MIXTURE PROBING IN FROZEN AUDIO ENCODERS: SHORTCUT BOUNDS AND A SUPPRESSION LAW

Anonymous authors
Paper under double-blind review

ABSTRACT

Linearly mixed audio is an attractive testbed for representation analysis because the ground-truth sources are known exactly. The standard instrument is a probe: train a classifier on a frozen encoder's embedding of a mixture to name a source inside it, and read its accuracy as evidence about what the encoder retains. We show that this instrument is biased in two ways that have not been reported, quantify what survives once both are removed, and identify the mechanism of what survives. First, if mixtures are built by drawing k *distinct* classes, which is the obvious construction, the masker labels alone determine an AUC of $0.5 + (k{-}1)/(2(n{-}1))$ with no audio at all. On a ten-class corpus at $k{=}5$ this reaches 0.722, exceeding the value our own probe measured there; drawing classes i.i.d. drives the bound to exactly 0.5. Second, when sources are paired by natural co-occurrence, a predictor given only the masker's class reaches AUC 0.968 against the audio probe's 0.913, so the probe is reading the loud source and inferring. Third, with both shortcuts closed, weak-source decodability still falls with the number of sources at a matched conditional ceiling, $\beta = -0.098$ AUC per $\log k$ (95% CI $[-0.133, -0.062]$; 63 cells, three encoders, three corpora). Fourth, corrupting an oracle time-frequency mask at a controlled rate shows this capacity law is entirely a suppression effect: a reader that can remove 5 dB of masker energy pays only 0.015 of the 0.201 AUC that k costs at 0 dB, and at 19 dB it pays 0.001. Naming one class among many is not intrinsically harder; separating it is. Part of the remaining loss is architectural rather than informational: pooling over the encoder's own frame tokens recovers up to $+0.077$ AUC at $k{=}5$ from the same frozen features, and this margin widens with training budget while the clip vector saturates. We release the diagnostic suite with a full account of the hypotheses we falsified and of the controls we withdrew, so that the retractions can be audited beside the results.

1 INTRODUCTION

A frozen audio encoder maps a waveform to a vector. When the waveform contains several concurrent sources, it is natural to ask which of them that vector still describes. The question matters for interpretability, where sparse dictionary methods assume that concepts compose in representation space (Elhage et al., 2022; Bricken et al., 2023), and it is unusually tractable in audio, because a mixture can be assembled from known isolated recordings and the correct composition law in the spectral domain is textbook (Nadas et al., 1988; Varga & Moore, 1990; Gales & Young, 1996; Moreno, 1996). Unlike the synthetic ground truth used in dictionary-learning benchmarks, here both *which* sources are present and *how* they combine are known here while the sources stay real recordings.

The standard instrument is a probe. One trains a classifier on the encoder's embedding of the mixture to predict a source present in it, and reads the resulting accuracy as evidence about encoding. This paper is about how that instrument fails, what survives its repair, and why.

Our first two results are that the probe can succeed without reading the target source at all. If the sampler that builds mixtures draws k distinct classes, observing which classes appear as maskers excludes $k{-}1$ candidates for the target, and we give the exact AUC this exclusion is worth (Section 4). On a small-vocabulary corpus at $k{=}5$ that value exceeds what our own probe achieved, so the cell carried no evidence about encoding whatsoever. The semantic analogue is larger: when sources are paired by real co-occurrence statistics, which is the setting one would choose to make a benchmark

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.

realistic, a predictor given only the masker's class label reaches an AUC higher than the audio probe it is meant to bound (Section 5). Matching the marginal class distribution across arms does not help, because the joint distribution is itself the shortcut.

With both shortcuts closed, a capacity effect survives: at a fixed target-to-rest level the weak source becomes harder to decode as the scene fills, and it does so faster than a conditional ceiling measured on the same mixtures declines (Section 6). We then ask what causes it, and measure rather than interpret. Corrupting an oracle ideal binary mask at a controlled rate sweeps the target-to-masker ratio the reader enjoys (Section 7): the number of sources is almost free at the favourable end of that axis and costs the entire effect at the unfavourable end. A short proposition shows the sweep is one axis rather than two, which retires a natural but vacuous control we had built, and a permutation control separates arrangement from level at matched energies. Finally, part of the loss is architectural, since pooling over the encoder's own frame tokens recovers a growing fraction of the weak source from the same frozen features and its advantage widens with budget (Section 8).

Contributions.

1. A closed-form label-only bound for mixture probing, verified by simulation, with a sampler that drives it to exactly 0.5 for every vocabulary size and every k (Section 4). The bound was live in our own design and we quantify its cost.
2. Evidence that natural source co-occurrence makes weak-source decodability an invalid measure of encoding, with the bound that detects it (Section 5).
3. A capacity measurement with both shortcuts closed and the ceiling controlled in matched units: $\beta = -0.098$ per $\log k$, robust to a level control, to per-run fixed effects and to two alternative specifications of the regression (Section 6).
4. A mechanism, measured rather than argued: the k penalty is a steep function of the masker suppression available to the reader, falling from 0.201 AUC at 0 dB to 0.015 at 5 dB and 0.001 at 19 dB (Section 7). We prove that no target-blind transformation can move along that axis, which both explains why a frozen encoder pays the full penalty and invalidates a class of controls.
5. Evidence that clip-level pooling is a recoverable part of the loss, with a budget sweep that distinguishes it from a small-sample artefact (Section 8).
6. A complete account of negative results: two pre-registered hypotheses, both falsified; a control and its derived statistic, both withdrawn under Proposition 2; a retracted decomposition; and a normalisation bug in our own code (Appendix F).

2 Related work

Probing and shortcut learning. Shortcut learning is well documented in vision and language (Geirhos et al., 2020; McCoy et al., 2019), and probing methodology has been criticised for conflating decodability with use (Hewitt & Liang, 2019; Belinkov, 2022). Control tasks (Hewitt & Liang, 2019) address probe capacity, whereas the shortcuts we identify sit upstream of the probe entirely, arising from how mixtures are constructed and from how sources co-occur in the world. Neither of the two, as far as we can establish, has been reported before.

Ground truth for dictionary learning. Tang et al. (2026) give a unified theory of sparse dictionary learning, characterise its non-identifiability and introduce a benchmark with full ground-truth access; related evaluations of sparse autoencoders share the constructed-ground-truth setting (Venhoff et al., 2024; Chanin, 2026; Cui et al., 2026). In that line the ground truth is synthetic and composition is additive in representation space by assumption. Ours is physical, which is what makes a label-only null model both definable and necessary.

Composition and separation. That embeddings of mixtures are not sums of embeddings has been observed qualitatively (Hawley & Tackett, 2024), and additivity has been induced in purpose-built autoencoders (Torres et al., 2026). The reference composition law, that log-mel spectra of a mixture are approximately a log-sum-exp of the sources', is decades old in robust speech recognition (Nadas et al., 1988; Varga & Moore, 1990; Gales & Young, 1996; Moreno, 1996; Acero et al., 2000); we claim no novelty for it, and Appendix E reports one methodological consequence, namely that an

embedding-space linearity test conducted in the wrong domain reports a nonlinearity the model does not have. Ideal binary masks and their oracle performance are likewise standard (Wang, 2005; Vincent et al., 2006); we use the ideal binary mask not as a separation target but as an instrument that supplies a controlled amount of target-specific information, and hence of masker suppression.

