Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

StemGround: Source-Interventional Auditing and Selective Response Alignment for Instrument Grounding in Music Foundation Model Readouts

Chao Guo
Sichuan Normal University
Junxiang Wang
dave_wangjx@aliyun.com
University of Electronic Science and Technology of China

Article

Keywords: music foundation models, instrument recognition, multitrack audio, source intervention, context dependence, selective response alignment, music information retrieval

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11083018/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

StemGround: Source-Interventional Auditing and Selective Response Alignment for Instrument Grounding in Music Foundation Model Readouts

Chao Guo[^1], Junxiang Wang[^2,*]

1.Music College of Sichuan Normal University , Chengdu, 610101, China.

2. Chengdu Research Institute, University of Electronic Science and Technology of China, Chengdu, 611731, China.

***Correspondence**:Junxiang Wang

Email:dave_wangjx@aliyun.com

Abstract

An accurate instrument tag can persist after the instrument is removed, exposing a distinction between recognition and source grounding. We introduce StemGround to measure this distinction and train source-aligned readouts using ground-truth multitrack recordings. Same-recording target removal, ordered gain changes, and target-by-context interventions specify which outputs should change and which label truths should be preserved. Loudness matching, whole-mixture gain, and unrelated-source controls test acoustic alternatives. Selective response alignment (SRA) implements these require- ments through a 263,171-parameter residual head, leaving the music encoder fixed. We audit MERT, MuQ, MusicFM, and a MuQ-MuLan zero-shot reference, then compare SRA with an ordinary head matched for capacity, input data, and selected training updates. Both heads receive labeled interven- tions. Improvements on artist-disjoint MoisesDB evaluation replicate on the frozen MUSDB18 official test: 50 tracks from 46 artists, with 13,361 interventions across 49 eligible tracks. Relative to ordinary training, SRA increases standardized target-removal efect by 0.664 and dose-order consistency by 0.043. It reduces target-absent false positives by 0.161, net context-induced false positives by 0.300, and standardized target-absent context efect by 0.756. All five endpoint-wise 95% artist-cluster con- fidence intervals favor SRA; cross-model natural-task and of-target noninferiority criteria are met. MusicFM exhibits a target-removal trade-of, and probability reliability worsens. StemGround estab- lishes a measurable and trainable connection between physical source changes and instrument-label responses, distinct from natural recognition accuracy and probability calibration.

Keywords: music foundation models, instrument recognition, multitrack audio, source intervention, context dependence, selective response alignment, music information retrieval

1 Introduction

Instrument-recognition systems assign source labels to recordings for music retrieval, annotation, and editing. Music foundation models make these labels accessible through readouts fitted to reusable audio representations [1–4]. Natural test accuracy evaluates whether the labels are recoverable, but not whether a decision follows the instrument's own sound. When bass and drums co-occur, a correct bass prediction can depend on bass evidence, drum evidence, or both. Removing bass while retaining the accompaniment distinguishes these possibilities at the system output.

The technical problem is selective response, not indiscriminate robustness. Removing a target changes its label truth; changing another source need not. Raising target gain defines an ordered physical manipulation, but does not imply a linear

change in classification probability. Accompaniment introduced while the target remains absent should not generate a target-positive decision. A useful grounding specification must therefore con- nect each source operation to both an expected target response and the labels that should remain correct.

Ground-truth multitracks allow these relations to be tested within the same recorded performance [5, 6]. Unlike cross-recording replacement, their recombination preserves the source realization and timing. Unlike an estimated separation, it does not require an additional separator to define the manipulated components. The remaining acoustic issue is explicit: changing a source also changes mixture level and masking. We address this through paired raw and loudness-matched ren- derings, whole-mixture gain controls, and removal of an unrelated source with a comparable level effect. These controls separate particular explana- tions rather than treating every perturbation as equivalent.

StemGround turns this design into a common language for evaluation and learning. Its audit measures target-removal magnitude, dose order, target-absent decisions, context-induced errors, and truth-aware off-target changes. Its selective response alignment (SRA) objective links the same source operations to absence margins, ordi- nal constraints, context suppression, and label preservation. A small residual head implements the objective on cached embeddings; the founda- tion encoder and original observational readout remain fixed. The methodological contribution is this operation-specific response specification, not a new source separator or foundation-model architecture.

We isolate the contribution of the joint objective using an ordinary residual head with the same capacity, natural and interventional examples, paired initialization, and selected train- ing updates. Ordinary training already receives the intervention labels. The comparison there- fore tests whether explicit response relations add value beyond condition-wise supervision on the same inputs. We first establish the audit on MoisesDB, then evaluate the learned heads on artist-disjoint internal data and the independent MUSDB18 official test without refitting. Together, these experiments support source-interventional auditing and readout alignment as a practical way to measure and improve instrument grounding while reporting natural-task retention, off-target behavior, and probability reliability separately.

2 Related work

2.1 Music representations and observational evaluation

MERT combines acoustic and musical predic- tion targets during self-supervised pretraining; MuQ uses Mel residual vector quantization and extends its representation to the MuQ-MuLan audio–text system; MusicFM studies scalable self- supervised learning for music informatics [1– 3]. MARBLE organizes representation evaluation across 18 tasks and 12 datasets [4]. These studies establish the utility of learned representations and motivate frozen-readout evaluation. StemGround asks a complementary question: once an instru- ment label is predictable, which physical source changes control the readout's output? We audit both linear and nonlinear observational readouts to distinguish representation–readout combina- tions, and include MuQ-MuLan as an audio–text zero-shot reference.

Diagnostic music benchmarks also distinguish recognition scores from evidence use. RUListening uses text-only and noise-input evaluations to iden- tify questions answerable without sufficient audio perception [7]. Beyond Binary Instrument QA constructs OpenMIC-derived tests of genre pri- ors, confusable instruments, longer context, and temporal localization [8]. Its findings show that similar binary-QA accuracy can conceal differ- ent response biases. These studies test question or benchmark conditions. StemGround intervenes on physical sources within a recording and turns the resulting label-response requirements into a readout-training objective.

2.2 Source-based explanations and counterfactual music analysis

Source-based explanation methods use musically meaningful components as perturbation units. audioLIME employs separated sources to con- struct listenable perturbations and explain indi- vidual predictions [9]. Its extended study uses

ground-truth MUSDB18 sources for a demonstration and Slakh experiments to expose deliberately induced instrument co-occurrence dependence [10]. Praher et al. test the fidelity of local explanations using targeted adversarial exam- ples in audio classification [11]. These studies establish source perturbations and co-occurrence diagnosis as prior art. StemGround connects operation-specific response quantities to a joint learning objective, evaluates multiple frozen music encoders, and tests the learned readout through matched training comparisons and independent replication. Its target is a trainable response spec- ification across recordings, rather than a local surrogate explanation of an individual prediction.

Learnable Counterfactual Attention trains a counterfactual branch to steer attention away from misleading regions for singer and genre classification [12]. Beyond Genre uses con- cept activation vectors to diagnose gender- and language-related associations across MERT, Whisper, MuQ, and MuQ-MuLan, and demon- strates gender-related post-hoc adjustment in a MuQ-MuLan Hip-Hop ranking example [13]. These approaches modify attention or concept- based decision directions. StemGround defines the intervention in the recorded mixture and supervises its label-specific output consequences.

Table 1 summarizes these methodological relationships.

2.3 Selective sensitivity and evidence-guided learning

LOEV retains augmentation-specific information in contrastive music representations [14]. Deng et al. examine embedding trajectories under gain, filtering, reverberation, and bitcrushing, showing that sensitivity to effect strength need not yield a globally linear representation [15]. StemGround assigns different roles to target-source gain and global gain: the first defines a target-response requirement, while the second tests a level-based explanation. Accordingly, SRA uses an ordinal target constraint rather than assuming linear gain equivariance.

Right for the Right Reasons provides a broader precedent for specifying the evidence a predictor should use, through input-gradient constraints [16]. SRA expresses evidence requirements through finite source-intervention pairs and chains, with explicit label-truth masks. In generative music, SMITIN uses probes to monitor and intervene in transformer activations [17]. SRA instead changes the readout of a frozen encoder for an existing recording.

2.4 Response alignment and probability calibration

Probability calibration concerns agreement between confidence and empirical correctness. Guo et al. examine calibration in neural classifiers, and Cheng and Vasconcelos address multi-label calibration through a proper asymmetric loss and label-pair regularization [18, 19]. Source-response alignment concerns how a specified label output changes between interventions with known source relationships. We evaluate natural ranking performance, threshold decisions, and probability reliability separately. SRA's losses operate on ordered differences and standardized decision margins; Brier score and expected calibration error (ECE) assess the resulting probabilities.

3 Source-interventional formulation

3.1 Evaluated system and source decomposition

Let m index a frozen music encoder f_m, and let $e_m(x)$ be its pooled representation of an audio waveform x. The output classes are bass, drums, and vocals; k indexes one of these $K = 3$ classes. For each supervised encoder–readout system, A_m is the feature standardizer fitted on natural train- ing data and then held fixed. The standardized representation is

$$\tilde{e}_m(x) = A_m(e_m(x)). \quad (1)$$

Both $e_m(x)$ and $\tilde{e}_m(x)$ are 1,024-dimensional vectors for the three supervised encoders. The transform uses the stored per-feature training means and scales and is applied once. For a fixed linear or multilayer-perceptron readout, the original scalar class score is

$$z^{(0)}_{m,k}(x) = g^{(0)}_{m,k}(\tilde{e}_m(x)). \quad (2)$$

Table 1 Methodological positioning of source-interventional response alignment.

Method	Intervention or supervision	Evaluated or learned object	Relationship to StemGround
Beyond Binary Instrument QA [8]	OpenMIC-based diagnostic task and answer-format changes	Instrument grounding in audio-language model responses	Diagnostic grounding precedent; StemGround measures same-recording target/context source efects and learns a selective readout.
audioLIME and extended study [9, 10]	Binary source perturbations; estimated and ground-truth sources	Local explanation and controlled co-occurrence diagnosis	Establishes source-level explanation; StemGround links paired response quantities to joint readout training and independent evaluation.
LCA [12]	Learned counterfactual attention	Singer and genre classifier	Attention-based correction; StemGround specifies physical source changes and their label responses.
Beyond Genre [13]	Concept activation vectors and adjustment of a genre-ranking direction	Gender/language association audit; gender-related ranking adjustment	Multi-model diagnosis and a MuQ-MuLan ranking-adjustment example; StemGround evaluates target/context source control.
LOEV [14]	Selective augmentation sensitivity	Contrastively learned representation	Selective sensitivity precedent; SRA acts on frozen-encoder outputs through a readout.
Right for the Right Reasons [16]	Annotated input-gradient constraints	Evidence-guided classifier	Specified-evidence supervision; SRA uses finite source operations and truth-aware margins.
StemGround	Ground-truth target, dose, context, and control interventions	Instrument readout audit and joint response alignment	Connects physical source operations, measurable label-response requirements, and matched readout training.

Here $g^{(0)}_{m,k}$ is the class-k output of the observational readout trained on natural recordings; the superscript (0) denotes this original, frozen system. Supervised scores are logits. The MuQ-MuLan zero-shot reference instead uses its fixed positive-minus-negative audio–text similar- ity score, specified in Section 6.2. The evaluated object is the complete scoring system, not the encoder alone. In the audit equations below, encoder, readout, method, seed, and loudness-arm indices are suppressed after fixing the system and arm under comparison.

A target window is a physical audio window paired with a target class; one physical window can support several target windows. Within a fixed target window, S_k is the sum of ground- truth stems assigned to target class k, C_k is the designated context source, and R_k comprises the remaining sources. These waveforms share the same time axis and length. Addition denotes sample-wise source recombination. The natural mixture is

$$x_0 = R_k + C_k + S_k. \tag{3}$$

The subscript 0 denotes the natural target level of 0 dB, not target absence. The non-target background is

$$B_k = R_k + C_k. \tag{4}$$

Thus B_k retains both the context and residual sources. For a finite relative target gain d, measured in dB, the dimensionless linear amplitude multiplier is

$$\alpha_d = 10^{d/20}. \tag{5}$$

We use the six finite gains _18, _12, _6, 0, +3, and +6 dB; 0 dB gives an amplitude multiplier of one. The corresponding target-dose mixture is

$$x_d = B_k + \alpha_d S_k. \tag{6}$$

Target removal is a separate condition in which the target multiplier is set to zero:

$$x_{\text{mute}} = B_k. \tag{7}$$

Mute is not a finite dB dose and does not silence the non-target sources. Equations (3)– (7) specify source recombinations before the level adjustments in Sections 4.1–4.2. When these condition symbols are used as arguments of a scoring function, they denote the corresponding rendered inputs for the specified loudness arm.

3.2 Response requirements

We operationalize source grounding through four response properties. Target sensitivity measures the score change when a target is present versus removed. Ordinal consistency evaluates whether scores follow increasing target gain over the specified dose range. Context selectivity tests whether non-target accompaniment creates tar- get evidence when the target is absent. Label selectivity concerns labels whose known truth remains unchanged by the operation. These are finite input–output requirements on the encoder– readout system, not assumptions about individual neurons or internal causal pathways.

These properties describe different aspects of the same decision. A strong target-removal effect can coexist with false positives after removal if both the present and absent scores lie above the decision threshold. Likewise, correct dose order does not establish that an absent target is rejected. The audit therefore retains separate magnitude, order, threshold, and preservation endpoints.

3.3 From audit to learning

The study separates measurement from adaptation. The original-system audit applies source interventions to readouts trained on natural recordings, exposing behavior that natural recognition scores leave unresolved. SRA subsequently uses development interventions to learn the spec- ified response relations. A matched ordinary head identifies the effect of the complete selective objec- tive, while the original readout anchors reten- tion. Artist-disjoint internal evaluation and frozen cross-dataset evaluation then test the learned sys- tem on their respective recording populations. Sections 4–6 define the interventions, objective, and comparisons in that order.

Figure 1 connects the response specification to the trainable readout and evaluation design.

