PRIOR-CENTRIC SUPPORT ALLOCATION FOR FEW-SHOT AUDIO RECOGNITION WITH AUDIO-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-shot audio recognition based on audio-language models requires the collaboration of transferable semantic priors derived from class texts and task-specific evidence from support sets. Nevertheless, existing methods tend to over-emphasize support memories or matching pathways, resulting in the entanglement of sparse support-set evidence with already strong semantic priors and leaving the adaptation mechanism insufficiently explicit. To address this issue, this paper formulates few-shot audio adaptation as a prior-centric support allocation problem, and accordingly proposes a Prior-Centric Support Allocation (PCSA) framework, which treats the support set as a control signal for episode-level inference rather than as a competing memory pathway. The framework consists of two coupled components: Prior-Centric Semantic Scaffold (PriorCore) and Support-Aware Transductive Refinement (SATR). Specifically, PriorCore constructs a stable cross-granular semantic scaffold from coarse and fine priors, while SATR converts support evidence into reliability signals and uses them to regulate transductive refinement over the current episode. Furthermore, this paper designs an explicit reliability-controlled prediction rule that delineates the functional boundaries between semantic priors and support evidence, making the division of labor explicit and analyzable. Extensive experiments on several few-shot audio recognition benchmarks illustrate that our method achieves strong performance under the adopted evaluation protocols, and comprehensive ablation and diagnostic analyses further confirm the effectiveness and robustness of the proposed framework.

1 INTRODUCTION

Few-shot audio classification aims to recognize unseen sound categories from only a handful of labeled support examples. In conventional supervised audio recognition, strong backbones such as PANNs (Kong et al., 2020), AST (Gong et al., 2021), HTS-AT (Chen et al., 2022), BEATs (Chen et al., 2023), and CAT (Liu et al., 2023) have substantially improved closed-set sound classification. However, these systems usually rely on abundant labels and fixed class vocabularies, which makes direct transfer to low-resource open-set scenarios difficult. Classical few-shot audio recognition therefore turns to episodic metric learning and constructs support-query matching rules from prototypes or pairwise similarities (Vinyals et al., 2016; Snell et al., 2017; Chou et al., 2019). Yet environmental sounds are temporally variable, often corrupted by background interference, and frequently semantically adjacent, so a few support examples do not always define a stable class boundary. In a frozen audio-language model, however, class text supplies a stable semantic anchor even when the local spectrogram geometry is noisy, so support evidence is best used to correct the current episode rather than to replace that anchor.

Audio-language pretraining changes this landscape. AudioCLIP (Guzhov et al., 2022), CLAP (Elizalde et al., 2023), and related language-audio pretraining methods (Wu et al., 2023; Elizalde et al., 2024) make it possible to inject open-vocabulary text semantics into downstream audio recognition. As a result, few-shot audio recognition is no longer governed only by support-query similarity. Each episode now contains two distinct sources of evidence: transferable semantic priors from class text and task-specific evidence from the support set. Few-shot audio adaptation thus hinges on coordinating these two sources, rather than relying on supporting evidence alone.

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.

This shift raises a more fundamental question. Once the semantic prior is already strong, what role should support examples still play? A common answer in recent work is to strengthen support adaptation by introducing more cache-style memories, finer local correspondences, or more elaborate support interactions. A representative published reference along this line is a multi-grained correspondence learner (Zhao et al., 2024). That direction is clearly effective, but it also points to a conceptual tension in current few-shot audio adaptation.

When several support mechanisms are stacked together, semantic priors, support memories, and support correction become entangled, making it difficult to identify what the support set is actually contributing. The method may become stronger, but the division of labor between prior semantics and sparse support evidence becomes less clear. For few-shot audio recognition, this matters alongside the final score: a useful method should not only perform well, but should also clarify how support evidence ought to be used once a reliable prior scaffold is already available.

In this paper, we approach the problem from the viewpoint of *prior-centric support allocation*. Instead of asking how to add another support branch, we ask how support evidence should be allocated after a prior-centric scaffold has already been formed. This perspective leads to *Prior-Centric Support Allocation* (PCSA). Conceptually, PCSA has two coupled components. *Prior-Centric Semantic Scaffold* (PriorCore) preserves a stable cross-granular semantic scaffold built from coarse and fine priors, while *Support-Aware Transductive Refinement* (SATR) converts support evidence into an explicit reliability signal and uses it to regulate episode-level transductive refinement. Figure 1 visualizes the resulting inference. The key idea is that support examples need not become another memory bank; instead, they can determine how strongly the model should refine the current episode.

The reported method therefore keeps the semantic scaffold explicit, estimates support reliability from that scaffold, and uses the resulting signal to regulate episode refinement. This choice keeps the paper centered on a single thesis rather than on stacking additional support-side machinery.

As summarized in Figure 1 and detailed in Section 3, this perspective yields a two-component method with a three-stage operational view: PriorCore first establishes the semantic organization of the episode, SATR then carries out support-aware transductive refinement over the current episode, and the control-and-final-prediction stage combines the propagated and pre-propagation beliefs.

Our main contributions can be summarized as follows:

- We formulate few-shot audio adaptation as a *prior-centric support allocation* problem and argue that, once strong semantic priors are available, the support set should primarily serve as a control signal for episode inference.
- We propose *Prior-Centric Support Allocation* (PCSA), composed of PriorCore and SATR. PriorCore supplies a stable cross-granular semantic scaffold, SATR estimates support reliability and performs transductive refinement, and the resulting reliability-controlled final prediction rule is written explicitly for interpretability and analysis, yielding a simple division of labor between prior semantics and support evidence.
- We conduct experiments on ESC50, UrbanSound8K, and FSDKaggle2018, where PCSA consistently improves the representative published references under the adopted comparison protocol, and extensive ablations and diagnostics verify the effectiveness of the proposed design.

