncertain sharing structure both cause prediction error, yet they need not respond to the same additional resource.

How many labels suffice at a given input and computation budget? A label fraction records the amount of supervision, while its usefulness also depends on the structure available to the learner. If a grouping is already reliable, labels can concentrate on estimating its responses. If important sharing relations remain uncertain, collecting labels helps only to the extent that the learning procedure can use their evidence. This raises a consequential distinction: a target may require better estimates within an existing representation, or access to prediction functions that an early structural decision excluded.

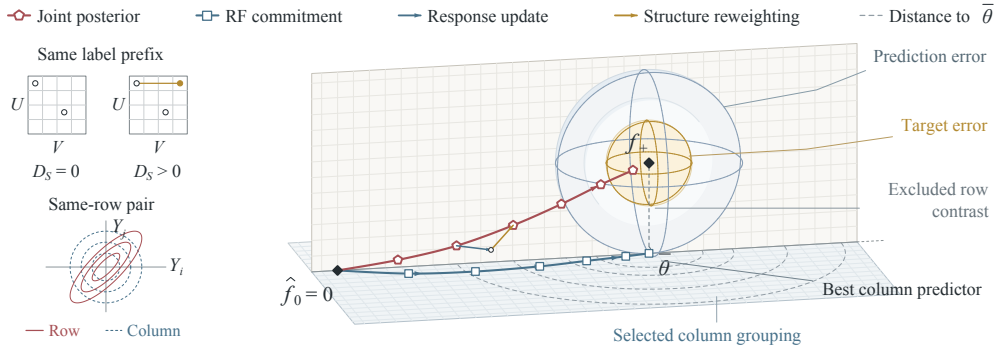


Figure 1: Labels refine responses and identify sharing. Left: D_S counts labeled pairs with different row/column sharing relations. Right: wrong commitment excludes directions needed to reach the target-error region. Joint updates responses and grouping weights from the same labels.

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.

Consider a simple task: inputs have a row and a column, and the unknown response is shared either within rows or within columns. Both possibilities require estimating equally many response values. Labels estimate these values, while patterns across labels reveal the grouping. Keeping only row membership permits better row estimates but cannot recover column distinctions. Keeping both memberships leaves that choice to supervision. This construction separates response complexity from evidence about sharing.

The choice can determine whether a target error is attainable at all. With four responses, unit response variance, noise variance 0.25, and uninformative contexts, every learner committed to one grouping before labels has average risk at least 0.375. A frozen representation retaining both groupings admits an explicit readout with risk at most 0.1 using 78 random labels. The task and observation rules are identical. Retention uses seven effective linear dimensions rather than four, with a corresponding estimation cost. Thus, freezing features need not close off supervised structure choice. Upper and lower bounds matching within a factor of three quantify the committed class's best risk and yield label, input, and record-access requirements.

Neural experiments probe this resource question through input coverage and encoder update permissions. Across electrocardiogram (ECG) diagnosis, image classification, and robot action prediction, more distinct pretraining inputs improve risk in all twelve primary comparisons with labels and training presentations fixed. From shared checkpoints and under fixed recipes, adaptation reaches specified mean-risk targets with one-tenth as many labels as frozen encoders on the measured ImageNet and robot grids. ECG risk improves without crossing an earlier main budget tier. Robotic execution tests assess the practical outcomes of these choices. Local risk prediction tests whether an input benefit persists at another label budget.

Our results provide three complementary answers. The finite model quantifies early commitment and shows how candidate retention changes target attainability (Eqs. 3 and 7). For a known representation and Gaussian task, explicit bounds turn label-direction geometry into computable budgets (Eq. 8). Neural studies measure label frontiers under fixed training recipes. Together, they distinguish the evidence labels provide from the prediction functions available to use it: label efficiency depends on which structural choices remain open to supervision.

2 Related Work

Unlabeled inputs help when their geometry supports the task. Analyses link label requirements to manifolds (Moscovich et al., 2017), reconstruction (Lee et al., 2021), latent classes (Arora et al., 2019), augmentation graphs (HaoChen et al., 2021), and redundant views (Tosh et al., 2021). Alignment and uniformity describe feature geometry (Wang & Isola, 2020); view selection controls retained task information (Tian et al., 2020). We separate response complexity from evidence about sharing.

Identifying latent factors requires assumptions (Locatello et al., 2019); augmentation pairs can identify invariant content under a specified model (von Kügelgen et al., 2021). Shared representations can reduce per-task supervision (Baxter, 2000; Maurer et al., 2016; Du et al., 2021), with guarantees governed by task diversity (Tripuraneni et al., 2020) and subspace estimation (Tripuraneni et al., 2021). Here labels resolve equally complex groupings within one task. Candidate retention keeps that choice open in fixed features, at a measurable estimation cost.

The supervised procedure also matters. Analyses account for contrastive inductive bias (Saunshi et al., 2022), SimCLR pretraining and fine-tuning (Zhang & Cao, 2026), two-stage representation symmetries (Tinati & Tu, 2026), and loss-dependent multi-objective sample complexity (Wegel et al., 2025). We bound the best risk among response estimators restricted to a grouping selected before labels.

Empirical pretraining uses contrastive objectives (Chen et al., 2020a; He et al., 2020), redundancy reduction (Zbontar et al., 2021), self-distillation (Caron et al., 2021), and image or point-cloud reconstruction (He et al., 2022; Pang et al., 2022). Large pretrained models (Chen et al., 2020b) and consistency-based pseudo-labeling (Sohn et al., 2020) improve low-label prediction through different procedures. CLOCS uses cardiac-signal views (Kiyasseh et al., 2021); R3M and VIP transfer video representations to robotic policies and rewards (Nair et al., 2023; Ma et al., 2023). Label efficiency depends on architecture, task, and adaptation protocol (Kornblith et al., 2019; Newell &

Deng, 2020; Raghu et al., 2019). Pretraining benchmarks and scaling studies vary data and model resources (Goyal et al., 2019; Zhai et al., 2022); feature distortion can limit fine-tuning under shift (Kumar et al., 2022). We match checkpoints and labels, fixing pretraining presentations when varying distinct inputs. Domain, granularity, and update-depth controls accompany target-risk frontiers under fixed recipes.

Experimental design optimizes A-optimal subsets by proportional volume sampling (Nikolov et al., 2019) or Bayesian objectives by regularized determinantal sampling (Dereziński et al., 2020). We bound risk under prescribed random masks. Error surfaces (Rosenfeld et al., 2020) and effective downstream data (Hernandez et al., 2021) motivate forecasting; our three-source comparison measures local input–label interaction at a held-out configuration.

3 Method

3.1 Separating response estimation from structure evidence

Each task has $r \geq 2$ latent responses and two candidate sharing structures. For $i \in [N]$ and $j \in [r] = \{1, \dots, r\}$, labels are noisy observations:

$$\begin{aligned} &\theta_j \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau^2), \quad U_i, V_i \overset{\text{iid}}{\sim} \text{Unif}[r], \quad Z_i \sim \mathcal{N}(g\Delta, 1), \\ &Y_i = f_g(U_i, V_i) + \xi_i, \quad \xi_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), \qquad f_g(u, v) = \begin{cases} \theta_u, & g = +1, \\ \theta_v, & g = -1. \end{cases} \end{aligned} \tag{1}$$

Here g selects row/column sharing; Z_i supplies structure evidence. The parameters $\tau^2, \sigma^2 > 0$ and $\Delta \geq 0$ are known. Before forming labels, inputs, responses, and noise are independent conditional on g. Increasing r adds responses to estimate; decreasing Δ at fixed r reduces structure evidence.

All N inputs are available. Masks are independent of inputs, responses, and noise, with either design:

$$\begin{aligned} &\text{Bernoulli:} \quad S = \{i : \zeta_i = 1\}, \quad \zeta_i \overset{\text{iid}}{\sim} \text{Bern}(\rho), \quad \mathbb{E}|S| = \rho N; \\ &\text{fixed size:} \quad S \sim \text{Unif}\{A \subseteq [N] : |A| = m\}, \quad 0 \leq m \leq N. \end{aligned}$$

Risk averages the response prior, training data, mask, and learner randomness. Only g is worst-case; test prediction receives a fresh uniform (u, v) without a new context:

$$R(\widehat{f}) = \max_{g \in \{+1, -1\}} \mathbb{E}_g \left[\frac{1}{r^2} \sum_{u,v=1}^{r} (\widehat{f}(u, v) - f_g(u, v))^2 \right]. \tag{2}$$

A representation-first (RF) learner chooses a prediction space before reading labels:

$$\begin{aligned} &\mathcal{F}_+ = \{(u, v) \mapsto h_u : h \in \mathbb{R}^r\}, \qquad \mathcal{F}_- = \{(u, v) \mapsto h_v : h \in \mathbb{R}^r\}, \\ &\text{RF}(B): \quad \widehat{g} = \widehat{g}(U_{1:N}, V_{1:N}, Z_{\mathcal{I}}), \quad |\mathcal{I}| \leq B \leq N, \quad \widehat{f} \in \mathcal{F}_{\widehat{g}}. \end{aligned}$$

Selection may be randomized and adaptive under the hard context cap. Any supervised response estimator is allowed within the selected space. RF fixes the prediction space available to supervised fitting; comparisons use the same task and observations.

3.2 Which resource limits the target risk?

Let $a = \sigma^2/\tau^2$ and $c = \tau^2(1 - 1/r)$. Oracle estimation and context testing contribute

$$R_{\text{or}} = \tau^2 \mathbb{E} \frac{a}{a + K}, \quad K \sim \text{Bin}(N, \rho/r), \qquad p_{\text{err}}(B) = Q(\Delta\sqrt{B}), \quad p_{\text{err}}(0) = \tfrac{1}{2},$$

where Q is the standard normal right tail. For $R^*_{\text{RF}(B)} = \inf_{\widehat{f} \in \text{RF}(B)} R(\widehat{f})$, App. A proves

$$\max\{R_{\text{or}},\ cp_{\text{err}}(B)\} \leq R^*_{\text{RF}(B)} \leq \min\{\tau^2,\ R_{\text{or}} + 2cp_{\text{err}}(B)\}. \tag{3}$$

The bounds match within a factor of three. Wrong-coordinate prediction loses response contrasts of prior variance c; more accurate fitting within that space cannot restore them. Context selection, shrinkage, and zero prediction attain the upper bound.

$$R^*_{\mathrm{RF}(B)} \le \varepsilon \implies R_{\mathrm{or}} \le \varepsilon \text{ and } cp_{\mathrm{err}}(B) \le \varepsilon,$$
$$R_{\mathrm{or}} \le \varepsilon - 2cp_{\mathrm{err}}(B) \implies R^*_{\mathrm{RF}(B)} \le \varepsilon.$$

The sufficient condition applies when its oracle target is positive and attainable. A specified procedure's first crossing gives an achievable budget, distinct from the class optimum.

For $B = N$, $0 < \varepsilon/\tau^2 \le 1/8$, $\rho > 0$, and $\Delta > 0$, the minimum input count obeys

$$N^*_{\mathrm{RF}} \asymp \max\left\{ \underbrace{\frac{r}{\rho}\left[\log\frac{\tau^2}{\varepsilon} + \frac{\sigma^2}{\varepsilon}\right]}_{\text{response coverage and estimation}} , \underbrace{\Delta^{-2}\log\frac{\tau^2}{\varepsilon}}_{\text{structure identification}} \right\}. \tag{4}$$

The comparison constants are universal. Increasing ρ reduces only the first demand. Average risk avoids the extra $\log r$ needed to cover every response simultaneously.

At exactly m random labels, the specified shrinkage rule instead separates as

$$K_m \sim \mathrm{Bin}(m, 1/r), \qquad R_{\mathrm{sh}}(N, m) = R_{\mathrm{or}}(m) + p_{\mathrm{err}}(N)A(m), \quad A(m) \ge 0.$$

More inputs reduce grouping error without changing oracle estimation. The label-dependent wrong-grouping penalty $A(m)$ determines how input gains transfer across label budgets.

With a pre-indexed random label list, one query reads a context or labeled record. In the same small-risk range, with sufficient availability and $\sigma^2/\varepsilon \ge \log(\tau^2/\varepsilon)$,

$$T^*_{\mathrm{RF,query}} \asymp r\frac{\sigma^2}{\varepsilon} + \Delta^{-2}\log\frac{\tau^2}{\varepsilon}, \qquad \Delta > 0. \tag{5}$$

A matching arithmetic implementation uses $O(r)$ storage; unindexed scanning adds $O(N)$. App. A.5 specifies availability.

3.3 When supervision changes what is attainable

Joint inference leaves both candidates available to the same labels (Fig. 2; Alg. 1). With observed training data $\mathcal{D}$, equal structure priors, branch means $\widehat{\theta}^+, \widehat{\theta}^-$, and all $B = N$ contexts,

$$\pi = \Pr(g = +1 \mid \mathcal{D}), \qquad \widehat{f}_{\mathrm{joint}}(u, v) = \pi\widehat{\theta}^+_u + (1-\pi)\widehat{\theta}^-_v.$$

Its exact risk decomposes into response estimation and remaining structural disagreement:

$$R^*_{\mathrm{joint}} = R_{\mathrm{or}} + \mathbb{E}[\pi(1-\pi)D], \qquad D = \frac{1}{r^2}\sum_{u,v}(\widehat{\theta}^+_u - \widehat{\theta}^-_v)^2. \tag{6}$$

The expectation uses the equal-prior mixture; symmetry gives optimality for maximum-direction risk. Label relations identify sharing: for distinct i, j,

$$\mathrm{Cov}(Y_i, Y_j \mid U_i = U_j,\ V_i \ne V_j,\ g) = \tau^2\mathbf{1}\{g = +1\}.$$

A frozen representation can retain this opportunity. With standard basis vectors e_j,

$$\phi_{\mathrm{both}}(u, v) = (e_u, e_v), \qquad \mathcal{F}_{\mathrm{both}} = \{\alpha_u + \beta_v : \alpha, \beta \in \mathbb{R}^r\}, \qquad \dim\mathcal{F}_{\mathrm{both}} = 2r - 1.$$

This is the minimum linear space containing both response families. For exactly $m \ge 1$ random labels, the explicit label-only readout of Prop. B1 attains

$$R(\widehat{f}_m) = \frac{C_r}{m}, \qquad C_r = 2(r-1)\tau^2 + (2r-1)\sigma^2. \tag{7}$$

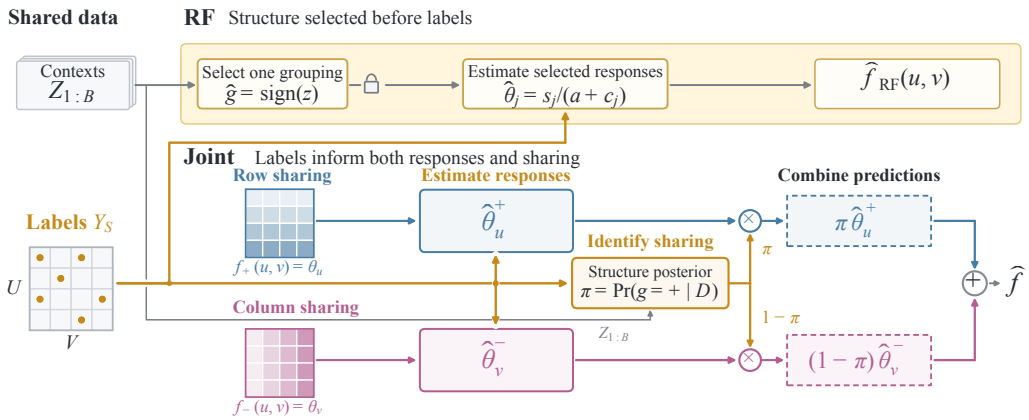


Figure 2: How labels estimate responses and reweight row and column sharing.

