What Do Audio Models Really Hear? Layer Selection and Mechanistic Structure in Sound Representations

Arnesh Batra
Indraprastha Institute of Information Technology Delhi (IIIT-Delhi)
New Delhi, India
arnesh23129@iiitd.ac.in

Abstract

Frozen audio encoders are reused across speech, sound, and music, and a downstream head usually reads a single layer, often the last. Intermediate layers of speech self-supervised (SSL) models are known to transfer better than the last one on many tasks. We ask how this depends on the pretraining family and the task, how cheaply a good layer can be found, and how task information is stored there. We compare seven frozen encoders, covering audio–text contrastive training (CLAP), weakly supervised speech recognition (Whisper), and five speech SSL objectives, on environmental sound, urban sound, music, and speech emotion, and check the findings on larger checkpoints and on the audio encoders of Qwen-Audio and SALMONN. (i) The best layer depends on encoder and task: speech SSL encoders give their best non-speech features in layers 0–4 of 12, and the last layer costs up to 38 accuracy points, whereas CLAP peaks at its top stages and Whisper in its middle-to-late layers. (ii) Choosing the layer whose unlabeled outputs have the highest isotropy and participation ratio reduces character error rate by 21.3% relative to the last layer in low-resource ASR and matches a middle-layer default on classification with an earlier exit, but fails on speech emotion; one labeled clip per class recovers 85% of the best layer's gain over the last layer. (iii) Truncating wav2vec2-base after the layer selected for ESC-50 raises accuracy, nearly halves latency, and cuts resident memory by 77%. Sparse autoencoders, transcoders, corruption tests, and steering show that CLAP stores class evidence in a few class-specific features, whereas Whisper spreads it over many features that change across layers and under noise. We release all per-layer measurements.

1 Introduction

Pretrained audio encoders such as wav2vec 2.0, HuBERT, WavLM, Whisper, and CLAP are often used as frozen feature extractors: a small head is trained on the output of one layer for a new task [57, 52]. Which layer to read is a real design choice. For speech SSL models it is well established that intermediate layers often carry more phonetic and lexical information than the last layer [38, 39], and similar effects are known for text and image models [35, 13]. Practitioners therefore commonly read a middle layer, choose a layer with a validation probe, learn a weighted sum of all layers [40, 57], or use labels to select task-discriminative subspaces across layers [6].

This paper asks three follow-up questions on a broader set of encoders and tasks. **Q1:** How does the best layer depend on the pretraining family and on the task, beyond speech SSL models on speech tasks? **Q2:** How well can a layer be chosen with no labels, or with a handful, compared with strong defaults such as the middle layer? **Q3:** How is task information stored at the chosen layer: in a few class-specific features or spread across many, and does it survive acoustic corruption? The questions

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

share one practical consequence. If a task is best served by layer ℓ, the layers after ℓ need not be computed for that task, so choosing a layer also decides how much of the encoder to run.

We study seven frozen encoders: CLAP-HTSAT (audio–text contrastive), the Whisper-small encoder (weakly supervised ASR), and five 12-layer speech SSL models with different training targets (HuBERT, WavLM, UniSpeech-SAT, wav2vec 2.0, data2vec), on environmental sound, urban sound, music genre, speech emotion, and low-resource ASR. For every (encoder, dataset, layer) triple we record probe accuracy, two geometry statistics, and robustness to corruption; at the selected layers we add sparse-autoencoder (SAE) features, transcoder routes, and steering and ablation effects, and we release all of these measurements. Our main findings are:

- **The best layer depends on encoder family and task (Sec. 4.1).** Speech SSL encoders peak in layers 0–4 on non-speech tasks, and the last layer is at least 10 points worse than the best in 16 of 28 encoder–task cells. CLAP peaks at its top stages, Whisper-S in its middle-to-late layers, and the audio encoders of Qwen-Audio and SALMONN at or near the top.
- **Label-free selection helps for content-type tasks, not for affect (Sec. 4.2).** A geometry rule lowers low-resource ASR character error rate (CER) from 62.4 to 49.1, versus 51.3 for a fixed middle layer. On classification it only matches middle-layer defaults, with an earlier exit.
- **Exiting early gives measured savings (Sec. 4.5).** Truncating wav2vec2-base after layer 2 cuts median latency by 47.5% while raising ESC-50 accuracy.
- **Encoders store the same task differently (Secs. 4.3–4.4).** CLAP concentrates class evidence in about eight SAE features; Whisper-S spreads it over many features that change across layers and under corruption.

Scope. This is an empirical study. Probes, CKA, isotropy, participation ratio, SAEs, transcoders, and steering are existing tools used without modification, and the layer-selection rules are deliberately simple baselines. We contribute a matched comparison across pretraining families and audio domains, a test of label-free and few-shot layer choice against strong defaults (including where it fails), measured early-exit costs, and the per-layer measurements with a procedure for using them (Sec. 4.6). Because released checkpoints differ in objective, data, architecture, and scale at once, statements about pretraining families are associations, not causal effects.

2 Related Work

Layer-wise analysis and layer choice. Linear probes on intermediate layers are the standard tool for locating information across depth [1, 51, 28], and in text and vision intermediate layers often transfer better than the last one [35, 13]. For speech SSL models, Pasad et al. [38, 39] showed that phonetic and word content peaks in intermediate layers and declines in the final layers, and compared these trends across SSL objectives. SUPERB and HEAR evaluate frozen encoders on speech and general audio [57, 52, 47]; SUPERB learns a softmax-weighted sum of all layers, as ELMo did for text [40]. LOES [6] showed across modalities that task-relevant information is distributed non-monotonically over depth and is not recovered by naive aggregation, and selects task-discriminative layer subspaces with a supervised spectral method. Choosing a layer with a validation probe is common practice, and we treat it, the middle layer, and a literature-based prior as baselines. We add a matched comparison across audio–text, ASR-supervised, and speech SSL encoders on non-speech as well as speech tasks, test how well a single layer can be chosen with few or no labels, and connect the selected layer to measured inference cost.

Representation geometry. Ethayarajh [22] used isotropy, i.e., how evenly representations spread over directions, to show that contextual word representations become more anisotropic with depth; related work studied and corrected anisotropy in word embeddings [36, 24]. Participation ratio is a standard effective-dimensionality measure in neuroscience and deep-network analysis [25, 34, 45, 21], and intrinsic dimension varies systematically across layers [2]. We use both statistics unchanged and only test whether their combination is a useful zero-label layer selector for audio. CKA compares representations across layers and models [31].

Sparse features, transcoders, and interventions. SAEs rewrite a layer's activations as sparse combinations of learned directions that are often easier to interpret [20, 30, 8, 50]; transcoders

approximate the computation from one layer to the next through a sparse bottleneck [18]. Activation steering adds a direction to a hidden state to change model behaviour [53, 33], and INLP and LEACE remove concepts by projection [44, 7]. In audio, NMF explanations [37] and AudioSAE [3] applied sparse decompositions to Whisper and HuBERT. Our feature ablations are inference-time edits, not machine unlearning [10].

Early exit and layer dropping. Early-exit models and layer-dropping studies show that later layers can often be skipped with little loss [56, 27], and DistilHuBERT compresses a speech SSL model by layer-wise distillation [11]. Here the exit follows directly from layer selection, without retraining the encoder.

3 Setup

Encoders and layer numbering. We use public checkpoints of CLAP-HTSAT (86M parameters) [55, 12], the encoder of Whisper-small [42] (Whisper-S; 244M parameters in the full model), and the Base versions of HuBERT [29], WavLM [14], UniSpeech-SAT [15], wav2vec 2.0 [4], and data2vec-Audio [5] (94–95M each). Apart from CLAP, each encoder has 12 transformer blocks with 768-dimensional outputs, and we number its layer outputs 0–12: layer 0 is the input to the first block (the output of the convolutional front end and positional embedding) and layer $\ell \geq 1$ is the output of block ℓ. Thus 0–12 covers the whole encoder and layer 12 is the final layer. CLAP-HTSAT is a hierarchical transformer with four stages; index 0 is its patch-embedding output and 1–4 are the stage outputs. These stage numbers are not comparable with speech-model layers, so CLAP is left out of averages over the six common-depth encoders. Layer outputs are mean-pooled over time (and frequency for CLAP).

Tasks. ESC-50 (2,000 clips, 50 classes) [41], UrbanSound8K (8,732 clips, 10 classes) [46], GTZAN (999 clips, 10 genres) [54], and CREMA-D (7,442 clips, 6 emotions) [9]. Because GTZAN has known duplicates and mislabels [48], we repeat the music analysis on FMA-small [17] and MagnaTagATune [32]. Splits are in Appendix A. Low-resource ASR covers six languages (Amharic, Welsh, Hausa, Kyrgyz, Swahili, Yoruba) with wav2vec2-large and data2vec encoders, scored by CER.

Probes. For each (encoder, dataset, layer) we train a linear classifier on the pooled, frozen layer output and report test accuracy, or macro-F1 for CREMA-D. The layer with the best full-data probe score is the *oracle layer*. Encoders stay frozen except in one fine-tuning check (Sec. 4.5).

Geometry statistics. Let $\{z_i^{(\ell)}\}_{i=1}^N$ be the pooled layer-ℓ outputs of N unlabeled clips, with covariance eigenvalues $\lambda_1, \ldots, \lambda_d$. The participation ratio [25]

$$\mathcal{PR}^{(\ell)} = \frac{\left(\operatorname{tr} \hat{\Sigma}^{(\ell)}\right)^2}{\|\hat{\Sigma}^{(\ell)}\|_F^2} = \frac{\left(\sum_k \lambda_k\right)^2}{\sum_k \lambda_k^2} \tag{1}$$

is the effective number of directions with substantial variance and lies between 1 and d. Isotropy, following Ethayarajh [22], is one minus the absolute mean cosine similarity between outputs,

$$\mathcal{I}^{(\ell)} = 1 - \left| \frac{2}{N(N-1)} \sum_{i<j} \frac{(z_i^{(\ell)})^\top z_j^{(\ell)}}{\|z_i^{(\ell)}\| \, \|z_j^{(\ell)}\|} \right|, \tag{2}$$

which is near 1 when outputs point in varied directions and near 0 when they share a common direction. Neither statistic uses labels.

Layer-selection rules. Five rules use no target labels: (a) the final layer; (b) a fixed middle layer (6 of 12); (c) a *family prior*: layer 6 for the speech SSL encoders, following Pasad et al. [38, 39], and layer 12 for Whisper-S, whose final encoder output is what its decoder reads; (d) two learned selectors, trained leave-one-encoder-out on the other encoders' layer-wise geometry (Appendix C.7); and (e) the *geometry rule*, which min–max normalizes $\mathcal{I}$ and $\mathcal{PR}$ over one encoder's layers and picks

$$\hat{\ell}_{\text{geo}} = \arg\max_{\ell} \; \tfrac{1}{2} \left[\widetilde{\mathcal{I}}^{(\ell)} + \widetilde{\mathcal{PR}}^{(\ell)} \right]. \tag{3}$$

The equal weighting was fixed in advance, not tuned. The *few-shot probe rule* trains probes on $s \in \{1, 5, 10, 20, 50\}$ labeled clips per class and picks the best validation layer. With more labels, a

Table 1: Best single-layer probe score per encoder and task, with the best layer in parentheses (layers 0–12, where 12 is the final layer; CLAP: stages 0–4). Accuracy, except macro-F1 for CREMA-D. Bold: best encoder per task. The last row is a linear probe on combined classical descriptors (MFCC, mel, spectral, chroma, rhythm).

Pretraining	Encoder	ESC-50	GTZAN	CREMA-D	US8K
Audio-text contrastive	CLAP-HTSAT	**.970** (4)	**.825** (3)	.604 (3)	**.890** (4)
Supervised ASR	Whisper-S	.860 (6)	.745 (8)	**.732** (12)	.850 (6)
Masked SSL	HuBERT-B	.715 (3)	.700 (2)	.676 (12)	.778 (1)
Denoising SSL	WavLM-B	.708 (3)	.690 (1)	.699 (6)	.774 (3)
Speaker-aware SSL	UniSpeech-SAT	.720 (2)	.695 (2)	.681 (12)	.785 (4)
Contrastive SSL	wav2vec2-B	.693 (2)	.715 (2)	.647 (2)	.751 (0)
Self-distillation	Data2Vec	.665 (2)	.690 (0)	.673 (0)	.730 (0)
Classical features	Combined	.628	.620	.569	.748

learned layer mixture can be restricted to the layers up to the selected layer $\hat{\ell}$,

$$z_{\text{prefix}} = \sum_{\ell=0}^{\hat{\ell}} \alpha_\ell z^{(\ell)}, \qquad \alpha_\ell = \frac{\exp a_\ell}{\sum_{j=0}^{\hat{\ell}} \exp a_j}. \tag{4}$$

This is SUPERB's weighted sum without the layers after $\hat{\ell}$, so those layers still need not run.

Feature-level tools. These are applied at the oracle layer of each encoder–dataset pair (settings in Appendix A.3). A *sparse autoencoder* (SAE) rewrites each layer output as a sparse, non-negative combination of 4,608 learned directions, called features (Top-k, k annealed from 130 to 75). A feature is *class-specific* if more than 70% of its activation mass falls on one class (we sweep 50–90%). The *minimum sufficient feature set* (MSFS-50) of a class is the number of its top-ranked features that must be zeroed before the probe's score on that class halves. A *transcoder* [18] predicts layer $\ell{+}1$'s output from layer ℓ's output through 16 active latents out of 3,840; its explained variance R^2 measures how well a sparse route summarizes that step, and we also report the share of active latents that are class-specific. *Feature genealogy* matches SAE features at adjacent layers by the overlap of their top-activating clips and counts features that persist, split, merge, die, or emerge. *Steering* adds αv to a layer output, with v the difference of two class means and $\alpha \in \{0.5, 1, 2, 4\}$; selectivity is the probe's gain on the target class minus its loss on other classes. *Robustness* tests apply additive noise, band-pass filtering, and frequency shift, and report the accuracy drop, the cosine drift of layer outputs, and the *feature drift* (Jaccard distance between the active SAE feature sets of the clean and corrupted input).

4 Results

4.1 Where is task information? (Q1)

Table 1 gives each encoder's best layer and score. Three patterns stand out.

Speech SSL encoders peak early on non-speech audio. On ESC-50, GTZAN, and UrbanSound8K, all five speech SSL encoders peak in layers 0–4, and Data2Vec peaks at layer 0 on three of the four tasks. Their final layers are much worse (Table 2): on ESC-50, wav2vec2-B drops from .693 at layer 2 to .315 at layer 12, and Data2Vec from .665 to .300. This extends the speech-task findings of Pasad et al. [38] to environmental sound and music.

CLAP and Whisper peak later, with small final-layer losses. CLAP peaks at stage 3 or 4 and is the strongest encoder on the three non-speech tasks; its last stage is within 2 points of its best on every task. Whisper-S peaks at layers 6–12 and loses at most 5 points at layer 12 (on GTZAN). For CLAP a late peak is expected, since its training aligns the top of the network with text descriptions of sound categories.

Speech emotion is the exception. On CREMA-D, Whisper-S, HuBERT-B, and UniSpeech-SAT peak at their final layer. This is also where label-free selection fails (Sec. 4.2).

Table 2: Largest final-layer losses (best-layer score minus final-layer score). Parentheses: share of the 12 blocks that must run to reach the best layer.

Encoder	Task	Best layer	Best	Final	Loss
wav2vec2-B	ESC-50	2 (17%)	.693	.315	.378
Data2Vec	ESC-50	2 (17%)	.665	.300	.365
wav2vec2-B	GTZAN	2 (17%)	.715	.430	.285
Data2Vec	GTZAN	0 (0%)	.690	.415	.275
wav2vec2-B	US8K	0 (0%)	.751	.505	.246
WavLM-B	ESC-50	3 (25%)	.708	.463	.245

Table 3: Correlation between encoders' per-depth CKA curves (colour version: Appendix Figure 3). HuB: HuBERT-B, WLM: WavLM-B, USAT: UniSpeech-SAT, w2v: wav2vec2-B, d2v: Data2Vec, Whs: Whisper-S.

	HuB	WLM	USAT	w2v	d2v	Whs
WLM	.98					
USAT	.99	.99				
w2v	.73	.84	.80			
d2v	.91	.83	.86	.42		
Whs	.77	.72	.78	.51	.61	
CLAP	−.63	−.58	−.64	−.44	−.49	−.97

Table 4: Best versus final layer for larger encoders (accuracy, %). The Qwen-Audio and SALMONN rows use only the audio encoder (for SALMONN, its Whisper-large-v2 speech branch); layer indices follow each checkpoint's exported hidden states. Full profiles and bootstrap intervals are in Appendix G.

Encoder	Task	Best layer: score	Final layer: score	Loss
WavLM-large	ESC-50	L9: 82.3	L24: 69.0	13.3
WavLM-large	GTZAN	L18: 74.0	L24: 64.0	10.0
Whisper-large-v3	UrbanSound8K	L11: 86.0	L32: 78.9	7.1
Whisper-large-v3	GTZAN	L17: 77.5	L32: 71.0	6.5
Qwen-Audio encoder	ESC-50	L26: 93.25	L31: 93.00	0.25
SALMONN encoder	ESC-50	L32: 87.25	L32: 87.25	0.00
SALMONN encoder	GTZAN	L29: 88.0	L32: 85.0	3.0

Overall, the final layer is at least 10 points below the best layer in 16 of the 28 encoder–task cells, and for 20 of the 24 common-depth cells the paired 95% bootstrap interval of the best-minus-final gain excludes zero. CKA depth profiles group the encoders the same way (Table 3): HuBERT-B, WavLM-B, and UniSpeech-SAT are nearly identical ($r \geq .98$), wav2vec2-B and Data2Vec are further away, and CLAP is anti-correlated with every speech encoder. Appendix A.4 gives a sufficient condition (Proposition 1) under which a pretraining target makes a downstream label linearly separable at a layer.

Cleaner music data and larger models. On FMA-small, the best layer beats the final layer by 10.8 points for wav2vec2-B (95% CI [7.1, 14.5]) and 10.3 [5.8, 14.4] for Data2Vec, while CLAP again peaks at its last stage. On MagnaTagATune, the validation-selected wav2vec2-B layer (layer 0) improves macro-mAP over the final layer by 4.8 points [4.1, 5.6], and the fixed middle layer by 1.5 [0.8, 2.2] (Appendix C.6). Among larger encoders (Table 4), WavLM-large and Whisper-large-v3 still have clearly non-final optima. For the 32-layer audio encoders of Qwen-Audio [16] and SALMONN [49], both derived from Whisper-large-v2, the best layer is at or within five blocks of the last, the final-layer penalty is at most 3 points, and participation ratio and isotropy still track probe accuracy across layers (rank correlation $\rho = .80$ and .82 for Qwen-Audio on ESC-50). Layer choice thus matters most for speech SSL encoders on non-speech audio and least for large ASR-derived encoders.

4.2 Choosing a layer with few or no labels (Q2)

Without labels. In low-resource ASR, transcripts are the scarce resource, so layers are selected from unlabeled target-language audio. The geometry rule lowers aggregate CER from 62.37 (final layer) to 49.09, a 21.3% relative reduction (Table 5), and improves 11 of 12 language–encoder settings, all 11 with paired 95% bootstrap intervals that exclude zero (10,000 utterance resamples; the exception is Data2Vec on Hausa, Appendix Table 22). It also beats the fixed middle layer (51.27) while selecting earlier layers (4.1 vs. 6.0 on average). On the 24 classification cells the comparison with strong defaults is closer: the geometry rule is 5.73 points below the oracle on average, against 5.99 for the middle layer, 5.91 for the family prior, and 15.64 for the final layer. Most of the gain thus comes from avoiding the final layer, which the defaults already do; the geometry rule's advantage is an earlier exit (layer 5.75 on average, about 48% of the stack). The learned selectors do not beat it, which is unsurprising with five training encoders per fold.

Table 5: Layer selectors that use no target labels. Classification: mean gap to the oracle layer in score points (lower is better) over the 24 common-depth encoder–task cells, and mean selected layer (of 12). ASR: aggregate CER over 12 language–encoder settings and mean selected layer. Learned selectors are trained leave-one-encoder-out. –: not evaluated or not applicable.

	Classification (24 cells)		Low-resource ASR (12)	
Selector (no target labels)	Gap to oracle ↓	Layer ↓	CER (%) ↓	Layer ↓
Final layer	15.64	12.0	62.37	12.0
Fixed middle layer (6)	5.99	6.0	51.27	6.0
Family prior (SSL: 6, Whisper: 12)	5.91	7.0	–	–
Learned, encoder summary	7.73	–	–	–
Learned, layer profile	6.19	–	–	–
Geometry rule (Eq. 3)	**5.73**	**5.75**	**49.09**	**4.1**

Table 6: Gains over final-layer extraction for the six common-depth encoders. "Final" is the final-layer score; other entries are selected-layer score minus Final. Geometry uses no labels; the probe rule uses s labeled clips per class. The bottom rows count cells whose paired 95% bootstrap interval for the gain lies entirely above or below zero; at 50 shots ESC-50 is excluded from the count because it has only 32 training clips per class.

		No labels	Few-shot probe: labeled clips per class				
Dataset	Final	Geometry	1	5	10	20	50
ESC-50	.502	+.148	+.210	+.213	+.218	+.222	+.225
CREMA-D	.637	−.027	+.023	−.001	+.010	+.029	+.028
GTZAN	.524	+.133	+.151	+.122	+.162	+.166	+.177
US8K	.650	+.113	+.114	+.108	+.108	+.115	+.105
Mean	.578	+.092	+.124	+.111	+.124	+.133	+.134
Interval above 0	–	15/24	16/24	17/24	17/24	17/24	12/18
Interval below 0	–	1/24	2/24	0/24	0/24	0/24	0/18

When can a label-free rule work? The rule sees the encoder and the unlabeled target audio, but not the task. It can adapt to the audio domain, but it cannot separate two tasks on the same audio, such as transcription and emotion recognition on the same speech. It works when the labels follow the dominant variation in the target audio (phonetic content for ASR, sound-source identity for event and music datasets) and the encoder's final layer has moved away from that variation, as speech SSL encoders do on non-speech audio. It fails otherwise (Appendix Table 24): on CREMA-D, where emotion is a low-variance property of speech, it is 2.7 points worse than the final layer on average, and on CLAP, whose last stage is already aligned with sound categories, it is worse than the last stage on all four tasks. Consistent with this, geometry and probe score are weakly correlated when pooled over all encoders and tasks (Pearson $r = .21$ for $\mathcal{PR}$, .27 for $\mathcal{I}$), moderately correlated within ASR (.38 and .42), and differ in sign across encoders (Appendix C.4). We therefore recommend the geometry rule only for content-type tasks without labels.

