arXiv:2601.07331v1 [cs.SD] 12 Jan 2026

SEE：用于量化大型音频语言模型中噪声干扰的信号嵌入能量

Yuanhe Zhang[1], ⋆, **Jiayu Tian**[2], ⋆, **Yibo Zhang**[1], **Shilinlu Yan**[1], **Liang Lin**[3], **Zhenhong Zhou**[4], **Li Sun**[1], **Sen Su**[1], [5], †北京邮电大学 [2]华北电力大学[3]信息工程研究所，中国科学院[4]南洋理工大学 [5]重庆邮电大学 {charmes-zhang, zhangyibo2023,lulu_land,lsun, susen}@bupt.edu.cn; jiayu@ncepu.edu.cn; linliang@iie.ac.cn; zhenhong001@e.ntu.edu.sg

摘要

大型音频语言模型（LALM）已广泛应用于实时场景，例如车载助手和在线会议理解。在实际应用中，音频输入常受到设备噪声和环境噪声的污染，导致性能下降。然而，现有关于 LALM 噪声问题的研究缺乏定量分析，主要依赖直觉和经验观察，因此未能理解其实际鲁棒性。为解决这一问题，我们提出**信**号嵌入**能**量（SEE），这是一种用于量化噪声强度对 LALM 输入影响的方法，可据此区分 LALM 在真实世界部署中的鲁棒性。SEE 基于从模型内部表征中提取的结构化激活子空间提出了一种新视角，相较于原始音频特征，该视角能够更准确地捕捉模型对噪声的感知。在各项实验中，SEE 与 LALM 性能呈现强相关性，相关系数达到 0.98。令人意外的是，传统音频降噪方法对 LALM 仅有微弱效果，在某些情况下甚至会提高 SEE 并损害性能。这表明，以语音为中心的降噪目标与现代 LALM 的噪声敏感性并不匹配。因此，我们提出一种基于 SEE 的缓解策略，用于对 LALM 输入进行降噪，其效果优于现有降噪方法。本文提出了一种用于量化 LALM 噪声的新指标，为真实世界部署中的鲁棒性改进提供指导。我们的代码已公开发布于 https://github.com/jyutian/SEEN。

1 引言

大型音频语言模型（LALM）正越来越多地应用于实时应用场景，

⋆ 表示共同贡献。 † 表示通讯作者。

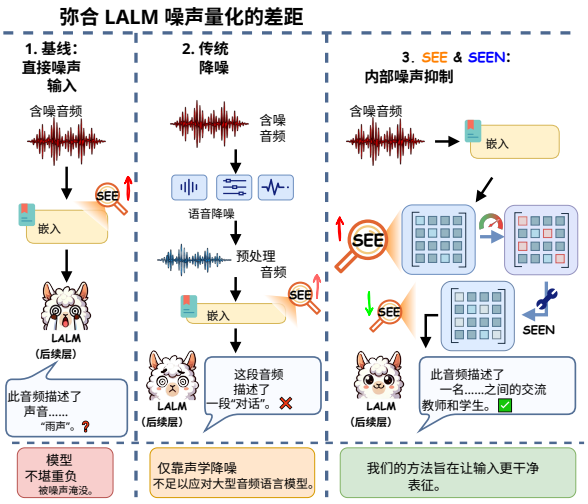

图1：LALM 中表示层面噪声鲁棒性的研究动机与概述。波形层面的去噪可以提升声学质量，但可能引入语义干扰；信号嵌入能量（SEE）可对其进行量化，并通过 SEEN 加以缓解。

例如车载助手，在这些场景中，音频是语言理解与交互的主要界面（Rubenstein 等人，2023；Radford 等人，2023；Li 等人，2025c）。在这些真实环境中，设备采样缺陷和环境噪声会污染输入波形，使原本干净的请求变成带有干扰的信号（Gelbart 和 Morgan，2002年；Deliyski 等人，2005年）。随后，这种受污染的信号被输入 LALM；同样会扭曲波形的噪声，也会降低模型的生成质量（Lin 等人，2025；Kumar 和 Mishra，2025）。虽然量化这种影响对 LALM 来说仍然颇具挑战，但现有研究通常依赖任务性能（Shen 等人，2024）作为替代指标，而这需要大规模评估，也难以用于缓解问题。

为了最大限度地提升基准测试表现，许多 LALM 都是在相对理想化的条件下进行训练和验证的（Chu 等人，2024b年；Zhang 等人，2023年）。在真实世界部署中，

