INVARIANCE DOES NOT GUARANTEE REPAIR IN FROZEN-FEATURE DETECTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Removing a nuisance subspace from frozen features makes a detector exactly invariant to the removed directions. Whether that repairs the group error that motivated the removal is a different question, which we study in authenticity detectors on frozen speech and image encoders. Deleting a rank-4 emotion subspace estimated by the same procedure on acted corpora raises false rejection of surprised genuine speech from 17.3% to 33.9% on a w2v-BERT 2.0 head and lowers it on a WavLM head. What reverses the outcome on that w2v-BERT head is where the eraser is estimated. With four or five labels and matched rank, every tested eraser estimated on acted corpora raises the error, the same constructions estimated on genuine target-domain speech lower it at a cost in neutral rejection, and that speech helps more still when it trains the detector instead. The removed rank, the residual decodability of the nuisance and its mean response are three diagnostics an erasure audit can report, and none of them decides between the two outcomes. For a fixed linear head, the removed part of that response gives the direction in which the response moves but not reliably that of the error. Controlled image experiments show the limit of the mean response as well. With the audited geometry fixed, a second training association changes the group error several-fold while the mean nuisance response remains similar. Repair therefore has to be judged on target-domain decisions, for the genuine groups and generated sources at stake, not on the erasure itself.

1 INTRODUCTION

Authenticity detectors are often a small head on a frozen pretrained encoder. The encoder's representation varies with factors that leave authenticity unchanged, such as a speaker's emotion or a photograph's processing history, and a detector that responds to such a factor can reject one genuine group more than others. A natural repair is to estimate the directions the factor occupies and project them out. The projection achieves invariance by construction, since afterwards the head cannot respond to the removed directions, whatever it does elsewhere.

Whether the group is better off is a separate question, and the answer can be no. On an anti-spoofing head over frozen w2v-BERT 2.0 features, projecting out a rank-4 emotion subspace estimated on acted corpora raises false rejection of surprised genuine speech from 17.3% to 33.9% (Fig. 1). Refitting the head on the projected features leaves it at 28.4%. Every model is read at a threshold chosen by the same development rule. On a WavLM head trained on the same data and audited by the same procedure, the same deletion lowers the error, from 66.0% to 49.6%. Invariance was achieved in both cases. The increase on the reference split is among the largest. Across 168 constrained allocations of the English speakers, fixed projection raises surprise rejection on the test pair in 36, and on held-out Mandarin speakers in 159 (§3.6).

The two outcomes raise the question this paper answers, whether any of three diagnostics an erasure audit can report, the removed rank, the construction with its residual probe accuracy, or the mean nuisance response, decides the group error that follows. Under one fixed threshold rule, none of them does. The sign of the in-subspace response gives the direction in which the deletion moves the mean response but, across speaker allocations, not reliably the direction of the error. What reverses the outcome on the reference w2v-BERT head is the data the eraser is estimated on. With four or five labels and matched rank, every eraser estimated on acted corpora raises the error and every eraser

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.

estimated on genuine target-domain speech lowers it, and that speech lowers the error further still when it trains the detector rather than the eraser. Invariance and audit summaries therefore cannot stand in for a reading of target-domain decisions.

We make two contributions. The first is a set of controlled speech comparisons under one threshold rule that change one element of an erasure at a time where the design allows, the estimation data, the training data, the rank, the construction, the head or the evaluation population. They show which of these elements move the target error and that the audit summaries do not determine it (§3). The second is an image design that fixes the encoder, the audit subspace and the evaluation images and varies training associations alone; there a nuisance's mean response remains similar while the group error moves several-fold (§4). Both rest on a sample-level accounting of which decisions an intervention changes (§2), reported for the genuine groups and generated sources at stake.

2 HEADS, ERASERS AND AUDIT SUMMARIES

Let f be a frozen encoder and $z = \mathrm{std}(f(x)) \in \mathbb{R}^d$ its embedding, standardized with the statistics of the head's training data. The head is an ℓ_2-regularized logistic regression $s(z) = w^\top z + b$ trained to convergence; larger scores mean fake, and a sample is rejected when $s \geq \tau$. A *context* c is an entity observed under several nuisance classes $e \in E$, a speaker recorded in several emotions or a photograph under several corruptions. From a parallel corpus we form within-context centroid displacements $\delta_{c,e} = \mu_{c,e} - \frac{1}{|E_c|}\sum_{e' \in E_c} \mu_{c,e'}$, and U_K is their top-K principal subspace. K is an audit budget fixed before the reported analyses, four for speech and five for images, and App. B.1 sweeps it. U_K captures several times more of held-out displacements than covariance-matched random subspaces or the leading principal directions of the downstream data, and it transfers to corpora excluded from its estimation (App. B.1).

Interventions. Deleting U_K projects the features onto its orthogonal complement, $P_\perp = I - U_K U_K^\top$. We either re-score the fixed head on the projected features, $s_{P_\perp}(z) = w^\top P_\perp z + b$, or refit the head on $P_\perp z$. Both keep the training-data standardization and the same regularization constant, which fixes the recipe, not the effective penalty. Every model, baseline or intervention, is read at a threshold chosen on a named development population by a named rule, recomputed for that model. In the speech experiments the rule is the equal-error threshold on the two development speakers' genuine and synthetic utterances. A deletion succeeds, in the sense used throughout, when the head is exactly invariant to U_K; this is not the removal of all nuisance information, and an emotion probe still decodes emotion above chance after every eraser we compare.

Which decisions change. With s_0, τ_0 the baseline score and threshold and s_I, τ_I those of intervention I, let

$$g_0(z) = s_0(z) - \tau_0, \qquad h_I(z) = [s_I(z) - \tau_I] - [s_0(z) - \tau_0]. \tag{1}$$

g_0 is the baseline signed margin and h_I the shift the intervention gives a sample. The change in a genuine group's false rejection is exactly

$$\Delta\mathrm{FR}_G = \Pr_G(g_0 < 0,\ g_0 + h_I \geq 0) - \Pr_G(g_0 \geq 0,\ g_0 + h_I < 0), \tag{2}$$

entries minus exits, and a generated source's change in acceptance is exits minus entries. The identity accounts exactly for the decision changes from the observed scores and thresholds. For fixed projection of a linear head the score loses exactly $q(z) = w^\top U_K U_K^\top z$, so $h_I(z) = -q(z) - (\tau_I - \tau_0)$ splits into a score term and a recalibration term. A refitted head has new weights and its s_I is measured directly.

Three diagnostics. We examine three diagnostics an erasure audit can report, the removed rank, the residual accuracy of a probe for the nuisance, and the signed mean response $R = w^\top \bar{d}$, where $\bar{d}$ averages $d_{c,e} = \mu_{c,e} - \mu_{c,e_0}$ over contexts and measures how far the nuisance class e moves the score relative to a reference class e_0. The part of R inside the subspace, $R_{\mathrm{in}} = w^\top U_K U_K^\top \bar{d}$, is what the audit localizes, and fixed projection turns R into exactly $R - R_{\mathrm{in}}$. For a general affine eraser $z \mapsto Az + c$ the removed response is $R_{\mathrm{rem}} = w^\top (I - A)\bar{d}$, which equals R_{in} when $A = P_\perp$; the offset c cancels in the contrast. Three steps separate these summaries from $\Delta\mathrm{FR}$. R_{in} is a contrast between two classes whereas q enters Eq. 2 at its absolute level, recalibration moves τ, and the error counts only the samples near the boundary, so that the mean of h_I can have the opposite sign to $\Delta\mathrm{FR}$. The speech interventions below are therefore reported as decisions, the false rejection of the genuine strata and the acceptance of the generated sources.

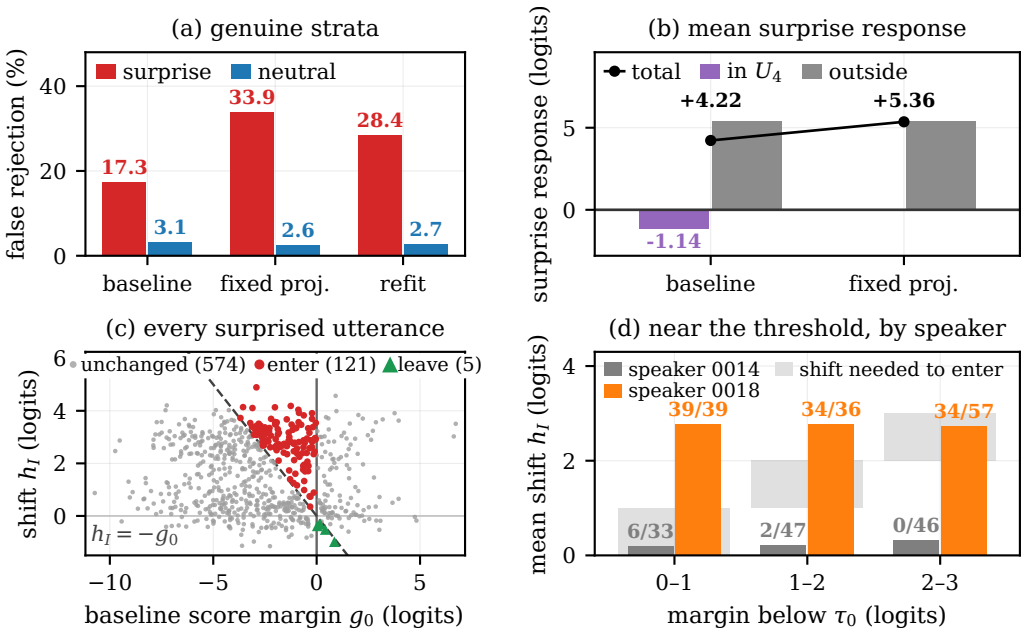

Figure 1: Deleting the emotion subspace raises surprise rejection on the w2v-BERT 2.0 head. The head is trained on ASVspoof 2019 and the TTS of six speakers and evaluated on 700 neutral and 700 surprised utterances of the two EmoFake test speakers; each model is read at its own development threshold. (a) False rejection of the two genuine strata. (b) Mean surprise response of the baseline head, the surprise minus neutral score contrast of §2, split inside and outside U_4, and of the fixed projected head. (c) Every surprised utterance by its baseline margin g_0 and its shift h_I under fixed projection; points above the line $h_I = -g_0$ with $g_0 < 0$ enter rejection. (d) Surprised utterances of each test speaker by distance below the threshold. Bars give the mean shift, labels the number entering rejection out of those in the bin, and the grey band the shift needed to enter.

3 One deletion, two outcomes

We now change one element of the erasure at a time, where the design allows, on a speech detector and read the false rejection of surprised genuine speech under one development threshold rule.

3.1 The reference case and the decisions behind it

Two anti-spoofing heads are trained on the same data, ASVspoof 2019 (Todisco et al., 2019) and the EmoSpoof-TTS utterances (Mahapatra et al., 2025) of six ESD speakers (Zhou et al., 2022), one on frozen WavLM-large (Chen et al., 2022) features and one on w2v-BERT 2.0 (Seamless Communication et al., 2023). Genuine training speech comes only from ASVspoof 2019, so authenticity is partly confounded with corpus, which §3.3 relieves. Each is audited with a rank-4 emotion subspace estimated by the same procedure from the same acted corpora (Livingstone & Russo, 2018; Cao et al., 2014), in that encoder's own coordinate. Both are evaluated on the two EmoFake (Zhao et al., 2024) test speakers of the reference split, which assigns six ESD speakers to training, two to development and two to test; the threshold is the equal-error point on the two development speakers, who enter no fit (App. A).

On w2v-BERT 2.0, false rejection of surprised genuine utterances rises from 17.3% to 33.9% under fixed projection and to 28.4% after refitting, while neutral rejection stays near 3% (Fig. 1a). On WavLM the same deletion lowers surprise rejection from 66.0% to 49.6%. The increase does not depend on the equal-error rule. At thresholds that accept at most 1, 5 or 10% of the development speakers' TTS utterances, fixed projection raises surprise rejection by 14.3 to 20.9 points and refitting by 8.7 to 12.3, and pooled test acceptance of TTS stays within 1.6 points of the baseline's (App. A). Re-estimating U_4 on resampled estimation speakers keeps the increase in 498 of 500 draws (App. B.1).

Table 1: Decisions changed by the deletion, for the two genuine strata of Fig. 1 and the three TTS sources, from a repeat run of that configuration evaluated on the two EmoFake test speakers at $\tau_0 = 10.885$. App. B.6 lists all five emotions and the EVC spoofs. Rates in %; in parentheses, the samples entering and leaving rejection. $\Delta\tau$ is the recalibration of the threshold in logits. The genuine columns report false rejection and the TTS columns acceptance.

		surprise, n=700	neutral, n=700	TTS accepted			
head	$\Delta\tau$	FR (in/out)	FR (in/out)	CosyVoice	F5-TTS	StyleTTS 2	pooled
baseline	–	17.1	3.0	30.8	0.4	2.4	11.2
fixed	-0.041	33.9 (122/5)	2.6 (2/5)	30.8 (49/49)	0.4 (1/2)	1.0 (33/0)	10.7 (83/51)
refit	-0.142	28.4 (83/4)	2.7 (1/3)	30.3 (23/11)	0.4 (1/2)	2.0 (12/1)	10.9 (36/14)

The removed directions offset the surprise response. On the harmed head the mean surprise response is $+4.22$ logits, of which -1.14 lies inside U_4 and $+5.36$ outside (Fig. 1b). The audited directions carry a component that opposes the contrast between surprise and neutral carried by the rest of the representation, so deleting them raises the response. On WavLM the in-subspace part is $+3.66$ of $+9.78$, and deleting it lowers the response. The sign of R_{in} thus gives the direction of the change in the mean response in both heads; the size of the error change depends on where the shifted samples sit relative to the threshold.

Which samples moved. Under fixed projection the score term moves surprised utterances up by 1.51 logits on average, against 0.37 for neutral ones, and recalibration adds 0.03. In decisions, 121 of the 700 surprised utterances enter rejection and 5 leave, against 1 and 5 for neutral ones, and 115 of the 121 entries start within 3 logits below the boundary (Fig. 1c). The increase is concentrated in speaker 0018. Surprise rejection moves from 20.6% to 21.4% for speaker 0014 and from 14.0% to 46.3% for speaker 0018. The two hold similar numbers of surprised utterances within one logit below the threshold, 33 and 39, but fixed projection shifts them by $+0.19$ and $+2.77$ logits on average, so 6 and 39 of them enter rejection (Fig. 1d).

Genuine rejection and source-specific acceptance. Table 1 reports one repeat run of the configuration for both genuine strata and each TTS source. The fake side gains a little. Pooled TTS acceptance moves from 11.2% to 10.7% and 10.9%, and StyleTTS 2 acceptance from 2.4% to 1.0% under fixed projection, against the 16.7-point rise in surprise rejection. CosyVoice acceptance is unchanged at 30.8% although 49 of its fakes enter rejection and 49 leave, so a net rate can hide many changed decisions. The mean shift can even disagree in sign with the error. In the repeat run of Table 1, the mean margin shift of neutral utterances is $+0.42$ logits, yet their false rejection falls, because five leave rejection and two enter (App. B.6).

3.2 Where the eraser is estimated

The subspace of Fig. 1 comes from eight emotions of acted English corpora recorded by other speakers. We estimated four constructions, paired PCA at rank 3 and 4, unpaired centroid PCA and LEACE as defined in §3.5, restricted to the four emotions both sources share and at matched rank, either on those corpora or on the genuine ESD speech of the six training speakers (Table 2). The two sources differ in speakers, utterance counts and recording conditions as well as in domain, so the intervention is the estimation dataset as a whole. Changing the estimation data reverses the outcome on this head. All four external constructions raise surprise rejection, to 22.0–37.7% under fixed projection, and all four in-domain constructions lower it, to 4.0–5.0%. The in-domain erasers also lower the 14 evaluation-only speakers' mean from 15.1% to 8.4–11.5%, at the cost of higher neutral rejection, 5.6–7.9% against 3.1%. The removed surprise response has opposite signs, -0.8 to -1.5 logits for the external erasers and $+2.8$ to $+3.6$ for the in-domain ones, matching the direction of every change in Table 2. With five labels the same split holds (Table 11). Across the 168 allocations of §3.6, all four in-domain constructions lower surprise rejection together in 154 allocations under fixed projection and in 159 after refitting, and all four external constructions raise it together in 60 and 75 (App. B.3).

