PRESERVE, THEN BOUND: CONDITIONALLY INFORMATIVE ACOUSTIC CONTEXT FOR FEW-SHOT ATC CALL-SIGN SPOTTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Acoustic context in few-shot keyword spotting is neither uniformly irrelevant nor consistently reliable. Speaker and channel cues can disambiguate similar lexical targets while a local source association persists, yet become misleading when that association changes. We introduce BioS2D, a framework that treats these cues as conditionally informative context while retaining lexical content as the primary evidence. BioS2D summarizes class-stable lexical content with class prototypes and preserves transient acoustic context in an instance-level support memory. A pairwise relation network derives a candidate-wise contextual applicability cue for each query–candidate pair, using joint class- and source-agreement supervision during training without requiring source labels at inference. This applicability cue and margin-based lexical uncertainty jointly modulate a bounded contextual residual, allowing contextual evidence to refine uncertain lexical comparisons without granting it unrestricted decision authority. The bounded update preserves the lexical top-1 prediction whenever the lexical margin exceeds the maximum possible context-induced score difference. We evaluate BioS2D on air traffic control call-sign spotting in ATCO2 and UWB-ATCC under class-disjoint 5-way 1-shot and 5-shot protocols with clean and acoustically degraded queries. Controlled ablations examine the contributions of instance memory, source-context supervision, contextual applicability, and lexical-uncertainty conditioning. By separating contextual applicability from decision authority, BioS2D provides a strong lexically anchored recognizer and extends it with transient acoustic evidence, refining ambiguous comparisons while preserving sufficiently decisive lexical predictions.

1 INTRODUCTION

Few-shot keyword spotting (KWS) enables previously unseen spoken targets to be enrolled from only a small number of examples (Yuan et al., 2024), extending recognition beyond fixed command inventories toward query-by-example (Wang et al., 2024), open-vocabulary (Jin et al., 2024; Kim & Hwang, 2025), and rapidly personalized settings (Jung et al., 2025a). This formulation is particularly relevant to air traffic control (ATC), where call signs are short, structured identifiers whose active inventory changes with traffic and whose lexical forms may differ by only a few letters or digits. Recent ATC corpora (Zuluaga-Gomez et al., 2023) and call-sign studies (Li et al., 2027) further expose radio degradation, non-native accents, limited supervision, and unseen identities. Existing few-shot recognizers therefore focus primarily on lexical similarity, while non-lexical variation is either absorbed into a shared representation or suppressed for robustness. Yet ATC communications suggest a less conventional possibility: speaker, channel, and background characteristics do not define a call sign, but may temporarily provide useful evidence for distinguishing lexically similar candidates.

Prior ATC research has improved call-sign recognition in two complementary ways. Speech-based approaches strengthen lexical inference through domain-adapted recognition (Zuluaga-Gomez et al., 2020), call-sign extraction (Blatt et al., 2022), or acoustic matching (Li et al., 2027), while context-aware systems incorporate surveillance-derived candidate lists (Kocour et al., 2021; Zuluaga-Gomez et al., 2021), flight information, or contextual biasing (Bhattacharjee et al., 2024) to favor operationally plausible identifiers. These approaches exploit either the spoken content or external opera-

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.

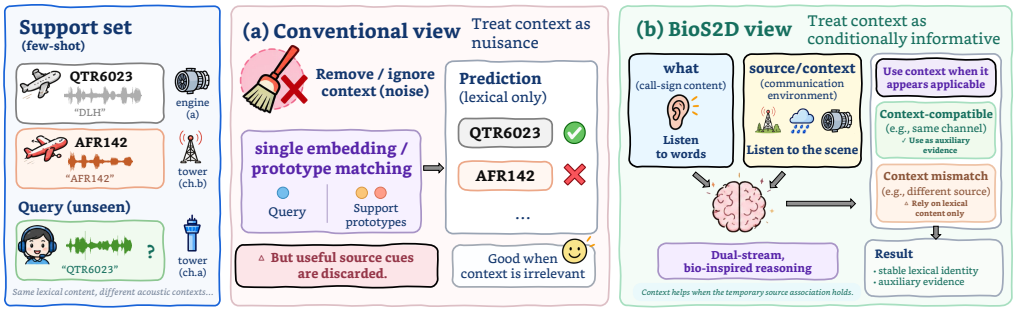

Figure 1: Conditional use of acoustic context in BioS2D. Lexical similarity provides the primary evidence for call-sign identity. The support set is represented by a lexical prototype and an instance-level context memory. For each query–candidate pair, BioS2D derives a candidate-wise contextual applicability cue and uses lexical uncertainty to bound the residual contribution of contextual evidence. Context can therefore refine ambiguous lexical comparisons, while its influence is progressively restricted as the lexical decision becomes more confident.

tional context. In parallel, robust KWS typically treats speaker, channel, and background variation as nuisance factors and seeks invariance (Chen et al., 2025) or adaptation (Xiao et al., 2025) to them. What remains largely unexplored is the context already carried by the enrolled speech itself. Within a local communication episode, non-lexical acoustics may remain associated with a particular query–support relation; once that association changes, the same acoustic similarity can become misleading. We refer to this regime as *conditionally informative acoustic context*. The resulting question is not whether such acoustics should always be retained or removed, but how they can be preserved when useful without turning transient associations into identity shortcuts.

This observation has two consequences for few-shot representation learning. First, lexical and contextual evidence have different lifetimes. Lexical content is stable across realizations of the same call-sign class, whereas source and channel conditions may vary across individual support utterances. Although self-supervised speech models (Chen et al., 2022a) encode linguistic, speaker, and acoustic information across their representations (Pasad et al., 2023; Ashihara et al., 2024), standard prototypical learning summarizes all support evidence at the class level (Snell et al., 2017). BioS2D instead uses an asymmetric dual-evidence representation: class-stable lexical evidence is summarized by a prototype, whereas transient contextual evidence is retained as an instance-level memory. Second, preserving context does not imply granting it unrestricted decision power. BioS2D therefore separates contextual applicability from contextual authority: a candidate-wise applicability cue indicates whether the retained context may be useful for a query–candidate relation, while lexical uncertainty bounds the authority granted to that context. Together, these choices implement a single principle—preserve potentially useful context, but keep lexical identity as the decision anchor.

We instantiate this principle in **BioS2D**, a brain-inspired dual-evidence metric-learning framework for few-shot ATC call-sign spotting. Rather than treating non-lexical acoustics as either nuisance or class evidence, BioS2D introduces the notion of conditionally informative acoustics and models a query–candidate-specific cue for their potential usefulness. Its prototype–memory representation further matches support-set aggregation to the lifetime of each evidence type, while lexically anchored bounded allocation restricts the influence of transient context as lexical confidence increases. Experiments on ATCO2 and UWB-ATCC show that BioS2D sets a strong lexical anchor for few-shot call-sign spotting, and that preserved context extends it: contextual acoustics contain information complementary to lexical similarity and refine ambiguous or degraded decisions without becoming the dominant decision cue. The remaining contextual headroom identifies applicability estimation as the lever for further gains.

2 Related Work

Flexible keyword spotting and call-sign recognition. Flexible KWS replaces a closed command classifier with enrollment: query-by-example systems learn acoustic similarity from exemplars (Wang et al., 2024), audio–text systems align speech with arbitrary keywords (Jung et al.,

2024; Jin et al., 2024; Jung et al., 2025b), and few-shot adaptation specializes a pretrained spotter from limited target speech (Kim et al., 2024; Jung et al., 2025a; Yuan et al., 2024; Kim & Hwang, 2025). All score candidates in a shared embedding space or against a class representative, so non-lexical evidence carries the same weight for every comparison in an episode, whether or not the acoustic conditions that produced it persist. Call-sign recognition has developed along a complementary line, documenting the phraseology, accents and radio degradation that make the task hard (Zuluaga-Gomez et al., 2020; 2023; Li et al., 2026), using surveillance data to bias decoding toward active flights (Kocour et al., 2021; Zuluaga-Gomez et al., 2021; Blatt et al., 2022), and augmenting with external knowledge (Hui et al., 2025)—all injecting side information from outside the audio. BioS2D instead uses a side channel already present in the enrolled speech, the acoustic trace of each support's source, and infers its relevance per query–class pair because that trace can expire.

Content and source information in speech representations. Self-supervised encoders (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022a) represent content, speaker identity and environment in one stream, and layer-wise probing shows these factors follow different trajectories through the network (de Seyssel et al., 2022; Pasad et al., 2023; Ashihara et al., 2024); speaker-aware models confirm that source characteristics survive brief utterances (Chen et al., 2021; 2022b; 2024). This factor structure gives a two-stream design a principled basis, echoing dual-stream speech enhancement (Rong et al., 2026a) and the functional separation of auditory streams (Hickok & Poeppel, 2007; Rauschecker & Scott, 2009). These analyses establish that the factors are separable. BioS2D adds the allocation step that decides when the contextual factor applies, assigning the streams different support-set operators and gating the contextual one.

Robust learning with conditionally useful context. Robust KWS conventionally seeks invariance to environment and domain (Ritter-Gutierrez et al., 2024; Chen et al., 2025; Xiao et al., 2025), and the wider literature learns domain-invariant features (Ganin et al., 2016; Kim et al., 2019), optimizes over groups (Sagawa et al., 2020; Creager et al., 2021; Liu et al., 2021), or guards against spurious structure in embeddings (Lu et al., 2025; Paduraru et al., 2026; Park et al., 2026). Transient source acoustics call for a third treatment: global invariance discards evidence that is locally predictive, while unrestricted fusion lets the same evidence act as a shortcut. Relevance-aware context selection in automatic speech recognition (Rong et al., 2026b) and uncertainty-calibrated speech learning (Kim & Lee, 2025) point to the middle ground, which BioS2D formalizes at the decision level: estimate applicability per query–candidate relation, then bound its influence so an incorrect estimate costs a bounded amount of accuracy.