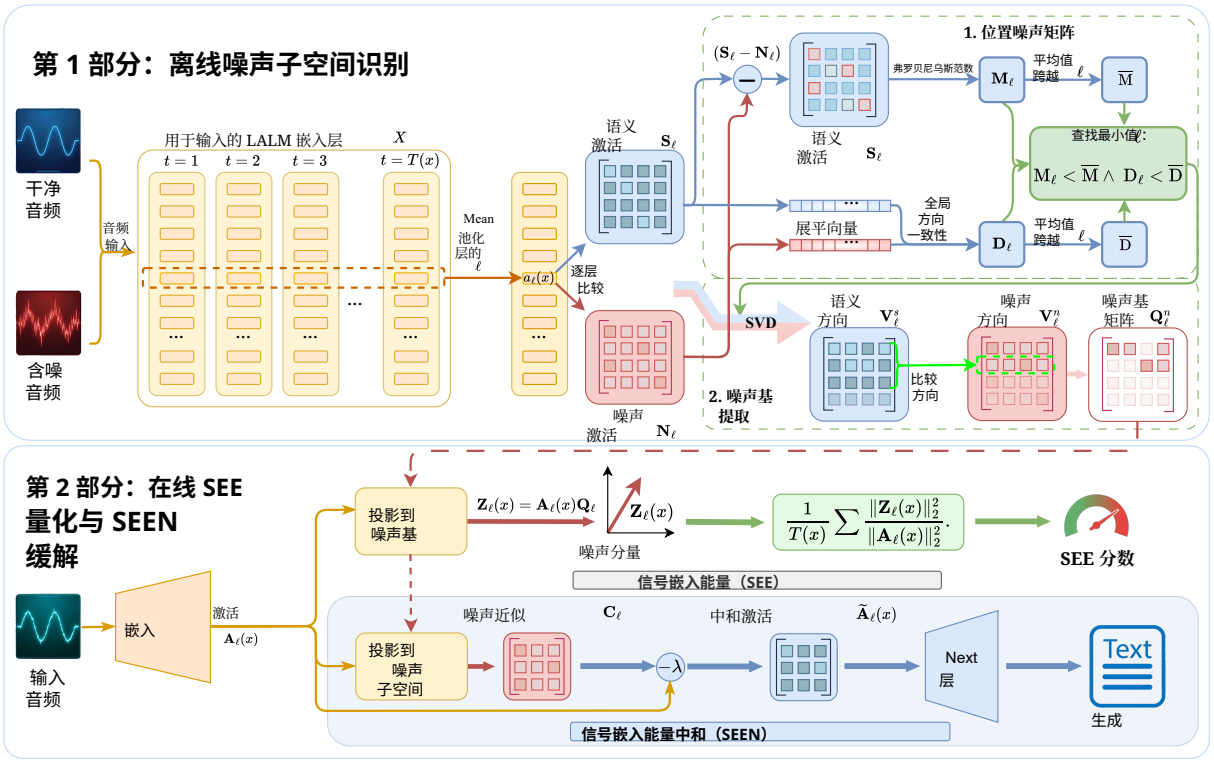

图 2：离线阶段，我们构建噪声激活矩阵，定位噪声主导层，并提取主要噪声方向以形成噪声子空间。在线阶段，SEE 量化投影到该子空间上的激活能量，而 SEEN 则通过移除嵌入层处投影的分量来缓解噪声。

然而，输入音频常常受到设备伪影和环境噪声的干扰，而由此对 LALM 造成的影响仍难以通过有原则的方式加以量化（Li 等人，2025b；Kumar 和 Mishra，2025；Hou 等人，2025）。尽管纳入更多真实世界数据或应用语音增强可能提升鲁棒性，但缺乏定量评估，使得干预效果难以评估，也难以可靠地比较模型（Goel 等人，2021年；Hu 等人，2024年）。现有研究往往依赖结果层面指标，例如模型在各类任务上的表现；这类指标需要大规模基准测试，对于诊断噪声影响或指导缓解措施所能提供的帮助有限（Yang 等人，2024年；Wang 等人，2025）。这些局限促使我们寻求一种模型感知准则，以直接量化大型音频语言模型内部的噪声干扰。

