Listening at the Right Layer: Pretraining and Effective Rank in Audio Models

Arnesh Batra [1 2]

Abstract

Frozen audio encoders are usually reused by taking a final embedding, but useful task evidence often lives in an earlier, lower-cost, or sparser part of the representation hierarchy. We study how useful readout depth varies across pretraining families, using layer probes together with activation geometry, sparse dictionary features, transcoder routes, and intervention tests. Across audio-text, ASR-supervised, masked, denoising, contrastive, speaker-aware, and self-distillation encoders, final-layer extraction loses at least 10 score points in about half of the evaluated encoder-task settings, with the largest gaps reaching 24–38 points. A zero-label selector based on isotropy and effective rank reduces low-resource ASR character error rate in 11 of 12 language-encoder settings, while few-shot probes recover 11–13 points over final-layer extraction on common-depth encoders. Sparse autoencoders and transcoders show when the selected readout is concentrated, distributed, stable, or editable. The result is a conservative, readout-centered view of audio representation reuse: pretraining family is associated with useful depth, effective-rank geometry provides a candidate-layer prior to validate, and sparse/routing analyses record how selected layers store and route task evidence.

1. Introduction

Parameter- and compute-efficient model use is often presented as a weight-space problem: factorize an update, prune parameters, or restrict adaptation to a small subspace. Frozen audio encoders expose a complementary activation-space problem. A downstream system must decide which hidden state to store, probe, or serve. If the relevant evidence appears before the final layer, the chosen readout defines both a compression boundary and an activation subspace whose effective rank, sparse features, and inter-layer routes can be audited.

This paper studies that boundary for audio. Prior work has shown that speech SSL encoders are depth-dependent (Pasad et al., 2021; Yang et al., 2021; Turian et al., 2022); we ask which layer or prefix should be kept when the goal is efficient transfer rather than post hoc visualization. The question is practical for long audio streams and multi-head systems: event tagging, music retrieval, speech affect, moderation, and transcription may need different states from the same frozen backbone. It is also mechanistic. wav2vec 2.0, HuBERT, WavLM, UniSpeech-SAT, Data2Vec, Whisper, and CLAP/HTS-AT are trained with different targets (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022b;c; Baevski et al., 2022; Radford et al., 2023; Wu et al., 2023; Chen et al., 2022a). Those targets should not be expected to place acoustic, phonetic, semantic, and paralinguistic evidence at the same depth.

We call the resulting artifact the *Audio Interpretability Atlas*: a readout-centered diagnostic panel over seven pretrained encoders and four benchmarks. The atlas connects layer probes, centered kernel alignment (CKA), participation ratio, isotropy, classical audio descriptors, sparse autoencoders (SAEs), transcoders, perturbation tests, steering, and targeted feature ablation on the same encoder-layer-task cells. The organizing claim is that a selected layer is not merely the winner of a probe sweep. It is a structured decision record: score, depth, effective rank, sparse feature budget, transcoder faithfulness, robustness, editability, and cost.

The connection to factorization is diagnostic rather than algorithmic. We do not propose a new low-rank adapter or claim that selected layers are always low-rank. Instead, the singular-value spectrum of a centered activation matrix gives participation ratio and related effective-rank diagnostics, prefix selection identifies a smaller portion of the frozen computation, and the SAE and transcoder objectives provide sparse summaries of hidden states and layer transitions. These tools support an audit record; they do not by themselves settle the layer choice, which still needs downstream validation.

Our contributions are:

[1]Indraprastha Institute of Information Technology Delhi, Delhi, India [2]StarkVision Research. Correspondence to: Arnesh Batra <arnesh23129@iiitd.ac.in>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

- We formulate frozen audio readout selection as a structured prefix-compression problem: choose a layer or prefix, then audit the effective-rank, sparse, and routing structure around that readout.
- We show pretraining-associated depth profiles across audio-text, ASR-supervised, masked, denoising, contrastive, speaker-aware, and self-distillation encoders. Final-layer extraction is a poor single-layer default for many non-speech tasks.
- We evaluate selection with and without labels. A zero-label geometry rule improves low-resource ASR in 11 of 12 settings; few-shot probes recover 11–13 points over final-layer extraction.
- We connect compression to interpretability through SAEs, transcoders, perturbations, steering, and feature ablation, identifying which selected layers are concentrated, distributed, stable, or editable.

2. Related Work

Audio representation depth. SUPERB and HEAR made frozen audio transfer measurable across speech, environmental sound, and music tasks (Yang et al., 2021; Turian et al., 2022). Layer-wise probing further showed that speech SSL encoders localize different information at different depths (Pasad et al., 2021). Our operating point is different from a full-depth learned layer mixture: we ask which state should be extracted when labels are absent, scarce, or when a system wants a shallow prefix that can be cached, served, or audited.

Objective-conditioned subspaces. Audio encoders differ by prediction target as much as by architecture. HuBERT, wav2vec 2.0, WavLM, UniSpeech-SAT, and Data2Vec emphasize different SSL targets (Hsu et al., 2021; Baevski et al., 2020; Chen et al., 2022b;c; Baevski et al., 2022); Whisper is optimized for transcription (Radford et al., 2023); and CLAP aligns audio with text (Wu et al., 2023). We treat these targets as sources of pretraining-associated subspaces. Appendix A.4 formalizes a sufficient condition: a readout is linearly useful when downstream between-class scatter lands in such a subspace while projected within-class scatter remains controlled.

Effective-rank geometry and activation reuse. Low-rank adapters reduce trainable weight degrees of freedom; our setting is different because the base encoder remains frozen and the readout layer is the object being selected. For a centered activation matrix $Z^{(\ell)} = U_\ell \Sigma_\ell V_\ell^\top$, a task head only sees the subspace made available by the chosen layer. The singular values in Σ_ℓ define the covariance spectrum used by participation ratio. This is where SVD enters: it is a way to compute the effective-rank statistic, not the main selection method or a claim that the representation itself is low-rank. We therefore treat depth selection, effective rank, and sparse dictionary size as coupled diagnostics for compression.

Sparse factors and mechanistic audits. Sparse autoencoders and transcoders have become useful tools for decomposing large-model activations and inter-layer computation (Elhage et al., 2022; Cunningham et al., 2023; Bricken et al., 2023; Templeton et al., 2024; Dunefsky et al., 2024). Audio-native work has begun to apply similar tools to Whisper and HuBERT (Aparin et al., 2026). We extend this lens across audio pretraining families and connect the sparse factorization to an actionable compression boundary: the layer chosen for a downstream workload.

