NeurIPS 2026

What Can Labels Alone Audit and Edit in Frozen Representations?

Quotient-first auditing and protected residual editing

Yirong Xiang
Air Force Medical Center

Yibo Zhou
Beihang University

The label-only setting

Frozen features and task labels are available for fitting



Waterbird / water background



Waterbird / land background

Fit using (X, Y)

X Frozen representation
Y Bird category

Evaluate with E

E Background

Which variation can we safely remove?

Labels identify the posterior quotient

$$\eta(x) = P(Y \mid X = x)$$

The coarsest deterministic representation that preserves all label-prediction information

$$x \sim x' \quad \Leftrightarrow \quad P(Y \mid x) = P(Y \mid x')$$

Finer semantic/background roles remain ambiguous

DMD protects a surrogate before deletion

1 Estimate the protected subspace

$$M = \mathrm{Cov}(\,\mathrm{E}[X \mid Y]\,)$$

$\hat{U}$ = leading eigenspace of M

Mean-stage rank ≤ K − 1

2 Rank residual covariance directions

$$Q = I - \hat{U}\hat{U}^{\top}$$

$$T_{\perp} = \sum_{\pi} (\Sigma_{\pi,\perp} - \bar{\Sigma}_{\perp})^2$$

$\Sigma_{\pi,\perp} = Q\Sigma_{\pi} Q$

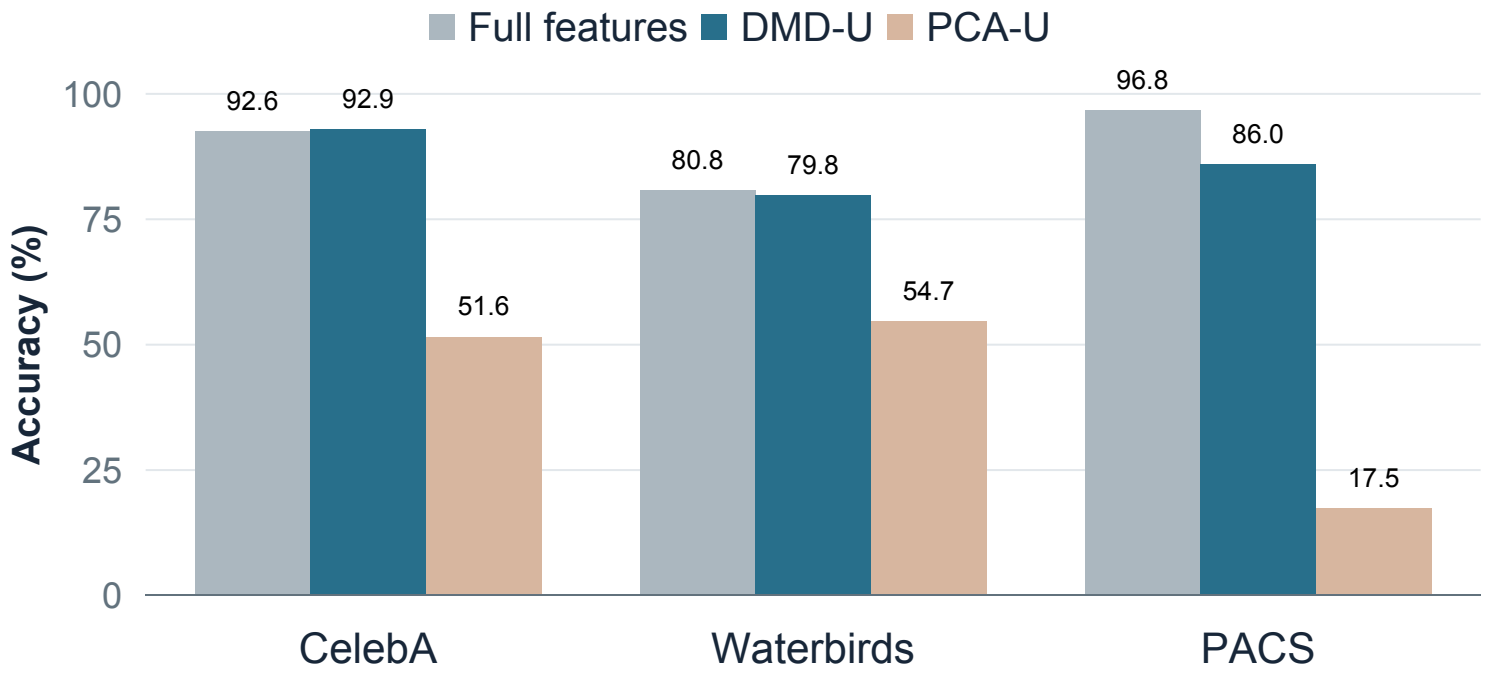
$\hat{P}$ = top-q projector of $T_{\perp}$