为量化语义干扰，我们提出**信**号**嵌**入**能**量（**SEE**），这是一种用于衡量 LALM 嵌入空间中噪声强度的指标。SEE 利用大型音频语言模型的激活空间，将输入嵌入中的语义成分与噪声成分区分开来。通过分析各层之间的方向和幅度差异，我们确定了噪声扭曲语义处理的主要位置。随后，我们测量噪声激活的累积强度，从而量化生成过程中的语义干扰。由于 SEE 捕捉的是嵌入层面的推理偏差，因此可直接用作衡量噪声对模型生成影响的指标。以 SEE 作为探针，我们发现 SEE 随噪声强度单调递增，并与生成质量高度相关，皮尔逊相关系数（Benesty 等人，2009）达到 0.98。此外，主流大型音频语言模型容易受到真实环境噪声的影响，而标准降噪流程往往无法降低 SEE，甚至可能使其升高，这表明声学降噪并不一定能减轻语义干扰。

在 SEE 的指导下，我们提出了一种无需重新训练即可缓解噪声的防御方法。具体而言，我们引入了**信**号**嵌**入**能**量**中**和（**SEEN**），该方法直接作用于模型的音频嵌入，以最小化噪声成分。由于 SEEN 针对的是嵌入干扰，而非波形能量，因此其缓解目标与 SEE 保持一致。在实验中，SEEN 的准确率比现有去噪方法提高了 6.7%。

我们的贡献总结如下：

• 我们提出了**SEE**，这是一种量化大型音频语言模型中扰动强度的指标，超越了传统声学测量方法。

• 我们提出**SEEN**，一种无需训练的门控机制，可优化音频嵌入并减少噪声条件下的误判。

• 我们开展了广泛的实验，以分析声学缓解的风险并评估本研究的有效性。

2 相关工作

2.1 大型音频语言模型

大型音频语言模型（LALMs）通过将预训练音频编码器与文本骨干网络相结合，将大语言模型（LLMs）的推理能力扩展至听觉领域（Radford 等，2023；Chen 等，2023）。从技术上看，大语言模型处理离散文本标记，而大型音频语言模型通常通过离散标记化（Lakhotia 等，2021；Zhang 等，2023）或跨模态适配器（Tang 等，2023；Chu 等，2023），将高维连续信号与语义空间对齐。这种架构转变使大型音频语言模型能够捕捉文本模型通常无法获取的副语言线索和环境语境，从而将应用场景扩展至复杂的以音频为中心的推理。代表性框架包括 AudioPaLM（Rubenstein 等，2023）、Qwen-Audio（Chu 等，2024a）、MiniCPM-o（Yao 等，2024）和 StepAudio（Wu 等，2025）。不同于级联式流程，大型音频语言模型的端到端特性有助于在各种声学任务中实现统一的指令遵循（Gong 等，2024；Peng 等，2024）。

2.2 噪声鲁棒性

传统语音鲁棒性研究主要通过多条件训练、数据增强（Ko 等，2015年；Park 等，2019年）或增强前端（Wang 和 Chen，2018年）来缓解声学失配（Li 等人，2015年），以降低词错误率（Watanabe 等，2018年）。相较之下，大型音频语言模型（LALM）的鲁棒性研究侧重于语义推理稳定性和指令遵循完整性（Li 等人，2025a；Hou 等人，2025）。鉴于非平稳噪声常会引发“语义幻觉”（Hou 等人，2025），研究重点已从声学保真度转向环境扰动下的逻辑一致性和任务可靠性（Wang 等人，2025；Xiong 等人，2025；Gopal 等人，2025）。

2.3 语音增强与去噪前端

语音增强范式已从统计信号处理转向数据驱动的深度学习。早期经典方法，如维纳滤波（Lim，1978年）、谱减法（Boll，2003年）和 MMSE-STSA（Ephraim 和 Malah，2003年），主要通过谱估计来处理平稳噪声。现代神经网络方法则通过谱映射（Xu 等人，2014年）、Conv-TasNet（Luo 和 Mesgarani，2019年）等端到端时域建模，以及包括生成对抗网络（Pas- cual 等人，2017年）和扩散模型（Lu 等人，2022年）在内的生成式框架，推动了这一领域的发展。然而，传统语音增强优先考虑任务性能所需的声学保真度，往往忽视了 LALM 所需的语义连贯性。因此，去噪伪影可能在改善信号指标的同时，损害多模态对齐和推理能力。

3 方法

本节将正式介绍一个框架，用于在内部嵌入激活层面分析并缓解 LALM 中的噪声干扰。我们首先介绍信号嵌入能量（SEE），这是一种通过语义方向分解量化噪声干扰的指标。随后，我们提出信号嵌入能量中和（SEEN），这是一种无需训练的策略，通过减去噪声成分来提升下游生成质量。

