Where to Read a Frozen Audio Encoder: Objective-Induced Geometry and Zero-Label Layer Selection

Arnesh Batra [1 2] **Aniket Khandelwal** [1 2 *] **Arush Gumber** [1 2 *] **Krish Thukral** [3 2 *]

Abstract

Pretrained audio encoders are increasingly reused across speech, environmental sound, music, and paralinguistic tasks, yet frozen-feature pipelines almost always read out the final layer by default. We show this default is often a major failure mode: across seven encoders, Whisper-S, CLAP-HTSAT, HuBERT-B, WavLM-B, UniSpeech-SAT, wav2vec2-B, and Data2Vec-Audio, and four audio domains, environmental sound (ESC-50), urban sound (UrbanSound8K), music genre (GTZAN), and speech emotion (CREMA-D), final-layer extraction can silently discard 24 to 38 score points relative to an earlier depth. The pretraining objective induces a predictable representation geometry across depth, and that geometry localizes the useful extraction layer well enough to select it with zero target-task labels: choosing the layer with highest average normalized isotropy and participation ratio improves character error rate on 11 of 12 low-resource ASR settings, a 17.0% average relative reduction in CER and 23.9% for wav2vec2-large across six languages. A counterintuitive regime is that low isotropy is useful when class scatter aligns with the low-rank objective subspace, so hard-unit and speaker-aware SSL encoders break the global "more isotropy is better" rule in a predictable way. QUICKLAYER, a weighted rule that adds a few-shot probe to the geometry score, recovers 83 to 89% of the avoidable last-layer gap, turning extraction depth into a measurable, label-cheap choice for any frozen audio pipeline.

[1]Indraprastha Institute of Information Technology Delhi (IIIT-Delhi) [2]StarkVision Research [3]Manipal University Jaipur. Correspondence to: Arnesh Batra <arnesh23129@iiitd.ac.in>.

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

1. Introduction

Frozen audio encoders are now standard across speech, sound, and music (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022b;c; Baevski et al., 2022; Radford et al., 2023; Chen et al., 2022a; Wu et al., 2023). The common recipe is to take the final layer, train a linear head, and report transfer accuracy. Benchmarks make it clear that the recipe works (Yang et al., 2021; Turian et al., 2022), but they hide a model-selection step. For wav2vec2-B and Data2Vec-Audio, the final layer loses .378 and .365 of ESC-50 accuracy relative to a layer near the input, and 11 of 12 low-resource ASR settings improve when extraction moves away from the final layer. Picking the layer is not a tuning detail. It is the largest single source of variance in frozen-feature transfer that we observe.

This paper makes the layer-selection step predictive rather than empirical. The central claim is that the pretraining objective induces a representation geometry across depth, and that geometry localizes the useful extraction layer well enough to choose it with zero target-task labels. The argument runs in three steps. (i) Final-layer extraction is a silent, common failure mode: across seven encoders and four diagnostic datasets, it can cost up to 38 points. (ii) Geometry, measured by participation ratio and isotropy, is a useful selection prior, and it organizes encoders into three falsifiable regimes. Regime 3, structured anisotropy, is the counterintuitive case: when masked-unit or speaker-aware objectives concentrate variance into a low-rank subspace, low isotropy is useful exactly when class scatter aligns with that subspace. (iii) The geometry prior already gives a zero-label selector that improves low-resource ASR by 17.0% average relative CER reduction, and combining it with a small few-shot probe (QUICKLAYER) recovers 83 to 89% of the avoidable gap to the full-data oracle layer.

The result holds across environmental sound (ESC-50), urban sound (UrbanSound8K), music genre (GTZAN), and speech emotion (CREMA-D) (Piczak, 2015; Salamon et al., 2014; Tzanetakis & Cook, 2002; Cao et al., 2014), so the same selection rule is usable by practitioners working with frozen encoders in any of these domains. Prior audio layer analyses showed that SSL speech encoders are strongly depth-dependent (Pasad et al., 2021), and prior geometry

work showed that anisotropy and effective dimensionality influence transfer (Ethayarajh, 2019; Raghu et al., 2021). We connect these to a practical, label-free selector and show that the predictive content of geometry is strong enough to use in deployment.

Contributions.

- A layer-wise transfer record for seven frozen encoders on four audio domains showing that final-layer extraction can lose 24 to 38 score points, which makes layer selection a primary model-selection choice, not a hyperparameter detail.
- A three-regime geometry taxonomy explaining when participation ratio and isotropy are useful as layer-selection priors and when structured anisotropy makes low isotropy the correct signal.
- A zero-label low-resource ASR result: selecting the geometry-best layer improves CER in 11 of 12 language-encoder pairs, a 17.0% average relative reduction in CER and 23.9% for wav2vec2-large across six languages.
- QUICKLAYER, a few-shot weighted selector that recovers 83 to 89% of the avoidable last-layer gap and is also a structured pruning point because layers after the selected one can be skipped.

2. Methods

Encoders and datasets. We freeze seven encoders: Whisper-S (244M parameters), CLAP-HTSAT (86M), HuBERT-B (95M), WavLM-B (95M), UniSpeech-SAT (94M), wav2vec2-B (95M), and Data2Vec-Audio (94M). All transformer encoders use 768-dimensional hidden states; CLAP-HTSAT exposes four hierarchical HTS-AT stages plus an input/stem representation, reported as indices 0–4 rather than speech-layer indices, while the other encoders expose layers 0–12. CLAP is omitted from common-depth QUICKLAYER aggregates, and its geometry comparisons are descriptive. Datasets are ESC-50 (50 classes), UrbanSound8K (10 classes), GTZAN (10 genres), and CREMA-D (6 emotions). We report accuracy except on CREMA-D, where we report macro-F1.

