COMPADAPT: WHEN AUDIO REPRESENTATION ADAPTATION HELPS FOR NEW SOUND COMBINATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio models encounter familiar sounds in combinations absent from downstream training. We introduce CompAdapt, a prototype-guided map that adapts a frozen audio representation. Separate recordings define class-mean prototypes; their sums guide a nonlinear residual map while an anchor keeps it close to the original representation. The fitted map preserves representation width, serves separately trained downstream models, and needs only audio at inference. Across four encoders on ESC-50, it improves mean separation of withheld class combinations by 0.16–0.89 dB at a common 60-epoch endpoint. With each separator's calibration-selected checkpoint, differences from the original representation range from -0.028 to $+0.042$ dB. An optional low-rank student retains selected full-map performance within ± 0.15 dB on all four encoders with 86.8–96.1% fewer coefficients. On natural SONYC-UST recordings, source-fitted maps improve macro average precision by 0.15–0.19 percentage points in three corrected comparisons. Together with alternative adaptors and cross-dataset reuse, these results show when a reusable representation map helps and how its benefit depends on downstream training. Code and numerical artifacts are available anonymously.

1 INTRODUCTION

Real-world audio often contains several sounds at once. A downstream training set may contain dogs, rain and keyboards without containing their joint mixture. The 50 classes of ESC-50 (Piczak, 2015) already form 19,600 three-class combinations. Learning from a limited set of mixtures therefore raises a question: can information about familiar classes help a model handle their other combinations? This is systematic generalisation, the reuse of known parts in new configurations (Fodor & Pylyshyn, 1988). Benchmarks designed around this question show that familiar components alone do not ensure success on withheld combinations (Lake & Baroni, 2018; Keysers et al., 2020).

Existing methods approach mixtures at different levels. Separators learn to recover waveforms, with source references or mixture-based objectives (Luo & Mesgarani, 2019; Wisdom et al., 2020). SoundBeam and AudioSep use a class label, example recording or text query to specify a source to extract (Delcroix et al., 2023; Liu et al., 2025). These methods build source knowledge into a separation system rather than a representation map shared by several tasks. SSLAM trains the encoder on mixtures (Alex et al., 2025); CARL keeps its encoder frozen but learns source-centred vectors (Sridhar & Cartwright, 2025). Adapters and low-rank adaptation (LoRA) make parameter-efficient changes inside a pretrained model (Houlsby et al., 2019; Hu et al., 2022). Our question is how to use known sound classes to adapt the output of an already-trained encoder, without changing its weights or representation width.

Our proposal is to describe each known class once and reuse that description in every mixture that contains it. A class prototype is the average encoder feature of separate recordings of the class, as in prototype-based classification (Snell et al., 2017), and the sum of the prototypes of a mixture's classes is its target. A target therefore exists for any mixture of known classes, including combinations absent from downstream training. We hypothesise that moving mixture features toward these targets, while keeping them close to the original features, makes new combinations easier for downstream models to handle. The sum is a training target, not an assumption that encoders are already additive (Chen et al., 2026).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

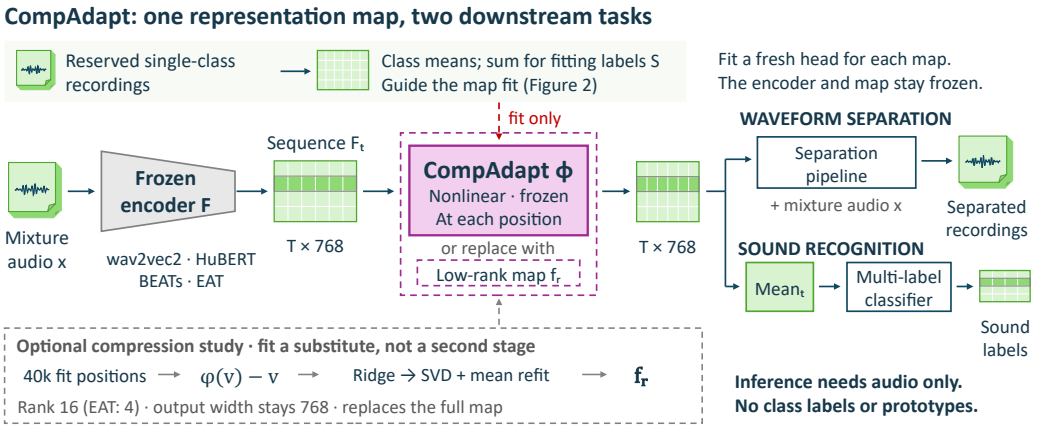

Figure 1: CompAdapt's nonlinear map is fitted using class prototypes, then frozen and reused by separately fitted downstream systems. Sound recognition pools after mapping. Optional compression learns the full map's displacements and replaces, rather than follows, the full map. Inference uses audio only.

We implement this proposal as **CompAdapt**, a small nonlinear residual map placed between a frozen audio encoder and the downstream model (Figure 1). It is fitted once from the prototype targets and then frozen; it keeps the size and format of the encoder's features and needs only audio at test time. For deployments that favour a smaller map, an optional compression step fits a low-rank affine *student* to reproduce the full map's changes.

We test the adapted features on separating a mixture into its source waveforms and on recognising which classes it contains. *In-combination* (IC) test mixtures reuse class combinations from downstream training, whereas *out-of-combination* (OOC) mixtures place the same known classes in withheld combinations; both use unseen recordings. We use four frozen encoders, the speech models wav2vec2 and HuBERT and the general audio models BEATs and EAT (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2023; 2024), reuse ESC-50 maps on natural urban recordings from SONYC-UST (Cartwright et al., 2019), and repeat separation on FSD50K (Fonseca et al., 2022).

The results distinguish adaptation from compression. CompAdapt improves separation at a common 60-epoch endpoint, while calibration-selected separators on the original representations perform similarly (Section 5). The student retains selected full-map performance for both speech encoders and both general-audio encoders, using rank 16 for the first three and rank 4 for EAT. It is therefore a smaller replacement for the fitted map. Sound-recognition comparisons with direct fitting and label-supervised adaptation test the same representations on a different task.

Our contributions are (i) CompAdapt, which adapts a frozen audio representation with separate recordings of known classes, without retraining the encoder or changing its interface; (ii) a protocol that fits the map before downstream training and reports fixed-budget and calibration-selected results for separation, recognition, alternative recipes and reuse; and (iii) a measurement of how much of the map's performance an optional low-rank student retains.

2 Related work

Mixture-aware and source-centred audio representations. SSLAM uses mixed audio in self-supervised pretraining: a student encoder receives the mixture, and its output is supervised by the average of a teacher encoder's features computed on each source recording separately (Alex et al., 2025). Its targets are therefore features of the particular recordings in each mixture, and the encoder is updated. CARL keeps an AudioMAE encoder frozen and trains a slot-based module that outputs one representation per source (Sridhar & Cartwright, 2025); AudioSlots and composable audio latents pursue related ideas with generative models (Reddy et al., 2023; Bindi & Esling, 2024).

CompAdapt keeps both the encoder and the representation format, and its targets are class prototypes shared by every mixture that contains those classes.

Systematic generalisation and compositional structure. Whether known sound classes can be handled in combinations absent from downstream training is a question of systematic generalisation, which asks whether known parts can be reused in new combinations (Fodor & Pylyshyn, 1988; Lake & Baroni, 2018; Keysers et al., 2020). Other work tests whether learned representations can be reconstructed from attributes or show additive structure in audio (Guo et al., 2025; Chen et al., 2026). These tests describe a fixed representation; we use class structure to change a frozen one and measure the effect through separation and recognition.

Adapting and compressing pretrained models. Adapters, LoRA, feature scaling and shifting (SSF) and bias-only tuning (BitFit) modify a pretrained model through a few parameters (Houlsby et al., 2019; Hu et al., 2022; Lian et al., 2022; Ben-Zaken et al., 2022), and linear probing followed by fine-tuning (LPFT), weight-space ensembling (WiSE-FT), L2-SP and side-tuning adapt a model while keeping its pretrained behaviour (Kumar et al., 2022; Wortsman et al., 2022; Li et al., 2018; Zhang et al., 2020). These methods specify how pretrained parameters are adapted; the fitting objective is a separate choice. CompAdapt uses class-prototype targets before downstream training. Section 7 compares representation-map versions of these recipes and in-weight LoRA fitted to the same prototype targets. Because even information-preserving coordinate changes can alter downstream training (Balles & Hennig, 2018; Zhang et al., 2025), every map receives its own newly trained downstream model.

3 CompAdapt: prototype-guided adaptation

CompAdapt changes the features of a frozen audio encoder before they reach a downstream model. We place it there for three reasons: the encoder is not retrained; its representation width and sequence interface are preserved; and the map can be fitted before choosing a downstream task. Separators and classifiers can keep their architectures, but their parameters are trained separately for each map. We first fit and freeze the map, then train the downstream models (Figure 2, Algorithm 1).

For an input mixture x, the encoder produces T feature vectors $F(x) = (F_1, \dots, F_T)$ with $F_t \in \mathbb{R}^d$ and $d = 768$; positions are time frames for wav2vec2 and HuBERT and flattened time–frequency patches for BEATs and EAT. The pooled representation of a recording is its mean feature, $z(x) = T^{-1} \sum_{t=1}^{T} F_t$. Our reference system uses the original, unmodified encoder features; we call it the *identity map* ("base" in the tables), because it passes $F(x)$ to the downstream model unchanged.

