ACOUSTIC-TO-KV REGRESSION FOR BIOACOUSTIC RECOGNITION IN SPEECH LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Animal-sound recognition requires distinguishing call types, callers, and species from acoustic evidence. Speech language models can encode useful distinctions between animal sounds yet fail to select the correct labels. On marmoset calls, a classifier fitted to frozen Qwen2.5-Omni features outperforms audio in-context learning given the same labelled recordings. We introduce Acoustic-to-KV Regression (AKR) to help the model use its acoustic features when forming its own answer. The key idea is to learn from labelled recordings how to adjust the model's internal states to favor the correct answer. Negative gradients of the correct-answer loss provide correction targets for audio-token key/value (K/V) states. A low-rank ridge regressor then learns to predict these targets from acoustic features. For each new recording, AKR adds the predicted correction to its K/V states before the original decoder generates a label, without backbone updates or query-time backpropagation. Across five recording-group evaluations of 1-kHz marmoset recognition, using support clips that become misclassified after filtering, AKR increases mean accuracy from 12.35% to 32.47% and macro-F1 from 5.92% to 24.54%, outperforming native generation and a fixed correction in every evaluation. Held-out analyses further show that acoustic features predict these correction targets more accurately than fixed-mean and shuffled-pairing controls. These results demonstrate that predicting internal corrections from acoustic features can improve a frozen model's own recognition.

1 INTRODUCTION

Animal-sound recognition often requires distinguishing sounds within the same species. Marmoset call types such as Phee, Trill, and Twitter differ in duration and frequency–time structure (DiMattina & Wang, 2006; Agamaite et al., 2015). Other tasks concern the identity of the caller, such as identifying which dog produced a bark. These distinctions support the study of vocal repertoires and the identification of sound sources (Lamothe et al., 2025; Hagiwara et al., 2022). For a speech language model, such tasks pose a concrete challenge: connecting acoustic distinctions to the appropriate text labels.

We examine this connection through six-way call-type recognition on MarmAudio (Lamothe et al., 2025), using two labelled recordings per class. A ridge classifier fitted to frozen Qwen2.5-Omni features (Xu et al., 2025) from these twelve recordings reaches 58.67% accuracy on 75 recording-disjoint queries. Audio in-context learning (ICL) provides the same recordings and labels as demonstrations in the model's input (Kong et al., 2024), but reaches only 17.33% on the same queries, predicting a single class throughout (Figure 1b). The frozen features therefore support useful classification with few labelled examples, yet the model's own prediction process benefits much less from the same examples. This gap motivates learning how to make the acoustic information in these features more useful to the model's own label predictions.

A complementary diagnostic examines what happens when the acoustic condition changes. We low-pass filter MarmAudio recordings at 1 kHz while keeping their class labels unchanged. On the same filtered queries, a classifier fitted directly to 1-kHz features reaches 77.29% accuracy, whereas one fitted on Full-band input and transferred without refitting reaches 34.98% (Figure 1c). Useful class information therefore remains accessible after filtering, but a decision rule learned under the original condition can fail to transfer. We refer to this readout-transfer gap as *spectral decision drift*.

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

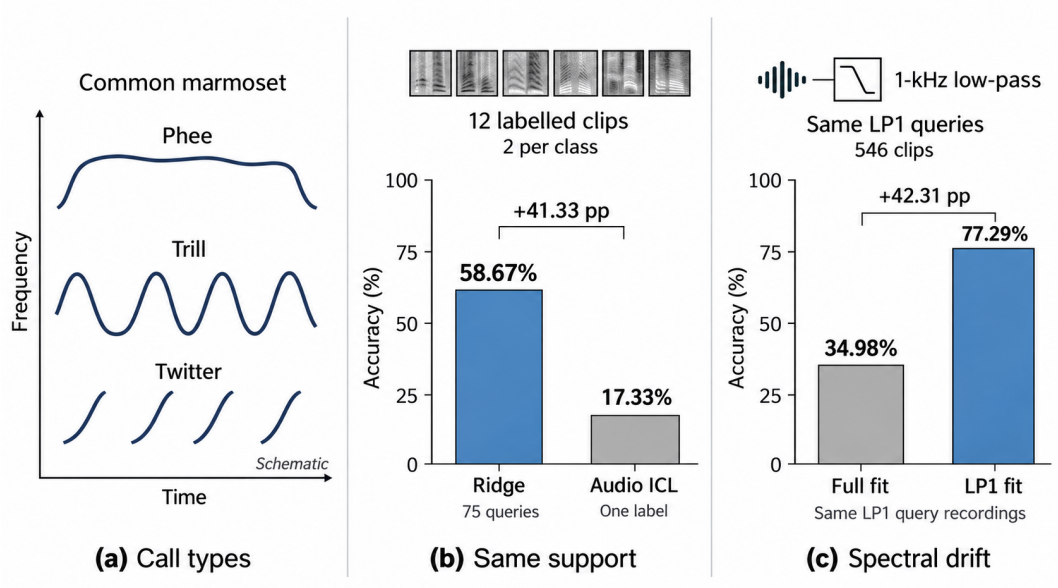


Figure 1: **Useful acoustic features do not always yield accurate label predictions.** (a) Schematic call structures within one species. (b) An explicit readout outperforms native audio ICL using the same labelled recordings. (c) Low-pass queries remain classifiable, but a readout fitted on Full-band input transfers poorly. Protocols are in Appendix B.

Similar gaps appear on BEANS Dogs caller identification and BEANS Watkins species recognition (Hagiwara et al., 2022).

Our goal is to use the available acoustic information to improve the original decoder's label predictions. A labelled support recording provides both a target answer and a way to estimate how the model's internal states should change to favor it. Specifically, negative gradients of the correct-answer loss with respect to audio-token key/value (K/V) states identify locally corrective directions. We use these directions as correction targets. Learning to predict them from acoustic features would make such adjustments available for new recordings whose labels are unknown.

We therefore introduce *Acoustic-to-KV Regression* (AKR) to learn this mapping. We first centre the support-derived correction targets and compress them into a low-rank basis. A ridge regressor then maps frozen acoustic features to coordinates in this basis. For a new recording, AKR extracts its acoustic features, predicts a recording-specific correction, and adds the reconstructed increment to its audio-token K/V states in a subsequent forward pass. The original decoder then generates the text label. All backbone weights remain frozen, and inference requires neither query labels nor query-time backpropagation.

We evaluate AKR through three complementary studies. In five recording-group evaluations of 1-kHz MarmAudio call-type recognition, using support clips that become misclassified after filtering, AKR increases mean accuracy from 12.35% to 32.47% and macro-F1 from 5.92% to 24.54%, outperforming both native generation and a fixed correction in every evaluation. On BEANS Dogs, matched-norm controls compare predicted, fixed, random, and token-structured corrections to examine the value of input conditioning. Separately, held-out support analyses show that acoustic regression predicts correction targets more accurately than fixed-mean and shuffled-pairing baselines.

Our contributions are as follows:

- **Acoustic-to-KV Regression.** We introduce a low-rank regression method that uses support-derived correct-answer gradients to learn recording-specific K/V corrections. The fitted mapping predicts internal adjustments for new recordings without backbone updates or query-time backpropagation.

- **Diagnostics of acoustic information use.** Matched-support comparisons reveal a gap between explicit classification and native audio ICL. Spectral comparisons further show that useful class information can remain accessible when a previously fitted readout transfers poorly.
- **Recognition gains and controlled analyses.** Five recording-group evaluations demonstrate improvements in native 1-kHz MarmAudio recognition. Complementary matched-norm controls and held-out gradient prediction analyses examine the value of input conditioning and the predictability of correction targets.

2 Related Work

Bioacoustic recognition and spectral robustness. BEANS benchmarks animal-sound classification and detection, while MarmAudio provides annotated vocalizations for studying within-species call types (Hagiwara et al., 2022; Lamothe et al., 2025). AVES learns transferable representations from unlabelled recordings, and NatureLM-audio connects bioacoustic recognition with language interaction (Hagiwara, 2022; Robinson et al., 2025). Beyond the Baseband incorporates higher-frequency information excluded by conventional low-rate inputs (Sarkar et al., 2026). Our study examines acoustic changes within the model-observable band and asks how the information available in a frozen generalist speech language model can better support its own label predictions.

Few-shot learning and acoustic information use. Audio Flamingo supports audio ICL and retrieval-based adaptation (Kong et al., 2024). MetaSICL and FSA-GRPO improve auditory models' use of demonstrations through post-training (Zheng et al., 2026b;a). Analyses of the speech–text modality gap further suggest that representation alignment alone is insufficient for stable downstream decisions (Hsu et al., 2026). Our matched-support comparisons examine how explicit readouts and native audio ICL use the same labelled recordings, while distinguishing information accessible to a supervised probe from information used in native predictions (Hewitt & Liang, 2019). AKR uses the labelled support to fit a reusable predictor of internal corrections, keeping the speech model frozen.

