音频语言模型在倾听吗？用于自适应音频引导的音频专用注意力头

Neta Glazer[1], *Lenny Aharon*[2],*Ethan Fetaya*[1]

[1] 巴伊兰大学，拉马特甘，以色列[2] 哥伦比亚大学，纽约，纽约州，美国neta.glazer@biu.ac.il, Lenny.Aharon@columbia.edu

arXiv:2603.06854v1 [cs.SD] 6 Mar 2026

摘要

多模态大语言模型可能表现出文本主导，过度依赖语言先验，而不是将预测建立在非文本输入之上。一个例子是大型音频语言模型（LALMs）：即使音频证据包含重要信息，模型也可能未能充分利用其中的决定性音频证据。为解决这一问题，我们利用机制可解释性识别出一小组音频专精注意力头，其音频注意力会产生“倾听”信号。我们表明，当音频证据影响模型输出时，该信号会增强，从而为标准提示下的音频参与度提供一种指标。借助这一定位结果，我们构建了音频－静音引导方向，并在推理时对最终表示实施激活干预，增强模型的音频效应。为展示这一干预的实用性，我们在 MMAU 上表明，在无需任何参数更新的情况下，该方法可使两个基于通义千问的大语言音频模型的准确率最高提升 +8.0 个百分点。

索引词：语音识别、可解释性、大型音频语言模型

1. 引言

大型音频语言模型（大语言音频模型，LALMs）将预训练音频编码器与基于解码器的大语言模型（LLM）结合，使模型能够根据自然语言提示对语音、环境声音和音乐进行指令遵循理解与推理 [1, 2, 3, 4]。在架构上，常见的融合策略包括：(i) 将音频编码器表征投影到大语言模型嵌入空间中，并将其作为一小段音频条件伪词元序列插入，通过自注意力与文本共同处理 [5, 2], ；以及 (ii) 通过额外添加的（通常为门控）交叉注意力适配器层，利用音频特征对大语言模型进行条件化，而不是将音频词元与文本流交错排列 [3]。

然而，在主要以文本训练的单一主干网络中处理这些混合词元流，会引入一种关键现象：文本主导。即使非文本模态提供了丰富的信息，模型也常常过度依赖语言线索。近期的系统性证据表明，这种失衡普遍存在于各个模态中，包括音频，其成因包括融合设计选择，以及非文本词元冗余导致的注意力稀释 [6, 7]。相关工作将这一现象称为语言先验偏差，并表明多模态输出往往更多受底层大语言模型的先验驱动，而非非文本输入本身 [8]。具体到音频语言领域，受控的不匹配研究表明，能力较强的大语言音频模型可能会受文本提示主导，实际上忽略与之矛盾的音频证据 [9, 10]。

机制可解释性已成为分析纯文本大语言模型内部计算的核心框架，旨在识别权重和激活值中能够因果性地驱动模型行为的局部机制，而不是依赖事后解释 [11, 12, 13]。近年来，这套工具开始拓展至多模态架构，包括大语言音频模型，使研究者能够在组件层面分析非文本模态如何融入模型，以及模态特有的失效模式会出现在哪里 [14, 15, 16, 17]。这一范式的关键在于对内部激活进行因果干预（例如消融或激活修补），以检验机制假设。基于这一思路，引导是指在推理时进行干预，通过修改内部激活来影响信息的处理方式 [18]。一项反复出现的发现是，Transformer组件，尤其是单个注意力头，呈现出稳定且专门化的计算作用，从而能够在注意力头层面进行因果干预，并推动基于激活值的引导方法，在推理过程中调节模型行为 [12, 19, 20]。

受大语言音频模型中的文本主导失效模式，以及多模态 Transformer 机制可解释性近期进展的启发，我们使用机制工具深入研究音频利用不足问题。具体而言，我们探讨头级信号是否能够指示模型何时正在利用音频，以及这些信号能否作为推理时转向的实用手段。

在本研究中，我们提出了两项主要贡献。首先，我们找出一小组音频专精注意力头，它们对音频的注意力能够预测正确性，从而提供实例级的“聆听”信号。其次，我们证明，在音频语言模型中，组件级的机制分析能够提供可操作的切入点：利用定位结果指导受控的推理时激活干预，我们增强了模型的音频效应，并提升了两个基于通义千问的大语言音频模型（Qwen2-Audio-7B [1]和 R1-AQA [21]）在 MMAU 上的表现，且无需进行任何参数更新。