Candidate retention changes target attainability, even without informative contexts:

$$m \geq \lceil C_r/\varepsilon \rceil \implies R(\widehat{f}_m) \leq \varepsilon,$$
$$\Delta = 0 \implies R^*_{\text{RF}(B)} \geq c/2.$$

For $(r, \tau^2, \sigma^2) = (4, 1, 0.25)$, $m = 78$ gives

$$R(\widehat{f}_m) \leq 0.1 < 0.375 = c/2.$$

Joint uses contexts and labels; the moment readout uses labels alone. Retention trades projection error for fitting cost (Table S1), while frozen features preserve supervised structure choice. The same frozen space also contains Joint, with weights determined by context and label evidence:

$$\widehat{f}_{\text{joint}} = \langle w_{\mathcal{D}}, \phi_{\text{both}} \rangle,$$
$$w_{\mathcal{D}} = (\pi\widehat{\theta}^+, (1 - \pi)\widehat{\theta}^-).$$

Algorithm 1 RF and Joint

Require: Contexts $Z_{1:B}$, group IDs, labels Y_S; known $r, \Delta, \tau^2, \sigma^2$.

$a = \sigma^2/\tau^2$, $z = \sum_{i=1}^{B} Z_i$.
if RF commitment **then**
  Set $\widehat{g} = \text{sign}(z)$; fair coin at $z = 0$.
  In $\widehat{g}$, count labels c_j and sum s_j.
  $\widehat{\theta}_j = s_j/(a + c_j)$.
  Return $\widehat{f}(u, v) = \widehat{\theta}_u$ if $\widehat{g} = +1$; $\widehat{f}(u, v) = \widehat{\theta}_v$ otherwise.
else *Joint posterior*
  For each $b \in \{+, -\}$, compute c_j^b, s_j^b.
  $\widehat{\theta}_j^b = s_j^b/(a + c_j^b)$.
  $\ell_b = -\frac{1}{2}\sum_j \log(1 + c_j^b/a) + \frac{1}{2\sigma^2}\sum_j (s_j^b)^2/(a + c_j^b)$.
  $\pi = [1 + \exp\{-2\Delta z - \ell_+ + \ell_-\}]^{-1}$.
  Return $\widehat{f}(u, v) = \pi\widehat{\theta}_u^+ + (1 - \pi)\widehat{\theta}_v^-$.
end if

For exactly m labels and a context cap B, the structural excess satisfies

$$R_{\text{joint}}(B, m) - R_{\text{or}}(m) \leq 2\tau^2 e^{-B\Delta^2/2}\Gamma_m.$$

Here Γ_m is the mean affinity between conditional label distributions. Sharing patterns first differ with nonvanishing probability at $m \asymp \sqrt{r}$; evidence strength depends on noise, and prediction still requires response estimation (App. B.5).

3.4 Label requirements beyond a single task family

Candidate retention determines which targets are attainable. We next bound the estimation budget for a known representation and Gaussian task. Fix possibly nonorthogonal $b_i \in \mathbb{R}^d$, prior precision $\Lambda \succ 0$, and prediction weight $W \succeq 0$. Let $\theta \sim \mathcal{N}(0, \Lambda^{-1})$ and $Y_i = b_i^\top \theta + \xi_i$ with independent $\xi_i \sim \mathcal{N}(0, 1)$. For masks independent of the parameter and noise, define

$$G = \sum_{i=1}^{N} b_i b_i^\top, \qquad R_X(S) = \text{tr}\left[W\left(\Lambda + \sum_{i \in S} b_i b_i^\top\right)^{-1}\right],$$
$$F(s) = \text{tr}[W(\Lambda + sG)^{-1}], \qquad \kappa = \max_i b_i^\top \Lambda^{-1} b_i.$$

For $F(1) < \varepsilon < F(0)$, solve $F(s_\varepsilon) = \varepsilon$. Under independent revelation, $\rho^*_X(\varepsilon)$ is the smallest reveal probability with $\mathbb{E}_S R_X(S) \le \varepsilon$:

$$s_\varepsilon \le \rho^*_X(\varepsilon) \le \frac{(1+\kappa)s_\varepsilon}{1+\kappa s_\varepsilon}. \tag{8}$$

The minimum expected label count is $N\rho^*_X$. Prop. C1 bounds the mask-averaged covariance by inverse convexity and linear prediction; the budget ratio is at most $1+\kappa$.

Uniform fixed-size masks replace κ by $N\kappa/(N-1)$ and round to integers ($N \ge 2$). The weighted design spectrum computes F without orthogonality. App. C gives proofs and overlapping graph measurements, alongside a general-prior risk conversion of Frankston et al. (2021).

4 Experiments

We measure how input coverage and encoder update permissions affect label budgets at fixed risk targets.

4.1 Tasks, protocols, and evaluation

Finite regressions isolate structural commitment, while neural interventions vary input coverage and update permissions. Neural tasks use ECG diagnosis (Gow et al., 2023; Wagner et al., 2022), ImageNet classification (ImageNet, 2012), and robotic picking, stacking, and insertion (Tao et al., 2025). Frozen trains learned-query attention and a two-layer nonlinear head with encoder parameters and running statistics fixed. Frozen can retain multiple candidate structures; RF commits to one. Adapt updates the encoder and readout from the same pretrained checkpoint and head initialization. Scratch uses the same architecture from random initialization. All pretraining uses the stated input pools.

Recipes are developed with the full development-label union and then fixed. Configuration budgets count fitting and selection records with their target bundles, separately from development supervision. Robot budgets count actions per task. Neural configurations use 52 paired seed groups and nested masks. The finite main grid uses five seed groups, while Fig. 3 uses its own simulation design (App. E.1). Input sweeps preserve labeled identities and pretraining presentations.

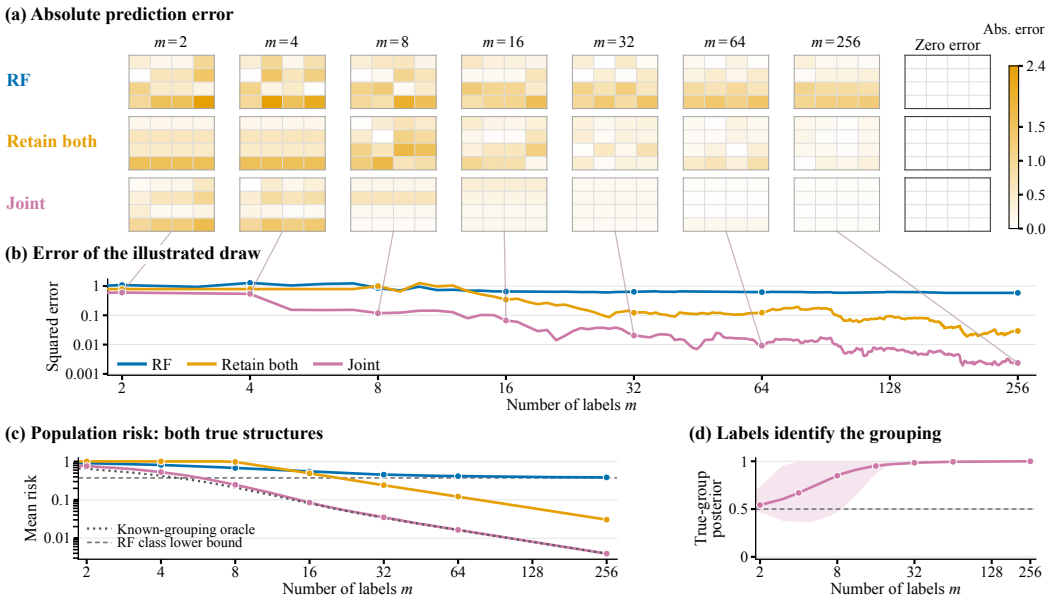


Figure 3: Labels recover responses and identify sharing in fixed candidate features.

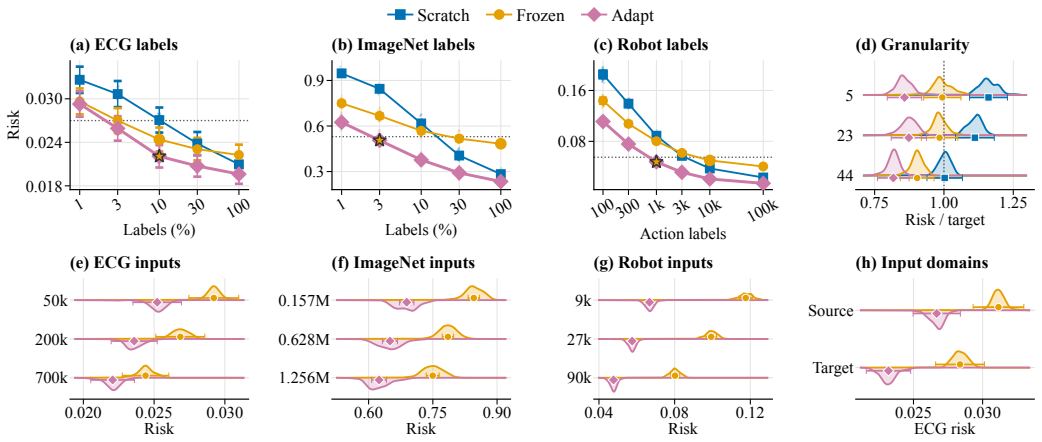


Figure 4: Risk across label budgets, input pools, update rules, and scientific controls.

Neural risk is observed-label loss: ECG label-mean Brier loss, summed ImageNet probability squared error, and seven-dimensional action mean squared error (MSE) averaged equally across tasks. Under a common input interface and test distribution, squared loss equals conditional-mean error plus a common irreducible term (App. D.3). Primary intervals combine approximate seed and shared-test uncertainty components with cluster and family adjustments. For tested budgets $\mathcal{M}$, mean risk $\widehat{R}_m$, and adjusted upper endpoint U_m, define

$$\widehat{m}_{\text{mean}}(\varepsilon) = \min\{m \in \mathcal{M} : \widehat{R}_m \le \varepsilon\}, \qquad \widehat{m}_{\text{adj}}(\varepsilon) = \min\{m \in \mathcal{M} : U_m \le \varepsilon\}. \tag{9}$$

The adjusted-interval frontier uses the upper endpoint, with no change to the mean-risk target. An empty set means no tested budget passes. Targets are 0.027, 0.530, and 0.055 for ECG, ImageNet, and robot actions. Input pool, architecture, and training allocation stay fixed within each label curve. App. D specifies the protocols and accounting.

4.2 How much can the same labels achieve?

In the finite model, stronger input evidence improves the use of fixed supervision. For $r = 8$ and 64 labels, increasing contexts from 64 to 8,192 lowers the larger directional RF simulation mean from 0.4223 to 0.0342, close to the exact oracle risk of 0.0346 (Fig. S1). A separate main-grid $r = 4$ setting with uninformative contexts shows the value of retaining structure choice: at 64 labels, Joint reaches mean risk 0.0155, below RF's 0.375 floor. An independent simulation in Fig. 3 connects response recovery to grouping identification. Its example conditions on incorrect RF commitment. Population curves include all RF decisions and use the larger directional mean. Retain both uses a label-only moment readout, while Joint uses the full posterior.

Broader input coverage also improves neural prediction with unchanged labels and presentation counts. All twelve primary adjacent-pool paired contrasts favor the larger pool. With 12,812 ImageNet labels, increasing distinct inputs from 156,943 to 1,255,544 lowers Frozen risk from 0.8457 to 0.7497 and Adapt risk from 0.6890 to 0.6245. The curves and seed distributions in Fig. 4 show the gains across measured pools. Relevance matters alongside coverage: matched-size target-domain ECG pretraining lowers risk by 8.9% for Frozen and 13.1% for Adapt.

4.3 Adaptation changes the label frontier

Adaptation improves risk at matched label budgets. Relative to Frozen, the reductions are 9.5% for ECG at 10% labels, 33.8% for ImageNet at 10%, and 40.5% for task-average robot risk at 1,000 actions per task. Adapt has lower mean risk and a lower adjusted upper endpoint at every primary label budget. Its measured frontiers therefore cover every target passed by Frozen (Fig. S3).

Table 1 separates mean and adjusted-interval frontiers on the tested grid. At these targets, adaptation uses one-tenth as many labels on ImageNet and robot actions under the mean rule. On ECG,

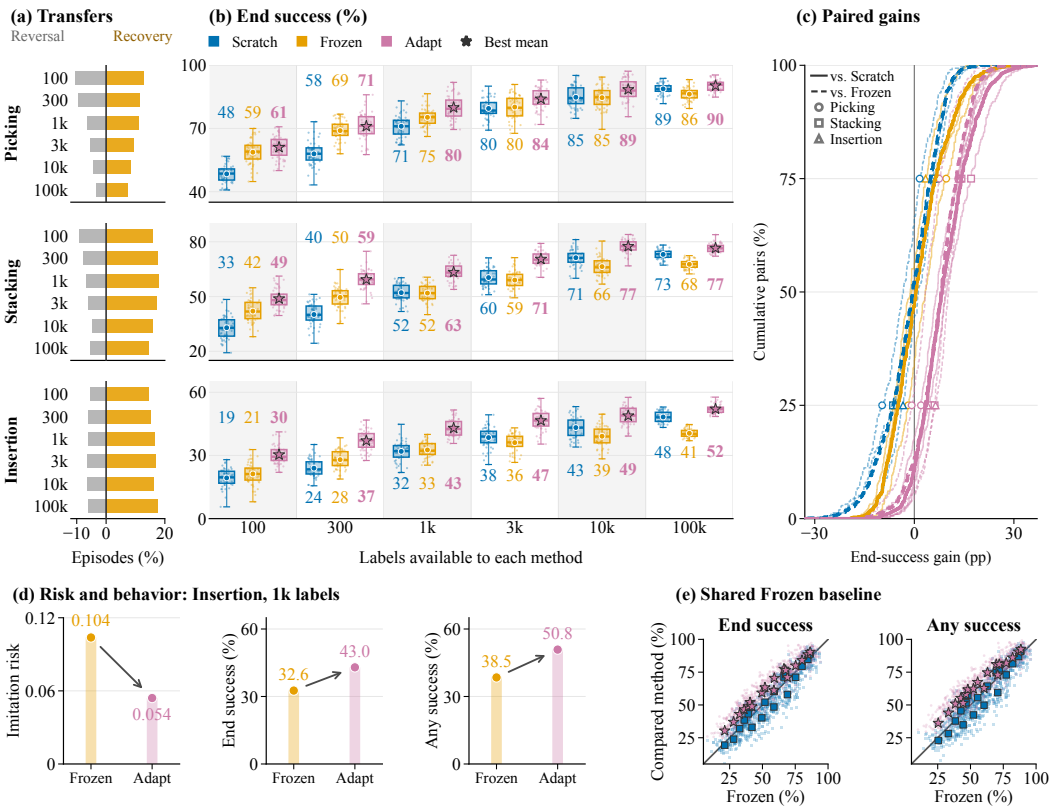


Figure 5: Paired robot outcomes link action-label budgets to endpoint and any-time success.

adaptation lowers risk while the first passing budget stays unchanged; its 5- and 23-target controls pass the interval rule at 10% with Adapt versus 30% with Frozen.