Why this is a structured-readout problem. The readout boundary factorizes a frozen system into a reused prefix, a discarded or deferred suffix, and a task head. The participation ratio estimates the effective rank of the selected activation cloud; isotropy measures whether linear directions are usable; SAEs factorize the state into sparse dictionary features; and transcoders factorize the transition from one layer to the next. The same selected layer can therefore be evaluated as an efficient readout, a measured activation subspace, a sparse feature set, and an intervention surface.

3. Methods

Encoders and tasks. We freeze Whisper-S (244M parameters), CLAP-HTSAT (86M), HuBERT-B (95M), WavLM-B (95M), UniSpeech-SAT (94M), wav2vec2-B (95M), and Data2Vec-Audio (94M). For Whisper we analyze encoder hidden states only. CLAP-HTSAT exposes four hierarchical HTS-AT stages plus an input/stem representation, reported as indices 0–4; the other encoders expose layers 0–12. The benchmark suite spans environmental sound, urban sound, music genre, and speech affect using ESC-50, UrbanSound8K, GTZAN, and CREMA-D (Piczak, 2015; Salamon et al., 2014; Tzanetakis & Cook, 2002; Cao et al., 2014). We report accuracy except on CREMA-D, where we report macro-F1. Dataset and split details are in Appendix A.

Layer probes and prefix compression. For each encoder-layer pair, we train a linear classifier on frozen pooled representations. We focus on single-layer readouts because they support label-free selection, few-shot validation, and early-exit compression. When labels are available, the selected layer $\hat{\ell}$ can also define a prefix mixture

$$z_{\text{prefix}} = \sum_{\ell=0}^{\hat{\ell}} \alpha_\ell z^{(\ell)}, \qquad \alpha_\ell = \frac{\exp a_\ell}{\sum_{j=0}^{\hat{\ell}} \exp a_j}. \tag{1}$$

This keeps the supervised weighting mechanism of learned layer mixtures while avoiding execution or storage of layers after $\hat{\ell}$ for that workload.

Zero-label geometry. Let $Z_c^{(\ell)} \in \mathbb{R}^{N\times d}$ be the centered activation matrix at layer ℓ, with rows $z_i^{(\ell)} - \bar{z}^{(\ell)}$. We compute the covariance eigenspectrum, equivalently obtainable from the singular values of $Z_c^{(\ell)}$, then report the participation ratio

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

If $\lambda_i^{(\ell)}$ are covariance eigenvalues and $p_i = \lambda_i^{(\ell)} / \sum_j \lambda_j^{(\ell)}$, then $\mathcal{PR}^{(\ell)} = 1/\sum_i p_i^2$. Thus PR is the rank of an equally weighted subspace with the same spectral concentration; it is a practical effective-rank proxy for geometry-based readout selection. Throughout this paper, "effective rank" refers to this participation-ratio estimate unless stated otherwise. We also compute isotropy

$$\mathcal{I}^{(\ell)} = 1 - \left| \frac{2}{N(N-1)} \sum_{i<j} \frac{(z_i^{(\ell)})^\top z_j^{(\ell)}}{\|z_i^{(\ell)}\|\,\|z_j^{(\ell)}\|} \right|. \tag{3}$$

The label-free QUICKLAYER mode min-max normalizes these two metrics within an encoder and selects

$$\hat{\ell}_{\text{geo}} = \arg\max_{\ell} \frac{1}{2}\left[\widetilde{\mathcal{I}}^{(\ell)} + \widetilde{\mathcal{PR}}^{(\ell)}\right]. \tag{4}$$

The rule is intentionally untuned: it proposes a candidate layer before labels exist. Appendix A.4 gives a sufficient condition linking pretraining-associated subspaces, downstream between-class scatter, and projected Fisher signal. The alignment quantity behind this link is

$$\alpha_{e,\ell}(y) = \frac{\operatorname{tr}(P_{e,\ell} S_B^{(\ell)} P_{e,\ell})}{\operatorname{tr} S_B^{(\ell)}}, \tag{5}$$

where $P_{e,\ell}$ projects onto the target-mean subspace induced by the pretraining target of encoder e. A layer is geometrically useful when this projected between-class scatter is large and the projected within-class scatter or perturbation drift is small.

Sparse and routing factorizations. For selected layers we train Top-k SAEs with a $6\times$ dictionary ($768 \to 4608$ features), annealing k from 130 to 75 active features:

$$f_\phi(h) = W_{\text{dec}}\,\text{TopK}(W_{\text{enc}}h + b_{\text{enc}}, k) + b_{\text{dec}}. \tag{6}$$

A feature is class-monosemantic when more than 70% of its activation mass falls on one class. Transcoders factorize transitions by predicting $h^{(\ell+1)}$ from $h^{(\ell)}$ through sparse latents and are scored by explained variance R^2. Robustness is measured by clean-to-corrupted probe drop, embedding cosine drift, and alive-feature drift. Steering and feature ablation test whether selected representations expose usable intervention handles.

4. Results

4.1. Useful Depth and Final-Layer Gaps

Table 1 shows that every neural encoder outperforms the classical descriptor baseline at its best readout, but no universal extraction depth exists. CLAP-HTSAT peaks at upper stages for environmental, urban, and music classification, consistent with audio-text alignment creating compact category features. Whisper-S peaks later, especially on CREMA-D, where layer 12 achieves .732 macro-F1. Speech SSL models often peak early or middle for non-emotion tasks, and Data2Vec is strongly front-loaded.

These profiles turn depth into a compression variable. Table 2 lists the largest cases where final-layer extraction discards useful evidence. On wav2vec2-B with ESC-50, layer 2 scores .693 while the final layer scores .315, a 37.8-point loss; for Data2Vec on ESC-50 the analogous gap is 36.5 points. The selected layer is therefore a candidate prefix boundary, not merely a plotting artifact.

CKA profile similarity supports the view that these boundaries are pretraining-associated rather than arbitrary probe noise. Figure 1 compares depth profiles across encoders: masked and denoising SSL models have closely related profiles, while CLAP's stage-wise profile separates from speech-focused encoders. Thus useful depth is a property of the pretrained representation family and target workload, not just parameter count.

CLAP-HTSAT has the strongest clean peak on three tasks: ESC-50 .970 at layer 4, UrbanSound8K .890 at layer 4, and GTZAN .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 .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

Table 1. Best-layer transfer score with optimal readout in parentheses. CREMA-D is macro-F1; all others are accuracy. CLAP indices are HTS-AT stage-wise readouts, not speech-transformer layer numbers.

Paradigm	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)
Traditional audio	Combined	.628	.620	.569	.748

Table 2. Largest final-layer losses. Loss is best-layer score minus final-layer score; keep is the fraction of analyzed 0–12 depth needed to reach the best layer.

Encoder	Task	Best layer	Best	Final	Loss
wav2vec2-B	ESC	2 (17%)	.693	.315	.378
Data2Vec	ESC	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	3 (25%)	.708	.463	.245

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 .315 versus .693 at layer 2; Data2Vec is optimal at layer 0 on most transfer tasks.

