What Can Labels Alone Audit and Edit in Frozen Representations?

Yirong Xiang*
Air Force Medical Center
yirong_xiang@163.com

Yibo Zhou†
Beihang University
ybzhou@buaa.edu.cn

Abstract

Frozen representations are often adapted using task labels alone, without environment, nuisance, or concept annotations. In this label-only regime, the first question is identifiability: what semantic target is actually determined by the joint law of representation and label? We show that labels identify only the posterior quotient, equivalently the coarsest deterministic representation that preserves all label information, and do not identify a finer semantic/background decomposition. This yields a quotient-first principle for label-only auditing and editing: first estimate a low-dimensional quotient-visible surrogate, then analyze residual class-dependent structure only after removing it. Under a structured linear second-order specialization, this leads to dual-moment decomposition (DMD). Empirically, stage one is a strong semantic audit and task-aware compression baseline across frozen CelebA, Waterbirds, and PACS features, while stage-two residual editing improves leakage–label tradeoffs only in some regimes. Additional checks on frozen sentence-encoder, decoder-only language-model, and larger vision–language representations show the same task-protection pattern beyond these vision settings. Overall, the results support a conservative recipe for label-only frozen-feature editing: preserve the quotient first, and trust residual edits when residual support and split stability are strong. Code is publicly available on GitHub.

1 Introduction

Large pretrained encoders are increasingly reused through cached embeddings or other frozen-feature interfaces, after which downstream adaptation often has access only to representations and task labels rather than the original pretraining pipeline. This frozen-feature regime is now common in probe-style adaptation and post hoc representation intervention, across both vision and language settings [Radford et al., 2021, Wu et al., 2024, El Banani et al., 2024, Bartoszcze et al., 2025, Nguyen et al., 2026]. Practitioners still want to answer familiar questions: where does task-relevant information live, what residual structure can be edited away, and when should such edits be trusted? But before asking how to recover semantics and nuisance, one must ask a prior question of observability: what target is actually determined by $P(X, Y)$ when labels are the only supervision?

Our starting point is that labels alone do *not* identify an arbitrary semantic/background decomposition of a frozen representation. The strongest semantic object canonically determined by the joint law of features and labels is the posterior quotient, namely the coarsest deterministic representation that preserves all label information. Beyond that quotient, finer role assignments are observationally ambiguous: the same within-quotient variation can be treated as semantic in one valid decomposition and as background in another.

This boundary suggests a workflow rather than a latent-recovery claim. We first estimate a low-dimensional quotient-visible surrogate and use it as a semantic audit/compression anchor. We then

*Primary corresponding author.
†Co-corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

study residual structure only *after* projecting out this quotient-visible surrogate. Under a structured linear second-order specialization, this leads to dual-moment decomposition (DMD): a first stage based on class means and a second stage based on post-quotient class-dependent covariance.

Across frozen features from CelebA [Liu et al., 2015], Waterbirds [Sagawa et al., 2020], and PACS [Li et al., 2017], extracted with ResNet-18 [He et al., 2016] and CLIP ViT-B/32 [Radford et al., 2021], we find a clear empirical asymmetry. Stage one is consistently strong and often tied with LDA, as expected in the finite-class linear setting studied here. Stage two can help substantially, but not uniformly: in several regimes quotient-first residualization yields a cleaner leakage–label tradeoff than raw covariance or generic projection-matched alternatives, while in others multiple strong methods are nearly tied. Across repeated reruns, support quality and split stability emerge as the most informative practical checks for deciding whether stage two is worth trusting.

On the editing side, our setting is related to concept erasure, adversarial attribute removal, post hoc representation editing, and recent representation-intervention methods [Bolukbasi et al., 2016, Elazar and Goldberg, 2018, Ravfogel et al., 2020, 2022, Belrose et al., 2023, Wu et al., 2024, Zou et al., 2024]. It is also adjacent to robustness and spurious-correlation methods that use observed or inferred environments [Arjovsky et al., 2019, Sagawa et al., 2020, Creager et al., 2021, Lin et al., 2022, Ahuja et al., 2021, Arefin et al., 2024, He et al., 2025, Lu et al., 2025, Wen et al., 2025, Zhou et al., 2026]. Unlike those lines of work, we assume only (X, Y) and treat post-quotient edits as observable residual operations rather than recovery of a uniquely identified nuisance factor.

Taken together, the paper makes three contributions. First, we make the identifiability boundary in the label-only frozen-representation setting explicit: labels determine a quotient-level semantic object, but not a finer semantic/background decomposition. Second, we use that boundary to formulate label-only editing as a protected-deletion problem: quotient-visible structure is preserved first, and a post-quotient residual objective is optimized only over admissible deletion subspaces. The individual mean- and covariance-based estimators are closely related to classical supervised dimension reduction; the methodological contribution is their identifiability-driven ordering and the resulting constrained editing formulation. Third, we show empirically that the first stage is already a strong audit/compression anchor, while the second stage is useful only in some regimes and is best trusted when residual support and split stability are strong.

2 Observable targets in the label-only regime

Let X be a frozen representation on a standard Borel space $\mathcal{X}$ and $Y \in [K] := \{1, \dots, K\}$ a task label; Δ^{K-1} denotes the probability simplex. The label-only learner observes only $P_{X,Y}$, or an i.i.d. sample from it. Define

$$\eta(x) := P(Y \mid X = x) \in \Delta^{K-1}.$$

A deterministic representation $T = t(X)$ is label-sufficient if $Y \perp X \mid T$. We compare deterministic representations by the factor order

$$T_1 \preceq T_2 \quad \Longleftrightarrow \quad T_1 = g(T_2) \text{ a.s. for some measurable } g.$$

Theorem 1 (Posterior quotient)**.** *The posterior representation $\eta(X)$ is label-sufficient:*

$$Y \perp X \mid \eta(X).$$

Moreover, if $T = t(X)$ is deterministic and label-sufficient, then

$$\eta(X) \preceq T.$$

Equivalently,

$$\sigma(\eta(X)) = \bigcap_{T:\, Y \perp X \mid T} \sigma(T) \quad \textit{modulo } P\textit{-null sets}.$$

Finally, for every finite action space $\mathcal{A}$ and bounded loss $\ell : \mathcal{A} \times [K] \to \mathbb{R}$,

$$\inf_{a:\mathcal{X}\to\mathcal{A}} \mathbb{E}[\ell(a(X), Y)] = \inf_{b:\Delta^{K-1}\to\mathcal{A}} \mathbb{E}[\ell(b(\eta(X)), Y)].$$

Conversely, if a deterministic statistic $T = t(X)$ has the same Bayes risk as X for every finite-action bounded label-prediction loss, then $\eta(X) \preceq T$.

Thus labels identify a quotient of representation space,

$$x \sim_\eta x' \quad \Longleftrightarrow \quad P(Y \mid X = x) = P(Y \mid X = x'),$$

not a semantic/background factorization.

A deterministic decomposition $(S, B) = (s(X), b(X))$ is label-valid if $Y \perp X \mid S$. The next result shows that any assignment of variation inside posterior fibers to "semantic" or "background" is observationally ambiguous.

Theorem 2 (Non-identifiability beyond the quotient)**.** *Let $\phi(X)$ be any statistic not measurable with respect to $\sigma(\eta(X))$. Define the role property*

$$\theta(S, B) := \mathbf{1}\{\phi(X) \textit{ is measurable with respect to } \sigma(S) \textit{ modulo null sets}\}.$$

For any label-only procedure $\widehat{\theta}$, measurable with respect to $P_{X,Y}$ or to an i.i.d. sample from it,

$$\sup_{\substack{(S,B):\, Y \perp X \mid S \\ P_{X,Y} \text{ fixed}}} \mathbb{P}\{\widehat{\theta} \neq \theta(S, B)\} \geq \frac{1}{2}.$$

The proof constructs two label-valid decompositions, $(\eta(X), \phi(X))$ and $((\eta(X), \phi(X)), 0)$, which induce the same observable law but disagree on the role of $\phi(X)$. A confidence-set version and all proofs are in Appendix A. Hence residual edits can be meaningful only as post-quotient observable operations, not as recovery of a uniquely identified nuisance factor.

3 Quotient-first residual editing

The observable within-quotient object, defined up to $P_{\eta(X)}$-null sets, is the posterior-fiber law

$$\mathfrak{R}(t) := P_{X \mid \eta(X) = t}.$$

By Theorem 1, for $P_{\eta(X),Y}$-almost every (t, y),

$$P_{X \mid \eta(X) = t, Y = y} = P_{X \mid \eta(X) = t}.$$

Thus labels reveal no further class-specific law inside an exact posterior fiber. In practice, however, the exact posterior quotient is unavailable. We therefore estimate a low-dimensional linear quotient surrogate and analyze class-dependent second moments that remain after projecting it out. These are observable surrogate-residual statistics, not residual label information beyond the exact quotient or a uniquely identified latent background factor. We therefore specialize to an observable second-order target in frozen-feature space. The use of class-conditional covariance information is related to second-order sufficient dimension reduction methods such as SAVE, principal Hessian directions, and directional regression [Cook and Weisberg, 1991, Li, 1992, Li and Wang, 2007]; the key distinction is that the second-order operator is applied after quotient-first residualization and treated as an observable residual edit rather than a means of latent-factor recovery.

Scope of the specialization. Theorems 1–2 are model-free observational statements about $P_{X,Y}$; the model below adds representation-level structure under which one post-quotient operation can be analyzed. It does not make U, V, S, or B unique and does not assume linearly separable classes or a perfect linear classifier; the key population assumption is that $U^\top X$ is label-sufficient.

Assume the representation-level model

$$X = US + VB + \varepsilon, \tag{1}$$

where $U \in \mathbb{R}^{p \times r}$ and $V \in \mathbb{R}^{p \times q}$ have orthonormal columns, $U^\top V = 0$, US is quotient-visible, VB is structured post-quotient residual variation, and ε is noise. We assume:

$$U^\top X \text{ is label-sufficient in the population specialization,}$$
$$\mathbb{E}[B \mid Y] = 0, \qquad \mathbb{E}[\varepsilon \mid Y] = 0,$$
$$\varepsilon \text{ is conditionally uncorrelated with } (S, B) \text{ given } Y, \quad \mathrm{Cov}(\varepsilon \mid Y) \text{ is independent of } Y,$$

the class-dependent second-order residual variation analyzed by DMD is summarized by $\mathrm{Cov}(B \mid Y)$.

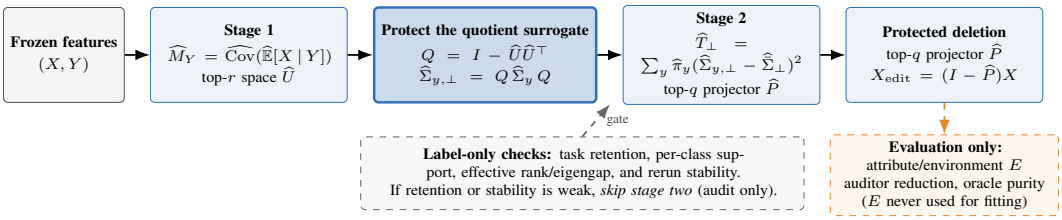


Figure 1: Quotient-first auditing and protected residual deletion. DMD is fit using (X, Y) only. Attribute or environment labels E, when available, are used only for post hoc evaluation and controlled support analyses.

In practice, U is estimated as a low-dimensional quotient-visible surrogate. Let

$$M_Y := \mathrm{Cov}(\mathbb{E}[X \mid Y]), \qquad \Sigma_y := \mathrm{Cov}(X \mid Y = y), \qquad \pi_y := P(Y = y).$$

After quotient residualization,

$$\Sigma_{y,\perp} := P_{U^\perp}\Sigma_y P_{U^\perp}, \qquad \bar{\Sigma}_\perp := \sum_y \pi_y \Sigma_{y,\perp},$$

and

$$T_\perp := \sum_{y=1}^{K} \pi_y(\Sigma_{y,\perp} - \bar{\Sigma}_\perp)^2.$$

Here $P_{U^\perp} := I - UU^\top$. For a symmetric matrix A, A^2 denotes AA. Eigenvalues $\lambda_j(\cdot)$ are ordered nonincreasingly, and $\sin\Theta(\cdot,\cdot)$ denotes the usual principal-angle distance between subspaces.

Theorem 3 (Dual-moment factorization and optimality)**.** *Under model (1) and the assumptions above, define*

$$A := \mathrm{Cov}(\mathbb{E}[S \mid Y]), \qquad \Gamma_y := \mathrm{Cov}(B \mid Y = y), \qquad \bar{\Gamma} := \sum_y \pi_y \Gamma_y,$$

and

$$G := \sum_y \pi_y(\Gamma_y - \bar{\Gamma})^2.$$

Then

$$M_Y = UAU^\top, \qquad T_\perp = VGV^\top.$$

Thus the mean stage identifies the mean-active quotient subspace $\mathrm{range}(UA)$*, and the residual stage identifies the covariance-active residual subspace* $\mathrm{range}(VG)$ *when the corresponding ranks and eigengaps are nondegenerate.*

Moreover, for any rank-q projector P with $PU = 0$, define

$$\mathcal{J}_U(P) := \sum_y \pi_y \|(\Sigma_{y,\perp} - \bar{\Sigma}_\perp)P\|_F^2 = \mathrm{tr}(PT_\perp).$$

If $P_\star$ is the top-q eigenspace of $T_\perp$ restricted to $U^\perp$, then

$$P_\star \in \arg\max_{\substack{P=P^\top=P^2,\ \mathrm{rank}(P)=q \\ PU=0}} \mathcal{J}_U(P), \qquad \max \mathcal{J}_U(P) = \sum_{j=1}^{q} \lambda_j(T_\perp).$$

If

$$\gamma_q := \lambda_q(T_\perp) - \lambda_{q+1}(T_\perp) > 0,$$

then

$$\mathcal{J}_U(P_\star) - \mathcal{J}_U(P) \geq \frac{\gamma_q}{2}\|P - P_\star\|_F^2.$$

The curvature inequality makes the residual eigengap the identifiability modulus of stage two. Small γ_q means that many residual subspaces are nearly equivalent under the observable objective, so the edit is unstable. Also, since $\mathbb{E}[X \mid Y]$ takes at most K values, $\mathrm{rank}(M_Y) \leq K - 1$; the first stage is therefore a low-dimensional quotient surrogate, not a recovery of every possible semantic direction.