Probing and layer selection. For each encoder–layer pair, we train a linear classifier on frozen pooled representations. We also evaluate a lightweight layer selector that combines few-shot probe accuracy with a geometry prior; the weighted rule and shot counts are reported in Section 3.3.

Geometry. For layer representations $Z^{(\ell)} = \{z_i^{(\ell)}\}_{i=1}^N$, we compute participation ratio and isotropy.

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

Table 1. Best-layer transfer score with optimal layer in parentheses. CREMA-D is macro-F1; all others are accuracy.

Paradigm	Encoder	ESC-50	GTZAN	CREMA-D	US8K
Audio-text contr.	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 SSL	UniSpeech-SAT	.720 (2)	.695 (2)	.681 (12)	.785 (4)
Contrastive SSL	wav2vec2-B	.693 (2)	.715 (2)	.647 (2)	.751 (0)
Self-distill.	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 the analyzed 0–12 depth needed 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

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

$\mathcal{PR}$ measures spectral diversity (it equals d for a uniform eigenvalue spectrum and approaches 1 when variance collapses), and $\mathcal{I}$ measures angular spread of representations around the origin. These two scalars are the geometry prior used by the zero-label and QUICKLAYER selectors.

3. Results

3.1. Useful Depth and Final-Layer Failure

Table 1 shows that every neural encoder outperforms the classical baseline at its best layer, but no universal extraction depth exists. CLAP-HTSAT is dominant for environmental, urban, and music classification, peaking at stages 3–4 in the 0–4 HTS-AT extraction pipeline. Whisper-S peaks later, especially on CREMA-D, where layer 12 achieves .732 macro-F1. Speech transcription supervision creates a late, decoder-oriented hierarchy that transfers well to emotion but is not the same hierarchy learned by general audio-text supervision.

The SSL encoders split by target type. HuBERT-B, WavLM-B, UniSpeech-SAT, and wav2vec2-B usually peak in early layers, consistent with masked or contrastive objectives that make lower acoustic/phonetic structure linearly accessible early. Data2Vec is the most front-loaded: layer 0 is optimal on three of four tasks. Table 2 shows that the final layer can lose 24 to 38 score points relative to an earlier extraction layer.

Final-layer extraction is therefore actively harmful for several common encoders. The layer at which a downstream

task becomes linearly available is an observed structural property of the pretraining recipe and data domain, not a consequence of parameter count.

3.2. Geometry Provides Priors, Not Universal Rules

Across all encoder-layer-task combinations, participation ratio and isotropy are positively associated with probe score:

$$\mathrm{Corr}(\mathcal{PR}, \mathrm{score}) = 0.211, \quad \mathrm{Corr}(\mathcal{I}, \mathrm{score}) = 0.266.$$

The stronger isotropy correlation is expected because angular spread controls how easily a linear classifier can isolate class directions. Yet the global trend hides three regimes rather than one universal rule.

Table 3 is intentionally retrospective. The experiments assign encoders to regimes from observed correlations, while the objective cues state a falsifiable hypothesis one should record before applying the rule to a new model. Regime 1 is the clean case: more angular spread usually means more linearly accessible task structure. Regime 2 needs objective context because Whisper's late ASR directions and WavLM's denoising directions can be useful even when global spread is only weakly informative. Regime 3 is the crucial qualification. HuBERT and UniSpeech-SAT make representations less globally isotropic because discrete-unit and speaker-aware targets concentrate variance into acoustic, phonetic, and speaker directions. That looks worse under a scalar isotropy score, but it is useful when downstream classes align with the concentrated subspace. The operative question is not "is the layer isotropic?" but "does the low-rank objective subspace contain the class scatter?"

Zero-label low-resource ASR. The same geometry-prior idea generalizes to sequence recognition. With no target-language labels, selecting the layer with the largest average normalized isotropy and participation ratio improves 11 of 12 language-encoder pairs, giving a 17.0% relative CER reduction on average and a 23.9% relative CER reduction for wav2vec2-large across all six languages. The Data2Vec-Hausa failure is important: a high-geometry layer can retain broad acoustic variation while missing language-specific sequence structure, so geometry should propose an exit and a small validation set should confirm it before pruning. This zero-label selector is the most direct evidence that the pretraining-induced geometry localizes the useful extraction layer for a task whose labels were never seen during selection.

3.3. QUICKLAYER: Few-Shot Selection

We treat final-layer extraction as the standard extraction baseline because it is the default used in many frozen-feature pipelines, and Table 2 shows that it can be actively harmful. QUICKLAYER is a low-cost alternative. For each candidate depth it fits a few-shot linear probe on cached frozen activations, forms a geometry score from normalized participation ratio and inverse anisotropy, and selects the largest weighted score, $0.3\, g_\ell + 0.7\, \widehat{\mathrm{Acc}}_\ell$. Table 4 compares the weighted rule with its geometry-restricted and probe-only counterparts on the six encoders with a common 0–12 stack.

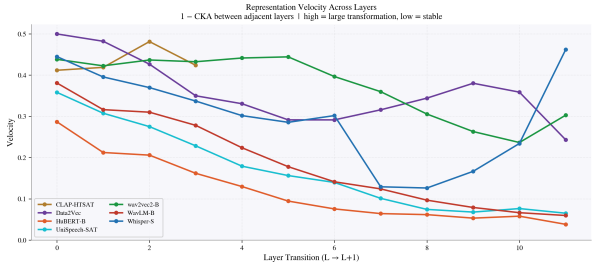

Figure 1. Representation velocity, $1 - \mathrm{CKA}$ between adjacent layers. SSL curves flatten after early acoustic layers, while Whisper-S keeps moving late as ASR supervision reshapes decoder-facing features.