The external erasers leave target-domain linear-probe accuracy nearly unchanged. A probe trained and tested on held-out Mandarin ESD speakers reaches 82.3% balanced accuracy without erasure, 80.7–82.4% after the external erasers and 76.5–77.5% after the in-domain ones. Exact invariance to the removed directions extends to every input; the guardedness LEACE achieves on its fitting

Table 2: Where the eraser is estimated, on the w2v-BERT 2.0 head of Fig. 1. The same four constructions, with the same four labels, neutral, angry, happy and sad, and the rank matched between the two sources, are estimated on the acted corpora of the paper, labelled external, or on the genuine ESD speech of the six training speakers, labelled in-domain; development and test speakers enter no fit. F and R are fixed projection and projected refit, each at its own development threshold. Rates in %. The 14 spk. columns give the mean false rejection of the 14 evaluation-only speakers of Fig. 2, and R_{rem} the surprise response the eraser removes from the baseline head, $w^\top(I - A)\bar{d}$ on the test pair, in logits. Five-label constructions and WavLM are in Tables 11 and 12.

			surprise FR		neutral FR		TTS accepted		14 spk. FR		
estimated on	eraser	rank	F	R	F	R	F	R	F	R	R_{rem}
none		–	17.3		3.1		11.2		15.1		–
external	paired PCA 3	3	29.4	24.4	3.1	2.7	10.4	10.8	20.3	18.6	−1.19
	paired PCA 4	4	37.7	30.0	2.7	2.7	11.0	11.7	19.2	19.0	−1.53
	unpaired PCA	3	30.0	23.1	3.1	2.4	9.7	11.1	19.1	17.5	−1.03
	LEACE	3	22.0	20.4	2.7	2.6	11.3	10.7	17.1	15.4	−0.76
in-domain	paired PCA 3	3	4.9	4.3	7.9	6.7	10.8	9.0	8.7	10.8	+3.28
	paired PCA 4	4	5.0	4.3	7.7	6.9	10.6	9.0	8.9	10.8	+3.31
	unpaired PCA	3	4.0	3.4	7.3	7.0	11.2	8.8	8.4	10.1	+3.56
	LEACE	3	4.6	3.9	5.6	6.9	11.7	8.9	11.5	10.8	+2.76

Table 3: The genuine speech of the six training speakers, G4, used to estimate the eraser or to train the detector, on w2v-BERT 2.0 with the weighted recipe of App. B.5, whose baseline is not that of Fig. 1. Eraser rows: fixed projection, ranges over the four constructions of each source with four labels, each at its own development-TTS threshold. Rates in %; refits and WavLM in Table 16.

genuine training speech	eraser	surprise FR	neutral FR	TTS accepted	14 speakers' FR
ASVspoof 2019 only	none	20.4	3.3	13.2	18.9
	external	25.3–32.7	3.3–4.1	11.7–13.1	19.5–21.5
	in-domain (G4)	3.1–4.4	7.7–10.1	11.8–13.0	9.9–14.1
plus G4	none	0.0	0.0	0.1	5.9
	external	0.0–0.7	0.0–0.4	0.1–0.4	6.1–10.1
	in-domain (G4)	0.0	0.0	0.1	5.9–6.3

distribution is not guaranteed to extend to a new distribution, and on the target speakers the external erasers remove directions that offset the surprise response rather than carry it. On WavLM every four-label construction lowers surprise rejection and the in-domain ones lower it more, to 31.6–46.7% against 51.6–63.9%, again at a cost in neutral rejection (App. B.3).

3.3 The same data as training data

Adding the six training speakers' own non-surprised genuine recordings, G4, to the genuine class relieves the corpus confounding of the main recipe and lowers the 14 evaluation-only speakers' mean rejection from 18.9% to 5.9% on w2v-BERT 2.0 under development-selected regularization, and the gain holds at every tested penalty (App. B.5). Deleting the original eight-label subspace from this improved head raises the 14 speakers' mean again, by 4.1 to 5.7 points across penalties, whereas refitting preserves the gain, changing it by -0.1 to $+0.5$.

Table 3 uses G4 in both roles under one weighted recipe. As eraser data, in-domain estimation brings surprise rejection from 20.4% to 3.1–4.4% and the 14 speakers' mean to 9.9–14.1%. As training data, G4 brings them to 0.0% and 5.9% with no eraser at all, and once G4 trains the detector every four-label eraser leaves surprise rejection at 0.7% or below. The genuine speech that makes an eraser work makes a better detector when it is used directly. The cost of adding it is on the fake side, where on WavLM the added recordings raise EmoFake EVC acceptance from 0.00% to 0.29%.

3.4 The removed rank

We repeated the reference configuration at every paired-PCA rank from 1 to 8 (Table 7). Rank 1 lowers surprise rejection slightly, to 15.3% and 15.6%. Every rank from 2 to 8 raises it under both interventions and raises the mean rejection of the 14 evaluation-only speakers of Fig. 2 as well. R_{in}

Table 4: Three erasers fitted to the same emotion labels on w2v-BERT 2.0, grouped by removed rank, 7 for LEACE. Means over five partitions of the estimation speakers: in each, one group of RAVDESS and CREMA-D speakers fits the eraser and the same fitted transform is applied to an eight-class emotion probe, trained and tested on two further disjoint groups, and to the detector. Probe BA: its balanced accuracy on held-out estimation-corpus speakers; false rejection and TTS acceptance are measured on the two EmoFake test speakers, each head at its own development-TTS threshold. Bold: the rank-4 contrast. The retrained baseline's 17.1% is a different run from the 17.3% of Fig. 1 (Table 6). All values in %; differences in the text are computed before rounding.

eraser	head	probe BA	FR surprise	FR neutral	TTS accepted
none	baseline	66.0	17.1	3.0	11.2
rank 4					
paired PCA	fixed	48.7	**30.9**	2.4	10.9
	refit	48.7	26.9	2.5	11.0
unpaired centroid PCA	fixed	41.5	**13.1**	3.3	11.2
	refit	41.5	14.1	3.1	10.2
rank 7					
paired PCA	fixed	41.4	26.3	4.0	10.8
	refit	41.4	24.0	2.9	11.3
unpaired centroid PCA	fixed	26.1	23.1	3.6	9.2
	refit	26.1	19.3	2.9	10.3
LEACE	fixed	27.2	23.8	4.1	13.1
	refit	27.2	18.7	2.8	10.0

is positive at rank 1 and negative from rank 2 on. On WavLM R_{in} is positive and surprise rejection falls at every rank.

3.5 The construction and its residual decodability

Three constructions are fitted to the same eight emotion labels. Paired PCA uses the within-speaker displacements of §2; unpaired centroid PCA uses the principal directions of the emotion centroids pooled over speakers; LEACE (Belrose et al., 2023) uses its fitted affine map, whose measured rank is 7. Each is estimated on five partitions of the estimation speakers fixed in advance, so these rows do not reuse the subspace of Fig. 1. Within a partition, disjoint speakers fit the eraser, train an emotion probe and test it, and one fitted transform is applied to the probe and to the detector (Table 4).

With eight labels and at rank 4, paired and unpaired PCA move surprise rejection in opposite directions, to 30.9% and 13.1% under fixed projection, while TTS acceptance differs by 0.25 pp. At rank 7, the rank LEACE removes, both raise it. Residual probe accuracy does not order the errors either. The two erasers that leave the lowest held-out probe accuracy, rank-7 unpaired PCA and LEACE, both keep surprise rejection above the baseline, and rank-4 unpaired PCA, which leaves a higher accuracy, lowers it. The rank-4 reversal recurs on ten Mandarin ESD speakers that enter no detector, threshold or eraser fit, where paired and unpaired fixed projection change surprise rejection by $+5.1$ and -2.4 pp, a difference of $+7.6$ pp with interval $[+2.5, +13.5]$ (App. B.2). On WavLM both rank-4 projections lower surprise rejection.

3.6 The evaluation population and the head

Speakers outside every detector, threshold and eraser fit. Applied unchanged to the ten Mandarin speakers of ESD and the four of Expresso (Nguyen et al., 2023), the projected refit raises rejection for all 14 speakers on w2v-BERT 2.0, by $+3.3$ pp on average with nine speakers above $+1$ pp, and lowers it for all 14 on WavLM, by -12.4 pp. Fixed projection gives $+3.1$ and -20.6 pp (Fig. 2). A speaker bootstrap with heads and thresholds fixed puts the refit means at $[+1.7, +5.1]$ and $[-15.2, -9.8]$ pp (App. B.4).

How typical is the reference split? The reference split is one of 168 allocations of the ten English ESD speakers to training, development and test roles (Fig. 3; App. B.10). On the allocations' own English test pairs, w2v-BERT 2.0 surprise rejection rises in 36 of 168 under fixed projection and in 72 after refitting, mostly where speaker 0018 is tested. All 21 allocations that test 0018 rise by at least 2 pp under both interventions, against 3 and 26 of the other 147. Only 5 fixed-projection and 11

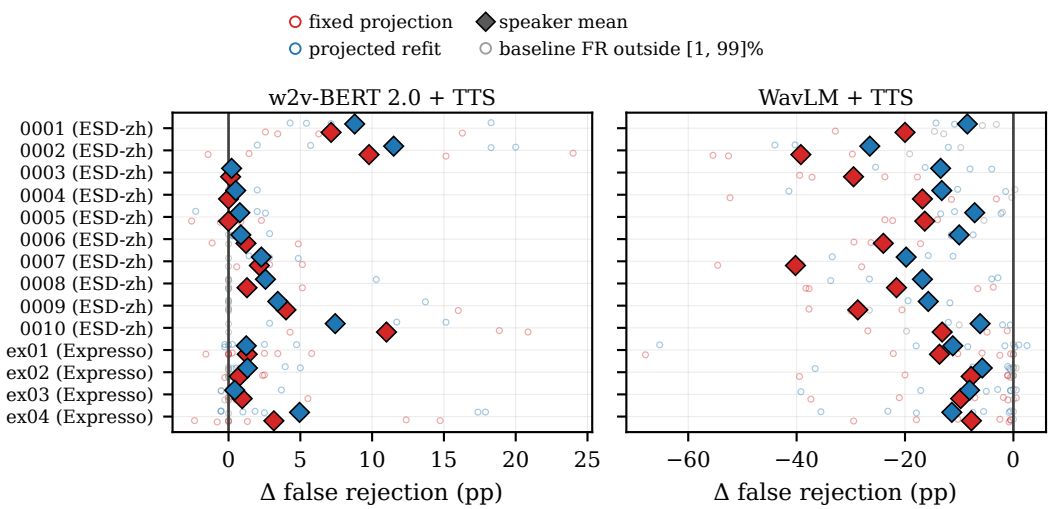

Figure 2: The two heads of Fig. 1, unchanged, on 14 speakers that enter no detector, threshold or eraser fit, ten Mandarin ESD speakers and four Expresso speakers whose read styles are not the acted emotions U_4 was estimated from. Red marks fixed projection and blue the projected refit. Small open marks are the 72 eligible conditions of speaker and emotion or speaker and style per encoder and intervention; diamonds are speaker means over their conditions. Conditions whose baseline false rejection lies outside $[1, 99]\%$ are drawn grey and kept in the means. Panels differ in horizontal scale.

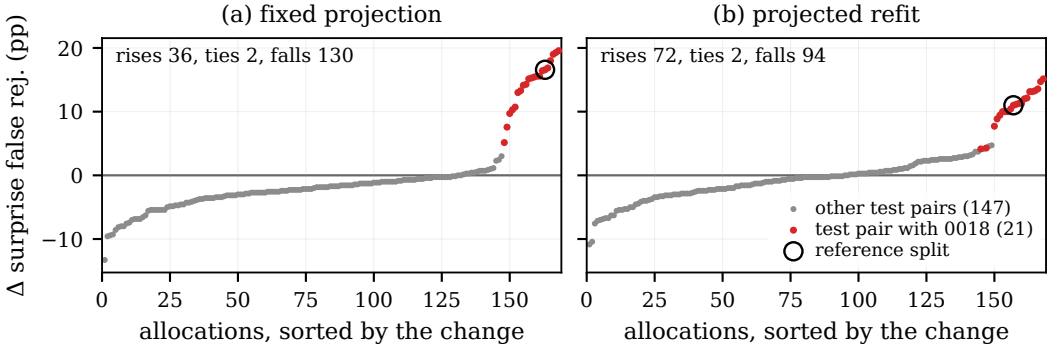

Figure 3: Where the reference split sits among the 168 speaker allocations of the w2v-BERT 2.0 grid, as the change in surprise false rejection on each allocation's own test pair under (a) fixed projection and (b) the projected refit, sorted, in whole utterances of 700. Red marks the 21 allocations whose test pair contains speaker 0018; the ring is the grid's own row for the reference split. Each allocation refits the detector, standardizer and subspace and changes the development and test speakers; allocations share speakers, so this is a descriptive distribution.

refit allocations exceed the reference split's increases. Scored on five held-out Mandarin speakers, fixed projection raises surprise rejection in 159 of the 168 allocations, by a median of 3.3 points, and the configurations raise the 14 evaluation-only speakers' mean in 33 under fixed projection and in 125 after refitting; WavLM rejection falls in 167 of 168 on the test pairs and in all 168 on the Mandarin speakers. How often a deletion helps thus depends on which population is read. The sign of R_{in} is a warning where it appears, not a certificate where it does not. Estimated on the Mandarin audit speakers and read on the held-out Mandarin speakers, a negative sign predicts an increase that occurs in all 120 allocations where it appears, but 39 of the 159 increases occur without it, and on the English test pairs no rule built on it exceeds the accuracy of always predicting the more frequent direction (App. B.10).

Other heads, adapted encoders and a newer source. With all but the head fixed, surprise rejection rises under both interventions for five ℓ_2 strengths and three one-hidden-layer perceptrons, though not for a two-hidden-layer head, and on WavLM its direction depends on the head family (App. B.7).

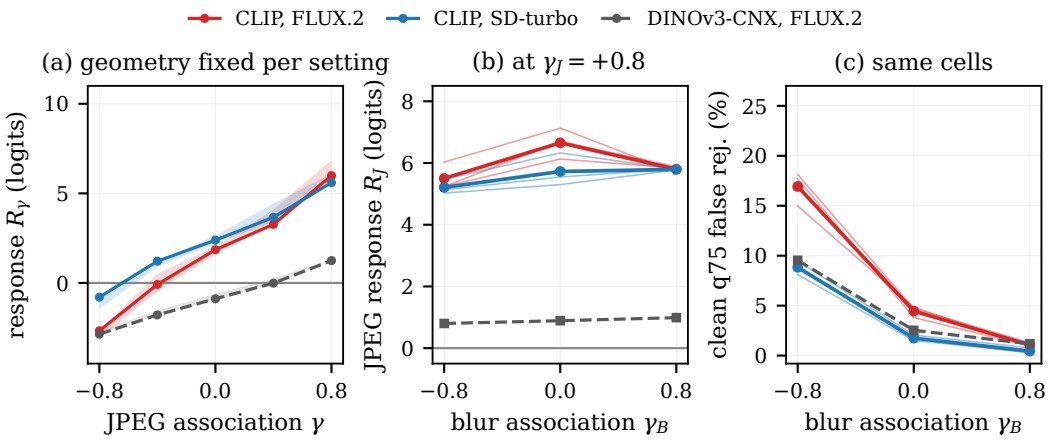

Figure 4: Training associations under a fixed audited geometry. (a) Signed mean response $R_\gamma = w_\gamma^\top \bar{d}$ on held-out audit images for CLIP heads trained against FLUX.2 and SD-turbo fakes, with DINOv3-ConvNeXt against FLUX.2 dashed; the line is assignment seed 0 and the band spans five seeds. (b) With the JPEG and blur associations assigned independently, the JPEG response at $\gamma_J{=}{+}0.8$ against the blur association, and (c) false rejection of the clean q75 genuine images in the same cells. Thin lines are three assignment seeds and thick lines their mean. All 540 trajectories of the full grid are in Fig. 5.

After LoRA adaptation (Hu et al., 2022) of the encoder with U_K re-estimated, the w2v-BERT 2.0 increase recurs in all three seeds under both interventions and WavLM keeps its decreases (App. B.8). With Qwen3-TTS (Qwen Team, 2026) clones as training source, w2v-BERT 2.0 surprise rejection rises from 24.9% to 54.4% under fixed projection and WavLM rejection falls (App. D).

4 The mean response under a fixed geometry in images

We vary the association between nuisance and label in the training set with the encoder, feature coordinate, audit subspace and evaluation images fixed; each association setting gets a newly trained head and its own threshold from the same development-EER rule.

The association moves the response. Every real and fake training image exists as a JPEG q75 and a q95 version. Exact per-class quotas give the q75 version to a fraction $(1+\gamma)/2$ of the fakes and $(1-\gamma)/2$ of the reals, $\gamma \in \{-0.8, \ldots, 0.8\}$, so positive γ ties low quality to fakes and negative γ ties it to reals while the overall q75 share stays at one half. For the CLIP (Radford et al., 2021) head trained against FLUX.2 (Black Forest Labs, 2026) fakes, R_γ rises from -2.7 to $+6.0$ logits across the range, and in all five seeds 97–98% of the change lies inside U (Fig. 4a). Across nine encoders, six training generators, five assignment seeds and two nuisances, JPEG quality and Gaussian blur, all 540 response trajectories rise strictly with γ, and 73–104% of each change lies inside the audit subspace. The sign reverses in 260 of 270 JPEG and 245 of 270 blur trajectories (App. C). Error follows the response only where the margin allows it. From $\gamma{=}{-}0.8$ to $+0.8$, CLIP/SD 1.5 goes from 0.7% to 5.0% false rejection of q75 reals while CLIP/SD-turbo (Sauer et al., 2024) stays below 0.8% with a response of similar size.

Two nuisances associated independently. We then give every training image four versions, JPEG q95 or q75, each clean or blurred, and set the two associations γ_J and γ_B independently by exact quotas, for SD-turbo and FLUX.2 fakes and three assignment seeds. The two audit subspaces are far from aligned, with a smallest principal angle of 58°, and the responses separate, since a regression over the grid gives R_J a slope of 5.83 logits per unit γ_J and 0.09 per unit γ_B on FLUX.2. The errors do not separate. At $\gamma_J{=}{+}0.8$, false rejection of the clean q75 genuine images is 16.9% when blur is associated with the real class, 4.5% when it is balanced and 1.1% when it is associated with the fake class (Fig. 4c). The seed-mean JPEG responses in the same cells are 5.50, 6.65 and 5.80 logits (Fig. 4b). The difference between the two extreme blur associations, $\gamma_B{=}{-}0.8$ minus $+0.8$, is -0.30 logits in response, with a paired image bootstrap interval of $[-0.40, -0.21]$, and $+15.9$ points in error, $[14.8, 17.0]$ (App. C.2). The same ordering holds for the seed means of all nine encoders

with either generator (Table 26). The JPEG response misses what the blur association changes, the reference level of the clean images. The mean signed margin $\mathbb{E}[s(z) - \tau]$ of the clean q95 reference images falls from -12.3 to -19.8 logits, from -1.0 to -1.7 development score deviations, while the JPEG contrast between the q75 and q95 versions of the same images stays between 5.6 and 6.8 logits (App. C.2).

