arXiv:2602.04924v1 [cs.LG] 4 Feb 2026

知道何时作答：面向可靠音视频问答的自适应置信度精炼

Dinh Phu Tran [*1] **Jihoon Jeong** [*2 3] **Saad Wazir** [1] **Seongah Kim** [1] **Thao Do** [1] **Cem Subakan** [† 2 3] **Daeyoung Kim** [† 1]

摘要

我们提出了可靠的音视频问答（$\mathcal{R}$- 音视频问答）的一种形式化问题表述，在该表述中，我们更倾向于弃答，而不是错误作答。尽管近期的音视频问答模型具有较高的准确率，但识别自身可能出错的时机，以及随之而来的弃答能力，仍是尚未得到充分研究的领域。为填补这一空白，我们探索了多种方法，并提出自适应置信度细化（ACR），这是一种轻量级方法，用于进一步提升 $\mathcal{R}$-音视频问答的性能。我们的关键洞见是，最大 Softmax 概率（MSP）只有在强校准条件下才是贝叶斯最优的，而深度神经网络通常无法满足这一条件，尤其是在多模态模型中。ACR 并不替代 MSP，而是将其保留为主要置信度信号，并在判断 MSP 不可靠时施加输入自适应的残差校正。ACR 引入了两个学习头：i）残余风险头，用于预测 MSP 未能捕捉到的低幅度正确性残差；ii）置信度门控头，用于确定 MSP 的可信度。我们的实验和理论分析表明，在分布内、分布外以及数据偏差设置下，ACR 在三种不同的音视频问答架构上均稳定优于现有方法，为 $\mathcal{R}$-音视频问答任务奠定了坚实基础。代码和检查点将在录用后提供 于此处

1. 引言

音视频问答（AVQA）是一项重要的多模态推理任务，有助于开发能够在日常活动中帮助用户（包括有感官障碍的用户）的助手。Kumar等人，2022年；Patel等人，2025年）。为了在这些情境中真正发挥作用，音视频问答模型必须具备高可靠性（L¨utkenh¨oner & Basel，2013年；Asan等人，2020年）。提高准确率对于构建可靠的音视频问答系统至关重要，但仅有准确率还不够。所有模型都有固有局限，也容易做出错误预测。在实际应用中，给出错误答案可能会造成严重后果，尤其是对无法独立核实回复的感官障碍用户而言。

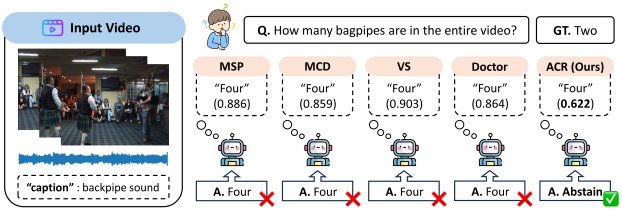


图 *1.* $\mathcal{R}$-音视频问答需要知道何时作答。在 MUSIC-AVQA 数据集的一个示例中，标准的基于置信度的基线方法（MSP、MCD、向量缩放和 Doctor）以高置信度给出错误答案，而我们的 ACR 能够有效识别不可靠预测，并给出较低的置信度分数。

避免给出错误答案的一种方法，是允许模型放弃预测，正如选择性预测研究所述（Geifman &El-Yaniv，2017年；2019年；Kamath等人，2020年；Xin等人，2021）。如图1所示，标准的基于置信度的基线方法（MSP、MCD、VS、Doctor）会以很高的置信度给出错误答案，这表明存在过度自信问题。相比之下，我们的 ACR 能有效降低置信度水平；当模型缺乏充分证据时，例如无法确定风笛的数量时，它会弃答，并提示模型回答“我不知道”，这与人类的决策方式相似。此外，以往关于选择性预测的研究主要关注单模态（Geifman & El-Yaniv，2019年）或双模态（Whitehead等人，2022年；Dancette 等人，2023）场景。相比之下，音视频问答涉及音频、视觉和语言输入之间的三模态交互。单独考虑时，每种模态都可能有效或属于分布内，但将它们组合起来可能颇具挑战，尤其是在音频和视觉信号对齐较弱甚至相互冲突时，这使得 $\mathcal{R}$-AVQA 更加困难。

[1]韩国科学技术院计算机学院，大韩民国 [2]拉瓦尔大学 [3]Mila-魁北克人工智能研究所。通讯作者：Dinh Phu Tran <phutx2000@kaist.ac.kr>。

预印本。*2026*年*2*月*6*日。

在本研究中，我们将音视频问答中的可靠性形式化为一个选择性预测问题（Geif-man & El-Yaniv，2017年），在此框架下，模型必须给出答案或选择不作答。该框架要求具备两项在音视频问答中尚未得到充分探索的能力：1）准确估计预测不确定性的能力；2）有效的答案弃答策略。在这一框架下，我们表明，现有音视频问答模型的可靠性仍有很大的提升空间。采用广泛使用的最大软最大概率（MSP）（Geif-man & El-Yaniv，2017年）方法时，我们发现，在严格的低风险约束下，覆盖率极低（为维持低风险而放弃回答过多问题）。例如，在MUSIC-AVQA 数据集上，最先进的 QA-TIGER 方法为将错误率维持在 1%，只能回答约**14%** 的问题，尽管其整体 AVQA 准确率超过 **77%**。在其他选择性预测方法中也观察到了类似的局限，这凸显出当前音视频问答模型在可靠性至关重要时的实际效用有限。

最后，我们提出自适应置信度细化（*ACR*），这是一种可学习的置信度估计框架，适用于 $\mathcal{R}$-音视频问答。尽我们所知，本研究首次提出了一种可学习的置信度估计方案，用于三模态音视频问答场景，专门用于判断模型何时应作答、何时应弃答。不同于依赖固定启发式规则或事后置信度估计的既有选择性预测方法，ACR 利用多模态特征和 softmax 前 logits，学习精炼 MSP 基线方法的置信度分数，以捕捉 MSP 常常忽略的不确定性信号，从而在复杂的跨模态交互下实现更可靠的置信度估计。我们将 ACR 与广泛使用的基线方法进行了全面评估，包括 MSP（Geifman& El-Yaniv，2017年）、蒙特卡洛丢弃（MCD）（Gal &Ghahramani，2016年）、基于校准的方法（向量缩放）（Guo 等人，2017年）以及 Doctor（Granese 等人，2021）。在三个音视频问答数据集和三种代表性骨干模型上，ACR 始终实现最先进的风险—覆盖率权衡，展现了其在构建可靠音视频问答系统方面的有效性和鲁棒性。

总而言之，我们的主要贡献如下：

1. 据我们所知，这是首项在选择性预测框架下分析和评估音视频问答模型可靠性的研究。2. 我们探索了多种引入弃答机制的方法，并提出自适应置信度细化（*ACR*），这是一种可学习的置信度估计方法，通过简单的线性融合，将 MSP 与学习得到的残余风险信号相结合，并由输入自适应加权机制进行调节。3. 我们为 $\mathcal{R}$-音视频问答任务建立了新的基准和强有力的基线方法，并证明 ACR 在三种种具有代表性的音视频问答骨干模型上均持续优于现有方法，涵盖分布内和分布偏移情形，包括分布外泛化和数据偏差场景。

2. 相关工作

音视频问答。近年来，音视频问答受到广泛关注，研究者提出了多种旨在提升答案准确率的方法（Schwartz 等人，2019年；Yun 等人，2021；Li 等，2022；2023；2024a；b；Kim 等，2025年）。近期方法（Li 等人，2024a；Kim 等，2025年）在多个标准基准上取得了出色的结果，包括 MUSIC-AVQA（Li 等，2022）、MUSIC-AVQA-R（Ma 等，2024）和 MUSIC-AVQA-v2.0（Liu 等，2024）数据集。然而，现有的音视频问答模型没有明确的弃答机制，并且被设计为针对给定音视频数据的每个输入问题都生成答案。为弥补这一不足，本文从可靠性的角度探索并构建音视频问答模型，将该任务界定为选择性预测问题，并赋予若干具有代表性的音视频问答模型（Li 等，2022；2024a；Kim 等，2025年）弃答的能力。我们的研究结果表明，尽管近期音视频问答模型准确率很高，但在可靠性设定下表现不佳。这凸显出提升其可靠性仍有很大空间。

选择性预测。在选择性预测中（Geifman & El-Yaniv，2017；Liu等，2019；Kamath等，2020），选择函数会为模型的每个预测分配一个置信度分数。选择性预测中一种常见的选择是最大软最大概率（MSP）（Geifman & El-Yaniv，2017）。通过将该分数与预定义的阈值进行比较，可以做出弃答决策：置信度低于阈值的预测会被拒绝，而高于阈值的预测则会被接受。选择性预测可分为外部方法和集成方法。外部方法在固定预测器之上构建 Selector，例如辅助预测头（Mielke等人，2022）、基于 LoRA 的适配（Chen等人，2023）或辅助视觉工具（Srinivasan等人，2024）。集成方法则在共享优化阶段联合训练预测器和 Selector，所使用的机制包括 SelectiveNet（Geifman & El-Yaniv，2019）或弃答类别（Liu等人，2019）。然而，从头开始进行联合训练往往会导致优化不稳定，需要专门的训练方案来同时优化这两个目标。此外，音视频问答模型通常使用针对音频、视频和文本的预训练编码器，这使得这些方法不适用于$\mathcal{R}$-音视频问答。不同于以往工作，我们没有用新的置信度估计器替代 MSP。我们提出了一种输入自适应校正机制，专注于优化 MSP；MSP 是选择性预测中的一种强基线方法（Jiang等人，2021）。更多相关工作见附录 C。

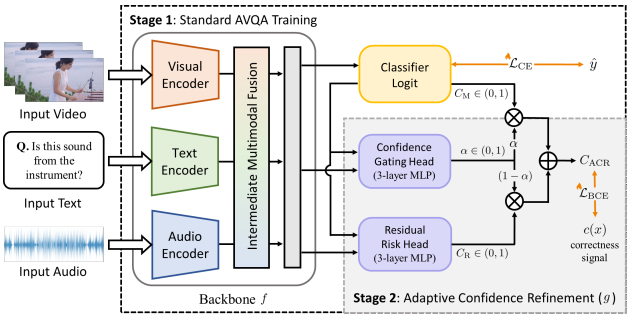


图2. 自适应置信度精炼框架概览，用于 $\mathcal{R}$-音视频问答。阶段一使用交叉熵损失训练标准的音视频问答模型。第二阶段冻结音视频问答模型，并训练两个额外的头部（紫色方块），通过二元正确性监督实现更可靠、随输入自适应的置信度。

3. 可靠的视听问答

3.1. 问题定义与符号表示

我们将标准的音视频问答任务扩展到可靠设置，在该设置中，模型可以选择回答或弃答。我们将这一设置称为可靠音视频问答($\mathcal{R}$-音视频问答)。

在标准音视频问答任务中，每个输入都由一对音视频数据和一个自然语言问题组成。形式化地，令 $\boldsymbol{x} = (\boldsymbol{v}, \boldsymbol{a}, \boldsymbol{q}) \in \mathcal{X}$ 表示一个视频 $\boldsymbol{v}$、一段音频 $\boldsymbol{a}$和一个问题 $\boldsymbol{q}$。答案空间为 $\mathcal{Y} = \{1, \ldots, K\}$，对应于预定义的候选答案集合；数据从分布$\mathcal{D}_{\mathcal{X},\mathcal{Y}}$中抽取，覆盖 $\mathcal{X} \times \mathcal{Y}$。选择性分类器定义为一对$(f, g)$，其中 $f\colon \mathcal{X} \to \mathbb{R}^K$ 是一个基础音视频问答模型，$g\colon \mathcal{X} \to \{0,1\}$ 是一个选择器函数，用于决定是进行预测还是弃权。形式化地，

$$(f,g)(\boldsymbol{x}) \triangleq \begin{cases} f(\boldsymbol{x}) & \text{if } g(\boldsymbol{x}) = 1, \\ \text{“abstain”} & \text{if } g(\boldsymbol{x}) = 0. \end{cases} \tag{1}$$

当 $g(\boldsymbol{x}) = 0$ 时，模型会拒绝回答输入的问题。这一点在音视频问答中尤为重要，因为不完整的跨模态证据或时间歧义可能导致模型自信地给出错误答案。在实践中，选择器函数 g 通常通过对实值置信度度量 $\gamma \in \mathbb{R}$ 应用可调阈值来实现：

$$g_{s,\gamma}(\boldsymbol{x}) \triangleq \mathbb{I}[s(\boldsymbol{x}) > \gamma], \tag{2}$$

其中 $s\colon \mathcal{X} \to \mathbb{R}$ 是一个用于估计预测 $f(\boldsymbol{x})$ 可靠性的置信度函数。理想的 $s(\boldsymbol{x})$ 会在 $f(\boldsymbol{x})$ 可能正确时给出较高分数，否则给出较低分数。

3.2. 评估指标

风险与覆盖率。 覆盖率是模型选择回答的输入问题所占的比例，而风险量化了模型在这部分问题上的预测误差(Geifman & El-Yaniv，2017年；Varshney 等，2022年）。

$$\text{Coverage:} \quad \mathcal{C}_{s,\gamma} \triangleq \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_{\mathcal{X},\mathcal{Y}}}[g_{s,\gamma}(\boldsymbol{x})], \tag{3a}$$

$$\text{Risk:} \quad \mathcal{R}_{s,\gamma} \triangleq \frac{\mathbb{E}_{(\boldsymbol{x},y) \sim \mathcal{D}_{\mathcal{X},\mathcal{Y}}}[\ell(f(\boldsymbol{x}), y) g_{s,\gamma}(\boldsymbol{x})]}{\mathcal{C}_{s,\gamma}}. \tag{3b}$$

如果 $f(\boldsymbol{x}) \neq y$，则 $\ell(f(\boldsymbol{x}), y)$ 等于 1，否则等于 0，表示 0/1损失（Geifman & El-Yaniv，2017年），用于衡量预测的 $f(\boldsymbol{x})$ 与真实答案 y 之间的准确率。风险 $\mathcal{R}_{s,\gamma}$ 表示模型选择回答的问题中的条件错误率。$\mathcal{C}@\mathcal{R}$ 指标定义为满足风险约束的最大覆盖率：

$$\mathcal{C}@\mathcal{R} \triangleq \max\{\mathcal{C}_{s,\gamma} \mid \mathcal{R}_{s,\gamma} \leq r\}. \tag{4}$$

公式 4 确保我们能够在指定的风险容忍度下，评估每种方法可能达到的最佳工作点 r。在实践中，用户愿意容忍的风险水平因应用场景而异。因此，我们使用目标风险水平下的覆盖率（$\mathcal{C}@\mathcal{R}$）来评估模型，包括 1%、5%、10% 和 20% 的风险。该指标直接体现了选择性模型的实际效用：在固定风险下，覆盖率越高，表示系统越可靠。我们还测量风险-覆盖率曲线下面积（AURC）（Geifman 等人，2018年），以评估所有工作点上的性能。为减轻对阈值 γ 在 g中的敏感性，我们报告每个风险水平下可达到的最大覆盖率，从而能够直接、公平地比较不同的选择函数。

期望校准误差。 选择性预测的目标是在覆盖率与风险之间实现平衡。该目标在很大程度上依赖于 Selector g，而 Selector 依赖于经过良好校准的函数 $s(\boldsymbol{x})$(Guo 等人，2017年）。为评估校准，我们使用期望校准误差（ECE），该指标量化模型预测的平均置信度与其平均准确率之间的差异。ECE 是评估$s(\boldsymbol{x})$可靠性的诊断工具。当模型经过良好校准时，置信度分数能够为正确概率提供可靠估计，从而使选择性风险随覆盖率变化而具有可预测性，并减少系统性过度自信。给定一个大小为N的测试集，将预测划分为 m分箱（$m \ll N$），其中B_j 表示分箱 j 中的索引集合，ECE 定义为：

$$\mathrm{ECE} \triangleq \sum_{j=1}^{m} \frac{|B_j|}{N} \Bigg| \underbrace{\frac{\sum_{i \in B_j} \mathbb{I}[\hat{y}_f(x_i) = y_i]}{|B_j|}}_{\text{Accuracy}} - \underbrace{\frac{\sum_{i \in B_j} s(x_i)}{|B_j|}}_{\text{Confidence}} \Bigg| \tag{5}$$

3.3. 选择函数基线

最大软最大概率。 一种广泛使用的选择性预测方法是最大软最大概率（MSP）（Geifman &El-Yaniv，2017年），这是一种简单且无需成本的方法，能够捕捉预测分布的峰值。MSP 置信度分数的公式如下：

$$C_{\mathrm{M}}(\boldsymbol{x}) = \max_{k \in \mathcal{Y}} \mathbb{P}(Y = k \mid \boldsymbol{x}) = \max_{k} \frac{\exp(l_k)}{\sum_j \exp(l_j)} \tag{6}$$

其中，l_k 是类别 k的对数几率。现在，我们将确立选择性预测中 MSP 最优性的理论条件。

定义 3.1 (校准)。 令 $\hat{y} = \text{Softmax}(f(\boldsymbol{x}))$ 为预测类别，$c(\boldsymbol{x}) = \mathbb{I}[\hat{y} = y^*] \in \{0, 1\}$ 为正确性：

i) 弱校准（边际校准）：如果置信度分数函数 $s(\boldsymbol{x}) \in [0, 1]$ 满足以下条件，则称其为弱校准：

$$\mathbb{P}[c(\boldsymbol{x}) = 1 \mid s(\boldsymbol{x}) = p] = p, \quad \forall p \in [0, 1].$$

总体层面性质：按置信度 p 分组的预测，其平均准确率为 p。

ii) 强校准（逐点校准）：如果置信度分数函数 $s(\boldsymbol{x})$ 满足以下条件，则称其为强校准：

$$s(\boldsymbol{x}) = \mathbb{P}[c(\boldsymbol{x}) = 1 \mid \boldsymbol{x}], \quad \forall \boldsymbol{x}.$$

样本层面性质：置信度反映了每个单独输入的实际正确概率。

命题 3.2 （均方误差降低意味着排序改进）。 设 $g^*(\boldsymbol{x}) = \mathbb{P}[c(\boldsymbol{x}) = 1 \mid \boldsymbol{x}]$ 为贝叶斯最优选择函数。对于任意置信度估计器 $s(\boldsymbol{x})$，将均方误差定义为 $\text{MSE}(s) = \mathbb{E}\big[\big(s(\boldsymbol{x}) - c(\boldsymbol{x})\big)^2\big]$。若估计器 s_1 的均方误差低于估计器 s_2，则 s_1 在均方误差意义上是对 g^* 更接近的近似：

$$\mathbb{E}\big[\big(s_1(\boldsymbol{x}) - g^*(\boldsymbol{x})\big)^2\big] < \mathbb{E}\big[\big(s_2(\boldsymbol{x}) - g^*(\boldsymbol{x})\big)^2\big]$$

因此，按 s_1 排序能更接近地逼近由 g^*诱导的贝叶斯最优排序。

注 3.1。 由于 $g^*(\boldsymbol{x})$ 诱导了选择性预测的最优排序（Geifman& El-Yaniv，2017年），任一估计器若能更好地逼近 g^*，便会产生更接近最优的排序，从而改进诸如AURC等指标。

定理 3.3（强校准下的 MSP 最优性）。考虑一个对每个输入 $\mathbb{P}(Y = k \mid \boldsymbol{x})$ 生成类别概率的分类器 $\boldsymbol{x}$。如果 *MSP* 置信度函数 $C_{\text{M}}(\boldsymbol{x}) = \max_k \mathbb{P}(Y = k|\boldsymbol{x})$ 是强校准的（定义 *3.1*），那么：*1*）*MSP* 等于贝叶斯最优选择函数：$C_{\text{M}}(\boldsymbol{x}) = g^*(\boldsymbol{x}) = \mathbb{P}[c(\boldsymbol{x}) = 1 \mid \boldsymbol{x}]$。*2*）按照 *MSP* 对示例进行排序，会得到选择性预测的贝叶斯最优排序。*3*）该排序在每个覆盖率水平下都能使风险最小，从而实现最小 *AURC*。

