arXiv:2602.03817v1 [cs.SD] 3 Feb 2026

音频时空融合的自适应证据加权

Oscar Ovanger [1] **Levi Harris** [2] **Timothy H. Keitt** [1]

摘要

许多机器学习系统可以利用多个证据来源来预测同一目标，但这些来源的可靠性和信息量往往因输入而异。在生物声学分类中，物种身份既可以根据声学信号推断，也可以根据地点和季节等时空上下文推断；贝叶斯推断虽为证据的乘法组合提供了理论依据，但在实践中，我们通常只能使用判别式预测器，而无法获得经过校准的生成模型。我们提出**融**合于**相**互独立的**条**件**假**设（**FINCH**），这是一种自适应对数线性证据融合框架，可将预训练音频分类器与结构化时空预测器相结合。FINCH 学习逐样本门控函数，根据不确定性和信息量统计来估计上下文信息的可靠性。由此形成的融合族包含纯音频分类器这一特例，并明确限制上下文证据的影响，从而得到一个风险受控的假设类，并提供可解释的纯音频回退方案。在多个基准测试中，FINCH 的表现始终优于固定权重融合和纯音频基线，即使上下文信息单独使用时作用有限，也能提升鲁棒性并改善误差权衡。凭借轻量、可解释的基于证据的方法，我们在 CBI 上取得最先进的性能，并在 BirdSet 的若干子集上达到具有竞争力或更优的表现。代码见：匿名代码仓库

1. 引言

集成学习研究如何组合针对同一目标变量的多个预测模型，以提升性能、鲁棒性或校准效果，使其优于任何单个模型（Kuncheva，2004；Kittler 等人，1998年）。装袋法、提升法和堆叠法等经典集成方法通常假设所有预测器都基于相同的数据分布进行训练，并且集成权重或组合规则可以与模型参数一同学习。



图 *1.* 暗眼灯草雀 （北美朱雀），本模型的命名由来。

然而，在许多现代应用中，预测模型通常经过预训练、彼此异构，并且在推理时保持固定。这些模型可能依赖不同的证据来源，在不同数据集上训练，并基于彼此不兼容的建模假设运行。因此，联合重新训练或微调往往不可行。相反，问题变成了如何以合理的方式融合固定预测器的输出。

我们考虑这样一种常见情形：多个预测模型为同一个潜在目标提供互补证据。令 y 表示一个类别标签，令 x 和 s 表示两个证据来源。条件独立性，

$$x \perp s \mid y, \tag{1}$$

提供了一种有用的理想化设定，为乘法式证据融合提供了依据。如果完整的生成分布已知，贝叶斯推断将得到

$$p(y \mid x, s) \propto p(x \mid y)\, p(s \mid y)\, p(y). \tag{2}$$

FINCH 采用相应的对数线性融合形式，同时适用于独立性假设仅近似成立的情形；偏差则可通过自适应有界加权加以缓解。

然而在实践中，生成模型往往不可用。相反，通常只能获得独立训练的判别式预测器 $p_\theta(y \mid x)$ 和 $p_\psi(y \mid s)$。在这种情况下，后验概率可表示为

$$p(y \mid x, s) \propto \frac{p_\theta(y \mid x)\, p(x)\, p_\psi(y \mid s)\, p(s)}{p(y)}, \tag{3}$$

其中，边缘分布 $p(x)$、$p(s)$ 以及隐含的先验分布 $p(y)$ 通常都是未知的。因此，仅凭判别模型无法直接计算真实的后验概率。

在这种情形下，一种常见的近似方法是使用对数线性模型或专家乘积形式，将判别预测器进行组合（Hinton，2002年；Genest & Zidek，1986年）。这些模型直接作用于后验概率输出，并定义如下形式的融合分布：

$$\log p(y \mid x, s) = \log p_\theta(y \mid x) + \log p_\psi(y \mid s) - \log Z(x, s), \tag{4}$$

其中，$Z(x, s)$ 是一个归一化常数，用于确保 $\sum_y p(y \mid x, s) = 1$。全文中，我们均以未归一化的对数空间形式表示融合规则；归一化常数由最终的 Softmax 隐式处理，因此在不影响类别间比较时予以省略。在温和公理条件下，对数线性融合可作为对数意见池获得决策理论上的依据（Heskes，1998年），并且在仅有判别模型可用时，可作为一种实用的替代方案。

现有方法在组合预测器时，通常采用固定权重或全局学习得到的权重。这隐含地假设每个证据来源的相对可靠性在整个输入空间中保持不变。在许多应用中，这一假设并不成立。某个证据来源的信息量可能因样本而显著变化，使用固定权重会导致性能下降，或使某个来源出现病态主导。

在本研究中，我们提出了一种自适应对数线性融合框架。该框架保留了经典专家乘积模型的结构与可解释性，同时允许按样本调节证据强度。我们假设各组成预测器均为固定且具有判别能力的模型，不会重新训练或重新校准其参数。相反，我们学习一个门控函数，按样本估计上下文证据的可靠性。由此得到的融合后验概率（忽略归一化因子）定义如下：

$$\log \tilde{p}_\omega(y \mid x, s) = \log p_\theta(y \mid x) + \omega(x, s)\, \log p_\psi(y \mid s), \tag{5}$$

其中 $\omega(x, s) \geq 0$ 是一个学习得到的加权函数。归一化后验概率 $p_\omega(y \mid x, s)$ 通过对 y 应用 softmax 得到。该公式在 $\omega(x, s) = 0$ 时还原为第一个模型的分类器，同时限制上下文证据的影响，并且无需重新训练基础模型即可实现自适应融合。

我们在生物声学物种分类场景下评估了这一框架，将最先进的音频分类器与从大规模观测数据中推导出的结构化时空先验分布相结合。实验表明，自适应加权始终优于固定权重融合和纯音频基线，尤其是在上下文信息具有异质性或单独使用时较弱的情况下。

2. 相关工作

我们将相关工作分为两个主题：（i）概率预测器的组合理论与机制（对数线性池化、专家乘积和门控）；（ii）生物声学应用，并区分仅音频基础编码器与融合时空上下文的系统。

2.1. 对数线性融合、意见池与专家乘积

在机器学习和统计学领域，多分类器组合已得到广泛研究（Kuncheva，2004年；Kittler 等，1998年）。经典集成方法（例如平均、投票、贝叶斯组合）依赖分类器多样性和误差相关性，通常会与模型参数联合学习权重，或通过验证调优权重。

最大熵模型和对数意见池将分类器融合形式化为对预测分布的对数线性聚合（Berger 等，1996；Genest 与 Zidek，1986年）。专家乘积模型通过相乘来组合分布：专家意见一致时会产生尖锐的后验概率，意见不一致时则会产生分散的后验概率（Hinton，2002年）。对数意见池在温和公理条件下，为加权对数线性组合提供了决策理论依据（Heskes，1998年）；当只有判别式预测器可用时，对数线性融合可视为贝叶斯推断的一种实用替代方案（Bishop，2006）。

大多数先前方法采用固定权重或全局学习得到的权重，隐含地假设专家可靠性在整个输入空间中保持不变。我们的工作则聚焦于对上下文证据进行逐样本加权，同时保留对数线性融合的可解释性。

2.2. 门控网络、混合专家与可靠性估计

混合专家模型引入门控网络，为多个预测器分配依赖于输入的权重（Jacobs 等人，1991；Jordan 和 Jacobs，1994）。这类模型通常采用端到端方式训练，同时联合优化专家和门控网络。

我们的设定有两个关键区别。首先，我们将各个预测器视为固定的，并假设它们分别基于不同的证据来源独立训练。其次，基于给定标签时各证据来源条件独立这一假设，我们在概率空间中采用乘法式融合（在对数空间中采用对数加法）。门控函数调节强度

的conte 上下文证据，而非在专家之间进行选择 rts.