4.2. Geometry as a Zero-Label Layer Prior

The zero-label selector in Equation 4 is most useful when annotations are absent but unlabeled target-domain audio is available. In low-resource ASR, selecting layers from unlabeled target-language audio reduces average CER from 62.37 to 49.09 across wav2vec2-large and Data2Vec settings, an average 21.3% relative reduction (Table 3). The same geometry signal transfers to classification: over the six common-depth encoders, label-free selection gains 9.2 points over final-layer extraction on average, with larger gains on broad acoustic tasks (Table 4). Thus effective rank and isotropy are a candidate-layer prior, not an explanation of pretraining or a replacement for task validation.

4.3. Layer Readout With or Without Labels

We use final-layer extraction as the single-layer reference because it is the readout obtained when no layer-selection step is performed. Learned layer mixtures are supervised full-depth readouts; QUICKLAYER instead asks which single

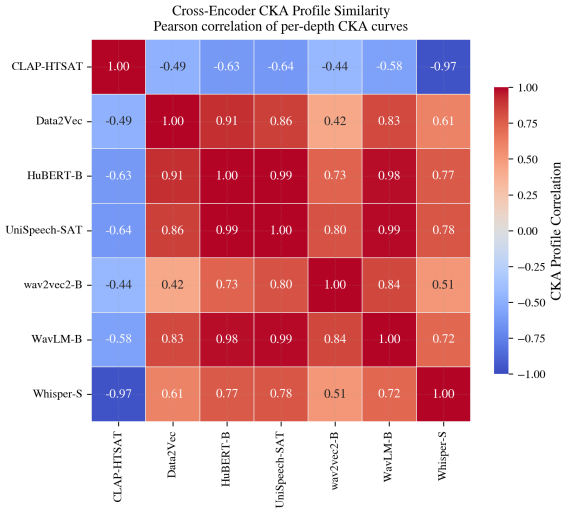

Figure 1. CKA depth-profile similarity across encoders. Related speech SSL objectives route information through depth in similar ways, while CLAP's stage-wise profile is separated from speech-focused encoders.

Table 3. Zero-label ASR layer selection by geometry. CER is averaged over Amharic, Welsh, Hausa, Kyrgyz, Swahili, and Yoruba; positive reduction is relative to final-layer extraction.

Encoder	Improved	Geo. CER	Final CER	Rel. red.
wav2vec2-large	6/6	55.42	72.85	23.9%
Data2Vec	5/6	42.76	51.88	17.6%
Mean	11/12	49.09	62.37	21.3%

layer should be extracted when labels are absent or scarce and pruning decisions matter. This is a different operating point from SUPERB-style supervised full-depth mixtures, but it is not incompatible with them: after selection, one can learn a mixture over layers $0, \ldots, \hat{\ell}$ using Equation 1 to improve the supervised readout without running the pruned suffix. Without labels it uses Equation 4; with labels it selects the candidate readout with the best few-shot validation score.

Table 4 answers a readout-selection question rather than an exact-oracle-indexing question. Without labels, geometry

Table 4. QUICKLAYER gains over final-layer extraction on common-depth encoders. Geometry is label-free; probe uses s labeled examples per class.

		No labels	With labels: probe gain by shots				
Dataset	Final	Geo.	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

improves over final-layer extraction by 9.2 points on average and by 11–15 points on broad acoustic tasks; CREMA-D marks the speech-affect boundary where later paralinguistic structure is often needed. This is why the label-free rule is best treated as a proposal mechanism, although it improves 11/12 low-resource ASR settings in Table 3. With labels, the few-shot probe mode is the stronger default: one example per class recovers 12.4 points and fifty examples recover 13.4 points.

Layer choice is also intervention-relevant in a diagnostic sense: class-difference vectors steer most selectively near the probe-optimal layer, late for Whisper-S and earlier for HuBERT-B and wav2vec2-B. We treat these 20–30 point steering effects as evidence that selected readouts expose useful linear handles that should be validated for the intended application, not as deployment-ready control policies.

4.4. Sparse Features and Transcoders Reveal Editable Structure

SAEs expose whether task information is concentrated in a few class-specific features or spread across polysemantic ones. Extending AudioSAE beyond Whisper and HuBERT (Aparin et al., 2026), we find that feature budget is an empirical pretraining-associated signature in this panel. CLAP-HTSAT often stores task evidence in compact label-aligned features; WavLM and HuBERT expose more distributed but robust acoustic structure; Whisper routes information faithfully but polysemantically through late speech states. Figure 2 shows two diagnostics. The minimum sufficient feature set (MSFS) counts how many SAE features must be removed to cause a 50% target-class score drop. CLAP reaches that threshold with about eight features, while Whisper-S and WavLM-B require broader sets. The entropy-rate plot separates rare class detectors from high-rate reusable features that require retain-set audits.

Transcoders add a sparse-routing view (Table 6). Across encoder-dataset combinations, the best inter-layer transitions explain most next-layer variance ($R^2 \approx 0.86$ to 0.998), but faithfulness is not the same as interpretability. UrbanSound8K elicits the highest monosemantic transcoder ratios

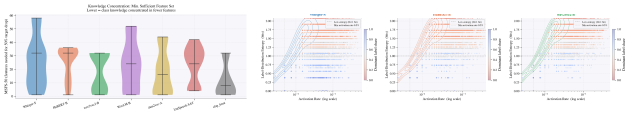

Figure 2. Sparse feature structure and editability. Left: feature budget needed for a 50% target-class drop. Right: feature entropy versus activation rate; rare low-entropy points are class-specific detectors, while high-rate features are reusable and require retain-set audits.

for every encoder, with CLAP-HTSAT reaching .774, while Whisper-S has zero class-monosemantic transcoder ratio under the current criterion. Thus a sparse route can be faithful but dense, or editable but more shift-sensitive; this distinction matters when the selected layer is used as a compression boundary.

The full transcoder metrics clarify 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. 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.

This separation has a simple linear-readout implication. 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,

$$\begin{aligned} &\left| g_c(h^{(\ell+1)}) - g_c(T_\psi(h^{(\ell)})) \right| \\ &\qquad \le \|w_c\|_2 \, \|h^{(\ell+1)} - T_\psi(h^{(\ell)})\|_2. \end{aligned} \tag{7}$$

High R^2 therefore bounds the residual that can perturb linear readout information. It does not show that the route decomposes into class-readable pieces, which is why the sparse-route ratio is reported separately from transcoder explained variance.

Interventions test whether these structures are actionable. Steering is most selective near the probe-optimal layer, and targeted SAE-feature ablation asks whether monosemantic features can suppress a target with lower retain-set movement than distributed encodings. These are feature-ablation diagnostics rather than training-data unlearning. They are useful because they connect the selected readout to an operational question: whether the selected layer exposes sparse or approximately linear handles that can be audited.