证明见附录A。在强校准条件下，MSP 与贝叶斯最优选择函数一致，因此是选择性预测的理想置信度分数。然而，这一条件在实践中很少满足，因为现代神经网络通常校准不佳（Guo 等人，2017年），尤其是在多模态模型中，导致 MSP 偏离 g^*。这种偏差会损害预测的相对排序，使一些错误预测排在正确预测之前，从而增加选择性风险。我们的方法通过学习纠正 MSP 的排序错误，而不是完全取代 MSP。

蒙特卡洛 Dropout。另一种常见的选择性预测方法是蒙特卡洛丢弃（MCD）（Gal &Ghahramani，2016年）。该方法在测试时启用丢弃，以衡量贝叶斯不确定性。给定输入 $\boldsymbol{x}$ 和网络 $f(\boldsymbol{x};\ \boldsymbol{\theta})$，MCD 通过采样丢弃掩码执行 K 随机前向传播 $\{\boldsymbol{\theta}^{(k)}\}_{k=1}^{K}$。每次传播都会得到一个预测分布 $p^{(k)}(\hat{y} \mid \boldsymbol{x})$。随后，通过取平均来近似贝叶斯后验预测分布：

$$\bar{p}(\hat{y} \mid \boldsymbol{x}) = \frac{1}{K} \sum_{k=1}^{K} p^{(k)}(\hat{y} \mid \boldsymbol{x}). \tag{7}$$

随后，MCD 的置信度分数可表示为$C_{\text{MC}}(\boldsymbol{x}) = \max_{y \in \mathcal{Y}} \bar{p}(y \mid \boldsymbol{x})$。虽然 MCD 比单次传播方法提供了更丰富的不确定性估计，但其 K次推理对于延迟敏感型 AVQA 而言代价过高。

校准。校准会调整模型的置信度分数，使预测概率准确反映正确性的可能性。较差的校准可能会对置信度排序产生负面影响（Kamath等人，2020年），从而降低选择性预测性能。在各种校准方法中，Platt 缩放（Platt 等人，1999年）最初是为二分类提出的，无法应用于多分类场景。温度缩放（Guo 等人，2017年）是另一种方法；然而，它保留了相对置信度排序，因此对选择性预测不起作用。最后，我们采用向量缩放（VS）（Guo 等人，2017年），通过对逻辑值进行类别特定仿射变换来调整置信度值和相对排序：

$$h\big(f(\boldsymbol{x})\big) = \text{softmax}\big(\mathbf{W} f(\boldsymbol{x}) + \mathbf{b}\big). \tag{8}$$

其中，$\mathbf{W}$ 是对角缩放矩阵，$\mathbf{b}$ 是偏置项，而h 是校准函数。我们对校准后的输出应用 MSP，以进行选择性预测。然而，向量缩放作用于输出 logit，并忽略中间特征中的不确定性信号。这凸显了利用多模态特征更好地建模预测正确性的必要性。

Doctor。我们进一步采用 Doctor（Granese 等人，2021）作为一种较新的选择性预测事后基线方法。Doctor 通过考察完整的预测分布，而不只是最大对数几率，来识别不确定样本。给定模型的 softmax 输出 $\mathbb{P}(Y \mid \boldsymbol{x}) = [p_1, \ldots, p_K]^\top$，Doctor 置信度分数定义为 $s(\boldsymbol{x}) = \sum_{k=1}^{K} p_k(\boldsymbol{x})^2$. 这种二次形式与基尼不纯度的补数密切相关，会惩罚由多类别混淆导致的高熵预测。与 MSP 相比，它能更好地对风险—覆盖率权衡进行排序。与 MSP 类似，Doctor 仅处理输出逻辑值，忽略了中间特征中丰富的不确定性信号，因此性能欠佳。

3.4. 自适应置信度精炼

我们提出自适应置信度细化（ACR），这是一种轻量级框架，通过细化基于MSP的置信度来增强选择性预测。ACR 并不取代 MSP，而是在保留其主要信号的同时，有选择地纠正其失效情形。ACR 1）将预测正确性建模为二分类任务；2）利用中间特征进行处理，从丰富的不确定性信号中学习，同时保留相较于 MSP 的改进在理论上的保证。

定义 3.4 (误差矩)。令 $C_{\mathrm{M}}(\boldsymbol{x})$ 表示 *MSP* 的输出，令 $C_{\mathrm{R}}(\boldsymbol{x})$ 表示残差风险估计器（*RRH* 头）的输出。我们定义如下误差：

$$\epsilon_{\mathrm{M}}(\boldsymbol{x}) = C_{\mathrm{M}}(\boldsymbol{x}) - c(\boldsymbol{x}), \quad \epsilon_{\mathrm{R}}(\boldsymbol{x}) = C_{\mathrm{R}}(\boldsymbol{x}) - c(\boldsymbol{x})$$

以及它们的二阶矩：

$$\begin{aligned} \sigma_{\mathrm{M}}^2 &= \mathbb{E}[\epsilon_{\mathrm{M}}^2] && \text{(MSE of MSP)} \\ \sigma_{R}^2 &= \mathbb{E}[\epsilon_{\mathrm{R}}^2] && \text{(MSE of Residual Risk Estimator)} \\ \sigma_{MR} &= \mathbb{E}[\epsilon_{\mathrm{M}}\epsilon_{\mathrm{R}}] && \text{(Error Cross-Moment)} \end{aligned}$$

定理 3.5 (置信度估计的融合收益)。如果误差交叉矩满足以下条件：

$$\sigma_{\mathrm{MR}} < \min(\sigma_{\mathrm{M}}^2, \sigma_{\mathrm{R}}^2) \tag{9}$$

那么存在唯一的最优固定融合权重 $\bar{\alpha}^* \in (0,1)$，使得融合置信度：

$$C_{\bar{\alpha}^*}(\boldsymbol{x}) = \bar{\alpha}^* C_{\mathrm{M}}(\boldsymbol{x}) + (1-\bar{\alpha}^*) C_{\mathrm{R}}(\boldsymbol{x})$$

其均方误差严格低于单独使用任一估计器。此外，根据命题 *3.2*，$C_{\bar{\alpha}^*}$ 对贝叶斯最优选择函数 g^*的近似更为准确，因此，与单独使用 *MSP* 相比，提升了预测的相对排序。最优权重为：

$$\bar{\alpha}^* = \frac{\sigma_{\mathrm{R}}^2 - \sigma_{\mathrm{MR}}}{\sigma_{\mathrm{M}}^2 + \sigma_{\mathrm{R}}^2 - 2\sigma_{\mathrm{MR}}}$$

证明见附录 A。公式15 中的条件确保残差风险估计器能够提供互补信息，从而纠正 MSP 无法准确对预测进行排序的区域中的错误。这意味着两个估计器的误差足够“分散”，使二者的组合能够减轻误差，并增强正确预测与错误预测之间的区分度。如附录B所示，我们的 ACR 在实证上满足这一条件。

设计原则。ACR在MSP基线方法的基础上加入了两个可学习组件：残差风险头（RRH）和置信度门控头（CGH），如图 2 所示。RRH利用中间多模态特征，对MSP之外的不确定性进行建模，尤其是过度自信的失败案例。CGH用于判断何时应用残差风险校正，从而实现输入自适应融合。不同于以往的工作在单模态或双模态设置中（Dong等人，2018年；Liu等人，2019年；Kamath等人，2020年），由于其三模态特性，音视频问答面临独特挑战，需要对多种模态之间的交互进行推理：视频、音频、问题和候选答案。因此，我们将骨干网络最后一层生成的多模态融合特征 f 以及softmax前logits作为两个头部的输入，使其能够获取预测答案及其支持证据。为简洁高效起见，RRH和CGH均采用输出经Sigmoid函数处理的多层感知机（MLP），将预测限制在 $[0,1]$。架构细节见附录 M。最终正确性分数表示为：

$$C_{\mathrm{ACR}}(\boldsymbol{x}) = \alpha(\boldsymbol{x}) C_{\mathrm{M}}(\boldsymbol{x}) + (1-\alpha(\boldsymbol{x})) C_{\mathrm{R}}(\boldsymbol{x}) \tag{10}$$

其中 $C_{\mathrm{R}}(\boldsymbol{x})$ 和 $\alpha(\boldsymbol{x})$ 分别是 RRH 和 CGH 的输出；$C_{\mathrm{M}}(\boldsymbol{x})$ 是基于 MSP 的置信度分数。总体而言，我们的 ACR 定义了一个全面的置信度估计框架，涵盖校准和基于选择器的策略。具体而言，移除 RRH 头后，ACR 仅作为校准方法发挥作用（Platt 等人，1999年）。相反，若移除 CGH 头，它便如以往研究（Geifman &El-Yaniv，2019年；White- head 等人，2022年）中那样发挥选择器的作用。我们在附录D中提供了更多消融实验和分析，表明简单的校准方法或基于选择器的方法不足以应对 $\mathcal{R}$-音视频问答。

定理 *3.6*（输入自适应置信度融合的最优性）。设自适应置信度分数为 $C_{\mathrm{ACR}}(\boldsymbol{x}) = \alpha^\dagger(\boldsymbol{x}) C_{\mathrm{M}}(\boldsymbol{x}) + (1-\alpha^\dagger(\boldsymbol{x})) C_{\mathrm{R}}(\boldsymbol{x})$，其中 $\alpha^\dagger(\boldsymbol{x}) \in [0,1]$ 是输入自适应融合权重。假设 $C_{\mathrm{M}}(\boldsymbol{x})$ 和 $C_{\mathrm{R}}(\boldsymbol{x})$ 是 $\boldsymbol{x}$ 的可测函数，且假设类 $\mathcal{H} = \{\alpha^\dagger : \mathcal{X} \to [0,1] \mid \alpha^\dagger$ 可测$\}$ 具有足够的表达能力，能够表示贝叶斯最优融合权重 $\alpha^*(\boldsymbol{x})$。使用正确性标签（*0/1*）上的二元交叉熵损失训练 *R-AVQA*模型时，任何风险最小化器都满足

$$C_{\mathrm{ACR}}(\boldsymbol{x}) = \mathbb{P}(c(\boldsymbol{x}) = 1 \mid \boldsymbol{x}) \quad a.e.$$

由于 $C_{\mathrm{ACR}} = g^*$ 几乎处处成立，按 C_{ACR} 排序会更接近选择性预测的贝叶斯最优排序，从而降低 *AURC*。如果逐点最优融合权重$\alpha^*(\boldsymbol{x}) = \operatorname{argmin}_\alpha \mathbb{E}\left[\left(C_\alpha(\boldsymbol{x}) - c(\boldsymbol{x})\right)^2 \mid \boldsymbol{x}\right]$ 随输入而变化（即 方差$[\alpha^*(\boldsymbol{x})] > 0$），那么任何固定权重 $\lambda \in [0,1]$ 到 C_λ 在均方误差和排序方面都不是最优的。

证明见附录 A，且 α 分布分析见附录E。其核心思想是，输入自适应融合扩展了置信度估计器的假设类，允许对 MSP 进行样本特定校正，从而比任何静态融合都更接近贝叶斯最优正确概率。附录 E表明**没有任何分布**会退化为固定值（即 $\alpha = \{0,1\}$），这证实了该方法实现的是真正的输入自适应融合，而非平凡解。

训练方案。ACR旨在提升置信度估计，同时通过

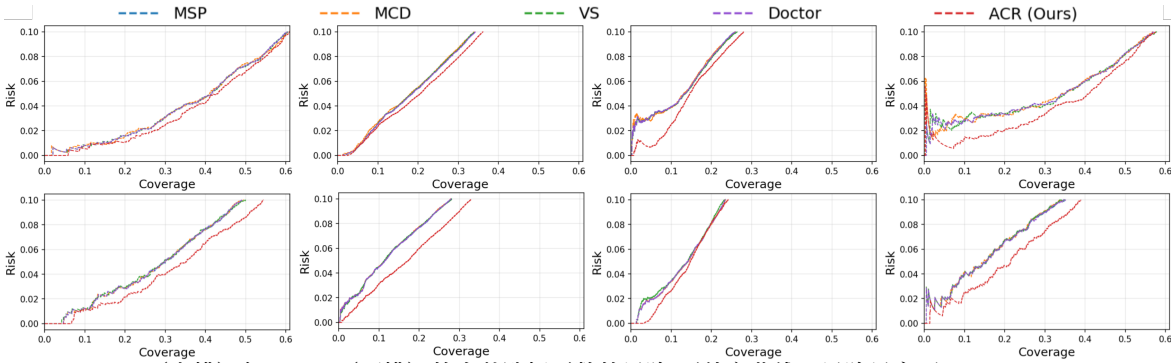


图 3. QA-TIGER（上排）与 TSPM（下排）的多种选择函数的风险-覆盖率曲线，风险最高至 10%。

两阶段训练流程，如图2所示。在阶段一，我们使用标准交叉熵损失训练基础音视频问答模型，然后冻结所有参数。在第二阶段，在冻结骨干网络的情况下，我们使用二元交叉熵损失联合训练RRH和CGH $\mathcal{L}_{\mathrm{BCE}}$。我们的训练目标是通过学习一个置信度估计器来近似正确性的后验概率$\mathbb{P}$(正确 | $\boldsymbol{x}$)。我们使用正确性信号 $\mathcal{L}_{\mathrm{BCE}}$ 进行优化$c(\boldsymbol{x})$，将（0/1标签）作为二元奖励，使我们的ACR与选择性预测的贝叶斯最优排序保持一致。冻结骨干网络可将置信度估计与基础预测解耦，从而在不受分类梯度干扰的情况下优化选择性预测。

$$\mathcal{L}_{\mathrm{BCE}} = -\mathbb{E}\big[c(\boldsymbol{x})\log C_{\mathrm{ACR}}(\boldsymbol{x}) - \big(1-c(\boldsymbol{x})\big)\log\big(1-C_{\mathrm{ACR}}(\boldsymbol{x})\big)\big]$$

二元交叉熵是一种严格适当评分规则，当预测置信度与正确性的条件概率相匹配时，可使期望风险最小化（Gneiting 和 Raftery，2007）。优化 $\mathcal{L}_{\mathrm{BCE}}$ 促使 C_{ACR}（$\boldsymbol{x}$）逼近概率 $\mathbb{P}$（c（$\boldsymbol{x}$）$=1 \mid x$）。由式 10，$\mathcal{L}_{\mathrm{BCE}}$ 关于 α（$\boldsymbol{x}$）的梯度与以下量成正比：

$$\frac{\partial \mathcal{L}_{\mathrm{BCE}}}{\partial \alpha} \propto \big(C_{\mathrm{M}}(x) - C_{\mathrm{R}}(x)\big)\cdot\big(C_{\mathrm{ACR}}(\boldsymbol{x}) - c(\boldsymbol{x})\big).$$

当 MSP 准确时，梯度会增大 $\alpha(\boldsymbol{x})$，从而强化其主导地位。若 MSP 过度自信但不正确，梯度会减小 $\alpha(\boldsymbol{x})$，从而激活残差校正。

4. 实验

4.1. 数据集和基线方法

我们在 MUSIC-AVQA（Li 等人，2022年）、MUSIC-AVQA-R（Ma 等人，2024）和 MUSIC-AVQA-v2.0（Liu等人，2024）数据集上进行实验。MUSIC-AVQA 是一个广泛使用的音视频推理数据集。MUSIC-AVQA-R 针对分布偏移下的稀有样本和分布外泛化，而 MUSIC-AVQA-v2.0 通过提高混合场景的多样性来缓解数据集偏差，并提供偏置集和平衡集，以评估模型鲁棒性。在我们的设置中，我们将 AVQA 模型设为 f，并将选择函数设为 g。我们在f上训练 AVQA 模型（

表 1. MUSIC-AVQA 数据集上所有选择函数的目标风险下的覆盖率（$\mathcal{C}@\mathcal{R}$）↑、AURC ↓和 ECE ↓。均为%。

模型 f	选择函数 g	$\mathcal{C}$@1%	$\mathcal{C}$@5%	$\mathcal{C}$@10%	$\mathcal{C}$@20%	AURC	ECE
QA-TIGER	MSP	14.29	41.53	60.47	91.83	8.67	8.83
	MCD	16.81	41.03	60.34	91.94	8.69	8.83
	VS	14.23	41.57	60.54	91.84	8.66	8.91
	Doctor	14.29	41.44	60.38	91.31	8.65	4.99
	ACR（我们的方法）	**17.21**	**43.07**	**61.39**	**91.96**	**8.41**	**4.88**
	Oracle	78.08	81.37	85.89	96.93	2.80	0.00
ST-AVQA	MSP	1.71	8.00	24.03	65.59	16.05	6.23
	MCD	1.71	8.06	23.43	65.03	16.07	6.22
	VS	1.74	8.11	23.44	65.88	16.07	6.18
	Doctor	1.71	7.98	23.86	65.80	16.08	5.01
	ACR（我们的方法）	**3.67**	**10.65**	**28.55**	**66.03**	**15.46**	**1.35**
	Oracle	71.70	74.08	78.19	87.97	4.90	0.00
TSPM	MSP	8.82	30.04	49.21	87.60	10.40	4.63
	MCD	8.73	29.98	49.58	87.43	10.40	4.46
	VS	6.63	29.67	50.00	87.87	10.41	4.68
	Doctor	8.82	29.90	49.03	86.87	10.47	4.35
	ACR（我们的方法）	**8.89**	**35.12**	**54.50**	**88.66**	**9.76**	**1.51**
	Oracle	76.91	80.14	84.60	95.17	3.11	0.00

训练集，并使用验证集对其进行评估，同时训练选择函数（g）。为 g使用独立训练数据，对于防止 AVQA 模型在训练集上过拟合并保持泛化能力至关重要。我们将测试集划分为两个子集：一个（20%）用于验证选择函数，另一个留出子集（80%）用于评估最终的选择性模型。更多详情请参阅附录L。

我们在具有不同架构设计的代表性音视频问答基线模型上，评估第 3 节讨论的选择函数。具体而言，我们考虑了 QA-TIGER（Kim 等人，2025年）、ST-AVQA（Li 等人，2022年）以及 TSPM（Li 等人，2024a），这些模型广泛应用于以往的音视频问答研究中。

4.2. 评估指标基准测试

本节对目标风险（$\mathcal{C}@\mathcal{R}$）下的最大覆盖率、风险-覆盖率曲线下面积（AURC）和期望校准误差（ECE）进行基准测试。所有结果均在第 4.1 节定义的新留出测试集划分上评估。

ACR 的表现优于其他方法。表 1、2 展示了 ACR 与其他方法的结果。排名前两位的结果以**粗体** 和下划线标出。表 2 仅报告了近期音视频问答模型（QA-TIGER、TSPM）在 MUSIC-AVQA-

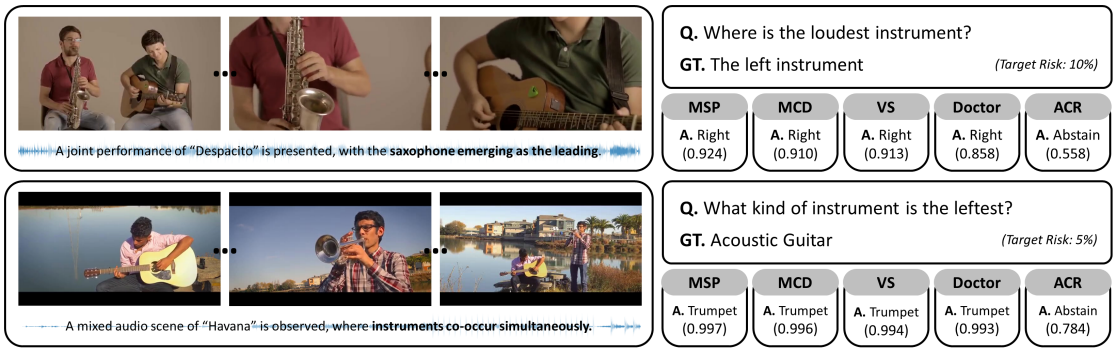


图 *4.* 基于 QA-TIGER 骨干网络的选择性预测示例。在这些示例中，MSP、MCD 和 VS 等标准置信度方法给出了过度自信但错误的答案，而我们提出的 ACR 能识别不可靠预测并正确选择弃答。括号中显示了不同目标风险水平下的置信度分数。