2 Related Work

2.1 Few-Shot Audio Recognition

Few-shot recognition originally developed around episodic metric learning, where support examples define either pairwise references or class prototypes (Vinyals et al., 2016; Snell et al., 2017). In audio, this line includes attentional similarity for transient sound events (Chou et al., 2019), analyses of few-shot protocol design (Wang et al., 2021), and benchmark construction for systematic evaluation (Heggan et al., 2022). Recent journal work further extends low-shot audio learning to neighboring regimes such as few-shot class-incremental audio classification (Li et al., 2023b), few-shot speaker identification (Li et al., 2023c), and multi-label few-shot audio tagging (Papaioannou et al., 2025). Related studies on acoustic-scene/event analysis, multi-task acoustic event detection, and label-efficient sound event detection (Stowell et al., 2015; Xia et al., 2019; Park et al., 2022)

highlight the same practical difficulties: overlapping events, noisy backgrounds, and limited annotation budgets. Taken together, this literature suggests that audio few-shot recognition is difficult not only because the support set is small, but also because environmental sounds exhibit strong temporal variability, background interference, and class overlap. These properties make the role of the support set fundamentally different from that in more static visual domains.

The classical literature also suggests what a strong audio few-shot method should ideally provide: stable class-level structure under sparse supervision, robustness to within-class variation, and effective use of episode geometry rather than independent query prediction. Our work inherits these goals, but re-examines them in the audio-language setting, where a strong semantic prior is already available before episodic adaptation even begins.

2.2 Audio-Language Adaptation

Audio-language pretraining has moved few-shot audio transfer away from purely metric backbones toward language-guided adaptation. Even before CLAP-style transfer, zero-shot audio classification via semantic embeddings (Xie & Virtanen, 2021) showed that textual labels and descriptions can provide useful side information for unseen sound classes. CLAP-style models (Elizalde et al., 2024; Wu et al., 2023; Elizalde et al., 2023) further strengthen this direction by providing transferable text-audio semantics, while Tip-Adapter, CLIP-Adapter, and related vision-language methods (Zhang et al., 2022; Gao et al., 2024; Zhou et al., 2022) show that pretrained priors can be strengthened with lightweight caches or prompts. More recent journal work also studies prompt-style extension of frozen multimodal backbones to audio (Liang et al., 2025). Treff-adapter (Liang et al., 2023) translated the lightweight adaptation recipe to the audio domain, and a recent multi-grained correspondence learner (Zhao et al., 2024) further developed a richer cross-granular correspondence pipeline.

This transfer trend is also enabled by broader audio resources and descriptive supervision. AudioSet (Gemmeke et al., 2017) expanded the event ontology used by later benchmarks and transfer studies, WavCaps (Mei et al., 2024) enriched caption-style supervision for audio-language modeling, and low-label audio adaptation has continued to be tested in neighboring settings such as stochastic-classifier class-incremental learning (Li et al., 2023a).

Most methods in this line emphasize richer correspondence modeling or richer support adaptation. Our work is positioned on top of the same audio-language foundation, but asks a different question: once a strong prior scaffold already exists, should support evidence still be accumulated as another memory mechanism, or should it mainly control how episode inference is performed? This difference in viewpoint is central to the method design in this paper.

2.3 Transductive Few-Shot Adaptation of Language-Guided Models

The same reliability-oriented concern has developed in parallel for vision-language models. Transductive zero-shot and few-shot CLIP (Martin et al., 2024), the transductive booster (Zanella et al., 2024), and the later transductive few-shot CLIP family (Silva-Rodríguez et al., 2024; Baklouti et al., 2025) exploit the unlabeled query set at test time while preserving the pretrained text prior, and information maximization for few-shot learning (Boudiaf et al., 2020) established that query-set structure can be exploited without changing the backbone. PCSA inherits this transductive perspective but applies it in the audio-language setting, where the semantic prior is strong while the local graph geometry is often corrupted by transient events, background noise, and class overlap. The contribution is therefore not a new transductive objective but a reliability-conditioned rule that decides how much of the already-computed transductive refinement should influence the base prior prediction.

2.4 Structured Episode Inference

Structured matching and transductive inference are natural tools for few-shot recognition. Sinkhorn regularization (Cuturi, 2013) makes optimal transport practical at scale, structured classifiers such as DeepEMD (Zhang et al., 2020) show that local structured comparison can be highly effective, graph-based transductive methods such as TPN (Liu et al., 2019) and LaplacianShot (Ziko et al., 2020) demonstrate the value of episode-level label propagation. These ideas are particularly relevant

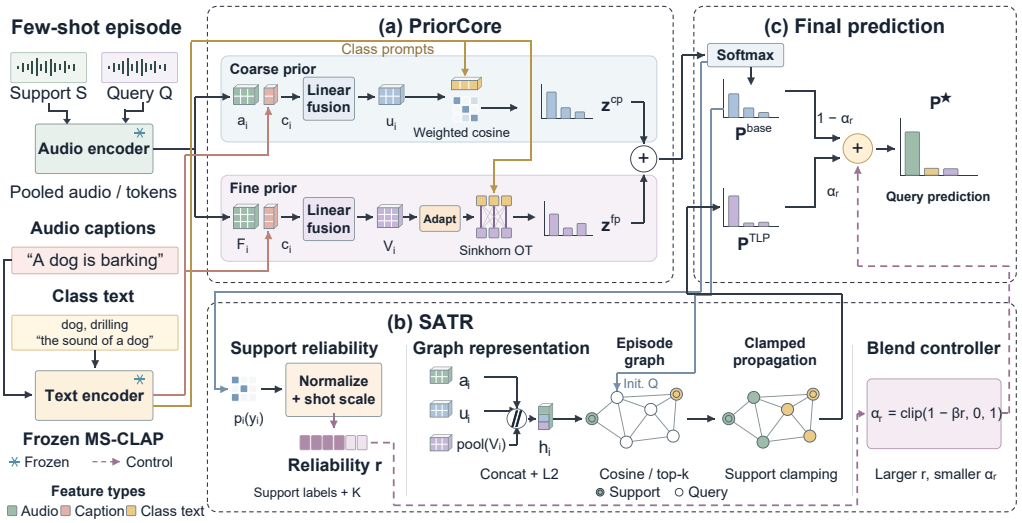

Figure 1: Overview of the proposed Prior-Centric Support Allocation (PCSA) framework. Conceptually, PCSA is composed of PriorCore and SATR. For readability, the full inference is organized into three operational stages: PriorCore, SATR, and control and final prediction. The third stage makes explicit how the SATR reliability signal determines the mixture weight and combines propagated and base predictions at decision time.

in audio, where local evidence can be noisy and query decisions often benefit from episode-level regularization.

