WHEN SHOULD A DOMAIN-GENERALIZED MODEL SPECIALIZE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Limited-label adaptation raises two questions: whether a fixed policy improves a domain-generalized model, and whether episode-specific policy selection provides additional value. The Cross-Fitted Domain-Generalization Specialization Assessment (CF-DGSA) separates fixed adaptation gain, held-out selection premium, and total gain over zero-shot inference. Its comparator is the accuracy-maximizing fixed policy on a labeled audit panel, while a routing margin affects only selection. Symmetric cross-fitting and a matched-half diagnostic distinguish held-out value from same-fold reuse. Selection-premium intervals project paired component intervals, with operating characteristics evaluated under Gaussian and paired-image models. Across nine configurations on PACS, Office-Home, and VLCS, four selection-premium intervals are positive, but only three corresponding total-gain intervals are positive. VLCS ResNet-50 has a $+0.80\,\mathrm{pp}$ selection premium and a $-3.00\,\mathrm{pp}$ total gain: improved selection among adaptation policies can still underperform zero-shot inference. Reproduced ERM++ models further distinguish beneficial fixed adaptation without resolved selection value, observed policy degeneracy, and harmful adaptation. The assessment separates evidence for adaptation from evidence for policy selection under explicit audit-label access.

1 INTRODUCTION

Domain generalization (DG) trains predictors for target domains excluded from source training (Gulrajani & Lopez-Paz, 2021). Residual target error motivates limited-label adaptation but does not establish its benefit. Likewise, variation among adaptation policies does not establish that episode-specific selection improves on one fixed policy. The practical alternatives are zero-shot inference with the unchanged source checkpoint, fixed adaptation, and episode-specific selection. Retrospective oracle gaps can obscure these alternatives by selecting and evaluating on the same finite data (Cawley & Talbot, 2010; Yu et al., 2024), while improvement over zero-shot inference differs from improvement over the best fixed adaptation policy.

CF-DGSA distinguishes fixed adaptation gain, G_{fixed}, from the cross-fitted selection premium, X_{xfit}, and total gain over zero-shot inference, G_{total}. Adaptation uses labeled support; a separate audit panel provides labeled target examples for fold-specific selection and opposite-fold evaluation. A matched-half diagnostic compares same-fold and opposite-fold evaluation without increasing the selection budget. The routing margin is the observed advantage required to select the more extensive policy, not a modification of the best-fixed comparator. This is an offline, label-assisted assessment, not a support-only deployment rule or an online or unlabeled test-time adaptation method.

The evaluation includes ResNet-18/50 baselines (He et al., 2016) and reproduced ERM++ checkpoints (Teterwak et al., 2025) on PACS, Office-Home, and VLCS. ERM++ reproduction is within $0.15\,\mathrm{pp}$ of published benchmark values, enabling assessment on competitive source models. The purpose is to determine what the observed adaptation results support, rather than to introduce another source-training or adaptation algorithm. A common assessment distinguishes improvements obtainable with one fixed adaptation rule from those requiring episode-specific policy selection. The contributions are:

- CF-DGSA separates fixed adaptation, episode-specific selection, and total gain using an audit-based best-fixed comparator independent of the routing margin.

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.

- Through symmetric cross-fitting and hierarchical resampling, the procedure estimates paired selection contrasts and constructs component-projected intervals.
- Across nine configurations, positive selection value can coexist with harmful adaptation, demonstrating why adaptation benefit and selection benefit require separate assessment.

2 Related Work

2.1 Domain Generalization

DomainBed established controlled multi-domain evaluation and identified carefully selected empirical risk minimization (ERM) as a strong DG baseline (Gulrajani & Lopez-Paz, 2021). ERM++ strengthened source-only training through optimization, model-selection, and averaging choices (Teterwak et al., 2025). CF-DGSA assesses subsequent adaptation without modifying source training; competitive ERM++ reproduction reduces the concern that adaptation behavior merely reflects weak source models. Recent CLIP studies address diverse-prompt DG (Wen et al., 2025) and frozen-model adaptation from few unlabeled target examples (Chi et al., 2025). Although CF-DGSA operates on paired policy outcomes, its validation here is limited to residual networks, and its labeled audit protocol differs from adaptation using only unlabeled target examples.

2.2 Evaluation Protocols and Policy Selection

DG comparisons depend on architecture, optimization, model selection, and access to target information (Gulrajani & Lopez-Paz, 2021; Yu et al., 2024). Reusing evaluation data for selection can bias reported performance (Cawley & Talbot, 2010). CF-DGSA assesses held-out selection value relative to fixed and zero-shot references, paralleling single-best and virtual-best references in algorithm selection (Rice, 1976; Lindauer et al., 2019). Each policy is instantiated from a checkpoint and support draw, whose paired lineage must be retained. Audit labels select policies but never update neural parameters.