3 METHOD

Given an N-way K-shot episode, let $\mathcal{S}_k = \{x_{k,i}\}_{i=1}^{K}$ denote the support set of call-sign class k, and let x_q denote a query utterance. During training, each utterance may additionally carry a source-context identifier $c(x)$, which is used only for auxiliary supervision.

BioS2D contains three main components. First, a brain-inspired dual-stream encoder extracts lexical content and non-lexical acoustic context from a shared frozen speech representation. Second, the two streams are organized asymmetrically across the support set: lexical embeddings are summarized by class prototypes, while contextual embeddings are retained as instance-level memory. Third, a lexically anchored fusion rule combines a candidate-wise contextual applicability cue with lexical uncertainty to bound the influence of auxiliary context on the final prediction. Figure 2 illustrates the overall framework.

3.1 BRAIN-INSPIRED DUAL-STREAM REPRESENTATION

Inspired by the functional division of auditory processing between identifying *what* is heard and retaining complementary non-lexical acoustic context, BioS2D constructs two asymmetric pathways over a frozen WavLM-Large encoder (Chen et al., 2022a). The content pathway targets lexical information associated with call-sign identity, while the context pathway preserves speaker-, source-, channel-, and background-related acoustic characteristics.

WavLM-Large provides 24 hidden-layer sequences. The content pathway aggregates later-layer representations, applies masked instance normalization adapted from Ulyanov et al. (2016), and

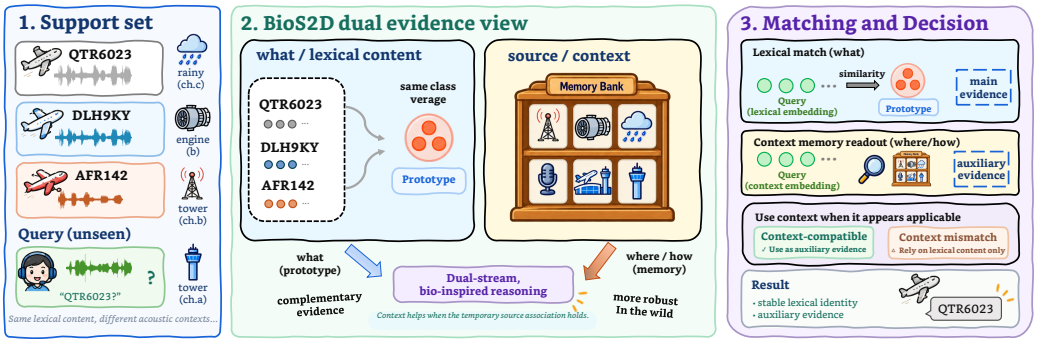

Figure 2: Overview of BioS2D. A brain-inspired dual-stream encoder produces lexical content and contextual representations. Support lexical embeddings are summarized as class prototypes, while contextual embeddings are retained as instance-level memory. Lexical prototype similarity provides the primary classification score. Context-memory matching provides auxiliary evidence, whose contribution is controlled by candidate-wise contextual applicability and lexical uncertainty through a bounded residual.

uses a lightweight Conformer (Gulati et al., 2020) to model lexical temporal structure. The context pathway aggregates earlier-layer representations and applies attentive statistics pooling (Okabe et al., 2018) to capture distributional source and channel characteristics.

The two pathways produce normalized embeddings

$$\mathbf{z}^c(x) = \frac{f_c(x)}{\|f_c(x)\|_2}, \qquad \mathbf{z}^e(x) = \frac{f_e(x)}{\|f_e(x)\|_2}, \tag{1}$$

where $\mathbf{z}^c(x)$ denotes the lexical/content representation and $\mathbf{z}^e(x)$ denotes the contextual representation. The two pathways use independent layer aggregation, temporal processing, and supervision, providing the two evidence streams used in few-shot inference.

3.2 Prototype–Memory Matching

Support examples from the same call-sign class share lexical identity but may differ in speaker, channel, and recording context. BioS2D therefore applies different support-set operators to the two streams.

Lexical prototype. Following prototypical networks (Snell et al., 2017), the content embeddings of class k are summarized by

$$\mathbf{p}_k^c = \frac{1}{K}\sum_{i=1}^{K} \mathbf{z}^c(x_{k,i}), \qquad s_{q,k}^c = \cos(\mathbf{z}^c(x_q), \mathbf{p}_k^c)\,. \tag{2}$$

The lexical score $s_{q,k}^c$ provides the primary call-sign evidence.

Context memory. Contextual conditions can vary across supports of the same class. We retain their embeddings individually as an instance-level context memory

$$\mathcal{M}_k^e = \{\mathbf{z}^e(x_{k,1}), \ldots, \mathbf{z}^e(x_{k,K})\}, \tag{3}$$

and compute query-to-support similarities

$$r_{q,k,i}^e = \cos(\mathbf{z}^e(x_q), \mathbf{z}^e(x_{k,i}))\,. \tag{4}$$

A shot-normalized LogMeanExp readout aggregates the instance-level matches:

$$s_{q,k}^e = \tau_m \log\left[\frac{1}{K}\sum_{i=1}^{K} \exp\left(\frac{r_{q,k,i}^e}{\tau_m}\right)\right]. \tag{5}$$

This operation retains strong compatibility with individual support instances while remaining permutation invariant over the support set.

3.3 LEXICALLY ANCHORED BOUNDED FUSION

The context-memory score provides auxiliary evidence whose usefulness is conditional on the current query–candidate relation. BioS2D does not treat a learned applicability score as a guarantee of a valid source-context match. Instead, it separates a candidate-wise applicability cue from the authority granted to context, using relative contextual evidence, contextual applicability, and lexical uncertainty to construct a bounded residual.

Relative contextual evidence. For each query, the context scores are normalized relative to the candidate set:

$$e_{q,k} = \tanh\left(\frac{s^e_{q,k} - \mu_q}{\tau_e\sqrt{\sigma_q^2 + \epsilon}}\right), \tag{6}$$

where μ_q and σ_q^2 are the mean and variance of $\{s^e_{q,j}\}_{j=1}^N$. Thus, $e_{q,k} \in [-1, 1]$ measures whether candidate k is unusually compatible with the query context relative to the remaining candidates.

Contextual applicability cue. To obtain a candidate-wise cue for the potential usefulness of contextual evidence for candidate k, we construct a pairwise relation representation

$$\rho(\mathbf{a}, \mathbf{b}) = [\mathbf{a}; \mathbf{b}; |\mathbf{a} - \mathbf{b}|; \mathbf{a} \odot \mathbf{b}]. \tag{7}$$

A lexical relation encoder compares $\mathbf{z}^c(x_q)$ with $\mathbf{p}^c_k$, while a context relation encoder compares $\mathbf{z}^e(x_q)$ with the entries of $\mathcal{M}^e_k$ and aggregates the instance-level relations. Their concatenated representation is mapped to

$$b_{q,k} \in [0, 1], \tag{8}$$

which is used as the candidate-wise contextual applicability cue.

Lexical uncertainty. Let

$$m_q = s^c_{q,(1)} - s^c_{q,(2)} \tag{9}$$

be the margin between the two highest lexical scores. We define

$$u_q = \mathrm{sg}\left[\exp\left(-\frac{m_q}{\tau_u}\right)\right], \tag{10}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. A small lexical margin produces a larger u_q, increasing the maximum authority that contextual evidence may receive.

The final score is

$$s_{q,k} = s^c_{q,k} + \underbrace{\alpha u_q \,\mathrm{sg}(b_{q,k}) e_{q,k}}_{\Delta_{q,k}}, \tag{11}$$

where $\Delta_{q,k}$ is the contextual residual. Since $b_{q,k} \in [0, 1]$ and $e_{q,k} \in [-1, 1]$,

$$|\Delta_{q,k}| \le \alpha u_q. \tag{12}$$

The contextual residual can change the score difference between two candidates by at most $2\alpha u_q$. Therefore,

$$m_q > 2\alpha u_q \quad \Longrightarrow \quad \arg\max_k s_{q,k} = \arg\max_k s^c_{q,k}. \tag{13}$$

Context consequently receives greater influence when lexical candidates are close, while sufficiently confident lexical decisions are preserved.

3.4 TRAINING OBJECTIVES

BioS2D is trained with an episodic classification objective and two auxiliary context objectives.

The primary classification loss $\mathcal{L}_{\mathrm{cls}}$ applies cross-entropy to the final scores $\{s_{q,k}\}_{k=1}^N$.

To structure the contextual embedding space, we use a supervised contrastive loss (Khosla et al., 2020), $\mathcal{L}_{\mathrm{ctx}}$, which pulls together context embeddings sharing the same training-time source-context identifier and separates different contexts.

The contextual applicability score is trained using

$$t_{q,k} = \mathbf{1}[y_q = k]\,\mathbf{1}\left[\exists x \in \mathcal{S}_k : c(x) = c(x_q)\right]. \tag{14}$$

Importantly, $t_{q,k}$ defines decision-level contextual applicability: contextual evidence is regarded as applicable to candidate k only when the candidate is lexically correct and its support set contains source-compatible evidence. A source-matched but incorrect competitor is therefore negative by design, since amplifying its contextual similarity would reinforce a wrong identity decision. We separately evaluate whether the learned cue retains source-related information after controlling for class compatibility in Section 4.3 and Appendix C.1.