4 StemGround audit

4.1 Source eligibility and rendering

Recordings are represented at 44.1 kHz in stereo for source recombination. Candidate windows are 10 s long, drawn from a 10 s grid with at least 20 s separation between selected windows for the same track–target pair. At most three windows are selected per track and target. Activity is mea- sured using 100 ms frames with a 50 ms hop and a _60 dBFS activity threshold. Eligible target and context sources are active in at least 60% and 50% of frames, respectively. The target-to-background ratio must be at least _24 dB, and an active resid- ual source outside target and context must exceed _70 dBFS. Windows are ranked by target activity, context activity, and start time before apply- ing the spacing constraint. These criteria select source interventions using audio and annotation properties rather than model responses.

Natural labels are positive at activity of at least 20%, negative at activity of at most 5%, and otherwise missing. Missing labels are excluded from truth-dependent losses and metrics. Non- muted dose conditions retain a present target label; source mute and factorial operations explic- itly update the affected labels.

The target, context, and residual components are recombined linearly. For each target window, the renderer computes the maximum peak across all 31 conditions and applies a single shared atten- uation if needed to respect a _1 dBFS peak ceiling. This common headroom operation pre- serves the intended differences between conditions. No limiter or compressor is applied.

The 31-condition design comprises 14 target-dose conditions (seven levels in each of two loud- ness arms), seven whole-mixture gain controls, two unrelated-source placebo conditions, and eight target-by-context factorial conditions. The context mapping is selected from development-group source coactivity and held fixed: bass uses drums as context, drums uses vocals, and vocals uses drums.

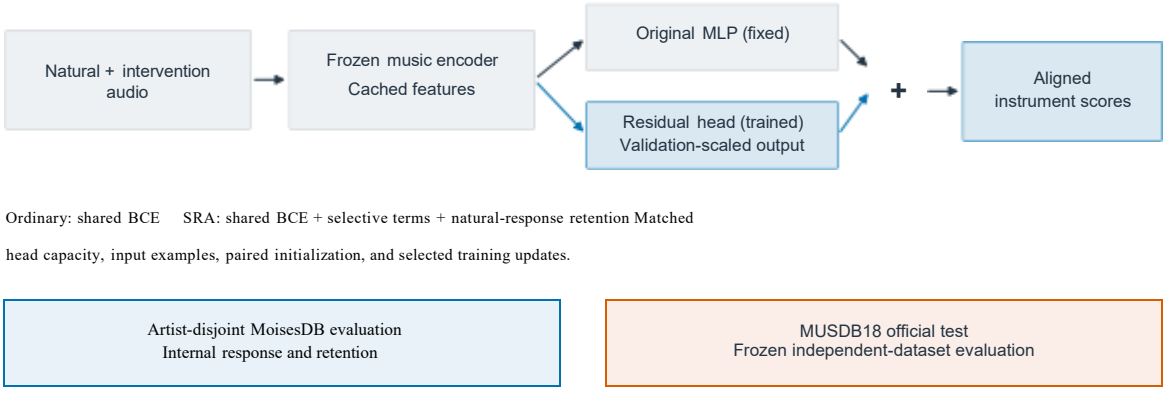
a Physical operations specify different output requirements

Same-recording ground-truth sources: target + context + residual

Source operation	Required response	Loss / audit endpoint
Target removal	Absent target rejected	Absence / TRE, TA-FPR
Ordered target gain	Ordered target scores	Ordinal / DOC
Context added; target absent	No false target evidence	Context / CIFP-net, CPE0
Known label truth unchanged	Correct label side preserved	Preservation / OTNER

Controls: raw / loudness-matched arms, whole-mixture gain, unrelated-source removal.

b Learn the response specification without changing the encoder



Ordinary: shared BCE SRA: shared BCE + selective terms + natural-response retention Matched head capacity, input examples, paired initialization, and selected training updates.

Fig. 1 Source-interventional response specification and selective readout learning. (a) Physical source operations define distinct target and truth-preserving response requirements, their training terms, and audit endpoints. TRE measures target sensitivity; TA-FPR measures rejection after removal. The preservation loss uses operation-specific label masks, whereas OTNER evaluates eligible non-target labels. (b) A trainable residual head adjusts the fixed observational MLP in original validation-logit units; the music encoder remains frozen. Both ordinary training and SRA receive labeled natural and interventional inputs. SRA adds selective terms and natural-response retention. The selected systems undergo artist-disjoint MoisesDB evaluation and frozen MUSDB18 replication.

4.2 Loudness and unrelated-source controls

In the raw arm, only the specified source opera- tion is applied before common headroom attenu- ation. In the LUFS-matched arm, a single linear gain matches the recombined mixture's integrated loudness to the natural 0 dB target condition; the factorial arm uses its target-present/context- present mixture as the reference. The accepted loudness error is 0.10 LU. An explicitly recorded RMS fallback handles cases in which integrated loudness is unavailable.

Whole-mixture gain controls hold source balance fixed. For each target dose, a scalar gain adjusts the natural mixture to the loudness of the corresponding raw target-dose mixture. Com- paring these responses with target-specific manip- ulation tests how much of the target effect is explained by overall level.

For the raw-arm target-removal comparison, let $父_{\text{gain}}$ be the whole-mixture gain control matched to the loudness of the raw target-muted mixture $父_{\text{mute}}$. Both inputs share the target window's common headroom attenuation. For a fixed scoring system and target window, the gain-adjusted source-removal contrast is

$$G_k = \frac{z_k(x_{\text{gain}}) - z_k(x_{\text{mute}})}{s_k}. \quad (8)$$

The positive reference scale s_k is the sam- ple standard deviation, with one degree of free- dom removed ($ddof = 1$), of the original scor- ing system's outputs on natural validation probe examples with valid class-k labels; reference-only examples are excluded. For supervised systems these outputs are logits; for the zero-shot refer- ence they are audio–text contrast scores. Ordinary and SRA heads use the scale of their correspond- ing original MLP. No evaluation data are used to estimate or update this scale. Thus Eq. (8) is dimensionless and equals the target-removal effect after subtracting the response to the correspond- ing whole-mixture level change. A positive value indicates a larger score reduction from removing the target than from changing overall level alone.

For the placebo control, the target and desig- nated context remain unchanged while an active unrelated source is removed. The renderer selects the candidate whose removal produces the closest overall level change to target removal. Selection uses source identity and measured level, with lexi- cal tie-breaking; it does not use model output. The primary matched subset requires a level-change mismatch of at most 1 dB. Match-failed cases remain identifiable and are analyzed separately. This control tests source specificity at compara- ble level changes without assuming that every recording contains a suitable matched unrelated source.

4.3 Target-by-context experiment

The factorial design independently varies target and context presence within a target window. In x_{ab}, the first binary index a denotes target presence and the second binary index b denotes context presence; 0 means absent and 1 means present. The residual source R_k is retained in all four conditions.

With both target and context absent, the input contains only the residual:

$$x_{00} = R_k. \quad (9)$$

With the target absent and context present, the input is

$$x_{01} = R_k + C_k. \quad (10)$$

With the target present and context absent, the input is

$$x_{10} = R_k + S_k. \quad (11)$$

With both target and context present, the input is

$$x_{11} = R_k + S_k + C_k. \quad (12)$$

These definitions specify source composi- tion before arm-specific loudness adjustment and shared headroom attenuation. The rendered x_{11} is the factorial reference condition. The indices describe the presence of the designated sources, not whether the entire recording is silent or whether every instrument label is absent.

For each window, adding the target while retaining context gives the direct target-score contrast

$$D_k^{(1)} = z_k(x_{11}) - z_k(x_{01}). \quad (13)$$

Adding the target without context gives the corresponding contrast

$$D_k^{(0)} = z_k(x_{10}) - z_k(x_{00}). \quad (14)$$

The superscripts in Eqs. (13)–(14) indicate context presence, not model versions. Each con- trast compares two conditions from the same target window and loudness arm.

In the following endpoint definitions, $\mathrm{E}_{\text{eval},k}$ denotes an empirical evaluation average for a fixed class, scoring system, and loudness arm. Eligi- ble paired windows are first aggregated within their tracks, and represented tracks receive equal weight. Section 6.3 specifies the additional class, model, and seed aggregation and artist-cluster inference. The target-absent context effect is

$$CPE_{0,k} = \mathrm{E}_{\text{eval},k}\left[z_k(x_{01}) - z_k(x_{00})\right]. \quad (15)$$

Here the subscript 0 means that the target is absent in both conditions. The effect is signed and expressed in score units. Its standardized version is

$$CPE_{0,k}^{\text{std}} = \frac{CPE_{0,k}}{s_k}. \quad (16)$$

The scale s_k is the fixed original validation-score standard deviation defined in Section 4.2. A positive CPE0 indicates that context increases

the target score while the target remains absent; whether this produces an erroneous positive decision additionally depends on the operating threshold.

The interaction contrast compares the two direct target effects:

$$INT_k = \mathbb{E}_{\text{eval},k}\left[D_k^{(1)} - D_k^{(0)}\right]. \tag{17}$$

A positive interaction means that the target has a larger score effect with context than without it; a negative interaction means the reverse. This is nonadditivity on the evaluated score scale, not an error criterion by itself. CPE0 and the decision endpoints below distinguish target-absent score changes from false target-positive predictions.

4.4 Target and context endpoints

The target-removal effect (TRE) compares the natural 0 dB target condition with target mute:

$$TRE_k = \mathrm{E}_{\text{eval},k}\left[z_k(x_0) - z_k(x_{\text{mute}})\right]. \tag{18}$$

Both inputs come from the same target window and loudness arm, including their common headroom coefficient. Positive TRE indicates a higher target score before removal. The dimensionless effect used for comparisons across encoders and aligned heads is

$$TRE_k^{\text{std}} = \frac{TRE_k}{s_k}. \tag{19}$$

The original validation scale s_k remains fixed for every method comparison and evaluation dataset.

Dose-order consistency (DOC) is defined per target window i, whose target class is k_i. Let $d_1 < \cdots < d_J$ be the ordered finite gains and let x_{i,d_a} be the rendered input at gain d_a in the fixed loudness arm. For dose indices $1 \le a < b \le J$, the score difference is

$$\Delta_{i,ab} = z_{k_i}(x_{i,d_b}) - z_{k_i}(x_{i,d_a}). \tag{20}$$

DOC assigns one point to a correctly ordered pair and half a point to a tie:

$$DOC_i = \frac{2}{J(J-1)} \sum_{1 \le a < b \le J} \left(\mathbf{1}[\Delta_{i,ab} > \epsilon] + \frac{1}{2}\mathbf{1}[|\Delta_{i,ab}| \le \epsilon]\right). \tag{21}$$

Here $\mathbf{1}[\cdot]$ is one when the enclosed condition is true and zero otherwise. The factor outside the sum divides by the number of distinct ordered dose pairs. The primary gains are −12, −6, 0, and +3 dB, so $J = 4$; the score-space tie tolerance is $\epsilon = 10^{-10}$. DOC lies in $[0, 1]$, reaches 1 when every higher-dose score exceeds the lower-dose score by more than this tolerance, and equals 0.5 for an all-tie sequence. Its expected value under random ordering is 0.5. The wider finite-dose range is a sensitivity analysis; mute is not a finite dB observation. Kendall correlation and adjacent-dose violations complement this pairwise measure.

For class k, let t_k denote the operating threshold of the system, method, and seed currently being evaluated. Its binary instrument-presence decision is

$$\hat{y}_k(x) = \mathbf{1}[z_k(x) \ge t_k]. \tag{22}$$

Each threshold is selected on natural validation data and frozen before evaluation. It is not necessarily shared between methods. Section 5 distinguishes the original threshold that anchors training losses from the final operating threshold.

The target-absent false-positive rate is

$$\text{TA-FPR}_k = \mathrm{E}_{\text{eval},k}\left[\hat{y}_k(x_{\text{mute}})\right]. \tag{23}$$

TA-FPR lies in $[0, 1]$. The evaluated target is absent in every eligible dose-mute condition in this denominator. This rate does not pool all target-absent factorial conditions with dose mute; those conditions supply the context endpoints.

Context-induced false positives (CIFP) compare x_{00} with x_{01} while the target remains absent. The new-error component is

$$CIFP_{\text{new},k} = \mathrm{E}_{\text{eval},k}\left[\mathbf{1}[\hat{y}_k(x_{00}) = 0 \wedge \hat{y}_k(x_{01}) = 1]\right] \tag{24}$$

The correction component records opposite transition:

$$CIFP_{\text{correction},k} \tag{25}$$

$$= \mathsf{E}_{\text{eval},k}\,[\mathbf{1}[\hat{y}_k(x_{00}) = 1 \;\wedge\; \hat{y}_k(x_{01}) = 0]]\,.$$

The symbol $\wedge$ means that both enclosed conditions must hold. Both components use all eligible target-absent context pairs under the same averaging rule, rather than conditioning their denominators on the initial prediction. The net contrast is

$$CIFP_{\text{net},k} = CIFP_{\text{new},k} - CIFP_{\text{correction},k}. \tag{26}$$

Thus CIFP-net is the change in target-positive decision rate when context is introduced while the target remains absent. The two components lie in $[0, 1]$ and their net difference lies in $[-1, 1]$. We report all three alongside the threshold- independent standardized CPE0 in Eq. (16).

4.5 Truth-aware of-target evaluation

Each operation is paired with a label-response specification. Target mute changes target truth; context removal changes the context label where that class is represented; placebo removal changes the removed source's label. Only non-target labels with known, unchanged truth enter the corre- sponding off-target comparison.

Let p index a directed before/after intervention pair, k_p its designated target class, and x_p^{before} and x_p^{after} its two rendered inputs. The eligible non-target label set is

$$U_p = \{\ell \in \{1,\dots,K\} \setminus \{k_p\} : y_{p,\ell}^{\text{before}} = y_{p,\ell}^{\text{after}} \in \{0, 1\}\}. \tag{27}$$

Here ℓ indexes an output class, and the two y values are the known label truths in the paired conditions. Missing or changed truths are excluded. For $\ell \in U_p$, write $y_{p,l}$ for their shared truth. If U_p is nonempty, the pair-level off-target new error rate (OTNER) is

$$OTNER_p = \frac{1}{|U_p|} \sum_{\ell \in U_p} \mathbf{1}\big[\hat{y}_\ell(x_p^{\text{before}}) = y_{p,\ell} \wedge \hat{y}_\ell(x_p^{\text{after}}) \neq y_{p,\ell}\big]. \tag{28}$$

OTNER lies in $[0, 1]$. The cardinality $|U_p|$ counts all eligible labels, including those already misclassified before the operation; only a transition from correct to incorrect contributes to the numerator. Pairs with no eligible labels are omit- ted, not assigned a zero error rate. Predictions use Eq. (22).