A raw second-order edit instead uses

$$T_{\mathrm{raw}} := \sum_y \pi_y (\Sigma_y - \bar{\Sigma})^2, \qquad \bar{\Sigma} := \sum_y \pi_y \Sigma_y,$$

before removing quotient-visible variation. Write

$$\Delta_y := \Sigma_y - \bar{\Sigma} = U A_y U^\top + U C_y V^\top + V C_y^\top U^\top + V B_y V^\top + R_y,$$

where A_y is semantic covariance variation, C_y is semantic–residual cross-variation, B_y is residual covariance variation, and R_y is remainder.

Theorem 4 (Contamination of raw covariance editing)**.** *Let $G_B := \sum_y \pi_y B_y^2$. Then*

$$T_{\mathrm{raw}} = V G_B V^\top + E_{\mathrm{sem}},$$

with

$$\|E_{\mathrm{sem}}\|_{\mathrm{op}} \le C \sum_y \pi_y \left(\|A_y\|_{\mathrm{op}}^2 + \|C_y\|_{\mathrm{op}}^2 + \|R_y\|_{\mathrm{op}}^2 + \|B_y\|_{\mathrm{op}} \|C_y\|_{\mathrm{op}} + \|B_y\|_{\mathrm{op}} \|R_y\|_{\mathrm{op}} \right).$$

Consequently, if $\widehat{P}_{\mathrm{raw}}$ is the top-q eigenspace of T_{raw} and P_B is the top-q eigenspace of $V G_B V^\top$, define

$$\delta_B := \lambda_q(V G_B V^\top) - \lambda_{q+1}(V G_B V^\top).$$

Whenever $\delta_B > \|E_{\mathrm{sem}}\|_{\mathrm{op}}$,

$$\|\sin\Theta(\widehat{P}_{\mathrm{raw}}, P_B)\|_{\mathrm{op}} \le \frac{\|E_{\mathrm{sem}}\|_{\mathrm{op}}}{\delta_B - \|E_{\mathrm{sem}}\|_{\mathrm{op}}}.$$

In particular, if semantic covariance energy dominates residual covariance energy, the leading raw-covariance directions may lie in $\mathrm{span}(U)$ *even when nonzero residual signal exists.*

Thus raw covariance editing can reduce leakage while damaging label prediction, because its leading directions may still contain quotient-visible structure. Quotient-first residualization removes the $U A_y U^\top$ and $U C_y V^\top + V C_y^\top U^\top$ channels before applying the residual objective.

4 Estimation, perturbation, and trust

Given samples $\{(x_i, y_i)\}_{i=1}^n$, DMD estimates

$$\widehat{M}_Y = \widehat{\mathrm{Cov}}\big(\widehat{\mathbb{E}}[X \mid Y]\big)$$

and takes its top-r eigenspace $\widehat{U}$ as a mean-active quotient surrogate; this empirical subspace need not equal the exact posterior quotient $\eta(X)$, and its adequacy as a protected subspace is assessed through held-out task retention. It then forms

$$\widehat{\Sigma}_{y,\perp} = P_{\widehat{U}^\perp} \widehat{\Sigma}_y P_{\widehat{U}^\perp}, \qquad \widehat{\bar{\Sigma}}_\perp = \sum_y \widehat{\pi}_y \widehat{\Sigma}_{y,\perp},$$

and

$$\widehat{T}_\perp = \sum_y \widehat{\pi}_y (\widehat{\Sigma}_{y,\perp} - \widehat{\bar{\Sigma}}_\perp)^2.$$

The top-q eigenspace of $\widehat{T}_\perp$ is used for residual analysis or editing. When needed, we write $\widehat{T}_{\widehat{U}^\perp}$ for this empirical operator computed with $\widehat{U}$, and $T_{U^\perp}$ for its population counterpart computed with U.

We state a finite-sample perturbation bound using effective rank. Let $\mu_y := \mathbb{E}[X \mid Y = y]$, and assume $L \ge 1$ and that each centered class conditional is covariance-sub-Gaussian:

$$\|\langle u, X - \mu_y \rangle\|_{\psi_2} \le L \sqrt{u^\top \Sigma_y u} \qquad \text{for all } u \in \mathbb{R}^p,$$

with

$$\max_y \left\{ \|\mu_y\|_2^2, \|\Sigma_y\|_{\mathrm{op}} \right\} \le B.$$

For $r_{\text{eff}}(\Sigma) := \operatorname{tr}(\Sigma)/\|\Sigma\|_{\text{op}}$, let

$$r_{\text{eff}} := \max_y r_{\text{eff}}(\Sigma_y), \qquad n_{\min} := \min_y n_y \geq 1,$$

and

$$\xi_n(\delta) := \sqrt{\frac{r_{\text{eff}} + \log(K/\delta)}{n_{\min}}} + \frac{r_{\text{eff}} + \log(K/\delta)}{n_{\min}}.$$

Theorem 5 (Finite-sample DMD perturbation)**.** *With probability at least* $1 - \delta$*,*

$$\|\widehat{M}_Y - M_Y\|_{\text{op}} \leq CBL^2\xi_n(\delta), \qquad \max_y \|\widehat{\Sigma}_y - \Sigma_y\|_{\text{op}} \leq CBL^2\xi_n(\delta).$$

Let P_U *and* $\widehat{P}_U$ *be the population and empirical stage-one projectors, and set*

$$\epsilon_U := \|\widehat{P}_U - P_U\|_{\text{op}}.$$

Then

$$\|\widehat{T}_{\widehat{U}^\perp} - T_{U^\perp}\|_{\text{op}} \leq CB^2L^2\xi_n(\delta) + CB^2L^4\xi_n^2(\delta) + CB^2\epsilon_U.$$

If

$$\gamma_U := \lambda_r(M_Y) - \lambda_{r+1}(M_Y) > 0$$

and $\gamma_U > 2CBL^2\xi_n(\delta)$*, then*

$$\epsilon_U \leq \frac{2CBL^2\xi_n(\delta)}{\gamma_U}.$$

If

$$\gamma_V := \lambda_q(T_\perp) - \lambda_{q+1}(T_\perp) > 0$$

and

$$\gamma_V > 2\left(CB^2L^2\xi_n(\delta) + CB^2L^4\xi_n^2(\delta) + CB^2\epsilon_U\right),$$

then

$$\|\sin\Theta(\widehat{V}, V)\|_{\text{op}} \leq \frac{2\left(CB^2L^2\xi_n(\delta) + CB^2L^4\xi_n^2(\delta) + CB^2\epsilon_U\right)}{\gamma_V}.$$

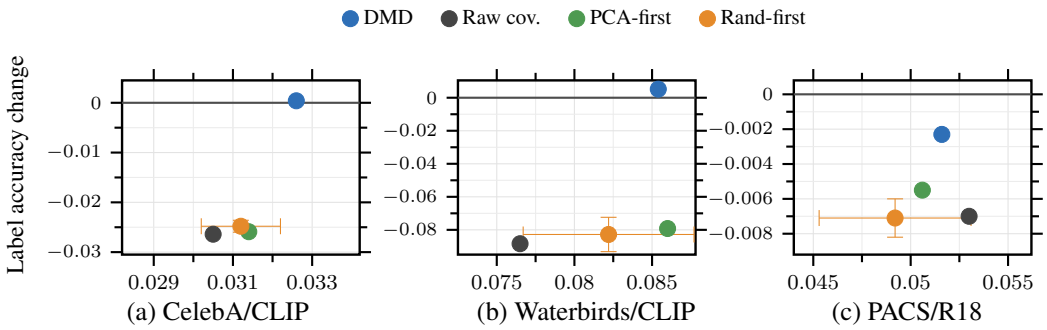
Theorem 5 explains the empirical asymmetry: stage one depends on a supervised mean eigengap, whereas stage two additionally requires covariance concentration, residual eigengap, and accurate quotient estimation.

The same perturbation view motivates our trust diagnostics. First, residual covariance is support-limited: if $X \mid Y = y \sim \sum_e \alpha_{y,e} P_{y,e}$, then the covariance noise floor scales with the smallest realized cell count $s_{\min} := \min_{y,e} n_{y,e}$; its relation to $n_{\min} \min_{y,e} \alpha_{y,e}$ and a formal bound are in Appendix A. Second, split stability estimates the perturbation-to-gap ratio: for rerun projectors $\widehat{P}_b$, Davis–Kahan gives $\|\widehat{P}_b - P_\star\|_F \lesssim \|E_b\|_{\text{op}}/\gamma_V$, and averaging projectors reduces variance at the usual $1/B$ rate. Third, permutation-null spectra are conditionally valid under $Y \perp X$ by label exchangeability and are used only as supporting rank diagnostics. Together these results give the usage rule used in the experiments: *preserve the quotient first; edit residuals only when support and split stability are strong.*

5 Experiments

Unless stated otherwise, all real-data results are reported as mean $\pm$ standard deviation over 10 outer random seeds. Within each outer seed, stability-oriented diagnostics are computed from repeated class-stratified reruns. DMD is always fit from (X, Y) alone. Observed environment labels, when available, are used only post hoc for evaluation, including environment-leakage probes, oracle residual metrics, and controlled support sweeps.

We evaluate frozen representations from CelebA [Liu et al., 2015], Waterbirds [Sagawa et al., 2020], and PACS [Li et al., 2017] using pretrained ResNet-18 [He et al., 2016] and CLIP ViT-B/32 [Radford et al., 2021] features. Main-text residual editing uses a fixed removal rank $q = 8$ to compare methods under the same edit budget. Baselines are raw covariance editing and projection-matched variants that remove a rank-r PCA, LDA, PLS, or random subspace before applying the same second-order construction. Permutation diagnostics, expanded baselines, and rank-sensitivity summaries are in the



(a) CelebA/CLIP (b) Waterbirds/CLIP (c) PACS/R18

Figure 2: Fixed-rank residual editing at $q = 8$. Points show mean performance over 10 outer seeds; error bars are omitted when visually negligible. In all three panels, the horizontal axis is environment leakage reduction. Quotient-first editing is substantially cleaner than raw covariance and generic projection-matched baselines on CelebA/CLIP, Waterbirds/CLIP, and PACS/R18. LDA-first is numerically almost identical to DMD and is omitted for readability.

appendix. For an edited representation X_{edit}, we report $\Delta\text{Leak} := \text{Leak}(X) - \text{Leak}(X_{\text{edit}})$, where Leak is held-out environment-probe accuracy; $\Delta\text{Acc} := \text{Acc}(X_{\text{edit}}) - \text{Acc}(X)$; and $\Delta\text{NLL} := \text{NLL}(X_{\text{edit}}) - \text{NLL}(X)$. Thus larger ΔLeak and ΔAcc are better, while smaller ΔNLL is better. When environment labels are available for evaluation only, oracle purity is $\text{tr}(P_{\text{edit}} P_{\text{oracle}})/q$, where P_{edit} is the removed rank-q projector and P_{oracle} is the oracle environment-residual projector; oracle gain means purity improvement over raw covariance at the same rank.

5.1 Stage one is a strong and stable semantic audit

Table 1: Stage-one auditing on frozen features. Accuracy is mean ± std over 10 outer seeds. R18 = ResNet-18; ViT = CLIP ViT-B/32. Standard deviations smaller than 10^{-3} are shown as < 0.001.

Setting	Full	DMD-U	LDA-U	PCA-U	Rand-U
CelebA/R18	0.907±<0.001	0.906±<0.001	0.906±<0.001	0.521±<0.001	0.532±0.017
Waterbirds/R18	0.755±<0.001	0.729±<0.001	0.729±<0.001	0.441±<0.001	0.522±0.043
PACS/R18	0.859±<0.001	0.724±<0.001	0.725±<0.001	0.202±<0.001	0.214±0.021
CelebA/ViT	0.926±<0.001	0.929±<0.001	0.929±<0.001	0.516±<0.001	0.530±0.020
Waterbirds/ViT	0.808±<0.001	0.798±<0.001	0.798±<0.001	0.547±<0.001	0.545±0.049
PACS/ViT	0.968±<0.001	0.860±<0.001	0.860±<0.001	0.175±<0.001	0.240±0.027

Table 1 summarizes stage-one semantic auditing across six dataset/backbone settings. The pattern is consistent across 10 seeds: a compact quotient-visible surrogate preserves task prediction far better than unsupervised or random low-rank alternatives and is essentially tied with LDA in every setting we study. The largest margins over PCA-U and Rand-U appear on Waterbirds and PACS, especially with CLIP features; representative values are shown in Table 1. After the total-covariance whitening used in our pipeline, the DMD mean-stage and LDA projectors have overlap $\text{tr}(P_{\text{DMD}} P_{\text{LDA}})/r = 1.000$ in all six settings (0.998–1.000 against PLS), so stage one should be read as the protected quotient surrogate rather than as a new supervised estimator.

