FREEZER: EXTRACTING MORE FROM FROZEN REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

A pretrained encoder may retain useful task information that its scoring rule does not fully use for the population being evaluated. We introduce FREEZER (Feature Reweighting and Embedding Enhancement with Zero Encoder Retraining), a post-hoc framework that uses dataset-level embedding statistics and supervision, when available, to adapt scoring while keeping the encoder fixed. For speaker verification, it retains cosine similarity after centering and covariance-based reweighting. For speech and image deepfake detection, it fits a regularized scoring direction and optionally blends the original detector score. On Common Voice 26, macro equal error rate (EER) across 51 locales falls from 3.03% to 2.17%, with improvements in 49 locales. A separate collective face-verification procedure reduces ArcFace IJB-C EER from 1.20% to 0.95%, using unlabeled evaluation embeddings and neighboring comparisons. Speech deepfake detection EER falls from 20.79% to 10.05% on the full CoSG release, with adaptation using only separate MLAAD and In-the-Wild data. Image deepfake detection with AIDE reduces EER from 10.29% to 7.53% on the CNN benchmark. Component studies show that supervised centering supplies most of the speaker-verification gain, while a fitted direction preserves most of the speech deepfake detection gain without blending. For image deepfake detection, blending helps some populations but hurts others. These findings demonstrate the value of adapting the scoring stage of frozen systems, while showing why gains must be evaluated across populations and operating points.

1 INTRODUCTION

A pretrained system's scoring geometry may not make the best use of its representations for the task and population where it is deployed. Useful information may be present in the embeddings, yet have insufficient influence on the decisions. This leaves room for improving the system without necessarily retraining or fine-tuning its encoder.

If a representation already contains useful information, improving how the system uses it offers a smaller intervention than retraining the encoder. Fine-tuning requires additional computation and suitable training data. It can also reduce generalization beyond the adaptation population; for example, feature distortion has been demonstrated when pretrained features are strong and distribution shift is substantial. (Kumar et al., 2022)

Our premise is that *the usefulness of variation depends on what it distinguishes*. For verification, variation between samples of the same class can obscure differences between classes. A useful direction must therefore be assessed through its inter-class separation relative to its intra-class variability. For deepfake detection, the variation associated with content or acquisition conditions can compete with evidence of authenticity, and cues associated with one generator may not transfer to another. The distribution of embeddings provides evidence for adjusting this balance.

Two established approaches address this opportunity: adapt how embeddings are compared, or fit a task-specific classifier on frozen features (Cumani et al., 2025; Ojha et al., 2023). Both use adaptation data to change the scoring without requiring a new encoder. We ask which adjustments help across verification and detection, what supplies their benefit, and whether they are transferred.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

FREEZER studies two related procedures. Verification estimates a center and regularized covariance transformation, then retains *cosine similarity* on the normalized transformed embeddings. Translation preserves Euclidean differences, but changes the directions being compared (Figure 1). Detection fits a direction from class labels or teacher-derived pseudo-labels, optionally blending the original detector score. The shared principle is to use evidence at the dataset-level to adjust how existing information influences decisions. We regularize covariance estimates, balance supervised detector fitting across classes and sources, and select configurations on source/development data.

We find improvements in verification and deepfake detection across speech and images. In our Common Voice 26 (CV26) evaluation, speaker-verification equal error rate (EER), averaged across 51 locales, falls from 3.03% to 2.17%, with improvements in 49 locales. A separate face-verification experiment uses other image comparisons as context and reduces ArcFace's IJB-C EER from 1.20% to 0.95% (Appendix D). For speech deepfake detection, fitting a scoring function on separate data reduces CoSG EER from 20.79% to 10.05%. For image deepfake detection, EER falls from 25.79% to 17.38% with CLIP on Synthbuster and from 10.29% to 7.53% with AIDE on the CNN benchmark.

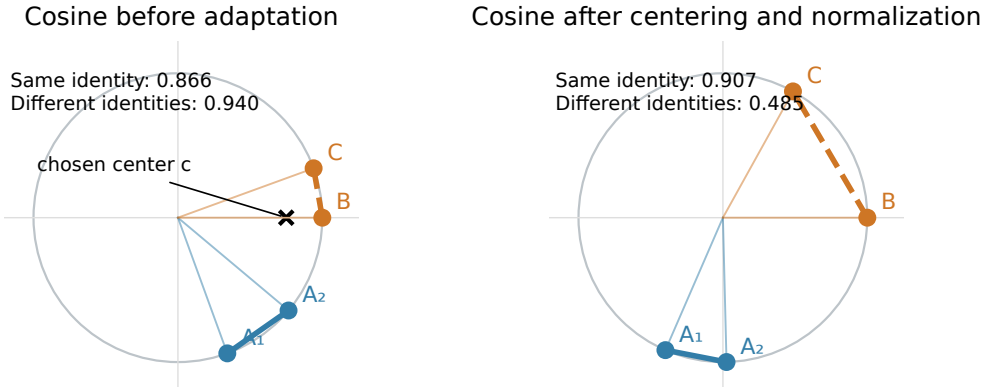

Figure 1: **Illustrative example.** In this exact two-dimensional construction, the genuine pair A_1, A_2 initially has lower cosine than the impostor pair B, C. Subtracting $c = (0.75, 0)$ and normalizing reverses their ordering. The construction illustrates a possible geometric correction.

The component studies show which adjustments contribute to these gains. Centering accounts for most of the CV26 improvement in our component comparisons. For speech detection, most of the benefit remains when the fitted scoring function is used on its own. For images, combining it with the original detector helps on CNN and Diffusion but hurts on Synthbuster. These findings show how dataset-informed scoring can improve the use of frozen representations, while showing that the useful adjustment depends on the task and population. The practical lesson is to evaluate this scoring stage before investing in encoder changes, and to test whether its gains transfer.

2 Related Work

The dependence of verification and detection performance on the evaluation population is well documented. SVeritas identifies substantial variation in speaker-verification performance across recording conditions, speaking styles and demographic groups (Baali et al., 2025). Studies of speech and image deepfake detection likewise report poor cross-population transfer and sensitivity to the data used to fit a detector (Müller et al., 2022; Kotyan et al., 2026). These findings motivate adapting systems to their intended application and evaluating whether the resulting improvements transfer.

Adapting verification geometry. One established approach is to adjust how representations are compared. Within-class covariance normalization reweights speaker features using within-speaker variation (Hatch et al., 2006); length normalization improves Gaussian speaker models (Garcia-Romero & Espy-Wilson, 2011). SimpleShot shows the value of mean subtraction and normalization for few-shot nearest-neighbor classification (Wang et al., 2019b). Our speaker study instead evaluates these operations for pair verification, including a supervised center fitted to a pairwise loss. For modern large-margin embeddings, Wang et al. (2022) revisit the cosine–PLDA comparison, while

Cumani et al. (2025) study preprocessing and calibration under mismatch. QMagFace uses model-specific quality estimates to adjust face comparisons (Terhörst et al., 2023). These precedents motivate adapting the scoring stage; our primary speaker procedure retains cosine.

Fitting decision rules on frozen features. UniversalFakeDetect uses nearest-neighbor classification and linear probing for image authenticity (Ojha et al., 2023). Wang et al. (2024) build a Gaussian classifier from class means and shared covariance, combining it with CLIP's original predictions. Our detection procedure similarly uses class statistics and score combination; its ridge direction is closely related to the shared-covariance discriminant. In particularly close work, Kotyan et al. (2026) compare Gaussian image-detector heads while controlling the encoder and fitting population. We separate the effects of covariance weighting, angular normalization and original-score blending, then examine how source composition affects transfer. Whitened CLIP uses covariance transformations for image and caption likelihood estimation (Betser et al., 2025), a different objective from our ranking evaluations.

Using comparison context. Cohort score normalization adjusts a trial using comparisons with reference recordings (Matějka et al., 2017). Our collective experiments instead use neighboring trials in the requested comparison graph. Their binary representations use iterative quantization (ITQ), which learns rotations to reduce quantization error (Gong & Lazebnik, 2011); the graph correction is an additional operation. Because endpoint fitting and neighboring comparisons supply extra information, we report this setting separately from independent pair scoring (Appendix D).

3 Feature Reweighting and Embedding Enhancement

FREEZER fits a verification map or a detection scoring function, called a *readout*, from source embeddings and available labels (Figure 2). Centering changes angular relationships, covariance weighting adjusts the influence of correlated variation, and a detection direction separates real from fake examples. Their usefulness depends on how well the source data represent the intended application.

For input a, let $x = f_\theta(a) \in \mathbb{R}^D$ and $z = x/\|x\|$. The encoder parameters θ stay fixed. Implementations floor normalization denominators; Appendix B gives numerical safeguards.

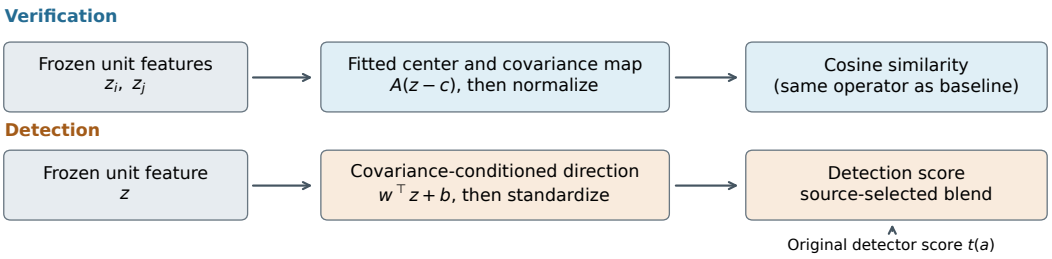

Figure 2: **Post-hoc adaptation.** Fit centers, covariance maps and detection directions from source embeddings; then apply the fixed rules to new inputs. Primary verification retains cosine; detection optionally blends the original score.

Verification: choose an origin and reweight directions. Subtract a center c, apply a linear map A, normalize, then compare by cosine:

$$q(z) = \frac{A(z-c)}{\|A(z-c)\|}, \qquad s(z_i, z_j) = q(z_i)^\top q(z_j). \tag{1}$$

The baseline has $c = 0$, $A = I$. Centering preserves Euclidean differences but changes normalized directions and hence cosine rankings (Figure 1). We estimate c from training means or fit it with balanced genuine/impostor pairs and a cosine softplus loss. The positive-definite metric $M = A^\top A$ reweights directions using a covariance or second-moment matrix C ($p \geq 0$, $\epsilon > 0$):

$$M = (C/\tau + \epsilon I)^{-p}, \qquad \tau = \max\{\mathrm{tr}(C)/D, 10^{-12}\}. \tag{2}$$

Positive p downweights directions with greater estimated variation. Within-identity covariance targets variation among examples of the same identity. Total covariance and uncentered second moments use

unlabeled embeddings and may also downweight useful identity differences. Regularization stabilizes the inverse weighting. The map A uses exponent $-p/2$; uniform metric scaling cancels in cosine.

Detection: fit a direction that separates classes. Ridge regression fits labels $y_i \in \{-1, +1\}$ (real/fake) with regularization $\lambda > 0$, accounting for correlations between coordinates. Weights $\omega_i \geq 0$ sum to one; $\mu, \bar{y}$ are weighted feature and label means. With weighted covariance $S_T = \sum_i \omega_i (z_i - \mu)(z_i - \mu)^\top$, the score $g(z) = w^\top z + b$ uses

$$w = (S_T + \lambda I)^{-1} \sum_i \omega_i (z_i - \mu)(y_i - \bar{y}), \quad b = \bar{y} - w^\top \mu. \tag{3}$$

Single-source fits give equal weight to each class; mixtures balance classes within and across sources. For intuition, correlated nuisance variation in two coordinates may cancel in their difference while retaining class separation. Covariance weighting can exploit such relationships; identifying their physical causes requires additional evidence.

With equal class weights and matching absolute λ, ridge and regularized linear discriminant analysis (LDA) have the same direction. This identity follows algebraically from the weighted fitting moments. A shared-covariance Gaussian model also yields a linear log likelihood ratio (Wang et al., 2024; Kotyan et al., 2026) (Appendix A).

The original detector score $t(a)$ may retain useful information absent from z. We therefore allow

$$s(a) = (1 - \alpha)\widehat{t}(a) + \alpha \widehat{g}(z). \tag{4}$$

Hats denote standardization by fitting-score means and standard deviations. Source validation selects α: zero retains the original score; one uses only the readout. Standardization makes blend scales comparable; positive affine scaling alone preserves standalone EER.

The original detector can also act as a *teacher*: its highest- and lowest-scoring examples supply two pseudo-labeled fitting groups. Their mean difference defines a contrast d; their embedding moments define a metric M.

For both pseudo-label and ridge fitting, we compare projection with angular normalization. Let $v = z - c_{\text{det}}$, using $c_{\text{det}} = 0$ for pseudo-label fitting and the weighted fitting mean for ridge. For ridge, d is the feature–label cross moment in Eq. 3 and $M = (S_T + \lambda I)^{-1}$. The scores are

$$g_{\text{dot}}(z) = v^\top M d, \qquad g_{\text{ang}}(z) = \frac{v^\top M d}{\sqrt{v^\top M v}}. \tag{5}$$

The dot form agrees with centered ridge up to a constant. The angular form divides this projection by the example's distance from the center in metric M; constant direction normalization cancels after standardization. Translation alone adds a common offset to a fixed linear score and leaves EER unchanged. Different M, d can yield the same $w = Md$, so one scoring direction can have several geometric decompositions.