Internal-state steering and gradient prediction. Internal-state interventions provide a way to control frozen models. KV-cache steering modifies attention states (Belitsky et al., 2025), while audio-specialist-head steering uses audio–silence activation contrasts to strengthen the influence of acoustic evidence (Glazer et al., 2026). HyperSteer learns to generate steering vectors conditioned on steering prompts and model activations (Sun et al., 2025). Synthetic-gradient methods predict backward signals to decouple parameter updates during training (Jaderberg et al., 2017). AKR learns an acoustic-to-correction mapping using negative correct-answer gradients from labelled support recordings as targets. A low-rank ridge regressor predicts these corrections for new recordings and applies them to audio-token K/V states, allowing the original decoder to use the adjusted states while all backbone weights remain frozen.

3 Acoustic-to-KV Regression

AKR learns a mapping from acoustic features to corrections of a frozen model's audio-token K/V states. Labelled support recordings provide correction targets for fitting this mapping. For a new recording, the fitted map predicts an increment that is applied before the original decoder generates its label (Figure 2). Let θ denote the frozen model parameters, p the task prompt, and $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{n}$ the labelled support set. For each recording, we extract a pooled audio-token representation $h_i = h_\theta(x_i, p) \in \mathbb{R}^d$.

3.1 Extract corrective targets from support answers

For each support recording, its known answer lets us estimate a local change to the internal states that favors that answer. Let $Z^s_{i,\ell,t}$ denote the output of a selected K/V projection at layer ℓ, audio position t, and type $s \in \{K, V\}$. We intervene before head reshaping and, for keys, rotary position encoding. Let $\mathcal{A}_i$ denote the expanded audio-token positions. For the tokenized class name $y_i = (y_{i,1}, \ldots, y_{i,m_i})$,

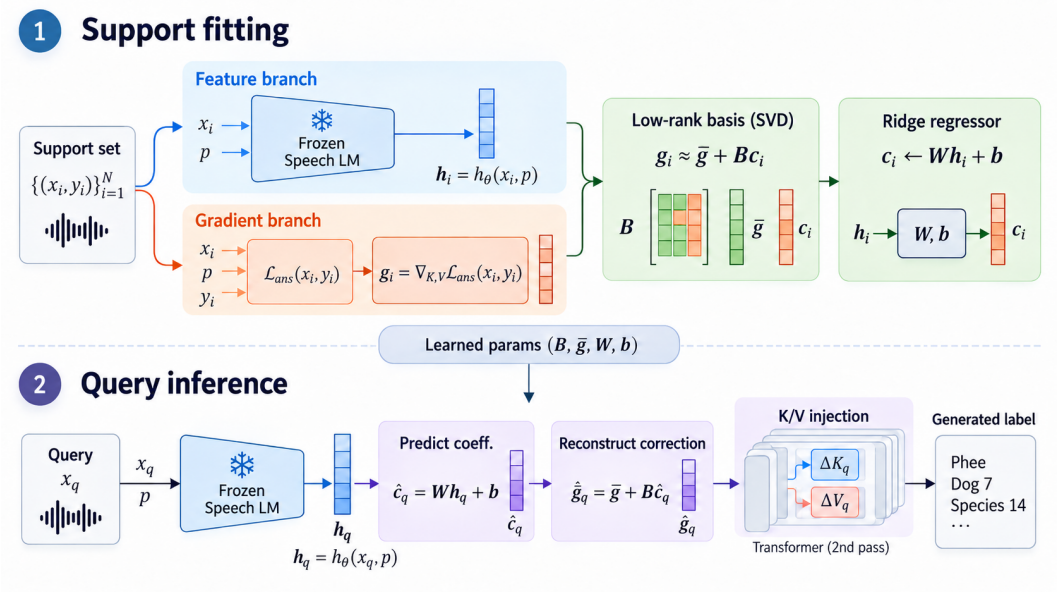


Figure 2: **Learn corrections from labelled support; predict them for new recordings.** During fitting, acoustic features provide regressor inputs, while negative answer-loss gradients provide K/V correction targets. For each query, a first forward pass supplies acoustic features; the predicted correction is applied during a second forward pass to produce the original decoder's answer. Backbone weights remain frozen throughout.

we compute

$$\mathcal{L}_{\text{ans}}(x_i, y_i; p) = -\frac{1}{m_i} \sum_{j=1}^{m_i} \log p_\theta(y_{i,j} \mid x_i, p, y_{i,<j}). \tag{1}$$

Only the class-name tokens contribute to this loss.

Negative gradients of this loss provide locally corrective directions at the selected projections. We average them over audio positions and concatenate the resulting K/V blocks:

$$g^s_{i,\ell} = -\frac{1}{|\mathcal{A}_i|} \sum_{t \in \mathcal{A}_i} \frac{\partial \mathcal{L}_{\text{ans}}}{\partial Z^s_{i,\ell,t}}, \qquad g_i = \text{concat}_{\ell,s}\, g^s_{i,\ell} \in \mathbb{R}^P. \tag{2}$$

Pooling gives a fixed-width correction target for recordings with different audio-token counts. The feature branch extracts h_i from the recording and prompt without the correct answer; the labelled branch supplies g_i as supervision. Computing these gradients leaves the backbone weights unchanged.

3.2 Fit low-rank acoustic regression

To make the correction targets easier to predict, we separate their shared mean from the variation across recordings. Stack the targets as rows of $G \in \mathbb{R}^{n \times P}$ and let $\bar{g} = n^{-1} \sum_i g_i$. A thin singular value decomposition of the centred targets gives

$$G - \mathbf{1}\bar{g}^\top = U\Sigma V^\top, \qquad B = V_{:,1:r}, \qquad a_i = B^\top(g_i - \bar{g}), \tag{3}$$

where $B \in \mathbb{R}^{P \times r}$ is orthonormal and $r \leq n - 1$. The coordinates a_i describe each recording's correction relative to the shared mean. This basis compresses the correction targets; it does not project the model's original K/V states.

We standardize acoustic features using means and scales computed only from the support set, obtaining z_i. An affine ridge regressor then predicts the correction coordinates from these features (Hoerl & Kennard, 1970):

$$(W^*, b^*) = \arg\min_{W,b} \left[\sum_{i=1}^{n} \|Wz_i + b - a_i\|_2^2 + \lambda \|W\|_F^2 \right]. \tag{4}$$

Here, $W \in \mathbb{R}^{r \times d}$ and $b \in \mathbb{R}^r$. Answer cross-entropy supplies the correction targets, while squared-error regression learns to predict their signed, real-valued coordinates. The fitted predictor consists of the feature normalizer, $\bar{g}$, B, W^*, and b^*.

3.3 Predict and apply a query-specific increment

The fitted map is reused for every query. A first forward pass extracts the query's acoustic feature from the recording and prompt. Applying the support-set normalizer gives z_q, from which we predict

$$\widehat{g}_q = \bar{g} + B(W^* z_q + b^*). \tag{5}$$

The mean term is shared across recordings, while the regression term adjusts the correction according to the query's acoustic features.

We split $\widehat{g}_q$ into its K/V blocks. During a second forward pass of the same recording and prompt, we add each block's predicted increment at audio positions, with step size η:

$$Z^{s\prime}_{q,\ell,t} = Z^s_{q,\ell,t} + \eta\, \mathbf{1}[t \in \mathcal{A}_q]\, \widehat{g}^s_{q,\ell}. \tag{6}$$

Within each selected projection, the increment is shared across audio positions but varies between recordings. The original decoder generates the label using the adjusted states. Query inference requires two forward passes, with no query labels or backward pass.

For the fixed-mean control, we remove the input-dependent term in Eq. 5 and use $\widehat{g}_q = \bar{g}$. This tests the value of predicting a recording-specific correction beyond applying one shared support-derived increment. Relative-norm controls instead rescale the broadcast increment in each active K/V block to have Frobenius norm equal to α times the norm of that block's original audio-token states. This is a separate scaling protocol from the raw step η in Eq. 6; zero increments remain zero. Appendix C provides the full algorithm, normalization details, and the local loss analysis. The analysis explains when a small aligned correction lowers answer loss; its effect on recognition is evaluated empirically.

4 Experiments

We first evaluate whether AKR improves the frozen model's own recognition, then examine the predictability of its correction targets and the value of input-dependent corrections. Each study uses a separate protocol suited to its evaluation goal.