一条相关研究探讨不确定性估计与选择性预测。贝叶斯近似和校准不确定性信号（Gal & Ghahramani，2016年；Kendall &Gal，2017年）以及基于置信度的方法（Hendrycks& Gimpel，2017年）常用于识别不可靠预测或分布外输入。选择性分类允许在置信度较低时弃权（Geifman& El-Yaniv，2017年）。相比之下，我们的方法利用可靠性信号，在概率融合模型中持续调节上下文证据的贡献，而不是弃权或丢弃某个专家。

2.3. 面向生物声学的音频基础模型（纯音频编码器）

现代生物声学分类器越来越多地依赖自监督或大规模预训练音频编码器，随后在下游数据集上进行轻量适配（例如线性探针或原型探针）（Lauha等，2022年；Wang等，2022年；Hagiwara，2022年；Choudhary等，2022年；Nolasco等，2023年；Rauch等，2025年）。近期的基础模型式方法为动物声音提供了强可迁移表示，包括BEATs（Chen等，2022年）和NatureLM-Audio（Robinson等，2025年），以及BEANS等大规模基准测试/分析工作，和对生物声学表示质量驱动因素的研究（Hagiwara等，2022年；Miron等，2025年）。Perch式模型强调面向鸟类与野生动物声学的大规模预训练，并为迁移学习提供强嵌入（Ghani等，2023年；van Merri¨enboer等，2025年）。

这些纯音频编码器功能强大，因为它们无需时空元数据即可在音频语料库上进行预训练，之后还能在不同任务和区域中重复使用。我们的方法明确保留了这种模块化特性：冻结音频编码器/分类器，并仅在融合层引入上下文信息。

2.4. 生物声学中的时空上下文：先验分布与联合模型

位置和季节等上下文信息通常通过基于物种分布模型得出的生态学先验分布加以利用（Fink 等人，2013；Madhusudhana 等人，2021），其中包括大规模的公民科学丰度模型（Fink 等人，2010；2014年；Johnston 等人，2015）。许多实用的生物声学系统会将声学预测与此类先验分布结合，方法包括启发式处理（例如事后过滤）或将元数据与学习得到的音频特征拼接。

BirdNET 是一种广泛使用的鸟类监测系统，通常与元数据约束（例如基于位置和日期的筛选或情境可行性）结合部署，并用于声学分类（Kahl等人，2021）。更进一步

通常而言，将音频和元数据输入单一预测器的联合方法 $p(y \mid x,\ s)$ 表达能力更强，但必须直接学习音频-情境交互，并且需要配对监督。正如附录A2中形式化阐述的那样，当证据来源之间存在较强的条件依赖时，这种联合建模便成为必要；而在（近似）条件独立的情况下，对数线性融合已足够，并可避免重新训练大型多模态架构的复杂性。

我们的研究针对一种常见情形：上下文信息虽有价值，却存在异质性。我们将强大的预训练音频分类器与结构化时空先验分布结合起来，并通过有界的逐样本门控函数实现融合，从而提升鲁棒性，同时保留仅音频回退能力和模块化预训练的优势。

3. 方法

3.1. 符号与证据来源

我们考虑一个多类分类问题，使用以下符号：

- $x \in \mathcal{X}$：音频观测，以 3 秒对数梅尔频谱图表示，
- $s \in \mathcal{S}$：时空上下文（纬度、经度和时间），
- $y \in \mathcal{Y}$：物种标签，
- $p_\theta(y \mid x)$：基于音频的分类器，
- $p_\psi(y \mid s)$：时空（生态）分类器，
- $p_\omega(y \mid x,\ s)$：融合分类器。

虽然我们的应用是生物声学分类，但所提出的融合框架通常适用于任何给定目标标签且具有多个条件独立证据来源的场景。

3.2. 对数线性融合族

我们采用如下形式的逐样本对数线性融合模型

$$\log \tilde{p}_\omega(y \mid x, s) = \log p_\theta(y \mid x) + \omega(x, s)\ \log p_\psi(y \mid s), \tag{6}$$

其中，$\tilde{p}_\omega$ 表示未归一化的对数分数。通过对 y应用 Softmax，可得到归一化后验概率 $p_\omega(y \mid x,\ s)$。标量 $\omega(x,\ s)\ \geq 0$ 控制时空模型在每个样本上的影响。我们使用带有可学习尺度的有界变换来参数化 $\omega(x,\ s)$：

$$\omega(x, s) = \omega_{\max}\, \sigma\big(g_\phi(u(x, s))\big) + \epsilon,$$

其中， $\sigma(\cdot)$表示 Sigmoid 函数， $\epsilon > 0$ 确保数值稳定性，而 $\omega_{\max}$ 是一个可训练参数，其约束范围为 $[\epsilon, 10]$。这种设计使模型能够根据数据自适应地学习合适的上下文影响范围，同时避免融合权重退化或无界。当 $\omega(x,\ s) = 0$ 时，融合分类器会精确还原纯音频模型 $p_\theta(y \mid x)$。

这种融合有意采用非对称方式：对于所有样本，音频分类器都会在对数空间中提供加性贡献；而上下文模型只有在被判定为具有信息量时，才会调节后验概率。这体现了两种来源互补的归纳偏置。音频观测通常具有较强的判别能力，但容易混淆声学相似物种；时空模型则通过编码生态可行性，提供粗略但稳健的约束。

3.3. 自适应门控：可靠性与信息量

为了自适应地对上下文证据加权，我们通过双层多层感知机参数化并学习门控函数 $\omega(x,\ s)$。门控网络以特征向量 $u(x,\ s)$ 为输入，该向量概括了不确定性、置信度和上下文结构：

- **音频特征**：

$$f_{\text{audio}}(x) = \left[\max_y p_\theta(y \mid x),\ H(p_\theta(y \mid x)),\ p_\theta(y_1 \mid x) - p_\theta(y_2 \mid x)\right],$$

其中 $y_1,\ y_2$ 表示排名前两位的预测类别。

- **时空特征**：

$$f_{\text{prior}}(s) = \left[\max_y p_\psi(y \mid s),\ H(p_\psi(y \mid s)),\ p_\psi(y_1 \mid s) - p_\psi(y_2 \mid s)\right].$$

- **元数据特征**：

$$f_{\text{meta}} = \left[\sin\left(\tfrac{2\pi d}{365}\right), \cos\left(\tfrac{2\pi d}{365}\right), \sin\left(\tfrac{2\pi h}{24}\right), \cos\left(\tfrac{2\pi h}{24}\right), \tfrac{\text{lat}}{90}, \tfrac{\text{lon}}{180}\right].$$

拼接后的特征向量 $u(x,\ s)$ 经过带有 ReLU 激活和 Dropout 的双层多层感知机 $g_\phi(u)$。其输出经过 sigmoid 压缩并由 $\omega_{\max}$ 缩放，得到 $\omega(x,\ s) \in [0, \omega_{\max}]$。

除门控参数外，我们还联合学习应用于音频逻辑值的温度参数 T，并引入一个小常数 ϵ 以确保数值稳定性。最终得到的融合模型为

$$\log \tilde{p}(y \mid x, s) = \frac{\log p_\theta(y \mid x)}{T} + \omega(x, s) \log\big(p_\psi(y \mid s) + \epsilon\big), \tag{7}$$