The linearized effect of feature ablation makes this target-

Table 5. Geometry-utility regimes. Correlations are computed over observed layers and datasets; $r_{\mathcal{PR}}$ and $r_{\mathcal{I}}$ denote correlations between probe score and participation ratio or isotropy. CLAP uses hierarchical stages, so its correlations are descriptive rather than directly depth-matched to 0–12 speech stacks.

Reg.	Objective cue	Observed assignment	Interpretation
R1	Broad contrastive, text-aligned, or teacher-state targets can reward high effective dimension.	CLAP-HTSAT, Data2Vec, wav2vec2-B; $r_{\mathcal{PR}} = .49$ to .72, $r_{\mathcal{I}} = .20$ to .76.	Spread is a strong 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 gated by ASR or corruption-stable subspaces.
R3	Hard discrete units or speaker-aware targets can intentionally compress variance.	HuBERT-B, UniSpeech-SAT; $r_{\mathcal{I}} = -.40/-.31$.	Anisotropy is useful if labels align with the compressed phonetic or speaker subspace.

Table 6. Sparse-feature, transcoder, and robustness signatures behind compression risk. MSFS is the number of SAE features whose ablation causes a 50% target-class drop. Mono. denotes class-monosemantic transcoder ratio on the strongest diagnostic transition.

Family	Sparse/routing evidence	Robustness signal	Selection implication
CLAP-HTSAT	MSFS ≈ 8; ESC-50 SAE monosemantic ratio .258; US8K transcoder mono. .774, $R^2 = .975$.	Strong clean transfer, but feature drift is higher than denoising SSL.	Good candidate for compact edits; audit off-target and shift behavior.
Whisper-S	Broad SAE feature sets; transcoder mono. 0 despite R^2 up to .998.	Largest controlled score drop in the corruption suite.	Keep later layers for speech affect; steer late and audit retain sets.
Masked + denoising SSL	Moderate selectivity; US8K transcoder mono. .296–.471.	Lowest alive-feature drift for WavLM and strong masked-SSL stability.	Prefer when expected shifts resemble the tested corruptions.
Data2Vec + w2v2	Useful layers often 0–2; final-layer losses up to .365/.378.	Early states preserve non-speech evidence before objective specialization.	Use pruned exits for non-speech transfer after validation.

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 feature set S changes a class margin by

$$\Delta M_{y,c}(S) = -\sum_{f \in S} a_f (w_y - w_c)^\top d_f. \tag{8}$$

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

4.5. Robustness Complements Clean Accuracy

Robustness adds an orthogonal view of layer quality. In the controlled perturbation suite (Table 7 and Figure 3), WavLM-B and the masked SSL family are most stable under the controlled corruption suite, while Whisper-S combines strong clean transfer with the largest controlled score drop. These controlled stress tests complement the clean readout and sparse-feature diagnostics.

This noise view is also layer- and subspace-dependent. Let $\tilde{\delta}_i^{(\ell)} = f_\ell(x_i + \delta_i) - f_\ell(x_i)$ be the representation drift caused by corruption, and let $\Gamma_i^{(\ell)}$ be the clean margin of a linear probe at layer ℓ. The prediction is unchanged if

$$\max_{c \neq y_i} \left| (w_{y_i} - w_c)^\top \tilde{\delta}_i^{(\ell)} \right| < \Gamma_i^{(\ell)}. \tag{9}$$

Thus total embedding drift is not sufficient: corruptions matter most when they move examples in the probe or pretraining-associated subspace used by the selected layer.

Whisper-S illustrates why multiple robustness metrics are needed: embeddings remain directionally similar under noise, while alive-feature drift identifies changes in the sparse features used by the classifier. CLAP shows a complementary pattern: strong clean transfer can rely on narrow label-concentrated features that shift under corruption. Thus robustness and clean readout selection should be audited together. Denoising SSL has the smallest average accuracy drop (.044), activation drift (.103), and alive-feature drift (.117), while supervised ASR has the largest score drop (.280) despite small global embedding drift. Appendix B contains the full perturbation and intervention diagnostics.

Table 7. Controlled-perturbation robustness by pretraining paradigm. Lower is better except for relative robustness.

Paradigm	Drop	Act.	Feat.	Rel.
Denoising SSL	.044	.103	.117	1.00
Masked SSL	.065	.195	.168	.89
Contrastive SSL	.083	.460	.138	.87
Audio-text	.101	.377	.207	.68
Supervised ASR	.280	.074	.407	.00

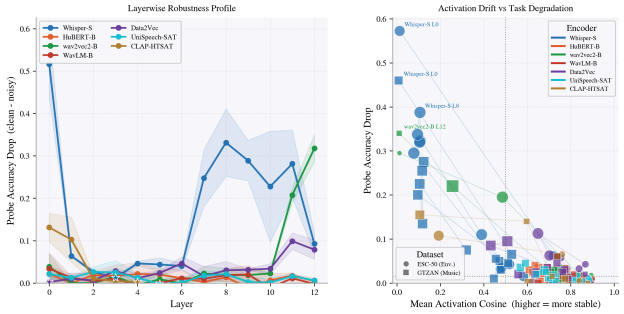

Figure 3. Distribution-shift fingerprints under acoustic corruption. The plot jointly summarizes clean-to-corrupted score drop, embedding drift, and alive-feature drift, separating global representation stability from sparse feature stability. Whisper-S keeps globally similar embeddings while its classifier-relevant sparse features change, producing the largest performance drop; WavLM-B and masked SSL encoders preserve both task score and feature activity more reliably.

4.6. Layer Selection as Pruning Evidence

Layer selection's primary operational role is pruning: it indicates how much of the analyzed stack is needed for a frozen-extraction workload. When an earlier layer matches or exceeds the final-layer score, later blocks become pruning candidates after target-domain validation. Table 2 gives the strongest cases, such as wav2vec2-B reaching its ESC-50 optimum after 17% of analyzed depth. This is a pruning signal rather than a hardware throughput claim: the exit is chosen because it preserves task evidence. With labels, Equation 1 can learn weights over the selected prefix while retaining the same pruning boundary. Appendix B gives the risk-cost form with robustness, feature rank, descriptor retention, and compute terms.

Table 8 is the operational version of the largest-loss rows: 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. The gain column is not a latency measurement; it is the score recovered by not using the final layer as the default extraction point.

Let m_ℓ be the downstream probe score at layer ℓ and let L be the final analyzed layer. For tolerance $\epsilon \geq 0$, the earliest

Table 8. Non-speech pruning evidence averaged over ESC-50, GTZAN, and UrbanSound8K. Gain is best-layer score minus final-layer score.