The before state is the source-present reference for that intervention pair. Each target window contributes 27 directed contrasts. Within each loudness arm, six compare dose zero with the other target doses, one compares dose zero with placebo removal, and three start from x_{11}. The remaining seven compare whole-gain controls with raw dose zero. Equation (28) defines a single-pair quantity; reported OTNER averages eligible contrast–label entries within each model and track, rather than weighting pair-level rates equally. Paired seed differences are averaged before this aggregation, and cross-model estimates give equal weight to models within tracks and then to tracks. Error corrections are counted separately.

For an eligible label and either condition v, the signed truth margin is

$$M_{p,\ell}^{v} = (2y_{p,\ell} - 1)\frac{z_\ell(x_p^v) - t_\ell}{s_\ell}. \tag{29}$$

The superscript v denotes either before or after, t_l is the evaluated system's frozen operat- ing threshold, and s_l is its original validation-score scale. A positive margin places the score strictly on the truth-consistent side of the threshold; a nega- tive margin places it on the opposite side. Zero is the decision boundary, at which Eq. (22) predicts presence. A decrease in this signed margin mea- sures erosion of correct-side evidence. The evalua- tion set U_p is distinct from the operation-specific training preservation mask defined in Section 5.3.

5 Selective response alignment

SRA learns a score-correction function over a fixed music representation. Source stems are used offline to construct labeled examples and relations between them. The learned function receives only an audio embedding: at inference, it needs nei- ther stems nor intervention labels. The ordinary comparator has the same inputs and architecture; SRA differs in the response relations used to train that function.

5.1 Standardized residual readout

SRA is fitted separately for each of the three supervised encoders, using its original observational MLP as the fixed anchor. The aligned score vector is

$$z_m(x) = z_m^{(0)}(x) + s_m \odot h_{\theta_m}(\tilde{e}_m(x)). \tag{30}$$

Here $z_m(x)$ and $z_m^{(0)}(x)$ are K-dimensional logit vectors, $\tilde{e}_m(x)$ is the standardized feature in Eq. (1), and s_m is the vector of positive original- MLP validation-logit standard deviations defined in Section 4.2. The symbol $\odot$ denotes element- wise multiplication. The residual output h_{θ_m} is dimensionless; multiplication by s_m converts it to original logit units. Only the residual parameters θ_m are trainable. The zero-shot reference is not adapted.

The residual head maps 1,024 inputs through 256 hidden units to three outputs. It uses Gaus- sian error linear unit (GELU) activation and dropout of 0.10. Its output layer is initialized to zero, so the initial combined system reproduces the original logits. Each head contains 263,171 trainable parameters. Encoder features, the origi- nal MLP, the feature standardizer, and logit scales remain fixed.

For each class, the reference scale uses the sample standard deviation ($ddof = 1$) of origi- nal logits on natural validation probe examples with valid labels, excluding reference-only exam- ples. Feature standardization and logit scaling are distinct operations: the former acts on the 1,024 input features, whereas the latter sets the units of the three output adjustments. Training uses cached features and does not require repeated foundation-model inference. Encoder subscripts are omitted in Sections 5.2–5.5, where one fixed encoder and residual head are considered.

5.2 Common supervised objective and ordinary comparator

Both ordinary training and SRA receive the same natural mixtures and labeled interventions. Let N and I be the indexed collections of natural and interventional fitting examples, respectively. The natural collection contains training probe- pool records and excludes reference-only scoring rows. Each x denotes a record together with its audio input; distinct record identities are retained even when their waveforms coincide. For class k, $y_k(x)$ is a binary presence label when known and is otherwise excluded from label-dependent terms.

The positive-class weight is computed from valid natural fitting labels:

$$p_k = \frac{n_k^-}{n_k^+}. \tag{31}$$

The counts n_k^- and n_k^+ are the numbers of valid negative and positive natural fitting examples for class k. Both classes of label are represented for each output. The sigmoid maps a logit v to a probability:

$$\sigma(v) = \frac{1}{1+\exp(-v)}. \tag{32}$$

For a valid example–class cell, the weighted binary cross-entropy (BCE) is

$$\ell_{\mathrm{BCE}}(x,k) = -p_k y_k(x)\log\sigma(z_k(x)) - (1-y_k(x))\log(1-\sigma(z_k(x))). \tag{33}$$

The logarithm is natural, and p_k multiplies only the positive-label term. The score $z_k(x)$ is the current residual readout in Eq. (30). Let V_N collect all valid example–class cells in N. Natural BCE averages these cells:

$$L_{\text{natural BCE}} = \frac{1}{|\mathcal{V}_{\mathcal{N}}|}\sum_{(x,k)\in\mathcal{V}_{\mathcal{N}}}\ell_{\mathrm{BCE}}(x,k). \tag{34}$$

For interventional inputs, let G_+ be the collection of target-window groups with at least one valid label cell, and let V_g collect the valid cells in group g. Each group contains the 31 rendered conditions for one target window, including both loudness arms and the controls. Intervention BCE gives equal weight to groups after averaging their valid cells:

$$L_{\text{intervention BCE}} = \frac{1}{|\mathcal{G}_+|}\sum_{g\in\mathcal{G}_+}\frac{1}{|\mathcal{V}_g|}\sum_{(x,k)\in\mathcal{V}_g}\ell_{\mathrm{BCE}}(x,k). \tag{35}$$

In Eqs. (34)–(35), $|\cdot|$ denotes collection cardinality. These denominators count valid cells and

represented groups, not sums of class weights. The shared supervised objective is

$$L_{\text{common}} = \frac{1}{2} L_{\text{natural BCE}} + \frac{1}{2} L_{\text{intervention BCE}}. \tag{36}$$

The ordinary comparator minimizes Eq. (36) using the same residual architecture as SRA. It therefore sees target absence, source-dose changes, context manipulations, and their updated labels.

Condition-wise BCE provides the same source-presence labels to both methods, but does not explicitly constrain the ordering of logits across a source-dose sequence or the direction of a paired context effect. SRA adds those relations and operation-specific correct-side margins. This distinction defines the method comparison without withholding intervention examples from ordinary training.

5.3 Joint selective objective

Let $t_k^{(0)}$ be the original MLP's validation-selected threshold. For the current residual readout, the standardized margin relative to this fixed thresh- old is

$$u_k(x) = \frac{z_k(x) - t_k^{(0)}}{s_k}. \tag{37}$$

Positive and negative u_k values lie above and below the original decision boundary, respec- tively. The original thresholds anchor training; the final ordinary and SRA operating thresholds are selected separately on validation data and then frozen. The scale s_k is the fixed original- MLP validation-logit standard deviation, so u_k is dimensionless.

Four indexed response collections are constructed from the fitting interventions. The collection A_{abs} contains target-absent entries (x, k) from dose mute, x_{00}, and x_{01}. The collection O contains ordered target-dose pairs (x_a, x_b, k) from the same target window and loudness arm, with x_a preceding x_b in the seven-level order: mute, −18, −12, −6, 0, +3, and +6 dB. The collection C_{ctx} contains target-absent context pairs (x_{00}, x_{01}, k) from the same target window and loudness arm. Finally, P contains intervention–label entries (x, ℓ) selected by the operation-specific truth- preservation mask described below. Both loudness arms contribute their eligible records; these collections retain their record and pair identities.

We use the positive-part operator $[v]_+$, equal to zero for nonpositive v and to v otherwise. This is the rectified linear unit (ReLU). The smooth absence penalty uses softplus:

$$\text{softplus}(v) = \log(1 + \exp(v)). \tag{38}$$

The joint SRA objective is

$$L_{\text{SRA}} = L_{\text{common}} + \lambda_{\text{tar}}(L_{\text{absence}} + L_{\text{ordinal}}) + \lambda_{\text{ctx}} L_{\text{context}} + \lambda_{\text{pres}} L_{\text{preserve}} + \lambda_{\text{ret}} L_{\text{retention}}. \tag{39}$$

The four nonnegative weights λ_{tar}, λ_{ctx}, λ_{pres}, and λ_{ret} control the target, context, preserva- tion, and natural-retention terms, respectively. The same target weight multiplies absence and ordinal penalties. The nonnegative margins ρ_{abs}, ρ_{ord}, ρ_{ctx}, and ρ_{pres}, and the nonnegative context tolerance δ, are expressed in standardized logit units. Section 6.5 specifies their selection and the retained configuration.

Target absence. The target is absent in every $(x, k) \in A_{\text{abs}}$. Its loss is

$$L_{\text{absence}} = \frac{1}{|\mathcal{A}_{\text{abs}}|} \sum_{(x,k)\in\mathcal{A}_{\text{abs}}} \text{softplus}\big(u_k(x) + \rho_{\text{abs}}\big). \tag{40}$$

This smooth loss pushes target-absent scores below the original threshold, including when a large removal effect still leaves the absent score positive. Softplus remains positive beyond the prescribed negative-side margin and decreases smoothly as $u_k(x)$ becomes more negative, for fixed ρ_{abs}.

Ordinal target response. The ordered-pair penalty is

$$L_{\text{ordinal}} = \frac{1}{|\mathcal{O}|} \sum_{(x_a,x_b,k)\in\mathcal{O}} \left[\rho_{\text{ord}} - \frac{z_k(x_b) - z_k(x_a)}{s_k}\right]_+. \tag{41}$$

Each seven-level chain contributes all 21 pairs within its loudness arm, including mute and the two outer stress doses. The term is zero for a pair whose standardized score increase reaches ρ_{ord}. It imposes a minimum ordered difference, not a fixed relationship between dB and probability. The primary DOC endpoint remains the four-level natural-range measure in Eq. (21).

Target-absent context response. For paired conditions in which the target remains absent, the context loss is

$$L_{\text{context}} = \frac{1}{|\mathcal{C}_{\text{ctx}}|} \sum_{(x_{00},x_{01},k)\in\mathcal{C}_{\text{ctx}}} \Big([u_k(x_{01}) - u_k(x_{00}) - \delta]_+ + [\max\{u_k(x_{00}), u_k(x_{01})\} + \rho_{\text{ctx}}]^+\Big). \quad (42)$$

The first penalty applies only when adding context raises the standardized target score by more than δ. The second applies one hinge penalty to the larger of the two target-absent margins, requiring both to reach the prescribed negative side of the original threshold. It is not the sum of two separate absence hinges. Small context effects and context-related score reductions remain com- patible with the specification.

Truth-preserving labels. For valid intervention–label entries selected by P, the preservation loss is

$$L_{\text{preserve}} = \frac{1}{|\mathcal{P}|} \sum_{(x,\ell)\in\mathcal{P}} [\rho_{\text{pres}} - (2y_\ell(x) - 1)u_\ell(x)]_+ . \quad (43)$$

The known binary truth $y_\ell(x)$ determines the required direction: the multiplier is positive for a present label and negative for an absent label. The loss is zero once the score reaches the correct-side margin ρ_{pres} relative to the original threshold.

The mask excludes the target for all dose changes, both target and context for facto- rial changes, and the removed class for placebo changes. Whole-mixture gain preserves all known labels. Consequently, the target label can enter preservation for placebo and global-gain con- trols, whereas OTNER always concerns non-target labels. This operation-specific selection differs both from a uniform non-target mask and from selecting every label whose binary truth happens to remain unchanged.

Natural-task retention. The final term penalizes changes to the original response vector on unmodified natural inputs:

$$L_{\text{retention}} = \frac{1}{K|\mathcal{N}|} \sum_{x\in\mathcal{N}} \left\| \left(z(x) - z^{(0)}(x)\right) \oslash s \right\|_2^2 . \quad (44)$$

The symbol $\oslash$ denotes element-wise division, $\|\cdot\|_2$ is the Euclidean norm, and $K = 3$ is the number of output classes. Equation (44) aver- ages squared standardized logit changes over all natural records and all output classes, including cells with missing truth labels. Natural BCE in Eq. (34) instead averages only valid label cells.

The response losses in Eqs. (40)–(43) use arith-metic means over their respective indexed entries or pairs; they do not inherit the target-window group weighting of intervention BCE. All four response collections are nonempty in the fitting data.

The joint objective therefore treats target changes, target absence, accompaniment, unaf-fected labels, and natural inputs according to their different roles.

Figure 2 makes the distinction from condition-wise supervision explicit. A label says whether a source is present in one input; an SRA relation specifies how scores for two or more related inputs should compare.

5.4 Optimization and operating thresholds

For a fixed training configuration, SRA optimizes only the residual parameters θ. The observational readout, feature standardizer, original thresholds $t^{(0)}$, and logit scales s are fixed. Zero initializa- tion of the residual output layer makes the initial combined readout equal to the original system. Ordinary training uses the same architecture and labeled examples, but minimizes only L_{common}.