表 *2.* MUSIC-AVQA-R 和 MUSIC-AVQA-v2 数据集上的比较。我们报告 MUSIC-AVQA-R 的头部（H）/尾部（T）结果，以及 MUSIC-AVQA-v2 的平衡（Bal）/偏置（Bias）结果。指标：目标风险下的覆盖率（$\mathcal{C}@\mathcal{R}$）↑、AURC ↓和 ECE ↓（均以 % 表示）。

数据集	模型 f	选择函数 g	$\mathcal{C}$@1%		$\mathcal{C}$@5%		$\mathcal{C}$@10%		$\mathcal{C}$@20%		AURC		ECE	
			H/Bal	T/Bias	H/Bal	T/Bias	H/Bal	T/Bias	H/Bal	T/Bias	H/Bal	T/Bias	H/Bal	T/Bias
MUSIC-AVQA-R	QA-TIGER	MSP	6.04	0.23	18.55	13.48	34.25	25.96	63.56	59.06	15.43	16.87	14.25	14.16
		MCD	5.65	0.47	18.09	13.12	33.96	25.99	63.63	58.57	15.44	16.94	14.08	14.13
		VS	**6.12**	0.24	18.39	13.55	34.07	26.43	63.30	59.52	15.49	16.77	14.25	14.07
		Doctor	6.04	0.23	18.56	13.49	34.24	26.01	63.52	59.31	15.41	16.81	11.87	11.95
		ACR（我们的方法）	6.11	**6.46**	**20.59**	**14.62**	**36.16**	**28.03**	**63.84**	**59.98**	**15.17**	**16.36**	**10.86**	**9.90**
		Oracle	68.41	68.86	71.29	71.76	75.25	75.75	84.66	85.22	5.88	5.71	0.00	0.00
	TSPM	MSP	0.80	1.97	11.51	14.22	28.07	23.49	65.33	55.37	16.00	17.95	5.46	8.31
		MCD	0.80	1.98	11.50	14.24	27.88	23.54	65.22	55.61	16.03	17.92	5.39	8.22
		VS	0.89	2.05	11.35	14.08	28.18	23.28	65.66	54.76	15.93	18.02	5.34	8.22
		Doctor	0.80	1.97	11.47	14.22	28.09	23.53	65.31	55.16	16.01	18.04	5.52	6.16
		ACR（我们的方法）	**3.55**	**6.15**	**17.25**	**14.75**	**32.94**	**24.19**	**65.86**	**55.77**	**15.34**	**17.62**	**3.75**	**5.99**
		Oracle	68.62	66.60	71.72	69.40	75.71	73.26	85.17	82.41	5.72	6.60	0.00	0.00
MUSIC-AVQA v2.0	QA-TIGER	MSP	2.38	1.45	34.98	35.14	56.75	57.02	88.04	88.98	9.84	9.89	13.79	13.72
		MCD	2.58	0.19	35.04	34.84	56.62	56.80	87.71	89.05	9.88	9.89	13.58	13.48
		VS	1.41	0.47	36.58	35.12	56.84	57.05	**88.11**	88.68	9.80	9.91	13.70	13.87
		Doctor	2.38	1.45	34.98	35.14	56.75	57.02	88.10	88.64	9.84	9.89	9.89	9.87
		ACR（我们的方法）	**10.64**	**8.88**	**39.79**	**40.90**	**58.63**	**57.59**	**88.11**	**89.53**	**9.23**	**9.23**	**3.52**	**2.99**
		Oracle	76.69	76.98	79.92	80.22	84.36	84.67	94.91	95.26	3.16	3.09	0.00	0.00
	TSPM	MSP	4.69	0.47	17.42	15.46	34.80	34.96	64.35	66.49	15.05	14.97	5.60	5.18
		MCD	4.44	0.45	17.62	14.83	35.02	34.85	64.23	66.43	15.07	15.00	5.45	5.16
		VS	4.70	0.45	17.38	15.46	34.43	35.05	64.19	66.41	15.05	15.00	5.58	5.05
		Doctor	4.69	0.47	17.46	15.49	34.81	35.16	64.00	65.72	15.08	14.96	3.87	4.68
		ACR（我们的方法）	**7.12**	**4.66**	**24.58**	**21.48**	**37.66**	**38.94**	**66.84**	**69.44**	**14.03**	**13.84**	**1.64**	**1.34**
		Oracle	71.12	71.95	74.12	74.98	78.24	79.15	88.02	89.04	4.89	4.60	0.00	0.00

R 和 MUSIC-AVQA-v2.0（在均衡集上训练）的结果，因篇幅所限，其他结果见附录 I。ACR 的表现始终优于其他强基线方法：最大软最大概率（MSP）（Geifman& El-Yaniv，2017年）、蒙特卡洛丢弃（MCD）（Gal & Ghahramani，2016年）、向量缩放（VS）（Guo 等人，2017年）以及 Doctor（Granese 等人，2021）。在低风险水平（例如，1%、5%）下，性能提升尤为明显，表明该方法在要求置信度估计极为精准的场景中表现出色。对于 MUSIC-AVQA-v2.0 上的 QA-TIGER，ACR 在 $\mathcal{C}$@1% 时取得最大提升：平衡测试集上的结果提升了 **+8.26%p**，偏置测试集上的结果提升了 **+7.43%p**（表 2）。与其他方法相比，ACR 的 ECE 显著更低，表明其校准效果更佳。在低风险水平下，图 3 显示，随着覆盖率提高，ACR 的风险远低于其他方法

方法，证明了其在安全关键条件下的优势。值得注意的是，其他方法在不同设置下的表现各异：分布内（表 1）、分布外和偏置场景（表 2），而 ACR 在所有设置下均优于这些方法，展现出其在分布偏移或数据集偏差下的泛化能力。ACR 明显优于 MSP，验证了我们的设计原则：MSP 是一种强大的基线方法，但通过纠正 MSP 的失败案例，可以进一步提升其可靠性。附录 E对我们的优势进行了定性解释：与 MSP 相比，ACR 能更好地区分正确样本和错误样本，尤其是在高置信度水平下接近关键决策边界的位置。

准确率与可靠性的相关性。我们观察到，在表1-2中，标准音视频问答准确率与可靠性之间存在很强的相关性：准确率越高，往往表现出越好的选择性预测行为。例如，问答-

TIGER 在标准音视频问答中是准确率最高的模型，在各种设置下的可靠性评估中也优于 TSPM 和 ST-AVQA。这表明，可靠性的提升源自改进预测器 f 或选择器 g。此外，附录 E显示，基础音视频问答准确率较高的骨干网络往往能学到更高的 α 值，这表明其可靠性更高。然而，本研究着重开发一种更有效的选择器 g，以直接提升模型的可靠性。

阈值能否泛化到测试时？ 到目前为止，我们报告了在预定义目标风险水平下可实现的最大覆盖率，这对应于一种理想化评估：阈值直接在测试集上选择。然而在实践中，弃答阈值必须在验证集上选择，然后在测试期间固定。为评估该设置的实用性，我们使用 QA-TIGER，将验证集选定的阈值应用于测试集并测量由此产生的测试时风险，从而评估这些阈值向测试划分的迁移效果。如附录 K所示，目标风险与实际测试风险之间的差异始终很小，表明所学阈值能够在不同划分之间可靠泛化。这一表现与选择性预测文献中的既有发现一致（Geifman& El-Yaniv，2019年；Whitehead等人，2022年）。

仍有改进空间。表1、2 表明，尤其在低风险水平下，覆盖率仍远未达到理想的预言机性能。例如，表1显示 $\mathcal{C}@1\%$时，选择函数与Oracle之间的差距超过60%。这一差距出现在低风险约束下，因为Oracle在样本选择时可获取真实正确性，而实际方法依赖于多模态作答中并不完美的置信度估计。因此，跨模态问题的不确定性估计误差会导致较为保守的拒绝率，并降低覆盖率。这凸显了音视频问答模型和选择函数仍有很大的改进空间，从而能够构建更可靠的音视频问答系统。

4.3. 定性分析

图4 展示了定性结果，显示ACR与其他方法在置信度估计上的差异。ACR的关键改进在于提升了置信度估计以及示例之间的相对排序，而非提高准确率（第二阶段冻结音视频问答模型，因此ACR不会影响准确率）。对于真正不确定或无法回答的问题，ACR会分配远低于其他方法的置信度分数，从而放弃作答。在第一个示例中，许多帧只显示单个表演者或乐器，而音频中则有多种乐器同时发声。在第二个示例中，歧义来自音频，其中乐器音量相近，而视频提供了更清晰的空间线索。音频与视觉线索之间的这种差异造成了跨模态歧义，可能误导音视频问答模型做出过度自信的预测。在此类情况下，多模态特征传达了不确定性信号，ACR通过降低的置信度有效捕捉并反映了这些信号。更多示例见附录J。

表 3. 在 MUSIC-AVQA 测试集上使用 $\mathcal{C}@\mathcal{R}\uparrow$、AURC $\downarrow$和 ECE $\downarrow$对 $f'(x)$ a, v, q对 QA-TIGER 的 ACR 输入进行消融实验。 avq 分别表示 softmax前 logits、音频、视频、问题和融合多模态表示。所有数值均为百分比。

特征	$\mathcal{C}$@1%	$\mathcal{C}$@5%	$\mathcal{C}$@10%	$\mathcal{C}$@20%	AURC	ECE
a	10.20	40.66	60.79	90.55	8.82	5.19
v	3.57	34.15	59.35	91.02	9.17	5.01
q	13.79	35.91	60.84	90.99	8.85	5.33
$f'(x)$	10.98	38.77	59.39	91.63	8.96	5.53
avq	12.54	41.63	60.99	90.89	8.75	4.96
$(f'(x)\ a)$	13.21	42.29	60.45	90.95	8.65	5.66
$(f'(x)\ v)$	10.39	39.03	60.01	91.80	8.81	4.96
$(f'(x)\ q)$	12.46	42.62	61.09	91.15	8.55	5.69
$(f'(x)\ avq)$	**17.21**	**43.07**	**61.39**	**91.96**	**8.41**	**4.88**
$(f'(x)\ a, v, q, avq)$	14.29	41.53	60.47	91.83	8.52	5.14

4.4. 消融研究

本节中，我们介绍 ACR 选择函数的消融实验。更多消融实验见附录 D。

多模态特征的影响。表 3凸显了多模态特征对于实现高性能的重要性。单独使用各类表示时，多模态特征（$f'(x)$、avq）始终比单模态特征（a， v， q）具有更高的覆盖率和更低的 AURC。然而，将所有模态输入 ACR（最后一行）会因跨模态错位而导致性能下降（Ma 等人，2023；Cai 等人，2025年），从而妨碍可靠的置信度估计。因此，我们仅将 ACR 应用于最终的融合多模态特征，以更好地缓解错位。这一结果表明，可靠的音视频问答本质上是一个多模态问题。结合 $f'(x)$ 和 avq 可获得最佳性能，因此我们在所有实验中均采用这一设置。

ACR 架构。附录D 展示了采用不同架构的 ACR 所得到的结果。为简便起见，我们对两个学习头采用相同的架构，并发现这种做法通常能取得最佳性能。我们发现，简化架构设计会降低性能，而增加复杂度并不总能带来显著提升。结合表 3中的结果可知，ACR 在很大程度上取决于输入特征和优化目标，而非仅仅取决于架构的复杂度。

5. 讨论与结论

标准的音视频问答（AVQA）设定不允许模型在不确定时弃答，尽管许多现实应用要求只有在风险较低时才进行预测（即 $\mathcal{C}$@1%,$\mathcal{C}$@5%）。为此，我们提出了一种新的 $\mathcal{R}$-AVQA 设定，将弃答置于错误预测之上。我们通过在选择性预测设定下评估多个音视频问答模型，建立了新的基准，并为 $\mathcal{R}$-AVQA 提出了自适应置信度细化（ACR），其表现显著优于其他方法。尽管 MSP 是一种

强有力的选择性预测基线方法，但它距离最优选择函数仍相差甚远。因此，ACR 通过精炼 MSP 来提升置信度估计，而非完全取代它。我们的理论分析和实证结果表明，与其他方法相比，ACR 能够在不同风险容忍度水平下持续提高覆盖率。总而言之，我们的研究为多模态推理中的可靠性分析奠定了坚实基础，并推动音视频问答系统不仅准确，而且具备自我认知且值得信赖。

致谢

本研究获得韩国政府（MSIT）资助的韩国国家研究基金会（NRF）项目（RS-2025-00573160）支持；获得产业通商资源部（MOTIR，韩国）资助的技术创新计划（RS-2025-25422280）支持；并获得大韩民国政府（科学技术信息通信部）资助的高性能计算支持项目（RQT-25-070104）支持。

Cem Subakan 获得 NSERC 探索资助 RGPIN 2023-05759 的支持。本研究还部分得益于 Calcul Qu´ebec（https://www.calculquebec.ca/）和加拿大数字研究联盟（https://alliancecan.ca/en）提供的支持。

本研究还获得现代汽车郑梦九全球奖学金的支持，该奖学金授予 Dinh Phu Tran（第一作者）和 Thao Do（第五作者）。

影响声明

本文旨在推动机器学习领域的发展。我们的工作可能会带来许多社会影响，但我们认为没有哪一项需要在此特别强调。

参考文献

Adam, K. D. B. J. 等. 一种随机优化方法。*arXiv* 预印本 *arXiv:1412.6980*，1412(6)，2014年。

Angelopoulos, A. N., Bates, S. 等. 保形预测：简明入门。机器学习基础与趋势®，16(4):494–591，2023。

Asan, O., Bayrak, A. E. 和 Choudhury, A. 医疗保健领域的人工智能与人类信任：聚焦临床医生。《医学互联网研究杂志》，22(6):e15154，2020年。

Cai, Y., Liu, Y., Gao, E., Jiang, T., Zhang, Z., Hengel, A. v. d., 和 Shi, J. Q. 论跨模态错位在多模态表示学习中的价值。*arXiv* 预印本 *arXiv:2504.10143*，2025年。

Chen, J., Yoon, J., Ebrahimi, S., Arik, S., Pfister, T.，以及 Jha, S. 通过自我评估进行适应，以提升大语言模型中的选择性预测。载于计算语言学协会 *EMNLP 2023* 论文集，第 5190–5213 页，2023。Dancette, C., Whitehead, S., Maheshwary, R., Vedantam, R., Scherer, S., Chen, X., Cord, M., 以及 Rohrbach, M. 通过向同伴学习，改进选择性视觉问答。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第 24049–24059 页，2023。

Dong, L., Quirk, C., 和 Lapata, M. 神经语义解析的置信度建模。载于 Gurevych, I. 和 Miyao, Y.（编），第*56*届计算语言学协会年会论文集（第*1*卷：长篇论文），第743–753页，澳大利亚墨尔本，2018年7月。计算语言学协会。doi：10.18653/v1/P18-1069。网址：https://aclanthology.org/P18-1069/。

Freund, Y., Mansour, Y., 和 Schapire, R.E. 平均分类器的泛化界。2004年。

Gal, Y. 和 Ghahramani, Z. 将丢弃作为贝叶斯近似：在深度学习中表示模型不确定性。载于国际机器学习会议，第 1050–1059 页。PMLR，2016年。

Geifman, Y. 和 El-Yaniv, R. 深度神经网络的选择性分类。神经信息处理系统进展，第 30 卷，2017年。

Geifman, Y. 和 El-Yaniv, R. SelectiveNet：一种集成拒绝选项的深度神经网络。载于国际机器学习会议，第 2151–2159 页。PMLR，2019年。

Geifman, Y.、Uziel, G. 和 El-Yaniv, R. 深度神经分类器的偏差降低不确定性估计。*arXiv* 预印本 *arXiv:1805.08206*，2018年。

Gemmeke, J. F.、Ellis, D. P.、Freedman, D.、Jansen, A.、Lawrence, W.、Moore, R. C.、Plakal, M. 和 Ritter, M. Audio Set：音频事件的本体与人工标注数据集。载于*2017*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*），第 776–780 页。IEEE，2017年。

Glenn, W. B. 等. 以概率形式表达的预报验证。《月度天气评论》，78(1):1–3，1950年。

Gneiting, T. 和 Raftery, A. E. 严格适当评分规则、预测与估计。美国统计学会杂志，102(477):359–378，2007。

Granese, F.、Romanelli, M.、Gorla, D.、Palamidessi, C. 和 Piantanida, P. Doctor：一种检测分类错误的简单方法。神经信息处理系统进展，34:5669–5681，2021。

Graves, A. 长短期记忆。使用循环神经网络进行监督式序列标注，第 37–45 页，2012年。

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. 现代神经网络的校准。载于国际机器学习会议，第 1321–1330 页。PMLR，2017年。

Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., 等。用于大规模音频分类的 CNN 架构。载于 *2017* 年 *IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 131–135 页。IEEE，2017年。

Jiang, Z.、Araki, J.、Ding, H. 和 Neubig, G. 我们如何知道语言模型何时真正“知道”？论语言模型在问答中的校准。计算语言学协会会刊，9：962–977，2021。

Kamath, A.、Jia, R. 和 Liang, P. 领域偏移下的选择性问答。载于 Jurafsky, D.、Chai, J.、Schluter, N. 和 Tetreault, J.（编），计算语言学协会第*58*届年会论文集，第5684–5696页，线上，2020年7月。计算语言学协会。doi: 10.18653/v1/2020.acl-main.503。网址: https://aclanthology.org/2020.acl-main.503/。

Kim, H.、Jung, I.、Suh, D.、Zhang, Y.、Lee, S. 和 Hong, S. 问题感知高斯专家用于音频-视觉问答。载于计算机视觉与模式识别会议论文集，第 13681–13690页，2025年。

Kumar, L. A., Renuka, D. K., Rose, S. L., Wartana, I. M. 等。基于深度学习的辅助技术：面向听力障碍者的视听语音识别。国际认知计算工程期刊，3：24–30，2022年。

Lakshminarayanan, B., Pritzel, A. 和 Blundell, C. 使用深度集成进行简单且可扩展的预测不确定性估计。神经信息处理系统进展，30，2017年。

Langley, P. 机器学习论文撰写。载于 Langley, P.（编），第*17*届国际机器学习会议论文集（*ICML 2000*），第1207–1216页，斯坦福，加利福尼亚州，2000年。Morgan Kaufmann。

Li, G., Wei, Y., Tian, Y., Xu, C., Wen, J.-R., and Hu, D. 学习在动态音视频场景中回答问题。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第19108–19118页，2022年。

Li, G., Hou, W. 和 Hu, D. 面向音视频问答的渐进式时空感知。载于第*31*届*ACM*国际多媒体会议论文集，第7808–7816页，2023。

Li, G., Du, H. 和 Hu, D. 借助关键语义感知线索提升音视频问答。载于第*32*届*ACM*国际多媒体会议论文集，第5997–6005页，2024a。

Li, Z., Guo, D., Zhou, J., Zhang, J., and Wang, M. 面向音视频问答的对象感知自适应正例学习. In *AAAI*人工智能会议论文集, 第38卷, 第3306–3314页, 2024b.

Liu, X., Dong, Z., and Zhang, P. 应对音乐问答中的数据偏差：构建用于无偏问答的平衡数据集. In *IEEE/CVF*冬季计算机视觉应用会议论文集, 第4478–4487页, 2024.

Liu, Z., Wang, Z., Liang, P. P., Salakhutdinov, R. R., Morency, L.-P., and Ueda, M. 深度赌徒：运用投资组合理论学习弃答. 神经信息处理系统进展, 32, 2019年.

L¨utkenh¨oner, B. and Basel, T. 针对特异性高但敏感性低的诊断测试进行预测建模：对疑似梅尼埃病患者进行甘油试验的研究. *PLoS One*, 8(11):e79315, 2013.