2.3 Performance Estimation and Statistical Inference

Out-of-distribution performance estimation concerns frozen-model accuracy (Miller et al., 2021; Baek et al., 2022; Mishra & Liu, 2025); CF-DGSA instead assesses subsequent labeled-support decisions. Sample splitting and cross-fitting separate selection from evaluation (Cawley & Talbot, 2010; Chernozhukov et al., 2018), whereas Multi-Condition Conformal Selection controls false discoveries under multiple criteria (Hao et al., 2026). Optimal policy values can be nonregular near ties (Luedtke & van der Laan, 2016). Expressing the selection premium as a minimum of paired contrasts permits component projection without solving nonregular inference generally. The contribution is a comparator-aware post-DG assessment with explicit label access and empirical operating characteristics, not cross-fitting or resampling in isolation.

3 Method

3.1 Problem Definition

Let $t \in \{1, \ldots, T\}$ index held-out targets, $s \in \{1, \ldots, n_t\}$ independently trained source checkpoints within target t, and $r \in \{1, \ldots, R_{t,s}\}$ labeled support draws. An evaluation episode is $e = (t, s, r)$, with source checkpoint $f_{0,t,s}$ and class-balanced support $S_{t,s,k,r}$ containing k labeled examples per class. For target class set $\mathcal{C}_t$, the support size is $b_t = k|\mathcal{C}_t|$; this is a label budget, not an optimizer batch size. Support is drawn from the cell's adaptation pool, disjoint from its frozen audit manifest. An episode is an offline evaluation unit, and class-balanced sampling is an experimental convention rather than a deployment acquisition model.

The candidate set is $\Pi^C = \{H, A\}$. Policy H freezes the encoder, keeps all Batch Normalization (BN) modules in evaluation mode (Ioffe & Szegedy, 2015), and updates only the classifier. Policy A updates the final residual block and classifier; earlier BN modules remain in evaluation mode, whereas final-block BN modules update their running statistics. Both use Adam (Kingma &

Ba, 2015) for 200 steps without target-validation early stopping. The predictor f_e^π uses an order-independent, policy-specific random stream $\xi_{e,\pi}$, so draw-level variation includes support composition and adaptation randomness. These operational configurations do not isolate weight adaptation from changes in BN behavior. Assessment performs no neural optimization after paired predictions are cached. Figure 1(a) shows the shared checkpoint and support and the two adaptation branches.

Each checkpoint–manifest cell has a fixed, class-stratified audit manifest $\mathcal{T}_{t,s}^{\text{test}}$ of size $N_{t,s}$ and L_{fold} prespecified two-fold partitions. Fold indices $j \in \{A, B\}$ are distinct from policy A. Equation (1) defines example-level accuracy, fold scores, and sample-size weights; class-balanced support does not imply class-macro averaging of audit outcomes.

$$M(f;\mathcal{T}) = \frac{1}{|\mathcal{T}|}\sum_{(x,y)\in\mathcal{T}} \mathbf{1}\{f(x)=y\},$$
$$Y_{e,\ell,j}^{\pi} = M(f_e^\pi;\mathcal{T}_{t,s,\ell,j}), \qquad w_{t,s,\ell,j} = \frac{|\mathcal{T}_{t,s,\ell,j}|}{N_{t,s}}. \tag{1}$$

Equation (2) gives full-manifest accuracy and target-equal aggregation, weighting cells equally within targets and draws equally within cells. Repeated partitions do not create independent audit samples. Unstarred quantities are empirical summaries; stars denote repeated-sampling targets.

$$\bar{Y}_e^\pi = \frac{1}{L_{\text{fold}}}\sum_{\ell=1}^{L_{\text{fold}}}\sum_{j\in\{A,B\}} w_{t,s,\ell,j} Y_{e,\ell,j}^{\pi},$$
$$\widehat{\mathbb{E}}_{\text{hier}}[Z] = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{n_t}\sum_{s=1}^{n_t}\frac{1}{R_{t,s}}\sum_{r=1}^{R_{t,s}} Z_{t,s,r}. \tag{2}$$

Beyond the b_t adaptation labels, directional selection uses one audit fold and evaluation uses the opposite fold. Bidirectional assessment accesses all $N_{t,s}$ audit labels, reused across partitions. Audit size is separate from the support budget k. The routing rule in Eq. (3) retains H unless A exceeds it by δ_{eq}. The primary margin is 0.005 accuracy units, or 0.5 pp; it is an operational preference, not a statistical equivalence test or a threshold on aggregate selection value.

$$\pi_\delta(v) = \begin{cases} A, & v^A - v^H > \delta_{\text{eq}}, \\ H, & \text{otherwise.} \end{cases} \tag{3}$$

3.2 Comparator-Aware Cross-Fitted Estimation

Zero-shot accuracy is evaluated on the same audit manifests as the adapted policies. The notation A_0^{std} denotes source-model reproduction accuracy, whereas $A_0 = A_0^{\text{audit}}$ is used in the assessment. Equation (4) defines policy means, the best-fixed value, and fixed adaptation gain. The best-fixed value is a panel-oracle comparator based on labeled audit outcomes, not source-validation policy selection. Fixed refers to a common adaptation rule across episodes, not an identical fitted predictor. The comparator is independent of the routing margin.

$$A_0 = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{n_t}\sum_{s=1}^{n_t} M(f_{0,t,s};\mathcal{T}_{t,s}^{\text{test}}),$$
$$Q_\pi = \widehat{\mathbb{E}}_{\text{hier}}[\bar{Y}_e^\pi], \qquad Q_{\text{fixed}}^{\max} = \max_{\pi\in\Pi^C} Q_\pi, \qquad G_{\text{fixed}} = Q_{\text{fixed}}^{\max} - A_0. \tag{4}$$