To balance the class/source combinations, the binary cross-entropy (BCE) is computed separately for true-class/same-context, true-class/different-context, competitor/same-context, and competitor/different-context pairs. The four stratum losses are then averaged:

$$\mathcal{L}_{\text{app}} = \frac{1}{4} \sum_{r \in \{\text{TS,TD,CS,CD}\}} \mathcal{L}_{\text{BCE}}^{(r)}. \tag{15}$$

The complete objective is

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{ctx}}\mathcal{L}_{\text{ctx}} + \lambda_{\text{app}}\mathcal{L}_{\text{app}}, \qquad \lambda_{\text{ctx}} = \lambda_{\text{app}} = 0.1. \tag{16}$$

The contextual objectives are evaluated on a balanced auxiliary episode stream constructed from the training data. The content representation is protected during the auxiliary update, and $\text{sg}(b_{q,k})$ prevents the classification objective from directly updating the applicability estimator. At inference, BioS2D uses only the support audio, support class labels, and query audio; source-context identifiers are required only during training.

4 Experiments

4.1 Experimental Setup

Datasets and evaluation. We evaluate BioS2D on call-sign subsets derived from ATCO2 (Zuluaga-Gomez et al., 2023) and UWB-ATCC (Šmídl et al., 2019). After filtering, the two subsets contain 50,860 utterances from 1,200 call-sign classes and 5,086 utterances from 1,333 classes, respectively. The primary evaluation considers 5-way 1-shot and 5-way 5-shot classification with five queries per class. Training, validation, and test partitions use disjoint call-sign classes and source contexts. Source-context identifiers are used only for auxiliary supervision during training and for constructing controlled diagnostics; they are never provided as model inputs at validation or test time. All systems are evaluated on the same precomputed episodes: two validation and five test manifests of 100 episodes each. BioS2D and its ablations are trained with three independent seeds, whereas external reference systems use the preregistered seed 1999.

Baselines and analysis protocol. We compare BioS2D with spectral few-shot classifiers on log-Mel or mel-frequency cepstral coefficient (MFCC) features, namely BCResNet (Kim et al., 2021), DSCNN (Zhang et al., 2017), ECAPA-TDNN (Desplanques et al., 2020), TCSN (Zhu et al., 2024), and pretrained SS-Flex (Li et al., 2027), and with WavLM-based references: last-layer, learned layer weighting, and ResNet15 (Tang & Lin, 2018) on WavLM representations (Chen et al., 2022a). For controlled analysis, the matched WavLM variant *Content only* removes the contextual pathway and contextual allocation, and further variants replace individual representation and allocation choices. Because the context memory is defined over multiple support instances, the analysis is conducted at 5 shots, where it additionally uses a higher-competition 10-way nearest-hard diagnostic and controlled source-relation episodes. All audio is converted to 16-kHz mono and cropped to at most two seconds; BioS2D uses a frozen WavLM-Large backbone, 256-dimensional pathway embeddings, and 40,000 training episodes. Implementation details, 1-shot analyses, and additional ablations are provided in the Appendix.

4.2 Main Results

Table 1 compares BioS2D with existing few-shot call-sign recognition systems on ATCO2 and UWB-ATCC under both 1-shot and 5-shot enrollment.

Table 1: Performance comparison across few-shot call-sign recognition models. Accuracy (%) is reported on ATCO2 and UWB-ATCC under 5-way 1-shot and 5-way 5-shot evaluation on the fixed test episodes. BioS2D reports the mean with the training-seed SD in small type. Best value per column in bold; the proposed model is shaded. The margin over the references is established by the lexical pathway; the contribution of context is isolated against the matched Content only model in Table 2.

	ATCO2		UWB-ATCC	
Model	1-shot	5-shot	1-shot	5-shot
Spectral few-shot references†				
BCResNet	90.13	96.65	90.45	81.97
DSCNN	71.00	88.86	69.98	70.61
ECAPA-TDNN	90.50	96.65	88.40	84.45
SS-Flex	91.82	96.78	66.04	80.05
TCSN	84.09	95.34	85.44	83.38
WavLM references†				
WavLM last	73.27	88.94	69.17	90.65
WavLM weighted	63.41	84.25	68.93	90.64
ResNet15	85.46	93.86	79.70	81.50
BioS2D	**93.87** ±0.22	**97.77** ±0.16	**96.47** ±0.10	**98.49** ±0.32

† One registered training seed.

As shown in Table 1, BioS2D achieves strong few-shot call-sign recognition performance across both corpora and enrollment regimes. It obtains the best accuracy in all four settings, reaching 97.77% on ATCO2 5-shot and 98.49% on UWB-ATCC 5-shot, 7.84 points above the strongest reference system on UWB-ATCC. This margin is established by the lexical pathway, which the matched Content only model shares (Table 2), so the contextual pathway is evaluated as an extension of a strong anchor against that model. At 5 shots the clean fixed episodes leave only a small set of lexically confusable queries, so the controlled analyses below complement clean accuracy with conditions that reopen this headroom: query-side degradation, the relation-balanced diagnostic, and the 10-way nearest-hard diagnostic.

4.3 Analysis

We examine whether non-lexical acoustic context provides useful information, ablate the asymmetric representation and the allocation components, analyze how much complementary information context retains and how reliably its applicability is identified, and test whether bounded allocation prevents unreliable contextual evidence from overriding confident lexical decisions.

Context as an extension of the lexical anchor. Contextual evidence can provide additional discrimination in some recognition conditions. On ATCO2 5-shot, BioS2D achieves 97.77% accuracy, higher than the 97.54% obtained by Content only (Table 2), and on the relation-balanced diagnostic its macro accuracy rises from 97.56% to 98.03%. On identical relation-balanced episodes, the paired difference to Content only is positive under clean and all six perturbed query conditions, from $+0.22$ to $+0.80$ points, with a 95% hierarchical bootstrap interval excluding zero under 20% band masking (Appendix B.3). These results indicate that non-lexical acoustics are not purely nuisance variation: removing them a priori can discard information that is predictive within a particular episode distribution. The extension depends on how much lexical headroom remains: on UWB-ATCC 5-shot, where clean accuracy is already close to ceiling, Content only reaches 98.52% against 98.49% for BioS2D (the 1-shot comparison is given in Appendix A.1). Context therefore acts as auxiliary evidence rather than stable class evidence, which motivates the treatment adopted in BioS2D: non-lexical information should not be globally suppressed, but neither should it be incorporated unconditionally into the class representation.

Ablation study. Table 2 ablates the main representation and allocation components of BioS2D, separately examining pathway specialization, support-set context representation, candidate-dependent applicability, and lexical-uncertainty-dependent authority.

Table 2: Ablation study of the main representation and allocation components of BioS2D, 5-way 5-shot. Accuracy (%) on the fixed test episodes: ATCO2 clean, with 20-dB Gaussian noise and with 2.5-kHz low-pass (LP) filtering on the query, and UWB-ATCC clean; and, on the relation-balanced ATCO2 diagnostic, macro accuracy (%) and the share (%) of decisions changed when context memories are reassigned to the wrong candidates. Means with the training-seed SD in small type; best accuracy per column in bold.

	Fixed episodes				Relation diagnostic	
	ATCO2			UWB-ATCC		
Variant	Clean	20 dB	LP 2.5 kHz	Clean	Macro	Changed
Independently trained content control						
Content only	97.54 ±0.19	97.44 ±0.25	96.54 ±0.05	98.52 ±0.17	97.56 ±0.25	0.00 ±0.00
Pathway assignment						
Swapped layer ranges	97.64 ±0.36	97.56 ±0.38	96.31 ±0.77	97.47 ±0.69	97.31 ±0.99	1.09 ±0.83
Mechanism ablations						
Mean context prototype	97.75 ±0.21	97.64 ±0.22	96.77 ±0.15	**98.53** ±0.29	97.88 ±0.38	0.76 ±0.21
Fixed applicability	97.69 ±0.17	97.66 ±0.18	96.65 ±0.25	**98.53** ±0.43	97.79 ±0.12	1.32 ±0.21
w/o lexical uncertainty	97.76 ±0.12	97.65 ±0.15	96.77 ±0.05	98.48 ±0.29	97.81 ±0.32	0.93 ±0.25
BioS2D	**97.77** ±0.16	**97.69** ±0.14	**96.88** ±0.12	98.49 ±0.32	**98.03** ±0.42	0.74 ±0.07

Content only is invariant to the memory reassignment by construction.

Effect of evidence-specific representation. Swapping the layer ranges of the two pathways yields 97.64% on ATCO2 and 97.47% on UWB-ATCC, against 97.77% and 98.49% for BioS2D. The benefit therefore depends on assigning each pathway the layer range that suits its evidence, not merely on access to both ranges. The effect is compressed on clean episodes, where both corpora approach the ceiling, and emerges once the channel degrades: under 10-dB noise on UWB-ATCC, swapping the ranges lowers accuracy from 96.30% to 90.89% (Appendix Table 5), and on ATCO2 with 2.5-kHz low-pass queries it falls to 96.31% against 96.88% for BioS2D.

Effect of the prototype–memory representation. Replacing the instance-preserving context memory with a class-level mean embedding yields 97.75% on ATCO2 and 98.53% on UWB-ATCC, and lowers the relation-diagnostic macro accuracy from 98.03% to 97.88%. The memory itself thus does not provide additional classifier capacity. Its role is structural: lexical evidence can be naturally summarized by a class prototype, whereas retaining individual context instances avoids imposing a class-level centroid assumption on contextual evidence whose association may vary across support utterances.

Effect of contextual applicability and lexical uncertainty. Fixed applicability assigns the retained context a uniform applicability factor; a mismatched context memory then changes 1.32% of its decisions, against 0.74% for BioS2D. Removing lexical uncertainty, which restricts contextual intervention when the lexical decision is already confident, raises this share to 0.93%. Together, the two ablations separate the roles of the allocation module: estimating whether contextual evidence may be relevant and determining how much authority that evidence should receive. Pathway-range and auxiliary-supervision ablations are reported in Appendix A.