Prototype targets. For each sound class c, we reserve single-class recordings $\mathcal{P}_c$ from recording groups separate from the fitting, calibration and test mixtures. The class prototype is their mean pooled feature, $a_c = |\mathcal{P}_c|^{-1} \sum_{u \in \mathcal{P}_c} z(u)$. A fitting mixture that contains the set S of $k = |S|$ classes receives the target

$$q_S = \sum_{c \in S} a_c.$$

For a dog–rain–keyboard mixture, the target is the sum of these three prototypes, whichever recordings were mixed, so targets are shared across combinations instead of being learned for each one. In the sound-recognition comparison, these labelled recordings are also supplied to every source classifier and the BCE-map fit, matching source-label access.

Prototype-guided map. The full CompAdapt map is a residual network with one hidden layer of 128 units:

$$\phi(z) = z + W_2 \,\mathrm{ReLU}(W_1 z + b_1) + b_2, \tag{1}$$

with trainable $W_1 \in \mathbb{R}^{h \times d}$, $b_1 \in \mathbb{R}^h$, $W_2 \in \mathbb{R}^{d \times h}$ and $b_2 \in \mathbb{R}^d$, $h = 128$; ReLU sets negative values to zero. We fit these parameters by minimising

$$\mathcal{L}_{\mathrm{map}} = \mathbb{E}_{\mathrm{fit}} \left[\|\phi(z) - q_S\|_2^2 + \beta \|\phi(z) - z\|_2^2 \right], \qquad \beta = 0.3, \tag{2}$$

where the expectation is over fitting mixtures. The first term moves each mixture toward the target for its classes. The second anchors the result to the original feature, so that, with the residual

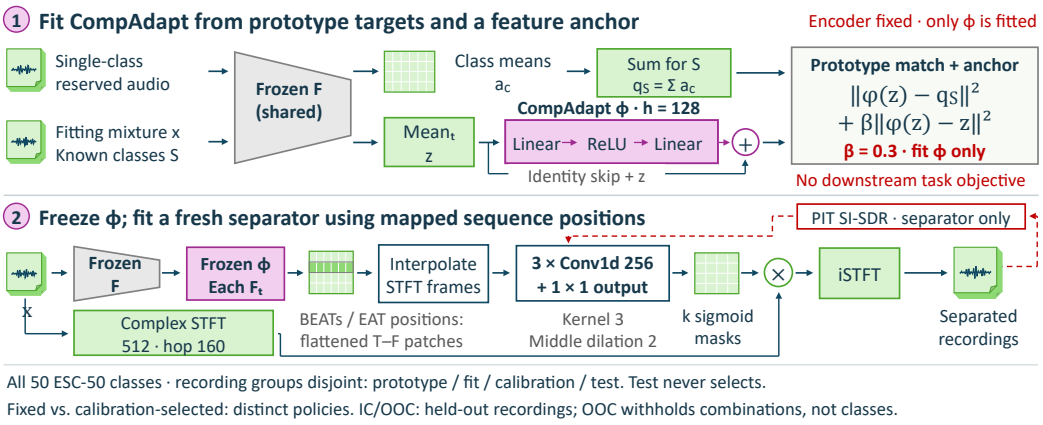

Figure 2: Fitting the CompAdapt map and using it for separation. (1) Class means give each fitting mixture its target q_S; only ϕ is fitted, with the prototype term and the anchor. (2) With ϕ frozen, a new separator predicts masks from the mapped sequence and is trained with permutation-invariant SI-SDR. Calibration mixtures, never test mixtures, select checkpoints.

connection, the map learns a moderate nonlinear change to an existing representation. No term uses separated waveforms, classifier outputs or any other downstream objective. The map is fitted on pooled vectors but applied to every position F_t; because ϕ is nonlinear, mapping each position and then averaging differs from mapping the average.

Downstream separation and recognition. Each map, including the identity map, receives its own newly trained downstream model. For separation, the mapped sequence is interpolated to the frame rate of the mixture's short-time Fourier transform (STFT), and a one-dimensional convolutional network predicts one mask per source. It is trained on the source recordings with the permutation-invariant, scale-invariant signal-to-distortion ratio (SI-SDR) loss, which lets sources appear in any order (Yu et al., 2017; Le Roux et al., 2019); architecture, mixtures, initialisation, training order and optimiser are the same for every map. For recognition, the mapped sequence is averaged and a multi-label classifier predicts which classes are present; on natural recordings, the source-fitted map is reused unchanged and only the classifier is fitted to target labels.

Optional compression: the low-rank student. To obtain a smaller map, we follow teacher–student distillation (Romero et al., 2015), with the full map as the teacher, and fit a low-rank affine *student* to reproduce how ϕ changes individual feature vectors. Its regression targets are the displacements $\phi(v) - v$ at 40,000 sequence positions $v = F_t$ sampled from each run's fitting mixtures; a ridge solution is truncated by singular-value decomposition and its mean correction refitted:

$$f_r(v) = v + M_r(v - \mu_X) + c_r, \qquad \operatorname{rank}(M_r) \le r, \tag{3}$$

where $\mu_X \in \mathbb{R}^d$ is the mean sampled input, $M_r \in \mathbb{R}^{d\times d}$ the retained linear change and $c_r \in \mathbb{R}^d$ the refitted shift. The student still outputs all d coordinates; the rank limits the change, not the width. Ranks fixed in an earlier validation study (Appendix A.4) are $r = 16$ for wav2vec2, HuBERT and BEATs and $r = 4$ for EAT, giving 26,128 or 7,684 stored coefficients instead of the full map's 197,504 parameters (86.77% or 96.11% fewer). At inference, the student replaces ϕ and needs neither ϕ nor the prototype recordings. As a reference, the *direct low-rank fit* has the same rank and width but is fitted to the displacement implied by the prototype and anchor terms, $(q_S - z)/(1+\beta)$, on a run's 576 pooled mixture vectors with a rank-constrained ridge solver. Since it differs from the student in rows, solver and target, comparing the two does not isolate the need for the full map. We also compare a direct fit using the student's exact sampled positions, ridge, rank and truncation (Appendix H).

Algorithm 1 Fitting and evaluating CompAdapt. The map and optional student use representation targets; comparison maps use their stated objectives. Map-only reuse fits a target downstream model. Complete-pipeline reuse performs no target fitting or selection.

Require: Frozen encoder F; single-class recordings; labelled mixtures; disjoint combination splits

1: Compute prototypes a_c and mixture targets $q_S = \sum_{c \in S} a_c$.
2: Fit ϕ on pooled fitting features by minimising Equation 2; freeze ϕ.
3: *Optional compression:* fit the student f_r to displacements $\phi(F_t) - F_t$.
4: Fit the comparison maps with their own targets.
5: **for** each map f (identity, ϕ, f_r and comparison maps) **do**
6: Freeze F and f; fit downstream candidates or reuse a complete frozen source pipeline.
7: For fitted candidates, select on calibration only and retain declared fixed endpoints.
8: **end for**
9: Seal all choices, then evaluate on IC/OOC mixtures or the natural-audio test set.
10: Report paired differences with the declared corrected tests.

4 Evaluation protocol

We judge each map by the downstream performance it enables, training a new downstream model for every map (Algorithm 1). For separation, let $U_D(f)$ be the mean scale-invariant signal-to-distortion ratio improvement (SI-SDRi) obtained with map f on evaluation set D: the SI-SDR of the separated sources minus that of the unprocessed mixture, in dB. The benefit of a map is its difference from the identity map:

$$\Delta_D(f) = U_D(f) - U_D(\text{identity}), \qquad J(f) = \Delta_{\text{ooc}}(f) - \Delta_{\text{ic}}(f). \tag{4}$$

Our main separation measure is Δ_{ooc}, the benefit on unseen combinations; J only compares the effects on OOC and IC mixtures. For recognition, U_D is macro average precision (AP), the average precision of each label averaged equally over labels, with differences in percentage points.

Main study. We use all 50 ESC-50 classes (Piczak, 2015) with frozen wav2vec2, HuBERT, BEATs and EAT encoders (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2023; 2024). The wav2vec2 export is the last layer of `WAV2VEC2_ASR_BASE_960H`, fine-tuned for automatic speech recognition; HuBERT is pretrained by masked prediction for speech. BEATs and EAT are general-audio encoders with flattened patch outputs. Each class supplies six prototype recordings, and each mixture combines three recordings. Recording groups are kept separate for prototype construction, fitting, checkpoint selection and testing. All classes occur in fitting, while the calibration and OOC class combinations are withheld from both map and downstream fitting. Each run has 576 fitting, 240 calibration, 576 IC and 480 OOC mixtures from pairwise-disjoint sets of 48 fitting, 20 calibration and 40 OOC class combinations; IC reuses the fitting combinations with test recordings. For each encoder, 16 paired runs, which vary the recording and combination draws and the initialisation, compare the identity map, the full CompAdapt map, its student and the direct low-rank fit.

Checkpoint selection. Separators train for 180 epochs, and the same 13 checkpoints are kept for every system. Each system selects its own checkpoint on the calibration mixtures only; test sets are never used for selection. We also report epochs 60 and 180, where all systems have equal training.

External study. An independent eight-run study repeats the separation evaluation with wav2vec2 and EAT on a verified 36-class subset of FSD50K with 1,499 recordings (Fonseca et al., 2022). ESC-50 overlaps and exact duplicates are removed, and uploader groups are kept separate across data roles. Mixture counts match the main study; separators train for 90 epochs with ten checkpoints. FSD50K mixtures combine field recordings, not isolated sound events (Appendix C).