The reversal of the response with the training association is not specific to the JPEG design or to linear heads. A sentence-matched speech construction reverses the response in all 18 conditions of encoder, emotion and TTS, and a converged perceptron head reproduces the monotone reversal in all 12 image trajectories of four encoder and generator pairs at three seeds (App. C).

5 Related work

INLP (Ravfogel et al., 2020) removes linearly decodable information by iterated nullspace projections, R-LACE (Ravfogel et al., 2022) learns a rank-constrained erasure through a convex relaxation of a linear minimax objective, and LEACE (Belrose et al., 2023) gives the least-squares affine edit that achieves linear guardedness, so that under any convex loss no affine predictor of the concept does better than a constant. SPLINCE (Holstege et al., 2025) removes a concept by an oblique projection chosen to preserve the covariance of the representation with a target label. These guarantees are stated for the representation and labels the eraser is fitted to. Downstream, on Bias in Bios, LEACE narrows a gender gap in true-positive rates at a small cost in accuracy, and refitting the classifier restores part of the gap (Belrose et al., 2023). Nuisance attribute projection in speaker recognition (Solomonoff et al., 2005) and its emotion variant (Bao et al., 2007) are the constructions closest to our paired PCA. ACES (Parekh, 2026) reports that partially removing an accent subspace worsens both the word error rate and the accent disparity of speech recognition on clean audio, and Khani & Liang (2021) show that removing spurious features can lower accuracy and affect groups disproportionately. We ask which audit quantities determine the group error of authenticity detectors.

EmoFake (Zhao et al., 2024) provides emotion-converted fakes, and Mahapatra et al. (2025) show that anti-spoofing models struggle with emotional TTS, releasing the EmoSpoof-TTS corpus our speech heads are trained on. Group robustness is sought in training, by group DRO (Sagawa et al., 2020), JTT (Liu et al., 2021) or DFR (Kirichenko et al., 2023), which retrains the last layer on a group-balanced set, the setting closest to our refitted heads. Others are data balancing (Idrissi et al., 2022), blur and compression augmentation (Wang et al., 2020) and content-matched synthetic data (Guillaro et al., 2025). Dataset shift degrades predictive calibration (Ovadia et al., 2019). Cross-generator and cross-corpus detection gaps are documented (Ojha et al., 2023; Corvi et al., 2023; Müller et al., 2022); subgroup thresholds narrow error gaps between text groups (Jung et al., 2025), and source-aware representations target cross-domain image detection (Zhang et al., 2026).

6 Conclusion

Invariance to a nuisance subspace is a property of the representation a head sees; the error a group suffers is a property of scores, a calibrated threshold and the samples near it; and the two are joined by the detector, the eraser and the evaluation population. On the reference w2v-BERT 2.0 head, with four or five labels and matched rank, every eraser fitted to acted emotional speech from other speakers raised surprise rejection, while the same constructions fitted to genuine target-domain speech lowered it; on WavLM, both estimation sources lowered it in the four-label comparison, and how often a deletion helped depended on the speakers evaluated. The removed rank, residual decodability and the total mean response did not determine the direction of these changes consistently, and in images a second association moved the group error while the mean response remained similar. The one input that reliably helped, genuine target-domain speech, helped more as training data than as eraser data in the recipes tested.

Repair therefore cannot be certified from erasure metrics on the fitting data; it has to be read as target-domain decisions for the genuine groups and generated sources at stake. Our evidence covers linear and small nonlinear heads on frozen or LoRA-adapted encoders and the tested nuisance definitions and pipelines (App. B.9); every erasure intervention holds the encoder fixed, including the checks after LoRA adaptation.

REPRODUCIBILITY STATEMENT

All PCA uses full SVD and all seeds are fixed. Every primary result uses ℓ_2-logistic heads trained to convergence with lbfgs at a tolerance of 10^{-8}. The perceptron heads of App. B.7 are three one-hidden-layer widths and one two-hidden-layer head with three seeds each; seeds are training repetitions and never count as independent speakers or deployments. The head-family, eraser and training-data comparisons retrain heads with the stored recipe; the retrained linear head's threshold differs in the third decimal and moves one utterance of 700, so Fig. 1 and its crossing counts are read from the stored scores, and Table 6 maps each speech result to its run. Logistic fits are not bit-reproducible across BLAS thread counts, so every analysis added after the original runs fixes the thread count; the replay behind Table 1 reproduces the stored summaries byte for byte. In the primary speech experiments, external eraser-estimation speakers are disjoint from detector-training, development and test speakers; the in-domain comparisons estimate erasers from the six training speakers and keep development and test speakers excluded. The analysis plan with its dated amendments, all result files with their hashes, the figure and table generators, the per-sample decision export and a release gate with explicit numerical and regeneration checks will be released with the code; App. A summarizes the analyses that bear on the paper only indirectly. Feature manifests reference the public corpora.

ETHICS STATEMENT

This work studies group-level errors of authenticity detectors that fall on emotional speakers, on photographs with a particular processing history and, in analyses summarized in App. A, on non-native writers, and it evaluates interventions together with their costs. All corpora are public research datasets used under their licences; no new human data were collected, and ICNALE learner essays are used only as features and are not redistributed. We synthesized speech in two places, zero-shot clones of the 24 RAVDESS actors with two public text-to-speech systems for an external operating-point test, and Qwen3-TTS clones of the ten English ESD speakers of EmoFake for the synthesis-source repetition (App. D). The synthesized audio was used only to extract the features and scores we analyse, is never presented as genuine recordings, and is not redistributed; the manifests that identify which public utterance prompted each clone will be released with the code. Reducing a disparity can trade against utility and can shift errors to other groups; we report those costs and recommend re-auditing operating points for each deployment domain before any use.

AI USE STATEMENT

We used large language model assistants in four ways, following the categories of the ICLR policy. *To aid or polish writing.* An assistant, Claude from Anthropic, drafted and edited prose; the authors reviewed and revised all text and are responsible for it. *For retrieval and discovery.* Assistants helped to find and check related work; the authors verified every source cited. *For research ideation or execution.* Assistants proposed follow-up analyses, drafted and ran analysis scripts under the authors' specifications, and reviewed the manuscript against the stored results; the authors set every design choice, verified the outputs and are responsible for the analyses reported. *For generating synthetic datasets.* The fake images from seven public image generators, the AI-written text from open-weight language models alongside public releases, and the voice clones described above are objects of study, not writing aids, and each is named where it is used.

REFERENCES

Huanjun Bao, Ming-Xing Xu, and Thomas Fang Zheng. Emotion attribute projection for speaker recognition on emotional speech. In *Proc. Interspeech*, pp. 758–761, 2007.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Proc. NeurIPS*, 2023.

Black Forest Labs. FLUX.2 [klein]: open-weight rectified-flow image generation models (base 4b checkpoint). `https://huggingface.co/black-forest-labs/FLUX.2-klein-base-4B`, 2026. Model card; Apache 2.0.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv:2504.13181*, 2025.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Trans. Affective Computing*, 5(4), 2014.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Topics Signal Process.*, 16(6), 2022.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv:2410.06885*, 2024.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP*, 2023.

Wei Deng, Siyi Zhou, Jingchen Shu, Jinchao Wang, and Lu Wang. IndexTTS: An industrial-level controllable and efficient zero-shot text-to-speech system. *arXiv:2502.05512*, 2025.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, Zhifu Gao, and Zhijie Yan. CosyVoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv:2407.05407*, 2024a.

Zhihao Du et al. CosyVoice 2: Scalable streaming speech synthesis with large language models. *arXiv:2412.10117*, 2024b.

Enrico Fini et al. Multimodal autoregressive pre-training of large vision encoders. In *Proc. CVPR*, 2025.

Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general AI-generated image detection. In *CVPR*, 2025.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is ChatGPT to human experts? comparison corpus, evaluation, and detection. *arXiv:2301.07597*, 2023.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proc. CVPR*, 2022.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Proc. ICLR*, 2019.

Floris Holstege, Shauli Ravfogel, and Bram Wouters. Preserving task-relevant information under linear concept removal. In *Proc. NeurIPS*, 2025.

Jeremy Howard. Imagenette: a subset of 10 easily classified classes from ImageNet. `https://github.com/fastai/imagenette`, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proc. ICLR*, 2022.

Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning (CLeaR)*, 2022.

Shin'ichiro Ishikawa. *The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English*. Routledge, 2023.

Minseok Jung, Cynthia Fuertes Panizo, Liam Dugan, Yi R. Fung, Pin-Yu Chen, and Paul Pu Liang. Group-adaptive threshold optimization for robust AI-generated text detection. *arXiv preprint arXiv:2502.04528*, 2025.

Fereshte Khani and Percy Liang. Removing spurious features can hurt accuracy and affect groups disproportionately. In *Proc. ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 196–205, 2021.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *Proc. ICLR*, 2023.

Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. StyleTTS 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. In *Proc. NeurIPS*, 2023.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. GPT detectors are biased against non-native english writers. *Patterns*, 2023.

Evan Zheran Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *Proc. ICML*, 2021.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. In *arXiv:1907.11692*, 2019.

Steven R. Livingstone and Frank A. Russo. The Ryerson audio-visual database of emotional speech and song (RAVDESS). *PLOS ONE*, 13(5), 2018.

Reza Lotfian and Carlos Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Trans. Affective Computing*, 10(4), 2019.

Aurosweta Mahapatra, Ismail Rasim Ulgen, Abinay Reddy Naini, Carlos Busso, and Berrak Sisman. Can emotion fool anti-spoofing? In *Proc. Interspeech*, 2025.

Nicolas M. Müller, Pavel Czempin, Franziska Diekmann, Adam Froghyar, and Konstantin Böttinger. Does audio deepfake detection generalize? In *Interspeech*, 2022.

Tu Anh Nguyen, Wei-Ning Hsu, Antony D'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, Felix Kreuk, Yossi Adi, and Emmanuel Dupoux. EXPRESSO: A benchmark and analysis of discrete expressive speech resynthesis. In *Proc. Interspeech*, pp. 4823–4827, 2023.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *CVPR*, 2023.

Omnilingual ASR team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, et al. Omnilingual ASR: Open-source multilingual speech recognition for 1600+ languages. *arXiv:2511.09690*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *NeurIPS*, 2019.

Swapnil Parekh. ACES: Accent subspaces for coupling, explanations, and stress-testing in automatic speech recognition. *arXiv preprint arXiv:2603.03359*, 2026.

Qwen Team. Qwen3-TTS technical report. *arXiv:2601.15621*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *Proc. ICML*, 2021.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proc. ACL*, 2020.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. Linear adversarial concept erasure. In *Proc. ICML*, 2022.

Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky: An improved text-to-image synthesis with image prior and latent diffusion. In *Proc. EMNLP: System Demonstrations*, pp. 286–295, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, 2022.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts. In *Proc. ICLR*, 2020.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *Proc. ECCV*, 2024.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, et al. Seamless: Multilingual expressive and streaming speech translation. *arXiv:2312.05187*, 2023. Introduces the w2v-BERT 2.0 speech encoder.

Oriane Siméoni et al. DINOv3. *arXiv:2508.10104*, 2025.

A. Solomonoff, W. M. Campbell, and I. Boardman. Advances in channel compensation for SVM speaker recognition. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pp. 629–632, 2005.

Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee. ASVspoof 2019: Future horizons in spoofed and fake audio detection. In *Proc. Interspeech*, 2019.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv:2502.14786*, 2025.

Jean-Marc Valin, Koen Vos, and Timothy B. Terriberry. Definition of the Opus audio codec. IETF RFC 6716, 2012. `https://www.rfc-editor.org/info/rfc6716`.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-generated images are surprisingly easy to spot... for now. In *CVPR*, 2020.

Chenfei Wu, Jiahao Li, Jingren Zhou, et al. Qwen-Image technical report. *arXiv:2508.02324*, 2025.

Z-Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, et al. Z-Image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv:2511.22699*, 2025.

Shan Zhang, Yongxin He, Mingming Zhang, Huiwen Tian, and Lei Ma. FiSeR: Fine-grained source representations for cross-domain AI image detection. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2606.00606.

Yan Zhao, Jiangyan Yi, Jianhua Tao, Chenglong Wang, and Yongfeng Dong. EmoFake: An initial dataset for emotion fake audio detection. In *Proc. CCL*, 2024.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Emotional voice conversion: Theory, databases and ESD. *Speech Communication*, 137, 2022.

A EXPERIMENTAL PROTOCOLS

Data and models. The speech data are ASVspoof 2019 LA (Todisco et al., 2019); EmoSpoof-TTS (Mahapatra et al., 2025), whose fakes come from CosyVoice (Du et al., 2024a), F5-TTS (Chen et al., 2024) and StyleTTS 2 (Li et al., 2023); EmoFake (Zhao et al., 2024); ESD (Zhou et al., 2022); RAVDESS (Livingstone & Russo, 2018); CREMA-D (Cao et al., 2014); Expresso (Nguyen et al., 2023); MSP-Podcast (Lotfian & Busso, 2019) for the population audit summarized below; the encoders WavLM-large (Chen et al., 2022), w2v-BERT 2.0 (Seamless Communication et al., 2023) and, in App. D, omniASR (Omnilingual ASR team et al., 2025); clones from Qwen3-TTS (Qwen Team, 2026), CosyVoice 2 (Du et al., 2024b) and IndexTTS (Deng et al., 2025); and the Opus codec (Valin et al., 2012). The image data are imagenette (Howard, 2019) with the corruptions of Hendrycks & Dietterich (2019); the encoders CLIP (Radford et al., 2021), MAE (He et al., 2022), DINOv2 (Oquab et al., 2024), SigLIP 2 (Tschannen et al., 2025), AIMv2 (Fini et al., 2025), DINOv3 (Siméoni et al., 2025) and Perception Encoder (Bolya et al., 2025); the generators SD-turbo and SDXL-turbo (Sauer et al., 2024), Stable Diffusion 1.5 (Rombach et al., 2022), FLUX.2 [klein] (Black Forest Labs, 2026), Qwen-Image (Wu et al., 2025) and Z-Image (Z-Image Team et al., 2025), and, for the blind check summarized below, Kandinsky 2.2 (Razzhigaev et al., 2023). Table 24 lists every checkpoint; the revision of each will be released with the code.

Features, coordinate and rank. Speech enters every encoder as 16 kHz mono audio, RMS-normalized and truncated at 20 s; the representation is the mean over frames of hidden states 9 and 23, counting the convolutional front end as state 0 and the transformer layers as 1 to 24, concatenated. Images are resized to 224×224 pixels before each encoder's own processor. Every embedding is standardized with the training statistics of the head's own task, so the subspace is a property of the encoder under this coordinate. Contexts are the 24 RAVDESS and 91 CREMA-D speakers and 1,200 imagenette photographs under five corruptions. The displacement matrix is mean-zero within every context by construction and its global mean is $0.000\sigma_1$, so centred and uncentred PCA coincide. Ranks are budgets fixed before the reported analyses, $K{=}4$ for speech and $K{=}5$ for images, one per corruption condition; the speech value was selected on development data under an earlier recipe and then frozen, and App. B.1 sweeps it.

Heads, thresholds and calibration. Heads are ℓ_2-logistic regressions in sklearn with $C{=}1$, trained to convergence with lbfgs at tolerance 10^{-8} and cross-checked with newton-cg to 4×10^{-4} in coefficient distance. Fake is the positive class and a sample is rejected when $s \geq \tau$. In speech, the two EmoFake development speakers, 0019 and 0020, enter no head fit and no subspace estimate and are the genuine side of every threshold; τ_{TTS} is each head's equal-error threshold on their genuine utterances against the TTS utterances of the same speakers, and everything is evaluated on the test speakers 0014 and 0018. The equal-error threshold is the first candidate of `roc_curve` minimizing $|\mathrm{FNR} - \mathrm{FPR}|$, with no interpolation. In images the threshold is the equal-error point on the detector's development split. Re-calibration after an intervention uses the same rule on the same development population. Table 5 replaces the equal-error rule by development acceptance targets. For each model, τ_β is the largest development fake score at which at most β of the development speakers' TTS utterances are accepted, for β of 1, 5 and 10%. On w2v-BERT 2.0 surprise rejection is 64.3, 24.6 and 7.1% for the baseline at the three targets, 82.1, 45.4 and 21.4% under fixed projection and 76.6, 34.9 and 15.9% after refitting, while pooled test acceptance of TTS stays within 1.6 points of the baseline's. On WavLM both interventions lower surprise rejection at every target. Realized test acceptance is not matched across models, so these are not comparisons at an identical deployment cost.

Runs behind the speech results. Several results read the w2v-BERT 2.0 configuration of Fig. 1 from different runs. Table 6 gives, for each, the run, how the subspace was estimated and how the rates are aggregated. Three are trainings of the same recipe and agree to one utterance of 700; the eraser comparison of Table 4 is a different estimand, a mean over five re-estimated partitions.

Uncertainty. The EmoFake test strata are conditional on their two genuine test speakers, and no interval is attached to them. The 14-speaker means use a speaker bootstrap with heads and thresholds held fixed; image strata use an image bootstrap. All intervals are 95% percentile intervals.