With a few labels. A probe trained on one labeled clip per class gains 12.4 points over the final layer on average, against 13.4 points with 50 clips per class (Table 6), and recovers 85% of the oracle's gain (Appendix C.7). The gain is significant in 16–17 of 24 cells at 1–20 shots; the CREMA-D entries are within noise and do not form a trend. The selected layer often differs from the oracle layer because neighbouring layers score almost the same, so the score gap is the relevant measure.

4.3 Is the selected layer robust? (Q3)

A layer that is accurate on clean audio can still be fragile. Under controlled corruption (Table 7), the denoising and masked-prediction SSL encoders lose the least accuracy (.044 and .065), while Whisper-S loses the most (.280) even though its layer outputs move the least (cosine drift .074). Feature drift explains the mismatch: Whisper-S has the largest feature drift (.407), so the particular sparse features its probe relies on switch on and off under corruption while the output vector as a whole barely changes. This fits the finding that Whisper's encoder represents background-noise conditions rather than ignoring them [26]. Robustness also varies with depth: under severe band-pass

Table 7: Robustness under controlled corruption (additive noise, band-pass filtering, frequency shift), by pretraining family. Lower is better for the accuracy drop and both drifts; relative robustness is scaled so the best family is 1.00. CLAP values are stage-wise and not depth-matched to the speech encoders.

Family	Accuracy drop	Cosine drift	Feature drift	Rel. robustness
SSL denoising (WavLM-B)	.044	.103	.117	1.00
SSL masked prediction	.065	.195	.168	.89
SSL contrastive	.083	.460	.138	.87
Audio-text contrastive (CLAP)	.101	.377	.207	.68
Supervised ASR (Whisper-S)	.280	.074	.407	.00

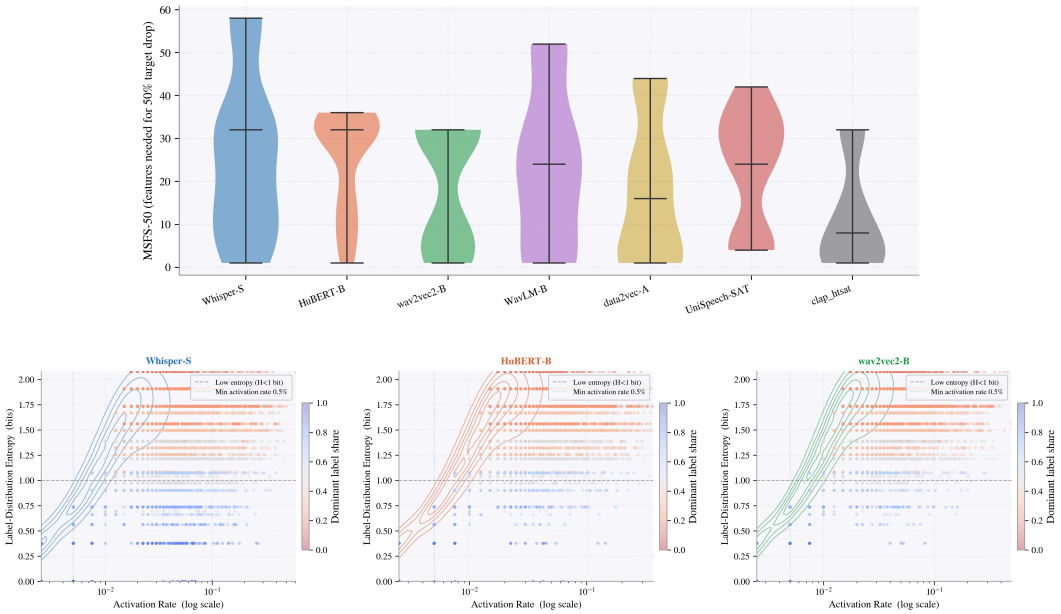

Figure 1: How class evidence is spread over SAE features at each encoder's best layer (ESC-50). **Top:** MSFS-50, the number of top-ranked features that must be zeroed to halve a class's probe score, over the 50 classes (bars: minimum, median, maximum; "data2vec-A" is Data2Vec). Lower means more concentrated. **Bottom:** each SAE feature of Whisper-S, HuBERT-B, and wav2vec2-B by how often it fires (x, log scale) and the entropy of the labels it fires on (y); colour is the share of its activation on its dominant class (blue: class-specific, red: shared), and contours show feature density. Rare, low-entropy features act as class detectors; frequent, high-entropy features are reused across classes. Whisper-S and HuBERT-B are dominated by shared features, matching their larger MSFS-50; wav2vec2-B has both kinds, matching its bimodal MSFS-50.

filtering the final layer loses 11.3 points less than layer 0, whereas a +4 semitone pitch shift affects all depths about equally (Appendix Figures 5 and 6). The most accurate clean layer is thus not always the most reliable one, which is why the exit procedure of Section 4.6 can include a corruption check.

4.4 How is task information stored at the selected layer? (Q3)

At each encoder's best layer we ask whether class evidence sits in a few class-specific features or is spread across many shared ones. Three measurements address this (Figure 1, Table 8).

Feature budget. Zeroing a median of eight ranked SAE features halves a CLAP class score on ESC-50. Whisper-S, HuBERT-B, WavLM-B, and UniSpeech-SAT need medians of 24–32 features and Data2Vec about 16, while wav2vec2-B is bimodal, with some classes removable by a few features and others needing many (Figure 1, top). The bottom panels show why: in Whisper-S and HuBERT-B most features fire often and on many classes, and class-specific features are rare.

Table 8: Feature-level measurements at each encoder's best layer. SAE share: fraction of SAE features that are class-specific (more than 70% of activation mass on one class), pooled over datasets. Transcoder R^2: range over the four datasets of the best transition's explained variance. Transcoder share: class-specific fraction of active transcoder latents on UrbanSound8K, where it is highest for every encoder. Persist and split: genealogy event rates between adjacent layers (common-depth encoders only).

Encoder	SAE share	Transcoder R^2	Transcoder share	Persist	Split
CLAP-HTSAT	.188	.867–.983	.774	–	–
Whisper-S	.136	.933–.998	.000	.001	.576
HuBERT-B	.135	.906–.957	.321	.031	.549
WavLM-B	.143	.865–.941	.296	.031	.537
UniSpeech-SAT	.119	.880–.950	.471	.041	.445
wav2vec2-B	.110	.935–.994	.152	.026	.587
Data2Vec	.156	.859–.983	.190	.029	.451

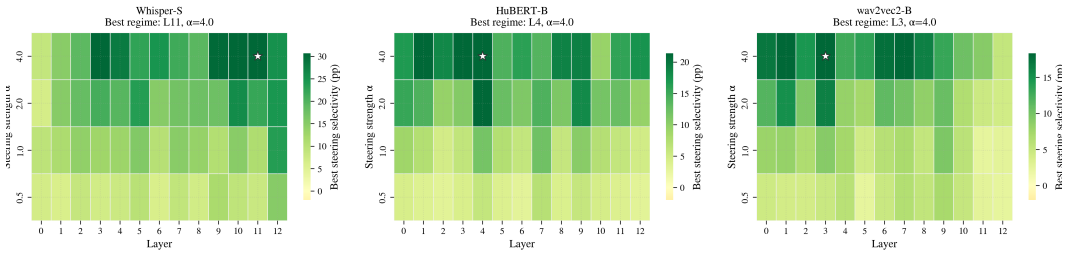

Figure 2: Steering selectivity by layer (x) and steering strength α (y) for Whisper-S, HuBERT-B, and wav2vec2-B. Each cell is the best selectivity (target-class gain minus off-target loss, in percentage points) over the evaluated class contrasts when the class-difference vector is added at that layer; the star marks the maximum. Selectivity peaks late for Whisper-S (layer 11) and early for HuBERT-B (layer 4) and wav2vec2-B (layer 3), close to the layers chosen by probing.

Layer-to-layer routes. Transcoders predict the next layer well for every encoder (R^2 from .86 to .998), but being predictable is not the same as being interpretable. On UrbanSound8K, 77% of CLAP's active transcoder latents are class-specific, compared with 15–47% for the speech SSL encoders and none for Whisper-S at the 70% criterion. Across all transitions, R^2 and the class-specific share are almost uncorrelated ($r = .12$).

Stability across layers. SAE features rarely persist unchanged between adjacent layers. Whisper-S has the lowest persistence (.001) and one of the highest split rates (.576).

Taken together for Whisper-S: many features must be removed to erase a class, its sparse routes are predictable but not class-specific, its features keep splitting across layers, and they switch under corruption (Sec. 4.3). We read this as class information held in a broad code that is rewritten from layer to layer, rather than in a small, stable set of class detectors. CLAP shows the opposite profile, with compact, class-specific features at its top stages. Those narrow features also drift under corruption (feature drift .207), so compactness does not imply robustness.

Sensitivity to thresholds. The contrasts do not hinge on the 70% and 50% cut-offs. When the class-specific threshold is swept from 50% to 90%, CLAP keeps a larger class-specific share of transcoder latents than Whisper-S at every threshold (from .889 to .611 for CLAP, from .812 to .188 for Whisper-S); the zero for Whisper-S in Table 8 refers to its single best transition at 70%. For feature ablation, every encoder reaches 25%, 50%, and 75% target drops within the ablation range, and none reaches 90% (Appendix D). The SAE results remain conditional on our dictionary size and sparsity.

Steering and ablation. Adding a class-difference vector is most selective at layer 11 for Whisper-S (about 30 points), layer 4 for HuBERT-B (about 20), and layer 3 for wav2vec2-B (about 18; Figure 2), close to each encoder's probe-selected layers (Whisper-S: 6–12; HuBERT-B: 1–3 on non-emotion tasks; wav2vec2-B: 0–2). The largest gains come with noticeable off-target movement (Appendix

Table 9: Measured effect of truncating wav2vec2-base after layer 2, its best ESC-50 layer (batch size 1). Accuracy: frozen-encoder ESC-50 probe. Memory and the first latency/throughput pair use 5-second clips; the last pair uses 30-second clips, for which resident memory also falls by 66.3%.

	Acc.	Memory (MiB)			5 s clips		30 s clips	
Encoder	(%)	Params	Resident	Peak	ms	clips/s	ms	clips/s
Exit at layer 2	69.25	89.6	90.2	181.2	4.59	216.7	25.34	39.46
Full (layer 12)	32.00	360.0	389.1	470.9	8.74	114.1	58.46	17.10
Change	+37.25	−75.1%	−76.8%	−61.5%	−47.5%	+89.8%	−56.7%	+130.8%

Table 10: Fields of a layer report, which has one row per encoder, dataset, and layer. "Labels" marks fields that need labeled target data. "Role" marks whether the procedure in Section 4.6 uses the field to choose an exit or whether the field only describes the chosen layer.

Field	What it measures	Labels	Role
Probe score (few-shot or full)	How linearly accessible the task is	yes	choose
Geometry score (Eq. 3)	Spread of unlabeled layer outputs	no	choose
Depth, latency, memory	Cost of running layers $0, \ldots, \ell$	no	choose
Score and drift under an expected shift	Reliability under corruption	partly	choose
Transcoder R^2 for $\ell \to \ell{+}1$	Faithfulness of the sparse route	no	choose
SAE fidelity, class-specific share, MSFS-50	Concentration of class evidence	yes	describe
Steering selectivity, ablation collateral	Whether class directions can be edited	yes	describe
CKA profile, feature genealogy	Similarity across depth and encoders	no	describe

Figure 14), and targeted feature ablation removes a class with less collateral change when the class is carried by few features (Appendix E). We use both as checks that the selected layer exposes usable linear directions, not as editing methods.

4.5 From the selected layer to a smaller encoder

If an earlier layer serves the task, the blocks after it need not run. We truncated wav2vec2-base after layer 2, its best ESC-50 layer (Table 9). Frozen ESC-50 accuracy rises from 32.0% to 69.3%, parameter memory falls by 75.1%, and median latency falls by 47.5% (5-second clips) and 56.7% (30-second clips). Elsewhere we use depth as the cost proxy; these measurements show that it turns into real savings for this model, though exact numbers depend on hardware and sequence length. The advantage also survived one controlled fine-tuning check (wav2vec2-base, ESC-50, 30 epochs end-to-end): the truncated model reaches 66.0% against 62.5% for the full model. Fine-tuning may move the useful depth elsewhere in other settings.

Is a prefix enough when labels are plentiful? A learned mixture over the layers up to the selected exit (Eq. 4) matches or beats the same mixture over all layers in 19 of 27 encoder–dataset cells (mean difference $+0.2$ points over five seeds; Appendix Table 18). It is better on ESC-50 ($+1.7$) and GTZAN ($+0.6$), equal on UrbanSound8K and FMA-small, and worse on CREMA-D (-1.4). With the exit at layer 5.75 on average, about half of the encoder is often, but not always, enough.

4.6 Using the measurements to choose an exit

The released measurements form one row per (encoder, dataset, layer) with the fields in Table 10; the rows for one encoder and dataset form its *layer report*. We use the report in three steps.

1. **Shortlist.** Fit few-shot probes (five clips per class) at every layer and keep the layers whose score is within a tolerance ϵ of the best (Eq. 16). Without labels, shortlist the top layers by geometry score.
2. **Apply the deployment constraint.** If a specific acoustic shift is expected, such as telephone-band audio, re-score the shortlisted layers on shifted copies of the labeled clips and keep the

most stable one (*perturbation-aware*). If compute matters most, take the earliest shortlisted layer whose transcoder route to the next layer is faithful (*routing priority*).

3. **Exit.** Read out the chosen layer and drop the blocks after it. If more labels arrive later, train the prefix mixture of Eq. 4 up to that layer.

On held-out ESC-50 data, the extra fields change 52–78% of five-shot exit decisions relative to the probe alone. For telephone-band audio, perturbation-aware selection improves accuracy by 0.46 points (95% CI [0.07, 1.00]) at the same label budget, and routing priority exits 0.51 layers earlier on average while improving accuracy by 1.29 points. We have not ablated the descriptive fields one by one, so their individual effect on the exit is not established; they serve users who want to understand, edit, or audit the chosen layer (Sec. 4.4). Appendix C.2 writes the trade-off as a utility over score, cost, and drift.

5 Discussion and Limitations

Practical guidance. For speech SSL encoders on non-speech audio, the final layer is a poor default and a middle layer a reasonable one; the geometry rule makes a similar choice with an earlier exit and works without transcripts for ASR. For speech emotion and for CLAP, late layers are preferable and label-free selection should not be trusted. One to five labeled clips per class usually suffice to confirm the choice, and the chosen layer is often early enough to make inference cheaper. The encoders also differ in how task information is stored, compactly in CLAP and broadly and unstably in Whisper, and the SAE, transcoder, genealogy, and corruption measurements agree on this contrast. Code and all per-layer measurements are released with the paper.

Limitations. (i) Released checkpoints differ in objective, data, architecture, and scale together, so family-level statements are associations. (ii) We study frozen encoders with linear readouts of one layer or a prefix; the fine-tuning check covers one model and task, and full fine-tuning or full-depth mixtures can prefer other depths, as CREMA-D shows. (iii) The geometry rule cannot tell apart tasks on the same audio and fails on speech emotion and on CLAP. (iv) Savings were measured for one model; elsewhere depth is the cost proxy. (v) SAE and transcoder results depend on dictionary size, sparsity, and thresholds, and the exit procedure was validated on ESC-50 only. (vi) Test sets are small (400 clips for ESC-50, 200 for GTZAN), so we report intervals where available (Appendix H).

Societal impact. Cheaper and better-understood audio encoders can help accessibility tools, acoustic safety monitoring, and on-device inference. The same diagnostics could make audio surveillance easier to tune for sensitive classes or settings. We present steering and ablation as diagnostics, and recommend reporting off-target effects, robustness, and dataset provenance before any user-facing deployment.

Acknowledgments and Disclosure of Funding

The author thanks the anonymous reviewers and the area chair for detailed feedback that substantially improved this paper.

References

[1] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations, Workshop Track*, 2017. URL https://arxiv.org/abs/1610.01644.

[2] Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/cfcce0621b49c983991ead4c3d4d3b6b-Abstract.html.

[3] Georgii Aparin, Tasnima Sadekova, Alexey Rukhovich, Assel Yermekova, Laida Kushnareva, Vadim Popov, Kristian Kuznetsov, and Irina Piontkovskaya. AudioSAE: Towards understanding of audio-processing models with sparse AutoEncoders. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3221–3254, Rabat, Morocco, March 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.eacl-long.149. URL https://aclanthology.org/2026.eacl-long.149/. arXiv preprint arXiv:2602.05027.

[4] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460, 2020. URL https://arxiv.org/abs/2006.11477.

[5] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 1298–1312. PMLR, 2022. URL https://arxiv.org/abs/2202.03555.

[6] Arnesh Batra, Arush Gumber, Aniket Khandelwal, Jashn Khemani, and Anubha Gupta. Uncovering the latent potential of deep intermediate representations. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 7020–7055. PMLR, 2026. URL https://openreview.net/forum?id=6up1qGJwYZ.

[7] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*, volume 36, pages 66044–66063, 2023. doi: 10.52202/075280-2884.

[8] 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*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

[9] 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 Transactions on Affective Computing*, 5(4):377–390, 2014. doi: 10.1109/TAFFC.2014.2336244.

[10] Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pages 463–480. IEEE, 2015. doi: 10.1109/SP.2015.35.

[11] Heng-Jui Chang, Shu-wen Yang, and Hung-yi Lee. DistilHuBERT: Speech representation learning by layer-wise distillation of hidden-unit BERT. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7087–7091. IEEE, 2022. doi: 10.1109/ICASSP43922.2022.9747490.

[12] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 646–650. IEEE, 2022. doi: 10.1109/ICASSP43922.2022.9746312. URL https://arxiv.org/abs/2202.00874.

[13] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1691–1703. PMLR, 2020. URL https://proceedings.mlr.press/v119/chen20s.html.

[14] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022. doi: 10.1109/JSTSP.2022.3188113. URL https://arxiv.org/abs/2110.13900.

[15] Sanyuan Chen, Yu Wu, Chengyi Wang, Zhengyang Chen, Zhuo Chen, Shujie Liu, Jian Wu, Yao Qian, Furu Wei, Jinyu Li, and Xiangzhan Yu. UniSpeech-SAT: Universal speech representation learning with speaker aware pre-training. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6152–6156. IEEE, 2022. doi: 10.1109/ICASSP43922.2022.9747077. URL https://arxiv.org/abs/2110.05752.

[16] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models, 2023. URL https://arxiv.org/abs/2311.07919.

[17] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. FMA: A dataset for music analysis. In *Proceedings of the 18th International Society for Music Information Retrieval Conference*, Suzhou, China, 2017. doi: 10.5281/zenodo.1414728.

[18] Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable LLM feature circuits. In *Advances in Neural Information Processing Systems*, volume 37, pages 24375–24410, 2024. doi: 10.52202/079017-0768.

[19] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

[20] 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. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

[21] Eric Elmoznino and Michael F. Bonner. High-performing neural network models of visual cortex benefit from high latent dimensionality. *PLOS Computational Biology*, 20(1):e1011792, 2024. doi: 10.1371/journal.pcbi.1011792.

[22] Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 55–65. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1006.

[23] Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2): 179–188, 1936. doi: 10.1111/j.1469-1809.1936.tb02137.x.

[24] Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1907.12009.

[25] Peiran Gao, Eric Trautmann, Byron Yu, Gopal Santhanam, Stephen Ryu, Krishna Shenoy, and Surya Ganguli. A theory of multineuronal dimensionality, dynamics and measurement. bioRxiv preprint 214262, 2017.

[26] Yuan Gong, Sameer Khurana, Leonid Karlinsky, and James R. Glass. Whisper-AT: Noise-robust automatic speech recognizers are also strong general audio event taggers. In *Proceedings of Interspeech*, pages 2798–2802, 2023. doi: 10.21437/Interspeech.2023-2193.

[27] Andrey Gromov, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Daniel A. Roberts. The unreasonable ineffectiveness of the deeper layers. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2403.17887.

[28] John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2733–2743, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275.

[29] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291. URL https://arxiv.org/abs/2106.07447.

[30] Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

[31] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

[32] Edith Law, Kris West, Michael I. Mandel, Mert Bay, and J. Stephen Downie. Evaluation of algorithms using games: The case of music tagging. In *Proceedings of the 10th International Society for Music Information Retrieval Conference*, pages 387–392, Kobe, Japan, 2009. doi: 10.5281/zenodo.1417647.

[33] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, pages 41451–41530, 2023. doi: 10.52202/075280-1797.

[34] Ashok Litwin-Kumar, Kameron Decker Harris, Richard Axel, Haim Sompolinsky, and L. F. Abbott. Optimal degrees of synaptic connectivity. *Neuron*, 93(5):1153–1164.e7, 2017. doi: 10.1016/j.neuron.2017.01.030.

[35] Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1112.

[36] Jiaqi Mu and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HkuGJ3kCb.

[37] Jayneel Parekh, Sanjeel Parekh, Pavlo Mozharovskyi, Florence d'Alché Buc, and Gaël Richard. Listen to interpret: Post-hoc interpretability for audio networks with NMF. In *Advances in Neural Information Processing Systems*, volume 35, pages 35270–35283, 2022. doi: 10.52202/068431-2556.

[38] Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. Layer-wise analysis of a self-supervised speech representation model. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, pages 914–921. IEEE, 2021. doi: 10.1109/ASRU51503.2021.9688093.

[39] Ankita Pasad, Bowen Shi, and Karen Livescu. Comparative layer-wise analysis of self-supervised speech models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE, 2023. doi: 10.1109/ICASSP49357.2023.10096149.

[40] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202.

[41] Karol J. Piczak. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 1015–1018. ACM, 2015. doi: 10.1145/2733373.2806390.

[42] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

[43] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems*, volume 34, pages 12116–12128, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/652cf38361a209088302ba2b8b7f51e0-Abstract.html.

[44] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.647.

[45] Stefano Recanatesi, Matthew Farrell, Madhu Advani, Timothy Moore, Guillaume Lajoie, and Eric Shea-Brown. Dimensionality compression and expansion in deep neural networks, 2019. URL https://arxiv.org/abs/1906.00443.

[46] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pages 1041–1044. ACM, 2014. doi: 10.1145/2647868.2655045.