Sound-recognition study. This study asks whether the same maps help a multi-label classifier identify which classes a recording contains. It compares five maps on the four encoders: identity, the full CompAdapt map, its student, the direct low-rank fit, and a *BCE map*, which has CompAdapt's residual architecture but is trained on source class labels with a binary cross-entropy (BCE) loss and the same anchor weight. Test data are controlled ESC-50 OOC mixtures and natural ten-second urban recordings from SONYC-UST-V2 (Cartwright et al., 2019; DCASE Community, 2020), with

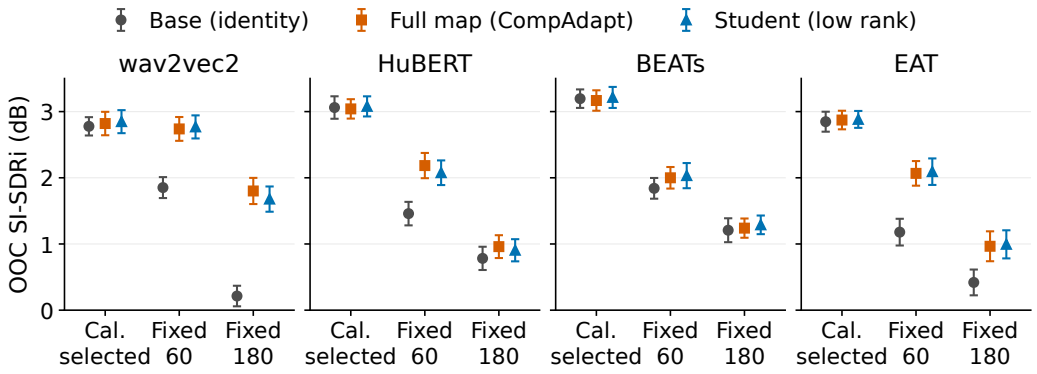

Figure 3: OOC separation with the identity map (base), the full CompAdapt map and its student, one panel per encoder: each system's calibration-selected checkpoint and fixed epochs 60 and 180. Points are mean SI-SDRi over 16 paired runs with ordinary 95% t intervals.

Table 1: Grouped ESC-50: paired OOC SI-SDRi differences (dB), 16 runs. Each map has its own calibration-selected separator; epoch 60 is descriptive. Full map is CompAdapt. †: retention within ± 0.15 dB (primary Holm4); $*$: positive contrast (secondary Holm12). Intervals are ordinary paired intervals; full inference: Appendix B.

	Calibration-selected checkpoints				Epoch 60
Encoder (student rank)	Full map − base	Student − base	Student − full map [90% CI]	Student − direct	Full map − base
wav2vec2 (16)	+0.042	+0.071	+0.028 [−0.016, +0.073]†	+0.139*	+0.886
HuBERT (16)	−0.021	+0.017	+0.038 [−0.006, +0.082]†	+0.132*	+0.726
BEATs (16)	−0.028	+0.018	+0.047 [−0.010, +0.103]†	+0.050	+0.159
EAT (4)	+0.025	+0.035	+0.010 [−0.031, +0.051]†	+0.010	+0.888

eight coarse labels and 13,538 fitting, 538 calibration and 664 test recordings from 48 test sensors. All source classifiers see the same mixture labels and labelled prototype recordings. Each map has six classifier candidates, a linear model or a 128-unit tanh multilayer perceptron (MLP), each with three regularisation strengths; we select by calibration macro AP over 16 paired runs.

Statistical tests. Comparisons are paired within each encoder. For the main separation study, the primary family has four student–full-map equivalence tests (margin ± 0.15 dB) and the secondary family 12 tests of gains over identity and the direct fit. For recognition, 16 primary tests compare the full map and student with identity, and 32 secondary tests cover student–full-map retention (margin ± 1 AP point) and comparisons with the direct and BCE maps. Holm correction at level 0.05 is applied within each prespecified family; the FSD50K study is tested separately. We report ordinary paired intervals alongside corrected decisions (Appendices B, C and D).

5 Separation of new sound combinations

We first use the adapted features to separate three-source mixtures. At the common 60-epoch endpoint, the full map improves mean OOC SI-SDRi by 0.886, 0.726, 0.159 and 0.888 dB for wav2vec2, HuBERT, BEATs and EAT, respectively, with a frozen encoder and the same separator training budget. IC mixtures show gains of similar size (0.254–0.830 dB; Appendix B), so at this endpoint the map helps familiar and new combinations alike. The student gains 0.192–0.917 dB. Figure 3 compares the fixed endpoints with calibration-selected checkpoints.

Table 2: Deployment on one RTX 3090: medians of 40 paired timing rounds. Map-only latency uses batch 8; complete waveform-to-waveform latency uses batch 1 with loaded ten-second, 16-kHz audio. Overhead uses unrounded times relative to identity.

	Map (ms)	Pipeline (ms)	
Encoder	Full / student	Identity / student	Overhead
wav2vec2	0.348 / 0.311	15.738 / 15.817	0.51%
EAT	0.344 / 0.307	14.984 / 15.259	1.83%

Choosing the downstream checkpoint. With independent calibration-based selection, full-map-minus-identity differences become $+0.042$, -0.021, -0.028 and $+0.025$ dB, and student differences range from $+0.017$ to $+0.071$ dB; none of these eight comparisons with identity passes the secondary correction. On EAT, selection improves the identity system by 1.669 dB relative to epoch 60 and the full-map system by 0.806 dB, so their difference falls from 0.888 to 0.025 dB. All selections occur by epoch 60, so identity receives no extra training. After its selected checkpoint, each of the 256 trajectories keeps improving on its fitting mixtures while its OOC score declines, and the mean decline to epoch 60 is smaller with the map on all four encoders (Appendix B). Thus the common 60-epoch policy benefits from adaptation, while calibration-selected systems perform similarly. Earlier clip-disjoint studies find that translations, rotations and random or constant targets also improve fixed-endpoint separation (Appendix F), so this gain cannot be attributed to the class assignment of the targets alone.

Compressing the map. At the calibration-selected checkpoints, student-minus-full-map differences are $+0.028$, $+0.038$, $+0.047$ and $+0.010$ dB for wav2vec2, HuBERT, BEATs and EAT, and all four 90% intervals lie inside the prespecified ± 0.15 dB margin, so all four equivalence tests pass (Table 1). Retention covers both speech and both general-audio encoders, including EAT at rank 4. The affine student can therefore replace the nonlinear map with 86.8–96.1% fewer coefficients and the same output width. This is selected-checkpoint retention; Appendix B gives the two fixed-endpoint exceptions. The student also exceeds the pooled direct low-rank fit by 0.139 dB on wav2vec2 and 0.132 dB on HuBERT after secondary correction, which favours distillation as a way to obtain a low-rank map; because the two routes also differ in training rows, target and solver, this is a comparison of fitting procedures.

Matched fitting and deployment. A follow-up uses the student's exact 40,000 sampled positions, rank, ridge and truncation for direct fitting. At calibration-selected checkpoints, student-minus-direct means are $+0.114$ dB for wav2vec2 (positive in seven of eight runs) and $+0.031$ dB for EAT. Their ordinary 95% intervals are $[-0.003, +0.231]$ and $[-0.057, +0.120]$; the individual corrected tests do not pass. Appendix H gives all controls and endpoints from these eight reused splits.

The deployment profile in Table 2 measures the practical cost of this replacement. At batch eight, student map latency is 10.7% lower on both encoders, with all 40 paired timing rounds favouring the student. In the full batch-one waveform pipeline, it adds only 0.080/0.275 ms (0.51/1.83%) relative to identity. Map-only savings and whole-pipeline overhead describe different computations.

6 Recognising sounds with adapted representations

Sound recognition tests another use of the same maps: identifying the sounds present instead of reconstructing their waveforms. Table 3 reports all five maps on controlled OOC mixtures and natural SONYC-UST recordings; each map selects its own classifier without test labels.

Comparison with direct and label-supervised fitting. On controlled OOC recognition, the student exceeds the direct low-rank fit by 1.631, 3.843, 5.020 and 2.724 AP points for wav2vec2, HuBERT, BEATs and EAT, and all four comparisons pass the secondary Holm correction. The full map and student also exceed the BCE map on HuBERT, BEATs and EAT, by 1.03–2.30 points across these six corrected comparisons. This pooled direct comparison differs in rows, target and solver, as in separation. Relative to identity, controlled-task differences range from -0.553 to $+0.302$ points; none passes the primary correction.

Table 3: Sound recognition: mean macro step AP (%), 16 paired runs with calibration-selected classifiers. Full map is CompAdapt; Student is its low-rank compression; Direct is the same-rank prototype fit; BCE is label-supervised. Bold marks positive full-map/student contrasts with identity after primary Holm16 correction. Complete inference: Appendix D.

Encoder	Identity	Full map	Student	Direct	BCE
Controlled ESC-50: OOC mixtures					
wav2vec2	12.432	12.734	12.693	11.062	12.474
HuBERT	21.671	21.366	21.294	17.451	20.265
BEATs	54.518	54.769	53.965	48.945	52.473
EAT	44.896	44.441	44.822	42.098	42.920
Natural SONYC-UST					
wav2vec2	37.485	37.423	37.456	37.114	37.397
HuBERT	46.289	46.271	46.287	45.547	46.206
BEATs	66.998	67.297	**67.151**	67.880	67.278
EAT	63.097	**63.247**	**63.288**	64.148	63.604

Table 4: Alternative post-encoder maps: mean OOC SI-SDRi differences from identity (dB), three-source mixtures and five runs in the earlier clip-disjoint ESC-50 protocol. Each map has a fresh separator and a validation-selected checkpoint within 60 epochs. Method definitions and the full numerical records are described in Appendix E.