Parameters are estimated on fitting data; separate source validation then selects a configuration. Speaker verification fits each locale's map and selects a common recipe by macro development EER. Detection selects regularization and blending by mean relative source-domain validation EER, including the original-score fallback; source labels also select pseudo-label variants (Appendix B). At inference, apply the fixed map and cosine, or the detector readout and blend. Linear ensembles sharing features and the original detector collapse to one dot product plus that score; angular variants additionally evaluate each example's metric radius.

Collective verification with comparison context. Collective verification also fits centering and iterative quantization (ITQ) rotations on unlabeled evaluation embeddings, forms binary pair scores, and standardizes them using neighboring trials. The rotations reduce quantization error; neighboring scores provide an endpoint-specific reference. This uses additional evaluation-population information and is assessed separately (Appendix D).

4 Experimental Setup

Three questions guide the experiments, using frozen checkpoints and matching evaluation examples. First, which components supply the verification gains? Center, covariance and quantization controls

test the importance of fitted directions (Section 5.1). Second, what supplies the detection gain? We compare normalization, blending, readout families and fitting populations using the same validation data (Section 5.2). Third, do the improvements transfer across datasets and operating points? We compare fixed recipes, transfer source thresholds, and control extraction and graph context (Section 5.3). Some controls are exploratory follow-ups motivated by earlier inspection of public benchmark results.

Table 1: Evaluation populations and data available for adaptation. Counts refer to scored examples; CNN/Diffusion cover preserved archives. "External" means fitting, selection and moment estimation use source data exclusively.

Population	Scored rows / aggregation	Adaptation access
CV26	524,087 recordings; 41.28M sampled pairs; 51-locale macro	Locale training; separate dev selection
MLAAD	19,614 recordings; 173 cells; six-language macro	Separate labeled calibration; excluded generators
ITW	12,160 of 31,779 recordings; 26 person/content groups	Separate labeled calibration/validation
CoSG	All 1,797 recordings; pooled	External
Synthbuster	4,980 images; nine-generator macro	Labeled calibration on separate stems
CNN / Diffusion	90,329 / 10,000 images; 13 / 8-population macro	CNN source fit/validation; Diffusion external
DF-Eval Test	788 audio files; 786 / 791 images	External; audio windows and images
Official verification	Vox O/E/H and complete face trial lists	Unlabeled endpoints and full trial graph

Representations and populations. Speaker verification uses 192-dimensional ReDimNet2 B6 embeddings (Yakovlev & Okhotnikov, 2026). Common Voice 26 (CV26) (Ardila et al., 2019) covers 51 locales. Dataset-provided speaker IDs are disjoint across training, development and test, including across locales. We call this identity-filtered evaluation the *strict test*. Trial balancing gives each eligible anchor speaker equal weight; representation at the other endpoint varies. Each recording uses the map fitted for its known locale. Face verification uses 512-dimensional ArcFace/AdaFace embeddings, with source fitting on VGGFace2 and evaluation on CALFW, CPLFW, AgeDB and IJB-C (Appendix D).

Detection uses frozen XLSR-SLS (Babu et al., 2021; Zhang et al., 2024) for speech, and CLIP-based UniversalFakeDetect (Radford et al., 2021; Ojha et al., 2023) and AIDE (Yan et al., 2024) for images. Speech readouts use 2,048-dimensional layer-6 temporal mean/standard-deviation features; the baseline is the released detector score. A separate 240-dimensional handcrafted acoustic view tests sensitivity to recording conditions. Image features have 768 dimensions for CLIP and 1,024 for AIDE, retaining CLIP's center crop and AIDE's released DCT-patch/whole-image preprocessing. We evaluate MLAAD, In-the-Wild (ITW), CodecFake+ CoSG, Synthbuster, CNN-generated and diffusion image populations, and Deepfake-Eval-2024 (Müller et al., 2024; 2022; Chen et al., 2025; Bammey, 2023; Wang et al., 2019a; Chandra et al., 2025).

Aggregation and source access. MLAAD macro EER averages generator-group EERs within each language, then equally across six languages. Each language–generator group is a *cell*. Two complementary fits exclude the corresponding scoring generators and source-content IDs. ITW fitting and scoring are separated by groups linked through person or content.

Speech sources are MLAAD and ITW; image sources are Synthbuster and CNN. A *specialist* fits and validates on one source; the primary *mixed-source* procedure uses both. To isolate the effect of fitting-source choice, *matched single-source* controls fit on one source while retaining validation on both. External datasets are reserved for evaluation.

Speech ensembles average twelve readouts; image ensembles average two per encoder. MLAAD and Synthbuster cells instead use fits that exclude their scoring generators. Matched controls share validation data and score standardization; specialist-versus-mixture comparisons change both. Appendix B gives folds and label counts.

Deepfake-Eval is evaluation-only: fitting, selection, moment estimation and pseudo-labeling all use separate source data. Its 788 audio Test recordings use either the first window or an average of first/center/end window scores. We score 786 of 791 Test images; five files remain unresolved (23 across all splits). Evaluation covers audio and images only.

Metrics and uncertainty. We report EER (lower is better), verification true-accept rate (TAR) at stated false-accept rates (FAR), and detection true-positive rate (TPR) at stated false-positive rates (FPR). Macro EER averages group EERs. Genuine pairs and fake media are positive. EER interpolates the ROC crossing where false-positive and false-negative rates match, keeping ties together. Relative reduction is $(e_{\text{base}} - e_{\text{new}})/e_{\text{base}}$. Destination ROC thresholds use destination scores; transferred thresholds are chosen on source data. Reported 95% intervals resample identity, linked person/content, or identical-content groups, as specified. Intervals are unadjusted for multiple comparisons. Some hold fits fixed; others refit calibration. Shared-endpoint trials are dependent. Comparisons match rows and frontends unless extraction is the intervention.

5 Results and Discussion

5.1 Verification across speech and images

Speaker verification with independent pair scoring. On strict Common Voice trials, cosine scoring after the development-selected transformation reduces macro EER from 3.03% to 2.17%, a 28.38% relative reduction. Forty-nine of 51 locales improve; Romanian and Tamil regress. A paired bootstrap over dataset-provided speaker IDs shared across locales gives a relative-reduction interval of $[26.89, 29.82]\%$, conditional on the fitted models and sampled trials.

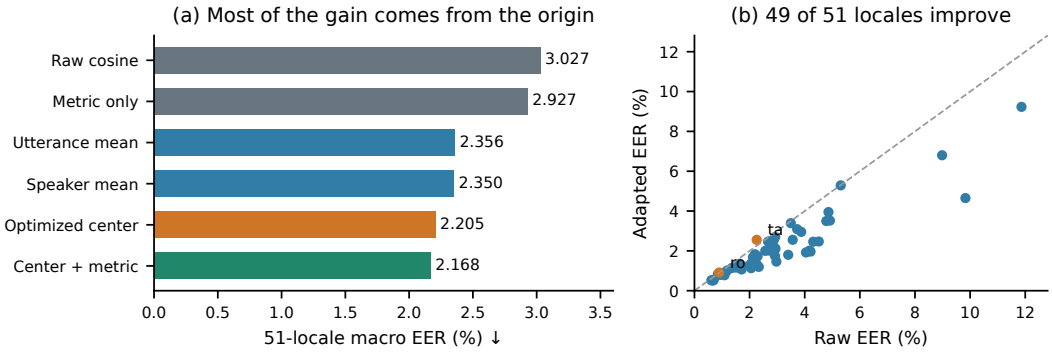

Figure 3: **Centering supplies most of the CV26 gain.** Left: strict-test macro EER for matched component comparisons. Right: one development-selected configuration, fitted separately for all 51 locales. Points below the diagonal improve; two locales regress.

Centering supplies most of the improvement (Figure 3). With a center fitted for each locale, subtracting the utterance mean gives 2.36% EER; weighting speaker centroids equally gives 2.35%. Optimizing the center with speaker labels reaches 2.20%, and adding the metric reaches 2.17%. Along this ablation sequence, the optimized center accounts for 95.7% of the raw-to-final EER reduction. Metric adaptation alone gives 2.93%. A Gaussian random-effects score fitted on the same pools and separately selected on development gives 2.72%. The Gaussian control uses a moments-based estimator, as detailed in Appendix B.

The fitted directions matter: rotating the center at fixed norm reduces its benefit, and randomizing covariance axes while preserving eigenvalues gives approximately 2.250–2.262%, worse than the fitted axes. Once the supervised center is fixed, however, total and shuffled-label covariance approximately tie the selected covariance. The incremental value of correct identity grouping is therefore unresolved. At low false-accept rates, TAR at FAR 10^{-4} changes from 67.527% to 67.580%, with an interval for the change crossing zero.

Face verification with comparison context. On all official IJB-C trials, ITQ plus graph scoring reduces ArcFace EER from 1.1965% to 0.9511% (20.51% relative) and AdaFace from 1.4828%

to 1.3243% (10.69%). The procedure uses unlabeled evaluation templates and neighboring trials, additional information beyond independent CV26 scoring. With the same graph correction, centered continuous features give 1.0817% ArcFace EER and random sign rotations 1.1147%, versus fitted ITQ's 0.9511%. Appendix D gives both encoders' controls and collective VoxCeleb results.

5.2 Deepfake detection across speech and images

Fitting on two sources yields improvements in both speech and image deepfake detection. The mixed-source speech readout reduces EER from 6.61% to 2.23% on the separate In-the-Wild (ITW) scoring groups, and from 20.79% to 10.05% on external CoSG. The mixed-source image procedure reduces Synthbuster EER from 25.79% to 17.38% with CLIP and from 56.56% to 33.22% with AIDE.

How much gain remains without the original score? We fix fitted directions and regularizers, then compare dot and angular scores, with and without the original detector score. *Head-only* scoring uses only the fitted readout, setting its blend weight to one. Each form is standardized on fitting data. Angular variants inherit dot-score regularizers and, when blended, blend weights. Removing blends also changes relative readout contributions when ensemble members have different weights (Appendix G).

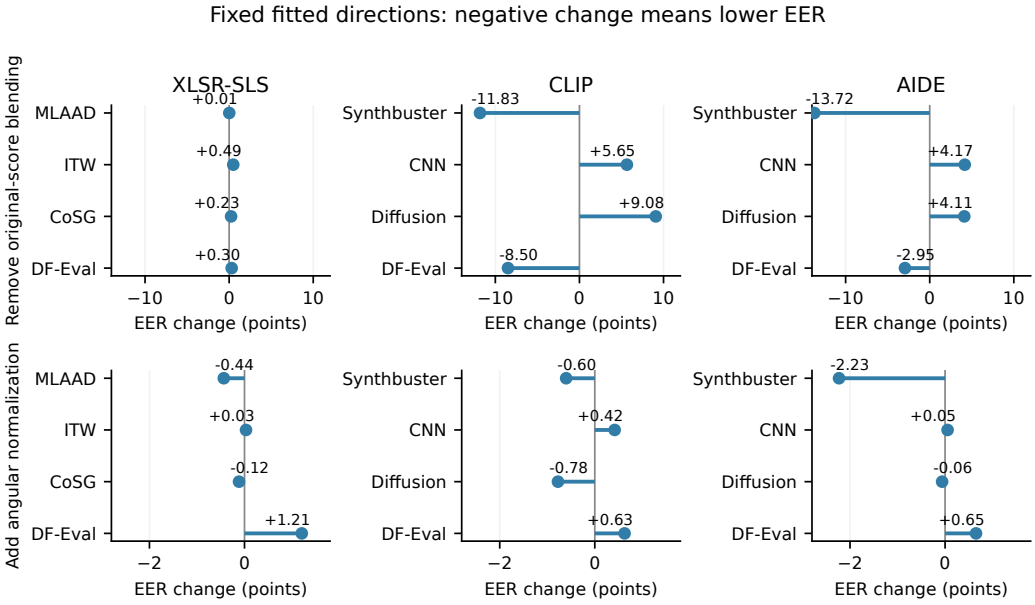

Figure 4: **Normalization and score blending affect transfer differently.** Top: change in EER after removing the original detector score from the dot-score blend. Bottom: change after adding angular normalization to the head-only dot score. Negative values indicate lower EER. All directions and regularizers are fixed; the bottom comparison changes only normalization and the resulting standardization estimated from fitting data. Each panel reports a separate intervention on the source-selected configuration.

For speech, the fitted direction alone retains nearly all the gain on MLAAD (1.6583% versus blended 1.6449%) and on the full CoSG release (10.28% versus 10.05%; original 20.79%). Paired resampling of identical-audio groups across all 1,797 CoSG recordings gives head-only minus original EER interval $[-13.06, -8.25]$ points, conditional on the fits. Thus most of the improvement survives without the original-score blend or angular normalization.

For images, blending helps some populations and hurts others. Removing the original CLIP score improves Synthbuster from 17.38% to 5.56%, but worsens CNN from 10.49% to 16.13% and Diffusion from 10.90% to 19.98%. AIDE shows the same tradeoff. One possible explanation

is that the original score retains useful distinctions, such as feature magnitude discarded by unit normalization.

Angular normalization helps on the primary populations but hurts on DF-Eval. Head-only EER falls from 1.66% to 1.22% on MLAAD, from 5.56% to 4.95% with CLIP on Synthbuster, and from 19.50% to 17.27% with AIDE on Synthbuster. On DF-Eval it worsens all three: speech 32.93% to 34.14%, CLIP 42.00% to 42.62%, and AIDE 40.33% to 40.98%. These comparisons keep the fitted directions and regularizers fixed.