[47] Joel Shor, Aren Jansen, Ronnie Maor, Oran Lang, Omry Tuval, Félix de Chaumont Quitry, Marco Tagliasacchi, Ira Shavitt, Dotan Emanuel, and Yinnon Haviv. Towards learning a universal non-semantic representation of speech. In *Proceedings of Interspeech*, pages 140–144, 2020. doi: 10.21437/Interspeech.2020-1242.

[48] Bob L. Sturm. The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use, 2013. URL https://arxiv.org/abs/1306.1461.

[49] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. SALMONN: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.13289.

[50] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

[51] Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1452.

[52] Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Björn W. Schuller, Christian J. Steinmetz, Colin Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, Max Henry, Nicolas Pinto, Camille Noufi, Christian Clough, Dorien Herremans, Eduardo Fonseca, Jesse Engel, Justin Salamon, Philippe Esling, Pranay Manocha, Shinji Watanabe, Zeyu Jin, and Yonatan Bisk. HEAR: Holistic evaluation of audio representations. In *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, volume 176 of *Proceedings of Machine Learning Research*, pages 125–145. PMLR, 2022. URL https://proceedings.mlr.press/v176/turian22a.html.

[53] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2023. URL https://arxiv.org/abs/2308.10248.

[54] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002. doi: 10.1109/TSA.2002.800560.

[55] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE, 2023. doi: 10.1109/ICASSP49357.2023.10095969. URL https://arxiv.org/abs/2211.06687.

[56] Ji Xin, Raphael Tang, Jaejun Lee, Yaoliang Yu, and Jimmy Lin. DeeBERT: Dynamic early exiting for accelerating BERT inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2246–2251. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.204.

[57] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. SUPERB: Speech processing universal performance benchmark. In *Proceedings of Interspeech*, pages 1194–1198, 2021. doi: 10.21437/Interspeech.2021-1775.

A Extended Experimental Details

A.1 Datasets and Splits

The benchmark suite was selected to stress different acoustic abstractions. ESC-50 tests compact environmental semantics; UrbanSound8K tests narrow urban events with strong local spectral-temporal signatures; GTZAN tests music genre and timbre/rhythm structure; and CREMA-D tests speech emotion, where prosody and speaker variation make late speech layers important. We use the fixed train/held-out partitions stored in the released artifact metadata, with no encoder fine-tuning. This makes every layer comparison a frozen-representation comparison rather than a downstream training-capacity comparison.

Table 11: Dataset summary used in all experiments.

Dataset	Task	Clips	Classes	Train	Test/held-out
ESC-50	Environmental sound	2000	50	1600	400
GTZAN	Music genre	999	10	799	200
CREMA-D	Speech emotion	7442	6	6136	1306
UrbanSound8K	Urban sound	8732	10	7079	1653

Table 12: Why each dataset is diagnostic for the study.

Dataset	Acoustic stress test	Main diagnostic signal	Interpretation note
ESC-50	Broad environmental semantics	CLAP forms compact class detectors; SSL models peak early	Compact taxonomy emphasizes layer-ordering and class-selective acoustic evidence.
UrbanSound8K	Narrow urban events	Highest transcoder monosemanticity across all encoders	Local event classes highlight narrow spectral-temporal routes.
GTZAN	Music genre and timbre/rhythm	CLAP and low/mid SSL layers retain useful music structure	Music labels combine instrumentation, rhythm, texture, and recording style.
CREMA-D	Speech affect and prosody	Whisper and HuBERT-family models often need late speech structure	Controlled acted speech isolates paralinguistic emotion cues.

A.2 Model Catalog

The model set is intentionally organized by pretraining recipe rather than by published aggregate rank. This is important because the paper's central claim is that objective family, data domain, and supervision source are strongly associated with useful depth, geometry, robustness, feature localization, and editability. Layer 0 denotes the first extracted hidden representation in the extraction pipeline. Whisper entries refer to encoder hidden states, not decoder states. CLAP-HTSAT exposes four hierarchical HTS-AT stages plus an input/stem representation, reported as indices 0-4. These stages do not align one-to-one with the 12-layer speech encoders; we therefore compare CLAP by extracted stage and empirical layer profiles, not by raw layer index.

Table 13: Studied encoders and analyzed depths.

Encoder	Objective family	Parameters	Width	Layers analyzed
Whisper-S	Weakly supervised ASR	244M	768	0-12
CLAP-HTSAT	Audio-text contrastive	86M	768	0-4 (stem+4 stages)
HuBERT-B	Offline-unit masked prediction	95M	768	0-12
WavLM-B	Masked prediction + denoising	95M	768	0-12
UniSpeech-SAT	Speaker-aware masked prediction	94M	768	0-12
wav2vec2-B	Contrastive masked prediction	95M	768	0-12
Data2Vec-Audio	Self-distillation	94M	768	0-12

A.3 Implementation Hyperparameter Ledger

The release stores effective run configurations with every generated CSV. The table below documents the settings used by the reported artifacts so that SAE feature budgets, monosemanticity ratios, transcoder faithfulness, and layer-selection gains can be reproduced from the documented configuration.

Table 14: Implementation hyperparameters from the released code and reported experiment configs.

Component	Settings
Datasets and splits	ESC-50 uses fold 5 as held-out test and the remaining folds for train; UrbanSound8K uses fold 10 as test and fold 9 as validation; GTZAN uses deterministic per-genre 80/10/10 splits so every split contains every genre; CREMA-D is grouped by actor id using deterministic 80/10/10 hash splits so speakers do not cross splits.
Audio preprocessing	Audio is loaded as mono, resampled to the checkpoint processor rate, and clipped according to dataset config: 5 s for ESC-50, UrbanSound8K, and CREMA-D; 30 s for GTZAN. Hidden states are extracted with 'output_hidden_states=True' and mean-pooled over time or time-frequency positions. Base encoders are always frozen.
Linear probes and metrics	Probes train on frozen representations extracted from each candidate layer. Accuracy is reported for ESC-50, UrbanSound8K, and GTZAN; macro-F1 is reported for CREMA-D. Layer selection has two reported modes: label-free geometry selection using normalized participation ratio and isotropy, and few-shot probe selection using shots $\{1, 5, 10, 20, 50\}$ per class. The main few-shot aggregate uses only the six common-depth encoders. FMA-small uses its official split; MagnaTagATune is evaluated on 4,332 test clips with macro-mAP.
SAE training	Top-k SAE with hidden multiplier 6, dictionary size 4608 for width-768 encoders, per-sample Top-k activation, $k_0 = 130$ annealed to $k = 75$ over the first 50% of steps, sparsity weight 5×10^{-5} with 10% warmup, 600 epochs, batch size 128, Adam learning rate 10^{-4}, cosine schedule with 5% warmup and 20% final decay, minimum learning-rate scale 0.05, decoder normalization, dead-feature revival window 75 with threshold 1, inference batch size 2048, seed 13. Activations are centered and standardized before SAE fitting.
Transcoder training	Sparse transcoder with hidden multiplier 5, Top-$k = 16$, 20 epochs, batch size 512, Adam learning rate 9×10^{-4}, sparsity weight 3×10^{-4}, seed 17. Metrics report reconstruction MSE, explained variance, active fraction, mean latent ℓ_1, and class monosemantic ratio.
Perturbations and interventions	Robustness sweeps use controlled additive noise, band-pass filtering, and frequency shift. Steering uses empirical class-mean differences with strengths $\{0.5, 1, 2, 4\}$. Feature ablation ranks class-associated SAE features by activation mass and reports target suppression, retain score, and off-target drift.

CLAP depth asymmetry. CLAP-HTSAT's extracted indices are four hierarchical audio-transformer stages plus an input/stem representation, not the first five layers of a 12-layer speech encoder. We therefore never use raw layer index as evidence that CLAP stage 3 is equivalent to speech layer 3, nor do we call stage 3 or 4 "early" within CLAP. The profile correlations below quantify the mismatch: CLAP's depth profiles have low average correlation with speech encoders, while masked/denoising SSL profiles are tightly aligned. Cross-encoder CLAP comparisons in the main text are functional stage comparisons.

Table 15: CLAP profile comparability. Values are mean probe-profile correlations across encoder pairs within each dataset.

Dataset	CLAP vs. speech encoders	Masked/denoising SSL pairs
CREMA-D	.036	.936
ESC-50	.130	.993
GTZAN	.095	.979
UrbanSound8K	.314	.988
Mean	.144	.974

Table 16: Objective-level hypotheses tested by the study.

Objective family	Expected representation pressure	Observed signature
Audio-text contrastive	Make broad semantic categories linearly accessible	Upper-stage CLAP peaks, small MSFS, strong clean transfer, but detector drift under shift.
Weak ASR supervision	Preserve decoder-useful speech content and prosody	Whisper peaks mid-to-late, steers late, and routes through faithful but polysemantic transcoder latents.
Masked or denoising SSL	Recover acoustic units while ignoring missing or corrupted frames	Early-to-middle useful layers and stronger robustness, especially for WavLM.
Self-distillation	Match teacher states without an explicit class or text target	Front-loaded Data2Vec profiles and large losses when extracting final layers for non-speech tasks.

A.4 Objective-Induced Subspaces

The claim that pretraining recipe is associated with useful depth can be made precise without assuming that the objective alone predicts every downstream task. Let t_e denote the target variable used by encoder e during pretraining: a masked discrete unit for HuBERT-like models, a contrastive latent for wav2vec2, a denoising target for WavLM, a decoder/transcript state for Whisper, a teacher state for Data2Vec, or a text-aligned semantic target for CLAP. At layer ℓ, define the target-mean subspace

$$\mathcal{S}_{e,\ell} = \operatorname{span}\left\{\mathbb{E}[h_e^{(\ell)} \mid t_e = t] - \mathbb{E}[h_e^{(\ell)}] \;:\; t \in \mathcal{T}_e\right\}, \tag{5}$$

with orthogonal projector $P_{e,\ell}$. For a downstream label y, the fraction of between-class scatter captured by this pretraining-induced subspace is

$$\alpha_{e,\ell}(y) = \frac{\operatorname{tr}\left(P_{e,\ell} S_B^{(\ell)} P_{e,\ell}\right)}{\operatorname{tr} S_B^{(\ell)}} \in [0,1]. \tag{6}$$

This α is the formal version of the paper's objective-alignment language: CLAP should have high α when labels are text-nameable sound categories; Whisper should have high α for speech-adjacent labels; and hard-unit SSL can have high α even when the full representation is anisotropic. In finite samples, if pretraining targets are available, $P_{e,\ell}$ can be estimated by forming centered target-conditional means and taking the top left singular vectors of that matrix, with shrinkage for rare targets. The present study does not estimate $P_{e,\ell}$ for the main claims; the projector is used to state sufficient conditions, while empirical regime labels come from measured probe/geometry correlations.

Proposition 1 (Objective-alignment lower bound for useful layers)**.** *Fix an encoder e, layer ℓ, and downstream label y. Let $P = P_{e,\ell}$ be the orthogonal projector onto the target-mean subspace $\mathcal{S}_{e,\ell}$ in Equation 5. Let $S_B^{(\ell)}$ and $S_W^{(\ell)}$ be the downstream between-class and within-class scatter matrices at that layer, with $\operatorname{tr} S_B^{(\ell)} > 0$. Define the objective-alignment coefficient*

$$\alpha_{e,\ell}(y) = \frac{\operatorname{tr}(P S_B^{(\ell)} P)}{\operatorname{tr} S_B^{(\ell)}}.$$

Assume that, on the projected subspace, within-class scatter is bounded by $P S_W^{(\ell)} P \preceq \sigma_\ell^2 P$. For any ridge parameter $\tau > 0$, define the projected ridge Fisher signal

$$\mathcal{F}_{P,\tau}^{(\ell)} = \operatorname{tr}\left[\left(P S_W^{(\ell)} P + \tau P\right)^{\dagger} P S_B^{(\ell)} P\right],$$

where the pseudoinverse is taken on $\operatorname{im}(P)$. Then

$$\mathcal{F}_{P,\tau}^{(\ell)} \geq \frac{\alpha_{e,\ell}(y)}{\sigma_\ell^2 + \tau} \operatorname{tr} S_B^{(\ell)}. \tag{7}$$

Consequently, among layers with comparable projected within-class scatter, larger alignment between downstream class scatter and the pretraining-induced subspace gives a larger certified lower bound on linear separability inside that subspace.

Proof. All matrices below are restricted to $\mathrm{im}(P)$, where P acts as the identity. Since $PS_W^{(\ell)}P \preceq \sigma_\ell^2 P$, adding τP gives

$$0 \prec PS_W^{(\ell)}P + \tau P \preceq (\sigma_\ell^2 + \tau)P.$$

For positive definite operators on the same subspace, Loewner order reverses under inversion, hence

$$\left(PS_W^{(\ell)}P + \tau P\right)^\dagger \succeq \frac{1}{\sigma_\ell^2 + \tau} P.$$

The matrix $PS_B^{(\ell)}P$ is positive semidefinite because $S_B^{(\ell)}$ is positive semidefinite and P is orthogonal. Taking the trace after multiplying by this positive semidefinite matrix preserves the inequality:

$$\mathcal{F}_{P,\tau}^{(\ell)} \geq \frac{1}{\sigma_\ell^2 + \tau} \operatorname{tr}\left(PS_B^{(\ell)}P\right).$$

Substituting the definition of $\alpha_{e,\ell}(y)$ yields Equation 7. □

The proposition does not say that pretraining objective alone determines performance. It states a sufficient condition: the objective creates candidate subspaces, and downstream labels are linearly easier when their between-class scatter lands in those subspaces while within-class scatter remains controlled. Empirical regime assignments are therefore checked from data rather than treated as theoretical guarantees.

A.5 SAE and Transcoder Objectives

Let $\mathcal{E} = \{e_\theta\}$ be a frozen audio encoder with layer-wise representations $\{h^{(\ell)}\}_{\ell=0}^{L}$, $h^{(\ell)} \in \mathbb{R}^{768}$. For each selected layer, the Top-k SAE decomposes activations as

$$f_\phi(h) = W_{\mathrm{dec}} \operatorname{TopK}(W_{\mathrm{enc}}h + b_{\mathrm{enc}}, k) + b_{\mathrm{dec}}, \tag{8}$$

where $W_{\mathrm{enc}} \in \mathbb{R}^{4608\times768}$ and $W_{\mathrm{dec}} \in \mathbb{R}^{768\times4608}$. We use a $6\times$ dictionary expansion and anneal sparsity from $k_0 = 130$ active features to $k = 75$. Here TopK denotes the nonnegative sparse-code operator: negative pre-activations are zeroed and the largest k positive activations are retained. The training objective combines reconstruction MSE with an ℓ_1 sparsity penalty. MSE is computed on the normalized activations used for SAE training, so the empirical reconstruction coefficient is

$$R^2_{\mathrm{SAE}} = 1 - \frac{\sum_i \|h_i - f_\phi(h_i)\|_2^2}{\sum_i \|h_i - \bar{h}\|_2^2}. \tag{9}$$

Across the SAE artifact table, reconstruction MSE has median 1.9×10^{-5} and mean 4.7×10^{-5}, giving near-unit reconstruction R^2 under this normalization; the median reconstructed-probe change is -0.018 score points, and the dead-feature fraction is below 2% in the stored best-layer SAE summaries. The $6\times$ dictionary and $k = 75$ final sparsity follow common Top-k SAE practice [30, 3], but we do not claim hyperparameter invariance. MSFS, monosemanticity, and genealogy claims should therefore be read as conditional on this dictionary size and sparsity schedule; the release includes configs for $3\times$ and $12\times$ sensitivity sweeps.

A feature f with nonzero total activation is class-monosemantic when a single label accounts for more than 70% of its activation mass:

$$\max_c \frac{\sum_i a_f(x_i)\mathbf{1}[y_i = c]}{\sum_i a_f(x_i)} > 0.70.$$

For each SAE feature we also measure activation rate and label entropy,

$$H^{(f)} = -\sum_{c=1}^{C} p(y = c \mid f \text{ active}) \log_2 p(y = c \mid f \text{ active}), \tag{10}$$

which separates rare class-selective detectors from frequent polysemantic features.

Transcoders predict layer $\ell + 1$ from layer ℓ through sparse latents:

$$T_\psi(h^{(\ell)}) = W_{\mathrm{out}} \operatorname{TopK}(W_{\mathrm{in}}h^{(\ell)} + b_{\mathrm{in}}, k_t) + b_{\mathrm{out}}, \tag{11}$$

with latent dimension 3840 and $k_t = 16$ for 12-layer encoders. CLAP-HTSAT uses the same active fraction over its shallower analysis depths. We score transcoder faithfulness by explained variance

$$R^2 = 1 - \frac{\sum_i \|h_i^{(\ell+1)} - T_\psi(h_i^{(\ell)})\|_2^2}{\sum_i \|h_i^{(\ell+1)} - \bar{h}^{(\ell+1)}\|_2^2}. \tag{12}$$

The best transition is the one maximizing R^2 subject to the highest monosemantic latent ratio. Full per-layer probe, SAE, transcoder, robustness, and feature-level records are stored in the supplemental CSV artifacts.

A.6 Steering and Ablation Protocols

For a binary class contrast at layer ℓ, the steering vector is the empirical class-mean difference

$$v^{(\ell)} = \frac{1}{n_1} \sum_{y_i=1} z_i^{(\ell)} - \frac{1}{n_0} \sum_{y_i=0} z_i^{(\ell)}. \tag{13}$$

We apply

$$\tilde{z}^{(\ell)}(x) = z^{(\ell)}(x) + \alpha v^{(\ell)}, \qquad \alpha \in \{0.5, 1.0, 2.0, 4.0\}, \tag{14}$$

and report selectivity as target-class gain minus off-target movement. For feature-level ablation, class-associated SAE features are ranked by activation mass and zeroed in code space before reconstruction; we report target suppression, retain-set score, and collateral drift.

For a steering strength α, the selectivity score used in the phase maps is

$$\mathrm{Sel}^{(\ell)}(\alpha) = \left[m_{\mathrm{target}}(\tilde{z}_\alpha^{(\ell)}) - m_{\mathrm{target}}(z^{(\ell)})\right] - \left[m_{\mathrm{retain}}(z^{(\ell)}) - m_{\mathrm{retain}}(\tilde{z}_\alpha^{(\ell)})\right], \tag{15}$$

where the first bracket is target-class gain and the second bracket is off-target movement. Positive selectivity means the intervention moves the target class more than it harms retain classes.

B Extended Related Work

Layer-wise audio analysis. Layer-wise probing has a long history in language and speech representation analysis [51, 38]. SUPERB made depth-dependent speech transfer especially visible by evaluating frozen speech encoders and supervised learned layer mixtures [57]; HEAR broadened frozen audio evaluation beyond speech [52]. Our study uses these results as a starting point rather than a point of contrast: depth dependence is not the novelty by itself. The contribution is to compare depth profiles across audio-text, ASR-supervised, masked, denoising, contrastive, speaker-aware, and self-distillation objectives, and then connect the selected readout to pruning, geometry, sparse features, routing, robustness, and interventions.

Layer mixtures, prefixes, and pruning. Learned layer weighting is a strong supervised readout because it can use evidence distributed across the full stack. Our operating point is different: when labels are absent, scarce, or when the downstream system wants to remove later blocks, a single selected layer or selected prefix is the actionable object. This makes single-layer selection complementary to full-depth mixtures rather than a replacement for them. The prefix mixture in Equation 4 is included for this reason: it preserves the supervised weighting mechanism while restricting computation and storage to layers up to the selected pruning boundary.

Objective-conditioned audio representations. Audio encoders differ not only in architecture but in the variable their pretraining objective makes predictable. wav2vec 2.0, HuBERT, WavLM, UniSpeech-SAT, and Data2Vec emphasize different speech SSL targets [4, 29, 14, 15, 5]; Whisper is trained for transcription [42]; and CLAP aligns audio with text [55]. The main text treats these targets as sources of different depth orderings and different sparse/routing signatures, while avoiding a causal claim that objective alone explains all differences among checkpoints.

Geometry, sparse features, and interventions. Representation geometry has been used to study anisotropy, transfer, and similarity across models [22, 31, 43]. We use participation ratio and isotropy as no-label layer priors, not as guarantees of downstream performance. Sparse interpretability work in language models develops SAEs, superposition analyses, monosemantic feature studies, and transcoders [19, 20, 30, 8, 50, 18]; audio-native work includes NMF-style explanations and AudioSAE [37, 3]. Our use of steering and ablation is diagnostic: it tests whether selected readouts expose approximately linear and sparse handles, while retaining the distinction between feature editing and training-data deletion [44, 7, 10].

C Extended Results

C.1 Layer-Wise Performance Details

The optimal transfer layer, the most selective intervention layer, and the most portable cross-dataset layer are associated with pretraining recipe in this panel. CLAP-HTSAT has the strongest clean peak on three tasks: ESC-50 0.970 at layer 4, UrbanSound8K 0.890 at layer 4, and GTZAN 0.825 at layer 3. Layer 0 is near-chance on the non-speech tasks because the HTS-AT patch projection has not yet formed semantic audio categories; the sharp layer-0-to-layer-4 rise is consistent with audio-text contrastive alignment.

Whisper-S has a monotone or plateauing profile, peaking at layers 6-12. CREMA-D reaches 0.732 macro-F1 at layer 12, reflecting that prosodic and emotional information aligns with later transcription-oriented abstractions. This late precision gives strong clean transfer and steering, while the perturbation results indicate that a representation organized for decoder cross-attention is not necessarily invariant to noise.

HuBERT-B, WavLM-B, and UniSpeech-SAT peak early on non-emotion tasks. The offline-unit masked-prediction objective is consistent with class-separable phonetic or spectro-temporal structure appearing by layers 1-4; deeper layers can reduce the linear accessibility of the boundaries needed by environmental and urban classification. WavLM retains more depth-wise transfer, consistent with denoising preserving acoustically rich features at depth. wav2vec2-B and Data2Vec-Audio show the largest last-layer losses: for wav2vec2-B on ESC-50, layer 12 gives 0.315 versus 0.693 at layer 2; Data2Vec is optimal at layer 0 on most transfer tasks.