Map recipe	wav2vec2	HuBERT	BEATs	EAT	MuQ
CompAdapt	−0.064	−0.011	+0.029	+0.153	−0.014
LPFT-style	−0.022	−0.056	+0.001	+0.036	−0.027
WiSE-FT-style	+0.000	+0.009	−0.053	+0.000	+0.012
L2-SP-style	+0.017	−0.098	−0.037	−0.170	−0.036
Side-tuning	−0.020	−0.224	+0.072	−0.169	−0.030
Supervised low-rank	−0.084	−0.185	+0.040	−0.028	−0.060

Reusing the map on natural audio. The ESC-50 maps are reused unchanged with a classifier trained on SONYC-UST. On this fixed test set, BEATs reaches 66.998% macro AP with identity and 67.151% with the student, and EAT reaches 63.097% with identity, 63.247% with the full map and 63.288% with the student. These three gains of 0.15–0.19 points pass the primary 16-test correction. Appendix D reports separate sensor-composition uncertainty. The direct fit gives the highest mean SONYC scores for BEATs and EAT, while the student exceeds it on HuBERT by 0.740 points after secondary correction, so the preferred fitting procedure differs between controlled and natural recordings.

7 Alternative adaptors and cross-dataset reuse

We next compare CompAdapt with other adaptation recipes and test reuse on another corpus.

7.1 Comparison with other adaptation recipes

An earlier ESC-50 study with the four encoders above and the music encoder MuQ (Zhu et al., 2025) compares CompAdapt with LPFT-style and WiSE-FT-style adaptation, L2-SP regularisation, side-tuning and a supervised low-rank map (Kumar et al., 2022; Wortsman et al., 2022; Li et al., 2018; Zhang et al., 2020). All act on encoder outputs, and each map receives a new separator with a 60-epoch budget and a validation-selected checkpoint. Table 4 shows the three-source results.

CompAdapt has the highest mean among these six recipes on EAT; the ordering differs on the other encoders, and LPFT and WiSE-FT are competitive without a class-prototype objective. CompAdapt exceeds an unregularised joint map-and-separator fit on all five encoders, by 0.10–0.62 dB, but for two-source mixtures a supervised recipe chosen on validation data in each run scores higher than CompAdapt on all five (Appendix E).

Table 5: FSD50K reuse: paired target OOC SI-SDRi differences (dB), eight runs and ordinary 95% intervals. Panels identify refitting and the identity reference. P/S: primary Holm8 / secondary Holm12; retention margin: ± 0.15 dB. Complete results: Appendix C and the numerical artifact.

Contrast	wav2vec2 (rank 16)	EAT (rank 4)	Test
Refit on FSD50K: FSD-fitted maps and heads; base = identity map with an FSD-trained head			
Full map − base	$+0.004$ $[-0.179, +0.186]$	$+0.118$ $[-0.052, +0.288]$	P gain
Student − base	$+0.017$ $[-0.095, +0.128]$	$+0.070$ $[-0.106, +0.246]$	S gain
Student − full map	$+0.013$ $[-0.144, +0.171]$	-0.048 $[-0.187, +0.092]$	P equiv.
Frozen ESC-50 map with a new FSD-trained head; base as above			
Full map − base	$+0.003$ $[-0.124, +0.130]$	$+0.052$ $[-0.123, +0.227]$	S gain
Student map − base	$+0.032$ $[-0.087, +0.152]$	-0.006 $[-0.176, +0.164]$	P gain
Frozen ESC-50 pipeline (map and head unchanged); base = unchanged ESC-50 identity pipeline			
Full map pipeline − base pipeline	$+0.090$ $[-0.047, +0.227]$	$+0.078$ $[-0.099, +0.255]$	S gain
Student pipeline − base pipeline	$+0.043$ $[-0.098, +0.184]$	$+0.051$ $[-0.114, +0.216]$	P gain

A separate five-run comparison fits CompAdapt and rank-8 in-weight LoRA on `wav2vec2-base` and `hubert-base-ls960` using the same prototype targets and 30-epoch separators (Hu et al., 2022). Mean OOC gains over identity are 0.348/0.764 dB for CompAdapt and 0.194/0.940 dB for LoRA; paired differences, intervals and implementation details are in Appendix E.

7.2 Reuse on FSD50K

The FSD50K study compares refitting both map and separator, reusing the ESC-50 map with a new target separator, and reusing the complete ESC-50 pipeline without target fitting or selection; Table 5 states each route's identity reference.

Target-refit mean gains are 0.004/0.118 dB for the full map and 0.017/0.070 dB for the student on wav2vec2/EAT (Table 5), and reusing the complete full-map pipeline gives 0.090 and 0.078 dB. None of these differences passes its corrected test. Two wav2vec2 secondary tests pass: the student retains the reused ESC-50 map's performance, and the target-trained student exceeds the direct low-rank fit by 0.168 dB. Absolute scores are in Appendix C.

8 Discussion and conclusion

CompAdapt reuses known-class prototypes to fit a nonlinear map after a frozen encoder. The map keeps the representation interface and needs no class labels at inference. It improves separation at the common 60-epoch endpoint; the EAT full map and the BEATs/EAT students also give corrected gains on natural sound recognition. Optional compression retains selected full-map separation performance with fewer coefficients and small measured pipeline overhead. These uses depend on the downstream learner: fixed-budget and calibration-selected performance answer different questions, and the preferred fitting procedure can differ between controlled and natural recordings.

AI use disclosure

We used generative AI tools for experiment design and critique, method implementation, data cleaning and reformatting, and simulated review. AI-assisted code constructed synthetic mixtures from recorded audio. We also used these tools to assist with literature search, software artifacts, improving writing, and proofreading. We did not use them for conceptual and theoretical development, mathematical claims and proofs, hypothesis refinement, translation, or qualitative/thematic analysis. All AI-assisted work was reviewed, with derivations and citations checked and code and numerical claims checked against tests and recorded outputs. We take responsibility for the final text, claims, code and data.

Reproducibility statement. Sections 3–4 specify the method and evaluation protocol; Appendices A–H give data processing, fitting and statistical details. The supplementary artifact provides code, configurations and numerical records for reproducing the reported comparisons.

REFERENCES

Tony Alex, Sara Ahmed, Armin Mustafa, Muhammad Awais, and Philip J. B. Jackson. SSLAM: Enhancing self-supervised models with audio mixtures for polyphonic soundscapes. In *International Conference on Learning Representations (ICLR)*, 2025.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Lukas Balles and Philipp Hennig. Dissecting adam: The sign, magnitude and variance of stochastic gradients. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 404–413, 2018. URL https://proceedings.mlr.press/v80/balles18a.html.

Elad Ben-Zaken, Yoav Goldberg, and Shauli Ravfogel. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

Giovanni Bindi and Philippe Esling. Unsupervised composable representations for audio. In *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR)*, 2024.

Mark Cartwright, Ana Elisa Mendez Mendez, Jason Cramer, Vincent Lostanlen, Graham Dove, Ho-Hsiang Wu, Justin Salamon, Oded Nov, and Juan Pablo Bello. SONYC urban sound tagging (SONYC-UST): A multilabel dataset from an urban acoustic sensor network. In *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2019. URL https://markcartwright.com/files/cartwright2019sonycust.pdf.

Chuyang Chen, Bea Steers, Brian McFee, and Juan Pablo Bello. Evaluating compositional structure in audio representations. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026. URL https://arxiv.org/abs/2603.13685.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In *International Conference on Machine Learning (ICML)*, 2023.

Wenxi Chen, Yuzhe Liang, Ziyang Ma, Zhisheng Zheng, and Xie Chen. EAT: Self-supervised pre-training with efficient audio transformer. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

DCASE Community. Urban sound tagging with spatiotemporal context. DCASE 2020 Challenge, Task 5, 2020. URL https://dcase.community/challenge2020/task-urban-sound-tagging-with-spatiotemporal-context.

Marc Delcroix, Jorge Bennasar Vázquez, Tsubasa Ochiai, Keisuke Kinoshita, Yasunori Ohishi, and Shoko Araki. SoundBeam: Target sound extraction conditioned on sound-class labels and enrollment clues for increased performance and continuous learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:121–136, 2023. doi: 10.1109/TASLP.2022.3221000. arXiv:2204.03895.

Jerry A Fodor and Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 1988.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. FSD50K: An open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022.

Zhijin Guo, Chenhao Xue, Zhaozhen Xu, Hongbo Bo, Yuxuan Ye, Janet B. Pierrehumbert, and Martha Lewis. Quantifying compositionality of classic and state-of-the-art embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 22130–22146, 2025. URL https://aclanthology.org/2025.findings-emnlp.1206/.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, et al. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning (ICML)*, 2019.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

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

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations (ICLR)*, 2020.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://iclr.cc/virtual/2022/oral/5946. Oral. arXiv:2202.10054.

Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning (ICML)*, 2018.

Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. SDR – half-baked or well done? In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

Xuhong Li, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2825–2834, 2018. URL https://proceedings.mlr.press/v80/li18a.html. arXiv:1802.01483.

Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. In *Advances in Neural Information Processing Systems*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/00bb4e415ef117f2dee2fc3b778d806d-Abstract-Conference.html.

Xubo Liu, Qiuqiang Kong, Yan Zhao, Haohe Liu, Yi Yuan, Yuzhuo Liu, Rui Xia, Yuxuan Wang, Mark D. Plumbley, and Wenwu Wang. Separate anything you describe. *IEEE Transactions on Audio, Speech and Language Processing*, 33:458–471, 2025. doi: 10.1109/TASLP.2024.3520017. AudioSep. First posted as arXiv:2308.05037 (Aug 2023).

Yi Luo and Nima Mesgarani. Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2019.

Karol J Piczak. ESC: Dataset for environmental sound classification. In *ACM International Conference on Multimedia (ACM MM)*, 2015.