Comparing fitting sources and readouts. To isolate the effect of fitting data, the ridge controls use the same validation data, folds, features, standardization and parameter grids. Table 2 also compares logistic regression and a class-mean contrast without covariance weighting fitted on the mixture. All use both domains' validation labels; the MLAAD-only row differs from the specialist selected on MLAAD alone.

Table 2: **Comparing fitting sources and readouts with shared validation.** EER (%); all fitted families use the same two-domain source validation. ITW uses separate labeled calibration and 12,160 component-disjoint scoring rows; CoSG is the full 1,797-recording external population. DF-Eval is all 788 audio Test files, first window.

Fitting population / readout	MLAAD	ITW	CoSG	DF-Eval
Original detector	21.070	6.609	20.785	40.700
MLAAD / ridge	5.234	5.439	16.513	37.160
ITW / ridge	22.127	3.545	20.193	39.825
Mixture / mean contrast	20.403	7.024	19.861	39.275
Mixture / logistic	2.860	3.672	12.137	33.698
Mixture / ridge	1.645	2.226	10.046	32.628

Mixed ridge has lower measured EER than either single-source ridge on all four populations. Its ITW reduction is 66.31% relative to baseline; bootstrapping 26 linked person/content groups with fixed fits gives an absolute EER-change interval of $[-8.784, -2.621]$ points. Its advantage over matched ITW-only ridge is less certain: the interval $[-3.817, 0.183]$ includes zero, as does the interval with source refitting (Appendix F). Logistic regression also improves substantially, while the mean contrast produces much smaller gains or a regression, supporting the value of fitting beyond the class-mean difference.

The mixture also improves CoSG at a matched 512-row fitting budget per head. Unique ensemble label counts still differ; Appendix H reports these controls and the matched image comparison with shared-covariance LDA.

Teacher-score pseudo-labels also improve MLAAD EER (21.07% to 18.45%), but labeled source validation still selects the rule. Appendix B.3 reports its covariance controls and explains which labels each procedure uses.

5.3 Transfer and operating-point behavior

The speech specialist transfers unevenly (Figure 5): it improves MLAAD but worsens ITW to 13.60%, from the original 6.61%. The mixture improves ITW relative to the specialist while increasing MLAAD EER to 1.6449%. It ties at 10.05% on full CoSG and worsens first-window DF-Eval (32.63% versus 31.51%). Both fitting and validation access change. The mixture improves 166/173 MLAAD language–generator groups over baseline, yet worsens 46 relative to the specialist.

CLIP and AIDE mixtures also increase EER relative to specialists on every Synthbuster generator while reducing CNN/Diffusion failures (Appendix C). Switching the source objective from relative to absolute EER improves some destinations and degrades others (Appendix F); each objective is retained across all destinations. CLIP still regresses on full Diffusion: 10.775% original versus 10.90% mixed EER. The mixture improves full CNN from 11.06% to 10.49%.

On Deepfake-Eval audio, averaging three fixed windows improves the original score from 40.70% first-window EER to 38.29%. With matching windows, the mixture reaches 30.63%, a 20.00%

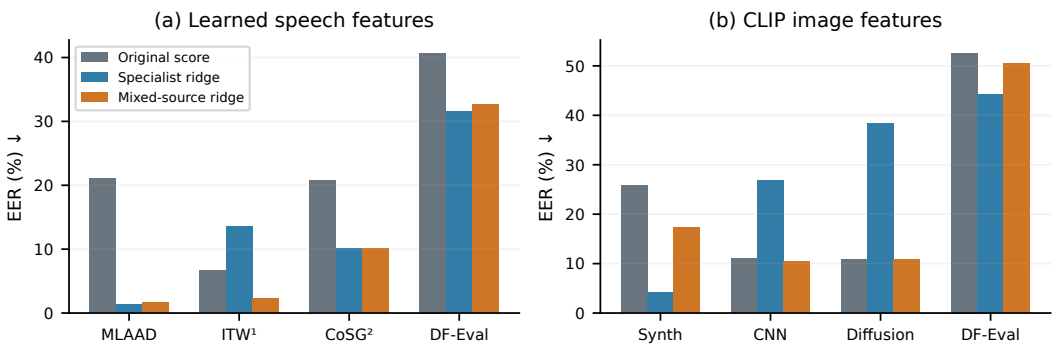

Figure 5: **Strong performance on one source can accompany poor transfer.** Methods are fixed across columns. Specialists use one fitting/validation source; mixtures use two. This compares complete recipes; Table 2 uses the same validation data for all fitted methods. Speech DF-Eval uses the first window; CLIP uses its original center crop. [1]Separate labeled ITW calibration; [2]all 1,797 released CoSG files.

reduction; the specialist reaches 29.32%. Threshold transfer remains difficult: at the threshold chosen for at most 1% source FPR, the first-window mixture has target FPR 2.845% and misses 81.571% of fakes. The transferred threshold therefore remains poorly matched to this population.

Face verification also depends on the population and operating point. Both encoders improve IJB-C EER and TAR at FAR 10^{-4}, but regress on AgeDB. On CPLFW, ArcFace EER improves from 8.27% to 6.77% while TAR at FAR 10^{-4} falls from 43.93% to 9.13%. With only 3,000 impostor trials, this operating point requires zero observed false accepts and is sensitive to extreme scores (Appendix D).

Attributing gains also requires a baseline with the same extraction procedure. On full CNN, five-crop extraction supplies 2.510 of the 3.505 EER-point improvement; the fitted head supplies 0.995 along that sequence (Appendix C).

6 Conclusion

FREEZER uses the distribution of frozen embeddings, with labels when available, to adapt how a system compares examples or detects fakes. It changes the scoring stage without retraining the encoder. The results show why this stage deserves attention: useful improvements can come from estimating a center, reweighting variation, or fitting a task-specific direction from a representative dataset.

The component comparisons reveal different sources of improvement. On CV26, supervised centering supplies most of the locale-macro EER reduction while retaining cosine. In speech deepfake detection, a fitted direction preserves most of the gain without the original-score blend. Image detection shows that this blend can help transfer to some populations while hurting others. For face verification, learned binary representations improve IJB-C results with additional comparison-graph context. Across tasks, the lesson is to validate the adaptation data and scoring choices against each intended population and operating point.

Limitations. Our evidence covers selected encoders and public benchmarks, including exploratory follow-ups. Identity overlap with pretraining data or between corpora remains unresolved, and DF-Eval image coverage is incomplete. Collective verification requires evaluation-population and graph access beyond independent pair scoring. Uncertainty estimates are conditional; most small image and normalization differences remain point estimates. Comparing accuracy and computational cost against encoder fine-tuning remains future work.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix A give the scoring rules and mathematical arguments. Appendix B describes feature extraction, fitting and validation rules, parameter grids and trial sampling; the remaining appendices report coverage, additional results and uncertainty.

The supplementary code bundle provides aggregate evidence, figure-generation scripts, reference implementations, correctness checks and a cached-feature detector fitting/selection/scoring pipeline. It supports mathematical checks and regeneration of reported displays. Rerunning the benchmark experiments also requires dataset access and the original checkpoints.

AI USE STATEMENT

Generative AI was used for code generation and manuscript revision. All research ideas, experimental design, results and conclusions are the authors' own; the first author reviewed and verified all AI-assisted text.

ETHICS STATEMENT

Speaker and face verification involve sensitive biometric information, and deepfake scores can wrongly implicate authentic media. We use existing datasets under their access conditions and report aggregate results rather than personal identities.

REFERENCES

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common Voice: A Massively-Multilingual Speech Corpus. *arXiv preprint arXiv:1912.06670*, 2019. URL https://arxiv.org/abs/1912.06670.

Massa Baali, Sarthak Bisht, Francisco Teixeira, Kateryna Shapovalenko, Rita Singh, and Bhiksha Raj. SVeritas: Benchmark for robust speaker verification under diverse conditions. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 9714–9731. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.516. URL https://aclanthology.org/2025.findings-emnlp.516/.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. *arXiv preprint arXiv:2111.09296*, 2021. URL https://arxiv.org/abs/2111.09296.

Quentin Bammey. Synthbuster: Towards detection of diffusion model generated images. Dataset release, 2023. URL https://zenodo.org/records/10066460.

Shai Ben-David, Tyler Lu, Teresa Luu, and David Pal. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 129–136, 2010. URL https://proceedings.mlr.press/v9/david10a.html.

Roy Betser, Meir Yossef Levi, and Guy Gilboa. Whitened clip as a likelihood surrogate of images and captions. In *International Conference on Machine Learning*, volume 267, pp. 4069–4095, 2025. URL https://proceedings.mlr.press/v267/betser25a.html.

Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. *arXiv preprint arXiv:1710.08092*, 2017. URL https://arxiv.org/abs/1710.08092.

Nuria Alina Chandra, Hannah Lee, Ryan Murtfeldt, Lin Qiu, Arnab Karmakar, Emmanuel Tanumihardja, Kevin Farhat, Ben Caffee, Changyeon Lee, Jongwook Choi, Sejin Paik, Aerin Kim, and

Oren Etzioni. Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. *arXiv preprint arXiv:2503.02857*, 2025. URL https://arxiv.org/abs/2503.02857.

Xuanjun Chen, Jiawei Du, Haibin Wu, Lin Zhang, I-Ming Lin, I-Hsiang Chiu, Wenze Ren, Yuan Tseng, Yu Tsao, Jyh-Shing Roger Jang, and Hung-yi Lee. CodecFake+: Codec-Based Resynthesized Data as a Proxy for Detecting CodecFake Speech. *arXiv preprint arXiv:2501.08238*, 2025. URL https://arxiv.org/abs/2501.08238.

Sandro Cumani, Anna Silnova, Sara Barahona, Ladislav Mošner, Oldřich Plchot, and Johan Rohdin. Analysis of the ABC classification backends for NIST SRE24. In *Interspeech*, pp. 3978–3982, 2025. doi: 10.21437/Interspeech.2025-146. URL https://www.isca-archive.org/interspeech_2025/cumani25_interspeech.html.

Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *arXiv preprint arXiv:1801.07698*, 2018. URL https://arxiv.org/abs/1801.07698.

Daniel Garcia-Romero and Carol Y. Espy-Wilson. Analysis of i-vector length normalization in speaker recognition systems. In *Interspeech*, 2011. URL https://www.isca-archive.org/interspeech_2011/garciaromero11_interspeech.html.

Yunchao Gong and Svetlana Lazebnik. Iterative quantization: A procrustean approach to learning binary codes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 817–824, 2011. doi: 10.1109/CVPR.2011.5995432. URL https://slazebni.cs.illinois.edu/publications/cvpr11_small_code.pdf.

Andrew O. Hatch, Sachin Kajarekar, and Andreas Stolcke. Within-class covariance normalization for SVM-based speaker recognition. In *Interspeech*, pp. 1471–1474, 2006. doi: 10.21437/Interspeech.2006-183. URL https://www.isca-archive.org/interspeech_2006/hatch06_interspeech.html.

Minchul Kim, Anil K. Jain, and Xiaoming Liu. AdaFace: Quality Adaptive Margin for Face Recognition. *arXiv preprint arXiv:2204.00964*, 2022. URL https://arxiv.org/abs/2204.00964.

Shashank Kotyan, Makoto Shing, Yuki Imajuku, Rujikorn Charakorn, and Tarin Clanuwat. Prior-conditioned gaussian discriminants for generalizable ai-generated image detection. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2608.18523. Accepted paper.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2202.10054.

Pavel Matějka, Ondřej Novotný, Oldřich Plchot, Lukáš Burget, Mireia Diez Sánchez, and Jan Černocký. Analysis of score normalization in multilingual speaker recognition. In *Interspeech*, pp. 1567–1571, 2017. doi: 10.21437/Interspeech.2017-803. URL https://www.isca-archive.org/interspeech_2017/matejka17_interspeech.html.

Brianna Maze, Jocelyn Adams, James A. Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K. Jain, W. Tyler Niggel, Janet Anderson, Jordan Cheney, and Patrick Grother. IARPA Janus Benchmark–C: Face dataset and protocol. In *International Conference on Biometrics*, pp. 158–165, 2018. doi: 10.1109/ICB2018.2018.00033. URL https://doi.org/10.1109/ICB2018.2018.00033.

Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. AgeDB: The first manually collected, in-the-wild age database. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017_workshops/w33/papers/Moschoglou_AgeDB_The_First_CVPR_2017_paper.pdf.

Nicolas M. Müller, Franziska Dieckmann, Pavel Czempin, Roman Canals, Konstantin Böttinger, and Jennifer Williams. Speech is silver, silence is golden: What do ASVspoof-trained models really learn? In *ASVspoof Workshop*, 2021. URL https://arxiv.org/abs/2106.12914.

Nicolas M. Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. Does audio deepfake detection generalize? In *Interspeech*, 2022. URL https://arxiv.org/abs/2203.16263.

Nicolas M. Müller, Piotr Kawa, Wei Herng Choong, Edresson Casanova, Eren Gölge, Thorsten Müller, Piotr Syga, Philip Sperl, and Konstantin Böttinger. MLAAD: The Multi-Language Audio Anti-Spoofing Dataset. *arXiv preprint arXiv:2401.09512*, 2024. URL https://arxiv.org/abs/2401.09512.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards Universal Fake Image Detectors that Generalize Across Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24480–24489, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Ojha_Towards_Universal_Fake_Image_Detectors_That_Generalize_Across_Generative_Models_CVPR_2023_paper.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint arXiv:2103.00020*, 2021. URL https://arxiv.org/abs/2103.00020.