$$X' = (I - \hat{P})X \qquad \text{with } \hat{P}\hat{U} = 0$$

The edit preserves the estimated $\hat{U}$ coordinates

One direction retains most CelebA accuracy

Task accuracy in frozen CLIP features



92.9%

CelebA / CLIP
1-D surrogate

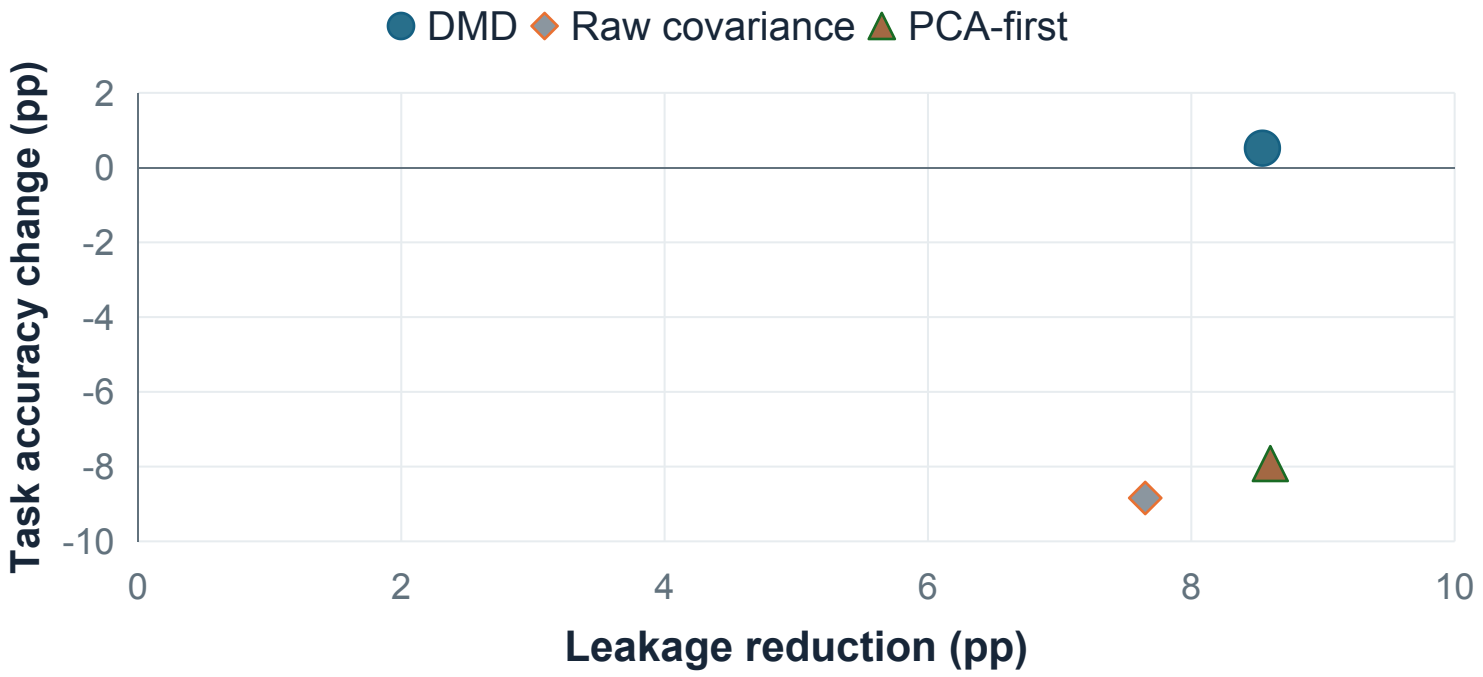
Full features: 92.6%

DMD-U and LDA-U are essentially tied across all six vision settings

CLIP ViT-B/32. Means over 10 seeds. Standard deviations < 0.1 percentage point.