Pradyumna Reddy, Scott Wisdom, Klaus Greff, John R. Hershey, and Thomas Kipf. AudioSlots: A slot-centric generative model for audio separation. In *ICASSP Workshop on Self-Supervision in Audio, Speech and Beyond (SASB)*, 2023.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6550.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems 30 (NIPS)*, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html. arXiv:1703.05175.

Sripathi Sridhar and Mark Cartwright. Compositional audio representation learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

Scott Wisdom, Efthymios Tzinis, Hakan Erdogan, Ron J Weiss, Kevin Wilson, and John R Hershey. Unsupervised sound separation using mixture invariant training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7949–7961, 2022. doi: 10.1109/CVPR52688.2022.00780. arXiv:2109.01903.

Dong Yu, Morten Kolbaek, Zheng-Hua Tan, and Jesper Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

Jeffrey O. Zhang, Alexander Sax, Amir Zamir, Leonidas Guibas, and Jitendra Malik. Side-tuning: A baseline for network adaptation via additive side networks. In *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pp. 698–714. Springer, 2020. doi: 10.1007/978-3-030-58580-8_41. Spotlight. arXiv:1912.13503.

Tianyue Zhang, Lucas Maes, Alan Milligan, Alexia Jolicoeur-Martineau, Ioannis Mitliagkas, Damien Scieur, Simon Lacoste-Julien, and Charles Guille-Escuret. Understanding Adam requires better rotation dependent assumptions. In *Advances in Neural Information Processing Systems*, 2025. URL https://neurips.cc/virtual/2025/poster/118665.

Haina Zhu et al. MuQ: Self-supervised music representation learning with mel residual vector quantization. *arXiv preprint*, 2025.

A NOTATION, DATA ROLES AND IMPLEMENTATION

This appendix gives the details needed to interpret and reproduce the main comparisons. M4 denotes the grouped ESC-50 separation study (Section 5), M5 the FSD50K reuse study (Section 7) and M6 the sound-recognition study (Section 6). Earlier studies keep their own codes (P2, R5, R6, M2, M3) and protocols. The numerical artifact retains all run scores, selection records, omitted tables and timing samples; `supplementary/README_CURRENT.md` gives reproduction commands.

Notation and abbreviations.

Symbol or term	Meaning
F; F_t; T; d	Frozen encoder; feature vector at sequence position t; number of positions; feature width ($d = 768$)
$z(x)$	Pooled feature of recording x: the mean of F_t over positions
$\mathcal{P}_c$; a_c	Reserved single-class recordings of class c; their mean pooled feature (class prototype)
S; k; q_S	Classes in a mixture; number of sources $k = \lvert S\rvert$; prototype-sum target $\sum_{c\in S} a_c$
ϕ; h; β	Full CompAdapt map (Equation 1); hidden width 128; anchor weight 0.3
f_r; r	Low-rank student (Equation 3); rank of its linear change (16, or 4 for EAT)
Identity (base)	Original encoder features passed unchanged to the downstream model
Direct fit	Same-rank affine map fitted directly to the prototype-and-anchor target on pooled rows
BCE map	Map with CompAdapt's architecture trained with binary cross-entropy on source class labels
IC; OOC	Test mixtures whose class combination was seen; was withheld during downstream fitting
SI-SDR; SI-SDRi	Scale-invariant signal-to-distortion ratio; its improvement over the mixture (dB)
AP; macro AP	Average precision; its mean over classes with equal weight
$\Delta_D(f)$; $J(f)$	Map-minus-identity score on evaluation set D; $\Delta_{\text{OOC}}(f) - \Delta_{\text{IC}}(f)$
PIT; STFT	Permutation-invariant training; short-time Fourier transform
LPFT; WiSE-FT; L2-SP	Linear probing then fine-tuning; weight-space ensembling; L2 penalty toward the starting point
TOST; Holm	Two one-sided tests of equivalence; Holm step-down correction for multiple tests

A.1 RECORDING GROUPS AND CLASS COMBINATIONS

The grouped studies pool the ESC-50 folds and split them anew by source recording (`src_file`) instead of using the standard five-fold protocol. Each class supplies six prototype, nine fitting, four calibration and five test clips, with at least two recording groups per class and role. The 48 fitting, 20 calibration and 40 OOC class-tuple sets are pairwise disjoint, and each family covers all 50 classes. Twelve mixtures per tuple give 576 fitting, 240 calibration and 480 OOC mixtures. IC mixtures reuse the fitting tuples with test recordings, so IC and OOC share test recordings but not fitting recordings. Class and class-pair frequencies can differ between tuple families.

The FSD50K subset has 1,499 recordings from 36 classes without parent–child label overlap, grouped by hashed uploader ID. Known ESC-50 overlaps and exact duplicates are removed before splitting, and the per-class allocation and mixture counts follow the ESC-50 study. Its source systems use seeds 101–108, independent of M4's seeds 34–49. An FSD50K recording can contain background events, so the separation target is the contributing recording, not an isolated sound event.

A.2 ENCODERS AND SEPARATOR

The wav2vec2 features are the last layer of `WAV2VEC2_ASR_BASE_960H`, which is fine-tuned for automatic speech recognition; HuBERT, BEATs and EAT also give 768-dimensional features. The earlier five-encoder comparison adds the music model MuQ, of width 1024 (Zhu et al., 2025), and the in-weight LoRA experiment uses the Hugging Face `wav2vec2-base` and `hubert-base-ls960` models. For a ten-second input, BEATs and EAT give flattened 62×8

Table 6: Class coverage and data roles in the current studies. Counts are the classes used in each experiment, not the full label vocabulary of the original corpus.

Study	Classes	Grouping or split	Downstream task
Grouped ESC-50 (M4)	All 50	Source-recording groups	Separation
FSD50K subset (M5)	36 feasible	Uploader groups	Separation
Controlled ESC-50 (M6)	All 50	Source-recording groups	Recognition
Natural SONYC-UST (M6)	8 coarse	Dataset train/validation/test	Recognition

and 64×8 time–frequency patch grids, with the frequency index changing fastest; the separator interpolates along this flattened axis without pooling patches within a time window. The artifact identifies the fixed exported encoders; a complete upstream checkpoint and exporter reconstruction is not available for the staged EAT and BEATs models.

ESC-50 clips are peak-normalised and stored at 48 kHz as int16 with peak 0.7, then resampled to 16 kHz and tiled from five to ten seconds; 862 of the 1,499 FSD50K recordings are shorter than five seconds and are zero-padded first. Each mixture has one shared gain in $[0.7, 1.0]$, so source energies are not equalised, and prototypes use unscaled class means. SONYC recordings are used at their natural ten-second length.

The separator has three width-256, kernel-3 one-dimensional convolutions with ReLU, the middle one with dilation two, and a 1×1 convolution that gives k sigmoid masks over 257 frequency bins; there are no normalisation layers. The mapped sequence is linearly interpolated to the STFT frame count (`align_corners=False`). Masks multiply the mixture's complex STFT (512-point Hann window, hop 160), and the inverse STFT gives the sources for permutation-invariant SI-SDR training.

A.3 Fitting the map and the downstream models

The CompAdapt map is fitted on the 576 pooled fitting-mixture vectors of each run as one full batch, for 300 Adam epochs at learning rate 10^{-3}; $h = 128$ and $\beta = 0.3$ are fixed across M4–M6. For separation, each frozen map receives an independently trained separator (learning rate 10^{-3}, batch size 48). M4 retains checkpoints at epochs 1, 3, 5, 10, 15, 20, 30, 45, 60, 90, 120, 150 and 180; M5 retains the first ten, through epoch 90. Within a comparison, mixtures, waveform targets, parameter seeds and training order are matched. Selection uses the unrounded calibration mean, with ties resolved toward the earlier epoch, and all choices are sealed before testing. Selected and fixed-endpoint scores come from the same training trajectories and are not counted as independent observations.

A.4 Low-rank student and direct fit

The student is fitted to 40,000 sequence positions F_t sampled from the fitting mixtures (time frames for wav2vec2 and HuBERT, flattened patches for BEATs and EAT), not to pooled vectors. Let the rows of $X \in \mathbb{R}^{n \times d}$ be these inputs and the rows of Y the full map's displacements $\phi(F_t) - F_t$. After centring both matrices, ridge regression gives

$$D = (X_c^\top X_c + \lambda I_d)^{-1} X_c^\top Y_c, \qquad \lambda = 10^{-3} \operatorname{tr}(X_c^\top X_c)/d.$$

We truncate D to rank r by Euclidean singular-value decomposition, refit the intercept and set $M_r = D_r^\top$ in Equation 3. Fitting is in float64 and stored factors are float32. Ranks come from an earlier ESC-50 validation study that tested ranks 0, 4, 16 and full rank, choosing the smallest $r \leq 16$ with mean validation $R^2 \geq 0.90$ between student and full-map outputs, including the identity component; this is not the fraction of the displacement explained. The ranks are fixed at 16/16/16/4 throughout M4–M6, and later test scores were not used to choose them. M4's post-fit BEATs mean is 0.8968, reported without changing the rank.

The direct fit uses the same 576 pooled mixture rows X and prototype targets Q. With $R = (Q - X)/(1 + \beta)$ and its mean shift retained, it solves

$$\min_{\operatorname{rank}(D) \leq r} \|X_c D - R_c\|_F^2 + \lambda \|D\|_F^2.$$

If $LL^\top = X_c^\top X_c + \lambda I_d$, we truncate $L^{-1}X_c^\top R_c$ and map the result back with $L^{-\top}$. The student and direct fit therefore differ in target, training rows and truncation, while sharing rank and output width.