4.1 Evaluation setup

Model and tasks. The main experiments use frozen Qwen2.5-Omni-7B (Xu et al., 2025). We study six-way marmoset call-type recognition on MarmAudio (MA-CT) (Lamothe et al., 2025) and ten-way dog caller identification on BEANS Dogs (BD-ID) (Hagiwara et al., 2022). MA-CT supports the recognition and correction-prediction studies; BD-ID provides the matched-norm intervention controls. Full denotes the model-observable input band, and LP1 denotes audio low-pass filtered at 1 kHz.

Recognition protocol. The main MA-CT/LP1 study constructs AKR and the fixed-mean control from 20 support clips, selected from recordings recognized correctly under Full but misclassified under LP1. For each of five draws, we first reserve ten query recording groups and then select support clips from disjoint recordings. All events in the held-out query groups are evaluated. Both correction methods use the same raw step size $\eta = 300$; their resulting correction norms are not matched.

Correction analyses. Correction predictability is evaluated on a separate, class-balanced MA-CT support set containing 48 clips from 32 recording groups. We use five-fold recording-group cross-validation, refitting the feature normalizer, correction basis, and regressor within each training fold. The BD-ID intervention controls use 20 support clips, two per dog, and 139 validation queries. Corrections are compared at matched active-block relative norms, with $\alpha \in \{0.003, 0.01, 0.03\}$.

Comparisons and metrics. Within each comparison, methods are evaluated on the same examples, and support-fitted methods use the same support recordings. Recognition is measured by accuracy and

Table 1: **Native recognition on MA-CT/LP1.** Mean ± sample SD across five recording-group draws, using support clips that become misclassified after filtering. Both correction methods use the raw step $\eta = 300$; their correction norms are not matched.

Method	Accuracy (%)	Macro-F1 (%)
Native generation	12.35 ± 4.16	5.92 ± 1.69
Fixed-mean correction	13.01 ± 10.20	6.74 ± 4.36
AKR	$\mathbf{32.47 \pm 7.50}$	$\mathbf{24.54 \pm 3.81}$

Table 2: **Recognition across recording-group draws.** Accuracy (%) on MA-CT/LP1. Methods share the query recordings within each draw. Different draws may overlap.

Draw	Queries	Native	Fixed mean	AKR
1	39	12.82	5.13	**30.77**
2	64	14.06	29.69	**42.19**
3	57	5.26	7.02	**28.07**
4	68	16.18	7.35	**23.53**
5	82	13.41	15.85	**37.80**

macro-F1 (Sokolova & Lapalme, 2009), with invalid answers counted as incorrect. Gradient NMSE measures correction-prediction error. For repeated recording-group draws, we report descriptive means and sample standard deviations; different draws may overlap. Dataset details, preprocessing, and protocol-specific settings are summarized in Appendix A.

4.2 Improving native recognition

Table 1 shows that AKR improves the original decoder's MA-CT/LP1 predictions. Mean accuracy increases from 12.35% to 32.47%, a gain of 20.12 percentage points, while macro-F1 rises from 5.92% to 24.54%. The fixed-mean correction reaches 13.01% accuracy, providing little average improvement over native generation. All evaluated labels are generated by the original decoder, with AKR supplying the internal correction.

The improvement holds in every evaluated draw: AKR exceeds both native generation and the fixed-mean correction in all five comparisons (Table 2). Its accuracy gain over native generation ranges from 7.35 to 28.12 percentage points.

Which predictions improve? We examine all 39 queries in the first indexed draw. Compared with the fixed-mean correction, AKR increases the number of correct answers from two to twelve. Paired predictions show twelve incorrect answers becoming correct and two correct answers becoming incorrect. The gains are concentrated in Phee and Twitter, while Infant Cry and Trill remain unresolved. Both methods produce four distinct labels and no invalid answers, so the improvement occurs without a change in output validity. The complete class breakdown and paired outcomes are provided in Appendix D.1.

Results under additional settings. Table 3 summarizes a separate MA-CT evaluation using 48 class-balanced support clips and 97 queries. Under this protocol, 7B/LP1 accuracy increases from 14.43% to 17.53%, while macro-F1 decreases from 8.38% to 7.47%. Accuracy remains unchanged on Full input and under both input conditions for the 3B model. Direct ridge readouts achieve higher accuracy in all four settings. Support selection, feature layers, and update scaling differ from the main recognition protocol, so these comparisons do not isolate their individual effects. Full accuracy and macro-F1 results are provided in Appendix E.

We next examine whether acoustic features predict correction targets on held-out recordings, testing the learned mapping underlying these interventions.

Table 3: **Recognition across model scales and input conditions.** Accuracy (%) using the same 48 class-balanced support clips and 97 query clips. Adaptation is fitted separately for each backbone and input condition. K/V corrections use relative strength $\alpha = .01$. Ridge is a direct classifier; the other methods use the original decoder.

Model / input	Native	Fixed mean	AKR	Ridge
7B / Full	26.80	25.77	26.80	81.44
7B / LP1	14.43	12.37	17.53	60.82
3B / Full	27.84	28.87	27.84	84.54
3B / LP1	16.49	15.46	16.49	69.07

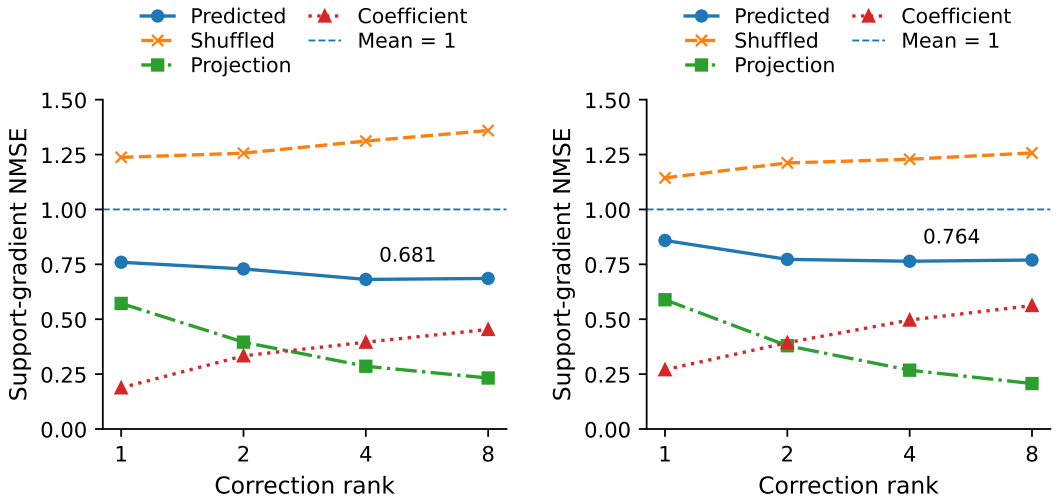


Figure 3: **Acoustic features predict held-out correction targets.** Prediction NMSE is below the fixed-mean reference of one and the shuffled-pairing control across all tested ranks. Projection and coefficient errors decompose total prediction error into target variation outside the retained basis and prediction error within it. All error terms use the same normalization. Numerical results and normalization details are provided in Appendix D.3.

4.3 Predicting corrections from acoustic features

AKR relies on acoustic features to predict a correction for a recording whose answer is unknown. We test this mapping on held-out recording groups, fitting the feature normalizer, correction basis, and regressor only on the training portion of each fold. Held-out answer gradients serve as evaluation targets. We normalize prediction error by the error of the training-fold mean correction, so a fixed-mean predictor has NMSE equal to one and lower values indicate more accurate prediction.

Prediction beyond the shared mean. Figure 3 shows that acoustic regression outperforms the fixed-mean predictor on both Full and LP1 inputs. At rank four, NMSE reaches 0.6813 on Full and 0.7641 on LP1, reducing error by 31.87% and 23.59%, respectively. To test whether this improvement depends on the correspondence between sounds and correction targets, we shuffle the training feature–gradient pairings while retaining the regression procedure. NMSE then rises to 1.3115 on Full and 1.2286 on LP1. Acoustic regression outperforms both controls at every tested rank, $r \in \{1, 2, 4, 8\}$, showing that the benefit extends beyond a single choice of correction dimension.

Balancing target compression and prediction. A larger correction basis retains more variation in the support targets, but also requires the regressor to predict more coefficients. The error decomposition in Figure 3 makes this trade-off visible. Projection error measures target variation that the retained basis cannot represent; coefficient error measures prediction error within that basis. On Full,

Table 4: **Correction controls at matched relative norms.** Accuracy (%) on the same 139 BD-ID validation queries. α specifies the correction norm relative to the original state norm within each active block. All three evaluated strengths are reported.