C.2 Layer Selection as Pruning Optimization

Layer selection is also a pruning signal. Let m_ℓ be the downstream probe score at layer ℓ and let L be the final analyzed layer. For tolerance $\epsilon \geq 0$, define the earliest near-oracle exit

$$\ell^\star_\epsilon = \min \left\{ \ell : m_\ell \geq \max_{j \leq L} m_j - \epsilon \right\}. \tag{16}$$

When $\ell^\star_\epsilon < L$, layers after $\ell^\star_\epsilon$ become candidates for skipping in that frozen-extraction workload, or representations can be stored at that depth instead of at the final layer. This is especially relevant for audio systems that run continuously or combine multiple downstream heads. A single full-depth encoder may be acceptable for offline evaluation, but server-side monitoring, on-device assistants, music search, moderation, and acoustic sensing often require many hours of audio, multiple target taxonomies, or ensembles of complementary encoders. In those settings, selecting an earlier readout can define a task-specific pruning boundary after validation.

Pruning preservation. For any downstream task with layer scores $\{m_\ell\}_{\ell=0}^{L}$, extracting at $\ell^\star_\epsilon$ is within ϵ of the best layer and improves over final-layer extraction whenever $m_{\ell^\star_\epsilon} > m_L$. The analyzed depth fraction kept is at most $(\ell^\star_\epsilon + 1)/(L + 1)$.

This preservation statement treats accuracy alone. For resource-aware inference, the relevant quantity is utility after compute cost. Let C_ℓ be the cumulative cost of running layers $0, \ldots, \ell$, with $C_\ell < C_L$ for $\ell < L$, and define

$$U_\lambda(\ell) = m_\ell - \lambda C_\ell, \tag{17}$$

where $\lambda \geq 0$ expresses the value of latency, memory, or energy.

Table 17: Non-speech pruning evidence. Scores are averaged over ESC-50, GTZAN, and Urban-Sound8K. Gain is best-layer score minus final-layer score.

Encoder	Best layers	Max kept depth	Best score	Gain over final
CLAP-HTSAT	3,4,4	5/5	.895	+.007
Whisper-S	6,8,6	9/13	.818	+.025
HuBERT-B	3,2,1	4/13	.731	+.112
WavLM-B	3,1,3	4/13	.724	+.175
UniSpeech-SAT	2,2,4	5/13	.733	+.168
wav2vec2-B	2,2,0	3/13	.720	+.303
Data2Vec-Audio	2,0,0	3/13	.695	+.288

Cost-aware dominance. A pruned exit $\ell < L$ has higher utility than the final layer when

$$m_\ell - m_L > -\lambda(C_L - C_\ell). \tag{18}$$

In particular, if $m_\ell \geq m_L$, any positive compute weight $\lambda > 0$ makes the lower-cost pruned exit weakly preferable, and strictly preferable whenever $C_\ell < C_L$.

Risk-cost dominance. Let

$$V(\ell) = m_\ell - \lambda_c C_\ell - \lambda_D D_\ell - \lambda_F F_\ell + \lambda_A A_\ell(\mathcal{G}),$$

where D_ℓ is embedding drift, F_ℓ is alive-feature drift, and $A_\ell(\mathcal{G})$ is the descriptor-retention score of Equation 41. A pruned exit $\ell < L$ dominates the final layer under this utility iff

$$\begin{aligned} m_\ell - m_L + \lambda_c(C_L - C_\ell) + \lambda_D(D_L - D_\ell) + \lambda_F(F_L - F_\ell) \\ + \lambda_A(A_\ell(\mathcal{G}) - A_L(\mathcal{G})) > 0. \end{aligned} \tag{19}$$

Thus a lower-cost layer may still be rejected when it loses robustness or descriptor support, and a slightly lower-scoring layer may be accepted when cost and drift improvements compensate for the score gap.

These results formalize the pruning rows in Table 2: when the final layer is worse than an earlier layer, executing fewer analyzed blocks can preserve or improve the single-layer readout score; when the earlier layer is only slightly worse, it can still be preferred under a latency or memory budget. Actual throughput depends on the implementation, so the richer utility adds robustness and descriptor audits before accepting a pruned exit.

Table 17 summarizes the applied pruning evidence. CLAP-HTSAT uses a shorter stage-wise analysis pipeline, so its best non-speech stages sit near the top of that hierarchy. Whisper-S can be cut around layer 6 for environmental and urban sound, with GTZAN peaking slightly later. The SSL speech encoders are more dramatic: for wav2vec2-B and Data2Vec-Audio, the useful non-speech representation is in the first three analyzed depths, and the final layer removes roughly 29-30 points of linear task signal. This means pruned exits are not only an efficiency option; they can also improve accuracy when later layers are less aligned with the target single-layer readout. The ensemble implication is also practical: combining two pruned encoders can preserve complementary pretraining biases while reducing analyzed depth. The layer reports therefore support three resource-aware choices: truncate one encoder after validation, choose between encoders under a latency budget, or ensemble multiple truncated encoders when their error profiles are complementary. Main-text Table 9 reports measured latency and memory for one such truncation.

Prefix versus full-depth mixtures. Table 18 compares the prefix mixture of Equation 4 with the same learned mixture over all layers. Both use the same downstream head and five fixed seeds, and all values are five-seed means. The prefix ends at the selected exit (layer 5.75 of 12 on average for the common-depth encoders). FMA-small has three encoder cells. For wav2vec2-base on ESC-50, the prefix reaches 64.30% $\pm$ 0.21 against 63.95% $\pm$ 0.57 for full depth. The claim is that a prefix is frequently sufficient, not that it always dominates; CREMA-D is the main exception.

C.3 Cross-Encoder CKA Profile Details

The CKA profile matrix separates objective families more cleanly than any single downstream metric. HuBERT-B, UniSpeech-SAT, and WavLM-B have near-identical depth profiles: HuBERT-UniSpeech

Table 18: Selected-prefix versus full-depth learned layer mixtures (five-seed means). "Prefix $\geq$ full" counts encoder cells where the prefix mean matches or exceeds the full-depth mean.

Dataset	Cells	Prefix $\geq$ full	Prefix mean	Full mean	Prefix $-$ full
CREMA-D	6	3/6	.6348	.6487	$-$.0139
ESC-50	6	5/6	.6338	.6163	+.0174
FMA-small	3	2/3	.4731	.4742	$-$.0011
GTZAN	6	4/6	.6700	.6643	+.0057
UrbanSound8K	6	5/6	.7121	.7118	+.0002
All	27	19/27	–	–	+.0020

Table 19: Readout shortcuts and required audits. Each shortcut omits a different axis of the per-layer measurements: depth, perturbation response, sparse routing, or retain-set behavior.

Shortcut	Our evidence	Required audit
Use the final layer	wav2vec2-B gains .378 on ESC-50 and Data2Vec gains .365 by moving to an earlier readout.	Report layer profiles, last-layer gap, and pruning depth.
Rank by clean score	Whisper-S transfers strongly and also shows the largest controlled corruption drop (.280) and alive-feature drift (.407).	Pair clean accuracy with perturbation and feature-drift metrics.
Trust high R^2 alone	Whisper-S transcoders reach high explained variance with zero class-monosemantic ratio under our criterion.	Report transcoder faithfulness and monosemanticity separately.
Edit labeled features only	Sparse label alignment can move nearby classes when concepts share acoustic descriptors.	Report target suppression, retain score, and off-target drift.

$r = 0.99$, HuBERT-WavLM $r = 0.98$, and UniSpeech-WavLM $r = 0.99$. Data2Vec remains broadly aligned with this masked prediction family ($r = 0.83$ to 0.91), while wav2vec2-B is more distant ($r = 0.73$ to 0.84), consistent with online contrastive training.

CLAP-HTSAT is anti-correlated with every speech-focused encoder in the recovered profile matrix ($r = -0.44$ to -0.64 with SSL models and $r = -0.97$ with Whisper-S). This does not mean CLAP and Whisper are both merely "supervised" or "not supervised"; it means their supervision domains impose opposite depth orders. Whisper builds speech-decoder information late, while CLAP aligns broad audio semantics to language within its stage-wise HTS-AT hierarchy. The two models can therefore be strong transfer encoders with different shift responses.

C.4 Geometry Regimes

Let $\tilde{\lambda}_k^{(\ell)} = \lambda_k^{(\ell)} / \operatorname{tr} \hat{\Sigma}^{(\ell)}$ be normalized covariance eigenvalues. The participation ratio can be written as a Rényi entropy:

$$\mathcal{PR}^{(\ell)} = \frac{1}{\sum_{k=1}^{d} (\tilde{\lambda}_k^{(\ell)})^2} = \exp\Big(H_2(\tilde{\boldsymbol{\lambda}}^{(\ell)}) \Big). \tag{20}$$

Thus $\mathcal{PR}$ is a spectral diversity measure: it is d for a uniform eigenvalue spectrum and approaches 1 when variance collapses onto one direction. Isotropy is complementary rather than redundant. A layer can have high $\mathcal{PR}$ but low isotropy if representations share a nonzero mean direction, since the mean pairwise cosine is approximately

$$\frac{1}{\binom{N}{2}} \sum_{i<j} \frac{(z_i^{(\ell)})^\top z_j^{(\ell)}}{\|z_i^{(\ell)}\| \, \|z_j^{(\ell)}\|} \approx \frac{\|\bar{z}^{(\ell)}\|^2}{\sigma^2 + \|\bar{z}^{(\ell)}\|^2}. \tag{21}$$

This explains why isotropy is empirically the stronger global predictor of linear probe performance.

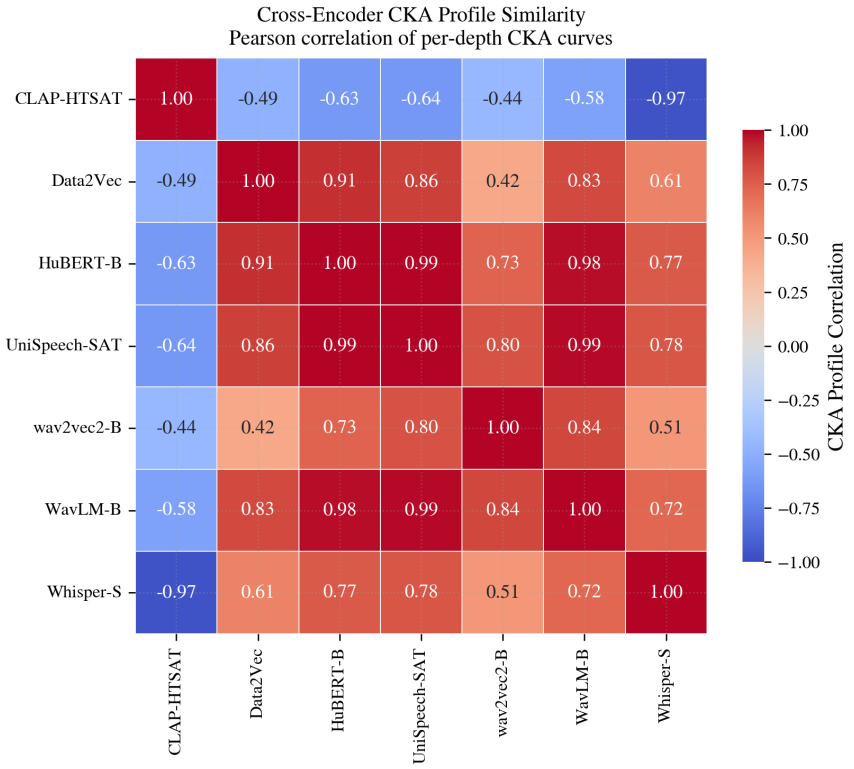

Figure 3: CKA depth-profile similarity across encoders. Each cell compares how probe-relevant information moves through depth rather than comparing one layer in isolation. Masked-prediction SSL encoders cluster tightly, Data2Vec remains near that family, wav2vec2-B is more distant because its contrastive target front-loads transfer, and CLAP-HTSAT follows a different audio-text depth ordering. Colour version of main-text Table 3.

Let $\mu_c^{(\ell)} = \mathbb{E}[z^{(\ell)} \mid y = c]$ and $\bar{\mu}^{(\ell)} = \sum_c \pi_c \mu_c^{(\ell)}$. Define

$$S_W^{(\ell)} = \sum_{c=1}^{C} \pi_c \, \mathbb{E}[(z^{(\ell)} - \mu_c^{(\ell)})(z^{(\ell)} - \mu_c^{(\ell)})^\top \mid y = c], \tag{22}$$

$$S_B^{(\ell)} = \sum_{c=1}^{C} \pi_c \, (\mu_c^{(\ell)} - \bar{\mu}^{(\ell)})(\mu_c^{(\ell)} - \bar{\mu}^{(\ell)})^\top. \tag{23}$$

The multiclass Fisher criterion is

$$\mathcal{F}^{(\ell)} = \operatorname{tr}\left[(S_W^{(\ell)})^{-1} S_B^{(\ell)} \right], \tag{24}$$

which measures class-discriminative signal-to-noise [23]. Proposition 1 is the algebraic result used by the paper: geometry helps when the high-spread or low-noise subspace is also the subspace that captures downstream between-class scatter. The tables below therefore treat geometry as a measurable prior, not as a proof of task performance.

Table 21 groups the encoders into three regimes by these correlations. The grouping is post hoc: regime labels are assigned from the observed correlations, and neither the geometry rule nor the probes use them. A practitioner applies Equation 3 directly to unlabeled target audio, with no regime assignment. The "objective cue" column records our hypothesis about why each objective might produce its regime; with seven encoders we cannot test that hypothesis, and we offer the table only as a description of the panel. The pooled correlations mix heterogeneous encoders and tasks; within the low-resource ASR setting where the rule is applied without labels, the correlations with ASR performance are higher ($r = .42$ for isotropy and .38 for participation ratio).

Table 20: Per-encoder Pearson correlations between geometry metrics and probe score across layers and datasets.

Regime	Encoder	$r_{\mathcal{PR}}$	$r_{\mathcal{I}}$
Strong positive	CLAP-HTSAT	$+.72$	$+.76$
Strong positive	Data2Vec	$+.65$	$+.50$
Strong positive	wav2vec2-B	$+.49$	$+.20$
Weak positive	WavLM-B	$+.20$	$-.04$
Weak positive	Whisper-S	$+.19$	$+.16$
Structured anisotropy	HuBERT-B	$-.03$	$-.40$
Structured anisotropy	UniSpeech-SAT	$-.02$	$-.31$
Global	All	$+.211$	$+.266$

Table 21: Post hoc grouping of encoders by geometry–score correlation. $r_{\mathcal{PR}}$ and $r_{\mathcal{I}}$ are correlations between probe score and participation ratio or isotropy over layers and datasets. The objective cue is a hypothesis, not a tested prediction. CLAP uses a stem plus four stages, so its correlations are descriptive rather than comparable to the 0–12 speech stacks.

Reg.	Objective cue	Observed assignment	Interpretation
R1	No hard unit bottleneck; broad contrastive, text-aligned, or teacher-state targets.	CLAP-HTSAT, Data2Vec, wav2vec2-B; $r_{\mathcal{PR}} = .49$ to $.72$, $r_{\mathcal{I}} = .20$ to $.76$.	Spread is a useful layer prior when class scatter uses many directions.
R2	Decoder-facing or denoising losses may make only some directions useful.	Whisper-S, WavLM-B; $r_{\mathcal{PR}} \approx .20$, $r_{\mathcal{I}} \approx .16/-.04$.	Global spread is less informative because utility is concentrated in ASR- or corruption-stable subspaces.
R3	Hard discrete units or speaker-aware targets can compress variance.	HuBERT-B, UniSpeech-SAT; $r_{\mathcal{I}} = -.40/-.31$.	Anisotropy can coexist with good probes if labels align with the low-rank phonetic or speaker subspace.

Table 22: Low-resource ASR: geometry-selected layer versus final layer, by character error rate (CER; lower is better). The geometry layer is chosen with no target-language transcripts as $\arg\max_\ell \frac{1}{2}(\widetilde{\mathcal{I}}_\ell + \widetilde{\mathcal{PR}}_\ell)$, the layer with highest average normalized isotropy and participation ratio. Relative CER reduction is $(\mathrm{CER}_{\mathrm{final}} - \mathrm{CER}_{\mathrm{geo}})/\mathrm{CER}_{\mathrm{final}}$; positive values are percentage reductions in CER; negative values mark boundary cases.

Language	Encoder	Geo. layer	Geo. CER	Final CER	Rel. gain
Amharic	wav2vec2-large	0	86.98	93.06	6.5%
Amharic	Data2Vec	0	47.06	89.68	47.5%
Welsh	wav2vec2-large	0	54.09	87.13	37.9%
Welsh	Data2Vec	2	43.16	43.79	1.4%
Hausa	wav2vec2-large	0	52.32	60.02	12.8%
Hausa	Data2Vec	10	42.10	37.09	-13.5%
Kyrgyz	wav2vec2-large	10	31.37	51.44	39.0%
Kyrgyz	Data2Vec	10	34.78	39.07	11.0%
Swahili	wav2vec2-large	0	51.38	80.42	36.1%
Swahili	Data2Vec	10	32.02	37.48	14.6%
Yoruba	wav2vec2-large	7	56.37	65.01	13.3%
Yoruba	Data2Vec	0	57.43	64.19	10.5%
Mean	wav2vec2-large	–	55.42	72.85	23.9%
Mean	Data2Vec	–	42.76	51.88	17.6%
Mean	All	–	49.09	62.37	21.3%

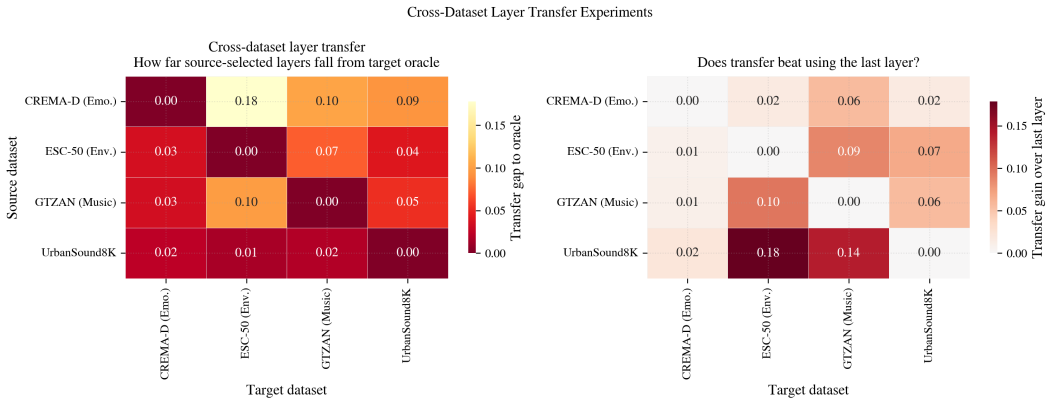

Figure 4: Cross-dataset layer transfer. Each source dataset selects a layer without seeing target labels, then that depth is evaluated on the remaining tasks. Source-selected layers often improve on last-layer extraction because early and middle layers retain reusable spectral-temporal structure; CREMA-D transfers more selectively because speech emotion pushes selection toward later paralinguistic layers.

Table 22 stress-tests the same geometry principle on sequence recognition. The result is strongest for wav2vec2-large: the geometry-selected layer improves CER in all six languages and is often layer 0, which means later blocks can be removed for those inference workloads. Across all rows, the geometry layer improves 11/12 settings and gives a 21.3% relative CER reduction on average; all 11 improvements have paired 95% utterance-bootstrap intervals (10,000 resamples) that exclude zero. A fixed layer 6 gives an aggregate CER of 51.27 (17.8% reduction), so geometry is better than the middle-layer default while selecting earlier layers (4.1 on average). Data2Vec is also positive on five languages; Hausa is the boundary row, with geometry CER 42.10 versus final-layer CER 37.09. Geometry supplies a label-free candidate exit, and a small labeled set can confirm it when available.

Discrete masked-prediction encoders explain why the geometry regimes are not a single monotone rule. If the target labels are offline k-means units, pretraining can encourage covariance of the form

$$\hat{\Sigma}^{(\ell)} \approx U_K \Lambda_K U_K^\top + \epsilon I_d, \tag{25}$$

where U_K spans the cluster-separating subspace. This can reduce isotropy, yet linear probes can still succeed when downstream between-class scatter is aligned with $\mathrm{col}(U_K)$. UniSpeech-SAT can add speaker-aware directions with the same effect. Regime 3 therefore shows why scalar geometry should be interpreted with the objective: the useful question is whether the task label aligns with the low-rank objective subspace. The main text therefore treats geometry as a label-free proposal and uses the few-shot probe rule when even a small labeled set is available.

C.5 Cross-Dataset Layer Transfer

UrbanSound8K is the most portable source for transfer because its selected layers sit early-to-middle, where general spectral-temporal structure is still available. CREMA-D is a more specialized source for environmental or music targets because emotion-optimized layers are often late and speech-specific. In the original transfer matrix, UrbanSound8K-selected layers improve ESC-50 and GTZAN over last-layer extraction without target labels, while CREMA-D-selected layers can incur large gaps when transferred to environmental sound. Early-to-middle layers are broadly acoustic, while late emotion or transcription layers are more task-specific.

C.6 Additional Music Benchmarks

GTZAN contains duplicated, mislabeled, and artist-overlapping excerpts [48], so we repeated the music analysis on FMA-small [17] (official split) and MagnaTagATune [32] (4,332 test clips, macro-mAP). Table 23 reports best-layer gains over the final layer with paired 95% bootstrap intervals. Both benchmarks reproduce the GTZAN pattern: speech SSL encoders expose useful music information well before their final layer, whereas CLAP peaks at its final analyzed stage on FMA-small.

Table 23: Music benchmarks beyond GTZAN. Gain is selected-layer score minus final-layer score (accuracy for FMA-small, macro-mAP for MagnaTagATune), with paired 95% bootstrap intervals.

Dataset	Encoder	Selected layer	Gain over final [95% CI]
FMA-small	wav2vec2-B	best layer	+.1075 [.0712, .1450]
FMA-small	Data2Vec	best layer	+.1025 [.0575, .1437]
FMA-small	CLAP-HTSAT	final stage	.0000
MagnaTagATune	wav2vec2-B	0 (validation-selected)	+.0484 [.0406, .0557]
MagnaTagATune	wav2vec2-B	6 (fixed middle)	+.0154 [.0081, .0221]

C.7 Layer-Selection Rules: Details

The label-free mode forms $g_\ell = \frac{1}{2}\widetilde{\mathcal{PR}}_\ell + \frac{1}{2}\widetilde{\mathcal{I}}_\ell$ and selects $\hat{\ell}_0 = \arg\max_\ell g_\ell$. This is the mode used for the low-resource ASR result, because it requires only unlabeled target-domain audio. The labeled mode fits a few-shot linear probe at every extracted layer and selects

$$\hat{\ell}_s = \arg\max_\ell \widehat{\mathrm{Acc}}_s^{(\ell)},$$

where $s \in \{1, 5, 10, 20, 50\}$ is the number of labeled examples per class. The two modes answer different questions: geometry proposes a layer before labels exist, while the probe identifies a layer once even a small validation set is available.