What was specified in advance. The analysis plan and its dated amendments will be released with the code; each was written with its inclusion rules before the corresponding results were computed. Prospectively specified and met were the exact-quota association manipulation and its seed repeti-

Table 5: Common development acceptance targets for the configuration of Fig. 1, $K{=}4$. For each arm, τ_β is the largest development fake score whose development acceptance, the share of the calibration speakers' TTS utterances scored below it, is at most β; the development-EER rule of the paper is shown for reference. Test acceptance is realized, not matched, and differs between models; this is therefore not a comparison at an identical deployment cost. Dev. FA and dev. FR are the realized development acceptance and genuine rejection; the other columns are test-pair rates in %.

			development		genuine FR			TTS accepted				EVC
encoder	rule	arm	FA	FR	surprise	neutral	other	CosyVoice	F5-TTS	StyleTTS 2	pooled	FA
w2v-BERT 2.0	dev. EER	baseline	6.39	6.4	17.3	3.1	3.5	30.7	0.4	2.4	11.2	0.00
		fixed	6.89	6.9	33.9	2.6	4.3	30.8	0.4	1.0	10.8	0.00
		refit	6.36	6.4	28.4	2.7	3.7	30.3	0.4	2.0	10.9	0.00
	dev. FA 1%	baseline	1.00	31.4	64.3	18.7	23.8	10.7	0.0	0.0	3.6	0.00
		fixed	1.00	32.9	82.1	20.6	31.0	9.6	0.0	0.0	3.2	0.00
		refit	1.00	32.2	76.6	17.9	26.0	9.6	0.0	0.0	3.2	0.00
	dev. FA 5%	baseline	5.00	9.5	24.6	4.7	5.6	25.8	0.2	1.6	9.2	0.00
		fixed	5.00	10.9	45.4	4.9	7.0	25.0	0.3	0.6	8.6	0.00
		refit	5.00	9.6	34.9	3.6	5.1	26.3	0.3	1.5	9.3	0.00
	dev. FA 10%	baseline	10.00	2.7	7.1	1.1	1.1	40.4	0.7	5.0	15.3	0.00
		fixed	10.00	3.5	21.4	1.6	2.0	38.6	0.7	2.1	13.8	0.00
		refit	10.00	3.1	15.9	1.0	1.4	39.6	0.5	4.5	14.9	0.00
WavLM	dev. EER	baseline	12.78	12.8	66.0	1.7	9.0	36.7	1.7	6.7	15.0	0.00
		fixed	9.29	9.3	49.6	3.4	5.9	33.3	0.8	1.9	12.0	0.01
		refit	9.78	9.8	51.3	3.6	6.2	32.8	0.7	3.0	12.2	0.01
	dev. FA 1%	baseline	1.00	56.2	98.4	39.3	69.3	3.2	0.0	0.0	1.1	0.00
		fixed	1.00	52.0	94.1	46.3	62.0	4.3	0.0	0.0	1.4	0.00
		refit	1.00	51.9	93.4	42.6	60.2	3.6	0.0	0.0	1.2	0.00
	dev. FA 5%	baseline	5.00	31.2	87.0	12.3	32.2	13.7	0.3	0.8	4.9	0.00
		fixed	5.00	23.2	72.9	14.1	19.4	17.0	0.2	0.2	5.8	0.01
		refit	5.00	24.8	75.3	13.6	20.1	14.8	0.2	0.3	5.1	0.00
	dev. FA 10%	baseline	10.00	17.8	73.6	3.3	13.4	29.0	1.0	4.2	11.4	0.00
		fixed	10.00	7.9	47.4	2.6	4.7	35.7	0.8	2.5	13.0	0.01
		refit	10.00	9.6	51.3	3.6	6.1	33.3	0.8	3.0	12.4	0.01

Table 6: The runs behind the speech results for the w2v-BERT 2.0 configuration of Fig. 1; surprise false rejection (%) on the two EmoFake test speakers. The original run, the repeat run and the reference-split row of the allocation grid are separate trainings of the same recipe and differ by at most one utterance of 700 in any rate; the recalibration part of the mean surprise margin shift rounds to $+0.03$ and $+0.04$ logits in the two runs that record it. Table 4 is a different estimand: its eraser is re-estimated on each of five partitions of the estimation speakers and its rows are means over them, so its paired-PCA rates are not re-runs of Fig. 1. Its rank-7 PCA rows come from a rerun on the same partitions that reproduces its other rows exactly. in/out = surprised utterances entering and leaving rejection under fixed projection; recal. = recalibration part of the mean surprise margin shift under fixed projection (logits).

result	run	subspace	aggregation	baseline	fixed	refit	in/out	recal.
Fig. 1	original run	all estimation speakers	two test speakers	17.3	33.9	28.4	121/5	+0.03
Table 4	retrained	per estimation partition	mean of 5 partitions	17.1	30.9	26.9	–	–
Table 1	repeat run	all estimation speakers	two test speakers	17.1	33.9	28.4	122/5	+0.04
Table 21	allocation grid	per allocation	reference split	17.3	33.9	28.3	121/5	–

tions, the recent-model replications with encoders and generation sources fixed before any result was seen, the speaker-allocation enumeration, and the converged-MLP replication. Specified in advance and not met were an external operating-point stress test on cloned RAVDESS actors and a blind-generator test, both summarized in the next paragraph. The rank sweep, the eraser comparisons, the estimation-data comparison, the training-data comparison, the allocation-specific configurations, the two-nuisance decomposition and the head-family, LoRA and processing checks were designed after the reference case had been read; each ran under a protocol written beforehand and first reproduces a stored result. Negative outcomes are kept.

Analyses outside the paper. Four analyses bear on the paper's question only indirectly and are summarized here; their files will be released with the code. *External operating-point stress test, specified in advance and not met.* On zero-shot clones of the 24 RAVDESS actors, the fixed heads score the clones below the actors' own recordings, with median AUROC 0.12–0.40 on w2v-BERT 2.0 and 0.20–0.26 on WavLM, so 99–100% of the clones pass the unchanged EVC-calibrated threshold, and

a threshold recalibrated on 12 actors rejects 64% of held-out genuine utterances while accepting 49–88% of the clones. The pre-specified criterion, a bound on clone acceptance together with a rise of more than 1 pp in the surprise minus neutral differential in at least 18 of 20 actor splits, held in 12 splits for the differential clause and was vacuous for the bound at these error rates. *Blind-generator test, specified in advance and not met.* Of three pass criteria frozen before any Kandinsky 2.2 image was scored, the sign-prediction rule met two and missed the third by 0.4 points; a smoothed direct audit added afterwards reaches the same 95.3%. *Text detection.* On the original RoBERTa (Liu et al., 2019) head trained on HC3 (Guo et al., 2023), deleting a learner-edit subspace estimated on ICNALE essays (Ishikawa, 2023) narrows the false-flag gap between the TOEFL and US essays of Liang et al. (2023) while GPT-3 acceptance rises from 45.8% to at least 78.6% over eleven invariant heads, and the cost persists in the ten unrestricted candidates of the same search. *MSP-Podcast.* With the ASV19-only heads and 160 podcast speakers that enter no fit, fixed projection lowers the difference between happy and neutral rejection by 4.7 points [3.4, 6.1] while raising neutral rejection from 11.8% to 16.8% until the threshold is recalibrated.

B SPEECH

B.1 THE AUDIT SUBSPACE AND ITS RANK

Capture and its null. Emotion subspaces are estimated on acted corpora from frozen features. On WavLM the rank-4 between-class capture is 0.55 against a label-shuffle null whose 99th percentile is 0.52; directions from half the speakers capture 65% of held-out speakers' displacements, against 0.2% for isotropic random subspaces of the same rank, and 29% of the unseen ESD corpus's; w2v-BERT 2.0 replicates at the same depth indices with held-out-speaker capture 0.51. In images, under the primary q90 pipeline, the CLIP subspace captures 0.32 of held-out-image displacements against 0.03 for the top-5 PCA of the downstream covariance and 0.02 for covariance-matched random subspaces. A correspondence-permutation null, which permutes the class label of an already-formed displacement inside its context and re-estimates the rank-K subspace, the span of the first K QR vectors of the class-centroid matrix in a fixed class order, from the permuted labels at every permutation, puts the observed values above the 99th percentile in both modalities, speech 0.55 against a null mean of 0.40 with a 99th percentile of 0.52 and $p{=}0.002$ over 1,000 within-speaker shuffles, and images under q90 0.31 against 0.11 with a 99th percentile of 0.15 and $p{=}0.001$ over 1,000 within-image shuffles. Capture measures what the subspace contains, not how removing it changes decisions.

Rank sweep. Table 7 repeats the configuration of Fig. 1 at every paired-PCA rank from 1 to 8, with the stored run reproduced exactly at $K{=}4$. Surprise rejection falls slightly at $K{=}1$, from 17.3% to 15.3% and 15.6%, and rises under both interventions at every rank from 2 to 8, most at $K{=}5$ with 35.4% and 29.0%. R_{in} is positive at $K{=}1$ and negative from $K{=}2$ on; on WavLM it is positive at every rank and every rank lowers surprise rejection. The stability-selected rank $K^{*}{=}2$, the last rank before the median largest principal angle to the full-sample subspace over 100 context-bootstrap draws first exceeds 20°, is one of these ranks. In a separate retraining of the configuration, surprise rejection rises from 17.1% to 23.9% under fixed projection and to 21.4% after refitting at $K^{*}{=}2$, against 33.9% and 28.3% at $K{=}4$.

Re-estimating the subspace at the same rank. With head, coordinate, threshold rule and evaluation population fixed, we resample the estimation contexts with replacement and re-estimate U_K (Table 8). Capture is stable, within ± 0.03 of the full-sample value in the central 95% of draws. The subspace is less so. The median largest principal angle to the full-sample U_K is 10° for WavLM and 25° for w2v-BERT 2.0, the least stable direction rather than a rotation of the whole subspace. The intervention conclusions are direction-stable. For the w2v-BERT 2.0 head the in-subspace response is negative in 499 of 500 draws and fixed projection raises surprise rejection in 498; on WavLM the direction of every reported change is the same in all draws. Magnitudes are not. The $+16.6$ pp rise has a re-estimation median of $+13.9$ and a central 95% range of $+4.6$ to $+21.1$, and speaker 0014's in-subspace response, -0.17 in the full sample, is negative in only 71.6% of draws, against 99.8% for speaker 0018. These are sensitivity frequencies under re-estimation of U_K alone.

Table 7: Rank sensitivity of the configuration of Fig. 1. Paired PCA of rank K from the same displacements as U_4; fixed projection (F) keeps the head and bias, the projected refit (R) retrains on the projected features; every arm at its own development-TTS EER threshold. $K{=}4$ reproduces the stored run of Fig. 1 exactly. Distortion: mean over the test speakers' genuine utterances of $\|U_K U_K^\top z\|/\|z\|$. R_{in}: the baseline head's in-subspace surprise response (logits). Rates in %; the last two columns are the change in the 14 evaluation-only speakers' equally weighted mean false rejection, in points.

				surprise FR		neutral FR		TTS FA		14 spk. ΔFR	
encoder	K	distort.	R_{in}	F	R	F	R	F	R	F	R
w2v-BERT 2.0	0	–	–	17.3		3.1		11.2		–	
	1	0.22	+0.18	15.3	15.6	3.1	3.1	11.5	11.0	−1.0	−0.6
	2	0.27	−0.80	23.9	21.4	2.7	2.4	11.2	11.0	+3.0	+2.5
	3	0.31	−0.49	20.3	20.3	2.6	2.9	12.1	11.3	+1.5	+1.9
	4	0.34	−1.14	33.9	28.4	2.6	2.7	10.8	10.9	+3.1	+3.3
	5	0.37	−1.21	35.4	29.0	2.9	2.7	10.8	11.0	+2.8	+3.0
	6	0.38	−1.13	32.1	28.4	2.9	2.7	10.9	11.0	+2.4	+2.9
	7	0.40	−0.42	27.0	24.3	4.1	3.1	10.5	11.0	+1.7	+2.6
	8	0.41	−0.45	26.7	23.6	4.0	3.0	10.7	11.6	+1.7	+2.6
WavLM	0	–	–	66.0		1.7		15.0		–	
	1	0.22	+0.07	64.9	64.4	1.7	1.7	14.6	14.3	−0.1	+0.3
	2	0.26	+2.52	59.6	59.0	4.4	3.3	12.3	12.3	−5.6	−3.5
	3	0.29	+4.08	48.9	49.9	4.6	3.9	13.1	12.9	−21.5	−13.1
	4	0.34	+3.66	49.6	51.3	3.4	3.6	12.0	12.2	−20.6	−12.4
	5	0.35	+3.86	49.9	51.3	3.7	3.6	11.6	12.0	−19.9	−11.8
	6	0.37	+4.13	51.3	51.4	4.6	3.7	11.4	11.9	−20.5	−12.0
	7	0.38	+4.43	51.6	51.9	4.6	3.4	9.6	10.6	−26.3	−16.7
	8	0.39	+6.62	37.0	35.1	7.0	4.7	8.0	9.6	−36.2	−24.6

Table 8: Re-estimating the audit subspace at the *same* K. Heads, coordinate, threshold rule and evaluation populations are the published ones and stay fixed; only the nuisance-estimation contexts are resampled with replacement (speech 115 speakers, vision 600 images). Brackets are central 95% ranges over draws, conditional on the fixed head and evaluation sample; they and the shares describe sensitivity to subspace re-estimation, not uncertainty about new speakers, heads or training data or the probability of harm in a new deployment. "Angle" is the median largest principal angle in degrees between a draw's U_K and the full-sample U_K: the least stable direction, not a rotation of the whole subspace. R_{in} is the in-subspace part of the response the intervention deletes (speech surprise−neutral on the two EmoFake test speakers, vision blur). The main effect is fixed-head projection. The heads are retrained from the stored recipe; the WavLM one starts at 66.1% surprise rejection, one utterance above the 66.0% of §3, so its full-sample change is -16.6 rather than -16.4 pp.

configuration	draws	capture [95%]	angle	λ_K/λ_{K+1}	R_{in} [95%]; same sign
speech, w2v-BERT 2.0 + TTS	500	0.284 [0.257, 0.302]	25 [13, 63]	1.21	−1.14 [−1.50, −0.32]; 99.8%
speech, WavLM + TTS	500	0.285 [0.264, 0.302]	10 [7, 17]	1.67	+3.66 [+2.84, +4.32]; 100.0%
vision, CLIP/SD-turbo, q90	500	0.314 [0.312, 0.313]	9 [7, 12]	1.45	+4.66 [+4.33, +4.97]; 100.0%
main effect of the intervention under the same resampling					
w2v-BERT, ΔFR surprise	full +16.6 pp; median +13.9; 95% [+4.6, +21.1]; all [−1.3, +24.0]; same direction 99.6%				
WavLM, ΔFR surprise	full −16.6 pp; 95% [−19.3, −11.9]; same direction 100.0%				
vision, ΔFR blur / contrast	full −1.5 / −4.7 pp; 95% [−1.7, −1.5] / [−4.8, −4.5]; same direction 100.0%				
what does not carry over					
speaker split	0018: R_{in} −2.11, negative in 99.8% of draws; 0014: −0.17, negative in 71.6%				

B.2 Erasers on one footing

Table 9 is the full comparison behind Table 4. Paired PCA, at ranks 4 and 7, is fitted to the rows $\delta_{c,e}$ of the eraser-fitting speakers, each speaker's emotion centroids minus their mean. Unpaired centroid PCA, at ranks 4 and 7, is fitted to the eight emotion centroids pooled over all eraser-fitting utterances and centred by their own mean, so 7 is its full rank. LEACE is fitted to the utterance-level features with one-hot emotion labels; its measured rank is 7. Five partitions of the estimation speakers were fixed before any outcome was read, stratified by corpus so that RAVDESS and CREMA-D speakers

appear in every split, with the eraser-fitting, probe-training and probe-test speakers disjoint. Each eraser is estimated once per partition, and that fitted transform is applied to both the probe representation and the detector features. Probe hyperparameters come from a speaker-disjoint validation split inside probe-training only, with the same budget for every representation. Balanced accuracy is a residual decodability measurement, not a quantity of information removed.

At rank 4, under fixed projection on w2v-BERT 2.0, paired minus unpaired surprise rejection is $+17.8$ pp on average and between $+15.4$ and $+20.3$ pp across the five partitions, while pooled TTS acceptance differs by -0.25 pp on average. At rank 7 both projections raise surprise rejection in every partition, and paired minus unpaired is $+3.1$ pp on average, between -2.0 and $+8.9$ pp. The eraser with the lowest probe accuracy is rank-7 unpaired PCA in all ten partitions of the two encoders, 26.1% on w2v-BERT 2.0 against 27.2% for LEACE, and both keep surprise rejection above the baseline under both interventions. After refitting, LEACE's surprise rejection exceeds that of the rank-4 unpaired subspace by $+4.5$ pp on average, between $+1.1$ and $+8.4$. On WavLM both rank-4 projections lower surprise rejection, the paired one more although it leaves more decodability; LEACE under fixed projection is the one construction that raises it there, to 70.5%, while also raising false acceptance from 15.0 to 23.2%, and its refit recovers both.