3 SETUP

Mixtures. Let $a_1, \ldots, a_k$ be isolated single-source recordings with class labels $y_1, \ldots, y_k$ from a vocabulary of size n. The last source is the *target*; the others are *maskers*. Each masker is set to a fixed integrated loudness $L = -40$ LUFS and the target is placed δ decibels below the *combined* level of the maskers, $L_{\text{rest}} = L + 10\log_{10}(k-1)$, so that δ has the same meaning at every k. The mixture is the waveform sum $m = \sum_{i<k} a_i + a_k$. The task throughout is to predict y_k, the class of the quiet source, from a representation of m, scored by macro one-vs-rest AUC on a clip-disjoint split in which a recording never appears in both halves.

Samplers. How the k classes are drawn is the subject of Section 4. DISTINCT draws k classes without replacement; IID draws them independently and uniformly with replacement. Every result reported below uses IID, except where we are explicitly quantifying the bias of DISTINCT.

Readers. A frozen encoder f maps a waveform to a clip vector $f(m)$; its tower also exposes frame tokens. We compare a logistic regression (LIN) and a 256-unit MLP (MLP) on $f(m)$; attention pooling (ATTN) and max-over-frames multiple instance learning (MAX) on the tokens; and a small convolutional network trained from scratch on $\mathrm{mel}(m)$ (CNN), which uses no encoder and exists so that claims about the encoder are not claims about probes.

Masks. The oracle ideal binary mask is $\mathcal{M}^\star = \mathbf{1}\big[|Z_{a_k}|^2 > |Z_{\text{rest}}|^2\big]$, where Z denotes a short-time Fourier transform. We corrupt it by flipping each bin independently with probability x, so $x = 0$ is perfect bin assignment and $x = \frac{1}{2}$ carries none. A masked reader sees $\mathrm{istft}(Z_m \odot \mathcal{M})$ and therefore only ever sees bins the mixture already contains. We write $\mathcal{E}_k(\mathcal{M})$ and $\mathcal{E}_{\neg k}(\mathcal{M})$ for the fractions of target and masker energy a mask retains, and define the *suppression* it achieves as $10\log_{10}\big(\mathcal{E}_k(\mathcal{M})/\mathcal{E}_{\neg k}(\mathcal{M})\big)$, its improvement in target-to-masker ratio over the unmasked mixture.

Encoders and corpora. Three publicly released pretrained encoders, used frozen throughout, spanning the training-objective axis: **CLAP** (contrastive), **AST** (supervised on AudioSet) and **AudioMAE** (self-supervised masked autoencoding). AudioMAE is run through AST's own feature extractor, so AST versus AudioMAE varies the objective alone. Corpora: **ESC-50** (2000 clips, 50 classes), **UrbanSound8K** (1930, 10) and **FSD50K** restricted to single-*source* clips (7890 clips, 42 classes with the ≥ 20 clips probing requires). Full protocol in Appendix A.

4 A LABEL-ONLY BOUND FOR MIXTURE PROBING

Consider a predictor that observes only the *set* of masker classes $\mathcal{S} = \{y_1, \ldots, y_{k-1}\}$, with no audio, and must score each candidate class for being the target.

Proposition 1 (Exclusion bound)**.** *Under* DISTINCT *sampling from a vocabulary of size n, the macro one-vs-rest AUC of the optimal label-only predictor is*

$$\mathrm{AUC}_{\text{label}} = \tfrac{1}{2} + \frac{k-1}{2(n-1)}. \tag{1}$$

Proof. Fix a class c. Because the k classes are distinct, a positive example ($y_k = c$) never has $c \in \mathcal{S}$, so the score $\mathbf{1}[c \notin \mathcal{S}]$ equals 1 on every positive. A negative example has $y_k \neq c$ and its $k-1$ maskers are drawn without replacement from the remaining $n-1$ classes, so $\Pr[c \in \mathcal{S} \mid y_k \neq c] = (k-1)/(n-1) =: p$. The score is binary, so ties count half, giving $\mathrm{AUC} = p + \frac{1}{2}(1-p)$, which is (1). Averaging over c leaves it unchanged. □

Table 1: Label-only AUC available from the masker set alone (Proposition 1). The DISTINCT bound rises with k, which is the axis along which capacity effects are measured. The last column is the same quantity under IID sampling.

		DISTINCT			IID
corpus	n	k=2	k=3	k=5	any k
ESC-50	50	0.510	0.520	0.541	0.500
FSD50K	42	0.512	0.524	0.549	0.500
UrbanSound8K	10	0.556	0.611	**0.722**	0.500

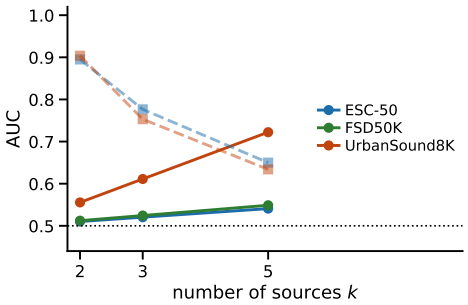

Figure 1: Solid: the DISTINCT bound of Eq. (1). Dashed: the measured probe, which approaches it and on UrbanSound8K falls below it. Under IID sampling the bound is the dotted line at 0.5.

Equation (1) is small when the vocabulary is large and k is modest, and is not small otherwise. Table 1 evaluates it; a Monte-Carlo simulation of the sampler actually used in our experiments reproduces it to three decimals in every cell (Appendix B).

The bound was live in our own design. On UrbanSound8K (n=10) at k=5 the bound is 0.722, while the probe on $f(m)$ we had measured reached 0.634. A cell whose measured AUC falls below what the class labels alone provide carries no evidence about encoding. Worse, the bound grows with k on every corpus, so the bias runs along exactly the axis a capacity study varies, and it inflates the high-k end where the effect is largest.

The repair, and what it cost. Under IID sampling $\Pr[y_k = c \mid \mathcal{S}] = 1/n$ regardless of $\mathcal{S}$, so the bound is exactly $\frac{1}{2}$ for all n and k. The price is that a mixture may contain two clips of the same class, which is realistic and unbiased. Enlarging the vocabulary is not an alternative: it shrinks (1) but does not remove its dependence on k, so the bias still runs along the axis of interest. Re-running the affected experiments moved the measured probe only slightly (ESC-50 at k=5: 0.650 → 0.678; UrbanSound8K: 0.634 → 0.641), so the capacity effect reported below was not an artefact of the leak. The possibility is nevertheless now removed rather than argued away.

5 NATURAL CO-OCCURRENCE IS THE SAME SHORTCUT, AND IT IS LARGER

Independent sources are an artificial choice; real scenes co-occur. We built mixtures from FSD50K single-source clips paired according to co-occurrence counts mined from FSD50K's own multi-label clips (88 usable class pairs, for instance *Knock+Wood* and *Motor vehicle+Engine*). The control arm draws the two classes independently from the marginal distribution that the co-occurrence arm induces, so the marginals match by construction and only the joint differs.

The co-occurrence arm looks dramatically easier (Table 2) and, more tellingly, is flat in δ: a target 20 dB down is read as easily as one 5 dB down, which no acoustic account predicts. Applying the logic of Proposition 1 empirically, that is, estimating $\Pr[y_k \mid \mathcal{S}]$ from the pairing distribution and scoring with it, gives AUC 0.968 in the co-occurrence arm against 0.620 in the control. The label-only bound therefore exceeds the audio probe it is supposed to bound (0.968 against 0.913). The probe is not reading the weak source; it is reading the loud one and inferring. The acoustics confirm this: the oracle ceiling and the mask coverage are essentially identical across arms, so correlation changes neither what the mixture contains nor what the encoder does with it. What it changes is what the metric measures, and by an amount that no acoustic quantity in the table anticipates.