Encoder	Best layers	Kept	Best	Gain
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

near-oracle exit is

$$\ell^\star_\epsilon = \min\left\{\ell : m_\ell \geq \max_{j \leq L} m_j - \epsilon\right\}. \tag{10}$$

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: 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.

The 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 $U_\lambda(\ell) = m_\ell - \lambda C_\ell$. A pruned exit $\ell < L$ has higher utility than the final layer when

$$m_\ell - m_L > -\lambda(C_L - C_\ell). \tag{11}$$

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 $C_\ell < C_L$. A slightly lower-scoring earlier layer can also be preferred when latency, memory, robustness, or sparse-feature stability compensate for the score gap.

The risk-aware version keeps the same structure while making the audit terms explicit. Let $\kappa_\ell = \mathcal{PR}^{(\ell)}/d$ be the normalized activation rank and define

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

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

$$\begin{aligned} m_\ell - m_L + \lambda_c(C_L - C_\ell) + \lambda_k(\kappa_L - \kappa_\ell) \\ + \lambda_D(D_L - D_\ell) + \lambda_F(F_L - F_\ell) + \lambda_A \Delta A_{\ell,L} > 0, \end{aligned} \tag{12}$$

where $\Delta A_{\ell,L} = A_\ell(\mathcal{G}) - A_L(\mathcal{G})$. 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, drift, rank, or descriptor-retention improvements compensate for the score gap.

The same interpretation connects layer selection to compression and effective-rank analysis. A selected readout defines a prefix, an activation subspace, and a task head. Its participation ratio estimates how many effective directions the head can use; its sparse feature budget estimates whether those directions are concentrated or distributed; and its perturbation behavior estimates whether the subspace is stable under expected shifts. Thus an early layer is not accepted merely because it is early. It is accepted when the task signal remains linearly accessible in a lower-cost state and the audit metrics do not flag unacceptable robustness or retain-set risk.

5. Discussion

Frozen audio encoders are best treated as hierarchies rather than single embedding functions. A downstream readout is useful when, under a given pretraining family, the relevant acoustic or semantic evidence is linearly available at that depth and later layers do not reshape it away from the downstream objective. This helps explain the observed associations: audio-text contrastive training yields compact upper-stage category features, ASR-supervised encoders route useful speech structure later, and several speech SSL encoders expose reusable acoustic structure early.

The atlas suggests a reporting standard for pruning frozen audio encoders. A clean score is more informative when paired with readout depth, last-layer gap, geometry score, controlled-perturbation drop, alive-feature drift, sparse localization, transcoder faithfulness, editing collateral, and compute cost. These metrics are complementary: Whisper can have faithful inter-layer prediction with polysemantic routing; CLAP can pair strong clean transfer with concentrated features that deserve perturbation audits; and geometry can propose useful zero-label ASR exits while speech-affect tasks benefit from labeled validation. Layer selection is therefore simultaneously a transfer, pruning, and audit decision.

This framing is especially relevant for long-stream and multi-head audio systems. If an early readout preserves target evidence, it can reduce executed blocks and stored representation depth; with labels, a prefix mixture can improve the head without requiring the pruned suffix. If a later readout is needed, as in speech affect, the atlas records why that cost is justified.

The central hypothesis is falsifiable: if a new objective changes useful depth, geometry, sparse-feature concentration, sparse routing, or robustness, that change should be visible in the readout record before downstream fine-tuning hides it. The next extension is to move from frozen classifiers to production audio systems: 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 tasks and shifts, including localization, retrieval, caption grounding, multilingual ASR, diarization, room acoustics, codec artifacts, far-field speech, and natural noise. The same readout record can support explicit early exits or layer-skipping policies by checking target score, descriptor support, perturbation stability, retain-set behavior, and compute cost.

The study is observational: objective, pretraining data, architecture, scale, and supervision source vary together. The benchmark suite is a diagnostic panel; probes, steering, and SAEs measure linearly accessible structure; robustness uses controlled perturbations; and pruning claims are based on analyzed depth rather than measured deployment latency. Applications should validate selected layers with target, retain-set, and expected-shift audits.

Societal impact. Better audio-representation audits can improve accessibility tools, acoustic safety systems, music search, model debugging, and efficient on-device inference. The same diagnostics could also make audio monitoring systems easier to tune for sensitive classes or environments. We therefore frame steering and ablation as diagnostic interventions, and recommend reporting retain-set behavior, off-target drift, perturbation robustness, and dataset provenance before user-facing deployment.

Conclusion. Audio encoders do not simply become more semantic with depth; they reorganize acoustic evidence in ways associated with objective, data domain, architecture, and supervision. The atlas records where useful readouts occur, their candidate pruning boundaries, and whether the selected representation is concentrated, distributed, stable, or editable. Practically, final-layer extraction should not be assumed to be the safe single-layer default: environmental sound, urban events, and music often prefer early or middle readouts, speech affect can require late readouts, and ASR can use geometry without transcripts.

References

Aparin, G., Sadekova, T., Rukhovich, A., Yermekova, A., Kushnareva, L., Popov, V., Kuznetsov, K., and Piontkovskaya, I. 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)*, pp. 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.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460, 2020. URL https://arxiv.org/abs/2006.11477.

Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., and Auli, M. 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*, pp. 1298–1312. PMLR, 2022. URL https://arxiv.org/abs/2202.03555.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N. L., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014. doi: 10.1109/TAFFC.2014.2336244.

Chen, K., Du, X., Zhu, B., Ma, Z., Berg-Kirkpatrick, T., and Dubnov, S. 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*, pp. 646–650. IEEE, 2022a. doi: 10.1109/ICASSP43922.2022.9746312. URL https://arxiv.org/abs/2202.00874.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Qian, Y., Seltzer, M. L., Wang, S., Chen, L., Meng, H., Yu, D., and Wei, F. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022b. doi: 10.1109/JSTSP.2022.3188113. URL https://arxiv.org/abs/2110.13900.

Chen, S., Wu, Y., Wang, C., Chen, Z., Chen, Z., Liu, S., Wu, J., Qian, Y., Wei, F., Li, J., Zeng, X., and Yu, D. UniSpeech-SAT: Universal speech representation learning with speaker aware pre-training. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6152–6156. IEEE, 2022c. doi: 10.1109/ICASSP43922.2022.9747077. URL https://arxiv.org/abs/2110.05752.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

Dunefsky, J., Chlenski, P., and Nanda, N. Transcoders find interpretable LLM feature circuits. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0768.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. 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.

Pasad, A., Chou, J.-C., and Livescu, K. Layer-wise analysis of a self-supervised speech representation model. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 914–921. IEEE, 2021. doi: 10.1109/ASRU51503.2021.9688093.