Table 9: Erasure constructions on one footing (App. B.2). Five partitions of the estimation speakers fixed before any outcome was read; within each, one fitted transform feeds both the probe representation and the detector features. The fixed arm scores $s_0(T(z))$ and the refit arm trains a new head on $T(Z_{\text{train}})$, so LEACE enters as its own affine map and is not forced to rank 4. "Probe" is residual multiclass decodability on held-out estimation-corpus speakers, not a quantity of information removed, as a mean over partitions with the standard deviation in brackets; "RAV s/n" is an auxiliary RAVDESS-only surprise-versus-neutral probe, a different question on a different scale. Rates are means over partitions, with the range for surprise. LEACE's measured rank is 7; the rank-7 PCA rows come from a rerun on the same partitions whose other rows reproduce this table exactly, and unpaired centroid PCA has no rank above 7. Distortion is the mean over the test speakers' genuine utterances of $\|T(z) - z\|/\|z\|$. These are re-estimated partitions, not the U_K estimated on all estimation speakers, of Fig. 1.

encoder	eraser	rank	distort.	probe BA (%)	RAV s/n	surprise FR (%)	neutral FR	TTS FA
w2v-BERT 2.0	no erasure	0	0.00	66.0 (3.8)	96.5	17.1 [17.1,17.1]	3.0	11.2
	paired PCA, rank 4	4	0.34	48.7 (2.7)	85.8	30.9 [27.9,34.4]	2.4	10.9
	projected refit	4	0.34	48.7 (2.7)	85.8	26.9 [24.3,29.0]	2.5	11.0
	unpaired centroid PCA, rank 4	4	0.30	41.5 (1.2)	82.1	13.1 [10.4,15.1]	3.3	11.2
	projected refit	4	0.30	41.5 (1.2)	82.1	14.1 [12.0,15.4]	3.1	10.2
	LEACE (affine)	7	0.43	27.2 (2.3)	80.2	23.8 [22.4,27.1]	4.1	13.1
	projected refit	7	0.43	27.2 (2.3)	80.2	18.7 [16.6,20.4]	2.8	10.0
	paired PCA, rank 7	7	0.40	41.4 (2.3)	83.1	26.3 [22.3,30.7]	4.0	10.8
	projected refit	7	0.40	41.4 (2.3)	83.1	24.0 [21.3,26.3]	2.9	11.3
	unpaired centroid PCA, rank 7	7	0.37	26.1 (2.1)	79.4	23.1 [19.9,26.7]	3.6	9.2
	projected refit	7	0.37	26.1 (2.1)	79.4	19.3 [17.4,21.0]	2.9	10.3
WavLM-large	no erasure	0	0.00	71.6 (4.7)	99.0	66.0 [66.0,66.0]	1.7	15.0
	paired PCA, rank 4	4	0.33	54.1 (5.3)	95.6	50.3 [47.7,52.4]	3.2	11.8
	projected refit	4	0.33	54.1 (5.3)	95.6	51.7 [49.1,53.1]	3.6	12.3
	unpaired centroid PCA, rank 4	4	0.32	44.7 (3.4)	87.5	56.6 [55.6,58.4]	2.8	8.8
	projected refit	4	0.32	44.7 (3.4)	87.5	59.6 [58.6,61.0]	3.3	11.1
	LEACE (affine)	7	0.60	30.1 (3.0)	85.0	70.5 [64.9,80.9]	9.4	23.2
	projected refit	7	0.60	30.1 (3.0)	85.0	55.9 [53.4,58.4]	3.1	12.0
	paired PCA, rank 7	7	0.38	44.8 (4.2)	86.2	51.5 [48.1,54.3]	4.0	10.3
	projected refit	7	0.38	44.8 (4.2)	86.2	52.5 [50.1,54.7]	3.5	11.1
	unpaired centroid PCA, rank 7	7	0.36	28.6 (3.6)	84.0	48.8 [45.3,51.1]	3.1	11.2
	projected refit	7	0.36	28.6 (3.6)	84.0	54.1 [52.3,55.6]	3.5	12.2

The same comparisons on the speakers of Fig. 2. A rerun rebuilt the 35 partition-level rows per encoder of the three-eraser comparison, reproducing every numeric column to 10^{-9}, and stored each transform, head and threshold. The stored objects then score the 14 speakers of App. B.4, which enter no threshold, head or eraser fit (Table 10). On w2v-BERT 2.0 the equal-rank reversal recurs on the ten Mandarin ESD speakers. Paired projection changes surprise rejection by $+5.1$ pp with interval $[+1.4, +9.6]$ and unpaired projection by -2.4 pp with interval $[-5.2, -0.5]$, and their difference is positive in all five partitions. The difference between the LEACE and unpaired refits is unresolved there, with an interval spanning zero and three of five partitions positive. On WavLM both comparisons keep the sign of the test pair.

Table 10: The two eraser comparisons of Table 4 on the speakers of Fig. 2: false rejection of the first eraser minus the second, in pp, averaged over the five partitions of the estimation speakers. The test-pair and Mandarin ESD columns use surprise. The 14-speaker column averages the emotion and style conditions within each speaker and then weights speakers equally, across 72 conditions in total; it includes the ten Mandarin ESD speakers, so the columns are not separate replications. The transforms, heads and development thresholds of the test-pair rows are applied unchanged; the transforms are fitted within these partitions, whereas Fig. 2 and the original-encoder rows of Table 20 use all estimation speakers. Differences are computed before rounding. The 95% intervals resample the ten Mandarin ESD speakers, or the ten ESD and four Expresso speakers separately, with one set of draws shared by every row, and are conditional on the five fitted transforms, the trained heads and the development calibration. *pos.* counts the partitions with a positive difference. No generated source accompanies these speakers, so acceptance costs come only from the test-pair fake pool of Table 4.

		two test speakers		Mandarin ESD, surprise			14 speakers	
encoder	comparison	diff.	pos.	diff.	95% interval	pos.	diff.	95% interval
w2v-BERT 2.0	paired − unpaired PCA, both rank 4, fixed	$+17.8$	5/5	$+7.6$	$[+2.5, +13.5]$	5/5	$+4.3$	$[+2.2, +6.6]$
w2v-BERT 2.0	LEACE (rank 7) − unpaired PCA (rank 4), refit	$+4.5$	5/5	-0.3	$[-1.7, +1.2]$	3/5	-0.1	$[-1.2, +0.9]$
WavLM	paired − unpaired PCA, both rank 4, fixed	-6.3	0/5	-13.8	$[-18.3, -9.7]$	0/5	-10.0	$[-12.1, -8.1]$
WavLM	LEACE (rank 7) − unpaired PCA (rank 4), refit	-3.7	0/5	-2.8	$[-4.2, -1.5]$	0/5	-1.1	$[-1.6, -0.6]$

B.3 WHERE THE ERASER IS ESTIMATED

Table 11 gives the full comparison behind Table 2, for both encoders and for four and five labels; Table 12 gives the diagnostics of the same erasers. Four labels give pooled centroids a rank of at most 3 and five labels at most 4, so paired PCA is run at ranks 3 and 4, unpaired centroid PCA at its full rank and LEACE at its measured rank; these rows are not compared with the eight-label constructions of Table 4. On the w2v-BERT 2.0 head, every external construction raises surprise rejection, to 22.0–37.7% under fixed projection and 20.4–30.0% after refitting, and every in-domain one lowers it, to 4.0–5.0% and 3.4–4.3%; with surprise among the labels the ranges are 21.6–36.4% and 21.1–29.4% against 4.0–5.9% and 2.9–3.9%. The in-domain erasers raise neutral rejection from 3.1% to 5.6–7.9% under fixed projection and leave pooled TTS acceptance at 10.6–11.7% against 11.2%.

LEACE removes all linear cross-covariance with the labels on its fitting data in both domains, yet a probe trained on Mandarin speakers still reaches 81.5% on held-out Mandarin speakers after the external LEACE and 77.3% after the in-domain one, against 82.3% without erasure. Over all four-label constructions the probe reads 80.7–82.4% after external and 76.5–77.5% after in-domain estimation, and with five labels 72.5–75.6% and 65.4–66.7% against 75.3% without erasure and 74.6% after the eight-label subspace of the paper; within a source its ordering of the four constructions follows the error in some sources and runs against it in others. The removed surprise response R_{rem} is computed on the test pair and is an after-the-fact reading. On WavLM every four-label construction lowers surprise rejection, the in-domain ones more, to 31.6–46.7% against 51.6–63.9% under fixed projection, at a cost in neutral rejection of 5.4–12.1% against 1.7%; with surprise among the external labels, LEACE raises it to 77.1% under fixed projection, and its removed response is negative, -3.6 logits.

Across allocations. Table 13 repeats the four-label comparison of Table 2 on all 168 allocations of App. B.10, each with its own head, standardization, thresholds and erasers, the in-domain erasers estimated on that allocation's six training speakers (`v090_indomain_alloc.py`; the official allocation reproduces Table 2 exactly). All four in-domain constructions lower surprise rejection on the allocation's test pair together in 154 allocations under fixed projection and in 159 after refitting, and on the five held-out Mandarin speakers in all 168 for every construction and arm except fixed LEACE, which lowers it in 149. The external constructions raise test-pair rejection in 87 to 123 allocations under fixed projection and in 90 to 106 after refitting, by medians of 0.3 to 1.9 points, and all four raise it together in 60 and 75 allocations; on the Mandarin speakers they raise it in 108 to 166 allocations under fixed projection, while external LEACE after refitting lowers it in all 168. The in-domain erasers cost a median 1.9 to 4.5 points of neutral rejection and lower pooled TTS acceptance by 0.7 to 1.8 points. The reversal of Table 2 is therefore general on its in-domain side and frequent but not universal on its external side, in line with §3.6.

Table 11: Where the eraser is estimated: outcomes, on the heads of Fig. 1. External: RAVDESS and CREMA-D; in-domain: the genuine ESD speech of the six training speakers; 4 labels are neutral, angry, happy and sad, 5 add surprise, 8 are the external subspace of the paper. F, R: fixed projection and projected refit, each at its own development-TTS threshold. Rates in %; 14 spk.: the 14 evaluation-only speakers' mean false rejection. Diagnostics of the same erasers in Table 12.

			surprise FR		neutral FR		TTS FA		14 spk. FR	
encoder	estimated on	eraser	F	R	F	R	F	R	F	R
w2v-BERT 2.0	none		17.3		3.1		11.2		15.1	
	external, 8 labels	paired PCA 4	33.9	28.4	2.6	2.7	10.8	10.9	18.1	18.4
	external, 4 labels	paired PCA 3	29.4	24.4	3.1	2.7	10.4	10.8	20.3	18.6
		paired PCA 4	37.7	30.0	2.7	2.7	11.0	11.7	19.2	19.0
		unpaired PCA	30.0	23.1	3.1	2.4	9.7	11.1	19.1	17.5
		LEACE	22.0	20.4	2.7	2.6	11.3	10.7	17.1	15.4
	in-domain, 4 labels	paired PCA 3	4.9	4.3	7.9	6.7	10.8	9.0	8.7	10.8
		paired PCA 4	5.0	4.3	7.7	6.9	10.6	9.0	8.9	10.8
		unpaired PCA	4.0	3.4	7.3	7.0	11.2	8.8	8.4	10.1
		LEACE	4.6	3.9	5.6	6.9	11.7	8.9	11.5	10.8
	external, 5 labels	paired PCA 3	21.6	21.1	1.9	2.4	12.2	11.4	18.3	17.9
		paired PCA 4	36.4	29.4	3.6	3.0	11.3	11.8	18.2	18.5
		unpaired PCA	26.4	22.4	3.0	2.4	9.3	10.6	17.9	17.2
		LEACE	28.9	21.9	2.7	2.4	12.1	10.9	18.5	15.5
	in-domain, 5 labels	paired PCA 3	5.9	3.9	8.0	6.9	10.3	9.0	9.0	10.7
		paired PCA 4	5.9	3.9	8.0	6.9	10.0	9.0	9.2	10.7
		unpaired PCA	4.1	2.9	7.0	6.6	11.1	8.8	8.4	9.7
		LEACE	4.0	3.4	6.1	6.4	11.7	9.1	11.4	10.3
WavLM	none		66.0		1.7		15.0		58.2	
	external, 8 labels	paired PCA 4	49.6	51.3	3.4	3.6	12.0	12.2	37.6	45.9
	external, 4 labels	paired PCA 3	54.0	55.7	4.0	3.7	11.8	11.7	40.3	47.4
		paired PCA 4	51.6	53.7	3.6	3.6	10.3	11.1	39.5	48.3
		unpaired PCA	62.6	62.1	3.0	2.0	13.4	13.4	53.5	55.5
		LEACE	63.9	62.9	3.1	1.7	12.7	14.0	58.1	58.3
	in-domain, 4 labels	paired PCA 3	32.1	26.9	12.0	8.9	9.7	9.6	29.8	38.6
		paired PCA 4	31.6	26.9	12.0	8.9	9.8	9.7	28.6	38.1
		unpaired PCA	36.0	30.6	12.1	9.0	9.5	9.3	32.7	40.1
		LEACE	46.7	34.1	5.4	8.6	16.2	9.9	50.6	41.7
	external, 5 labels	paired PCA 3	52.4	54.3	4.9	4.0	11.8	11.6	39.2	46.4
		paired PCA 4	50.9	52.9	3.6	3.6	10.4	10.9	39.0	47.8
		unpaired PCA	56.3	61.4	1.7	2.3	9.0	11.2	45.2	52.9
		LEACE	77.1	63.3	6.0	2.0	21.8	11.6	62.8	56.1
	in-domain, 5 labels	paired PCA 3	28.3	21.1	10.0	7.6	9.2	9.5	33.1	38.6
		paired PCA 4	27.6	21.3	10.7	8.3	9.3	9.5	29.4	37.1
		unpaired PCA	32.1	26.9	12.0	8.6	9.2	9.1	32.3	38.7
		LEACE	40.1	26.7	6.7	8.6	16.3	10.0	46.1	39.5

B.4 Speakers outside every detector, threshold and eraser fit

Both reference-split heads are reused exactly as stored, weights, standardizer, U_4 and the projected refit, and a regression check re-derives the stored ESD-en scores from them before any new corpus is scored. The new speakers are the ten Mandarin speakers of ESD, 0001–0010, and the four Expresso speakers, whose read styles, whisper, laughing, enunciated, confused, emphasis, happy and sad, are not the acted emotions U_4 was estimated from. Fig. 2 uses each head's own development threshold transferred unchanged and the speaker as the unit. There are 72 eligible conditions per encoder and intervention, each speaker's conditions are averaged with equal weight and the 14 speakers are then weighted equally. The projected refit gives $+3.31$ pp on w2v-BERT 2.0 (median speaker mean $+1.80$; 14 of 14 speaker means positive, 9 above $+1$ pp; ESD-zh $+3.84$, Expresso $+1.98$) and -12.40 pp on the linear WavLM head (median -11.27; 0 of 14 positive; ESD-zh -13.72, Expresso -9.09). Fixed projection gives $+3.06$ and -20.60 pp. A percentile bootstrap over speakers within each corpus, holding the stored heads, the thresholds and the eligible conditions fixed, puts the two refit means at $[+1.69, +5.11]$ and $[-15.20, -9.83]$ pp; only four speakers are Expresso, and the conditions are not independent replications. Read for surprise alone, the ten Mandarin ESD

Table 12: Where the eraser is estimated: diagnostics of the erasers of Table 11. Cov.: residual feature–label cross-covariance on the fitting data relative to before. Probe: balanced accuracy (%) of a probe trained on Mandarin ESD 0001–0005 and tested on 0006–0010 with the same labels; the baseline row uses five labels, and without erasure the four-label probe reaches 82.3% on w2v-BERT 2.0 and 89.1% on WavLM. Energy: share of the Mandarin within-speaker displacement energy removed. R_{rem}: the surprise response the eraser removes from the baseline head on the test pair, $w^{\top}(I - A)\bar{d}$ (logits).

encoder	estimated on	eraser	rank	cov.	probe	energy	R_{rem}
w2v-BERT 2.0	none		–	–	75.3	–	–
	external, 8 labels	paired PCA 4	4	–	74.6	0.24	−1.14
	external, 4 labels	paired PCA 3	3	0.22	82.4	0.15	−1.19
		paired PCA 4	4	0.17	80.7	0.16	−1.53
		unpaired PCA	3	0.00	80.9	0.14	−1.03
		LEACE	3	0.00	81.5	0.25	−0.76
	in-domain, 4 labels	paired PCA 3	3	0.25	77.0	0.31	+3.28
		paired PCA 4	4	0.21	76.5	0.37	+3.31
		unpaired PCA	3	0.00	77.5	0.35	+3.56
		LEACE	3	0.00	77.3	0.22	+2.76
	external, 5 labels	paired PCA 3	3	0.33	74.3	0.16	−0.87
		paired PCA 4	4	0.28	72.5	0.19	−1.26
		unpaired PCA	4	0.00	74.1	0.16	−0.78
		LEACE	4	0.00	75.6	0.54	−1.52
	in-domain, 5 labels	paired PCA 3	3	0.32	66.7	0.32	+3.15
		paired PCA 4	4	0.29	66.2	0.37	+3.19
		unpaired PCA	4	0.00	65.4	0.37	+3.64
		LEACE	4	0.00	65.8	0.21	+3.07
WavLM	none		–	–	79.0	–	–
	external, 8 labels	paired PCA 4	4	–	76.5	0.26	+3.66
	external, 4 labels	paired PCA 3	3	0.15	86.8	0.13	+3.46
		paired PCA 4	4	0.12	87.0	0.15	+3.65
		unpaired PCA	3	0.00	87.2	0.13	+1.40
		LEACE	3	0.00	86.0	0.44	+0.73
	in-domain, 4 labels	paired PCA 3	3	0.18	83.3	0.35	+7.66
		paired PCA 4	4	0.15	85.2	0.39	+7.69
		unpaired PCA	3	0.00	82.3	0.37	+7.30
		LEACE	3	0.00	82.7	0.20	+4.48
	external, 5 labels	paired PCA 3	3	0.23	76.7	0.15	+3.83
		paired PCA 4	4	0.20	76.7	0.17	+3.76
		unpaired PCA	4	0.00	77.1	0.15	+2.37
		LEACE	4	0.00	77.0	0.58	−3.58
	in-domain, 5 labels	paired PCA 3	3	0.31	75.6	0.34	+7.84
		paired PCA 4	4	0.27	75.8	0.39	+8.05
		unpaired PCA	4	0.00	74.5	0.37	+7.67
		LEACE	4	0.00	74.3	0.19	+5.63

speakers, 350 surprised utterances each, give 14.0% baseline rejection and 19.9% and 20.0% after fixed projection and refitting on w2v-BERT 2.0, where fixed projection raises rejection for seven speakers, leaves two unchanged and lowers it for one; on WavLM it falls from 78.2% to 48.7% and 60.9%, for all ten speakers under both interventions.