Table 2: FSD50K, marginal-matched arms, CLAP. The co-occurrence arm is far more decodable, yet its oracle ceiling and its ideal-binary-mask coverage are indistinguishable from the control, so the mixture the encoder receives is not acoustically different between arms.

	probe AUC on $f(m)$			oracle ceiling		IBM coverage	
δ (dB)	co-occur	independent	Δ	co-occur	indep.	co-occur	indep.
5	0.906	0.807	+0.099	0.972	0.967	0.951	0.956
10	0.913	0.778	+0.135	0.962	0.972	0.913	0.925
20	0.912	0.710	+0.202	0.955	0.941	0.824	0.834

Table 3: Weak-source AUC on $f(m)$ and the conditional ceiling (grey), at δ=10 dB, two seeds, IID sampling. The ceiling is flat in k on every run while the probe falls. Absolute values depend on probe budget (Table 8); only within-row comparisons are meaningful.

		probe on $f(m)$			conditional ceiling		
encoder	corpus	k=2	k=3	k=5	k=2	k=3	k=5
CLAP	ESC-50	0.782	0.694	0.628	0.995	0.994	0.992
CLAP	UrbanSound8K	0.719	0.627	0.601	0.985	0.982	0.984
CLAP	FSD50K	0.735	0.635	0.610	0.959	0.945	0.946
AST	ESC-50	0.848	0.753	0.656	0.995	0.995	0.991
AST	UrbanSound8K	0.768	0.638	0.609	0.988	0.987	0.987
AudioMAE	ESC-50	0.692	0.624	0.566	0.963	0.965	0.960
AudioMAE	UrbanSound8K	0.620	0.556	0.573	0.919	0.923	0.911

Remark 1 (Recommendation)**.** *Any probing study on real, co-occurring audio should report the masker-label-only AUC beside its headline number. Where the two are close, the headline is uninformative. Our controlled results are immune by construction, because classes are drawn independently, but that is a property of this design and not of probing in general. The measurement the experiment was built for, namely whether correlation changes what the encoder retains, requires stratifying by masker class so that the masker carries no information; with 88 pairs most strata contain one or two target classes, so we leave it open rather than report a number we cannot defend.*

6 THE CAPACITY EFFECT, WITH BOTH SHORTCUTS CLOSED

We now fix δ and vary k under IID sampling. Comparing a probe AUC against an energy-coverage fraction would mix two scales, since AUC has a chance floor of 0.5 and coverage a floor of 0. We therefore compare the probe against a *conditional ceiling* in the same units: the same probe, on the same mixtures, applied to the oracle-masked mixture. The mask selects only bins the mixture already contains, so the ceiling answers whether the target's identity is still present, and any shortfall below it is information that is present and not encoded.

A slope with the ceiling controlled. Over 63 cells (encoder $\times$ corpus $\times k \times \delta$, with $\delta \in \{10, 20, 30\}$) we fit the specification

$$\mathrm{AUC}_{f(m)} = \alpha_0 + \alpha_1 \mathrm{AUC}_{\mathrm{ceil}} + \beta \log k + (\text{controls}), \tag{2}$$

obtaining $\beta = -0.098$ with bootstrap 95% CI $[-0.133, -0.062]$ and $R^2 = 0.524$. Adding a δ control gives -0.103 and adding per-run fixed effects gives -0.104 ($R^2 = 0.812$), so the slope holds within a single encoder-corpus pair. The equivalent specification on the gap gives $+0.099$ $[+0.064, +0.136]$ and a coverage-based specification over 119 cells gives -0.077 $[-0.103, -0.051]$; Appendix C states all three. The conditional ceiling is nearly flat across cells (mean 0.949, s.d. 0.036), so the effect is not the information leaving the mixture. All three encoders give a negative slope with intervals excluding zero: AST -0.126, CLAP -0.092, AudioMAE -0.080.

Across corpora the slope is -0.121 on ESC-50, -0.071 on FSD50K and -0.070 on UrbanSound8K. ESC-50 stands apart from the other two and we were unable to explain this; Section 9 reports the two hypotheses we pre-registered and falsified rather than a third that we did not test.

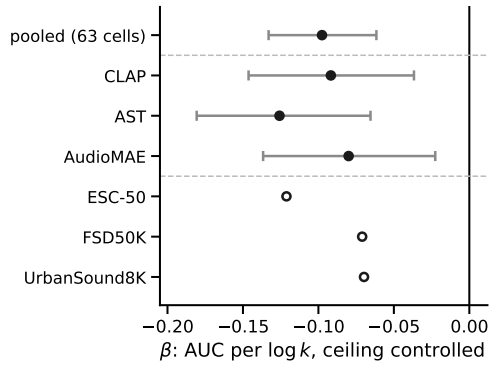

Figure 2: The capacity slope of Eq. (2), pooled and per encoder, with bootstrap intervals over cells. Per-corpus values (hollow) are shown without intervals because the three corpora contribute unequal cell counts.

7 Mechanism: the capacity law is a suppression law

The obvious reading of Section 6 is that scenes with more sources are intrinsically harder. This section shows that reading is false and replaces it with a measurement.

7.1 Suppression and target-specific information are one quantity

It is tempting to treat how much a mask says about bin ownership and how much it improves the target-to-masker ratio as two variables that a suitable control could vary separately. They cannot be separated, for an elementary reason.

Proposition 2 (Target-blind transformations cannot suppress)**.** *Let $\mathcal{M}$ be a binary time-frequency mask whose selection is independent of the target's spectrogram and exchangeable across bins, and write $\rho = \mathbb{E}[|\mathcal{M}|]/(FT)$. Then $\mathbb{E}[\mathcal{E}_k(\mathcal{M})] = \mathbb{E}[\mathcal{E}_{\neg k}(\mathcal{M})] = \rho$, so the expected retained energies are equal and the mask achieves zero suppression in expectation.*

Proof. $\mathcal{E}_k(\mathcal{M}) = \sum_{tf} \mathcal{M}_{tf} |Z_{a_k}|^2_{tf} / \sum_{tf} |Z_{a_k}|^2_{tf}$. By independence $\mathbb{E}[\mathcal{M}_{tf} |Z_{a_k}|^2_{tf}] = \mathbb{E}[\mathcal{M}_{tf}]\, |Z_{a_k}|^2_{tf}$, and by exchangeability $\mathbb{E}[\mathcal{M}_{tf}] = \rho$ for every bin, so $\mathbb{E}[\mathcal{E}_k(\mathcal{M})] = \rho$. The identical argument gives $\mathbb{E}[\mathcal{E}_{\neg k}(\mathcal{M})] = \rho$. (For the ratio itself, concentration of both sums over $FT \approx 10^5$ bins makes the difference between the ratio of expectations and the expectation of the ratio negligible; measured suppression for all target-blind arms is 0.0 ± 0.1 dB, Table 4.) □

Remark 2 (A control we withdraw)**.** *Proposition 2 invalidates a control we had built.* RAND-$\mathcal{E}$ *was a random mask retaining bins until its* target *energy matched that of $\mathcal{M}^\star$, intended to hold energy fixed while removing target-specific information. Matching the target energy of a target-blind mask forces the masker energy to match too, and Table 4 confirms it:* RAND-$\mathcal{E}$ *retains* 0.79 *of the target and* 0.79 *of the masker, so it is the unmasked mixture with a random subset of bins deleted. The observation that $f(m)$ matches* RAND-$\mathcal{E}$ *to within* 0.006 *AUC is true but vacuous, comparing the mixture with a thinned copy of itself. We withdraw it, along with the "effective flip rate" statistic built on the same logic, which for the same reason could only ever report that a reader of the raw mixture obtains no suppression, a fact true by construction.*

7.2 The k penalty is bought off by a few decibels of suppression

With this settled, the corruption sweep is a single well-defined axis, and it is informative. Table 4 reports every arm at k=5 with both energies; Table 5 reports the sweep across k.