The table answers a model-selection question, not an oracle-indexing question. Let $\rho = \sum_{e,d}(S_{e,d}(\hat{\ell}) - S_{e,d}(L)) / \sum_{e,d}(S_{e,d}(\ell^\star) - S_{e,d}(L))$, where L is the last layer, $\ell^\star$ is the full-data best layer, and $\hat{\ell}$ is the selected layer. Then ρ measures how much of the avoidable last-layer loss has been recovered. The weighted rule obtains $\rho = .83$ with one example per class and $\rho = .89$ with 50 examples per class, even though exact Hit@1 is only .167 and .333, respectively. For frozen-feature extraction, choosing a high-scoring neighboring layer is nearly as useful as naming the exact oracle layer.

The geometry-restricted selector already recovers 10.8 to 12.9 points over final-layer extraction, showing that representation shape can remove many bad exits before the small probe is applied. The probe-only selector becomes the strongest option when more labels are available. The weighted rule is the operating point we use: it remains within half a point of the probe-only score, while still biasing the search toward layers with broad, isotropic acoustic structure. At 50 shots, it adds +.221 on ESC-50, +.163 on GTZAN, +.105 on UrbanSound8K, and +.030 on CREMA-D. Broad acoustic tasks benefit from escaping late speech-specialized layers, while CREMA-D gains less because paralinguistic labels often require later speech structure.

3.4. Selected Layer as a Pruning Point

Layer selection is also structured pruning. For each downstream score profile $\{m_\ell\}$ and tolerance $\epsilon \geq 0$, define the earliest near-oracle exit $\ell^\star_\epsilon = \min\{\ell : m_\ell \geq \max_j m_j - \epsilon\}$. When $\ell^\star_\epsilon < L$ the analyzed layers after $\ell^\star_\epsilon$ can be skipped for that frozen-extraction workload. For a layer-cost profile C_ℓ with $C_\ell < C_L$ and any compute weight $\lambda \geq 0$, the early exit dominates the final layer in utility $U_\lambda(\ell) = m_\ell - \lambda C_\ell$ whenever $m_\ell - m_L > -\lambda(C_L - C_\ell)$; if $m_\ell \geq m_L$ any positive λ already makes the lower-cost early exit weakly preferable.

Table 3. Geometry-utility regimes, reported as empirical diagnostics. 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. Objective cues state what the loss suggests before inspecting correlations; the observed assignments are data-driven. CLAP has four hierarchical stages plus an input/stem representation, so its correlations are descriptive rather than directly comparable to 0–12 speech stacks.

Reg.	Objective cue	Observed assignment	Interpretation
R1	No hard unit bottleneck; 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 weaker 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 low-rank phonetic/speaker subspace.

Table 4. QUICKLAYER selector variants on the six common-depth encoders; CLAP-HTSAT is excluded because it exposes four hierarchical stages plus an input/stem representation. Gain is selected-layer score minus last-layer score, averaged over ESC-50, CREMA-D, GTZAN, and UrbanSound8K. H@1 is exact full-data oracle-layer recovery.

	Geometry pool		Probe		Weighted	
Shots	G↑	H@1	G↑	H@1	G↑	H@1
1	+0.108	0.167	**+0.124**	**0.208**	+0.121	0.167
5	+0.109	**0.167**	+0.111	**0.167**	**+0.114**	0.125
10	+0.121	**0.208**	**+0.124**	0.167	+0.122	0.125
20	+0.129	0.333	**+0.133**	**0.375**	+0.128	0.292
50	+0.129	0.417	**+0.134**	**0.542**	+0.130	0.333

Table 5. Exit policies by audio type. Gains are relative to final-layer extraction.

Audio type	Layer evidence	Exit policy
Events and scenes	CLAP stage 4; wav2vec2 gains .378/.246 on ESC-50/US8K at L2/L0.	Early-exit speech SSL.
Music and timbre	CLAP stage 3; wav2vec2/Data2Vec gain .285/.275 at L2/L0.	Prefer CLAP or low/mid SSL layers.
Speech affect	Whisper-S L12 gives .732; WavLM-B L6 gives .699.	Keep mid/late speech layers.
Low-resource ASR	Zero-label geometry reduces CER 23.9% for wav2vec2-large, 17.0% average.	Propose exit by geometry, then confirm CER.

The accuracy and cost terms point the same direction in every Paper A row of Table 2: the earlier extraction layer is faster *and* more accurate.

Figure 1 shows the structure behind these exits. Representation velocity, the per-layer $1 - \mathrm{CKA}$ between adjacent layers (Kornblith et al., 2019), is high near the input of every encoder while acoustic state is still being assembled, then flattens. Whisper-S keeps moving late as ASR supervision reshapes decoder-facing features, which matches its late best layer on CREMA-D. SSL encoders flatten earlier, which is why early-exit pruning recovers so much for them on non-speech targets. Table 5 turns this into an exit policy by audio type.

4. Discussion

Three findings are robust across the evaluated encoders and tasks. First, final-layer extraction is often suboptimal for frozen audio transfer, with losses of 24–38 points relative to earlier layers. Second, pretraining objectives induce predictable geometric structure across depth, and participation ratio plus isotropy provide practical layer-selection priors. Third, geometry-guided selection improves low-resource ASR even without target labels and combines naturally with few-shot probing. The selected layer also serves as an efficient early-exit point for frozen inference.

Limitations. The study is broad but observational: objective, pretraining data, and supervision source are entangled, so we report associations rather than isolated causal effects of objective alone. GTZAN and CREMA-D are compact benchmark tasks. The geometry prior is a prior, not a guarantee: the Data2Vec-Hausa CER row in the low-resource ASR result shows that a high-geometry layer can still miss language-specific sequence structure, which is why we recommend geometry to propose an exit and a small validation set to confirm it.

5. Conclusion

Audio encoders do not simply become more semantic with depth. Useful layers reflect the pretraining objective, and the objective leaves a geometric fingerprint across depth that is readable from frozen activations alone. Final-layer extraction can lose 24–38 points; geometry can already propose ASR exits with no target labels; and a small few-shot probe closes most of the avoidable gap. The result turns layer selection from a tuning artifact into a measurable, label-cheap step of the frozen-feature pipeline. An anonymized release accompanies the submission.