The present paper takes a narrower route than larger structured-matching pipelines. We do not build a new heavy matching branch into the reported method. Instead, the support set estimates reliability, and that reliability controls how strongly transductive refinement should influence the final prediction. In this sense, the method is closer to a support-aware episode inference rule than to another support-side matching pipeline layered on top.

3 Method

3.1 Problem Setup

Each episode contains an N-way support set $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{NK}$ and an unlabeled query set $\mathcal{Q} = \{\mathbf{x}_j\}_{j=1}^{NQ}$. The model also has access to textual class names and class descriptions, which are encoded by the text branch of a frozen audio-language encoder. Given a frozen audio-language model, the goal is to predict query labels from support audio, class text, and within-episode geometry. Throughout this paper, the backbone remains frozen. A fresh identity-initialized linear adapter is created for each episode and trained on the current support set, and the same two small audio-caption fusion layers are updated on the same support data and persist across episodes. The contribution is therefore not a new pretraining recipe, but a prior-centric organization of support information. PCSA builds on the frozen audio-language encoder and the multi-grained cross-granular correspondence idea developed for few-shot audio adaptation (Zhao et al., 2024). It retains that coarse/fine prior scaffold but removes the cache-style support inventory of the prior adaptor: support evidence no longer enters through a separate memory branch, but enters once through the reliability controller of SATR.

As illustrated in Figure 1, PCSA has two substantive components, PriorCore and SATR; the figure organizes the full inference into three operational stages and makes explicit how the SATR reliability signal combines the propagated and base predictions at decision time.

3.2 PRIOR-CENTRIC SEMANTIC SCAFFOLD

Let $\mathbf{a}_i \in \mathbb{R}^d$ be the pooled audio feature of sample i, $\mathbf{F}_i \in \mathbb{R}^{L\times d}$ its fine-grained token features, and $\mathbf{c}_i \in \mathbb{R}^d$ its caption feature from the frozen audio-language encoder. For each candidate class c, the same frozen text encoder produces one or more class-prompt embeddings, which are normalized and summarized as a class anchor $\mathbf{t}_c$. A lightweight episode adapter g_θ consists of a support adapter plus audio-caption fusion and token-level fusion layers. The support adapter is initialized and trained for the current episode, while the same fusion layers are updated on each episode's support data. Training minimizes the cross-entropy between the prior logits below and the support labels.

The coarse prior compares the pooled adapter-fused representation with class prompts:

$$\mathbf{z}_i^{cp} = \sigma \cdot \frac{g_\theta^{\text{coarse}}(\mathbf{a}_i, \mathbf{c}_i)}{\|g_\theta^{\text{coarse}}(\mathbf{a}_i, \mathbf{c}_i)\|_2}^{\top} \frac{\mathbf{t}_c}{\|\mathbf{t}_c\|_2}, \tag{1}$$

where σ is the frozen model logit scale, and $\mathbf{z}_i^{cp}$ collects one entry per candidate class c in the episode vocabulary. The fine prior compares adapted token-level audio-caption features with class prompts through optimal transport:

$$\mathbf{z}_i^{fp} = \sigma_f \cdot \mathrm{OT}\big(g_\theta^{\text{fine}}(\mathbf{F}_i, \mathbf{c}_i), \mathbf{t}_c\big), \tag{2}$$

where σ_f is the fine-prior scale, set to $\log(1/0.07)$ in all reported runs, and $\mathrm{OT}(\cdot,\cdot)$ denotes an entropy-regularized transport solved by Sinkhorn iterations with temperature 0.2, entropy weight 0.1, and 100 solver steps. The two priors are complementary: the coarse term preserves pooled audio-text alignment, while the fine term preserves cross-granular local correspondence. PriorCore defines the final semantic scaffold as

$$\mathbf{z}^{\text{prior}} = \mathbf{z}^{cp} + \mathbf{z}^{fp}. \tag{3}$$

Both branches are logits over the same frozen text vocabulary, so additive log-space aggregation introduces no free weights of its own.

PriorCore deliberately excludes cache-style support memories. It isolates the semantic part of the pipeline and forces support-side information to enter through SATR as explicit episode-level refinement rather than through another memory bank.

3.3 SUPPORT-AWARE TRANSDUCTIVE REFINEMENT

SATR operates on top of PriorCore and decides how strongly the current support episode should affect inference once the semantic scaffold is available. The support set answers "how much should this episode be propagated?" rather than "how many extra support branches should be added?"

Support reliability from the current scaffold. Let $\mathbf{p}_i = \mathrm{softmax}(\mathbf{z}_i^{\text{support}})$ be the support prediction under the same prior branches used at query time. The normalized support confidence is

$$r_i = \mathrm{clip}\left(\frac{\mathbf{p}_i(y_i) - 1/N}{1 - 1/N}, 0, 1\right), \tag{4}$$

and its shot-aware aggregate is

$$r = \left(1 - e^{-K/\tau}\right) \cdot \frac{1}{NK} \sum_{i=1}^{NK} r_i, \tag{5}$$

where τ is a shot-scale constant. The score r is high when support labels are already explained by the scaffold and the number of shots is informative.

Reliability-aware transductive label propagation. Let $\mathcal{P}^{\text{base}} = \mathrm{softmax}(\mathbf{z}^{\text{prior}})$ be the pre-propagation probabilities. For graph construction, each sample is represented by the concatenation of its pooled audio feature, its adapted coarse fusion feature, and the pooled fine-grained audio feature:

$$\mathbf{h}_i = \mathrm{L2Norm}\big([\mathbf{a}_i;\ g_\theta^{\text{coarse}}(\mathbf{a}_i, \mathbf{c}_i);\ \mathrm{pool}(g_\theta^{\text{fine}}(\mathbf{F}_i, \mathbf{c}_i))]\big). \tag{6}$$

The pooling operation is mean-pooling over token-level fused features when they are available and otherwise uses the cached pooled fine audio feature. For benchmarks with per-audio captions, the