B.5 Added genuine data and the training objective

Table 14 is the comparison behind §3.3. Comparison A asks what additional target-domain genuine data does, with the authenticity-class totals and the fake-source mixture held fixed and only the genuine mixture moved, $(1 - \rho)$ on ASVspoof 2019 genuine and ρ on G4, the six training speakers' ESD genuine in neutral, angry, happy and sad, with ρ=0.5 fixed in advance. Comparison B asks what a different objective does with exactly those samples. Group-balanced weights and group DRO change only the weights inside the emotion-annotated block. The penalty is selected on development only, by one rule for every arm, the worst development loss over the four trained emotions on each authenticity side; it picked opposite edges of the pre-fixed grid in the two comparison-A arms,

Table 13: The four-label comparison of the main text's estimation-source table on all 168 speaker allocations of App. B.10, w2v-BERT 2.0, each allocation with its own head, standardization, thresholds and erasers, the in-domain erasers estimated on that allocation's six training speakers. For each construction, the number of the 168 allocations in which surprise false rejection on the allocation's own test pair rises under the external eraser or falls under the in-domain eraser, decided on utterance counts of 700, with the median change in points; the last two columns give the median change in neutral false rejection under the in-domain erasers. F is fixed projection and R projected refit, each at its own development threshold. `indomain_alloc_all168_per_construction.csv`, `indomain_alloc_all168_costs.csv`.

	external, surprise FR rises				in-domain, surprise FR falls				in-domain, neutral FR	
	F		R		F		R		F	R
construction	n	median	n	median	n	median	n	median	median	median
paired PCA 3	98	$+1.0$	90	$+0.4$	160	-12.4	160	-11.6	$+2.3$	$+1.9$
paired PCA 4	87	$+0.3$	99	$+1.7$	159	-12.0	160	-11.3	$+2.6$	$+2.2$
unpaired PCA	96	$+1.2$	106	$+1.9$	160	-12.5	162	-11.9	$+3.2$	$+2.9$
LEACE	123	$+1.3$	106	$+1.1$	156	-7.0	163	-12.1	$+4.5$	$+2.9$

so Table 15 repeats the comparison at every penalty. Surprise never selects anything, and neither the test speakers nor the 14 evaluation-only speakers enter any selection. Group DRO inherits the augmented ERM's $\lambda{=}10^{-5}$ and compares three step sizes; its outer loop reached its step cap at all three.

The added data help substantially and the erasure contrast survives that help. The improvement is measured against the $\rho{=}0$ arm of the same normalization, whose surprise rejection on w2v-BERT 2.0 is 20.4%, and not against the head of Fig. 1, which uses a different weighting convention. The augmented arms draw on genuine recordings from the same corpus and recording condition as the evaluation genuine speech, so part of the gain is domain match; the 14 evaluation-only speakers, from Mandarin ESD and Expresso, are the harder test, and the gain there is smaller. The objective arms stay close to the augmented ERM, so under the evaluated regularization and optimization settings they did not materially improve these outcomes. At every penalty the added data lower the 14 evaluation-only speakers' mean rejection, from 14.9–18.9% to 5.5–6.3% on w2v-BERT 2.0 and from 57.7–62.5% to 25.4–37.0% on WavLM. On the $\rho{=}0$ recipe the w2v-BERT 2.0 increase holds at every penalty, with surprise rejection rising by 11.1 to 16.7 points under fixed projection and 9.0 to 11.6 after refitting. On the $\rho{=}0.5$ recipe fixed projection raises the evaluation-only speakers' mean at every penalty on both encoders, by 4.1 to 5.7 and 4.8 to 11.8 points, while the projected refit changes it by -0.1 to $+0.5$ and -5.3 to $+1.0$. Table 16 crosses the same recipe, without and with G4, with the four external and the four in-domain erasers; it is the full version of Table 3.

B.6 The decision report

Table 1 comes from the repeat run of Table 6, the linear w2v-BERT 2.0 head with $K{=}4$; the baseline, fixed and refitted heads are each read at their own development threshold. We replayed that run at its original thread count with a flag that writes the scores, thresholds and decisions of every genuine test utterance and TTS fake; the replay's summary files are byte-identical to the stored ones, so the export belongs to the same run. Table 17 recomputes every row from the per-sample export and matches the run's summary. Entries and exits are counted into and out of rejection, so for genuine rows the change in false rejection equals entries minus exits divided by n, and for fake rows the change in acceptance has the opposite sign. This run is not the stored run behind Fig. 1. Its baseline surprise rejection is 17.1% rather than 17.3%, and its fixed-projection counts are 122 entries and 5 exits rather than 121 and 5. The EmoFake EVC spoofs were not part of the export; their acceptance, 0.0% in every model, comes from the stored summary of the same configuration.

Which shift reaches which utterance. A diagnostic permutes the fixed-projection shifts of Fig. 1 10,000 times within each speaker and emotion, keeping both marginal distributions (Table 18). For surprised speech the observed increase of $+16.6$ points lies inside the permutation range; shuffled pairings also produce large increases, concentrated in speaker 0018. For neutral speech the permutations give $+0.3$ against the observed -0.6, lower than 99.8% of the permutations, so the utterances near the boundary receive the shifts that carry them out of rejection.

Table 14: Added genuine data and training objective, separated (App. B.5). One linear head class, one coordinate, the original external U_K, one evaluation population, one development threshold rule, and a weighted *mean* log loss plus $\lambda\|\theta\|^2$. Held fixed across arms are the authenticity-class weighting and the selection procedure, *not* the selected penalty: the common development-only grid lands on 10^{-2} at ρ=0 and 10^{-5} at ρ=0.5 on both encoders, and group DRO does not repeat that search. Comparison A moves only the genuine mixture, $(1-\rho)$ on ASVspoof 2019 bona fide and ρ on G4 (the six training speakers' ESD genuine in neutral/angry/happy/sad), with ρ=0.5 fixed in advance. Comparison B runs on exactly those samples and changes only the weights *inside* the emotion-annotated block. G5 adds genuine surprise and is a different target-label access condition. The last column is the 14 evaluation-only speakers' equally weighted mean cell false rejection, genuine only. The manuscript's historical unweighted head uses a different weighting convention and is not crossed with these arms.

		genuine FR (%)			fake acceptance (%)		14 speakers
encoder	training arm	surprise	neutral	other	TTS	EVC	mean FR (%)
w2v-BERT 2.0	ERM, ρ=0	20.4	3.3	3.1	13.17	0.00	18.9
	fixed projection	31.6	3.1	3.5	12.28	0.00	19.1
	projected refit	29.4	2.6	3.2	12.14	0.00	20.3
	ERM, ρ=0.5 (adds G4)	0.0	0.0	0.1	0.14	0.00	5.9
	fixed projection	1.7	0.6	1.2	0.08	0.00	11.6
	projected refit	0.0	0.0	0.1	0.13	0.00	5.8
	group-balanced, ρ=0.5	0.0	0.0	0.1	0.14	0.00	5.9
	group DRO, ρ=0.5	0.0	0.0	0.1	0.14	0.00	5.9
	ERM, ρ=0.5 with G5	0.0	0.0	0.2	0.19	0.00	5.6
WavLM-large	ERM, ρ=0	70.7	1.6	11.4	20.09	0.00	62.5
	fixed projection	55.4	3.9	7.0	15.14	0.01	45.0
	projected refit	56.6	4.1	7.6	15.17	0.01	51.6
	ERM, ρ=0.5 (adds G4)	1.0	0.0	0.0	0.01	0.29	25.4
	fixed projection	4.7	0.0	0.0	0.11	0.57	37.2
	projected refit	0.3	0.0	0.0	0.08	0.68	20.1
	group-balanced, ρ=0.5	1.0	0.0	0.0	0.01	0.29	25.4
	group DRO, ρ=0.5	1.0	0.0	0.0	0.01	0.29	25.4
	ERM, ρ=0.5 with G5	0.4	0.0	0.0	0.01	0.31	21.4

Table 15: The genuine mixture of Table 14 crossed with its whole penalty grid. Same objective, weights, fake-source mixture, U_4, populations and threshold rule; †marks the penalty the development-only rule selects for that mixture, whose rows reproduce Table 14 exactly. B, F, R: baseline, fixed projection, projected refit. Rates in %; the last three columns are the 14 evaluation-only speakers' equally weighted mean false rejection.

			surprise FR			TTS FA			14 spk. FR		
encoder	ρ	λ	B	F	R	B	F	R	B	F	R
w2v-BERT 2.0	0	10^{-5}	16.9	33.6	28.1	11.14	10.77	10.99	14.9	18.1	18.2
	0	10^{-4}	17.6	34.1	29.1	11.35	10.71	10.86	15.3	18.0	18.5
	0	10^{-3}	17.9	33.9	29.4	11.74	11.08	11.33	16.3	18.0	18.7
	0	$10^{-2\dagger}$	20.4	31.6	29.4	13.17	12.28	12.14	18.9	19.1	20.3
	0.5	$10^{-5\dagger}$	0.0	1.7	0.0	0.14	0.08	0.13	5.9	11.6	5.8
	0.5	10^{-4}	0.0	1.6	0.0	0.21	0.10	0.15	5.5	11.2	5.9
	0.5	10^{-3}	0.0	1.1	0.0	0.29	0.11	0.17	5.6	9.7	6.2
	0.5	10^{-2}	0.0	1.7	0.0	0.31	0.15	0.29	6.3	10.7	6.5
WavLM	0	10^{-5}	65.4	49.4	51.3	14.84	11.77	11.99	57.7	37.1	45.5
	0	10^{-4}	66.9	51.0	52.1	15.65	12.27	12.39	59.2	39.0	46.9
	0	10^{-3}	68.4	54.0	54.0	17.38	12.96	13.35	60.7	41.6	49.3
	0	$10^{-2\dagger}$	70.7	55.4	56.6	20.09	15.14	15.17	62.5	45.0	51.6
	0.5	$10^{-5\dagger}$	1.0	4.7	0.3	0.01	0.11	0.08	25.4	37.2	20.1
	0.5	10^{-4}	1.1	5.3	1.0	0.01	0.11	0.03	27.2	37.9	28.2
	0.5	10^{-3}	1.3	4.4	1.4	0.06	0.14	0.06	28.5	36.9	28.2
	0.5	10^{-2}	2.1	5.4	1.7	0.11	0.15	0.11	37.0	41.8	32.7

B.7 HEAD FAMILY ON ONE COORDINATE

Table 19 is the grid behind the head-family sentence of §3.6, five ℓ_2 strengths fitted once each and four perceptron architectures at three seeds each, per encoder, with the coordinate, the head training

Table 16: G4 used to estimate the eraser or to train the detector. The weighted recipe of Table 14 without and with G4 in the genuine training class, each at its development-selected penalty, crossed with the four external and the four in-domain constructions of Table 11; ranges over the four. F, R: fixed projection, projected refit. Rates in %. The no-erasure baselines reproduce Table 14 exactly.

encoder	detector training	eraser		surprise FR	neutral FR	TTS FA	14 spk. FR
w2v-BERT 2.0	without G4	none		20.4	3.3	13.17	18.9
		external, 4 labels	F	25.3–32.7	3.3–4.1	11.67–13.07	19.5–21.5
			R	23.7–29.7	2.3–3.3	12.38–12.75	17.9–20.8
		in-domain, 4 labels	F	3.1–4.4	7.7–10.1	11.79–13.03	9.9–14.1
			R	1.7–2.6	7.0–8.0	10.90–11.57	11.3–12.8
	with G4	none		0.0	0.0	0.14	5.9
		external, 4 labels	F	0.0–0.7	0.0–0.4	0.06–0.39	6.1–10.1
			R	0.0	0.0	0.13–0.17	5.8–6.1
		in-domain, 4 labels	F	0.0	0.0	0.11–0.14	5.9–6.3
			R	0.0	0.0	0.10–0.14	5.9–6.1
WavLM	without G4	none		70.7	1.6	20.09	62.5
		external, 4 labels	F	56.7–71.0	3.0–5.7	12.86–17.03	46.2–62.5
			R	58.3–67.4	1.7–4.6	13.90–17.03	52.3–63.6
		in-domain, 4 labels	F	35.3–46.4	5.7–16.3	11.64–19.97	35.3–56.2
			R	28.1–34.0	11.7–12.4	11.42–13.11	43.6–46.6
	with G4	none		1.0	0.0	0.01	25.4
		external, 4 labels	F	0.9–7.7	0.0	0.11–0.51	29.8–47.0
			R	0.3–0.9	0.0	0.04–0.08	20.6–24.2
		in-domain, 4 labels	F	0.6–0.9	0.0	0.01–0.03	21.4–24.9
			R	0.3–1.0	0.0	0.01–0.06	19.8–25.6

Table 17: The filled report of Table 1, with aggregate EVC results reported separately: every genuine group of the two EmoFake test speakers and every fake source, each head at its own development-TTS threshold. Rate = false rejection for genuine rows and acceptance for fake rows (%); in/out = samples entering and leaving rejection relative to the baseline, so Δ in points is $100\,(\text{in}-\text{out})/n$ for genuine rows and $100\,(\text{out}-\text{in})/n$ for fake rows; margin = mean h_i in logits. Every row except EVC* is recomputed from the per-sample export and matches the run's stored summary. *The EVC spoofs are not in the per-sample export: their acceptance comes from the stored summary of the same configuration, and no sample identifiers, scores or crossings exist for them.

			fixed projection				refit			
side	unit	n	rate	Δ	in/out	margin	rate	Δ	in/out	margin
genuine	angry	700	6.4	+1.3	13/4	+0.52	5.3	+0.1	4/3	+0.07
	happy	700	5.9	+1.4	12/2	+0.94	5.0	+0.6	5/1	+0.43
	neutral	700	2.6	−0.4	2/5	+0.42	2.7	−0.3	1/3	−0.04
	sad	700	0.7	−0.1	0/1	+0.56	0.9	+0.0	0/0	+0.36
	surprise	700	33.9	+16.7	122/5	+1.55	28.4	+11.3	83/4	+1.08
	all genuine	3500	9.9	+3.8	149/17	+0.80	8.5	+2.3	93/11	+0.38
fake	CosyVoice	2399	30.8	+0.0	49/49	−0.08	30.3	−0.5	23/11	+0.06
	F5-TTS	2400	0.4	+0.0	1/2	−0.36	0.4	+0.0	1/2	−0.03
	StyleTTS 2	2400	1.0	−1.4	33/0	+0.87	2.0	−0.5	12/1	+0.02
	all TTS	7199	10.7	−0.4	83/51	+0.15	10.9	−0.3	36/14	+0.02
	EmoFake EVC*	–	0.0	+0.0	–	–	0.0	+0.0	–	–

Table 18: Which shift reaches which utterance. Stored scores of Fig. 1, fixed projection: the observed change in false rejection (points) against 10,000 permutations of the shifts h_i within each speaker and emotion, which keep the distributions of g_0 and h. A score-level diagnostic, not a realizable intervention. Mean shift and crossings of the reporting run (Table 17): neutral +0.42 logits, 2 in / 5 out, ΔFR -0.43; sad +0.56 logits, 0 in / 1 out, ΔFR -0.14; surprise +1.55 logits, 122 in / 5 out, ΔFR +16.71.

emotion	speakers	n	observed	permuted mean	permuted 95%
surprise	both speakers	700	+16.57	+17.70	[+16.29, +19.14]
surprise	0014	350	+0.86	+2.59	[+0.86, +4.29]
surprise	0018	350	+32.29	+32.81	[+30.57, +35.14]
neutral	both speakers	700	−0.57	+0.32	[−0.29, +0.86]
neutral	0014	350	−1.14	+0.66	[−0.57, +1.71]
neutral	0018	350	+0.00	−0.01	[−0.29, +0.00]

data and U_K held at the main setting. On w2v-BERT 2.0 surprise rejection rises in every logistic setting, by $+15.1$ to $+18.3$ pp under fixed projection and $+11.0$ to $+12.1$ after refitting, and at every one-hidden-layer width, where the seed medians are $+3.0$ to $+6.4$ and $+5.0$ to $+5.7$ pp. For any head g, $g(P_\perp z)$ is exactly invariant to the removed directions, but only for a linear head is the score change the additive term $-w^\top U_K U_K^\top z$; for a nonlinear head the refit row, a head trained from scratch on $P_\perp z$, is what the claim rests on. On w2v-BERT 2.0 the refit $(256,)$ head raises surprise rejection from 4.9% to 10.7% while its EmoFake EER improves, the median going from 4.7% to 3.3%, and neutral rejection stays at 0.7%; on WavLM the refit perceptron is also a worse detector (median EmoFake EER 21.5→29.8%), so part of its extra rejection there is lost discrimination. Two exceptions limit the pattern. On w2v-BERT 2.0 the two-layer $(256, 64)$ head is the one setting whose test-pair surprise rejection does not rise, yet its 14-speaker means are still positive for 10 of 14 speakers under fixed projection and 12 of 14 after refitting. On WavLM the direction splits by head class and by population at once. Every logistic setting improves the test-pair surprise stratum and every perceptron worsens it, while on the 14 evaluation-only speakers every family improves.