Stefan Smeu, Dragos-Alexandru Boldisor, Dan Oneata, and Elisabeta Oneata. Circumventing shortcuts in audio-visual deepfake detection datasets with unsupervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18815–18825, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Smeu_Circumventing_Shortcuts_in_Audio-visual_Deepfake_Detection_Datasets_with_Unsupervised_Learning_CVPR_2025_paper.html.

Philipp Terhörst, Malte Ihlefeld, Marco Huber, Naser Damer, Florian Kirchbuchner, Kiran Raja, and Arjan Kuijper. QMagFace: Simple and accurate quality-aware face recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023. URL https://openaccess.thecvf.com/content/WACV2023/html/Terhorst_QMagFace_Simple_and_Accurate_Quality-Aware_Face_Recognition_WACV_2023_paper.html.

Qiongqiong Wang, Kong Aik Lee, and Tianchi Liu. Scoring of large-margin embeddings for speaker verification: Cosine or plda? In *Interspeech*, pp. 600–604, 2022. URL https://www.isca-archive.org/interspeech_2022/wang22r_interspeech.html.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-generated images are surprisingly easy to spot... for now. *arXiv preprint arXiv:1912.11035*, 2019a. URL https://arxiv.org/abs/1912.11035.

Yan Wang, Wei-Lun Chao, Kilian Q. Weinberger, and Laurens van der Maaten. SimpleShot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv:1911.04623*, 2019b. URL https://arxiv.org/abs/1911.04623.

Zhengbo Wang, Jian Liang, Lijun Sheng, Ran He, Zilei Wang, and Tieniu Tan. A hard-to-beat baseline for training-free clip-based adaptation. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2402.04087.

Ivan Yakovlev and Anton Okhotnikov. ReDimNet2: Scaling Speaker Verification via Time-Pooled Dimension Reshaping. *arXiv preprint arXiv:2603.11841*, 2026. URL https://arxiv.org/abs/2603.11841.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A Sanity Check for AI-generated Image Detection. *arXiv preprint arXiv:2406.19435*, 2024. URL https://arxiv.org/abs/2406.19435.

Qishan Zhang, Shuangbing Wen, and Tao Hu. Audio deepfake detection with self-supervised xls-r and sls classifier. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6765–6773, 2024. doi: 10.1145/3664647.3681345. URL https://doi.org/10.1145/3664647.3681345.

Tianyue Zheng and Weihong Deng. Cross-pose LFW: A database for studying cross-pose face recognition in unconstrained environments. Technical Report 18-01, Beijing University of Posts and Telecommunications, 2018. URL https://www.whdeng.cn/CPLFW/Cross-Pose-LFW.pdf.

Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017. URL https://arxiv.org/abs/1708.08197.

A STATISTICAL FOUNDATIONS AND RANKING CHANGES

The appendices provide the mathematical arguments, experimental recipes and additional controls behind the main findings. Use the guide below to locate a particular question.

Reader question	Section
Why can these operations change decisions?	A
What data and parameters define each procedure?	B
How are transfer populations and coverage defined?	C
What changes when the full trial graph is available?	D
Are gains sensitive to input perturbations?	E
What happens under refitting or a transferred threshold?	F
What do fixed-direction and fitting-budget controls show?	G, H

A.1 POPULATION STATISTICS AND TASK-CONDITIONED DECISIONS

Verification and detection use different statistical relationships. Under specified class-conditional distributions, their likelihood-ratio scores are, respectively,

$$\ell_{\mathrm{ver}}(z_i, z_j) = \log \frac{p(z_i, z_j \mid H_{\mathrm{same}})}{p(z_i, z_j \mid H_{\mathrm{different}})}, \qquad \ell_{\mathrm{det}}(z) = \log \frac{p(z \mid y = +1)}{p(z \mid y = -1)}. \tag{6}$$

Verification compares shared-identity hypotheses; detection compares classes. These likelihood ratios provide statistical reference rules for the fitted scores. Speaker PLDA models the identity hypotheses (Garcia-Romero & Espy-Wilson, 2011); our adapted cosine instead fits a restricted family of pair scores.

What an unlabeled distribution can tell us. The law of total covariance separates overall variation into two terms:

$$\mathrm{Cov}(z) = \mathbb{E}_y[\mathrm{Cov}(z \mid y)] + \mathrm{Cov}_y(\mathbb{E}[z \mid y]). \tag{7}$$

Here y denotes identity or detection class, with expectations under the fitting weights. This decomposition depends on the class assignments as well as marginal covariance. For example, place equal mass on the four unit vectors $(\pm 1, \pm 1)/\sqrt{2}$. Assigning classes by the sign of either coordinate leaves the unlabeled distribution unchanged but changes the separating direction.

Unlabeled reweighting can therefore help or hurt. Identity labels supply pair relationships and within-identity scatter; class labels supply detection contrasts. Teacher-derived groups can inherit the teacher's errors. Their usefulness must be assessed through task performance; attributing a direction to a physical nuisance requires additional evidence.

When a linear detection direction is justified. Consider a Gaussian working model with class means $\mu_\pm$ and a shared positive-definite covariance Σ. Expanding its log density ratio gives

$$\ell_{\mathrm{det}}(z) = (\mu_+ - \mu_-)^\top \Sigma^{-1} z - \tfrac{1}{2}(\mu_+^\top \Sigma^{-1} \mu_+ - \mu_-^\top \Sigma^{-1} \mu_-). \tag{8}$$

This Gaussian approximation motivates a linear detection direction for normalized features (Wang et al., 2024; Kotyan et al., 2026). Unequal covariances or multimodal classes generally require richer rules. The ridge/LDA identity below instead follows algebraically from the fitting moments and concerns score rankings.

For a fixed linear detector, subtracting a common center adds only the constant $-w^\top c$ and leaves rankings and EER unchanged. Refitting ridge after a pure translation also leaves predictions unchanged when an unpenalized intercept is included consistently. Cosine and angular detection have example-dependent denominators, so centering can change their rankings.

Transfer depends on the stability of the fitted task relationships across populations. Assumptions connecting source and target tasks are central to domain-adaptation theory (Ben-David et al., 2010). Fitting uses source labels; transfer is evaluated separately from source validation.

A.2 Centering changes example-dependent projections

Centering changes a pair's cosine through its projections onto the center. For unit inputs, write $s = z_i^\top z_j$ and $a_i = c^\top z_i$. The exact center-only score is

$$s_c = \frac{s - a_i - a_j + \|c\|^2}{\sqrt{1 - 2a_i + \|c\|^2}\sqrt{1 - 2a_j + \|c\|^2}}. \tag{9}$$

If $a_i = a$ for every example, the denominator is a common positive constant and s_c is a positive affine function of s. Rankings, ties and EER are unchanged. This gives a concrete counterexample to the explanation that more angular spread must improve verification.

Let $r_i = \|z_i - c\|$ and $u_i = (z_i - c)/r_i$. Differentiating gives

$$\nabla_c s_c = \frac{s_c u_i - u_j}{r_i} + \frac{s_c u_j - u_i}{r_j}, \qquad \nabla_c s_c\big|_{c=0} = (s-1)(z_i + z_j). \tag{10}$$

A supervised loss combines these pair-dependent directions with equal weight for genuine pairs P and impostor pairs N:

$$\begin{aligned}\mathcal{L}(c, b) = {} & \tfrac{1}{2}\mathbb{E}_P \operatorname{softplus}[-20(s_c - b)] + \tfrac{1}{2}\mathbb{E}_N \operatorname{softplus}[20(s_c - b)] \\ & + \eta \|c - c_0\|^2 .\end{aligned} \tag{11}$$

Here b is a fitted score threshold, c_0 is the reference speaker mean, η controls its penalty, and $\operatorname{softplus}(u) = \log(1 + e^u)$. The supervised center therefore depends on labeled pair relationships across the fitting dataset, beyond its mean.

The main illustration uses four unit points at angles $-70, -40, 0, 20$ degrees and center $(0.75, 0)$. Genuine/impostor cosines change from $0.8660/0.9397$ to $0.9071/0.4850$. This construction illustrates the ranking reversal for those two comparisons.

A.3 Ridge, a metric direction, and regularized LDA

For fitting embeddings arranged as rows of Z and labels y, the no-intercept ridge solution is $w = (Z^\top Z + \lambda I)^{-1} Z^\top y$. It is exactly a metric direction Md with $d = Z^\top y$. For the fitting second moment $C = Z^\top Z/n$ and $p = 1$ in Eq. 2, the induced metric is $n\tau(Z^\top Z + n\tau\epsilon I)^{-1}$. Corresponding absolute regularization therefore produces the same score up to a positive scalar. An intercept or centering changes these normal equations.

For the centered weighted fit of Eq. 3, suppose each class has weight $1/2$. Write its means as m_+ and m_-, $\delta = m_+ - m_-$, and $\mu = (m_+ + m_-)/2$. The weighted cross moment is $\delta/2$ and total covariance is $S_T = S_W + \delta\delta^\top/4$. Let $K = S_W + \lambda I \succ 0$. Sherman–Morrison gives

$$w_{\text{ridge}} = \frac{\frac{1}{2}K^{-1}\delta}{1 + \frac{1}{4}\delta^\top K^{-1}\delta}. \tag{12}$$

The positive denominator makes ridge and regularized LDA produce the same ranking at matching absolute regularization. Fitting-score standardization removes their intercept and positive scale differences, so identical blending weights also give identical scores. Standard-deviation floors can break the identity in degenerate cases; saved scores agree to numerical precision.

This equivalence depends on the fitting convention. Switching trace normalization from *total* to *within* covariance changes the absolute regularization. Unequal class priors, alternative shrinkage, nonlinear features and different score scaling also change the comparison. Our mixture assigns equal total weight to each class, so the identity applies to its weighted fitting distribution.

A.4 Why a covariance change may leave standalone EER unchanged

Let $M = K^{-1}$ for $K \succ 0$, and consider adding a rank-one covariance term aligned with a fixed direction d: $M_\beta = (K + \beta dd^\top)^{-1}$, $\beta \geq 0$. Define $u(z) = z^\top Md/\sqrt{z^\top Mz}$ and $a = d^\top Md$. Then

$$u_\beta(z) = \frac{u(z)}{1 + \beta a}\left(1 - \frac{\beta u(z)^2}{1 + \beta a}\right)^{-1/2}. \tag{13}$$

Cauchy–Schwarz gives $u(z)^2 \le a$. The derivative with respect to u is positive, so rankings and ties are preserved. A nonlinear monotone change may nevertheless change rankings after blending with another score. Fractional inverse powers, changing ridge scales, and finite noisy endpoint directions fall outside this identity.

For equally sized high/low-score fitting groups with means $m \pm \delta$, their uncentered moment is $W + mm^\top + \delta\delta^\top$, with W the biased within scatter. The implemented unbiased within covariance rescales W by $n/(n-2)$. Comparing the uncentered moment with within-group covariance therefore changes both mean terms and covariance scaling. Flipping labels while keeping the fitting examples fixed leaves the uncentered moment unchanged but changes the estimated contrast. These identities motivate varying the contrast and moment separately.

A.5 A fixed linear ensemble needs only one head

For head j, expanding the standardized blend of $a_j^\top z + b_j$ and the teacher gives $w_j^\top z + \beta_j t(a) + \gamma_j$. Averaging yields $\bar{w}^\top z + \bar{\beta} t(a) + \bar{\gamma}$ exactly, even when $t(a)$ is nonlinear in z. The averaged correction is therefore linear for fixed fitted heads; changing the calibration population changes its coefficients.

B Data, fitting information, and implementation

The following recipes specify source fitting, validation and scoring. Configurations are saved before evaluation; changes to source objectives, fitting counts or normalization are reported as separate controls.

B.1 Primary Common Voice protocol

Encoder and extraction. The frozen checkpoint is ReDimNet2 B6, `b6-vb2+vox2_v0-lm.pt`, pretrained on VoxBlink2 and VoxCeleb2. Extraction uses mono 16 kHz audio, at most the first 15 seconds, and batch size one. A 16-recording re-extraction check across eight locales finds a maximum absolute difference of 1.63×10^{-4} between normalized embedding coordinates under the available decoder/PyTorch stack.

Identity separation. Blocking every development/test native ID from every training locale removes 5,021 of 108,950 training IDs, leaving 103,929. The evaluated 51-locale development subset contains 528,782 recordings and 41,194 native IDs; strict test contains 524,087 recordings and 98,639 IDs, with IDs disjoint from development. IDs from other locales also enter the exclusion list. Separation is defined by dataset IDs; physical-person overlap across IDs and with encoder pretraining remains unresolved.

Center and covariance fitting. Each locale has three training pools capped at 16 recordings per speaker, seeds 12345, 112345 and 212345. The speaker mean weights training-speaker centroids equally; the utterance mean weights training recordings equally. Center candidates use penalties $\eta \in \{0, 0.1, 1, 10\}$ toward the filtered speaker mean, 150 Adam steps at learning rate 0.01, and 2,048 genuine plus 2,048 impostor identity-balanced pairs per step. The threshold starts at 0.2 and is fitted jointly. Centers are projected to norm at most one; candidate centers average the three fits.

For each center, training vectors are translated and renormalized. Within-speaker covariance uses $n - k$ degrees of freedom, where n is the recording count and k the speaker count; we average the three pool estimates. Oracle approximating shrinkage (OAS) targets a trace-matched identity or a globally filtered pooled within covariance. The feature-axis exponent is 0.1875 (metric exponent 0.375), with eigenvalue floor $\max\{10^{-6}\,\mathrm{tr}(C)/192, 10^{-12}\}$. Equal-locale development macro EER selects one of 23 configurations including raw cosine. The selected center has zero penalty and identity-target shrinkage. The selected transformations are fixed before strict-test scoring, with hashes recorded for verification.