Contextual information and applicability. Because the standard 5-way 5-shot task is close to saturation, we use a 10-way diagnostic built from lexically similar candidate classes to measure the discriminative information retained by each evidence score (Table 3). The relative contextual evidence $e_{q,k}$ read from the context memory achieves an area under the ROC curve (AUROC) of 0.640, compared with 0.573 for lexical similarity, confirming that the preserved context contains complementary information not fully captured by lexical matching. This information is only partially converted into recognition gains: a zero-harm oracle reveals $+3.82$ percentage points of potential improvement, $+1.56$ of it within the residual bound, whereas the learned contextual residual realizes -0.09 points. The contextual information is therefore present and largely admissible under the bound, and determining when it can be used reliably is where the method can grow further. Consistently, on the relation-balanced diagnostic applicability separates source relations at 0.589 AUROC once class compatibility is controlled, above 0.550 for the lexical score, so we treat con-

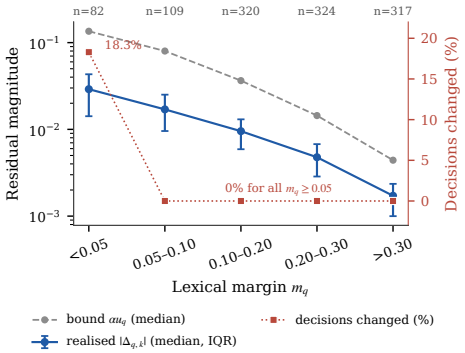

Figure 3: Contextual authority as a function of lexical confidence on the 10-way diagnostic. Dashed: median residual bound αu_q; solid: median realized residual $|\Delta_{q,k}|$ with interquartile range; squares: share of decisions changed (right axis).

Table 3: Contextual information and allocation on the 5-shot ATCO2 10-way nearest-hard and relation-balanced diagnostics (three seeds). 10-way AUROCs use the middle tertile of lexical similarity; coverage is the share of queries satisfying Eq. (13).

Diagnostic	10-way	Balanced
Lexical-score AUROC	0.573	–
Contextual-evidence AUROC	0.640	–
Applicability AUROC	0.634	–
Class-controlled applicability	–	0.589
Zero-harm oracle (pp)	+3.82	–
Oracle within αu_q (pp)	+1.56	–
Realized gain (pp)	−0.09	+0.03
Coverage (%)	80.0	94.3

textual applicability as an auxiliary cue rather than a calibrated estimate of whether context is valid evidence.

Effect of bounded allocation. Figure 3 examines how contextual authority changes with lexical confidence. As the lexical margin m_q increases, both the residual bound αu_q and the realized residual magnitude decrease monotonically, and decisions change only for $m_q < 0.05$. Context therefore receives its largest decision budget precisely in lexically ambiguous cases, while its influence is progressively suppressed as the lexical decision becomes more confident. The realized residual remains below the permitted bound in every margin range, and 80.0% of 10-way queries (94.3% on the relation-balanced diagnostic) satisfy the preservation condition. This confirms that contextual evidence operates as a limited correction to uncertain lexical comparisons rather than as a competing source of class identity. As shown in Appendix C.2, the bound still leaves substantial correction capacity in the most ambiguous regime, identifying applicability estimation, rather than the residual budget itself, as the dominant remaining bottleneck.

Robustness under acoustic degradation. Under query-side Gaussian noise, band masking, and low-pass filtering with clean supports, BioS2D remains above Content only across all six ATCO2 perturbations, by 0.22–0.34 percentage points, and obtains the best accuracy among all variants with 20-dB noise and 2.5-kHz low-pass queries (Table 2). On identical relation-balanced episodes the paired difference to Content only is positive under clean and all six perturbed conditions (Appendix B.3). The perturbation settings are given in Appendix D.

Overall, BioS2D combines a strong lexical anchor with non-lexical acoustic context that contains complementary task information and helps where lexical evidence is ambiguous.

5 Conclusion

In this work, we presented BioS2D, a few-shot ATC call-sign spotting framework that treats non-lexical acoustics as conditionally informative context rather than as variation that should be either universally suppressed or universally trusted. BioS2D separates class-stable lexical evidence from instance-transient context through an asymmetric prototype–memory representation, and introduces bounded evidence allocation to keep lexical information as the decision anchor. This design preserves potentially useful source-related information while limiting its influence when the lexical decision is already reliable, providing a conservative alternative to both nuisance suppression and unrestricted context fusion. Experiments on ATCO2 and UWB-ATCC show that BioS2D provides a strong lexically anchored recognizer and that non-lexical context extends it under ambiguous or degraded conditions, while the measured oracle headroom indicates that better applicability estimation can extend these gains further.

AI USE STATEMENT

In this work, we used generative AI tools for literature retrieval, for language editing, for drafting sections of the manuscript, for parts of the code, and for the cartoon iconography in the figures. We have reviewed all AI-assisted work: retrieved references were checked against primary sources, generated code was tested for correctness, and every passage of text and figure was revised and approved by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses a licensed air traffic control dataset (ATCO2).

REPRODUCIBILITY STATEMENT

The supplementary material contains the complete training, validation and test code, the configuration of every reported model and ablation, the fixed validation and test episode manifests, and the code for the relation-balanced and 10-way diagnostics and the query-side perturbation conditions. The evaluation protocol is described in Section 4.1, the construction of the relation-balanced diagnostic in Appendix B.1, and the perturbation settings in Appendix D; all reported results average three independently trained seeds unless marked as single-seed references.

Two dependencies lie outside the release. UWB-ATCC must be downloaded by the user from its public source. ATCO2 is distributed under a licence that requires each user to obtain access independently; once that access is granted, the released code and manifests reproduce the reported experiments without modification.

REFERENCES

Takanori Ashihara, Marc Delcroix, Takafumi Moriya, Kohei Matsuura, Taichi Asami, and Yusuke Ijima. What Do Self-Supervised Speech and Speaker Models Learn? New Findings from a Cross Model Layer-Wise Analysis. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10166–10170, April 2024. doi: 10.1109/ICASSP48485.2024.10446422.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html.

Mrinmoy Bhattacharjee, Iuliia Nigmatulina, Amrutha Prasad, Pradeep Rangappa, Srikanth Madikeri, Petr Motlicek, Hartmut Helmke, and Matthias Kleinert. Contextual biasing methods for improving rare word detection in automatic speech recognition. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12652–12656, 2024. doi: 10.1109/ICASSP48485.2024.10447465.

Alexander Blatt, Martin Kocour, Karel Veselý, Igor Szöke, and Dietrich Klakow. Call-Sign Recognition and Understanding for Noisy Air-Traffic Transcripts Using Surveillance Information. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8357–8361, May 2022. doi: 10.1109/ICASSP43922.2022.9746301.

Sanyuan Chen, Yu Wu, Chengyi Wang, Zhengyang Chen, Zhuo Chen, Shujie Liu, Jian Wu, Yao Qian, Furu Wei, Jinyu Li, and Xiangzhan Yu. UniSpeech-SAT: Universal Speech Representation Learning with Speaker Aware Pre-Training, October 2021. arXiv:2110.05752 [cs.CL].

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, October 2022a. ISSN 1941-0484. doi: 10.1109/JSTSP.2022.3188113.

Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, Qian Chen, Shiliang Zhang, and Junjie Li. ERes2NetV2: Boosting Short-Duration Speaker Verification Performance with Computational Efficiency. In *Interspeech 2024*, pp. 3245–3249. ISCA, September 2024. doi: 10.21437/Interspeech.2024-742.

Yongqiang Chen, Qianhua He, Zunxian Liu, Mingru Yang, and Wenwu Wang. Noise-robust feature extraction for keyword spotting based on supervised adversarial domain adaptation training strategies. *Speech Communication*, 175:103323, November 2025. ISSN 0167-6393. doi: 10.1016/j.specom.2025.103323.

Zhengyang Chen, Sanyuan Chen, Yu Wu, Yao Qian, Chengyi Wang, Shujie Liu, Yanmin Qian, and Michael Zeng. Large-Scale Self-Supervised Speech Representation Learning for Automatic Speaker Verification. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6147–6151, May 2022b. doi: 10.1109/ICASSP43922.2022.9747814.

Elliot Creager, Joern-Henrik Jacobsen, and Richard Zemel. Environment Inference for Invariant Learning. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 2189–2200. PMLR, July 2021. URL https://proceedings.mlr.press/v139/creager21a.html.

Maureen de Seyssel, Marvin Lavechin, Yossi Adi, Emmanuel Dupoux, and Guillaume Wisniewski. Probing phoneme, language and speaker information in unsupervised speech representations. In *Interspeech 2022*, pp. 1402–1406, 2022. doi: 10.21437/Interspeech.2022-373.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Interspeech 2020*, pp. 3830–3834, October 2020. doi: 10.21437/Interspeech.2020-2650. arXiv:2005.07143 [eess.AS].

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1):2096–2030, January 2016. ISSN 1532-4435.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020*, pp. 5036–5040, 2020. doi: 10.21437/Interspeech.2020-3015.

Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5):393–402, May 2007. ISSN 1471-0048. doi: 10.1038/nrn2113.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. ISSN 2329-9304. doi: 10.1109/TASLP.2021.3122291.

Yi Hui, Yang Yang, Shengsheng Qian, and Kaiquan Cai. Knowledge-augmented encoder for few-shot deep intent recognition in air traffic control. *Knowledge-Based Systems*, 320:113524, June 2025. ISSN 0950-7051. doi: 10.1016/j.knosys.2025.113524.

Sichen Jin, Youngmoon Jung, Seungjin Lee, Jaeyoung Roh, Changwoo Han, and Hoonyoung Cho. CTC-aligned Audio-Text Embedding for Streaming Open-vocabulary Keyword Spotting. In *Interspeech 2024*, pp. 332–336, 2024. doi: 10.21437/Interspeech.2024-706.