并通过对 y 应用 Softmax，得到 $p(y \mid x, s)$。

避免门控塌缩。为防止自适应门控塌缩到平凡解（例如 $\omega(x, s) \equiv 0$），我们在训练过程中引入方差正则化项：

$$\mathcal{L}_{\text{var}} = -\lambda_{\text{var}} \operatorname{Var}_{(x,s)\sim\mathcal{B}}[\omega(x, s)],$$

其中，方差是在每个小批量 $\mathcal{B}$ 上计算的。该项促使 $\omega(x,\ s)$ 在不同输入之间取非恒定值，从而促进真正的自适应性，而不是全局抑制上下文证据。我们将 λ_{var} 设置得足够小，并依赖基于验证的模型选择，使正则化项促进非平凡的自适应性，同时不会在上下文有帮助时压制对 $\omega(x,\ s) \approx 0$ 的经验偏好。

3.4. 决策理论安全性与鲁棒性

所提出的融合框架旨在对模型设定错误或较弱的上下文模型保持鲁棒性。其一项关键结构属性是可恢复性：对于任意输入（$x,\ s$），将$\omega(x,\ s) = 0$ 设定为相应值，即可精确恢复纯音频分类器。因此，该融合族包含一个明确的仅音频回退机制，并不要求依赖上下文信息。

通过对融合权重 $\omega(x,\ s) \le \omega_{\max}$设置上界，可以进一步控制时空模型的影响，从而防止上下文证据病态地主导后验概率。因此，即使上下文预测具有很高的置信度但实际上错误，也无法压倒强有力的声学证据。

自适应门控能够实现对上下文信息的选择性纳入。当时空模型提供的信息量不足时（例如分布接近均匀或熵较高），学习到的门控值会趋近于零。当上下文证据具有信息量且与音频预测一致时，提高 $\omega(x,\ s)$会通过抑制生态学上不合理的替代选项，使后验概率分布更加尖锐。

从决策理论视角来看，融合分类器对应于在有界的、依赖样本的对数似然项组合下最小化期望对数损失。可恢复性保证了纯音频假设在模型类中始终可用，而有界门控则在实践中减轻了对上下文证据的有害过度依赖。形式化条件见附录。

最后，在整个融合训练过程中冻结音频分类器，可以保留底层声学决策函数，确保预测中的任何变化都仅由融合机制引起，而非音频模型本身的修改。

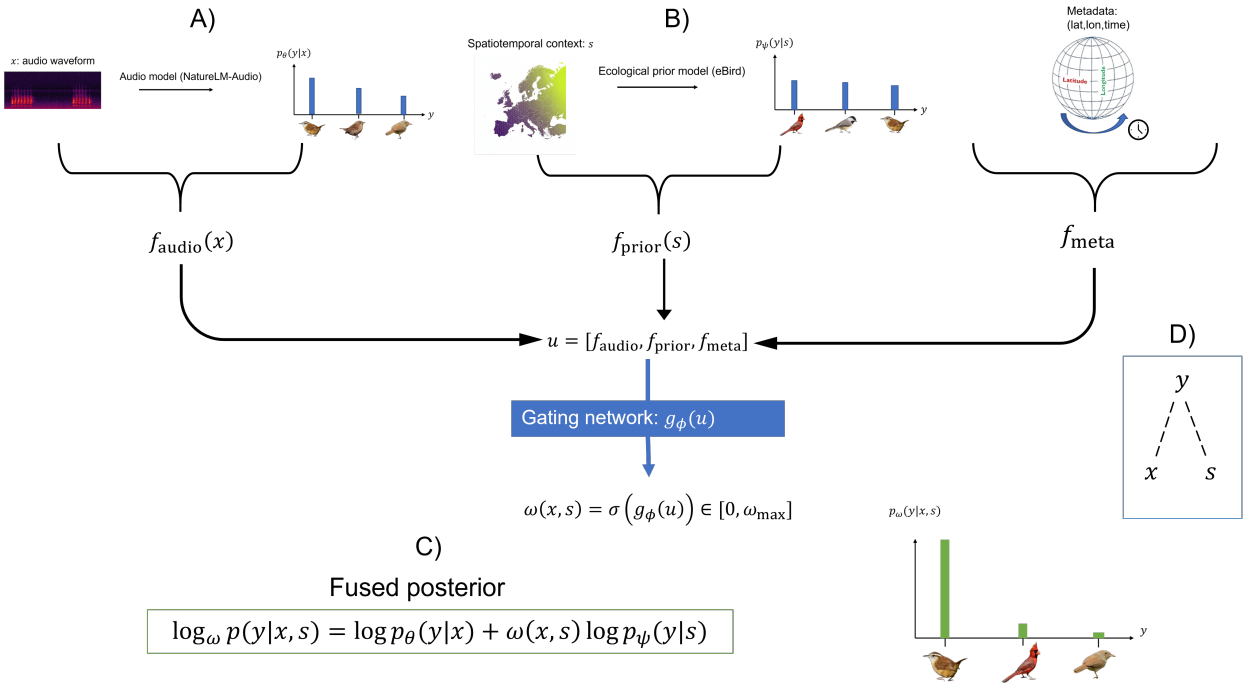


图 2. **FINCH：自适应对数线性证据融合。(A)** 音频编码器和线性头根据输入频谱图生成声学后验 $p_\theta(y \mid x)$。**(B)** 时空模型将上下文 s （位置和时间）映射为先验分布 $p_\psi(y \mid s)$。**(C)** 门控网络根据不确定性和信息量特征计算每个样本的非负权重 $\omega(x, s)$，生成未归一化的对数分数 $\log\ \tilde{p}_\omega(y \mid x, s) = \log\ p_\theta(y \mid x) + \omega(x, s) \log\ p_\psi(y \mid s)$，再通过对 y 进行 Softmax 归一化，得到 $p_\omega(y \mid x, s)$。当 $\omega(x, s) = 0$ 时，便可恢复纯音频模型。**(D)** 有向无环图说明了这一假设：给定标签 y 后，x 和 s 条件独立，这为乘法融合提供了依据。

4. 实验设置

4.1. 数据集与时空先验分布

我们在两个大规模生物声学基准上评估了该方法：康奈尔鸟鸣识别（CBI）数据集和 BirdSet。CBI 源自 eBird 公民科学语料库，并作为 Kaggle 竞赛的一部分发布（Howard et al.，2020年）。该数据集包含标注了物种标签及相关元数据（包括纬度、经度和日期）的短音频录音。CBI 共包含 264 种鸟类。BirdSet 是用于鸟类音频分类的大规模基准，采用标准化的数据划分和评估协议（Rauchetal.，2025年）。

我们采用 eBird 状况与趋势项目（AdaSTEM）提供的物种级每周时空丰度先验分布（Fink 等人，2010；2014年；Johnston 等人，2015），并按地点和日期进行查询。为提高训练效率，先验分布会预先计算，并在推理时直接使用。详情见附录 A5。

每个基准的上下文模型。我们根据数据集采用不同的时空模型。在 CBI 上，$p_\psi(y \mid s)$ 由外部 eBird Status & Trends（AdaSTEM）先验分布根据地点和日期查询得到（附录 A5）。