2. 相关工作

多模态大语言模型中的文本主导。 大语言音频模型（大语言音频模型）通过音频前端和多模态融合机制（如词元注入）扩展了指令跟随大语言模型，使其能够对音频和文本进行联合推理[1, 3]。然而，这类系统存在一项可靠性隐患：文本主导（或语言先验偏差），即语言线索压过了有价值的非文本证据 [6, 22]。这一现象已在多模态场景中得到系统性记录：即使强烈的语言先验与感知证据相冲突，模型仍会默认依赖语言先验，进而导致虚假相关、模态利用不足，以及无法恰当地

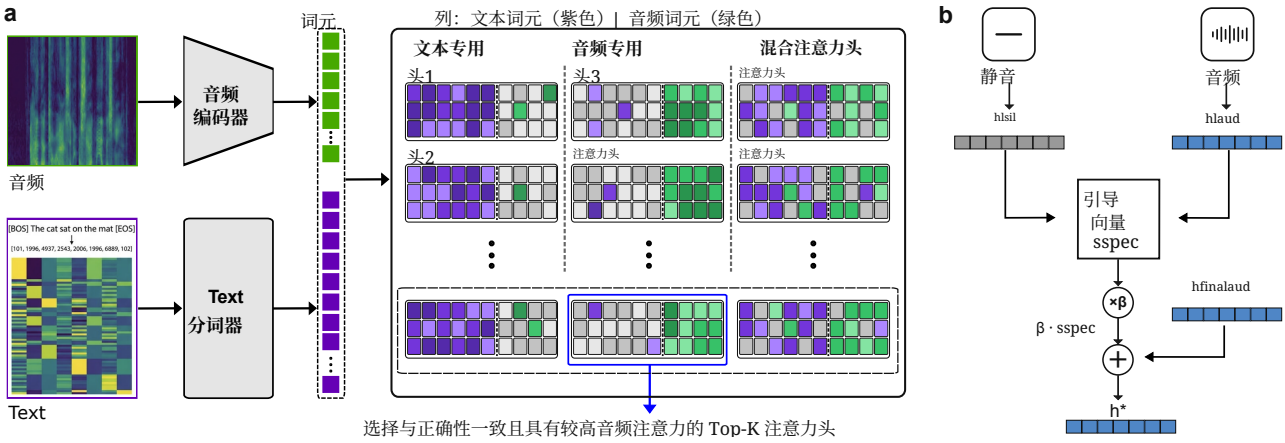

图 1：专用头引导（***SGS***）。***(a)*** 我们通过计算每个注意力头的音频注意力占比，并选取在校准集上其音频注意力对正确性预测能力最强的 *Top-K* 个注意力头，来识别音频专精注意力头。***(b)*** 我们分别进行音频和时长匹配的静音前向传播，并在专家层集合 $\mathcal{L}$（包含已发现注意力头的层）上聚合残差差异，形成层局部化的转向方向$(\mathbf{h}_\ell^{aud} - \mathbf{h}_\ell^{sil})$。我们将该方向按 β 缩放，并将其加到最终表征中，得到用于预测的 $\mathbf{h}^*$。

将预测建立在非文本信号之上 [8, 23, 24, 25]。在音频领域，受控的音频与文本不一致研究直接揭示了模态仲裁失效：即使文本指令与声学真实情况直接矛盾，模型仍倾向于遵循文本指令 [9, 10]。此外，语音情感评估表明，许多大语言音频模型的行为更像僵化的转录器，而非主动的聆听者，无法将声学韵律与词汇内容区分开来 [26]。

机制可解释性。机制可解释性提供了一系列工具，用于定位信息在 Transformer 计算过程中被表示的位置以及被表示和使用的方式，从而将模型行为与其内部机制联系起来，而非依赖事后给出的解释 [27]。一个反复出现的发现是，单个组件——尤其是注意力头——往往具有专门且可复用的功能角色 [20],，从而能够进行有针对性的组件级干预。近年来，这些工具也被扩展到多模态 Transformer，包括识别与模态相关的注意力头并研究其在不同任务中的因果作用的分析 [16, 17, 28]。在音频和 LALM 领域，早期研究运用机制分析，探查声学证据如何在模型中传播，并定位与音频相关的计算过程 [14]。另一条互补研究路线探索无需训练的推理时干预，利用此类定位来提升多模态落地能力，例如针对音频模型的向量引导方法 [29, 30]。另一条互补研究路线探索无需训练的推理时干预，利用此类定位来提升多模态落地能力，包括通过向内部表征添加学习得到的或对比式方向来进行激活引导 [31, 32]。