Ma, H., Zhang, Q., Zhang, C., Wu, B., Fu, H., Z hou, J. T., and Hu, Q. 多模态学习的校准。载于国际机器学习会议，第23429–23450页。PMLR，2023。

Ma, J., Hu, M., Wang, P., Sun, W., Song, L., Pei, H., Liu, J., and Du, Y. 看、听、回答：克服音视频问答中的偏差。神经信息处理系统进展，37:9507–9531，2024。

Mielke, S. J., Szlam, A., Dinan, E., and Boureau, Y.-L. 通过语言校准降低对话代理的过度自信。计算语言学协会会刊，10:857–872，2022年。

Patel, P.、Pampaniya, S.、Ghosh, A.、Raj, R.、Karuppaih, D. 和 Kandasamy, S. 通过机器学习提升无障碍体验：视觉与听力障碍技术综述。*IEEE Access*，2025年。

Platt, J. 等。支持向量机的概率输出及其与正则化似然方法的比较。大间隔分类器进展，10(3):61–74，1999年。

Radford, A.、Kim, J. W.、Hallacy, C.、Ramesh, A.、Goh, G.、Agarwal, S.、Sastry, G.、Askell, A.、Mishkin, P.、Clark, J. 等。通过自然语言监督学习可迁移的视觉模型。在国际机器学习会议，第 8748–8763 页。PMLR，2021。

Schwartz, I.、Schwing, A. G. 和 Hazan, T. 视听场景感知对话的简单基线。在*IEEE/CVF*计算机视觉与模式识别会议论文集，第 12548–12558 页，2019年。

Shazeer, N.、Mirhoseini, A.、Maziarz, K.、Davis, A.、Le, Q.、Hinton, G. 和 Dean, J. 规模惊人的神经网络：稀疏门控专家混合层。*arXiv* 预印本 *arXiv:1701.06538*，2017年。

Srinivasan, T.、Hessel, J.、Gupta, T.、Lin, B. Y.、Choi, Y.、Thomason, J.与Chandu, K. R. 选择性“选择性预测”：减少视觉-语言推理中不必要的弃答。*arXiv*预印本 *arXiv:2402.15610*，2024.

Varshney, K. R. 带拒绝选项的集成分类的风险界。见*2011*年*IEEE*统计信号处理研讨会（*SSP*），第769–772页。IEEE，2011年。

Varshney, N.、Mishra, S.与Baral, C. 在分布内、分布外和对抗设置下若干任务中对选择性预测方法的研究。*arXiv*预印本 *arXiv:2203.00211*，2022年。

Vovk, V.、Gammerman, A.与Shafer, G. 随机世界中的算法学习。施普林格，2005年。

Whitehead, S., Petryk, S., Shakib, V., Gonzalez, J., Darrell, T., Rohrbach, A., and Rohrbach, M. 可靠的视觉问答：宁可弃答，也不答错。载于欧洲计算机视觉会议，第 148–166 页。施普林格，2022年。

Xin, J., Tang, R., Yu, Y., and Lin, J. 弃答的艺术：自然语言处理中的选择性预测与误差正则化。载于计算语言学协会年会与国际自然语言处理联合会议第十一届会议论文集（第 *1* 卷：长篇论文），第 1040–1051 页，2021。

Yun, H., Yu, Y., Yang, W., Lee, K., and Kim, G. Pano-avqa：360度视频中的基于定位的音视频问答。载于*IEEE/CVF*国际计算机视觉会议论文集，第 2031–2041 页，2021。

Zadrozny, B.和Elkan, C. 将分类器分数转换为准确的多类别概率估计。载于第八届*ACM SIGKDD*国际知识发现与数据挖掘会议论文集，第694–699页，2002。

知道何时作答：面向可靠音视频问答的自适应置信度精炼

附录

索引

附录 A 包含定理 3.3、3.5和 3.6的完整证明。

附录 B 展示了满足定理 3.5中条件的实证结果，确保与 MSP 相比，我们的方法始终能在示例之间实现更低的均方误差和更优的相对排序。

附录C 对相关工作进行了更详细的综述。

附录D 进一步讨论了我们的消融研究。

附录E进一步分析和讨论了 MSP 以及我们结合 α 分布分析的 ACR。

附录F讨论了我们提出的 ACR 的计算复杂度和开销。

附录 G 展示了三个音视频问答模型（QA-TIGER、ST-AVQA 和 TSPM）在原始 MUSIC-AVQA 测试集上的复现结果。

附录 H 展示了针对 MSP 使用数据增强的实验。

附录 I 报告了在 MUSIC-AVQA-R 和 MUSIC-AVQA-v2 上的额外性能（在偏置数据集上训练，并在平衡测试集和偏置测试集上进行评估）。

附录J 提供了更多定性结果。

附录K 评估了阈值的泛化能力。

附录L进一步介绍了实验设置中使用的数据集划分。

附录M 全面介绍了模型架构、超参数和实现细节。

附录N 报告了 TSPM 基线方法在 MUSIC-AVQA 数据集上的均值和标准差，这些数据是通过五个随机种子计算得出的。

附录O 探讨了校准与保形预测之间的关系，背景是我们的可靠音视频问答框架和近期文献。

附录 P 进一步讨论了我们工作的广泛影响，并概述了未来研究的潜在方向。

附录 Q 概述了我们工作的更广泛影响，包括潜在的社会和伦理考量 s.

A. 证明

A.1. 定理的证明3.3

本节将给出正文中定理 3.3 （附录中的定理 A.2 ）的完整证明。**假设 A.1** （最大后验预测）。 我们假设预测器输出最大后验（*MAP*）类别：

$$\hat{y} = \arg\max_{k \in \mathcal{Y}} \mathbb{P}(Y = k \mid \boldsymbol{x}),$$

这也是基于 *softmax* 的分类器中的标准做法。

定理 *A.2*（强校准下的 *MSP* 最优性）。考虑一个分类器 $f(\boldsymbol{x})$，它输出类别概率$\mathbb{P}(Y = k|\boldsymbol{x})$，并预测 $\hat{}y = arg\,max_k \mathbb{P}(Y = k|\boldsymbol{x})$（假设 *A.1*）。如果 *MSP* 置信度函数 $C_{\mathrm{M}}(\boldsymbol{x}) = max_k \mathbb{P}(Y = k|\boldsymbol{x})$ 是强校准的（定义 *3.1*），那么：*1*）*MSP* 等于贝叶斯最优选择函数：$C_{\mathrm{M}}(\boldsymbol{x}) = g^*(\boldsymbol{x}) = \mathbb{P}[c(\boldsymbol{x}) = 1 \mid \boldsymbol{x}]$。*2*）按 *MSP* 对样本排序会产生选择性预测的贝叶斯最优排序。*3*）此排序在每个覆盖率水平上都能使风险最小，从而实现最低 *AURC*。

证明。选择性预测等价于二分类：区分正确（$c(\boldsymbol{x}) = 1$）与错误（$c(\boldsymbol{x}) = 0$）预测。贝叶斯最优分类器使用似然比，该比值与后验概率单调相关：

$$s^*(\boldsymbol{x}) = \mathbb{P}[c(\boldsymbol{x}) = 1 \mid \boldsymbol{x}] = \mathbb{P}[\hat{y} = y^* \mid \boldsymbol{x}] \tag{11}$$

在强校准（定义3.1 和假设A.1）条件下，对于预测类别为$\hat{}y = \operatorname{argmax}_k \ \mathbb{P}(Y = k \mid \boldsymbol{x})$：

$$C_{\mathrm{M}}(\boldsymbol{x}) = \max_k \mathbb{P}(Y = k \mid \boldsymbol{x}) = \mathbb{P}(Y = \hat{y} \mid \boldsymbol{x}) \tag{12}$$

由于$\hat{}y$ 是预测类别，如果模型预测正确（$\hat{}y = y^*$），那么 $\mathbb{P}(Y = \hat{}y \mid \boldsymbol{x}) = \mathbb{P}(Y = y^* \mid \boldsymbol{x})$。在强校准条件下，这等于真实后验概率：

$$C_{\mathrm{M}}(\boldsymbol{x}) = \mathbb{P}[\hat{y} = y^* \mid \boldsymbol{x}] = s^*(\boldsymbol{x}) \tag{13}$$

由于 $s^*(\boldsymbol{x})$等于预测正确的真实概率，按 s^* 排序便能完美地根据样本预测正确的可能性对样本排序。对于任意一对样本 $\boldsymbol{x}_1$， $\boldsymbol{x}_2$：

$$s^*(\boldsymbol{x}_1) > s^*(\boldsymbol{x}_2) \iff \mathbb{P}[c(\boldsymbol{x}_1) = 1] > \mathbb{P}[c(\boldsymbol{x}_2) = 1] \tag{14}$$

这种排序在每个覆盖率水平下都能将风险降至最低：按照 s^* 选取排名前k 的样本，可以得到最有可能预测正确的 k 样本。因此，根据定义，MSP $= s^*$ 等于贝叶斯最优选择函数，并实现最小 AURC。

A.2. 定理3.5的证明

本节将给出正文中定理3.5 （附录中的定理A.3 ）的完整证明。

定理 A.3 （置信度估计的融合收益）。 如果误差交叉矩满足以下条件：

$$\sigma_{\mathrm{MR}} < \min(\sigma_{\mathrm{M}}^2, \sigma_{\mathrm{R}}^2) \tag{15}$$

*th*则存在唯一的最优固定融合权重 $\bar{\alpha}^* \in$ （0，1），使得融合置信度 :

$$C_{\bar{\alpha}^*}(\boldsymbol{x}) = \bar{\alpha}^* C_{\mathrm{M}}(\boldsymbol{x}) + (1 - \bar{\alpha}^*) C_{\mathrm{R}}(\boldsymbol{x})$$

取得了严格低于任一单独估计器的均方误差。此外，根据命题*3.2*， $C_{\bar{\alpha}^*}$ 能够更接近地逼近贝叶斯最优选择函数 g^*，从而相比单独使用 *MSP*，改善了预测的相对排序。最优权重为：

$$\bar{\alpha}^* = \frac{\sigma_{\mathrm{R}}^2 - \sigma_{\mathrm{MR}}}{\sigma_{\mathrm{M}}^2 + \sigma_{\mathrm{R}}^2 - 2\sigma_{\mathrm{MR}}}$$

证明。定义 $C_\alpha = \alpha C_{\mathrm{M}} + (1-\alpha) C_{\mathrm{R}}$ 的均方误差（为简洁起见，省略输入 $\boldsymbol{x}$）：

$$J(\alpha) = \mathbb{E}[(C_\alpha - c)^2] = \mathbb{E}[(\alpha\epsilon_{\mathrm{M}} + (1-\alpha)\epsilon_{\mathrm{R}})^2] \tag{16}$$

$$= \alpha^2\mathbb{E}[\epsilon_{\mathrm{M}}^2] + (1-\alpha)^2\mathbb{E}[\epsilon_{\mathrm{R}}^2] + 2\alpha(1-\alpha)\mathbb{E}[\epsilon_{\mathrm{M}}\epsilon_{\mathrm{R}}] \tag{17}$$

$$= \alpha^2\sigma_{\mathrm{M}}^2 + (1-\alpha)^2\sigma_{\mathrm{R}}^2 + 2\alpha(1-\alpha)\sigma_{\mathrm{MR}} \tag{18}$$

一阶条件。

令 $\frac{dJ}{d\alpha} = 0$:

$$2\alpha\sigma_{\mathrm{M}}^2 - 2(1-\alpha)\sigma_{\mathrm{R}}^2 + 2(1-2\alpha)\sigma_{\mathrm{MR}} = 0 \tag{19}$$

重新整理:

$$\alpha(\sigma_{\mathrm{M}}^2 + \sigma_{\mathrm{R}}^2 - 2\sigma_{\mathrm{MR}}) = \sigma_{\mathrm{R}}^2 - \sigma_{\mathrm{MR}} \tag{20}$$

因此:

$$\bar{\alpha}^* = \frac{\sigma_{\mathrm{R}}^2 - \sigma_{\mathrm{MR}}}{\sigma_{\mathrm{M}}^2 + \sigma_{\mathrm{R}}^2 - 2\sigma_{\mathrm{MR}}} \tag{21}$$

二阶条件。

$$\frac{d^2J}{d\alpha^2} = 2(\sigma_{\mathrm{M}}^2 + \sigma_{\mathrm{R}}^2 - 2\sigma_{\mathrm{MR}}) = 2\mathbb{E}[(\epsilon_{\mathrm{M}} - \epsilon_{\mathrm{R}})^2] > 0 \tag{22}$$

当估计器不同时，分母严格为正，从而确认 $\bar{\alpha}^*$ 是最小值。

内部解条件。对于 $\bar{\alpha}^* \in (0,\ 1)$:

- $\bar{\alpha}^* > 0$: 要求分子 > 0，即 $\sigma^2{}_{\mathrm{R}} > \sigma_{\mathrm{MR}}$
- $\bar{\alpha}^* < 1$: 要求分子 $<$ 分母，即 $\sigma_{\mathrm{R}}^2 - \sigma_{\mathrm{MR}} < \sigma_{\mathrm{M}}^2 + \sigma_{\mathrm{R}}^2 - 2\sigma_{\mathrm{MR}}$，化简为 $\sigma_{\mathrm{MR}} < \sigma_{\mathrm{M}}^2$

当 $\sigma_{MR} < \min(\sigma_M^2,\ \sigma_R^2)$时，两个条件均成立。

严格改进。 在内部极小值 $\bar{\alpha}^* \in (0,1)$ 处:

$$J(\bar{\alpha}^*) < J(0) = \sigma_{\mathrm{R}}^2 \quad \text{and} \quad J(\bar{\alpha}^*) < J(1) = \sigma_{\mathrm{M}}^2 \tag{23}$$

因此， $J(\bar{\alpha}^*) < \min(\sigma_{\mathrm{M}}^2, \sigma_{\mathrm{R}}^2)$。

对排序的启示。任意置信度估计器 s 相对于正确性 $c(\boldsymbol{x})$ 的均方误差（MSE）可分解为（Glenn 等人，1950年）：

$$\mathbb{E}[(s-c)^2] = \mathbb{E}[(s-s^*)^2] + \mathbb{E}[s^*(1-s^*)] \tag{24}$$

其中， $s^*(\boldsymbol{x}) = \mathbb{P}[c(\boldsymbol{x}) = 1 \mid \boldsymbol{x}]$ 是贝叶斯最优选择函数。由于第二项是常数（不可约不确定性），因此，缩减量 $J(\bar{\alpha}^*) < \sigma_{\mathrm{M}}^2$ 意味着:

$$\mathbb{E}[(C_{\bar{\alpha}^*} - s^*)^2] < \mathbb{E}[(C_{\mathrm{M}} - s^*)^2] \tag{25}$$

因此，融合置信度 $C_{\bar{\alpha}^*}$ 比单独使用 MSP 更接近 s^* 的 L^2近似值，从而改善了选择性预测的相对排序。

A.3. 定理3.6的证明

本节将给出主论文中定理3.6 的完整证明（附录中的定理 A.4 ）。

定理 *A.4*（输入自适应置信度融合的最优性）。设自适应置信度分数为 $C_{\mathrm{ACR}}(\boldsymbol{x}) = \alpha^{\dagger}(\boldsymbol{x}) C_{\mathrm{M}}(\boldsymbol{x}) + (1-\alpha^{\dagger}(\boldsymbol{x})) C_{\mathrm{R}}(\boldsymbol{x})$，其中 $\alpha^{\dagger}(\boldsymbol{x}) \in [0,1]$ 是输入自适应融合权重。假设 $C_{\mathrm{M}}(\boldsymbol{x})$和 $C_{\mathrm{R}}(\boldsymbol{x})$是 $\boldsymbol{x}$ 的可测函数，且假设类 $\mathcal{H} = \{\alpha^{\dagger} : \mathcal{X} \to [0,1] \mid \alpha^{\dagger}$ 可测函数$\}$ 具有足够的表达能力，能够表示贝叶斯最优融合权重 $\alpha^{*}(\boldsymbol{x})$。使用正确性标签（*0/1*）上的二元交叉熵损失训练 *R-AVQA*模型时，任何风险最小化器都满足

$$C_{\mathrm{ACR}}(\boldsymbol{x}) = \mathbb{P}(c(\boldsymbol{x}) = 1 \mid \boldsymbol{x}) \quad a.e.$$

由于 $C_{\mathrm{ACR}} = g^{*}$ 几乎处处成立，按照 C_{ACR} 排序会更接近选择性预测的贝叶斯最优排序，从而降低 *AURC*。如果逐点最优融合权重 $\alpha^{*}(\boldsymbol{x}) = \operatorname{argmin}_{\alpha} \mathbb{E}\left[\left(C_{\alpha}(\boldsymbol{x}) - c(\boldsymbol{x})\right)^{2} \mid \boldsymbol{x}\right]$ 在不同输入之间变化（即方差$[\alpha^{*}(\boldsymbol{x})] > 0$），那么任何固定权重 $\lambda \in [0,1]$ 到 C_{λ} 在均方误差和排序方面都严格次优。

证明。 二元交叉熵是一种严格适当评分规则（Gneiting 和 Raftery，2007）。因此，当预测置信度等于正确性的条件概率时，其期望风险达到唯一最小值：

$$C_{\mathrm{ACR}}(\boldsymbol{x}) = \mathbb{P}(c(\boldsymbol{x}) = 1 \mid \boldsymbol{x}) = s^{*}(\boldsymbol{x}) \quad \text{a.e.}$$

由于 $C_{\mathrm{ACR}} = s^{*}$ 几乎处处成立，按照 C_{ACR} 排序会更接近贝叶斯最优选择函数所给出的排序。根据定理 3.3中的论证，该排序在每个覆盖率水平上都能使风险最小，从而获得更低的 AURC，也更接近 AURC 的最小值。

将 $\alpha^{\dagger}(\boldsymbol{x})$ 限定为常数，会把假设类约束为 C_{M} 与 C_{R} 的固定凸组合。允许 $\alpha^{\dagger}(\boldsymbol{x})$ 依赖于 $\boldsymbol{x}$，则会严格扩大此函数类。如果贝叶斯最优融合权重随 $\boldsymbol{x}$ 变化，那么几乎处处都不存在能够匹配最优解的常数 λ，因此该解是次优的。

要理解固定融合为何会导致排序次优，可以考虑以下情形：当逐点最优权重 $\alpha^{*}(\boldsymbol{x})$ 随输入变化时，任何固定的 λ 都必然会产生：

$$C_{\lambda}(\boldsymbol{x}) \neq s^{*}(\boldsymbol{x}) \quad \text{on a set of positive measure} \tag{26}$$

这种差异会导致排序错误：存在一些样本对（$\boldsymbol{x}_1$，$\boldsymbol{x}_2$），使得 C_{λ} 相对于其真实正确性概率对它们的排序有误，从而使 AURC 高于贝叶斯最优下限。
因此，与固定融合相比，输入自适应融合能严格改善期望风险和排序表现。

B. 定理 3.5 的实证验证

我们通过分析 MSP 与残差风险头（RRH）错误之间的关系，验证定理 3.5 的条件。融合收益的关键条件是：

$$\sigma_{\mathrm{MR}} = \mathbb{E}[\epsilon_{\mathrm{M}} \cdot \epsilon_{\mathrm{R}}] < \min\left(\sigma_{\mathrm{M}}^{2}, \sigma_{\mathrm{R}}^{2}\right) \tag{27}$$

其中 ϵ_{M} 和 ϵ_{R} 分别表示 MSP 和 RRH 在排序正确性方面的误差，而 σ_{M}^{2}，σ_{R}^{2} 则分别表示它们的误差方差。该条件表明，当两个估计器误差之间的相关性低于各自误差幅度时，融合便能带来收益。换言之，当 MSP 和 RRH 产生互补误差（相关性较低）时，融合可以利用各自的优势，提升整体选择性预测性能。表 4 展示了在三个 AVQA 主干网络和三个数据集上进行的全面验证。我们发现，该条件普遍成立：所提出的 ACR 始终满足融合收益条件，且优势显著。这些结果为定理 3.5中的理论分析提供了实证支持，并表明 ACR 的性能提升源于通过互补的不确定性信号系统性地降低置信度估计误差。

C. 更多相关工作

校准。校准是不确定性估计的另一种视角，用于评估模型的置信度水平与其实际准确率之间的一致性。当我们考虑这样一种模型时，校准与选择性预测之间的区别便显而易见：该模型对一定比例的输入预测准确，却为每个预测赋予相同的置信度水平