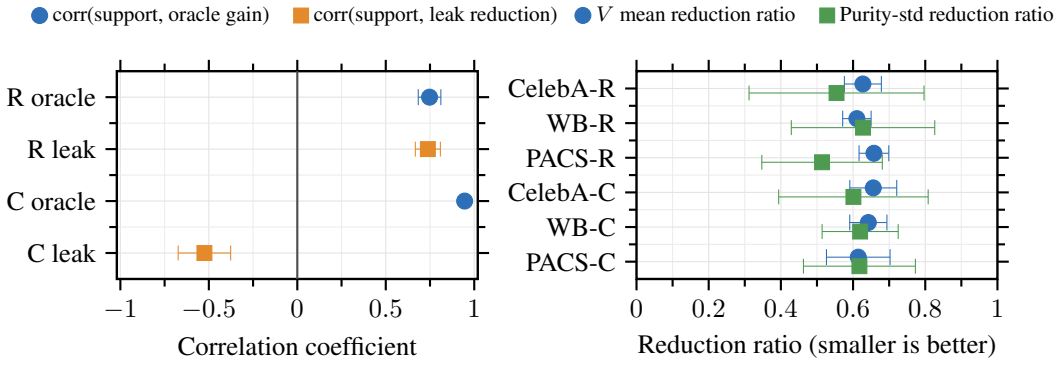
In several settings the quotient surrogate is extremely compact yet nearly lossless. On CelebA / ResNet-18, DMD-U attains 0.906 accuracy versus 0.907 for the full representation. On CelebA / CLIP, the one-dimensional surrogate slightly exceeds full-feature held-out accuracy (0.929 vs. 0.926), although its NLL is slightly worse, which is consistent with viewing stage one as a supervised compression step rather than a claim of dominance over the full representation.

5.2 Residual editing and stage-two trust

We next compare residual editing at a fixed removal rank $q = 8$, which avoids conflating purity gains with different editing strengths. Figure 2 shows three representative regimes, and Table 2 reports exact results for all six settings. Across 10 seeds, DMD and LDA-first are numerically almost identical, reflecting that both provide similar supervised low-dimensional quotient-visible anchors in these finite-class linear settings. We therefore attribute the observed gains primarily to the quotient-first

Table 2: Fixed-rank residual editing summary ($q = 8$). Mean $\pm$ std over 10 outer seeds. R18 = ResNet-18; ViT = CLIP ViT-B/32. Expanded baselines and sensitivity summaries are reported in the appendix. Standard deviations smaller than 10^{-3} are shown as < 0.001.

Setting	Method	ΔLeak	ΔAcc	Oracle purity
CelebA / R18	DMD	$0.013 \pm < 0.001$	$-0.002 \pm < 0.001$	$0.908 \pm < 0.001$
	Raw cov.	$0.013 \pm < 0.001$	$-0.045 \pm < 0.001$	$0.297 \pm < 0.001$
	PCA-first	$0.012 \pm < 0.001$	$-0.049 \pm < 0.001$	$0.281 \pm < 0.001$
Waterbirds / R18	DMD	$0.009 \pm < 0.001$	$-0.001 \pm < 0.001$	$0.508 \pm < 0.001$
	Raw cov.	$0.012 \pm < 0.001$	$-0.030 \pm < 0.001$	$0.435 \pm < 0.001$
	PCA-first	$0.003 \pm < 0.001$	$-0.025 \pm < 0.001$	$0.428 \pm < 0.001$
PACS / R18	DMD	$0.052 \pm < 0.001$	$-0.002 \pm < 0.001$	$0.920 \pm < 0.001$
	Raw cov.	$0.053 \pm < 0.001$	$-0.007 \pm < 0.001$	$0.688 \pm < 0.001$
	PCA-first	$0.051 \pm < 0.001$	$-0.006 \pm < 0.001$	$0.691 \pm < 0.001$
CelebA / ViT	DMD	$0.033 \pm < 0.001$	$0.000 \pm < 0.001$	$0.831 \pm < 0.001$
	Raw cov.	$0.031 \pm < 0.001$	$-0.026 \pm < 0.001$	$0.262 \pm < 0.001$
	PCA-first	$0.031 \pm < 0.001$	$-0.026 \pm < 0.001$	$0.263 \pm < 0.001$
Waterbirds / ViT	DMD	$0.085 \pm < 0.001$	$0.005 \pm < 0.001$	$0.730 \pm < 0.001$
	Raw cov.	$0.077 \pm < 0.001$	$-0.088 \pm < 0.001$	$0.509 \pm < 0.001$
	PCA-first	$0.086 \pm < 0.001$	$-0.079 \pm < 0.001$	$0.540 \pm < 0.001$
PACS / ViT	DMD	$0.064 \pm < 0.001$	$0.000 \pm < 0.001$	$0.971 \pm < 0.001$
	Raw cov.	$0.065 \pm < 0.001$	$-0.002 \pm < 0.001$	$0.971 \pm < 0.001$
	PCA-first	$0.055 \pm < 0.001$	$0.000 \pm < 0.001$	$0.969 \pm < 0.001$



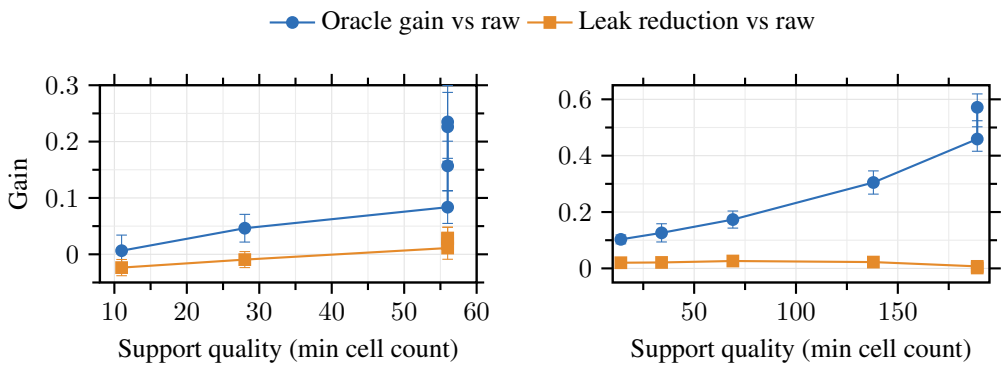
(a) Support diagnostics on Waterbirds

(b) Bagging improves stability

Figure 3: When is stage two worth trusting? Left panel: in controlled Waterbirds support sweeps, support quality strongly predicts oracle-side gains across 10 seeds, especially for CLIP; circles denote oracle-side gains and squares denote leakage-reduction correlations. Right panel: bagging consistently reduces stage-two subspace instability and metric variability across all six dataset/backbone settings. In panel (a), C denotes CLIP ViT-B/32 and R denotes ResNet-18; "oracle" abbreviates oracle-side gain.

residualization step. The mean-stage estimator should be viewed as a convenient instantiation of this principle, closely related to classical supervised dimension reduction.

On CelebA / CLIP, DMD achieves environment-leakage reduction 0.033 with no measurable label loss and oracle purity 0.831, whereas raw covariance achieves 0.031, -0.026, and 0.262, and PCA-first achieves 0.031, -0.026, and 0.263. On Waterbirds / CLIP, DMD reduces environment leakage by 0.085 while slightly improving label accuracy $(+0.005)$, compared with raw covariance at 0.077 leakage reduction with a large label drop (-0.088), and PCA-first at 0.086 leakage reduction with label drop (-0.079). The corresponding oracle purities are 0.730 for DMD, 0.509 for raw covariance, and 0.540 for PCA-first. PACS / ResNet-18 shows the same qualitative pattern. DMD attains leakage reduction 0.0516 ± 0.0001 with label change -0.0023 ± 0.0001 and oracle purity 0.9199, whereas



(a) Waterbirds / ResNet-18 (b) Waterbirds / CLIP ViT-B/32

Figure 4: Detailed Waterbirds support sweeps. Support quality strongly predicts oracle-side gains across both backbones, but its relationship with leakage reduction is less uniform.

raw covariance attains 0.0530, -0.0070, and 0.6876, and PCA-first attains 0.0506, -0.0055, and 0.6908. Waterbirds / ResNet-18 is more mixed: DMD yields slightly less leakage reduction than raw covariance (0.0093 vs. 0.0117) but much smaller label damage (-0.0014 vs. -0.0304) and higher oracle purity (0.5077 vs. 0.4352), so the tradeoff is still cleaner even though no single scalar dominates. PACS / CLIP is the main near-tie regime. There, DMD, LDA-first, and raw covariance are all very strong: DMD achieves 0.0644 ± 0.0002 leakage reduction with no measurable label loss and oracle purity 0.9705, while raw covariance achieves 0.0653 leakage reduction, label change -0.0015, and oracle purity 0.9706. We therefore do not claim uniform dominance: paired bootstrap intervals at $q = 8$ show the DMD-versus-raw auditor-reduction difference is unresolved in all six settings, while task accuracy and NLL favor DMD in five of six. The supported claim is narrower: quotient-first residual editing yields a cleaner leakage–label tradeoff than generic second-order alternatives, especially where label damage matters.

For diagnostics, split stability is $1 - \|P_b - P_{b'}\|_F^2/(2q)$, averaged over rerun pairs. In the controlled Waterbirds sweeps, support quality is $\min_{y,e} n_{y,e}$. Because this quantity uses the evaluation-only label E, it is a benchmark diagnostic; its label-only analogues are task retention, per-class size, and rerun stability. Figures 3 and 4 show that support strongly predicts oracle-side gains, especially for CLIP, although its relationship with leakage reduction is less uniform.

Across all six settings, bagging reduces mean stage-two subspace distance to roughly 0.61–0.66 of its single-run value and reduces oracle-purity variability to roughly 0.51–0.63. Bridge-style scalar diagnostics are less consistent, and the Appendix C abstention gate is nearly non-binding. We therefore recommend stage one as an audit or compression anchor and stage two only when residual structure is well supported and stable across reruns.

5.3 Checks beyond frozen-vision setting

We additionally evaluated frozen sentence-encoder, decoder-only language-model, and larger vision–language representations; protocols and confidence intervals are in Appendix C.3. On CivilComments with BGE-large-en-v1.5 and on Waterbirds with OpenCLIP ViT-L/14 representations, both DMD and raw covariance reduce held-out MLP-auditor recoverability at the same rank $q = 8$, while DMD has paired task-accuracy advantages over raw covariance of $0.0539\,[0.0462, 0.0618]$, and $0.0294\,[0.0224, 0.0361]$, respectively. These checks extend the task-protection pattern beyond the original small frozen-vision benchmarks, but they do not establish complete nonlinear erasure or end-to-end language-model editing.

6 Discussion

The paper's main message is deliberately scoped to the label-only setting. In the label-only frozen-feature setting, the observable semantic target is a quotient-level object rather than a uniquely identified semantic/background factorization. That boundary matters methodologically: it suggests protecting an estimated quotient-visible surrogate first and interpreting any subsequent residual edit as a post-quotient operation rather than as recovery of a nuisance variable.

The experiments reinforce that asymmetry. Stage one is consistently strong and often indistinguishable from classical supervised reduction, which supports its role as an audit anchor rather than as a standalone novelty claim. Stage two can be useful, but only in some regimes, and its practical value depends much more on support and split stability than on any single scalar residual proxy.

The limitations are clear. The structured second-order model is representation-level and deliberately restrictive: DMD targets linear subspaces and class-dependent second moments, and may miss leakage encoded only through nonlinear geometry or higher-order moments. Accordingly, reduced recoverability under the tested probe families is not a certificate of complete nuisance removal. Theorems 1–2 apply separately to any fixed frozen checkpoint, but our factorization and perturbation results do not describe how the quotient evolves during fine-tuning or end-to-end optimization. The broader sentence-encoder, decoder-only language-model, and vision–language checks improve the empirical scope, but we have not benchmarked DMD at ImageNet scale or established complete nonlinear erasure; at larger scale, randomized or matrix-free eigensolvers and sketched or streaming moment estimates are natural computational extensions. Stronger selective abstention rules and richer residual families remain future work within the same observable-target discipline.

7 Related work

Our work is related to classical sufficiency, comparisons of experiments, and supervised dimension reduction, including sliced inverse regression, sliced average variance estimation, principal Hessian directions, directional regression, supervised principal components, partial least squares, and information-theoretic label-relevant compression [Blackwell, 1953, Li, 1991, Cook and Weisberg, 1991, Li, 1992, Cook, 1998, Li and Wang, 2007, Cook, 2007, Bair et al., 2006, Tishby et al., 2000]. The second-order SDR literature is closest to our residual stage, but our target differs: after removing a quotient-visible surrogate, we treat the remaining covariance structure as an observable post-quotient residual operation, not as recovery of a uniquely identified latent nuisance factor. Our work is also related to probing and exposing behaviorally relevant structure in frozen representations [Burns et al., 2023, Marks and Tegmark, 2023, El Banani et al., 2024, Bartoszcze et al., 2025, Zhou et al., 2024, 2025b].

On the intervention side, our setting is related to linear concept erasure, adversarial attribute removal, post hoc representation editing, and activation- or representation-level interventions [Bolukbasi et al., 2016, Elazar and Goldberg, 2018, Gonen and Goldberg, 2019, Ravfogel et al., 2020, 2022, Belrose et al., 2023, Meng et al., 2022, 2023, Zou et al., 2023, Turner et al., 2023, Rimsky et al., 2024, Wu et al., 2024, Zou et al., 2024, Nguyen et al., 2026]. It is also adjacent to spurious-correlation, robustness, and invariance methods that use limited side information or observed/inferred environments [Arjovsky et al., 2019, Sagawa et al., 2020, Creager et al., 2021, Lin et al., 2022, Arefin et al., 2024, He et al., 2025, Lu et al., 2025, Zhou et al., 2023, 2025a, Zhou, 2022]. Unlike these lines of work, we assume only (X, Y) and emphasize that labels identify a posterior quotient rather than a full semantic/background decomposition.

8 Conclusion

Labels alone identify a posterior quotient of a frozen representation, not a full semantic/background factorization. This motivates a conservative workflow: preserve a quotient-visible surrogate first, and treat residual editing as a post-quotient operation whose value must be checked rather than assumed. Across our settings, stage one is a strong audit/compression baseline, while stage two helps only in some regimes and is most credible when residual support and split stability are strong. Additional annotations or assumptions may justify stronger semantic recovery; labels alone do not.