3. 预备知识与记号

词元流与音频索引。 音频语言转换器在单个 n 词元序列上运行，其中包含文本词元和音频词元。我们用 $\mathcal{I}_{\text{audio}} \subset \{1, \ldots, n\}$ 表示音频词元索引的集合。

多头自注意力。给定输入 x，自注意力在层 ℓ 和头 h 中生成一个注意力矩阵 $\mathbf{A}_{\ell,h}(x) \in \mathbb{R}^{n \times n}$，其各行之和为 1，其中 $\mathbf{A}_{\ell,h}[i,j](x)$ 表示从查询位置 i 到键位置 j 的注意力权重。

来自最终提示词元的音频注意力。令 i_{final} 表示提示中的最终词元索引（生成前的最后一个位置）。对于头 (ℓ, h)，我们计算

$$a_{\ell,h}(x) = \sum_{j \in \mathcal{I}_{\text{audio}}} \mathbf{A}_{\ell,h}[i_{\text{final}}, j](x). \tag{1}$$

由于注意力行的总和为一，$a_{\ell,h}(x) \in [0,1]$ 表示从查询位置 i_{final} 指向音频词元的注意力占比。

残差流状态。 设 $\mathbf{h}_\ell(x) \in \mathbb{R}^{d_{\text{model}}}$ 表示经过第 ℓ层后位置 i_{final} 的残差流表示，设 $\mathbf{h}_{\text{final}}(x)$ 表示该位置的最终层表示。

音频消融。 为隔离音频的影响，我们采用时长匹配的静音基线。对于每个示例 x，我们定义 x^{aud} （原始音频）和 x^{sil} （以相同时长的零值替代音频），其对应的残差流状态分别为 $\mathbf{h}_\ell^{\text{aud}}(x)$ 和$\mathbf{h}_\ell^{\text{sil}}(x)$。

引导干预。 给定一个方向和强度 β，我们通过在前向传播过程中修改内部激活值来实施引导，然后使用模型的语言建模头计算预测。

4. 方法

我们的方法分为两个阶段。首先，我们找出一小组音频专精注意力头，这些注意力头对音频的注意力最能预测校准划分中的正确性，从而在注意力头层面定位音频参与度。其次，我们利用这一定位结果构建受专用头约束的音频–静音引导方向，并在推理时施加受控的激活干预。

4.1. 发现音频专用注意力头

音频注意力信号。我们使用最终提示位置的音频注意力质量 i_{final}, $a_{\ell,h}(x)$（公式 1）作为衡量音频参与度的头级度量。

专精度评分和选择。 使用由多项选择题构成的留出校准集 $\mathcal{D}_{\text{cal}}$，我们定义二元正确性标签 $y(x) = \mathbb{1}[\hat{c}(x) = c^*(x)] \in \{0,1\}$，其中$\hat{c}(x)$是模型的预测选项， $c^*(x)$是真实标注

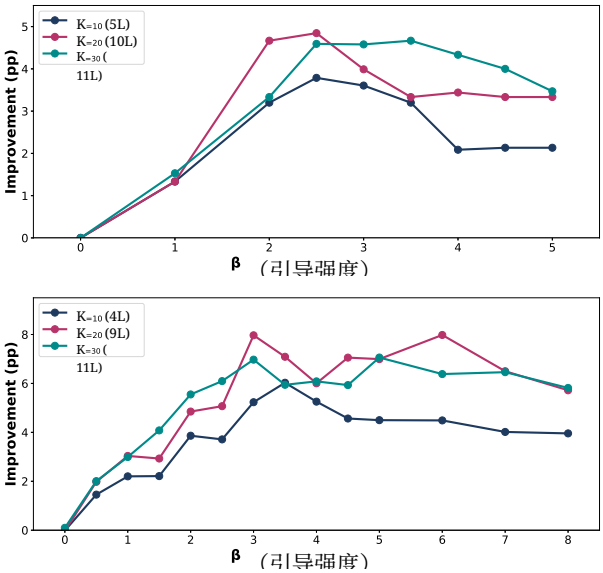

图 2： *R1-AQA*（上）和 *Qwen2-Audio-7B*（下）上的引导强度 β 和专家数量 K 对性能的影响。曲线展示了不同 *Top-K* 专家头集合带来的性能提升百分点（*pp*）；每个 K 都会形成一个专家层集合 $\mathcal{L}$，并在 $\mathcal{L}$内应用层局部引导。