表 4. 定理的实证验证 3.5。我们提出的 ACR 始终满足公式 27中的条件，从而确保我们的方法严格降低均方误差并改进选择性预测。

数据集	数据集描述		QA-TIGER	ST-AVQA	TSPM
MUSIC-AVQA	N/A	σ_{M}^2	0.1542	0.1864	0.1564
		σ_{R}^2	0.4517	0.2225	0.1960
		σ_{MR}	**0.0966**	**0.1648**	**0.1384**
MUSIC-AVQA-R	Head	σ_{M}^2	0.2044	0.2038	0.1833
		σ_{R}^2	0.3847	0.2557	0.2323
		σ_{MR}	**0.1183**	**0.1865**	**0.1721**
	Tail	σ_{M}^2	0.2092	0.2375	0.1983
		σ_{R}^2	0.4168	0.3574	0.2516
		σ_{MR}	**0.1176**	**0.2166**	**0.1850**
MUSIC-AVQA-v2.0	训练平衡 , 测试平衡	σ_{M}^2	0.1754	0.1897	0.1848
		σ_{R}^2	0.7513	0.2951	0.2318
		σ_{MR}	**0.0535**	**0.1597**	**0.1595**
	训练平衡 , 测试有偏	σ_{M}^2	0.1737	0.1875	0.1830
		σ_{R}^2	0.7535	0.3060	0.2388
		σ_{MR}	**0.0524**	**0.1581**	**0.1601**
	训练有偏 , 测试平衡	σ_{M}^2	0.1912	0.2099	0.2039
		σ_{R}^2	0.7433	0.2587	0.2426
		σ_{MR}	**0.0528**	**0.1684**	0.1636
	训练有偏 , 测试有偏	σ_{M}^2	0.1676	0.1827	0.1808
		σ_{R}^2	0.7763	0.2647	0.2403
		σ_{MR}	**0.0531**	**0.1535**	**0.1488**

。例如，如果模型对 x% 的输入预测准确，却始终给出 x% 的置信度评分，而不考虑输入内容，那么它便是完美校准的。然而，该模型无法区分正确预测与错误预测，这意味着它无法为放弃预测提供有用的信号。

在单模态分类中，温度缩放（Guo 等人，2017年）和 Platt 缩放（Platt 等人，1999年）等技术通过调整分类器输出，使其基于验证集生成校准概率，从而提升校准效果。本研究考察向量缩放（VS）（Guo 等人，2017年）这一 Platt 缩放的多类别扩展方法在提升选择性预测性能方面的有效性。

我们认为，多模态推理任务（例如音视频问答（AVQA））的可靠性取决于两个基本组成部分：(i) 经过良好校准的置信度估计，即实现更低的期望校准误差（ECE）；以及 (ii) 识别并拒绝不可靠预测的能力，也就是说，知道何时提供答案以及何时应当避免作答。我们的结果表明，与其他方法（包括最大软最大概率、蒙特卡洛丢弃和向量缩放）相比，我们提出的策略自适应置信度精炼（ACR）显著改善了这两个组成部分。

蒙特卡洛丢弃。基于集成的技术已被用于选择性预测，其中拒绝预测的决策通常基于集成的不确定性统计量（Freund 等人，2004年；Varshney，2011年）。然而，这些方法需要多个预训练模型，由于训练规模足够大的集成模型成本高昂，因此不切实际。

我们提出使用蒙特卡洛丢弃（MCD）（Gal & Ghahramani，2016年；Geifman& El-Yaniv，2019年），作为提升选择性预测性能的一种高效替代方案。MCD 通过在推理期间启用丢弃并进行多次随机前向传播来估计预测不确定性。该方法无需训练多个相互独立的模型，即可扰动网络的预测。尽管 MCD 被广泛用于不确定性估计问题，但只需根据预测的最大 Softmax 概率应用拒绝阈值，就能轻松将其用于这一目的。实证研究表明，与确定性模型相比，MCD 能够提升校准性和鲁棒性，从而在某些情况下改善选择性预测性能。然而，虽然 MCD 通常能改善绝对不确定性估计，但并不能保证为选择性预测实现最优的样本排序（Geifman & El-Yaniv，2017年；

Lakshminarayanan 等人，2017年）。具体而言，MCD 不确定性分数所产生的排序，对于最小化选择性风险而言仍可能并非最优，尤其是在整个风险-覆盖率谱上进行评估时。因此，与更简单的基于置信度的方法相比，MCD 可能仅带来有限优势，尤其是在推理成本和可扩展性方面。

在我们的实验中，我们使用 MCD 作为一种基于不确定性的稳健基线方法，用于选择性预测。按照标准做法，我们在推理期间启用丢弃层，并汇总多次随机前向传播得到的预测。与先前研究的结论一致，我们发现，尽管 MCD 能够提升校准效果，但其选择性预测性能往往受限于推理成本以及不够完善的置信度排序。这一局限促使我们探索更高效的替代方案，着重直接优化置信度分数。不过，我们也注意到，MCD 的性能并非在所有情境下都能稳定提升，包括分布内（ID）和分布外（OOD）情形以及有偏设置。

用于选择性预测的 Doctor 分数。传统的选择性预测依赖最大软最大概率（MSP）来衡量置信度。然而，MSP 常因容易产生过度自信，且未能利用预测分布“尾部”所包含的信息而受到诟病。为解决这些局限，近期一项研究Granese 等人（2021）提出了 DOCTOR，这是一种后处理、无需训练的方法，将关注重点从最大对数几率转向 softmax 向量 $\mathbf{p}(\boldsymbol{x})$ 的整体分布。该方法通过以下二次和定义置信度分数：

$$\mathcal{C}_{\mathrm{DOC}}(\boldsymbol{x}) = \sum_{k=1}^{K} p_k(\boldsymbol{x})^2, \tag{28}$$

该方法隐式利用了基尼不纯度与模型预测不确定性之间的关系。在概率单纯形上，该指标充当一种距离度量，会惩罚高熵分布；与 MSP 相比，这类分布中的模型概率质量更分散地分布在多个类别上。因此，DOCTOR 是一种稳健且计算高效的选择性分类基线方法，尤其适用于无法访问内部特征表示或训练时统计量的黑盒场景。然而，Doctor 分数与 MSP 存在相同的局限：它仅基于答案逻辑值进行计算，因此通常会忽略来自中间表示的宝贵不确定性信号。

D. 关于消融研究的更多讨论

我们进行了广泛的消融研究，以验证所提出组件的有效性。在消融研究部分的所有实验中，包括论文正文以及附录中针对 MUSIC-AVQA 数据集开展的实验，我们均使用 QA-TIGER 骨干网络。

表5. 在 MUSIC-AVQA 测试集上，采用 QA-TIGER 的两个学习头——残余风险头和置信度门控头——的不同架构

架构	$\mathcal{C}$@1%	$\mathcal{C}$@5%	$\mathcal{C}$@10%	$\mathcal{C}$@20%	AURC	ECE
单层线性结构	10.24	38.72	59.87	90.41	8.93	6.69
双层多层感知机	14.96	40.17	60.83	90.98	8.56	5.12
三层多层感知机（本文方法）	**17.21**	**43.07**	**61.39**	**91.96**	**8.41**	**4.88**
四层多层感知机	16.74	42.77	61.30	91.68	8.47	4.92

ACR 架构。如第4.4节所述，我们在 MUSIC-AVQA 数据集上，针对 QA-TIGER 模型的 ACR 采用多种架构进行了实验。为简便起见，在包括主要实验和消融研究在内的所有实验中，我们对两个学习头采用了相同的架构。我们发现，这种一致性通常能取得最佳性能。我们的 ACR 采用三层多层感知机（MLP）。在表5中，我们观察到，为 ACR 使用更简单的单层线性结构会导致性能显著下降，尤其是在低风险水平下。相反，采用四层多层感知机的更复杂设计在所有指标上的表现均略低于我们所选的设计。综合这些结果与表3中的结果，我们得出结论：输入表示和训练目标至关重要。因此，提升学习到的选择函数性能的努力应主要集中在这些因素上，而不是设计更复杂的架构。

两个学习头的有效性。我们进行了实验，以验证双头设计以及修正基于MSP的置信度、而非完全取而代之的策略。如表 6 所示，同时使用两个头能带来更出色的

Table 6. 关于 ACR 学习头的消融研究。**RRH**：残余风险头，**CGH**：置信度门控头 d.

组件		$\mathcal{C}@\mathcal{R}\uparrow$				AURC ↓	ECE ↓
RRH	CGH	$\mathcal{C}$@1%	$\mathcal{C}$@5%	$\mathcal{C}$@10%	$\mathcal{C}$@20%		
✗	✓	14.57	41.94	60.77	91.63	8.61	7.82
✓	✗	13.78	40.11	60.31	91.24	8.72	9.14
✓	✓	**17.21**	**43.07**	**61.39**	**91.96**	**8.41**	**4.88**

最优选择性预测性能，并在多项指标上均有体现。这些指标包括目标风险水平下的最大覆盖率、恢复曲线下面积（AURC）以及期望校准误差（ECE）。值得注意的是，移除残差风险头（RRH）会导致性能显著下降，这凸显了它在修正与MSP方法提供的基础置信度信号相关的错误案例方面发挥着不可或缺的作用。

总体而言，我们的设计实现了**校准**与**选择**方法的升级与整合，使我们能够同时发挥两种方法的优势。具体而言，当移除 RRH 组件后，ACR 仅作为一种校准方法发挥作用（Platt 等人，1999年）。相反，如果移除 CGH 组件，它就会像以往的工作一样作为 Selector 运行（Geifman & El-Yaniv，2017年；Whitehead等人，2022年）。通过消融研究，我们得出以下几点认识：

1. 移除残余风险头（RRH）后，我们的框架简化为一种基于校准的方法。尽管该变体在所有评估指标上均较纯 MSP基线略有提升（如表 1 所示），但提升幅度十分有限。这表明，仅使用基于MSP的置信度很难取得显著改进。

2. 移除置信度门控头（CGH）后，我们的框架转变为一种基于选择器的方法。有趣的是，尽管既有研究（Geifman &El-Yaniv，2019年；Whitehead等人，2022年）表明，基于选择器的设计显著优于 MSP，但我们的研究结果显示，该变体的性能甚至可能低于 MSP。这一差异凸显出，在其他选择性预测场景中有效的技术无法直接应用于音视频问答（音视频问答）。此外，这也强调了音视频问答作为一种三模态多模态推理任务所面临的独特挑战，其本质上不同于单模态或双模态设置。

自适应加权融合的有效性。表 8 分析了融合权重 α 对 $\mathcal{R}$-音视频问答性能的影响，并将固定融合规则与我们的输入自适应融合机制进行比较。随着 α 从 0.1 增至 0.8，各覆盖率水平和校准指标上的性能均持续提升，这表明 MSP 是一种强有力的主要置信度信号。然而，当 α 进一步增至 0.9 时，性能会下降或趋于饱和，这说明单一的全局融合权重无法兼顾简单和困难的输入。相比之下，采用输入自适应融合的 ACR 在所有指标上均取得最佳结果，包括 $\mathcal{C}@\mathcal{R}$、AURC 和 ECE。这证实了最优融合权重会因输入而异，从而使 $\alpha(\boldsymbol{x})$ 能够动态调整，并且其选择性风险表现严格优于任何固定融合策略。

ACR 与基于选择器的策略的比较。如第 3.4 节所述，我们的 ACR 提出了一种综合性的置信度估计框架，融合了校准和基于选择器的策略。我们在论文主体中明确将我们的方法与一种称为向量缩放的校准方法进行了比较。本节进一步将 ACR 与两种知名的基于选择器的方法进行比较：SelectiveNet（Geifman& El-Yaniv，2019年）和 Selector（Whitehead等人，2022年）。SelectiveNet 是一种端到端框架，采用三头架构联合优化特征、预测和选择分数，同时在训练期间施加特定的“覆盖率”约束。然而，这种联合训练可能不适用于 $\mathcal{R}$-音视频问答任务，因为音视频问答模型通常使用预训练模型作为音频、视频和文本编码器。因此，它们很少从头开始训练所有部分。因此，在 $\mathcal{R}$-音视频问答场景中应用专为从头训练而设计的 SelectiveNet，可能会损害选择性预测性能。另一方面，Selector 方法充当外部“裁判”，利用视觉和文本特征之间的对齐关系，在视觉问答（VQA）中进行选择性预测。与在 $\mathcal{R}$-音视频问答任务中应用 SelectiveNet 相比，这种方法的训练过程更加稳定。

表 7 使用 QA-TIGER 骨干网络，在 MUSIC-AVQA 数据集上从所有指标对我们的 ACR 与 SelectiveNet 和 Selector 进行了直接比较。结果表明，在 $\mathcal{R}$-AVQA 场景中仅采用 SelectiveNet，会导致选择性预测性能较差，甚至不如朴素的 MSP 方法。尽管 Selector 方法

表 7. 消融实验：比较我们的 ACR 与基于选择器的方法。指标：所有选择函数在 MUSIC-AVQA 数据集上的目标风险下的覆盖率（$\mathcal{C}@\mathcal{R}$）↑、AURC ↓和 ECE ↓。均以 % 为单位。

模型 f	选择函数 g	$\mathcal{C}$@1%	$\mathcal{C}$@5%	$\mathcal{C}$@10%	$\mathcal{C}$@20%	AURC	ECE
QA-TIGER	SelectiveNet（Geifman & El-Yaniv，2019年）	9.72	37.74	56.91	87.26	9.92	11.31
	Selector（Whitehead等人，2022年）	12.27	40.31	58.83	89.75	9.02	10.19
	ACR（我们的方法）	**17.21**	**43.07**	**61.39**	**91.96**	**8.41**	**4.88**

表8. 输入自适应融合机制的有效性。指标：MUSIC-AVQA 数据集上所有选择函数的目标风险下的覆盖率（$\mathcal{C}@\mathcal{R}$）↑、AURC ↓和ECE ↓。均为%。

α 值	$\mathcal{C}$@1%	$\mathcal{C}$@5%	$\mathcal{C}$@10%	$\mathcal{C}$@20%	AURC	ECE
0.1	10.69	38.77	57.41	88.94	9.83	10.64
0.5	12.66	40.11	59.91	89.48	9.62	9.75
0.7	14.02	41.42	59.86	89.83	8.78	9.67
0.8	15.73	42.08	61.17	90.82	8.46	6.79
0.9	14.47	41.99	60.71	90.35	8.49	8.11
ACR（我们的方法，输入自适应融合）	**17.21**	**43.07**	**61.39**	**91.96**	**8.41**	**4.88**

的表现优于 SelectiveNet，但仍不及我们的 ACR 方法。这表明，像 Selector 那样依赖单个额外头部，无法充分捕捉音视频问答任务中固有的、更具挑战性的跨模态不确定性信号。相比之下，我们的方法始终优于基于选择器的策略和 MSP 方法，表明其具有鲁棒性，适用于 $\mathcal{R}$-音视频问答任务。

E. MSP 与 ACR 的补充分析和讨论

我们对最大软最大概率（MSP）与所提出的自适应置信度细化（ACR）在音视频问答（AVQA）选择性预测中的表现进行了比较分析。比较重点包括分布、校准、正确与错误预测之间的区别，以及选择性预测的质量。我们使用 TSPM 骨干网络，在全部三个数据集上进行评估。

E.1. 分布分析

图 5 展示了两种方法对正确和错误预测所得置信度分数的分布特征。

分布重叠与区分度。对于 **MSP 基线方法（第一列）**，直方图（图 5a）显示，正确预测（绿色）与错误预测（红色）之间存在显著重叠。正确预测和错误预测的平均置信度分数分别为 0.838 和 0.681，二者仅相差 0.16。图 5b 中的核密度估计图进一步揭示了这一问题：在 0.6 至 0.9 的置信度范围内，两者明显重叠，难以区分正确和错误的预测。该重叠区域涵盖了两种分布中的相当一部分，因此 MSP 作为可靠性指标的效用受到限制。此外，图 5c 中的箱线图也证实了这种重叠：两组的四分位距彼此交叠，且正确预测中有若干离群值低至 0.2 至 0.4。这表明，MSP 在评估错误预测子集时可能容易产生过度自信。**我们提出的 ACR（第二列）**展现出显著改善的区分度。图 5a 中的直方图显示，正确预测和错误预测的平均置信度分数分别为 0.803 和 0.627，二者相差 0.176，较 MSP 提升了 12%。值得注意的是，图 5b 中的核密度估计图呈现出更加明显的双峰分布。正确预测的分布在 1.0 附近有一个更尖锐的峰值，而错误预测则在较低置信度区域（0.4–0.7）呈现出更高的密度。尤其是，在决策边界区域（0.6–0.8）的重叠减少了；该置信度范围内的选择性预测决策最为敏感，这表明判别能力有所增强。

方差与分布形状。箱线图比较（图 5c）凸显了分布特征的关键差异。MSP 在正确预测中表现出较高的方差，四分位距约为

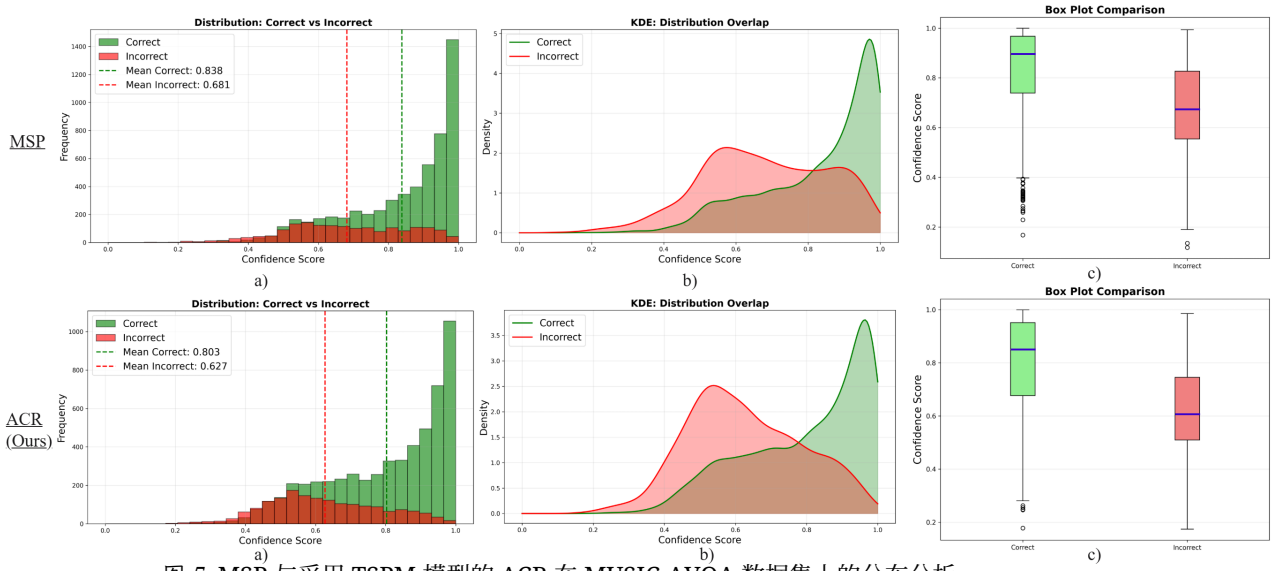
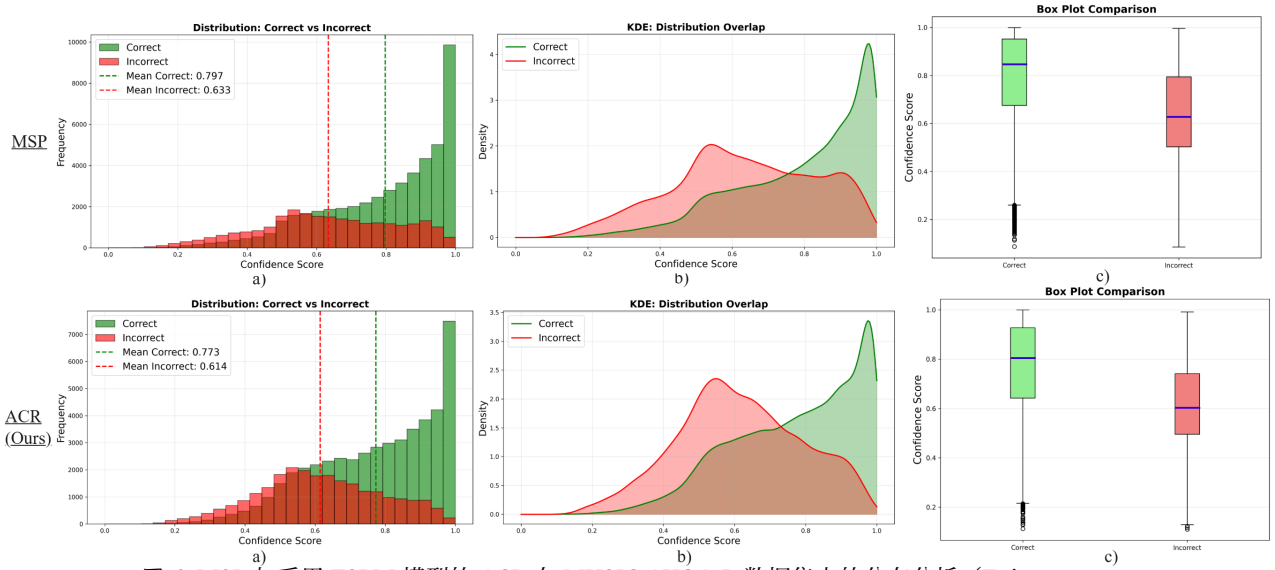


图 5. MSP 与采用 TSPM 模型的 ACR 在 MUSIC-AVQA 数据集上的分布分析。



*Fi*图 *6.* MSP 与采用 TSPM 模型的 ACR 在 MUSIC-AVQA-R 数据集上的分布分析（Tai l).