在 BirdSet 上，无法获取 AdaSTEM 先验分布；因此，我们在 BirdSet 训练集划分上训练一个轻量级纯元数据MLP，利用门控网络所用的相同时间/地点编码，得到学习得到的上下文预测器 $p_\psi(y \mid s)$（附录 A1）。

4.2. 音频编码器与分类头

在音频表征学习中，我们使用从 NatureLM-Audio 模型中提取的 BEATs 编码器（Robinson 等人，2025年）。尽管 NatureLM-Audio 是针对音频语言任务进行端到端训练的，但先前研究表明，将由此得到的 BEATs 编码器作为冻结的特征提取器使用时，在生物声学基准测试上可达到最先进的性能（Miron 等人，2025）。

在整个训练过程中，音频编码器的所有参数均保持冻结。该编码器生成维度为 4096 的嵌入，随后沿时间维度进行均值池化，并输入线性分类头

$$y = Ah(x) + b, \tag{8}$$

其中 $A \in \mathbb{R}^{C \times 4096}$ 和 $b \in \mathbb{R}^C$，C 表示物种数量。对于 CBI，$C = 264$。

4.3. 融合模型与训练流程

为将表征学习与证据融合解耦，我们采用三阶段训练流程，期间音频编码器始终保持冻结。

阶段 1：纯音频训练。我们首先基于冻结的音频嵌入训练一个线性分类器，得到仅音频预测器 $p_\theta(y \mid x)$，从而建立强大的声学基线。

阶段 2：固定权重融合。随后，我们引入一个固定的标量融合权重 $\omega \geq 0$，并将其应用于所有样本，以结合音频和时空证据。此阶段将学习一个全局校准融合基线。

阶段 3：自适应门控融合。最后，我们用一个自适应门控网络（第 3.3 节）替代全局标量权重，该网络会预测一个非负、依赖于样本的融合权重 $\omega(x, s)$。门控网络初始化为与固定权重解相匹配，并通过训练实现逐样本自适应。

表 4 总结了各训练阶段的模型组件和可训练参数。

5. 结果

5.1. 主要基准测试结果（汇总）

表 2 总结了在 CBI 和 BirdSet 子集上的性能，并将 FINCH 与强大的纯音频基线及现有生物声学系统进行了比较。对于 CBI，FINCH 使用源自 eBird Status & Trends（AdaSTEM）的外部时空先验；对于 BirdSet，则使用在 BirdSet 训练集划分上训练得到的纯元数据 MLP 先验。

CBI。 在 CBI 上，按照线性探测协议，FINCH 在所有评估方法中取得了最高测试准确率，通过自适应融合将准确率从纯音频基线的 0.806 提高到 0.826。固定权重融合仅带来微小提升（0.808），这表明当上下文可靠性随输入而变化时，经过全局校准的融合并不足够。时空先验单独使用时表现不佳（准确率 $= 0.030$），这证实观察到的提升源自选择性整合，而非单纯依赖上下文。尽管 FINCH 只在冻结的预测器之上引入了轻量级门控网络，但在相同协议下，其表现优于纯音频模型和联合音频—上下文系统。</output>

BirdSet. 在 BirdSet 子集上，FINCH 在多项评估指标上达到或优于强纯音频基线，包括检索（AUROC）、检测（cmAP），以及分类（Top-1准确率）。鉴于上下文预测器只是一个在有限监督下训练的简单纯元数据多层感知机，这些结果尤为显著。尽管改进幅度因子集和指标而异，FINCH 始终表明，自适应证据加权能够利用弱且异构的上下文信号，同时不会相对于纯音频模型降低性能。

5.2. 定性分析：实践中的自适应融合

为补充汇总结果，我们在图中展示了两个具有代表性的测试时示例，3说明 FINCH 如何自适应地平衡声学证据与时空证据。在第一个示例中，纯音频分类器为真实物种赋予的概率较低，将几个声学上相似的备选类别排在更前面，而时空模型则高度确信，且给出的地理定位十分具体。FINCH 赋予较大的门控权重，通过增强可靠的上下文证据，恢复了正确类别。在第二个示例中，声学模型和时空先验单独使用时都无法得出正确结果：先验分散在多个物种上，而音频模型则十分确信但预测错误。此时，自适应融合抑制了声学上占主导地位、但在生态学上不合理的类别，尽管两个边缘预测器都失效，仍得出了正确的后验预测。值得注意的是，这两个案例中学得的门控权重存在显著差异，表明固定权重融合无法同时得到这两种结果。这些示例凸显了自适应可靠性估计在权衡误差方面的作用，也与附录A2中的条件独立性分析相吻合。

6. 讨论

本研究表明，对近似条件独立的证据源进行自适应对数线性融合，能够相对于独立预测器和固定权重融合方案持续取得改进。在大规模生物声学基准上，FINCH 在评估子集上优于强纯音频基线，证明了在弱且异构的上下文信号下自适应证据融合的有效性。

这些增益源于按样本选择性地整合上下文证据，而非在各模态之间强制进行全局校准。

所提出框架的一个关键特性是其决策理论安全性。在这一 formulation 下， 仅音频 分类器作为融合族中的显式后备方案。当上下文证据缺乏信息或具有误导性时，学习到的门控函数会被鼓励抑制其影响，从而降低相对于似然模型发生性能下降的风险。这种行为在总体指标和定性示例中均得到了实证体现，

表 *1.* **CBI 和 BirdSet 子集上的主要结果。** 对于 BirdSet，我们报告四个子集（PER、NES、UHH、SSW）上的检索AUROC（R）、检测cmAP（m）和 Top-1准确率（A），格式为 *R/m/A*。对于 CBI，我们报告 Top-1准确率以及对应的训练方案（“方法”）。FINCH 在 CBI 上使用外部 AdaSTEM先验，在 BirdSet 上使用学习得到的仅元数据MLP先验。

模型	CBI		BirdSet（子集指标：R / m / A）			
	方法	Acc	PER（R/m/A）	NES（R/m/A）	UHH（R/m/A）	SSW（R/m/A）
Audio ProtoPNet-5	–	–	0.790 / 0.300 / 0.590	0.930 / 0.380 / 0.520	0.870 / 0.310 / 0.490	0.970 / 0.420 / 0.660
BirdMAE-L	–	–	0.820 / **0.350** / 0.600	0.910 / **0.410** / 0.520	0.820 / 0.300 / 0.420	0.930 / 0.410 / 0.700
BirdMAE-L (PP)	–	–	0.820 / 0.310 / 0.590	0.930 / 0.380 / 0.470	0.830 / 0.300 / 0.360	0.940 / 0.380 / 0.620
Perch 2.0 – 峰值选择	LP	0.785	0.802 / 0.255 / 0.528	0.948 / 0.382 / 0.504	0.901 / 0.375 / 0.462	**0.974** / 0.455 / 0.782
Perch 2.0 – 随机	LP	0.792	0.786 / 0.232 / 0.535	**0.953**/ 0.403 / 0.562	0.912 / 0.380 / 0.595	0.973 / **0.469** /**0.789**
仅音频 ($p_\theta(y \mid x)$)	LP	0.806	–	–	–	–
仅先验 ($p_\psi(y \mid s)$)	LP	0.030	–	–	–	–
FINCH（本文方法）	LP	**0.826**	**0.824**/ 0.232 / 0.429	0.936 / 0.245 / **0.679**	**0.927/ 0.536 /0.747**	0.642 / 0.025 / 0.688