The original thresholds $t^{(0)}$ anchor the response losses; the final operating thresholds τ determine binary predictions. These are distinct quantities. The training configuration and operat- ing thresholds are selected on the validation set and retained unchanged for evaluation. Section 6.5

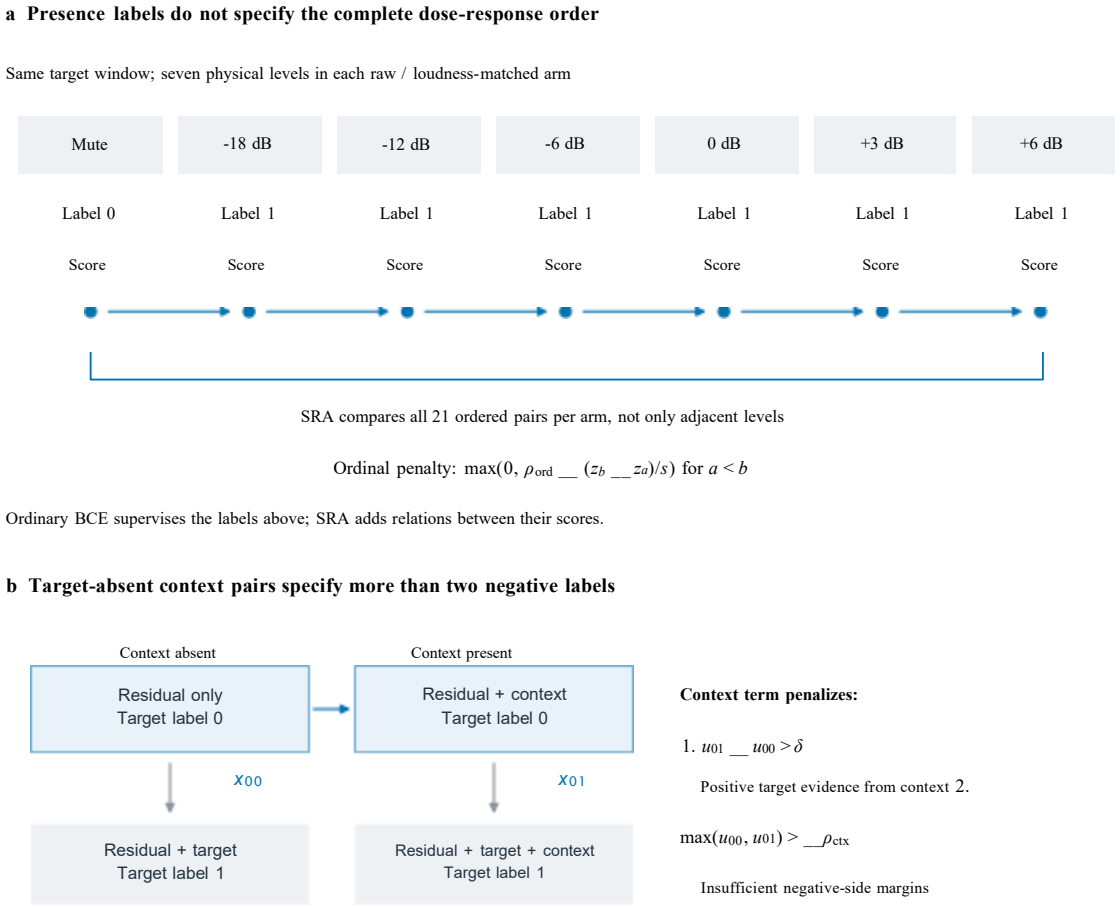


Fig. 2 Response relations added to condition-wise supervision. (a) Ordinary BCE and SRA share the same presence labels: mute is negative and every non-muted dose is positive. SRA additionally penalizes insufficient score separation for all 21 ordered pairs within each seven-level dose chain and loudness arm. The arrows denote the ordering relation, not measured score magnitudes; the training pairs are not limited to adjacent levels. (b) Both target-absent factorial mixtures have a negative target label. The context term separately penalizes context-induced score increases above its tolerance and failure to attain negative-side margins. Here u is the standardized margin relative to the original threshold. Gray boxes show the target-present counterparts. These are training penalties, not claims that every constraint is exactly satisfied.

specifies the matched optimization budget and shared selection procedure.

5.5 Learning and inference algorithms

Algorithm 1 presents SRA for one fixed encoder and one residual head. The indexed fitting collec- tions N and I, their valid-label cells, and target- window groups are defined in Section 5.2; the four response collections are defined in Section 5.3. All original logits and standardized encoder features are fixed. The algorithm optimizes the residual head, while Section 6.5 specifies validation selec- tion and pairing with the ordinary comparator.

Here j indexes optimization updates and d_θ is the gradient with respect to residual parame- ters only. Each update uses one residual forward pass on N and one on I; all losses reuse those logits. The original scores remain fixed. Missing labels are omitted from BCE and truth preser- vation, whereas natural retention averages all K output cells. Ordinary training follows the same update procedure with all four selective weights

Algorithm 1 Selective response alignment for residual readout learning

Require: Natural and interventional fitting collections N, I with standardized features, labels, and source- operation identities; fixed original logits $z^{(0)}$, original thresholds $t^{(0)}$, and positive scales s; four loss weights λ, four margins ρ, and context tolerance δ; number of full-batch updates T, AdamW settings, and positive gradient-norm bound c_{clip}.

Ensure: Residual readout h_θ defining the aligned logits in Eq. (30).

Construct the response collections $A_{\text{abs}}, O, C_{\text{ctx}}, P$ from source operations and known label truths.

Compute positive-class weights from natural fitting labels using Eq. (31).

Initialize θ with a zero residual output layer; initialize the AdamW optimizer.

for $j = 1, \ldots, T$ **do**

  Compute the aligned logits for the complete N and I collections using Eq. (30), with residual dropout active.

  Form the standardized original-threshold margins using Eq. (37).

  Evaluate the shared supervised loss, Eq. (36), using the valid-cell natural average in Eq. (34) and the group-balanced intervention average in Eq. (35).

  Evaluate target-absence and ordered-dose penalties using Eqs. (40)–(41).

  Evaluate target-absent context and truth-preservation penalties using Eqs. (42)–(43).

  Evaluate natural-response retention over all natural output cells using Eq. (44).

  Assemble the joint loss L_{SRA} using Eq. (39).

  Compute $d\theta \leftarrow \nabla_\theta L_{\text{SRA}}$; clip its Euclidean norm to c_{clip}.

  Update θ by one AdamW step using the clipped gradient.

end for

return h_θ.

set to zero, reducing the objective to Eq. (36). Architecture, dropout, optimizer settings, paired initialization, and validation selection are specified in Sections 5.1 and 6.5.

Algorithm 2 applies the selected readout to an audio input. The fixed feature extractor Φ includes the encoder, preprocessing, and pooling specified in Section 6.2, so its output is the representation $e_m(x)$ in Section 3.1. The transform A is the stored natural-training feature standard- izer. Encoder subscripts are suppressed for this single-system inference procedure.

The selected parameters $\theta^\star$ are fixed before evaluation. The residual vector r has $K = 3$ dimensionless components; $z^\star$ and τ are in logit units. No labels or stems are needed at inference. Ground-truth stems construct the training and audit inputs, not an additional input to the learned readout. Feature extraction can be reused when several heads score the same recording. The experiments evaluate each model and seed sep- arately and aggregate their results according to Section 6.3; Algorithm 2 does not define a seed ensemble.

6 Experimental setup

6.1 Data and evidence roles

MoisesDB is divided into training, validation, and internal evaluation sets [5]. The initial ros- ter contains 47 credited artist IDs: 27 assigned to training, eight to validation, and 12 to inter- nal evaluation. Artist-name normalization identi- fied one artist represented under different credit strings in both the training and internal evalua- tion sets. The corresponding evaluation track was excluded, leaving 11 evaluation artists. The eli- gible intervention population is smaller than the natural-task population because source, context, and window criteria must all be met (Table 2).

The MoisesDB training and validation sets together contribute 95 tracks from 29 eligible cred- ited artist units and 833 target windows. The internal evaluation set contributes 24 tracks from 11 artists and 205 target windows. Together, these yield 32,178 intervention instances from 1,038 target windows. For SRA fitting, the training intervention subset contains 73 tracks from 22 eli- gible artist units, 638 target windows, and 19,778 instances. The validation set supplies the remain- ing 195 development target windows and 6,045

Algorithm 2 Inference with the aligned readout

Require: Audio input x; fixed feature extractor Φ, standardizer A, original readout $g^{(0)}$, and positive scales s; selected residual head $h_{\theta^\star}$ and class operating thresholds τ.

Ensure: Aligned logits $z^\star(x)$ and instrument-presence decisions $\hat{y}(x)$.

1: Extract and standardize the representation: $\tilde{e} \leftarrow A(\Phi(x))$, as in Eq. (1).
2: Evaluate the fixed original readout in evaluation mode: $z^{(0)} \leftarrow g^{(0)}(\tilde{e})$.
3: Evaluate the selected residual head with dropout disabled: $r \leftarrow h_{\theta^\star}(\tilde{e})$.
4: Form aligned logits: $z^\star \leftarrow z^{(0)} + s \odot r$, as in Eq. (30).
5: Apply the frozen operating thresholds: $\hat{y}_k \leftarrow \mathbf{1}[z^\star_k \geq \tau_k]$ for $k = 1, \ldots, K$, using Eq. (22) with $t_k = \tau_k$.
6: **return** $(z^\star, \hat{y})$.

instances. The natural examples used for resid- ual training and validation number 941 and 273, respectively.

The excluded evaluation track is "Seattle," credited to "The Unfortunates "; removing the trailing space matches the training-set artist credit. Exclusion removes eight natural probe windows and nine target windows. All reported MoisesDB internal evaluation results were recomputed on the retained observations using the saved pre- dictions, the original estimators, and 2,000 artist- bootstrap draws for the reported intervention and retention intervals. No head, feature scale, decision threshold, or validation-selected configuration was refitted. After whitespace and case normalization, the retained evaluation artist names have zero overlap with both training and validation. This set remains an internal evaluation: its original-system audit preceded SRA development.

Natural-task retention uses a broader MoisesDB internal evaluation population of 78 tracks and 612 eligible probe samples from 11 artists, distinct from the 24 intervention-eligible tracks. The independent evaluation is the MUSDB18 standard-AAC official test [6]. It contains 50 tracks from 46 artists; 49 tracks provide eligible interventions. There are 431 target windows cor- responding to 270 distinct physical windows: 142 bass, 145 drums, and 144 vocals windows. These generate 13,361 conditions. A separate 12-track pilot uses the MUSDB18 training partition and is excluded from the official-test denominators.

All cross-dataset method conclusions concern bass, drums, and vocals, the shared target taxon- omy. Dataset-provided source assignments deter- mine the interventions. MUSDB18's standard AAC stems are used as released, rather than the HQ edition.

6.2 Encoders, observational readouts, and zero-shot reference

The supervised audit uses MERT-v1-330M, MuQ-large-msd-iter, and the MSD version of MusicFM [1–3, 20–22]. Each frozen encoder is evaluated with a linear classifier and an observational MLP trained on natural mixtures from the MoisesDB training set, excluding the original 30-track pilot's ten credited artist IDs. The selected feature lay- ers are 6 for MERT, 3 for MuQ, and 6 for MusicFM. All provide 1,024-dimensional pooled features. Encoder inputs are converted to mono at 24 kHz. MERT processes two 5 s chunks per 10 s window; MuQ and MusicFM process a 10 s win- dow. Hidden states are averaged over time, and MERT's two chunk vectors are averaged.

Natural probe features are standardized using training data only. Linear regularization is selected from $C \in \{0.01, 0.1, 1, 10\}$. Observational MLPs use 256 hidden units, GELU, dropout 0.10, learning rate 0.001, weight decay 0.0001, and batches of 64, with a maximum of 200 epochs and validation patience of 20. Layer and readout selec- tion use natural macro-average precision on the validation set. Class thresholds maximize natural validation F1, choosing the smallest threshold on a tie.

MuQ-MuLan-large supplies a frozen zero-shot reference with two positive prompts and two nega- tive prompts per class [23]. The templates describe a recording with or without the instrument, or whether the instrument is audible. Its score is positive-prompt similarity minus negative-prompt similarity; its operating threshold is selected on the validation set. The seven original model– readout identities are therefore six supervised combinations and one zero-shot system. SRA uses

Table 2 Data coverage and experimental roles.

A. Source-intervention populations.

Dataset and partition	Credited artist IDs	Tracks	Physical windows	Target windows	Interventions	Role
MoisesDB training + validation	29	95	499	833	25,823	Development audit
MoisesDB training	22	73	387	638	19,778	SRA/ordinary fitting; subset of the preceding row
MoisesDB internal evaluation	11	24	115	205	6,355	Artist-disjoint internal evaluation
MUSDB18 training-partition pilot	12	12	62	108	3,348	Separate pilot evaluation
MUSDB18 official test	46	49	270	431	13,361	Independent frozen test

B. Natural-task retention populations.

Dataset and partition	Credited artist IDs	Tracks	Natural probe windows
MoisesDB internal evaluation	11	78	612
MUSDB18 official test	46	50	395

Notes. Each target window generates 31 interventions. Development counts retain their original credited identifiers; the MoisesDB internal evaluation set contains 11 distinct normalized artist names and excludes the training-overlapping track described in Section 6.1. The natural probe pool excludes reference-only scoring windows. One MUSDB18 vocals label is missing, giving 394 valid natural vocals examples. Nested subsets, multiple targets per physical window, and repeated model/seed evaluations are not additive independent samples. The official test uses standard AAC stems and excludes the pilot.

the original MLP of each of the three supervised encoders as its fixed anchor; the zero-shot system remains an audit reference.

For the aligned heads, each model–method–seed combination selects its class thresholds on the validation set using the same maximum-F1 rule. These thresholds, the selected recipe, and all learned parameters are carried unchanged into the MoisesDB internal evaluation set and the MUSDB18 official test. The original validation logit scale remains the common reference for standardized effect sizes.

6.3 Statistical estimation and primary comparisons

Intervention conditions are paired within a target window. For method comparisons, seed-specific responses are paired and then averaged within audio units. Window responses are aggregated to tracks, with the appropriate available class and model dimensions averaged within the track. The primary point estimate is track-equal; repeated windows, seeds, or conditions do not increase its independent sample count.

Confidence intervals use 2,000 percentile-bootstrap draws of whole artist clusters, preserving each artist's observations and paired method comparisons. The bootstrap seed is 20260829. Cross-model results summarize the three evaluated encoders.

Natural macro-mAP is recomputed within each bootstrap draw. Let j index a natural probe example, let track(j) identify its track, and let $n_{\mathrm{track}(j),k}$ count that track's natural probe examples with valid class-k labels. Each valid example–class entry receives weight

$$w_{j,k} = \frac{1}{n_{\mathrm{track}(j),k}}. \tag{45}$$

The count is positive for every included entry, so each represented track contributes total weight one to a class-specific AP calculation. Class-specific average precision (AP) is calculated separately for each seed, then averaged across classes, seeds, and, for cross-model results, encoders. Neither per-track AP scores nor logits across seeds are averaged before this calculation.

The five primary response comparisons are SRA versus ordinary in the LUFS-matched arm: standardized TRE, natural-range DOC, TA-FPR,

CIFP-net, and standardized CPE0. Results are oriented so positive values indicate improvement: SRA minus ordinary for TRE and DOC, and ordinary minus SRA for the three reduction endpoints. Raw-arm results assess sensitivity to loudness matching.