Acknowledgements and disclosures

Funding. This work received no third-party funding.

Competing interests. The authors declare no competing interests.

References

K. Ahuja, E. Caballero, D. Zhang, J.-C. Gagnon-Audet, Y. Bengio, I. Mitliagkas, and I. Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. In *Advances in Neural Information Processing Systems*, volume 34, pages 3438–3450, 2021.

M. R. Arefin, Y. Zhang, A. Baratin, F. Locatello, I. Rish, D. Liu, and K. Kawaguchi. Unsupervised concept discovery mitigates spurious correlations. In *Proceedings of the 41st International Conference on Machine Learning*, pages 1672–1688, 2024.

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

E. Bair, T. Hastie, D. Paul, and R. Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137, 2006.

L. Bartoszcze, S. Munshi, B. Sukidi, J. Yen, Z. Yang, D. Williams-King, L. Le, K. Asuzu, and C. Maple. Representation engineering for large-language models: Survey and research challenges. *arXiv preprint arXiv:2502.17601*, 2025.

N. Belrose, D. Schneider-Joseph, S. Ravfogel, R. Cotterell, E. Raff, and S. Biderman. Leace: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*, volume 36, pages 66044–66063, 2023.

D. Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2): 265–272, 1953.

T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357, 2016.

D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 491–500, 2019. doi: 10.1145/3308560.3317593.

C. Burns, H. Ye, D. Klein, and J. Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*, 2023.

M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.

R. D. Cook. *Regression Graphics: Ideas for Studying Regressions through Graphics*. John Wiley & Sons, New York, 1998. ISBN 978-0-471-19365-8.

R. D. Cook. Fisher lecture: Dimension reduction in regression. *Statistical Science*, 22(1):1–26, 2007.

R. D. Cook and S. Weisberg. Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association*, 86(414):328–332, 1991.

E. Creager, J.-H. Jacobsen, and R. Zemel. Environment inference for invariant learning. In *Proceedings of the 38th International Conference on Machine Learning*, pages 2189–2200, 2021.

M. El Banani, A. Raj, K.-K. Maninis, A. Kar, Y. Li, M. Rubinstein, D. Sun, L. Guibas, J. Johnson, and V. Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21795–21806, 2024.

Y. Elazar and Y. Goldberg. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, 2018.

H. Gonen and Y. Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, 2019.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Q. He, K. Xu, and A. Yao. Eva: Erasing spurious correlations with activations. In *International Conference on Learning Representations*, 2025.

B. Li and S. Wang. On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479):997–1008, 2007. doi: 10.1198/016214507000000536.

D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5542–5550, 2017.

K.-C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

K.-C. Li. On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma. *Journal of the American Statistical Association*, 87(420):1025–1039, 1992. doi: 10.1080/01621459.1992.10476258.

Y. Lin, S. Zhu, L. Tan, and P. Cui. Zin: When and how to learn invariance without environment partition? In *Advances in Neural Information Processing Systems*, volume 35, pages 24529–24542, 2022.

Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

S. Lu, J. Chai, and X. Wang. Mitigating spurious correlations in zero-shot multimodal models. In *International Conference on Learning Representations*, 2025.

S. Marks and M. Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372, 2022.

K. Meng, A. S. Sharma, A. J. Andonian, Y. Belinkov, and D. Bau. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*, 2023.

D. V. Nguyen, H. M. Vu, N. Y. Pham, L. Zhang, and T. M. Nguyen. Activation steering with a feedback controller. In *International Conference on Learning Representations*, 2026.

Qwen, A. Yang, B. Yang, B. Zhang, et al. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2024.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763, 2021.

S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, and Y. Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, 2020.

S. Ravfogel, M. Twiton, Y. Goldberg, and R. Cotterell. Linear adversarial concept erasure. *arXiv preprint arXiv:2201.12091*, 2022.

N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of EMNLP-IJCNLP*, pages 3982–3992, 2019.

N. Rimsky, N. Gabrieli, J. Schulz, M. Tong, E. Hubinger, and A. Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, 2024.

S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *International Conference on Learning Representations*, 2020.

N. Tishby, F. C. N. Pereira, and W. Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

A. M. Turner, L. Thiergart, G. Leech, D. Udell, J. J. Vazquez, U. Mini, and M. MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

T. Wen, Z. Wang, Q. Zhang, and Q. Lei. Elastic representation: Mitigating spurious correlations for group robustness. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, pages 541–549, 2025.

Z. Wu, A. Arora, Z. Wang, A. Geiger, D. Jurafsky, C. D. Manning, and C. Potts. Reft: Representation finetuning for language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 63908–63962, 2024.

S. Xiao, Z. Liu, P. Zhang, and N. Muennighoff. C-pack: Packaged resources to advance general chinese embedding, 2023.

Y. Zhou. Rethinking reconstruction autoencoder-based out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7379–7387, 2022. doi: 10.1109/CVPR52688.2022.00723. URL https://doi.org/10.1109/CVPR52688.2022.00723.

Y. Zhou, H.-M. Hu, J. Yu, Z. Xu, W. Lu, and Y. Cao. A solution to co-occurence bias: Attributes disentanglement via mutual information minimization for pedestrian attribute recognition. In E. Elkind, editor, *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 1831–1839. International Joint Conferences on Artificial Intelligence Organization, 8 2023. doi: 10.24963/ijcai.2023/203. URL https://doi.org/10.24963/ijcai.2023/203. Main Track.

Y. Zhou, H.-M. Hu, Y. Xiang, X. Zhang, and H. Wu. Pedestrian attribute recognition as label-balanced multi-label learning. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 61964–61978. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zhou24j.html.

Y. Zhou, H.-M. Hu, J. Yu, H. Wu, S. Pu, and H. Wang. A solution to co-occurrence bias in pedestrian attribute recognition: Theory, algorithms, and improvements. *International Journal of Computer Vision*, 133(7):4712–4726, 2025a. doi: 10.1007/s11263-025-02405-7. URL https://doi.org/10.1007/s11263-025-02405-7.

Y. Zhou, B. Li, H.-M. Hu, X. Zhang, D. Zhang, and H. Wang. Heterogeneous feature re-sampling for balanced pedestrian attribute recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2706–2722, 2025b. doi: 10.1109/TPAMI.2025.3526930. URL https://doi.org/10.1109/TPAMI.2025.3526930.

Y. Zhou, B. Li, H.-M. Hu, H. Wang, X. Zhang, and R. Zhang. Is spurious correlation removal always learnable?, 2026. URL https://arxiv.org/abs/2606.12930.

A. Zou, L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, X. Yin, M. Mazeika, A.-K. Dombrowski, S. Goel, N. Li, M. J. Byun, Z. Wang, A. Mallen, S. Basart, S. Koyejo, D. Song, M. Fredrikson, J. Z. Kolter, and D. Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A. Zou, L. Phan, J. Wang, D. Duenas, M. Lin, M. Andriushchenko, R. Wang, Z. Kolter, M. Fredrikson, and D. Hendrycks. Improving alignment and robustness with circuit breakers. In *Advances in Neural Information Processing Systems*, volume 37, pages 83345–83373, 2024.

A Additional theory and proofs

A.1 Scope of the claims

All claims are made at the level of a frozen learned representation. The structured model

$$X = US + VB + \varepsilon$$

is therefore a model of representation geometry, not a claim that raw inputs admit a unique semantic/background decomposition. The identifiability statements are observational statements about the joint law $P(X, Y)$ in the label-only regime.

A.2 Proof of Theorem 1

Proof. Let $\eta(X) = \mathbb{P}(Y \mid X)$.

First, $\eta(X)$ is label-sufficient. For any bounded measurable function f,

$$\mathbb{E}[f(Y) \mid X] = \mathbb{E}[f(Y) \mid \eta(X)],$$

because the conditional law of Y given X is exactly encoded by $\eta(X)$. Equivalently,

$$Y \perp\!\!\!\perp X \mid \eta(X).$$

Now let $T = t(X)$ be deterministic and label-sufficient. Then

$$Y \perp\!\!\!\perp X \mid T,$$

so the conditional law of Y given X factors through T. For every class y,

$$\mathbb{P}(Y = y \mid X) = \mathbb{P}(Y = y \mid T) \qquad \text{a.s.}$$

Hence there exists a measurable map g such that

$$\eta(X) = g(T) \qquad \text{a.s.}$$

Thus $\eta(X) \preceq T$. Taking the intersection over all deterministic label-sufficient T gives the sigma-field identity, modulo null sets.

For the decision-theoretic statement, define the conditional risk

$$L(a, t) := \sum_{y=1}^{K} t_y \, \ell(a, y), \qquad t \in \Delta^{K-1}.$$

For any decision rule $a(X)$,

$$\mathbb{E}[\ell(a(X), Y) \mid X] = L(a(X), \eta(X)).$$

Because $\mathcal{A}$ is finite, a minimizer of $L(\cdot, \eta(X))$ can be selected measurably using a fixed tie-breaking rule. Thus no rule depending on X can improve on a rule depending only on $\eta(X)$, which gives the equality of Bayes risks. The converse is the deterministic version of Blackwell sufficiency for predicting Y: if T has the same Bayes risk as X for every finite-action bounded label-prediction loss, then the conditional law of Y given X factors through T, hence $\eta(X) \preceq T$. □

A.3 Proof of Theorem 2

Proof. Let $\phi(X)$ be measurable and not measurable with respect to $\sigma(\eta(X))$. Consider two deterministic decompositions of the same observable representation:

$$(S_0, B_0) := (\eta(X), \phi(X)), \qquad (S_1, B_1) := ((\eta(X), \phi(X)), 0).$$

Both are label-valid. The first is label-valid by Theorem 1; the second is a refinement of the first semantic statistic and is therefore also label-sufficient.

The two decompositions induce exactly the same observable joint law $P_{X,Y}$, but they disagree on the role of $\phi(X)$: in the first world $\phi(X)$ is not measurable with respect to $\sigma(S_0) = \sigma(\eta(X))$, whereas in the second world it is measurable with respect to $\sigma(S_1) = \sigma(\eta(X), \phi(X))$. Any label-only procedure has the same distribution under these two observationally equivalent worlds. Therefore, for any decision rule $\widehat{\theta}$, at least one of the two worlds has error probability at least $1/2$. This proves the minimax lower bound. □

Corollary 6 (No informative valid certificate beyond the quotient)**.** *Let $\mathcal{C}$ be a label-only confidence-set procedure based on an i.i.d. sample from $P_{X,Y}$ for the role of $\phi(X)$, with uniform coverage at least $1-\alpha$ over all label-valid decompositions inducing the same $P_{X,Y}$. If $\alpha < 1/2$, then $\mathcal{C}$ must contain both role assignments for some observationally equivalent decomposition. Hence no label-only certificate can be simultaneously uniformly valid and uniformly informative beyond $\sigma(\eta(X))$.*

Proof. Apply the two-world construction in the proof of Theorem 2. If a confidence set excluded one of the two roles in both worlds with probability exceeding α, its coverage would fail in one of the two observationally indistinguishable worlds. Thus any uniformly valid certificate must be non-informative for at least one such pair. □

A.4 Rank limitation of the mean stage

Proposition 7 (Rank limitation)**.** *For discrete $Y \in [K]$,*

$$\operatorname{rank}(\operatorname{Cov}(\mathbb{E}[X \mid Y])) \le K - 1.$$

Proof. The random vector $\mathbb{E}[X \mid Y]$ takes at most K distinct values and therefore lies in an affine subspace of dimension at most $K-1$. Its covariance matrix has rank at most that affine dimension. □

This is why the first stage should be interpreted as a low-dimensional quotient surrogate rather than recovery of every possible semantic direction.

A.5 Proof of Theorem 3

Proof. Under the structured model,

$$X = US + VB + \varepsilon, \qquad U^\top V = 0,$$

with $\mathbb{E}[B \mid Y] = 0$, $\mathbb{E}[\varepsilon \mid Y] = 0$, and class-invariant noise covariance. Hence

$$\mathbb{E}[X \mid Y] = U\mathbb{E}[S \mid Y],$$

and therefore

$$M_Y = \operatorname{Cov}(\mathbb{E}[X \mid Y]) = U \operatorname{Cov}(\mathbb{E}[S \mid Y])\, U^\top = UAU^\top.$$

For the second moment, write

$$\Gamma_y := \operatorname{Cov}(B \mid Y = y), \qquad \bar{\Gamma} := \sum_y \pi_y \Gamma_y.$$

Expanding $\Sigma_y = \operatorname{Cov}(X \mid Y = y)$ gives

$$\Sigma_y = U \operatorname{Cov}(S \mid Y = y) U^\top + V\Gamma_y V^\top + UC_{SB,y}V^\top + VC_{SB,y}^\top U^\top + \Sigma_\varepsilon,$$

where the noise term is class-invariant; if the noise has class-invariant cross-covariances with (S, B), those terms are absorbed into Σ_ε. Projection onto $U^\perp$ removes every term passing through U:

$$\Sigma_{y,\perp} = P_{U^\perp}\Sigma_y P_{U^\perp} = V\Gamma_y V^\top + P_{U^\perp}\Sigma_\varepsilon P_{U^\perp}.$$

Averaging over classes,

$$\bar{\Sigma}_\perp = V\bar{\Gamma}V^\top + P_{U^\perp}\Sigma_\varepsilon P_{U^\perp}.$$

Thus

$$\Sigma_{y,\perp} - \bar{\Sigma}_\perp = V(\Gamma_y - \bar{\Gamma})V^\top,$$

and hence

$$\begin{aligned} T_\perp &= \sum_y \pi_y (\Sigma_{y,\perp} - \bar{\Sigma}_\perp)^2 \\ &= V\left(\sum_y \pi_y (\Gamma_y - \bar{\Gamma})^2\right) V^\top = VGV^\top. \end{aligned}$$

It remains to prove the variational statement. For any rank-q projector P with $PU = 0$,

$$\begin{aligned}\mathcal{J}_U(P) &= \sum_y \pi_y \|(\Sigma_{y,\perp} - \bar{\Sigma}_\perp)P\|_F^2 \\ &= \sum_y \pi_y \operatorname{tr}\big(P(\Sigma_{y,\perp} - \bar{\Sigma}_\perp)^2\big) = \operatorname{tr}(PT_\perp).\end{aligned}$$

The Ky Fan maximum principle gives

$$\max_{\substack{P=P^\top=P^2,\ \operatorname{rank}(P)=q \\ PU=0}} \operatorname{tr}(PT_\perp) = \sum_{j=1}^{q} \lambda_j(T_\perp),$$

attained by the top-q eigenspace $P_\star$ of $T_\perp$ restricted to $U^\perp$.