References

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.

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.

Ethayarajh, K. 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*, pp. 55–65. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1006.

Fisher, R. A. 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.

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.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

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.

Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., and Dosovitskiy, A. Do vision transformers see like convolutional neural networks?, 2021. URL https://arxiv.org/abs/2108.08810.

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.

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. Datasets, Encoders, and Implementation Details

A.1. Datasets and Splits

The four datasets 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 6. 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

A.2. Model Catalog

The model set is organized by pretraining recipe rather than by published aggregate rank, because Paper A's claim is that objective family, data domain, and supervision source are strongly associated with useful depth and geometry. Layer 0 denotes the first extracted hidden representation in the artifact pipeline. 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 7. 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. Probing Hyperparameters

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 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. Probes train on cached frozen representations. Accuracy is reported for ESC-50, UrbanSound8K, and GTZAN; macro-F1 is reported for CREMA-D. QUICKLAYER uses shots $\{1, 5, 10, 20, 50\}$ per class, geometry score g_ℓ from normalized participation ratio and inverse anisotropy, weighted score $0.3\, g_\ell + 0.7\, \widehat{\mathrm{Acc}}_\ell$, and only the six common-depth encoders in the main aggregate.

A.4. 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 weak average correlation with speech encoders, while masked/denoising SSL profiles are tightly aligned. Cross-encoder CLAP comparisons in the main paper should be read as functional stage comparisons.

Table 8. 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

B. Layer-Wise Performance Details

The optimal transfer layer and the most portable cross-dataset layer are both governed by the pretraining objective. 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 the front-loaded signature of 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, consistent with a representation organized for decoder cross-attention.

HuBERT-B, WavLM-B, and UniSpeech-SAT peak early on non-emotion tasks. The offline-unit masked-prediction objective forces class-separable phonetic or spectro-temporal structure to appear by layers 1–4; deeper layers often over-smooth the boundaries needed by environmental and urban classification. 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 three of four tasks.

B.1. Objective-Level Hypotheses

Table 9 states the objective-level hypotheses that the layer-depth observations test. The "observed signature" column is restricted to depth and geometry consequences; non-Paper-A signatures from the source row entries have been removed.

Table 9. Objective-level hypotheses tested by the layer-depth and geometry analysis. Restricted to depth/geometry signatures.

Objective family	Expected representation pressure	Observed depth signature
Audio-text contrastive	Make broad semantic categories linearly accessible.	Upper-stage CLAP peaks; strong clean transfer concentrated at stages 3–4.
Weak ASR supervision	Preserve decoder-useful speech content and prosody.	Whisper peaks mid-to-late, with the late best layer on CREMA-D.
Masked or denoising SSL	Recover acoustic units while ignoring missing or corrupted frames.	Early-to-middle useful layers.
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.

C. 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 $r = 0.99$, 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.

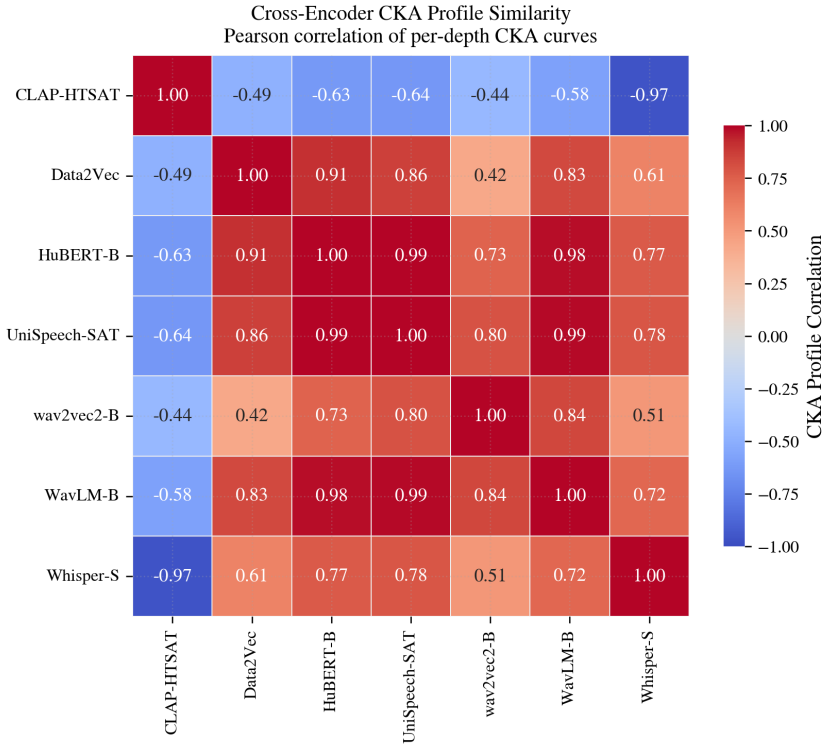

Figure 2. CKA depth-profile similarity across encoders. Each cell compares how probe-relevant information moves through depth rather than comparing one layer in isolation. Masked-prediction SSL encoders cluster tightly, Data2Vec remains near that family, wav2vec2-B is more distant because its contrastive target front-loads transfer, and CLAP-HTSAT follows a different audio-text depth ordering. This is the appendix version of the structural claim made in the main text: useful layer depth is organized by pretraining objective.

D. Objective-Induced Subspaces

The claim that pretraining objective organizes 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

$$\mathcal{S}_{e,\ell} = \operatorname{span}\left\{ \mathbb{E}[h_e^{(\ell)} \mid t_e = t] - \mathbb{E}[h_e^{(\ell)}] \;:\; t \in \mathcal{T}_e \right\}, \tag{3}$$

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

$$\alpha_{e,\ell}(y) = \frac{\operatorname{tr}\left(P_{e,\ell} S_B^{(\ell)} P_{e,\ell}\right)}{\operatorname{tr} S_B^{(\ell)}} \in [0,1]. \tag{4}$$

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. Paper A 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.