Retention uses the original observational system as the prespecified reference. Natural-task noninferiority requires the lower 95% confidence limit for the macro-mAP difference to be at least -0.01. Off-target noninferiority requires the upper limit for the OTNER increase to be at most +0.005. OTNER aggregates the specified valid contrasts across raw, loudness-matched, and whole-gain control conditions; it is not restricted to the primary LUFS response arm. Its denominator comprises all eligible unchanged-truth non-target entries, including entries already misclassified before the operation. We also report error fixes and signed-margin erosion.

The original-system audit uses two-sided, centered artist-bootstrap tests against 0 for TRE and CIFP-net and against 0.5 for natural-range DOC. Each system–loudness-arm family contains these three class-aggregated tests; Holm correction controls the familywise significance level at 0.05. TA-FPR is reported with artist-cluster confidence intervals rather than a point-null test. The SRA results report paired effects with unadjusted endpoint-wise 95% confidence intervals, not simultaneous coverage across endpoints. Probability reliability is assessed descriptively through track-equal Brier score and ten-bin ECE, using sigmoid-transformed logits on valid natural labels.

6.4 Controls and computational implementation

The completed controls include whole-mixture gain contrasts, matched unrelated-source effects, low-level and pooled log-mel readouts, and artist-grouped low-level cross-fitting. The cross-fit regressions predict system scores (logits for the supervised readouts) from acoustic features across rendered conditions; $R^2 \geq 0.90$ is the prespecified warning level for near-complete prediction. These analyses complement the direct factorial experiment.

Implementation uses Python 3.10, PyTorch 2.7.1 with CUDA 11.8, and model-specific dependency environments. The server provides three working NVIDIA RTX 3090 GPUs with 24 GB memory each. Intervention embedding extraction is performed once per frozen encoder; ordinary and SRA training reuse cached embeddings. Deployment adds one 263,171-parameter residual MLP to the original encoder–readout system. The foundation forward pass is unchanged.

The official-test input audit passed all 127 recorded checks, with no item requiring review. Checks cover split mapping, class and window eligibility, source recombination, clipping, level matching, and inference inputs.

6.5 Optimization and validation selection

The selected weight tuple is $(\lambda_{\text{tar}}, \lambda_{\text{ctx}}, \lambda_{\text{pres}}, \lambda_{\text{ret}}) = (1.0, 0.5, 0.5, 0.5)$. The standardized margins and context tolerance are 0.10. This weight configuration and the model state at update 60 are shared across MERT, MuQ, and MusicFM.

Training uses AdamW with learning rate 3×10^{-4}, weight decay 10^{-4}, and gradient-norm clipping at 5. Each epoch performs one full-batch optimizer update on cached features. The three seeds are 20260829, 20260830, and 20260831. Ordinary and SRA heads start from equal parameter tensors with independent storage and use the same prepared data and update schedule.

Model selection on the validation set first requires each model's three-seed mean change in natural class-macro mean average precision (macro-mAP) to be at least -0.01 and OTNER change to be at most +0.005 relative to original. These are development point-estimate tolerances. Eligible candidates are compared on TRE, DOC, TA-FPR, and CIFP-net against ordinary training at the same checkpoint. For these comparisons, response windows are first averaged separately within each model–seed–target-class– loudness-arm cell. The corresponding SRA and ordinary cell means are then compared, giving 54 cells for each of the four selection endpoints and 216 oriented values in total. Endpoint means average those cells equally. Selection retains the Pareto-nondominated candidates and prioritizes the number of nonadverse endpoint means, the number of positive cell–endpoint values, and mean standardized improvement. Remaining ties favor a

smaller mean OTNER change, a larger mean natu- ral macro-AP change, less total structured weight, an earlier checkpoint, and then the recipe name in lexical order.

For the ranking tie-breaker, each oriented endpoint improvement is z-standardized across all eligible recipe–checkpoint candidates using the population standard deviation (*ddof* = 0). The four resulting values are averaged. This candidate-ranking normalization is separate from the fixed validation-logit scale used in the losses and response metrics.

Six complete SRA weight recipes are eval- uated at updates 20, 40, 60, and 80; ordinary training follows its single common-BCE objec- tive. The selected comparison matches capacity, data, and training updates, while the development search includes six selective recipe trajectories. The selected recipe and checkpoint are shared across models and seeds. The experimental treat- ment is the complete joint objective in Eq. (39).

The four ranking coordinates are oriented TRE, DOC, TA-FPR, and CIFP-net improve- ments; CPE0 is an evaluation endpoint, not a fifth selection coordinate. Among nondominated candidates, priorities are: more nonadverse mean endpoints; more positive values among the 216 cell–endpoint comparisons; larger mean endpoint z-score; smaller mean OTNER change; larger mean natural macro-AP change; smaller total structured weight; earlier update; and lexical recipe name. The implementation uses a 10^{-12} comparison tolerance. Each z-score uses all eli- gible candidates with *ddof* = 0; a zero-variance coordinate contributes zero.

Natural AP on the validation set is the class mean of window-level AP on valid probe examples. Validation response means follow the cell aggre- gation above. These development selection sum- maries are distinct from the track-equal, artist-bootstrap evaluation estimands in Section 6.3. Each saved candidate has its own validation-selected maximum-F1 thresholds before its selec- tion statistics are computed. The selected thresh- olds are retained, not refitted after official-test scoring.

The procedure runs nine ordinary trajectories and 54 SRA trajectories, with four saved check- points per trajectory. The selected model state uses this same weight configuration at update 60 across models and seeds. It selects neither an indi- vidual best seed nor a separately optimized recipe for each encoder. Candidate readouts are retained intermediate states of the learning trajectories in Algorithm 1; selecting a state does not retrain the head.

7 Results

7.1 Natural recognition and source grounding measure diferent behavior

The audit of the MoisesDB internal evaluation set shows a clear separation between natural recog- nition and target-absent behavior (Table 3). The six trained observational readouts achieve natu- ral macro-mAP values of approximately 0.9956–0.9982. Despite this performance, target removal leaves substantial false-positive rates in the LUFS-matched arm: across trained readouts, the rates range from 0.7536 to 1.0000 for bass, 0.7778 to 0.9583 for drums, and 0.0972 to 0.4931 for vocals. Natural classification ranks these systems within a narrow high-performance range, while source intervention reveals markedly different decision profiles.

The systems nevertheless respond directly to their targets. In the LUFS-matched arm, stan- dardized TRE ranges from approximately 0.4735 to 2.1141 across the 18 supervised model–readout– class cells, and every corresponding confidence interval is positive. DOC ranges from approxi- mately 0.8079 to 1.0000, with all 18 confidence intervals above 0.5. The raw arm independently satisfies these same 18-of-18 directional criteria for TRE and DOC. The readouts detect target changes while often retaining a target-positive decision after removal.

7.2 Acoustic controls and target-absent context efects

Across the 18 MoisesDB internal-evaluation supervised model–readout–class cells, whole-mixture gain adjustment retained 83.5%–98.1% of the mute-related target effect, with a median of 95.2%. Most of the target-removal response persisted after accounting for the corresponding overall level change.

Table 3 Natural recognition and source-interventional behavior of the original systems

A. Natural-task macro-mAP and LUFS-matched intervention estimates. In each paired cell, the first value corresponds to the MoisesDB internal evaluation set and the second to the MUSDB18 official test set.

Frozen encoder and original readout	Natural macro-mAP: MoisesDB / MUSDB18	DOC: MoisesDB / MUSDB18	TA-FPR: MoisesDB / MUSDB18	CIFP-net: MoisesDB / MUSDB18
MERT, linear	0.9982 / 0.9825	0.9834 / 0.9681	0.7662 / 0.6995	0.2176 / 0.2472
MERT, MLP	0.9977 / 0.9864	0.9537 / 0.9556	0.7338 / 0.7211	0.1968 / 0.1995
MuQ, linear	0.9976 / 0.9887	0.9757 / 0.9577	0.6852 / 0.8005	0.1505 / 0.1100
MuQ, MLP	0.9960 / 0.9867	0.9313 / 0.9053	0.6620 / 0.7540	0.1748 / 0.0907
MusicFM, linear	0.9981 / 0.9866	0.9776 / 0.9796	0.5394 / 0.6814	0.2199 / 0.2336
MusicFM, MLP	0.9956 / 0.9874	0.9458 / 0.9537	0.6157 / 0.7313	0.0949 / 0.0964
MuQ-MuLan, zero-shot	0.9321 / 0.9617	0.6314 / 0.6691	0.9676 / 0.9830	—0.0093 / 0.0420

B. Class-aggregated original-system tests surviving the specified Holm correction.

Dataset	Loudness arm	TRE versus 0	DOC versus 0.5	CIFP-net versus 0
MoisesDB internal evaluation	LUFS-matched	7/7	7/7	6/7
MoisesDB internal evaluation	Raw	7/7	7/7	6/7
MUSDB18 official test	LUFS-matched	7/7	7/7	7/7
MUSDB18 official test	Raw	7/7	7/7	7/7

Notes. MERT uses layer 6, MuQ layer 3, and MusicFM layer 6. The evaluated object is the frozen encoder plus its specified readout. Natural macro-mAP averages three class-specific average precisions, computed on natural probe windows with track-equal weights. Intervention columns report class-aggregated, track-equal estimates. DOC is dose-order consistency; TA-FPR is the target-absent false-positive rate; CIFP-net is newly induced context errors minus context corrections. A positive TRE and ordered dose response can coexist with a high target-absent false-positive rate. TA-FPR has confidence intervals but no prespecified point-null test and is excluded from the Holm family. Each Holm family comprises the three class-aggregated null tests for one system and loudness arm. On the MoisesDB internal evaluation set, the only CIFP-net non-rejection in each arm is the zero-shot reference. Appendix Table A1 supplies class-specific TRE, CIFP decomposition, and CPE0 estimates for this evaluation set.

Context interventions identify an error mechanism beyond incomplete target suppression. In the MoisesDB internal evaluation set, adding the designated context under target absence produces a positive class-aggregated CIFP-net for all six supervised readouts in both loudness arms, surviving the specified Holm correction (Table 3B). The zero-shot system does not reject this context- error null on the MoisesDB internal evalua- tion set. Appendix A separates newly induced errors from corrections and reports threshold- independent CPE0, showing which class-specific score changes accompany the decision effects.

The acoustic controls provide complementary measurements. None of the 63 acoustic cross- fit cells in the MoisesDB internal evaluation set reached the near-complete-prediction threshold of $R^2 \geq 0.90$. The maximum was 0.7778. The unrelated-source matching criterion was met for 158 of 205 target windows (77.07%), with lower coverage for vocals. Whole-mixture level responses thus accounted for a small fraction of the removal contrast, while the acoustic regressions remained below their prespecified warning threshold.

Figure 4 separates whole-mixture level effects from source removal and evaluates how well the specified acoustic descriptors predict system scores across artists. In the MoisesDB internal evaluation, all 18 supervised readout–class cells retain a positive gain-adjusted removal response, ranging from 0.4635 to 2.1254 validation-logit standard deviations. The maximum held-artist R^2 on MUSDB18 is 0.7745, comparable to 0.7778 on the MoisesDB internal evaluation set. These measurements show that the target-source response persists after whole-gain adjustment and that the specified low-level descriptors do not fully predict the system scores across artists.

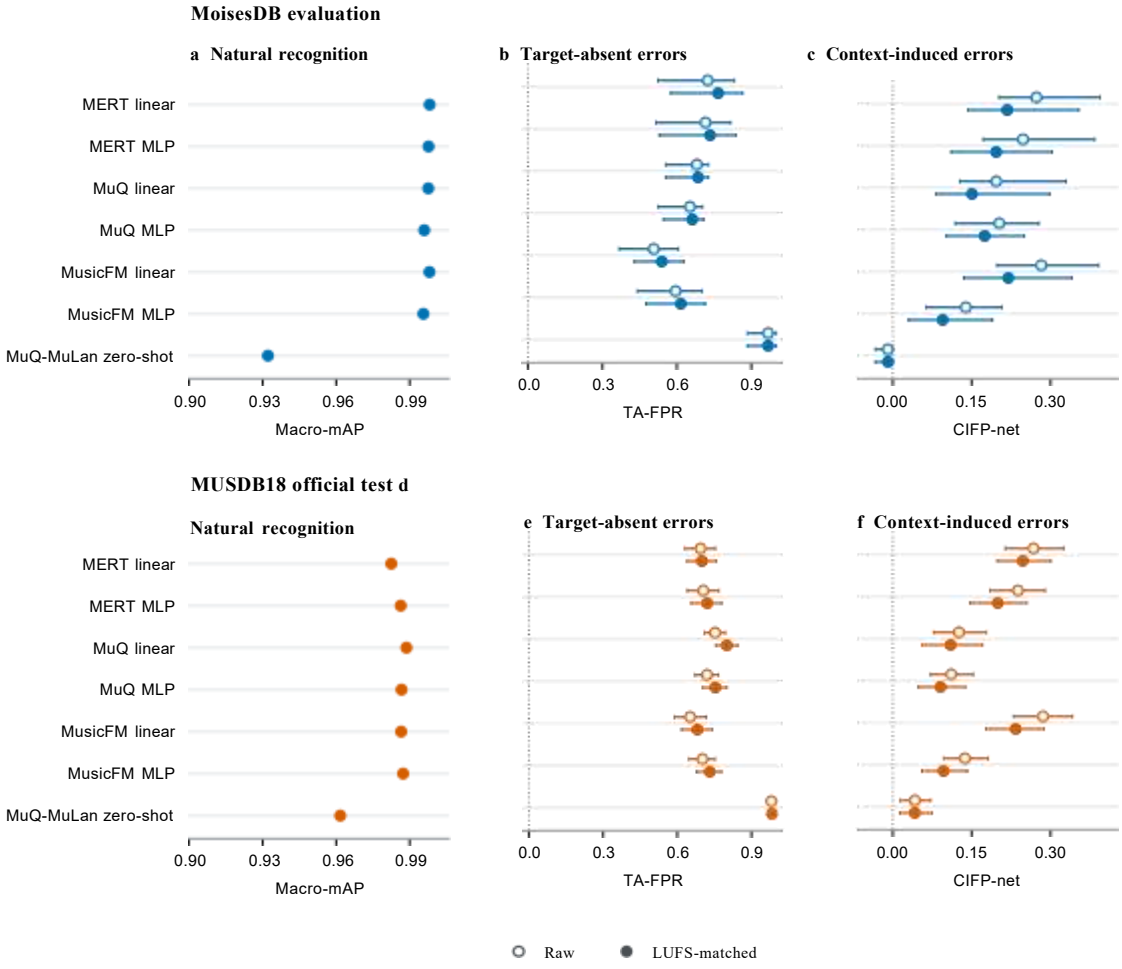


Fig. 3 High natural recognition coexists with target-absent and context-induced errors. (a–c) MoisesDB internal evaluation. (d–f) MUSDB18 official test. Rows retain all seven original encoder–readout systems; SRA is not applied in this figure. Natural macro-mAP uses the natural probe populations in Table 2B and is shown as a point estimate. TA-FPR and CIFP- net use the intervention populations in Table 2A, with open symbols for raw audio and filled symbols for LUFS-matched audio. Error bars are endpoint-wise 95% intervals from 2,000 artist-cluster bootstrap draws; they are not multiplicity- adjusted intervals. Point estimates are track-equal, and repeated system evaluations do not add independent observations. The two datasets contain 11 and 46 evaluation artist clusters, respectively. Axes are shared within each column; the macro- mAP axis starts at 0.90. Dotted lines mark zero on the two error endpoints. The figure displays coexistence of recognition and grounding behavior, not an estimated correlation between metrics.