Learned zero-label selectors. Table 5 also reports two learned selectors under strict leave-one-encoder-out evaluation: for each held-out encoder, the selector is trained only on the other encoders and receives no labels, probe scores, or selected layer from the target encoder. The *encoder-summary* predictor maps summary statistics of an encoder's geometry profile to a depth; it beats the geometry rule in only 10 of 24 cells (mean gap 7.73 points). The *layer-profile* predictor is trained on the per-layer geometry of the other encoders (65 layer cells per dataset) and reaches a mean gap of 6.19 points, still above the fixed rule's 5.73. The remaining 5.73-point gap to the oracle leaves room for better learned selectors, but six encoders are too few to train one that generalizes.

The operational question is how much performance the selector recovers over the default final layer. Let $m_{d,e}^{(L_e)}$ be the score at the final available layer for dataset d and encoder e, and let $m_{d,e}^{(q_s)}$ be the score at the layer selected by method q with s labeled examples per class. The absolute improvement over final-layer extraction is

$$G_s(q) = \frac{1}{|\mathcal{D}||\mathcal{E}|} \sum_{d,e} \left(m_{d,e}^{(q_s)} - m_{d,e}^{(L_e)} \right), \tag{26}$$

with relative lift

$$\gamma_s(q) = \frac{G_s(q)}{\frac{1}{|\mathcal{D}||\mathcal{E}|} \sum_{d,e} m_{d,e}^{(L_e)}}. \tag{27}$$

The denominator is .611 in our aggregate, so the gain values are directly interpretable as recovered score over the extraction rule many users would otherwise use. A complementary recovered-gap fraction compares the selected layer with the full-data oracle,

$$R_s(q) = \frac{\sum_{d,e} \left(m_{d,e}^{(q_s)} - m_{d,e}^{(L_e)} \right)}{\sum_{d,e} \left(m_{d,e}^{(\star)} - m_{d,e}^{(L_e)} \right)}. \tag{28}$$

For the few-shot probe mode, $R_s \approx .85, .76, .85, .91, .92$ for one, five, ten, twenty, and fifty shots per class. Table 25 reports the underlying artifact diagnostics for the probe mode. These diagnostics are stricter than last-layer gain because they compare the selected layer with the full-data oracle layer for the same encoder and dataset. The main Table 6 avoids candidate-pool asymmetry by averaging only the six encoders with 13 analyzed depths. Appendix Tables 24 and 25 include the all-encoder artifact records for completeness; CLAP-HTSAT's stem-plus-four-stage pool has a different random Hit@1 baseline (0.20 instead of $1/13$), so candidate counts contextualize hit rates. This is also why we emphasize score gap and gain over final-layer extraction rather than exact Hit@1 alone.

Few-shot selector stability. If $|\widehat{a}_\ell - a_\ell| \le \eta$ for every layer, and $\hat{\ell} = \arg\max_\ell \widehat{a}_\ell$ while $\ell^\star = \arg\max_\ell a_\ell$, then

$$a_{\ell^\star} - a_{\hat{\ell}} \le 2\eta. \tag{29}$$

If, additionally, each few-shot estimate is σ^2/s-sub-Gaussian around a_ℓ and there are $L+1$ candidate layers, then with probability at least $1-\delta$,

$$a_{\ell^\star} - a_{\hat{\ell}} \le 2\sigma\sqrt{\frac{2\log(2(L+1)/\delta)}{s}}. \tag{30}$$

This bound follows from uniform concentration of the few-shot probe estimates and a union bound over candidate layers. It explains why exact H@1 is a strict diagnostic in the one-shot regime: if neighboring layers have similar true scores, a noisy probe may choose a different layer with little score change.

Table 25 reports gain and score gap together. Exact layer recovery can stay low when adjacent early-to-middle layers are statistically close, while the selected layer still recovers most of the available final-layer gap. ESC-50 is the cleanest case, with a zero score gap at 50 shots and a +.1932 gain over the final layer. GTZAN improves from a .0271 to a .0050 gap as shot count increases. CREMA-D is more specialized because its oracle depths are late and speech-specific, while UrbanSound8K has several useful neighboring layers. The probe rule is not meant to find a mathematically privileged depth; it recovers, at low cost, performance that is lost when final-layer extraction is used as the default.

Appendix Table 17 turns the numeric gains into pruning evidence, so we do not repeat a second use-case table here. For environmental, urban, and music-like targets, a strong default is to keep the layer where general acoustics are still linearly available. For paralinguistic speech, later layers become more plausible. For low-latency or memory-limited systems, the selected layer is also a pruning boundary: later transformer blocks can be skipped when they do not improve the target score.

The steering phase maps (main-text Figure 2) provide an intervention check on layer choice. In the stored sweeps, Whisper-S is most selective at a late layer (around layer 11, approximately 30 percentage points selectivity at the strongest steering strength), HuBERT-B peaks earlier (around layer 4, approximately 20 points), and wav2vec2-B peaks at an early layer (around layer 3, approximately 15 points). These maxima align with the probe/geometry optima, indicating that selected layers expose approximately linear handles for controlled perturbation of frozen representations. They should not be interpreted as reliable control policies. At the largest steering strengths, target gains of 20-30 percentage points can require meaningful off-target movement, which is why the main text treats steering as an intervention diagnostic and Figure 14 reports the target/off-target tradeoff.

C.8 Perturbation Analysis

Let f_ℓ be the encoder map to layer ℓ and let $\tilde{\delta}^{(\ell)} = f_\ell(x+\delta) - f_\ell(x)$ be the representation perturbation induced by input corruption. Under the structured covariance approximation of Equation (25), the task-relevant component is

$$\tilde{\delta}_\parallel^{(\ell)} = U_K^{(\ell)} (U_K^{(\ell)})^\top \tilde{\delta}^{(\ell)}. \tag{31}$$

Probe degradation is governed by $\|\tilde{\delta}_\parallel^{(\ell)}\|$, not by total perturbation energy. Later layers are often more robust because they compress task-relevant subspaces and encode invariances to low-level spectral noise. Embedding stability is measured as

$$\rho^{(\ell)} = \frac{1}{N}\sum_i \frac{f_\ell(x_i)^\top f_\ell(x_i+\delta_i)}{\|f_\ell(x_i)\|\,\|f_\ell(x_i+\delta_i)\|}. \tag{32}$$

When perturbations do not strongly change representation norms,

$$\rho^{(\ell)} \approx 1 - \frac{1}{2N}\sum_i \left\| \frac{f_\ell(x_i+\delta_i)}{\|f_\ell(x_i+\delta_i)\|} - \frac{f_\ell(x_i)}{\|f_\ell(x_i)\|} \right\|^2. \tag{33}$$

For SAE feature stability, let $A_i^{(\ell)} = \{f : a_f(x_i) > 0\}$ be the Top-k active feature set and $A_{i,\delta}^{(\ell)}$ the active set after corruption. Alive-feature drift is

$$F^{(\ell)} = \frac{1}{N}\sum_i \left(1 - \frac{|A_i^{(\ell)} \cap A_{i,\delta}^{(\ell)}|}{|A_i^{(\ell)} \cup A_{i,\delta}^{(\ell)}|}\right). \tag{34}$$

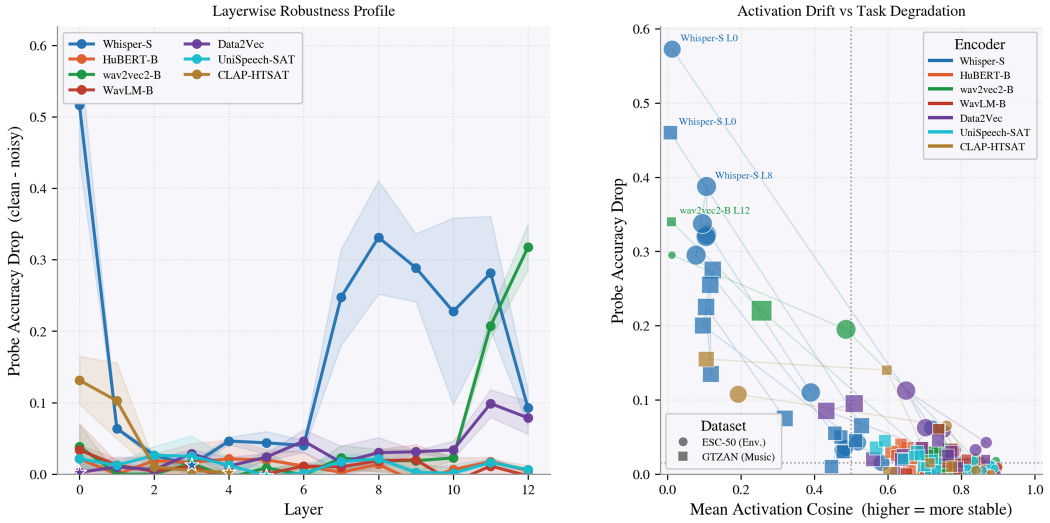

Figure 5: Per-encoder robustness under acoustic corruption, summarizing the clean-to-corrupted score drop, the cosine drift of layer outputs, and the feature drift of active SAE features (main-text Table 7). Whisper-S keeps globally similar outputs while the sparse features its probe uses change, giving the largest score drop. WavLM-B and the masked-prediction encoders preserve both score and active features more reliably.

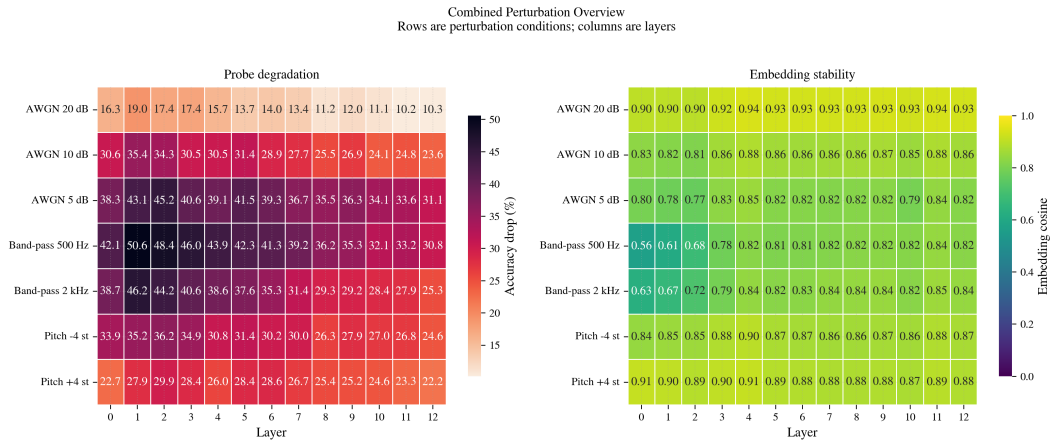

Figure 6: Controlled perturbation analysis across additive noise, band-pass filtering, and frequency shift. The panels separate score drop, representation drift, and sparse-feature drift, revealing structure that clean accuracy alone cannot see. Later layers are generally more stable to additive and spectral corruptions, but the clean optimum often occurs earlier, so robust application use requires choosing the layer that balances task score with perturbation stability.

It is therefore an active-set Jaccard drift, not a continuous activation-norm metric. For a linear probe with class weights w_c and biases b_c, define the clean margin of example i with true class y_i at layer ℓ as

$$\Gamma_i^{(\ell)} = \min_{c \neq y_i} \left[(w_{y_i} - w_c)^\top z_i^{(\ell)} + b_{y_i} - b_c \right]. \tag{35}$$

Projected drift and probe robustness. If $\Gamma_i^{(\ell)} > 0$ and the corruption-induced drift $\tilde{\delta}_i^{(\ell)}$ satisfies

$$\max_{c \neq y_i} \left| (w_{y_i} - w_c)^\top \tilde{\delta}_i^{(\ell)} \right| < \Gamma_i^{(\ell)}, \tag{36}$$

then the linear probe prediction for example i is unchanged. Moreover, if probe directions lie in a task subspace U and $w_{y_i} - w_c = UU^\top (w_{y_i} - w_c)$, then only the projected drift $UU^\top \tilde{\delta}_i^{(\ell)}$ affects the condition. This is simply the linear margin condition after adding the corrupted representation drift.

The perturbation plots follow the same subspace logic. A model can have small global cosine drift while predictions change if drift lands in the probe subspace, as in Whisper-S alive-feature drift. Conversely, masked and denoising SSL can tolerate large irrelevant perturbations when the projected drift is small.

The perturbation figure shows the expected interaction: band-pass filtering at 500 Hz produces the largest layer-robustness gain because it removes low-level spectral content, while frequency shifting produces smaller gains because harmonic structure is encoded across multiple depths. For perturbation type p, we summarize this interaction as a layer-robustness gain

$$\mathcal{G}_p = \Delta\,\mathrm{score}_p^{(0)} - \Delta\,\mathrm{score}_p^{(L)}, \tag{37}$$

where $\Delta\,\mathrm{score}_p^{(\ell)}$ is the clean-to-corrupted score drop at layer ℓ. The original artifact gives $\mathcal{G}_{500\mathrm{Hz}} \approx 11.3$ points for severe band-pass filtering, but only about 0.5 points for a $+4$ semitone pitch shift, matching the intuition that spectral filtering is mostly abstracted away with depth while pitch structure remains encoded across the hierarchy.

D Sparse Feature and Transcoder Appendix

The SAE analyses ask how task information is distributed over features. The feature-entropy landscape separates rare, low-entropy detectors from frequent polysemantic features. A low-rate feature with $H^{(f)} < 1$ bit behaves like a class-selective acoustic detector; a high-rate, high-entropy feature is reused across many labels and must be interpreted jointly with other features. This is why the paper reports both monosemanticity and feature-count interventions.

The minimum sufficient feature set at threshold τ is

$$\text{MSFS-}\tau = \min\{|S| : \mathrm{Acc}_{\mathrm{ablate}(S)} \leq (1 - \tau/100)\,\mathrm{Acc}\}. \tag{38}$$

Low MSFS means the task signal is concentrated in a few features; high MSFS means the same label information is redundant or distributed. CLAP-HTSAT has the lowest MSFS-50 (median of eight features; main-text Figure 1), consistent with narrow, label-aligned sparse features. Whisper-S and WavLM-B require broader sets, reflecting dense or redundant encodings.

Threshold sensitivity of transcoder and ablation contrasts. We also swept the dominant-label threshold for transcoder latents from 50% to 90%. CLAP has a larger class-concentrated fraction than Whisper-S at every threshold, falling from .889 to .611 for CLAP and from .812 to .188 for Whisper-S. The zero class-monosemantic ratio reported for Whisper-S in Tables 8 and 27 refers to its single best transition at the 70% criterion; the threshold-robust statement is the consistently higher concentration for CLAP. For feature ablation we evaluated MSFS-25, MSFS-50, MSFS-75, and MSFS-90 rather than only MSFS-50. All encoders reach the 25%, 50%, and 75% target drops within the evaluated ablation range, and none reaches 90%, so cross-encoder comparisons should be made at a fixed operating point between 25% and 75%.

Linear readout effect of SAE ablation. Let a linear probe logit be $g_c(h) = w_c^\top h + b_c$ and let the SAE reconstruction be $\hat{h} = W_{\mathrm{dec}} a + b_{\mathrm{dec}}$. If features S are ablated by setting $a_S = 0$, then the logit change satisfies

$$\left| g_c(\hat{h}) - g_c(\hat{h}_{\setminus S}) \right| \leq \left\| W_{\mathrm{dec},S}^\top w_c \right\|_2 \left\| a_S \right\|_2. \tag{39}$$

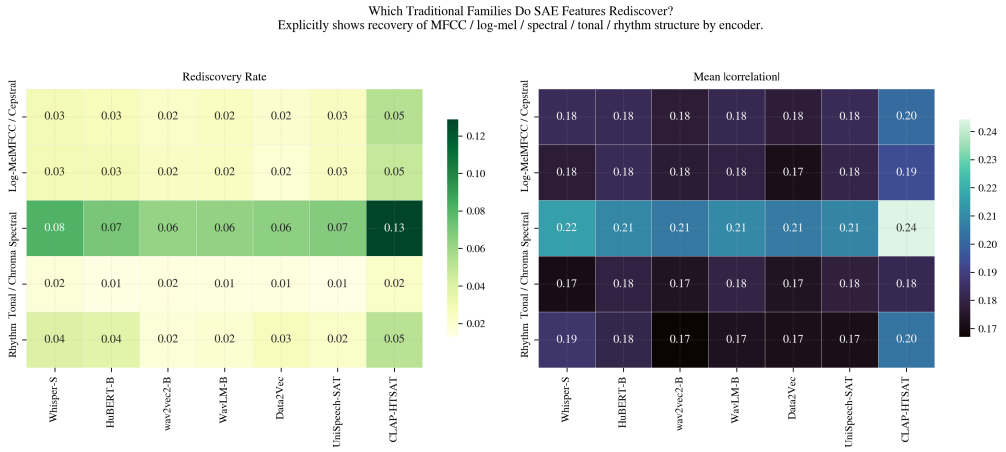

Figure 7: Alignment between SAE features and classical audio descriptor families. The rediscovery score compares SAE activations with MFCC/cepstral, mel, spectral, tonal/chroma, rhythm, and combined descriptors. Spectral descriptors are the most universally rediscovered family, while CLAP shows the strongest overall bridge from learned semantic features back to traditional acoustic measurements.

This is the Cauchy-Schwarz bound for the decoder contribution removed by ablation.

This explains why monosemanticity alone is insufficient: a feature must be both class-selective and aligned with the probe readout direction to be a strong intervention handle. MSFS combines these ingredients empirically by asking how many ranked features must be removed before the target score drops.

Traditional-feature rediscovery compares SAE activations with MFCC/cepstral, log-mel, spectral, tonal/chroma, rhythm, and combined descriptor families. For a traditional family $\mathcal{T}$ and SAE feature f, we compute the maximum absolute Pearson correlation between $a_f(x_i)$ and any descriptor in $\mathcal{T}$. Spectral descriptors are the most universally rediscovered family, while CLAP-HTSAT shows high bridge scores across the measured families. These bridge scores connect CLAP's high clean accuracy with compact label-aligned detectors grounded in familiar acoustic primitives. The feature-level bridge and the layer-level descriptor-retention audit used in the main pruning objective are

$$B_f(\mathcal{T}) = \max_{g \in \mathcal{T}} \left| \mathrm{Corr}(a_f(x_i), g(x_i)) \right|, \tag{40}$$

$$A_\ell(\mathcal{G}) = \frac{1}{|\mathcal{G}|} \sum_{\mathcal{T} \in \mathcal{G}} R^2_{\mathrm{ridge}} \left(h_i^{(\ell)} \to \phi_{\mathcal{T}}(x_i) \right), \tag{41}$$

where $\phi_{\mathcal{T}}$ is the descriptor vector for family $\mathcal{T}$. B_f asks whether an individual sparse feature has a classical acoustic interpretation; A_ℓ asks whether an extraction layer still retains the descriptor family required by the target application.

The full table clarifies the distinction between faithfulness and interpretability. Whisper-S has near-perfect R^2 on multiple tasks, but zero class-monosemantic routing under the current criterion and the largest mean latent ℓ_1 values. Its layer-to-layer computation is predictable but dense. UrbanSound8K produces the highest monosemantic ratios for every encoder because its urban categories occupy narrow acoustic regions; CLAP-HTSAT reaches .7739 on US8K but collapses to zero on CREMA-D and GTZAN, revealing task-conditional sparse routing. The lower-R^2 best transitions do not simply coincide with low monosemanticity: ESC-50 CLAP has $R^2 = .8667$ and Mono=.0263, while GTZAN Data2Vec has $R^2 = .8586$ and Mono=.0139; across all stored transitions the correlation between explained variance and monosemantic ratio is only .123. Faithfulness and interpretability are therefore separate axes, not one metric in disguise.

Transcoder error and probe-logit error. Let $T_\psi(h^{(\ell)})$ approximate $h^{(\ell+1)}$ and let a downstream logit at layer $\ell + 1$ be $g_c(h) = w_c^\top h + b_c$. For each example,

$$\left| g_c(h^{(\ell+1)}) - g_c(T_\psi(h^{(\ell)})) \right| \leq \|w_c\|_2 \, \|h^{(\ell+1)} - T_\psi(h^{(\ell)})\|_2. \tag{42}$$

The bias cancels, leaving a Cauchy-Schwarz bound between probe weight and transcoder residual.

High R^2 therefore bounds the residual that can perturb linear readout information. It does not certify that the route is monosemantic. Tables 27 and 29 therefore report separate axes: R^2 measures faithfulness, while the monosemantic ratio measures whether the faithful route decomposes into class-readable pieces.

E Editing and Targeted Feature Ablation

We evaluate representation editing in two forms. Latent steering adds a scaled class-difference vector to layer representations and measures target-class gain minus off-target drop. SAE feature ablation removes high-scoring class-associated features and measures target suppression and retain-set movement. The experiments support a diagnostic interpretation: sparse monosemantic features are useful handles for concept-level intervention, but these are not machine-unlearning experiments and they do not remove training examples from a model. For safety-sensitive settings, feature editing should be reported with retain-set accuracy, off-target drift, and robustness under perturbation.

For an ablation set S, we compute a target suppression score

$$\Delta_{\text{target}}(S) = \text{score}_{\text{clean,target}} - \text{score}_{\text{ablate}(S),\text{target}}$$

and a retain-set movement score over non-target labels,

$$\Delta_{\text{retain}}(S) = \text{score}_{\text{clean,retain}} - \text{score}_{\text{ablate}(S),\text{retain}}\,.$$

A useful edit has high Δ_{target} and low Δ_{retain}. Monosemantic features generally occupy this regime; distributed encodings require larger S and move toward higher collateral movement. The plots diagnose feature localization rather than formal guarantees of deletion.