Youngmoon Jung, Seungjin Lee, Joon-Young Yang, Jaeyoung Roh, Chang Woo Han, and Hoon-Young Cho. Relational Proxy Loss for Audio-Text based Keyword Spotting. In *Interspeech 2024*, pp. 327–331, 2024. doi: 10.21437/Interspeech.2024-229.

Youngmoon Jung, Jinyoung Lee, Seungjin Lee, Myunghun Jung, Yong-Hyeok Lee, and Hoon-Young Cho. Text-Aware Adapter for Few-Shot Keyword Spotting. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, April 2025a. doi: 10.1109/ICASSP49660.2025.10890609.

Youngmoon Jung, Yong-Hyeok Lee, Myunghun Jung, Jaeyoung Roh, Chang Woo Han, and Hoon-Young Cho. Adversarial Deep Metric Learning for Cross-Modal Audio-Text Alignment in Open-Vocabulary Keyword Spotting. In *Interspeech 2025*, pp. 2645–2649. ISCA, August 2025b. doi: 10.21437/Interspeech.2025-378.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673. Curran Associates, Inc., 2020. URL https://papers.neurips.cc/paper_files/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.

Byeonggeun Kim, Simyung Chang, Jinkyu Lee, and Dooyong Sung. Broadcasted residual learning for efficient keyword spotting. In *Interspeech 2021*, pp. 4538–4542, 2021. doi: 10.21437/Interspeech.2021-383.

Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning Not to Learn: Training Deep Neural Networks With Biased Data. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9004–9012, June 2019. doi: 10.1109/CVPR.2019.00922.

Dohyun Kim and Jiwook Hwang. Fully End-to-end Streaming Open-vocabulary Keyword Spotting with W-CTC Forced Alignment. In *Interspeech 2025*, pp. 519–523, 2025. doi: 10.21437/Interspeech.2025-193.

Eungbeom Kim and Kyogu Lee. Uncertainty-Aware Self-Training for CTC-Based Automatic Speech Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23): 24330–24338, April 2025. ISSN 2374-3468. doi: 10.1609/aaai.v39i23.34610.

Ilseok Kim, Ju-Seok Seong, and Joon-Hyuk Chang. Few-Shot Keyword-Incremental Learning with Total Calibration. In *Interspeech 2024*, pp. 5083–5087, 2024. doi: 10.21437/Interspeech.2024-1823.

Martin Kocour, Karel Veselý, Alexander Blatt, Juan Zuluaga Gomez, Igor Szöke, Jan Černocký, Dietrich Klakow, and Petr Motlicek. Boosting of Contextual Information in ASR for Air-Traffic Call-Sign Recognition. In *Interspeech 2021*, pp. 3301–3305, 2021. doi: 10.21437/Interspeech.2021-1619.

Dongdong Li, Jianwei Song, Jianwei Wang, and Zhe Wang. ATCCaps: A Call-Sign-Aware Speech Dataset for Air Traffic Control Recognition, June 2026. arXiv:2606.22399 [cs.SD].

Dongdong Li, Zhenbin Lu, Zhiwei Liu, Jianwei Song, and Zhe Wang. Few-shot callsign keyword spotting in air traffic communications with self-supervised pretraining and flexible convolution. *Aerospace Science and Technology*, 180:113532, January 2027. ISSN 1270-9638. doi: 10.1016/j.ast.2026.113532.

Evan Z. Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just Train Twice: Improving Group Robustness without Training Group Information. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 6781–6792. PMLR, July 2021. URL https://proceedings.mlr.press/v139/liu21f.html.

Shenyu Lu, Junyi Chai, and Xiaoqian Wang. Mitigating Spurious Correlations in Zero-Shot Multimodal Models. In *International Conference on Learning Representations (ICLR)*, May 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2649b1d435f02524812ccbcae0dfe48b-Abstract-Conference.html.

Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda. Attentive statistics pooling for deep speaker embedding. In *Interspeech 2018*, pp. 2252–2256, 2018. doi: 10.21437/Interspeech.2018-993.

Cristian D. Paduraru, Antonio Barbalau, Radu Filipescu, Andrei Nicolicioiu, and Elena Burceanu. Bridging Explainability and Embeddings: BEE Aware of Spuriousness. In *International Conference on Learning Representations (ICLR)*, April 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/2f4837a74fe73997ad9bfd3b655f0f3d-Abstract-Conference.html.

Subeen Park, Joowang Kim, Hakyung Lee, Sunjae Yoo, and Kyungwoo Song. Spurious Correlation-Aware Embedding Regularization for Worst-Group Robustness. In *International Conference on Learning Representations (ICLR)*, April 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/fd512d61c862e3614686f26a0e09de71-Abstract-Conference.html.

Ankita Pasad, Bowen Shi, and Karen Livescu. Comparative Layer-Wise Analysis of Self-Supervised Speech Models. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, June 2023. doi: 10.1109/ICASSP49357.2023.10096149.

Josef P. Rauschecker and Sophie K. Scott. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature Neuroscience*, 12(6):718–724, June 2009. ISSN 1546-1726. doi: 10.1038/nn.2331.

Fabian Ritter-Gutierrez, Kuan-Po Huang, Dianwen Ng, Jeremy H.M. Wong, Hung-Yi Lee, Eng Siong Chng, and Nancy F. Chen. Noise Robust Distillation of Self-Supervised Speech Models via Correlation Metrics. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pp. 495–499, April 2024. doi: 10.1109/ICASSPW62465.2024.10627241.

Xiaobin Rong, Qinwen Hu, Mansur Yesilbursa, Kamil Wojcicki, and Jing Lu. PASE: Leveraging the Phonological Prior of WavLM for Low-Hallucination Generative Speech Enhancement. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(39):32826–32834, March 2026a. ISSN 2374-3468. doi: 10.1609/aaai.v40i39.40562.

Yiming Rong, Yixin Zhang, Ziyi Wang, Deyang Jiang, Yunlong Zhao, Haoran Wu, Shiyu Zhou, and Bo Xu. Speech-Aware Long Context Pruning and Integration for Contextualized Automatic Speech Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40 (39):32835–32842, March 2026b. ISSN 2374-3468. doi: 10.1609/aaai.v40i39.40563.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization, April 2020. arXiv:1911.08731 [cs.LG].

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4080–4090, Red Hook, NY, USA, December 2017. Curran Associates Inc. ISBN 978-1-5108-6096-4. URL https://dl.acm.org/doi/10.5555/3294996.3295163.

Raphael Tang and Jimmy Lin. Deep Residual Learning for Small-Footprint Keyword Spotting. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5484–5488, April 2018. doi: 10.1109/ICASSP.2018.8462688.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization, 2016. arXiv:1607.08022.

Zhenyu Wang, Shuyu Kong, Li Wan, Biqiao Zhang, Yiteng Huang, Mumin Jin, Ming Sun, Xin Lei, and Zhaojun Yang. Query-by-Example Keyword Spotting Using Spectral-Temporal Graph Attentive Pooling and Multi-Task Learning. In *Interspeech 2024*, pp. 322–326, 2024. doi: 10.21437/Interspeech.2024-100.

Yang Xiao, Tianyi Peng, Yanghao Zhou, and Rohan Kumar Das. AdaKWS: Towards Robust Keyword Spotting with Test-Time Adaptation. In *Interspeech 2025*, pp. 5408–5412. ISCA, August 2025. doi: 10.21437/Interspeech.2025-15.

Junming Yuan, Ying Shi, LanTian Li, Dong Wang, and Askar Hamdulla. Few-Shot Keyword Spotting from Mixed Speech. In *Interspeech 2024*, pp. 5063–5067, 2024. doi: 10.21437/Interspeech.2024-2296.

Yundong Zhang, Naveen Suda, Liangzhen Lai, and Vikas Chandra. Hello Edge: Keyword spotting on microcontrollers, 2017. arXiv:1711.07128.

Hai Zhu, Xin Wang, Kun Wang, and Huayi Zhan. Temporal Convolution Shrinkage Network for Keyword Spotting. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10161–10165, April 2024. doi: 10.1109/ICASSP48485.2024.10446268.

Juan Zuluaga-Gomez, Petr Motlicek, Qingran Zhan, Karel Veselý, and Rudolf Braun. Automatic Speech Recognition Benchmark for Air-Traffic Communications. In *Interspeech 2020*, pp. 2297–2301. ISCA, October 2020. doi: 10.21437/Interspeech.2020-2173.

Juan Zuluaga-Gomez, Iuliia Nigmatulina, Amrutha Prasad, Petr Motlicek, Karel Veselý, Martin Kocour, and Igor Szöke. Contextual Semi-Supervised Learning: An Approach to Leverage Air-Surveillance and Untranscribed ATC Data in ASR Systems. In *Interspeech 2021*, pp. 3296–3300. ISCA, August 2021. doi: 10.21437/Interspeech.2021-1373.

Juan Zuluaga-Gomez, Karel Veselý, Igor Szöke, Alexander Blatt, Petr Motlicek, Martin Kocour, Mickael Rigault, Khalid Choukri, Amrutha Prasad, Seyyed Saeed Sarfjoo, Iuliia Nigmatulina, Claudia Cevenini, Pavel Kolčárek, Allan Tart, Jan Černocký, and Dietrich Klakow. ATCO2 corpus: A Large-Scale Dataset for Research on Automatic Speech Recognition and Natural Language Understanding of Air Traffic Control Communications, June 2023. arXiv:2211.04054 [cs.CL].

Luboš Šmídl, Jan Švec, Daniel Tihelka, Jindřich Matoušek, Jan Romportl, and Pavel Ircing. Air traffic control communication (ATCC) speech corpora and their use for ASR and TTS development. *Language Resources and Evaluation*, 53(3):449–464, September 2019. ISSN 1574-0218. doi: 10.1007/s10579-019-09449-5.