Equation (5) selects on each fold and evaluates on the opposite fold for C_e, or on the same fold for S_e. Both directions use evaluation-fold weights. The two quantities use the same fitted predictors, audit partitions, and fold-specific selections. Their difference therefore concerns where a selected

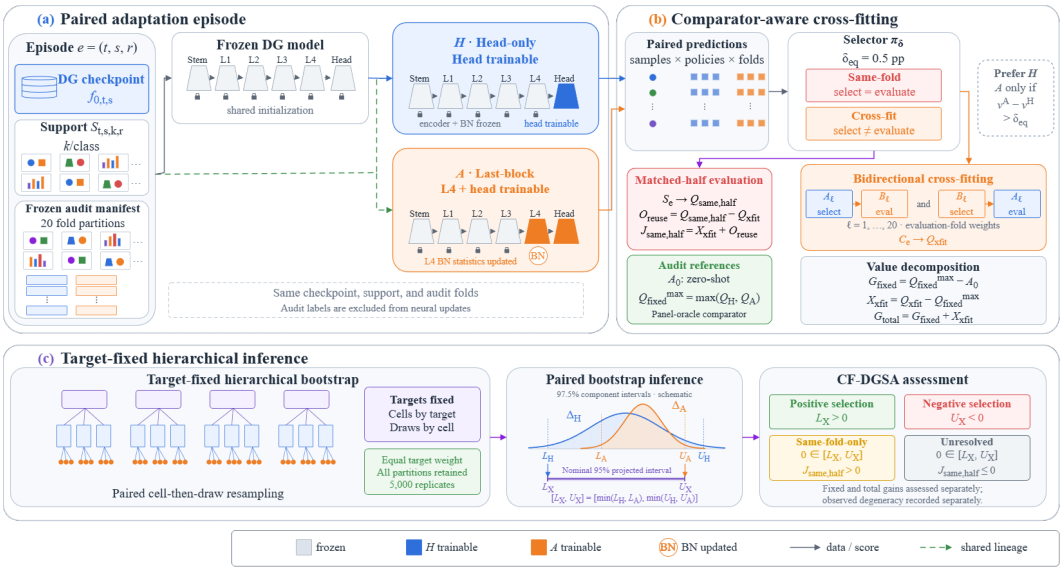

Figure 1: CF-DGSA assessment: (a) paired head-only and final-block adaptation; (b) comparator-aware cross-fitting with separate selection and evaluation paths; (c) target-fixed hierarchical resampling and component-projected inference. Audit labels are excluded from neural updates.

policy is evaluated, while preserving the label budget of each selection decision.

$$\widehat{\pi}_{e,\ell,j} = \pi_\delta\big((Y^{\pi}_{e,\ell,j})_{\pi\in\Pi^C}\big), \qquad j \in \{A, B\},$$

$$C_e = \frac{1}{L_{\text{fold}}} \sum_{\ell=1}^{L_{\text{fold}}} \left(w_{t,s,\ell,B} Y^{\widehat{\pi}_{e,\ell,A}}_{e,\ell,B} + w_{t,s,\ell,A} Y^{\widehat{\pi}_{e,\ell,B}}_{e,\ell,A} \right),$$

$$S_e = \frac{1}{L_{\text{fold}}} \sum_{\ell=1}^{L_{\text{fold}}} \left(w_{t,s,\ell,A} Y^{\widehat{\pi}_{e,\ell,A}}_{e,\ell,A} + w_{t,s,\ell,B} Y^{\widehat{\pi}_{e,\ell,B}}_{e,\ell,B} \right). \tag{5}$$

Equation (6) separates fixed adaptation, selection value, total improvement, and same-fold reuse. A positive selection premium need not imply beneficial adaptation when the fixed comparator underperforms zero-shot inference. Figure 1(b) traces audit-fold selection and the same-fold and opposite-fold evaluation paths.

$$\begin{aligned} Q_{\text{xfit}} &= \widehat{\mathbb{E}}_{\text{hier}}[C_e], \qquad Q_{\text{same,half}} = \widehat{\mathbb{E}}_{\text{hier}}[S_e], \\ X_{\text{xfit}} &= Q_{\text{xfit}} - Q^{\max}_{\text{fixed}}, \qquad O_{\text{reuse}} = Q_{\text{same,half}} - Q_{\text{xfit}}, \\ J_{\text{same,half}} &= Q_{\text{same,half}} - Q^{\max}_{\text{fixed}} = X_{\text{xfit}} + O_{\text{reuse}}, \\ G_{\text{total}} &= Q_{\text{xfit}} - A_0 = G_{\text{fixed}} + X_{\text{xfit}}. \end{aligned} \tag{6}$$

For one episode and partition, suppress those indices and let m_j denote the observed $A - H$ margin and I_j indicate that fold j selects A. Direct subtraction yields Eq. (7). Equal-fold nonnegativity follows from threshold monotonicity; magnitude, rather than positivity alone, is diagnostic. Unequal folds introduce directional weighting, so the gap need not be nonnegative. Comparing against a full-manifest same-test selector would also change selection-set size.