Finally, if

$$\gamma_q = \lambda_q(T_\perp) - \lambda_{q+1}(T_\perp) > 0,$$

the standard spectral projector curvature inequality gives

$$\operatorname{tr}(P_\star T_\perp) - \operatorname{tr}(PT_\perp) \geq \frac{\gamma_q}{2}\|P - P_\star\|_F^2$$

for every rank-q projector P in $U^\perp$. Since $\mathcal{J}_U(P) = \operatorname{tr}(PT_\perp)$, this proves the claimed curvature bound. □

A.6 Proof of Theorem 4

Proof. Let

$$\Delta_y = UA_yU^\top + UC_yV^\top + VC_y^\top U^\top + VB_yV^\top + R_y.$$

Group the displayed terms into the semantic block $UA_yU^\top$, the cross block $UC_yV^\top + VC_y^\top U^\top$, the residual block $VB_yV^\top$, and the remainder block R_y. Because $U^\top V = 0$, the square of the pure residual block is

$$(VB_yV^\top)^2 = VB_y^2V^\top.$$

Therefore

$$T_{\text{raw}} = \sum_y \pi_y \Delta_y^2 = V\left(\sum_y \pi_y B_y^2\right)V^\top + E_{\text{sem}} = VG_BV^\top + E_{\text{sem}},$$

where E_{sem} collects all terms containing A_y, C_y, or R_y, and all cross-products with those terms.

Using submultiplicativity of the operator norm and $2ab \leq a^2 + b^2$, all products among A_y, C_y, R_y are bounded by a constant multiple of

$$\|A_y\|_{\text{op}}^2 + \|C_y\|_{\text{op}}^2 + \|R_y\|_{\text{op}}^2.$$

The remaining products involving the residual block are bounded by

$$\|B_y\|_{\text{op}}\|C_y\|_{\text{op}} + \|B_y\|_{\text{op}}\|R_y\|_{\text{op}}.$$

Thus, for a universal constant C,

$$\|E_{\text{sem}}\|_{\text{op}} \leq C\sum_y \pi_y\left(\|A_y\|_{\text{op}}^2 + \|C_y\|_{\text{op}}^2 + \|R_y\|_{\text{op}}^2 + \|B_y\|_{\text{op}}\|C_y\|_{\text{op}} + \|B_y\|_{\text{op}}\|R_y\|_{\text{op}}\right).$$

The eigenspace perturbation bound follows from Davis–Kahan applied to

$$T_{\text{raw}} = VG_BV^\top + E_{\text{sem}}.$$

If P_B is the top-q eigenspace of $VG_BV^\top$ and $\widehat{P}_{\text{raw}}$ is the top-q eigenspace of T_{raw}, then define

$$\delta_B := \lambda_q(VG_BV^\top) - \lambda_{q+1}(VG_BV^\top).$$

Whenever $\delta_B > \|E_{\text{sem}}\|_{\text{op}}$, Davis–Kahan gives

$$\|\sin\Theta(\widehat{P}_{\text{raw}}, P_B)\|_{\text{op}} \leq \frac{\|E_{\text{sem}}\|_{\text{op}}}{\delta_B - \|E_{\text{sem}}\|_{\text{op}}}.$$

If the semantic covariance block has larger leading energy than the residual block and cross/remainder terms are small, the leading eigenspace of T_{raw} can lie in $\operatorname{span}(U)$, which proves the final qualitative claim. □

A.7 Proof of Theorem 5

Proof. Let

$$h_n(\delta) := r_{\text{eff}} + \log(K/\delta).$$

Standard effective-rank concentration under the covariance-sub-Gaussian assumption gives, uniformly over classes,

$$\max_y \|\widehat{\Sigma}_y - \Sigma_y\|_{\text{op}} \le CBL^2 \left[\sqrt{\frac{h_n(\delta)}{n_{\min}}} + \frac{h_n(\delta)}{n_{\min}} \right] = CBL^2 \xi_n(\delta).$$

The corresponding class-mean bound is

$$\max_y \|\widehat{\mu}_y - \mu_y\|_2 \le CL \sqrt{\frac{B\, h_n(\delta)}{n_{\min}}}.$$

Since $\max_y \|\mu_y\|_2 \le \sqrt{B}$, expanding the covariance of the estimated class means yields

$$\|\widehat{M}_Y - M_Y\|_{\text{op}} \le CBL^2 \xi_n(\delta).$$

For i.i.d. sampling, the class-proportion error obeys $\|\widehat{\pi} - \pi\|_1 \lesssim \sqrt{(K + \log(1/\delta))/n}$ and is absorbed by the same bound because $n \ge K n_{\min}$; under fixed stratification with $\widehat{\pi} = \pi$, this term vanishes.

Let P_U and $\widehat{P}_U$ be the population and empirical stage-one projectors, and set

$$\epsilon_U := \|\widehat{P}_U - P_U\|_{\text{op}}.$$

For each class,

$$\begin{aligned} &\|P_{\widehat{U}^\perp} \widehat{\Sigma}_y P_{\widehat{U}^\perp} - P_{U^\perp} \Sigma_y P_{U^\perp}\|_{\text{op}} \\ &\quad \le CBL^2 \xi_n(\delta) + CB\epsilon_U. \end{aligned}$$

Since $T_\perp$ is quadratic in centered residual covariances, the identity $A^2 - B^2 = A(A-B) + (A-B)B$ gives

$$\|\widehat{T}_{\widehat{U}^\perp} - T_{U^\perp}\|_{\text{op}} \le CB^2 L^2 \xi_n(\delta) + CB^2 L^4 \xi_n^2(\delta) + CB^2 \epsilon_U.$$

Applying Davis–Kahan to M_Y gives

$$\epsilon_U \le \frac{2CBL^2 \xi_n(\delta)}{\gamma_U}$$

under the stated stage-one gap condition. Applying Davis–Kahan again to the residual operators gives

$$\|\sin\Theta(\widehat{V}, V)\|_{\text{op}} \le \frac{2\left(CB^2 L^2 \xi_n(\delta) + CB^2 L^4 \xi_n^2(\delta) + CB^2 \epsilon_U\right)}{\gamma_V}$$

under the stated residual eigengap condition. □

A.8 Support, stability, and permutation diagnostics

Proposition 8 (Support-limited residual visibility)**.** *Suppose that within each class*

$$X \mid Y = y \sim \sum_{e=1}^{m} \alpha_{y,e} P_{y,e}.$$

Let $\Sigma_{y,e}$ be the covariance of $P_{y,e}$, and assume its centered distribution is covariance-sub-Gaussian with parameter L. Define

$$s_{\min} := \min_{y,e} n_{y,e}, \qquad B_{\text{cell}} := \max_{y,e} \|\Sigma_{y,e}\|_{\text{op}},$$

$$r_{\text{cell}} := \max_{y,e} \frac{\text{tr}(\Sigma_{y,e})}{\|\Sigma_{y,e}\|_{\text{op}}}.$$

Assume $s_{\min} \geq 1$. Conditional on the realized cell counts, with probability at least $1-\delta$,

$$\max_{y,e} \|\widehat{\Sigma}_{y,e} - \Sigma_{y,e}\|_{\mathrm{op}} \leq CL^2 B_{\mathrm{cell}} \left[\sqrt{\frac{r_{\mathrm{cell}} + \log(Km/\delta)}{s_{\min}}} + \frac{r_{\mathrm{cell}} + \log(Km/\delta)}{s_{\min}} \right].$$

Thus poorly supported class–component cells raise the residual covariance noise floor and can destroy the residual eigengap. Moreover, if $\alpha_{\min} := \min_{y,e} \alpha_{y,e}$ and $n_{\min}\alpha_{\min} \gtrsim \log(Km/\delta)$, then

$$s_{\min} \geq \tfrac{1}{2} n_{\min}\alpha_{\min}$$

with high probability under random mixture sampling.

Proof. Conditional on the realized cell counts, apply effective-rank sub-Gaussian covariance concentration to each component cell and take a union bound over the Km cells. This gives the displayed bound with the smallest realized count $s_{\min}$. Under random mixture sampling, a Chernoff bound and a union bound over the cells give $n_{y,e} \geq n_y \alpha_{y,e}/2$ simultaneously whenever $n_{\min}\alpha_{\min} \gtrsim \log(Km/\delta)$, proving the final claim. □

Proposition 9 (Split stability and projector bagging)**.** *Let $P_\star$ be the top-q projector of $T_\perp$ with eigengap $\gamma_V > 0$. Suppose rerun operators satisfy*

$$\widehat{T}_b = T_\perp + E_b, \qquad \|E_b\|_{\mathrm{op}} \leq \gamma_V/4.$$

Let $\widehat{P}_b$ be the top-q projector of $\widehat{T}_b$. Then

$$\|\widehat{P}_b - P_\star\|_F \leq \frac{C\|E_b\|_{\mathrm{op}}}{\gamma_V}.$$

For independent reruns b, b',

$$\mathbb{E}\|\widehat{P}_b - \widehat{P}_{b'}\|_F^2 \leq \frac{C}{\gamma_V^2}\mathbb{E}\|E_b\|_{\mathrm{op}}^2.$$

Moreover, for the averaged projector

$$\bar{P}_B = \frac{1}{B}\sum_{b=1}^{B} \widehat{P}_b,$$

$$\mathbb{E}\|\bar{P}_B - \mathbb{E}\widehat{P}_b\|_F^2 = \frac{1}{B}\mathbb{E}\|\widehat{P}_b - \mathbb{E}\widehat{P}_b\|_F^2.$$

Proof. The first two inequalities are Davis–Kahan perturbation bounds applied to each rerun. The final identity is the standard variance reduction identity for the average of independent projectors. □

Proposition 10 (Permutation-null validity)**.** *Condition on the observed feature vectors $\{x_i\}_{i=1}^n$. Under the null hypothesis $Y \perp\!\!\!\perp X$, the conditional distribution of any eigenvalue statistic computed from $\widehat{M}_Y$, or from the corresponding residual operator, is invariant under label permutations. Therefore an empirical $(1-\alpha)$ permutation envelope controls the conditional type-I error of selecting a spurious supervised rank at level α.*

Proof. Under $Y \perp\!\!\!\perp X$, the labels are conditionally exchangeable given the features. Therefore permuting labels leaves the conditional null distribution of any label-dependent statistic unchanged. The usual permutation-test argument then gives conditional level-α control. □

B Implementation details

B.1 Cached-feature setting

All reported experiments operate on cached frozen representations. The label-only workflow therefore starts from precomputed feature matrices together with task labels and, when available for evaluation only, observed environment labels.

B.2 Feature preprocessing

For all frozen representations, features are first ℓ_2-normalized, then standardized using training-split statistics, and finally reduced by PCA with whitening. The PCA dimension is dataset-specific:

- CelebA: 128
- Waterbirds: 64
- PACS: 128

The same fitted preprocessing transformation is applied to validation and test splits without leakage.

B.3 Ranks and estimators

For the settings highlighted in the main text, we use:

- stage-one rank $r = 1$ for CelebA and Waterbirds,
- stage-one rank $r = 4$ for PACS,
- stage-two removal rank $q = 8$ for the fixed-rank comparisons in the main text.

Editing frontiers vary the removal strength over a small grid of residual ranks.

For PACS, the main experiments use $r = 4$ as a fixed compact quotient surrogate. The permutation-null diagnostic in Appendix Figure 5 often flags a broader PACS mean-stage spectrum, up to approximately $K - 1 = 6$ directions. We use this diagnostic only to visualize the supervised spectrum, not to tune the main operating rank. Thus the fixed $r = 4$ experiments and the selected-rank diagnostic serve different purposes.

B.4 Covariance regularization

Class-conditional covariance estimates use the shrinkage operator

$$\Sigma \mapsto (1 - \alpha)\Sigma + \alpha\tau I, \qquad \tau = \mathrm{tr}(\Sigma)/p,$$

with shrinkage coefficient $\alpha = 0.10$.

B.5 Probes and evaluation

Held-out task accuracy and NLL are computed using balanced multinomial logistic regression after standardization of the representation used by the probe. Environment leakage is the held-out accuracy of a quadratic environment probe: representations are optionally reduced, expanded with a quadratic feature map, standardized, and then fit with the same balanced logistic-regression family. We report $\Delta\mathrm{Leak}$, $\Delta\mathrm{Acc}$, and $\Delta\mathrm{NLL}$ relative to the unedited representation as defined in Section 5. No large hyperparameter sweeps are used in the reported experiments; the emphasis is on comparing representations under a fixed probe family.

Probe implementation details. Unless otherwise noted, all task and environment probes use the same balanced logistic-regression family throughout the paper. The representation passed to each probe is standardized using training-split statistics only. For the quadratic environment probe, we first construct the specified quadratic feature expansion and then fit the same probe family without performing a large hyperparameter sweep. This fixed-probe protocol is used to keep comparisons across representations and editing methods aligned.

B.6 Seeds, reruns, and bagging

Unless stated otherwise, all real-data results are summarized over 10 outer random seeds. Within each outer seed, stability-oriented diagnostics use 8 repeated class-stratified reruns, which also form the basis of the bagging summaries reported in the paper.