7.3 SRA improves responses on the MoisesDB internal evaluation set

SRA improves all five primary cross-model response endpoints relative to the matched ordi- nary head on the MoisesDB internal evaluation set (Table 4). Standardized TRE increases by 0.6617, with a 95% artist-cluster interval of [0.4643, 0.9235]. DOC increases by 0.0304 [0.0070, 0.0435]. TA-FPR falls by 0.1682 [0.0648, 0.2218], CIFP-net falls by 0.2831 [0.2064, 0.3333], and standardized CPE0 falls by 0.7498 [0.5811, 0.8898]. All five intervals favor SRA.

The matched comparison identifies a gain from the complete SRA objective beyond ordi- nary training on the same labeled interventions. It spans both threshold-independent responses

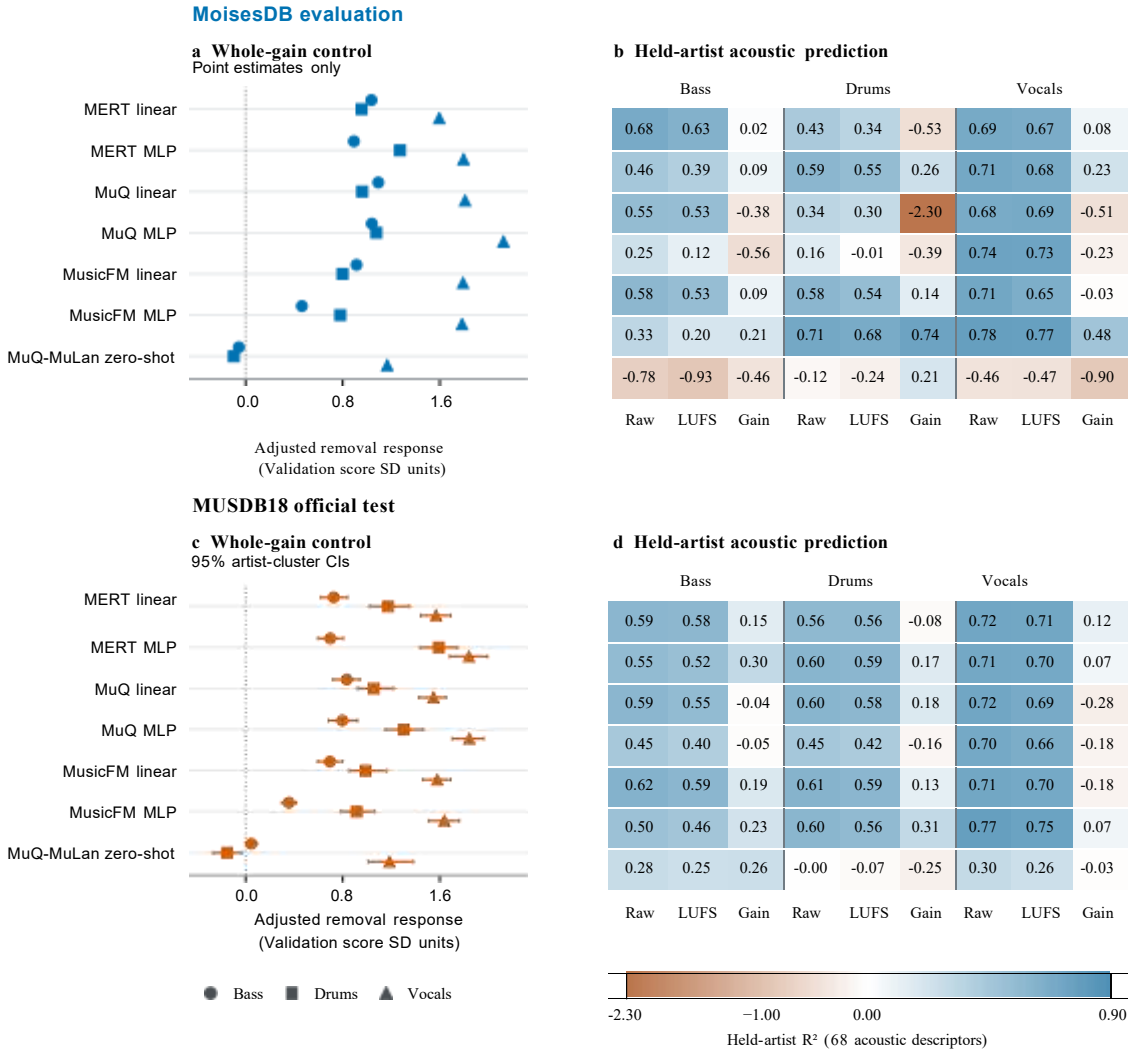


Fig. 4 Acoustic controls distinguish source-removal responses from whole-mixture gain and low-level score prediction. (a,c) Gain-adjusted target-removal response for bass, drums, and vocals in each original system. The plotted contrast is G_k in Eq. (8), comparing the whole-mixture gain control with the raw target-muted mixture from the same target window. Positive values indicate a remaining target-removal response after whole-gain adjustment. The fixed original validation-score standard deviation s_k supplies the common unit; supervised readouts use logits, and MuQ-MuLan uses its positive-minus- negative similarity score. The MoisesDB internal evaluation set provides 21 point estimates; MUSDB18 provides 21 estimates with 95% artist-cluster intervals. (b,d) Held-artist R^2 for predicting each original system's scores from four scalar acoustic descriptors and 64 pooled log-mel features. Outer leave-one-artist-out prediction uses inner artist-grouped ridge selection. Each heatmap retains all 63 system–class–arm cells, including negative R^2 values; Raw, LUFS, and Gain denote the raw, loudness-matched, and whole-mixture gain-control arms. The MoisesDB regressions contain 10 eligible artists for bass and 11 for drums and vocals; MUSDB18 regressions contain 46. The common color scale includes negative values and ends at the prespecified near-complete-prediction warning level of 0.90. Heatmaps are descriptive cross-fit diagnostics, not additional hypothesis tests.

(TRE, DOC, and CPE0) and target-absent decisions (TA-FPR and CIFP-net). Final threshold selection alone cannot account for improvements in the three threshold-independent endpoints.

Natural-task and off-target retention satisfy the cross-model criteria relative to the original system. The macro-mAP difference is +0.00028 [_0.00061, 0.00497], above the _0.01 noninferi- ority limit. The OTNER difference is _0.00365 [_0.01163, 0.00068], below the +0.005 upper limit. These retention comparisons use the original sys- tem as their anchor, separately from the ordinary- head response comparison.

7.4 Frozen MUSDB18 replication

The independent official test reproduces the audit pattern and the method effect. The original-system audit passes all 42 Holm-adjusted tests formed by seven model–readout identities, two loudness arms, and three class-aggregated endpoints. None of the 63 low-level cross-fit cells reaches $R^2 \geq 0.90$. Placebo matching succeeds for 305 of 431 target windows (70.77%), yielding 4,270 of 6,034 comparison rows across seven systems and two arms. The official test thus supports the same audit contrasts on a separate recording collection and its standard AAC source format.

The frozen SRA systems improve all five primary cross-model endpoints relative to ordinary training (Table 5). Standardized TRE increases by 0.6637 [0.5919, 0.7371], and DOC increases by 0.0429 [0.0345, 0.0523]. TA-FPR decreases by 0.1610 [0.1253, 0.1978], CIFP-net by 0.3002 [0.2659, 0.3322], and standardized CPE0 by 0.7561 [0.7050, 0.8073]. Each interval lies entirely in the favorable direction. The effect pattern closely matches that on the MoisesDB internal evaluation set despite the change in recording collection and eligible artists.

The independent test also meets the cross-model retention criteria. Natural macro-mAP changes by +0.00399 [0.00011, 0.00841] relative to original. OTNER changes by _0.00030 [_0.00511, 0.00323], satisfying the +0.005 upper limit. Relative to ordinary, the corresponding differences are +0.00078 [_0.00159, 0.00341] for macro-mAP and +0.00082 [_0.00104, 0.00274] for OTNER. The latter comparisons show that the substantial intervention-response gains occur alongside small overall differences in these retention metrics.

Figure 5displays the paired response effects for each encoder and the cross-model estimand across the two datasets.

7.5 Model-specific response and retention

The aggregate result contains a distinct model-level pattern (Table 6). On MUSDB18, MERT and MuQ improve all five response endpoints relative to ordinary training. MusicFM improves DOC, TA-FPR, CIFP-net, and standardized CPE0, but its standardized TRE difference is _0.2083 [_0.3186, _0.0996]. MusicFM's TRE is lower than ordinary's but remains higher than its origi- nal MLP's by approximately 1.6483 standardized units.

All three encoders meet the external natu- ral macro-mAP noninferiority criterion relative to original. For off-target retention, MERT's external OTNER difference is +0.00337 [_0.00079, 0.00839], whose upper interval limit exceeds +0.005. The MERT-specific criterion is therefore not established, while the cross-model criterion is met. In the MoisesDB internal evaluation set, MuQ has the analogous individual result: +0.00161 [_0.00812, 0.01456]. Reporting the model-level intervals identifies variation that the pooled estimate alone does not show.

Figure 6 separates natural-task retention from off-target retention and shows the prespecified bounds for every encoder.

7.6 Probability reliability difers from response alignment

On MUSDB18, descriptive track-equal macro Brier scores are 0.0953 for original, 0.0771 for ordinary, and 0.1626 for SRA. The corresponding ECE values are 0.1287, 0.0930, and 0.2173 (Table 6). Probability reliability is worse for SRA, even as its source-interventional response endpoints improve and natural ranking performance is retained.

Thus, retained ranking performance and improved source responses coexist with worse numerical probability scores. The distinction is consequential: average precision, ordered paired scores, threshold decisions, and probability reliability assess different properties of the output.

Table 4 Selective response alignment on the MoisesDB internal evaluation set

A. Cross-model response efects. Values are oriented improvements with 95% artist-cluster confidence intervals.

LUFS-matched endpoint	SRA minus ordinary, oriented	SRA minus original, oriented
Standardized TRE increase	+0.6617 [0.4643, 0.9235]	+1.9686 [1.4321, 2.2836]
DOC increase	+0.0304 [0.0070, 0.0435]	+0.0540 [0.0264, 0.0654]
TA-FPR reduction	+0.1682 [0.0648, 0.2218]	+0.5185 [0.2581, 0.6535]
CIFP-net reduction	+0.2831 [0.2064, 0.3333]	+0.2566 [0.1449, 0.3218]
Standardized CPE0 reduction	+0.7498 [0.5811, 0.8898]	+0.7110 [0.4719, 0.8541]

B. Cross-model retention relative to the original readout.

Endpoint	Signed diference, SRA __ original [95% CI]	Prespecified retention criterion	Result
Natural macro-mAP	+0.00028 [__0.00061, 0.00497]	Lower CI ≥ __0.010	Met
OTNER	__0.00365 [__0.01163, 0.00068]	Upper CI ≤ +0.005	Met

Notes. Ordinary and SRA use the same 263,171-parameter residual architecture, input rows, paired seeds, full-batch sequence, optimizer, learning rate, and 60 selected training updates. SRA adds the structured response-alignment terms. Corrected response efects use 11 artist credits and 24 intervention-eligible tracks; natural retention uses the same 11 credits and 78 tracks. For intervention-response and of-target contrasts, three seeds are averaged within the paired audio unit before track-equal aggregation. Natural macro-mAP is computed separately for each seed within each bootstrap draw before averaging, as specified in Section 6.3. Both analyses use 2,000 artist-cluster bootstrap replicates. The original readout provides the prespecified retention anchor, while the ordinary head provides the capacity-, data-, and selected-update-matched method control. Positive oriented response values favor SRA; negative OTNER diferences favor SRA. Table 2 describes the identity-based exclusion; no head, threshold, recipe, or checkpoint was changed for this reanalysis.

Table 5 Frozen replication on the MUSDB18 official test

A. Cross-model response efects. Values are oriented improvements with 95% artist-cluster confidence intervals.

LUFS-matched endpoint	SRA minus ordinary, oriented	SRA minus original, oriented
Standardized TRE increase	+0.6637 [0.5919, 0.7371]	+1.8064 [1.6796, 1.9384]
DOC increase	+0.0429 [0.0345, 0.0523]	+0.0563 [0.0444, 0.0700]
TA-FPR reduction	+0.1610 [0.1253, 0.1978]	+0.4717 [0.4199, 0.5261]
CIFP-net reduction	+0.3002 [0.2659, 0.3322]	+0.2576 [0.2193, 0.2958]
Standardized CPE0 reduction	+0.7561 [0.7050, 0.8073]	+0.7920 [0.7121, 0.8650]

B. Cross-model retention relative to the original readout.

Endpoint	Signed diference, SRA __ original [95% CI]	Prespecified retention criterion	Result
Natural macro-mAP	+0.00399 [0.00011, 0.00841]	Lower CI ≥ __0.010	Met
OTNER	__0.00030 [__0.00511, 0.00323]	Upper CI ≤ +0.005	Met

Notes. Heads, selected recipe, selected training update, thresholds, preprocessing, and validation-logit scales were fixed before official-test scoring; no MUSDB18 head or threshold refit was performed. Response and of-target analyses cover 46 credited artist IDs and 49 intervention-eligible tracks. Natural retention covers 46 IDs and all 50 official-test tracks. Statistical aggregation and orientation are identical to Table 4. The full test contains 431 target windows and 13,361 intervention instances. The separate 12-track training-partition pilot is not pooled with this test.