Table 1: First passing label budgets: Frozen → Adapt.

Task	Target risk	Mean rule	Adjusted-interval rule
ECG (44 targets)	0.027	3% → 3%	10% → 10%
ImageNet	0.530	30% → **3%**	100% → **3%**
Robot (per task)	0.055	10,000 → **1,000**	10,000 → **1,000**

Update depth controls which layers labels can change. At 10% ImageNet labels, risk falls from 0.5688 with Frozen to 0.4662 with last-block updates and 0.3767 with full adaptation. Opening earlier encoder layers improves prediction under these fixed training recipes.

4.4 Closed-loop robotic performance

Lower action error is accompanied by better robotic execution. For insertion at 1,000 actions, Adapt reduces risk from 0.104 to 0.054, while endpoint success rises from 32.6% to 43.0% and any-time success from 38.5% to 50.8%. Fig. 5 shows paired recoveries and reversals alongside seed-level outcomes. Adapt has the highest mean endpoint success in all 18 task-budget settings. Full profiles in App. F.4 relate offline risk to both execution metrics, which measure different aspects of performance.

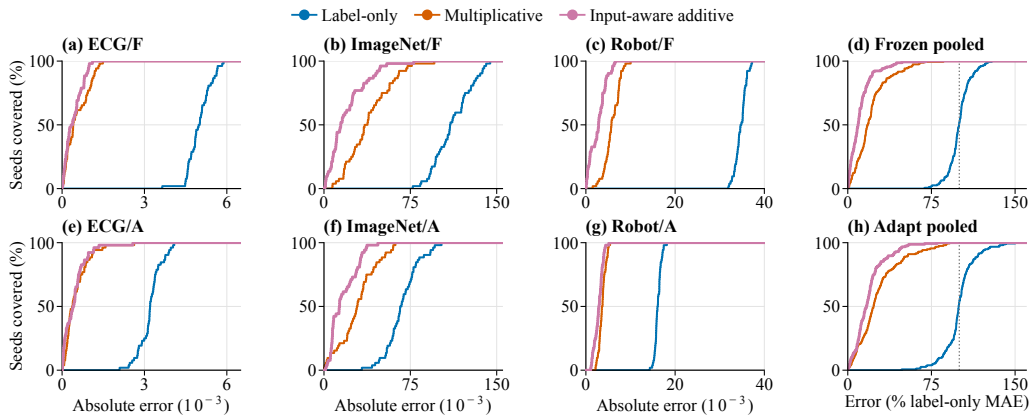


Figure 6: Risk-forecast error coverage across tasks, readouts, and three prediction rules.

4.5 Does an input gain transfer across label budgets?

For $N_s < N_l$ and $m_l < m_h$, define source risks $x = \widehat{R}(N_s, m_l)$, $y = \widehat{R}(N_l, m_l) > 0$, $z = \widehat{R}(N_l, m_h)$, and held-out risk $t = \widehat{R}(N_s, m_h)$. Three coefficient-free forecasts are

$$\widetilde{R}_{\text{add}} = x + z - y, \qquad \widetilde{R}_{\text{mult}} = xz/y, \qquad \widetilde{R}_{\text{label}} = z. \tag{10}$$

The additive rule was prespecified; the multiplicative rule is a retrospective comparison. Additive lowers mean absolute risk-forecast error (MAE) by 75.5%–92.1% against label-only and has lower MAE than the multiplicative rule in all six groups (Fig. 6). Crossed seed/test-cluster intervals favor additive in six and five groups, respectively, with ECG Adapt versus the multiplicative rule spanning zero (App. F.2). The input–label interaction $I = (t - z) - (x - y)$ measures how the input benefit changes between label budgets, so the signed forecast error is $e_{\text{add}} = -I$. App. A.6 characterizes this interaction for the specified RF rule.

5 Conclusion and Limitations

Labels help estimate responses and identify which inputs should share them. Our finite model shows that early commitment to a grouping can impose a risk floor, while retaining both candidates in frozen features lets a supervised readout overcome it. For known Gaussian tasks, we derive computable label-budget bounds from the design geometry. Neural experiments show complementary benefits from broader input coverage and encoder adaptation. Under fixed recipes, adaptation reaches specified mean-risk targets with one-tenth as many labels as frozen encoders on the measured ImageNet and robot grids. Together, these results link label efficiency to the structural choices that remain open to supervision.

These findings also suggest directions for extending the theory and evaluation.

Theoretical scope. The guarantees cover the two-candidate model and known Gaussian designs. Extending them to learned candidate families remains open.

Retention cost. Additional directions can increase estimation cost for the studied moment readouts, creating a trade-off between structural flexibility and label requirements.

Empirical resolution. Label savings depend on training recipes and discrete budget grids. Denser sampling near target crossings would refine the measured frontiers.

Local forecasting. Risk forecasts are evaluated at one held-out combination of input and label budgets per task and protocol. Testing additional combinations would clarify the range of transfer.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing, grammar checking, and literature retrieval. We did not use generative AI tools for substantive research tasks, and other disclosure tasks are not applicable to this work. We reviewed all AI-assisted work and verified relevant information against the original sources. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study uses publicly available datasets. This work does not raise ethical concerns.

REPRODUCIBILITY STATEMENT

The relevant code is provided as supplementary material. Sec. 3 and Apps. A–C specify the models, assumptions, and algorithms. App. D details the data, training, evaluation, and hardware, while Apps. E and F provide extended results. Dataset access follows the original providers' terms.

REFERENCES

S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi. A theoretical analysis of contrastive unsupervised representation learning. In *ICML*, volume 97 of *PMLR*, pp. 5628–5637, 2019.

J. Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.

M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pp. 9650–9660, 2021.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, volume 119 of *PMLR*, pp. 1597–1607, 2020a.

T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton. Big self-supervised models are strong semi-supervised learners. In *NeurIPS*, volume 33, pp. 22243–22255, 2020b.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255, 2009.

M. Dereziński, F. Liang, and M. W. Mahoney. Bayesian experimental design using regularized determinantal point processes. In *AISTATS*, volume 108, pp. 3197–3207, 2020.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

S. S. Du, W. Hu, S. M. Kakade, J. D. Lee, and Q. Lei. Few-shot learning via learning the representation, provably. In *ICLR*, 2021.

K. Frankston, J. Kahn, B. Narayanan, and J. Park. Thresholds versus fractional expectation-thresholds. *Annals of Mathematics*, 194(2):475–495, 2021.

B. Gow, T. Pollard, L. A. Nathanson, A. Johnson, B. Moody, C. Fernandes, N. Greenbaum, J. W. Waks, P. Eslami, T. Carbonati, A. Chaudhari, E. Herbst, D. Moukheiber, S. Berkowitz, R. Mark, and S. Horng. MIMIC-IV-ECG: Diagnostic electrocardiogram matched subset. PhysioNet, 2023. URL https://physionet.org/content/mimic-iv-ecg/1.0/. Version 1.0.

P. Goyal, D. Mahajan, A. Gupta, and I. Misra. Scaling and benchmarking self-supervised visual representation learning. In *ICCV*, pp. 6391–6400, 2019.

J. Z. HaoChen, C. Wei, A. Gaidon, and T. Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *NeurIPS*, volume 34, pp. 5000–5011, 2021.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pp. 9729–9738, 2020.

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pp. 16000–16009, 2022.

D. Hernandez, J. Kaplan, T. Henighan, and S. McCandlish. Scaling laws for transfer. arXiv preprint arXiv:2102.01293, 2021. URL https://arxiv.org/abs/2102.01293.

ImageNet. ImageNet large scale visual recognition challenge 2012, 2012. URL https://www.image-net.org/challenges/LSVRC/2012/.

D. Kiyasseh, T. Zhu, and D. A. Clifton. CLOCS: Contrastive learning of cardiac signals across space, time, and patients. In *ICML*, volume 139 of *PMLR*, pp. 5606–5615, 2021.

S. Kornblith, J. Shlens, and Q. V. Le. Do better ImageNet models transfer better? In *CVPR*, pp. 2661–2671, 2019.

A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *ICLR*, 2022.

J. D. Lee, Q. Lei, N. Saunshi, and J. Zhuo. Predicting what you already know helps: Provable self-supervised learning. In *NeurIPS*, volume 34, pp. 309–323, 2021.

F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *ICML*, volume 97 of *PMLR*, pp. 4114–4124, 2019.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Y. J. Ma, S. Sodhani, D. Jayaraman, O. Bastani, V. Kumar, and A. Zhang. VIP: Towards universal visual reward and representation via value-implicit pre-training. In *ICLR*, 2023.

A. Maurer, M. Pontil, and B. Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

A. Moscovich, A. Jaffe, and B. Nadler. Minimax-optimal semi-supervised regression on unknown manifolds. In *AISTATS*, volume 54 of *PMLR*, pp. 933–942, 2017.

S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta. R3M: A universal visual representation for robot manipulation. In *CoRL*, volume 205 of *PMLR*, pp. 892–909, 2023.

A. Newell and J. Deng. How useful is self-supervised pretraining for visual tasks? In *CVPR*, pp. 7345–7354, 2020.

A. Nikolov, M. Singh, and U. T. Tantipongpipat. Proportional volume sampling and approximation algorithms for A-optimal design. In *SODA*, pp. 1369–1386, 2019.

Y. Pang, W. Wang, F. E. H. Tay, W. Liu, Y. Tian, and L. Yuan. Masked autoencoders for point cloud self-supervised learning. In *ECCV*, volume 13662 of *LNCS*, pp. 604–621, 2022.

T. Pollard, B. E. Moody, L.-W. H. Lehman, B. J. Gow, C. Fernandes, C. Xie, A. Johnson, R. G. Mark, and T. Heldt. PhysioNet as a global platform for biomedical research. *Nature Health*, 1(8): 792–795, 2026.

M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio. Transfusion: Understanding transfer learning for medical imaging. In *NeurIPS*, volume 32, 2019.

J. S. Rosenfeld, A. Rosenfeld, Y. Belinkov, and N. Shavit. A constructive prediction of the generalization error across scales. In *ICLR*, 2020.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

N. Saunshi, J. T. Ash, S. Goel, D. Misra, C. Zhang, S. Arora, S. Kakade, and A. Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In *ICML*, volume 162 of *PMLR*, pp. 19250–19286, 2022.

K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, volume 33, pp. 596–608, 2020.

N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek. Deep learning for ECG analysis: Benchmarks and insights from PTB-XL. *IEEE Journal of Biomedical and Health Informatics*, 25(5): 1519–1528, 2021.

S. Tao, F. Xiang, A. Shukla, Y. Qin, X. Hinrichsen, X. Yuan, C. Bao, X. Lin, Y. Liu, T.-K. Chan, Y. Gao, X. Li, T. Mu, N. Xiao, A. Gurha, V. N. Rajesh, Y. W. Choi, Y.-R. Chen, Z. Huang, R. Calandra, R. Chen, S. Luo, and H. Su. Demonstrating GPU parallelized robot simulation and rendering for generalizable embodied AI with ManiSkill3. In *Robotics: Science and Systems*, 2025.

Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola. What makes for good views for contrastive learning? In *NeurIPS*, volume 33, pp. 6827–6839, 2020.

M. Tinati and S. Tu. On the asymptotics of self-supervised pre-training: Two-stage M-estimation and representation symmetry. In *COLT*, volume 336 of *PMLR*, pp. 6197–6309, 2026.

C. Tosh, A. Krishnamurthy, and D. Hsu. Contrastive learning, multi-view redundancy, and linear models. In *ALT*, volume 132 of *PMLR*, pp. 1179–1206, 2021.

N. Tripuraneni, M. I. Jordan, and C. Jin. On the theory of transfer learning: The importance of task diversity. In *NeurIPS*, volume 33, pp. 7852–7862, 2020.

N. Tripuraneni, C. Jin, and M. I. Jordan. Provable meta-learning of linear representations. In *ICML*, volume 139 of *PMLR*, pp. 10434–10443, 2021.

J. von Kügelgen, Y. Sharma, L. Gresele, W. Brendel, B. Schölkopf, M. Besserve, and F. Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *NeurIPS*, volume 34, pp. 16451–16467, 2021.

P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7:154, 2020.

P. Wagner, N. Strodthoff, R.-D. Bousseljot, W. Samek, and T. Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. PhysioNet, 2022. URL https://physionet.org/content/ptb-xl/1.0.3/. Version 1.0.3.

T. Wang and P. Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, volume 119 of *PMLR*, pp. 9929–9939, 2020.

T. Wegel, G. So, J. Park, and F. Yang. On the sample complexity of semi-supervised multi-objective learning. In *NeurIPS*, volume 38, pp. 128421–128476, 2025.

J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny. Barlow Twins: Self-supervised learning via redundancy reduction. In *ICML*, volume 139 of *PMLR*, pp. 12310–12320, 2021.

X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer. Scaling vision transformers. In *CVPR*, pp. 12104–12113, 2022.

H. Zhang and Y. Cao. Understanding the benefits of SimCLR pre-training in two-layer convolutional neural networks. In *AISTATS*, volume 300 of *PMLR*, pp. 4825–4833, 2026.