B.7 Observed environment labels

Observed environment labels, when available, are used strictly after fitting the label-only workflow: for oracle-side evaluation, group-wise reporting, and controlled support analyses only. They are not used to fit DMD, select the main stage-one rank, or construct the reported stage-two edit subspaces.

B.8 Code, artifacts, and compute resources

Code is publicly available on GitHub under the MIT license. The package contains scripts for preprocessing, fitting DMD and baseline methods, reproducing the main tables and figures, and running the repeated-rerun diagnostics reported in the paper. All experiments use public benchmark datasets and standard frozen backbones as described in the main text. The repository includes instructions for obtaining the datasets, extracting or loading frozen features, and reproducing the evaluation pipeline used in the paper.

Experiments were run on cached frozen representations rather than by end-to-end finetuning the pretrained backbone. A typical run used a single NVIDIA V100 GPU with approximately 32GB GPU memory. The full 10-seed evaluation for a single dataset/backbone setting required approximately 30 minutes, and the complete experimental sweep required approximately 6 hours. We report these values to indicate the order of compute needed for reproduction. The current implementation forms dense covariance operators after the stated PCA preprocessing; for substantially larger feature dimensions, the same objectives can be implemented with randomized or matrix-free eigensolvers and sketched or streaming covariance estimates, which we have not benchmarked at ImageNet scale.

C Additional experimental results

This appendix complements the main-text results in three ways. First, it reports the full multi-metric stage-one and stage-two tables underlying the condensed main-text summaries. Second, it provides the rank-diagnostic and synthetic illustrations that support the quotient-first interpretation. Third, it reports rank, shrinkage, and probe sensitivity summaries that complement the fixed-rank main-text comparison and help interpret the regime dependence of stage two.

Unless stated otherwise, all real-data values below are reported as mean $\pm$ standard deviation over 10 outer seeds. Within each outer seed, stability-oriented diagnostics are computed from repeated class-stratified reruns.

C.1 Full stage-one semantic-auditing results

Table 3 reports the full stage-one results, including held-out accuracy, NLL, and ECE, for DMD-U, LDA-U, PLS-U, PCA-U, and random low-rank projections. These numbers reinforce the main-text picture: stage one is consistently strong and essentially tied with LDA-U across all six dataset/backbone settings.

The most striking gaps remain those against unsupervised or random alternatives. On Waterbirds / ResNet-18, DMD-U reaches 0.7286 ± 0.0002 accuracy versus 0.4408 for PCA-U and 0.5220 ± 0.0429 for Rand-U. On Waterbirds / CLIP, DMD-U reaches 0.7979 ± 0.0002 versus 0.5466 for PCA-U and 0.5447 ± 0.0486 for Rand-U. On PACS / CLIP, DMD-U reaches 0.8603 ± 0.0005, while PCA-U collapses to 0.1754 and Rand-U reaches only 0.2397 ± 0.0269.

These results also clarify the role of the first stage. In several settings, LDA-U and PLS-U are numerically indistinguishable from DMD-U, and on PACS / CLIP, PLS-U is slightly better than DMD-U on both accuracy and NLL. This is fully compatible with the paper's interpretation: the role of stage one is to provide a compact quotient-visible semantic anchor, not to claim a uniquely superior first-stage estimator in every finite-class linear regime.

C.2 Full fixed-rank residual-editing summary

Table 4 reports the full fixed-rank residual-editing summary at $q = 8$, including DMD, LDA-first, PLS-first, raw covariance editing, PCA-first, and random-first. This table is the most direct numerical companion to the main-text stage-two claims.

Table 3: Full stage-one auditing results. Values are mean ± std over 10 outer seeds. DMD-U is consistently strong and essentially tied with LDA-U; PLS-U behaves similarly and is slightly stronger on PACS/CLIP. Standard deviations smaller than 10^{-3} are shown as < 0.001 when this convention is used.

Setting	Method	Accuracy	NLL	ECE
CelebA / ResNet-18	Full	$0.907 \pm < 0.001$	$0.236 \pm < 0.001$	$0.004 \pm < 0.001$
	DMD-U	$0.906 \pm < 0.001$	$0.255 \pm < 0.001$	$0.014 \pm < 0.001$
	LDA-U	$0.906 \pm < 0.001$	$0.255 \pm < 0.001$	$0.014 \pm < 0.001$
	PLS-U	$0.906 \pm < 0.001$	$0.255 \pm < 0.001$	$0.014 \pm < 0.001$
	PCA-U	$0.521 \pm < 0.001$	$0.692 \pm < 0.001$	$0.002 \pm < 0.001$
	Rand-U	0.532 ± 0.017	0.690 ± 0.004	0.009 ± 0.005
Waterbirds / ResNet-18	Full	$0.755 \pm < 0.001$	$0.640 \pm < 0.001$	$0.137 \pm < 0.001$
	DMD-U	$0.729 \pm < 0.001$	$0.680 \pm < 0.001$	$0.151 \pm < 0.001$
	LDA-U	$0.729 \pm < 0.001$	$0.680 \pm < 0.001$	$0.151 \pm < 0.001$
	PLS-U	$0.729 \pm < 0.001$	$0.680 \pm < 0.001$	$0.151 \pm < 0.001$
	PCA-U	$0.441 \pm < 0.001$	$0.718 \pm < 0.001$	$0.109 \pm < 0.001$
	Rand-U	0.522 ± 0.043	0.691 ± 0.009	0.035 ± 0.023
PACS / ResNet-18	Full	$0.859 \pm < 0.001$	$0.395 \pm < 0.001$	$0.010 \pm < 0.001$
	DMD-U	$0.724 \pm < 0.001$	$0.670 \pm < 0.001$	$0.010 \pm < 0.001$
	LDA-U	$0.725 \pm < 0.001$	$0.670 \pm < 0.001$	$0.010 \pm < 0.001$
	PLS-U	$0.725 \pm < 0.001$	$0.670 \pm < 0.001$	$0.011 \pm < 0.001$
	PCA-U	$0.202 \pm < 0.001$	$1.908 \pm < 0.001$	$0.002 \pm < 0.001$
	Rand-U	0.214 ± 0.021	1.891 ± 0.018	0.017 ± 0.005
CelebA / CLIP ViT-B/32	Full	$0.926 \pm < 0.001$	$0.183 \pm < 0.001$	$0.016 \pm < 0.001$
	DMD-U	$0.929 \pm < 0.001$	$0.205 \pm < 0.001$	$0.012 \pm < 0.001$
	LDA-U	$0.929 \pm < 0.001$	$0.205 \pm < 0.001$	$0.012 \pm < 0.001$
	PLS-U	$0.929 \pm < 0.001$	$0.205 \pm < 0.001$	$0.012 \pm < 0.001$
	PCA-U	$0.516 \pm < 0.001$	$0.693 \pm < 0.001$	$0.008 \pm < 0.001$
	Rand-U	0.530 ± 0.020	0.690 ± 0.003	0.012 ± 0.007
Waterbirds / CLIP ViT-B/32	Full	$0.808 \pm < 0.001$	$0.457 \pm < 0.001$	$0.070 \pm < 0.001$
	DMD-U	$0.798 \pm < 0.001$	$0.470 \pm < 0.001$	$0.065 \pm < 0.001$
	LDA-U	$0.798 \pm < 0.001$	$0.470 \pm < 0.001$	$0.065 \pm < 0.001$
	PLS-U	$0.798 \pm < 0.001$	$0.470 \pm < 0.001$	$0.065 \pm < 0.001$
	PCA-U	$0.547 \pm < 0.001$	$0.687 \pm < 0.001$	$0.007 \pm < 0.001$
	Rand-U	0.545 ± 0.049	0.684 ± 0.014	0.030 ± 0.021
PACS / CLIP ViT-B/32	Full	$0.968 \pm < 0.001$	$0.101 \pm < 0.001$	$0.011 \pm < 0.001$
	DMD-U	$0.860 \pm < 0.001$	$0.384 \pm < 0.001$	0.030 ± 0.001
	LDA-U	$0.860 \pm < 0.001$	$0.384 \pm < 0.001$	$0.030 \pm < 0.001$
	PLS-U	$0.870 \pm < 0.001$	$0.359 \pm < 0.001$	$0.029 \pm < 0.001$
	PCA-U	$0.175 \pm < 0.001$	$1.924 \pm < 0.001$	$0.014 \pm < 0.001$
	Rand-U	0.240 ± 0.027	1.866 ± 0.025	0.030 ± 0.015

The clearest advantages of quotient-first residualization remain on CelebA and Waterbirds with CLIP features. On CelebA / CLIP, DMD achieves environment-leakage reduction 0.0326 with label change $+0.0004$ and oracle purity 0.8314, while raw covariance achieves 0.0305, -0.0264, and 0.2623, and PCA-first achieves 0.0314, -0.0259, and 0.2628. On Waterbirds / CLIP, DMD achieves 0.0854 ± 0.0001 leakage reduction with a slight label improvement $(+0.0052 \pm 0.0001)$ and oracle purity 0.7303 ± 0.0002, while raw covariance attains 0.0765, -0.0884, and 0.5086, and PCA-first attains 0.0860, -0.0792, and 0.5402.

PACS / ResNet-18 shows the same qualitative pattern: DMD attains oracle purity 0.9199 at label change -0.0023 ± 0.0001, while raw covariance and PCA-first are both around 0.69 oracle purity with larger label damage. Waterbirds / ResNet-18 is more mixed but still favorable to quotient-first editing as a tradeoff: DMD has slightly less leakage reduction than raw covariance (0.0093 vs. 0.0117), but much smaller label loss (-0.0014 vs. -0.0304) and higher oracle purity (0.5077 vs. 0.4352). PACS / CLIP remains the clearest near-tie regime, where DMD, LDA-first, PLS-first, and raw covariance are all extremely strong and nearly indistinguishable on several metrics.

Table 4: Full fixed-rank residual-editing summary at $q = 8$. Values are mean $\pm$ std over 10 outer seeds. DMD, LDA-first, and PLS-first are nearly identical in most settings; raw covariance and PCA-first are often competitive in leakage reduction but typically incur substantially larger label damage in the regimes highlighted in the main text. Abbr.: R18 = ResNet-18; ViT = CLIP ViT-B/32. Standard deviations smaller than 10^{-3} are shown as < 0.001 when this convention is used.

Setting	Method	ΔLeak	ΔAcc	ΔNLL	Oracle purity
CelebA/R18	DMD	0.013 ± 0.001	-0.002 ± 0.001	0.010 ± 0.001	0.908 ± 0.001
	LDA-first	0.013 ± 0.001	-0.002 ± 0.001	0.010 ± 0.001	0.908 ± 0.001
	PLS-first	0.013 ± 0.001	-0.002 ± 0.001	0.010 ± 0.001	0.908 ± 0.001
	Raw cov.	0.013 ± 0.001	-0.045 ± 0.001	0.085 ± 0.001	0.297 ± 0.001
	PCA-first	0.012 ± 0.001	-0.049 ± 0.001	0.093 ± 0.001	0.281 ± 0.001
	Rand-first	0.012 ± 0.002	-0.045 ± 0.001	0.084 ± 0.003	0.297 ± 0.006
Waterbirds/R18	DMD	0.009 ± 0.001	-0.001 ± 0.001	0.004 ± 0.001	0.508 ± 0.001
	LDA-first	0.009 ± 0.001	-0.002 ± 0.001	0.004 ± 0.001	0.508 ± 0.001
	PLS-first	0.009 ± 0.001	-0.002 ± 0.001	0.004 ± 0.001	0.508 ± 0.001
	Raw cov.	0.012 ± 0.001	-0.030 ± 0.001	0.056 ± 0.001	0.435 ± 0.001
	PCA-first	0.003 ± 0.001	-0.025 ± 0.001	0.040 ± 0.001	0.428 ± 0.001
	Rand-first	0.012 ± 0.004	-0.030 ± 0.008	0.057 ± 0.021	0.436 ± 0.031
PACS/R18	DMD	0.052 ± 0.001	-0.002 ± 0.001	0.012 ± 0.001	0.920 ± 0.001
	LDA-first	0.052 ± 0.001	-0.002 ± 0.001	0.012 ± 0.001	0.920 ± 0.001
	PLS-first	0.052 ± 0.001	-0.002 ± 0.001	0.012 ± 0.001	0.920 ± 0.001
	Raw cov.	0.053 ± 0.001	-0.007 ± 0.001	0.026 ± 0.001	0.688 ± 0.001
	PCA-first	0.051 ± 0.001	-0.006 ± 0.001	0.023 ± 0.001	0.691 ± 0.001
	Rand-first	0.049 ± 0.004	-0.007 ± 0.001	0.025 ± 0.002	0.690 ± 0.018
CelebA/ViT	DMD	0.033 ± 0.001	$+0.000 \pm 0.001$	0.004 ± 0.001	0.831 ± 0.001
	LDA-first	0.033 ± 0.001	$+0.000 \pm 0.001$	0.004 ± 0.001	0.831 ± 0.001
	PLS-first	0.033 ± 0.001	$+0.000 \pm 0.001$	0.004 ± 0.001	0.831 ± 0.001
	Raw cov.	0.031 ± 0.001	-0.026 ± 0.001	0.072 ± 0.001	0.262 ± 0.001
	PCA-first	0.031 ± 0.001	-0.026 ± 0.001	0.071 ± 0.001	0.263 ± 0.001
	Rand-first	0.031 ± 0.001	-0.025 ± 0.001	0.069 ± 0.003	0.272 ± 0.011
Waterbirds/ViT	DMD	0.085 ± 0.001	$+0.005 \pm 0.001$	-0.007 ± 0.001	0.730 ± 0.001
	LDA-first	0.085 ± 0.001	$+0.005 \pm 0.001$	-0.007 ± 0.001	0.730 ± 0.001
	PLS-first	0.085 ± 0.001	$+0.005 \pm 0.001$	-0.007 ± 0.001	0.730 ± 0.001
	Raw cov.	0.077 ± 0.001	-0.088 ± 0.001	0.184 ± 0.001	0.509 ± 0.001
	PCA-first	0.086 ± 0.001	-0.079 ± 0.001	0.169 ± 0.001	0.540 ± 0.001
	Rand-first	0.082 ± 0.006	-0.083 ± 0.010	0.173 ± 0.021	0.527 ± 0.028
PACS/ViT	DMD	0.064 ± 0.001	0.000 ± 0.001	0.002 ± 0.001	0.971 ± 0.001
	LDA-first	0.064 ± 0.001	0.000 ± 0.001	0.002 ± 0.001	0.971 ± 0.001
	PLS-first	0.065 ± 0.001	0.000 ± 0.001	0.002 ± 0.001	0.971 ± 0.001
	Raw cov.	0.065 ± 0.001	-0.002 ± 0.001	0.002 ± 0.001	0.971 ± 0.001
	PCA-first	0.055 ± 0.001	0.000 ± 0.001	0.002 ± 0.001	0.969 ± 0.001
	Rand-first	0.061 ± 0.004	-0.001 ± 0.001	0.003 ± 0.001	0.961 ± 0.012

Table 5: Checks on broader frozen representations. Auditor reduction is the decrease in held-out one-hidden-layer MLP balanced accuracy for predicting the evaluation-only attribute E; task change is edited-minus-unedited accuracy, so larger is better; the final column is the paired DMD-minus-raw task-accuracy advantage.