3.1 噪声实质识别

符号说明。 我们将具有 L 个可观测激活函数层的 LALM 记为，其中层的索引为 $\ell \in \{1, \ldots, L\}$。对于音频输入 $x \in X$，块 ℓ 处的激活是一个矩阵 $\mathbf{A}_\ell(x) \in \mathbb{R}^{T(x) \times d_\ell}$，其中 $T(x)$ 表示时间步数， d_ℓ 表示层 ℓ 的隐藏维度。

我们使用两个对齐输入集，分别包含一个语义（干净）请求集 $X^s = \{x_1^s, x_2^s, \cdots, x_m^s\}$和一个纯噪声集 $X^n = \{x_1^n, x_2^n, \cdots, x_m^n\}$，其中 m 表示在下游应用环境中收集的信号帧数量。

我们使用阈值 α 保留占主导地位的奇异方向；如果某个方向与任何占主导地位的语义方向的余弦相似度低于 $\delta = 0.1$，则将其视为“纯噪声”方向；并使用 $\varepsilon > 0$ 作为一个较小的常数，以确保数值稳定性。

分离噪声方向。由于音频输入表示为帧级令牌序列，且序列长度各不相同 $T(x)$。为平衡样本长度差异，我们在层 ℓ沿时间维度进行均值池化：

$$\mathbf{a}_\ell(x) = \frac{1}{T(x)} \sum_{t=1}^{T(x)} \mathbf{A}_\ell(x)_{t,:} \in \mathbb{R}^{d_\ell}, \quad (1)$$

其中$\mathbf{A}_\ell(x)_{t,:}$表示第 t 个帧词元的激活向量。

接下来，我们将 M 个样本的池化向量堆叠起来，以形成联合表示空间：

$$\begin{aligned} \mathbf{S}_\ell &= \left[\mathbf{a}_\ell(x_1^s) \quad \cdots \quad \mathbf{a}_\ell(x_M^s)\right]^\top \in \mathbb{R}^{M \times d_\ell}, \\ \mathbf{N}_\ell &= \left[\mathbf{a}_\ell(x_1^n) \quad \cdots \quad \mathbf{a}_\ell(x_M^n)\right]^\top \in \mathbb{R}^{M \times d_\ell}. \end{aligned} \quad (2)$$

其中，$\mathbf{S}_\ell$ 是语义（干净）激活矩阵，$\mathbf{N}_\ell$ 是第 ℓ 层的噪声激活矩阵。

位置噪声矩阵。为了确定噪声主要影响编码倾向的位置，我们从两个互补的方面刻画干扰：幅度和方向。

为了确定噪声从何处开始改变语义处理，我们使用弗罗贝尼乌斯范数（Böttcherand Wenzel，2008年）作为整体差异能量，并使用差异矩阵进行计算：

$$\mathrm{M}_\ell = \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{d_\ell} (\mathbf{S}_\ell(i,j) - \mathbf{N}_\ell(i,j))^2} \quad , \quad (3)$$

其中，$\mathbf{S}_\ell(i,j)$（分别地，$\mathbf{N}_\ell(i,j)$）表示第 i 个干净（分别为噪声）输入在第 ℓ 层的池化激活向量在特征维度 j 上的值。随后，我们使用 $\mathrm{vec}(\cdot)$ 将矩阵展平为 $\mathbb{R}^{Md_\ell \times 1}$ 中的向量，并计算全局方向一致性：

$$\mathrm{D}_\ell = \frac{\|\mathrm{vec}(\mathbf{S}_\ell)^\top \mathrm{vec}(\mathbf{N}_\ell)\|_2}{\|\mathrm{vec}(\mathbf{S}_\ell)\|_2 \, \|\mathrm{vec}(\mathbf{N}_\ell)\|_2 + \varepsilon}, \quad (4)$$

我们用 $\|\cdot\|_2$ 表示向量的欧几里得（L2）范数。其平均值 $\overline{\mathrm{M}} = \frac{1}{L}\sum_{\ell=1}^{L} \mathrm{M}_\ell$和 $\overline{\mathrm{D}} = \frac{1}{L}\sum_{\ell=1}^{L} \mathrm{D}_\ell$ 可以得到。

利用这两个指标，可以确定噪声监测的主要位置：

$$\ell^\star = \min\{\ell \mid \mathrm{M}_\ell > \overline{\mathrm{M}} \wedge \mathrm{D}_\ell > \overline{\mathrm{D}}\}, \quad (5)$$

$\ell^\star$ 通常出现在模型的后续层中。令 $\mathcal{L}^\star = \{\ell^\star, \ell^\star + 1, \ldots, L\}$ 表示保留基并用于识别的层集合。