其中，自适应融合会在上下文可靠时纠正置信度较高的音频错误，而在上下文不可靠时避免病态支配。

门控机制的设计涉及若干建模选择，我们采取务实而非穷尽的处理方式。在这项工作中，门控网络基于从预测分布中得出的精简不确定性与置信度摘要统计量运行。我们也可以探索其他特征构造方式或表达能力更强的门控架构，但结果表明，即使是轻量级模型，也足以捕捉上下文可靠性的显著异质性。

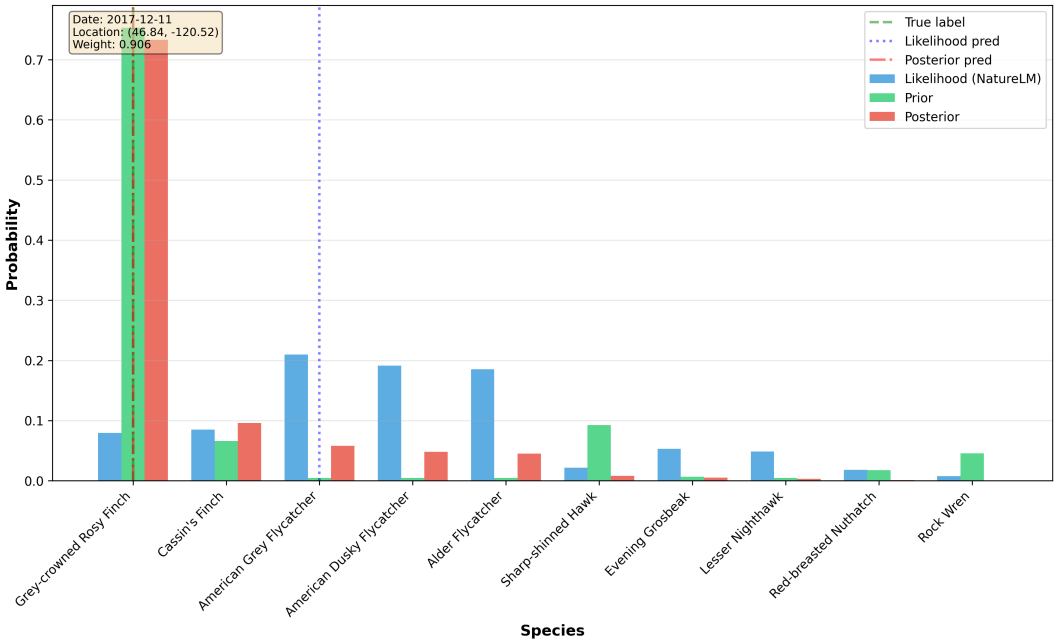
我们还注意到，本研究使用的时空先验分布单独发挥作用时相对较弱。对于 264 个类别，随机猜测的准确率约为 0.38%，而先验分布的准确率约为 3%。这一差距证实，观察到的性能提升源于选择性整合，而非单纯依赖上下文。更强的生态模型或许能进一步放大自适应融合的效益，但当前结果表明，即使是粗略的上下文信号，只要以稳妥的方式加以整合，也能显著改善预测。

CBI 数据集和时空先验分布均源自更广泛的 eBird 生态系统，因此自然会引发关于潜在数据纠缠的问题。值得注意的是，AdaSTEM 先验是在独立的公民科学观测清单上训练的，这些清单汇总自大范围的空间和时间窗口，并未纳入 CBI 中使用的音频记录或竞赛数据划分。此外，该先验分布的单独准确率很低，这表明它发挥的是一种较弱的生态可行性信号，而非物种标签的替代指标。

一个重要的局限涉及假阳性。尽管情境信息能够抑制不符合生态学规律的预测，但此处采用的基于逻辑回归的先验分布会在空间和时间上进行插值，从而为观测稀少或没有观测的区域赋予非零概率。因此，假阳性有所减少，但并未完全消除；更严格的可行性约束可能会以牺牲数据集偏移下的鲁棒性为代价。

最后，对数线性融合所依据的条件独立假设本质上很强，在真实数据中并不能严格成立。我们强调，此处将条件独立用作建模动机，而非严格假设。我们的实证分析表明，音频特征与时空上下文之间存在残差依赖，但这种依赖较弱，且在不同类别间具有异质性。这一情形与 FINCH 的预期运行条件相符：独立性在足够程度上成立，从而为乘法融合提供动机，而自适应门控可在无需显式建模音频—上下文交互的情况下缓解偏差。当条件依赖较强时，完全联合模型可能更为可取，但它们会带来显著更高的训练与数据需求。

更广泛地说，FINCH 最好被视为一种通用融合框架，而非生物声学专用解决方案。该形式化表述适用于任何这样的场景：多个判别性预测器就共享目标提供互补且近似独立的证据；将该方法扩展到其他领域，仍是未来工作中一个有前景的方向。



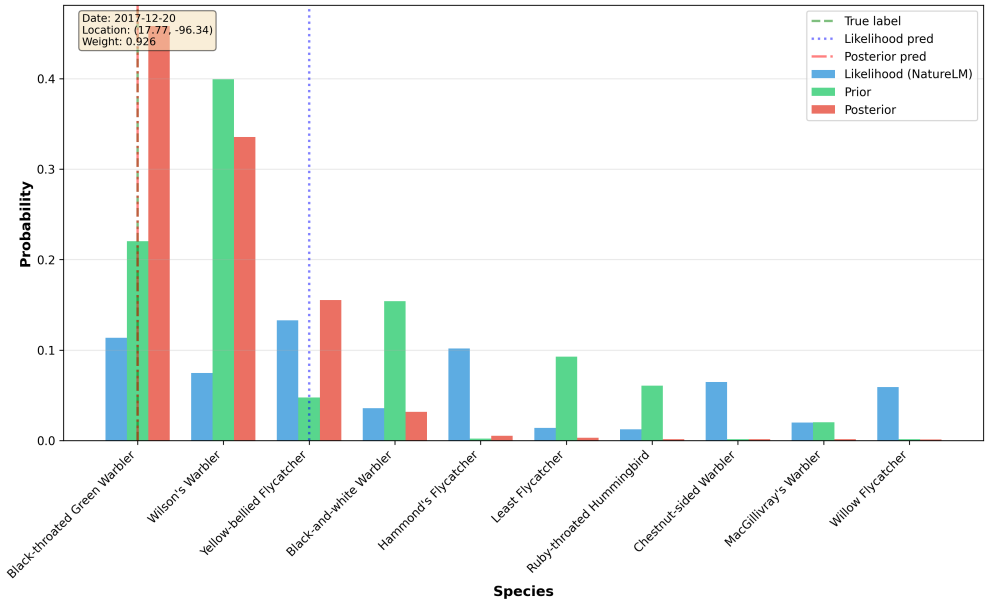


图*3.* **测试时自适应融合的启发性示例。上：**声学模型对真实类别（灰冠玫瑰雀）赋予较低概率，并将若干声学相似物种排在更前。时空模型置信度很高且具有明确的地理特异性，融合后验以较高的门控权重正确恢复了先验分布的最高预测。**下：**无论是声学模型（最高预测：黄腹霸鹟）还是时空先验（在多种林莺上较为分散），单独来看都不正确。然而，自适应融合抑制了声学上占主导但生态学上不合理的类别，从而得到正确的后验预测（黑喉绿林莺）。在这两种情况下，学习得到的门控权重在不同输入之间差异显著，这说明了固定权重融合为何无法恢复这些结果。