Correction	$\alpha = .003$	$\alpha = .01$	$\alpha = .03$
Fixed mean	2.88	2.88	5.04
AKR (pooled)	2.88	2.88	**12.95**
Ordered tokenwise	2.88	7.19	10.79
Token-permuted field	2.88	3.60	3.60
Matched random field	2.88	2.88	2.88

increasing rank from four to eight reduces projection error but increases coefficient error, leaving total NMSE slightly higher, from 0.6813 to 0.6858. LP1 follows the same pattern. Thus, retaining more correction directions does not automatically make the resulting targets easier to predict from limited support.

Directional alignment. The predicted corrections also show positive mean cosine similarity with the held-out targets. At rank four, cosine similarity is 0.4092 on Full and 0.3580 on LP1, indicating partial alignment with the measured corrective directions. Increasing rank to eight raises these values to 0.4475 and 0.4208, despite the slightly higher NMSE. The two metrics capture different aspects of prediction: cosine measures directional agreement, while NMSE reflects discrepancies in both direction and magnitude. Complete results are reported in Appendix D.3.

These results show that acoustic features predict variation in support-derived correction targets beyond a shared mean. We next test whether input-dependent corrections improve native recognition when intervention strength is controlled.

4.4 Input conditioning and correction structure

We test the value of input conditioning using the separate BD-ID control protocol. The pooled AKR correction is compared with a shared mean and a random field, with each active block rescaled to the same relative strength α. Ordered tokenwise and token-permuted corrections additionally test whether the arrangement of increments across audio positions affects recognition.

Input-dependent corrections at controlled strength. At $\alpha = .03$, pooled AKR reaches 12.95% accuracy, compared with 5.04% for the fixed mean and 2.88% for the matched random field (Table 4). The gain over the fixed mean is 7.91 percentage points. At this strength, selecting a correction from the recording's acoustic features is more effective than applying a shared or random update at the same per-block relative norm. At the two smaller strengths, pooled, fixed-mean, and random corrections all yield 2.88% accuracy. The benefit therefore depends on intervention strength within this comparison.

Does tokenwise structure help? Pooled correction shares one vector across the audio positions of each K/V block, whereas tokenwise correction allows these positions to receive different increments. The ordered variant outperforms its token-permuted counterpart at $\alpha = .01$ and $\alpha = .03$, reaching 7.19% and 10.79% accuracy, respectively, compared with 3.60% for both permuted conditions. This comparison indicates that preserving the arrangement of tokenwise corrections can matter. However, tokenwise control does not consistently outperform pooling: it performs better at $\alpha = .01$ but worse at $\alpha = .03$. Finer positional control therefore offers no uniform advantage over the pooled correction in this study.

Additional analyses of regression feature depth and correction rank are provided in Appendix D.2. Class-routed alternatives, which select a class-specific correction, are reported separately in Appendix F.

5 Discussion and Limitation

AKR provides a way to translate labelled support into input-dependent state corrections while retaining the model's original decoder. Support answers define local correction targets, and acoustic regression predicts corresponding updates for new recordings. The recognition results demonstrate that this procedure can improve the model's own label predictions. Direct ridge readouts remain more accurate for fixed-label classification in the separate extension (Appendix E). The contribution of AKR is therefore a demonstrated route for learning internal adaptation of the original model.

The main recognition gain is obtained with support selected to target errors introduced by low-pass filtering. Benefits are smaller under the class-balanced extension, and transferring correction maps between Full and LP1 does not outperform shuffled controls in accuracy. These findings make the conditions for effective correction an important part of the problem. The introductory spectral diagnosis concerns the transfer of external readouts: it motivates studying how acoustic information is used, but does not identify a unique cause of native prediction errors.

The results motivate examining when accurate correction prediction translates into better discrete decisions. Support selection, feature layers, and update scaling change together across the current protocols; isolating their effects would help explain when the learned map is most useful. Evaluation across additional model families would further test the generality of this approach. MarmAudio also lacks clip-level caller identities, so the present recording-disjoint evaluation should be complemented by studies that explicitly hold out individual callers.

6 Conclusion

We introduced AKR, a method for predicting internal K/V corrections from acoustic features using correct-answer gradients as supervision. A low-rank regressor fitted to labelled support recordings supplies input-dependent updates for a frozen model's original decoder. Across five recording-group evaluations of 1-kHz marmoset recognition, using support selected for filtering-induced errors, AKR raises mean accuracy from 12.35% to 32.47% and improves macro-F1. Held-out gradient prediction and matched-norm controls provide complementary evidence for correction predictability and the value of input conditioning. These results demonstrate that support-derived internal corrections can be predicted from acoustic features and used to improve the model's own recognition.

AI USE STATEMENT

Generative AI tools assisted manuscript drafting and revision,Quantitative results, software checks, and scientific experiments are treated separately.

ETHICS STATEMENT

This work reuses existing animal-sound datasets and conducts no new animal experiments. Recordings are subject to their dataset-specific licensing and redistribution terms. Recognition concerns supplied call-type, caller, and species labels; it does not establish translations, intentions, or validated communicative responses. Ecological or welfare deployment would require population- and environment-specific validation.

REPRODUCIBILITY STATEMENT

Appendix A maps datasets, sample selection, and settings to each research question. Appendices B–F give diagnostic definitions, the algorithm, complete supporting results, extensions, and variant identities. The source package contains the manuscript, figure assets, bibliography, and scalar data snapshots used for reported plots. Raw audio, model weights, and large intermediate tensors are not bundled. Author-facing source and editorial records are separate from the anonymous manuscript.

REFERENCES

James A. Agamaite, Chia-Jung Chang, Michael S. Osmanski, and Xiaoqin Wang. A quantitative acoustic analysis of the vocal repertoire of the common marmoset (Callithrix jacchus). *The Journal of the Acoustical Society of America*, 138(5):2906–2928, 2015. doi: 10.1121/1.4934268.

Max Belitsky, Dawid J. Kopiczko, Michael Dorkenwald, M. Jehanzeb Mirza, James R. Glass, Cees G. M. Snoek, and Yuki M. Asano. KV cache steering for controlling frozen LLMs. arXiv:2507.08799, 2025. URL https://arxiv.org/abs/2507.08799.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Christopher DiMattina and Xiaoqin Wang. Virtual vocalization stimuli for investigating neural representations of species-specific vocalizations. *Journal of Neurophysiology*, 95(2):1244–1262, 2006. doi: 10.1152/jn.00818.2005.

Neta Glazer, Lenny Aharon, and Ethan Fetaya. Are audio-language models listening? audio-specialist heads for adaptive audio steering. arXiv:2603.06854, 2026. URL https://arxiv.org/abs/2603.06854.

Masato Hagiwara. AVES: Animal vocalization encoder based on self-supervision. arXiv:2210.14493, 2022. URL https://arxiv.org/abs/2210.14493.

Masato Hagiwara, Benjamin Hoffman, Jen-Yu Liu, Maddie Cusimano, Felix Effenberger, and Katie Zacarian. BEANS: The benchmark of animal sounds. arXiv:2210.12300, 2022. URL https://arxiv.org/abs/2210.12300.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition, 2009. doi: 10.1007/978-0-387-84858-7. URL https://hastie.su.domains/ElemStatLearn/.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP-IJCNLP*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.

Ming-Hao Hsu, Xueyao Zhang, Xiaohai Tian, Jun Zhang, and Zhizheng Wu. Anatomy of the modality gap: Dissecting the internal states of end-to-end speech LLMs. arXiv:2603.01502, 2026. URL https://arxiv.org/abs/2603.01502.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Max Jaderberg, Wojciech Marian Czarnecki, Simon Osindero, Oriol Vinyals, Alex Graves, David Silver, and Koray Kavukcuoglu. Decoupled neural interfaces using synthetic gradients. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1627–1635, 2017. URL https://proceedings.mlr.press/v70/jaderberg17a.html.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. Audio Flamingo: A novel audio language model with few-shot learning and dialogue abilities. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 25125–25148. PMLR, 2024. URL https://proceedings.mlr.press/v235/kong24a.html.

Charly Lamothe, Manon Obliger-Debouche, Paul Best, Régis Trapeau, Sabrina Ravel, Thierry Artières, Ricard Marxer, and Pascal Belin. A large annotated dataset of vocalizations by common marmosets. *Scientific Data*, 12:782, 2025. doi: 10.1038/s41597-025-04951-8.

David Robinson, Marius Miron, Masato Hagiwara, Benno Weck, Sara Keen, Milad Alizadeh, Gagan Narula, Matthieu Geist, and Olivier Pietquin. NatureLM-audio: an audio-language foundation model for bioacoustics. arXiv:2411.07186, version 2, 2025. URL https://arxiv.org/abs/2411.07186.

Eklavya Sarkar, Marius Miron, David Robinson, Gagan Narula, Milad Alizadeh, Ellen Gilsenan-McMahon, Emmanuel Chemla, Olivier Pietquin, and Matthieu Geist. Beyond the baseband: Adaptive multi-band encoding for full-spectrum bioacoustics classification. arXiv:2604.27936, 2026. URL https://arxiv.org/abs/2604.27936.

Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009. doi: 10.1016/j.ipm.2009.03.002.

Jiuding Sun, Sidharth Baskaran, Zhengxuan Wu, Michael Sklar, Christopher Potts, and Atticus Geiger. HyperSteer: Activation steering at scale with hypernetworks, 2025. URL https://arxiv.org/abs/2506.03292.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report. arXiv:2503.20215, 2025. URL https://arxiv.org/abs/2503.20215.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706, 2021. URL https://proceedings.mlr.press/v139/zhao21c.html.

Haolong Zheng, Siyin Wang, Xulin Fan, Zengrui Jin, and Mark Hasegawa-Johnson. FSA-GRPO: Teaching auditory LLMs to use few-shot demonstrations. *arXiv preprint arXiv:2606.02615*, 2026a. doi: 10.48550/arXiv.2606.02615. URL https://arxiv.org/abs/2606.02615.

Haolong Zheng, Siyin Wang, Zengrui Jin, and Mark Hasegawa-Johnson. MetaSICL: Globalizing auditory LLMs for underserved speakers and languages via meta speech in-context learning. *arXiv preprint arXiv:2601.18904*, 2026b. doi: 10.48550/arXiv.2601.18904. URL https://arxiv.org/abs/2601.18904.

A Data and Evaluation Protocols

This appendix separates dataset identity from experimental role. MA-CT, BD-ID, and BW-SP abbreviate existing recognition tasks, not new benchmarks. The same dataset appears in several protocols because explicit-readout diagnosis, deployed correction, and gradient prediction answer different questions. Sample counts and baseline scores are not interchangeable across those protocols.

A.1 Datasets, classes, and labels

Table 5: **Evaluated datasets and recognition targets.** Counts describe the subsets used here, not each source collection's full size. Training/validation/test counts follow the released BEANS splits.

Dataset	Label denotes	Evaluated data	Role
MarmAudio (MA-CT) (Lamothe et al., 2025)	Six call types in one species	546 expert-agreed clips	Motivation, RQ1, RQ3, and fixed-setting extensions
BEANS Dogs (BD-ID) (Hagiwara et al., 2022)	Ten individual dogs	415/139/139 clips	Matched-support diagnosis and RQ2 controls
BEANS Watkins (BW-SP) (Hagiwara et al., 2022)	31 marine-mammal species	1,017/339/339 clips	Spectral readout diagnosis and full-training references

The MA-CT subset contains recordings affirmed by all four annotators in the technical-validation release. Its six labels are Infant Cry, Phee, Seep, Trill, Tsik, and Twitter. The Phee, Trill, and Twitter contours in Figure 1a are schematic representations of spectrotemporal structure (DiMattina & Wang, 2006; Agamaite et al., 2015), not measured spectrograms or statements about communicative meaning. MarmAudio does not assign a caller identity to each clip. Recording-disjoint evaluation therefore does not establish performance on unseen callers.

A support example is one recording and its task label. In MA-CT, two examples per class mean two clips of each call type: twelve clips from one species, not two animals. In BD-ID, an individual is the class, so two examples are two clips from the same named dog; the full two-shot support contains twenty clips. In BW-SP, two examples per class would mean two clips per species, or 62 clips, without requiring distinct individuals within a species. This last definition describes a support convention, not a claim that the completed study includes a 62-support Watkins AKR experiment.

Semantic labels and consistently remapped symbolic labels are evaluated separately. A–F mappings on MA-CT and A–J mappings on BD-ID associate the same acoustic classes with alternative output symbols. These controls test the demonstrated association rather than familiarity with category names. They do not change the underlying recognition labels for scoring.

A.2 One protocol per scientific comparison

RQ1 selection. The main recognition protocol selects twenty total support clips, one per source recording, from the available pool answered correctly at Full but incorrectly at LP1. Query recording groups are drawn first from all 93 groups, before checking failure eligibility. Ten query groups are drawn and all their events are evaluated. Support and query recordings are disjoint within each draw; the five draws may overlap with one another. This error-enriched support selection is not the class-balanced k-per-class protocol used for ICL or the later fixed-setting study.

RQ2 identity. The conditioning table uses the recorded continuous-field Dogs validation comparison, with three active-block relative strengths. Ordered tokenwise and token-permuted fields remain part of that comparison. A separate class-routed A–J sweep is not substituted into these rows. The historical feature-depth/rank grid also has its own query set and is reported separately in Appendix D.2.

Table 6: **Protocol-to-question map.** Full-data probe fitting, small-support ICL, error-enriched correction, and class-balanced extensions have different sample budgets. Within a row, the indicated comparators share evaluation identities.

Purpose	Support or training data	Evaluation data	Comparators / endpoint
Matched support	MA-CT: $k = 1, 2, 4, 8, 16$ per class; nested support	Same 75 recording-disjoint queries	Ridge/centroid; measured ICL at $k = 1, 2$; accuracy
Spectral diagnosis	Condition-specific training features; MA-CT grouped out-of-fold; BEANS released splits	MA-CT 546 out-of-fold predictions; BD-ID 139 and BW-SP 339 test clips	Native, condition-fitted, and transferred readouts
RQ1 recognition	MA-CT: 20 total error-enriched clips, one per support recording	Five draws of 39, 64, 57, 68, and 82 LP1 queries	Native, fixed mean, continuous AKR; accuracy/macro-F1
RQ2 controls	BD-ID: 2 clips per dog, 20 total	Same 139 validation queries at each strength	Fixed, pooled, ordered, permuted, random; accuracy
RQ3 prediction	MA-CT: 48 class-balanced clips from 32 recording groups	Five held-out folds within support	Predicted, mean, and shuffled targets; NMSE
Transfer/scale	Same 48 class-balanced MA-CT support clips; 7B/3B fitted separately	Same 97 recording-separated query clips for Full/LP1	Native, fixed, AKR, ridge; transfer also includes shuffled fields

RQ3 and extension identity. The completed fixed-setting study records seed 20260914 and eight support clips per call type. Its prediction analysis holds out support recording groups; its recognition extensions use 97 separate queries. The held-out fold sizes are 9, 9, 13, 11, and 6. A new normalizer, basis, and ridge map are fitted within each training fold. No recognition-query gradient is used in this analysis. The same support and query identities are reused across Full/LP1 and 7B/3B in the extensions, but each model has its own features and correction targets. These data had been inspected previously; the study is a fixed-setting extension, not a new untouched test.

Table 7: **Recorded fitting and generation settings.** Feature indices are the stored backend indices, not names for the intervention layers. Unreported historical layer lists are not inferred from the separate development grid or later extension.

Protocol	Setting
RQ1	Feature index 0; rank 4; ridge penalty 10; raw $\eta = 300$; draw seeds 20250813–20250817
RQ2	Active-block relative strengths .003, .01, .03; same recorded support and query set across fields
RQ3	7B feature index 24; K/V layers 7, 14, 21, 24; diagnostic ranks 1, 2, 4, 8; ridge penalty 10; five recording-group folds
7B extension	Feature index 24; K/V layers 7, 14, 21, 24; rank 4; ridge penalty 10; $\alpha = .01$
3B extension	Feature index 32; K/V layers 9, 18, 27, 31; rank 4; ridge penalty 10; $\alpha = .01$
Recognition	Frozen Thinker, Talker disabled; BF16; batch size 1; greedy text generation
Extension hardware	One RTX 4090; 24-GB placement profile; direct Thinker with SDPA attention

The extension uses a fixed numerical setting, not historical validation-selected parameters. Relative depth gives different layer indices at different scales; no 7B tensors are copied into 3B. The execution record lists Torch 2.13.0+cu130 and Transformers 4.52.3. Large intermediate arrays, raw audio, and model weights are not included in the manuscript package. The source snapshot does not enumerate every historical intervention layer or every per-draw sample identity in the manuscript itself; the corresponding saved run artifacts remain the authority for those details.

A.3 Audio processing and scoring

Filtering is applied at each recording's source sampling rate before anti-alias resampling to 16 kHz. Full therefore denotes the model-observable 0–8 kHz baseband, not the original 0–48 kHz range

of a 96-kHz MarmAudio recording. LP1/2/4/6/8 denote low-pass cutoffs of 1/2/4/6/8 kHz. LP8 is a filtered control, distinct from unfiltered Full. Per-condition peak normalization is disabled; RMS-matched controls are separate. Low-pass filtering changes the available spectrum without relabelling the vocalization or shifting its entire pitch trajectory.