The linearized effect of feature ablation makes the target-retain tradeoff explicit. Let an SAE reconstruction be $\hat{z} = b + \sum_f a_f d_f$, where d_f is decoder feature f, and let a linear probe logit be $g_c(z) = w_c^\top z + b_c$. Ablating S changes a class margin by

$$\Delta M_{y,c}(S) = -\sum_{f \in S} a_f (w_y - w_c)^\top d_f. \tag{43}$$

Selective ablation condition. Suppose the ablated SAE activations are nonnegative and features in S are target-aligned for class y in the sense that $(w_y - w_c)^\top d_f \geq \beta_f > 0$ for the target contrast, and are retain-neutral for non-target contrasts in the sense that $|(w_r - w_{c'})^\top d_f| \leq \kappa_f$. Then target margin reduction is at least

$$\sum_{f \in S} a_f \beta_f, \tag{44}$$

while any retain margin changes by at most

$$\sum_{f \in S} a_f \kappa_f. \tag{45}$$

Thus edits are selective when target alignment is large and retain alignment is small.

This is the mathematical reason MSFS, monosemanticity, and retain-set movement plots must be reported together. A feature can be sparse but not target-aligned, or target-aligned but also used by retain classes. Good editing requires both conditions.

F Additional Figures

This section collects visual diagnostics that are too detailed for the main paper but important for interpreting the study. Captions state the intended reading; the quantitative genealogy summary appears in Table 30.

The genealogy plots in Figures 17 through 23 trace sparse features through depth. They are feature-lineage diagrams rather than causal explanations by themselves: a thick branch indicates that many features at the next layer inherit activation mass from a previous feature, while merges indicate that

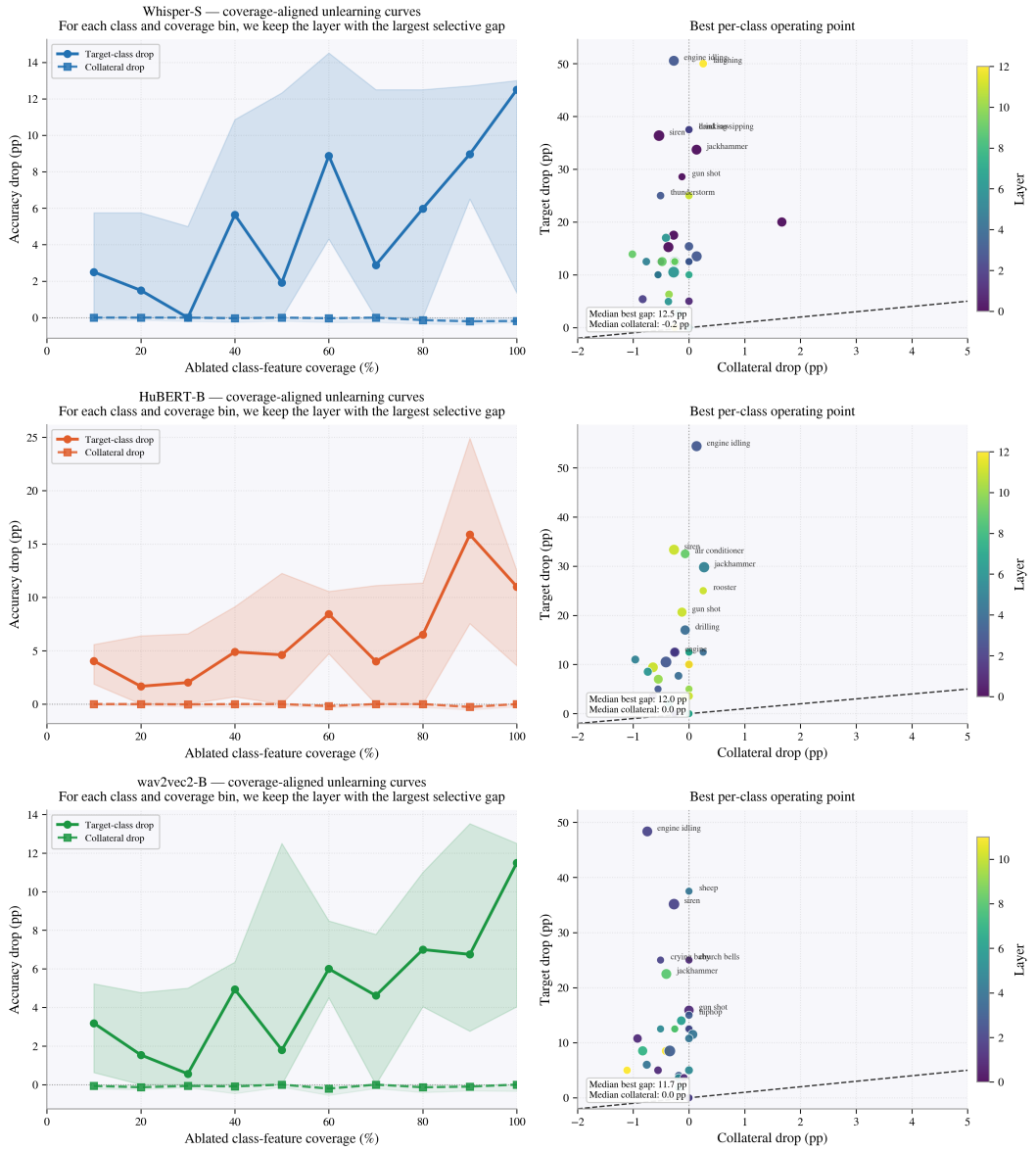

Figure 8: Top-k probing and targeted feature-ablation diagnostics. The plot tracks how probe performance changes as increasingly many target-associated sparse features are kept or removed. A steep target drop with limited retain-set movement indicates a localized concept handle; a shallow or noisy curve indicates that the concept is distributed across polysemantic features.

Table 24: Full label-free geometry-rule results by dataset and encoder (single run). The selector uses no labels, transcripts, or validation targets: it chooses the layer with the highest normalized participation-ratio/isotropy score. Gain is selected-layer score minus final-layer score; gap is full-data-oracle score minus selected-layer score; exact equals one only when the label-free layer is also the supervised oracle layer. The table shows why geometry is useful before labels exist, but also why it should be audited on speech-affect tasks such as CREMA-D.

Dataset	Encoder	Geo L	Oracle L	Final L	Geo score	Oracle	Final	Gain	Gap	Exact
CREMA-D	CLAP-HTSAT	1	3	4	.520	.604	.586	-.066	.084	0
CREMA-D	Data2Vec	12	0	12	.553	.673	.553	.000	.121	0
CREMA-D	HuBERT-B	1	12	12	.612	.676	.676	-.065	.064	0
CREMA-D	UniSpeech-SAT	1	12	12	.611	.681	.681	-.070	.070	0
CREMA-D	wav2vec2-B	11	2	12	.553	.647	.514	+.039	.093	0
CREMA-D	WavLM-B	1	6	12	.646	.699	.660	-.014	.053	0
CREMA-D	Whisper-S	4	12	12	.682	.732	.732	-.050	.050	0
ESC-50	CLAP-HTSAT	3	4	4	.932	.970	.970	-.038	.038	0
ESC-50	Data2Vec	7	2	12	.442	.665	.300	+.143	.222	0
ESC-50	HuBERT-B	7	3	12	.662	.715	.585	+.078	.052	0
ESC-50	UniSpeech-SAT	5	2	12	.705	.720	.500	+.205	.015	0
ESC-50	wav2vec2-B	6	2	12	.535	.692	.315	+.220	.158	0
ESC-50	WavLM-B	4	3	12	.692	.707	.462	+.230	.015	0
ESC-50	Whisper-S	6	6	12	.860	.860	.845	+.015	.000	1
GTZAN	CLAP-HTSAT	2	3	4	.740	.825	.805	-.065	.085	0
GTZAN	Data2Vec	0	0	12	.690	.690	.415	+.275	.000	1
GTZAN	HuBERT-B	7	2	12	.655	.700	.575	+.080	.045	0
GTZAN	UniSpeech-SAT	5	2	12	.655	.695	.515	+.140	.040	0
GTZAN	wav2vec2-B	7	2	12	.585	.715	.430	+.155	.130	0
GTZAN	WavLM-B	8	1	12	.620	.690	.510	+.110	.070	0
GTZAN	Whisper-S	3	7	12	.735	.745	.695	+.040	.010	0
US8K	CLAP-HTSAT	3	4	4	.881	.890	.890	-.009	.009	0
US8K	Data2Vec	0	0	12	.730	.730	.506	+.224	.000	1
US8K	HuBERT-B	3	1	12	.763	.778	.697	+.065	.015	0
US8K	UniSpeech-SAT	3	4	12	.756	.785	.681	+.076	.029	0
US8K	wav2vec2-B	2	0	12	.721	.751	.505	+.216	.030	0
US8K	WavLM-B	4	3	12	.768	.774	.674	+.094	.007	0
US8K	Whisper-S	4	6	12	.842	.850	.839	+.003	.008	0

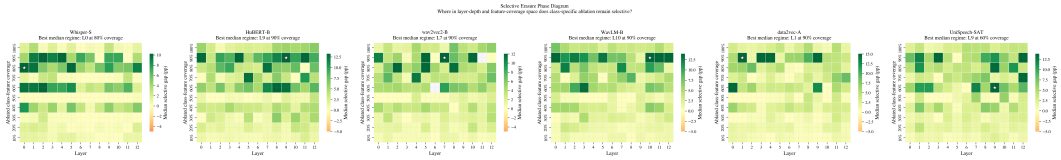

Figure 9: Feature-ablation phase diagram over target suppression and retain-set interference. Points in the desirable region suppress the target class while preserving off-target classes; points with high retain-set movement show why feature ablation is a diagnostic intervention rather than a machine-unlearning or deletion guarantee.

Table 25: Few-shot probe rule, full dataset-wise results including CLAP-HTSAT. Hit@1 is exact oracle-layer recovery. Gain is selected-layer score minus final-layer score; score gap is the absolute metric deficit to the oracle. Candidate count is below 13 because this appendix table includes CLAP-HTSAT, which exposes four hierarchical stages plus an input/stem representation while the other encoders expose layers 0-12.

Dataset	Shots	Hit@1	Gain	Score gap	Cand.
CREMA-D	1	.1429	+.0219	.0224	11.8571
CREMA-D	5	.1429	+.0019	.0424	11.8571
CREMA-D	10	.1429	+.0108	.0334	11.8571
CREMA-D	20	.1429	+.0272	.0171	11.8571
CREMA-D	50	.1429	+.0236	.0206	11.8571
ESC-50	1	.2857	+.1796	.0136	11.8571
ESC-50	5	.4286	+.1829	.0104	11.8571
ESC-50	10	.4286	+.1864	.0068	11.8571
ESC-50	20	.7143	+.1900	.0032	11.8571
ESC-50	50	1.0000	+.1932	.0000	11.8571
GTZAN	1	.4286	+.1321	.0271	11.8571
GTZAN	5	.0000	+.1050	.0543	11.8571
GTZAN	10	.1429	+.1386	.0207	11.8571
GTZAN	20	.5714	+.1421	.0171	11.8571
GTZAN	50	.7143	+.1543	.0050	11.8571
UrbanSound8K	1	.4286	+.0974	.0121	11.8571
UrbanSound8K	5	.4286	+.0923	.0172	11.8571
UrbanSound8K	10	.2857	+.0923	.0172	11.8571
UrbanSound8K	20	.2857	+.0985	.0110	11.8571
UrbanSound8K	50	.4286	+.0898	.0197	11.8571

Table 26: Sensitivity of SAE class-monosemantic feature ratios to the dominant-label threshold. Ratios are computed on each dataset's selected SAE layer and then pooled by encoder. CLAP remains the most concentrated across thresholds, so the qualitative sparse-feature claim is not tied to the 70% cutoff.

Encoder	Share $\geq 60\%$	Share $\geq 70\%$	Share $\geq 80\%$	Mean dominant share
CLAP-HTSAT	.300	.188	.100	.513
Data2Vec-Audio	.255	.156	.075	.485
WavLM-B	.232	.143	.072	.475
Whisper-S	.227	.136	.060	.474
HuBERT-B	.226	.135	.065	.470
UniSpeech-SAT	.208	.119	.055	.463
wav2vec2-B	.193	.110	.050	.456

Table 27: Best-transcoder monosemantic ratio by dataset and encoder. UrbanSound8K elicits the highest monosemantic routing for every encoder.

Encoder	CREMA-D	ESC-50	GTZAN	US8K
CLAP-HTSAT	.000	.026	.000	**.774**
UniSpeech-SAT	.115	.057	.015	**.471**
HuBERT-B	.190	.057	.000	**.321**
WavLM-B	.086	.089	.011	**.296**
wav2vec2-B	.050	.093	.000	**.152**
Data2Vec	.021	.000	.014	**.190**
Whisper-S	.000	.000	.000	.000

Table 28: Best-transcoder transition by encoder and dataset.

Encoder	CREMA-D	ESC-50	GTZAN	US8K
CLAP-HTSAT	$3 \to 4$	$0 \to 1$	$0 \to 1$	$3 \to 4$
Whisper-S	$10 \to 11$	$10 \to 11$	$2 \to 3$	$10 \to 11$
HuBERT-B	$11 \to 12$	$0 \to 1$	$0 \to 1$	$0 \to 1$
UniSpeech-SAT	$1 \to 2$	$0 \to 1$	$0 \to 1$	$2 \to 3$
WavLM-B	$11 \to 12$	$0 \to 1$	$0 \to 1$	$0 \to 1$
wav2vec2-B	$10 \to 11$	$10 \to 11$	$11 \to 12$	$10 \to 11$
Data2Vec	$11 \to 12$	$11 \to 12$	$6 \to 7$	$11 \to 12$

Table 29: Full best-transcoder metrics. R^2 is explained variance of the next layer from sparse latents; Mono is the fraction of active latents whose activation mass is dominated by one class.

Dataset	Encoder	Src	Tgt	Recon. MSE	R^2	Mean ℓ_1	Mono
CREMA-D	CLAP-HTSAT	3	4	.0746	.9832	.0829	.0000
CREMA-D	Data2Vec	11	12	.0002	.9802	.0085	.0213
CREMA-D	HuBERT-B	11	12	.0021	.9470	.0078	.1899
CREMA-D	UniSpeech-SAT	1	2	.0009	.9402	.0057	.1154
CREMA-D	wav2vec2-B	10	11	.0090	.9940	.0263	.0500
CREMA-D	WavLM-B	11	12	.0017	.9409	.0047	.0855
CREMA-D	Whisper-S	10	11	.0054	.9978	.1080	.0000
ESC-50	CLAP-HTSAT	0	1	.0061	.8667	.0173	.0263
ESC-50	Data2Vec	11	12	.0009	.9442	.0107	.0000
ESC-50	HuBERT-B	0	1	.0018	.9059	.0038	.0571
ESC-50	UniSpeech-SAT	0	1	.0020	.8999	.0047	.0565
ESC-50	wav2vec2-B	10	11	.0224	.9589	.0195	.0932
ESC-50	WavLM-B	0	1	.0017	.8847	.0040	.0886
ESC-50	Whisper-S	10	11	.0381	.9895	.1265	.0000
GTZAN	CLAP-HTSAT	0	1	.0042	.8896	.0167	.0000
GTZAN	Data2Vec	6	7	.0027	.8586	.0054	.0139
GTZAN	HuBERT-B	0	1	.0015	.9079	.0036	.0000
GTZAN	UniSpeech-SAT	0	1	.0020	.8804	.0046	.0154
GTZAN	wav2vec2-B	11	12	.0015	.9354	.0067	.0000
GTZAN	WavLM-B	0	1	.0016	.8651	.0037	.0112
GTZAN	Whisper-S	2	3	.0037	.9326	.0138	.0000
US8K	CLAP-HTSAT	3	4	.1551	.9746	.0798	.7739
US8K	Data2Vec	11	12	.0005	.9830	.0140	.1905
US8K	HuBERT-B	0	1	.0010	.9568	.0041	.3214
US8K	UniSpeech-SAT	2	3	.0013	.9502	.0064	.4714
US8K	wav2vec2-B	10	11	.0089	.9740	.0141	.1524
US8K	WavLM-B	0	1	.0011	.9410	.0040	.2958
US8K	Whisper-S	10	11	.0172	.9955	.1311	.0000

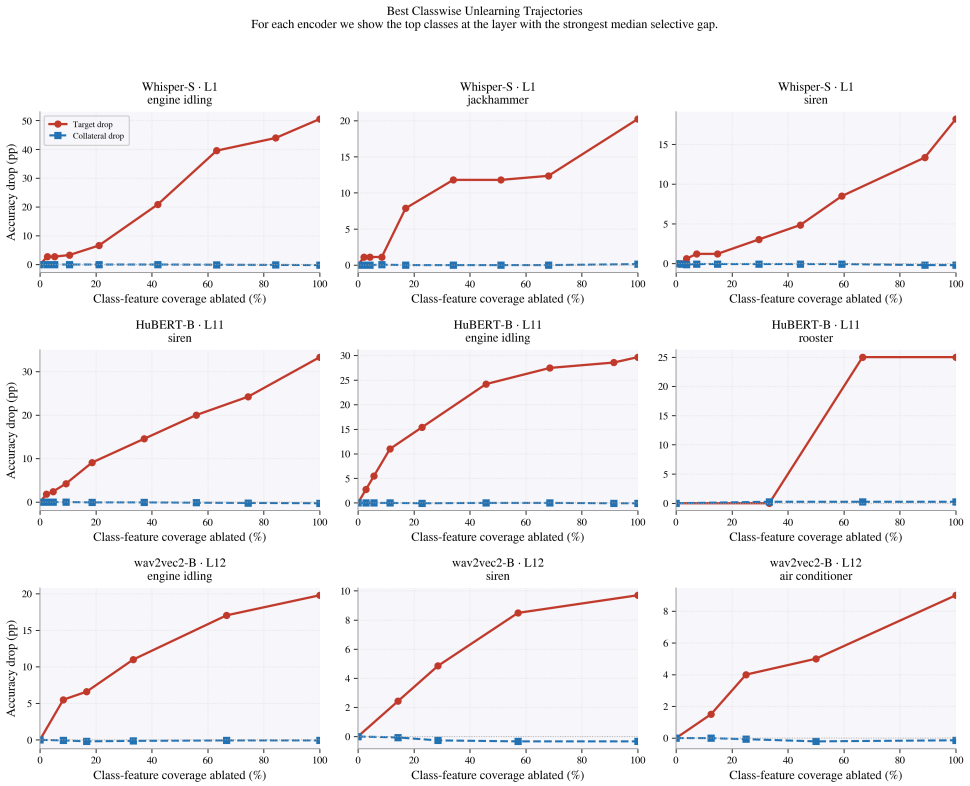

Figure 10: Top class-level ablation examples. The grid highlights classes whose SAE features are most removable under the target-suppression criterion and classes where removal spills into acoustically related labels. These examples connect the aggregate phase diagram to concrete sound categories.

a later feature combines earlier acoustic or semantic factors. These plots explain why the same monosemanticity threshold can mean different things for different objectives.

We quantify the visual patterns with three lineage statistics. For SAE feature f at layer ℓ and feature g at layer $\ell+1$, let $\mathrm{IoU}(f,g)$ be the intersection-over-union of their top-activation example sets. The parent of g is

$$p(g) = \arg\max_f \mathrm{IoU}(f,g).$$

Lineage persistence, branching, and merging can then be summarized as

$$\mathrm{Persist}_{\ell\to\ell+1} = \frac{1}{|\mathcal{F}_{\ell+1}|} \sum_{g\in\mathcal{F}_{\ell+1}} \max_f \mathrm{IoU}(f,g), \tag{46}$$

$$\mathrm{Branch}(f) = \left|\{g : p(g) = f,\ \mathrm{IoU}(f,g) > \tau\}\right|, \tag{47}$$

$$\mathrm{Merge}(g) = \left|\{f : \mathrm{IoU}(f,g) > \tau\}\right|. \tag{48}$$

The event table below is computed for the common-depth encoders with adjacent 0-12 transitions. CLAP is visualized separately because its hierarchical stages do not yield the same event records as the speech-style stacks.

Whisper-S has almost no exact persistence and high split rate, indicating that late ASR layers keep redistributing sparse evidence rather than preserving stable class detectors. UniSpeech-SAT has the highest merge rate, consistent with speaker-aware objectives combining content and speaker factors. CLAP's visual stage genealogy remains compact, matching its small MSFS and strong clean transfer, but we treat that claim as stage-level qualitative evidence rather than the common-depth event statistic above.

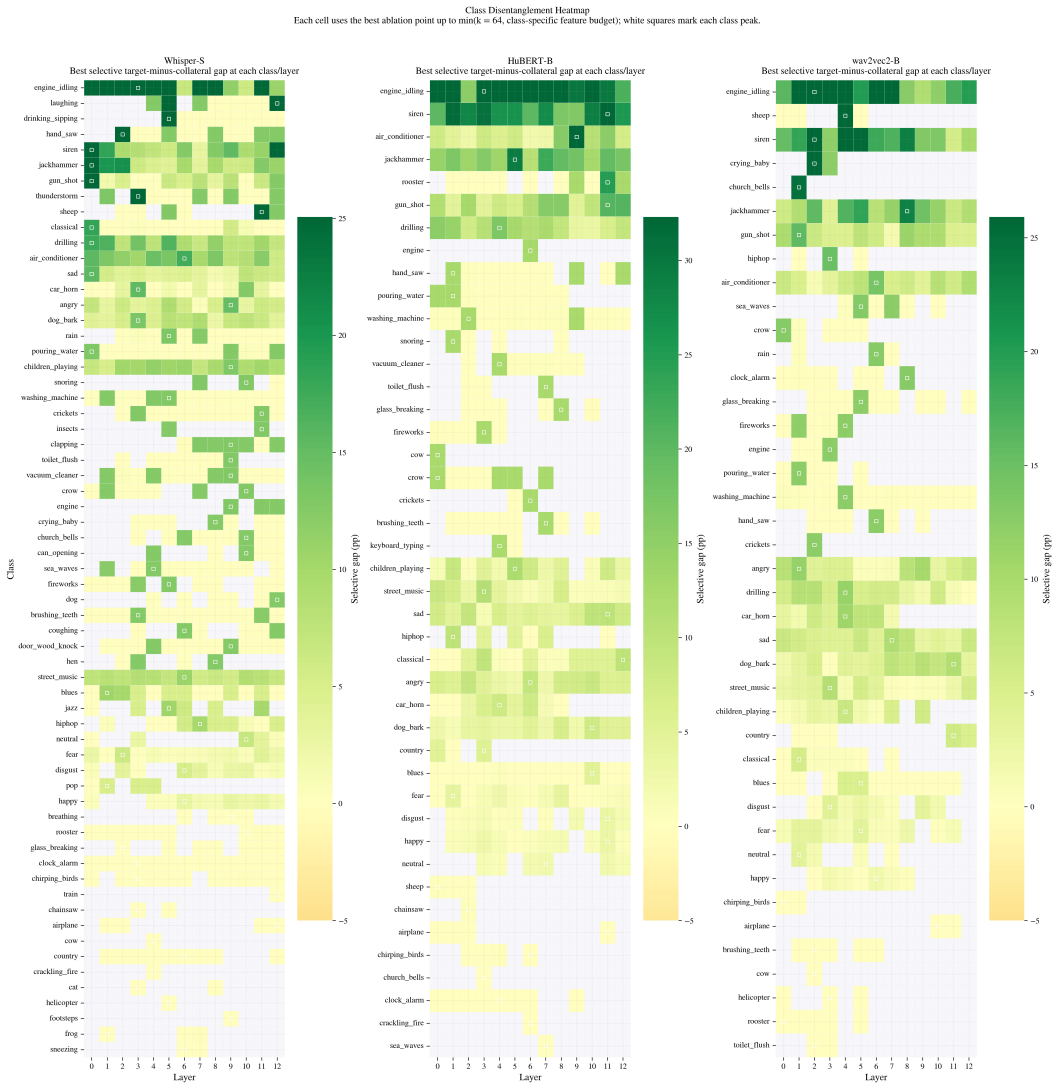

Figure 11: Class disentanglement heatmap across model layers and classes. The heatmap decomposes aggregate probe performance into per-class separability, revealing classes that become clean intervention targets and classes that stay entangled even when the dataset-level score is high.