Three readings follow, and one common reading does not.

Adding sources is free to a reader that can separate. At x=0 the AUC runs from 0.998 to 0.997 across $k \in \{2, 3, 4, 5\}$, a penalty of 0.001 at 18.9 dB of suppression; at x=0.5 the same range spans 0.866 to 0.665, a penalty of 0.201 at 0 dB. Naming one class among k is therefore not intrinsically

Table 4: Every mask arm at k=5, CLAP/ESC-50, δ=10 dB. Target-blind arms all sit at 0 dB, as Proposition 2 requires them to.

mask	AUC	$\mathcal{E}_k$	$\mathcal{E}_{\neg k}$	supp.
$\mathcal{M}^\star$ (x=0)	0.997	0.79	0.013	17.8
x=0.35	0.834	0.59	0.354	2.2
x=0.5	0.665	0.50	0.500	0.0
RAND-$\mathcal{E}$	0.694	0.79	0.791	0.0
RAND-n	0.662	0.15	0.150	0.0
SHUFFLE	0.762	0.79	0.013	17.8
none, $f(m)$	0.683	1.00	1.000	0.0

Table 5: Weak-source AUC by flip rate, CLAP/ESC-50, δ=10 dB, 6000 mixtures per cell. The last column is the k penalty; "supp." is the suppression the mask achieves, averaged over k.

x	supp.	k=2	3	4	5	pen.
0	18.9	0.998	0.998	0.997	0.997	0.001
0.05	11.2	0.997	0.995	0.995	0.994	0.003
0.1	8.5	0.994	0.993	0.991	0.990	0.005
0.2	5.3	0.982	0.973	0.967	0.969	0.015
0.35	2.3	0.930	0.871	0.846	0.834	0.096
0.5	0.0	0.866	0.738	0.692	0.665	0.201
SHUF	18.9	0.759	0.755	0.767	0.762	0.012
$f(m)$	0.0	0.895	0.764	0.712	0.683	0.212

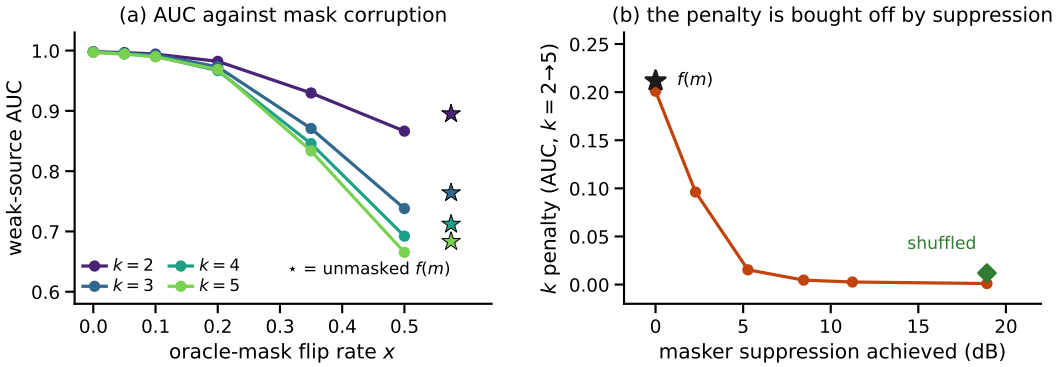

Figure 3: (a) Weak-source AUC against oracle-mask flip rate, one curve per k. Stars at the right edge give the unmasked encoder, which by Proposition 2 sits at the zero-suppression end of the axis by construction rather than by measurement. (b) The k penalty as a function of the suppression the reader achieves. The shuffled arm (diamond) has the oracle mask's suppression and its near-zero penalty, yet scores 0.235 lower in AUC.

harder as k grows: there is no residual difficulty once the target is separated. The penalty is also steeply nonlinear in suppression, since 5.3 dB removes 93% of it and 8.5 dB removes 98%.

This is a statement about the task, not a measurement of the encoder. Proposition 2 places any reader of the unmasked mixture at 0 dB by construction, so the encoder's position on the axis is not evidence. What the axis establishes is the shape of the penalty, and hence what would have to change for the capacity law to disappear: a frozen encoder pays 0.212 AUC over k=2 $\rightarrow$ 5 because it performs no target-conditional suppression, and a few decibels would suffice. The capacity law is the price of not separating, not a degradation of representation quality.

Level and arrangement are separable, and only level carries the k dependence. SHUFFLE permutes the time-frequency positions of the $\mathcal{M}^\star$-masked spectrogram, preserving its multiset of values and therefore both energies and its suppression; only the arrangement differs. It is the one control here not blocked by Proposition 2, being derived from the oracle mask rather than chosen independently of the target. At matched 18.9 dB the coherent arrangement reads 0.997 and the shuffled one 0.762, so roughly a quarter of what the oracle mask buys requires structure that survives no permutation. Yet SHUFFLE has a k penalty of 0.012, close to the oracle's 0.001 and far from the encoder's 0.212: arrangement is worth a large constant and contributes nothing at all to the way the penalty depends on the number of sources.

Remark 3 (Why the oracle gap must not be used to size the loss)**.** *An earlier version of this work measured the distance from $f(m)$ to a probe on the oracle-masked mixture and found it large (+0.37 AUC at k=5). That gap is a separation gap: an ideal binary mask hands the probe a cleanly separated target, and a frozen encoder is not a source separator, so losing to an oracle separator is*

Table 6: Readers of identical mixtures, δ=10 dB, IID sampling, 8000 mixtures per cell for CLAP and 6000 otherwise. FRAME is the better of attention and maximum pooling over the encoder's own tokens; CNN uses no encoder; CTRL is the same CNN on the oracle-masked mixture, a capacity control at the same budget. Every reader is listed in Table 10.

		clip vector LIN			FRAME − clip			CTRL		
encoder	corpus	k=2	3	5	k=2	3	5	k=2	3	5
CLAP	ESC-50	0.899	0.769	0.678	+0.040	+0.074	+0.077	0.978	0.979	0.978
CLAP	UrbanSound8K	0.879	0.740	0.641	+0.013	+0.060	+0.061	0.996	0.996	0.994
AST	ESC-50	0.913	0.794	0.685	−0.018	+0.025	+0.039	0.968	0.971	0.967
AudioMAE	ESC-50	0.769	0.659	0.603	+0.085	+0.097	+0.046	0.968	0.971	0.967
AudioMAE	UrbanSound8K	0.739	0.637	0.555	+0.049	+0.048	+0.022	0.994	0.995	0.996

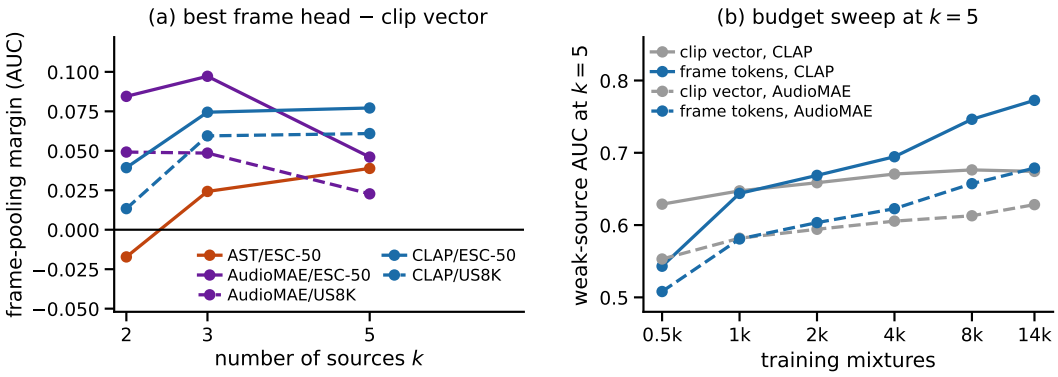

Figure 4: (a) Advantage of the best frame-level head over the clip vector at δ=10 dB, by encoder and corpus. (b) Budget sweep at k=5 on ESC-50 (CLAP solid, AudioMAE dashed): the frame-token head starts below the clip vector, overtakes it, and keeps climbing, while CLAP's clip vector flattens after 8000 mixtures.

not a finding. The conditional ceiling establishes that the identity is present in the mixture, which is what Section 6 uses it for. It must not be read as the encoder's share of the shortfall.