Preserve, Then Bound: Conditionally Informative Acoustic Context for Few-Shot ATC Call-Sign Spotting

Supplementary Material

CONTENTS

1 Introduction **1**

2 Related Work **2**

3 Method **3**

3.1 Brain-Inspired Dual-Stream Representation 3

3.2 Prototype–Memory Matching 4

3.3 Lexically Anchored Bounded Fusion 5

3.4 Training Objectives 5

4 Experiments **6**

4.1 Experimental Setup 6

4.2 Main Results 6

4.3 Analysis 7

5 Conclusion **9**

A Additional Ablation Studies **17**

A.1 Matched Content Control Across Enrollment Settings 17

A.2 Evidence-Specific Layer Aggregation 17

A.3 Auxiliary Context Supervision 17

A.4 Context Support Representation 17

B Source-Relation Stress Test **20**

B.1 Diagnostic Construction 20

B.2 Full Relation-Balanced Results 20

B.3 Paired Differences Against the Content Control 20

B.4 Natural vs. Controlled Relation Distribution 20

C Contextual Applicability and Utilization **22**

C.1 Class-Controlled Applicability 22

C.2 Available vs. Realized Contextual Gain 22

C.3 Context-Memory Intervention 22

C.4 Allocation Across Protocols 23

C.5 Preservation Analysis 23

D Acoustic Perturbation Settings **24**

E Case Study **25**

A Additional Ablation Studies

This section reports additional ablations that complement the core component analysis in the main paper. We focus on representation choices that are directly related to the separation of lexical and contextual evidence but are omitted from the main text for space.

A.1 Matched Content Control Across Enrollment Settings

Table 4 compares lexical-only and context-preserving models on the fixed test episodes under both enrollment settings, and Figure 4 expresses the 5-shot comparison relative to the Content-only model.

On ATCO2 1-shot, Content only achieves 94.00%, whereas BioS2D obtains 93.87%. With a single support instance per class the context memory reduces to one embedding, so the instance-level design has no multiple supports to preserve; the contextual analysis in the main text is accordingly conducted at 5 shots.

A.2 Evidence-Specific Layer Aggregation

We further examine whether the lexical and contextual pathways benefit from different pretrained representation ranges. Table 5 compares pathway-specific aggregation and swapped pathway ranges against the Content only control.

Swapping the pathway ranges degrades performance in most settings, with the largest reduction observed on UWB-ATCC under 10-dB Gaussian noise (90.89% against 96.30%). This result supports the functional asymmetry between the two evidence pathways, although it does not imply that one particular layer range is universally optimal.

A.3 Auxiliary Context Supervision

We additionally examine the auxiliary supervision used by the context pathway. *w/o context SupCon* removes the supervised contrastive (SupCon) loss $\mathcal{L}_{\mathrm{ctx}}$; *Fixed, w/o source-context supervision* additionally fixes the applicability factor, so no source-context identifier is used in training; and the *Recording-context variant* defines the source-context identifier at recording level rather than by speaker and one-hour window.

Fixed-episode accuracy is largely insensitive to removing the auxiliary supervision, staying within 0.15 points of BioS2D. In contrast, the relation-balanced diagnostic is more sensitive: macro accuracy falls from 98.03% to 97.59–97.87%, and removing all source-context supervision doubles the share of decisions a mismatched memory changes (1.53% against 0.74%).

A.4 Context Support Representation

Table 7 compares class-mean context prototypes with the instance-preserving context memory.

The two operators are identical in the 1-shot setting because each candidate contains only one support instance. Their difference appears when multiple support utterances are available.

The comparable performance in the 5-shot setting indicates that the memory should not be interpreted as an additional high-capacity classifier. Its purpose is instead to avoid imposing a class-level centroid assumption on contextual evidence whose relation may differ across support instances.

Table 4: Comparison of lexical-only and context-preserving models on the fixed test episodes. Accuracy (%) is the mean with the training-seed SD in small type. Best value per column in bold; the proposed model is shaded.

	ATCO2		UWB-ATCC	
Model	1-shot	5-shot	1-shot	5-shot
Independently trained content control				
Content only	**94.00** ±0.23	97.54 ±0.19	96.27 ±0.52	**98.52** ±0.17
Pathway assignment				
Swapped layer ranges	–	97.64 ±0.36	–	97.47 ±0.69
BioS2D	93.87 ±0.22	**97.77** ±0.16	**96.47** ±0.10	98.49 ±0.32

– marks a variant trained at 5-shot only.

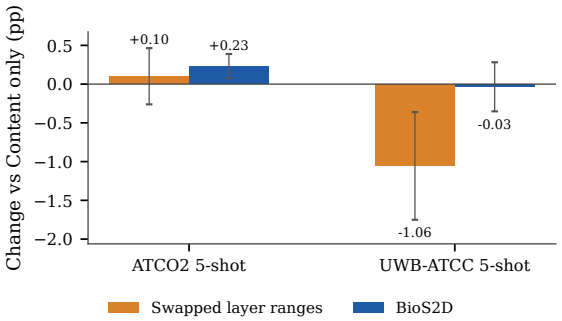

Figure 4: Conditional utility of contextual evidence. Each bar reports the 5-shot accuracy change relative to the matched Content-only model on the same corpus. Positive and negative changes across conditions illustrate that contextual information can be predictive without being uniformly reliable.

Table 5: Ablation of evidence-specific WavLM layer aggregation, 5-shot. Accuracy (%) on the fixed test episodes (clean and with 10-dB Gaussian noise on the query), macro accuracy on the relation-balanced ATCO2 diagnostic, and accuracy on the 10-way nearest-hard ATCO2 diagnostic. Means with the training-seed SD in small type; best value per column in bold.

	ATCO2	UWB-ATCC		ATCO2 diagnostics	
Variant	Clean	Clean	Gaussian 10 dB	Relation macro	10-way
Independently trained content control					
Content only	97.54 ±0.19	**98.52** ±0.17	96.20	97.56 ±0.25	**92.01** ±1.67
Pathway assignment					
Swapped layer ranges	97.64 ±0.36	97.47 ±0.69	90.89	97.31 ±0.99	91.49 ±3.94
BioS2D	**97.77** ±0.16	98.49 ±0.32	**96.30**	**98.03** ±0.42	91.15 ±0.52

Table 6: Auxiliary context supervision ablations, 5-shot. Accuracy (%) on the fixed test episodes, macro and Lexical-hard Misleading accuracy (%) on the relation-balanced ATCO2 diagnostic, and the share (%) of decisions changed under memory reassignment. Means with the training-seed SD in small type; best accuracy per column in bold.

	Fixed episodes		Relation diagnostic		
Variant	ATCO2	UWB-ATCC	Macro	Misleading	Changed
Mechanism ablations					
w/o context SupCon	**97.77** ±0.14	98.50 ±0.27	97.71 ±0.33	95.58 ±0.26	0.75 ±0.22
Fixed, w/o source-context supervision	97.62 ±0.18	**98.59** ±0.17	97.59 ±0.39	95.58 ±1.70	1.53 ±0.20
Recording-context variant	97.75 ±0.16	–	97.87 ±0.44	96.28 ±0.83	0.91 ±0.11
BioS2D	**97.77** ±0.16	98.49 ±0.32	**98.03** ±0.42	**96.36** ±0.52	0.74 ±0.07

– marks a variant not trained in that cohort.

Table 7: Comparison of contextual support-set representations, 5-shot; at 1-shot the two operators coincide. Accuracy (%) on the fixed test episodes, macro and Lexical-hard Misleading accuracy (%) on the relation-balanced ATCO2 diagnostic, and the share (%) of decisions changed under memory reassignment. Means with the training-seed SD in small type.

	Fixed episodes		Relation diagnostic		
Context operator	ATCO2	UWB-ATCC	Macro	Misleading	Changed
Mechanism ablations					
Mean context prototype	97.75 ±0.21	**98.53** ±0.29	97.88 ±0.38	96.10 ±0.69	0.76 ±0.21
BioS2D	**97.77** ±0.16	98.49 ±0.32	**98.03** ±0.42	**96.36** ±0.52	0.74 ±0.07

B SOURCE-RELATION STRESS TEST

B.1 DIAGNOSTIC CONSTRUCTION

Natural test episodes contain relatively few cases in which contextual similarity actively favors an incorrect candidate. We therefore construct a relation-balanced diagnostic to stress-test contextual shortcut behavior.

For a query q and candidate support set S_k, a candidate is regarded as source matched when at least one support utterance shares the query source:

$$\exists x \in S_k : \; c(x) = c(q).$$

We define four relations according to whether this condition holds for the correct class and for a competitor:

We combine these four relations with two candidate-selection strategies. *Random* draws competitors from the eligible pool, whereas *Lexical-hard* draws them from the eight classes with the smallest normalized string edit distance to the query call sign. The top-8 rule is selected on validation data and fixed before test evaluation.

This produces eight equally weighted diagnostic cells. In the 5-shot setting, each cell contains 77 focal queries per manifest; in the 1-shot setting, each cell contains 65.

B.2 FULL RELATION-BALANCED RESULTS

Table 9 reports the complete 5-shot diagnostic.

BioS2D obtains a macro accuracy of 98.03%, compared with 97.56% for Content only. The largest descriptive improvements over Content only occur in the Lexical-hard Conflicting and Unmatched cells.

More importantly, deliberately amplifying Misleading relations does not induce a substantial collapse. In the Lexical-hard Misleading condition, BioS2D achieves 96.36%, compared with 96.02% for the content control. We therefore interpret this experiment as evidence of limited contextual harm rather than precise source-relation recovery.

