CAUSALGUARD: COVERAGE-AWARE COUNTERFACTUAL REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Counterfactual predictive consistency is informative only when interventions preserve task semantics, explore relevant nuisance variation, and reach the deployed prediction pathway. We introduce CausalGuard, which combines an input-dependent feature partition, counterfactual re-encoding, and compatibility-constrained donor weighting. Recombined features pass through the same mask and classifier as factual features, avoiding consistency that holds merely because a perturbed branch is discarded. A constrained policy balances prediction difficulty and directional diversity, while a frozen audit separates acceptance, label preservation, and prediction disagreement. We derive a conditional risk-transfer bound that exposes source reweighting, semantic mismatch, and unsupported target variation alongside prediction instability. Under source-only selection, the reported five-run results are $90.3 \pm 0.5\%$ on PACS, $81.2 \pm 0.4\%$ on VLCS, and $75.6 \pm 0.5\%$ on Office-Home. The controlled intervention audit reports $96.8 \pm 0.7\%$ label preservation and $3.4 \pm 0.5\%$ disagreement. These results motivate evaluating intervention validity and coverage jointly; they do not establish latent causal identification or unrestricted out-of-distribution robustness.

1 INTRODUCTION

A model can appear invariant for the wrong reason. Suppose its prediction is $g(c)$, while an augmentation modifies only a discarded branch s. Then $g(c)$ remains unchanged for every intervention on s, even if c contains the entire shortcut. A zero consistency penalty therefore need not establish robustness. This distinction matters in domain generalization (DG), where no target examples are available for fitting or model selection and source correlations alone may not identify the stable predictive mechanism.

CausalGuard asks when feature-space interventions make predictive stability informative. Its first component is an input-dependent mask that produces retained and replaceable features. The second is a shared re-encoding path: after recombination with a donor, the complete candidate feature is masked again and classified by the deployed predictor. Candidate changes can consequently affect the same computation used at inference. Re-encoding removes one architectural source of vacuity; it does not guarantee that optimization learns the correct factors.

Two further problems remain. A semantically conservative donor pool may contain only near-identity transformations, while a diverse pool may corrupt the class. CausalGuard therefore filters candidate pairs with a frozen semantic teacher and optimizes donor weights for both prediction difficulty and directional diversity under concentration constraints. A separate frozen audit measures label preservation, acceptance, disagreement, and donor diversity. These measurements distinguish stability under substantial valid interventions from stability obtained through rejection or policy collapse.

Our contributions are: (i) a shared re-encoding construction with an explicit bypass negative control; (ii) compatibility-constrained donor selection that exposes the validity–coverage trade-off; and (iii) a conditional risk-transfer analysis and component-wise evaluation. The scientific claim concerns their interaction, not the novelty of masking, consistency, or standard probability inequalities individually. We report gains against the evaluated controls on three visual DG benchmarks and retain a narrow interpretation: the learned partition is operational, and robustness extends only to the tested shifts and the assumptions stated in the analysis.

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.

2 Related work

Generalization and evaluation. IRM seeks invariant optimal prediction across environments (Arjovsky et al., 2019); its limitations illustrate why invariance requires informative variation (Rosenfeld et al., 2021). Risk extrapolation controls variation in environment risks (Krueger et al., 2021), while GroupDRO optimizes worst-group risk (Sagawa et al., 2020). DomainBed establishes the importance of consistent selection and tuning (Gulrajani & Lopez-Paz, 2021); SWAD uses weight averaging to seek flat minima (Cha et al., 2021). MIRO regularizes representations with a pre-trained reference (Cha et al., 2022). These approaches motivate strong matched controls rather than attributing every gain to causal structure.

Pair construction and causal assumptions. MixStyle mixes feature statistics (Zhou et al., 2021), causal matching uses object-level correspondences (Mahajan et al., 2021), and causal balancing modifies sampling (Wang et al., 2023). Causally inspired regularization can exploit graph-implied restrictions (Salaudeen & Koyejo, 2024). CausalGuard instead examines endogenous feature recombination, a shared re-encoding path, and constrained donor weights. Our evaluated baselines are not an exhaustive comparison to this literature; direct matched comparisons with causal matching and MIRO remain important for assessing the scope of the contribution.

Identification and recent developments. Causal representation learning aims beyond predictive transfer (Schölkopf et al., 2021). Unsupervised disentanglement requires inductive assumptions (Locatello et al., 2019), and real-system sanity checks challenge interpretations based only on synthetic success (Gamella et al., 2025). Recent identification results address general environments under nonparametric mixing (Ng et al., 2025) and intervention-dependent causal abstractions (Li et al., 2025); time-aware representations address evolving domains (He et al., 2025). These are distinct information and structural regimes from pooled-source feature recombination. Intervention-induced representation divergence (Grant et al., 2026) motivates checking off-manifold candidates. CARL explicitly targets cross-modal causal-structure preservation (Li et al., 2026); CausalGuard does not claim that its mask preserves a latent graph.

3 CausalGuard

3.1 Problem and shared prediction path

Let P be the pooled source distribution over features and labels, and Q an unseen target distribution with the same C classes. A shared backbone F_θ maps an image to $u \in \mathbb{R}^d$. Target examples are excluded from fitting and selection; source-domain identities define benchmark splits but are not inputs to the primary CausalGuard objective. A motivating structural model is

$$C_0 = f_c(U_c), \quad Y = f_y(C_0, U_y), \quad S = f_s(C_0, E, U_s), \quad X = f_x(C_0, S, U_x). \tag{1}$$

Nuisance interventions alter S while retaining content and label-generating noise. The algorithm constructs candidate feature interventions, not guaranteed structural counterfactuals. Acceptance by a semantic filter is operational evidence, not causal identification.

Define an input-dependent partition and a probability-floor predictor:

$$m_\phi(u) = \sigma(a_\phi(u)/T_m), \quad c_\phi(u) = m_\phi(u) \odot u, \quad s_\phi(u) = (1 - m_\phi(u)) \odot u, \tag{2}$$

$$p(u) = (1 - C\tau)\,\mathrm{softmax}(h_\psi(c_\phi(u))) + \tau \mathbf{1}, \qquad 0 < \tau < 1/C. \tag{3}$$

For an anchor i and donor j, form $v_{ij} = c_\phi(u_i) + s_\phi(u_j)$ and evaluate $p(v_{ij})$. Crucially, this applies m_ϕ again to the complete candidate, as shown in Figure 1. It is different from applying the classifier directly to the unchanged $c_\phi(u_i)$. Self-recombination satisfies $v_{ii} = u_i$ and is an implementation check; self-pairs are excluded from intervention training.

For a constant binary mask, interventions confined to discarded coordinates can still be vacuous after re-encoding. More generally, $J_c(u) = \mathrm{diag}(m(u)) + \mathrm{diag}(u) J_m(u)$ explains how an input-dependent mask allows candidate changes to reach retained features. This is a computational possibility, not a guarantee of semantic recovery (Appendix B).

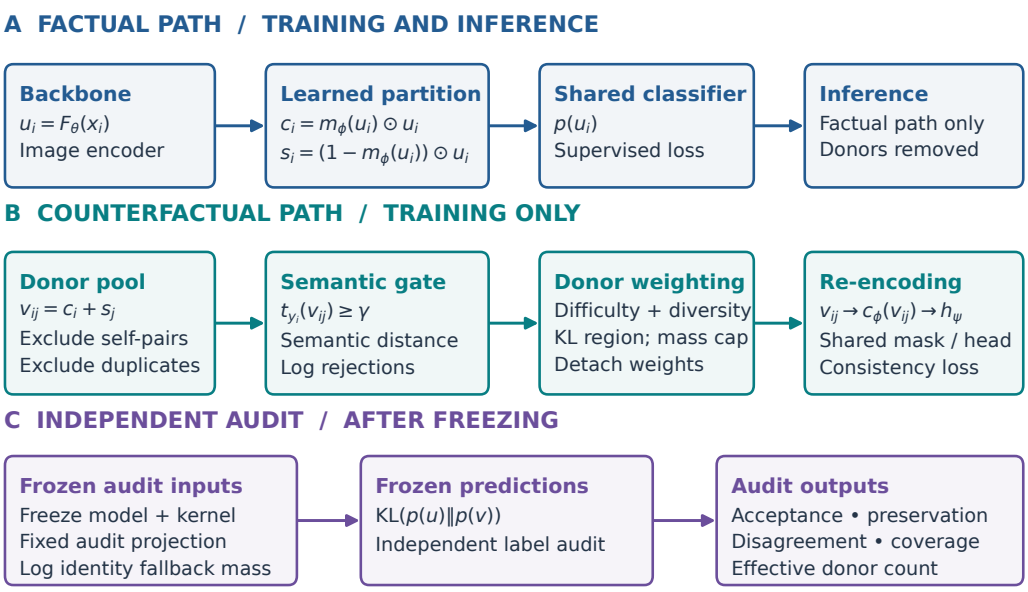


Figure 1: CausalGuard architecture. Factual prediction, gated counterfactual training, and independent auditing after freezing are separated. Donor weights are detached during the student update; inference uses the factual path.

3.2 Compatibility-constrained donor selection

Source-only donor pools exclude self-pairs and known duplicate captures. Same-class sampling reduces obvious conflicts but does not ensure instance-level validity. A frozen source-trained semantic teacher t, with feature map b, accepts a candidate when

$$a_{ij} = \mathbf{1}\{t_{y_i}(v_{ij}) \geq \gamma\}\, \mathbf{1}\{\|b(v_{ij}) - b(u_i)\|_2 \leq r_b\}. \tag{4}$$

Both functions act on backbone features. The compatibility teacher differs from the moving-average prediction reference used for optimization. Teacher shortcuts can cause false acceptance, so controlled interventions with known labels provide a separate validity audit.

Let q_i^0 be uniform over accepted donors, $\Delta_{ij} = s_\phi(u_j) - s_\phi(u_i)$, and $z_{ij} = V^\top \Delta_{ij}$ for a source-derived projection $V \in \mathbb{R}^{d \times r}$ fixed during policy optimization. Define $G_i(q) = \sum_j q_{ij} z_{ij} z_{ij}^\top$. We solve approximately

$$\begin{aligned} \max_{q_i} \quad & \sum_j q_{ij} d_{ij} + \beta \log\det(G_i(q) + \epsilon I), \\ \text{subject to} \quad & \sum_j q_{ij} = 1, \quad 0 \leq q_{ij} \leq q_{\max}, \quad q_{ij} = 0 \text{ if } a_{ij} = 0, \\ & D_{\mathrm{KL}}(q_i \| q_i^0) \leq \rho_q. \end{aligned} \tag{5}$$

Here d_{ij} is detached prediction divergence. The KL trust region limits departure from the accepted-pool reference, and the cap limits donor concentration. Log determinant rewards directional spread, not semantic validity or coverage of every target mode. Source-fitted feature scaling and a fixed projection prevent arbitrary rescaling from masquerading as improved coverage.