8 PART OF THE LOSS IS POOLING, AND IT IS RECOVERABLE

If the encoder performs no suppression, is any of the remaining loss architectural rather than informational? We compare readers that share the same frozen features and differ only in how they pool across time frames.

The capacity control passes. The from-scratch CNN reads the oracle-masked mixture at 0.967 or better in every cell at the same budget as every other reader, so the architecture is adequate for mixture-shaped data and is not starved of data. This removes the sample-efficiency objection to the CNN column by measurement rather than by argument.

Frame pooling recovers a growing fraction. On CLAP/ESC-50 the best frame-level head beats the clip vector by +0.040, +0.074 and +0.077 at $k = 2, 3, 5$: the same frozen features with the one-vector-per-clip constraint dropped, so no sample-efficiency confound applies. A weak source tends to dominate a few frames, and averaging the clip into one vector is part of where it goes. The effect reproduces on all three encoders (Figure 4a), though which head wins is encoder-dependent. On UrbanSound8K the from-scratch CNN outperforms the frozen AudioMAE embedding outright once the scene fills, so $f(m)$ is not the practical frontier there either.

Not a small-sample artefact. Probe AUC moves with budget by more than the margins of interest, so we sweep the budget with the test set held fixed (Figure 4b, Table 8). At k=5 the margin is *negative* at the smallest budgets on both encoders, crosses zero by 2000 mixtures and grows monotonically thereafter, reaching +0.098 on CLAP and +0.051 on AudioMAE at 14000. On CLAP the clip vector

additionally saturates, gaining nothing between 8000 and 14000 mixtures ($0.676 \rightarrow 0.674$) while the frame-token head climbs ($0.746 \rightarrow 0.773$): the clip vector is at a ceiling of its own and the tokens are not. A margin produced by data starvation would shrink as data is added, so this advantage is a property of the representation, not of the estimator. At k=2 the picture is mixed, the CLAP margin flattening after 2000 mixtures while AudioMAE's keeps growing, which is why we read the argument in the crowded regime.

Where pooling does not help. Sweeping δ on CLAP/ESC-50, the margin for $k = 2, 3, 5$ is $+0.042/+0.070/+0.056$ at δ=5 and $+0.040/+0.074/+0.077$ at δ=10, but collapses to $+0.065/+0.025/-0.024$ at δ=20. Once the target is buried in the signal itself no frame is dominated by it and frame pooling has nothing to recover, so the effect is specific to the regime in which the signal still carries the source at all.

9 What we could not explain, and what we withdrew

We report the failures because they bound what the positive results can be taken to mean; Appendix F gives the full list with numbers.

β differs by corpus, and we registered two explanations before testing them. Under *spectral sparsity* FSD50K, the sparsest by Gini coefficient and active-bin fraction, should show the largest $|\beta|$; its $|\beta|$ instead lies between the other two. Under *class count*, after the sampler was fixed FSD50K with 42 probing classes and UrbanSound8K with 10 give -0.071 and -0.070. Noise-floor depth and temporal flatness order the corpora monotonically but β is a step rather than a gradient, so a graded relation is unsupported and cannot be separated from coincidence with three corpora. ESC-50 remains an unexplained outlier and we do not patch it with a third story. Separately, Remark 2 withdraws RAND-$\mathcal{E}$ and the statistic derived from it, and an earlier $75/25$ split of the mask effect into assignment and denoising shares using RAND-n is retracted because RAND-n differs from $\mathcal{M}^\star$ in two variables at once, so the split was never identified; we retain RAND-n in Table 4 only to exhibit the size of that confound.

10 Limitations

Sources are combined by waveform addition without reverberation, which is what buys the ground truth and is also a departure from real scenes. Our controlled results draw classes independently, which makes them immune to the shortcut of Section 5 and also means they say nothing about correlated scenes, where we have shown the standard metric fails. Proposition 2 means the mask sweep supports statements about the target-to-masker ratio but cannot attribute the effect to knowing which bins as distinct from having a better ratio; SHUFFLE is the only control that escapes this, and it varies arrangement rather than information. The oracle mask is an instrument for supplying suppression, not a claim that an unprivileged reader could construct it, and how close such a reader could come is open. Absolute AUCs move with probe budget, so only within-table comparisons are meaningful, and three encoders and three corpora suffice to show the phenomena are not specific to one model but not to characterise their dependence on architecture or data.

11 Conclusion

Mixture probing admits two shortcuts that let a probe succeed without reading the source it is scored on, one from the sampler and one from natural co-occurrence, and both were live in a careful design, ours. With both closed, weak-source decodability falls with scene density at a matched ceiling, and corrupting an oracle mask shows this law to be a suppression law: a few decibels of target-conditional suppression make the number of sources almost free, and none makes it cost the whole effect. Part of what remains is architectural, since frame pooling recovers a growing fraction from the same frozen features. Probing studies on mixtures should report the label-only bound beside the headline number.

Reproducibility statement

Every number in the paper is produced by a released script and read from a JSON artefact. Appendix A gives the full protocol (corpora, level control, encoder checkpoints, probe hyperparameters, mask construction and resynthesis); Appendix B gives the simulation check for Proposition 1; Appendix C states all three regression specifications and the bootstrap procedure; Appendix G lists the pipeline in execution order and the environment. The sampler is selected by an environment variable recorded in each artefact, and the artefacts produced before the sampler was fixed are retained so that the effect of that change can be audited rather than taken on trust.

Ethics statement

This work uses three public audio corpora of environmental and urban sounds and three publicly released model checkpoints. It involves no human subjects and no personally identifying information. The diagnostic tools we release are intended to make evaluation of audio representations more conservative and we are not aware of a foreseeable harmful application.

AI use statement

We used a large language model to check and polish the writing of this paper, and to assist with code, specifically the plotting and figure-generation scripts. All experiments were run on our own hardware, and every number reported here is computed from stored test-set logits of models we trained and was verified by the authors against the underlying data files.

References

Alex Acero, Li Deng, Trausti Kristjansson, and Jerry Zhang. Hmm adaptation using vector taylor series for noisy speech recognition. In *Proc. Int. Conf. on Spoken Language Processing*, October 2000. URL https://www.microsoft.com/en-us/research/publication/hmm-adaptation-using-vector-taylor-series-for-noisy-speech-recognition/.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, March 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, October 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/. Published October 4, 2023.

David Chanin. Are sparse autoencoder benchmarks reliable?, 2026. URL https://arxiv.org/abs/2605.18229.

Jingyi Cui, Qi Zhang, Yifei Wang, and Yisen Wang. On the limits of sparse autoencoders: A theoretical framework and reweighted remedy, 2026. URL https://arxiv.org/abs/2506.15963.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition, 2022. URL https://arxiv.org/abs/2209.10652.