For the averaged within covariance C, let $t = \mathrm{tr}(C)$, $u = \mathrm{tr}(C^2)$ and n be the integer mean of the three pool degrees of freedom. The shrinkage coefficient is

$$\rho = \mathrm{clip}_{[0,1]} \frac{(1 - 2/D)u + t^2}{(n + 1 - 2/D)\max\{u - t^2/D, 10^{-30}\}}, \quad D = 192.$$

The covariance is $(1-\rho)C+\rho T$, with $T=(t/D)I$ or the global raw within covariance scaled to trace t. The pooled-target variant reuses this coefficient as a fixed shrinkage rule. The 23 candidates comprise raw cosine, the across-locale speaker mean alone or with identity-target OAS, two zero-center maps (identity or pooled target), and six locale-specific centers crossed with three maps (identity map, identity-target OAS, or pooled-target shrinkage). The six centers are the utterance mean, the speaker mean, and the four supervised centers with $\eta \in \{0, 0.1, 1, 10\}$. Development selects the unpenalized supervised center with identity-target OAS, at 2.6992% macro EER. The component table below reports its strict-test result and the main alternatives.

Balanced trial sampling. Every anchor ID with at least two recordings contributes 256 genuine and 256 impostor comparisons. Genuine pairs are drawn without replacement when possible, otherwise with replacement. For impostors, we sample another ID uniformly within the locale, including IDs with one recording, then sample one of its recordings uniformly. This produces 20,037,632 development and 41,284,608 strict-test comparisons. All methods share them. The optimized sampler was checked for exact array equality with the original across 36 configurations.

Uncertainty and duplicate sensitivity. Five hundred paired Poisson bootstraps share one weight per global native ID across locales. Genuine comparisons use w_i and impostors $w_i w_j$. The EER-change interval is $[-0.9063, -0.8067]$ points. This accounts for trial identity dependence while conditioning on fitting, configuration selection and the locale inventory. Exact-vector duplicate sensitivity removes 130 evaluated development and 420 strict-test comparisons, rebalances anchor mass, and leaves both selection and the 28.38% reduction essentially unchanged. This sensitivity check covers exact embedding duplicates.

Table 3: Common Voice components with development and strict-test macro EER (%). Selection uses development data; strict-test columns report the component comparisons.

Configuration	Development	Strict test
Raw	3.7253	3.0269
Across-locale speaker mean	3.6792	2.9597
Across-locale mean + OAS	3.9871	3.2229
Metric only	3.6423	2.9265
Local utterance mean	2.9991	2.3563
Local speaker mean	2.9925	2.3495
Speaker mean + OAS	2.9567	2.3098
Optimized center	2.7524	2.2045
Optimized center + OAS (selected)	2.6992	2.1678
Optimized center + pooled-target shrinkage	2.7009	2.1644
Gaussian diagonal-within (separately selected)	3.4022	2.7224

Gaussian comparison. The Gaussian control estimates within covariance W and between-identity covariance B from the same pools, subtracting the centroid estimation contribution and clipping B to positive semidefinite. Its same-identity joint covariance has diagonal blocks $B+W$ and off-diagonal blocks B; different-identity covariance is block diagonal. The full quadratic log likelihood ratio retains self terms. Within-covariance diagonal shrinkage is selected from $0, 0.5, 1$, with fixed ridge $0.001\,\mathrm{tr}(W)/192$. The estimator fits these covariances by moments. Its quadratic agrees with direct joint-precision calculations, and its paired EER difference from the geometric selection is $[0.5099, 0.6105]$ points.

B.2 Detection features and source fitting

The frozen XLSR-SLS implementation supplies 24 transformer outputs. Layer-6 temporal mean and population standard deviation form 2,048 features. Audio is decoded at 16 kHz; the first 64,600 samples are scored, repeating shorter signals. The acoustic view comprises means and standard deviations of 40 linear-frequency cepstral coefficients and their first and second differences (240 dimensions), using pre-emphasis 0.97, 400-sample Hann windows, FFT size 512, hop 160, 80 linear

Table 4: EER (%) for all 51 strict-test locales. One configuration is selected on development data, fitted separately for each locale and fixed before strict-test scoring.

Locale	Raw	Adapted	Locale	Raw	Adapted	Locale	Raw	Adapted
ab	3.873	2.954	fr	1.111	0.930	ps	2.056	1.132
ady	2.734	2.478	fy-NL	2.196	1.830	pt	2.972	1.463
ar	2.912	2.152	gl	1.187	1.013	ro	0.894	0.912
ba	4.309	2.461	hu	1.517	1.160	ru	1.710	1.067
be	0.666	0.513	it	0.685	0.527	rw	11.863	9.230
ca	2.121	1.679	ja	9.830	4.651	sv-SE	2.040	1.343
ckb	1.738	1.239	ka	3.725	3.090	sw	8.983	6.800
cs	2.334	1.196	kab	4.783	3.500	ta	2.257	2.555
cy	3.496	3.390	kbd	3.565	2.557	th	4.169	2.011
de	0.613	0.521	kmr	2.921	1.749	tr	4.045	1.921
el	2.711	2.014	ky	4.862	3.948	ug	1.634	1.351
en	0.932	0.794	lg	4.914	3.521	uk	1.355	1.141
eo	2.873	2.530	lv	2.569	2.002	ur	5.307	5.288
es	0.699	0.543	mhr	4.511	2.471	yue	2.255	1.337
et	0.850	0.831	mn	2.930	2.698	zh-CN	2.059	1.216
eu	2.670	2.467	nl	1.091	0.777	zh-HK	4.212	1.966
fa	3.401	1.803	pl	2.291	1.728	zh-TW	2.940	2.109

triangular filters, a 10^{-10} log-power floor, and orthonormal DCT. This acoustic view supplies features additional to the XLSR representation.

CLIP UniversalFakeDetect provides 768 features; AIDE provides 1,024. The original image frontend and classifier are retained unless explicitly replaced by the five-crop condition. CLIP uses an RGB 224×224 center crop without resizing and CLIP channel normalization. AIDE uses four extreme DCT-selected 32×32 patches and the whole image, each resized to 256×256 with tensor antialiasing disabled and ImageNet normalization. Its features are the GELU activations before the final classifier. Original scores are oriented toward the synthetic class; AIDE uses fake minus real logits. The released UniversalFakeDetect head has SHA-256 prefix `47710745713bcc9`; AIDE uses the official `progan_train.pth` checkpoint (`ce3a9d66c124e4c2`). Exact checkpoint hashes remain unavailable for XLSR-SLS, the CLIP backbone and ArcFace buffalo_l.

The pseudo-label rule forms two fitting groups from the highest and lowest 5% of teacher scores, with at least 32 examples per side and at most half the available pool. We call these the high- and low-score groups. Source development chooses the rule, powers and blends. The learned speech condition uses the uncentered moment of these groups, power one, an angular score and blend 0.75. The acoustic condition uses within-group covariance, power 0.5 and blend 0.25. Calibration pools contain at most 4,096 rows in that setting; the supervised ensembles below use different label budgets. CLIP's source-selected moment/dot rule uses power one and blend 0.5; For AIDE, source validation selects the original teacher score.

The MLAAD cohort acquired for the primary results has 39,024 recordings, six languages, 60 generator names and 173 generator/language cells. Of these, 19,614 recordings are scored; the acquisition count includes data used for fitting and validation. Its source-content IDs are disjoint from the additional 77,653-recording cohort, but languages and generators recur. Two complementary fits exclude their corresponding scoring generators while retaining content-disjoint calibration/evaluation assignments. Generator exclusion is defined by names; technology families may still overlap.

Synthbuster has 1,000 source-image stems, each linking a real image to its generated variants. Hash grouping assigns 502 stems to calibration and 498 to scoring, keeping all variants of a stem together. Each scoring generator cell has 498 real and 498 fake images; the same real images recur across cells, creating dependence. The supervised generator-excluded fits use these fixed assignments with complementary generator sets. Macro EER averages the nine generator-cell EERs.

B.3 TEACHER-DERIVED AND LABELED DIRECTIONS

The high- and low-score groups estimate a direction without ground-truth fitting labels. Source labels still select the rule. On the primary MLAAD scoring cohort, whose content is disjoint from the additional 77,653-recording cohort, the fixed learned-feature rule changes EER from 21.0701% to 18.4491%; using the same high/low-group contrast with an identity metric instead gives 27.0329%. Using the dot score without the angular denominator reaches approximately 18.276%. The gain persists with the covariance-weighted dot score $z^\top Md$. The acoustic view behaves differently: its identity-metric score is better than its weighted score. The usefulness of covariance weighting therefore depends on the feature view.

Labeled calibration uses class structure in the fitting distribution to estimate a scoring direction. Under the generator/content-excluded protocol, a specialist full-feature ridge head reaches 1.3832% on MLAAD and 4.0830% on CLIP/Synthbuster, compared with respective original scores of 21.0701% and 25.7921%. These comparisons change both the fitting supervision and the generator-exclusion protocol.

B.4 SUPERVISED SPECIALIST AND MIXTURE SELECTION

Ridge uses unit inputs, class-balanced or domain–class-balanced weights, a weighted center and unpenalized intercept. The absolute ridge is $\lambda = \ell \operatorname{tr}(S_T)/D$ for $\ell \in \{0.01, 0.1, 1, 10\}$, with numerical protection for a degenerate trace. Blends are $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$. All scales are fitted before scoring. Labels stored as $0/1$ are mapped to $-1/+1$ for ridge. Standardization subtracts the fitting-score mean and divides by its fitting-score standard deviation, floored at 10^{-6}. The single-source specialist uses unweighted score-standardization moments over its fitting rows; the mixture uses the domain–class weights. Comparisons of complete recipes therefore change more than source data alone.

Allowed calibration generator names are ordered by a fixed hash and distributed over three inner folds. Source-content IDs independently receive a hash modulo three. An inner fit excludes both the validation generator group and validation contents. Validation uses authentic examples from the content fold plus fake examples in both that content and generator fold. The hash assignments fix validation membership within the permitted calibration pool. Final fitting uses all allowed calibration rows. Specialist selection minimizes mean generator validation EER; a separately reported conservative variant minimizes the worst quarter. Both include the original teacher.

Mixtures add the same 4,096 ITW calibration rows to each speech fit, or 4,000 CNN source-fit rows to each image fit. ITW inner validation uses three stratified component-grouped folds with seed 20260919. Image validation uses the separate fixed 4,000 CNN rows. Each of the four source–class groups receives $1/4$ of the total fitting weight. For candidate h, let $e_d(h)$ be source-domain EER averaged over the specified validation blocks and then folds. The primary selection risk is

$$R(h) = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \frac{e_d(h)}{\max\{e_d(t), 10^{-3}\}}, \qquad e_d \text{ in percent.} \tag{14}$$

Each domain has the same original-score denominator across candidates. A nearly perfect baseline can therefore give a domain high relative weight. The 10^{-3}-percentage-point floor belongs to the primary recipe; the one-percentage-point floor and absolute-EER objectives in Appendix F are separate controls. A second configuration requires mean EER in each source domain to be at or below its baseline.

Ties prefer lower blend, then stronger regularization, then a deterministic key. External scores equally average twelve speech models (six languages, two complementary generator-exclusion fits) or two image models. Each model is an ensemble member. Per speech head, MLAAD fitting uses 1,106–3,330 labels; across heads there are 18,442 unique MLAAD labels. Per image head, Synthbuster uses 2,059–2,438 labels, totaling 4,096 unique labels across heads. Adding source-dev labels used in selection is necessary for any complete information budget. Evaluation of a MLAAD cell uses only its corresponding generator-excluded model.

Table 5: EER (%) after transfer, with the original encoder frontends. Each specialist and mixed-source procedure is fixed across evaluation datasets. CoSG uses the full release; CNN and Diffusion use the full preserved archives; MLAAD, ITW and Synthbuster retain separated scoring subsets. Speech DF-Eval uses the first window. Validation data differ between these procedures; the controls in Table 2 use the same validation data.

Encoder	Dataset	Original	Specialist	Mixture
XLSR-SLS	MLAAD	21.0701	1.3832	1.6449
XLSR-SLS	ITW	6.6092	13.6015	2.2264
XLSR-SLS	CoSG	20.7852	10.0462	10.0462
XLSR-SLS	DF-Eval	40.7002	31.5098	32.6284
CLIP	Synthbuster	25.7921	4.0830	17.3806
CLIP	CNN (13 groups)	11.0629	26.8323	10.4855
CLIP	Diffusion (8 groups)	10.7750	38.4000	10.9000
CLIP	DF-Eval	52.5988	44.2623	50.4918
AIDE	Synthbuster	56.5596	11.9366	33.2218
AIDE	CNN (13 groups)	10.2941	30.3643	7.5297
AIDE	Diffusion (8 groups)	2.7125	31.6750	2.5625
AIDE	DF-Eval	48.1967	41.7879	43.2787

B.5 Matched source/readout experiment

The protocol for Table 2 was fixed before its scoring, after earlier public outcomes had been inspected. It reuses the exact twelve calibration memberships, inner generator/content folds, ITW component folds, ridge grid, and weighted score-standardization procedure. Five families are always reported: MLAAD-only ridge, ITW-only ridge, mixed ridge, mixed logistic and mixed mean contrast. Every family is selected by the same two-domain relative-EER criterion. Single-source fitting balances its classes; mixed fitting balances domain–class cells. Consequently all families have access to both validation label sets while only specified fitting rows estimate coefficients.