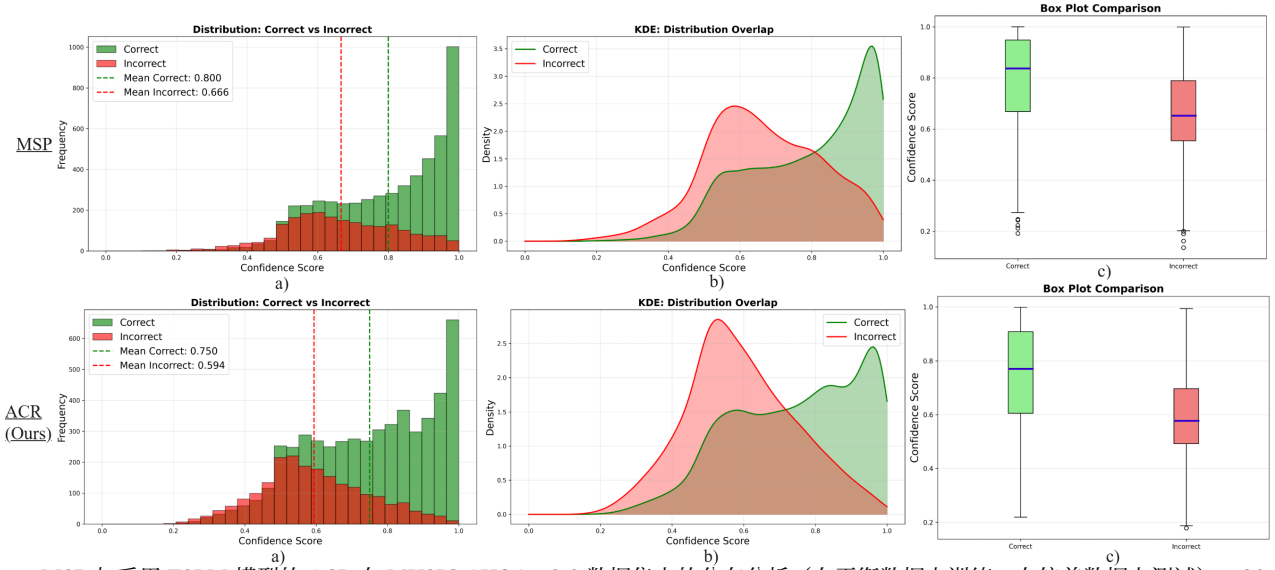


图 7. MSP 与采用 TSPM 模型的 ACR 在 MUSIC-AVQA-v2.0 数据集上的分布分析（在平衡数据上训练，在偏差数据上测试）。20

0.75-1.0，且包含大量低置信度离群值。这表明置信度校准不一致。相比之下，ACR 对正确预测给出的置信度水平更集中，离群值也更少，说明其置信度估计更可靠。对于错误预测，ACR 的中位数更低（~ 0.7 与 ~ 0.8相比），这有助于更好地区分错误预测与正确预测。

在图 6和 7中，MUSIC-AVQA-R（尾部）和MUSIC-AVQA-v2.0（在平衡集上训练，在偏置集上测试）也呈现出类似的模式。分布分析表明，ACR 在区分正确预测和错误预测方面优于 MSP，均值差提高了 12%，且分布重叠减少。这些改进提升了选择性预测能力，使实际视频问答场景中的风险—覆盖率权衡更为有效。然而，挑战依然存在：两种方法仍有大量高置信度错误预测，箱线图中的离群值便体现了这一点。这表明不确定性量化仍有待进一步改进。未来的工作将致力于缩小模糊区域，并通过不确定性量化领域的先进技术提升校准效果。

E.2. 校准与选择性预测分析

图 8 对置信度校准和选择性预测能力进行了详细评估，从校准准确度、累积分布特征以及风险—覆盖率权衡三个关键维度，对比了MSP方法（第一列）与我们提出的ACR方法（第二列）。

校准图分析。在图 **8**a所示的MSP基线（第一列）校准图中，我们观察到不同置信度分箱存在系统性的校准失准，这些分箱持续偏离完美校准对角线。虽然最大的分箱（由黄色圆圈标示，约包含4,000个样本）显示出近乎完美的校准，但这反映的是多数类的行为，即模型往往过度自信。值得注意的是，中等范围的分箱（置信度在0.4至0.7之间）表现出更明显的对角线偏离，凸显了在选择性预测决策最为重要的关键区间存在较差的校准。此外，较低置信度水平下较小的分箱（以较深的紫色表示）包含的样本较少，这表明MSP很少会为错误预测分配低置信度。相比之下，**我们的ACR方法**如图 8a第二列所示，展现出更优的校准特征。首先，各分箱在整个置信度范围内分布更加均匀，在0.5至0.7区间内具有较多的样本（以较深的颜色表示）。这表明ACR利用了更广泛的置信度水平，而不是将预测聚集在少数区间。其次，尽管中等范围的分箱仍存在轻微的校准失准，但与MSP相比，其偏离对角线的程度有所减小。高置信度分箱（同样由黄色圆圈标示，约包含4,000个样本）保持了出色的校准；更重要的是，ACR在错误预测集中的较低置信度分箱中生成了更多样本。这种分布表明，ACR能够更好地为不确定的预测分配适当较低的置信度，这对于实现有效的选择性预测至关重要。总的来说，尽管两种方法都表现出校准失准的迹象，但ACR展现出更加均匀的分箱分布，并且能够更好地利用置信度范围。

累积分布函数分析。如图 8b 所示，累积分布函数（CDF）曲线深入揭示了区分正确与错误预测的能力。对于**MSP方法**，红色曲线（代表错误预测）与绿色曲线（代表正确预测）分离较晚；在置信度达到约 0.5 之前，两条曲线始终相互接近，之后才开始分离。相比之下，我们的 **ACR** 则表现出更早、更明显的分离。错误预测（红色）的分布在较低置信度水平处上升更陡，表明更多错误预测被赋予了恰当的低置信度分数。当置信度为 0.6 时，MSP方法仅过滤掉约 36% 的错误预测，同时还会将约 10% 的正确预测误分类。相比之下，我们的 ACR 方法表现出显著更强的区分能力：约 50% 的错误预测低于此阈值，同时保留了约 84% 的正确预测（只有 16% 的正确预测低于 0.6 的置信度）。重要的是，ACR 在决策关键区域（0.6-0.9）的整个选择范围内始终保持这一优势，而非仅在个别点上表现出优势。

风险-覆盖率曲线下面积。AURC指标以单一数值概括了所有覆盖率水平下的选择性预测性能，数值越低表示性能越好，如图 8c所示。对于MSP，AURC为10.40%，风险-覆盖率曲线显示，随着覆盖率降低，风险逐渐下降。相比之下，我们的ACR取得了更低的AURC分数9.76%，较MSP **提升了6.15%** 。更重要的是，该曲线初始斜率更陡，表明其早期拒绝更加有效。

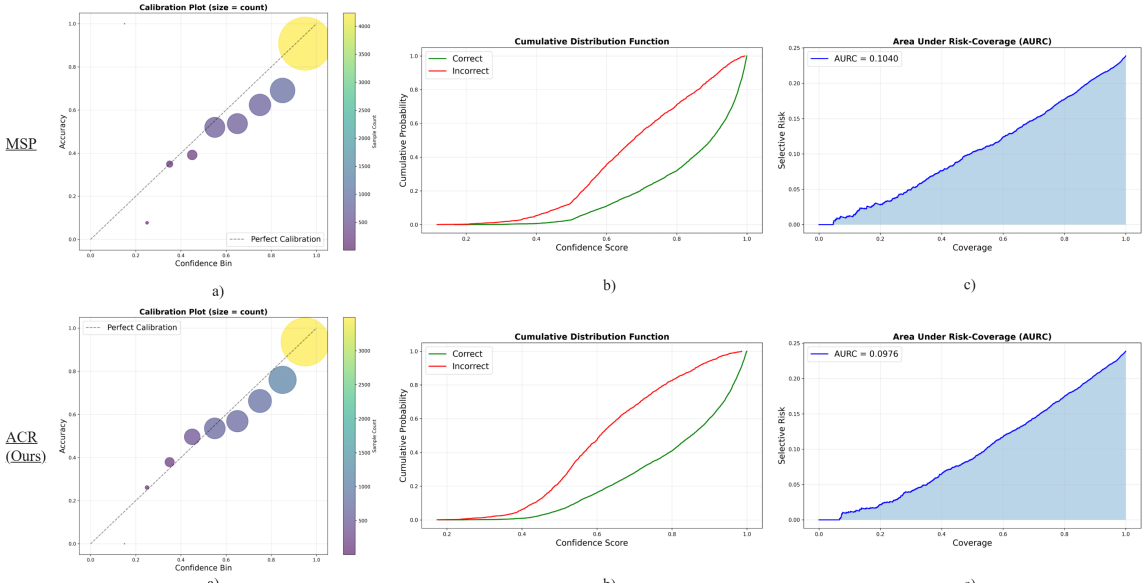


图 *8.* MUSIC-AVQA 数据集上 MSP 与采用 TSPM 模型的 ACR 的校准和 AURC 分析。

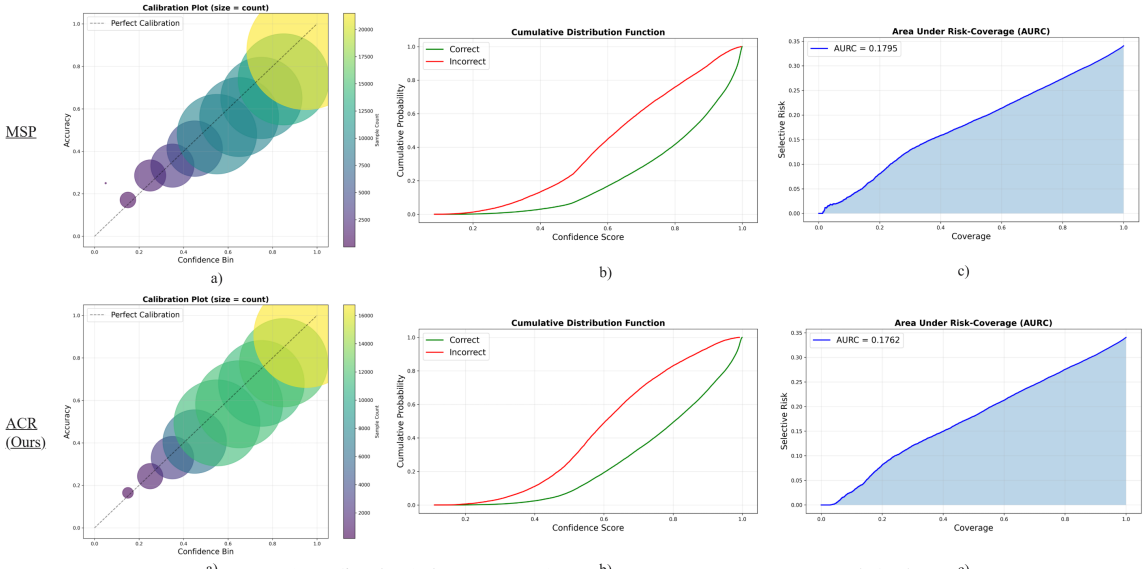


图 *9.* MUSIC-AVQA-R 数据集（尾部）上 MSP 与采用 TSPM 模型的 ACR 的分布分析。

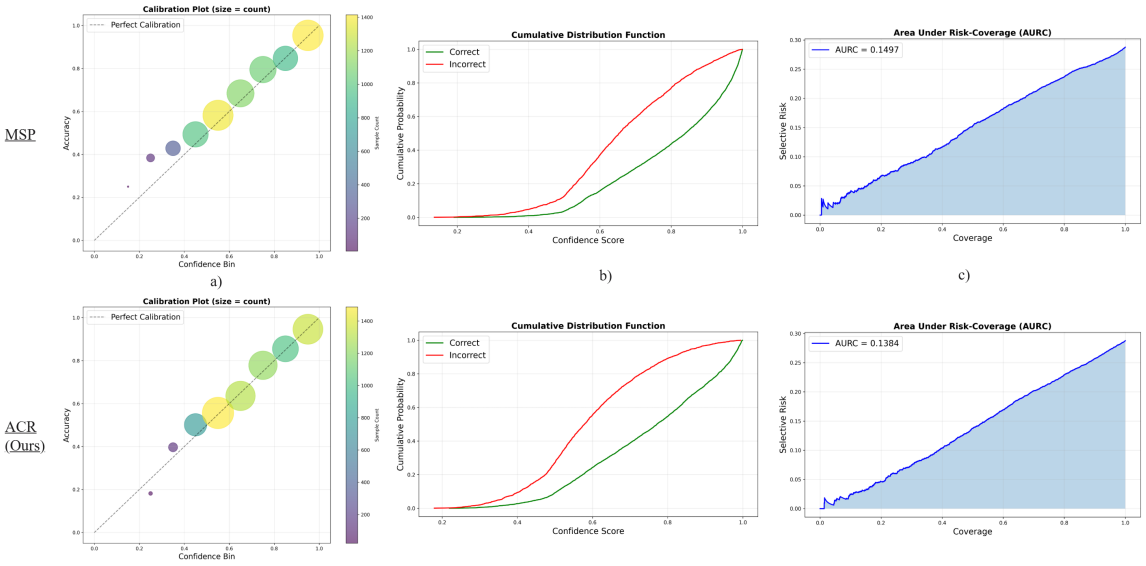


图 *10.* 分布分析 MUSIC-AVQA-v2.0 数据集上 MSP 与采用 TSPM 模型的 ACR 的分析（在平衡数据上训练，在偏差数据上测试 ）.

E.3. 分离质量的统计指标

为了全面评估 在各种置信度估计方法下，正确预测与错误预测之间的分离质量 n
方法，我们采用四项 互补的统计指标。

1. Cohen's d：该指标以合并标准差对均值差异进行标准化，得出不受量纲影响的效应量，同时考虑组内变异性。这使得不同研究之间可以进行比较；数值高于 0.8 表示差异较大，且在实践中具有显著意义。

2. Wasserstein距离：该指标量化将一个置信度分布转换为另一个所需的最小"运输成本"。它是一种稳健且对称的距离度量，可直接根据原始置信度尺度进行解读，并且与累积分布函数（CDF）之间的面积密切相关。

3. KL 散度（Kullback-Leibler 散度）：该指标衡量分布之间的信息论差异，反映它们的相异程度。它对尾部行为尤为敏感，但不具备对称性。

4. 风险-覆盖率曲线下面积（AURC）：该指标通过衡量所有覆盖率水平上的平均错误率来量化选择性预测性能；数值越低，表示通过基于置信度拒绝样本来降低风险的能力越强。

这些指标共同构成了一个全面的框架，用于评估不同置信度估计方法的有效性。我们在表 9 中报告了 MSP 和我们的 ACR 选择函数的这些指标。可以看出，在衡量正确预测与错误预测之间分离质量的四项代表性指标上，我们的 ACR 始终优于 MSP，从而在选择性预测场景中取得更好的表现。

表 *9.* TSPM 主干网络上分离质量的汇总统计指标。

数据集	选择函数 g	Cohen's d ↑	Wasserstein距离 ↑	KL 散度 ↑	AURC ↓
MUSIC-AVQ	MSP	0.9794	0.1569	11.11	10.40
	ACR	**1.0547**	**0.1758**	**14.22**	**9.76**
MUSIC-AVQ-R	MSP	0.8827	0.1589	8.88	17.95
	ACR	**0.9074**	**0.1640**	**10.22**	**17.62**
MUSIC-AVQA-v2.0	MSP	0.8160	0.1335	8.15	14.97
	ACR	**0.9139**	**0.1558**	**11.12**	**13.84**

E.4. 学习到的 $\alpha(x)$ 分布分析及其与定理3.6

图 11 展示了学习到的自适应融合权重的分布，记为 $\alpha^*(x)$，涵盖六种实验配置。该分析揭示了系统性模式，从而在经验上支持定理 3.6。对于基础 AVQA 准确率更高的 QA-TIGER 骨干网络（上排），学习到的权重往往取值更高，均值范围为 $\alpha = 0.858$ 至 0.932。这表明 Selector 主要依赖 MSP 的置信度，同时也会整合来自 RRH 的学习到的残余风险信号。相反，基础 AVQA 准确率较低的 TSPM 主干网络（下排）呈现出更宽的权重分布，均值更低（$\alpha = 0.828$ 至 0.870），且方差显著更高。这在配置 (c) 中尤为明显，其中权重几乎覆盖了 $[0.4, 1.0]$ 的整个取值范围。

这种依赖于架构的行为与定理 3.6 的预测完全一致。当 MSP 的质量发生变化时（QA-TIGER 较高，TSPM 较低），近最优逐点融合权重 $\alpha^*(x)$会相应调整。这在所有设置中都产生了 $\mathrm{Var}[\alpha^\star(x)] > 0$ 的方差。重要的是，没有任何分布坍缩为固定值（即 $\alpha = 0$ 或 $\alpha = 1$），这证实该网络学习到的是真正的输入自适应融合，而非诉诸平凡解。TSPM 上观察到的显著更大方差进一步表明，当 MSP 可靠性较低时，Selector 会按样本自适应地降低其影响，这正是定理 3.6 为最优逼近贝叶斯选择器 $g^*(x)$所建议的做法。

此外，代表均值 α 的绿色竖线始终位于红色参考线（表示等权重，$\alpha = 0.5$）与最右侧分箱之间。这表明，学习到的策略能够智能地平衡 MSP 的校准强度与 RRH 提供的互补风险信号，而非简单地取均值或完全依赖

任一组件。总的来说，这些实证证据证实了所提出的自适应融合机制具有非平凡性，并能感知架构差异。它直接支持定理3.6 中的主张：当最优权重随不同输入而变化时，自适应融合优于固定权重方案。