Table 30: Sparse-feature genealogy event rates for the six common-depth encoders. Rows aggregate the available dataset/encoder lineage records. High split rate means one feature fans out into several descendants; high merge rate means later features combine multiple earlier concepts.

Encoder	Persist	Split	Merge	Die	Emerge
Whisper-S	.001	.576	.015	.191	.216
wav2vec2-B	.026	.587	.039	.161	.188
HuBERT-B	.031	.549	.046	.166	.209
WavLM-B	.031	.537	.047	.184	.200
Data2Vec-Audio	.029	.451	.047	.223	.249
UniSpeech-SAT	.041	.445	.072	.208	.234

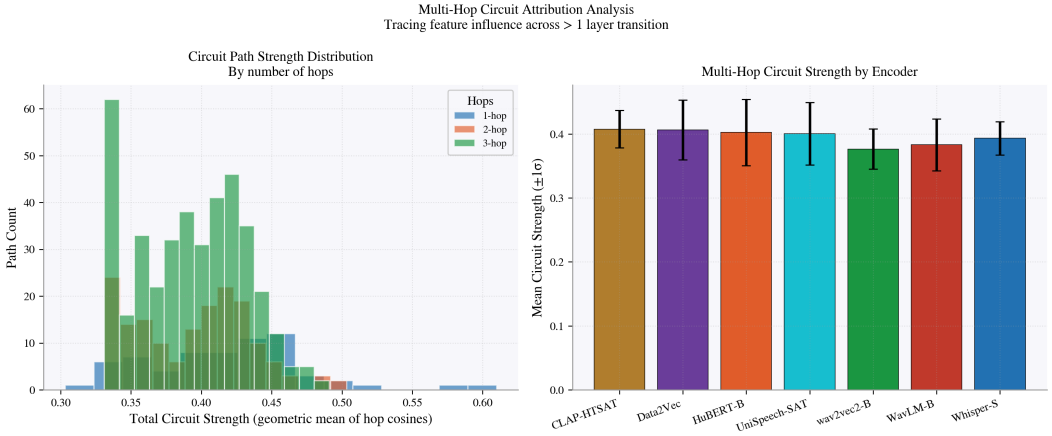

Figure 12: Sparse transcoder attribution graph. Nodes are sparse latent features and edges indicate attribution through an inter-layer transcoder. Dense hubs carry broad acoustic state, while narrow branches mark class-selective routes that are plausible ablation or steering targets.

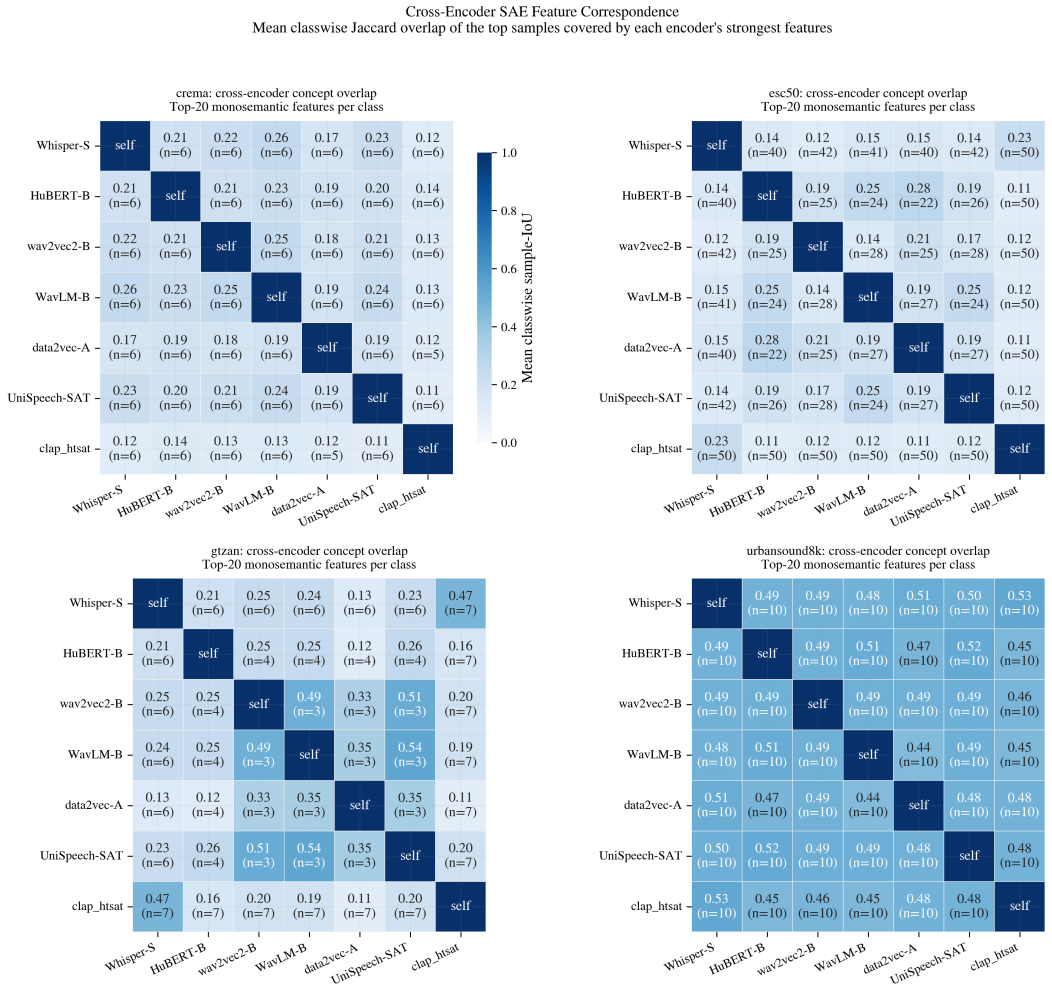

Figure 13: Cross-encoder feature overlap measured by activation-profile IoU. High overlap means two encoders activate sparse features on similar examples, while low overlap means the encoders carve the same dataset into different internal concepts. Overlap is strongest within objective families and weakest between CLAP and speech-focused encoders.

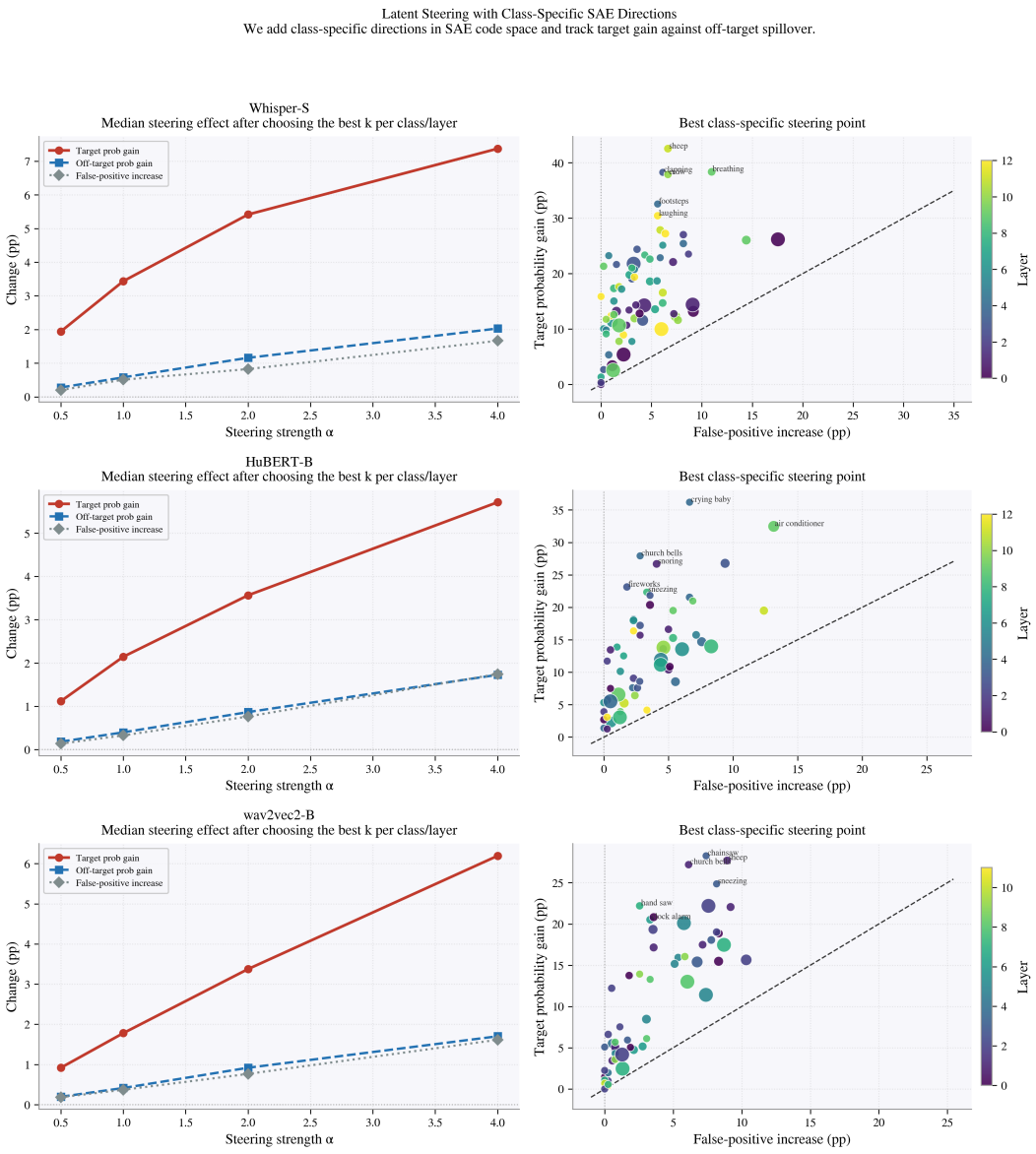

Figure 14: Latent steering tradeoff between target gain and off-target movement. The desirable region increases the target class while leaving retain classes unchanged. Points with large off-target movement expose broad semantic directions that are readable by probes and require retain-set validation as editing handles.

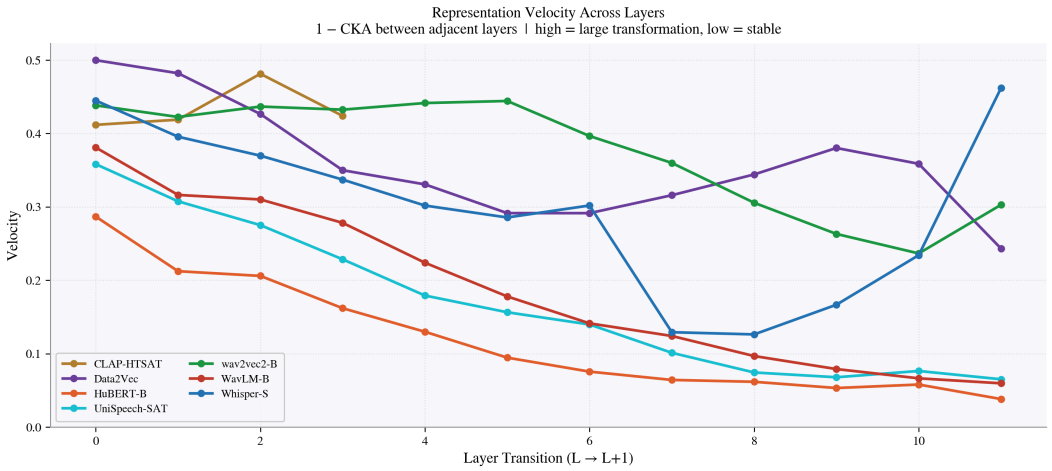

Figure 15: Representation velocity across layers and encoders. High velocity marks depths where the encoder rapidly rewrites acoustic state, often near objective-specific bottlenecks; low velocity marks plateaus where neighboring layers have similar downstream utility.

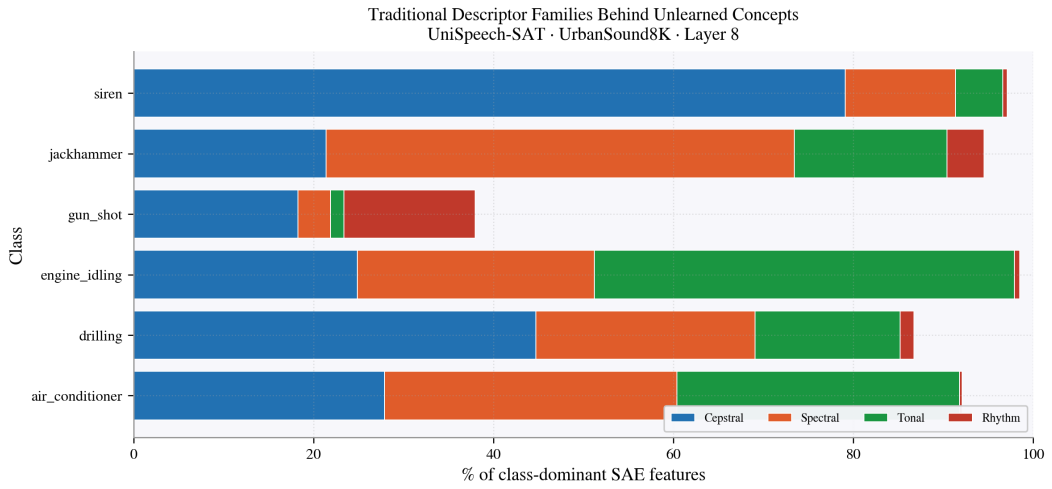

Figure 16: Relationship between concept ablation and traditional acoustic feature families. Concepts that align strongly with spectral or cepstral families are easier to interpret and can share evidence with nearby sound classes, so this bridge helps explain both target suppression and retain-set movement.

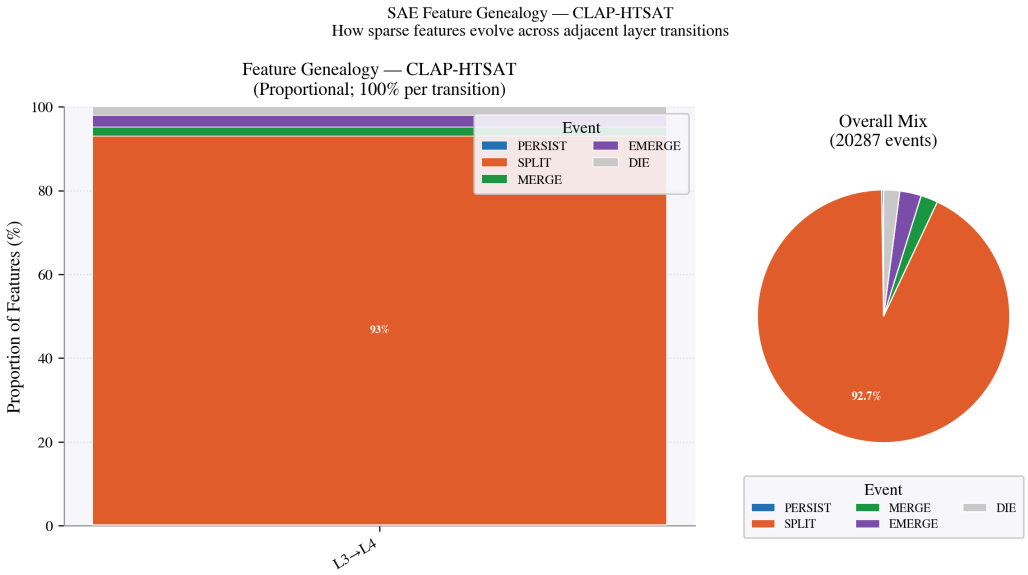

Figure 17: Feature genealogy for CLAP-HTSAT. Compact branches indicate that audio-text alignment forms label-aligned features that remain usable for class-level intervention diagnostics.

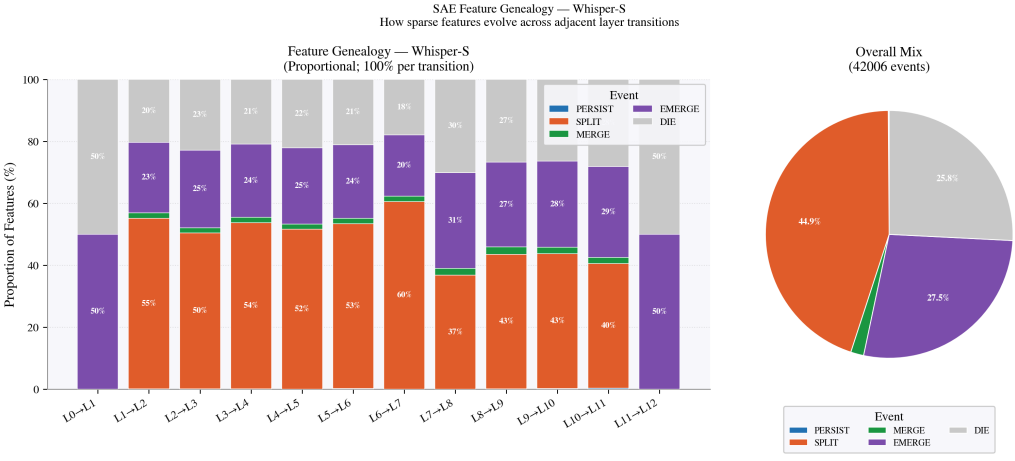

Figure 18: Feature genealogy for Whisper-S. Longer chains and repeated merges are consistent with late decoder-aligned, polysemantic routing.

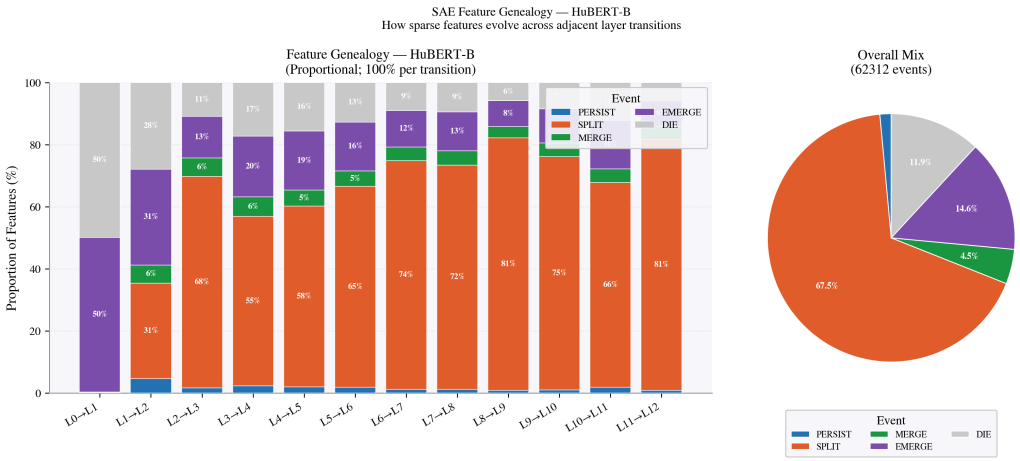

Figure 19: Feature genealogy for HuBERT-B. Discrete-unit pretraining produces clustered feature families that specialize early and then recombine.

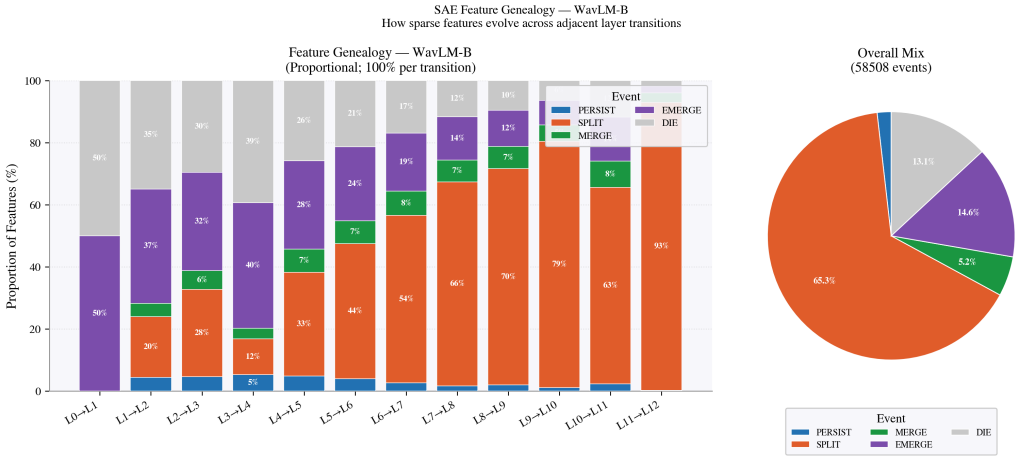

Figure 20: Feature genealogy for WavLM-B. Denoising encourages stable acoustic lineages that persist under perturbation more reliably than Whisper-S features.