A Finite-Risk and Resource Guarantees

A.1 Oracle and structural lower bounds

Risks average the Gaussian response prior and maximize over grouping directions. Masks are independent of inputs, responses, and noise. RF may randomize and inspect contexts adaptively under a deterministic cap, but selects its grouping before labels, independently of the responses. Test prediction receives (u, v) without a new context. Write $a = \sigma^2/\tau^2$, $c = \tau^2(1 - 1/r)$, and

$$R_{\mathrm{or}}(N, r, \rho) = \tau^2 \mathbb{E}_{K\sim \mathrm{Bin}(N,\rho/r)} \frac{a}{a+K}, \qquad p_{\mathrm{err}}(B) = Q(\Delta\sqrt{B}). \tag{A1}$$

Theorem 1 (finite-risk bounds). For the model in Sec. 3.1, $r \geq 2$, $\tau^2, \sigma^2 > 0$, $\Delta \geq 0$, $\rho \in [0, 1]$, and integers $0 \leq B \leq N$,

$$\max\{R_{\mathrm{or}}, cp_{\mathrm{err}}(B)\} \leq \inf_{\widehat{f}\in \mathrm{RF}(B)} R(\widehat{f}) \leq \min\{\tau^2, R_{\mathrm{or}} + 2cp_{\mathrm{err}}(B)\}.$$

The upper bound is at most three times the lower. With exactly $0 \leq m \leq N$ uniformly selected labels, replace the oracle count by $K \sim \mathrm{Bin}(m, 1/r)$.

Proof of necessity. Given the true grouping, let c_j and s_j be its label count and sum. Conjugacy gives

$$\mathbb{E}[\theta_j \mid \mathcal{D}, g] = \frac{s_j}{a + c_j}, \qquad \mathrm{Var}(\theta_j \mid \mathcal{D}, g) = \frac{\tau^2 a}{a + c_j}. \tag{A2}$$

Each marginal count is $\mathrm{Bin}(N, \rho/r)$, so averaging these variances gives (A1). Revealing g only enlarges every learner's information, including Joint's. For wrong column commitment under row truth, even knowing every response leaves

$$\inf_h \frac{1}{r^2} \sum_{u,v=1}^{r} \big(h(v) - \theta_u\big)^2 = \frac{1}{r} \sum_{u=1}^{r} (\theta_u - \bar{\theta})^2, \qquad \bar{\theta} = \frac{1}{r} \sum_{u=1}^{r} \theta_u. \tag{A3}$$

Its minimizing function is constant and its prior mean is c; row/column exchange gives the other direction. The wrong-selection event is independent of θ. Group identities have identical laws under both directions, while the B contexts have laws $\mathcal{N}(\pm\Delta, 1)^{\otimes B}$. Their likelihood ratio is $\exp(2\Delta \sum_i Z_i)$, whose symmetric test has optimal average error $Q(\Delta\sqrt{B})$. Maximum error is at least average error. Padding any shorter adaptive inspection with independent unused contexts cannot increase information beyond B such observations. Multiplying the testing lower bound by c proves structural necessity.

A.2 Exact shrinkage risk and the threefold upper bound

The specified RF rule selects the sign of the context sum (an independent fair coin at $B = 0$), then applies (A2). Its risk R_{sh} differs from the class optimum $R^*_{\mathrm{RF}(B)}$. Under wrong column commitment with k labels, row counts M_j are multinomial with probabilities $1/r$, and $\widehat{h} = (\sum_j M_j\theta_j + \sum_{i=1}^{k} \xi_i)/(a+k)$. For an independent uniform test row, $\mathbb{E}[\theta_U \widehat{h}] = \tau^2 k/[r(a+k)]$ and $\mathbb{E}\sum_j M_j^2 = k^2/r + k(1 - 1/r)$. Expanding the squared error gives excess risk $c\,h(k)$ over correct-group shrinkage. Since context selection is independent of counts and labels,

$$h(k) = \frac{k(2a + k + 1)}{(a + k)^2}, \qquad R_{\mathrm{sh}} = R_{\mathrm{or}} + cp_{\mathrm{err}}(B)\mathbb{E}h(K). \tag{A4}$$

Here $K \sim \mathrm{Bin}(N, \rho/r)$ for Bernoulli labels and $K \sim \mathrm{Bin}(m, 1/r)$ for exactly m labels. For $a > 0$ and integer $k \geq 0$, $2(a + k)^2 - k(2a + k + 1) = 2a^2 + 2ak + k(k - 1) \geq 0$, so $0 \leq h(k) \leq 2$. Thus $R_{\mathrm{sh}} \leq R_{\mathrm{or}} + 2cp_{\mathrm{err}}(B)$. Zero prediction has risk τ^2; choosing the smaller known procedure risk uses neither the unknown grouping nor test labels. The resulting upper bound is at most $3\max\{R_{\mathrm{or}}, cp_{\mathrm{err}}(B)\}$, proving Thm. 1. At $N = 0$ or $\rho = 0$, oracle and shrinkage risks equal τ^2.

A.3 FINITE LABEL INVERSIONS

Fix $N, r, B, \Delta, \sigma^2, \tau^2$ and $0 < \varepsilon < \tau^2$. Define

$$\rho^*_{\mathrm{or}}(u) = \inf\{\rho \in [0,1] : R_{\mathrm{or}}(N, r, \rho) \le u\}, \qquad \rho^*_{\mathrm{RF}}(\varepsilon) = \inf\{\rho \in [0,1] : R^*_{\mathrm{RF}(B)} \le \varepsilon\}, \quad \text{(A5)}$$

with $\inf \varnothing = +\infty$. Binomial coupling makes the oracle risk continuous and nonincreasing; independent label discarding makes the class-optimal RF risk nonincreasing. If $cp_{\mathrm{err}}(B) > \varepsilon$ or $R_{\mathrm{or}}(N, r, 1) > \varepsilon$, the RF target is infeasible. Otherwise,

$$\rho^*_{\mathrm{or}}(\varepsilon) \le \rho^*_{\mathrm{RF}}(\varepsilon) \le \rho^*_{\mathrm{or}}\big(\varepsilon - 2\tau^2(1 - 1/r)p_{\mathrm{err}}(B)\big), \quad \text{(A6)}$$

The sufficient side is useful only with positive argument and finite inverse. Evaluate the oracle's finite binomial sum and invert in one dimension. For the tighter procedure-specific frontier, scan (A4) and refine its first crossing.

A.4 SAMPLE AND FIXED-LABEL REQUIREMENTS

Let $t = \varepsilon/\tau^2$, $p = \rho/r \le 1/2$, and $\mu = Np$. The zero-count event, Jensen's inequality, and splitting at $K < \mu/2$ give

$$\max\left\{(1-p)^N, \frac{a}{a+\mu}\right\} \le \frac{R_{\mathrm{or}}}{\tau^2} \le \min\left\{1, e^{-\mu/8} + \frac{2a}{a+\mu}\right\}. \quad \text{(A7)}$$

Indeed, $\Pr(K < \mu/2) \le e^{-\mu/8}$ and on the complement $a/(a+K) \le 2a/(a+\mu)$; the outer minimum covers $\mu = 0$. Necessity gives $\mu \ge a(1/t - 1)$ and, since $-\log(1-p) \le 2p$, $\mu \ge \frac{1}{2}\log(1/t)$. For sufficiency with $\Delta > 0$, take

$$\mu \ge \max\{8\log(4/t), 8a/t\}, \qquad N \ge 2\Delta^{-2}\log(2/t). \quad \text{(A8)}$$

Each term of the oracle upper bound is at most $t/4$, and $Q(x) \le \frac{1}{2}e^{-x^2/2}$ gives $p_{\mathrm{err}}(N) \le t/4$, proving risk at most $\tau^2 t$ by (3). The necessary condition $Q(\Delta\sqrt{N}) \le t/(1 - 1/r)$ has squared inverse tail comparable to $\log(1/t)$ for $r \ge 2, t \le 1/8$. Together these prove (4); a sum and maximum of two positive resource terms differ by at most two.

At fixed N, the label-fraction order is $r[\log(1/t) + a/t]/N$ when the structural upper contribution is at most $\varepsilon/2$ and the sufficient fraction lies in $[0,1]$. Exactly m labels selected independently of input values have iid uniform identities and marginal count $\mathrm{Bin}(m, 1/r)$. Oracle risk is then independent of N, while context error can decrease with N.

A.5 RECORD-ACCESS GUARANTEES

Let $t \in (0, 1/8]$. The interface supplies the label count and a random index of revealed records. One query reads a context or a labeled record with identities and response; information can be reused but no new labels or unrevealed values requested. Query cost is a deterministic hard cap and risk averages the task model. For $\Delta > 0$, $a/t \ge \log(1/t)$, and the availability conditions below,

$$T^*_{\mathrm{RF,query}} \asymp r\frac{\sigma^2}{\varepsilon} + \Delta^{-2}\log\frac{\tau^2}{\varepsilon}. \quad \text{(A9)}$$

Even with g known, at most L labels give $\sum_j c_j \le L$. Gaussian posterior covariance and convexity imply risk at least $\tau^2 a/(a + L/r)$, including adaptive choices from observed records. Consequently

$$L \ge ra(1/t - 1) \quad \text{(A10)}$$

is necessary. Separately, structural necessity requires $B \gtrsim \Delta^{-2}\log(1/t)$, where the implicit constant is universal. For sufficiency use

$$B = \left\lceil 2\Delta^{-2}\log(4/t)\right\rceil, \qquad L = \left\lceil r\max\{8\log(8/t), 16a/t\}\right\rceil. \quad \text{(A11)}$$

Assume $N \ge B$ and $\rho N \ge 2L$. If fewer than L labels exist, output zero; otherwise inspect B contexts and the first L indexed labels and apply shrinkage. The shortage probability is at most $e^{-\rho N/8} \le e^{-L/4}$ and is independent of responses. Conditional on availability, the L group identities

are iid uniform. Eq. (A7) bounds the correct-group contribution by $\tau^2 t/4$, the context contribution obeys $2cp_{\mathrm{err}}(B) \le \tau^2 t/4$, and shortage costs at most $\tau^2 e^{-L/4} \le \tau^2 t/2$. Total risk is at most ε.

Arithmetic costs $O(B+L+r)$ and storage $O(r)$. When $a/t \ge \log(1/t) \ge \log 8$, labels also cover the $O(r)$ output cost and the two necessary terms match the upper cap up to constants, proving (A9). Outside that noise-dominated regime, (A10) and (A11) remain separate necessary/sufficient bounds. A missing label index adds a scan of up to N records.

A.6 Input-label interaction and allocation prediction

Fix $a > 0, r \ge 2, \tau^2$, exactly m random labels, and all $N \ge m$ contexts. Put $q = 1/r$, $c = \tau^2(1-q)$, $K_m \sim \mathrm{Bin}(m, q)$, and $p_N = Q(\Delta\sqrt{N})$. Eq. (A4) gives

$$A(m) = c\,\mathbb{E}h(K_m), \qquad h(k) = \frac{k(2a+k+1)}{(a+k)^2}, \qquad R_{\mathrm{sh}}(N,m) = R_{\mathrm{or}}(m) + p_N A(m). \tag{A12}$$

For $0 \le m_l < m_h \le N_s < N_l$, the additive prediction uses three source risks to predict $R_{\mathrm{sh}}(N_s, m_h)$. Its signed error is

$$\begin{aligned} e_{\mathrm{add}} &= R_{\mathrm{sh}}(N_s,m_l) + R_{\mathrm{sh}}(N_l,m_h) - R_{\mathrm{sh}}(N_l,m_l) - R_{\mathrm{sh}}(N_s,m_h) \\ &= -I = (p_{N_s} - p_{N_l})[A(m_l) - A(m_h)]. \end{aligned} \tag{A13}$$

Oracle terms cancel. The mixed-difference identity $e_{\mathrm{add}} = -I$ also holds empirically; the product formula and exact finite sum concern the specified RF rule.

Proposition A1 (shape and sign). The sequence $A(m)$ increases from zero to one maximum, possibly attained at two adjacent integers, then strictly decreases to c. For fixed a, r,

$$A(m) = c + \frac{cr}{m} + o(m^{-1}), \qquad A(m+1) - A(m) = -\frac{cr}{m^2} + o(m^{-2}). \tag{A14}$$

For $\Delta > 0$, two budgets on the increasing/decreasing branch underpredict/overpredict, respectively, with equality at a two-point peak. Straddling budgets have no fixed sign. At $\Delta = 0$, $e_{\mathrm{add}} = 0$.

Proof. Write $g(t) = 1 - q + qt$. Since $h(k) = 1 + (a+k)^{-1} - a(a+1)(a+k)^{-2}$, the integral representations of these reciprocals give

$$\frac{A(m)}{c} - 1 = \int_0^1 t^{a-1}[1 + a(a+1)\log t] g(t)^m \, dt.$$

Set $J_m = [A(m+1) - A(m)]/c$ and $\omega(t) = -q(1-t)t^{a-1}[1 + a(a+1)\log t]$. Then $J_m = \int_0^1 \omega(t) g(t)^m \, dt$. The kernel is positive below $t_0 = e^{-1/[a(a+1)]}$ and negative above it, whereas g strictly increases. Thus

$$J_{m+1} - g(t_0) J_m = \int_0^1 \omega(t)[g(t) - g(t_0)] g(t)^m \, dt < 0.$$

Once $J_m \le 0$, all later differences are negative; $J_0 = 2q/(a+1) > 0$. For eventual negativity and the limits, split both integrals at $1/2$. The lower part is exponentially small because $g(t) \le 1 - q/2$ and $t^{a-1}(1 + |\log t|)$ is integrable. On the upper part, put $t = 1 - u/m$ and use $g(t)^m \le e^{-qu}$. Dominated convergence gives $m(A(m)/c - 1) \to 1/q$ and $m^2 J_m \to -1/q$. This proves (A14), the single sign change, and the error signs in (A13).

B Joint Posterior and Comparison

B.1 Posterior, risk, and exact evaluation

Give g a uniform prior and use the counts, sums, branch means, and mixture weight from Alg. 1, with all N contexts. Gaussian conjugacy gives branch variances $v_j^{\pm} = \tau^2 a/(a + c_j^{\pm})$; empty groups have mean zero and variance τ^2. Conditional on the training data, the response variance at (u, v) is

$$\pi v_u^+ + (1-\pi) v_v^- + \pi(1-\pi)(\widehat{\theta}_u^+ - \widehat{\theta}_v^-)^2. \tag{B1}$$

The tower property, conditioning on the true branch, makes the first two terms average to R_{or}; the last gives (6). Disagreement requires no r^2-pair enumeration:

$$D = \frac{1}{r}\sum_u (\widehat{\theta}_u^+)^2 + \frac{1}{r}\sum_v (\widehat{\theta}_v^-)^2 - 2\left(\frac{1}{r}\sum_u \widehat{\theta}_u^+\right)\left(\frac{1}{r}\sum_v \widehat{\theta}_v^-\right). \tag{B2}$$

Swapping U, V and negating Z, g preserves the model and exchanges posterior branches. Joint therefore has equal directional risks. Its equal-prior Bayes optimality implies minimax optimality over g with the response prior fixed, hence

$$R_{\mathrm{or}} \le R^*_{\mathrm{joint}} \le R^*_{\mathrm{RF}(N)} \le R_{\mathrm{sh}}(B), \qquad 0 \le B \le N. \tag{B3}$$

Exact evaluation costs $O(N+m+r)$ arithmetic, $O(r)$ storage, and $O(1)$ per test pair, at fixed scalar precision. Contexts can be omitted at $\Delta = 0$; prediction is zero at $m = 0$. The optional determinant certificate costs $O(r^3)$ arithmetic and $O(r^2)$ storage. These are evaluation costs, distinct from access sufficient for a risk target in App. A.