Task protection at the same edit budget

Waterbirds / CLIP ViT-B/32, removal rank q = 8



+0.52 pp

DMD task accuracy change

−8.84 pp

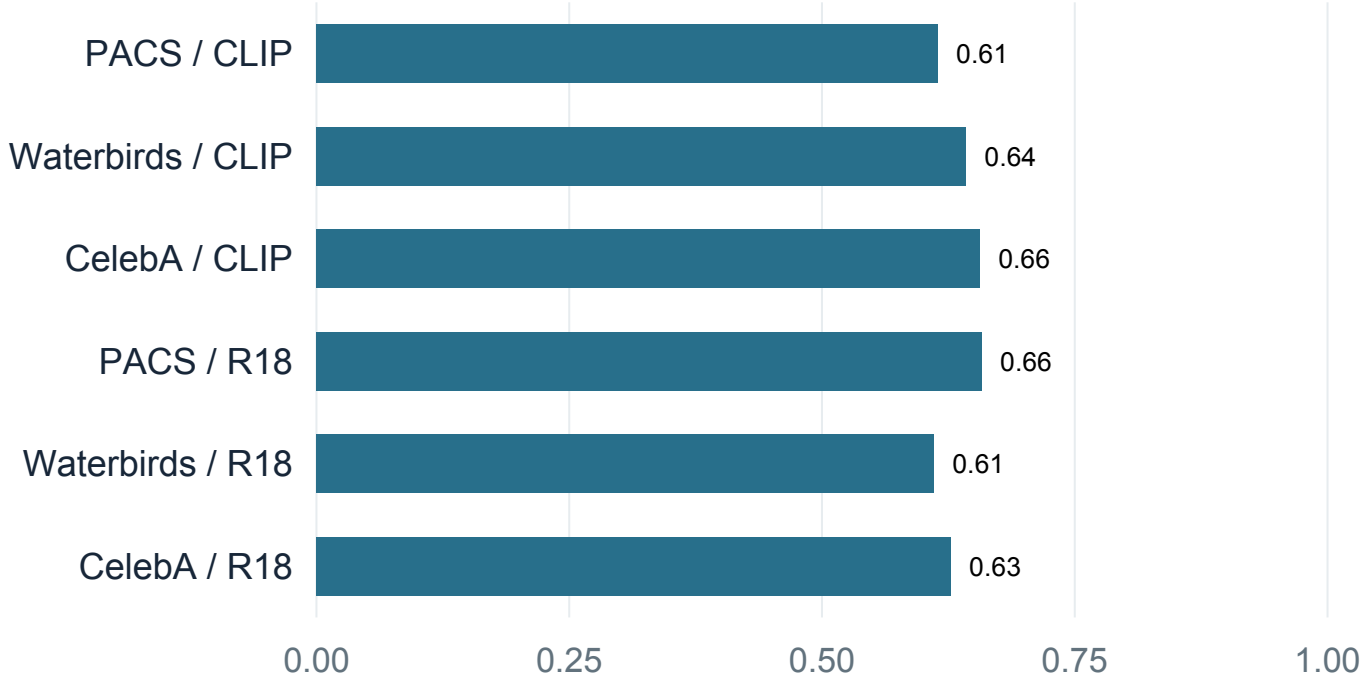
Raw covariance

Leakage-reduction differences remain unresolved. PACS / CLIP is near-tied.

Means over 10 seeds. pp = percentage points. Higher is better on both axes.

Residual edits need support and stability

Stage-two subspace distance after projector bagging



0.61–0.66×

remaining subspace distance

Check task retention
and stability across reruns

Single-run reference = 1. Smaller is better.

Environment-cell support uses evaluation annotations. R18 = ResNet-18.

Protect the quotient first

Labels identify a quotient.
A finer decomposition needs more information.

Stage one provides an audit and compression anchor.
Stage two needs evidence from support and stability.

Code and experiments

Yirong Xiang yirong_xiang@163.com
Yibo Zhou ybzhou@buaa.edu.cn

NEURAL INFORMATION PROCESSING SYSTEMS