Accuracy counts invalid or multiple-label outputs as incorrect. Macro-F1 complements overall accuracy with class-balanced performance (Sokolova & Lapalme, 2009). When paired predictions are available, a repair is a previously incorrect answer becoming correct, and a damage is a previously correct answer becoming incorrect. With R repairs and H damages over N queries, the accuracy change is $(R - H)/N$. Repeated recording draws use descriptive means and sample SD, not confidence intervals that treat overlapping observations as independent. Single-seed extensions have no between-seed error bars, and shared transfer/scale operating points are counted once.

B Exploratory Diagnostics

These measurements support the observations in Figure 1; they precede the three method-evaluation questions. Matched-support ICL measures supervision use, while spectral probes measure readout transfer. Neither diagnostic is an objective optimized by AKR.

B.1 Matched-support recognition

For a C-class task, let $\mathcal{S}_k$ contain k labelled recordings per class and $\mathcal{Q}$ contain held-out recordings from the same label set. An explicit readout $\widehat{f}_{\mathcal{S}_k}$ fits support feature–label pairs. Audio ICL instead receives the corresponding audio–answer demonstrations in its context. With a fixed prompt, the gap is

$$G(k) = \mathrm{Acc}_{\mathcal{Q}}(\widehat{f}_{\mathcal{S}_k} \circ h_\theta) - \mathrm{Acc}_{\mathcal{Q}}\big(f_\theta^{\mathrm{ICL}}(\cdot; \mathcal{S}_k, p)\big). \tag{7}$$

Both routes keep the speech model frozen, but only the explicit readout fits classifier parameters. The comparison matches recordings and labels, not optimization or computational cost. The concrete two-shot MA-CT episode supplies two clips of each of its six call types and evaluates both routes on the same 75 recording-disjoint queries.

Table 8: **Matched-support MA-CT recognition.** Accuracy (%) on the same 75 queries. Larger support sets follow the nested order.

k/class	Support clips	Audio ICL	Centroid	Ridge
1	6	8.00	29.33	46.67
2	12	17.33	46.67	58.67
4	24	18.94	50.67	73.33
8	48	27.23	78.67	84.00
16	96	26.13	76.00	82.67

The one- and two-shot gaps are 38.67 and 41.33 percentage points, with paired 95% intervals [25.33, 52.00] and [28.00, 54.67]. The ridge curve plateaus near eight examples per class in this evaluation; it is not a fitted scaling law. BD-ID gives a separate matched-support example: one clip per dog yields 35.25% for ridge and 7.19% for audio ICL on 139 validation queries.

Candidate scoring uses label-string log-likelihoods rather than free decoding, following likelihood-based few-shot evaluation (Brown et al., 2020). Audio demonstrations follow the audio-ICL setting studied by Kong et al. (2024); this is an evaluation on Qwen, not an Audio Flamingo backbone experiment. Sequence-sum and mean-token candidate scores are distinct because label strings have different lengths. Nearest-centroid classification compares query features with support class centroids, while ridge learns a linear feature-to-class map (Hoerl & Kennard, 1970; Hastie et al., 2009).

Table 9: **Zero-support candidate scoring on all 546 MA-CT clips.** Full-input accuracy and macro-F1 (%). These rows do not use the 75-query matched-support protocol.

Readout	Accuracy	Macro-F1
Free generation	29.49	21.94
Bare candidates, token mean	30.95	20.26
Bare candidates, sequence sum	33.88	25.36
Definitions, token mean	22.34	10.46
Definitions, sequence sum	18.86	6.58

Prompt and label preferences can affect few-shot predictions (Zhao et al., 2021). In the recorded A–F controls, six counterbalanced mappings give 17.33% one-shot candidate accuracy, within the shuffled-association range 15.33–17.78%. Formatting or candidate scoring alone therefore does not establish effective acoustic-to-symbol learning in that protocol. These rows remain separate from semantic-label ICL and from the new gradient-prediction study.

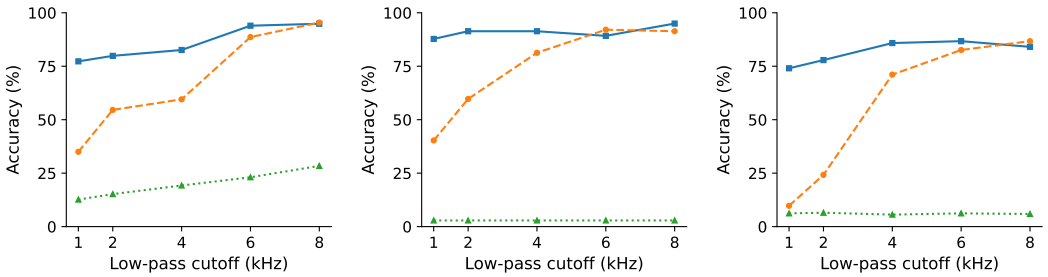
B.2 SPECTRAL READOUT TRANSFER

Let F_c be a low-pass filter and $h_c(x) = h_\theta(F_c x, p)$. Fit f_c on condition-c training features and f_F on Full training features, then evaluate both on identical filtered queries:

$$D(c) = \mathrm{Acc}_{\mathcal{Q}}(f_c \circ h_c) - \mathrm{Acc}_{\mathcal{Q}}(f_F \circ h_c). \tag{8}$$

This operational transfer deficit is the quantity called spectral decision drift. Negative values are retained when the transferred readout performs better. It is not an information-theoretic decomposition or a direct measurement of a boundary inside the native decoder.

MA-CT uses nested recording-grouped out-of-fold evaluation over its 546 expert-agreed clips; each recording is excluded from fitting the readout used to predict it. BD-ID and BW-SP use their declared fixed splits. The two-shot MA-CT example does not supply the training budget for these spectral probes.



(a) MA-CT **(b) BD-ID** **(c) BW-SP**

Figure 4: **Retained discrimination and readout transfer.** Condition-fitted and Full-trained readouts score identical filtered queries; native generation is shown separately. The LP8 filtered control is not unfiltered Full.

Table 10: **Complete spectral readout values.** Each triplet is native / condition-fitted / Full-trained transferred accuracy (%).

Cutoff	MA-CT	BD-ID	BW-SP
1 kHz	12.64 / 77.29 / 34.98	2.88 / 87.77 / 40.29	6.19 / 74.04 / 9.73
2 kHz	15.20 / 79.85 / 54.58	2.88 / 91.37 / 59.71	6.49 / 77.88 / 24.19
4 kHz	19.23 / 82.60 / 59.52	2.88 / 91.37 / 81.29	5.60 / 85.84 / 71.09
6 kHz	23.08 / 93.96 / 88.64	2.88 / 89.21 / 92.09	6.19 / 86.73 / 82.60
8 kHz	28.39 / 94.87 / 95.42	2.88 / 94.96 / 91.37	5.90 / 84.07 / 86.73

Bidirectional and shared-setting controls. Transfer also fails when a low-pass-trained readout receives Full features. Table 11 reports each source readout's matched accuracy and transfer accuracy. Shared-setting controls fix feature layer and ridge regularization across sources while still fitting readout coefficients on each source. Their deficits show that different layer/regularization choices are not sufficient to explain the pattern.

Table 11: **Bidirectional readout transfer.** Accuracy (%) on the declared BD-ID/BW-SP test splits. The source-matched reference and target-matched matrix reference below are different comparisons.

Task	Source → target	Source-matched	Transferred	Shared-setting transfer
BD-ID	Full → LP1	92.81	40.29	40.29
BD-ID	LP1 → Full	87.77	38.85	30.22
BW-SP	Full → LP1	88.20	9.73	9.73
BW-SP	LP1 → Full	74.04	20.35	23.01

For the complete matrices, write $M_{s,t}$ for accuracy of a readout fitted at source bandwidth s and evaluated at target t. The target-referenced deficit is

$$D_{s,t} = M_{t,t} - M_{s,t}. \tag{9}$$

Each column uses the same target recordings. Its diagonal is zero by definition. For BD-ID, Full→LP1 gives $87.77 - 40.29 = 47.48$ points; for BW-SP it gives $74.04 - 9.73 = 64.31$ points. These are readout-transfer measurements, not AKR repair scores.



(a) BD-ID **(b) BW-SP**

Figure 5: **Target-referenced spectral deficits.** Each cell is Eq. 9 in percentage points. Positive values mean that transferring a source readout underperforms fitting at the target condition.

Additional signal checks. On a separate balanced 300-event MA-CT subset, cyclic audio substitution scores 16.67% against the original labels and 30.33% against the substituted recording's labels; the latter equals the canonical unmodified-audio accuracy. Silence scores 15.00% and favours a small label subset. RMS matching raises LP1 native accuracy from 12.33% to 18.33% on that subset. Attenuation contributes, but these controls do not make all spectral effects an amplitude effect. Across five cutoffs, MA-CT's transfer deficit and correction-target rotation have Spearman $\rho = 1.00$ and Pearson $r = .969$. Both vary with bandwidth, so this is an association rather than causal mediation. The more extensive score-transform plots in the source materials are redundant views of these measurements, not additional experiments.