B.2 Candidate retention and a common moment identity

Let $\mathcal{V}$ be a d-dimensional space, fixed or selected from contexts independently of responses and labeled coordinates. For an orthonormal basis ψ_j under the uniform input law and $m \ge 1$ iid labeled pairs, define

$$\widehat{\beta}_j = \frac{1}{m}\sum_i Y_i \psi_j(X_i), \qquad \widehat{f}_{\mathcal{V},m}(x) = \sum_j \widehat{\beta}_j \psi_j(x). \tag{B4}$$

Its conditional mean is $P_{\mathcal{V}} f_g$. Let $L_{\mathcal{V},g}$ be the prior-averaged projection error and $L_{\mathcal{V}} = \max_g L_{\mathcal{V},g}$. Orthogonality of projection bias and coefficient variance gives

$$R(\widehat{f}_{\mathcal{V},m}) = L_{\mathcal{V}} + \frac{(d-1)\tau^2 + d\sigma^2 + L_{\mathcal{V}}}{m}. \tag{B5}$$

Indeed, the variance numerator is $\mathbb{E}[Y^2 \sum_j \psi_j(X)^2] - \mathbb{E}\|P_{\mathcal{V}} f_g\|_X^2 = d(\tau^2+\sigma^2) - (\tau^2 - L_{\mathcal{V},g})$: the prior second moment is pointwise τ^2, and the expected squared feature norm is d. Average any allowed context choice, then maximize this increasing function of $L_{\mathcal{V},g}$.

Proposition B1 (candidate retention). The frozen map $\phi_{\mathrm{both}}(u,v) = (e_u, e_v)$ contains both target families. Its additive function space has dimension $2r-1$, since only constants belong to both coordinate spaces. This is also the minimum dimension of any linear space containing both. Its reproducing kernel gives an explicit label-only readout:

$$H_{u,v}(U,V) = r\mathbf{1}\{U=u\} + r\mathbf{1}\{V=v\} - 1, \quad \widehat{f}_m(u,v) = \frac{1}{m}\sum_i Y_i H_{u,v}(U_i,V_i). \tag{B6}$$

This readout is linear in the frozen features and has conditional mean f_g. Substituting $d = 2r-1$ and $L_{\mathcal{V}} = 0$ into (B5) proves (7). Use zero at $m = 0$. For Bernoulli label count M, risk is $\tau^2 \Pr(M=0) + C_r \mathbb{E}[\mathbf{1}\{M>0\}/M]$, tending to zero at fixed $r, \rho > 0$ as N grows. The exact Joint predictor also belongs to the same fixed space, with weights $(\pi\widehat{\theta}^+, (1-\pi)\widehat{\theta}^-)$ computed from permitted contexts and labels; both rules predict from (u,v) without a new context.

B.3 Labeled relations and distribution affinity

For fixed labeled identities, let $F_U, F_V \in \mathbb{R}^{m\times r}$ be one-hot matrices and $C_\pm = \sigma^2 I_m + \tau^2 F_{U/V} F_{U/V}^\top$. Their Gaussian affinity is

$$\gamma_S = \frac{\det(C_+)^{1/4}\det(C_-)^{1/4}}{\det((C_+ + C_-)/2)^{1/2}} \in (0,1], \tag{B7}$$

with $\gamma_\varnothing = 1$. It bounds joint risk by

$$R^*_{\mathrm{joint}} \le R_{\mathrm{or}} + 2\tau^2 e^{-N\Delta^2/2}\mathbb{E}_{U,V,S}\gamma_S. \tag{B8}$$

Proof. For conditional branch densities $p_\pm$, the mixture-weighted variance factor is $p_+p_-/[2(p_+ + p_-)] \le \sqrt{p_+p_-}/4$. Context affinity is $e^{-N\Delta^2/2}$. The normalized geometric mean of label densities has covariance $H = 2(C_+^{-1} + C_-^{-1})^{-1} \preceq 2C_\pm$. Each branch mean is linear in labels and has variance at most τ^2 under its own branch, hence at most $2\tau^2$ under H. Thus $\mathbb{E}_H D \le 8\tau^2$ by $(x-y)^2 \le 2x^2 + 2y^2$, proving the bound.

For evaluation, set $G = \begin{pmatrix} \mathrm{diag}(c^+) & C_{UV} \\ C_{UV}^\top & \mathrm{diag}(c^-) \end{pmatrix}$, where C_{UV} is the table of labeled pair counts. The determinant lemma gives

$$\log \gamma_S = \frac{1}{4}\sum_j \log(1 + c_j^+/a) + \frac{1}{4}\sum_j \log(1 + c_j^-/a) - \frac{1}{2}\log\det(I_{2r} + G/(2a)). \tag{B9}$$

Define the two sharing matrices and their pairwise disagreements by

$$(G_U)_{ij} = \mathbf{1}\{U_i = U_j\}, \quad (G_V)_{ij} = \mathbf{1}\{V_i = V_j\}, \quad D_S = \sum_{i<j} |(G_U)_{ij} - (G_V)_{ij}|. \tag{B10}$$

The label likelihood ratio is identically one exactly when $D_S = 0$. Indeed, the sharing matrices have unit diagonals, so this condition is equivalent to $C_+ = C_-$. Nondegenerate centered Gaussians are determined by covariance; if $D_S > 0$, their densities differ and $\gamma_S < 1$. Repetitions common to both candidates provide no distinguishing relation.

A random pair disagrees with probability $q = 2(r-1)/r^2$, hence

$$\mathbb{E}D_S = \binom{m}{2} q = \frac{m(m-1)(r-1)}{r^2}. \tag{B11}$$

Let $q_{m,k}$ be the probability of identical equality patterns with k blocks, with $q_{0,0} = 1$ and out-of-range terms zero. The next pair joins an existing paired block or creates a new block in both coordinates, giving

$$q_{m+1,k} = \frac{k}{r^2} q_{m,k} + \frac{(r-k+1)^2}{r^2} q_{m,k-1}, \qquad \Pr(E_m) = \sum_k q_{m,k}, \quad E_m = \{D_S = 0\}. \tag{B12}$$

On E_m, labels supply no grouping evidence. This event is independent of contexts and has the same probability under either direction, so optimal equal-prior grouping error satisfies

$$p_{\text{joint}}^{\text{group}}(B, m) \ge \Pr(E_m) Q(\Delta\sqrt{B}). \tag{B13}$$

This concerns grouping tests, not prediction by mixtures or retained spaces. The no-repeat event in both coordinates has probability $[(r)_m/r^m]^2$ for $m \le r$, where $(r)_m = r!/(r-m)!$, and is a subset of E_m.

Distinct disagreement indicators are pairwise independent: disjoint pairs use independent identities, and pairs sharing one record are conditionally independent with constant probability q. Thus, with $\lambda = \binom{m}{2} q$, $\mathrm{Var}(D_S) = \lambda(1-q)$. Cauchy–Schwarz and Markov give

$$\frac{\lambda}{\lambda + 1 - q} \le \Pr(D_S > 0) \le \min\{1, \lambda\}. \tag{B14}$$

Relations vanish in probability for $m = o(\sqrt{r})$, appear with nonvanishing probability at order $\sqrt{r}$, and occur with probability tending to one when $m/\sqrt{r} \to \infty$. Noise still controls their strength.

For a disagreeing pair, put $\eta = \tau^2/(\tau^2 + \sigma^2)$. Its normalized covariances are the identity and a correlation matrix with off-diagonal η, giving

$$\gamma_{\text{pair}} = \frac{(1-\eta^2)^{1/4}}{(1-\eta^2/4)^{1/2}}, \qquad \mathbb{E}\gamma_S = 1 - q(1 - \gamma_{\text{pair}}) \quad (m = 2). \tag{B15}$$

For $D_S > 0$, select a disagreeing pair from the identities. Marginalization increases affinity by Cauchy–Schwarz, so $\gamma_S \le \gamma_{\text{pair}}$. Therefore

$$\mathbb{E}\gamma_S \le \Pr(E_m) + [1 - \Pr(E_m)]\gamma_{\text{pair}} \le 1 - (1 - \gamma_{\text{pair}})\frac{\lambda}{\lambda + 1 - q}. \tag{B16}$$

The recursion is sharper than the closed bound; both are exact for $m = 2$. Many relations may lower affinity further, but dependent pairs' affinities cannot simply be multiplied.

B.4 Information, rank, and linear-readout controls

Put $p_B = p_{\mathrm{err}}(B)$ and $\|h\|_X^2 = r^{-2}\sum_{u,v} h(u,v)^2$. Conditional on (g,θ), information and linear-readout errors are $\inf_h \|h\circ\phi - f_g\|_X^2$ and $\inf_w \|w^\top\phi - f_g\|_X^2$. Average over responses and context-based representation choices, then maximize over g. These quantify projection error before finite-label fitting.

Table S1 fixes width at $2r$. Commitment pads the selected indicator with zeros, giving rank r and wrong-choice loss c. Retaining both gives rank $2r-1$ and represents either truth exactly.

Table S1: Representation controls.

Features	Width	Rank	Info.	Linear
Committed	$2r$	r	cp_B	cp_B
Retained	$2r$	$2r-1$	0	0
Recoded	$2r$	$2r-1$	0	c

To preserve information, width, and rank while changing linear readability, index groups by $0,\dots,r-1$ and set

$$\phi_{\mathrm{rec}}(u,v) = (e_{(u+v) \bmod r}, e_v). \tag{B17}$$

This recoding is invertible and has rank $2r-1$. Column truth remains linear; under row truth, either recoded coordinate averages θ to $\bar{\theta}$. Orthogonality of the two coordinate contrast spaces makes the best additive predictor constant, with prior residual c. Its information error is zero but worst-direction linear error is c. At $r=4, \tau^2=1, \Delta=0$, committed, retained, and recoded linear errors are $0.375, 0, 0.750$. These projection errors separate information retention from linear readability.

For $r \le d \le 2r-1$, retain the constant, all $r-1$ contrasts of a context-selected grouping, and $d-r$ fixed orthogonal contrasts of the other. This space $\mathcal{V}_d$ can be embedded at the same width and has projection error

$$L_d = \frac{\tau^2(2r-1-d)}{r} p_B. \tag{B18}$$

Correct selection loses nothing. Wrong selection omits $2r-1-d$ contrasts, each with prior loss τ^2/r. The isotropic response prior makes their orientation immaterial. Substitution into (B5) yields

$$R(\widehat{f}_{\mathcal{V}_{d+1},m}) - R(\widehat{f}_{\mathcal{V}_d,m}) = -\frac{\tau^2 p_B}{r} + \frac{\tau^2+\sigma^2-\tau^2 p_B/r}{m}. \tag{B19}$$

An added contrast helps when $(m+1)p_B > r(1+a)$. The difference is independent of d, so the best dimension among these moment readouts is r below the threshold, $2r-1$ above it, and arbitrary at equality. For $p_B = 0$ choose r; at $m=0$ use zero, also selectable by known risk. This tradeoff concerns linear dimension, separately from nonlinear recoverability or encoder updates.

B.5 A finite resource envelope for joint inference

For exactly $1 \le m \le N$ random labels, Bayes optimality, direction symmetry, and the zero, affinity, and retained-space procedures give

$$R^*_{\mathrm{joint}}(N,m) \le \min\left\{\tau^2,\ R_{\mathrm{or}}(m) + 2\tau^2 e^{-N\Delta^2/2}\mathbb{E}\gamma_S,\ C_r/m\right\}. \tag{B20}$$

Here $R_{\mathrm{or}}(m)$ uses $\mathrm{Bin}(m,1/r)$ and $\Gamma_m = \mathbb{E}\gamma_S$ averages the fixed-size identity design. Use (B16) for an explicit bound, or (B9) for a conditional design certificate. With only B contexts, replace N in the exponential by B. At $m=0$ risk is τ^2. Retention guarantees risk at most ε with $\lceil C_r/\varepsilon\rceil$ available labels.

If $R_{\mathrm{or}}(m) < \varepsilon$, the affinity branch is sufficient when

$$\frac{B\Delta^2}{2} - \log\Gamma_m \ge \log\frac{2\tau^2}{\varepsilon - R_{\mathrm{or}}(m)}. \tag{B21}$$

For $\Delta > 0$, an integer context budget is therefore

$$B \ge \left\lceil \frac{2}{\Delta^2}\left[\log\frac{2\tau^2\Gamma_m}{\varepsilon - R_{\mathrm{or}}(m)}\right]_+ \right\rceil, \qquad B \le N, \quad m \le N. \tag{B22}$$

Use (B21) directly at $\Delta = 0$. The design average, not one observed affinity, is required for an unconditional bound. For Bernoulli labels, condition on the count and average. Finally, the cellwise shrinkage comparator used in experiments has

$$R_{\mathrm{cell}} = \tau^2 \mathbb{E}_{K \sim \mathrm{Bin}(N, \rho/r^2)} \frac{a}{a+K}, \qquad R^*_{\mathrm{joint}} \le R_{\mathrm{cell}}. \tag{B23}$$

At fixed m, use $K \sim \mathrm{Bin}(m, 1/r^2)$. Joint's Bayes optimality and symmetry prove the comparison.

C Risk-Level Thresholds and Gaussian Label Budgets

C.1 Design risk and increasing families

Fix inputs X, a response prior, and $W \succeq 0$. Before observing label values, define

$$R_X(S) = \inf_{\widehat{f}} \mathbb{E}[\|\widehat{f}(X, Y_S) - f\|_W^2 \mid X], \qquad \|v\|_W^2 = v^\top W v. \tag{C1}$$

The expectation averages targets and noise. The prior and loss stay fixed across masks. Extra labels can be ignored, so $\mathcal{F}_t = \{S : R_X(S) \le t\}$ is increasing. If $R_0 = R_X(\varnothing) < \infty$, a Bernoulli reveal mask satisfies

$$\overline{R}_X(\rho) = \mathbb{E} R_X(S_\rho) = \int_0^{R_0} \Pr(S_\rho \notin \mathcal{F}_t)\, dt. \tag{C2}$$

C.2 From set thresholds to mean risk

For a nontrivial family, let $\min \mathcal{F}_t$ denote its inclusion-minimal members and put $\ell_t = \max\{2, \max_{T \in \min \mathcal{F}_t} |T|\}$. Its fractional expectation threshold and median success threshold obey

$$q_f(t) = \sup \left\{ p \in [0,1] : \exists w_A \ge 0,\ \sum_{A \subseteq T} w_A \ge 1 \ (T \in \min \mathcal{F}_t),\ \sum_A w_A p^{|A|} \le \tfrac{1}{2} \right\}, \tag{C3}$$

$$q_f(t) \le p_c(t) \le C q_f(t) \log \ell_t. \tag{C4}$$

The first inequality follows by averaging the fractional cover; the second is the universal-constant theorem of Frankston et al. (2021). Applying that theorem yields the following mean-risk consequence.