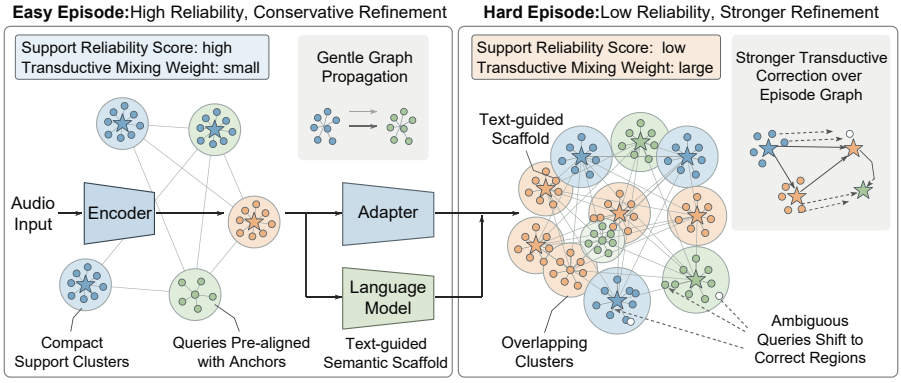

Figure 2: Schematic view of PCSA as a reliability-controlled inference rule. Left: when the prior scaffold already organizes the episode well, support reliability is high and the controller applies milder modulation. Right: when the episode becomes class-dense and locally entangled, reliability is lower and transductive correction is stronger.

normalized caption embedding $\mathbf{c}_i$ may be appended as an additional graph feature; ESC50 uses this caption-augmented form, while FSDKaggle2018 and UrbanSound8K use the three-term form. As in MGCL (Zhao et al., 2024), the per-clip caption is generated by a pretrained encoder-decoder captioner (an HTSAT audio encoder with a BART language decoder, applied without training) rather than taken from benchmark annotations, so it carries no label information beyond what the audio itself contains; PCSA reuses the same caption assets as MGCL. This graph representation is used only inside SATR and does not add a new evidence branch beyond the representations already used by the adapter. We compute cosine affinity over the support-query graph, clamp negative affinities to zero, remove self-loops, keep the top-k neighbors per row, symmetrize, and apply degree normalization. Let $\hat{\mathbf{A}}$ be the normalized affinity matrix and let $\mathbf{Y}^{(0)}$ concatenate support one-hot rows with query rows initialized from $\mathcal{P}^{\text{base}}$. Transductive label propagation iterates

$$\mathbf{Y}^{(t+1)} = \eta \hat{\mathbf{A}} \mathbf{Y}^{(t)} + (1-\eta)\mathbf{Y}^{(0)}, \tag{7}$$

while clamping support rows to their ground-truth one-hot vectors after each step. The query block after the final iteration is $\mathcal{P}^{\text{TLP}}$.

The reliability controller maps the scalar support reliability to the propagation weight

$$\alpha_r = \text{clip}(1 - \beta r, 0, 1), \tag{8}$$

and the final distribution is the bounded interpolation

$$\mathcal{P}^{\star} = \alpha_r \mathcal{P}^{\text{TLP}} + (1-\alpha_r)\mathcal{P}^{\text{base}}, \tag{9}$$

where β is the TLP blend coefficient. Here η controls propagation inertia, β sets the maximum controller influence, and α_r is the actual blend weight between the propagated and base predictions. In the reported method $\beta = 0.35$ and $r \in [0,1]$, so $\alpha_r \in [0.65, 1.0]$; the controller keeps propagation active and modulates its strength instead of switching between two modes.

3.4 Controller Interpretation and Final Prediction

Substituting Eq. (5) into Eq. (8) makes the allocation behavior explicit:

$$\alpha_r = 1 - \beta\,(1 - e^{-K/\tau})\,\bar{r}_s, \tag{10}$$

where $\bar{r}_s = \frac{1}{NK}\sum_i r_i$. The controller is therefore monotone in support confidence and shot count, and Eq. (9) keeps the prediction inside the classwise interval spanned by $\mathcal{P}^{\text{TLP}}$ and $\mathcal{P}^{\text{base}}$. The support-reliability step, affinity construction, and iterative propagation add only lightweight per-episode computation.

Figure 2 and Eq. (9) also give a direct role-separation interpretation. For episodes sharing the same prior scaffold $\mathbf{z}^{\text{prior}}$, the base prediction $\mathcal{P}^{\text{base}}$ is identical, and support-side variation affects the

output only through $\mathcal{P}^{\text{TLP}}$ and the scalar controller α_r. PCSA does not hard-code different modules for different benchmark settings: the same rule applies mild modulation when the scaffold already explains the episode and stronger modulation when local support-query geometry is less trustworthy. This setting-invariant interpretation is consistent with the cross-setting improvements in Section 4.

4 Experimental Results

4.1 Datasets

ESC50 The ESC50 dataset (Piczak, 2015) contains 2,000 environmental audio clips of 50 semantic classes, with 40 examples per class. In the released evaluation protocol of Liang *et al.* (Liang et al., 2023), the class vocabulary is repeatedly partitioned into five validation folds of 15 classes each, which makes ESC50 a stable, near-saturated benchmark for transfer.

UrbanSound8K UrbanSound8K (Salamon et al., 2014) contains urban acoustic events from 10 official classes. Compared with ESC50 it is acoustically less clean, so it stresses robustness when the semantic prior is less well aligned with the episode, and its validation folds differ substantially in difficulty.

FSDKaggle2018 FSDKaggle2018 (Fonseca et al., 2018) contains 11,073 clips annotated with 41 AudioSet-style labels. Its five validation folds hold 12 classes each and are of comparable difficulty, and it offers the most headroom for methodological comparison, so the most detailed ablations and parameter studies are centered on it.