Logistic regression minimizes weighted binary log loss plus $\lambda \|w\|^2 /2$ with an unpenalized intercept. It uses the same four covariance-scaled absolute penalties and blend grid, L-BFGS tolerance 10^{-8} and at most 2,500 iterations. The candidate counts match; effective regularization depends on both the penalty and the loss. All retained fits converge. The mean-contrast head uses the weighted positive-minus-negative means directly, with the same blend selection.

Conditional uncertainty. Two thousand paired evaluation person/component bootstrap resamples, with 26 ITW components, give mixture-minus-reference EER intervals: original $[-8.784, -2.621]$, matched ITW ridge $[-3.817, 0.183]$, matched MLAAD ridge $[-6.251, -2.006]$, and mixed logistic $[-2.681, -0.282]$ percentage points. These intervals describe evaluation variation conditional on the source fits and selections.

C Transfer, aggregation, and coverage

The specialist's 1.3832% MLAAD result uses single-domain validation. Table 2's MLAAD-only fit uses joint MLAAD/ITW validation, reaches 5.2341%, and transfers better to ITW without fitting on ITW. The comparison jointly changes validation access and score standardization.

CoSG. There are 1,797 recordings in 1,761 exact decoded-audio groups; the split used for the threshold diagnostic has 872 calibration and 925 evaluation rows. The split separates exact decoded content; authentic-speaker overlap remains unresolved. The external-transfer table scores all 1,797 released recordings using models fitted entirely on separate sources. The transferred-threshold diagnostic retains the 925-row evaluation subset; its baseline refers to that smaller population.

Deepfake-Eval. The dataset revision is `964763ccdbb659ca1cf40dcc8ba31476c3bb14a2`. All 1,980 audio recordings are represented, including 788 official Test rows. Test image coverage is 786/791 and all-split coverage 1,952/1,975. Twenty-three

metadata filenames remain unresolved. A known exact duplicate with conflicting supplied labels is retained. All its rows, including those flagged Train, are evaluation-only in this study.

Three-window audio averages scores from starts 0, $\lfloor (L - 64600)/2 \rfloor$ and $\max(0, L - 64600)$ for sufficiently long signals, repeating shorter signals as above. The rule was fixed after individual-window results were known, making this an exploratory coverage control. Median duration is approximately 51.37 seconds and the maximum around 5,633 seconds; three short windows sample only part of longer recordings.

Three-window teacher/specialist/mixture EERs are 38.2932/29.3217/30.6346%; first-window results appear in the transfer table. The specialist mean-window AUROC is approximately 0.7608 versus teacher 0.6513. Several image results remain near chance.

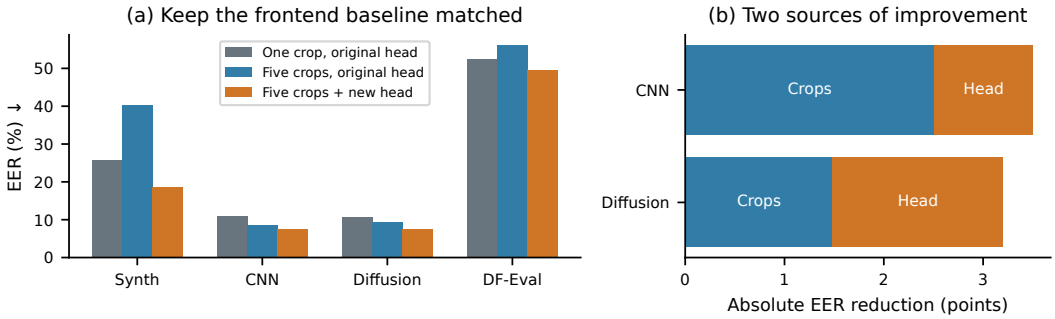

Figure 6: **Five-crop extraction supplies most of the CLIP improvement on CNN.** Five-crop averaging improves the original CLIP head on CNN/Diffusion but worsens it on Synthbuster/DF-Eval. On full CNN, 2.510 of the net 3.505 EER points come from preprocessing; the new head supplies 0.995. These contributions are measured by changing the frontend first, then fitting the new head.

Preprocessing can supply much of the apparent head gain. Against the matching five-crop baseline, the head reduces CNN EER by 11.63% and Diffusion EER by 18.44%. Figure 6 separates these gains from the preprocessing contribution.

The fixed frontend resizes the short side to 256 and takes five 224×224 crops. It averages unnormalized frozen features for the original classifier; the head uses the unit-normalized aggregate. Five encoder evaluations are required. A source-only gate compares center-crop and five-crop source validation using common original-baseline denominators. Its mean risk is 0.8553 versus 0.6384. In both directions, five-crop increases Synthbuster validation error (16.226 to 22.864; 15.615 to 19.451) while decreasing CNN error (0.183/0.217 to 0.067). CNN's original validation EER is only 0.2%, making relative-risk weighting sensitive to small changes. The frontend selected on source validation is used for every evaluation dataset.

Five-crop teacher versus adapted EER is 40.4061 versus 18.7416 on Synthbuster, 8.5526 versus 7.5579 on full CNN, 9.2875 versus 7.5750 on full Diffusion, and 56.0656 versus 49.5082 on Deepfake-Eval images. These use a different frontend from the center-crop transfer table.

D COLLECTIVE VERIFICATION: EQUATIONS, CONTROLS, AND FAILURES

D.1 OBSERVED COLLECTIVE GAINS AND THEIR LIMITS

Collective verification treats recordings or image templates as nodes and requested trials as edges. ITQ fits binary representations (Gong & Lazebnik, 2011); graph correction standardizes a score using neighboring comparisons. Table 6 compares scores computed from the same extracted features. The controls below test their dependence on this additional context.

Figure 7 shows why EER alone is insufficient: IJB-C improves on both metrics, whereas CPLFW's lower EER accompanies worse low-FAR performance. The full face table and failure controls below quantify this distinction.

Table 6: **Verification using unlabeled trial embeddings and neighboring comparisons.** EER (%). Vox uses the matched v22 extraction; both Vox and IJB-C use unlabeled endpoint adaptation and complete official trial-graph context.

Dataset / encoder	Raw	ITQ + graph	Relative reduction
Vox1-O / ReDimNet2	0.1754	0.0213	87.87%
Vox1-E / ReDimNet2	0.3401	0.1735	48.99%
Vox1-H / ReDimNet2	0.6812	0.2923	57.09%
IJB-C / ArcFace	1.1965	0.9511	20.51%
IJB-C / AdaFace	1.4828	1.3243	10.69%

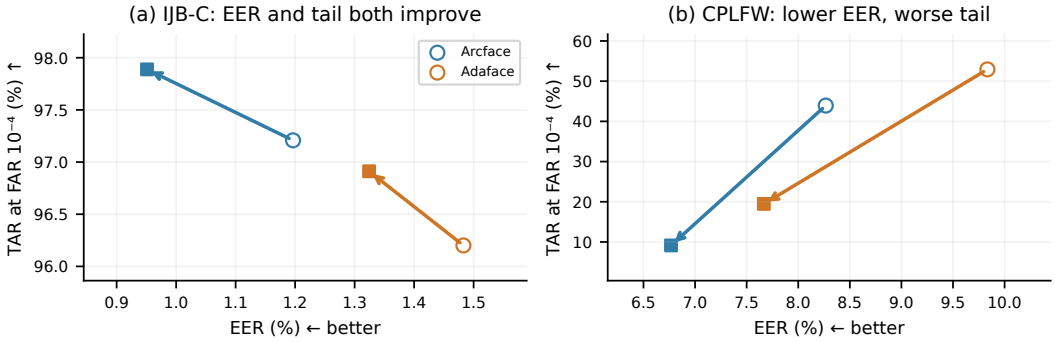

Figure 7: **Lower EER can accompany worse performance at low false-accept rates.** Open circles show raw scores; filled squares show ITQ fitted on trial endpoints followed by graph scoring. Both encoders improve IJB-C EER and low-FAR TAR; CPLFW moves toward lower EER but substantially worse low-FAR TAR. Arrows connect observed results.

ITQ fits its representation from embedding rows (Gong & Lazebnik, 2011). Cohort normalization instead adjusts scores using reference recordings, whose selection matters under language mismatch (Matějka et al., 2017). Our correction uses requested neighboring trials as its reference. The independent source-cohort control below tests this distinction.

D.2 Representation and contextual score

For centered, normalized fitting rows X, each ITQ block alternates $B = \mathrm{sign}(XR)$ and an orthogonal Procrustes update minimizing $\|B - XR\|_F^2$. An orthogonal rotation alone preserves cosine; its effect here comes through sign quantization. The binary score is the mean product of corresponding $\{-1, +1\}$ bits across concatenated blocks: matching bits contribute positively and mismatches negatively. Multiple fixed-seed scores are averaged.

For edge $e = (i, j)$ with score s_e, let $N_i \setminus e$ be the other incident edges. The global score mean m and variance v stabilize estimates for nodes with few comparisons. A pseudo-count κ gives

$$\mu_{i,-e} = \frac{\sum_{r \in N_i \setminus e} s_r + \kappa m}{|N_i| - 1 + \kappa}, \tag{15}$$

$$\sigma^2_{i,-e} = \max\left\{ \frac{\sum_{r \in N_i \setminus e} s_r^2 + \kappa(v + m^2)}{|N_i| - 1 + \kappa} - \mu^2_{i,-e},\ fv \right\}, \tag{16}$$

$$s_e^{\mathrm{graph}} = \tfrac{1}{2}\left(\frac{s_e - \mu_{i,-e}}{\sigma_{i,-e}} + \frac{s_e - \mu_{j,-e}}{\sigma_{j,-e}} \right). \tag{17}$$

The transferred reference uses $\kappa = 2$, $f = 10^{-3}$; global variance has numerical floor 10^{-8}. The queried edge is excluded from local moments but remains in the global prior. The separate five-fold edge audit removes evaluation edges from the entire context used in that diagnostic. A separate control also excludes parallel edges.

D.3 VoxCeleb protocol and graph sensitivity

The provenance tag v22 identifies the matched extraction and hard-ITQ scoring configuration. The rule uses 6,144 bits from 32 independent 192-dimensional rotations, averaging signed-Hamming scores over seeds 12345, 112345 and 212345, with 25 Procrustes iterations per block and unique-endpoint weighting. Five speaker-disjoint Vox2 folds select the configuration; unlabeled trial endpoints then fit the rotations. This fitting is transductive; scoring uses the graph-corrected binary similarity alone. Table 6 uses complete-tie EER interpolation; baseline comparisons use the same crossing convention and extraction pipeline.

Graph controls. The gain depends on which comparisons surround each trial. The audit verifies official-label alignment, independent ROC interpolation, exact score replay and invariance of computed scores to permuting trial labels. All EER triples below follow the O/E/H order. Removing reverse/parallel duplicates gives approximately 0.0215/0.1746/0.2915. Five-fold edge exclusion gives 0.0266/0.2239/0.3812. Endpoint median graph degrees are 16/8/8, with median genuine degree 8/4/4: neighboring genuine examples are abundant. Rewiring yields approximately 0.0574/0.3411/0.7014; reducing genuine context to 1% yields 0.5296/0.6027/0.8726. Replacing target-neighbor scores at the same degree also erases much of the gain. These topology diagnostics use trial labels to analyze which neighboring comparisons supply the gain.

Independent source-cohort comparison. A separate source-cohort experiment uses 523,776 development pairs from 1,024 recordings and 256 identities, with 1,536 genuine pairs. Development selects adaptive symmetric score normalization (AS-Norm), with $k = 50$ and blend one, and official O/E/H EER regresses from its own matched 0.1701/0.3391/0.6788 baseline to 0.2074/0.3622/0.7365.

D.4 Face fitting and full outcomes

Feature and rotation fitting. ArcFace and AdaFace (Deng et al., 2018; Kim et al., 2022) supply 512-dimensional embeddings. Experiments use custom VGGFace2 pairs and the unchanged CALFW, CPLFW, AgeDB and IJB-C trials (Cao et al., 2017; Zheng et al., 2017; Zheng & Deng, 2018; Moschoglou et al., 2017; Maze et al., 2018). The face ITQ transfer uses twelve 512-dimensional blocks (6,144 bits), 25 Procrustes iterations and seeds 12345, 112345 and 212345. Source fits use 64,000 VGGFace2 images from 2,000 training identities; endpoint fits use each benchmark's unique unlabeled image/template features. Inputs are normalized, the fitting mean is subtracted, and rows are normalized again. Exact decoded-image hashes merge repeated endpoints for the pair datasets. When fitting on trial endpoints, both the mean and rotations are estimated from unlabeled embeddings.

IJB-C extraction and template aggregation. IJB-C retains all 469,375 images, 89,524 template/media groups, 23,124 templates and 15,658,489 official pairs (19,557 genuine). Five-landmark alignment gives 112×112 crops. AdaFace IR50 MS1MV2 uses BGR; ArcFace buffalo_l ResNet50 uses RGB; both scale to $[-1, 1]$. Original and flipped raw embeddings are summed, normalized and weighted by detector confidence; media means are summed within template and normalized. Post-hoc adaptation follows template aggregation. All official pairs are retained. Encoder-pretraining and source-identity overlap remain unresolved.

Rotation controls. Fitted ITQ rotations outperform both controls on IJB-C. With the same graph score, centered continuous embeddings, matched random sign rotations and fitted ITQ give EERs of 1.0817/1.1147/0.9511 for ArcFace and 1.4726/1.4522/1.3243 for AdaFace, respectively. Continuous-feature graph scores also improve CALFW/CPLFW substantially.