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

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision, 2023. URL https://arxiv.org/abs/2212.04356.

Salamon, J., Jacoby, C., and Bello, J. P. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 1041–1044. ACM, 2014. doi: 10.1145/2647868.2655045.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Turian, J., Shier, J., Khan, H. R., Raj, B., Schuller, B. W., Steinmetz, C. J., Malloy, C., Tzanetakis, G., Velarde, G., McNally, K., Henry, M., Pinto, N., Noufi, C., Clough, C., Herremans, D., Fonseca, E., Engel, J., Salamon, J., Esling, P., Manocha, P., Watanabe, S., Jin, Z., and Bisk, Y. HEAR: Holistic evaluation of audio representations. In *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, volume 176 of *Proceedings of Machine Learning Research*, pp. 125–145. PMLR, 2022. URL https://proceedings.mlr.press/v176/turian22a.html.

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

Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., and Dubnov, S. 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*. IEEE, 2023. doi: 10.1109/ICASSP49357.2023.10095969. URL https://arxiv.org/abs/2211.06687.

Yang, S.-w., Chi, P.-H., Chuang, Y.-S., Lai, C.-I. J., Lakhotia, K., Lin, Y. Y., Liu, A. T., Shi, J., Chang, X., Lin, G.-T., Huang, T.-H., Tseng, W.-C., Lee, K.-t., Liu, D.-R., Huang, Z., Dong, S., Li, S.-W., Watanabe, S., Mohamed, A., and Lee, H.-y. SUPERB: Speech processing universal performance benchmark. In *Proceedings of Interspeech*, pp. 1194–1198, 2021. doi: 10.21437/Interspeech.2021-1775.

A. Extended Experimental Details

A.1. Datasets, Encoders, and Extraction

This appendix keeps the experimental and mechanistic details needed to audit the layer-selection claims: frozen layer readouts, geometry, sparse features, transcoder routes, robustness, and risk-cost pruning.

Table 9. Dataset summary. Accuracy is reported for ESC-50, UrbanSound8K, and GTZAN; CREMA-D uses macro-F1.

Dataset	Task	Clips	Classes	Train	Test
ESC-50	Env. sound	2000	50	1600	400
GTZAN	Music genre	999	10	799	200
CREMA-D	Emotion	7442	6	6136	1306
UrbanSound8K	Urban sound	8732	10	7079	1653

Table 10. Studied encoders and analyzed readout grids. CLAP-HTSAT uses stage-wise HTS-AT readouts rather than 12 speech-transformer blocks.

Encoder	Objective family	Params	Depths
Whisper-S	Weakly supervised ASR	244M	0–12
CLAP-HTSAT	Audio-text contrastive	86M	0–4
HuBERT-B	Offline-unit masked prediction	95M	0–12
WavLM-B	Masked prediction + denoising	95M	0–12
UniSpeech-SAT	Speaker-aware masked prediction	94M	0–12
wav2vec2-B	Contrastive masked prediction	95M	0–12
Data2Vec-Audio	Self-distillation	94M	0–12

Audio is loaded as mono, resampled to the checkpoint processor rate, and clipped to 5 s for ESC-50, UrbanSound8K, and CREMA-D and 30 s for GTZAN. Hidden states are extracted with `output_hidden_states=True` and mean-pooled over time or time-frequency positions. All base encoders are frozen. Linear probes train separately on each candidate layer. Top-k SAEs use a $6\times$ dictionary expansion, $k_0 = 130$ annealed to $k = 75$, 600 epochs, batch size 128, Adam learning rate 10^{-4}, and seed 13. Transcoders use hidden multiplier 5, Top-$k = 16$, 20 epochs, batch size 512, Adam learning rate 9×10^{-4}, and seed 17.

A.2. QUICKLAYER Modes

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.

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

$$G_s(q) = \frac{1}{|\mathcal{D}||\mathcal{E}|} \sum_{d,e} \left(m_{d,e}^{(q_s)} - m_{d,e}^{(L_e)} \right). \tag{13}$$

A complementary recovered-gap fraction compares the selected layer with the full-data oracle,

$$R_s(q) = \frac{\sum_{d,e} \left(m_{d,e}^{(q_s)} - m_{d,e}^{(L_e)} \right)}{\sum_{d,e} \left(m_{d,e}^{(\star)} - m_{d,e}^{(L_e)} \right)}. \tag{14}$$

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 11. QUICKLAYER with-label probe mode, dataset-wise artifact diagnostics. Hit@1 is exact oracle-layer recovery; gain is selected-layer score minus final-layer score; score gap is the deficit to the oracle.

Dataset	Shots	Hit@1	Gain	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
US8K	1	.4286	+.0974	.0121	11.8571
US8K	5	.4286	+.0923	.0172	11.8571
US8K	10	.2857	+.0923	.0172	11.8571
US8K	20	.2857	+.0985	.0110	11.8571
US8K	50	.4286	+.0898	.0197	11.8571

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.

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

$$a_{\ell^\star} - a_{\hat{\ell}} \le 2\eta. \tag{15}$$

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 $1-\delta$,

$$a_{\ell^\star} - a_{\hat{\ell}} \le 2\sigma\sqrt{\frac{2\log(2(L+1)/\delta)}{s}}. \tag{16}$$

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

A.3. Geometry, Effective Rank, and Fisher Signal

Let $\tilde{\lambda}_k^{(\ell)} = \lambda_k^{(\ell)} / \operatorname{tr} \hat{\Sigma}^{(\ell)}$ be normalized covariance eigenvalues. The participation ratio used as effective rank is a Rényi-2 spectral diversity:

$$\mathcal{PR}^{(\ell)} = \frac{1}{\sum_k (\tilde{\lambda}_k^{(\ell)})^2} = \exp\Big(H_2(\tilde{\boldsymbol{\lambda}}^{(\ell)})\Big). \tag{17}$$

It equals d for a uniform d-dimensional spectrum and approaches 1 when variance collapses onto one direction. For class means $\mu_c^{(\ell)}$ and class priors π_c, the between- and within-class scatter matrices define

$$S_B^{(\ell)} = \sum_c \pi_c (\mu_c^{(\ell)} - \bar{\mu}^{(\ell)})(\mu_c^{(\ell)} - \bar{\mu}^{(\ell)})^\top, \tag{18}$$

$$S_W^{(\ell)} = \sum_c \pi_c \mathbb{E}[(z^{(\ell)} - \mu_c^{(\ell)})(z^{(\ell)} - \mu_c^{(\ell)})^\top \mid y = c]. \tag{19}$$

The Fisher signal

$$\mathcal{F}^{(\ell)} = \operatorname{tr}\Big[(S_W^{(\ell)})^{-1} S_B^{(\ell)}\Big] \tag{20}$$

connects geometry to layer utility: effective rank is helpful when the spread or low-noise subspace also captures downstream between-class scatter.