4.2 Implementation Details

All experiments use the frozen MS-CLAP backend. Each dataset is evaluated in the four standard settings, 5-way 1-shot, 5-way 10-shot, 12-way 1-shot, and 12-way 10-shot, with 100 tasks per protocol fold; 5-way and 12-way episodes are both sampled from the per-fold validation vocabulary described above, and UrbanSound8K follows the five-class and twelve-slot protocols defined by MGCL, whose twelve-slot vocabulary repeats two of the ten official classes in each fold. The rows other than PCSA in Table 1 are as reported under the MGCL protocol (Zhao et al., 2024), and, unlike PCSA, those methods do not use the unlabeled query set. The encoder remains frozen throughout and is never updated; what is trained is the per-episode support adapter together with the two small audio-caption fusion layers that persist across episodes, so the comparison isolates prior-centric inference and support allocation rather than changing the pretraining recipe. Unless otherwise stated, the trainable components take 20 gradient steps per episode, each step over the full support set as a single batch, using AdamW and a cosine-annealed learning rate of 10^{-4}. The reported PCSA setting keeps the two prior branches, disables the two cache branches, uses transductive label propagation with $\eta = 0.70$, top-$k = 5$, and 10 steps, and sets the support controller to $\tau = 4.0$ and $\beta = 0.35$. Class prompts follow the template "this is a sound of *class name*" together with the benchmark-provided category phrases. The main comparison, ablation, cross-seed, and parameter tables use the 100-task-per-fold protocol; the episode diagnostics use the same sampled episodes as the main protocol. Transductive label propagation applies row normalization only after the final step, and the fine-prior transport solver uses an entropy-regularized assignment with temperature 0.2, 100 iterations, entropy weight 0.1, and a similarity scale of $\log(1/0.07)$ before fusion.

4.3 Main Result

As shown in Table 1, PCSA attains the best average on all three datasets under the adopted protocol, raising the average accuracy over the strongest comparison row from 99.03 to 99.43 on ESC50, from 86.14 to 90.46 on UrbanSound8K, and from 91.74 to 93.53 on FSDKaggle2018. Over the five folds the PCSA averages have 95% intervals of at most ± 1.3 points on ESC50 and FSDKaggle2018 and up to ± 7.8 points on UrbanSound8K, whose folds differ in difficulty, and the paired per-episode tests are reported in Section 4.4. Because the compared rows are inductive whereas PCSA also uses the unlabeled query set, the comparison additionally absorbs the effect of transductive inference, so Section 4.4 also isolates the contribution of the reliability controller inside the same transductive pipeline. The gain tracks unresolved acoustic ambiguity rather than support size alone: UrbanSound8K, whose ten-class vocabulary is acoustically confusable, benefits most, while

Table 1: Comparison of audio classification accuracy (%) with representative previous few-shot audio recognition methods on ESC50, UrbanSound8K, and FSDKaggle2018 under different few-shot settings. PCSA is evaluated under the protocol of Section 4; the compared rows are as reported under the MGCL protocol (Zhao et al., 2024) and are inductive. w/o CLAP: Non-CLAP backbone model, CLAP-base: Basic CLAP model, CLAP-adapt: CLAP model adapted for audio recognition; entries marked – are not reported. For UrbanSound8K the four settings draw on different class subsets, so its average is a reference across settings of differing difficulty.

Dataset	Type	Method	5-way Acc		12-way Acc		Avg.
			1-shot	10-shot	1-shot	10-shot	
ESC50	w/o CLAP	MatchNet (Vinyals et al., 2016)	–	86.83	–	71.81	–
	w/o CLAP	ProtoNet (Snell et al., 2017)	–	88.18	–	77.70	–
	CLAP-base	CLAP-zeroshot (Elizalde et al., 2023)	96.84	96.84	93.65	93.65	95.25
	CLAP-adapt	Tip-Adapter-F (Zhang et al., 2022)	96.97	97.52	93.97	95.58	96.01
	CLAP-adapt	Treff-adapter (Liang et al., 2023)	97.44	98.53	94.83	96.29	96.77
	CLAP-adapt	MGCL (Zhao et al., 2024)	98.63	99.54	98.42	99.51	99.03
	CLAP-adapt	PCSA (ours)	**99.39**	**99.77**	**98.76**	**99.78**	**99.43**
UrbanSound8K	CLAP-adapt	Tip-Adapter-F (Zhang et al., 2022)	81.98	83.48	72.35	75.20	78.25
	CLAP-adapt	Treff-adapter (Liang et al., 2023)	82.27	87.26	73.19	80.12	80.71
	CLAP-adapt	MGCL (Zhao et al., 2024)	85.24	90.16	81.26	87.90	86.14
	CLAP-adapt	PCSA (ours)	**89.19**	**93.48**	**86.59**	**92.59**	**90.46**
FSDKaggle2018	CLAP-base	CLAP-zeroshot (Elizalde et al., 2023)	88.76	88.76	80.24	80.24	84.50
	CLAP-adapt	Tip-Adapter-F (Zhang et al., 2022)	88.98	90.75	80.83	83.92	86.12
	CLAP-adapt	Treff-adapter (Liang et al., 2023)	89.37	92.31	81.58	86.94	87.55
	CLAP-adapt	MGCL (Zhao et al., 2024)	90.67	94.88	88.06	93.33	91.74
	CLAP-adapt	PCSA (ours)	**93.47**	**95.93**	**90.36**	**94.38**	**93.53**

ESC50, where the frozen text encoder already separates the classes, moves only slightly. The subsequent analyses focus on FSDKaggle2018 and ask where and why the gain appears. Because every comparison below uses the same episode sequence, the per-episode differences are paired rather than marginal.

4.4 Ablation Experiments

Table 2 isolates the role-changing design choices on FSDKaggle2018 under one consistent protocol. All rows keep the coarse prior as the common semantic anchor and lower bound. In the cache-memory operating point, the fine prior and SATR are active but support evidence is still accumulated through the cache branches; replacing that cache-memory pathway with PCSA raises the average from 92.36 to 93.53 ($+1.17$ points), improving every setting. Because the two operating points are evaluated on the same episode sequence, their per-episode results pair one-to-one: the improvement is $+0.50$, $+0.56$, $+1.61$, and $+2.02$ points across the four settings, and a paired t-test over the five fold means ($n = 5$; fold means are used because episodes within a fold share the validation class subset) gives $p = 0.13$, 0.0003, 0.006, and 0.0002, so the difference is significant in the settings with 10 shots or with 12 classes. The fine prior also remains important: adding it to the cache-free coarse scaffold with SATR active raises the average from 92.97 to 93.53 ($+0.56$ points), again with positive cell-level changes.