Low-FAR and graph failures. An extreme official CPLFW impostor (row 1,766) controls the zero-false-accept threshold. Source-selected shrinkage moves ArcFace ITQ+graph EER/TAR from 6.7667/9.1333 to 7.6667/23.3667; raw TAR is 43.9333. The development constraint on TAR used FAR 10^{-3}. At destination FAR 10^{-4}, the corrected score remains below baseline; the official row is retained.

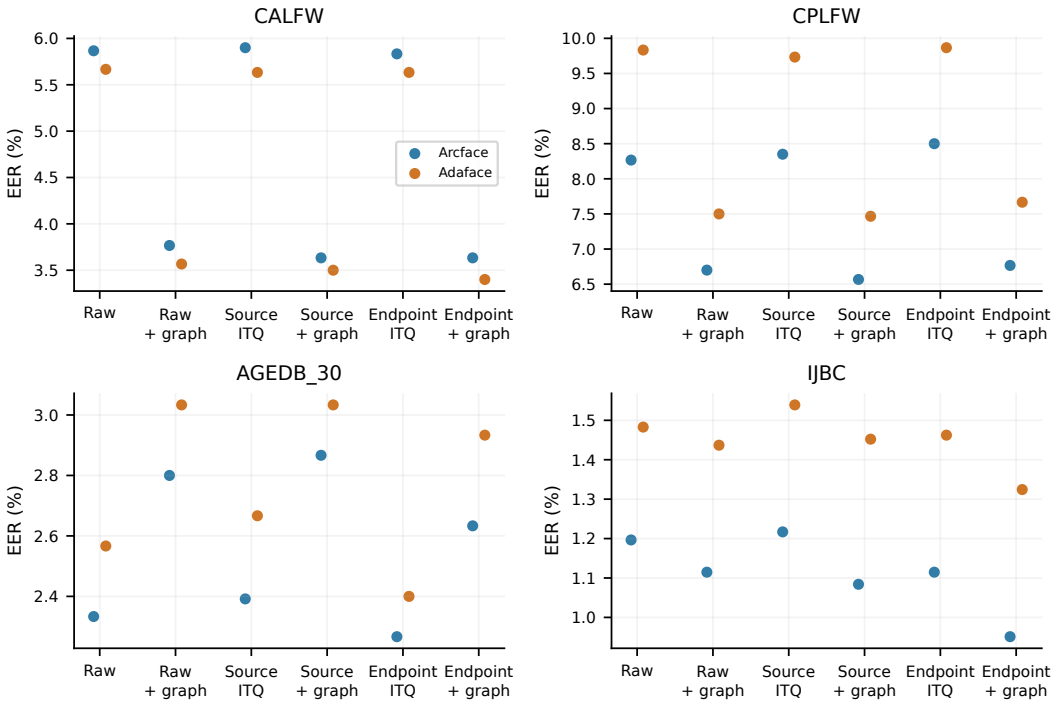

Figure 8: All eight face encoder/benchmark combinations under the transferred ITQ/graph recipe. The figure compares fitting on source data or trial endpoints, with independent or collective scoring. Both encoders regress on AgeDB.

Table 7: Face EER and TAR at low FAR (%). ITQ fits on unlabeled trial endpoints; graph scoring also uses neighboring comparisons. Raw and adapted scores use the same official trials.

Encoder	Benchmark	Raw EER	ITQ+graph EER	Raw TAR	ITQ+graph TAR
				FAR 10^{-4}	
ArcFace	CALFW	5.8667	3.6333	89.467	91.767
ArcFace	CPLFW	8.2667	6.7667	43.933	9.133
ArcFace	AgeDB-30	2.3333	2.6333	93.667	85.500
ArcFace	IJB-C	1.1965	0.9511	97.208	97.888
AdaFace	CALFW	5.6667	3.4000	89.600	91.000
AdaFace	CPLFW	9.8333	7.6667	52.933	19.467
AdaFace	AgeDB-30	2.5667	2.9333	93.500	91.233
AdaFace	IJB-C	1.4828	1.3243	96.201	96.912

AgeDB includes 281 repeated unordered edges covering 543 rows, all negative. Excluding all parallel counterparts from context leaves the endpoint-ITQ graph EER unchanged at 2.6333/2.9333 for ArcFace/AdaFace. Endpoint degree and class are associated in these constructed pair lists; IJB-C is much denser, with median degree approximately 809 and about 0.125% genuine edges. Graph behavior therefore depends on properties beyond the fraction of genuine edges.

Independent-pair adaptation. A separate VGGFace2 experiment fits a metric for independent pair scoring. It uses 480,000 test pairs (96,000 genuine), 500 scoring identities with 48 images each, and the 2,000-identity source pool above. At fixed zero center, within-class metric adaptation reduces ArcFace 5.7896 to 5.0542 and AdaFace 6.5698 to 6.0167. Its performance at low false-accept rates is weaker, and the original angular metric does not improve IJB-C. A separate Gaussian score can improve IJB-C but is a distinct score family. An external source identity-centroid cohort gives 5.7781/6.5628 on the custom pairs, supplying a much smaller gain.

E SENSITIVITY TO WAVEFORM PERTURBATIONS

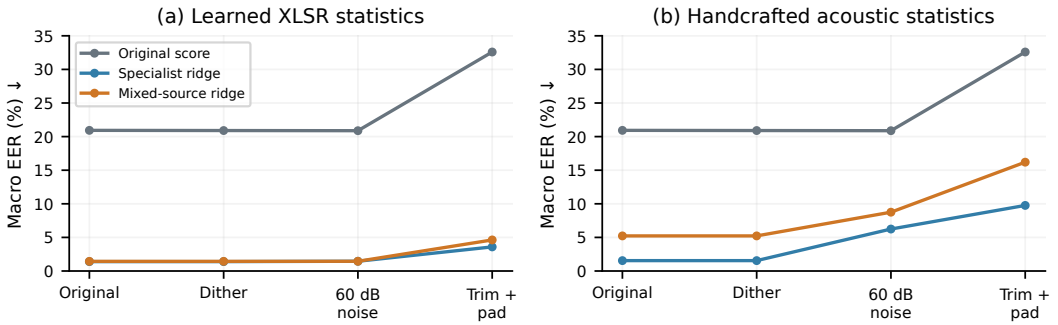

Figure 9: Fixed fitted heads under matched cohort-1 perturbations. The learned view tolerates small numerical/noise changes better than the acoustic view. Trimming and fixed padding affect both. Fitted parameters and selected recipes stay fixed across perturbations.

Models fitted on the later MLAAD cohort (cohort 2) score the earlier, content-disjoint cohort (cohort 1). Each recording is scored by the model assigned to its language and generator-exclusion group. The scoring generator is excluded from its corresponding fit; generators absent from both fitting inventories use the declared average. Baseline macro EER is 20.9301 original, 20.9022 with micro-dither, 20.8785 with 60 dB noise, and 32.5885 after trimming/fixed padding. Learned specialist ridge gives 1.3980/1.3980/1.4490/3.5874; learned mixture 1.4344/1.4344/1.4461/4.6220. Acoustic specialist ridge gives 1.5452/1.5452/6.2441/9.7596. These comparisons preserve waveform-condition membership and scoring rules.

The learned-feature gain survives small numerical perturbations, while acquisition shortcuts remain a possible influence (Müller et al., 2021; Smeu et al., 2025). The acoustic view is more sensitive to added noise; both views are sensitive to trimming and padding.

F REFITTING, SOURCE OBJECTIVES, AND TRANSFERRED THRESHOLDS

These follow-ups vary fitting samples, selection objectives and thresholds while retaining source memberships. Earlier public results motivated the questions; each protocol was fixed before computation. CNN/Diffusion use full preserved archives; speech subsets are specified below. Each procedure is retained across destinations.

F.1 DOES THE SUPERVISED SPEECH GAIN SURVIVE SOURCE REFITTING?

To test sensitivity to the fitting sample, we fix the twelve selected language/generator-exclusion configurations and refit their means, covariance, ridge directions and score standardizers. One hundred paired bootstrap draws resample 8,371 MLAAD original-content groups and 28 ITW calibration groups linked by person or content. The mixture and matched ITW-only fits share each source draw. A separate draw resamples the 26 disjoint ITW evaluation components. Each draw rebalances domain/class weights and retains both classes. Before resampling, the implementation reproduces the saved coefficients and scores.

With the evaluation population fixed, the 95% percentile intervals for mixed-minus-original and mixed-minus-ITW-only EER are $[-4.688, -3.352]$ and $[-1.787, -0.523]$ percentage points. Also resampling the evaluation groups gives $[-8.288, -2.299]$ and $[-3.851, 0.421]$, respectively. The gain over the original score remains supported. The advantage over matched ITW-only fitting becomes inconclusive when evaluation variation is included.

These estimates use 100 draws and condition on the observed source generators and selected hyperparameters. Model selection and the encoder remain fixed; physical-person overlap between corpora remains unresolved.

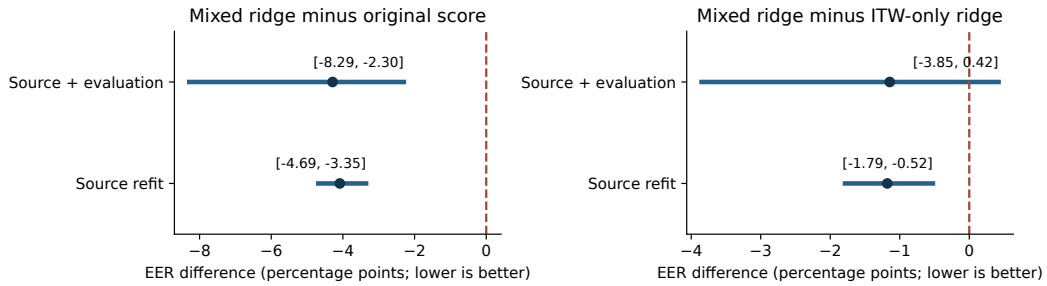

Figure 10: The gain over the original speech detector persists when fitting and evaluation groups are resampled. Lines are bootstrap percentile intervals; dots are bootstrap medians. Including variation in the evaluation groups makes the advantage over matched single-source fitting inconclusive.

F.2 Can a source threshold be transferred unchanged?

For each fixed speech configuration, we refit on the original three source folds and average the twelve heads on the corresponding ITW validation fold. Each of the 4,096 calibration rows is predicted by models that excluded its component during fitting. Because the same folds also select configurations, the out-of-fold predictions remain conditional on that selection.

We save three thresholds per method: source FPR at most 1% or 5%, and the source threshold minimizing the difference between FPR and false-negative rate (FNR). Scores at or above a threshold are classified as fake. Table 8 shows what happens when the 1% source rule is applied unchanged. Methods share the rule for choosing a threshold, but their resulting target FPRs differ. Deepfake-Eval uses the first audio window here.

Table 8: FPR and FNR (%) at thresholds selected for at most 1% source FPR and transferred unchanged. CoSG uses the 925-row subset; DF-Eval uses the first audio window.

	Original detector		ITW-only ridge		Mixed ridge	
Population	FPR	FNR	FPR	FNR	FPR	FNR
ITW	0.166	52.235	1.207	5.715	0.997	3.480
CoSG subset	3.761	85.412	5.310	61.522	6.195	24.736
DF-Eval audio Test	3.282	92.447	3.939	89.426	2.845	81.571

F.3 Matched image heads and source-objective sensitivity

The matched image controls retain the original CLIP/AIDE frontends, two complementary Synthbuster generator splits, three generator/content-excluded inner folds, CNN memberships, weighted standardizers and parameter grids. All families use both validation domains, including single-source fits. Ridge and logistic use the same absolute penalties, whose effective strength depends on the loss; the mean contrast omits covariance weighting.

For mixed ridge, we compare two additional source-only selection objectives: mean absolute domain EER, and mean relative domain EER with denominators floored at one percentage point. The original relative-EER implementation reproduces its saved choices and scores. Figure 11 reports every alternative.

Changing only the objective can improve some destinations while degrading others. For CLIP, absolute-EER selection changes EER from 17.381 to 8.255% on Synthbuster and 50.492 to 41.967% on Deepfake-Eval, but from 10.486 to 11.880% on full CNN and 10.900 to 12.538% on full Diffusion. The relative performance of ridge and logistic regression also varies across destinations. EERs near 40% remain weak forensic performance despite improving on the original score.