Setting / frozen representation	DMD auditor reduction	Raw auditor reduction	DMD task ΔAcc	Raw task ΔAcc	Paired DMD task advantage
CivilComments / all-mpnet-base-v2	$0.0515\,[0.0407, 0.0631]$	$0.0643\,[0.0498, 0.0786]$	$-0.0010\,[-0.0042, 0.0022]$	$-0.0625\,[-0.0703, -0.0539]$	$+0.0615\,[0.0522, 0.0697]$
CivilComments / Qwen2.5-1.5B	$0.0331\,[0.0239, 0.0426]$	$0.0323\,[0.0241, 0.0398]$	$+0.0001\,[-0.0026, 0.0031]$	$-0.0563\,[-0.0640, -0.0490]$	$+0.0564\,[0.0487, 0.0647]$
CivilComments / BGE-large-en-v1.5	$0.0537\,[0.0414, 0.0668]$	$0.0455\,[0.0361, 0.0543]$	$-0.0023\,[-0.0054, 0.0011]$	$-0.0562\,[-0.0636, -0.0491]$	$+0.0539\,[0.0462, 0.0618]$
Waterbirds / OpenCLIP ViT-L/14	$0.0436\,[0.0303, 0.0556]$	$0.0331\,[0.0204, 0.0447]$	$-0.0059\,[-0.0100, -0.0017]$	$-0.0353\,[-0.0419, -0.0288]$	$+0.0294\,[0.0224, 0.0361]$

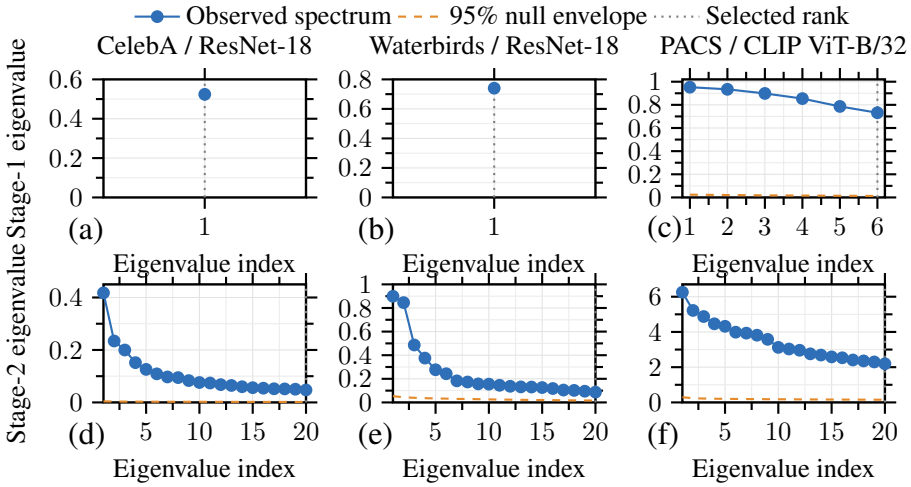


Figure 5: Permutation-null rank diagnostics. Representative observed spectra of the stage-one and stage-two operators together with permutation-null envelopes. Panels A and B contain a single stage-one eigenvalue because CelebA and Waterbirds are binary-label tasks, so the mean-stage rank is at most $K - 1 = 1$. Panel C shows the broader multiclass PACS stage-one spectrum. The second row shows the broader residual-stage spectra.

Taken together, Table 4 supports the narrower empirical conclusion emphasized in the main paper: quotient-first residual editing often yields a cleaner leakage–label tradeoff than raw covariance or projection-matched generic baselines, but the advantage is not uniform across every dataset/backbone combination.

C.3 Broader frozen-backbone checks

For CivilComments [Borkan et al., 2019], the task label is thresholded toxicity and the evaluation-only attribute is a lexical identity-mention indicator, not demographic identity. Frozen representations are extracted from all-mpnet-base-v2 [Reimers and Gurevych, 2019], Qwen2.5-1.5B [Qwen et al., 2024], and BGE-large-en-v1.5 [Xiao et al., 2023]. We use fixed cohorts of 40,000 training and 10,000 test examples, removal rank $q = 8$, quotient rank $r = 1$, and covariance shrinkage $\alpha = 0.10$. Auditor reduction is the unedited-minus-edited balanced accuracy of a held-out one-hidden-layer MLP predicting the evaluation-only attribute. For Waterbirds, we additionally evaluate frozen OpenCLIP ViT-L/14 representations [Cherti et al., 2023]. The Waterbirds/OpenCLIP check follows the same fixed-rank protocol as the original vision experiments. Brackets report paired 95% confidence intervals from 1,000 stratified two-way bootstrap draws over fitting seeds and held-out examples within joint (Y, E) cells; all methods share each draw, the cohorts are fixed, and bagging reruns are not treated as independent observations.

These checks support a narrower claim than complete attribute erasure: DMD substantially improves task retention relative to raw covariance at the same deletion rank, while the paired DMD-versus-raw auditor-reduction difference is not statistically significant in these settings, and none of these probes certifies removal of all nonlinear or higher-order attribute information.

C.4 Permutation-null rank diagnostics

Figure 5 visualizes the rank diagnostics behind the heuristic stage-one and stage-two choices used in the paper. The stage-one picture is especially clean. On CelebA / ResNet-18 and Waterbirds / ResNet-18, the observed leading eigenvalue of $\widehat{M}_Y$ is far above the permutation-null envelope and the selected rank is exactly one, which matches the binary-task structure. On PACS, the stage-one operator is broader, with selected rank around six in both ResNet-18 and CLIP features, again consistent with the multiclass label structure.

The stage-two picture is qualitatively different. The residual operator $\widehat{T}_\perp$ is broad across many more indices, and the permutation-null selected rank often saturates the maximum rank examined in our current diagnostic grid. We do *not* interpret this as evidence that a very large removal rank should be used by default. Instead, it reinforces the practical point made in the main text: the residual stage is typically less sharply low-rank than the mean stage, which is one reason fixed-rank comparisons and stability diagnostics are more informative than a single automatic residual-rank rule.

This figure is best read as supporting evidence for the asymmetry between stages: stage one often looks sharply low-rank and stable, while stage two is broader and more regime-dependent.

C.5 Synthetic illustrations of the observable-target view

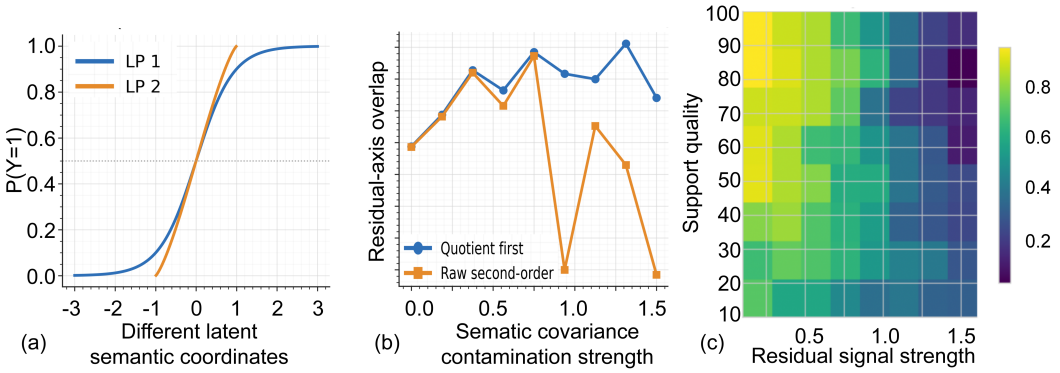


Figure 6: Synthetic illustrations of the observable-target theory. **Left:** different latent parameterizations can induce the same posterior quotient. **Middle:** quotient-first residualization remains cleaner than raw second-order editing under semantic covariance contamination. **Right:** stage-two gains are largest when residual signal and support are both strong.

Figure 6 provides three small synthetic illustrations of the paper's conceptual claims.

The left panel illustrates quotient-level invariance. Two different latent parameterizations induce the same posterior curve over labels, even though the latent coordinates themselves are not the same. This illustrates the quotient interpretation: labels determine posterior classes, not a unique latent semantic axis.

The middle panel illustrates why quotient-first residualization is structurally helpful under semantic covariance contamination. As contamination strength increases, the raw second-order direction can degrade substantially, while quotient-first residualization remains cleaner over a broad range of contamination levels. In our current synthetic construction, the quotient-first overlap is consistently at least competitive with the raw second-order overlap and often materially better.

The right panel gives a coarse phase-style view over residual signal strength and support quality. As expected, the stage-two gain depends on both ingredients. The pattern is not perfectly monotone at every grid point, so we interpret the plot as an illustration rather than as evidence for a universal phase transition. The main message is narrower and consistent with the real-data results: strong support and sufficiently exposed residual structure are both important for useful stage-two behavior.

C.6 Rank, shrinkage, and probe sensitivity

We evaluated removal-rank frontiers over $q \in \{1, 2, 4, 8, 16, 32\}$, together with protected-rank and covariance-shrinkage stress tests. For CelebA and Waterbirds we use $r \in \{1, 2, 4\}$, for PACS we use

Table 6: Sensitivity-analysis summary. The accuracy budget is $\Delta\mathrm{Acc} \geq -0.005$. Auditor reduction refers to held-out recoverability of the evaluation-only attribute.

Analysis	Result
Removal-rank frontier	$q \in \{1, 2, 4, 8, 16, 32\}$ for all six original vision settings.
DMD under the accuracy budget	Every vision setting has at least one tested $q > 0$ satisfying $\Delta\mathrm{Acc} \geq -0.005$.
Raw covariance and PCA-first under the same budget	No tested $q > 0$ satisfies the budget on either CelebA backbone.
Protected-rank/shrinkage stress test at $q = 8$	$r \in \{1, 2, 4\}$ for CelebA/Waterbirds, $r \in \{2, 4, 6\}$ for PACS, and $\alpha \in \{0, 0.05, 0.10, 0.20, 0.50\}$.
Quadratic-auditor reduction	Positive mean reduction in all $90 = 6 \times 15$ setting–configuration pairs.
Seed–configuration runs satisfying the accuracy budget	$150/150$ in four settings; $148/150$ on Waterbirds/ResNet-18; $104/150$ on PACS/ResNet-18.
Probe-family check	Logistic, quadratic, and one-hidden-layer MLP auditors show positive point-estimate reduction in all $18 = 6 \times 3$ checks.
DMD versus raw auditor reduction at $q = 8$	Paired 95% confidence intervals contain zero in all six settings.
DMD versus raw task accuracy and NLL at $q = 8$	Paired confidence intervals favor DMD in five of six settings; PACS/CLIP is unresolved.

Table 7: Support-sweep summary on Waterbirds. Support quality strongly predicts oracle-side gains across 10 seeds, especially on CLIP, but its relationship with leakage reduction is less uniform.

Metric	ResNet-18	CLIP ViT-B/32
$\mathrm{corr}(\text{support}, \text{oracle gain})$	0.748 ± 0.064	0.946 ± 0.013
$\mathrm{corr}(\text{support}, \text{leak reduction})$	0.738 ± 0.071	-0.525 ± 0.148
Top-half support minus bottom-half oracle gain	0.061 ± 0.016	0.311 ± 0.022
Top-half support minus bottom-half leak reduction	0.021 ± 0.007	-0.012 ± 0.004
Controlled total per class	556	689

$r \in \{2, 4, 6\}$, and in all cases $\alpha \in \{0, 0.05, 0.10, 0.20, 0.50\}$. For the binary tasks, $r > 1$ deliberately stress-tests over-protection beyond the single mean-active direction. Auditor reduction is measured under logistic, quadratic, and one-hidden-layer MLP probe families.

The supported advantage is safer use of the same deletion budget rather than uniformly stronger measured erasure: across the tested ranks and configurations, quotient-first editing more often preserves the task while still reducing tested attribute recoverability.