选项。我们根据$a_{\ell,h}(x)$与正确性之间的关联，对每个头（ℓ, h）进行评分：

$$\rho_{\ell,h} = \mathrm{corr}\Big(\{a_{\ell,h}(x)\}_{x\in\mathcal{D}_{\text{cal}}}, \{y(x)\}_{x\in\mathcal{D}_{\text{cal}}}\Big), \tag{2}$$

其中，corr 表示皮尔逊相关系数（对于二元 y，等价于点二列相关）。我们将专家集合 $\mathcal{H}_{\text{spec}}$ 定义为按$|\rho_\ell$ 排序的 Top-K 注意力头，$_h|$（我们使用 K=20）。计算实例级聆听分数时，我们对注意力头的贡献进行带符号（按 $|\rho|$加权）聚合，因此，与正确性负相关的注意力头会产生相反方向的贡献。

聚合专家参与度。 对于任意示例 x，我们通过带符号聚合来概括专家参与度

$$A_{\text{spec}}(x) = \frac{1}{\sum_{(\ell,h)\in\mathcal{H}_{\text{spec}}} |\rho_{\ell,h}|} \sum_{(\ell,h)\in\mathcal{H}_{\text{spec}}} \rho_{\ell,h}\, a_{\ell,h}(x), \tag{3}$$

并将 $A_{\text{spec}}(x)$ 用作实例级聆听指标。

验证协议。 我们仅使用 $\mathcal{D}_{\text{cal}}$ 选择 $\mathcal{H}_{\text{spec}}$，并在一个不重叠的评估划分上报告所有分析结果。我们进行两项健全性检查：(i) $A_{\text{spec}}(x)$ 能够预测正确性（例如使用 AUC 衡量），且优于匹配的随机头基线；(ii) 在模型预测于音频条件化运行和音频消融运行之间发生变化的示例中，$A_{\text{spec}}(x)$ 高于预测未发生变化的示例。

4.2. 层引导

令 $\mathcal{H}_{\text{spec}} = \{(\ell_1, h_1). ..,(\ell_K, h_K)\}$ 为专家头集合，并定义专家层集合

$$\mathcal{L} = \{\ell : \exists h \text{ such that } (\ell, h) \in \mathcal{H}_{\text{spec}}\}. \tag{4}$$

对于每个 $\ell \in \mathcal{L}$，令 $n_\ell = \big|\{h : (\ell, h) \in \mathcal{H}_{\text{spec}}\}\big|$ 并设置 $w_\ell = n_\ell/K$ (因此 $\sum_{\ell\in\mathcal{L}} w_\ell = 1$)。

转向方向（层局部化）。对于输入 x，我们运行两次前向传播（x^{aud} 和 x^{sil}），并在最终提示位置 i_{final} 提取残差流状态 $\mathbf{h}_\ell^{\text{aud}}(x)$，$\mathbf{h}_\ell^{\text{sil}}(x)$。我们定义

$$\mathbf{s}(x) = \sum_{\ell\in\mathcal{L}} w_\ell \Big(\mathbf{h}_\ell^{\text{aud}}(x) - \mathbf{h}_\ell^{\text{sil}}(x)\Big). \tag{5}$$

推理时转向。 我们通过以下方式引导最后一层的表示

$$\mathbf{h}^*(x) = \mathbf{h}_{\text{final}}^{\text{aud}}(x) + \beta\, \mathbf{s}(x), \tag{6}$$

并通过语言建模头，从 $\mathbf{h}^*(x)$ 计算预测结果。

4.3. 头级转向

为了检验这些增益是否依赖层局部引导，我们还考虑一种直接的头级干预基线。

逐头注意力输出。 令 $\mathbf{u}_{\ell,h}(x) \in \mathbb{R}^{d_h}$ 表示第 ℓ层中注意力头 h在位置 i_{final}处、注意力输出投影 $W_O^{(\ell)}$ 之前的输出。我们定义逐头音频特定差分：

$$\Delta\mathbf{u}_{\ell,h}(x) = \mathbf{u}_{\ell,h}^{\text{aud}}(x) - \mathbf{u}_{\ell,h}^{\text{sil}}(x). \tag{7}$$