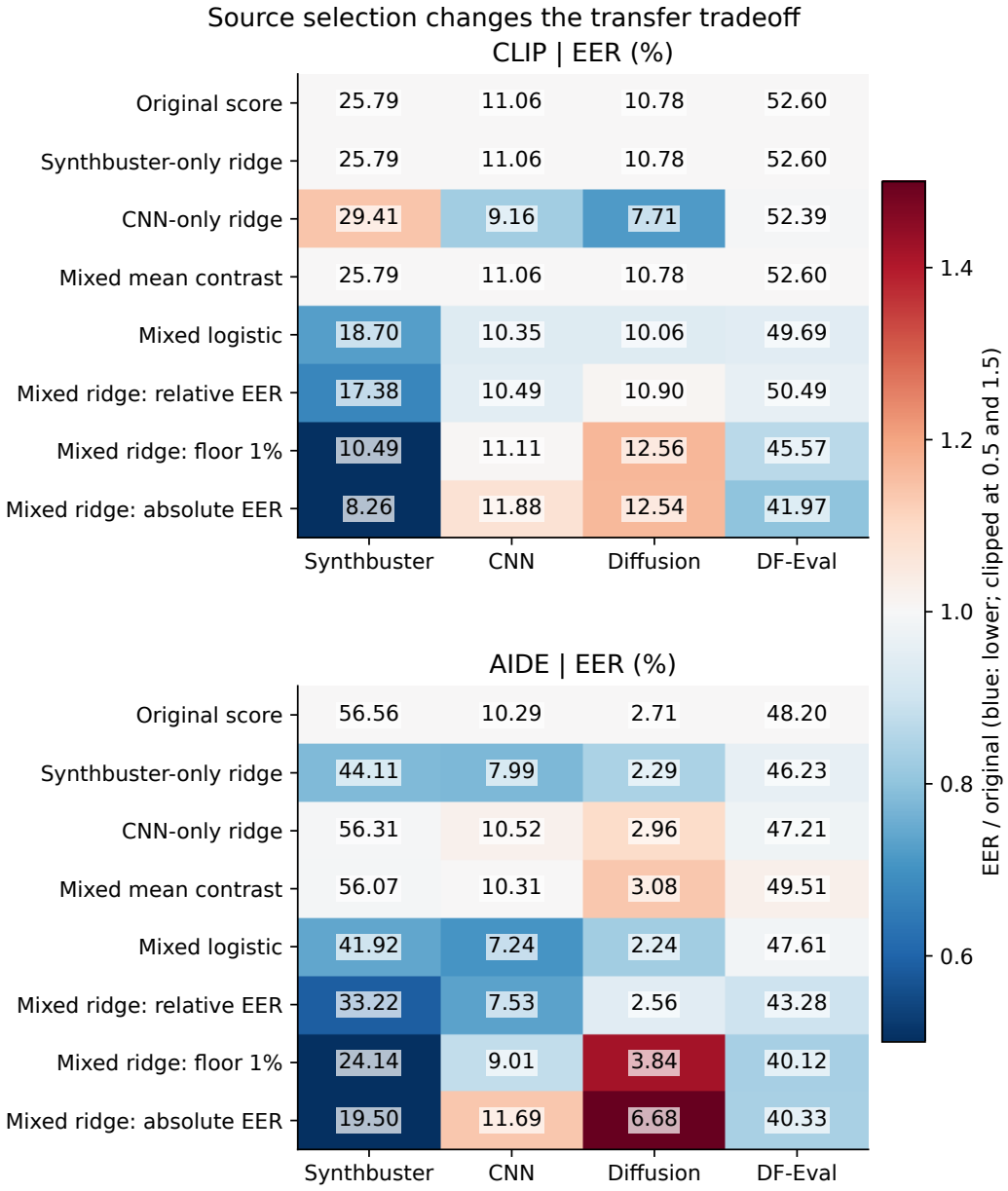

Figure 11: **Changing the source selection objective changes transfer performance.** Entries are EER percentages, with color showing the ratio to each matched original score. Both encoders retain their original frontends. Synthbuster averages EER across generators; CNN and Diffusion average all 13 and 8 full-archive scoring populations. Deepfake-Eval uses 786 resolvable official-Test images. Single-source fits use validation data from both domains; specialists validate on their fitting domain alone.

F.4 CONDITIONAL SENSITIVITY OF FACE BENCHMARK SCORES

For CALFW, CPLFW and AgeDB, decoded-image hashes identify trial endpoints, and genuine trials connect images into identity components. All labeled impostor trials connect different components. The components reflect observed genuine links; a person may span multiple components.

In 2,000 paired Poisson(1) draws, a trial within one component receives its component weight; a trial between components receives the product of their weights. ITQ fits and graph scores remain fixed. The intervals describe sensitivity to the observed components, excluding fitting and graph-topology uncertainty. The resampling unit is the observed component. Subject-level uncertainty for IJB-C remains unestimated because subject annotations were unavailable.

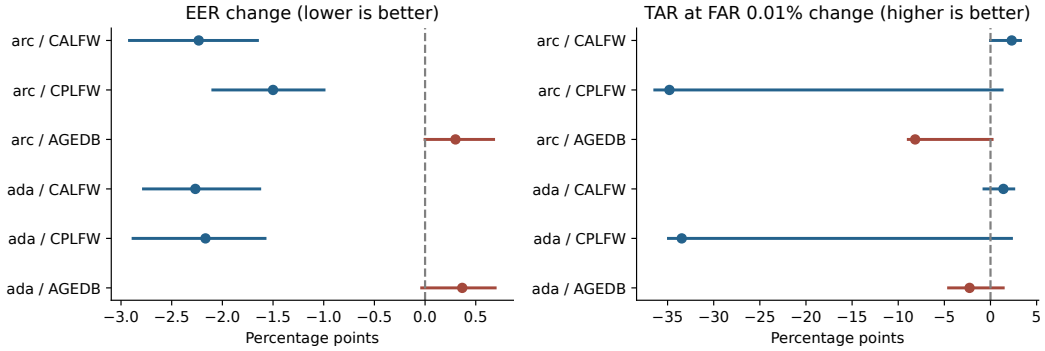

Figure 12: Changes from raw scoring to ITQ fitted on trial endpoints plus graph scoring. Dots are observed changes; lines are percentile intervals obtained by resampling linked image groups while holding fits and graph scores fixed. CALFW/CPLFW EER gains persist under this diagnostic, while AgeDB lacks a reliable improvement. CPLFW's low-FAR intervals are wide, consistent with its extreme official impostor controlling the original zero-false-accept threshold. The trial is retained.

G FIXED-DIRECTION INTERVENTIONS AND FULL-POPULATION UNCERTAINTY

We vary normalization and blending for the twelve speech readouts and two per image encoder from the selected mixed ridge configuration. All sixteen give nonzero weight to the fitted head and retain their selected regularizers. For weighted source mean μ, covariance C and absolute ridge λ, write $K = C + \lambda I$, $d = \sum_i \omega_i (z_i - \mu) y_i$ and $w = K^{-1} d$. The readouts are

$$h_{\mathrm{dot}}(z) = (z - \mu)^\top w, \qquad h_{\mathrm{ang}}(z) = \frac{(z - \mu)^\top w}{\sqrt{(z - \mu)^\top K^{-1} (z - \mu)}}.$$

The direction normalization is a constant absorbed by source standardization. Before inversion, eigenvalues are floored at 10^{-15}; squared radii are floored at 10^{-12}. Both readouts are standardized with the original domain/class weights on fitting rows.

Head-only scoring sets every member's $\alpha = 1$; the blended conditions retain every member's original α. When members have different original blends, this intervention also changes their relative head contributions.

Generator exclusions and external averaging are unchanged. These controls inherit dot-blend hyperparameters. Results across the three encoders share some scoring populations.

The CoSG uncertainty calculation uses all 1,797 released recordings and 1,761 decoded-audio groups. In 2,000 paired Poisson group-weight draws, dot-blend minus original EER has 95% percentile interval $[-13.093, -8.917]$ points; head-only dot minus original has $[-13.059, -8.251]$. Head-only angular minus dot has $[-0.450, 0.224]$ and angular-blend minus dot-blend has $[-0.214, 0.945]$. The intervals support the large improvement over the original score, while the small differences due to normalization remain unresolved. They hold source fits and selections fixed and resample decoded-content groups. Speaker dependence beyond those groups remains unresolved.

Table 9: Fixed-direction diagnostic, EER (%). Dot and angular columns set every member blend to one. Blend columns use the inherited source-selected α. The angular diagnostic retains the selected regularizer and blend. CNN/Diffusion use the full preserved archives and CoSG the full release; DF-Eval uses official Test with first-window audio.

Encoder	Population	Original	Dot	Dot blend	Angular	Angular blend
XLSR-SLS	MLAAD	21.070	1.658	1.645	1.222	1.386
XLSR-SLS	ITW	6.609	2.714	2.226	2.746	2.331
XLSR-SLS	CoSG	20.785	10.277	10.046	10.162	10.277
XLSR-SLS	DF-Eval	40.700	32.931	32.628	34.136	33.479
CLIP	Synthbuster	25.792	5.556	17.381	4.953	17.224
CLIP	CNN	11.063	16.132	10.486	16.550	10.585
CLIP	Diffusion	10.775	19.975	10.900	19.200	10.675
CLIP	DF-Eval	52.599	41.996	50.492	42.623	51.143
AIDE	Synthbuster	56.560	19.500	33.222	17.269	34.092
AIDE	CNN	10.294	11.695	7.530	11.747	7.689
AIDE	Diffusion	2.712	6.675	2.562	6.613	2.400
AIDE	DF-Eval	48.197	40.333	43.279	40.984	44.906

H Controls for fitting counts, classical heads, and scoring budgets

H.1 Source diversity at matched per-member fitting counts

We fit each source condition with 128, 256 or 512 total labeled rows per member, using three sampling seeds and the same validation data from both domains. Mixtures allocate half the fitting rows to each source, with balanced classes within each source. These controls use mean absolute source-domain EER for selection. The 512-row condition therefore differs from Table 2 in both source criterion and row budget; the table below compares the budget-controlled procedures with each other.

Table 10: Equal fitting rows per member: EER mean [minimum, maximum] across three sampling seeds. Ranges summarize variation across the sampled seeds. Source selection uses absolute EER; Table 2 uses relative EER.

Fitting population	Full CoSG	DF-Eval audio Test
MLAAD, 512 rows	15.075 [14.393, 15.704]	37.199 [36.761, 37.637]
ITW, 512 rows	20.457 [20.301, 20.554]	39.825 [39.606, 40.044]
Mixture, 256 + 256 rows	11.547 [11.085, 12.125]	34.256 [32.326, 35.650]

At 512 rows per member, the twelve-member ensemble uses 4,608 unique MLAAD rows in the MLAAD-only condition, 512 unique ITW rows in the ITW-only condition, and 2,560 unique rows (2,304 MLAAD plus 256 ITW) in the mixture. Reusing ITW examples across members produces different ensemble totals despite equal per-member fitting counts. Validation labels are additional and shared across conditions.

Across 108 frozen 512-row members, mixed fitting improves mean member EER over MLAAD-only by 1.29–2.23 points on full CoSG and 2.36–4.98 on official DF-Eval across three seeds. All fitting uses source data; source diversity, relevance and quality vary together.

H.2 A matched Gaussian-family control

The image comparison retains nearest-class-mean, diagonal Gaussian, shared-covariance linear discriminant, and class-specific quadratic discriminant alternatives, along with ridge and logistic readouts. It uses the same unit features, source memberships and teacher-blend access. Selecting the Gaussian family by lowest mean absolute source EER chooses shared-covariance LDA for both CLIP and AIDE in each of the two complementary generator-exclusion fits. With identical absolute regularization and source score standardization, it reproduces the absolute-objective ridge result

predicted by Appendix A. This compares source-selected classical scoring families under a common protocol.

Selecting for low false-positive rates. A source objective can prioritize reducing missed fakes at a low false-positive rate. Selecting by the worst source-domain false-negative rate (FNR), at empirical source FPR at most 1%, changes full CoSG EER from the relative-objective ridge's 10.0462% to 9.8152%. TPR at 1% target FPR changes from 16.8636% to 28.2492%. Each destination ROC determines the threshold for these operating points. On images this objective trades CLIP Synthbuster EER 8.2552% (absolute-EER ridge) for 5.6671%, but changes full CNN from 11.8797% to 15.6038% and full Diffusion from 12.5375% to 18.3500%. Prioritizing low-FPR performance thus changes the transfer tradeoff. Earlier CoSG uncertainty intervals apply to the 925-row subset; the point estimates here use 1,797 rows.

H.3 Bounding contextual corrections and adding views

We also test a bounded graph correction selected on source data. It reduces reliance on local statistics when an endpoint has few comparisons, then clips the standardized correction. The ArcFace correction selected to protect low-FAR performance uses blend one, degree scale 16 and cap 0.5; AdaFace falls back to blend zero. For ArcFace, endpoint ITQ+graph versus this bounded correction gives EER 3.6333 versus 5.6667 on CALFW, 6.7667 versus 8.4000 on CPLFW, 2.6333 versus 2.3000 on AgeDB, and 0.9511 versus 0.9562 on IJB-C. CPLFW TAR at FAR 10^{-4} rises from 9.13% to 31.60%, still below raw 43.93%. Bounding the correction reduces extreme scores at the cost of discrimination on some benchmarks. It still uses global and neighboring scores and remains collective.

Additional audio windows. Extra views have separate costs and failure modes. On DF-Eval audio Test, three-window mean scoring improves EER for 44 of 45 correlated readouts, but worsens TPR at 1% FPR for 18 (23 improve, four tie). For the mixed ridge, TPR changes from 11.480% to 12.085%: 38 versus 40 of 331 fake files, with 457 authentic files defining the FPR. The 45 readouts share data and fitting procedures.

Table 11: Compute and information costs differ by recipe. Relative encoder-call counts use the same input and backbone; wall-clock latency also depends on implementation.

Recipe	Extraction	Fitting	Scoring
Center + metric	Same cached embeddings	Means, center optimization, scatter eigendecomposition	Dense map plus cosine
Collapsed detector	Same cached feature view	Source-fitted linear heads; one recorded 134 s CPU run	One dot product plus teacher
Five-crop / three-window	Five / three encoder calls	Same head family, matching frontend	Average fixed-view scores/features
ITQ + graph	Same embeddings	Rotations; endpoint fitting where declared	Binary scores and full graph moments

The angular diagnostic additionally evaluates a dense metric radius; the head-only dot score needs one vector dot product. Frozen-encoder adaptation still requires feature extraction, the stated source labels and, for collective scoring, a comparison graph.