A.4. Objective-Induced Subspaces

Let t_e denote the target variable used by encoder e during pretraining: a masked 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

$$\mathcal{S}_{e,\ell} = \operatorname{span}\Big\{\mathbb{E}[h_e^{(\ell)} \mid t_e = t] - \mathbb{E}[h_e^{(\ell)}] : t \in \mathcal{T}_e\Big\}, \tag{21}$$

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

$$\alpha_{e,\ell}(y) = \frac{\operatorname{tr}(P_{e,\ell} S_B^{(\ell)} P_{e,\ell})}{\operatorname{tr} S_B^{(\ell)}} \in [0,1]. \tag{22}$$

Proposition A.1 (Objective-alignment lower bound)**.** *Fix encoder e, layer ℓ, and downstream label y. Let $P = P_{e,\ell}$. Let $S_B^{(\ell)}$ and $S_W^{(\ell)}$ be downstream between-class and within-class scatter matrices, with $\operatorname{tr} S_B^{(\ell)} > 0$. Assume $P S_W^{(\ell)} P \preceq \sigma_\ell^2 P$. For any $\tau > 0$, define*

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

Then

$$\mathcal{F}_{P,\tau}^{(\ell)} \ge \frac{\alpha_{e,\ell}(y)}{\sigma_\ell^2 + \tau} \operatorname{tr} S_B^{(\ell)}. \tag{23}$$

Proof. Restrict all matrices to $\operatorname{im}(P)$, where P acts as the identity. The assumption implies $P S_W^{(\ell)} P + \tau P \preceq (\sigma_\ell^2 + \tau) P$. Loewner order reverses under inversion on the same positive definite subspace, so $(P S_W^{(\ell)} P + \tau P)^\dagger \succeq (\sigma_\ell^2 + \tau)^{-1} P$. Multiplying by the positive semidefinite matrix $P S_B^{(\ell)} P$, taking the trace, and substituting Equation 22 gives Equation 23. □

The proposition is a sufficient condition, not a causal claim that objective alone determines transfer. It explains why a pretraining-associated subspace can make a layer linearly useful when downstream between-class scatter lands in that subspace and projected within-class scatter remains controlled.

A.5. SAE and Transcoder Objectives

For each selected layer, the Top-k SAE decomposes activations as

$$f_\phi(h) = W_{\text{dec}} \operatorname{TopK}(W_{\text{enc}} h + b_{\text{enc}}, k) + b_{\text{dec}}, \tag{24}$$

where $W_{\text{enc}} \in \mathbb{R}^{4608\times 768}$ and $W_{\text{dec}} \in \mathbb{R}^{768\times 4608}$. Here TopK zeroes negative pre-activations and keeps the largest k positive activations. SAE reconstruction quality is measured by

$$R^2_{\text{SAE}} = 1 - \frac{\sum_i \|h_i - f_\phi(h_i)\|_2^2}{\sum_i \|h_i - \bar{h}\|_2^2}. \tag{25}$$

Across stored best-layer SAE summaries, reconstruction MSE has median 1.9×10^{-5} and mean 4.7×10^{-5} under the activation normalization used for training; the median reconstructed-probe change is -0.018 score points, and the dead-feature fraction is below 2%.

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

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

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

$$T_\psi(h^{(\ell)}) = W_{\text{out}}\,\text{TopK}(W_{\text{in}}h^{(\ell)} + b_{\text{in}}, k_t) + b_{\text{out}}, \quad (26)$$

with latent dimension 3840 and $k_t = 16$ for 12-layer encoders. Transcoder faithfulness is

$$R^2 = 1 - \frac{\sum_i \|h_i^{(\ell+1)} - T_\psi(h_i^{(\ell)})\|_2^2}{\sum_i \|h_i^{(\ell+1)} - \bar{h}^{(\ell+1)}\|_2^2}. \quad (27)$$

B. Extended Results

B.1. Pruning and Risk-Cost Selection

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

$$\ell_\epsilon^\star = \min\left\{\ell : m_\ell \geq \max_{j \leq L} m_j - \epsilon\right\}. \quad (28)$$

If $\ell_\epsilon^\star < L$, layers after $\ell_\epsilon^\star$ are candidates for skipping in that frozen-extraction workload. With cumulative cost C_ℓ, resource-aware utility is

$$U_\lambda(\ell) = m_\ell - \lambda C_\ell. \quad (29)$$

A pruned exit $\ell < L$ has higher utility than the final layer when

$$m_\ell - m_L > -\lambda(C_L - C_\ell). \quad (30)$$

For risk-aware selection, let $\kappa_\ell = \mathcal{PR}^{(\ell)}/d$ be the normalized activation rank and define

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

where D_ℓ is embedding drift, F_ℓ is alive-feature drift, and $A_\ell(\mathcal{G})$ is descriptor-family retention. Then ℓ dominates the final readout under this utility iff

$$\begin{aligned} m_\ell - m_L + \lambda_c(C_L - C_\ell) + \lambda_k(\kappa_L - \kappa_\ell) \\ + \lambda_D(D_L - D_\ell) + \lambda_F(F_L - F_\ell) + \lambda_A \Delta A_{\ell,L} > 0, \end{aligned} \quad (31)$$

where $\Delta A_{\ell,L} = A_\ell(\mathcal{G}) - A_L(\mathcal{G})$.

The pruning table is the operational version of 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. This is a validation signal, not a hardware latency measurement; throughput depends on the implementation.

Table 12. Non-speech pruning evidence averaged over ESC-50, GTZAN, and UrbanSound8K. Gain is best-layer score minus final-layer score.

Encoder	Best layers	Kept	Best	Gain
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

Table 13. Largest final-layer losses in the large/fused run. Loss is best-layer score minus final-layer score.

Encoder	Task	Best L	Final L	Best	Final	Loss
WavLM-large	ESC-50	9	24	.823	.690	.133
WavLM-large	GTZAN	18	24	.740	.640	.100
Whisper-large-v3	US8K	11	32	.860	.789	.071
Whisper-large-v3	GTZAN	17	32	.775	.710	.065
WavLM-large	US8K	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

B.2. Large/Fused Checkpoints and Bootstrap Uncertainty

B.3. Sparse, Robustness, and Intervention Audits

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

$$\text{MSFS-}\tau = \min\{|S| : \text{Acc}_{\text{ablate}(S)} \leq (1 - \tau/100)\,\text{Acc}\}. \quad (32)$$

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 (about eight features), consistent with narrow audio-text label-aligned sparse features. Whisper-S and WavLM-B require broader sets, reflecting dense or redundant encodings.