M.J.F. Gales and S.J. Young. Robust continuous speech recognition using parallel model combination. *IEEE Transactions on Speech and Audio Processing*, 4(5):352–359, 1996. doi: 10.1109/89.536929.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine*

Intelligence, 2(11):665–673, November 2020. doi: 10.1038/s42256-020-00257-z. URL https://doi.org/10.1038/s42256-020-00257-z.

Scott H. Hawley and Austin R. Tackett. Operational latent spaces, 2024. URL https://arxiv.org/abs/2406.02699.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks, 2019. URL https://arxiv.org/abs/1909.03368.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference, 2019. URL https://arxiv.org/abs/1902.01007.

Pedro Moreno. *Speech Recognition in Noisy Environments*. PhD thesis, 08 1996.

A. Nadas, D. Nahamoo, and M.A. Picheny. Speech recognition using noise-adaptive prototypes. In *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, pp. 517–520 vol.1, 1988. doi: 10.1109/ICASSP.1988.196633.

Yiming Tang, Harshvardhan Saini, Zhaoqian Yao, Zheng Lin, Yizhen Liao, Jingyi Cui, Yisen Wang, Mengnan Du, and Dianbo Liu. A unified theory of sparse dictionary learning in mechanistic interpretability: Piecewise biconvexity and spurious minima, 2026. URL https://arxiv.org/abs/2512.05534.

Bernardo Torres, Manuel Moussallam, and Gabriel Meseguer-Brocal. Learning linearity in audio consistency autoencoders via implicit regularization, 2026. URL https://arxiv.org/abs/2510.23530.

A.P. Varga and R.K. Moore. Hidden markov model decomposition of speech and noise. In *International Conference on Acoustics, Speech, and Signal Processing*, pp. 845–848 vol.2, 1990. doi: 10.1109/ICASSP.1990.115970.

Constantin Venhoff, Ani Calinescu, Philip Torr, and Christian Schroeder de Witt. SAGE: Scalable ground truth evaluations for large sparse autoencoders, 2024. URL https://openreview.net/forum?id=sknUS8X9q0.

E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469, 2006. doi: 10.1109/TSA.2005.858005.

DeLiang Wang. *On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis*, pp. 181–197. Springer US, Boston, MA, 2005. ISBN 978-0-387-22794-8. doi: 10.1007/0-387-22794-6_12. URL https://doi.org/10.1007/0-387-22794-6_12.

SUPPLEMENTARY MATERIAL

A EXPERIMENTAL PROTOCOL

A.1 CORPORA AND SINGLE-SOURCE SELECTION

ESC-50: 2000 clips, 50 classes, 5 s, 44.1 kHz, from the official archive. **UrbanSound8K**: 1930 usable clips over 10 classes after discarding clips shorter than 2.5 s or quieter than -50 LUFS. **FSD50K**: labels carry the full AudioSet ontology chain, so a clip labelled *Electric_guitar, Guitar, Plucked_string_instrument, Musical_instrument, Music* contains one sound and not five. We therefore define a clip to be single-*source* when its label set contains exactly one *leaf*, that is, one mid with no child also present. This yields 7890 clips over 79 classes from ten parquet shards, of which 42 classes survive the ≥ 20-clip requirement for probing, against roughly 32 clips per shard by a naive single-label test. All probing statistics use the 42-class vocabulary, which is also the n used in Table 1.

A.2 LEVEL CONTROL

Every source is peak-normalised and then set to a target integrated loudness with `pyloudnorm`. Maskers sit at $L = -40$ LUFS and the target at $L_{\text{rest}} - \delta$ with $L_{\text{rest}} = L + 10\log_{10}(k-1)$, so δ is the target-to-combined-rest ratio and is comparable across k. Mixtures are peak-limited to 0.99 only when they would otherwise clip; the clip rate is below 0.1%. Mixtures are stored as recipes (source identifiers, gains, seed) and regenerated on demand rather than written to disk.

A.3 ENCODERS

CLAP `laion/clap-htsat-unfused`: 48 kHz, 64-mel decibel log-mel; we take the audio tower's `pooler_output` (d=768, pre-projection, without ℓ_2 normalisation) and its 2×32 patch grid as frame tokens. Pre-projection features are used throughout because the projection head is ℓ_2-normalised, which destroys the scale information any composition analysis needs (Appendix E). **AST** `MIT/ast-finetuned-audioset-10-10-0.4593`: 16 kHz, 128-mel Kaldi filterbank, normalised as $(x-\mu)/(2\sigma)$; `pooler_output` and 1212 patch tokens pooled to 128 groups over frequency. **AudioMAE** `gaunernst/vit_base_patch16_1024_128.audiomae_as2m`: a ViT-B/16 over a 1024×128 filterbank, self-supervised on AudioSet-2M without labels, run through AST's own feature extractor so that AST versus AudioMAE varies the training objective alone. The checkpoint is loaded into a self-contained ViT implementation with `strict=True` because `timm` hard-imports `torchvision`, for which no wheel compatible with our torch build exists on Python 3.13. Sanity check: clean-clip ESC-50 class AUC is 0.985 for AudioMAE and 1.000 for AST.

A.4 PROBES

All probes use a clip-disjoint split in which 30% of *recordings* are reserved and a recording never appears in both halves. Scores are macro one-vs-rest AUC over classes with at least three test items. LIN is logistic regression (`max_iter`=400) on standardised features; MLP is a 256-unit MLP with early stopping; ATTN is a learned attention pooling over frozen tokens ($d \to 128 \to 1$ attention logits, softmax, weighted sum, linear classifier); MAX is max-over-frames multiple instance learning on the same tokens; CNN is four stride-2 convolutional blocks ($32, 64, 128, 128$ channels) with batch normalisation, global average pooling, dropout 0.3 and a linear head, trained with AdamW (10^{-3}, cosine schedule, label smoothing 0.05, early stopping with patience 12 on a validation split carved out of the training half rather than out of the test half).

A.5 MASKS AND RESYNTHESIS

STFT with `nperseg`$=1024$ and `noverlap`$=768$. The oracle mask is $\mathcal{M}^\star = \mathbf{1}\left[|Z_{a_k}|^2 > |Z_{\text{rest}}|^2\right]$. Masked audio is $\text{istft}(Z_m \odot \mathcal{M})$, so no reader ever sees information absent from the mixture. Ideal-binary-mask coverage is the fraction of target energy lying in bins where the target dominates, restricted to bins within 60 dB of the mixture peak; the unrestricted value agrees to within 0.01 everywhere, so near-silence does not inflate it. SHUFFLE applies a single random

permutation to the time-frequency positions of the $\mathcal{M}^\star$-masked spectrogram, which preserves its multiset of complex values and therefore both energies exactly.

A.6 COMPUTE

All experiments run on a single RTX 4090 with 20 CPU cores in an isolated virtual environment (torch 2.5.1+cu121). The full suite is roughly 60 GPU-hours, dominated by the mask-corruption sweep, which encodes nine masked variants of every mixture.

B THE LABEL-ONLY BOUND

Simulation check. Proposition 1 is verified by simulating the exact sampler used in the experiments (4×10^4 draws per cell) and scoring the optimal label-only predictor. The simulated and closed-form values agree to three decimals in all nine cells of Table 1, and the IID sampler simulates to 0.500 ± 0.003 everywhere.

Why the i.i.d. sampler and not a larger vocabulary. Equation (1) shrinks as n grows, so one could reduce the leak by enlarging the label set. That hides it rather than removing it: the bound still rises with k, so the bias still runs along the axis of interest, and a study that varies k would still see an inflated high-k end. The IID sampler removes the leak exactly, at every n and k, at the sole cost of permitting repeated classes within one mixture.