Lemma D.1 (Objective Alignment Controls Projected Separability)**.** *Assume a linear readout is restricted to $\mathcal{S}_{e,\ell}$ and projected within-class scatter is bounded as $P_{e,\ell} S_W^{(\ell)} P_{e,\ell} \preceq \sigma^2 P_{e,\ell}$ on that subspace. Then the projected Fisher signal satisfies*

$$\mathcal{F}_P^{(\ell)} = \operatorname{tr}\left[\left(P_{e,\ell} S_W^{(\ell)} P_{e,\ell}\right)^{\dagger} P_{e,\ell} S_B^{(\ell)} P_{e,\ell} \right] \geq \frac{\alpha_{e,\ell}(y)}{\sigma^2} \operatorname{tr} S_B^{(\ell)}. \tag{5}$$

Proof. On $\mathcal{S}_{e,\ell}$, the covariance bound implies $(PS_WP)^\dagger \succeq \sigma^{-2}P$, where $P = P_{e,\ell}$ and the pseudoinverse is restricted to the range of P. Therefore $\mathcal{F}_P^{(\ell)} \geq \sigma^{-2}\operatorname{tr}(PS_B^{(\ell)}P)$. Substituting the definition of $\alpha_{e,\ell}(y)$ gives the result. □

The lemma does not say that the pretraining objective alone determines performance. It says that an objective creates candidate subspaces, and downstream labels are easy when their between-class scatter lands there. This is why regime assignment is empirical and the objective is treated only as a hypothesis to be checked.

E. 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:

$$\mathcal{PR}^{(\ell)} = \frac{1}{\sum_{k=1}^{d}(\tilde{\lambda}_k^{(\ell)})^2} = \exp\Big(H_2(\tilde{\boldsymbol{\lambda}}^{(\ell)})\Big). \tag{6}$$

$\mathcal{PR}$ 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, not 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

$$\frac{1}{\binom{N}{2}}\sum_{i<j}\frac{(z_i^{(\ell)})^\top z_j^{(\ell)}}{\|z_i^{(\ell)}\|\,\|z_j^{(\ell)}\|} \approx \frac{\|\bar{z}^{(\ell)}\|^2}{\sigma^2 + \|\bar{z}^{(\ell)}\|^2}. \tag{7}$$

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

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

$$S_W^{(\ell)} = \sum_{c=1}^{C}\pi_c\,\mathbb{E}[(z^{(\ell)} - \mu_c^{(\ell)})(z^{(\ell)} - \mu_c^{(\ell)})^\top \mid y = c], \tag{8}$$

$$S_B^{(\ell)} = \sum_{c=1}^{C}\pi_c\,(\mu_c^{(\ell)} - \bar{\mu}^{(\ell)})(\mu_c^{(\ell)} - \bar{\mu}^{(\ell)})^\top. \tag{9}$$

The multiclass Fisher criterion is $\mathcal{F}^{(\ell)} = \operatorname{tr}[(S_W^{(\ell)})^{-1}S_B^{(\ell)}]$ (Fisher, 1936), which measures class-discriminative signal-to-noise.

Lemma E.1 (Regularized Geometry Lower-Bounds Fisher Separability)**.** *For $\tau \geq 0$, let $\hat{\Sigma}_\tau^{(\ell)} = \hat{\Sigma}^{(\ell)} + \tau I$ and $S_{W,\tau}^{(\ell)} = S_W^{(\ell)} + \tau I$. Define the regularized participation ratio*

$$\mathcal{PR}_\tau^{(\ell)} = \frac{\Big(\operatorname{tr}\hat{\Sigma}_\tau^{(\ell)}\Big)^2}{\|\hat{\Sigma}_\tau^{(\ell)}\|_F^2}.$$

If $S_{W,\tau}^{(\ell)} \preceq \gamma\hat{\Sigma}_\tau^{(\ell)}$ for $\gamma > 0$, then the ridge Fisher criterion $\mathcal{F}_\tau^{(\ell)} = \operatorname{tr}[(S_{W,\tau}^{(\ell)})^{-1}S_B^{(\ell)}]$ obeys

$$\mathcal{F}_\tau^{(\ell)} \geq \frac{\sqrt{\mathcal{PR}_\tau^{(\ell)}}}{\gamma\,\operatorname{tr}\hat{\Sigma}_\tau^{(\ell)}}\,\operatorname{tr}S_B^{(\ell)}. \tag{10}$$

When $\tau = 0$ and the covariances are nonsingular, this reduces to the unregularized Fisher bound used in the main-paper intuition.

Proof. The Loewner order reverses under inversion, so $(S_{W,\tau}^{(\ell)})^{-1} \succeq \gamma^{-1}(\hat{\Sigma}_\tau^{(\ell)})^{-1}$. Since $S_B^{(\ell)} \succeq 0$, $\mathcal{F}_\tau^{(\ell)} \geq \gamma^{-1}\operatorname{tr}[(\hat{\Sigma}_\tau^{(\ell)})^{-1}S_B^{(\ell)}]$. For positive definite A and positive semidefinite B, $\operatorname{tr}(A^{-1}B) \geq \lambda_{\max}(A)^{-1}\operatorname{tr}B$. Finally, $\lambda_{\max}(\hat{\Sigma}_\tau^{(\ell)}) \leq \|\hat{\Sigma}_\tau^{(\ell)}\|_F = \operatorname{tr}\hat{\Sigma}_\tau^{(\ell)}/\sqrt{\mathcal{PR}_\tau^{(\ell)}}$. Substitution gives the bound. □