SRA improves the measured source-response pro- file; Table 6C characterizes its accompanying probability-score change.

8 Discussion

8.1 Recognition, sensitivity, and source selectivity are distinct

The audit resolves a question that natural accu- racy alone cannot answer. A readout can recognize

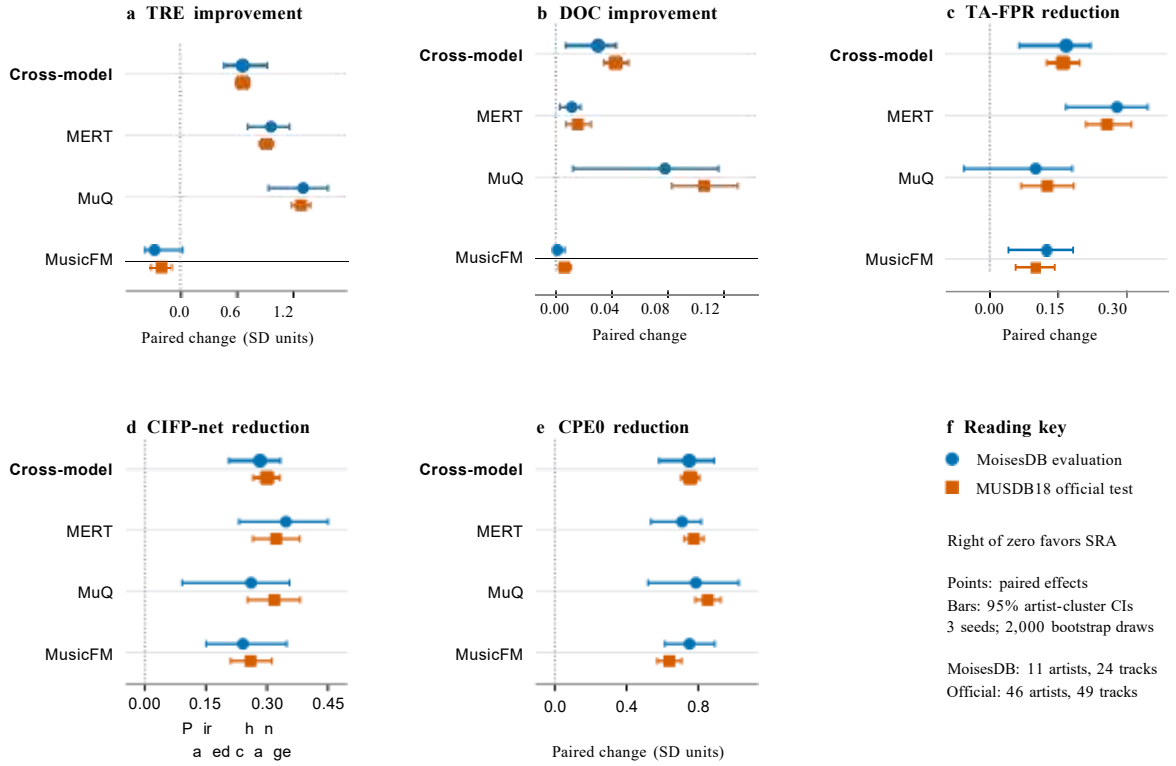


Fig. 5 Paired response efects of SRA relative to ordinary training. (a–e) Primary LUFS-matched endpoints for the three encoders and the cross-model aggregation. Positive values favor SRA; TRE and CPE0 use units of the original validation logit standard deviation, while the other endpoints are proportion diferences. Blue circles denote MoisesDB internal evaluation (11 artists, 24 tracks, 205 target windows); orange squares denote MUSDB18 official test (46 artists, 49 eligible tracks, 431 target windows). Points show track-equal paired estimates after within-unit seed aggregation; bars are endpoint-wise 95% percentile artist-cluster bootstrap intervals from 2,000 draws, without simultaneous-interval adjustment. All model results, including negative TRE efects, are shown. Panel f supplies the reading key.

the target, respond strongly to its removal, and still classify the target as present afterward. The context experiment adds a second distinction: accompaniment can introduce new target-positive decisions even when the target is absent in both compared signals. Direct target evidence and con- text dependence can therefore coexist. The finding is not that these models fail to hear their tar- get sources, but that hearing the target does not ensure selective attribution to it.

The intervention design makes this distinction experimentally accessible. Raw source removal retains the same background waveform; the facto- rial contrast varies the designated context under target absence. Loudness matching and whole- mixture gain controls address different conse- quences of level change. An unrelated-source con- trol asks whether a comparable level change has the same target response. This combination links interpretable waveform operations to specific deci- sion failures rather than reducing grounding to one perturbation score.

8.2 Relational supervision improves a frozen readout

The matched ordinary head is central to inter- preting SRA. Both methods use the same labeled natural and interventional inputs, so their differ- ence is not explained by access to source-removal examples. SRA adds a joint specification of ordi- nal target change, target-absent rejection, context response, and correct margins on preserved labels. The replicated gains support the complete objec- tive as the tested learning intervention.

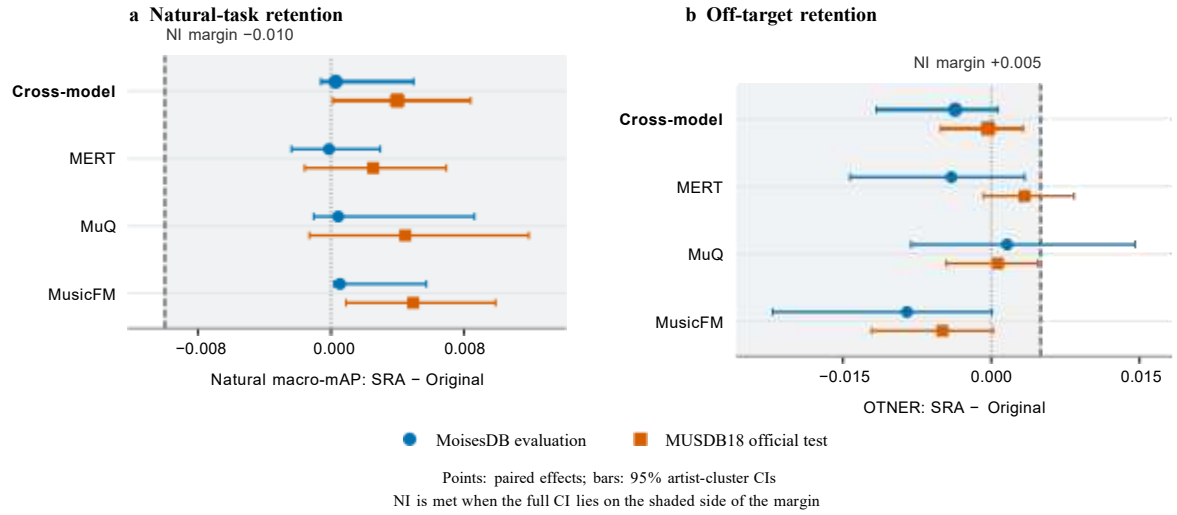


Fig. 6 Retention relative to the original observational readout. (a) Natural macro-mAP diference, SRA minus original; noninferiority requires a lower confidence limit of at least —0.010. (b) Of-target new error rate (OTNER) diference; noninferiority requires an upper confidence limit of at most +0.005. Blue circles represent the MoisesDB internal evaluation set, and orange squares represent MUSDB18 official test. Natural evaluation uses 78 and 50 tracks; of-target evaluation uses 24 and 49 eligible tracks, respectively, clustered within 11 and 46 artists. Points are track-equal estimates; bars are endpoint-wise 95% percentile artist-cluster bootstrap intervals from 2,000 draws. Natural AP is computed per seed before averaging; of-target contrasts average seeds within audio units. Each panel retains all model-specific intervals.

The fixed-encoder design identifies a useful place to act on the problem. The cached representations support a more source-aligned readout than either the original observational MLP or ordinary residual training. SRA learns that map- ping with 263,171 additional parameters and no encoder updates. Feature extraction is shared and cached during training, and inference adds only the residual head to the existing system.

8.3 A response profile is more useful than a single success score

MusicFM shows why increasing a removal contrast is not synonymous with improving every decision property. Its TRE gain is smaller than ordinary's, while its dose order and all three absence/context endpoints improve. Similarly, cross-model reten- tion does not imply that every individual encoder meets the same interval criterion. Table 6 exposes those differences without replacing the prespec- ified aggregate comparison with a best-model result.

The probability results distinguish response alignment from probability calibration [18, 19]. Standardized margins and ordered score differ- ences constrain behavior across inputs; they do not directly optimize agreement between sigmoid probabilities and empirical frequencies. SRA's higher Brier score and ECE are observed out- comes, not a consequence inferred solely from the loss design. For binary tagging, the reported thresholds and error rates are the relevant oper- ating behavior. Applications interpreting proba- bilities as confidence must also account for the measured reliability change.

8.4 Experimental scope

The evaluated source vocabulary is bass, drums, and vocals. These classes provide a consistent mapping between MoisesDB and MUSDB18 and define the population of interventions addressed by the study. The source operations preserve the recording realization and manipulate level or pres- ence; they do not sample arbitrary edits to timbre, performance, or arrangement.

The unrelated-source control depends on the availability of a suitable energy-matched source, with primary coverage of 77.07% on the MoisesDB internal evaluation set and 70.77% on the official

Table 6 Model-level response and retention profiles

A. MUSDB18 official-test response efects, SRA versus the ordinary head. Positive oriented values favor SRA; brackets show 95% artist-cluster confidence intervals.

LUFS-matched endpoint	MERT	MuQ	MusicFM
Standardized TRE increase	+0.9156 [0.8368, 0.9918]	+1.2838 [1.1853, 1.3821]	—0.2083 [—0.3186, —0.0996]
DOC increase	+0.0158 [0.0074, 0.0253]	+0.1063 [0.0833, 0.1305]	+0.0066 [0.0024, 0.0116]
TA-FPR reduction	+0.2581 [0.2118, 0.3117]	+0.1266 [0.0690, 0.1848]	+0.0983 [0.0559, 0.1415]
CIFP-net reduction	+0.3228 [0.2654, 0.3808]	+0.3182 [0.2522, 0.3818]	+0.2596 [0.2109, 0.3122]
Standardized CPE0 reduction	+0.7759 [0.7213, 0.8305]	+0.8537 [0.7838, 0.9269]	+0.6386 [0.5713, 0.7074]

B. Model-specific retention, SRA versus the original readout.

Dataset	Encoder	Natural macro-mAP diference [95% CI]	Natural criterion	OTNER diference [95% CI]	Of-target criterion
MoisesDB internal evaluation	MERT	—0.00012 [—0.00238, 0.00296]	Met	—0.00403 [—0.01430, 0.00339]	Met
MoisesDB internal evaluation	MuQ	+0.00043 [—0.00104, 0.00863]	Met	+0.00161 [—0.00812, 0.01456]	Not established
MoisesDB internal evaluation	MusicFM	+0.00054 [0.00019, 0.00573]	Met	—0.00852 [—0.02210, 0.00008]	Met
MUSDB18 official	MERT	+0.00254 [—0.00158, 0.00694]	Met	+0.00337 [—0.00079, 0.00839]	Not established
MUSDB18 official	MuQ	+0.00447 [—0.00130, 0.01191]	Met	+0.00067 [—0.00454, 0.00468]	Met
MUSDB18 official	MusicFM	+0.00495 [0.00090, 0.00991]	Met	—0.00495 [—0.01212, 0.00018]	Met

C. Natural probability scores on MUSDB18. Values are descriptive class- and model-macro averages; ordinary and SRA additionally average the three seeds.

Readout	Brier score	ECE, 10 equal-width bins
Original MLP	0.0953	0.1287
Ordinary residual head	0.0771	0.0930
SRA residual head	0.1626	0.2173

Notes. Natural noninferiority requires the lower confidence limit to be at least __0.010; of-target noninferiority requires the upper limit to be at most +0.005. "Not established" indicates that the interval crosses the prespecified retention bound, not that a positive error increase is statistically established. MusicFM has a smaller TRE increase than the ordinary head but improves the other four primary response endpoints. Relative to its original readout, its standardized TRE improvement is +1.6483 [1.5132, 1.7745]. Brier and ECE increase under SRA, identifying a probability-reliability trade-of alongside the response-alignment gains.

test. MUSDB18 supplies coarse source stems with- out directional bleed annotations. These dataset properties determine the interpretation of source absence and the coverage of the control analy- sis. The measured effects concern the specified recombinations of the released stems.

The evidence concerns the three evaluated frozen encoders and their fitted readouts, with track-equal estimates and artist-cluster uncer- tainty. The MoisesDB evaluation set is artist-disjoint from training and validation; the sep- arately entered MUSDB18 official test provides replication after model selection. This division keeps the statistical and experimental populations aligned with the algorithmic claim.

9 Conclusion

StemGround makes instrument grounding an explicit connection between physical source operations and label-response requirements. Its audit shows that high natural recognition performance and direct target sensitivity can coexist with target-absent and context-induced errors. SRA trains against those requirements using a lightweight residual readout over a fixed music encoder.

Relative to ordinary training on the same labeled inputs, SRA improves all five cross-model response endpoints on the MoisesDB internal evaluation and frozen MUSDB18 official test. Both datasets meet the cross-model natural- task and off-target retention criteria. The model- specific and probability results distinguish the supported response gains from properties that do not improve uniformly. The contribution is a reproducible evaluation-and-learning framework for source-selective instrument decisions, demon- strated across recorded datasets without modify- ing the foundation encoders.