B.3 PAIRED DIFFERENCES AGAINST THE CONTENT CONTROL

Table 10 compares BioS2D with Content only on identical relation-balanced episodes, clean and under the six query-side perturbations. The paired difference is positive under all seven conditions, from $+0.22$ to $+0.80$ percentage points, and the 95% interval excludes zero under 20% band masking. Candidate pairs within an episode and recurring queries across manifests are not independent, so the intervals resample training seeds and joint source–call-sign clusters.

B.4 NATURAL VS. CONTROLLED RELATION DISTRIBUTION

The balanced diagnostic deliberately over-represents difficult relations. In the random ATCO2 5-shot test episodes, Misleading and Conflicting relations together account for 2.66% of queries (332 of 12,500 per training seed), and in the classification stream used for training, Misleading relations account for 0.47% of sampled queries.

By contrast, the auxiliary source-supervision sampler increases the frequency of Misleading relations to 20.74% of its queries. This difference explains why average natural-test accuracy provides limited evidence about shortcut behavior and motivates the relation-balanced evaluation.

Table 8: Source-relation definitions used in the diagnostic.

Relation	Correct	Competitor	Interpretation
Helpful	Match	No match	Supports correct class
Misleading	No match	Match	Supports wrong class
Conflicting	Match	Match	Supports both sides
Unmatched	No match	No match	No usable relation

Table 9: Accuracy (%) in the eight cells of the relation-balanced ATCO2 diagnostic, 5-shot. Lexical-hard competitors are the query's nearest lexical neighbours and are listed first. Each cell holds 77 focal queries in each of five manifests. Best value per column in bold.

	Lexical-hard competitors				Random competitors			
Model	H	M	C	U	H	M	C	U
Spectral few-shot references†								
BCResNet	94.29†	93.25†	96.10†	91.43†	97.66†	96.62†	98.96†	96.10†
DSCNN	79.48†	79.22†	84.68†	75.58†	87.79†	87.27†	91.17†	82.86†
ECAPA-TDNN	93.25†	94.81†	97.40†	88.31†	96.88†	94.81†	98.96†	97.14†
SS-Flex	94.29†	94.81†	96.88†	91.43†	97.92†	98.44†	98.70†	96.62†
TCSN	93.51†	90.13†	92.47†	88.05†	96.10†	94.29†	98.70†	94.55†
WavLM references†								
WavLM last	75.58†	64.16†	65.45†	69.61†	80.00†	72.99†	77.66†	74.03†
WavLM weighted	69.61†	61.04†	65.71†	66.49†	77.66†	67.79†	75.84†	71.69†
ResNet15	88.57†	81.04†	85.45†	83.38†	94.29†	89.35†	94.55†	91.43†
Independently trained content control								
Content only	98.53 ±0.83	96.02 ±0.75	96.80 ±0.40	94.46 ±2.08	99.13 ±0.40	98.53 ±0.83	98.96 ±0.26	98.10 ±0.15
Pathway assignment								
Swapped layer ranges	97.84 ±0.40	95.24 ±1.82	96.80 ±2.08	94.89 ±0.83	98.96 ±0.52	97.66 ±1.70	98.96 ±0.78	98.10 ±0.40
Mechanism ablations								
Mean context prototype	98.79 ±0.30	96.10 ±0.69	97.23 ±0.54	95.15 ±0.54	**99.48** ±0.45	98.35 ±0.40	**99.39** ±0.15	98.53 ±0.75
w/o context SupCon	98.53 ±0.54	95.58 ±0.26	97.14 ±0.52	94.55 ±0.90	99.39 ±0.40	98.96 ±0.78	99.31 ±0.15	98.18 ±0.94
Fixed applicability	98.87 ±0.40	**96.71** ±0.91	97.14 ±1.19	94.46 ±1.67	99.39 ±0.30	98.27 ±0.60	**99.39** ±0.30	98.10 ±0.30
Fixed, w/o source loss	98.70 ±0.26	95.58 ±1.70	96.19 ±1.05	93.77 ±0.78	99.39 ±0.15	**99.31** ±0.75	**99.39** ±0.15	98.35 ±0.54
w/o lexical uncertainty	98.87 ±0.60	95.76 ±0.75	97.14 ±0.52	94.89 ±1.08	**99.48** ±0.00	98.53 ±0.91	**99.39** ±0.15	98.44 ±0.69
Recording-context variant	98.53 ±0.54	96.28 ±0.83	97.23 ±0.40	95.06 ±1.13	99.39 ±0.40	98.79 ±1.05	**99.39** ±0.15	98.27 ±0.79
BioS2D	**98.96** ±0.26	96.36 ±0.52	**97.49** ±0.30	**95.32** ±1.13	**99.48** ±0.45	98.61 ±0.60	**99.39** ±0.15	**98.61** ±0.83

H/M/C/U: the correct class, a competitor, both, or neither shares the query's source. † One registered training seed.

Table 10: Paired BioS2D-minus-Content-only macro accuracy (pp) on the relation-balanced ATCO2 5-shot diagnostic, clean and under the six query-side perturbations. Intervals are 95% hierarchical bootstrap intervals over training seeds and joint source–call-sign clusters; p is two-sided, and Holm-adjusted over the prespecified family of 14 5-shot comparisons.

Query condition	Δ	95% interval	p	Holm p
Clean	$+0.47$	$[-0.09, +1.05]$	0.101	1.000
Gaussian 20 dB	$+0.40$	$[-0.19, +0.98]$	0.167	1.000
Gaussian 10 dB	$+0.49$	$[-0.06, +1.12]$	0.081	0.967
Band mask 10%	$+0.44$	$[-0.03, +0.93]$	0.065	0.845
Band mask 20%	$+0.80$	$[+0.26, +1.32]$	0.005	0.073
Low-pass 3.4 kHz	$+0.22$	$[-0.47, +0.85]$	0.466	1.000
Low-pass 2.5 kHz	$+0.23$	$[-0.56, +1.02]$	0.535	1.000

C Contextual Applicability and Utilization

C.1 Class-Controlled Applicability

The auxiliary applicability target combines class correctness and source agreement. As a result, a model can obtain a high aggregate applicability score partly by exploiting class compatibility.

We therefore repeat the analysis after controlling the candidate class role.

Among same-source candidates the task reduces to class correctness, which both scores resolve almost perfectly (0.980 and 0.977 AUROC). Once class correctness is controlled by comparing the correct candidate across source relations, however, applicability drops to 0.589 AUROC, against 0.550 for the lexical score. The result indicates that source-related information is present, but precise source-context matching is substantially more difficult to recover.

Figure 5 further separates contextual interventions into corrections and harms. Under Random competitors, both are extremely rare, indicating that context has little effect when lexical matching already resolves most episodes. Under Lexical-hard competitors, interventions become more frequent. Helpful relations yield more corrections than harms, whereas Misleading and Conflicting relations expose occasional harmful interventions.

This result is consistent with the controlled applicability analysis: the model has learned source-related information but does not recover source-context matching perfectly. Importantly, even in these deliberately stress-tested conditions, corrected and harmed predictions each account for less than approximately 0.6% of a cell. Thus, the role of the bounded allocator is not to eliminate every contextual error, but to restrict the decision-level consequences of an imperfect applicability signal.

C.2 Available vs. Realized Contextual Gain

We further separate the amount of contextual information available to the model from the amount actually used by the learned residual.

On the relation-balanced 5-shot diagnostic, lexical-only prediction reaches 97.998% macro accuracy, whereas the complete contextual residual reaches 98.030%, corresponding to a realized gain of only $+0.032$ percentage points.

In contrast, on the 10-way diagnostic, where the realized gain is -0.09 percentage points, a zero-harm oracle that applies contextual correction only when it does not destroy a correct lexical decision reveals approximately $+3.82$ percentage points of available headroom.

The large oracle–realized gap suggests that information availability is not the primary bottleneck; reliable applicability estimation is.

Figure 6 localizes the oracle–realized gap by lexical difficulty. The available contextual headroom is strongly concentrated in low-margin queries. For $m_q < 0.05$, the unbounded oracle provides roughly 20 percentage points of potential gain, and most of this gain remains available even when the intervention is restricted by the BioS2D bound.

This distinction is important. The small realized gain cannot be explained primarily by an overly restrictive residual budget: the bounded oracle retains substantial capacity precisely where lexical evidence is most ambiguous. Instead, BioS2D fails to consistently select the beneficial contextual intervention within that budget. The result therefore identifies applicability estimation, rather than bounded authority itself, as the main remaining limitation.

C.3 Context-Memory Intervention

To determine whether the context memory actually contributes to prediction, we cyclically reassign candidate memories to incorrect classes while leaving lexical scores unchanged.

On the relation-balanced 5-shot diagnostic, the intervention changes 0.74% of final predictions, while the Content only control is invariant by construction. Thus, contextual memory is used by the classifier, but its influence is concentrated in a small subset of close lexical decisions.

Table 11: Applicability discrimination before and after controlling for class compatibility, on Lexical-hard episodes of the relation-balanced ATCO2 diagnostic (5-shot). AUROC averaged over three training seeds.