SATR also improves the coarse-prior-only operating point from 91.47 to 92.97 ($+1.50$ points), so support-conditioned refinement contributes even without the fine prior. Under these comparisons, PCSA is the strongest configuration in every reported setting, and the largest average upgrade appears in the 12-way evaluations, where the class vocabulary is most crowded. We notice that the benefit of the reliability controller itself is concentrated in the same regime: replacing it with the fixed-α row of Table 2 leaves the 5-way results unchanged but costs 0.29 and 0.69 points in 12-way 1-shot and 12-way 10-shot. We attribute this to the reliability-conditioned modulation of Eq. (9): a crowded class set makes the scaffold less confident, so the propagated distribution carries more of the decision. The controller therefore behaves as intended rather than as a uniform mixing rule: its contribution concentrates in the settings where the class vocabulary is densest, and it is close to inactive where the prior scaffold already resolves the episode.

4.5 Robustness Analysis

Repeating the full FSDKaggle2018 PCSA protocol with seeds 0, 1, 2, and 42 gives seed-level averages 93.29, 93.63, 93.42, and 93.53 (standard deviation 0.15). The same holds across the control-

Table 2: Component ablation on FSDKaggle2018 (accuracy %). "Cache", "Fine", and "SATR" denote whether the cache-memory branches, fine-prior branch, and support-aware transductive refinement are enabled. "w/ fixed α" replaces the reliability controller of Eq. (9) with a constant blend weight ($\alpha = 1$, i.e. the propagated distribution is used as is). Δ Avg. is relative to the coarse-prior-only operating point. Bold marks the best value in each column.

Configuration	Cache	Fine	SATR	5-way Acc		12-way Acc		Avg.	Δ Avg.
				1-shot	10-shot	1-shot	10-shot		
Coarse prior only	Off	Off	Off	90.78	93.27	89.00	92.84	91.47	–
Prior + cache + SATR	On	On	On	92.97	95.37	88.74	92.36	92.36	+0.89
Coarse + SATR	Off	Off	On	92.55	95.54	89.99	93.80	92.97	+1.50
PCSA w/ fixed α	Off	On	On	93.44	95.74	90.06	93.69	93.23	+1.76
PCSA	Off	On	On	**93.47**	**95.93**	**90.36**	**94.38**	**93.53**	**+2.06**

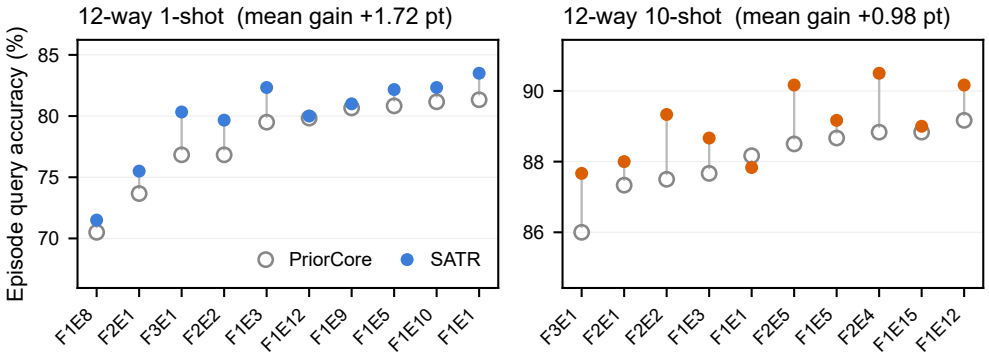

Figure 3: Episode-level SATR diagnostics on FSDKaggle2018 over the episodes of the main protocol (100 tasks per fold, 500 episodes per setting). Both panels show the ten hardest sampled 12-way episodes, selected by PriorCore accuracy and labeled F⟨fold⟩E⟨index⟩, as paired PriorCore and SATR accuracies for 12-way 1-shot (left) and 12-way 10-shot (right).

parameter grid: over $\beta \in \{0.2, 0.35, 0.5\}$, $\tau \in \{2, 4, 8\}$, and $k \in \{3, 5, 7\}$, the FSDKaggle2018 average varies by at most 0.10 points around the reported operating point.

4.6 Episode Diagnostics

Figure 3 probes episode-level behavior on FSDKaggle2018 over the episodes of the main protocol. Both panels show the ten hardest sampled 12-way episodes, selected by PriorCore accuracy: refinement raises their mean accuracy from 78.12% to 79.83% under 12-way 1-shot and from 88.07% to 89.05% under 12-way 10-shot. Over all sampled 12-way 1-shot episodes the mean change is $+0.64$ points and 75% of the episodes improve, which is the setting in which the paired test of Section 4.4 is significant. Refinement is therefore not a uniform post-processing bonus: it concentrates where the scaffold is least sufficient.

5 Conclusion

PCSA shows that support evidence is most effective when it regulates episode-level inference rather than acting as another memory pathway. The design is intentionally minimal: PriorCore keeps class-level semantics explicit, and SATR uses support reliability to decide when propagation should intervene. Under the adopted protocol, PCSA attains the best average on all three datasets, and the ablations and diagnostics confirm that both blocks contribute while gains concentrate where prior evidence is least sufficient. This highlights the value of reorganizing existing evidence rather than adding another support-heavy pathway. Because the encoder stays frozen and the mechanism depends only on support reliability and graph-based refinement, the same prior-centric view is in principle applicable to other audio-language backbones without retraining the backbone, although we have only evaluated it on MS-CLAP here.

AI USE STATEMENT

We used generative AI tools as an assistive aid during this work, including language editing and polishing of the manuscript and coding assistance in the implementation and evaluation pipeline. Generative AI tools were not used to fabricate experimental results, datasets, baselines, or citations, and all reported numbers were produced by the described implementation and evaluation protocol. All AI-assisted text and code were reviewed, tested, and validated by the authors, who take responsibility for the final content of this work, including text, claims, and artifacts.

ETHICS STATEMENT

This work uses publicly available audio datasets and frozen pretrained audio-language models, and it does not collect new human-subject data. Some of the benchmark recordings are crowdsourced field recordings of everyday environments; we use them only as distributed by their original providers under their stated terms. We are not aware of any immediate privacy, safety, or discrimination risks specific to the proposed few-shot recognition method.