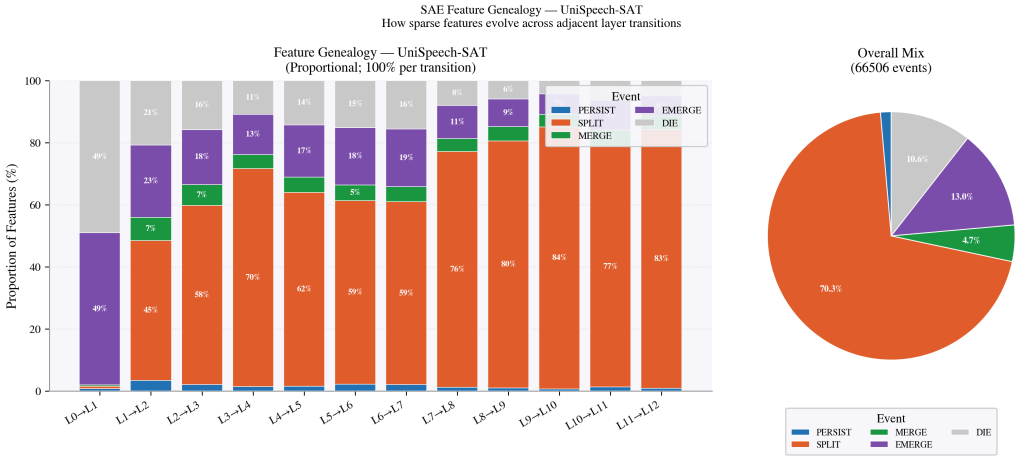

Figure 21: Feature genealogy for UniSpeech-SAT. Speaker-aware masked prediction creates intertwined content and speaker feature branches.

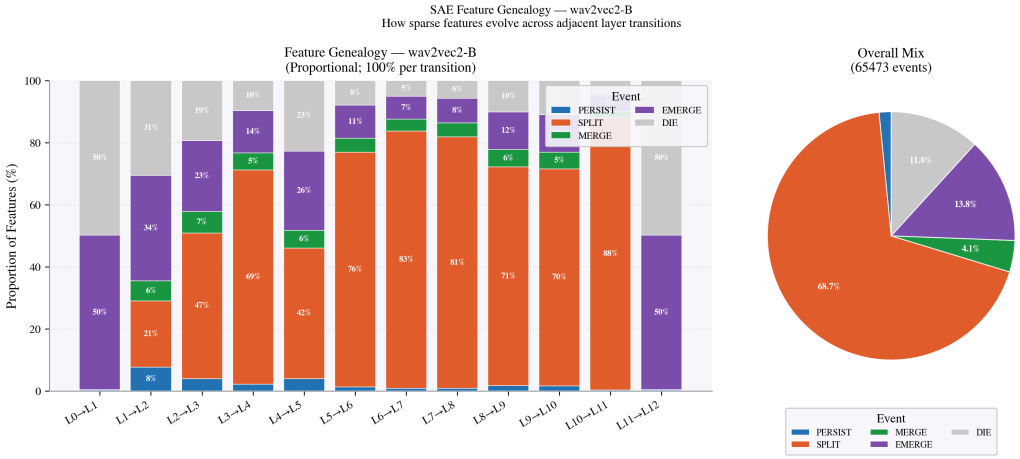

Figure 22: Feature genealogy for wav2vec2-B. Contrastive pretraining yields early acoustic branches that become less useful when later layers over-specialize.

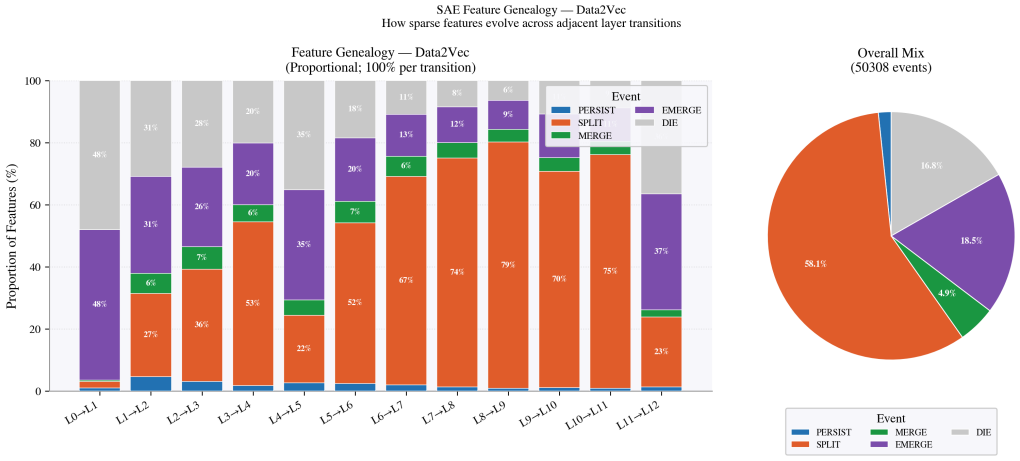

Figure 23: Feature genealogy for Data2Vec-Audio. Self-distillation produces a front-loaded genealogy, matching the frequent layer-0 optimum in probe results.

G Large-Model and Bootstrap Checks

This appendix section adds two checks that are useful for interpreting the main study. First, we repeat the layer-profile analysis on a larger/fused set: Whisper-large-v3, WavLM-large, and CLAP-HTSAT-fused. Second, we report bootstrap uncertainty for peak depth and depth-profile shape. These results are not used to change the main-paper claims; they test whether the same qualitative structure appears when model scale and sampling variation are made explicit. For Whisper-large-v3, as for Whisper-S, all reported readouts are encoder states.

G.1 Large-Model Layer Profiles

Table 31 shows that scale improves absolute scores while preserving depth dependence. WavLM-large peaks at intermediate layers across the reported transfer tasks, Whisper-large-v3 peaks mid-to-late depending on task, and CLAP-HTSAT-fused peaks at the upper analyzed HTS-AT stages. For CLAP, this is a stage-wise statement: the fused CLAP pipeline exposes four hierarchical stages plus an input/stem representation, not a 0-12 speech-transformer stack.

Table 32 isolates the largest best-layer versus final-layer gaps in the same large/fused run. Larger encoders still have non-final optimal readout depths: WavLM-large loses 13.3 points on ESC-50 and

Table 31: Best layers for the large/fused checkpoints. Brackets give 95% evaluation-bootstrap intervals. CREMA-D uses macro-F1; the other datasets use accuracy. CLAP indices are stage-wise indices in the fused HTS-AT extraction pipeline.

Encoder	ESC-50	UrbanSound8K	GTZAN	CREMA-D
CLAP-HTSAT-fused	L4: .958 [.938, .975]	L4: .871 [.850, .893]	L4: .810 [.750, .865]	L3: .607 [.579, .632]
WavLM-large	L9: .823 [.785, .858]	L10: .819 [.794, .844]	L18: .740 [.685, .800]	L16: .774 [.752, .797]
Whisper-large-v3	L20: .880 [.845, .913]	L11: .860 [.838, .883]	L17: .775 [.720, .830]	L31: .751 [.729, .772]

Table 32: Largest final-layer losses in the large/fused run. Loss is best-layer score minus final-layer score.

Encoder	Task	Best layer	Final layer	Best	Final	Loss
WavLM-large	ESC-50	9	24	.823	.690	.133
WavLM-large	GTZAN	18	24	.740	.640	.100
Whisper-large-v3	UrbanSound8K	11	32	.860	.789	.071
Whisper-large-v3	GTZAN	17	32	.775	.710	.065
WavLM-large	UrbanSound8K	10	24	.819	.757	.062
WavLM-large	CREMA-D	16	24	.774	.722	.051
Whisper-large-v3	ESC-50	20	32	.880	.845	.035
CLAP-HTSAT-fused	CREMA-D	3	4	.607	.580	.027

10.0 points on GTZAN if forced to use the final layer, while Whisper-large-v3 loses 7.1 points on UrbanSound8K and 6.5 points on GTZAN. Scale can raise the curve without removing the need to select a readout layer.

G.2 Audio Encoders of Audio-Language Models

We extended the frozen-encoder analysis to the audio encoder of Qwen-Audio [16] and to SALMONN's Whisper-large-v2 speech branch [49]. SALMONN's later fusion and generation path is outside this encoder-layer comparison. Main-text Table 4 gives the best and final layers. In the Qwen-Audio encoder, the ESC-50 optimum is at layer 26 (93.25%) versus 93.00% at the final layer 31, and participation ratio and isotropy are rank-correlated with probe accuracy across layers ($\rho = .80$ and .82). In the SALMONN speech branch, ESC-50 prefers the terminal state (layer 32, 87.25%), whereas GTZAN peaks three blocks earlier (layer 29, 88.0% versus 85.0%). The selected GTZAN layer is also concentrated: 69.2% of its SAE features are class-specific, and four features suffice for a 50% target drop. Under corruption, Qwen-Audio's layer 26 retains more accuracy than layer 31 under band-limiting and pitch shift, and SALMONN's layer 29 is more robust than its final layer under band-limiting. In these large ASR-derived encoders the best layer is near the top, so the final-layer penalty is small; the case for layer selection is weaker here than for speech SSL encoders on non-speech audio.

G.3 Large-Model Few-Shot Selection

The large/fused checkpoints also clarify how to read the few-shot probe rule. The candidate pool is heterogeneous: CLAP contributes a short stage-wise profile, WavLM-large contributes 25 readout depths, and Whisper-large-v3 contributes 33. For this reason, exact Hit@1 is less informative than score gap. Table 33 reports the few-shot probe mode at one and fifty shots per class. Even when Hit@1 is low, the score deficit to the full-data oracle remains small, typically two to four points and below one point for one-shot CREMA-D. This is the same distinction emphasized in the main text: the method is a low-cost recovery rule over final-layer extraction, not a claim that the exact oracle depth is always identifiable.

G.4 Bootstrap Peak-Depth Uncertainty

The bootstrap analysis resamples datasets and summarizes the normalized depth at which each large/fused encoder peaks. Normalized depth is computed on each encoder's own analyzed grid, so CLAP values are stage-normalized rather than speech-layer fractions. Table 34 and Figure 24 make

Table 33: Few-shot probe rule on the large/fused run. Gap is the selected-layer score deficit to the full-data oracle, averaged over the three large/fused encoders. Candidate count is the average number of analyzed depths.

Dataset	H@1, 1 shot	Gap, 1 shot	H@1, 50 shots	Gap, 50 shots	Cand.
CREMA-D	.667	.0051	.000	.0221	21.0
ESC-50	.333	.0275	.667	.0000	21.0
GTZAN	.000	.0350	.333	.0183	21.0
UrbanSound8K	.000	.0199	.000	.0214	21.0

Table 34: Bootstrap peak-depth summary for large/fused checkpoints. Depth is normalized within each encoder's analyzed grid.

Encoder	Peak depth mean [95% CI]	Peak score mean [95% CI]	Interpretation
CLAP-HTSAT-fused	.94 [.81, 1.00]	.811 [.673, .921]	Upper stage-wise CLAP peak.
WavLM-large	.55 [.40, .71]	.789 [.757, .821]	Intermediate denoising-SSL readout.
Whisper-large-v3	.62 [.41, .86]	.817 [.763, .870]	Mid-to-late ASR-supervised readout.

two points. First, CLAP-HTSAT-fused peaks near the top of its stage-wise hierarchy, with mean normalized peak depth .94 and 95% interval [.81, 1.00]. Second, WavLM-large and Whisper-large-v3 peak in broader middle-to-late regions, with intervals that reflect task variation rather than a single universal readout depth.

H Limitations

The study uses released checkpoints, so objective, training data, architecture, scale, and supervision source vary together. We therefore describe the observed depth profiles as pretraining-family associations rather than isolated causal effects. The audio encoders of Qwen-Audio and SALMONN are an out-of-panel check, but both derive from Whisper-large-v2, so they do not add an independent model lineage for every family.

The experiments use frozen encoders and lightweight readouts. This is the intended setting for label-free selection, few-shot validation, stored features, and pruning, and it is complementary to full fine-tuning and full-depth supervised layer mixtures. Our fine-tuning check covers one encoder–task pair (wav2vec2-base on ESC-50), and full-depth mixtures remain better on CREMA-D. When labels and compute are plentiful, these methods can still be the right choice; the prefix mixture targets the case where supervised weighting should remain compatible with pruning.

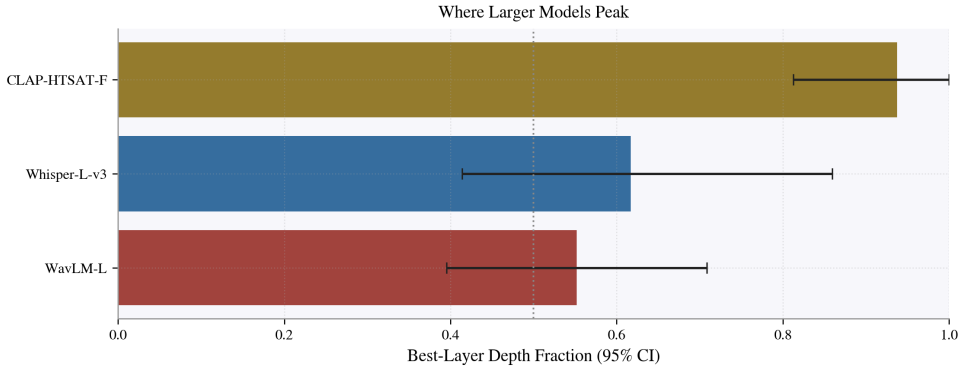

Figure 24: Bootstrap peak-depth summary for the large/fused checkpoints. CLAP is shown on its own stage-normalized grid, while WavLM-large and Whisper-large-v3 use their deeper readout grids. The intervals show that the qualitative depth ordering is not an artifact of a single dataset split.

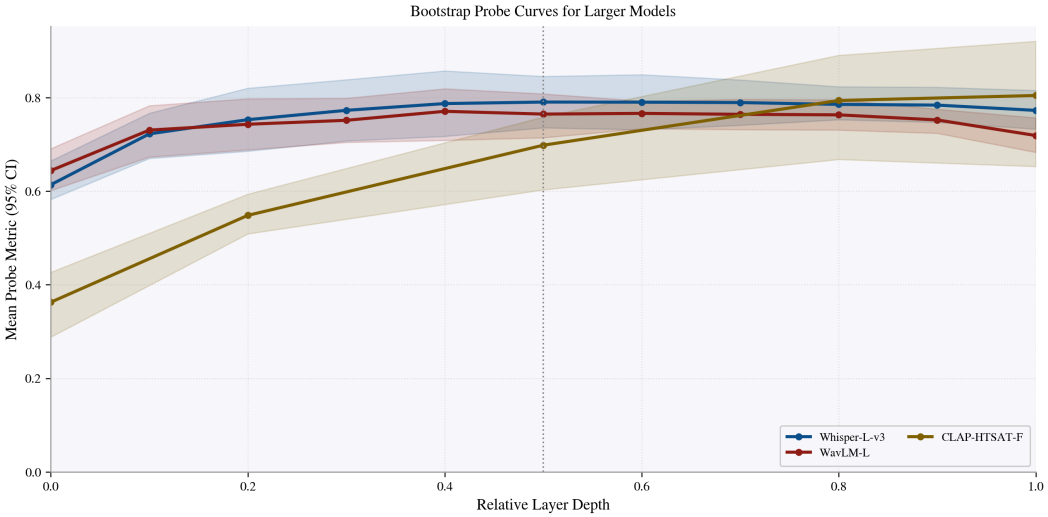

Figure 25: Bootstrap probe-depth profiles for the large/fused checkpoints. CLAP's profile rises toward the upper analyzed stages, WavLM-large has a broad middle-depth plateau, and Whisper-large-v3 remains strong from middle to late layers. Across these larger/fused checkpoints, scale raises absolute scores while depth dependence remains visible.

Latency and memory were measured for one truncation (wav2vec2-base after layer 2). Elsewhere, pruning results use depth as the cost proxy; real latency, memory, and energy depend on hardware, batching, sequence length, and implementation.

The geometry rule cannot distinguish tasks defined on the same audio. It is worse than the final layer on speech emotion and on CLAP, and its advantage over a middle-layer default on classification is small. Steering and feature ablation expose linear directions rather than deployment control policies or training-data deletion. SAE and transcoder summaries depend on the chosen dictionary, sparsity, threshold, and seed. The exit procedure of Section 4.6 was validated on held-out ESC-50 only, and we have not ablated its descriptive fields individually.

I Future Work and Release Scope

The central hypothesis is falsifiable: if a new objective changes useful depth, geometry, sparse-feature concentration, genealogy, or robustness, the per-layer measurements should reveal that change before downstream fine-tuning hides it. Our audio-language checks cover only the audio encoders of Qwen-Audio and SALMONN. The next extension is to move from frozen classifiers to full production audio systems: the fused paths of audio-language models, retrieval models, captioners, streaming ASR, diarization, source separation, codec encoders, and multimodal assistants. These systems add new readout choices such as cross-attention states, pooled retrieval tokens, decoder conditioning states, and codec latents, but the core question remains: which internal state should be read, stored, pruned, or edited for a workload?

Future evaluations should broaden both tasks and shifts. Event recognition and music genre are useful probes; long-form systems also require temporal event localization, retrieval, caption grounding, acoustic scene understanding, speaker and language change, multilingual ASR, paralinguistic attributes, assistive-listening audio, and continuous affect. Robustness should add room impulse responses, reverberation, codec artifacts, microphone nonlinearities, clipping, gain changes, overlapping speakers, background music, device/channel mismatch, far-field speech, and naturally recorded noisy corpora. Each extension can use the same layer report: peak layer, final-layer gap, geometry profile, robustness drift, sparse-feature localization, transcoder faithfulness, editing collateral, and compute cost.

A second direction is explicit compression. The present study treats early readout as a post hoc pruning signal for frozen feature extraction; future work can convert it into explicit early-exit encoders,

layer-skipping policies, cascaded ensembles, and adaptive compute controllers for streaming audio. The same measurements can supervise those controllers by checking whether the proposed exit preserves target score, descriptor support, perturbation stability, and retain-set behavior, connecting interpretability diagnostics to latency, energy, memory, and bandwidth budgets.

Finally, sparse features should be validated through human listening studies, counterfactual audio edits, activation patching, path-level transcoder interventions, threshold and dictionary-size sweeps, and cross-corpus replication. Transcoder faithfulness should be connected to downstream logit preservation as well as next-layer variance. The release is organized so these extensions can reuse the same probe, geometry, SAE, transcoder, steering, ablation, and bootstrap-error-plot scripts.

J Reproducibility Notes

Except for the single fine-tuning check in Section 4.5, all experiments use frozen encoders and linear downstream heads or learned layer mixtures. The public release contains scripts for dataset preparation, layer-wise activation extraction, linear probing, geometry/CKA computation, SAE and transcoder training, robustness perturbations, steering, ablation, table generation, figure generation, and bootstrap error plots. It also contains the derived CSV/JSON summaries used to produce the paper tables, with pointers to public checkpoints and datasets rather than redistributed raw audio. The supplemental manifest records source URLs, checkpoint and dataset versions, and license or terms-of-use pointers for each external asset.

The reported experiments were run on NVIDIA RTX A4500 20GB GPUs. The current reproduction estimate is approximately 150-250 A4500 GPU-hours plus CPU time for probing and table generation: 20-40 GPU-hours for activation extraction, 80-120 for SAE fitting, 40-70 for transcoders, and 10-20 for perturbation/steering sweeps. Linear probes, geometry, CKA summaries, table generation, and bootstrap summaries run on CPUs after extraction. Probe, robustness, and layer-selection uncertainty is reported with nonparametric bootstrap intervals over held-out clips (paired between selected and final layers; 10,000 utterance resamples for ASR). Prefix and full-depth mixtures are averaged over five fixed seeds. SAE, transcoder, and editing uncertainty uses example resampling and seed variation where reruns are available.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes].

   Justification: The abstract and introduction state the broad encoder and benchmark scope and describe only the analyses reported in the main paper and appendix.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes].

   Justification: Section 5 and Appendix H discuss the study scope, frozen-readout setting, pruning-cost interpretation, and validation requirements for geometry, steering, SAEs, and transcoders.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

Answer: [Yes].

Justification: The appendix states the assumptions behind each mathematical diagnostic, including geometry/Fisher separability, early-exit pruning, QUICKLAYER stability, projected-drift robustness, transcoder probe-error control, monosemantic ablation, and the decision-record utility. Derivations are presented as conditional bounds rather than standalone theory claims.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

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

Answer: [Yes].

Justification: Sections 3 and Appendix J specify the frozen encoders, datasets, metrics, layer extraction, SAE/transcoder setup, and available supplemental artifacts.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes].

   Justification: Appendix J states that an anonymized code and artifact release accompanies the submission, including scripts for activation extraction, probing, geometry/CKA, SAE/transcoder training, robustness, steering, ablation, and table/figure generation. Raw audio is accessed through the cited public datasets rather than redistributed.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

   Answer: [Yes].

   Justification: Section 3 and Appendix A provide dataset sizes, split counts, model metadata, metrics, and SAE/transcoder hyperparameters.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

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

   Answer: [Yes].

   Justification: Appendix J specifies the error-plot protocol: nonparametric bootstrap confidence intervals over held-out clips for probe, robustness, and QUICKLAYER metrics, with example resampling and available seed-to-seed variation for SAE, transcoder, and editing summaries. The accompanying release includes scripts to generate the corresponding error plots.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes].

   Justification: Appendix J states that the experiments used NVIDIA RTX A4500 20GB GPUs, gives approximate A4500 GPU-hour ranges for each major experiment family, and notes that probing, geometry, table generation, and bootstrap summaries run on CPU after extraction.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes].

   Justification: The study analyzes public models and datasets, preserves submission anonymity, and discusses limitations and dual-use concerns in Section 5.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes].

    Justification: Section 5 discusses positive uses for accessibility, safety analysis, debugging, and efficient inference, and negative uses involving sensitive audio monitoring or unsafe concept editing.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A].

    Justification: The paper does not release a new high-risk model or scraped dataset; it studies existing public audio encoders and public benchmark datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes].

Justification: The paper cites all model families and datasets used; Appendix J states that reproduction uses the public checkpoints and public datasets under their existing terms.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

Answer: [Yes].

Justification: The paper introduces derived interpretability artifacts and documents their fields in Appendix J and the supplemental CSV/table files.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

Answer: [N/A].

Justification: This work does not conduct new crowdsourcing or new human-subjects data collection; it analyzes existing public datasets.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

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

Answer: [N/A].

Justification: No new human-subjects study is performed; the analysis uses existing public audio benchmarks and frozen public model checkpoints.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

Answer: [N/A].

Justification: LLMs are not an important, original, or non-standard component of the scientific method; all reported results are from audio encoder analyses.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.