Diagnostic	Applicability	Lexical score
Class correctness among same-source candidates	0.980	0.977
Source relation within the correct candidate	0.589	0.550

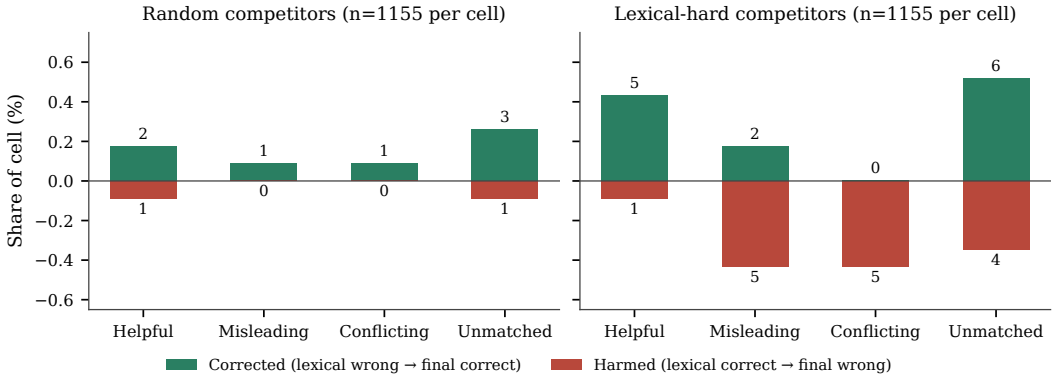

Figure 5: Prediction corrections and harms induced by contextual allocation under the relation-balanced diagnostic. Positive bars denote cases where the lexical prediction is incorrect but the final BioS2D prediction is correct; negative bars denote the reverse transition. Numbers indicate the corresponding query counts. Contextual intervention is rare under random competitors and becomes more active under lexically hard candidates. Misleading and Conflicting relations can induce harmful interventions, confirming that contextual applicability is imperfect, while their absolute frequency remains small.

C.4 Allocation Across Protocols

Figure 7 repeats the margin analysis of Figure 3 on every protocol whose released predictions store the contextual residual: the 10-way nearest-hard and random diagnostics, the relation-balanced diagnostic with clean and noisy queries, and UWB-ATCC transfer. The residual shrinks monotonically with the lexical margin in eleven of the twelve protocol–shot combinations; the exception, UWB-ATCC transfer at 5 shots, has only 12 queries below $m_q = 0.05$. In every combination, decisions change only for $m_q < 0.05$. The allocation pattern is therefore a property of the bounded residual rather than of a single evaluation set.

C.5 Preservation Analysis

For the bounded residual,

$$|\Delta_{q,k}| \leq \alpha u_q.$$

Let m_q denote the lexical top-1 margin. If

$$m_q > 2\alpha u_q,$$

the contextual residual cannot change the lexical top-1 prediction.

In the clean 5-shot relation-balanced evaluation, approximately 94.3% of queries satisfy this sufficient condition (80.0% on the 10-way diagnostic). The high coverage explains why misleading context produces little aggregate damage, while also limiting the amount of contextual information that can affect already confident decisions.

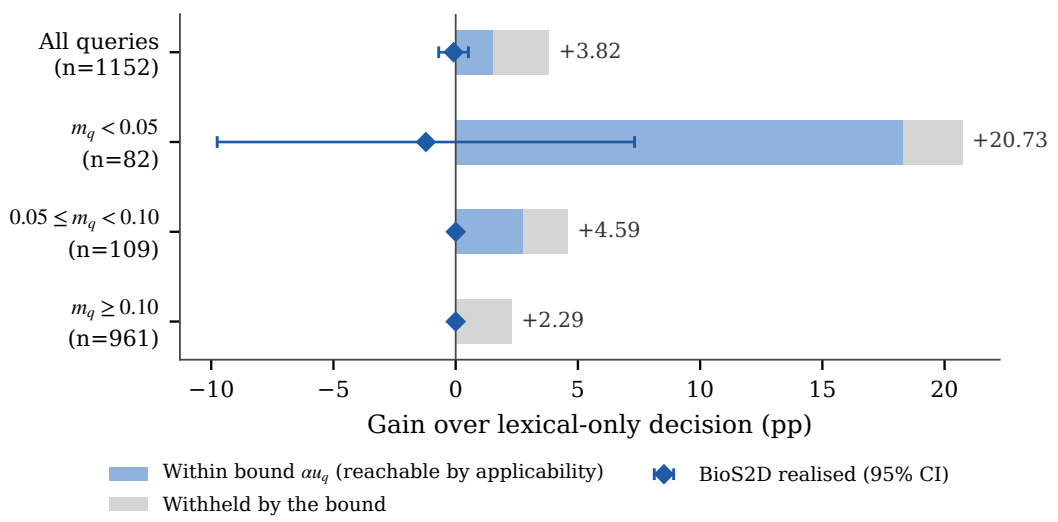

Figure 6: Available and realized contextual gains as a function of lexical margin. Queries are grouped by lexical top-1 margin m_q. The unbounded oracle measures the potential gain when contextual scores can be modified without the BioS2D residual constraint, while the bounded oracle is restricted by the same αu_q budget used by BioS2D. Most contextual headroom is concentrated in highly ambiguous lexical decisions. The small gap between the two oracle variants in the lowest-margin region shows that the residual bound retains substantial correction capacity, whereas the much smaller realized gain identifies applicability estimation as the dominant bottleneck.

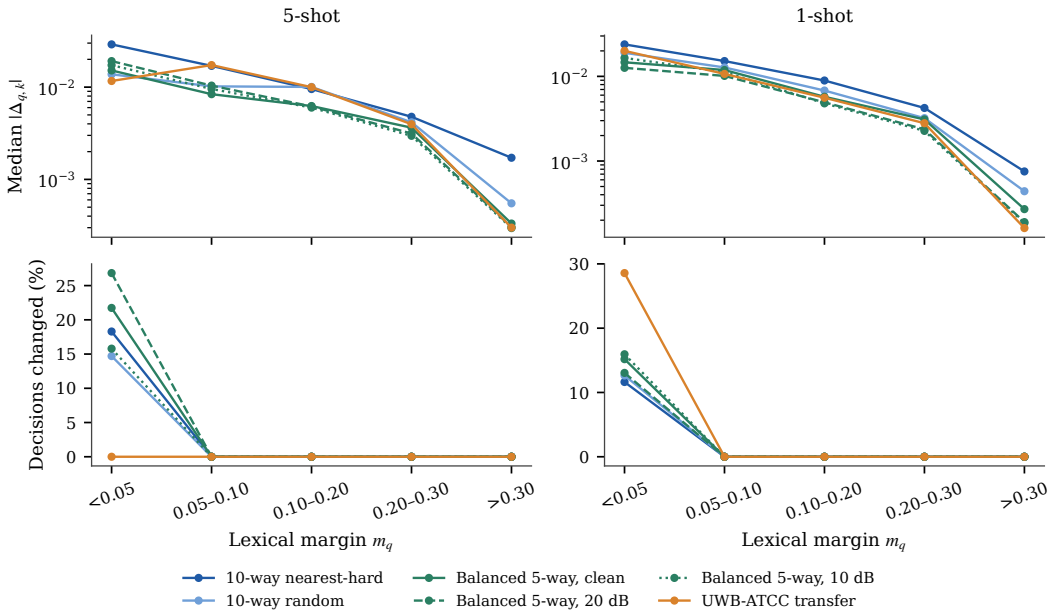

Figure 7: Residual magnitude (top, log scale) and share of decisions changed (bottom) against the lexical margin m_q, for six protocols at 5 shots (left) and 1 shot (right). Each curve pools three training seeds.

D Acoustic Perturbation Settings

We evaluate query-only acoustic degradation while keeping all support utterances clean. Perturbations are applied to the waveform before feature extraction, and matched models receive the same perturbation realization.

We consider: (i) Gaussian noise at 20 and 10 dB, (ii) 10% and 20% band masking with 30-dB attenuation, and (iii) low-pass filtering at 3.4 and 2.5 kHz.

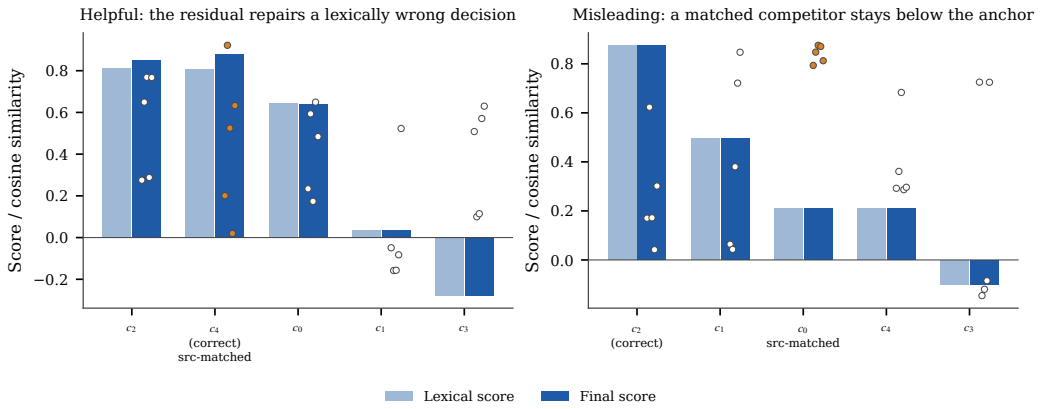

Figure 8: Detailed case study of contextual allocation. Left: a Helpful episode in which context resolves an ambiguous lexical decision. Right: a Misleading episode in which strong contextual similarity for an incorrect candidate is prevented from overriding the lexical anchor. For each candidate, we show the lexical score, individual query–support context similarities (filled for the source-matched candidate), and final score.

E CASE STUDY

Figure 8 presents two representative episodes from the relation-balanced test set.

In the Helpful example, the lexical scores of the leading candidates are close. One support instance of the correct class exhibits a strong contextual match to the query, producing a positive residual that resolves the ambiguous lexical comparison.

In the Misleading example, an incorrect candidate has the strongest contextual match. However, the lexical decision is sufficiently confident that the bounded residual cannot overturn the correct prediction. This example illustrates why contextual similarity alone should not be interpreted as class identity.

These examples illustrate the intended role of contextual evidence in BioS2D: context can alter a close lexical comparison when useful relational evidence is available, but it is prevented from acting as an unrestricted alternative definition of call-sign identity.