Corollary E.2 (Condition-Number Refinement). *Let $\kappa_\tau^{(\ell)} = \lambda_{\max}(\hat{\Sigma}_\tau^{(\ell)})/\lambda_{\min}(\hat{\Sigma}_\tau^{(\ell)})$ be the regularized covariance condition number. Under the assumptions of Lemma E.1,*

$$\mathcal{F}_\tau^{(\ell)} \geq \frac{1 + (d-1)/\kappa_\tau^{(\ell)}}{\gamma \operatorname{tr} \hat{\Sigma}_\tau^{(\ell)}} \operatorname{tr} S_B^{(\ell)}. \tag{11}$$

Proof. Since every eigenvalue of $\hat{\Sigma}_\tau^{(\ell)}$ is at least $\lambda_{\max}(\hat{\Sigma}_\tau^{(\ell)})/\kappa_\tau^{(\ell)}$, $\operatorname{tr} \hat{\Sigma}_\tau^{(\ell)} \geq \lambda_{\max}(\hat{\Sigma}_\tau^{(\ell)})(1 + (d-1)/\kappa_\tau^{(\ell)})$. Thus $\lambda_{\max}(\hat{\Sigma}_\tau^{(\ell)})^{-1} \geq [1 + (d-1)/\kappa_\tau^{(\ell)}]/\operatorname{tr} \hat{\Sigma}_\tau^{(\ell)}$. Substituting into the same Loewner argument used in Lemma E.1 proves the claim. □

Isotropy helps not only by increasing effective dimension, but also by lowering the condition number. When $\kappa_\tau^{(\ell)} \approx 1$, the Fisher lower bound approaches the ideal $d/(\gamma \operatorname{tr} \hat{\Sigma}_\tau)$ scaling; when κ_τ is large, the bound falls back toward the weaker largest-eigenvalue control.

Table 10. 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	$+.211$	$+.266$

Regime labels are assigned from the observed correlations in Table 10; they are not used as inputs to the probes. The non-circular protocol for a future encoder is to record the objective-level cue before computing these correlations: hard discrete or speaker-aware prediction should suggest structured anisotropy; decoder-facing ASR or denoising objectives should suggest objective-gated geometry; and contrastive, audio-text, or self-distillation objectives without a hard unit bottleneck should more often reward broad effective dimension. The observed taxonomy fits these seven encoders but remains a falsifiable diagnostic for new encoders rather than a theorem about all audio pretraining objectives.

E.1. Structured Anisotropy

Discrete masked-prediction encoders are the main exception to the global geometry trend. If the target labels are offline k-means units, convergence encourages covariance of the form

$$\hat{\Sigma}^{(\ell)} \approx U_K \Lambda_K U_K^\top + \epsilon I_d, \tag{12}$$

where U_K spans the cluster-separating subspace. This reduces isotropy, but probe accuracy can remain high when downstream between-class scatter lies in the same subspace, $S_B^{(\ell)} \subset \operatorname{col}(U_K)$. UniSpeech-SAT adds speaker-aware directions $V_S \Omega_S V_S^\top$, further reducing isotropy without reducing linear separability. This structured anisotropy explains why HuBERT-B and UniSpeech-SAT can be useful despite negative isotropy–accuracy correlation.

Lemma E.3 (Structured Anisotropy Helps When Labels Align). *Let $U \in \mathbb{R}^{d \times K}$ have orthonormal columns and let $P = UU^\top$. Suppose a masked-unit objective induces covariance $\hat{\Sigma} = U\Lambda U^\top + \epsilon I_d$, with $\Lambda \succ 0$ and $0 < \epsilon \ll \lambda_{\min}(\Lambda)$. Define $B_U = U^\top S_B U$ and $W_U = U^\top S_W U$. If the projected within-class scatter satisfies $W_U \preceq \gamma(\Lambda + \epsilon I_K)$, then the Fisher signal available to linear readouts inside the objective subspace satisfies*

$$\mathcal{F}_U = \operatorname{tr}(W_U^{-1} B_U) \geq \gamma^{-1} \operatorname{tr}\left[(\Lambda + \epsilon I_K)^{-1} B_U\right]. \tag{13}$$

If $\alpha_U = \operatorname{tr}(PS_BP)/\operatorname{tr} S_B$, then the aligned subspace contributes an α_U fraction of between-class scatter. Structured anisotropy is useful when α_U is large and misleading when α_U is small.

Proof. The projected criterion is the Fisher criterion after restricting readout directions to $\mathrm{col}(U)$. The Loewner bound reverses under inversion on this subspace, so $W_U^{-1} \succeq \gamma^{-1}(\Lambda + \epsilon I_K)^{-1}$. Multiplying by $B_U \succeq 0$ and taking the trace proves Equation (13). The identity $\mathrm{tr}(B_U) = \mathrm{tr}(U^\top S_B U) = \mathrm{tr}(P S_B P)$ gives the scatter fraction α_U. □

For a candidate subspace size K, the spectral concentration score

$$q_K^{(\ell)} = \frac{\sum_{k=1}^{K} \lambda_k^{(\ell)}}{\mathrm{tr}\,\hat{\Sigma}^{(\ell)}}, \qquad \ell_K^\star = \arg\max_\ell q_K^{(\ell)}, \tag{14}$$

is a sharper zero-label prior than scalar isotropy in Regime 3 because it explicitly looks for the concentrated phonetic or speaker-aware directions that hard-unit objectives create. When K is unknown or the target may not align with that subspace, QUICKLAYER uses the few-shot probe to override the geometry prior.

Regime 3 is therefore not a failure of geometry. It is a failure of one scalar geometry metric to know the task. HuBERT-B and UniSpeech-SAT compress many examples toward shared unit or speaker directions, lowering pairwise angular spread. When a downstream label can be read from those directions, this compression improves sample efficiency. When the label requires information discarded by the unit or speaker target, the same anisotropy becomes harmful. This is why the main paper treats geometry as a prior to be combined with a small probe rather than as a standalone ranking rule.