映射到残差流。 将 $\Delta\mathbf{u}_{\ell,h}(x)$ 视为列向量，通过 $W_O^{(\ell)}$ 中相应的切片将其映射到残差流空间：

$$\Delta\mathbf{c}_{\ell,h}(x) = W_O^{(\ell)}[:, hd_h : (h+1)d_h]\Delta\mathbf{u}_{\ell,h}(x) \in \mathbb{R}^{d_{\text{model}}}. \tag{8}$$

˜干预。 令 $\mathbf{h}_\ell(x)$ 表示第 ℓ层中注意力子层之后（且在 MLP 之前）位置 i_{final}处的隐藏状态。定义 $\mathcal{H}_{\text{spec}}(\ell) = \{h : (\ell, h) \in \mathcal{H}_{\text{spec}}\}$。对于每个$\ell \in \mathcal{L}$，我们在注意力子层的输出处将头级干预加到残差流中：

$$\tilde{\mathbf{h}}_\ell^*(x) = \tilde{\mathbf{h}}_\ell^{\text{aud}}(x) + \beta \cdot \frac{1}{|\mathcal{H}_{\text{spec}}(\ell)|} \sum_{h\in\mathcal{H}_{\text{spec}}(\ell)} \Delta\mathbf{c}_{\ell,h}(x), \tag{9}$$

然后照常继续前向传播，以获得最终预测。

5. 实验设置

我们在大规模多任务音频理解（MMAU）基准测试 [33], 上进行评估。该基准包含与多项选择题配对的音频片段，涵盖三个领域：语音、环境声音和音乐，并涉及 27 项技能，采用标准化数据划分。我们报告带标签的 MMAU test-mini 数据划分（1,000 个示例）上的准确率，并按领域（语音/声音/音乐）进一步细分结果。

基线与干预。 我们评估：(i) 不进行干预；(ii)最佳单层音频–静音引导；(iii)使用专用注意力头或同等规模的随机头对照进行头级转向（通过$W_O^{(\ell)}$ 映射逐头输出）；(iv)用于层引导的匹配随机头对照（相同的 K 和相同的流程）；以及 (v)专用头引导（头引导层）转向，该方法对包含专用注意力头的各层中的音频–静音残差状态差异进行汇总，并按专用头密度加权（$w_\ell = n_\ell/K$）。

模型。 我们研究了两种基于通义千问的大语言音频模型：Qwen2-Audio-7B-Instruct [1] 和 R1-AQA [21], ，后者是基于通义千问骨干网络构建、经强化学习优化的音频问答模型。这两种模型都将音频作为音频条件化词元输入

大语言模型词元序列中，从而通过标准自注意力实现联合音频-文本推理。我们将所有方法用于多项选择任务，并选择标签对应的下一个词元 logit 最高的选项。

评估协议。我们在 4 选 1 的多项选择设置下评估 MMAU。对于每个包含选项（A– D）的问题，我们根据最终提示位置处各选项标签对应的下一词元对数几率进行评分 i_{final}（生成前），并预测得分最高的选项。我们报告准确率（预测与真实标签一致的问题所占比例）。使用 McNemar检验评估配对比较的统计显著性。我们在留出的校准划分 $\mathcal{D}_{\text{cal}}$ 上选择引导强度 β，并报告 MMAU test-mini 数据划分上的最终结果；该数据划分从未用于超参数调优。

专家注意力头选择。我们从校准划分 $\mathcal{D}_{\text{cal}}$ 中提取所有 $L \times H$ 个注意力头的注意力权重（通义千问2-Audio-7B-Instruct 和 R1-AQA：L=32、 H=32，即 1024 个注意力头；两者均基于 Qwen-7B骨干网络）。对于每个注意力头，我们计算 $a_{\ell,\,h}(x)$（公式 1）及其与正确性的相关性 $\rho_{\ell,\,h}$，并依据$|\rho_{\ell,\,h}|$ 选取得分最高的 K 个注意力头（我们使用 K=20）。