Feasibility requires $k_i q_{\max} \geq 1$, where k_i is the number of accepted donors. An infeasible pool cannot be repaired merely by more optimization steps; it must be recorded and handled explicitly. Empty pools contribute supervised loss only. In the theoretical complete kernel, an identity fallback can assign the missing mass, but its zero divergence is reported separately from nonidentity intervention stability. The fraction of eligible anchors and effective donor count $1/\sum_j q_{ij}^2$ are therefore reported with consistency.

Table 1: Alternating training computation. The executed run configuration specifies solver tolerances, schedules, and enabled loss terms.

Step	Operation
1	Encode source anchors and eligible same-class donors; compute factual supervised loss for every anchor.
2	Construct $v_{ij} = c_i + s_j$ and evaluate the frozen teacher's compatibility rule. Exclude duplicate captures and self-pairs.
3	Check accepted-pool feasibility under the donor cap. Record excluded anchors separately from accepted intervention statistics.
4	Freeze candidate statistics and update donor weights under the simplex, cap, and KL constraints.
5	Recompute gradient-bearing candidate predictions through the shared mask and classifier; detach policy weights and reference predictions.
6	Update the student with Eq. (8).
7	Log acceptance, effective donor count, displacement norms, and constraint residuals; retain all factual examples in the supervised update.

3.3 Optimization and deployment

For anchors $\mathcal{I}_+$ with a feasible accepted pool, use

$$\mathcal{L}_{\mathrm{cf}} = \frac{1}{|\mathcal{I}_+|} \sum_{i \in \mathcal{I}_+} \sum_j q_{ij} D_{\mathrm{KL}}\big(\mathrm{sg}(\bar{p}(u_i)) \| p(v_{ij})\big). \tag{6}$$

$\bar{p}$ is a detached prediction reference. Set this term to zero if $\mathcal{I}_+$ is empty. Supervised factual loss uses all anchors. An accepted-candidate label loss, when enabled, uses the anchor label. Mask-budget and binarization terms are

$$\mathcal{L}_b = \big(\mathbb{E}_i[d^{-1} \textstyle\sum_k m_k(u_i)] - b_0\big)^2, \qquad \mathcal{L}_h = \mathbb{E}_i[d^{-1} \textstyle\sum_k m_k(u_i)(1 - m_k(u_i))], \tag{7}$$

$$\mathcal{L} = \mathcal{L}_{\mathrm{sup}} + \lambda_{\mathrm{cf}} \mathcal{L}_{\mathrm{cf}} + \lambda_y \mathcal{L}_{\mathrm{cf-label}} + \lambda_b \mathcal{L}_b + \lambda_h \mathcal{L}_h. \tag{8}$$

The notation b_0 distinguishes the scalar mask budget from semantic map b. Neither mask regularizer establishes disentanglement. Feature reconstruction alone is also uninformative because $c(u) + s(u) = u$.

Each update encodes anchors and donors, constructs candidates, applies the frozen compatibility gate, updates detached policy weights, and re-encodes accepted candidates before a student step. A warm start and source-selected consistency ramp reduce rejection from initially random predictions. Candidate forwards must use the same normalization convention as factual forwards. At inference, donor construction, the semantic teacher, and the policy solver are removed; the backbone, input-dependent mask, and classifier remain. Their latency is not identically that of ERM.

The update sequence is summarized in Table 1. Candidate-policy optimization and representation learning use distinct gradient paths. In particular, differentiating through the donor solver would change the stated alternating objective.

4 Coverage-qualified risk transfer

Fix the trained feature extractor, predictor, and candidate kernel before analysis. Let $\mu(du, dy, dv) = P(du, dy)q(dv \mid u, y)$. Let a nonnegative weight w satisfy $\mathbb{E}_\mu w = 1$ and $w \le \kappa$. The reweighted pair law represents target-relevant candidate interventions. Its factual marginal is P_w, and its candidate marginal with the retained label is Q_w. Suppose $\mathrm{TV}(P_w, P) \le \rho$. A corrected candidate law Q_w^* can relabel semantically invalid candidates, with a coupling under which the probability of label disagreement is at most ξ. Finally, assume $\mathrm{TV}(Q, Q_w^*) \le \eta$.

These quantities are assumptions or audit targets, not automatically observed by the algorithm. In particular, directional diversity does not estimate η for an unknown target. For a bounded loss $0 \le \ell(p, y) \le B$ satisfying

$$|\ell(p, y) - \ell(p', y)| \le L\, \mathrm{TV}(p, p'), \tag{9}$$

define $\varepsilon = \mathbb{E}_\mu D_{\mathrm{KL}}(p(u) \| p(v))$.

Theorem 1 (Conditional transfer bound)**.** *Under the preceding assumptions,*

$$R_Q(p) \le R_P(p) + L\sqrt{\kappa\varepsilon/2} + B(\rho + \xi + \eta). \tag{10}$$

Proof. Pinsker's inequality and Jensen's inequality give $\mathbb{E}_{w\mu}\,\mathrm{TV}(p(u), p(v)) \le \sqrt{\mathbb{E}_\mu[wD_{\mathrm{KL}}]/2} \le \sqrt{\kappa\varepsilon/2}$. Loss Lipschitzness transfers risk from candidates to factual examples under $w\mu$. The difference between factual expectations under P_w and P is at most $B\rho$. Relabeling contributes at most $B\xi$, and transferring from Q^*_w to Q contributes at most $B\eta$. Summing these inequalities proves the result. □

The bound explains four distinct failure modes. Small consistency loss is insufficient if target interventions have poor support, requiring large κ or η; semantic violations contribute ξ; changes in content composition contribute ρ. The expression is intentionally conditional and can be vacuous. It is not a distribution-free guarantee for natural-image targets.

For $\ell(p, y) = 1 - p_y$, take $B = L = 1$. Smoothed cross-entropy has $B = \log(1/\tau)$ and $L = 1/\tau$. Neither choice directly yields the same statement for discontinuous argmax accuracy. A margin-based conversion is provided in Appendix F. Independent finite-sample auditing and the support limitations of zero KL are developed in Appendices D–E.

The result decomposes uncertainty rather than eliminating it. Full-rank candidate covariance does not establish a density-ratio bound or identify the target discrepancy η. The complete proof is in Appendix C; Appendix G explains the distinction between geometric span and distributional support. The theorem also uses same-model frozen divergence, not the adaptive moving-average training objective.

4.1 Independent finite-sample audit

Training consistency is adaptively optimized, so it cannot be inserted directly into a fixed-model concentration bound. Freeze a portfolio of M predictors and their intervention kernels before drawing n independent source anchors. Draw a candidate independently from each frozen kernel conditional on its anchor. Define the empirical factual risk $\widehat{R}$ and same-model divergence $\widehat{\varepsilon}$ as anchor averages. With the probability floor in Eq. (3), divergence is bounded by $K = \log(1/\tau)$.

For $a_n = \sqrt{\log(2M/\delta)/(2n)}$, Hoeffding's inequality and a union bound over two statistics and M models give, simultaneously with probability at least $1 - \delta$,

$$R_Q \le \widehat{R} + Ba_n + L\sqrt{\frac{\kappa}{2}(\widehat{\varepsilon} + Ka_n)} + B(\rho + \xi + \eta). \tag{11}$$

The independence condition concerns audit anchors, not the number of candidate pairs. Multiple candidates per anchor may be averaged, but they do not multiply the number of independent observations. A shared donor bank must be fixed independently, with the argument conditioned on that bank. Models may share audit anchors because independence across models is unnecessary for a union bound.

The portfolio may be selected using this simultaneous bound, but adding models adaptively after inspecting the audit requires a new correction or fresh data. Moreover, concentration reduces sampling uncertainty only: it does not identify κ, ρ, ξ, or η. A narrow interval on observed KL cannot certify robustness to unsupported shifts. Appendix D provides the proof and semantic-audit qualifications.

4.2 What directional coverage can constrain

A local calculation clarifies the role of diversity. Let f denote logits and approximate candidate changes by $f(u + \Delta) - f(u) \approx J_f(u)\Delta$. Conditional on an anchor, define $G_u = \mathbb{E}[\Delta\Delta^\top \mid u]$. Then the expected squared linearized logit change is

$$\mathbb{E}[\|J_f(u)\Delta\|_2^2 \mid u] = \mathrm{tr}(J_f(u)G_uJ_f(u)^\top). \tag{12}$$

If $G_u \succeq \lambda\Pi$ for an orthogonal projector Π, this quantity is at least $\lambda\|J_f(u)\Pi\|_F^2$. Thus, for linear logits, small squared change combined with a positive coverage eigenvalue constrains sensitivity

Table 2: Held-out-domain accuracy (%, mean ± SD over five runs). Higher accuracy is better.

Method	PACS	VLCS	Office-Home
ERM	86.2±0.7	78.5±0.6	71.8±0.8
IRM	85.7±0.9	78.9±0.7	71.4±0.9
GroupDRO	86.5±0.8	79.1±0.6	72.0±0.8
SWAD	88.1±0.6	79.6±0.5	73.8±0.7
Mask-only	87.0±0.7	79.0±0.6	72.6±0.8
Gaussian feature perturbation	87.3±0.7	79.2±0.6	72.9±0.7
Uniform accepted donors	88.4±0.6	80.0±0.5	74.1±0.6
Filtered, no diversity	88.8±0.5	80.3±0.5	74.5±0.6
Unfiltered diverse donors	87.9±0.7	79.7±0.6	73.6±0.7
CausalGuard	**90.3±0.5**	**81.2±0.4**	**75.6±0.5**

in the explored subspace. A singular G_u leaves null directions unconstrained, regardless of total displacement energy. For nonlinear logits, a finite-step statement additionally requires control of the Taylor remainder; predictive KL is not identical to squared logit change.

This distinction explains why the policy uses a diversity term but the transfer theorem retains an explicit support discrepancy. Covariance describes second-order geometry, while the theorem requires a relationship between full probability laws. Even isotropic candidate covariance can miss disconnected target modes, and large displacements can be semantically invalid. Appendix K.3 visualizes equal-energy spectra and the exact feasibility boundary of the donor policy. These are analytical illustrations, not additional benchmark experiments.

5 Experiments

5.1 Protocol and comparators

We evaluate PACS (Li et al., 2017), VLCS (Fang et al., 2013), and Office-Home (Venkateswara et al., 2017) using held-out-domain evaluation and source-only selection following DomainBed (Gulrajani & Lopez-Paz, 2021). The experimental protocol uses an ImageNet-initialized ResNet-50 (He et al., 2016), 224×224 inputs, and five independently initialized runs. The source validation rule and pretraining checkpoint are shared across comparators. CausalGuard does not consume source-domain labels in its training objective; IRM and GroupDRO use their required group/environment information.

ERM and SWAD provide primary reference points. Mask-only, Gaussian feature perturbation, uniform accepted donors, filtering without diversity, and unfiltered diverse donors test alternative explanations for gains. Comparisons are interpreted within the stated training and selection protocol; additional per-epoch work means that matching epochs alone does not establish equal wall-clock cost. WILDS (Koh et al., 2021) is a relevant future extension, not a completed evaluation reported here.