$$\begin{aligned} m_j &= Y^A_j - Y^H_j, \qquad I_j = \mathbf{1}\{m_j > \delta_{\text{eq}}\}, \\ S - C &= (I_A - I_B)(w_A m_A - w_B m_B), \\ S - C &= \tfrac{1}{2}(I_A - I_B)(m_A - m_B) \ge 0 \qquad \text{if } w_A = w_B = \tfrac{1}{2}. \end{aligned} \tag{7}$$

3.3 Hierarchical Inference and Conditional Decisions

Mathematical maxima at finer levels provide the ordered decomposition in Eq. (8). The target and cell contrasts are nonnegative empirical oracle opportunities. The final term is a signed remain-

der combining support and adaptation randomness, selection error, and the transition from same-manifest references to held-out evaluation. This decomposition is descriptive, neither causal nor permutation invariant, and does not isolate a pure support effect.

$$
\begin{aligned}
Q_T &= \frac{1}{T}\sum_{t=1}^{T}\max_{\pi\in\Pi^C}\frac{1}{n_t}\sum_{s=1}^{n_t}\frac{1}{R_{t,s}}\sum_{r=1}^{R_{t,s}}\bar{Y}^{\pi}_{t,s,r},\\
Q_{TS} &= \frac{1}{T}\sum_{t=1}^{T}\frac{1}{n_t}\sum_{s=1}^{n_t}\max_{\pi\in\Pi^C}\frac{1}{R_{t,s}}\sum_{r=1}^{R_{t,s}}\bar{Y}^{\pi}_{t,s,r},\\
D_T &= Q_T - Q^{\max}_{\text{fixed}}, \qquad D_{S|T} = Q_{TS} - Q_T,\\
D_{R,\text{xfit}|T,S} &= Q_{\text{xfit}} - Q_{TS}, \qquad X_{\text{xfit}} = D_T + D_{S|T} + D_{R,\text{xfit}|T,S}.
\end{aligned}
\tag{8}
$$

Let m_e denote the population $A-H$ accuracy margin and $\pi^\star_e$ the better policy. Equation (9) decomposes the value X_g of a selector g beyond the population best-fixed policy into available opportunity and regret. The first term follows from the absolute-difference representation of a two-policy maximum; the second is the loss from selecting the inferior policy. Policy ordering must change to create opportunity, and sufficiently accurate selection must recover it. The expectation includes selector randomness; this is not a finite-manifest unbiasedness theorem.

$$
X_g = \frac{\mathbb{E}|m_e| - |\mathbb{E}m_e|}{2} - \mathbb{E}\left[|m_e|\mathbf{1}\{g(e)\neq\pi^\star_e\}\right]. \tag{9}
$$

The repeated-sampling law $\mathcal{P}$ fixes the observed targets, source datasets, adaptation pools, and cell-construction protocol. A checkpoint and its manifest form a joint cell, within which support and adaptation randomness vary. Equation (10) takes the population maximum after expectation. In the hierarchical bootstrap (Efron & Tibshirani, 1993), targets remain fixed and equally weighted; cells are resampled within targets, followed by draws within selected cells. All policy outcomes, zero-shot scores, and repeated partitions are retained together. This represents uncertainty within the observed target panel, not over unseen domains. Figure 1(c) depicts the resampling hierarchy and interval projection.

$$
\begin{aligned}
Q^\star_\pi &= \mathbb{E}_{\mathcal{P}}[Q_\pi], \quad Q^\star_{\text{xfit}} = \mathbb{E}_{\mathcal{P}}[Q_{\text{xfit}}], \quad A^\star_0 = \mathbb{E}_{\mathcal{P}}[A_0],\\
Q^{\max,\star}_{\text{fixed}} &= \max_\pi Q^\star_\pi, \qquad X^\star = Q^\star_{\text{xfit}} - Q^{\max,\star}_{\text{fixed}}.
\end{aligned}
\tag{10}
$$

Selection-premium inference uses the paired mean contrasts in Eq. (11). Each component receives a two-sided 97.5% percentile interval using the 1.25th and 98.75th bootstrap percentiles. If each component has at most 1.25% error in each tail, a union bound gives at least 95% simultaneous coverage, which transfers to the projected minimum. The two lower-endpoint failure events give at least 97.5% lower-endpoint validity; component independence is unnecessary. Projection does not establish component calibration, so small-cell behavior is evaluated empirically. With three to five cells per target, increasing bootstrap replicates improves Monte Carlo precision, not independent checkpoint–manifest information.

$$
\Delta_\pi = Q_{\text{xfit}} - Q_\pi, \quad X_{\text{xfit}} = \min_{\pi\in\{H,A\}}\Delta_\pi, \quad [L_X, U_X] = \left[\min_\pi L_\pi, \min_\pi U_\pi\right]. \tag{11}
$$