The student stores $2dr + r + 2d$ values (two factors, singular values, input centre and output shift): 26,128 for $r = 16$ and 7,684 for $r = 4$ when $d = 768$. The full map has 197,504 parameters including biases. These counts describe stored coefficients. Appendix H separately measures latency and live memory.

B GROUPED SEPARATION RESULTS (M4)

M4 has 64 encoder/run cells and 256 separator trajectories. Its primary family tests student–full-map equivalence within ± 0.15 dB on the four encoders; its secondary family has twelve comparisons with identity and with the direct fit. The margin was fixed before the study as an engineering tolerance, not a perceptual threshold. Table 7 gives the complete calibration-selected inference.

Table 7: Complete prespecified inference in the grouped ESC-50 study (M4): calibration-selected checkpoints, OOC SI-SDRi (dB), 16 paired runs per encoder. For the equivalence tests the raw p is the larger of the two one-sided p values. Intervals are ordinary and unadjusted; Pass is the Holm-corrected decision within each family. The ± 0.15 dB margin is an engineering retention tolerance fixed before the study, not a perceptual threshold.

Encoder	Comparison	Mean	Interval	Raw p	Holm p	Pass
Primary: student − full map equivalence within ±0.15 dB (TOST; 90% interval; Holm over four)						
wav2vec2	Student − full map	+0.0284	[−0.0165, +0.0733]	1.3×10^{-4}	3.9×10^{-4}	yes
HuBERT	Student − full map	+0.0379	[−0.0061, +0.0819]	2.3×10^{-4}	4.5×10^{-4}	yes
BEATs	Student − full map	+0.0466	[−0.0101, +0.1034]	0.0030	0.0030	yes
EAT	Student − full map	+0.0095	[−0.0315, +0.0505]	1.2×10^{-5}	4.8×10^{-5}	yes
Secondary: positive paired difference (two-sided t test; 95% interval; Holm over twelve)						
wav2vec2	Full map − base	+0.0422	[−0.0887, +0.1731]	0.502	1.000	no
wav2vec2	Student − base	+0.0706	[−0.0448, +0.1861]	0.212	1.000	no
wav2vec2	Student − direct	+0.1393	[+0.0587, +0.2198]	0.0022	0.026	yes
HuBERT	Full map − base	−0.0206	[−0.0970, +0.0558]	0.574	1.000	no
HuBERT	Student − base	+0.0173	[−0.0560, +0.0906]	0.622	1.000	no
HuBERT	Student − direct	+0.1318	[+0.0543, +0.2092]	0.0025	0.027	yes
BEATs	Full map − base	−0.0281	[−0.1173, +0.0610]	0.511	1.000	no
BEATs	Student − base	+0.0185	[−0.1046, +0.1416]	0.753	1.000	no
BEATs	Student − direct	+0.0500	[−0.0439, +0.1440]	0.274	1.000	no
EAT	Full map − base	+0.0252	[−0.0511, +0.1015]	0.492	1.000	no
EAT	Student − base	+0.0347	[−0.0528, +0.1222]	0.411	1.000	no
EAT	Student − direct	+0.0097	[−0.0517, +0.0712]	0.740	1.000	no

All 256 selected checkpoints occur at or before epoch 60. From the selected checkpoint to epoch 180, every trajectory improves on its fitting mixtures while its OOC score declines. Only the selected, epoch-60 and epoch-180 test points were computed, so Figure 4 is a calibration curve, not an all-epoch OOC curve.

Fixed endpoints. These results are descriptive; no multiplicity family is defined for them. The full map's OOC gains over identity are 0.886, 0.726, 0.159 and 0.888 dB at epoch 60 and 1.587, 0.177, 0.032 and 0.546 dB at epoch 180 for wav2vec2, HuBERT, BEATs and EAT; among these ordinary 95% intervals, only the BEATs epoch-180 interval includes zero. The student gains 0.917, 0.618, 0.192 and 0.913 dB at epoch 60. Student–full-map 90% intervals lie within ± 0.15 dB at both endpoints except for HuBERT at epoch 60 (-0.108 dB, $[-0.162, -0.055]$) and wav2vec2 at epoch 180 (-0.123 dB, $[-0.199, -0.046]$), so fixed-endpoint equivalence is not shown on all encoders.

Familiar and new combinations. At epoch 60, the full map's IC gains are 0.830, 0.630, 0.254 and 0.830 dB, close to its OOC gains, so the fixed-endpoint benefit is not specific to withheld combinations. At calibration-selected checkpoints, full-map-minus-identity differences lie between -0.03 and $+0.05$ dB for both IC and OOC mixtures, and every 95% interval includes zero. Because J is

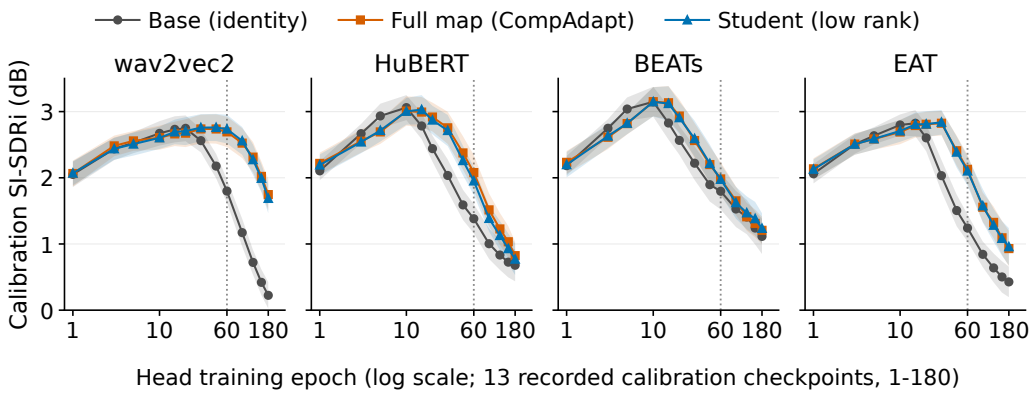

Figure 4: Calibration SI-SDRi at the 13 retained checkpoints in M4 for the identity map (base), the full map and the student. Points are means over 16 runs, shading is an ordinary 95% t interval and the dotted line marks epoch 60. Each run selects its own checkpoint; selection is not made from these mean curves.

a difference of two gains, a positive J alone need not mean a positive OOC benefit, so we report its components.

C EXTERNAL SEPARATION AND REUSE (M5)

M5 has 16 encoder/run cells, 144 separator trajectories and nine evaluated target pipelines per run. Table 8 gives absolute scores for every pipeline. Target-refit and new-head comparisons use an FSD-trained identity separator as reference; complete frozen-pipeline comparisons use the unchanged ESC-50 identity pipeline on FSD50K audio. Table 5 shows 95% intervals, while retention decisions use 90% intervals.

Table 8: Absolute OOC SI-SDRi (dB) for every FSD50K pipeline: means ± ordinary 95% half-widths over eight runs. Map and head specify the fitting corpus (FSD: FSD50K subset; ESC: ESC-50; a dash denotes identity). Full map is CompAdapt. Reused pipelines keep their ESC-50 heads and checkpoint choices. Epoch is the median selected checkpoint for wav2vec2/EAT. Scores measure separation relative to the input mixture; Table 5 gives paired pipeline differences.

Pipeline	Map	Head	wav2vec2	EAT	Epoch
Evaluated on FSD50K held-out triples					
Base	–	FSD	3.4309 ± 0.267	3.7405 ± 0.301	15 / 15
Full map	FSD	FSD	3.4344 ± 0.360	3.8585 ± 0.403	45 / 20
Student	FSD	FSD	3.4475 ± 0.321	3.8109 ± 0.375	45 / 20
Direct fit	FSD	FSD	3.2797 ± 0.353	3.6254 ± 0.295	30 / 25
Full map, new head	ESC	FSD	3.4338 ± 0.317	3.7925 ± 0.324	37.5 / 20
Student map, new head	ESC	FSD	3.4632 ± 0.341	3.7346 ± 0.297	30 / 20
Base pipeline	–	ESC	3.3208 ± 0.298	3.2264 ± 0.280	15 / 10
Full map pipeline	ESC	ESC	3.4111 ± 0.397	3.3039 ± 0.357	45 / 20
Student pipeline	ESC	ESC	3.3637 ± 0.389	3.2775 ± 0.315	52.5 / 17.5
Source pipelines evaluated on ESC-50 held-out triples					
Base	–	ESC	2.8980 ± 0.200	3.0080 ± 0.278	15 / 10
Full map	ESC	ESC	2.9047 ± 0.231	3.0119 ± 0.267	45 / 20
Student	ESC	ESC	2.8809 ± 0.240	3.0092 ± 0.215	52.5 / 17.5

None of the eight primary tests passes after Holm correction. Two of the twelve secondary tests pass, both on wav2vec2: student–full-map retention after reusing the ESC-50 map with a new FSD50K separator (difference $+0.0294$ dB, adjusted $p = 0.0391$), and the advantage of the target-fitted student over the direct fit ($+0.1679$ dB, adjusted $p = 0.0450$). The complete test table is retained in the numerical artifact. The eight runs vary fitting within one fixed corpus and construction. The target-fitted direct map scores below the FSD-trained identity reference on both encoders (Table 8); the student's advantage over that map is a comparison of fitting procedures, not a corrected gain over identity.

D SOUND-RECOGNITION PROTOCOL AND INFERENCE (M6)

Data and label access. The controlled task predicts the 50 ESC-50 labels for group-separated OOC mixtures. For every map, including identity, source classifiers are fitted on both the mixtures and the labelled prototype recordings, so the prototype-based and BCE routes have the same access to source labels; prototype recordings remain separate from the mixtures used to fit the CompAdapt map. The natural task uses the eight coarse SONYC-UST-V2 labels. The dataset separates training and validation sensors and shifts its test set in time, although test sensors can overlap with training or validation sensors. The ESC-50 map is reused unchanged and only the classifier is fitted, so this is target-supervised recognition with a frozen source map, not a held-out-combination split or a zero-shot pipeline.