5.2 Benchmark performance

Table 2 reports mean accuracy and standard deviation. Figure 2 displays the same values for the main reference methods. CausalGuard improves over ERM by 4.1, 2.7, and 3.8 percentage points on PACS, VLCS, and Office-Home, respectively; the corresponding gains over SWAD are 2.2, 1.6, and 1.8 points. These are differences between reported means, not paired confidence intervals.

The benchmark means alone do not establish performance on every held-out domain, calibration, or statistical significance. Five runs characterize optimization variability under the experimental setup, not uncertainty over all possible deployment domains. Paired inference additionally requires matched seed-level baseline outcomes. We therefore distinguish descriptive gains from inferential claims and avoid deriving paired error bars from marginal standard deviations.

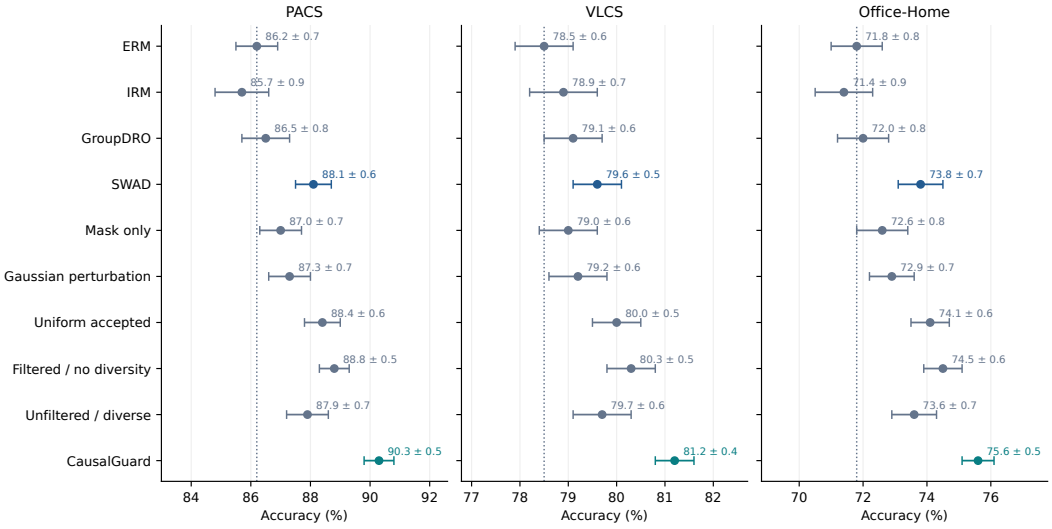


Figure 2: Benchmark comparison for all ten methods reported in the manuscript. Points and error bars show mean ± standard deviation over five independent runs; dashed vertical lines identify the ERM means.

Table 3: Controlled intervention diagnostics (mean ± SD where supplied). Acceptance, preservation, and disagreement are percentages; donor count is dimensionless.

Variant	Accept	Preserve	Disagree ↓	Eff. donors
Full CausalGuard	74.6±2.1	96.8±0.7	3.4±0.5	6.8±0.4
No re-encoding	75.1±2.3	96.5±0.8	5.9±0.7	6.7±0.5
Uniform accepted donors	74.9±2.0	96.7±0.7	4.8±0.6	4.9±0.3
No compatibility filter	100.0	88.4±1.4	8.7±0.9	7.3±0.5
No diversity term	75.0±2.2	96.9±0.6	4.6±0.5	3.6±0.3
Fixed random mask	68.7±2.8	93.1±1.1	7.5±0.8	5.2±0.4
No counterfactual loss	73.8±2.4	96.4±0.8	10.6±1.0	6.5±0.4

5.3 Mechanism diagnostics

Table 3 isolates re-encoding, compatibility, diversity, partition learning, and counterfactual regularization. Removing the compatibility filter increases acceptance to 100% while reducing preservation to 88.4% and increasing disagreement to 8.7%. Removing diversity lowers effective donor count from 6.8 to 3.6 while preserving similar semantic validity. Removing counterfactual loss increases disagreement to 10.6%. These diagnostics support different roles for validity, donor spread, and predictive regularization; they do not prove recovery of true nuisance variables.

5.4 Validity–coverage trade-off

Figure 3 plots the threshold sweep (Appendix I). The reported operating point $\gamma = 0.60$ accepts 74.6% of candidates with 96.8% preservation and normalized coverage 0.78. Raising γ to 0.90 increases preservation to 98.5% but lowers acceptance to 29.5% and coverage to 0.33. Thus, improving conditional validity through rejection can remove much of the intervention support. The sweep is descriptive; target accuracy is not used to select its operating point.

Audit measurements depend on their denominator. Acceptance must be reported over all eligible anchors or pairs, preservation over accepted interventions with independent semantic labels, and disagreement over a clearly specified intervention subset. A teacher's own acceptance decision cannot also serve as independent evidence that the accepted label is correct. Controlled positive and negative interventions (Appendix I.3) provide a stronger diagnostic than natural-image feature similarity alone.

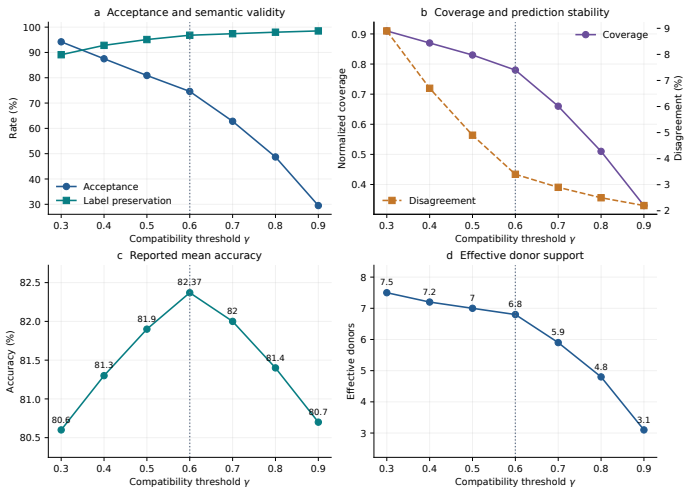


Figure 3: Threshold sweep from Table 5: acceptance and preservation, coverage and disagreement, mean accuracy, and effective donor count. Lines connect the seven reported settings; no additional observations or uncertainty are introduced.

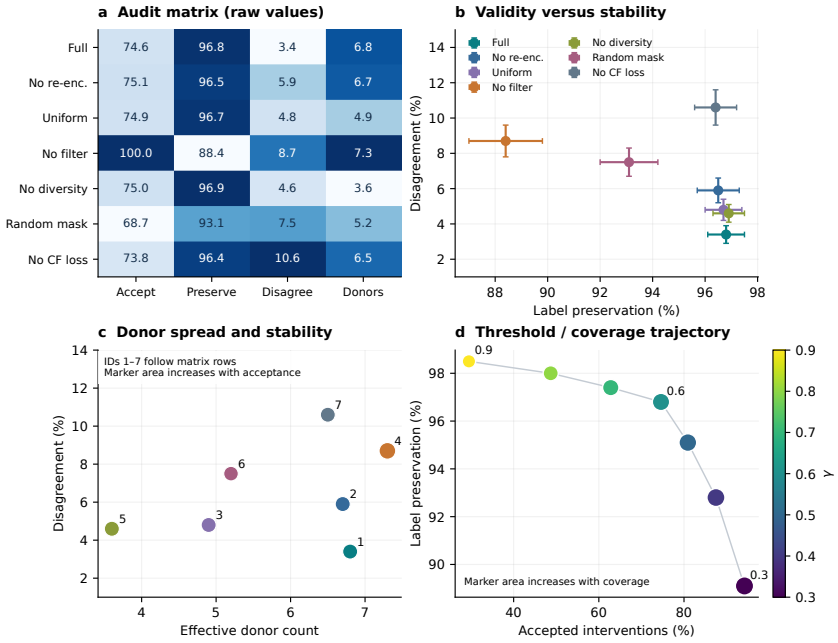


Figure 4: Mechanism diagnostics from the reported ablation and compatibility-threshold experiments. Matrix annotations retain raw values while colors are scaled within columns. Error bars show reported marginal standard deviations, not joint confidence regions; marker areas encode the quantities indicated in each panel.

5.5 Joint interpretation of intervention diagnostics

Figure 4 combines the ablation diagnostics with the threshold trajectory. The views answer different questions. The matrix exposes which component changes which measurement; the validity–stability plane distinguishes semantic corruption from predictive instability; the donor-count view tests whether broad donor support alone is sufficient; and the threshold trajectory shows the cost of obtaining higher conditional validity by rejecting more interventions. No single panel determines the quality of an intervention policy.

For example, removing the counterfactual loss leaves preservation near the full model's level but increases disagreement substantially. This separates the existence of semantically admissible transformations from the predictor's response to them. Conversely, the no-filter variant exposes many donors but also changes task semantics more often. Its high acceptance cannot be interpreted as better supported invariance. The no-diversity control retains strong preservation while concentrating on fewer effective donors; whether those missing directions matter depends on the target family rather than donor count alone.

The plotted uncertainty comprises marginal SDs, not confidence ellipses or paired significance tests. The seven-setting threshold trajectory is descriptive, not a continuously measured surface. Appendix K expands the diagnostics and specifies every normalization and derived quantity.

Graphical methodology view. Figure 5 specifies the training computation and the independent frozen audit. It makes candidate feature flow, detached donor weights, shared re-encoding, and audit prerequisites explicit. Appendix K.4 distinguishes this method schematic from the records required to document an executed implementation.

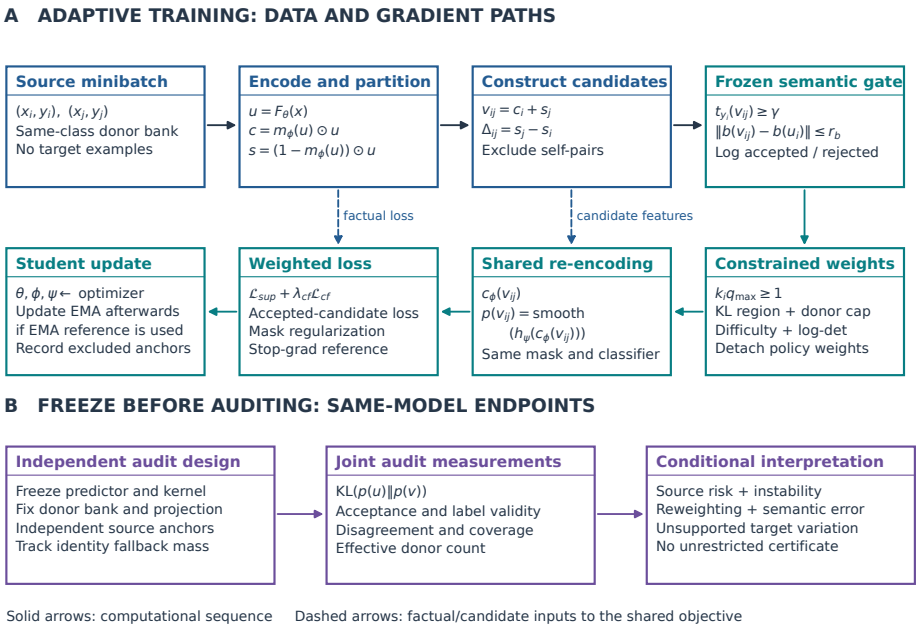


Figure 5: Training and frozen-audit computation. Solid arrows show the computational sequence; dashed arrows supply factual and candidate inputs. Donor weights are detached before the student update.