Empirical co-occurrence bound. For Section 5 we estimate $\Pr[y_k \mid \mathcal{S}]$ directly from the pairing distribution with add-ϵ smoothing, score each candidate class with it, and compute the same macro one-vs-rest AUC used everywhere else. This is an achievable value for a label-only predictor and hence a lower bound on what the best such predictor obtains, which is the direction that matters: the true shortcut is at least this large.

C REGRESSION SPECIFICATIONS

We report β under three specifications so that no single functional-form choice carries the claim:

$$\text{(S1)} \quad \mathrm{AUC}_{f(m)} = \alpha_0 + \alpha_1 \mathrm{AUC}_{\text{ceil}} + \beta \log k \tag{3}$$

$$\text{(S2)} \quad \mathrm{AUC}_{\text{ceil}} - \mathrm{AUC}_{f(m)} = \alpha_0 + \beta \log k \tag{4}$$

$$\text{(S3)} \quad \mathrm{AUC}_{f(m)} = g(\text{IBM coverage}) + \beta \log k, \quad g \text{ quadratic} \tag{5}$$

S1 is primary: both terms are in AUC units, so no link function is required, and the ceiling is measured per cell under that cell's own conditions. S2 states the same claim on the gap. S3 is a coverage-based robustness check retained from an earlier draft; it requires choosing g, and a misspecified g would leak into β, which is why it is not primary. Controls are added in turn: δ as a linear term, then per-encoder-corpus fixed effects. Confidence intervals are bootstrap over cells with 4000 resamples. Results are in Table 7. Note that β in S2 is unchanged to twelve decimals by the δ control and by fixed effects, because the design is balanced: every run contributes the same k and δ values, so $\log k$ is orthogonal to both sets of dummies and only R^2 moves.

Table 7: All regression specifications, IID sampling, 63 cells (119 for S3).

spec.	controls	β	95% CI	R^2
S1	none	-0.098	$[-0.133, -0.062]$	0.524
S1	$+\delta$	-0.103		0.734
S1	$+\delta$, run fixed effects	-0.104		0.812
S2	none	$+0.099$	$[+0.064, +0.136]$	0.299
S2	$+\delta$	$+0.099$		0.577
S2	$+\delta$, run fixed effects	$+0.099$		0.711
S3	none	-0.077	$[-0.103, -0.051]$	0.700

D ADDITIONAL RESULTS

D.1 GENERALITY OF THE MASK SWEEP

The corruption sweep was run at $x \in \{0, 0.5\}$ and k=5 on AST and AudioMAE as well (2500 mixtures per cell). Both reproduce the qualitative pattern of Table 5: AST reads 0.998 at x=0 and 0.642 at x=0.5 against 0.673 unmasked, and AudioMAE reads 0.984 and 0.586 against 0.579 unmasked. On both encoders SHUFFLE sits between the two (0.738 and 0.698) at the oracle's suppression, and both target-blind arms sit at 0.0 dB as Proposition 2 requires.

Table 8: Budget sweep on ESC-50, δ=10 dB, IID sampling, test set held fixed; only the training half grows. The frame-pooling margin is negative at the smallest budgets on every run and crosses zero by 2000 mixtures. At k=5 it then grows monotonically on both encoders, and CLAP's clip-vector probe saturates ($0.676 \rightarrow 0.674$ from 8000 to 14000) while its frame-token head does not ($0.746 \rightarrow 0.773$). At k=2 the margin instead flattens on CLAP, which is why we read the budget argument at k=5.

		CLAP			AudioMAE		
	mixtures	LIN	ATTN	margin	LIN	ATTN	margin
k=5	500	0.629	0.543	−0.086	0.553	0.508	−0.045
	1000	0.647	0.644	−0.004	0.582	0.581	−0.001
	2000	0.659	0.669	+0.010	0.594	0.603	+0.009
	4000	0.671	0.695	+0.024	0.605	0.623	+0.017
	8000	0.676	0.746	+0.070	0.613	0.657	+0.044
	14000	**0.674**	**0.773**	**+0.098**	0.628	0.679	**+0.051**
k=2	500	0.782	0.746	−0.035	0.689	0.667	−0.022
	1000	0.804	0.827	+0.023	0.708	0.700	−0.008
	2000	0.835	0.872	+0.038	0.728	0.745	+0.017
	4000	0.874	0.909	+0.036	0.759	0.785	+0.026
	8000	0.900	0.933	+0.033	0.777	0.840	+0.063
	14000	0.916	0.945	+0.029	0.792	0.896	+0.105

D.2 THE CAPACITY GRID AT FINER RESOLUTION

Table 9 sweeps k and δ on one run (CLAP/ESC-50) at three seeds under IID sampling, alongside the ideal-binary-mask coverage of the target, which is how much of it the *signal* still carries. The clean-target ceiling for this run is 1.000 and every cell now lies above the label-only bound of 0.500; under the DISTINCT sampler five of the twenty cells did not. Reading the δ=10 column, coverage falls from 0.873 to 0.784 while the probe falls from 0.787 to 0.604, which is the effect Eq. (2) quantifies against a ceiling rather than against coverage. At $\delta \geq 20$ dB the target is buried in the signal as well as in the embedding, so those columns do not isolate the encoder.

Table 9: Weak-source AUC (mean ± s.d. over three seeds) and, in the lower block, IBM coverage of the target in the signal. One run, CLAP on ESC-50, IID sampling.

k	δ=5	δ=10	δ=15	δ=20	δ=30
weak-source AUC					
2	$0.861_{\pm.012}$	$0.787_{\pm.020}$	$0.722_{\pm.017}$	$0.653_{\pm.032}$	$0.560_{\pm.013}$
3	$0.795_{\pm.020}$	$0.685_{\pm.008}$	$0.608_{\pm.012}$	$0.580_{\pm.032}$	$0.514_{\pm.015}$
4	$0.765_{\pm.007}$	$0.641_{\pm.031}$	$0.587_{\pm.003}$	$0.532_{\pm.007}$	$0.526_{\pm.012}$
5	$0.768_{\pm.024}$	$0.604_{\pm.021}$	$0.546_{\pm.007}$	$0.509_{\pm.019}$	$0.521_{\pm.028}$
IBM coverage of the target in the signal					
2	0.926	0.873	0.805	0.714	0.532
3	0.904	0.838	0.737	0.628	0.381
4	0.891	0.807	0.690	0.569	0.288
5	0.888	0.784	0.663	0.514	0.256

D.3 ALL READERS, ALL CELLS

Table 10 is the unabridged version of Table 6.