Table 19: Head family on one coordinate (App. B.7). Every effect is the paired difference inside one (encoder, family, hyperparameters, seed), taken before anything is averaged. Logistic heads are fitted once per C; perceptron entries are the median over three seeds with the seed range in brackets, and seeds are training repetitions, not speakers or deployments. Reference column: change in surprise false rejection on the two evaluation speakers, each head at its own development threshold. New-speaker columns: each speaker's equally weighted mean over its eligible conditions, then speakers weighted equally, with the count of positive speaker means. The $(256,)$ rows reproduce the stored 256-unit one-hidden-layer check exactly. On w2v-BERT 2.0 that check's median EmoFake EER falls from 4.7% to 3.3% after refitting while its surprise rejection rises from 4.9% to 10.7%.

		reference-split Δsurprise FR (pp)		14 evaluation-only speakers, ΔFR (pp)			
encoder	head	fixed	refit	fixed	> 0	refit	> 0
w2v-BERT 2.0	ℓ_2 logistic, C=0.01	+15.1	+12.1	+1.5	11/14	+2.2	10/14
	C=0.1	+16.4	+11.6	+2.6	12/14	+2.9	13/14
	C=1 (primary)	+16.6	+11.0	+3.1	12/14	+3.3	14/14
	C=10	+17.0	+11.7	+3.1	12/14	+3.3	14/14
	C=100	+18.3	+11.4	+3.7	13/14	+3.3	13/14
	MLP $(128,)$, 3 seeds	+6.4 [+6.1,+6.9]	+5.1	+11.6	13/14	+9.4	14/14
	$(256,)$, 3 seeds	+6.0 [+4.1,+6.6]	+5.7	+12.6	14/14	+13.7	14/14
	$(512,)$, 3 seeds	+3.0 [+1.9,+3.1]	+5.0	+6.8	13/14	+11.9	14/14
	$(256, 64)$, 3 seeds	-0.7 [-1.6,-0.1]	0.0	+2.0	10/14	+8.1	12/14
WavLM-large	ℓ_2 logistic, C=0.01	-15.1	-14.3	-19.0	0/14	-11.5	0/14
	C=0.1	-14.6	-14.6	-19.9	0/14	-11.9	0/14
	C=1 (primary)	-16.4	-14.7	-20.6	0/14	-12.4	0/14
	C=10	-15.6	-13.9	-20.5	0/14	-12.0	0/14
	C=100	-16.6	-14.1	-20.9	0/14	-12.5	0/14
	MLP $(128,)$, 3 seeds	+8.1 [+7.7,+13.4]	+15.1	-2.0	7/14	-5.7	2/14
	$(256,)$, 3 seeds	+3.7 [-2.4,+3.9]	+11.4	-8.3	0/14	-6.1	1/14
	$(512,)$, 3 seeds	+5.3 [+4.9,+12.7]	+12.1	-11.6	0/14	-10.5	1/14
	$(256, 64)$, 3 seeds	+7.6 [-0.1,+8.1]	+11.6	-9.1	2/14	-7.7	2/14

B.8 TASK-ADAPTED ENCODERS

This check asks whether the increase in surprise rejection needs an encoder that never saw the authenticity task. For each encoder and three seeds we trained rank-8 LoRA adapters (Hu et al., 2022) on the query and value projections of the zero-indexed transformer blocks 0 to 22, which produce hidden states 1 to 23 in the indexing of App. A and so the states 9 and 23 the head reads, jointly with a zero-initialized linear head on the original standardization. Training used one epoch of the central mixture with an unweighted cross-entropy and no emotion term, AdamW at learning rates 10^{-4} for the adapters and 10^{-3} for the head, 16-utterance accumulation and a linear warm-up and decay; the end-of-epoch checkpoint is used without selection. Each checkpoint was then fixed and merged, every set was re-extracted, and the standardization and the rank-4 context-paired U_K were re-estimated on its features. The adapters change all 46 targeted weight matrices, by 0.75 to 0.80% of their Frobenius norm on average.

Table 20 lists the two comparisons for every seed. The fixed comparison starts from the head trained jointly with the adapter, whose test EER on w2v-BERT 2.0 is 2.9 to 4.4%; the refit comparison starts from a logistic head refitted to convergence on the adapted features, at 2.1 to 2.9%. On w2v-BERT 2.0 surprise rejection rises in every seed under both comparisons, by $+3.7$ to $+5.7$ pp and

Table 20: Erasure on LoRA-adapted encoders. Each checkpoint was trained once with the head, then fixed; the standardization and U_K were re-estimated on its features before erasure. *Fixed vs joint* projects the head trained jointly with the adapter and compares it with that head; *refit vs converged* compares the projected refit with an unprojected logistic head refitted to convergence on the same features. The *frozen* rows apply the same code to the original features, where both comparisons start from the retrained converged head, a different run from Fig. 1. EER, development and test, belongs to the reference head of the row; every arm is recalibrated at its own development EER threshold. Surprise false rejection and TTS acceptance are on the two test speakers, the last two columns on the ten Mandarin ESD speakers with a 95% speaker-bootstrap interval. Seeds are training repetitions, not independent populations. All values in % or pp.

			reference EER	two test speakers			Mandarin ESD, surprise	
encoder	encoder state	comparison	dev. / test	surprise FR	diff.	TTS accepted	diff.	95% interval
w2v-BERT 2.0	frozen	fixed vs converged	6.4 / 9.2	17.1 → 33.9	+16.7	11.2 → 10.7	+5.9	[+1.5, +11.0]
		refit vs converged	6.4 / 9.2	17.1 → 28.4	+11.3	11.2 → 10.9	+6.0	[+2.1, +10.7]
	LoRA 0	fixed vs joint	2.0 / 3.0	5.7 → 11.1	+5.4	3.8 → 3.1	+1.1	[+0.3, +2.1]
		refit vs converged	1.5 / 2.4	3.9 → 7.4	+3.6	3.0 → 2.3	+1.4	[+0.4, +2.6]
	LoRA 1	fixed vs joint	2.7 / 4.4	7.1 → 12.9	+5.7	7.2 → 6.0	+1.7	[+0.7, +3.0]
		refit vs converged	1.4 / 2.9	6.6 → 8.7	+2.1	3.4 → 3.0	+1.8	[+0.7, +2.9]
	LoRA 2	fixed vs joint	1.8 / 2.9	1.7 → 5.4	+3.7	5.2 → 4.4	+0.8	[+0.3, +1.3]
		refit vs converged	1.4 / 2.1	2.0 → 4.3	+2.3	3.4 → 2.7	+1.3	[+0.4, +2.4]
WavLM	frozen	fixed vs converged	12.8 / 17.0	66.0 → 49.6	−16.4	15.0 → 11.9	−29.5	[−37.0, −22.5]
		refit vs converged	12.8 / 17.0	66.0 → 51.3	−14.7	15.0 → 12.2	−17.3	[−23.5, −12.0]
	LoRA 0	fixed vs joint	9.9 / 15.5	71.6 → 44.1	−27.4	11.4 → 7.9	−39.9	[−50.2, −29.3]
		refit vs converged	8.3 / 15.2	74.1 → 45.6	−28.6	9.7 → 6.9	−34.5	[−44.5, −24.9]
	LoRA 1	fixed vs joint	9.1 / 14.3	66.0 → 37.0	−29.0	11.0 → 6.4	−42.5	[−52.7, −31.9]
		refit vs converged	7.8 / 13.7	62.9 → 31.7	−31.1	10.5 → 7.5	−36.4	[−47.4, −25.8]
	LoRA 2	fixed vs joint	8.9 / 14.7	74.0 → 54.3	−19.7	9.0 → 4.8	−29.0	[−35.3, −21.7]
		refit vs converged	7.8 / 14.4	69.6 → 41.3	−28.3	9.8 → 7.0	−33.1	[−44.5, −22.1]

+2.1 to +3.6 pp on the two test speakers, and on the Mandarin ESD speakers by +0.8 to +1.8 pp with every interval above zero; the absolute increases are smaller than on the frozen encoder. On WavLM every comparison lowers surprise rejection. The check covers one adapter recipe; full-parameter fine-tuning and training the encoder after the deletion is inserted are not tested.

B.9 Speech-processing nuisances

We also erased a non-emotional nuisance made by signal processing, Opus coding at 12 kb/s (Valin et al., 2012) and a 300 to 3400 Hz telephone band limit applied identically to genuine and generated speech, with a subspace estimated from clean and processed versions of the same estimation-corpus utterances and the clean-trained heads of §3. Erasure lowers the aggregate processed-genuine rejection of the two test speakers in seven of eight encoder and intervention configurations; the sole increase, +1.6 pp for WavLM under Opus with fixed projection, comes with a resampling-only control that also moves WavLM's decisions, so the codec contribution is not isolated. The plan, all outcomes and the random-subspace controls will be released with the code.

B.10 Speaker allocations

The grid draws the development pair from the four speakers of the EmoFake development and test sets, 0014, 0018, 0019 and 0020, and the test pair from the remaining eight; the other six train the head. This gives $\binom{4}{2}\binom{8}{2} = 6 \times 28 = 168$ allocations. Table 21a gives the change in surprise rejection on each allocation's test pair, each model at its own development threshold. On w2v-BERT 2.0 the change is positive in 36 allocations under fixed projection and in 72 after refitting, with means of −0.5 and +0.4 pp. The reference split's fixed-projection increase of +16.6 pp is exceeded by 5 allocations and its refit increase of +11.0 pp by 11. On WavLM the change is negative in 167 of the 168 allocations under both interventions. The grid's refit rates come from its own cache, whose reference-split refit rate of 28.29% is a separate run from the 28.4% of Fig. 1. The test speakers are the largest single component of the variance of $\Delta\mathrm{FR}$ over the grid in 12 of 16 cells of encoder, emotion and rule, with a share of 9–90%, ahead of the development pair, the training set and the residual.

The same configurations on common speakers. The 168 configurations differ in training set, coordinate, subspace, heads and thresholds, and each was applied, with its own objects, to the 14 evaluation-only speakers (Table 21b). On w2v-BERT 2.0 fixed projection raises the 14 speakers' mean in 33 configurations, with a median change of -0.9 points, and refitting in 125, with a median of $+0.4$; the reference split's $+3.1$ and $+3.3$ points are exceeded by one and two configurations. On the held-out Mandarin speakers 0006–0010 fixed projection raises surprise rejection in 159 of the 168 configurations, by a median of 3.3 points and by -3.2 to $+8.4$ points in all (Table 21c). On WavLM every configuration lowers the 14 speakers' mean under both interventions, and fixed projection lowers held-out Mandarin surprise rejection in all 168.

Table 21: The 168 speaker allocations of Fig. 3 read on three populations, three different estimands that are not pooled. Changes in false rejection (points) from the baseline head of the same allocation, each head at its own development-TTS threshold. (a) Surprise on the allocation's own English test pair; signs are decided on utterance counts out of 700. (b) The allocation-specific configurations scored on the 14 evaluation-only speakers, each speaker's mean over eligible conditions averaged over speakers; above: allocations whose change exceeds the reference split's; corr.: correlation with the change in (a). (c) Surprise of the held-out Mandarin speakers 0006–0010, 1,750 utterances, fixed projection only. Each configuration uses its own standardizer, subspace, heads and thresholds, so it changes several factors at once, and the allocations share training sets, calibration pairs and test speakers, so no binomial interval is attached. Reference: the reference split of the main text.

(a) English test pairs, surprise

encoder	intervention	rises	ties	falls	≥1 pp	≥2 pp	mean / median	5th / 95th pct.	reference
w2v-BERT 2.0	fixed projection	36	2	130	26	24	-0.5 / -1.7	-7.5 / 15.5	$+16.6$
	projected refit	72	2	94	54	47	0.4 / -0.3	-6.3 / 11.3	$+11.0$
WavLM	fixed projection	1	0	167	0	0	-12.7 / -12.8	-21.4 / -4.1	-16.4
	projected refit	1	0	167	0	0	-13.1 / -13.5	-21.0 / -3.9	-14.7

(b) 14 evaluation-only speakers, mean over conditions

encoder	intervention	rises	median	range	reference	above	corr.
w2v-BERT 2.0	fixed projection	33/168	-0.9	$[-5.1, +3.2]$	$+3.1$	1	0.79
	projected refit	125/168	$+0.4$	$[-1.9, +4.3]$	$+3.3$	2	0.55
WavLM	fixed projection	0/168	-20.7	$[-25.5, -14.6]$	-20.6	77	0.49
	projected refit	0/168	-12.3	$[-16.7, -6.1]$	-12.4	89	0.64

(c) held-out Mandarin 0006–0010, surprise

encoder	intervention	rises	ties	falls	median	range	reference
w2v-BERT 2.0	fixed projection	159	0	9	$+3.3$	$[-3.2, +8.4]$	$+7.7$
WavLM	fixed projection	0	0	168	-36.4	$[-49.2, -25.4]$	-33.6

Does the sign of R_{in} sort the allocations? Deleting U_K removes the in-subspace part of the response, so the sign of R_{in} suggests a direction for the change in surprise rejection. Table 22 computes it from three sets of speakers and reads it on two populations. On w2v-BERT 2.0, among the 166 allocations whose rejection count changes on the test pair, the test pair's own R_{in} gives the right direction in 124 and flags 30 of the 36 that worsen, but also 36 of the 130 that improve, while a rule that always predicts a decrease is right in 130; as a detector of worsening allocations it has a recall of 83.3%, a precision of 45.5% and a balanced accuracy of 77.8%. Computed on five Mandarin audit speakers, 0001–0005, in each configuration's coordinate, the same rule reaches a balanced accuracy of 87.7% on the held-out Mandarin speakers, with a recall of 75.5% and a precision of 100%, but its accuracy of 76.8% is below the 94.6% of always predicting an increase; on the English test pairs it reaches 63.1%. A pre-intervention signal from independent speakers thus carries information about the direction for speakers of the same corpus and language, and little across corpora. On WavLM R_{in} is positive in all 168 allocations, so the rule always predicts a decrease and misses the one allocation that worsens on its test pair. The check does not extend to refitted heads.

Table 22: Sign rules for fixed projection across the 168 allocations: predict an increase in surprise false rejection when $w_0^\top U_K U_K^\top \bar{d} < 0$, that is, when the in-subspace response R_{in} computed from $\bar{d}$ is negative. $\bar{d}$ comes from the Mandarin audit speakers 0001–0005, from the allocation's own test pair, whose rule therefore uses the contrast of the speakers it predicts and is not target-free, or from the training speakers; each rule is evaluated on the allocation's English test pair or on the held-out Mandarin speakers 0006–0010. Allocations whose rejection count does not change are left out. w2v-BERT 2.0 rows; majority: accuracy of always predicting the more frequent direction, a decrease on the test pairs and an increase on held-out Mandarin. On WavLM R_{in} is positive in all 168 allocations, so every rule predicts a decrease; one allocation worsens on the test pairs and none on held-out Mandarin. Rates in %.

$\bar{d}$ from	evaluated on	n	worse	recall	precision	bal. acc.	acc.	majority
Mandarin audit speakers	test pair	166	36	91.7	28.0	63.1	47.0	78.3
Mandarin audit speakers	held-out Mandarin	168	159	75.5	100.0	87.7	76.8	94.6
test pair	test pair	166	36	83.3	45.5	77.8	74.7	78.3
test pair	held-out Mandarin	168	159	42.1	100.0	71.1	45.2	94.6
training speakers	test pair	166	36	55.6	29.4	59.3	61.4	78.3
training speakers	held-out Mandarin	168	159	43.4	100.0	71.7	46.4	94.6

C IMAGES

C.1 TRAINING ASSOCIATION MANIPULATED

Every real and fake training image of the common-pipeline benchmark exists as a JPEG q75 and a q95 version (App. C.3); one seeded uniform per image orders the images, and within each class the $\mathrm{round}(n_{\text{class}}\, p)$ images with the smallest uniform take the q75 version, $p = (1+\gamma)/2$ for fakes and $(1-\gamma)/2$ for reals, so the q75 sets are nested across γ and the overall q75 fraction is exactly $1/2$ at every γ. The coordinate is the $\gamma{=}0$ training standardization; the audit subspace is the $K{=}5$ context-paired PCA of q75$-$q95 displacements of 600 independent validation images; the response is measured on 600 others; development and test sets contain both versions of every image and do not change with γ; one converged head per γ, thresholds by development EER.

The full grid. Fig. 5 shows the grid of nine encoders, six training generators, five assignment seeds and two nuisances, JPEG quality and Gaussian blur of severity 3, with the blur audit subspace estimated outside all training data. Every trajectory rises strictly with γ, and 260 of 270 JPEG and 245 of 270 blur trajectories reverse sign. In its first subset of 16 JPEG encoder and generator pairs, the change lies inside the audit subspace for 96–101% of its size for CLIP, MAE and SigLIP 2 and 73–77% for DINOv2, against 0–11% for covariance-matched random directions. Fixed projection and refitting hold the JPEG response below 0.78 logits in absolute value for every pair and seed, but under blur, deletion can cost detection accuracy, since fixed projection raises the EER at $\gamma{=}{+}0.8$ in 7 of the 20 runs. The $\gamma{=}0$ response is not zero, a residual of the originals' processing that the design does not isolate.