Total gain uses a paired two-sided 95% percentile interval. Fixed-gain intervals are conventional 95% percentile summaries that recompute the maximum in each replicate; they do not inherit the projection argument or a boundary-uniform guarantee. A positive lower endpoint supports the corresponding gain, a negative upper endpoint supports the opposite comparison, and an interval containing zero leaves its sign unresolved. A same-fold-only diagnostic applies when the selection interval contains zero but the same-half premium is positive. Observed policy degeneracy, in which every fold selects the same policy, is recorded separately. These are marginal configuration-level assessments, not a simultaneous nine-configuration error guarantee.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

The evaluation uses PACS (Li et al., 2017), Office-Home (Venkateswara et al., 2017), and VLCS (Fang et al., 2013) with ImageNet-pretrained ResNet-18/50 (Deng et al., 2009; He et al., 2016) and reproduced ERM++ checkpoints (Teterwak et al., 2025). Each target has three checkpoint–manifest cells, except the VLCS ResNet baselines, which have five. Each cell uses ten support draws at $k = 10$ labels per class and 20 bidirectional partitions; intervals use 5,000 hierarchical bootstrap replicates. Audit manifests contain 334–787, 485–889, and 284–676 examples, respectively, disjoint from support. Adaptation uses learning rate 10^{-3} and the 200-step schedule in Section 3.1. Resolution is 128×128 for ResNet-18 and 224×224 otherwise. Routing sensitivity varies $\delta_{\mathrm{eq}} \in \{0, 0.001, 0.0025, 0.005\}$ without retraining or changing the fixed comparator.

The Gaussian score model centers policy scores at 0.60, offsets them symmetrically by half the episode-level $A - H$ margin, and adds independent zero-mean Gaussian fold noise. Margins combine target effects with independent zero-mean Gaussian cell and draw effects. The main grid uses target margins $(-\alpha, \alpha, \alpha, -\alpha)$, no cell variation, and draw standard deviation one-quarter of fold noise. Low, median, high, and stress values of σ_{fold} are 0.0045, 0.0065, 0.0088, and 0.04. Each setting uses four targets, three cells per target, ten draws per cell, one fold pair, and 10,000 trials. The amplitude is adjusted independently for each noise level to target effects of $\{0, 0.25, 0.5, 0.75, 1, 1.5, 2\}$ pp, approximately holding population effect fixed. With Φ denoting the standard normal cumulative distribution function, the held-out population premium is

$$X^{\star}_{\mathrm{heldout}} = \mathbb{E}\left[m_e \Phi\left(\frac{m_e - \delta_{\mathrm{eq}}}{\sqrt{2}\sigma_{\mathrm{fold}}}\right)\right] - \max\{\mathbb{E}[m_e], 0\}. \tag{12}$$

Positive claims require a lower endpoint above zero; their frequency measures power at positive effects and type-I error under a sign null. Nominal-zero grid points retain draw variation and differ from the separate median-noise exact null. Coverage measures interval containment of the population value, whereas lower-endpoint validity measures whether the lower bound does not exceed it. Small positive effects within 0.1 pp are near-null alternatives; claims beyond that threshold require a lower endpoint above it. One-way holdout omits fold reversal, and the same-test oracle reuses selection outcomes rather than implementing the matched-half diagnostic. The comparison does not impose a common type-I error.

The paired-image study thresholds uniform draws at the latent policy scores, sharing each episode–image draw between policies with probability 0.8 and using independent draws otherwise, with conditional independence across pairs given their latent margins. It retains the four-target, three-cell, ten-draw hierarchy and reuses 200 image pairs per episode across 20 balanced partitions. In accuracy units, exact null and selector boundary have constant margins of 0 and 0.005 with no latent variation. Target opportunity and nonzero cell use margins $(-0.02, 0.02, 0.02, -0.02)$ with cell/draw standard deviations $(0.003, 0.004)$ and $(0.00325, 0.001625)$, respectively. Draw opportunity has zero target margins, no cell variation, and draw standard deviation 0.02. Each regime uses 500 trials; population values are estimated independently, taking the maximum after averaging.

4.2 EXPERIMENTAL RESULTS

At median noise, 10,000 exact-null trials produce no positive claims, with a two-sided 95% Wilson upper endpoint of 0.038% (Wilson, 1927); the median nominal-zero point also produces none. Near-null positive-claim rates at low, high, and stress noise are 0.01%, 0.01%, and 0.03%. Power is 1.00 at an achieved median-noise effect of 0.981 pp and 0.4403 at the approximately matched stress effect of 1.005 pp. Across the 28-point grid, minimum coverage is 98.29% and minimum lower-endpoint validity is 99.94%. Reading power alongside the achieved effect distinguishes the magnitude of the population selection premium from its detectability using finite fold scores. The stress regime retains a similar positive premium, but the interval procedure resolves it less frequently. Figure 2 separates method comparison from noise sensitivity using achieved held-out effects. These results characterize the evaluated data-generating processes, not uniform error control.

Across the paired-image regimes, coverage is 97.0% at the exact null, 98.6% at the selector boundary, 99.2% for target opportunity, 99.8% for draw opportunity, and 98.6% for nonzero cell variation.