C METHOD AND IMPLEMENTATION DETAILS

Algorithm 1 AKR: support fitting and label-free query inference

Require: Frozen model θ, labelled support $\mathcal{S}$, rank r, ridge penalty λ, raw step η

for each $(x_i, y_i) \in \mathcal{S}$ **do**
  Extract h_i from (x_i, p) without the answer label.
  Compute answer loss and pooled negative K/V target g_i using Eqs. 1–2.
end for
Fit support feature normalizers and compute $\bar{g}, B, a_i$ using Eq. 3.
Fit W, b with Eq. 4; retain all fitted objects unchanged for evaluation.
for each unlabelled query x_q **do**
  Extract and standardize h_q in a first forward pass.
  Predict $\widehat{g}_q$ with Eq. 5 and unflatten it into K/V blocks.
  Run the same recording/prompt again with audio-position increments from Eq. 6.
  Generate the original decoder's answer; use the label only for subsequent scoring.
end for

C.1 TARGETS, COORDINATES, AND PARAMETER SHAPES

The answer-token cross-entropy defines an error signal for the frozen speech model. Its negative state gradient defines a real-valued correction target. Fitting W, b then minimizes squared coordinate error, not class cross-entropy or a policy-distillation loss. For n support recordings, feature width d, gradient width P, and retained rank r, the shapes are $G \in \mathbb{R}^{n\times P}$, $B \in \mathbb{R}^{P\times r}$, $W \in \mathbb{R}^{r\times d}$, and $b \in \mathbb{R}^r$. The standardizer uses $z_{ij} = (h_{ij} - \mu_j)/\sigma_j$ with support-only statistics, replacing very small scales by one. Ridge penalizes W, not its intercept. One support recording has no centred target variation, so its conditional correction reduces to the fixed mean.

The selected hooks operate on projection outputs before head reshaping. Key increments precede rotary position encoding; values do not undergo that key rotation. Audio pooling excludes text and answer positions from the target and from direct additive injection. Subsequent attention can still propagate the effect to the answer. A shared basis compresses correction targets; the original states remain present in the additive update.

C.2 PROJECTION AND COEFFICIENT-PREDICTION ERROR

For analysis, let $a_q = B^\top(g_q - \bar{g})$ denote coordinates of a labelled held-out gradient. With $B^\top B = I$, orthogonality gives

$$\begin{aligned}\|g_q - \widehat{g}_q\|_2^2 = \ & \left\|(I - BB^\top)(g_q - \bar{g})\right\|_2^2 \\ & + \|a_q - (Wz_q + b)\|_2^2 .\end{aligned} \tag{10}$$

The first term measures target variation omitted by the basis; the second measures prediction error in retained coordinates. A small projection residual does not establish predictability. RQ3 evaluates both terms on held-out support groups, where labels may be used for analysis; labelled recognition-query gradients are not required at deployment. The decomposition is exact for the defined Euclidean coordinates and does not imply an accuracy guarantee.

C.3 LOCAL LOSS INTERPRETATION

Let T_i broadcast a pooled block vector across a recording's audio positions. Eq. 2 gives $T_i^\top \nabla_Z \mathcal{L}_i = -|\mathcal{A}_i| g_i$. For a small raw step,

$$\mathcal{L}_i(Z + \eta T_i \widehat{g}_i) - \mathcal{L}_i(Z) = -\eta |\mathcal{A}_i| \langle g_i, \widehat{g}_i \rangle + O(\eta^2). \tag{11}$$

Positive alignment favours locally lower loss. Finite updates, sequential attention effects, numerical precision, and discrete decoding can change the observed answer. This explains the support target; it does not justify estimating a labelled recognition-query gradient during inference or guarantee that every predicted update is helpful.

C.4 RELATIVE-NORM CONTROLS AND VARIANT IDENTITIES

For an active layer/projection/scope block, relative scaling is

$$\Delta Z_{\mathrm{rel}} = \begin{cases} \alpha \dfrac{\|Z_{\mathrm{active}}\|_F}{\|\Delta Z\|_F} \Delta Z, & \|\Delta Z\|_F > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

The ratio is recorded per active block. Equal per-block ratios do not imply equal total energy when the number of layers or scope changes. These controls do not retroactively norm-match RQ1's raw-step comparison. Zero direction or zero strength is a strict no-op.

Continuous AKR predicts real-valued coefficients without first choosing a class. A class-routed dictionary instead predicts a class and retrieves its support-gradient centroid. Pooled targets share a block vector across audio positions; ordered tokenwise targets retain position-resolved fields. These choices define different variants whose results are not interchanged. Appendix F contains the separate class-routed and Oracle results.

Table 12: **Class-level changes in the first recording-group draw.** Correct predictions on all 39 queries. Recall changes are measured in percentage points.

Call type	Queries	Fixed correct	AKR correct	Recall change
Infant Cry	9	0	0	0.00
Phee	5	0	5	+100.00
Seep	7	0	1	+14.29
Trill	4	0	0	0.00
Tsik	4	2	1	-25.00
Twitter	10	0	5	+50.00
Total	39	2	12	+25.64

Table 13: **Regression feature depth and correction capacity.** Macro-F1 (%) on the separate BD-ID development set. Each cell uses the same support and development queries. Feature indices specify regression inputs.

Rank	Feature 0	Feature 8	Feature 16	Feature 28
1	1.18	1.25	0.71	0.71
2	1.33	1.30	0.73	0.74
4	4.19	1.25	3.95	3.58
8	7.80	10.22	8.36	10.32
16	7.94	9.84	9.90	10.09

D Additional experimental results

D.1 Recording-group results and class-level changes

The five rows of Table 2 correspond, in order, to seeds 20250813–20250817 and query counts 39, 64, 57, 68, and 82.

The mean paired accuracy differences are 20.12 ± 8.02 points against native generation and 19.46 ± 5.15 against fixed mean (sample SD across draws). Each difference is positive in every draw. This comparison reports the recorded raw-step operating point, not a sweep over newly selected parameters.

Class-level changes in the first draw. The first indexed draw, seed 20250813, contains 39 queries. Table 12 compares AKR with the fixed-mean correction across all six call types. The paired outcomes are zero both-correct, twelve fixed-wrong/AKR-correct, two fixed-correct/AKR-wrong, and twenty-five both-wrong. These counts sum to 39.

Phee and Twitter account for ten of the twelve correct AKR predictions, while Infant Cry and Trill remain unresolved. Both methods produce four distinct labels and no invalid answers. Their prediction counts for Infant Cry, Phee, Seep, Trill, Tsik, and Twitter are respectively 0/11/1/0/24/3 for fixed mean and 0/10/5/0/13/11 for AKR. Macro-F1 increases from 2.38% to 23.79%.

D.2 Conditioning controls and feature/rank analysis

Table 4 reports the continuous-correction comparison using the same twenty support clips and 139 validation queries at all three relative strengths. At $\alpha = .03$, pooled AKR exceeds the fixed-mean correction by 7.91 percentage points. Ordered tokenwise correction outperforms its permutation at $\alpha = .01$ and $\alpha = .03$, but does not consistently outperform pooling.

A separate development grid uses twenty support clips and 64 development queries to vary the regression feature index and retained correction rank (Table 13). Feature indices specify the regression inputs, not the K/V intervention sites. Development selection prioritizes macro-F1, followed by accuracy, lower rank, and earlier feature index. The selected setting is feature 28 with rank eight. This grid is separate from the main recognition, matched-norm control, and held-out gradient-prediction protocols.

D.3 Correction prediction, normalization, and decomposition

Let $\mathcal{H}_f$ contain the held-out support clips in fold f, $\bar{g}_{-f}$ be the training-gradient mean, and $\widehat{g}_{-f}(h_i)$ the predictor fitted on that training portion. The normalization is

$$\mathrm{NMSE} = \frac{\sum_f \sum_{i \in \mathcal{H}_f} \|g_i - \widehat{g}_{-f}(h_i)\|_2^2}{\sum_f \sum_{i \in \mathcal{H}_f} \|g_i - \bar{g}_{-f}\|_2^2}. \tag{13}$$

Numerators and denominators are summed across folds before division; the result is not an unweighted average of separately normalized fold errors. The same denominator is used for projection and coefficient errors. Their sum equals total prediction error up to rounding. No group crosses a training/held-out boundary, and normalization, SVD, and ridge are all refitted on each fold's training clips.

Table 14 reports all results. Cosine averages the 48 nonzero held-out support gradients; projection and coefficient terms follow Eq. 10.