BCE map. The BCE map is fitted on 876 pooled examples (576 mixtures and 300 reserved single-class recordings) with the same 768–128–768 residual architecture as the CompAdapt map and an auxiliary 50-output linear classifier on features standardised with source-fitting statistics. Its loss is multi-label binary cross-entropy plus 0.3 times the mean squared displacement. Map and auxiliary classifier train jointly for 300 Adam epochs at learning rate 10^{-3}, and the auxiliary classifier is discarded before the common classifier comparison.

Table 9: Recognition primary family. Differences and ordinary paired intervals are AP percentage points. T: full CompAdapt map; S: student; I: identity; D: direct; B: BCE. Intervals are ordinary paired 95% intervals. Holm correction covers all 16 comparisons. Pass requires a positive mean and an adjusted $p < 0.05$.

Encoder	Contrast	Δ	Interval	p_{Holm}	Pass
wav2vec2/ESC	S−I	$+0.261$	$[-0.060, +0.582]$	1.000	No
wav2vec2/ESC	T−I	$+0.302$	$[+0.099, +0.505]$	0.083	No
wav2vec2/SONYC	S−I	-0.029	$[-0.106, +0.048]$	1.000	No
wav2vec2/SONYC	T−I	-0.063	$[-0.260, +0.134]$	1.000	No
HuBERT/ESC	S−I	-0.377	$[-0.870, +0.115]$	1.000	No
HuBERT/ESC	T−I	-0.305	$[-0.801, +0.190]$	1.000	No
HuBERT/SONYC	S−I	-0.002	$[-0.100, +0.096]$	1.000	No
HuBERT/SONYC	T−I	-0.019	$[-0.151, +0.114]$	1.000	No
BEATs/ESC	S−I	-0.553	$[-1.335, +0.228]$	1.000	No
BEATs/ESC	T−I	$+0.251$	$[-0.295, +0.797]$	1.000	No
BEATs/SONYC	S−I	$+0.152$	$[+0.067, +0.237]$	0.025	Yes
BEATs/SONYC	T−I	$+0.299$	$[+0.044, +0.553]$	0.292	No
EAT/ESC	S−I	-0.073	$[-0.716, +0.569]$	1.000	No
EAT/ESC	T−I	-0.454	$[-1.146, +0.238]$	1.000	No
EAT/SONYC	S−I	$+0.190$	$[+0.108, +0.272]$	0.003	Yes
EAT/SONYC	T−I	$+0.150$	$[+0.059, +0.241]$	0.043	Yes

Classifiers and metric. Each map's classifier is linear or a one-hidden-layer tanh MLP with 128 units, with an L2 coefficient in $\{10^{-2}, 10^{-4}, 10^{-6}\}$. Centring and scaling are fitted separately for each map on fitting data; coordinates with standard deviation at most 10^{-8} use scale one. The highest unrounded calibration macro AP selects among the six candidates, with ties preferring linear and then the larger L2 coefficient, and all choices are sealed before test features are accessed. Fitting uses L-BFGS-B ($\texttt{ftol}= 10^{-8}$, $\texttt{gtol}= 10^{-6}$, $\texttt{maxiter}= 5000$, $\texttt{maxls}= 40$, history ten). Of 3,840 candidates, 609 MLP fits stop at the 5,000-iteration budget and are marked budget-limited rather than converged; all 640 selected classifiers converged, and all 320 natural-audio selections are linear. The metric is macro step AP: for each class, precision weighted by the increase in recall at each score threshold, averaged over classes. It differs from the official SONYC coarse-label trapezoidal area under the precision–recall curve; both are in the artifact. Reported means average 16 separately fitted pipelines, not ensemble predictions.

Test families. The primary family contains the full-map and student comparisons with identity on both tasks and all four encoders (16 tests; Table 9, where T denotes the full map). The secondary family contains student–full-map retention with a ± 1 AP-point margin and the full-map/BCE, student/BCE and student/direct comparisons (32 tests). Gain tests are two-sided paired t tests that require a positive mean; retention uses the larger of the two one-sided p values. Holm correction is applied to each complete family.

Secondary results. The complete secondary tables are retained numerically. On controlled mixtures, the student exceeds the direct fit for all four encoders (1.631–5.020 AP points), and the full map and student exceed the BCE map on HuBERT, BEATs and EAT (1.03–2.30 points). On SONYC, the student exceeds the direct fit on HuBERT by 0.740 points, whereas for BEATs and EAT it is below the direct fit by 0.729 and 0.860 points. These two differences are in the opposite direction to the prespecified student-superiority test, which is therefore not passed. Retention within ± 1 point passes for wav2vec2 and HuBERT on ESC-50 and for all four encoders on SONYC.

Sensor composition. Run-level intervals describe variation across fitted systems on one fixed test set. To see how the natural-audio results depend on which sensors are sampled, we reweight the 48 test sensors with 1,000 paired positive-exponential weight draws (seed 92623), recompute weighted step AP for each run and average the paired differences. The resulting descriptive 95% intervals include zero for all three primary natural-audio gains: $[-0.096, +0.633]$ for the BEATs student, and $[-0.089, +0.409]$ and $[-0.027, +0.423]$ for the EAT full map and student. They exclude zero

for the HuBERT student–direct difference ($[+0.360, +1.201]$) and, in the negative direction, for the BEATs and EAT student–direct differences ($[-1.270, -0.166]$ and $[-1.800, -0.096]$).

Affine maps and linear classifiers. An affine student commutes with mean pooling: averaging transformed features equals transforming their average. A following linear classifier can therefore absorb the student into its weights, so the student changes the classifier's parameterisation and fitting, not the set of affine functions it can represent. Fit-based centring also cancels a constant translation. The nonlinear full map does not generally commute with pooling.

E Alternative adaptation recipes and in-weight LoRA

Representation-map recipes (P2). P2 uses five encoders (wav2vec2, HuBERT, BEATs, EAT and MuQ), seeds 0–4 and two- and three-source mixtures under the earlier clip-disjoint ESC-50 protocol, separate from M4's recording-grouped design. Each cell has 576 fitting mixtures, 144 validation mixtures from tuples held out of fitting, and 480 OOC test mixtures from different tuples. Every map receives a new separator with a 60-epoch budget, and checkpoints and hyperparameters are chosen on the validation mixtures. LPFT-style fitting trains a head on unchanged features and then fits map and head jointly (Kumar et al., 2022). WiSE-FT-style fitting interpolates the residual map with identity using coefficients 0, 0.25, 0.5, 0.75 and 1; only the map is interpolated (Wortsman et al., 2022). L2-SP penalises departure from the initial map with weights 10^{-4}, 10^{-3} and 10^{-2} (Li et al., 2018). Side-tuning learns a blend of the unchanged features and a side network (Zhang et al., 2020). The supervised low-rank map is a rank-16 update in representation space trained with the separation loss, not in-weight LoRA. The complete comparison also includes an unregularised joint map-and-separator fit, a class-balanced version of it, mixup, joint separation and prototype fitting, and a MixIT-trained map (Wisdom et al., 2020). All maps are frozen before their reporting separators are trained; their total adaptation cost differs, for example because LPFT includes a probe stage. In Table 4, "CompAdapt" is the full map.

The complete two- and three-source tables are retained numerically. For three-source mixtures, CompAdapt exceeds the unregularised joint fit on all five encoders, by 0.10–0.62 dB. For two-source mixtures, its mean OOC difference from identity is negative on all five encoders (-0.260 to -0.021 dB). Compared with the supervised recipe chosen on validation data in each run, CompAdapt is lower on all five encoders for two-source mixtures (by 0.012–0.249 dB), and for three-source mixtures it is higher on four encoders (by 0.013–0.139 dB) and lower by 0.076 dB on wav2vec2.

In-weight LoRA. This separate study uses five seeds and two-source ESC-50 mixtures from pre-generated banks, scored on the stored novel-combination subset, with the encoders run as ordinary Hugging Face models. Rank-8 LoRA updates are inserted into the attention query, key, value and output projections, with scale 16 and no dropout (Hu et al., 2022), and fitted to the squared prototype-target loss for 20 epochs (Adam, learning rate 3×10^{-4}, batch size 8). The post-encoder CompAdapt map fits its prototype-and-anchor objective for 200 epochs at learning rate 10^{-3}. Class prototypes are means of clean recordings from the original folds 1–3. Both adapted encoders are frozen before new separators are trained for 30 epochs. Absolute OOC SI-SDRi is 1.847, 2.196 and 2.042 dB for identity, CompAdapt and LoRA on wav2vec2, and 0.929, 1.692 and 1.869 dB on HuBERT. The paired CompAdapt-minus-LoRA differences are $+0.154$ dB $[+0.012, +0.297]$ and -0.176 dB $[-0.362, +0.010]$ (ordinary 95% intervals), with Wilcoxon $p = 0.125$ and 0.0625.

F Target and coordinate controls

Earlier clip-disjoint studies test what else can produce fixed-endpoint gains: target controls (R5), representation and rank controls (R6), training-coordinate controls (M2) and independent calibration (M3). Numbers are OOC SI-SDRi differences in dB from each study's identity map with a matched separator. These protocols can include OOC classes absent from separator fitting (Appendix G), and their complete outcomes are retained numerically.