致谢

影响声明

本研究提出了一个通用框架，可在近似条件独立的前提下自适应融合多个来源的概率证据。虽然我们的实验聚焦于生物声学物种分类，但所提出的方法适用于众多领域，在这些领域中，不同类型的预测器能够提供互补信息。

在生态监测和生物多样性评估中，更有效地整合声学信息与上下文信息，有助于更准确地大规模分析物种分布和时间趋势。同时，自动化系统生成的预测不应被视为物种存在的确凿证据，尤其是在保护或政策相关场景中，除非经过适当的人工验证。

更广泛而言，所提出的融合机制旨在通过限制不可靠或误导性上下文信号的影响，并保留安全回退至主要证据的机制，增强系统的稳健性。这一设计选择有助于降低在现实环境中部署决策支持系统时，过度自信或存在偏差的辅助模型所带来的风险。

参考文献

Berger, A. L., Della Pietra, S. A., and Della Pietra, V. J. 自然语言处理的最大熵方法. 麻省理工学院出版社，1996年。

Bishop,C. M. 模式识别与机器学习. 施普林格，2006年。

Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., and Wei, F. BEATs：采用声学分词器进行音频预训练，2022年。网址 https://arxiv.org/abs/2212.09058。

Choudhary, S., Karthik, C. R., Lakshmi, P. S., and Ku- mar, S. Lean：轻量高效的音频分类网络。载于*2022*年电气电子工程师学会第*19*届印度理事会国际会议（*INDICON*），第1–6页。电气电子工程师学会，2022年11月。doi: 10.1109/indicon56171.2022. 10039921。网址 http://dx.doi.org/10.1109/INDICON56171.2022.10039921。

Fink, D.、Hochachka, W. M.、Zuckerberg, B.、Winkler, D. W.、Shaby, B.、Munson, M. A.、Hooker, G.、Riedewald, M.、Sheldon, D. 与 Kelling, S.。面向大范围调查数据的时空探索模型。《生态学应用》，20(8):2131–2147，2010。

Fink, D.、Damoulas, T.、Bruns, N. E.、La Sorte, F. A.、Hochachka, W. M. 与 Kelling, S.。自适应时空探索模型揭示鸟类出现的复杂模式。《生态学应用》，23(5):1067–1086，2013。

Fink, D.、Damoulas, T.、Bruns, N. E.、La Sorte, F. A.、Hochachka, W. M.、Gomes, C. P. 和 Kelling, S. 众包与生态学相遇：半球尺度的时空物种分布模型。《*AI Magazine*》，35(2)：19–30，2014年。doi：https://doi.org/10.1609/aimag.v35i2.2533。
U
R
L
https://onlinelibrary.wiley.com/doi/abs/10.1609/aimag.v35i2.2533。

Gal, Y. 和 Ghahramani, Z. 将 Dropout 作为贝叶斯近似：在深度学习中表示模型不确定性。载于国际机器学习会议，2016年。

Geifman, Y. 和 El-Yaniv, R. 深度神经网络的选择性分类。载于神经信息处理系统进展，2017年。

Genest, C. 和 Zidek, J. V. 概率分布的组合：一篇评论与注释书目。统计科学，1(1):114–135，1986年。

Ghani, B.、Denton, T.、Kahl, S. 和 Klinck, H. 全球鸟鸣嵌入助力生物声学分类的卓越迁移学习。科学报告，13(1): 22876，2023年。

Hagiwara, M. AVES：基于自监督的动物发声编码器，2022年。网址 https://arxiv.org/abs/2210.14493。

Hagiwara, M.、Hoffman, B.、Liu, J.-Y.、Cusimano, M.、Effenberger, F. 和 Zacarian, K. BEANS：动物声音基准，2022年。网址 https://arxiv.org/abs/2210.12300。

Hendrycks, D. 和 Gimpel, K. 一种用于检测神经网络中误分类样本和分布外样本的基线方法。载于国际学习表征会议，2017年。

Heskes, T. 对数意见池中加权因子的选择。神经信息处理系统进展，10，1998年。

Hinton, G. E. 通过最小化对比散度训练专家乘积。《*Neural Computation*》，14(8):1771– 1800，2002年。

Howard, A., Klinck, H., Dane, S., Kahl, S., tom denton, and Denton, T. 康奈尔鸟鸣识别。https://kaggle.com/competitions/birdsong-recognition，2020年。Kaggle。

Jacobs, R. A., Jordan, M.I., Nowlan, S. J., and Hinton, G. E. 局部专家自适应混合。《*Neural Computation*》，3(1):79–87，1991年。

Johnston, A., Fink, D., Reynolds, M. D., Hochachka, W. M., Sullivan, B. L., Bruns, N. E., Hallstein, E., Merrifield, M. S., Matsumoto, S. 和 Kelling, S. 丰度模型改进保护资源的时空优先排序。《生态学应用》,25(7):1749–1756,2015. doi: https://doi.org/10.1890/14-1826.1.URL
h
t
t
p
s://
esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/14-1826.1.

Jordan, M. I. 和 Jacobs, R. A. 分层专家混合与期望最大化算法。《*Neural Computation*》, 6(2): 181–214, 1994.

Kahl, S., Wood, C. M., Eibl, M. 和 Klinck, H. BirdNET: 一种用于鸟类多样性监测的深度学习解决方案。《生态信息学》, 61:101236, 2021.

Kendall, A. 和 Gal, Y. 计算机视觉的贝叶斯深度学习需要哪些不确定性？载于神经信息处理系统进展，2017年。

Kittler, J.、Hatef, M.、Duin, R. P. W. 和 Matas, J. 关于分类器组合。电气电子工程师学会模式分析与机器智能汇刊，20(3):226–239，1998年。

Kuncheva, L.I. 模式分类器组合：方法与算法。Wiley, 2004年。

Lauha, P., Somervuo, P., Lehikoinen, P., Geres, L., Richter, T., Seibold, S., 和 Ovaskainen, O. 只需少量本地数据，领域专用神经网络便能提升自动鸟鸣识别效果。生态学与进化方法, 13(12):2799–2810，2022年10月。ISSN 2041-210X。doi: 10.1111/ 2041-210x.14003。URLhttp://dx.doi.org/10.1111/2041-210X.14003。

Madhusudhana, S., Shiu, Y., Klinck, H., Fleishman, E., Liu, X., Nosal, E.-M., Helble, T., Cholewiak, D., Gille- ˇspie, D., Sirovi´c, A., 和 Roch, M. A. 利用时间背景改善动物叫声序列的自动检测。英国皇家学会界面期刊, 18(180):20210297，2021年7月。ISSN 1742-5662。doi: 10.1098/rsif.2021.0297。URLhttp://dx.doi.org/10.1098/rsif.2021.0297。

Miron, M., Robinson, D., Alizadeh, M., Gilsenan-McMahon, E., Narula, G., Chemla, E., Cusimano, M., Effenberger, F., Hagiwara, M., Hoffman, B., Keen, S., Ki m, D., Lawton, J., Liu, J.-Y., Raskin, A., Pietquin, O., 和 Geist, M. 生物声学编码的关键因素，2025年。URL https://arxiv.org/abs/2508.11845。

Nolasco, I., Singh, S.,Morfi, V., Lostanlen, V., Strandburg-Peshkin,A., Vida˜na-Vila, E., Gill, L., Pamuła, H., Whitehead, H., Kiskin, I., Jensen, F. H., Morford, J., Emmerson, M. G., Versace, E., Grout, E., Liu, H., Ghani, B., 和 Stowell, D. 从五个示例中学习检测动物声音。生态信息学, 77:102258，2023年11月。ISSN 1574-9541。doi: 10.1016/j.ecoinf.2023.102258。
U
RLhttp://dx.doi.org/10.1016/j.ecoinf.2023.102258。