Table 14: **Complete correction-prediction results.** All NMSE columns use the fixed-mean reference of one. Projection and coefficient errors sum to prediction error up to rounding. Shuffled breaks training feature–gradient pairings. Cosine measures alignment with held-out targets.

Input	Rank	Projection	Coefficient	Prediction	Shuffled	Cosine
Full	1	0.5719	0.1875	0.7595	1.2373	0.1986
	2	0.3957	0.3334	0.7291	1.2563	0.2931
	4	0.2858	0.3955	0.6813	1.3115	0.4092
	8	0.2323	0.4535	0.6858	1.3595	0.4475
LP1	1	0.5887	0.2700	0.8588	1.1439	0.1756
	2	0.3794	0.3932	0.7726	1.2118	0.2967
	4	0.2679	0.4962	0.7641	1.2286	0.3580
	8	0.2071	0.5625	0.7696	1.2574	0.4208

A true-class-centroid control reaches NMSE .3572 on Full and .2950 on LP1, with no missing-class fallback. It uses each held-out support label to select the centroid and is therefore a label-using diagnostic, not a deployable query method. It describes target label structure without implying that the regressor receives query labels. Prediction NMSE and final recognition need not agree: one measures squared error in update coordinates, the other measures a discrete decoder decision after a finite perturbation. Appendix E evaluates that downstream behaviour separately.

E TRANSFER, MODEL SCALE, AND DIRECT READOUTS

The class-balanced fixed-setting study tests how the correction map behaves beyond the main error-enriched operating point. Full and LP1 use the same 48 support clips and 97 query clips. Rank four, ridge penalty ten, and $\alpha = .01$ are fixed; settings and model-specific indices are listed in Appendix A. The numerical settings do not copy tensors between model sizes. The data had been inspected previously, and the single-seed comparisons do not establish generalization across model families.

E.1 CORRECTION-MAP TRANSFER

Fit one continuous predictor on Full support and another on LP1 support, then apply each unchanged to Full and LP1 queries. Normalization, mean, basis, and coefficients remain fixed within a source-trained map. This tests the correction map, unlike Appendix B.2's transfer of external classifiers.

Table 15: **Complete 7B correction-map transfer.** Accuracy (%) on the same 97 queries. Column labels are support condition → query condition. The diagonal cells are shared with the 7B scale comparison, not independent repetitions.

Method	Full→Full	Full→LP1	LP1→Full	LP1→LP1
Native	26.80	14.43	26.80	14.43
Fixed mean	25.77	12.37	24.74	12.37
Shuffled field	26.80	15.46	24.74	14.43
Continuous AKR	26.80	15.46	24.74	17.53
Frozen ridge	81.44	17.53	38.14	60.82

Table 16: **Transfer metrics and paired decision changes.** Accuracy/macro-F1 are percentages; R/H count repaired/damaged answers against native on the same query condition. All invalid rates are zero. Native Full is 26.80/16.53; native LP1 is 14.43/8.38.

Support → query	Method	Accuracy	Macro-F1	R	H
Full→Full	Fixed mean	25.77	16.06	0	1
	Shuffled field	26.80	16.26	1	1
	Continuous AKR	26.80	15.86	2	2
	Frozen ridge	81.44	78.21	56	3
Full→LP1	Fixed mean	12.37	4.17	0	2
	Shuffled field	15.46	8.36	1	0
	Continuous AKR	15.46	6.89	2	1
	Frozen ridge	17.53	12.28	12	9
LP1→Full	Fixed mean	24.74	14.82	0	2
	Shuffled field	24.74	14.38	1	3
	Continuous AKR	24.74	13.93	2	4
	Frozen ridge	38.14	33.80	25	14
LP1→LP1	Fixed mean	12.37	6.07	0	2
	Shuffled field	14.43	4.86	2	2
	Continuous AKR	17.53	7.47	4	1
	Frozen ridge	60.82	57.12	48	3

On LP1 queries, the matched AKR map yields 17 correct answers, the Full-trained map 15, and native generation 14. On Full queries, the LP1-trained map gives 24 against 26 native. Both off-diagonal AKR accuracies equal their shuffled-field counterparts; equal accuracy does not mean identical predictions, as the paired counts show. Matched LP1 has a local advantage over shuffled fields, but this does not establish cross-spectral invariance.

E.2 WITHIN-FAMILY SCALE COMPARISON

Only 7B/LP1 increases accuracy, by three net correct answers, while macro-F1 decreases. Both 3B conditions preserve the number of correct predictions; their small macro-F1 changes reflect a changed

Table 17: **Backbone-grouped comparison under identical supervision.** Accuracy and macro-F1 (%) on MA-CT, with the same support/query identities and fixed numerical settings. Each model fits its own map. Ridge is a direct classifier, not a KV correction.

		MA-CT / Full		MA-CT / LP1	
Backbone	Method	Acc.	F1	Acc.	F1
Qwen2.5-Omni 7B	Native	26.80	16.53	14.43	8.38
	Frozen ridge	**81.44**	**78.21**	**60.82**	**57.12**
	Fixed-mean KV	25.77	16.06	12.37	6.07
	Continuous AKR	26.80	15.86	17.53	7.47
Qwen2.5-Omni 3B	Native	27.84	18.60	16.49	6.06
	Frozen ridge	**84.54**	**82.37**	**69.07**	**65.36**
	Fixed-mean KV	28.87	19.36	15.46	5.32
	Continuous AKR	27.84	18.68	16.49	6.35

Table 18: **Native-to-AKR changes in the scale study.** Each row's outcome counts sum to 97; all invalid rates are zero. Deltas are percentage points.

Model	Input	Both right	Repair	Damage	Both wrong	ΔAcc.	ΔF1
7B	Full	24	2	2	69	+0.00	-0.66
7B	LP1	13	4	1	79	+3.09	-0.91
3B	Full	27	0	0	70	+0.00	+0.08
3B	LP1	16	0	0	81	+0.00	+0.29

error distribution. Frozen ridge is substantially stronger in this fixed-label task. These results do not demonstrate uniform state-repair gains or cross-family generalization. They also do not repeat RQ1's support selection, feature layer, or raw-step normalization. Their smaller effects cannot be attributed to any one of those changes from the available comparison.

F Additional Baselines and Variants

These references clarify the available alternatives and intervention boundaries without being substituted for the continuous-AKR RQs. Full-training baselines have a different supervision budget; class-routed and Oracle methods have different prediction or label-access rules.

F.1 Full-training recognition references

AVES-bio provides frozen animal-vocalization features (Hagiwara, 2022). Our classifiers fit its features on the declared training sets. The LoRA reference updates rank-eight adapters on query/value projections for one epoch (Hu et al., 2022); it is a budget-specific adaptation, not an exhaustive LoRA search. Qwen, AVES, and the MarmAudio specialist also differ in accessible input spectrum and task supervision.

Table 19: **Full-input references.** Accuracy (%); supervised probe and LoRA rows are not matched-support comparisons against native generation. The MarmAudio specialist uses the original high-rate spectrum. A dash denotes an unreported cell.

Reference	MA-CT	BD-ID	BW-SP
Qwen2.5-Omni-7B, native	29.49	2.88	6.19
Qwen-7B frozen probe	95.42	92.81	88.20
AVES-bio frozen probe (Hagiwara, 2022)	93.41	83.45	85.84
Qwen-7B full-train LoRA, one epoch (Hu et al., 2022)	96.56	25.18	31.56

Native model size also interacts with bandwidth in the historical paired sweep: 3B/7B score 17.22%/12.64% at LP1 but 20.33%/29.49% at Full. These native references are distinct from the AKR-adapted scale study and do not imply that one model size is uniformly preferable.

F.2 Class-routed prompt transfer and stability

A class-routed dictionary selects a class-specific gradient centroid using a classifier prediction. Its Dogs A–J prompt experiment reuses the same support gradients and router and changes only query wording or candidate order. Both prompt results below belong to that variant, not to the continuous regressor.

Table 20: **Class-routed pooled prompt transfer.** Complete results on 139 validation queries with $\alpha = .03$. Accuracy/macro-F1 (%); R/H are changes relative to native.

Query prompt	Native Acc./F1	Class-routed Acc./F1	R	H
Paraphrase	12.95 / 7.58	30.94 / 24.97	33	8
Reversed candidates	10.79 / 2.01	18.71 / 13.45	22	11

The net gains are 17.99 and 7.91 points, with paired intervals [9.35, 26.62] and [0.00, 15.83]. Reversal gives fewer repairs and more damaged answers than paraphrase. Neither result establishes prompt invariance for continuous AKR.

The ordered variant becomes less accurate as strength rises and reaches 99.28% invalid responses at .03. This result is retained together with the favourable prompt-transfer outcomes of the same broader variant family. It is not used to assign failure rates to the continuous pooled method.