Corollary C2 (mean-risk conversion from Frankston et al. (2021)). If $R_X([N]) < \varepsilon/2$ and $2\varepsilon < R_0$, set $\rho^*_X(\varepsilon) = \inf\{\rho : \overline{R}_X(\rho) \le \varepsilon\}$, $p_0 = \min\{1, C q_f(\varepsilon/2) \log \ell_{\varepsilon/2}\}$, and $h = \lceil \log_2(2R_0/\varepsilon) \rceil$. Then

$$q_f(2\varepsilon) \le \rho^*_X(\varepsilon) \le 1 - (1-p_0)^h. \tag{C5}$$

Proof. Mean risk at most ε implies success at level 2ε with probability at least $1/2$ by Markov, proving necessity. Each p_0-mask succeeds at level $\varepsilon/2$ with probability at least $1/2$. Their independent union has reveal probability $1 - (1-p_0)^h$ and fails only if all h masks fail, giving risk at most $\varepsilon/2 + R_0 2^{-h} \le \varepsilon$. The case $p_0 = 1$ follows from full-label feasibility. Writing $R_{\mathrm{full}} = R_X([N])$, integration also gives

$$\overline{R}_X(\rho) \ge R_{\mathrm{full}} + \frac{1}{2} \int_{R_{\mathrm{full}}}^{R_0} \mathbf{1}\{\rho < q_f(t)\}\, dt. \tag{C6}$$

The bound converts set thresholds into a mean-risk guarantee conditional on X, with an unspecified universal constant C.

C.3 Known Gaussian tasks

For $N \ge 1$ and a fixed known representation, let $\theta \sim \mathcal{N}(0, \Lambda^{-1})$, $\Lambda \succ 0$, and $Y_i = b_i^\top \theta + \xi_i$, with iid unit Gaussian noise independent of θ, and masks independent of the parameter and noise. For $W \succeq 0$, write

$$V_S = \left(\Lambda + \sum_{i \in S} b_i b_i^\top \right)^{-1}, \quad R_X(S) = \mathrm{tr}(W V_S), \quad G = \sum_i b_i b_i^\top, \quad \nu_i = b_i^\top \Lambda^{-1} b_i. \tag{C7}$$

The posterior formula is standard. Under determinant-weighted sampling, Dereziński et al. (2020, Lemma 5) upper-bound expected inverse information by $(\Lambda + \sum_i p_i b_i b_i^\top)^{-1}$, with p_i distribution parameters rather than marginals. Our ordinary Bernoulli or uniform fixed-size masks instead give an inverse-convexity lower bound. Linear prediction supplies the upper bound below, without optimizing subsets.

C.4 Computable covariance bounds

Proposition C1 (finite-mask covariance bounds). For independent inclusion with probability ρ,

$$(\Lambda + \rho G)^{-1} \preceq \mathbb{E} V_{S_\rho} \preceq \left(\Lambda + \sum_i \frac{\rho\, b_i b_i^\top}{1 + (1-\rho)\nu_i}\right)^{-1}. \tag{C8}$$

For a uniformly selected set of size m, $N \geq 2$, and $\rho = m/N$, the same lower bound holds and

$$\mathbb{E} V_{S_m} \preceq \left(\Lambda + \sum_i \frac{\rho\, b_i b_i^\top}{1 + (1-\rho)N\nu_i/(N-1)}\right)^{-1}. \tag{C9}$$

Both endpoints $m = 0, N$ are exact; for $N = 1$ they are the only fixed-size cases. Taking the trace against W preserves both inequalities.

Proof. Let $C = \Lambda^{-1}$ and let B have rows $b_i^\top$. The variational identity $v^\top M^{-1} v = \sup_u (2v^\top u - u^\top M u)$ gives the lower bound by exchanging expectation and supremum as an inequality. For $\rho > 0$, form the observed vector $z_i = \zeta_i Y_i/\rho$, assigning zero to missing labels. Write $z = B\theta + \eta$. The residual is uncorrelated with θ; under independent masking,

$$\mathrm{Cov}(\eta) = D_\rho := \mathrm{diag}\left(\frac{1 + (1-\rho)\nu_i}{\rho}\right).$$

The best linear predictor from z has error covariance $C - CB^\top(BCB^\top + D_\rho)^{-1}BC = (\Lambda + B^\top D_\rho^{-1} B)^{-1}$. Conditional expectation given (S, Y_S) has no larger error covariance, by orthogonal projection. This proves (C8) without assuming η is independent or Gaussian.

For uniform fixed size, $\mathbb{E}\zeta_i\zeta_j = m(m-1)/[N(N-1)]$ for $i \neq j$, giving

$$\mathrm{Cov}(\eta) = D_m - \frac{1-\rho}{\rho(N-1)} BCB^\top \preceq D_m, \qquad D_m = \mathrm{diag}\left(\frac{1 + (1-\rho)N\nu_i/(N-1)}{\rho}\right).$$

Use the linear predictor optimized for residual covariance D_m. Its actual error covariance is no larger than $(\Lambda + B^\top D_m^{-1} B)^{-1}$, and the posterior mean is at least as good. At $\rho = 0$ both bounds equal the prior covariance. The bounds use the first two moments of the stated mask designs.

C.5 Spectral inversion and a nonorthogonal example

Diagonalize $\Lambda^{-1/2} G \Lambda^{-1/2} = U\,\mathrm{diag}(\lambda_j) U^\top$. With u_j its eigenvectors, define

$$w_j = u_j^\top \Lambda^{-1/2} W \Lambda^{-1/2} u_j, \qquad F(s) = \sum_{j=1}^{d} \frac{w_j}{1 + s\lambda_j}, \qquad \kappa = \max_i \nu_i. \tag{C10}$$

No commutativity or orthogonality of the label directions is required. Since $w_j, \lambda_j \geq 0$, (C8) implies

$$F(\rho) \leq \overline{R}_X(\rho) \leq F\left(\frac{\rho}{1 + (1-\rho)\kappa}\right) \leq [1 + (1-\rho)\kappa] F(\rho). \tag{C11}$$

The final inequality holds termwise. Here $F(0) = R_0$ and $F(1) = R_{\mathrm{full}}$, but $F(\rho)$ is a lower bound rather than the mean risk in general.

If $R_{\mathrm{full}} < \varepsilon < R_0$, there is a unique $s_\varepsilon \in (0, 1)$ with $F(s_\varepsilon) = \varepsilon$: the strict endpoint inequality ensures some $w_j\lambda_j > 0$, so F strictly decreases. Monotone coupling of masks and inversion of (C11) give

$$s_\varepsilon \leq \rho^*_X(\varepsilon) \leq \frac{(1+\kappa)s_\varepsilon}{1 + \kappa s_\varepsilon}, \qquad \lceil N s_\varepsilon \rceil \leq m^*_X(\varepsilon) \leq \left\lceil \frac{N(1+\kappa_N)s_\varepsilon}{1 + \kappa_N s_\varepsilon} \right\rceil, \tag{C12}$$

Here $\kappa_N = N\kappa/(N-1)$ ($N \geq 2$), and m_X^* averages uniform size-m masks. If $\varepsilon \geq R_0$, no labels are needed; if $\varepsilon < R_{\text{full}}$, the target is impossible. Use $s_\varepsilon = 1$ at $\varepsilon = R_{\text{full}} < R_0$; when $R_0 = R_{\text{full}}$, labels cannot help. The Bernoulli ratio is $(1+\kappa)/(1+\kappa s_\varepsilon) \leq 1+\kappa$, sharpenable by Eqs. C8–C9. These are mask-averaged guarantees. Preprocessing costs $O(Nd^2 + d^3)$, evaluation $O(d)$, and bisection to width δ, $O(d\log(1/\delta))$. The calculation bypasses q_f.

For a connected simple graph with n vertices and N edges, let Q be an orthonormal basis of $\mathbf{1}^\perp$, $f = Q\theta$, and $\theta \sim \mathcal{N}(0, \tau^2 I)$. Observe edge differences $Y_e = f_u - f_v + \xi_e$ with noise variance σ^2, and use loss $\|\widehat{f} - f\|^2/(n-1)$. After dividing observations by σ, $b_e = \sigma^{-1}Q^\top(e_u - e_v)$, $\Lambda = \tau^{-2}I$, $W = I/(n-1)$, and

$$F(s) = \frac{\tau^2}{n-1}\sum_{j=2}^{n}\frac{1}{1+s(\tau^2/\sigma^2)\lambda_j(L)}, \qquad \kappa = 2\tau^2/\sigma^2, \tag{C13}$$

Here L is the graph Laplacian. Overlapping edges give nonorthogonal label directions. On the complete graph, $F(s) = \tau^2/[1 + sn\tau^2/\sigma^2]$. For $n = 101$, $N = 5050$, and $\tau^2 = \sigma^2 = 1$, $s_{0.1} = 9/101$ and Eq. C12 yields $450 \leq m_X^*(0.1) \leq 1146$, a computable interval for the required label budget.

D Data and Training Protocols

D.1 Data partitions and label access

ECG pretraining uses 700,000 MIMIC-IV-ECG v1.0 waveforms, excluding reports, diagnostic text, and extracted clinical measurements (Gow et al., 2023). Targets use PTB-XL v1.0.3 (Wagner et al., 2020; 2022), also distributed through PhysioNet (Pollard et al., 2026). Patient-disjoint folds 1–8, 9, and 10 (Strodthoff et al., 2021) provide 17,111 fitting, 2,146 selection, and 2,158 test records; the test set has 1,900 patient clusters. The main target has 44 diagnostic indicators, with 23-subclass and five-superclass controls.

ImageNet uses all 1,000 ILSVRC2012 classes (ImageNet, 2012; Deng et al., 2009; Russakovsky et al., 2015). An identity-based split of 1,281,167 development images reserves 25,623 for selection and 1,255,544 for fitting and input-only pretraining. The official 50,000-image validation set serves as the test set. Masks use identities without reading classes or repairing class-coverage gaps. The small input pool includes the designated label prefix for the held-out configuration.

Table S2: Visible label counts.

Labels	Fit	Select	Total
ECG			
1%	171	22	193
3%	514	64	578
10%	1,711	215	1,926
30%	5,133	644	5,777
100%	17,111	2,146	19,257
ImageNet			
1%	12,556	256	12,812
3%	37,666	769	38,435
10%	125,555	2,562	128,117
30%	376,663	7,687	384,350
100%	1,255,544	25,623	1,281,167

ManiSkill 3.0.1 provides three tasks: PickCube-v1, StackCube-v1, and PegInsertionSide-v1 (Tao et al., 2025). One task-conditioned policy serves all three. Each task has 90,000 fitting, 10,000 selection, and 10,000 test observations. Trajectories are split before label selection; at most ten predetermined observations are retained per trajectory. Input expansion favors new trajectories. Uncertainty uses 1,000 test-trajectory clusters per task. Additional labels reveal actions from this collection, without new expert trajectories.

Within each seed group, all methods share nested random fitting and selection prefixes. Their sum is the label budget. ECG/ImageNet fractions use the largest target development pool as denominator, regardless of input-pool size (Table S2). Robot budgets are 100, 300, 1,000, 3,000, 10,000, and 100,000 actions per task, split 90%/10% between fitting/selection; the three-task policy uses three times the plotted count.

D.2 Encoders, readouts, and training

Table S3 collects architecture and regular-schedule settings. ECG inputs are 12-lead, ten-second waveforms at 500 Hz in millivolts, without per-record amplitude normalization, using 50-sample patches. ImageNet uses 224-by-224 crops and ViT-L/16 (Dosovitskiy et al., 2021); masked autoencoding uses normalized patch targets (He et al., 2022). ECG and image decoders use four blocks of width 256 and eight blocks of width 512, respectively. Batch sizes are global, presentations count pretraining inputs, and pool passes use the largest pool. Input sweeps hold total presentations fixed.

Table S3: Architecture and training.

Setting	ECG	ImageNet	Robot
Blocks	12	24	12
Width	768	1,024	768
Heads	12	16	12
Fit updates	20,000	300,000	80,000
Batch size	512	1,024	512
Pretrain updates	136,719	490,447	105,469
Presentations	70,000,128	502,217,728	54,000,128
Pool passes	100	400	200
Mask fraction	0.60	0.75	0.60

Robot inputs contain 2,048 visible XYZRGB points in world coordinates, joint positions/velocities, and public task or goal information. The point encoder forms 64 groups of 32 neighbors. Point-MAE-style pretraining reconstructs local coordinates with color weight 0.1 (Pang et al., 2022). Hidden object poses, rewards, success flags, segmentation, and demonstration-stage indicators are excluded; the privileged-state reference deliberately uses simulator state.

The supervised objectives are binary cross-entropy for ECG, categorical cross-entropy for ImageNet, and canonical-action MSE for robots. Classification training objectives differ from the squared risks used for evaluation. All primary models use learned-query attention and a two-layer head with Gaussian error linear unit (GELU) activations. Frozen fixes encoder parameters and running statistics; Adapt shares its pretrained encoder and head initialization but updates both. Pretraining ends after a fixed input-only schedule. Downstream selection uses only that configuration's visible selection labels; the final step is predetermined at 1% ECG labels.

AdamW (Loshchilov & Hutter, 2019) uses cosine decay, warmup, bfloat16, and gradient-norm clipping at one. Initial encoder rates are 1.5×10^{-4} (pretraining), 5×10^{-5} (Adapt), and 2×10^{-4} (Scratch); Frozen/Adapt head rates are 10^{-3}. Matrix-valued parameters have weight decay 0.05, excluding biases and normalization. ResNet1D34/ResNet50 controls (He et al., 2016) use 20,000/120,000 ECG/ImageNet updates. Last-block adaptation updates the final block, normalization, and head. FixMatch uses confidence 0.95, equal labeled/unlabeled batch counts, consistency weight one, and stronger augmentation.