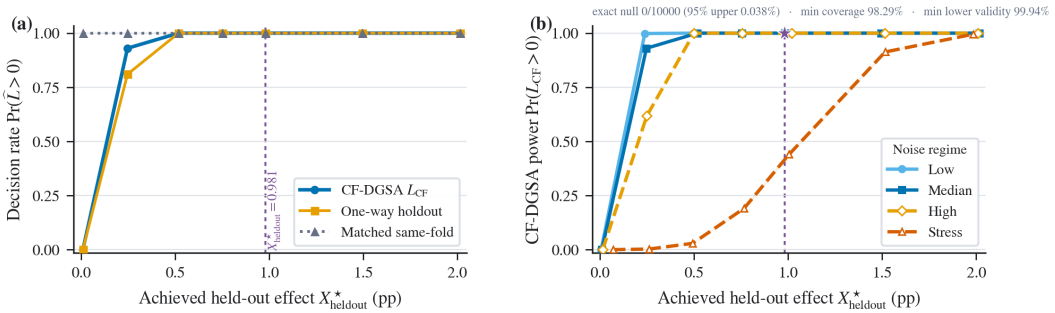

Figure 2: Operating characteristics: (a) method comparison at median noise; (b) CF-DGSA noise sensitivity. Bars are 95% Wilson intervals for positive-claim frequencies over 10,000 trials. Nominal-zero grid points are near-null effects, not the separate exact null.

Table 1: Fixed adaptation gain, selection premium, and total gain at $k = 10$.

Dataset	Base	A_0	G_{fixed}	X_{xfit}	G_{total}	Outcome
PACS	ResNet-18	71.93	$+5.76\,[4.73, 6.76]$	$+0.10\,[-0.13, 0.41]$	$+5.87\,[4.75, 6.99]$	F
	ResNet-50	80.61	$+3.98\,[2.90, 5.09]$	$+0.87\,[0.31, 1.52]$	$+4.84\,[3.47, 6.32]$	S+/T+
	ERM++	89.78	$+1.87\,[1.32, 2.44]$	$-0.02\,[-0.08, 0.08]$	$+1.84\,[1.31, 2.41]$	F
Office-Home	ResNet-18	55.09	$+4.33\,[3.81, 4.87]$	$+1.68\,[1.35, 2.02]$	$+6.01\,[5.53, 6.50]$	S+/T+
	ResNet-50	70.11	$+3.52\,[2.94, 4.09]$	$+0.68\,[0.39, 0.98]$	$+4.20\,[3.56, 4.83]$	S+/T+
	ERM++	74.44	$+3.63\,[3.27, 4.00]$	$0.00\,[0.00, 0.00]$	$+3.63\,[3.27, 4.00]$	D
VLCS	ResNet-18	71.77	$-5.33\,[-6.27, -4.38]$	$+0.19\,[-0.05, 0.49]$	$-5.14\,[-6.05, -4.20]$	T−
	ResNet-50	79.91	$-3.80\,[-4.74, -2.92]$	$+0.80\,[0.35, 1.30]$	$-3.00\,[-3.97, -2.03]$	S+/T−
	ERM++	77.84	$-2.95\,[-3.93, -1.93]$	$+0.17\,[-0.04, 0.46]$	$-2.77\,[-3.81, -1.72]$	T−

F: positive fixed-gain interval with unresolved selection. D: observed policy degeneracy. S/T signs: selection/total interval signs. T− alone: negative total gain with unresolved selection.

Lower-endpoint validity is 100% in all five 500-trial evaluations. These tests incorporate paired outcomes and repeated partitions but do not establish calibration for every real-data distribution or audit size.

Reproduced ERM++ accuracies are 89.86%, 74.55%, and 78.14% on PACS, Office-Home, and VLCS, versus published values of 89.80%, 74.70%, and 78.00% (Teterwak et al., 2025); these differ from the audit scores in Table 1. This check documents competitive source performance, not DG-training superiority. Four selection-premium intervals are positive—Office-Home ResNet-18 and all three ResNet-50 baselines—but only three corresponding total-gain intervals are positive. VLCS ResNet-50 has $X_{\text{xfit}} = +0.80\,\text{pp}\ [0.35, 1.30]$ and $G_{\text{total}} = -3.00\,\text{pp}\ [-3.97, -2.03]$: selection outperforms fixed adaptation without recovering zero-shot accuracy. All VLCS total-gain intervals are negative.

Table 1 reports A_0 in percent and gains in percentage points, with nominal 95% intervals in brackets. Fixed and total gains use percentile intervals; selection premiums use component projection. Outcome abbreviations are defined beneath the table. Rounding may produce 0.01 pp discrepancies in additive identities.

For ERM++, PACS supports fixed head-only adaptation, with a 1.87 pp fixed gain and an unresolved selection premium of -0.02 pp $[-0.08, 0.08]$. An interval containing zero does not establish a zero population premium. Office-Home is policy-degenerate: every fold selects H, fixed gain is 3.63 pp, and the selection premium, same-half premium, and reuse diagnostic are zero. Its zero-width interval reflects observed resampling support, not future dominance. On VLCS, fixed and selected adaptation remain harmful. Across the eight nondegenerate configurations, O_{reuse} ranges from 0.09 to 0.47 pp, consistent with the equal-fold identity in Section 3.2.