6 Limitations and Conclusion

An operational mask is coordinate-dependent and can carry information through its gating function. Feature recombination can leave the backbone manifold, and a frozen teacher can share the student's shortcuts. The theorem remains conditional on semantic mismatch, reweighting, and target-support discrepancy; these quantities are not identified by high directional diversity. CausalGuard does not address unseen classes, arbitrary label-mechanism changes, or target interventions outside candidate support. Its current comparative evidence also does not establish superiority over every recent pair-based or pretrained-model method.

CausalGuard combines shared re-encoding, compatibility-constrained donor diversity, and frozen auditing. The results motivate evaluating stability jointly with intervention quality; low consistency alone cannot certify unrestricted transfer.

AI USE STATEMENT

Generative AI assistance was used in developing and refining the conceptual framework, methodological exposition, mathematical claims and proofs, experimental design, and interpretation; it was also used to reorganize and edit the manuscript, check internal consistency and references, and produce vector diagrams and plotting code from author-supplied tabular values. It was not used to execute the reported training experiments or generate additional benchmark observations. The authors are responsible for verifying the scientific claims, numerical provenance, and final submission.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant Risk Minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. SWAD: Domain Generalization by Seeking Flat Minima. In *Advances in Neural Information Processing Systems*, 2021.

Junbum Cha, Kyungjae Lee, Sungrae Park, and Sanghyuk Chun. Domain Generalization by Mutual-Information Regularization with Pre-trained Models. In *European Conference on Computer Vision*, pp. 440–457, 2022. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/html/3660_ECCV_2022_paper.php.

Chen Fang, Ye Xu, and Daniel N. Rockmore. Unbiased Metric Learning: On the Utilization of Multiple Datasets and Web Images for Softening Bias. In *IEEE International Conference on Computer Vision*, pp. 1657–1664, 2013. doi: 10.1109/ICCV.2013.208.

Juan L. Gamella, Simon Bing, and Jakob Runge. Sanity Checking Causal Representation Learning on a Simple Real-World System. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 18143–18169. PMLR, 2025.

Satchel Grant, Simon Jerome Han, Alexa Tartaglini, and Christopher Potts. Addressing Divergent Representations from Causal Interventions on Neural Networks. In *International Conference on Learning Representations*, 2026.

Ishaan Gulrajani and David Lopez-Paz. In Search of Lost Domain Generalization. In *International Conference on Learning Representations*, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Zhuo He, Shuang Li, Wenze Song, Longhui Yuan, Jian Liang, Han Li, and Kun Gai. Learning Time-Aware Causal Representation for Model Generalization in Evolving Domains. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 22523–22547. PMLR, 2025.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard L. Phillips, Irena Gao, et al. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-Distribution Generalization via Risk Extrapolation (REx). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 5815–5826. PMLR, 2021. URL https://proceedings.mlr.press/v139/krueger21a.html.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, Broader and Artier Domain Generalization. In *IEEE International Conference on Computer Vision*, 2017.

Xiusi Li, Sékou-Oumar Kaba, and Siamak Ravanbakhsh. On the Identifiability of Causal Abstractions. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258, pp. 3241–3249. PMLR, 2025.

Yulong Li, Xiwei Liu, Barrett Tang, Zhixiang Lu, Ming Hu, Yichen Li, Haochen Xue, Peixin Guo, Jionglong Su, Yutong Xie, Eran Segal, and Imran Razzak. CARL: Preserving Causal Structure in Representation Learning. In *International Conference on Learning Representations*, 2026.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *International Conference on Machine Learning*, pp. 4114–4124. PMLR, 2019.

Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain Generalization Using Causal Matching. In *International Conference on Machine Learning*, pp. 7313–7324. PMLR, 2021.

Ignavier Ng, Shaoan Xie, Xinshuai Dong, Peter Spirtes, and Kun Zhang. Causal Representation Learning from General Environments under Nonparametric Mixing. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258, pp. 3700–3708. PMLR, 2025.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The Risks of Invariant Risk Minimization. In *International Conference on Learning Representations*, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. In *International Conference on Learning Representations*, 2020.

Olawale Salaudeen and Sanmi Koyejo. Causally Inspired Regularization Enables Domain General Representations. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238, pp. 3124–3132. PMLR, 2024.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward Causal Representation Learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep Hashing Network for Unsupervised Domain Adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Xinyi Wang, Michael Saxon, Jiachen Li, Hongyang Zhang, Kun Zhang, and William Yang Wang. Causal Balancing for Domain Generalization. In *International Conference on Learning Representations*, 2023.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain Generalization with MixStyle. In *International Conference on Learning Representations*, 2021.

A NOTATION AND INFORMATION ACCESS

Table 4 summarizes the notation used below.

Table 4: Notation used throughout the manuscript.

Symbol	Meaning
C_0, S, E	Latent content, nuisance, and environment variables in the motivating SCM
C, d, r	Number of classes, feature dimension, and audit projection dimension
P, Q	Source and target distributions over features and labels after fixing the backbone
u, v	Factual and counterfactual feature representations in $\mathbb{R}^d$
m, c, s	Learned mask, retained branch, and replaced branch
$p, \bar{p}, t$	Deployed predictor, optimization reference, and frozen compatibility teacher
q, q^0	Intervention kernel and reference donor distribution
μ, w, κ	Source-pair distribution, normalized reweighting function, and maximum density ratio
P_w, Q_w, Q_w^*	Weighted factual, candidate, and semantically corrected candidate distributions
ρ, ξ, η	Factual marginal shift, semantic mismatch, and target-support discrepancy
τ, B, L	Probability floor, loss bound, and probability-Lipschitz constant
G, V	Intervention second-moment matrix and fixed audit projection

The notation P denotes the feature-label distribution induced by pushing an image-label distribution through the backbone. Because the backbone evolves during optimization, the induced feature distribution also changes. The theoretical analysis is therefore applied to the fixed representation obtained after optimization and does not constitute a stepwise guarantee over the evolving feature geometry.

The optimization reference $\bar{p}$ is distinct from the compatibility teacher t. The former provides a stable reference for counterfactual consistency optimization, whereas the latter remains frozen and determines candidate compatibility. The independent audit evaluates the deployed predictor directly by comparing $p(u)$ and $p(v)$, thereby avoiding dependence on either teacher as the final prediction endpoint.

Information-access protocol. The primary CausalGuard model uses source images and class labels during training and does not use target-domain samples or source-domain labels in either the learning objective or donor-selection policy. Benchmark domain labels are used only to construct the prescribed held-out-domain evaluation splits. All compute-matched comparisons use the same pretrained backbone and corresponding external pretraining information.

Class labels alone do not define structural counterfactuals. Same-class observations may differ in pose, shape, subclass, background, or other task-relevant characteristics. Accordingly, the intervention mechanism is interpreted as label-conditioned feature recombination rather than direct manipulation of a known physical causal mechanism. Stronger structural interpretations require additional causal knowledge or intervention ground truth.

The reweighting function is normalized such that $\mathbb{E}_\mu[w] = 1$, ensuring that $w\mu$ defines a valid probability distribution. Total variation is defined as

$$\mathrm{TV}(P, Q) = \sup_A |P(A) - Q(A)|,$$

which equals one half of the L_1 distance when the corresponding densities exist.

Kernel normalization under rejection. For each anchor with a nonempty accepted donor set, the donor kernel is normalized conditionally over its accepted candidates. This anchor-wise normalization prevents samples with larger accepted pools from receiving disproportionate probability mass. For anchors without an accepted donor, the complete intervention kernel uses an identity fallback $v = u$, preserving a valid probability measure over the full source support.

For the finite-sample theorem, the audit evaluates this complete intervention kernel while separately reporting diagnostics over nonidentity interventions. This distinction exposes the contribution of fallback mass and prevents uninformative identity pairs from being interpreted as evidence of counterfactual robustness.

B WHY NAIVE BRANCH INVARIANCE CAN BE VACUOUS

Proposition 2 (Excluded-branch invariance)**.** *Let a predictor be $f(c, s) = g(c)$. For every perturbation kernel on s, $\mathbb{E}D_{\mathrm{KL}}(f(c, s)\|f(c, \tilde{s})) = 0$, regardless of whether c contains a shortcut.*

Proof. Both arguments of the divergence are identically $g(c)$ for every sampled pair. Hence each divergence and its expectation are zero. No property of the data-generating process has been used. □

For example, let source observations contain an object indicator a and a background indicator b, with $a = b = y$ on all training examples. A mask can retain only b. Perturbing a separately discarded variable leaves predictions unchanged, but when the target flips b the predictor fails. The consistency objective has provided no evidence for preferring the object indicator.

Re-encoding avoids this particular computational tautology. With $c(u) = m(u) \odot u$, the Jacobian is

$$J_c(u) = \mathrm{diag}(m(u)) + \mathrm{diag}(u)J_m(u). \tag{13}$$

For a nuisance perturbation Δ, a first-order expansion gives

$$p(u + \Delta) - p(u) = J_p(u)\Delta + o(\|\Delta\|). \tag{14}$$

Even if Δ lies in coordinates weakly retained at the anchor, the second term of J_c allows a changed mask to affect the retained representation. This makes the consistency loss nontrivial, but it does not guarantee that its gradients identify the correct factors.

Fixed-mask limiting case. If a binary mask is constant and donor perturbations occur entirely outside its retained coordinates, re-encoding also becomes invariant by construction. This limit is useful as a negative control. A zero loss in this configuration should not be interpreted as successful discovery; its quality depends entirely on which coordinates were retained before the perturbations were generated.

Implementation check. For an arbitrary nondegenerate initialized model, select a donor producing a nonzero candidate displacement and compare factual and candidate predictions. Verify that the candidate branch reaches the shared mask and classifier, and that the loss has gradients through the candidate computation. This check identifies wiring errors. It is not an empirical test of causal validity.

A further subtlety is mask leakage. The gate can encode information about the whole input through its weights, even if only a few output coordinates remain. Sparsity measures the number or magnitude of active coordinates, not the information content of the gating function. Probe-based and controlled-intervention diagnostics are therefore necessary even when masks look sharply binary.

A second collapse pathway. If the classifier outputs the same distribution for every input, every candidate divergence also vanishes. Supervised loss discourages this solution, but severe class imbalance or an excessive consistency coefficient can still make partial collapse attractive. Measure per-class accuracy, predictive entropy, and variation of logits across distinct content values. A low-rank logit representation is not itself a failure; the relevant question is whether task-relevant distinctions have been suppressed. This is why content-changing negative controls accompany nuisance-changing invariance tests.