D.3 Primary and auxiliary outcomes

For test count n, predicted probabilities $\widehat{p}_{ij}$, binary ECG targets y_{ij}, and image classes $y_i \in \{1, \ldots, 1000\}$,

$$\widehat{R}_{\text{ECG}} = \frac{1}{nq} \sum_{i=1}^{n} \sum_{j=1}^{q} (\widehat{p}_{ij} - y_{ij})^2, \qquad \widehat{R}_{\text{image}} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{1000} (\widehat{p}_{ij} - \mathbf{1}\{y_i = j\})^2. \tag{D1}$$

The main ECG task has $q = 44$; image loss sums, rather than averages, over classes. Auxiliary scores are ECG macro-averaged area under the receiver operating characteristic curve (AUROC) and average precision (AP), and ImageNet top-1/top-5 accuracy (Table S6). They complement the squared-risk targets. For robot task h, with n_h records and canonical actions $y_{hi}, \widehat{y}_{hi} \in \mathbb{R}^7$,

$$\widehat{R}_{\text{robot}} = \frac{1}{3} \sum_{h=1}^{3} \frac{1}{7n_h} \sum_{i=1}^{n_h} \|\widehat{y}_{hi} - y_{hi}\|_2^2. \tag{D2}$$

Tasks receive equal weight. The seven Panda delta-pose coordinates encode translation, rotation, and gripper control. Rotation uses radial projection onto the unit ball; offline and online action conversions agree. For a fixed data-generating distribution and an independent test draw,

$$\mathbb{E}\|\widehat{f}(X) - Y\|^2 = \mathbb{E}\|\widehat{f}(X) - \mathbb{E}[Y \mid X]\|^2 + \mathbb{E}\|Y - \mathbb{E}[Y \mid X]\|^2.$$

The second term is common under the same input interface and test distribution, but may change for privileged-state inputs. Experiments report observed loss without identifying this term.

D.4 Seed selection, uncertainty, and decisions

An independent precision pilot selected 52 repetitions using 40 risks and 29 contrasts, with standard-deviation multipliers 1, 1.5, and 2. This was the smallest count satisfying all 126 precision requirements at the largest multiplier. Pilot runs contribute to development supervision and are excluded from the reported test results.

For each primary risk, $n_{\mathrm{rep}} = 52$, s_{seed} is the standard deviation of seed-level risks, and s_{test} is the cluster standard error of seed-averaged loss. Clusters are patients, images, and robot trajectories, with equal robot-task weighting. At $\alpha = 0.05$ and comparison-family size F, the interval half-width is

$$H_F = t_{n_{\mathrm{rep}}-1,\,1-\alpha/(4F)} \frac{s_{\mathrm{seed}}}{\sqrt{n_{\mathrm{rep}}}} + z_{1-\alpha/(4F)} s_{\mathrm{test}}. \tag{D3}$$

Here t and z are Student and standard-normal quantiles. The two approximate components split the nominal error budget equally and add their half-widths. Risk families contain 38 ECG, 43 ImageNet, and 100 robot comparisons, including designated auxiliary configurations. Paired contrasts use $F = 308$ and apply (D3) to seed differences and paired test-cluster errors.

Primary targets are 0.027 (ECG), 0.530 (ImageNet), and 0.055 (robot average). ECG subclass/superclass targets are 0.040/0.108; picking/stacking/insertion targets are 0.039/0.060/0.065. Targets were fixed from development-selection results before formal testing, between fitted constant-prediction and full-label pilot risks. In two-seed full-label controls, quadrupling the training schedule improved risk by 0.4%–1.4%. Passing requires the combined upper endpoint at or below target, failing requires the lower endpoint above it, and crossing intervals are undetermined.

Risk-prediction errors use pointwise 95% paired Student intervals with 51 degrees of freedom, conditional on the shared test panel. App. F.2 propagates test-cluster and seed variation through the four configurations. Closed-loop summaries distinguish seed and initial-condition uncertainty, with 2,000-resample cluster-bootstrap diagnostics.

D.5 Development supervision, hardware, and reproducibility

Table S4 reports distinct labeled records within each role; the roles overlap. Development includes fitting/selection and covers the full target pools. Units are ECG records with diagnostic targets, image/classes, or robot observation/action bundles; robot totals cover three tasks. Granularity/interface controls reuse annotations. Forecast source and target risks share one test panel disjoint from development; its labels are used only for risk estimation. The three source risks calibrate the forecast, and the fourth evaluates it. Configuration frontiers use recipes fixed after development.

Table S4: Label access by role.

Role	ECG	ImageNet	Robot
Development	19,257	1,281,167	300,000
Fitting	17,111	1,255,544	270,000
Selection	2,146	25,623	30,000
Risk analysis	2,158	50,000	30,000
Union	21,415	1,331,167	330,000

Experiments used eight NVIDIA H200 GPUs.

Within each seed group, Frozen/Adapt share labels, checkpoints, and protocol-specified head initialization and augmentations; checkpoints are reused across label budgets. Each group uses distinct random seeds. The uncertainty analysis accounts for the shared test records.

D.6 Closed-loop evaluation

Panda delta-pose policies use CPU physics and two fixed cameras, with 50/50/100 control steps for picking/stacking/insertion. Each policy/task uses 250 initial-condition episodes, or 1,000 at designated boundary and full-label settings; paired policies share initializations. A separate two-seed, 250-episode development evaluation gives expert any-time/endpoint success of 96.4%/96.0%,

95.0%/94.0%, and 85.4%/83.4% for the three tasks. These development results assess action conversion and execution separately from the policy comparisons. Any-time success means the task is ever achieved; endpoint success requires achievement at the final step, not uninterrupted success. Both differ from offline action error.

E Extended Controlled Experiments

E.1 Task families and paired evaluation

The 124 conditions form 49 coupled families, with five seed groups and 64 draws per direction: 31,360 base tasks, 79,360 paired evaluations, plus 40,960 draws for 64 access conditions. The main grid is $N = 256, 1024, 4096$; $r = 4, 16$; $\rho = 0.05, 0.20$; $\Delta = 0.025, 0.10$; context fractions $0, 1/4, 1$; and $\tau^2 = 1, \sigma^2 = 0.25$. Extensions fix 64 labels, weaken contexts, remove data, or change noise.

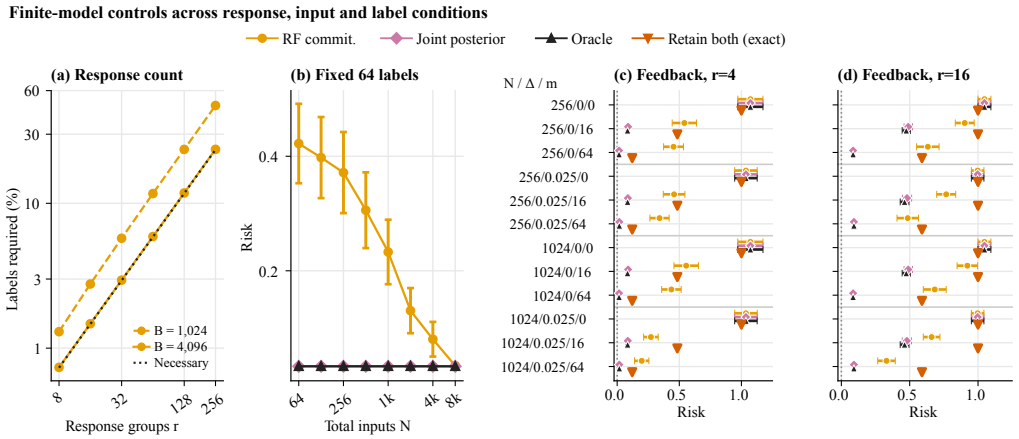


Figure S1: Finite-model risk separates response count, context evidence, and label feedback.

Oracle, RF, Joint, and cellwise shrinkage share tasks, masks, and noise. Fixed-label and access sweeps preserve labeled records and vary context prefixes, respectively. Risk and adjusted Student endpoints maximize separately over directions, without simultaneous configuration coverage. Formula comparisons and risk contrasts use separate multiplicity families.

Fig. 3 independently uses $r = 4, N = 256, \tau^2 = 1, \sigma^2 = 0.25, \Delta = 0$: 8,192 draws per direction with nested label prefixes 0–256. RF uses all contexts and shrinkage; Retain both chooses the moment rule or zero by known risk; Joint is exact. Among incorrect-RF draws, the example is nearest the coordinatewise median of three log-risk profiles at seven budgets, scaled by the interquartile range (IQR) floored at 0.15. Error cells share scale 0–2.4. Population means include every draw and maximize over directions; direction-specific 97.5% Student endpoints give approximate pointwise 95% coverage. The pooled posterior panel shows true-group probability means and 10th–90th percentiles.

E.2 Finite risks and representation controls

Fig. S1 includes all response/input curves and 24 feedback conditions with 72 method intervals. Triangles give exact retained-space moment risks with zero available. The fixed-label input scan uses $r = 8, \Delta = 0.03, \tau^2 = 1, \sigma^2 = 0.25, m = 64$ and all contexts. Its exact oracle risk is 0.0346; the simulation estimate is 0.0332. Label-budget inversions follow App. A.3, with exact shrinkage sums determining procedure-specific frontiers. Exact representation controls are given in App. B.4.

Fig. S2(a–c) evaluates the partition recursion at $r = 4, 16, 64$, distinguishing-pair affinity across noise, and the moment identity at $r = 8, \tau^2 = 1, a = 0.25, p_B = 0.1, d = 8, 11, 15$. Identical

partitions mean no label evidence, not prediction failure. Raw moment risks cross at $m = 99$, illustrating the projection/estimation tradeoff.

E.3 Calibration, rare events, and access limits

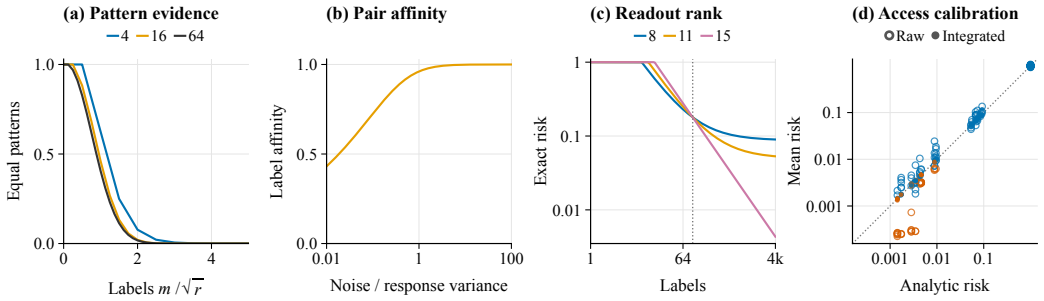


Figure S2: Label-pattern evidence, readout rank, and access-risk calibration in the finite model.

Raw Monte Carlo intervals fall below the analytic risk in 30 access comparisons, where only 0.490–1.056 wrong selections are expected per direction. Integrating over context decisions on the same tasks yields intervals containing the analytic risk in all 128 comparisons, without changing learner information. Fig. S2(d) shows the 30 rare-event discrepancies in orange and the other 98 comparisons in blue. Raw and integrated intervals use comparison families of 3,560 and 128, respectively. Of 64 configurations, 32 satisfy both access-availability conditions and meet the risk target in both directions; the remainder examine shortages or truncation. Joint-versus-RF-floor comparisons give 24 crossings, 20 noncrossings, and four undetermined outcomes; the affinity bound is supported in 36 comparisons, with 12 undetermined.

F Extended Empirical Results

F.1 Complete label curves and measured frontiers

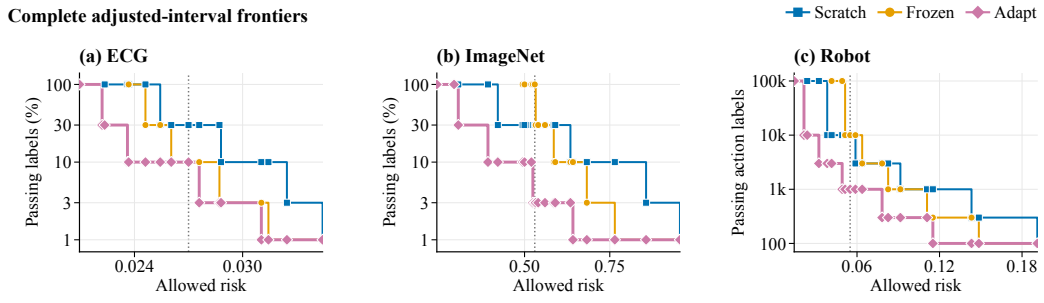


Figure S3: Adjusted-interval label frontiers for ECG, ImageNet, and robot prediction.

Table S5 gives complete largest-pool curves, with 52-seed means and adjusted intervals. Table 1 gives Frozen/Adapt crossings; Scratch passes both rules at 30% for ECG/ImageNet and 10,000 robot actions per task. Fig. S3 uses unrounded upper-endpoint breakpoints without interpolation. Adapt improves on Frozen at every primary budget; full-label Scratch also improves on Frozen.

Table S5: Complete label curves at the largest input pools: means and combined intervals.

Task / labels	Scratch	Frozen	Adapt
ECG: 1%	0.0326 [0.0308, 0.0344]	0.0296 [0.0279, 0.0314]	0.0293 [0.0275, 0.0310]
3%	0.0307 [0.0289, 0.0325]	0.0270 [0.0253, 0.0287]	0.0259 [0.0242, 0.0276]
10%	0.0271 [0.0253, 0.0288]	0.0244 [0.0228, 0.0260]	0.0221 [0.0205, 0.0236]
30%	0.0238 [0.0222, 0.0254]	0.0231 [0.0215, 0.0246]	0.0207 [0.0193, 0.0222]
100%	0.0210 [0.0196, 0.0223]	0.0223 [0.0208, 0.0237]	0.0196 [0.0183, 0.0209]
ImageNet: 1%	0.9468 [0.9378, 0.9558]	0.7497 [0.7345, 0.7648]	0.6245 [0.6071, 0.6419]
3%	0.8448 [0.8331, 0.8565]	0.6668 [0.6509, 0.6827]	0.5057 [0.4870, 0.5244]
10%	0.6168 [0.5991, 0.6346]	0.5688 [0.5519, 0.5857]	0.3767 [0.3607, 0.3927]
30%	0.4055 [0.3895, 0.4216]	0.5168 [0.5009, 0.5327]	0.2915 [0.2774, 0.3056]
100%	0.2832 [0.2732, 0.2933]	0.4829 [0.4714, 0.4945]	0.2343 [0.2259, 0.2427]
Robot: 100	0.1852 [0.1794, 0.1911]	0.1439 [0.1393, 0.1486]	0.1112 [0.1073, 0.1152]
300	0.1392 [0.1352, 0.1433]	0.1076 [0.1043, 0.1109]	0.0757 [0.0732, 0.0782]
1,000	0.0888 [0.0861, 0.0915]	0.0802 [0.0778, 0.0826]	0.0477 [0.0462, 0.0492]
3,000	0.0572 [0.0554, 0.0589]	0.0618 [0.0599, 0.0638]	0.0313 [0.0303, 0.0323]
10,000	0.0373 [0.0364, 0.0383]	0.0500 [0.0487, 0.0513]	0.0209 [0.0203, 0.0214]
100,000	0.0232 [0.0226, 0.0238]	0.0405 [0.0395, 0.0415]	0.0140 [0.0136, 0.0143]

Table S6 gives 52-seed means, with standard deviations in parentheses, all expressed as percentages. Adapt improves every auxiliary mean.