Table 2 reports target- and cell-level opportunities and the signed support-realization and cross-fitting remainder in percentage points. Office-Home ResNet-18 has 1.91 pp of target-level oppor-

Table 2: Ordered hierarchical decomposition of the selection premium.

Dataset	Base	X_{xfit}	D_T	$D_{S\|T}$	$D_{R,\mathrm{xfit}\|T,S}$
PACS	ResNet-18	$+0.10$	$+0.00$	$+0.05$	$+0.05$
	ResNet-50	$+0.87$	$+0.63$	$+0.10$	$+0.13$
	ERM++	-0.02	$+0.00$	$+0.00$	-0.02
Office-Home	ResNet-18	$+1.68$	$+1.91$	$+0.00$	-0.23
	ResNet-50	$+0.68$	$+0.72$	$+0.00$	-0.04
	ERM++	0.00	0.00	0.00	0.00
VLCS	ResNet-18	$+0.19$	$+0.00$	$+0.15$	$+0.05$
	ResNet-50	$+0.80$	$+0.45$	$+0.05$	$+0.30$
	ERM++	$+0.17$	$+0.00$	$+0.02$	$+0.15$

tunity and a $-0.23\,\mathrm{pp}$ remainder. PACS ResNet-50 combines target- and cell-level opportunities of 0.63 and $0.10\,\mathrm{pp}$ with a $0.13\,\mathrm{pp}$ remainder. VLCS ResNet-50 has $0.45\,\mathrm{pp}$ of target-level opportunity despite negative total gain. For Office-Home ResNet-18, the negative remainder means that $Q_{\mathrm{xfit}} < Q_{TS}$ even though the selector exceeds the panel-level best-fixed comparator. For VLCS ResNet-50, the positive decomposition terms sum to a selection benefit that remains insufficient to offset the negative fixed adaptation gain. These descriptive contrasts establish neither causal attribution nor component-wise significance.

Leave-one-cell-out recomputation preserves all X_{xfit} and G_{total} point-estimate signs. Maximum absolute changes are $0.85\,\mathrm{pp}$ in Office-Home ResNet-18 selection premium and $2.31\,\mathrm{pp}$ in VLCS ERM++ total gain, which remains negative. This does not establish interval-decision stability.

Figure 3(a) illustrates policy disagreements: H alone is correct in the first three examples, while neither policy is correct in the VLCS example. These cases illustrate failure modes rather than estimate their prevalence. Panel (b) shows the selection premiums and component-projected intervals. Every ResNet-50 baseline has a positive selection interval, ResNet-18 is mixed, and all ERM++ intervals contain zero. The VLCS ResNet-50 result must also be read with its negative total-gain interval.

4.3 Ablations and Sensitivity Analysis

Figure 2(a) provides a directional holdout diagnostic. At the achieved median-noise effect of $0.247\,\mathrm{pp}$, CF-DGSA has a positive-claim rate of 0.9303, compared with 0.8104 for one-way holdout. At the exact null, the corresponding rates are 0 and 0.0008, whereas the same-test oracle is positive in all 10,000 trials. The oracle's higher frequency is therefore not uniformly superior valid power. These compare the complete recorded rules, not an isolated fold-reversal effect or methods at matched type-I error.

All ERM++ assessment patterns and the four positive selection-premium configurations persist across routing margins. Lower margins route more near-tie folds to A without changing the comparator. Office-Home ERM++ never selects A, even at zero margin, so its degeneracy is not induced by the routing preference. Office-Home ResNet-18 changes from 1.69 to $1.68\,\mathrm{pp}$ as the margin increases to $0.5\,\mathrm{pp}$, with positive intervals throughout. This does not establish sensitivity to other support or audit budgets.

Gaussian amplitude calibration leaves the rejection threshold unchanged, so the curves evaluate one decision rule across approximately matched effects. This does not establish uniformly superior power over independently calibrated alternatives.

5 Conclusion

Whether a domain-generalized model should specialize requires separate comparisons with zero-shot inference and the best fixed adaptation policy. CF-DGSA separates these through fixed gain, cross-fitted selection premium, and total gain, using a panel-oracle comparator independent of the routing margin. Symmetric cross-fitting separates selection from evaluation, while Gaussian and