C COMPLETE RISK-TRANSFER PROOF AND INTERPRETATION

Write $\nu = w\mu$, a probability law on (u, y, v). Define $R_{Q_w} = \mathbb{E}_\nu \ell(p(v), y)$ and $R_{P_w} = \mathbb{E}_\nu \ell(p(u), y)$. For losses in $[0, B]$, integration against the signed difference of two probability mea-

sures gives

$$|R_A - R_D| \le B\,\mathrm{TV}(A, D). \tag{15}$$

This form uses nonnegative bounded losses; using a bound on absolute value alone would give a different constant.

The candidate-versus-factual difference satisfies

$$\begin{aligned} R_{Q_w} - R_{P_w} &\le \mathbb{E}_\nu |\ell(p(v), y) - \ell(p(u), y)| && (16)\\ &\le L\mathbb{E}_\nu\,\mathrm{TV}(p(v), p(u)) && (17)\\ &\le L\mathbb{E}_\nu \sqrt{\tfrac{1}{2} D_{\mathrm{KL}}(p(u)\|p(v))} && (18)\\ &\le L\sqrt{\tfrac{1}{2}\mathbb{E}_\mu[w D_{\mathrm{KL}}(p(u)\|p(v))]} && (19)\\ &\le L\sqrt{\kappa\varepsilon/2}. && (20) \end{aligned}$$

The third line is Pinsker's inequality. The fourth follows from concavity of the square root under the normalized law ν. The final step uses the nonnegativity of KL and the pointwise bound on w.

Under a coupling of Q_w and Q_w^* that preserves the feature while changing the label with probability at most ξ,

$$R_{Q_w^*} - R_{Q_w} \le B\Pr(Y^* \ne Y) \le B\xi. \tag{21}$$

Combining this with Eq. (15) for (Q, Q_w^*) and (P_w, P) yields Theorem 1.

Why source reweighting appears. A target may contain the same valid intervention pairs but emphasize a different subset of source content. Even a predictor perfectly invariant to nuisance can have higher risk if difficult classes become more frequent. The ρ term captures this source-composition change. Omitting it would silently assume identical factual marginals after target reweighting.

What can be measured. The source risk and synthetic consistency are directly estimable. Semantic error can be measured on interventions with known labels or a separately annotated audit sample. Density-ratio and target-discrepancy terms require a specified target family or additional information; they are not identified from pooled source data alone. The theorem should therefore be used to structure experiments and failure analysis, not to report a numerical deployment certificate with unknown terms set to zero.

Tightness perspective. Every additive term can dominate in an appropriate example: concentrated target weights increase κ, class-changing swaps increase ξ, shifted class composition increases ρ, and unrepresented nuisance support increases η. A small training penalty addresses only one component.

Choosing a target coupling. The weight w operates on pairs, so different couplings can yield the same candidate target marginal with different factual marginals and density-ratio bounds. An application may optimize over a family of admissible couplings to obtain a sharper analytical statement, but choosing a coupling after observing target labels introduces additional estimation and selection questions. The present theorem assumes a coupling satisfying the stated properties; it does not provide a source-only algorithm for finding the best coupling for an unseen target.

In particular, setting $w = 1$ is a simplifying assumption that forces $P_w = P$ and $\kappa = 1$. It is justified only when the target-relevant intervention mixture agrees with the source candidate mixture, not merely because uniform donors are convenient computationally.

D Independent finite-sample auditing

Training loss is adaptively optimized and cannot be inserted directly into a concentration bound for a fixed predictor. Freeze a portfolio of M candidate predictors and their kernels before drawing an independent audit sample. Conditional on all training information, draw n independent anchors $(u_i, y_i) \sim P$ and one independent candidate $v_i \sim q(\cdot \mid u_i, y_i)$ per anchor for each frozen candidate model. Models may share anchors; independence across models is unnecessary for a union bound.

Let

$$\widehat{R} = \frac{1}{n}\sum_i \ell(p(u_i), y_i), \qquad \widehat{\varepsilon} = \frac{1}{n}\sum_i D_{\mathrm{KL}}(p(u_i)\|p(v_i)). \tag{22}$$

With the probability floor in Eq. (3), $p_y(v) \geq \tau$ and

$$0 \leq D_{\mathrm{KL}}(p(u)\|p(v)) \leq K := \log(1/\tau). \tag{23}$$

The upper bound follows because $\sum_y p_y \log p_y \leq 0$ and $-\sum_y p_y \log p_y(v) \leq \log(1/\tau)$.

Proposition 3 (Finite portfolio audit)**.** *With probability at least $1-\delta$, simultaneously for all M fixed candidate models,*

$$R_Q \leq \widehat{R} + Ba_n + L\sqrt{\frac{\kappa}{2}(\widehat{\varepsilon} + Ka_n)} + B(\rho + \xi + \eta), \quad a_n = \sqrt{\frac{\log(2M/\delta)}{2n}}. \tag{24}$$

Proof. Hoeffding's inequality gives an upper deviation of Ba_n for risk and Ka_n for KL, each failing with probability at most $\delta/(2M)$. Union over two statistics and M models, then substitute into Theorem 1. □

The portfolio may be selected using the audit bound because the guarantee is simultaneous over the predeclared finite set. It may not be expanded indefinitely in response to audit outcomes without another correction or fresh data. An audit used for hyperparameter tuning is no longer an untouched final audit for later modified models.

If several donors are evaluated per anchor, averaging their bounded losses still gives a bounded anchor-level statistic. Concentration can use the number of independent anchors, not the product of anchors and donors. Shared donor banks introduce dependence unless the bank is fixed independently and the argument is conditioned on it. Record these choices explicitly.

The result remains conditional on κ, ρ, ξ, η. Finite-sample precision for observed consistency does not resolve structural uncertainty about those quantities. Reporting a narrow confidence interval for KL alongside an unbounded support mismatch would be mathematically precise but scientifically incomplete.

Independent semantic estimates. If a separate sample of m weighted-validity judgments could be drawn from $w\mu$, Hoeffding's inequality would upper-bound ξ by its empirical mismatch rate plus $\sqrt{\log(1/\delta_\xi)/(2m)}$. In practice, w is often unknown, so an unweighted annotation sample supports a different estimand. Importance weighting requires known or estimated ratios and its own uncertainty analysis. These distinctions are essential when combining a measured consistency interval with a semantic-error estimate.

All confidence levels should be allocated before inspecting the audit. Adding bounds with individually stated 95% coverage does not automatically produce a joint 95% guarantee; a union-bound allocation or another simultaneous procedure is needed.

E SUPPORT, ZERO DIVERGENCE, AND NON-IDENTIFIABILITY

Lemma 4 (Zero expected divergence)**.** *If $\mathbb{E}_{Pq} D_{\mathrm{KL}}(p(u)\|p(v)) = 0$, then $p(u) = p(v)$ for Pq-almost every pair. This does not imply equality for all possible interventions.*

Proof. A nonnegative measurable function with zero expectation is zero almost everywhere. KL vanishes exactly when the two probability distributions agree. The statement is restricted to the measure under which the expectation is taken. □

Continuity can extend equality to the *support* of the pair measure: if a continuous discrepancy were positive at a support point, it would remain positive in some neighborhood with positive probability, contradicting zero expectation. Continuity does not extend equality beyond that support.

A support counterexample. Let a scalar nuisance s take value zero in all source examples and candidate interventions. Set $p(Y = 1 \mid s) = \sigma(as)$ for a large positive a, and let the label mechanism assign class zero everywhere. Candidate KL is zero. A target concentrated at $s = 1$ receives a near-certain wrong prediction, while a source at $s = 0$ does not reveal that behavior. One can replace the source logit by a strongly negative intercept to make source risk arbitrarily small while choosing a large enough to make target risk high. Thus no unrestricted OOD bound follows from source risk and sampled KL alone.

Observational ambiguity. Consider source binary variables with $A = B = Y$. One structural explanation makes A stable and B environment-dependent; another exchanges their roles. They induce the same observed source distribution. A source-only learner cannot choose between them without further variation, inductive assumptions, or intervention information. A mask is an inductive bias, not additional observational evidence.

Selection does not create identification. Choosing diverse candidates from learned features may explore more directions, but the meaning of those directions remains learned from the same data. If all real examples share the same shortcut, arbitrary feature extrapolation need not discover which change preserves the label. This is why the proposal distinguishes operational candidate validity from structural validity and includes oracle interventions only as an explicitly privileged experimental condition.

Implication for claims. The final paper may claim robustness over tested transformations if supported by experiments. It may claim a conditional transfer theorem under the stated assumptions. It should not claim recovery of the true SCM, causal sufficiency of retained features, or invariance to every nuisance intervention on the basis of zero empirical KL or a visually interpretable mask.

Diagnostic design consequence. Include a test where the target intervention is intentionally outside the training candidate support. The expected outcome is a weakened relationship between synthetic consistency and target risk. This is not a defect to conceal: it verifies that the stated scope corresponds to observable behavior. Also test an in-support target generated from the same intervention mixture, where the theory's assumptions are more plausible. Comparing the two cases gives a sharper assessment than averaging them into a single robustness score.

F From probability stability to classification stability

The main theorem concerns losses Lipschitz in predictive probabilities. Argmax error is discontinuous at decision boundaries: an arbitrarily small probability change can flip a tied prediction. A margin condition provides a qualified connection.

Let $a(u) = \arg\max_k p_k(u)$ with a fixed tie-breaking rule and define the top-two margin

$$\gamma(u) = p_{a(u)}(u) - \max_{k \neq a(u)} p_k(u). \tag{25}$$

For probability vectors, each coordinate difference is bounded by total variation. If $\mathrm{TV}(p(u), p(v)) < \gamma(u)/2$, then the top class cannot change: its decrease plus a competitor's increase is less than the original gap.

For any $t > 0$, under the weighted pair law ν,

$$\Pr_{\nu}(a(u) \neq a(v)) \leq \Pr_{P_w}(\gamma(u) \leq 2t) + \Pr_{\nu}(\mathrm{TV}(p(u), p(v)) \geq t) \tag{26}$$

$$\leq \Pr_{P_w}(\gamma(u) \leq 2t) + \frac{\kappa\varepsilon}{2t^2}. \tag{27}$$

The last inequality applies Markov's inequality to squared total variation and then Pinsker. A first-moment alternative gives $\sqrt{\kappa\varepsilon/2}/t$; either may be tighter in a given regime.

Candidate classification error is at most factual classification error plus prediction disagreement. After the same semantic correction and distribution transfers used in the main proof,

$$R_Q^{01} \leq R_P^{01} + \rho + \xi + \eta + \Pr_{P_w}(\gamma(u) \leq 2t) + \frac{\kappa\varepsilon}{2t^2}. \tag{28}$$

Here the margin probability is under P_w; replacing it by its source counterpart costs up to another ρ. The inequality may exceed one and should then be clipped at the trivial bound one, without implying that the unclipped expression is informative.

Reporting implications. Measure accuracy and probability divergence separately. A model can improve KL consistency without changing accuracy if most examples have large margins. It can also preserve accuracy while substantially worsening calibration. Plot the margin distribution for correct and incorrect examples, and report the fraction near the decision boundary. Do not select the margin threshold using target labels.