双次前向运行与激活缓存。所有引导变体对每个示例均使用两次前向传播：一次音频条件前向，以及一次时长匹配的静音前向。为提高效率，我们缓存 $\mathcal{D}_{\text{cal}}$上所有层在位置 i_{final} 处的残差流状态$\mathbf{h}_\ell(x)$；对于头级基线，我们还在同一位置额外缓存逐头注意力输出$\mathbf{u}_{\ell,h}(x)$（在 $W_O^{(\ell)}$ 之前）。随后，我们将计算限制在由已发现的注意力头所确定的专家层集合$\mathcal{L}$ （第4.2节）。

6. 结果

听觉信号。专家听觉分数 $A_{\text{spec}}(x)$能够预测正确性，并且显著高于匹配的随机注意力头对照组。在预测选项于音频条件化运行与音频消融运行之间发生变化的示例上，该分数也会升高（$p < 0.001$），表明它能够追踪音频何时影响模型的决策。

准确率提升。头引导层操控将MMAU test-mini准确率在通义千问2-Audio上从49.20%→57.25%提升（+8.05个百分点），并在R1-AQA上从64.50%→69.40%提升（+4.90个百分点），优于最佳单层基线（表1）。头级专家基线取得了不可忽视的提升，但仍弱于层引导（表1）。

表 1：*MMAU test-mini*（*1,000* 个示例）上的准确率（%）。头引导层操控优于基线。

方法	Qwen2-Audio	R1-AQA
基线（不进行干预）	49.20	64.50
头级转向（随机注意力头）	49.42	64.02
最佳单层引导	53.82	65.21
头级引导（专家）	54.30	67.80
头引导层操控	**57.25**	**69.40**

领域细分。 各领域的提升趋势一致（表2）。在 Qwen2-Audio 上，语音的增幅最大（+14.1 个百分点），其次是声音（+4.9 个百分点）和音乐（+5.1 个百分点）。在 R1-AQA 上，声音的增幅最大（+7.5 个百分点），语音（+3.3 个百分点）和音乐（+3.9 个百分点）的提升则较小。

选择与敏感性。为隔离专家选择的影响，我们将层引导与匹配的

表2： *MMAU test-mini* 上各领域的准确率（%）；总体列报告增幅，单位为百分点（*pp*）。

	Qwen2-Audio		R1-AQA	
领域	基线	专家	基线	专家
语音	42.04	56.14	57.36	60.68
声音	54.95	59.84	69.37	76.89
音乐	50.98	56.07	66.77	70.67
总体	**49.20**	**57.25 (+8.05)**	**64.50**	**69.40 (+4.90)**

表 3： 使用专家选择的注意力头进行层引导，与匹配的随机注意力头集合相比：相对于基线的提升（百分点）。随机注意力头采用与专家相同的流程来确定层集合 $\mathcal{L}_{rand}$。

	Qwen2-Audio		R1-AQA	
K	随机	专家	随机	专家
10	+3.1%	+7.2%	+1.0%	+4.7%
20	+2.2%	+8.0%	+2.0%	+4.9%
30	+3.1%	+7.4%	+1.3%	+4.6%

对照组进行比较：在其中，我们均匀随机抽取 K 个注意力头，使用与专家相同的流程推导出相应的层集合 $\mathcal{L}_{\text{rand}}$ （以及权重 w_ℓ），并应用完全相同的层引导转向干预。与专家相比，匹配的随机注意力头集合在 K上带来的提升要小得多（表 3），这表明性能增益来自已发现的注意力头，而非对任意层进行引导。图 2 显示出一致的运行区间：性能在适中时达到峰值（ K≈20 通常接近最优），而 β过大时性能会下降，这表明存在过度转向。随着 K 增加，推导出的专家层集合仍保持稀疏（表 4）。

表 4： 作为 K 的函数而诱导出的专家层集合 $\mathcal{L}$ （见第 *4.2* 节）。“新增层”列出增加 K时新纳入 $\mathcal{L}$ 的层； $|\mathcal{L}|$ 为最终集合大小。

	Qwen2-Audio		R1-AQA	
K	新增层	$\lvert\mathcal{L}\rvert$	新增层	$\lvert\mathcal{L}\rvert$
10	19, 21, 22, 23	4	16、17、18、19、22	5
20	14, 18, 20, 24, 26	9	20, 24, 27, 28, 30	10
30	25, 27	11	26	11