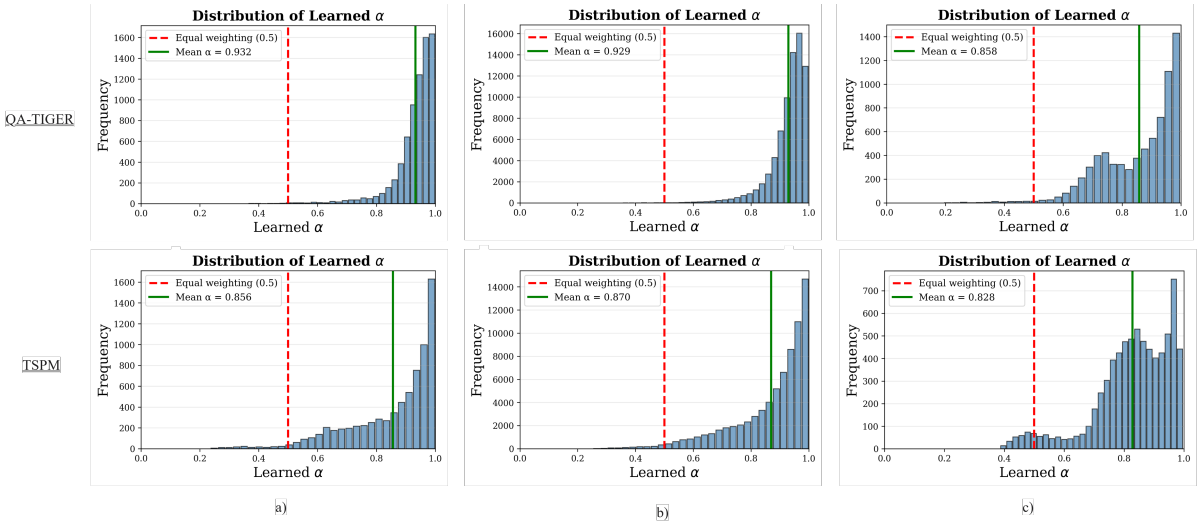


图 *11.* 基于 QA-TIGER 和 TSPM 主干网络的学习所得 α 分布分析：a) MUSIC-AVQA 数据集；b) MUSIC-AVQA-R 数据集（尾部）；以及 c) MUSIC-AVQA-v2.0 数据集（在均衡集上训练，在偏置集上测试）。

F. 计算复杂度与开销

自适应置信度细化（ACR）旨在改进标准音视频问答任务的选择性预测，同时避免引入显著的计算负担。ACR 完全在置信度估计阶段运行，不会修改音视频问答骨干模型或其推理路径。

参数复杂度。ACR 引入了两个轻量级组件：残差风险头（RRH）和置信度门控头（CGH）。二者均采用三层多层感知机（MLP）实现，隐藏维度记为 d。这两个头部引入的额外参数总数为 $\mathcal{O}(d^2)$。近期大多数音视频问答（AVQA）模型采用 Transformer 架构，其中包含视觉、音频和文本编码器的多个层，以及多模态融合层（如交叉注意力）。对于具有 L 个 Transformer 层的 AVQA 模型，这些基础 AVQA 模型的参数总数为 $\mathcal{O}(d^2 L)$，其中 $L \gg 6$。因此，与基础 AVQA 模型相比，我们新增头部所带来的额外参数可以忽略不计。例如，将我们的方法应用于 QA-TIGER 时，额外参数总量仅约占基础 AVQA 模型总体规模的 **3%** 。这表明，相对于基础 QAGE 架构，ACR 带来的参数开销可以忽略不计。

推理复杂度。 在推理过程中，ACR仅需通过浅层三层多层感知机（MLP）进行两次前向传播，并执行一次标量融合操作。每个样本产生的额外计算成本为 $\mathcal{O}(d^2)$，与多模态Transformer推理所涉及的 $\mathcal{O}(d^2 L)$ 复杂度相比微不足道。因此，ACR几乎不会引入延迟开销，也不会影响实时或批量推理场景下的吞吐量。

训练复杂度。 ACR采用两阶段训练策略。第一阶段使用标准交叉熵损失训练音视频问答骨干模型，之后将其冻结。第二阶段仅使用基于正确性标签的二元交叉熵训练RRH和CGH组件。第二阶段计算效率高、收敛速度快，并且无需对骨干模型进行重复微调。

内存开销。 ACR仅需存储额外的头部参数以及由骨干网络生成的中间特征表示。无需额外的内存缓冲区或集成模型。

ACR 可在现有音视频问答模型中实现，且只需极少的计算和内存开销，无需重新训练或进行架构变更。这一特性使 ACR 尤其适用于大规模或资源受限的部署场景，在这些场景中，可靠性至关重要。

G. 音视频问答模型的复现实验结果

表 *10.* 音视频问答模型在原始 MUSIC-AVQA 测试集上的复现准确率。所有数值均以 % 表示

音视频问答模型	原始准确率	复现准确率	Δ准确率
QA-TIGER（Kim 等人，2025年）	77.62	77.52	-0.10
ST-AVQA（Li 等人，2022年）	71.52	71.26	-0.26
TSPM（Li 等人，2024a）	76.79	76.51	-0.28

本节展示了我们在 $\mathcal{R}$-AVQA 基准测试中对实验所用所有音视频问答模型复现的结果，包括 QA-TIGER（Kim 等人，2025年）、ST-AVQA（Li 等人，2022年）和 TSPM（Li 等人，2024a）。汇总结果见表 10。我们重点关注具有公开实现和开源数据预处理流程的模型，以确保结果可复现并促进公平比较。所有模型均严格按照原论文中的描述进行训练和评估。如表 10所示，我们复现的结果与相应论文报告的性能高度一致，证实了实现的准确性以及实验设置的可靠性。

H. 与数据增强的比较

表 *11.* 目标风险下的覆盖率（$\mathcal{C}@\mathcal{R}$）↑、AURC ↓和 ECE ↓，适用于 MUSIC-AVQA 数据集上的所有选择函数。均以百分比表示。

模型 f	选择函数 g	$\mathcal{C}@1\%$	$\mathcal{C}@5\%$	$\mathcal{C}@10\%$	$\mathcal{C}@20\%$	AURC	ECE
QA-TIGER	MSP	14.29	41.53	60.47	91.83	8.67	8.83
	MSP-Augment	<u>15.41</u>	<u>41.98</u>	<u>60.77</u>	<u>91.90</u>	<u>8.56</u>	<u>6.23</u>
	ACR（我们的方法）	**17.21**	**43.07**	**61.39**	**91.96**	**8.41**	**4.88**

在我们的实验中，我们单独划分出一个验证集，用于验证音视频问答模型并训练选择函数，称为 Val-f（见表 16 和 17）。另一种做法是利用这些数据进一步扩充音视频问答训练集，以期提升基础模型的准确率。由于选择性预测的性能本质上与准确率相关（如第 4.2 节所述），这种数据增强也可能有助于基于 MSP 方法的选择。我们在 MUSIC-AVQA 数据集上使用 QA-TIGER 开展实验，以观察应用上述数据增强后，MSP 相较于我们提出的 ACR 方法的性能变化。

表 11 对比了这两种策略。我们发现，将额外数据用于训练选择器，在低风险水平下提升覆盖率方面始终更有效。尽管扩充音视频问答训练数据确实能提高整体准确率，但只有当风险容忍度接近基础模型的错误率时，这些提升才主要转化为更高的覆盖率。相比之下，直接改进选择函数（正如我们的方法所做的那样），在可靠性至关重要的低风险场景中具有显著优势。

I. 补充定量结果

由于篇幅限制，我们无法在正文中列出所有表格结果。本节报告其余所有表格结果，包括表 12和表 13。表 12 展示了 ST-AVQA 骨干网络在 MUSIC-AVQA-R 和 MUSIC-AVQA-v2.0（在均衡集上训练）上的性能，而表 13 报告了全部三种 AVQA 骨干网络在 MUSIC-AVQA-v2.0 数据集（在偏置集上训练）上的性能。

与主论文一致，我们评估了不同风险容忍度水平下的覆盖率，以及 AURC 和 ECE。在所有指标上，我们的方法 ACR 始终优于所有竞争方法，展现出强大的泛化能力和鲁棒性。值得注意的是，ACR 在低风险场景（例如 $\mathcal{C}$@1%、$\mathcal{C}$@5%）中表现出尤为显著的提升，这表明在安全关键情境中，当弃答优于做出错误预测时，ACR 更加可靠。

表*12*。 MUSIC-AVQA-R 和 MUSIC-AVQA-v2 数据集上的比较。对于 MUSIC-AVQA-R，我们报告头部（H）/尾部（T）结果；对于 MUSIC-AVQA-v2，我们报告平衡（Bal）/偏置（Bias）结果。指标：目标风险下的覆盖率（$\mathcal{C}@\mathcal{R}$）↑、AURC↓和 ECE ↓（均以 % 表示）。

数据集	模型 f	选择函数 g	$\mathcal{C}$@1%		$\mathcal{C}$@5%		$\mathcal{C}$@10%		$\mathcal{C}$@20%		AURC		ECE	
			H/Bal	T/Bias	H/Bal	T/Bias	H/Bal	T/Bias	H/Bal	T/Bias	H/Bal	T/Bias	H/Bal	T/Bias
MUSIC-AVQA-R	ST-AVQA	MSP	0.10	0.50	0.29	1.48	14.13	2.94	44.88	7.00	21.35	36.53	5.92	6.59
		MCD	0.10	0.44	0.42	1.39	14.19	2.74	44.85	6.76	21.34	36.68	6.02	6.56
		VS	0.11	0.44	0.25	1.53	13.99	3.04	44.66	7.47	21.42	36.03	5.91	6.40
		Doctor	6.04	0.23	18.56	13.49	34.24	26.01	63.52	59.31	15.41	16.81	11.87	11.95
		ACR（我们的方法）	**0.44**	**1.03**	**3.17**	**2.44**	**14.80**	**4.15**	**44.95**	**10.03**	**21.14**	**35.02**	**2.60**	**6.07**
		Oracle	63.45	51.33	66.12	53.50	69.79	56.47	78.52	63.56	7.98	14.78	0.00	0.00
MUSIC-AVQA v2.0	ST-AVQA	MSP	4.70	4.26	13.07	11.61	28.47	28.14	61.43	62.66	16.25	16.00	6.62	6.31
		MCD	4.76	4.24	13.23	11.29	28.69	28.67	61.29	62.80	16.28	16.01	6.32	6.11
		VS	4.75	4.30	13.68	11.62	28.93	28.43	61.41	62.86	16.24	15.99	6.54	6.14
		Doctor	4.70	4.26	13.01	11.55	28.45	27.97	62.35	63.94	16.24	15.99	3.95	4.26
		ACR（我们的方法）	**7.44**	**6.29**	**15.03**	**15.93**	**30.53**	**32.02**	**62.90**	**64.49**	**15.76**	**15.41**	**3.90**	**3.52**
		Oracle	69.30	69.80	72.22	72.74	76.23	76.78	85.76	86.38	5.54	5.36	0.00	0.00

J. 更多定性结果

图 12 展示了不同 AVQA 骨干网络（包括 QA-TIGER、ST-AVQA 和 TSPM）在多种选择策略下的更多定性示例。这些示例进一步说明了标准基于置信度的基线方法存在的一种普遍失效模式：面对多模态歧义时，它们往往会对错误预测赋予较高置信度。

在所有案例中，即使问题涉及细微的听觉比较（例如相对响度）、统计同时发声源的数量或识别时间顺序事件，MSP、MCD 和 VS 仍表现出严重的过度自信。这类问题需要对音视频证据进行细粒度推理，并且特别容易受到跨模态不一致的影响，例如视觉上显著但声学上模糊的线索，或相互重叠的声源。

相比之下，ACR 始终会分配较低的置信度分数，并在证据不足或相互矛盾时弃答。值得注意的是，无论底层骨干模型如何，这种行为都能观察到，这表明 ACR 捕捉到了不同架构共有的不确定性信号。这些结果进一步表明，ACR 的主要优势在于可靠的置信度估计和选择性弃答。

K. 阈值泛化

如第 4.2 节所述，我们通过测量最终的测试风险，评估针对特定目标风险水平在验证集上选出的阈值能否迁移至测试划分。我们以 QA-TIGER 作为代表性模型，比较 MSP 与我们的自适应置信度细化机制 ACR。表 14显示，两者产生的风险差异均极小：在 MUSIC-AVQA 和 MUSIC-AVQA-v2.0（在平衡集上训练，在偏差集上测试）数据集上，两种选择函数的差异均不超过0.39%p。值得注意的是，当 1% 为目标风险时，风险差异甚至更小。

我们还观察到，使用从验证集选出的阈值所达到的覆盖率，与相同风险水平下可实现的最大覆盖率之间存在差异，我们将其记为 ΔC。这些结果表明，基于风险的阈值在不同数据划分之间具有良好的泛化能力，但无法直接比较风险水平完全相同时的覆盖率。

L. 更多数据划分详情

我们使用 MUSIC-AVQA（Li 等人，2022年）、MUSIC-AVQA-R（Ma 等人，2024）和 MUSIC- AVQA-v2.0（Liu等人，2024）数据集开展实验，这些数据集均包含音视频问答对。表 15给出了问答统计数据的对比。表 16 和 17 总结了我们实验中使用的数据划分。对于 MUSIC-AVQA 和 MUSIC-AVQA-v2.0 数据集，我们还对其原始测试集进行了进一步划分，因为我们需要一个额外的数据集来验证选择函数 g，并评估风险与覆盖率。这些划分旨在确保不同数据集之间不共享任何视频，也不共享与之相关的音频流或问答标注。

表 *13.* 目标风险下的覆盖率（$\mathcal{C}@\mathcal{R}$）↑、AURC ↓和 ECE ↓，在 MUSIC-AVQA v2.0 上的结果（在偏置数据上训练；在平衡（Bal）和偏差（Bias）数据集上测试）。单位均为 %

模型 f	选择函数 g	$\mathcal{C}@1\%$		$\mathcal{C}@5\%$		$\mathcal{C}@10\%$		$\mathcal{C}@20\%$		AURC		ECE	
		Bal	Bias	Bal	Bias	Bal	Bias	Bal	Bias	Bal	Bias	Bal	Bias
QA-TIGER	MSP	3.63	5.55	28.71	36.81	46.67	58.94	82.22	93.15	11.51	9.17	13.58	11.16
	MCD	3.10	4.40	27.99	36.66	46.04	58.96	81.56	**93.48**	11.50	9.20	15.02	12.40
	VS	2.80	5.22	28.43	36.80	46.90	58.70	**82.37**	93.24	11.52	9.16	15.23	12.65
	Doctor	3.63	5.55	28.70	36.81	46.65	58.84	**82.37**	92.72	11.51	9.17	11.26	9.14
	ACR（我们的方法）	**3.88**	**12.55**	**30.13**	**38.94**	**47.23**	**59.58**	81.52	92.34	**11.30**	**8.80**	**7.56**	**2.78**
	Oracle	75.08	78.42	78.24	81.72	82.59	86.26	92.91	97.04	3.62	2.71	0.00	0.00
ST-AVQA	MSP	0.55	2.29	4.48	15.01	14.60	34.85	49.64	68.78	19.80	14.78	11.03	7.51
	MCD	0.55	2.40	4.23	14.57	14.48	**35.36**	49.56	69.06	19.82	14.80	11.03	7.61
	VS	0.57	2.38	4.76	14.95	14.64	34.95	49.63	68.70	19.76	14.80	10.76	7.35
	Doctor	0.55	2.26	4.47	14.97	14.60	34.95	49.53	68.75	19.78	14.76	4.89	3.06
	ACR（我们的方法）	**1.96**	**5.06**	**7.91**	**16.09**	**17.36**	34.84	**50.02**	**69.09**	**19.41**	**14.42**	**4.09**	**2.22**
	Oracle	66.40	70.80	69.19	73.78	73.04	77.88	82.17	87.62	6.69	5.00	0.00	0.00
TSPM	MSP	1.83	1.91	9.96	15.05	21.27	38.40	54.12	71.15	18.10	14.20	9.49	6.90
	MCD	1.85	1.95	9.95	15.43	21.18	38.67	54.10	71.16	18.12	14.21	9.45	6.81
	VS	1.81	1.88	10.28	15.95	21.97	38.24	53.95	71.23	18.04	14.27	9.57	7.10
	Doctor	1.83	1.91	9.89	15.21	21.15	38.53	54.25	70.45	18.14	14.11	5.51	3.58
	ACR（我们的方法）	**6.13**	**4.17**	**13.64**	**23.62**	**26.27**	**42.43**	**59.62**	**72.45**	**16.98**	**13.10**	**4.65**	**2.40**
	Oracle	69.79	72.28	71.69	75.33	75.67	79.51	85.13	89.45	5.74	4.49	0.00	0.00

表*14.* 使用 QA-TIGER 将弃答阈值 γ 从验证集迁移至测试集的泛化能力。 $\Delta\mathcal{R}$ 和 $\Delta\mathcal{C}$ 表示在验证期间为目标风险 $\mathcal{R}$ 选择的 γ 与 γ 之间的风险和覆盖率百分点（%p）差异，其中 γ 具有最大$\mathcal{C}@\mathcal{R}$。

数据集	选择函数 g $\mathcal{R}=$	$\Delta\mathcal{R}$ 1%	5%	10%	20%	$\Delta\mathcal{C}$ 1%	5%	10%	20%
MUSIC-AVQA	MSP	+0.19	+0.15	+0.37	+0.36	+0.13	+0.20	+1.86	+1.12
	ACR（我们的方法）	-0.06	+0.29	+0.39	-0.04	+0.28	+0.25	+1.78	-0.24
MUSIC-AVQA-v2.0	MSP	+0.08	+0.24	+0.34	-0.24	+0.18	+2.69	+2.74	-0.53
	ACR（我们的方法）	+0.07	+0.25	+0.33	-0.38	+0.24	+1.38	+2.03	-0.66

不同 t splits。这一策略与先前的研究一致（Whitehead等人，2022年；Dancette 等人，2023 ）

需要注意的是，留出测试集（在表 16 和17中称为“测试集”）在任何组件的训练或验证阶段都不会被访问，包括基础预测器 f 和选择函数 g。MUSIC-AVQA-R 数据集与 MUSIC-AVQA 数据集使用相同的训练集和验证集，仅新增了一个测试集。因此，我们将使用在 MUSIC-AVQA 数据集上训练的预训练模型来评估该测试集，无需为此数据集创建新的数据划分。该测试集仅用于最终评估。除非另有说明，所有报告结果均基于此留出测试集。

M. 更多模型细节

本节将全面介绍我们实验所用的模型架构、超参数和实现细节。

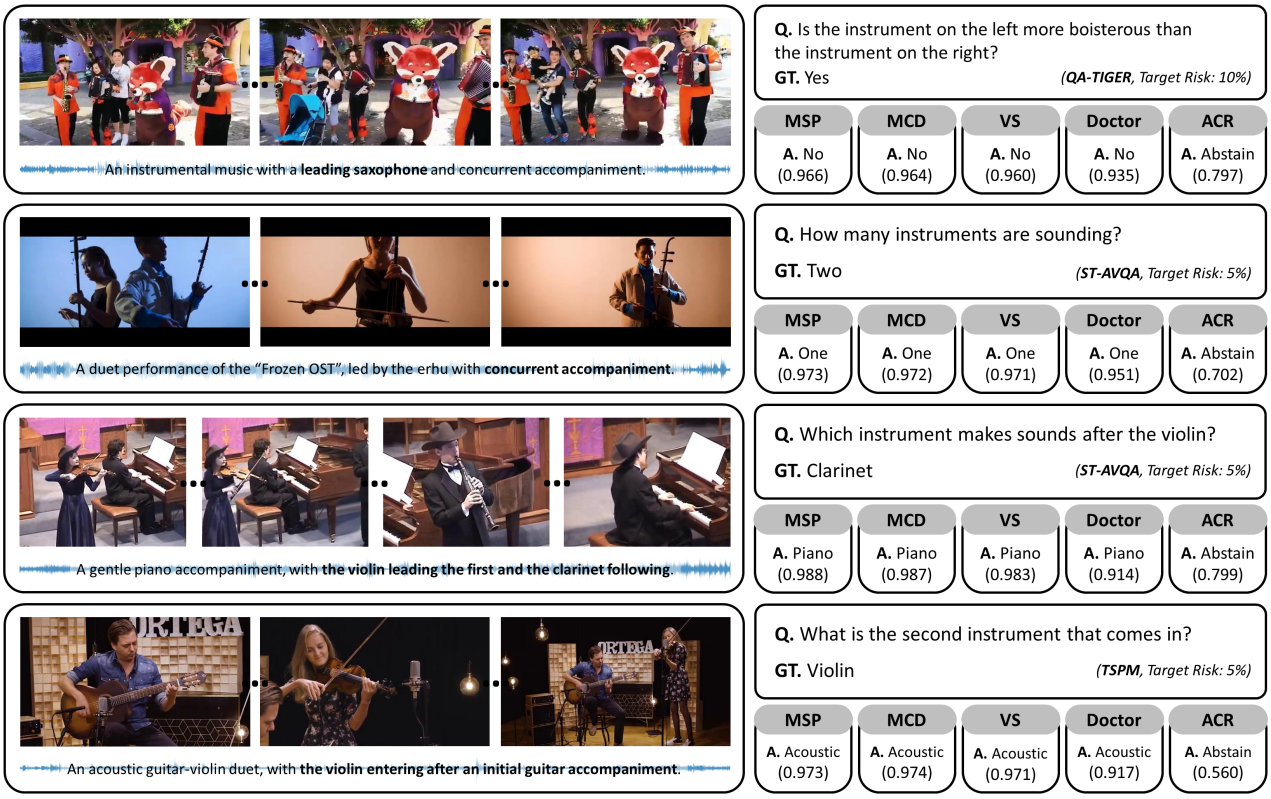


图 *12.* 基于各种骨干模型的选择性预测的其他定性示例。对于每种模型基线方法（QA- TIGER、ST-AVQA、TSPM），图中展示了其在不同选择策略下对应的选择性预测结果，如图所示。