Temperature sensitivity. Increasing softmax temperature changes probability divergence and margins without changing argmax predictions. Therefore temperature must be fixed or selected on source validation data consistently across methods. A consistency score is not comparable across independently tuned temperatures unless that distinction is made explicit. The probability floor used for the finite-sample theorem likewise changes the evaluated probability model and must be recorded with every audit.

Cross-entropy constants. For probabilities bounded below by τ, the mean-value theorem gives $|\log p_y - \log p'_y| \le |p_y - p'_y|/\tau$. Since a singleton-coordinate difference is at most total variation, the claimed Lipschitz constant follows. The resulting $1/\tau$ factor can make the transfer bound loose when the probability floor is very small. Choosing a larger floor solely to tighten the bound changes the predictive distribution and can worsen log-loss; it is not a free improvement. Report the unsmoothed and smoothed evaluation conventions separately when both are used.

G Coverage geometry and its limitations

A displacement distribution can have large average norm while varying in only one direction. Let $z_{ij} = V^\top \Delta_{ij}$ and $G = \mathbb{E} z z^\top$. Its trace measures total squared displacement; its smallest eigenvalue measures the least explored direction within the chosen projection. Neither statistic identifies the semantic meaning of that projection.

Figure 6 summarizes the distinction between displacement magnitude and directional span.

The proposed policy uses $\log\det(G+\epsilon I)$, which rewards spreading mass across directions. Because scaling all displacements increases this objective, candidate norms must be bounded or standardized using source-fitted statistics. Otherwise the optimizer may favor unrealistic magnitude over meaningful diversity. Standardization parameters are fixed before target evaluation.

A normalized diagnostic is

$$\widetilde{G} = \frac{G}{\operatorname{tr}(G)}, \qquad r_{\text{eff}} = \exp\left(-\sum_k \lambda_k(\widetilde{G}) \log \lambda_k(\widetilde{G})\right). \tag{29}$$

For strictly positive trace, with $0 \log 0 = 0$, this is an entropy-based effective rank. Report the actual eigenvalues and trace as well, since effective rank alone hides tiny intervention magnitudes. When all displacements vanish, mark coverage undefined or zero by a declared convention.

Relation to the theorem. A large smallest eigenvalue can help in a local linear sensitivity analysis, but does not establish the density-ratio bound $w \le \kappa$ or small total-variation discrepancy η. Full-rank covariance is compatible with missing modes or large distributional mismatch. The paper must not equate a geometric coverage score with a target-support certificate.

A controlled test can isolate geometry: hold average displacement magnitude fixed, vary the rank of known nuisance interventions, and evaluate on withheld directions. Repeat with semantically invalid directions mixed in. This directly examines whether the proposed filter and diversity objective improve the relevant trade-off.

Proposition 5 (Linear directional coverage)**.** *For logits $f(a,s) = Aa + Bs$ and nuisance displacement covariance $G = \mathbb{E}\Delta\Delta^\top \succeq \lambda I$ with $\lambda > 0$,*

$$\mathbb{E}\|f(a, s+\Delta) - f(a,s)\|_2^2 = \operatorname{tr}(BGB^\top) \ge \lambda \|B\|_F^2. \tag{30}$$

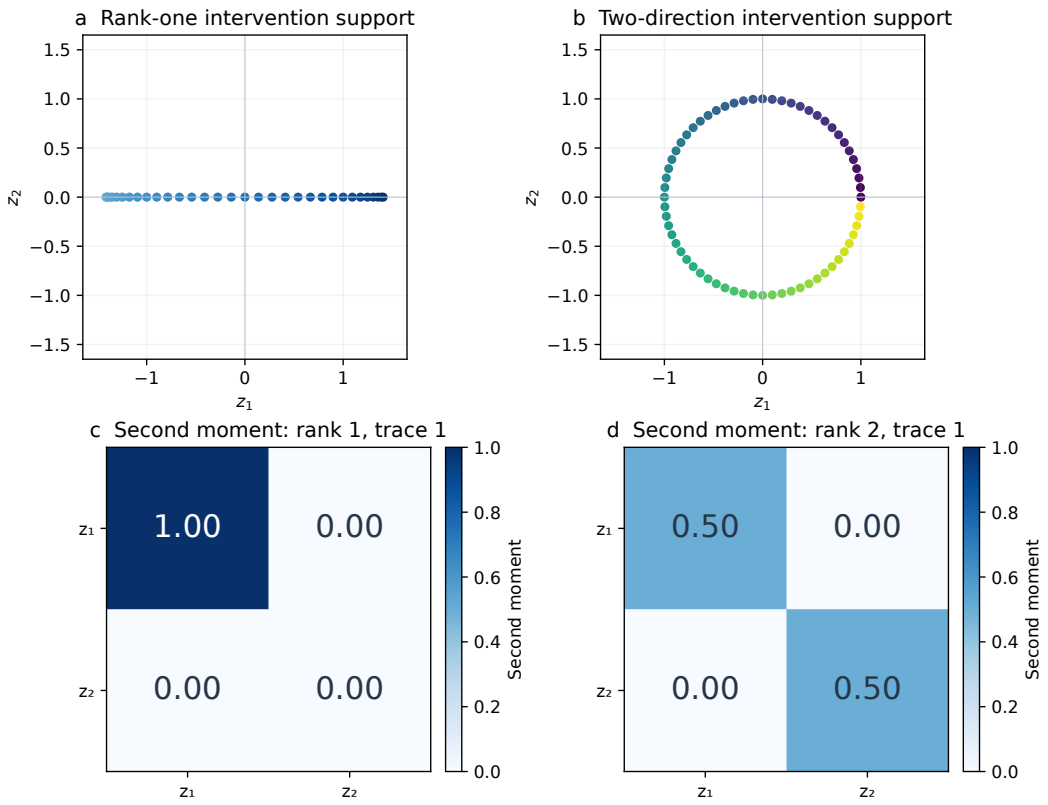


Figure 6: Analytical illustration of intervention coverage. Two deterministic 64-angle constructions have equal second-moment trace one but ranks one and two. Bottom panels report their second-moment matrices. These constructed points illustrate directional support and are not experimental observations.

Proof. Expand the squared norm as $\Delta^\top B^\top B\Delta$ and take the trace. The inequality follows because $G - \lambda I$ is positive semidefinite. If G is singular, a nonzero row of B in its nullspace can remain unconstrained. □

This elementary result motivates directional tests for linear logits. It does not identify nuisance directions in a learned nonlinear representation, and softmax probabilities additionally ignore common logit offsets.

H Policy feasibility and reproducible implementation

For a pool of k accepted candidates, a capped simplex is nonempty if and only if $kq_{\max} \geq 1$. Necessity follows by summing the caps. Sufficiency follows because the uniform vector has mass $1/k \leq q_{\max}$ and zero divergence from the uniform reference. A policy implementation must distinguish no accepted candidates from a nonempty but infeasible pool. Skipping counterfactual updates for such anchors, expanding a source-only candidate pool, or modifying the cap are different algorithms and must be recorded explicitly.

With fixed detached candidates, $G(q)$ is affine in q, the regularized log determinant is concave for $\epsilon > 0$, and the difficulty term is linear. The feasible capped-simplex and KL constraints are convex. Maximization is therefore a concave optimization problem in donor weights alone; this does not make joint representation training convex. Its unconstrained objective derivative is

$$\frac{\partial}{\partial q_j}\left[\sum_l q_l d_l + \beta \log\det(G(q) + \epsilon I)\right] = d_j + \beta z_j^\top (G(q) + \epsilon I)^{-1} z_j. \tag{31}$$

Use a linear solve or Cholesky factorization rather than an explicit matrix inverse. Log the solver tolerance, iteration count, cap residual, normalization residual, and KL residual. The manuscript's training procedure requires detached policy weights during the student step.

Candidate construction costs $O(B_a J d)$ after backbone encoding. A two-layer mask of hidden width h costs $O(B_a J d h)$ on candidates, while projected moment formation costs $O(B_a J r^2)$ and each dense factorization costs $O(r^3)$ per anchor. These are analytical scaling statements, not measured wall times. The mask adds $2dh + h + d$ parameters and remains active at inference.

I Additional intervention diagnostics

Table 5 is the numerical source for Figure 3. The coverage column is a normalized diagnostic; it is not an estimate of $1 - \eta$ in Theorem 1. Accuracy from the sweep is reported descriptively and does not define the source-validation selection criterion.

Table 5: Compatibility-threshold sweep.

γ	Accept %	Preserve %	Directional coverage	Disagree %↓	Mean Acc.	Eff. donors
0.30	94.2	89.1	0.91	8.9	80.6	7.5
0.40	87.5	92.8	0.87	6.7	81.3	7.2
0.50	80.9	95.1	0.83	4.9	81.9	7.0
0.60	74.6	96.8	0.78	3.4	82.37	6.8
0.70	62.8	97.4	0.66	2.9	82.0	5.9
0.80	48.7	98.0	0.51	2.5	81.4	4.8
0.90	29.5	98.5	0.33	2.2	80.7	3.1

I.1 Donor policies

Table 6 and Figure 7 compare the policy-level summaries. They do not provide per-class donor-mass distributions. In particular, broad effective donor support and high entropy are complementary diagnostics rather than interchangeable measures.

Table 6: Donor-policy diagnostics. Effective donor count is $1/\sum_j q_j^2$.

Policy	Accept %	Eff. donors	Nuisance distance	Max. mass
Uniform all donors	100.0	9.8	0.84	0.10
Compatible uniform	74.9	4.9	0.57	0.24
Difficulty only	74.7	4.2	0.61	0.29
Diversity only	75.2	7.1	0.78	0.18
Full constrained policy	74.6	6.8	0.75	0.19

I.2 Mask temperature

Table 7 and Figure 8 report the temperature sweep. Lower temperature does not monotonically improve accuracy or seed agreement. These aggregate diagnostics do not establish disentanglement or identify a latent causal partition.

Table 7: Mask and partition diagnostics.