7. 讨论

我们的结果凸显了机制可解释性作为理解并改进音频语言模型的实用工具。头级注意力分析提供了实例级的音频参与度指标，并定位出音频相关计算所集中的一组稀疏专用头。利用音频—静音反事实，我们表明，选择性地干预这些层可以放大模型的音频效应，并在无需参数更新的情况下带来稳定的准确率提升（在 MMAU 上最高达 +8 个百分点）。总体而言，这表明大语言音频模型中的文本主导是一种可诊断且可引导的失效模式，而可解释性可为构建更可靠接地的多模态系统提供可操作的定位信号。

8. 参考文献

[1] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin 等，“Qwen2-Audio 技术报告”，arXiv 预印本 arXiv:2407.10759，2024。[2] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma 与 C. Zhang，“SALMONN：面向大语言模型的通用听觉能力”，载于国际学习表征会议（ICLR），2024，另见 arXiv:2310.13289。 [在线]。可获取：https://openreview.net/forum?id=14rn7HpKVk[3] S. Ghosh, Z. Kong, S. Kumar, S. S, J. Kim, W. Ping, R. Valle, D. Manocha 与 B. Catanzaro，“Audio Flamingo 2：具备长音频理解与专家推理能力的音频语言模型”，载于 ICML 会议论文集，2025，第 19 358– 19 405 页。 [在线]。可获取：https://proceedings.mlr.press/v267/ ghosh25b.html[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser 与 I. Polosukhin，“Attention is All You Need”，《神经信息处理系统进展》，第 30 卷，2017。[5] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou 与 J. Zhou，“Qwen-audio：通过统一的大规模音频语言模型推进通用音频理解”，arXiv 预印本 arXiv:2311.07919，2023。[6] H. Wu, M. Tang, X. Zheng 与 H. Jiang，“当语言凌驾一切：揭示多模态大语言模型中的文本主导”，arXiv 预印本 arXiv:2508.10552，2025。[7] L. Aharon, K. Lee, K. Sikka, S. Chettih, C. Hurwitz, L. Paninski 与 M. R. Whiteway，“面向数据高效的多视角动物姿态估计的不确定性感知框架”，arXiv 预印本 arXiv:2510.09903，2025。[8] Y. Zhang, Y. Shi, W. Yu, Q. Wen, X. Wang, W. Yang, Z. Zhang, L. Wang 与 R. Jin，“通过对语言先验施加惩罚来消除多模态大语言模型的偏差”，2024。 [在线]。可获取：https://arxiv.org/abs/2403.05262[9] C. Wang, G. Deng, X. Yang, H. Qiu 与 T. Zhang，“当音频与文本不一致：揭示大型音频语言模型中的文本偏差”，载于《2025 年自然语言处理经验方法会议论文集》，2025，第 4878–4888 页。[10] J. Billa，“当音频大语言模型不倾听：一项关于模态仲裁的跨语言研究”，arXiv 预印本 arXiv:2602.11488，2026。[11] N. Nanda, L. Chan, T. Lieberum, J. Smith 与 J. Steinhardt，“通过机制可解释性衡量顿悟的进展指标”，arXiv 预印本 arXiv:2301.05217，2023。[12] N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly 等，“Transformer 电路的数学框架”，Transformer Circuits Thread，第 1 卷，第 1 期，第 12 页，2021。[13] K. Meng, D. Bau, A. Andonian 与 Y. Belinkov，“定位并编辑 GPT 中的事实关联”，《神经信息处理系统进展》，第 35 卷，第 17 359–17 372 页，2022。[14] N. Glazer, Y. Segal-Feldman, H. Segev, A. Shamsian, A. Buchnick, G. Hetz, E. Fetaya, J. Keshet 与 A. Navon，“超越转写：自动语音识别中的机制可解释性”，arXiv 预印本 arXiv:2508.15882，2025。[15] H. Futami, S. Arora, Y. Kashiwagi, E. Tsunoo 与 S. Watanabe，“在多任务口语理解模型中寻找任务专用子网络”，arXiv 预印本 arXiv:2406.12317，2024。[16] M. Golovanevsky, W. Rudman, V. Palit, R. Singh 与 C. Eickhoff，“视觉语言模型注意到了什么？一种面向无高斯噪声的文本图像扰动与评估的机制可解释性流程”，arXiv 预印本 arXiv:2406.16320，2024。[17] L. Basile, V. Maiorca, D. Doimo, F. Locatello 与 A. Cazzaniga，“注意力头追踪：探究多模态 Transformer 中的注意力专门化”，arXiv 预印本 arXiv:2510.21518，2025。[18] A. M. Turner, L. Thiergart, G. Leech, D. Udell, U. Mini 与 M. MacDiarmid，“激活加法：引导语言模型"