The same manipulation in speech. A unit is a quartet, one ESD speaker's recording of one sentence in neutral and in emotion e together with the EmoSpoof-TTS version of the same speaker and sentence in both emotions, kept only when the two real transcripts match. At γ a fraction $(1{+}\gamma)/2$ of each training speaker's quartets contributes its real-neutral and TTS-e recordings and the rest their real-e and TTS-neutral recordings, under exact per-speaker nested quotas, so speakers, sentences, emotion prevalence and class sizes are fixed and only the association between emotion and label moves. All 18 conditions of encoder, emotion and TTS and all 90 seed trajectories are strictly monotone with $R_{-0.8} < 0 < R_{+0.8}$, at ΔR of 2.3 to 5.4 logits, and R_0 ranges over -1.5 to $+1.3$. Only 10–58% of ΔR lies inside U_K, against 96–101% for three of the four original vision encoders, and the response almost never crosses a threshold, since these heads separate their own TTS source at an EER of at most 0.39% and emotional false rejection stays below 3.9% as a condition mean. EmoSpoof-TTS has no surprise, so this replicates the association effect across emotions and modality, not the result of §3.

Converged nonlinear heads. We repeated the association manipulation with a one-hidden-layer MLP of 256 units on four encoder and generator pairs, CLIP with SD-turbo, SD 1.5 and FLUX.2 and SigLIP 2 with FLUX.2, trained with adam over three training seeds with the quota assignment held at the primary run's, specifying before the fits that every response trajectory be strictly monotone

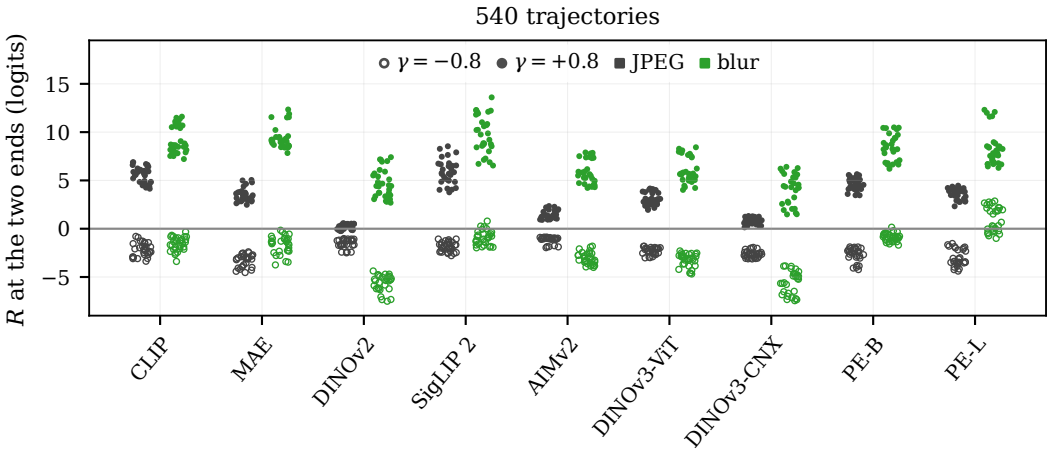

Figure 5: Every response trajectory of the association grid, nine encoders, six training generators, five assignment seeds and two nuisances, by its response at $\gamma{=}-0.8$ as an open mark and at $\gamma{=}+0.8$ as a filled mark. A sign reversal is an open mark below zero with its filled mark above.

in γ and change sign. All 12 trajectories are, with the linear heads' error pattern; all 120 fits, the baseline and the refit at each of five values of γ for the 12 trajectories, met their stopping criterion on the training loss, a tolerance of 10^{-6} with no validation split, at a training loss $\leq 1.2 \times 10^{-4}$.

C.2 Two nuisances associated independently

Every training image exists in four versions, JPEG q95 or q75, each clean or blurred with Gaussian blur of severity 3. Two independent exact-quota assignments set γ_J and γ_B over $\{-0.8, 0, +0.8\}$, so at every cell the q75 fraction and the blurred fraction are both exactly $1/2$ and only their associations with the label change. Training fakes come from SD-turbo and FLUX.2-klein; three assignment seeds; the two audit subspaces U_J and U_B, of rank 5 each, are estimated once outside all training data and held fixed. The two subspaces are far from aligned, with a smallest principal angle of 58.3–$59.4°$, and U_B captures 0.051–0.054 of the JPEG displacement and U_J 0.064–0.065 of the blur displacement, against 0.46 and 0.39 for their own. Regressing each response on both associations gives $\partial R_J/\partial\gamma_J = 4.72$ on SD-turbo and 5.83 on FLUX.2 against $\partial R_J/\partial\gamma_B = 0.07$ and 0.09. At $\gamma_J{=}+0.8$ the false rejection of the clean q75 reals is 8.83%, 1.75% and 0.45% for SD-turbo and 16.93%, 4.45% and 1.06% for FLUX.2 as γ_B goes from -0.8 to $+0.8$, while the seed-mean R_J stays between 5.21 and 5.79 logits on SD-turbo and between 5.50 and 6.65 on FLUX.2; the detector's EER moves with γ_B too, from 1.9% to 3.5–3.6% on SD-turbo and from 4.5% to 7.0–7.3% on FLUX.2. A paired image bootstrap of 2,000 draws, each resampling the 2,725 evaluation reals and the 600 audit-evaluation images once and applying the same draw to every cell with thresholds fixed, puts the seed-mean difference between $\gamma_B{=}-0.8$ and $+0.8$ at 15.9 points $[14.8, 17.0]$ in clean q75 false rejection and -0.30 logits $[-0.40, -0.21]$ in R_J on FLUX.2, and at 8.4 points $[7.5, 9.3]$ and -0.58 logits $[-0.65, -0.51]$ on SD-turbo; the ratio of the two false rejections is 15.9-fold $[12.3, 21.9]$ on FLUX.2. Table 23 repeats the design with the margins of the evaluation reals. At $\gamma_J{=}+0.8$ the contrast between the q75 and q95 versions of the same clean images, 5.6 to 6.8 logits on FLUX.2 and 5.2 to 5.8 on SD-turbo, tracks R_J and hardly moves with the blur association, while the mean signed score margin of the clean q95 reals falls from -12.3 to -19.8 and from -12.9 to -19.2 logits, or by 0.7 and 0.6 development score deviations.

C.3 Processing pipeline

The native benchmark's classes reach the encoder with different processing histories, JPEG-coded photographs against freshly written PNGs. The control decodes every image of both classes, resizes and re-encodes through one terminal pipeline, JPEG q90 as the primary with PNG, q75 and q95 also extracted, with detector, subspace and threshold re-estimated inside each condition; native, common PNG and common q90 are never pooled. A common lossless re-encoding leaves the blur failure largely intact, at 34–91%; the lossy step removes most of it, while the geometry remains,

Table 23: Why the clean q75 rejection moves at γ_J=+0.8 (CLIP, seed means). R_J, R_B: audit responses of §4. Contrast: mean over the evaluation reals of s(q75 clean) − s(q95 clean) on the same images, a different quantity from R_J. Margins: mean $s - \tau$ of the clean q95 and clean q75 reals (logits). SD: standard deviation of the development scores; the last two columns divide by it.

generator	γ_B	FR q75	R_J	R_B	contrast	q95 margin	q75 margin	SD	q95 margin/SD	contrast/SD
FLUX.2	−0.8	16.9	5.50	-1.78	5.60	-12.29	-6.69	12.05	-1.02	0.47
	+0.0	4.5	6.65	5.23	6.77	-19.33	-12.56	13.49	-1.43	0.50
	+0.8	1.1	5.80	9.91	5.86	-19.82	-13.96	11.68	-1.70	0.50
SD-turbo	−0.8	8.8	5.21	-0.88	5.24	-12.89	-7.65	11.46	-1.12	0.46
	+0.0	1.7	5.73	3.96	5.77	-18.38	-12.61	12.83	-1.43	0.45
	+0.8	0.5	5.79	8.44	5.83	-19.19	-13.36	11.45	-1.68	0.51

with held-out capture 0.32 against 0.35. Matching the pipeline does not end the problem. The heads trained against FLUX.2 fakes still falsely flag 18.3% of blurred and 25.7% of contrast-altered held-out reals on CLIP, and fixed projection removes these, to 0.8% and 2.3% on CLIP, at a same-source cost that refitting recovers, the CLIP EER going 0.26→0.70→0.30%. The cost is cross-generator discrimination, reported over the refreshed grid in App. D. The SD-trained heads detect FLUX.2 fakes at 6–19% EER without ever seeing them, so these evaluations are not at chance.

D MODEL REFRESH: RECENT ENCODERS AND GENERATION SOURCES

The reference speech analysis and the image association and two-nuisance experiments were extended to recent frozen encoders and recent generation sources, run after the original analyses were frozen, with the same pipelines, rules and seeds. Table 24 lists every checkpoint of the paper; each representation follows the model's documented pooling and prompt, and speech layers keep the relative depths of the reference configuration.

Table 24: Every checkpoint of the paper. Speech enters every encoder as 16 kHz mono audio, RMS-normalized and truncated at 20 s; images are resized to 224×224 pixels before each encoder's own processor. The revision of every checkpoint will be released with the code.

role	checkpoint	representation
speech	`microsoft/wavlm-large`; `facebook/w2v-bert-2.0`	transformer layers 9 and 23 of 24, mean over frames, concatenated
speech	`omniASR_W2V_1B`; `omniASR_W2V_300M`	the same relative depths, mean over frames
speech TTS	`Qwen/Qwen3-TTS-12Hz-1.7B-Base`	voice clones of the ten English ESD speakers (6 train / 2 development / 2 test)
speech TTS	`FunAudioLLM/CosyVoice2-0.5B`; `IndexTeam/IndexTTS-1.5`	RAVDESS clones of the external stress test (App. A)
vision	`openai/clip-vit-base-patch16`; `google/siglip2-base-patch16-224`	pooled output of the vision tower
vision	`facebook/vit-mae-base`; `facebook/dinov2-base`	CLS token (MAE with mask ratio 0)
vision	`apple/aimv2-large-patch14-224`	mean of last-layer patch tokens
vision	`facebook/dinov3-vitb16-pretrain-lvd1689m`; `facebook/dinov3-convnext-base-pretrain-lvd1689m`	CLS token / pooled output
vision	`facebook/PE-Core-B16-224`; `facebook/PE-Core-L14-336`	image embedding through the model's own transform, which resizes its input to 336 px for L14
vision gen.	`stabilityai/sd-turbo`; `stabilityai/sdxl-turbo`	1 step, guidance 0; 512 px
vision gen.	`stable-diffusion-v1-5/stable-diffusion-v1-5`	DPM++, 25 steps, guidance 7.5; 512 px
vision gen.	`black-forest-labs/FLUX.2-klein-base-4B`	50 steps, guidance 4.0; 512 px
vision gen.	`Qwen/Qwen-Image-2512`; `Tongyi-MAI/Z-Image-Turbo`	1024 px, 6,000 images each
vision gen.	`kandinsky-community/kandinsky-2-2-prior`; `kandinsky-community/kandinsky-2-2-decoder`	1,500 images of the blind check (App. A)

Reproducibility of the refresh. The extraction and generation code pins every revision. Logistic fits are not bit-reproducible across BLAS thread counts, so every refresh analysis was run at a fixed thread count, across which the speech rates move by at most one utterance of 700. Every result file used here will be listed with its SHA256 in the code release, where partial or duplicate runs are marked as superseded and never aggregated.

Speech. Table 25 repeats the reference analysis for w2v-BERT 2.0 and WavLM and two training TTS sources, the original EmoSpoof-TTS and Qwen3-TTS clones of the same ten English ESD speakers and emotions, 12,000 utterances with each clone prompted by a different same-speaker, same-emotion ESD utterance. In both cases six speakers train the head, two calibrate it and two are tested. The linear w2v-BERT head reproduces the stored reference run to one utterance, 122 entries against the stored 121. Both sources use the same speaker split and fitting procedure; the standardization, U_4 and the development and test fake pools are re-derived for each source, so the Qwen3-TTS rows repeat the analysis rather than change one variable. Neither training source gives a head that dominates on both fake pools. The EmoSpoof-trained w2v-BERT head reaches AUROC 0.967 on EmoSpoof and 0.469 on Qwen3-TTS, the Qwen3-TTS-trained head 0.770 and 0.849, and the WavLM heads 0.893 and 0.385, then 0.786 and 0.687. For the w2v-BERT head trained on Qwen3-TTS, 207 of 700 surprised utterances enter rejection under fixed projection and none leave, and the mean margin shift of the surprised utterances, $+2.40$ logits, splits into $+1.80$ from the score and $+0.59$ from recalibration. Two further encoders, omniASR-W2V 1B and 300M, detect poorly, with AUROC 0.46–0.70, and their results do not consistently support the effect, since under fixed projection their surprise rejection on the test speakers changes by -18.4 to $+1.3$ points. All four rows will be released with the code.

Table 25: Speech refresh with linear heads, $K=4$, the development-EER threshold and the two EmoFake test speakers. Rates in % and responses in logits; crossings are surprised utterances entering and leaving rejection under fixed projection. (b) The speaker-equal mean change in false rejection over the 14 evaluation-only speakers in points, with a 10,000-draw speaker bootstrap interval. The refresh is a separate run from App. B.4, and its bootstrap resamples the 14 speakers jointly, whereas App. B.4 resamples each corpus separately.

(a) test-pair decisions and response

				surprise FR			response		crossings	
encoder	TTS	AUROC	EER	baseline	fixed	refit	R_{in}	R	enter	leave
w2v-BERT 2.0	EmoSpoof	0.967	9.2	17.1	33.9	28.4	−1.14	+4.22	122	5
	Qwen3-TTS	0.849	23.5	24.9	54.4	42.6	−2.01	+4.70	207	0
WavLM	EmoSpoof	0.893	17.0	66.0	49.6	51.3	+3.66	+9.78	3	118
	Qwen3-TTS	0.687	34.0	77.7	57.3	61.6	+3.62	+10.04	1	144

(b) change over the 14 evaluation-only speakers

encoder	TTS	fixed projection	refit
w2v-BERT 2.0	EmoSpoof	+3.1 [+1.4, +5.1]	+3.4 [+1.7, +5.3]
	Qwen3-TTS	+10.4 [+7.4, +14.1]	+6.6 [+4.3, +9.1]
WavLM	EmoSpoof	−20.6 [−26.2, −15.4]	−12.4 [−15.5, −9.6]
	Qwen3-TTS	−19.2 [−24.9, −13.8]	−12.6 [−16.1, −9.4]

Images. Table 26 summarizes every encoder on the association grid of six training generators and five assignment seeds per nuisance, the two-nuisance design at $\gamma_J{=}{+}0.8$, same-source deletion on the static JPEG q90 heads and the 30 cross-generator transfers of each encoder. The Qwen-Image and Z-Image fakes were generated for this refresh with the older generators' prompt and seed protocol, 6,000 images each. PE-Core-L14's deletion removes the residual blur and noise failures at no same-source cost, yet its mean cross-generator EER rises from 1.05% to 1.59%, and to 1.36% after refitting. Over the nine encoders fixed projection raises the mean cross-generator EER of every encoder but only 221 of the 270 individual transfers, and after refitting 184 of 270 remain above baseline.

Table 26: Vision refresh per encoder. (a) Monotone and sign-reversing trajectories out of 30 per nuisance, the range of the share of ΔR inside U, and the clean q75-real false rejection in % at $\gamma_J{=}{+}0.8$ for three values of γ_B in the two-nuisance design, as seed means. (b) Deletion on the static JPEG q90 heads. Same-source EER in % as a mean over six generators, the cross-generator EER as a mean over 30 transfers, and the number of transfers whose EER rises.

(a) association and two nuisances

	JPEG		blur			FLUX.2, γ_B			SD-turbo, γ_B		
encoder	mono.	rev.	mono.	rev.	in U (%)	-0.8	0	$+0.8$	-0.8	0	$+0.8$
CLIP	30	30	30	30	94–104	16.9	4.5	1.1	8.8	1.7	0.5
MAE	30	30	30	30	96–103	22.6	4.9	1.2	10.5	2.3	0.7
DINOv2	30	20	30	30	73–94	6.3	1.7	0.6	4.6	0.9	0.3
SigLIP 2	30	30	30	26	94–102	17.7	3.9	1.1	10.5	2.5	0.8
AIMv2	30	30	30	30	91–102	5.8	1.1	0.3	3.1	0.5	0.2
DINOv3-ViT	30	30	30	30	88–98	10.2	2.5	0.5	5.6	1.3	0.3
DINOv3-CNX	30	30	30	30	74–82	9.5	2.5	1.2	6.1	1.7	0.7
PE-Core-B	30	30	30	29	95–101	11.6	1.9	0.2	4.8	0.7	0.1
PE-Core-L	30	30	30	10	98–101	7.4	0.6	0.1	1.0	0.1	0.0

(b) deletion on the static JPEG q90 heads

	same-source EER			cross-generator EER			transfers rising	
encoder	baseline	fixed	refit	baseline	fixed	refit	fixed	refit
CLIP	0.29	0.39	0.31	5.96	6.80	6.15	27	21
MAE	0.47	1.08	0.50	12.87	13.87	13.04	21	17
DINOv2	0.80	2.54	0.89	17.57	21.99	18.85	30	27
SigLIP 2	0.29	0.57	0.36	6.68	7.60	6.78	26	19
AIMv2	0.26	0.82	0.27	11.24	12.05	10.95	19	16
DINOv3-ViT	0.26	0.55	0.24	12.27	13.46	12.32	23	20
DINOv3-CNX	0.83	5.94	0.93	19.01	21.60	18.98	26	18
PE-Core-B	0.08	0.10	0.08	3.64	5.10	4.13	27	24
PE-Core-L	0.00	0.01	0.01	1.05	1.59	1.36	22	22