Declarations

Data and code availability. MoisesDB is available through its official research-data access page at https://music.ai/research/ [24], subject to its research-use license. The MUSDB18 stan- dard AAC corpus is distributed through its official dataset record at https://doi.org/10.5281/zenodo .1117372 [6]. Foundation-model resources are iden- tified in [20–23]. Recordings, rendered audio, and foundation-model weights are obtained through their respective upstream access routes and are not redistributed in the StemGround package.

The accompanying StemGround reproducibil- ity package contains original code, executed configurations, split and intervention manifests, the MoisesDB internal evaluation set roster and exclusion record, saved inference scores, trained readouts, aggregate results, analysis scripts, and full-precision figure source data. It includes 18 selected ordinary/SRA residual heads and plot- ting code for Figures 1–6. Original project soft- ware is licensed under MIT; this license does not relicense non-code materials or upstream assets. Saved results support table and figure regenera- tion without audio or foundation-model inference, and the analysis code supports score-based statis- tical reanalysis. Audio-to-score reconstruction and head training additionally require the recorded upstream resources. Repository access: [Insert the verified repository record, actual release version, and permanent URL/DOI after author-authorized deposit; the current package is prepared locally. Non-code reuse terms require author confirma- tion].

Funding. [Funding sources, grant numbers, and the funders' roles to be supplied by the authors; state the absence of specific funding only if appli- cable.]

CRediT authorship contribution statement. [Author names and verified CRediT roles to be supplied.]

Declaration of competing inter- est. [Authors to provide the applicable competing-interest declaration.]

Declaration of generative AI and AI-assisted technologies in manuscript preparation. OpenAI Codex was used to assist with English drafting, language editing, manuscript organization, literature-source checking, plotting-code preparation, and checking reported val- ues against saved experimental outputs. Figures were rendered programmatically from the stated method or saved statistical results; generative image synthesis was not used. The scientific claims and empirical results are attributed to the study and its human authors. [The authors must review the complete manuscript and approve this decla- ration before submission.]

Appendix A Class-specific source and context efects

Table A1 complements the aggregate original-system audit in Table 3. It reports LUFS-matched target-removal effects on the MoisesDB internal evaluation set, new context-induced errors, corrections, and threshold-independent context effects for every audited model–readout–class combina- tion.

Table A1 Class-specific original-system efects on the MoisesDB internal evaluation set. Confidence intervals use the recorded credited-artist bootstrap.

Original system	Target	Standardized TRE [95% CI]	CIFP - new	CIFP-correction	CIFP - net	Standardized CPE0 [95% CI]
MERT, linear	Bass	1.0580 [0.8352, 1.1749]	0.4493	0.0000	0.4493	1.0002 [0.8653, 1.3113]
MERT, linear	Drums	0.9479 [0.8511, 1.1466]	0.1250	0.0278	0.0972	0.0616 [—0.1363, 0.3015]
MERT, linear	Vocals	1.5825 [1.5246, 1.6525]	0.1528	0.0278	0.1250	0.2057 [0.1167, 0.3814]
MERT, MLP	Bass	0.9139 [0.7445, 1.0072]	0.4493	0.0000	0.4493	1.1834 [1.0826, 1.5186]
MERT, MLP	Drums	1.2830 [1.1487, 1.4375]	0.0972	0.0694	0.0278	—0.0832 [—0.4300, 0.0535]
MERT, MLP	Vocals	1.8002 [1.5978, 1.9531]	0.1458	0.0139	0.1319	0.2147 [0.1387, 0.4009]
MuQ, linear	Bass	1.0950 [0.9480, 1.2157]	0.3188	0.0000	0.3188	0.8077 [0.5345, 1.3568]
MuQ, linear	Drums	0.9558 [0.8353, 1.2591]	0.0694	0.0139	0.0556	0.3191 [0.2138, 0.5293]
MuQ, linear	Vocals	1.7843 [1.5988, 2.1031]	0.1042	0.0139	0.0903	0.1808 [0.1051, 0.3819]
MuQ, MLP	Bass	1.0326 [0.8603, 1.1379]	0.4493	0.0000	0.4493	1.0609 [0.8207, 1.5556]
MuQ, MLP	Drums	1.0797 [0.9563, 1.4295]	0.0833	0.0208	0.0625	0.2208 [—0.0554, 0.4416]
MuQ, MLP	Vocals	2.1141 [1.9713, 2.4252]	0.0556	0.0139	0.0417	0.0632 [0.0079, 0.1729]
MusicFM, linear	Bass	0.9196 [0.7858, 0.9883]	0.4493	0.0000	0.4493	0.8526 [0.6214, 1.4819]
MusicFM, linear	Drums	0.7898 [0.6957, 1.0660]	0.1875	0.0000	0.1875	0.3338 [0.1396, 0.4876]
MusicFM, linear	Vocals	1.7544 [1.6382, 1.9110]	0.0694	0.0278	0.0417	0.2137 [0.1382, 0.3860]
MusicFM, MLP	Bass	0.4735 [0.4260, 0.5020]	0.1739	0.0000	0.1739	0.7374 [0.5979, 1.1198]
MusicFM, MLP	Drums	0.7710 [0.7076, 0.9111]	0.0972	0.0417	0.0556	0.2360 [—0.0915, 0.4125]
MusicFM, MLP	Vocals	1.7699 [1.5406, 1.9105]	0.0833	0.0208	0.0625	0.1985 [0.1397, 0.3581]
MuQ-MuLan, zero-shot	Bass	—0.0523 [—0.1255, 0.1324]	0.0000	0.0000	0.0000	—0.1839 [—0.3226, —0.0078]
MuQ-MuLan, zero-shot	Drums	—0.0884 [—0.2246, 0.0150]	0.0139	0.0000	0.0139	0.7421 [0.1808, 0.9377]
MuQ-MuLan, zero-shot	Vocals	1.1711 [0.6155, 1.3533]	0.0000	0.0417	—0.0417	—0.2098 [—0.4105, 0.0376]

Notes. In the MoisesDB internal evaluation set, bass uses 10 artist credits, 23 tracks, and 69 target windows; drums use 11 credits, 24 tracks, and 69 target windows; vocals use 11 credits, 24 tracks, and 67 target windows. CIFP-net equals CIFP-new minus CIFP- correction before rounding. The class-specific estimates in this appendix need not average numerically to the class-aggregated Table 3 estimates because target eligibility difers across tracks. Positive CPE0 indicates that adding the designated context increases the target score while the target is absent. This threshold-independent response is distinct from whether that change crosses the target threshold.

References

[1] Li, Y., Yuan, R., Zhang, G., Ma, Y., Chen, X., Yin, H., Xiao, C., Lin, C., Ragni, A., Benetos, E., Gyenge, N., Dannenberg, R., Liu, R., Chen, W., Xia, G., Shi, Y., Huang, W., Wang, Z., Guo, Y., Fu, J.: MERT: Acous- tic music understanding model with large- scale self-supervised training. In: The Twelfth International Conference on Learning Repre- sentations (2024). https://openreview.net/forum?id=w3YZ9MSlBu

[2] Zhu, H., Zhou, Y., Chen, H., Yu, J., Ma, Z., Gu, R., Luo, Y., Tan, W., Chen, X.: MuQ: Self-supervised music representation learning with mel residual vector quantization. IEEE Transactions on Audio, Speech and Language Processing **33**, 3653–3664 (2025) https://doi.org/10.1109/TASLPRO.2025.3602320

[3] Won, M., Hung, Y.-N., Le, D.: A foundation model for music informatics. In: ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1226–1230 (2024). https://doi.org/10.110

9/ICASSP48485.2024.10448314

[4] Yuan, R., Ma, Y., Li, Y., Zhang, G., Chen, X., Yin, H., Le, Z., Liu, Y., Huang, J., Tian, Z., Deng, B., Wang, N., Lin, C., Benetos, E., Ragni, A., Gyenge, N., Dannenberg, R., Chen, W., Xia, G., Xue, W., Liu, S., Wang, S., Liu, R., Guo, Y., Fu, J.: MARBLE: Music audio representation benchmark for universal evaluation. In: Advances in Neural Information Processing Systems, vol. 36, pp. 39626–39647. Curran Associates, Inc., Red Hook, NY, USA (2023). https://doi.org/10.52202/075280-1722 . https://proceedings.neurips.cc/paper_files/paper/2023/hash/7cbeec46f979618beafb4f46d8f39f36-Abstract-Datasets-and-Benchmarks.html

[5] Pereira, I., Araújo, F., Korzeniowski, F., Vogl, R.: MoisesDB: A dataset for source separation beyond 4-stems. In: Proceedings of the 24th International Society for Music Information Retrieval Conference, pp. 619–626 (2023). https://doi.org/10.5281/zenodo.10265363 . https://archives.ismir.net/ismir2023/paper/000073.pdf

[6] Rafii, Z., Liutkus, A., Sto…ter, F.-R., Mimilakis, S.I., Bittner, R.: The MUSDB18 Corpus for Music Separation. Zenodo. [dataset] Version 1.0.0; standard AAC STEMS release (2017). https://doi.org/10.5281/zenodo.1117372

[7] Zang, Y., O'Brien, S., Berg-Kirkpatrick, T., McAuley, J., Novack, Z.: Are you really listening? boosting perceptual awareness in Music-QA benchmarks. In: Proceedings of the 26th International Society for Music Information Retrieval Conference, pp. 247–261 (2025). https://doi.org/10.5281/zenodo.17706385

[8] Lee, Y., Shin, J., Kim, H., Shim, K.: Beyond Binary Instrument QA: Probing Instrument Grounding in Music Audio-Language Models. Learning to Listen: ICML 2026 Workshop on Machine Learning for Audio; author manuscript, version 1, 30 June 2026 (2026). https://doi.org/10.48550/arXiv.2606.31338 . https://arxiv.org/abs/2606.31338v1

[9] Haunschmid, V., Manilow, E., Widmer, G.: audioLIME: Listenable explanations using source separation. In: 13th International Workshop on Machine Learning and Music, ECML-PKDD 2020, pp. 20–23 (2020). https://f5f218df-5007-42f6-905b-7f72220c6756.filesusr.com/ugd/23ae7b_dc0fe6ba05a14250bc3a6f148058a09e.pdf

[10] Haunschmid, V., Manilow, E., Widmer, G.: Towards Musically Meaningful Explanations Using Source Separation. Version 1; author manuscript (2020). https://doi.org/10.48550/arXiv.2009.02051 . https://arxiv.org/abs/2009.02051v1

[11] Praher, V., Prinz, K., Flexer, A., Widmer, G.: On the veracity of local, model-agnostic explanations in audio classification: Targeted investigations with adversarial examples. In: Proceedings of the 22nd International Society for Music Information Retrieval Conference, pp. 531–538 (2021). https://archives.ismir.net/ismir2021/paper/000066.pdf

[12] Lin, Y.-X., Lin, J.-C., Wei, W.-L., Wang, J.-C.: Learnable counterfactual attention for music classification. IEEE Transactions on Audio, Speech and Language Processing **33**, 570–585 (2025) https://doi.org/10.1109/TASLPRO.2025.3527143

[13] Gebhardt, R.B., Kuhle, A., Bektur, E.: Beyond genre: Diagnosing bias in music embeddings using concept activation vectors. In: Proceedings of the 26th International Society for Music Information Retrieval Con- ference, pp. 279–286 (2025). https://doi.org/10.5281/zenodo.17706391

[14] Guinot, J., Quinton, E., Fazekas, G.: Leave-One-EquiVariant: Alleviating invariance-related information loss in contrastive music representations. In: ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1–5 (2025). https://doi.org/10.1109/ICASSP49660.2025.10890270

[15] Deng, V., Wang, C., Richard, G., McFee, B.: Investigating the sensitivity of pre-trained audio embeddings to common effects. In:

ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1–5 (2025). https://doi.org/10.1109/ICASSP49660.2025.10888929

[16] Ross, A.S., Hughes, M.C., Doshi-Velez, F.: Right for the right reasons: Training differentiable models by constraining their explanations. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, pp. 2662–2670 (2017). https://doi.org/10.24963/ijcai.2017/371 . https://www.ijcai.org/proceedings/2017/371

[17] Koo, J., Wichern, G., Germain, F.G., Khurana, S., Le Roux, J.: SMITIN: Self-monitored inference-time INtervention for generative music transformers. IEEE Open Journal of Signal Processing **6**, 266–275 (2025) https://doi.org/10.1109/OJSP.2025.3534686

[18] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 1321–1330 (2017). https://proceedings.mlr.press/v70/guo17a.html

[19] Cheng, J., Vasconcelos, N.: Towards calibrated multi-label deep neural networks. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recog- nition, pp. 27589–27599 (2024). https://doi.org/10.1109/CVPR52733.2024.02605 . https://openaccess.thecvf.com/content/CVPR2024/html/Cheng_Towards_Calibrate d_Multi-label_Deep_Neural_Networks_CVP R_2024_paper.html

[20] M-A-P: MERT-v1-330M. Hugging Face model repository. [software/model] Revision 5240c2708a5acaee1007f43fb9735c7dcd0b78c9. https://huggingface.co/m-a-p/MERT-v 1-330M/tree/5240c2708a5acaee1007f43fb9735c7dcd0b78c9 Accessed 2026-09-09

[21] OpenMuQ: MuQ-large-msd-iter. Hugging Face model repository. [software/model] Revision 0562a57814f6f8bbd9fdea0a25921a2fce1a841a. https://huggingface.co/OpenMuQ/MuQ-large-msd-iter/tree/0562a57814f6f8bbd9fdea0a25921a2fce1a841a Accessed 2026-09-09

[22] Won, M.: MusicFM, MSD checkpoint. Hugging Face model repository. [software/model] Revision 4513b38bc25ad1d227b1980819b9691ba97f4d87; MSD checkpoint, not FMA. https://huggingface.co/minzwon/MusicFM/tree/4513b38bc25ad1d227b1980819b9691ba97f4d87 Accessed 2026-09-09

[23] OpenMuQ: MuQ-MuLan-large. Hugging Face model repository. [software/model] Revision 2e01c796b71dca71b45251384c04cd7b237c9020. https://huggingface.co/OpenMuQ/MuQ-MuLan-large/tree/2e01c796b71dca71b45251384c04cd7b237c9020 Accessed 2026-09-09

[24] Music AI: MoisesDB. Official dataset access page. [dataset]. https://music.ai/research/ Accessed 2026-09-09