[19] C. Olsson、N. Elhage、N. Nanda、N. Joseph、N. DasSarma、T. Henighan、B. Mann、A. Askell、Y. Bai、A. Chen 等，“上下文学习与归纳头”，*arXiv* 预印本 *arXiv:2209.11895*，2022。[20] E. Voita、D. Talbot、F. Moiseev、R. Sennrich 和 I. Titov，“分析多头自注意力：专用注意力头承担了主要工作，其余部分可以剪枝”，载于计算语言学协会第*57*届年会论文集，2019年，第5797–5808页。[21] G. Li、J. Liu、H. Dinkel、Y. Niu、J. Zhang 和 J. Luan，“强化学习优于监督微调：音频问答案例研究”，*arXiv* 预印本 *arXiv:2503 11197*，2025。.[22] X. Zheng、C. Liao、Y. Fu、K. Lei、Y. Lyu、L. Jiang、B. Ren、J. Chen、J. Wang、C. Li 等，“多模态大语言模型深受模态偏差影响”，*arXiv* 预印本 *arXiv:2505.18657*，2025。[23] S. Leng、H. Zhang、G. Chen、X. Li、S. Lu、C. Miao 和 L. Bing，“通过视觉对比解码缓解大型视觉语言模型中的物体幻觉”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2024年，第13 872–13 882页。[24] A. Favero、L. Zancato、M. Trager、S. Choudhary、P. Perera、A. Achille、A. Swaminathan 和 S. Soatto，“通过视觉信息锚定控制多模态幻觉”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2024年，第14 303–14 312页。[25] C.-Y. Kuan 和 H.-y. Lee，“大型音频语言模型真的能听懂吗？通过多任务评估和逐步音频推理应对幻觉”，载于*ICASSP 2025-2025 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2025年，第1–5页。[26] J. Chen、Z. Guo、J. Chun、P. Wang、A. Perrault 和 M. Elsner，“音频大语言模型真的在聆听，还是只是在转录？测量其对词汇与声学情绪线索的依赖”，*arXiv* 预印本 *arXiv:2510.10444*，2025。[27] K. Wang、A. Variengien、A. Conmy、B. Shlegeris 和 J. Steinhardt，“真实场景中的可解释性：GPT-2 *00593*small中的间接宾语识别电路”，2022*2211*，*arXiv* 预印本 *arXiv:* 。.[28] K. Li、O. Patel、F. Vi´egas、H. Pfister 和 M. Wattenberg，“推理时干预：从语言模型中 eliciting 真实回答”，神经信息处理系统进展，第36卷，第41 451–41 530页，2023。[29] A. M. Turner、L. Thiergart、G. Leech、D. Udell、J. J. Vazquez、U. Mini 和 M. MacDiarmid，“通过激活工程引导语言模型，2024”，网址 *https://arxiv. org/abs/2308.10248*，第2308卷，2024。[30] T.-E. Lin、K.-Y. Lee 和 H.-Y. Lee，“自适应向量引导：一种无需训练、逐层干预的方法，用于缓解大型音频与多模态模型中的幻觉”，*arXiv* 预印本 *arXiv:2510.12851*，2025。[31] N. Rimsky、N. Gabrieli、J. Schulz、M. Tong、E. Hubinger 和 A. Turner，“通过对比式激活添加引导 Llama 2”，载于计算语言学协会第*62*届年会论文集（第一卷：长篇论文），2024年，第15 504–15 522页。[32] A. M. Turner、L. Thiergart、G. Leech、D. Udell、J. J. Vazquez、U. Mini 和 M. MacDiarmid，“通过激活工程引导语言模型”，*arXiv* 预印本 *arXiv:2308.10248*，2023。[33] S. Sakshi、U. Tyagi、S. Kumar、A. Seth、R. Selvakumar、O. Nieto、R. Duraiswami、S. Ghosh 和 D. Manocha，“MMAU：大规模多任务音频理解与推理基准”，*arXiv* 预印本 *arXiv:2410.19168*，2024。

without optimization,” 2024.