Table 10: Every reader on identical mixtures, δ=10 dB, IID sampling. ORACLE is LIN on the oracle-masked mixture (the conditional ceiling).

			clip vector		frame tokens		mel	controls	
encoder	corpus	k	LIN	MLP	ATTN	MAX	CNN	CTRL	ORACLE
CLAP	ESC-50	2	0.899	0.910	0.931	0.939	0.804	0.978	0.998
CLAP	ESC-50	3	0.769	0.786	0.837	0.843	0.696	0.979	0.998
CLAP	ESC-50	5	0.678	0.702	0.755	0.741	0.624	0.978	0.998
CLAP	UrbanSound8K	2	0.879	0.917	0.892	0.890	0.881	0.996	0.998
CLAP	UrbanSound8K	3	0.740	0.777	0.800	0.795	0.780	0.996	0.998
CLAP	UrbanSound8K	5	0.641	0.662	0.702	0.688	0.671	0.994	0.998
AST	ESC-50	2	0.913	0.922	0.887	0.895	0.794	0.968	0.998
AST	ESC-50	3	0.794	0.816	0.780	0.819	0.690	0.971	0.998
AST	ESC-50	5	0.685	0.710	0.681	0.724	0.600	0.967	0.997
AudioMAE	ESC-50	2	0.769	0.790	0.794	0.854	0.791	0.968	0.990
AudioMAE	ESC-50	3	0.659	0.690	0.708	0.756	0.685	0.971	0.981
AudioMAE	ESC-50	5	0.603	0.595	0.629	0.649	0.601	0.967	0.985
AudioMAE	UrbanSound8K	2	0.739	0.784	0.771	0.788	0.869	0.994	0.982
AudioMAE	UrbanSound8K	3	0.637	0.626	0.685	0.670	0.777	0.995	0.977
AudioMAE	UrbanSound8K	5	0.555	0.540	0.577	0.543	0.678	0.996	0.978

D.4 TWO FURTHER CONTROLS ON THE CAPACITY EFFECT

Two controls accompany every conditional-ceiling cell. First, the target presented *alone* at exactly its in-mixture level is decodable at 0.95 or better in every cell for CLAP and AST and at 0.95 on average for AudioMAE, so being quiet is not by itself what costs the identity. Second, a crude mean-and-standard-deviation pooling of the mixture's own log-mel reaches only 0.48 to 0.63, far below $f(m)$, so the encoder is not merely passing its input through: it is organising it, and the capacity effect is a loss relative to that organisation rather than relative to nothing.

D.5 CORPUS DESCRIPTORS

Table 11 collects the descriptors used in Section 9. Neither spectral sparsity nor class count orders β.

Table 11: Corpus descriptors against β (400 isolated clips per corpus).

corpus	probing classes	Gini	active-bin frac.	p_{10}/peak (dB)	temporal flatness	β
ESC-50	50	0.980	0.290	-117.4	0.360	-0.121
FSD50K	42	0.986	0.220	-105.2	0.476	-0.071
UrbanSound8K	10	0.966	0.452	-91.3	0.707	-0.070

E A METHODOLOGICAL ASIDE: THE COMPOSITION OPERATOR AND ITS DOMAIN

A separate literature asks whether embeddings compose linearly under mixing, usually by testing $f(m) \approx w_1 f(a_1) + w_2 f(a_2)$ in representation space. Our setup makes it cheap to show that the answer depends on a choice that is rarely stated. Rather than change encoder, which would confound front end, data, objective and architecture at once, we sweep the composition operator on a single unchanged model. For k=2 we compose the sources' mel-power spectrograms with a weighted power mean of order p, $M_p = (w_1 M_1^p + w_2 M_2^p)^{1/p}$, convert back to the encoder's native input domain and measure $\cos(f(M_p), f(m))$. Here $p \to 0$ is the geometric mean, a linear combination in

a decibel input space; $p = 1$ is power addition, physically exact for uncorrelated sources; and $p \to \infty$ is the maximum operator of the log-max model (Nadas et al., 1988).

Cosines must be anchored before they can be compared. A raw cosine is not comparable across encoders. Two *unrelated* mixtures already sit at $\cos = 0.970$ in AudioMAE's space, leaving a usable range of 0.021 between chance and the noise ceiling, against 0.632 for CLAP and 0.631 for AST. An unanchored sweep would report AudioMAE as almost perfectly linear in every domain while carrying almost no information. We therefore report the normalised cosine $(\cos - \text{chance})/(\text{ceiling} - \text{chance})$ with anchors taken per encoder-corpus pair. This is the same failure mode as an ℓ_2-norm confound of Appendix A, one level further up the pipeline.

Result. On CLAP/ESC-50 the raw cosine rises smoothly and concavely with p: 0.641 at p=0, 0.848 at p=0.3, 0.905 at p=1, and a plateau at 0.907 to 0.911 for $p \geq 1.5$ including $p \to \infty$. The mel-synthesis anchor, which is the real mixture's own mel passed through the same synthesis path, is 0.918: discarding phase in $|Z|^2$, resampling and the feature extractor cost roughly 0.03 of cosine by themselves. The best composition therefore sits within 0.007 of the best any member of this family could reach, so at k=2 for uncorrelated sources the operator contributes essentially no residual.

Two things follow and one does not. First, composition in the encoder's own decibel input space is the outlier and every compressive order beats it. Second, against the honest baseline, which is the amplitude-weighted combination in embedding space at 0.845, moving to the right domain buys 0.07 of cosine: real, and small. What does *not* follow is that mel power is the uniquely correct domain, since the curve is flat from $p \approx 1$ to $p \to \infty$ and the model cannot separate power addition from the maximum. The practical consequence is a caveat rather than a phenomenon: an embedding-space linearity test conducted in the wrong domain reports a nonlinearity the model does not have. We do not build on it further.

F NEGATIVE RESULTS, IN FULL

1. **Spectral sparsity** (pre-registered, falsified). Prediction: sparser sources imply larger $|\beta|$. FSD50K is the sparsest on both Gini and active-bin fraction and has an intermediate $|\beta|$.
2. **Class count** (falsified after the sampler was fixed). FSD50K with 42 probing classes and UrbanSound8K with 10 give $\beta = -0.071$ and -0.070. Before the fix these were -0.072 and -0.044, and the apparent ordering was an artefact of the label leak, which is largest on the smallest vocabulary and therefore reshapes the ordering.
3. **Maximum-likeness predicts weak-source loss** (not supported). We tested whether an encoder whose power-mean sweep keeps rising toward the maximum operator discards weak sources harder. The correlation over six runs is $r = -0.50$, the sign the prediction wants, but it moved substantially when we corrected the AST normalisation and switched to centred cosines, that is, it is unstable under choices that should not matter. We report it and claim nothing from it.
4. **A withdrawn control and the statistic built on it.** RAND-$\mathcal{E}$ and the effective-flip-rate statistic, withdrawn under Proposition 2 (Remark 2). The earlier framing that the encoder *loses* assignment information as k grows was also withdrawn: adding k=4 and seed error bars showed the statistic moved by only 0.038 across k=2 to 5, and the proposition then showed the statistic to be uninformative in principle as well as in practice.
5. **A retracted decomposition.** The RAND-n-based $75/25$ assignment-versus-denoising split, superseded because RAND-n differs from $\mathcal{M}^\star$ in two variables at once.
6. **An experiment that could not measure what it was built for.** The co-occurrence study was designed to test whether source correlation changes what the encoder retains. It cannot, because the label-only shortcut dominates its metric. It is reported as a methodological result only.
7. **A normalisation bug we found and fixed.** AST's feature extractor divides by 2σ; an earlier inverse used σ, which rescales the power-mean order by exactly two. Verified by re-running: the old curve equals the new one at half the order at every grid point, while p=0 and $p \to \infty$ are unaffected by the error.

G REPRODUCIBILITY DETAILS

The pipeline in execution order: `step1--3` (corpora, mixture recipes, encoding); `step14` (conditional ceiling); `step19` (readers); `step20` (mask corruption, including SHUFFLE, RAND-$\mathcal{E}$ and RAND-n); `step21` (budget sweep); `step22` and `step26` (corpus descriptors); `step23` and `step24` (FSD50K single-source selection and the co-occurrence arms); `step25` (label-only bound and its simulation); `step18` (regressions); `step30` (figures); `step31--32` (value substitution and assembly). The sampler is selected by the environment variable `INDEP` and its value is recorded in every artefact, so the sampler used for any reported number can be checked without re-running. Artefacts produced before the sampler was fixed are retained under `results/leaky_backup/`, which is how the before-and-after comparison in Section 4 is computed.