C.7 Detailed Waterbirds support sweeps

In these sweeps, support quality is the minimum class-environment cell count, $\min_{y,e} n_{y,e}$, after the controlled subsampling used to vary support. Figure 4 and Table 7 report the detailed Waterbirds support sweeps. On both backbones, increasing support quality improves oracle-side gains substantially. The effect is especially strong on CLIP, where the oracle-purity gain against raw covariance grows markedly as support increases, matching the very high aggregate correlation $\mathrm{corr}(\text{support}, \text{oracle gain}) = 0.946 \pm 0.013$ reported in Table 7. The ResNet-18 sweep is noisier but still strongly positive, with corresponding correlation 0.748 ± 0.064. By contrast, the leak-reduction relationship is not monotone across backbones. On ResNet-18, better support tends to help both oracle-side gains and leakage reduction; on CLIP, better support helps oracle-side gains dramatically while leakage reduction itself is less aligned and can even weaken. This is why the main text deliberately phrases support quality as a predictor of whether useful residual structure is being exposed, not as a universal scalar score for every evaluation metric.

C.8 Diagnostics beyond support and stability

We use the following compact diagnostic definitions. The eigengap ratio is λ_q/λ_{q+1} for the residual operator at the evaluated rank. Mass above null is the fraction of residual spectral mass above the 95% permutation-null envelope. The cross-setting support score is a normalized version of $\min_{y,e} n_{y,e}$, scaled so that larger values indicate better class-environment coverage. Split stability is the rerun-pair average of $1 - \|P_b - P_{b'}\|_F^2/(2q)$. In bridge diagnostics, P denotes the scalar proxy score used by

Table 8: DMD residual diagnostics across settings. These are the stage-two diagnostics reported for DMD after quotient-first residualization. Split stability and support quality are the most interpretable diagnostics in our experiments; eigengap and spectral-mass summaries are included for completeness. Standard deviations smaller than 10^{-3} are shown as < 0.001 when this convention is used.

Setting	Eigengap ratio	Mass above null	Split stability	Support quality
CelebA / ResNet-18	1.318 ± 0.000	0.236 ± 0.000	0.996 ± 0.001	0.859 ± 0.000
Waterbirds / ResNet-18	1.295 ± 0.000	0.347 ± 0.000	0.970 ± 0.005	0.967 ± 0.003
PACS / ResNet-18	1.190 ± 0.000	0.146 ± 0.000	0.978 ± 0.005	0.846 ± 0.003
CelebA / CLIP ViT-B/32	1.185 ± 0.000	0.134 ± 0.000	0.995 ± 0.002	0.760 ± 0.045
Waterbirds / CLIP ViT-B/32	1.070 ± 0.000	0.309 ± 0.000	0.846 ± 0.047	0.934 ± 0.015
PACS / CLIP ViT-B/32	1.094 ± 0.000	0.112 ± 0.000	0.885 ± 0.077	0.857 ± 0.008

the bridge diagnostic, λ is the associated residual eigenvalue, OG is oracle gain, and LR is leakage reduction. In the abstention table, Pass is the gate pass rate, Harm-A/Harm-G are the always-on/gated rates of failing the diagnostic success condition, and q_{eval} is the evaluation removal rank.

Table 9: Bagging and stability summary. Bagging consistently reduces stage-one and stage-two subspace instability. The table reports mean reduction ratios together with mean pairwise single-run and bagged subspace distances. CA = CelebA, WB = Waterbirds. U-red/V-red = reduction ratio for U/V; PU-S and PU-B = pairwise U distances for single and bagged runs; PV-S and PV-B are the corresponding V distances. R18/ViT denote ResNet-18 and CLIP ViT-B/32, respectively. Standard deviations smaller than 10^{-3} are shown as < 0.001 when this convention is used.

Setting	U-red	V-red	PU-S	PU-B	PV-S	PV-B
CA/R18	0.624 ± 0.032	0.627 ± 0.051	$< 0.001 \pm < 0.001$	$< 0.001 \pm < 0.001$	$< 0.001 \pm < 0.001$	$< 0.001 \pm < 0.001$
WB/R18	0.636 ± 0.021	0.611 ± 0.039	$0.005 \pm < 0.001$	$0.002 \pm < 0.001$	$0.004 \pm < 0.001$	$0.001 \pm < 0.001$
PACS/R18	0.649 ± 0.027	0.658 ± 0.041	0.007 ± 0.001	$0.003 \pm < 0.001$	$0.003 \pm < 0.001$	$0.001 \pm < 0.001$
CA/ViT	0.651 ± 0.021	0.656 ± 0.065	$< 0.001 \pm < 0.001$	$< 0.001 \pm < 0.001$	$< 0.001 \pm < 0.001$	$< 0.001 \pm < 0.001$
WB/ViT	0.645 ± 0.023	0.642 ± 0.052	$0.005 \pm < 0.001$	$0.002 \pm < 0.001$	0.004 ± 0.001	$0.001 \pm < 0.001$
PACS/ViT	0.633 ± 0.040	0.614 ± 0.088	0.006 ± 0.001	$0.002 \pm < 0.001$	$0.002 \pm < 0.001$	$0.001 \pm < 0.001$

Setting	Label-std reduction ratio	Purity-std reduction ratio
CelebA / ResNet-18	0.3151 ± 0.4509	0.5542 ± 0.2424
Waterbirds / ResNet-18	0.1552 ± 0.2958	0.6276 ± 0.1987
PACS / ResNet-18	0.1162 ± 0.2320	0.5141 ± 0.1669
CelebA / CLIP ViT-B/32	0.1000 ± 0.3162	0.6010 ± 0.2071
Waterbirds / CLIP ViT-B/32	0.3064 ± 0.1468	0.6196 ± 0.1054
PACS / CLIP ViT-B/32	0.6553 ± 0.4548	0.6176 ± 0.1551

The remaining appendix tables complete the empirical picture.

Table 8 reports the DMD-side residual diagnostics underlying the discussion in Section 5. These values are often sensible within a given setting: for example, support quality is highest on Waterbirds / ResNet-18, while split stability is lower on Waterbirds / CLIP and PACS / CLIP than in most ResNet-18 settings. However, the same table also shows why we do not treat eigengap ratios or spectral-mass-above-null scores as universal trust scores. Across settings, these quantities vary far less consistently than support quality and split stability, even when practical stage-two usefulness differs substantially.

Table 9 shows that bagging consistently reduces both stage-one and stage-two subspace instability across all six settings. The reduction is especially visible for the stage-two subspaces, where the V-mean reduction ratio is typically around 0.61–0.66. This supports the main-text recommendation to inspect split stability before placing strong trust in stage-two edits.

Table 10 and the failure-case summaries show that bridge-style scalar proxies are mixed. In some settings, proxy-side gains correlate moderately with oracle-side gains; in others, the relationships are weak or unstable. This reinforces the paper's practical conclusion that support and stability are the more reliable trust indicators in the label-only settings we study.

Finally, Table 11 confirms that the current abstention gate is largely non-binding. Its pass rate is exactly one in most settings and only slightly below one on PACS / CLIP. We therefore treat abstention as a direction for future work rather than as a main empirical result.

Table 10: Bridge-style scalar diagnostics are mixed. Correlations are reported over 10 outer seeds. No single scalar proxy behaves as a universal trust score across all settings. Abbr.: ρ = correlation; P = proxy; OG = oracle gain; LR = leak reduction; R18 = ResNet-18; B32 = CLIP ViT-B/32.

Setting	$\rho(P, \mathrm{OG})$	$\rho(P, \mathrm{LR})$	$\rho(\log \lambda, \mathrm{OG})$	$\rho(\log \lambda, \mathrm{LR})$
CA / R18	0.417 ± 0.276	-0.105 ± 0.332	-0.173 ± 0.403	-0.108 ± 0.325
WB / R18	0.441 ± 0.439	0.281 ± 0.355	-0.127 ± 0.414	-0.146 ± 0.444
PACS / R18	0.285 ± 0.309	0.211 ± 0.307	-0.126 ± 0.370	0.003 ± 0.335
CA / ViT	0.563 ± 0.248	-0.012 ± 0.301	-0.109 ± 0.517	-0.025 ± 0.313
WB / ViT	-0.191 ± 0.336	0.190 ± 0.315	-0.071 ± 0.348	-0.004 ± 0.354
PACS / ViT	0.285 ± 0.330	0.032 ± 0.335	0.013 ± 0.443	0.025 ± 0.478

Table 11: Abstention summary. The current gate is nearly non-binding: its pass rate is exactly 1.0 in five of the six settings and 0.985 ± 0.024 on PACS/CLIP. We therefore treat abstention as a promising direction rather than a mature contribution. Abbr.: DS = dataset; BB = backbone; R18 = ResNet-18; V32 = CLIP ViT-B/32; Pass = gate pass rate; Harm-A/Harm-G = always-on/gated rates of failing the diagnostic success condition; OG-A = always-on mean oracle gain vs. raw; LR-A = always-on mean leak reduction. Standard deviations smaller than 10^{-3} are shown as < 0.001 when this convention is used. In this table, values are computed over the abstention/gating diagnostic pool rather than the fixed full-training $q = 8$ operating point in Table 4. In particular, the abstention protocol uses subsampled fits and $q_{\mathrm{eval}} = 3$. LR-A is the always-on leakage reduction relative to the unedited full-representation reference, whereas OG-A is the oracle-purity gain relative to raw covariance within the same diagnostic protocol. Thus these values are not directly comparable to the fixed-rank ΔLeak and ΔAcc values in Table 4.

DS	BB	Rates			Always-on mean	
		Pass	Harm-A	Harm-G	OG-A	LR-A
CelebA	R18	1.000 ± 0.000	0.740 ± 0.021	0.740 ± 0.021	0.683 ± 0.007	-0.018 ± 0.002
Waterbirds	R18	1.000 ± 0.000	0.095 ± 0.060	0.095 ± 0.060	0.035 ± 0.007	0.047 ± 0.007
PACS	R18	1.000 ± 0.000	0.695 ± 0.050	0.695 ± 0.050	0.402 ± 0.016	0.013 ± 0.002
CelebA	ViT	1.000 ± 0.000	0.750 ± 0.000	0.750 ± 0.000	0.618 ± 0.012	-0.060 ± 0.003
Waterbirds	ViT	1.000 ± 0.000	0.350 ± 0.078	0.350 ± 0.078	0.155 ± 0.016	0.018 ± 0.007
PACS	ViT	0.985 ± 0.024	0.665 ± 0.047	0.650 ± 0.062	-0.019 ± 0.006	0.010 ± 0.002

D Broader impact

This work is methodological and encourages more conservative claims in label-only representation editing. The main risk is over-interpreting a post hoc edit as removing a uniquely identified nuisance factor; our recommended mitigation is to preserve the quotient first and to rely on residual edits only when support and stability diagnostics are strong.

A positive impact is that the work may reduce over-claiming in fairness-sensitive or safety-sensitive representation editing by clarifying what is and is not identifiable from labels alone. It may help practitioners avoid removing task-relevant information while auditing residual leakage. A negative impact is that users may still over-interpret residual edits as nuisance removal, especially in demographic or fairness applications where labels are incomplete proxies.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state the paper's three core claims: the label-only identifiability boundary, the structured DMD specialization, and the empirical conclusions about semantic auditing, quotient-first editing, and trust diagnostics. These claims are supported by the theoretical results in Sections 2–3, the estimation and usage discussion in Section 4, the empirical evidence in Section 5, and the appendix material in Sections A–C.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper explicitly limits its claims to the representation level, discusses that stage two is conditional rather than universally reliable, and emphasizes the structured nature of the linear second-order specialization; see Sections 1, 4, 5, and especially Section 6. The discussion also notes that the paper does not claim latent semantic recovery or universally valid nuisance discovery.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The assumptions for the structured linear specialization are stated in Section 3. Appendix A provides proofs of the identifiability, factorization, optimality, and contamination results, together with a standard perturbation proof for the finite-sample recovery statement. Implementation and empirical supporting material appear in Appendices B and C.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 5 and Appendices B–C describe the datasets, frozen backbones, preprocessing, rank choices, metrics, rerun protocol, and evaluation setup used in the experiments. Code is publicly available at https://github.com/Candice-x1/What-Can-Labels-Alone-Audit-and-Edit-in-Frozen-Representations-, with scripts and instructions for reproducing the main reported results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper uses public benchmark datasets and standard frozen backbones, and the paper provides a public code repository (https://github.com/Candice-x1/What-Can-Labels-Alone-Audit-and-Edit-in-Frozen-Representations-) with scripts and instructions.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

Justification: Section 5 and Appendix B specify the datasets, preprocessing, rank choices, evaluation metrics, probe family, rerun protocol, and support-sweep construction used in the experiments.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports results as mean $\pm$ standard deviation over 10 outer seeds and also includes repeated-rerun stability summaries and variability-oriented diagnostics in the trust analysis; see Section 5 and Appendix C.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix B includes a compute-resources description for the experiments, including the use of cached frozen representations together with the reported hardware, memory, and runtime information for reproduction.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: To the best of our knowledge, the research conforms to the NeurIPS Code of Ethics. The work is conducted on existing benchmark datasets and frozen pretrained representations and does not involve human-subject experimentation.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section D in the appendix discusses both potential positive and negative societal impacts, including the value of discouraging over-claiming in label-only editing and the risks of misuse or over-interpretation in fairness-sensitive applications.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release any high-risk generative models, scraped datasets, or dual-use assets of the kind targeted by this question. The work studies analysis and editing of frozen benchmark representations.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original papers and sources for all datasets, backbones, and pretrained representations used in the experiments. The work uses public benchmark datasets and standard pretrained backbones for research purposes, and does not redistribute the original datasets or pretrained model weights. We followed the access and usage terms associated with the respective assets to the best of our knowledge.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The public repository at https://github.com/Candice-x1/What-Can-Labels-Alone-Audit-and-Edit-in-Frozen-Representations- contains reproduction code and scripts, documented with instructions for preprocessing, running DMD and baselines, and reproducing results. The code is released under the MIT license.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, participant studies, or research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subject research, so IRB approval or equivalent review is not applicable.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs were used only for grammar, spelling, wording, and formatting assistance during manuscript preparation. They were not used as an original or non-standard component of the proposed method, and the theoretical claims, proofs, experimental design, implementation, and reported results do not depend on LLM-generated scientific content.