表*15.* MUSIC AVQA数据集的问题统计比较

数据集	#QA	#类型	关键特征
MUSIC-AVQA	45.8K	33	9种问题类型
MUSIC-AVQA2	54K	33	平衡模板
MUSIC-AVQA-R	211K	33	25×改写；465个词汇。

M.1. 音视频问答模型

我们所有实验中使用的音视频问答模型均严格采用其原始开源代码。本研究选取了三种音视频问答架构（QA-TIGER[1] (Kim 等人，2025年）、ST-AVQA[2] (Li 等人，2022年）和 TSPM[3] (Li 等人，2024a），用于评测 $\mathcal{R}$-音视频问答基准并评估我们提出的方法。我们选择了架构各异、采用不同多模态融合策略的音视频问答模型进行基准测试，这些策略融合了音频、视觉和文本模态。多样性有助于稳健地评估所有选择函数（尤其是我们提出的 ACR）的有效性和泛化能力。训练音视频问答模型时，我们采用其原始论文和 GitHub 仓库中的超参数，如表 18所示。所有模型均将音视频问答模型（f）视为分类任务，并使用交叉熵损失进行训练。我们简要介绍实验中使用的模型和设置。

QA-TIGER (Kim 等人，2025年）。QA-TIGER 是音视频问答任务的先进方法，并获得了 CVPR 2025 的 Highlight 论文认可。它采用渐进式问题嵌入，并且

[1]https://github.com/AIM-SKKU/QA-TIGER [2]
https://github.com/GeWu-Lab/MUSIC-AVQA [3]
https://github.com/GeWu-Lab/TSPM

表 16. MUSIC-AVQA（Li 等人，2022年）数据集的数据划分详情

来源	划分名称	用途	% 来源
MUSIC-AVQA 训练集	训练	训练 f	100%
MUSIC-AVQA 验证	验证-f	验证 f，训练 g	100%
MUSIC-AVQA 测试	验证-g	验证 g，评估风险-覆盖率	20%
	Test	用于测试的留出测试集（f， g）	80%

表 17. MUSIC-AVQA-v2.0（Liu等人，2024）数据集的数据划分详情

来源	划分名称	用途	% 来源
MUSIC-AVQA-v2.0 训练	训练	训练 f	100%
MUSIC-AVQA-v2.0 验证	验证-f	验证 f，Tran g	100%
MUSIC-AVQA-v2.0 测试	验证-g	验证 g，评估风险-覆盖率	20%
	Test	用于测试的留出测试集（f, g）	80%

通过专家混合（MoE）（Shazeer 等人，2017年）架构对连续时间动态进行建模。音频嵌入由在 AudioSet（Hershey 等人，2017年）上预训练的 VGGish 模型提取，用于生成 1 秒片段的帧级特征；视觉嵌入则采用 CLIP（Radford 等人，2021）生成 [CLS] 基于 token 的帧级表示。问题的文本嵌入由 CLIP 的文本编码器生成，产生句子级 [EOT] token 和词级特征。我们使用 MUSIC-AVQA 和 MUSIC-AVQA-v2.0 的原始训练集训练该模型。

ST-AVQA (Li 等人，2022年)。ST-AVQA 是音视频问答任务的早期研究之一，并入选 2022 年 CVPR 口头报告。ST-AVQA 采用后期问题融合，并使用由分层空间定位模块和时间定位模块组成的两阶段架构，以提高 AVQA 准确率。音频嵌入由预训练的 VGGish 模型处理 1 秒片段得到，再通过线性投影将原始 128 维输出转换为 512 维特征。视觉嵌入则使用预训练的 ResNet-18 从每个片段中采样的帧（每秒 1 帧）提取，得到能够捕捉空间细节的 512 维特征图。问题的文本嵌入由长短期记忆网络（Graves，2012年）基于投影后的 512 维词嵌入生成，最终隐藏状态作为问题表示，注入时间模块作为查询，并在融合阶段进行逐元素相乘，以引导注意力并增强跨模态融合。我们在与 QA-TIGER 类似的数据集上训练该模型。

TSPM (Li 等人，2024a）。TSPM 是一种较新的音视频问答模型，在多个数据集上取得了令人瞩目的表现，为音视频问答研究引入了一种新方法。TSPM 利用中间的、由问题引导的渐进式感知，并采用强大的提示定位。音频嵌入由预训练的 VGGish 模型从 1 秒片段中提取，每个片段生成 128 维特征。视觉嵌入由冻结的 CLIP（ViT-L/14）模型提取，生成 512 维的帧级 [CLS] 标记，这些标记来自按 1 fps 采样的帧和块级标记特征；随后将它们合并，以减少冗余，同时保留关键语义。问题的文本嵌入通过 CLIP 的文本编码器处理经分词的词语，生成 512 维的句子级特征。

M.2. 选择函数

我们在此详细介绍向量缩放（VS）和自适应置信度细化（ACR）选择函数。由于最大软最大概率（MSP）、蒙特卡洛丢弃（MCD）和 Doctor 均无需额外训练，因此不在此介绍。训练每个选择函数时，我们会冻结音视频问答模型的权重 f。

校准。校准的输入是 AVQA模型的未归一化答案logits（即 softmax 之前的值，或 softmax 层之前的答案表示），输出则是校准后的logits。由于我们采用向量缩放（Platt 等人，1999年；Guo 等人，2017年），因此将 AVQA模型的逻辑值输入带有对角权重矩阵和

表18. 我们实验中使用的各模型的超参数。所有这些超参数均直接引用自原始论文。

超参数	QA-TIGER（Kim 等人，2025年）	ST-AVQA（Li 等人，2022年）	TSPM（Li 等人，2024a）
批量大小	32	64	64
隐藏维度	512	512	512
优化器	Adam（等，2014）	Adam（等，2014）	Adam（等，2014）
Adam ϵ	$1e$-8	$1e$-8	$1e$-8
Adam β_1	0.95	0.90	0.90
Adam β_2	0.999	0.999	0.999
学习率	$1e$-4	$1e$-4	$1e$-4
学习率调度器	StepLR	StepLR	StepLR
步长	8	8	8
丢弃	0.1	0.1 & 0.5	0.1
# 训练轮次	15	30	30

偏置项。在训练期间，经过线性层后，我们应用 Sigmoid激活函数；与（Guo 等人，2017年）不同，我们将所得激活值作为输入，使用 AVQA标签计算二元交叉熵损失。我们使用 Adam优化器训练线性层，学习率为 8e-5。隐藏维度设为 512。我们沿用原始 AVQA 主干网络的相同超参数，如表18所示。测试时，我们使用该线性层的输出作为校准后的logits，应用 softmax，并采用与 MSP 相同的弃权程序。

自适应置信度精炼。我们的自适应置信度细化（ACR）的输入包括中间多模态表示（即基础音视频问答模型分类器头的输入）和未归一化答案logits（即紧接在Softmax层之前的答案表示）。我们的ACR包含两个学习头（残余风险头和置信度门控头），其架构细节、前向传播和第二阶段训练见表 19 和算法 1、2。我们提供了自适应置信度精炼（ACR）的高层次 PyTorch 风格伪代码。在置信度精炼期间，音视频问答骨干模型被冻结。为确保可复现性，我们还在补充材料中提供了实际的 PyTorch 实现。为兼顾简洁性和效率，我们对两个头均采用相同的三层多层感知机（MLP）架构。我们在第二阶段使用 Adam优化器，以 8e-5 的学习率联合训练这两个头。隐藏维度设置为 $d' = 512$（中间多模态特征维度）+ 42 （逻辑值维度）。我们保留与原始音视频问答骨干网络相同的超参数，如表 18所示。在测试期间，我们根据公式10计算 ACR 的置信度分数。

表 19. 残差风险头（RRH）和置信度门控头（CGH）的架构。$\mathbf{h}$ 和$\mathbf{z}$ 分别表示融合多模态表示和答案逻辑值。

层	配置
输入	[d维特征$\mathbf{h}$，42维特征$\mathbf{z}$]
FC-1	线性层($d' \to d'$) +ReLU + 丢弃(p=0.1)
FC-2	线性层($d' \to d'$) +ReLU + 丢弃(p=0.1)
全连接层-3（输出）	线性层($d' \to 1$) + Sigmoid函数

算法 1 自适应置信度细化（ACR）：前向传播

需要： 输入样本 $\boldsymbol{x} = (a, v, q)$
需要： 冻结的 AVQA 主干网络 $f(\cdot)$
需要： 残余风险头 RRH(·)
需要： 置信度门控头 CGH(·)
1：获取答案逻辑值和融合表示：

$$(\mathbf{z}, \mathbf{h}) \leftarrow f(\boldsymbol{x})$$

2：计算 MSP 置信度：

$$C_{\mathrm{M}}(\boldsymbol{x}) \leftarrow \max_k \operatorname{softmax}(\mathbf{z})_k$$

3：预测残余风险置信度：

$$C_{\mathrm{R}}(\boldsymbol{x}) \leftarrow \mathrm{RRH}(\mathbf{h})$$

4：预测置信度门控：

$$\alpha(\boldsymbol{x}) \leftarrow \mathrm{CGH}(\mathbf{h})$$

5：融合置信度：

$$C_{\mathrm{ACR}}(\boldsymbol{x}) \leftarrow \alpha(\boldsymbol{x})\ C_{\mathrm{M}}(\boldsymbol{x}) + (1 - \alpha(\boldsymbol{x}))\, C_{\mathrm{R}}(\boldsymbol{x})$$

6: **返回** $C_{\mathrm{ACR}}(\boldsymbol{x})$

算法 2 使用二元交叉熵训练 ACR

要求： 训练集 $\{(x_i, y_i)\}_{i=}^{N}$ 1
要求： 冻结的音视频问答主干网络 $f(\cdot)$
要求： RRH 和 CGH 参数 $\theta_{\mathrm{R}}, \theta_{\mathrm{G}}$
要求： 二元交叉熵损失 $\mathcal{L}_{\mathrm{BCE}}$
1: **对**每个小批次 $\mathcal{B}$ **执行**
2: 计算逻辑值和表示：

$$(\mathbf{z}_i,\ \mathbf{h}_i) \leftarrow f(x_i),\ \ \forall x_i \in \mathcal{B}$$

3: 计算 MSP 置信度：

$$C_{\mathrm{M}}(x_i) \leftarrow \max_k \operatorname{softmax}(\mathbf{z}_i)_k$$

4: 预测残差置信度并进行门控：

$$C_{\mathrm{R}}(x_i) \leftarrow \mathrm{RRH}(\mathbf{h}_i), \qquad \alpha(x_i) \leftarrow \mathrm{CGH}(\mathbf{h}_i)$$

5: 融合置信度：

$$C_{\mathrm{ACR}}(x_i) \leftarrow \alpha(x_i) C_{\mathrm{M}}(x_i) + (1 - \alpha(x_i)) C_{\mathrm{R}}(x_i)$$

6: 计算正确性标签：

$$c(x_i) \leftarrow \mathbb{I}[\hat{y}_i = y_i]$$

7: 通过最小化以下内容来更新参数：

$$\mathcal{L}_{\mathrm{BCE}}(C_{\mathrm{ACR}}(x_i)\ c(x_i))$$

8: **结束循环**

N. TSPM 的标准差

在本节中，为评估结果稳定性，我们给出了 MUSIC-AVQA 数据集上 TSPM 的均值和标准差。这些结果与主论文表 1 中报告的结果略有不同，后者基于单个随机种子。表20 报告了五个随机种子下的性能，显示不同选择函数之间的方差较低（标准差为 0.04 至 0.05）。鉴于多随机种子评估的计算成本，我们验证了 QA-TIGER 和 ST-AVQA 表现出

不同随机种子下的稳定性相近，因此，对于这些架构，我们报告单次运行的结果，如表1所示。这确保我们的结论并非随机初始化造成的假象。

表*20*。MUSIC-AVQA 数据集上采用 TSPM 主干网络时，不同选择函数对应的风险-覆盖率、AURC 和 ECE 指标的均值与标准差。所有数值均以 % 表示。

模型 f	选择函数 g	$\mathcal{C}$@1%	$\mathcal{C}$@5%	$\mathcal{C}$@10%	$\mathcal{C}$@20%	AURC	ECE
TSPM	MSP	8.60 ± 0.66	31.06 ± 0.69	49.64 ± 0.27	87.56 ± 0.30	10.46 ± 0.05	4.45 ± 0.23
	MCD	8.77 ± 0.29	29.90 ± 0.46	50.25 ± 0.48	87.30 ± 0.54	10.47 ± 0.05	4.20 ± 0.44
	VS	6.70 ± 0.34	30.13 ± 0.69	50.42 ± 0.36	87.54 ± 0.35	10.46 ± 0.05	4.02 ± 0.43
	Doctor	8.63 ± 0.57	31.02 ± 0.61	49.57 ± 0.33	87.64 ± 0.31	10.47 ± 0.05	2.19 ± 0.20
	ACR（我们的方法）	8.92 ± 0.48	35.53 ± 0.38	54.22 ± 0.50	88.67 ± 0.34	9.74 ± 0.04	1.64 ± 0.11

O. 与校准和保形预测的关系

与校准的关系。不同于校准方法（Platt 等人，1999年；Zadrozny 和 Elkan，2002；Guo 等人，2017年），这些方法通过学习逻辑值的单调变换来最小化期望校准误差（ECE）；自适应置信度排序（ACR）则通过在中间特征上训练判别式选择器，直接优化选择性预测。校准侧重于概率匹配，确保具有特定置信度的样本在相应比例的情况下预测正确。相比之下，选择性预测要求实现最优排序，即正确预测的置信度高于错误预测。这两个目标可能并不一致：经过良好校准的模型仍可能无法对样本进行最优排序，而 ACR 则可以利用输出层之外更丰富的特征表示重新排列样本。

与保形预测的关系。我们的自适应置信度细化（ACR）框架基于一项与保形预测（Vovk 等人，2005；Angelopoulos 等人，2023）共有的基本原则：两种方法都使用留出的校准集来改进可能未经校准的模型输出，并生成更可靠的不确定性估计。在保形预测中，根据校准集计算的非符合度分数用于构建预测集，从而提供形式上的覆盖率保证（例如，真实标签以至少 $1-\alpha$ 的概率包含在预测集中）。类似地，ACR 使用留出的选择集 D_{select} 来训练选择器头，使其学习基础模型的自然错误模式，而不是简单地记住训练样本。然而，这两种方法的目标和保证有所不同。保形预测为预测集提供无分布覆盖率保证，而 ACR 则侧重于学习一种具有判别性的置信度函数，专为基于排序的选择性预测而设计，以在特定风险水平下最大化覆盖率。我们的方法可视为一种经过学习、针对特定任务的替代方案，它以牺牲形式保证为代价，换取选择性预测中更优的经验性能。重要的是，ACR 与保形方法相辅相成：选择器生成的精炼置信度分数可作为保形框架中的增强型非符合度分数，为未来研究提供了一个值得探索的方向。

P. 补充讨论与未来工作

选择性预测旨在通过允许模型在很可能出错时放弃给出预测，提升机器学习系统的可靠性。我们的工作通过探索多模态不确定性下的置信度估计，并为音视频问答提出输入自适应置信度融合策略，为实现这一目标做出了贡献。据我们所知，我们是首个正式定义可靠的 音视频问答问题的研究。通过处理模态间的错位和异质置信度信号，我们的方法在不改变底层任务预测器的情况下，改善了选择性风险与覆盖率之间的权衡。这对于安全关键型或面向用户的多模态人工智能应用尤其有益，例如辅助系统、教育工具和多媒体检索；在这些场景中，不准确的答案可能会误导用户并损害他们的信任。

然而，如果弃答对某些输入类型或用户群体的影响过大，选择性预测系统可能会造成信息获取不平衡，尤其是在不同模态的数据质量可能存在差异的多模态情境下。虽然我们的方法无法消除这些风险，但旨在通过改善模态错位或意见不一致情况下的置信度排序，减少任意或过度自信的失效。我们强调，在部署选择性模型时，透明说明弃答行为和失效模式的重要性。总体而言，这项工作加深了对多模态学习中可靠性的理解，并鼓励进一步研究有原则的不确定性估计，以及选择性预测系统的负责任部署。

当前的自适应置信度细化（ACR）设计用于答案空间固定且有限的分类式音视频问答模型。最大软最大概率（MSP）和残余风险头都依赖于针对封闭候选答案集合定义的逻辑值。然而，ACR目前是为离散答案选择而实现的；不过，其基本原则——输入自适应置信度精炼——并不局限于分类任务。未来一个有前景的探索方向是扩展ACR以支持自由形式生成。这可能包括在词元或序列层面进行残余置信度校正，并利用生成概率、熵或自洽性信号。我们将在未来工作中对这些扩展进行全面研究。

尽管ACR显著改善了正确预测与错误预测之间的区分，我们在附录E中的分析表明，仍有一部分错误预测获得了较高的置信度分数。这些失败案例通常源于复杂的多模态歧义，例如相互冲突的音视频线索或时间上较为细微的要求；在这些情况下，基础音视频问答模型和Selector都难以检测出不可靠性。

未来工作的一个重要方向，是将自适应置信度精炼从封闭集音视频问答扩展到具备开放式生成能力的多模态大语言模型（MLLMs）。在此类场景中，置信度估计不能依赖固定的答案空间，而必须考虑推理链中的不确定性、词元或序列层面的置信度，以及各生成输出之间的自洽性信号。

Q. 更广泛的影响

可靠的不确定性估计对于部署可信人工智能至关重要，尤其是在安全关键型应用中，错误预测可能带来严重后果。我们的自适应置信度细化（ACR）框架通过使模型能够更准确地识别其预测何时可能不正确，从而实现有依据的弃答或转交人类判断，有助于达成这一目标。

在音视频问答领域，增强型选择性预测可显著提升无障碍应用的效果，例如为视障或听障用户提供的辅助系统。这样可以确保系统仅在有把握时才作答，而不是冒险提供误导性信息。

学习弃答这一原则适用于多个高风险领域，包括医学诊断、自主导航和内容审核。在这些领域，错误预测的代价远高于将问题交由人类专业人员处理的代价。不过，我们也认识到这一方法存在潜在风险。实践者可能过度依赖置信度分数，而未充分了解其局限性；置信度估计器在不同人口群体或内容子群体中的表现也可能不一致，进而导致弃答率不平等。

我们敦促实践者在部署前评估 ACR 在不同子群体中的表现，并在决策时将置信度分数视为众多因素之一，而非衡量正确性的绝对标准。我们的工作并未引入生成有害内容的新能力，而是提供工具，使现有人工智能系统更加可靠，并能更好地认识自身局限。