Alternative targets (R5). On wav2vec2 and EAT with three-source mixtures, covariance-matched random targets and the constant target $k\mu$, where μ is the mean class prototype, give gains equivalent

to the full prototype target within the prespecified margin. In this setting, the class assignment of the targets is therefore not needed for the fixed-endpoint gain.

Translation and mean shift (R6, M2). At 60 epochs, a literal translation improves OOC separation by 0.88 dB on wav2vec2 and 0.73 dB on HuBERT (R6), and a training-mean shift by 0.900 and 0.946 dB on wav2vec2 and EAT (M2); M2's primary 180-epoch replacement family remains unresolved. For a translation-only map, the prototype objective has the closed-form optimum

$$b^* = \frac{\mathbb{E}_{\text{fit}}[q_S - z]}{1 + \beta}.$$

The full map's mean displacement obeys the same relation at a stationary output bias: differentiating the objective with respect to b_2 gives $(1 + \beta)\mathbb{E}_{\text{fit}}[\phi(z) - z] = \mathbb{E}_{\text{fit}}[q_S - z]$. Across 40 fitted maps (five encoders, eight runs), the observed mean displacement differs from this shift by 0.3–0.8%. This concerns the mean change only, not the complete nonlinear map; at rank zero, the student is a translation, not necessarily the identity.

Coordinate changes (R6, M2). An orthogonal feature map preserves distances and information but can change optimisation. Three dense orthogonal maps improve EAT by 1.16, 1.24 and 1.26 dB at 60 epochs (R6). For an orthogonal matrix Q and input Qv, replacing the first-layer weight A by $AQ^\top$ keeps the initial predictor unchanged, since $(AQ^\top)(Qv) = Av$; this function-matched rotation improves EAT by 1.392 dB at 60 epochs and 1.068 dB at 180 epochs (M2), although later Adam updates need not agree.

Independent calibration (M3). M3 uses eight EAT runs (seeds 26–33) with clip-disjoint calibration and test pools. A rotation chooses its checkpoint from 13 candidates, while a baseline portfolio chooses among identity, centring, standardisation and the prototype map on the same calibration set. The prespecified selected rotation-minus-portfolio OOC difference is $+0.0797$ dB (95% interval $[-0.0370, +0.1965]$, $p = 0.150$). Calibration-based selection raises the identity map's OOC score from 1.095 dB at epoch 60 to 2.777 dB. The portfolio selects the prototype map in three runs, standardisation in three, centring in one and identity in one.

Test-label controls (R6). Oracle code placement supplies the true test classes and is not a deployable variant. It gives OOC gains of 0.13–0.80 dB, with larger IC gains on every encoder.

G CLASS COVERAGE IN THE EARLIER STUDIES

The prototype banks in R6, M2 and M3 contain all 50 classes, but their random tuple constructors expose only 39–42 classes during separator fitting. Table 10 lists the OOC classes absent from fitting and the number of OOC tuples composed only of fitting-known classes. These results are not relabelled as the later all-class protocol.

Restricting M3's OOC evaluation to fitting-known classes, without changing calibration choices, gives a selected rotation-minus-portfolio difference of $+0.0675$ dB (95% interval $[-0.0402, +0.1752]$) and a fixed-60 rotation-minus-identity difference of $+1.2174$ dB ($[+1.0040, +1.4309]$). The earlier pools are clip-disjoint, not recording-group-disjoint: recovered metadata for M3 show that 17.6–23.6% of unique OOC clips and 42.1–62.5% of OOC mixtures involve a recording ID also present in fitting (a mixture counts if any of its sources matches). M4 groups by source recording instead. Class counts, recording overlaps and staged model identities are in `supplementary/class_coverage.csv`, `m3_recording_sets.csv`, `m3_recording_pairs.csv`, `m3_recording_test_overlap.csv` and `export_identifiers.json`.

Table 10: Readout-fitting class coverage in the earlier R6/M2/M3 campaigns, not the all-50-class M4 protocol. "Known OOC" counts novel tuples whose three classes all occur in fitting mixtures, out of 40. Each tuple has twelve mixtures. Class IDs are the fixed staged dataset labels.

Campaign	Seed	Fit classes	OOC classes	OOC classes absent fit	Known OOC
R6	0	42	40	–	40/40
R6	1	42	42	–	40/40
R6	2	42	40	–	40/40
R6	3	41	41	44	37/40
R6	4	40	41	19,45	37/40
R6	5	39	41	32,33,42	32/40
R6	6	40	42	33,39	37/40
R6	7	42	39	–	40/40
M2	18	42	41	–	40/40
M2	19	41	39	2	36/40
M2	20	40	36	1,3	28/40
M2	21	39	38	37,40,43	36/40
M2	22	40	39	25,37	32/40
M2	23	41	37	49	37/40
M2	24	41	40	20	37/40
M2	25	41	41	35	37/40
M3	26	41	38	44	36/40
M3	27	40	41	23,47	33/40
M3	28	41	40	9	38/40
M3	29	39	41	11,20,46	35/40
M3	30	42	41	–	40/40
M3	31	42	40	–	40/40
M3	32	41	41	34	33/40
M3	33	41	39	48	39/40

H MATCHED COMPONENT FITS AND DEPLOYMENT

This follow-up uses wav2vec2 and EAT on eight existing grouped ESC-50 splits (seeds 34–41). The identity, full-map, student and pooled-direct systems retain their M4 fits and scores. Four additional maps receive new separators with the same initialization, first 60 epoch orders, batch size and calibration grid. This adds 64 separator fits; it does not add independent repeats of the reused results.

Matched controls. The analytic shift is $(\overline{q} - \overline{z})/(1 + \beta)$. The mean-target control replaces every fitting target by $\overline{q}$. The shuffled-target control assigns each fitting tuple the target of a class-disjoint tuple while preserving the exact target multiset and mean. Both controls use the full map's fitting recipe followed by the same student distillation. The matched-frame direct fit uses the student's 40,000 sampled positions, rank, ridge and Euclidean-SVD truncation, but its displacement target is $(q_S - v)/(1 + \beta)$ rather than $\phi(v) - v$.

Table 11: Matched component follow-up: mean OOC SI-SDRi (dB), eight existing grouped ESC-50 splits. Fixed is epoch 60; selected uses calibration within the same 60-epoch budget. The first four methods reuse M4 results. Mean/shuffled targets keep the teacher and student fitting recipe; matched-frame direct uses the student's sampled positions, rank, ridge and truncation.

Method	wav2vec2		EAT	
	Fixed	Selected	Fixed	Selected
Unmodified	1.807	2.745	1.214	2.880
CompAdapt	2.703	2.782	2.142	2.892
Student	2.736	2.838	2.181	2.941
Pooled direct	2.318	2.665	2.290	2.871
Analytic shift	2.625	2.753	2.313	2.862
Mean-target student	2.809	2.822	2.272	3.003
Shuffled-target student	2.835	2.838	2.147	2.964
Matched-frame direct	2.668	2.724	2.368	2.909

All methods select from epochs 1, 3, 5, 10, 15, 20, 30, 45 and 60 using calibration SI-SDRi. Table 11 retains both selected and fixed-60 scores. At selection, the student exceeds matched-frame direct fitting by 0.114 dB on wav2vec2 and 0.031 dB on EAT. At fixed60, the EAT matched-frame fit

exceeds the student by 0.187 dB. Mean and shuffled targets match or exceed the correctly assigned targets in several comparisons, so the fitted representation change need not derive its benefit from class assignment alone.

The two prespecified families each contain ten student comparisons: identity, analytic shift, mean target, shuffled target and matched-frame direct, across two encoders. Tests use eight paired run means, ordinary 95% intervals and Holm correction. The fixed60 student-minus-identity gains, 0.930/0.967 dB, pass correction; the other eight fixed60 comparisons and all ten selected comparisons do not. All differences and intervals are retained numerically, including the positive selected differences reported in the main text.

Deployment profile. One RTX 3090 processes the first eight calibration mixtures from seed 34 as ten-second, 16-kHz float32 tensors already on the GPU. The map, cached-representation-to-waveform pipeline, and complete waveform-to-waveform pipeline are measured separately at batches one and eight, with ten warmup calls and forty paired timing rounds in randomized method order. Timings synchronize the GPU and exclude model loading and finite-value checks. They measure this workload, not variation over independent corpora or hardware.

Table 12: Measured deployment cost on one RTX 3090: medians of 40 paired timing rounds on ten-second, 16-kHz inputs. Map-only times use batch 8; the complete waveform-to-waveform pipeline uses batch 1 and includes the encoder, map, separator and reconstruction. Parameters count the representation map only.

Encoder	Map	Coefficients	Map (ms)	Pipeline (ms)
wav2vec2	Unmodified	0	0.025	15.738
wav2vec2	CompAdapt	197,504	0.348	15.770
wav2vec2	Student	26,128	0.311	15.817
EAT	Unmodified	0	0.025	14.984
EAT	CompAdapt	197,504	0.344	15.216
EAT	Student	7,684	0.307	15.259

At batch eight, student map latency is lower than full-map latency in all forty paired rounds for both encoders. The full waveform pipeline includes the encoder, map, separator, STFT and inverse STFT; its latency is similar for the full and compressed maps, rather than showing the same proportional speedup. Relative to identity, the student's batch-one additions are 0.0798 ms (0.51%) for wav2vec2 and 0.2745 ms (1.83%) for EAT.

The student reduces peak live allocation by 0.653/0.724 MiB relative to the full map and adds approximately 0.100/0.030 MiB over the identity waveform pipeline. Counters include the fixed eight-recording input bank, even at forward batch one. Stored map coefficients occupy 0.100/0.029 MiB, versus 0.753 MiB for the full map. Fitting costs and all timing samples remain in the numerical artifact; reused historical fitting is not assigned zero cost.