Setting	Active mask %	Seed agreement	Retained energy	CF disagreement	Mean Acc.
$T_m = 2.0$	58.2	0.71	0.83	5.2	81.4
$T_m = 1.0$	51.4	0.76	0.79	4.1	82.0
$T_m = 0.5$	46.8	0.79	0.76	3.4	82.37
$T_m = 0.25$	41.7	0.74	0.70	3.2	81.8

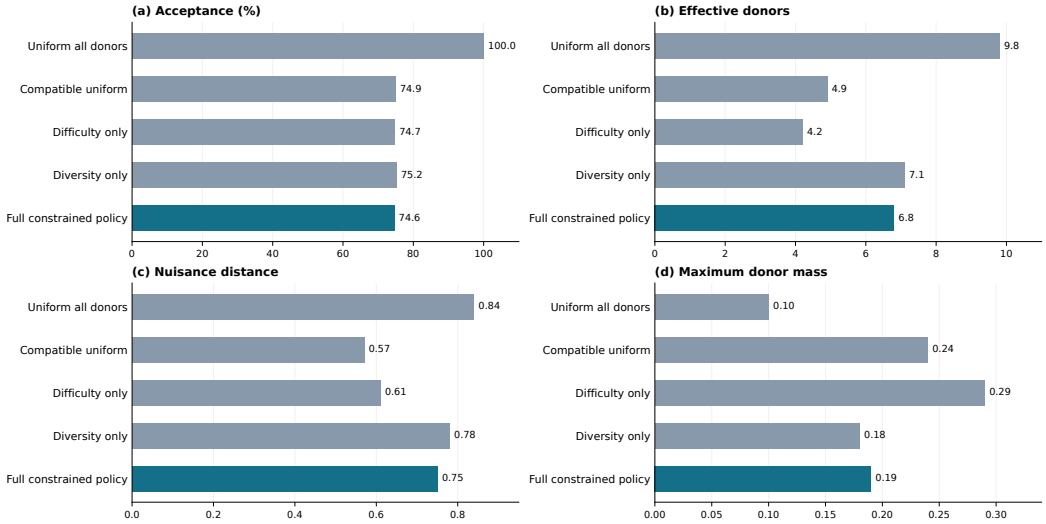


Figure 7: Donor-policy diagnostics for five reported policies. Panels show acceptance, effective donor count, nuisance distance, and maximum donor mass using the values in Table 6.

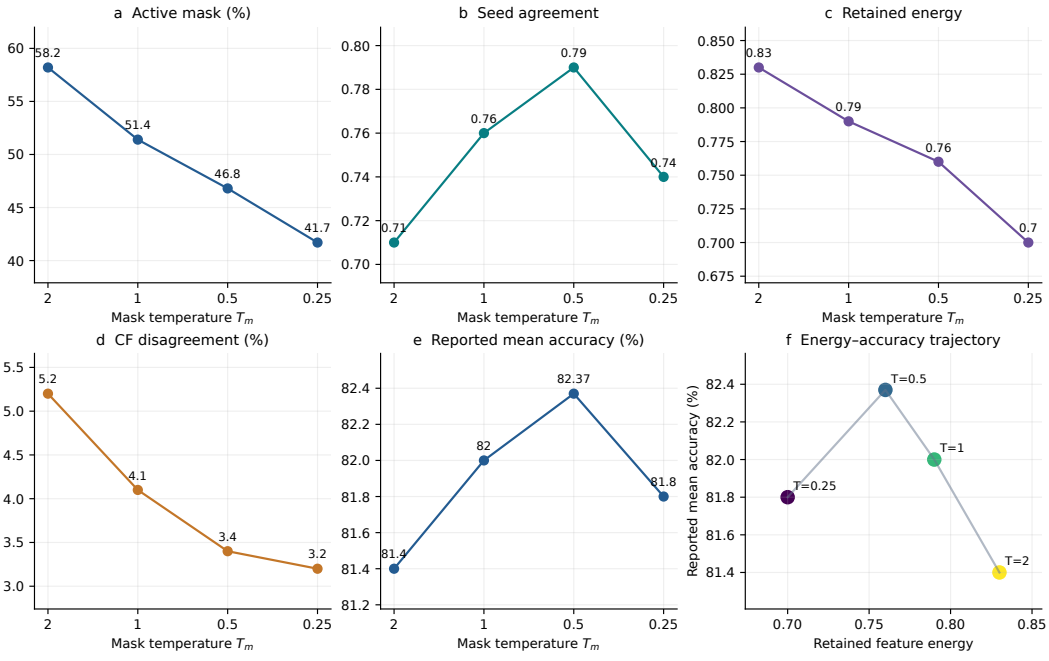


Figure 8: Mask-temperature diagnostics at four supplied settings. Five panels reproduce active-mask fraction, seed agreement, retained energy, counterfactual disagreement, and reported mean accuracy. The final panel combines the reported energy and accuracy values; connecting lines do not add observations.

I.3 CONTROLLED INTERVENTION FAMILIES

Table 8 and Figure 9 distinguish nuisance-changing controls from class-changing negative controls. Low acceptance for negative controls is desirable. Preservation and disagreement must be interpreted relative to their respective label definitions and evaluation denominators.

Table 8: Controlled intervention audit.

Intervention family	Semantics	Accept %	Preserve %	Disagree %	Coverage
Style-only positive control	Preserve	82.5	98.1	2.7	0.81
Texture-only positive control	Preserve	78.9	97.3	3.1	0.76
Background positive control	Preserve	76.4	96.5	3.8	0.73
Class-content negative control	Change	18.6	21.4	67.2	0.69
Random cross-class donor	Change	11.2	14.8	74.5	0.88

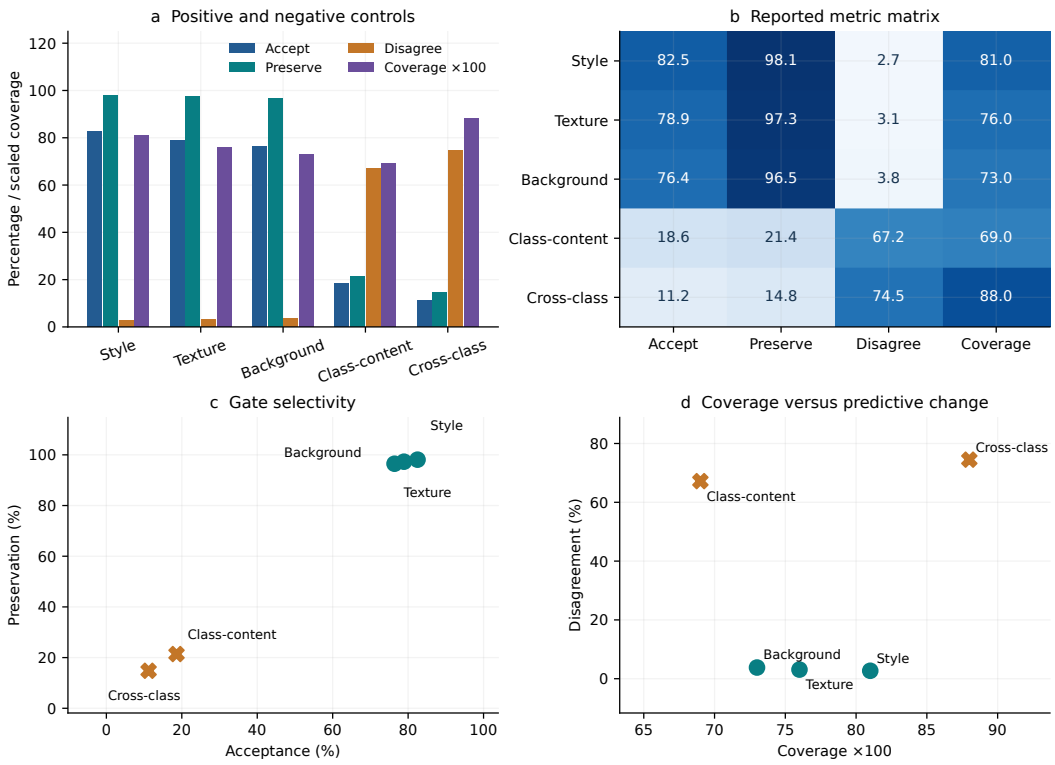


Figure 9: Controlled-intervention diagnostics for five supplied intervention families. Bars and the annotated matrix display reported summaries; joint plots relate acceptance to label preservation and coverage to disagreement. Coverage is multiplied by 100 only where indicated for display. Class-content and cross-class controls are shown separately from nuisance interventions.

J REPORTING DEFINITIONS AND REMAINING EVALUATION SCOPE

For benchmark b, domain d, and run r, let a_{bdr} be held-out-domain accuracy. A domain-macro score is $\bar{a}_{br} = D_b^{-1} \sum_d a_{bdr}$. Report its mean and sample standard deviation over runs, preserving domain-wise values and the selected checkpoint for each run. A three-benchmark macro score averages the three benchmark scores with equal benchmark weight; it is not class-balanced accuracy. Worst-domain accuracy additionally requires an explicit domain set and aggregation order.

For matched runs, paired differences are $d_r = \bar{a}_{br}^{\mathrm{CG}} - \bar{a}_{br}^{\mathrm{base}}$. A conventional five-run interval is $\bar{d} \pm t_{4,0.975} s_d/\sqrt{5}$, subject to its sampling assumptions. Marginal method standard deviations do not determine s_d because the covariance is unknown. Consequently, no paired interval is inferred from Table 2. Reporting intervals over seeds also does not quantify uncertainty over a population of new domains.

Calibration is a separate outcome from accuracy (Guo et al., 2017). With bins B_m, expected calibration error is $\sum_m (|B_m|/n)|\operatorname{acc}(B_m) - \operatorname{conf}(B_m)|$. Its binning, sample weights, and dataset aggregation must be specified to compare methods. Reliable calibration plots require individual predicted probabilities or bin counts and outcomes; aggregate ECE values cannot reconstruct a reliability diagram.

Future comparisons should include matched causal matching, MixStyle, and MIRO configurations, as well as fixed-wall-time controls. Representation projections and qualitative counterfactual panels require actual embeddings or audit examples. Synthetic points or illustrative images cannot substitute for those empirical diagnostics. Such extensions should be evaluated independently of target-based tuning.

K JOINT DIAGNOSTICS AND ANALYTICAL SENSITIVITY

The following figures separate three evidence types: reported experimental summaries, quantities derived algebraically from those summaries, and analytical examples with explicitly chosen constants. A graphical interface schematic is included to communicate how these diagnostics can be organized; it is not a claim that an audit software system was implemented or evaluated.

K.1 THRESHOLD RESPONSE ACROSS SIX COMPLEMENTARY VIEWS

Figure 10 uses exactly the seven rows of Table 5. Panels (a)–(c) preserve the raw scales. Panel (d) normalizes each metric across threshold settings as $(x - \min x)/(\max x - \min x)$ to compare response shapes. This operation does not put the metrics on a common scientific utility scale: a high preservation rate, a high rejection rate, and a high disagreement rate have different meanings. Panel (e) places the reported accuracy against coverage; it does not infer causation from this association.

Panel (f) is a derived finite difference. Between adjacent thresholds γ_j and γ_{j+1}, define

$$T_j = \frac{V(\gamma_{j+1}) - V(\gamma_j)}{A(\gamma_j) - A(\gamma_{j+1})}, \tag{32}$$

where A and V are percentages. This ratio measures preservation gain per percentage point of acceptance lost in the supplied sweep. It is defined only when its denominator is positive. It is not a derivative estimate with known sampling uncertainty and does not select a deployment threshold. Selection requires a source-validation objective and a task-specific cost for rejection, independent of target outcomes.