Table S6: Auxiliary prediction scores across the complete label grid.

		AUROC / Top-1 (%)			AP / Top-5 (%)		
Task	**Labels**	Scratch	Frozen	Adapt	Scratch	Frozen	Adapt
ECG (44)	1%	67.3 (1.5)	74.7 (1.6)	**75.6** (1.4)	10.1 (0.5)	16.4 (1.0)	**17.2** (0.9)
	3%	77.2 (1.4)	84.2 (1.4)	**85.8** (1.2)	15.3 (1.0)	24.4 (1.3)	**27.3** (1.5)
	10%	86.4 (0.9)	90.0 (0.7)	**92.3** (0.5)	25.9 (1.6)	33.8 (1.8)	**41.0** (1.7)
	30%	91.0 (0.6)	91.8 (0.5)	**93.7** (0.4)	36.6 (1.7)	39.1 (1.8)	**46.6** (1.8)
	100%	93.7 (0.3)	92.7 (0.3)	**94.7** (0.3)	46.3 (1.1)	42.0 (1.2)	**50.8** (1.2)
ImageNet	1%	14.4 (1.3)	38.2 (2.3)	**50.4** (2.4)	35.2 (2.6)	64.5 (3.3)	**74.4** (2.5)
	3%	27.5 (1.9)	46.5 (2.2)	**61.1** (2.3)	52.2 (2.9)	71.5 (2.6)	**81.7** (2.5)
	10%	51.5 (2.5)	55.6 (2.2)	**72.0** (1.8)	74.5 (2.4)	78.6 (2.0)	**88.1** (1.5)
	30%	69.6 (1.8)	60.2 (1.9)	**78.7** (1.5)	87.0 (1.7)	82.0 (1.7)	**91.8** (1.1)
	100%	79.6 (0.7)	63.2 (0.9)	**83.2** (0.6)	92.7 (0.9)	84.6 (1.6)	**94.3** (0.6)

F.2 Input diversity and held-out prediction

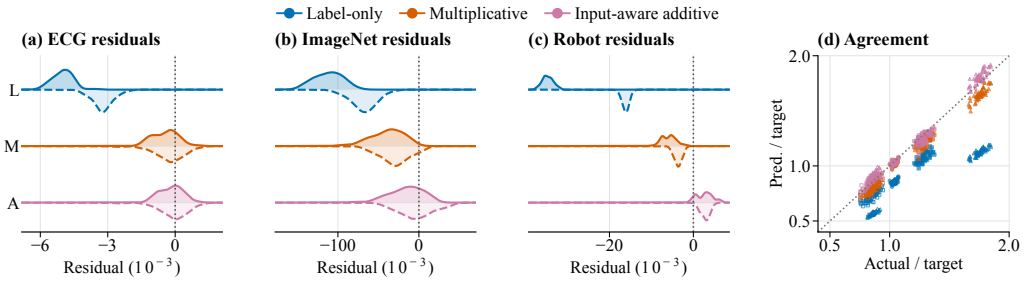


Figure S4: Signed forecast residuals and predicted-versus-observed risk across three tasks.

The held-out coordinate is (N_s, m_h), with $(N_s, N_l; m_l, m_h)$ equal to $(50{,}000, 700{,}000; 1{,}926, 5{,}777)$ for ECG, $(156{,}943, 1{,}255{,}544; 12{,}812, 128{,}117)$ for ImageNet,

and $(9{,}000, 90{,}000; 1{,}000, 3{,}000)$ per robot task. Frozen/Adapt share coordinates, and all four configurations share a test panel.

For source risks (x, y, z), additive $A = x + z - y$ preserves a common additive baseline, while raw multiplicative $M = xz/y$ obeys

$$M - A = \frac{(x - y)(z - y)}{y}. \qquad \text{(F1)}$$

Both coefficient-free rules use the three source risks directly. The observed risks satisfy $y > 0, x > y, z < y$, hence $M < A$.

Table S7 gives smaller-pool risks; the corresponding largest-pool rows of Table S5 use ECG 10%, ImageNet 1%, and 1,000 robot labels per task. Input sweeps hold these labels and presentations fixed. ECG inputs are source waveforms; robot counts exclude selection records. All twelve adjacent-pool paired intervals favor larger pools.

Table S7: Smaller input-pool risks.

Inputs	Frozen	Adapt
ECG	0.0292	0.0252
50,000	[0.0275, 0.0310]	[0.0235, 0.0269]
	0.0269	0.0236
200,000	[0.0251, 0.0286]	[0.0220, 0.0252]
ImageNet	0.8457	0.6890
156,943	[0.8323, 0.8591]	[0.6723, 0.7058]
	0.7852	0.6498
627,772	[0.7719, 0.7985]	[0.6321, 0.6675]
Robot	0.1176	0.0667
9,000	[0.1141, 0.1211]	[0.0647, 0.0687]
	0.0993	0.0575
27,000	[0.0965, 0.1022]	[0.0558, 0.0593]

Table S8 uses thousandths of risk; L/A/M mean label-only/additive/multiplicative, and F/A in task labels mean Frozen/Adapt. P/C identify paired-seed/crossed-resampling intervals. Negative contrasts favor additive. Primary 95% intervals pair 52 seeds. Sensitivity crosses 2,000 paired-seed and test-cluster resamples, recomputing predictors with shared cluster weights. Clusters are patients, class-stratified images, or within-task trajectories with equal task weights. One ImageNet Frozen source lacks record-level test losses, leaving 51 complete seeds for crossed resampling there and 52 elsewhere. All crossed A–L intervals favor additive, as do five A–M intervals; ECG Adapt spans zero, $[-0.000103, +0.000011]$.

Table S8: Risk-prediction errors and sensitivity.

		MAE			Difference [interval]	
Task	Rule	L	A	M	A − M	A − L
ECG F	P	5.004	0.417	0.555	-0.139 [-0.198, -0.079]	-4.588 [-4.697, -4.479]
	C				-0.139 [-0.205, -0.069]	-4.588 [-4.765, -4.334]
ECG A	P	3.216	0.458	0.518	-0.060 [-0.107, -0.013]	-2.758 [-2.917, -2.598]
	C				-0.060 [-0.103, 0.011]	-2.758 [-2.916, -2.502]
ImageNet F	P	111.870	19.892	38.940	-19.048 [-21.432, -16.664]	-91.979 [-95.907, -88.050]
	C				-19.268 [-21.405, -16.574]	-92.098 [-96.123, -88.070]
ImageNet A	P	67.997	16.631	29.142	-12.511 [-17.524, -7.498]	-51.366 [-55.847, -46.884]
	C				-12.511 [-17.359, -7.503]	-51.366 [-56.145, -46.594]
Robot F	P	34.706	2.733	5.874	-3.141 [-4.116, -2.167]	-31.973 [-32.696, -31.250]
	C				-3.141 [-4.113, -2.112]	-31.973 [-32.703, -31.114]
Robot A	P	16.146	2.856	3.667	-0.812 [-1.285, -0.339]	-13.290 [-13.665, -12.915]
	C				-0.812 [-1.293, -0.317]	-13.290 [-13.678, -12.901]

Additive predictions underestimate ECG/ImageNet and overestimate robot risk on average, with residual $-I$ (App. A.6). At the ECG Frozen target, additive/multiplicative/label-only reproduce observed point-risk decisions in 50/48/0 seeds. The margin condition $|\widetilde{R} - t| < |t - \varepsilon|$ guarantees the correct point-risk decision.

In Fig. 6, lower error at fixed coverage is better; pooled errors divide by each case's label-only MAE, while task panels use thousandths of risk. Fig. S4 uses Frozen above Adapt, negative residuals for underprediction, filled/open readout symbols, and large mean symbols. All densities use 52 seeds, a common bandwidth equal to $52^{-1/5}$ times the median group standard deviation within a comparison,

common height scaling, and zero reflection for nonnegative risks. The densities summarize variation across seeds; coverage curves are unsmoothed.

F.3 Update depth, domain, objective, and granularity

Table S9 reports eight controls, with negative risk differences favoring the alternative. Update depth, objective, and backbone change the measured risks, so requirements remain recipe-dependent.

Table S9: Paired update-depth, pretraining-domain, objective, and backbone controls.

Task	Labels	Reference	Alternative	Risk difference [interval]	Gain (%)
ImageNet	10%	Frozen	Last block	-0.1026 [-0.1141, -0.0911]	18.0
ImageNet	10%	Last block	Adapt	-0.0895 [-0.1011, -0.0779]	19.2
ECG	10%	Source F	Target F	-0.0028 [-0.0033, -0.0022]	8.9
ECG	10%	Source A	Target A	-0.0035 [-0.0042, -0.0028]	13.1
ImageNet	3%	Adapt	FixMatch	-0.0176 [-0.0294, -0.0058]	3.5
ImageNet	10%	Adapt	FixMatch	-0.0144 [-0.0241, -0.0047]	3.8
ImageNet	10%	Scratch ViT	Scratch CNN	-0.0216 [-0.0355, -0.0078]	3.5
ECG	30%	Scratch Tr.	Scratch CNN	-0.0019 [-0.0024, -0.0015]	8.1

Table S9's target-domain ECG comparisons match 17,111 inputs, presentations, and 10% labels, and both paired intervals favor target-domain pretraining. Fig. 4(d,h) combines domain and granularity controls, normalizing each granularity by its own target. Table S10 reports 10% and 30% budgets for 23 and five targets. Adapt passes both targets at 10%; Frozen passes at 30% after an undetermined 10%, while Scratch remains undetermined at 30%. Compare within target definitions because prevalence and averaging differ.

Table S10: ECG granularity controls: means and combined intervals.

Targets	Labels	Scratch	Frozen	Adapt
23	10%	0.0445 [0.0417, 0.0473]	0.0394 [0.0367, 0.0420]	0.0349 [0.0324, 0.0375]
23	30%	0.0383 [0.0357, 0.0408]	0.0368 [0.0343, 0.0393]	0.0327 [0.0303, 0.0350]
5	10%	0.1254 [0.1180, 0.1327]	0.1074 [0.1001, 0.1146]	0.0926 [0.0859, 0.0993]
5	30%	0.1040 [0.0969, 0.1110]	0.0985 [0.0916, 0.1053]	0.0848 [0.0784, 0.0911]

F.4 Task risk and complete closed-loop outcomes

Table S11 reports boundary and full-label task risks in canonical action coordinates. At 1,000 labels per task, Adapt passes each task target and their average-risk target. Table S12 reports full-label success percentages; parentheses give standard errors over initial conditions. State is the privileged-input reference. Success uses App. D.6's definitions.

Table S11: Robot task risks, with combined intervals.

Labels / task	Model	Picking	Stacking	Insertion
1,000	Scratch	0.0626 [0.0596, 0.0657]	0.0981 [0.0937, 0.1025]	0.1057 [0.1012, 0.1102]
1,000	Frozen	0.0495 [0.0470, 0.0520]	0.0874 [0.0833, 0.0914]	0.1038 [0.0994, 0.1082]
1,000	Adapt	0.0359 [0.0342, 0.0377]	0.0531 [0.0506, 0.0555]	0.0541 [0.0517, 0.0566]
3,000	Scratch	0.0368 [0.0351, 0.0386]	0.0626 [0.0599, 0.0653]	0.0720 [0.0690, 0.0750]
3,000	Frozen	0.0348 [0.0332, 0.0364]	0.0669 [0.0638, 0.0699]	0.0839 [0.0804, 0.0873]
3,000	Adapt	0.0219 [0.0209, 0.0230]	0.0348 [0.0333, 0.0364]	0.0372 [0.0355, 0.0388]
100,000	Scratch	0.0116 [0.0111, 0.0120]	0.0246 [0.0236, 0.0256]	0.0334 [0.0323, 0.0346]
100,000	Frozen	0.0174 [0.0168, 0.0181]	0.0421 [0.0405, 0.0437]	0.0620 [0.0598, 0.0642]
100,000	Adapt	0.0078 [0.0074, 0.0081]	0.0153 [0.0146, 0.0160]	0.0188 [0.0181, 0.0195]

Table S12: Full-label success rates.

Policy	Pick: any	Pick: final	Stack: any	Stack: final	Insert: any	Insert: final
Scratch	90.93 (0.60)	88.81 (0.72)	77.57 (0.91)	73.09 (1.04)	57.30 (1.11)	48.27 (1.20)
Frozen	88.38 (0.68)	86.32 (0.79)	71.92 (1.00)	67.67 (1.09)	48.00 (1.12)	40.54 (1.17)
Adapt	92.42 (0.54)	90.32 (0.67)	81.04 (0.85)	76.51 (1.00)	61.78 (1.10)	52.07 (1.22)
State	97.92 (0.25)	95.74 (0.51)	92.73 (0.52)	87.70 (0.82)	80.51 (0.87)	67.88 (1.21)

Figs. 5 and S5 reuse the same policies and initializations for paired outcomes and full success curves. Boxes give IQR/median and full seed ranges; large symbols mark means. Offline intervals combine seed and cluster variation, while closed-loop bands use seeds only. The final column of Fig. S5 contrasts Frozen/Scratch in yellow/blue and endpoint/any-time success with filled/open marks.

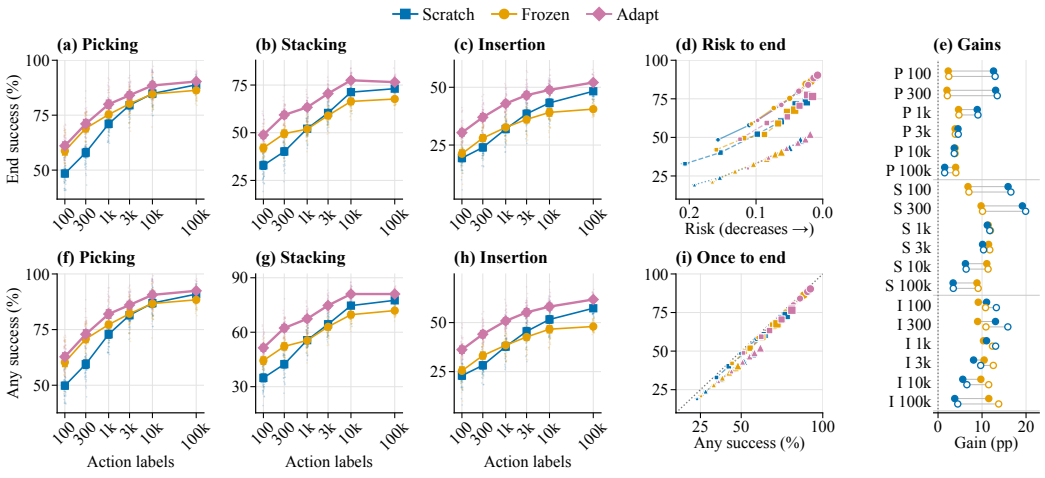


Figure S5: Robot label-budget profiles relate offline risk to endpoint and any-time success.