F. Low-Resource ASR: Full Table

Table 11 reports the full zero-label low-resource ASR result that anchors the main-text claim. The geometry layer is chosen with no target-language labels 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, and the single loss is marked explicitly.

Table 11. Low-resource ASR geometry selection using only character error rate (CER; lower is better).

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	73.87	37.09	99.2% loss
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	–	48.05	51.88	7.4%
Mean	All	–	51.74	62.37	17.0%

This stress-tests the 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 17.0% relative CER reduction on average. Data2Vec is mixed, especially on Hausa, so the rule is not "always trust geometry." The Hausa row is not explained by script (Hausa is written in Latin script in the benchmark); the most plausible mechanism in the artifacts is objective mismatch: the high-geometry Data2Vec layer has broad acoustic spread, but the final layer preserves sequence information needed for that language's phonotactics. This is why geometry should propose an early exit and a small validation set should confirm it.

G. QUICKLAYER: Formalism and Full Table

For each layer, the weighted selector first forms $g_\ell = \frac{1}{2}\widetilde{\mathcal{PR}}_\ell + \frac{1}{2}(1 - \widetilde{\mathcal{A}}_\ell)$, where $\mathcal{A}_\ell$ is the anisotropy statistic used in the artifact tables. It then scores

$$s_\ell = \alpha g_\ell + (1-\alpha)\widehat{\mathrm{Acc}}^{(\ell)}, \quad \alpha = 0.3,$$

where $\widehat{\mathrm{Acc}}^{(\ell)}$ is a few-shot linear probe score. The geometry term is a low-variance prior over candidate depths; the probe term dominates the fixed weighted score. The method is most useful on tasks with large last-layer gaps, and less useful on CREMA-D where many encoders already peak late.

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

$$G_s(q) = \frac{1}{|\mathcal{D}||\mathcal{E}|}\sum_{d,e}\left(m_{d,e}^{(q_s)} - m_{d,e}^{(L_e)}\right), \tag{15}$$

with relative lift $\gamma_s(q) = G_s(q)/\bar{m}^{(L)}$. The denominator $\bar{m}^{(L)} = .611$ in our aggregate, so the weighted selector obtains .715, .709, .715, .721, and .722 for one, five, ten, twenty, and fifty shots per class respectively. A complementary recovered-gap fraction compares the selected layer with the full-data oracle,

$$R_s(q) = \frac{\sum_{d,e}\left(m_{d,e}^{(q_s)} - m_{d,e}^{(L_e)}\right)}{\sum_{d,e}\left(m_{d,e}^{(\star)} - m_{d,e}^{(L_e)}\right)}. \tag{16}$$

For the weighted selector, $R_s = .83, .78, .83, .88, .89$ for one, five, ten, twenty, and fifty shots per class. These diagnostics are stricter than last-layer gain because they compare the selected layer with the full-data oracle layer.

Proposition G.1 (Few-Shot Selector Stability)**.** *Let $s_\ell = \alpha g_\ell + (1-\alpha)\widehat{a}_\ell$ and $s_\ell^\star = \alpha g_\ell + (1-\alpha)a_\ell$, where $\widehat{a}_\ell$ is the observed few-shot probe estimate and a_ℓ is its expectation. If $|\widehat{a}_\ell - a_\ell| \le \eta$ for every layer, and $\hat{\ell} = \arg\max_\ell s_\ell$ while $\ell^\star = \arg\max_\ell s_\ell^\star$, then*

$$s_{\ell^\star}^\star - s_{\hat{\ell}}^\star \le 2(1-\alpha)\eta. \tag{17}$$

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 $1-\delta$,

$$s_{\ell^\star}^\star - s_{\hat{\ell}}^\star \le 2(1-\alpha)\sigma\sqrt{\frac{2\log(2(L+1)/\delta)}{s}}. \tag{18}$$

Proof. Because $\hat{\ell}$ maximizes the observed score, $s_{\hat{\ell}} \ge s_{\ell^\star}$. The uniform error bound implies $s_\ell^\star - (1-\alpha)\eta \le s_\ell \le s_\ell^\star + (1-\alpha)\eta$ for every ℓ. Therefore $s_{\ell^\star}^\star - (1-\alpha)\eta \le s_{\ell^\star} \le s_{\hat{\ell}} \le s_{\hat{\ell}}^\star + (1-\alpha)\eta$, which proves the claim. For the high-probability statement, a union bound over $L+1$ sub-Gaussian layer estimates gives $\max_\ell |\widehat{a}_\ell - a_\ell| \le \sigma\sqrt{2\log(2(L+1)/\delta)/s}$ with probability at least $1-\delta$. Substitute this for η in the deterministic bound. □

The proposition explains why QUICKLAYER remains useful in the one-shot regime: the geometry prior does not need to be perfect; it only has to stabilize the ranking while the probe estimate is noisy. As shots increase, η shrinks and the probe term naturally dominates.

Table 12 shows why score gap is the most application-relevant diagnostic: exact layer recovery can stay low when adjacent early-to-middle layers are statistically close, but the absolute metric deficit can still be small. ESC-50 is the cleanest case, with a score gap below one point at 50 shots. GTZAN improves from .0393 to .0164 as shot count increases. CREMA-D remains difficult because its oracle depths are late and speech-specific, while UrbanSound8K has several useful neighboring layers.

H. Non-Speech Early-Exit Pruning

Layer selection is also a model-compression result. For tolerance $\epsilon \ge 0$, define the earliest near-oracle exit $\ell_\epsilon^\star = \min\{\ell : m_\ell \ge \max_{j \le L} m_j - \epsilon\}$. When $\ell_\epsilon^\star < L$, all layers after $\ell_\epsilon^\star$ can be skipped for that frozen-extraction workload, or activations can be cached at that depth instead of at the final layer.