Rauch, L.、Schwinger, R.、Wirth, M.、Heinrich, R.、Huseljic, D.、Herde, M.、Lange, J.、Kahl, S.、Sick, B.、Tomforde, S. 和 Scholz, C. Birdset：用于鸟类生物声学音频分类的大规模数据集，2025年。网址 https://arxiv.org/abs/2403.10380。

Robinson, D., Miron, M., Hagiwara, M., Weck, B., Keen, S., Alizadeh, M., Narula, G., Geist, M., and Pietquin, O. Naturelm-audio：面向生物声学的音频语言基础模型，2025年。网址 https://arxiv.org/abs/2411.07186。

van Merri¨enboer, B., Dumoulin, V., Hamer, J., Harrell, L., Burns, A., and Denton, T. Perch 2.0：生物声学的麻鹅启示，2025年。网址 https://arxiv.org/abs/2508.04665。

Wang, H.、Xu, Y.、Yu, Y.、Lin, Y.和Ran, J. 一种基于声学特征的大量鸟类物种识别高效模型。《动物》，12(18): 2434，2022年9月。ISSN 2076-2615. doi: 10.3390/ani12182434.
URLhttp://dx.doi.org/10.3390/ani12182434.

表 2. 大规模生物声学分类数据集上的模型结果：BirdSet（Rauch等人，2025年）、CBI（Howard等人，2020年）。（方法）“预训练”：预训练分类头，“微调”：全量微调，“线性探测”：线性探针，“提示调优”：原型探针，“0”：零样本。（BirdSet）与Perch2.0（van Merri¨enboer等人，2025年）保持一致，我们报告ROC曲线下面积（AU-ROC）和类别平均精度（cmAP）。（BEANS）同样，此处我们记录模型准确率（Acc）和平均AP（mAP）。对于CBI，FINCH使用外部AdaSTEM先验；对于BirdSet，FINCH使用在BirdSet 训练集划分上训练的学习得到的仅元数据MLP先验。最佳结果以**粗体**突出显示。次佳结果以下划线标出。

	BirdSet				CBI	
	方法	AUROC	cmAP	Acc	方法	Acc
Audio ProtoPNet-5	Pre	0.896	0.423	0.623	–	–
BirdMAE-L	FT	0.886	**0.440**	0.601	–	–
BirdMAE-L	PP	0.886	0.409	0.521	–	–
AVES-Bio	–	–	–	–	FT	0.598
BirdNet	–	–	–	–	FT	0.702
BioLingual	–	–	–	–	FT	0.744
NatureLM-Audio	–	–	–	–	0	0.778
Perch 2.0 - 峰值选择	Pre	0.907	0.430	0.619	LP	0.785
					PP	0.789
Perch 2.0 - 随机	Pre	**0.908**	0.431	**0.665**	LP	0.792
仅音频 ($p_\theta(y \mid x)$)	FT	0.833	0.261	0.636	LP	<u>0.806</u>
仅先验 ($p_\psi(y \mid s)$)	LP	0.466	0.027	0.461	LP	0.030
FINCH（我们的方法）	FT	0.832	0.260	<u>0.636</u>	LP	**0.826**

附录 A1：补充结果 我们报告了补充实验，以完善主要结果，并进一步刻画 FINCH 在采用其他上下文模型和融合策略时的表现。

BirdSet 学得的上下文先验分布。 BirdSet 无法访问外部生态学先验分布，例如 eBird Status & Trends。因此，在 BirdSet 实验中，我们直接根据时空元数据学习上下文预测器 $p_\psi(y \mid s)$。具体而言，我们使用与第 3.3 节相同的特征表示对元数据进行编码，

$$f_{\text{meta}}(s) = \left[\sin(2\pi d/365), \cos(2\pi d/365), \sin(2\pi h/24), \cos(2\pi h/24), \tfrac{\text{lat}}{90}, \tfrac{\text{lon}}{180}\right],$$

并在 BirdSet 训练集划分上训练一个小型多层感知机（MLP），仅根据 $f_{\text{meta}}(s)$预测物种标签。训练完成后，该模型被视为上下文预测器 $p_\psi(y \mid s)$，并在 FINCH 融合训练期间保持冻结，以匹配独立训练的专家这一设置。

表2 报告了 BirdSet 子集在多项评估指标上的结果。尽管仅使用元数据的先验分布单独来看效果较弱，自适应融合使 FINCH 在若干子集上达到或超过了强大的纯音频基线，尤其是在 Top-1准确率方面。这些结果表明，即使上下文信息是通过有限监督学习得到，且不同数据集之间存在显著异质性，FINCH 仍然有效。

CBI 上的固定权重与自适应融合。 我们进一步分析了在 CBI 数据集的一个子集上，将音频预测与 eBird 时空先验结合时，固定融合权重所产生的影响（表 3）。随着融合权重 ω的增加，性能单调下降，这反映出先验在不加区分地应用时信息量较低且熵较高。相比之下，自适应融合通过仅在上下文证据具有信息量时选择性地纳入该证据，显著优于所有固定权重设置。该实验凸显了全局融合权重的一项关键失效模式，并推动采用依赖输入的可靠性估计。

ω	Acc ↑	mAP ↑
0.0	64.28	0.333
0.2	63.91	0.328
0.4	64.12	0.332
0.8	60.22	0.325
1.6	59.76	0.316
2.0	57.72	0.311
自适应的	**66.10**	**0.352**

表 *3.* 不同 ω取值的消融研究。我们报告了在 CBI 的 1000 个样本上训练的模型的准确率和 mAP。固定 ω 用一个标量替换式 6 中的门控网络。（自适应）则学习动态加权特征证据。

附录A2：门控融合的决策理论安全性 我们表明，所提出的门控融合机制在决策理论上是安全的，符合风险约束的含义：它将纯音频分类器作为一个可恢复的特殊情况，因此不会被迫依赖具有误导性的上下文信息。

令 $\ell(y, \hat{p}) = -\log \hat{p}(y)$ 表示对数损失。考虑未归一化融合分数

$$\log \tilde{p}_\omega(y \mid x, s) = \log p_\theta(y \mid x) + \omega(x, s)\, \log p_\psi(y \mid s), \qquad \omega(x, s) \geq 0,$$

并通过对 y 应用 Softmax 定义归一化预测器 $p_\omega(y \mid x,\ s)$。

这一表述的关键结构性质是可恢复性：对于任意输入 $(x,\ s)$,

$$\omega(x, s) = 0 \quad \Longrightarrow \quad p_\omega(y \mid x, s) = p_\theta(y \mid x).$$

因此，纯音频分类器始终包含在门控融合所诱导的假设类中。

因此，对于任意数据生成分布，以及在 $(x,\ s)$ 上逐点评估的任意损失函数，$\omega(x,\ s)$ 的最优选择满足

$$\inf_{\omega(x,s)\geq 0} \mathbb{E}\big[\ell\big(y, p_\omega(\cdot \mid x, s)\big)\big] \ \leq\ \mathbb{E}\big[\ell\big(y, p_\theta(\cdot \mid x)\big)\big]\,.$$

具体而言，只要对于给定输入，时空预测器 $p_\psi(y \mid s)$ 不提供信息或具有误导性，选择$\omega(x,\ s) = 0$ 所达到的期望风险就与纯音频模型相同。