REPRODUCIBILITY STATEMENT

We describe the datasets, frozen encoder, episodic training procedure, and evaluation protocol in the Experiments section. The method is an inference-stage modification over a frozen audio-language backbone. The released code includes the episode sampler, graph construction, reliability controller, and evaluation scripts needed to reproduce the reported accuracy values, together with a documentation file listing the datasets, pretrained weights, caption and prompt metadata, and cached-feature schema that have to be prepared locally. An anonymized repository containing the code is available at `https://anonymous.4open.science/r/pcsa-code/`.

REFERENCES

Ghassen Baklouti, Maxime Zanella, and Ismail Ben Ayed. Language-aware information maximization for transductive few-shot clip. *arXiv preprint arXiv:2509.00305*, 2025.

Malik Boudiaf, Imtiaz Ziko, Jérôme Rony, José Dolz, Pablo Piantanida, and Ismail Ben Ayed. Information maximization for few-shot learning. *Advances in Neural Information Processing Systems*, 33:2445–2457, 2020.

Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Htsat: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 646–650. IEEE, 2022.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. Beats: audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 5178–5193, 2023.

Szu-Yu Chou, Kai-Hsiang Cheng, Jyh-Shing Roger Jang, and Yi-Hsuan Yang. Learning to match transient sound events using attentional similarity for few-shot sound recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 26–30. IEEE, 2019.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Benjamin Elizalde, Soham Deshmukh, and Huaming Wang. Natural language supervision for general-purpose audio representations. In *ICASSP 2024-2024 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 336–340. IEEE, 2024.

Eduardo Fonseca, Manoj Plakal, Frederic Font, Daniel PW Ellis, Xavier Favory, Jordi Pons, and Xavier Serra. General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline. *arXiv preprint arXiv:1807.09902*, 2018.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International journal of computer vision*, 132(2):581–595, 2024.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780. IEEE, 2017.

Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. In *Proc. Interspeech 2021*, pp. 571–575, 2021.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 976–980. IEEE, 2022.

Calum Heggan, Sam Budgett, Timothy Hospedales, and Mehrdad Yaghoobi. Metaaudio: A few-shot audio classification benchmark. In *International Conference on Artificial Neural Networks*, pp. 219–230. Springer, 2022.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

Yanxiong Li, Wenchang Cao, Jialong Li, Wei Xie, and Qianhua He. Few-shot class-incremental audio classification using stochastic classifier. In *Proc. Interspeech 2023*, pp. 4174–4178, 2023a.

Yanxiong Li, Wenchang Cao, Wei Xie, Jialong Li, and Emmanouil Benetos. Few-shot class-incremental audio classification using dynamically expanded classifier with self-attention modified prototypes. *IEEE Transactions on Multimedia*, 26:1346–1360, 2023b.

Yanxiong Li, Hao Chen, Wenchang Cao, Qisheng Huang, and Qianhua He. Few-shot speaker identification using lightweight prototypical network with feature grouping and interaction. *IEEE Transactions on Multimedia*, 25:9241–9253, 2023c.

Jinhua Liang, Xubo Liu, Haohe Liu, Huy Phan, Emmanouil Benetos, Mark David Plumbley, and Wenwu Wang. Adapting language-audio models as few-shot audio learners. In *Proc. Interspeech 2023*, pp. 276–280, 2023. doi: 10.21437/interspeech.2023-1082.

Jinhua Liang, Xubo Liu, Wenwu Wang, Mark D Plumbley, Huy Phan, and Emmanouil Benetos. Acoustic prompt tuning: Empowering large language models with audition capabilities. *IEEE Transactions on Audio, Speech and Language Processing*, 33:949–961, 2025.

Xiaoyu Liu, Hanlin Lu, Jianbo Yuan, and Xinyu Li. Cat: Causal audio transformer for audio classification. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

Ségolène Martin, Yunshi Huang, Fereshteh Shakeri, Jean-Christophe Pesquet, and Ismail Ben Ayed. Transductive zero-shot and few-shot clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28816–28826, 2024.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354, 2024.

Charilaos Papaioannou, Emmanouil Benetos, and Alexandros Potamianos. Lc-protonets: Multi-label few-shot learning for world music audio tagging. *IEEE Open Journal of Signal Processing*, 6:138–146, 2025.

Sangwook Park, David K Han, and Mounya Elhilali. Cross-referencing self-training network for sound event detection in audio mixtures. *IEEE transactions on multimedia*, 25:4573–4585, 2022.

Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1015–1018, 2015.

Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 1041–1044, 2014.

Julio Silva-Rodríguez, Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. A closer look at the few-shot adaptation of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23681–23690, 2024. doi: 10.1109/CVPR52733.2024.02235.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D Plumbley. Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia*, 17 (10):1733–1746, 2015.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

Yu Wang, Nicholas J Bryan, Justin Salamon, Mark Cartwright, and Juan Pablo Bello. Who calls the shots? rethinking few-shot learning for audio. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 36–40. IEEE, 2021.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Xianjun Xia, Roberto Togneri, Ferdous Sohel, Yuanjun Zhao, and Defeng Huang. Multi-task learning for acoustic event detection using event and frame position information. *IEEE Transactions on Multimedia*, 22(3):569–578, 2019.

Huang Xie and Tuomas Virtanen. Zero-shot audio classification via semantic embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1233–1242, 2021.

Maxime Zanella, Benoît Gérin, and Ismail B Ayed. Boosting vision-language models with transduction. *Advances in Neural Information Processing Systems*, 37:62223–62256, 2024.

Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 12200–12210. IEEE, 2020.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pp. 493–510. Springer, 2022.

Shengwei Zhao, Linhai Xu, Yuying Liu, and Shaoyi Du. Multi-grained correspondence learning of audio-language models for few-shot audio recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9244–9252, 2024. doi: 10.1145/3664647.3681389.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348, 2022.

Imtiaz Ziko, Jose Dolz, Eric Granger, and Ismail Ben Ayed. Laplacian regularized few-shot learning. In *International conference on machine learning*, pp. 11660–11670. PMLR, 2020.

A Controller Analysis

A.1 Controller Interpretation and Complexity

The allocation behavior of PCSA becomes more explicit if we separate the shot-dependent factor from the average support confidence. Let

$$\bar{r}_s = \frac{1}{NK} \sum_{i=1}^{NK} r_i. \tag{11}$$

Equation (10) follows by substituting Eq. (5) into Eq. (8). This form makes the core design principle explicit: support evidence does not create another competing classifier, but changes how aggressively the current episode should be refined once the semantic scaffold has already been formed.