Table 12. QUICKLAYER weighted-score layer selection, full draft artifact table. Hit@1 is exact oracle-layer recovery. Geo-rank is the average rank of the full-data oracle layer under the geometry score; score gap is the absolute metric deficit to the oracle. Candidate count is the average number of layers considered, which is below 13 because CLAP-HTSAT exposes four hierarchical stages plus an input/stem representation while the other encoders expose layers 0–12.

Dataset	Shots	Hit@1	Geo-rank	Score gap	Cand.	Pool
CREMA-D	1	.1429	5.1429	.0238	11.8571	11.8571
CREMA-D	5	.1429	5.1429	.0370	11.8571	11.8571
CREMA-D	10	.1429	5.1429	.0310	11.8571	11.8571
CREMA-D	20	.1429	5.1429	.0240	11.8571	11.8571
CREMA-D	50	.1429	5.1429	.0184	11.8571	11.8571
ESC-50	1	.1429	5.5714	.0164	11.8571	11.8571
ESC-50	5	.1429	5.5714	.0150	11.8571	11.8571
ESC-50	10	.2857	5.5714	.0111	11.8571	11.8571
ESC-50	20	.2857	5.5714	.0129	11.8571	11.8571
ESC-50	50	.4286	5.5714	.0089	11.8571	11.8571
GTZAN	1	.2857	7.4286	.0393	11.8571	11.8571
GTZAN	5	.1429	7.4286	.0471	11.8571	11.8571
GTZAN	10	.1429	7.4286	.0271	11.8571	11.8571
GTZAN	20	.5714	7.4286	.0186	11.8571	11.8571
GTZAN	50	.4286	7.4286	.0164	11.8571	11.8571
UrbanSound8K	1	.2857	7.7143	.0133	11.8571	11.8571
UrbanSound8K	5	.2857	7.7143	.0184	11.8571	11.8571
UrbanSound8K	10	.1429	7.7143	.0213	11.8571	11.8571
UrbanSound8K	20	.2857	7.7143	.0124	11.8571	11.8571
UrbanSound8K	50	.2857	7.7143	.0209	11.8571	11.8571

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

Proof. The first claim follows directly from the definition of $\ell_\epsilon^\star$. If $m_{\ell_\epsilon^\star} > m_L$, then the early exit has higher downstream score than the final layer. The depth fraction follows because only layers $0, \ldots, \ell_\epsilon^\star$ are required for extraction. □

For resource-aware inference, define the cost-aware utility $U_\lambda(\ell) = m_\ell - \lambda C_\ell$, where C_ℓ is the cumulative cost of running layers $0, \ldots, \ell$ and $\lambda \geq 0$ expresses the value of latency, memory, or energy.

Corollary H.2 (Early Exit Can Strictly Dominate the Final Layer)**.** *An early exit $\ell < L$ has higher application utility than the final layer when $m_\ell - m_L > -\lambda(C_L - C_\ell)$. In particular, if $m_\ell \geq m_L$, any positive compute weight $\lambda > 0$ makes the lower-cost early exit weakly preferable, and strictly preferable whenever $C_\ell < C_L$.*

Proof. $U_\lambda(\ell) - U_\lambda(L) = m_\ell - m_L + \lambda(C_L - C_\ell)$. The early exit dominates exactly when this is positive. The special case follows because $C_L - C_\ell > 0$. □

Table 13. Non-speech early-exit evidence. Scores are averaged over ESC-50, GTZAN, and UrbanSound8K. 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

Table 13 is the strongest applied argument against final-layer extraction. 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. Pruning is not only a speed optimization; it can be an accuracy optimization when the pretraining objective over-specializes late layers.

I. Cross-Dataset Layer Transfer

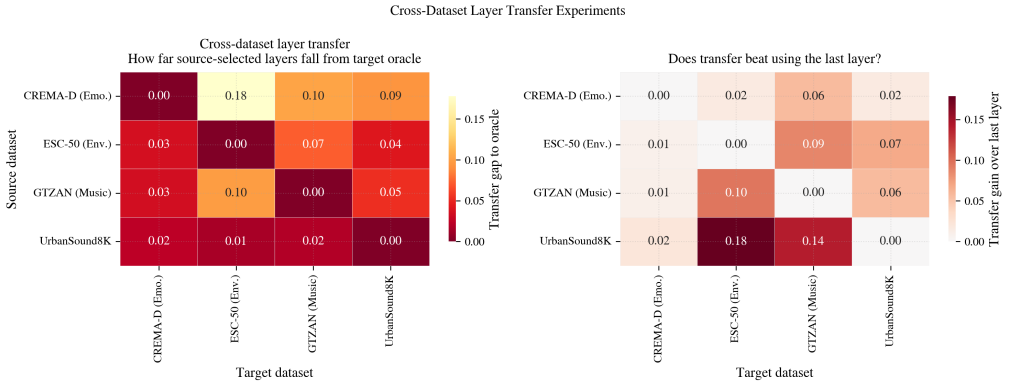

Figure 3. Cross-dataset layer transfer. Each source dataset selects a layer without seeing target labels, then that depth is evaluated on the remaining tasks. Source-selected layers often improve on last-layer extraction because early and middle layers retain reusable spectral-temporal structure; CREMA-D transfers less well because speech emotion pushes selection toward later paralinguistic layers.

UrbanSound8K is the most reliable source for transfer because its selected layers sit early-to-middle, where general spectral-temporal structure is still available. CREMA-D is a weaker 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. This supports the central hierarchy: early-to-middle layers are broadly acoustic, while late emotion or transcription layers are more task-specific. UrbanSound8K-selected layers gain 11.4 points over last-layer extraction while staying within 1.8 points of the best target-supervised depth; CREMA-D-selected layers trail the best target depth by 12.4 points because emotion supervision favors later, speech-specific abstractions.