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. Embedding stability is measured as

$$\rho^{(\ell)} = \frac{1}{N}\sum_i \frac{f_\ell(x_i)^\top f_\ell(x_i + \delta_i)}{\|f_\ell(x_i)\|\,\|f_\ell(x_i + \delta_i)\|}. \quad (33)$$

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

Table 14. 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]
CLAP-HTSAT-fused	.94 [.81, 1.00]	.811 [.673, .921]
WavLM-large	.55 [.40, .71]	.789 [.757, .821]
Whisper-large-v3	.62 [.41, .86]	.817 [.763, .870]

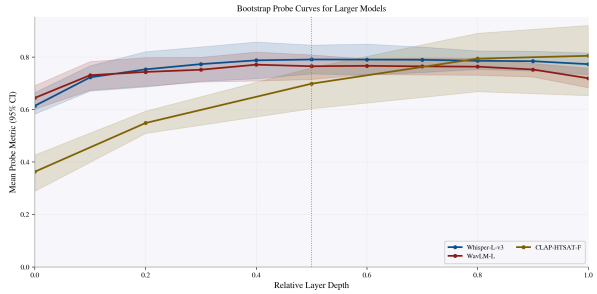

Figure 4. Bootstrap probe-depth profiles for large/fused checkpoints. Scale raises absolute scores, but depth dependence remains visible.

corruption. Alive-feature drift is

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

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

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

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. This is the linear margin condition after adding the corrupted representation drift. If probe directions lie in a task subspace U, so $w_{y_i} - w_c = UU^\top (w_{y_i} - w_c)$, then only the projected drift $UU^\top \tilde{\delta}_i^{(\ell)}$ affects the condition. This is why a model can have small global cosine drift but still lose score when noise moves examples along classifier-relevant directions.

For a linear probe $g_c(h) = w_c^\top h + b_c$ and SAE reconstruction $\hat{h} = W_{\text{dec}} a + b_{\text{dec}}$, ablating feature set S gives

$$\left| g_c(\hat{h}) - g_c(\hat{h}_{\setminus S}) \right| \leq \left\| W_{\text{dec},S}^\top w_c \right\|_2 \left\| a_S \right\|_2. \tag{37}$$

Thus a feature must be both class-selective and aligned with the probe readout direction to be a strong intervention handle; monosemanticity alone is not sufficient.

Table 15. 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.

Encoder	$\geq 60\%$	$\geq 70\%$	$\geq 80\%$	Mean 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 16. 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	$11 \to 12$	$11 \to 12$	$6 \to 7$	$11 \to 12$

C. Limitations and Reproducibility

The study is observational: objective, pretraining data, architecture, scale, and supervision source vary together. The benchmark panel is diagnostic rather than exhaustive, and pruning evidence is based on analyzed depth rather than measured end-to-end latency. Geometry proposes a no-label layer and should be validated when labels become available. Steering and feature ablation are inference-time diagnostics, not deployment policies or training-data deletion. SAE and transcoder summaries are conditional on dictionary size, sparsity, threshold, and seed.

All reported experiments use frozen encoders and linear downstream heads. The intended release contains scripts for activation extraction, probing, geometry/CKA computation, SAE and transcoder training, perturbation analysis, steering, ablation, table generation, and bootstrap summaries. The reported experiments were run on NVIDIA RTX A4500 20GB GPUs; the reproduction estimate is approximately 150–250 A4500 GPU-hours plus CPU time for probing and table generation.

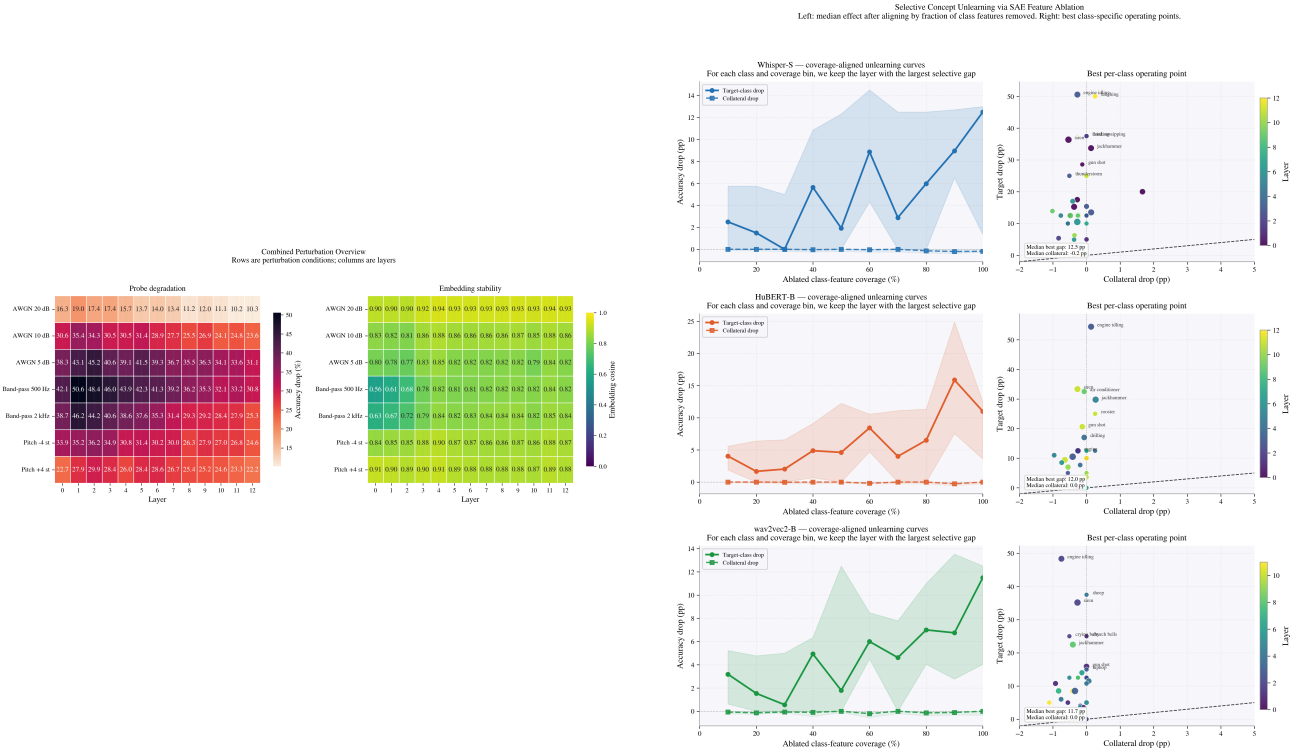

Figure 5. Intervention checks. Left: controlled perturbations separate clean score, embedding drift, and alive-feature drift. Right: targeted feature ablation diagnoses whether sparse features can suppress a target class without excessive retain-set movement.

Table 17. 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