Equation (10) also reveals three useful analytical properties. First, for fixed K, the controller is monotone in every support-confidence term:

$$\frac{\partial \alpha_r}{\partial r_i} = -\frac{\beta(1 - e^{-K/\tau})}{NK} \leq 0. \tag{12}$$

Better-explained support therefore always reduces the correction strength, which matches the intended interpretation that reliable support should require less intervention from transductive refinement. Second, for a fixed mean confidence $\bar{r}_s$, the controller is shot-aware:

$$\frac{\partial \alpha_r}{\partial K} = -\frac{\beta \bar{r}_s}{\tau} e^{-K/\tau} \leq 0. \tag{13}$$

Thus the same average support quality is treated as more trustworthy when more labeled shots are available, so the resulting refinement becomes less aggressive. Third, because Eq. (9) is a convex combination of two valid class distributions, the final prediction remains bounded classwise:

$$\min\big(P_{qc}^{\text{base}}, P_{qc}^{\text{TLP}}\big) \leq P_{qc}^{\star} \leq \max\big(P_{qc}^{\text{base}}, P_{qc}^{\text{TLP}}\big), \tag{14}$$

for every query q and class c. This boundedness ensures that PCSA cannot produce an unstructured correction outside the range already implied by the base and propagated predictions.

The same decomposition also clarifies the computational footprint. Let $M = NK + NQ$ denote the total number of samples in the episode and let d_g be the graph-feature dimension used by TLP. PriorCore itself introduces no new heavy branch beyond the already available prior logits. SATR adds three lightweight costs: support-reliability estimation of order $O(NKN)$, pairwise affinity construction of order $O(M^2 d_g)$ before sparsification, and iterative propagation of order $O(TkMN)$ for T propagation steps and row-wise top-k neighborhoods. Under standard few-shot protocols, M is small and the backbone remains frozen, so the added cost is concentrated in the episode graph rather than in retraining a large support-side module. This is consistent with the paper's positioning of PCSA as an inference-organization method rather than a new heavy adaptation stack.

B Parameter Sensitivity

The parameter study uses FSDKaggle2018. It provides auxiliary operating-region evidence around the selected full-protocol setting and is not needed for the main comparison claim.

Table 3 summarizes the operating region around the selected FSDKaggle2018 setting. The center rows repeat the selected full-protocol mainline, and the nearby one-parameter variants show that the method remains stable when the TLP blend, shot scale, or neighborhood size is adjusted. Across these variants the average spans only 93.50–93.63, a range smaller than the cross-seed spread (0.15 points) and far inside the fold-level interval, so the variants are not separable by average accuracy. The reported setting is used because it is the center of the grid and attains the best accuracy in the two 10-shot settings, including the most crowded 12-way 10-shot case.

C Extended Discussion

C.1 Prior-Centric Role Separation

PCSA assigns distinct roles to the two evidence sources already present in CLAP-based few-shot audio recognition. PriorCore keeps the semantic scaffold explicit, while SATR reintroduces support

Table 3: Sensitivity of PCSA to the main control parameters on FSDKaggle2018. The center rows ($\beta = 0.35$, $\tau = 4$, and $k = 5$) are the selected full-protocol mainline repeated for reference; the off-center rows are nearby one-parameter variants from auxiliary operating-region studies.

Parameter	Value	5-way Acc		12-way Acc		Avg.
		1-shot	10-shot	1-shot	10-shot	
TLP blend β	0.2	93.76	95.83	90.85	93.88	93.58
	0.35	93.47	95.93	90.36	94.38	93.53
	0.5	93.71	95.82	90.68	94.07	93.57
Shot scale τ	2	93.60	95.91	90.74	93.95	93.55
	4	93.47	95.93	90.36	94.38	93.53
	8	93.76	95.90	90.86	94.00	93.63
TLP neighborhood k	3	93.62	95.89	90.63	93.85	93.50
	5	93.47	95.93	90.36	94.38	93.53
	7	93.78	95.86	90.83	94.02	93.62

information as a controller of episode refinement. Once those roles are separated, the source of improvement becomes easier to interpret: the contribution is not simply to delete modules, but to reallocate what support evidence is allowed to do once the semantic scaffold is explicit, so that support evidence mainly decides how strongly episode refinement should intervene.

C.2 Why the Gain Concentrates on Harder Regimes

The largest margins appear when the episode is least trustworthy: fewer support examples, more competing classes, or noisier acoustic conditions. This is why the gains are modest on near-saturated ESC50 but much larger on UrbanSound8K and FSDKaggle2018. The same ordering holds within FSDKaggle2018 at the episode level: refinement contributes most where the prior scaffold is least decisive on its own, and its contribution shrinks as the scaffold becomes sufficient for the episode.

C.3 Mechanistic Interpretation and Boundary Conditions

The evidence should be read as mechanism-oriented diagnostics rather than as a full causal proof. These diagnostics also show the boundary conditions of the controller. The gain is not uniform: it is largest on the episodes where the scaffold is least decisive, close to zero where the scaffold already separates the classes, and slightly negative on average in the densest 12-way 10-shot setting. This behavior is consistent with a controller that reallocates refinement toward ambiguous local neighborhoods instead of providing a uniform post-processing bonus.

C.4 Scope and Evaluation

The episode-level diagnostics are reported on the FSDKaggle2018 12-way settings, where the class vocabulary is densest and the prior scaffold is least often decisive on its own. Denser support can coexist with local ambiguity once the class set becomes larger and acoustically closer. A full-protocol table remains necessary because it tells us whether the method is useful on average under the released benchmark, but the average alone cannot explain which type of episode is being improved. In our case, the full 100-task FSDKaggle2018 protocol gives a moderate but consistent gain, whereas the episode-level view shows that the hardest sampled episodes benefit more strongly than the protocol mean would suggest. This is especially relevant in audio, where difficulty is often driven by transient local similarity, background interference, and semantic overlap rather than by support scarcity alone. The deepest ablations and diagnostics are therefore centered on FSDKaggle2018, while ESC50 and UrbanSound8K serve as cross-dataset transfer evidence under the same released evaluation protocol.