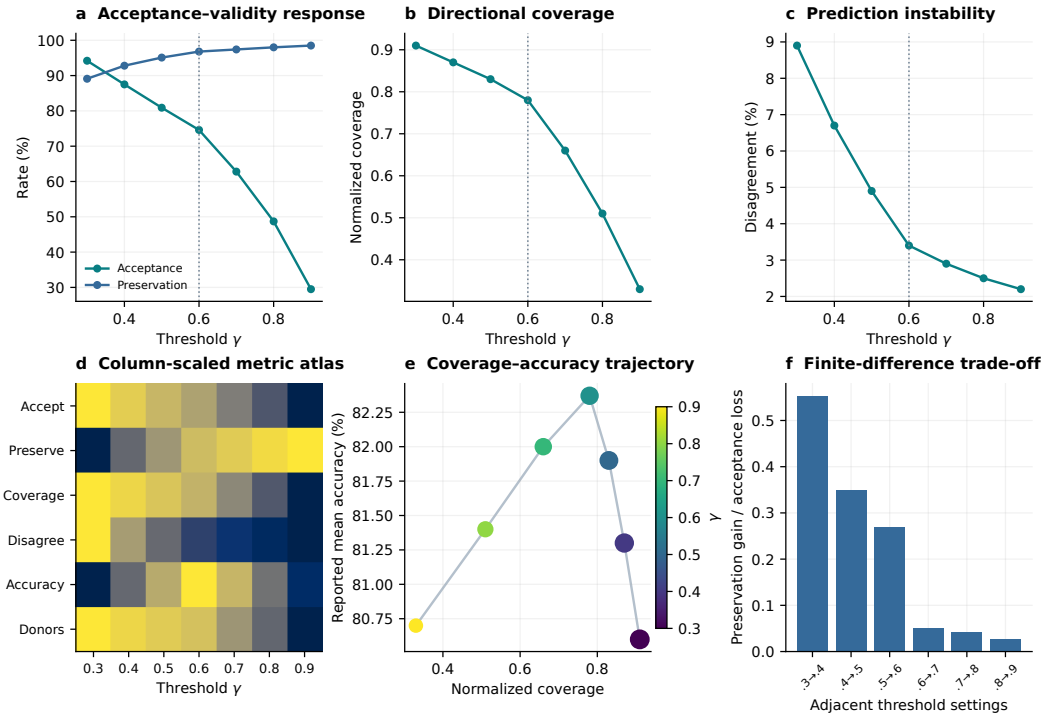


Figure 10: Threshold diagnostic atlas using the same seven reported settings. The metric matrix uses within-column scaling, while the trajectory retains reported coverage and accuracy. The final panel derives adjacent-setting preservation gains divided by acceptance losses and introduces no additional measured points.

K.2 Sensitivity of the conditional risk bound

Figure 11 evaluates the right-hand sides of the analytical bounds, not measured model performance. Throughout the population-bound panels, $B = L = 1$ and $R_P = 0.1$. Panels (a) and (b) use $\rho + \xi + \eta = 0.08$ while varying ε and κ. Panel (c) sets $\kappa = 4$, $\rho = 0.03$, and $\xi = 0.02$, and varies the unsupported-target term. Bounds above one are displayed without clipping to expose vacuity; the trivial upper bound one remains available for a loss in $[0, 1]$.

Panel (d) uses Eq. (11) with $\widehat{R} = 0.1$, $\widehat{\varepsilon} = 0.01$, $\kappa = 4$, $\tau = 0.01$, $\delta = 0.05$, and discrepancy sum 0.08. The hypothetical probability floor is valid only for a class count $C < 100$ and is not an asserted setting of the benchmark runs. Increasing portfolio size M increases the simultaneous confidence penalty. Increasing the number of independent anchors reduces it, but does not reduce an unknown support mismatch.

Panel (e) isolates the $\kappa\varepsilon/(2t^2)$ term in Eq. (27) at $\kappa = 4$. It omits the margin-distribution term and therefore is not a complete classification-risk bound. Panel (f) decomposes four illustrative population bounds. The reference uses instability contribution $\sqrt{0.004/2}$ and $(\rho, \xi, \eta) = (0.03, 0.02, 0.03)$. Each comparison changes one component only: $\eta = 0.35$, instability contribution $\sqrt{0.04/2}$, or $\xi = 0.20$.

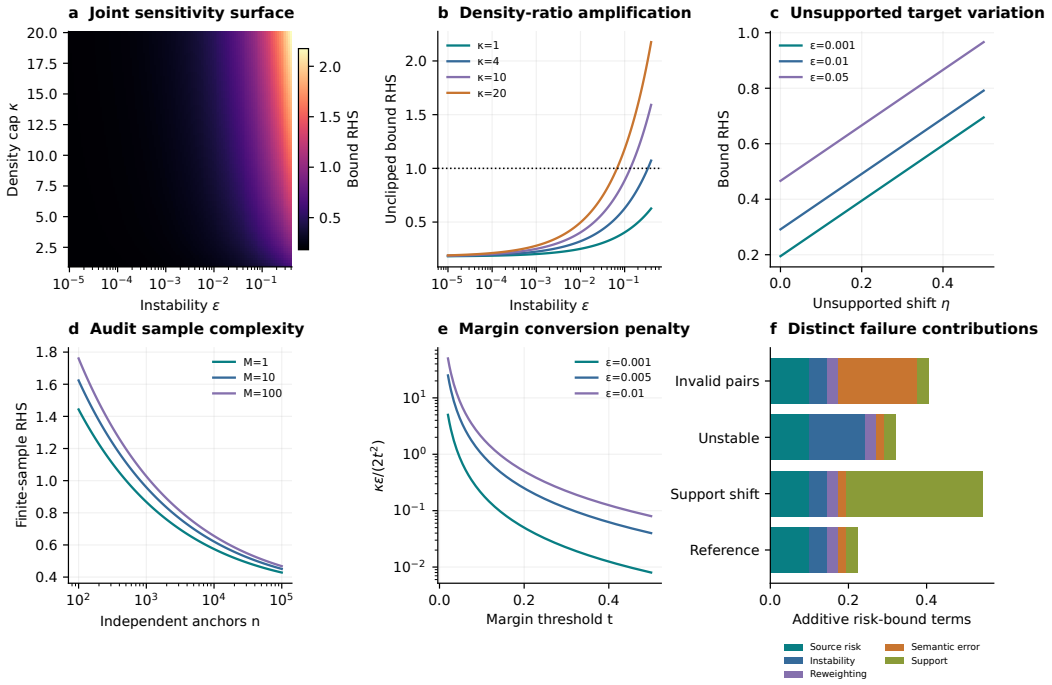


Figure 11: Analytical sensitivity atlas for the conditional risk-bound expressions. Surfaces and curves vary assumed parameters; sample-complexity and margin panels illustrate the corresponding mathematical terms. The stacked scenarios are constructed examples, not fitted experimental risk estimates or certified target guarantees.

K.3 Policy feasibility and equal-energy geometry

Figure 12(a) displays the exact condition $kq_{\max} \geq 1$ for the capped simplex. The uniform vector proves sufficiency when q^0 is uniform over the k accepted donors, since its KL divergence from the reference is zero. The boundary is a mathematical constraint; it is not a measured rate of successful policy optimization.

Panels (b) and (c) compare deterministic four-dimensional spectra $(1, 0, 0, 0)$, $(1/2, 1/2, 0, 0)$, and $(1/4, 1/4, 1/4, 1/4)$, each with trace one. Their log determinants are computed exactly as

$\sum_l \log(\lambda_l + \epsilon)$. Equal trace therefore does not imply equal span, and regularization does not turn missing directions into observed interventions. Panel (d) uses two-dimensional spectra $(1, 0)$, $(3/4, 1/4)$, and $(1/2, 1/2)$ with unit sensitivity direction $b = (\cos\alpha, \sin\alpha)$. The quadratic form $b^\top Gb$ demonstrates how equal-energy interventions can constrain different sensitivity directions unequally.

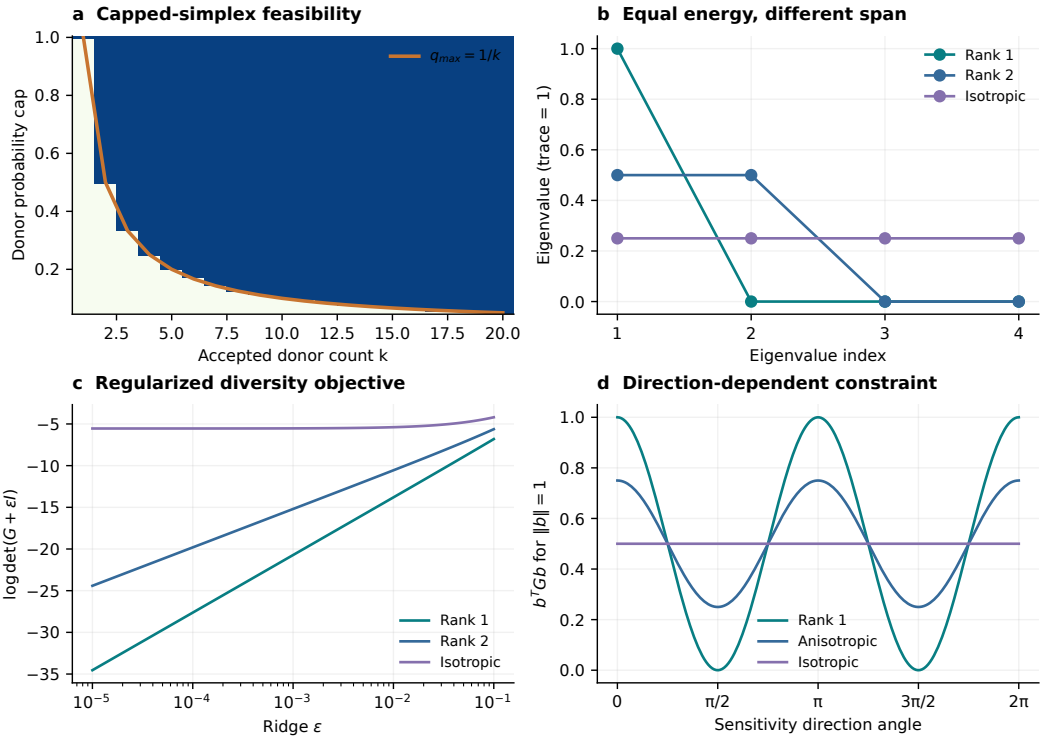


Figure 12: Analytical donor-geometry atlas. Capped-simplex feasibility follows the donor-count and mass-cap condition; equal-trace spectra demonstrate different directional support. Regularized log-determinants and directional quadratic forms are calculated from the illustrated matrices. All panels are mathematical illustrations, not additional benchmark results.

K.4 Technical workflow and audit records

Figure 5 separates candidate construction, compatibility gating, constrained donor weighting, shared re-encoding, and student updates. The lower part fixes the predictor, intervention kernel, donor bank, and audit projection before measuring instability and directional coverage. Dashed arrows identify factual and candidate inputs to the shared objective. Optional EMA use must agree with the executed configuration.

The diagram describes the method, not a screenshot of implemented software or an additional experimental result. Reproducible execution additionally requires immutable run identifiers, split hashes, the executed configuration, denominator definitions, and the underlying per-pair audit records. The analytical examples in Figures 6, 11, and 12 do not supply those experimental records.