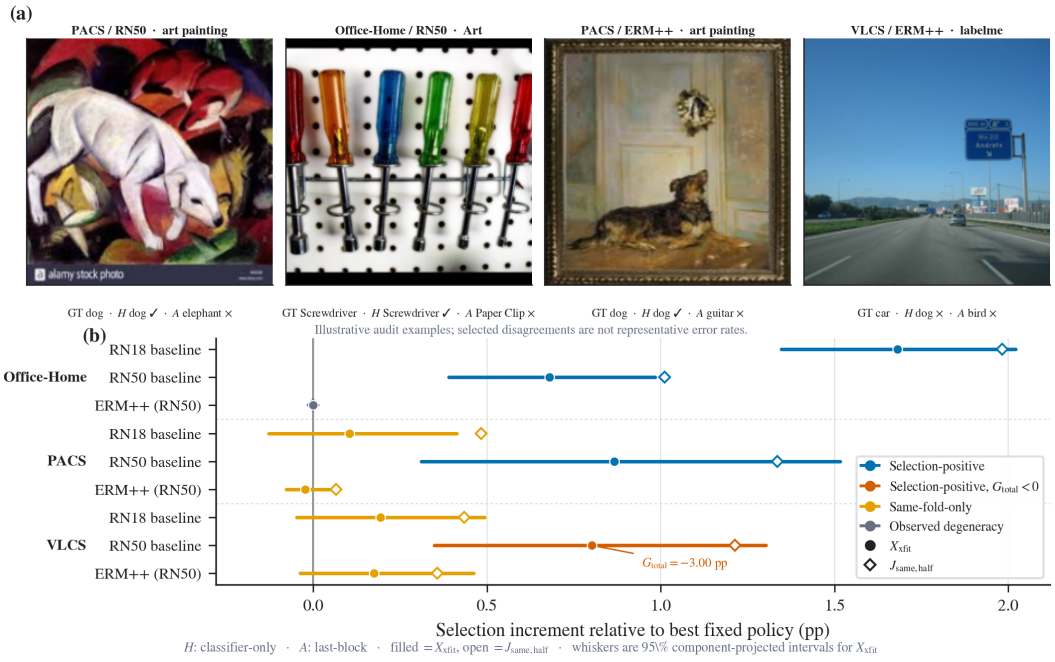

Figure 3: Policy disagreements and configuration-wise selection value. (a) Illustrative examples with ground-truth and H/A predictions. (b) Selection premiums with nominal 95% component-projected intervals; the reference line marks zero. Examples are illustrative, and total-gain conclusions follow Table 1.

paired-image studies characterize the interval procedure under specified noise and dependence structures.

The empirical findings distinguish the three alternatives introduced at the outset. Reproduced ERM++ models support fixed head-only adaptation on PACS without resolved selection value, exhibit observed policy degeneracy on Office-Home, and favor zero-shot inference on VLCS. Four selection-premium intervals are positive, but only three corresponding total-gain intervals are positive. The VLCS ResNet-50 result demonstrates that better selection among adaptation policies can still underperform zero-shot inference.

The findings concern two policies, ten support labels per class, additional audit labels, and the observed targets. Limited cell counts constrain uncertainty estimation, and component projection remains conditional on component-interval calibration. Cross-configuration trends are descriptive because architecture, resolution, and source training vary together. Priorities for further evaluation are BN-controlled comparisons, additional independent cells, and support-budget sensitivity with a fixed audit protocol. Extension to vision–language models and broader distribution shifts requires matched target-information access. Within this scope, fixed adaptation should be justified by improvement over zero-shot inference, and episode-specific selection should additionally be justified by improvement over the best fixed adaptation policy.

REFERENCES

Christina Baek, Yiding Jiang, Aditi Raghunathan, and J. Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. In *NeurIPS*, 2022.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *JMLR*, 11:2079–2107, 2010.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Zhixiang Chi, Li Gu, Huan Liu, Ziqiang Wang, Yanan Wu, Yang Wang, and Konstantinos Plataniotis. Learning to adapt frozen clip for few-shot test-time domain adaptation. In *International Conference on Learning Representations*, volume 2025, pp. 66359–66380, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1993.

Chen Fang, Ye Xu, and Daniel N. Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *ICCV Workshops*, 2013.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *ICLR*, 2021.

Qingyang Hao, Wenbo Liao, Bingyi Jing, and Hongxin Wei. Multi-condition conformal selection. In *ICLR*, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on Machine Learning*, pp. 448–456, 2015.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017.

Marius Lindauer, Jan N. van Rijn, and Lars Kotthoff. The algorithm selection competitions 2015 and 2017. *Artificial Intelligence*, 272:86–100, 2019. doi: 10.1016/j.artint.2018.10.004.

Alexander R. Luedtke and Mark J. van der Laan. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *The Annals of Statistics*, 44(2):713–742, 2016. doi: 10.1214/15-AOS1384.

John P. Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization. In *ICML*, pp. 7721–7735, 2021.

Aayush Mishra and Anqi Liu. ODD: Overlap-aware estimation of model performance under distribution shift. In *UAI*, pp. 3031–3047, 2025.

John R. Rice. The algorithm selection problem. In *Advances in Computers*, volume 15, pp. 65–118. Academic Press, 1976. doi: 10.1016/S0065-2458(08)60520-3.

Piotr Teterwak, Kuniaki Saito, Theodoros Tsiligkaridis, Kate Saenko, and Bryan Plummer. ERM++: An improved baseline for domain generalization. In *WACV*, pp. 8514–8524, 2025.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, 2017.

Changsong Wen, Zelin Peng, Yu Huang, Xiaokang Yang, and Wei Shen. Domain generalization in clip via learning with diverse text prompts. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9559–9569. IEEE, 2025.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Han Yu, Xingxuan Zhang, Renzhe Xu, Jiashuo Liu, Yue He, and Peng Cui. Rethinking the evaluation protocol of domain generalization. In *CVPR*, pp. 21897–21908, 2024.

AI USE STATEMENT

Generative AI tools were used to assist with the translation of papers. The authors designed the study, executed the experiments, and verified the implementations, statistical analyses, numerical results, and references. The authors retain full responsibility for the submission.