反之，如果上下文信息对某些输入具有非负的信息增益，那么就存在一些输入 $(x,\ s)$，使得 $\omega(x,\ s) > 0$ 相较于纯音频基线，能够严格降低期望对数损失。自适应门控网络能够学习这一选择性纳入规则的数据驱动近似。

因此，自适应门控融合 n 严格泛化了仅音频和固定权重融合，同时保留了安全的回退方案。

附录 A3：对数线性融合何时精确（条件独立情形） 我们阐明对数线性融合规则，并说明它与联合音频–上下文模型之间的关系。本分析旨在明确对数线性融合在理论上精确成立的条件；并不假设这些条件在实践中普遍成立。

假设给定类别标签 y时，音频特征 x 和时空上下文 s 条件独立，即

$$p(x, s \mid y) = p(x \mid y)\, p(s \mid y).$$

根据贝叶斯定理，后验概率满足

$$p(y \mid x, s) = \frac{p(x, s \mid y)\, p(y)}{p(x, s)} \propto p(x \mid y)\, p(s \mid y)\, p(y).$$

将该表达式改写为关于现有预测器的形式,

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{p(x)}, \qquad p(y \mid s) = \frac{p(s \mid y)p(y)}{p(s)},$$

可得

$$p(y \mid x, s) \propto \frac{p(y \mid x)\, p(y \mid s)}{p(y)}.$$

等价地，在相差一个与 y无关的加性常数的意义下，

$$\log p(y \mid x, s) = \log p(y \mid x) + \log p(y \mid s) - \log p(y) + \text{const.}$$

条件的影响 **al依赖**。如果违反条件独立，联合分布可以写成 s

$$p(x, s \mid y) = p(x \mid y)\, p(s \mid y)\, \kappa_y(x, s), \qquad \kappa_y(x, s) := \frac{p(x, s \mid y)}{p(x \mid y) p(s \mid y)}.$$

后验概率变为

$$p(y \mid x, s) \propto \frac{p(y \mid x)\, p(y \mid s)}{p(y)}\ \kappa_y(x, s),$$

或者，以对数形式（在相差一个类别无关常数的意义下），

$$\log p(y \mid x, s) = \log p(y \mid x) + \log p(y \mid s) - \log p(y) + \log \kappa_y(x, s) + \text{const.}$$

关键的是，log $\kappa_y(x,\ s)$ 显式依赖于 y，因此通常无法表示为 log $p(y \mid s)$ 的类别无关标量重加权。

对模型设计的启示。 当条件依赖较强时，要准确捕捉 $\kappa_y(x,\ s)$，需要一个联合预测器，同时接收 $(x,\ s)$。这类联合模型的表达能力严格更强，但需要配对监督，并显式学习音频—上下文交互。

相反，当条件独立性近似成立时，独立训练的预测器进行对数线性融合已足够，且（归一化后）为最优。FINCH 的设计基于这一情形：它保留条件独立性所隐含的后验概率形式，同时引入自适应的可靠性感知加权，以应对模型失配和异质不确定性。

附录 A4：条件依赖性的实证检验 上述理论依据依赖于条件独立假设。现在，我们通过实证评估这一假设在实践中成立的程度。

固定类别标签 y，我们通过训练一个线性预测器，检验音频嵌入 $e(x)$ 是否包含有关时空上下文统计量 s的信息

$$\hat{s} = f(e(x))$$

仅从音频嵌入预测时空先验。作为基线，我们将其与始终输出平均先验的常数预测器进行比较。

在具有足够样本的43个物种中，线性预测器在留出数据上并未持续优于基线。尽管置换检验表明，观察到的差异不太可能纯粹由偶然产生（$p < 10^{-3}$），但效应量较小（Cohen's $d = -0.261$），且解释方差可忽略不计或为负（均值 $R^2 = -0.34$）。仅有16.3%的类别相对基线呈现正向改进，且不足12%达到 $R^2 > 0.05$。

综合来看，这些结果表明，音频与时空信息并非严格条件独立，但其依赖较弱，且在不同类别间具有异质性。这些发现支持采用自适应门控：尽管条件独立并不精确，但偏离足够微弱且多变，因此有界的、依赖样本的融合机制优于固定模型或完全联合模型。

附录 A5：时空先验分布详情 时空先验分布基于 eBird 状况与趋势项目生成的每周相对丰度估计。这些估计采用自适应时空探索模型（AdaSTEM）生成，该模型由一组局部回归模型构成，旨在处理大规模公民科学数据中采样努力和检测偏差的异质性（Fink 等人，2010；2014年；Johnston 等人，2015）。丰度图以全球 27km $\times$ 27km 网格形式提供，时间分辨率为每周。每个丰度值表示在特定地点和时间由专家观察者检测到的个体预期数量

。原始值以无符号整数存储，取值范围为 $[0, 255]$，并线性缩放至 $[0, 1]$。

为避免引入训练时的瓶颈，我们为所有训练样本预先计算时空先验查找表。对于每个带有元数据（纬度$_i$，经度$_i$，日期$_i$）的音频片段，我们查询所有物种 y_j 对应的每周丰度先验$p(y_j \mid \text{纬度}_i, \text{经度}_i, \text{日期}_i)$。所得的 ($N_{\text{samples}}$，$N_{\text{species}}$) 形状矩阵存储在 HDF5 文件中，并按样本 ID 建立索引。训练期间，各批次通过索引查找获取先验值，平均访问时间为$0.22\,\mu$秒，而在线查询约需 0.01秒。在推理时，先验值直接从丰度图中查询，因为延迟并非性能关键因素。

附录 A6：训练配置与优化细节 除非另有说明，所有训练阶段均采用相同的优化器和训练配置。优化使用 AdamW 优化器，学习率为10^{-3}，权重衰减为 10^{-2}。模型训练 30 个训练轮数，批量大小为 96。

我们构建了一个分层验证集划分，其中包含 10% 的训练数据，并按物种标签进行分层。模型选择依据验证准确率进行。
采用余弦学习率调度，并在前 10% 的训练步数内进行线性预热。训练在图形处理器上使用混合精度运算（bfloat16）进行。
每个训练轮次结束后都会保存检查点。对于每个训练阶段，保留验证准确率最高的检查点用于评估。

阶段 2 从阶段 1 中表现最佳的检查点初始化，阶段 3 则从相应的阶段 2 检查点初始化。自适应门控网络的初始化方式使其输出与阶段 2 中学习到的固定标量融合权重相匹配。这种分阶段初始化能够稳定训练。

在所有阶段中，音频编码器始终保持冻结。在阶段 2 和阶段 3 中，音频分类头也保持冻结，仅更新融合参数。

所有实验均使用 PyTorch 实现。为确保可复现性，所有运行均固定随机种子。除非另有说明，除底层音频编码器固有的数据增强外，不应用其他数据增强。

表 4。汇总　了融合模型架构和训练配置。参数量假设为 $C = 264$ 个物种　.

组件	阶段 A	阶段 B
音频编码器	冻结	冻结
编码器架构	BEATs (NatureLM-Audio)	
编码器输出维度	4096	
音频分类器	$4096 \times C + C$	$4096 \times C + C$
温度参数	1	1
Epsilon 参数	1	1
标量权重	1	–
门控网络	–	2,945
可训练参数总数	≈ 1.08M	≈ 1.09M
优化器	AdamW	
学习率	1×10^{-3}	
批量大小	96	
训练轮数	30	
学习率调度	带预热的余弦	
精度	bfloat16	