噪声基提取。对于每个 $\ell \in \mathcal{L}^\star$，语义矩阵 $\mathbf{S}_\ell$ 和噪声矩阵 $\mathbf{N}_\ell$ 分别通过奇异值分解（SVD）（Stewart，1993年）分解为 $\mathbf{S}_\ell = \mathbf{U}_\ell^s \mathbf{\Sigma}_\ell^s (\mathbf{V}_\ell^s)^\top$ 和$\mathbf{N}_\ell = \mathbf{U}_\ell^n \mathbf{\Sigma}_\ell^n (\mathbf{V}_\ell^n)^\top$。

这里，$\mathbf{U}_\ell^s, \mathbf{U}_\ell^n \in \mathbb{R}^{M \times M}$ 张成样本空间，$\mathbf{V}_\ell^s$, $\mathbf{V}_\ell^n \in \mathbb{R}^{d_\ell \times d_\ell}$张成隐藏空间，而$\mathbf{\Sigma}_\ell^s, \mathbf{\Sigma}_\ell^n \in \mathbb{R}^{M \times d_\ell}$ 按降序存储奇异值。重要的是，由于我们的目标是识别隐藏空间中的方向，因此我们关注 V 中的右奇异向量。每一列$\mathbf{v}_{\ell,j}^s \in \mathbb{R}^{d_\ell}$（分别为 $\mathbf{v}_{\ell,j}^n$）表示一个主方向，沿该方向的激活分别彼此独立地包含语义信息（噪声信息）。

设 $\sigma_{\ell,j}^s$（分别为 $\sigma_{\ell,j}^n$）表示与$\mathbf{v}_{\ell,j}^s$（分别为 $\mathbf{v}_{\ell,j}^n$）相关的第 j 个奇异值。较大的奇异值表示在层 ℓ 中包含更强相似性信息的方向。我们通过阈值 α保留主要的语义 $\mathcal{I}_\ell^s = \{j \mid \sigma_{\ell,j}^s > \alpha\}$ 和噪声信息 $\mathcal{I}_\ell^n = \{j \mid \sigma_{\ell,j}^n > \alpha\}$。在后续步骤中，我们仅使用相应的右奇异向量 $\{\mathbf{v}_{\ell,j}^s\}_{j \in \mathcal{I}_\ell^s}$和 $\{\mathbf{v}_{\ell,j}^n\}_{j \in \mathcal{I}_\ell^n}$来构建 SEE 的纯噪声子空间。

对于每个噪声方向索引 $j \in \mathcal{I}_\ell^n$，我们计算它与每个语义方向 $k \in \mathcal{I}_\ell^s$ 的余弦相似度，并取最大绝对相似度进行汇总：

$$\mathbf{m}_{\ell,j} = \max_{k \in \mathcal{I}_\ell^s} (\cos(\mathbf{v}_{\ell,j}^n, \mathbf{v}_{\ell,k}^s)). \quad (6)$$

如果某个主导噪声方向与所有主导语义方向都近乎正交，则保留该方向：$\mathcal{J}_\ell = \left\{ j \in \mathcal{I}_\ell^n \,\middle|\, \mathbf{m}_{\ell,j} < \delta \right\}$。

定义一个二值掩码向量 $s_\ell \in \{0,1\}^{d_t}$，其中 $s_\ell[j] = 1$ 当且仅当 $j \in \mathcal{J}_\ell$，并定义一个对角掩码矩阵$\mathbf{M}_\ell = \mathrm{diag}(s_\ell) \in \mathbb{R}^{d_\ell \times d_\ell}$。由此得到噪声基矩阵：

$$\mathbf{Q}_\ell = \mathbf{V}_\ell^n \mathbf{M}_\ell \in \mathbb{R}^{d_\ell \times r_\ell}. \quad (7)$$

3.2 SEE 计算

对于每个音频输入 x，我们将词元激活投影到每个 $\ell \in \mathcal{L}^\star$ 对应的噪声方向上：

$$\mathbf{Z}_\ell(x) = \mathbf{A}_\ell(x) \mathbf{Q}_\ell \in \mathbb{R}^{T(x) \times r_\ell}. \quad (8)$$

对于每个帧词元 t，我们将层 SEE 分量定义为其能量：

$$\mathrm{SEE}_\ell(x) = \frac{1}{T(x)} \sum_{t=1}^{T(x)} \frac{\|\mathbf{Z}_\ell(x)_{t,:}\|_2^2}{\|\mathbf{A}_\ell(x)_{t,:}\|_2^2 + \varepsilon}. \quad (9)$$

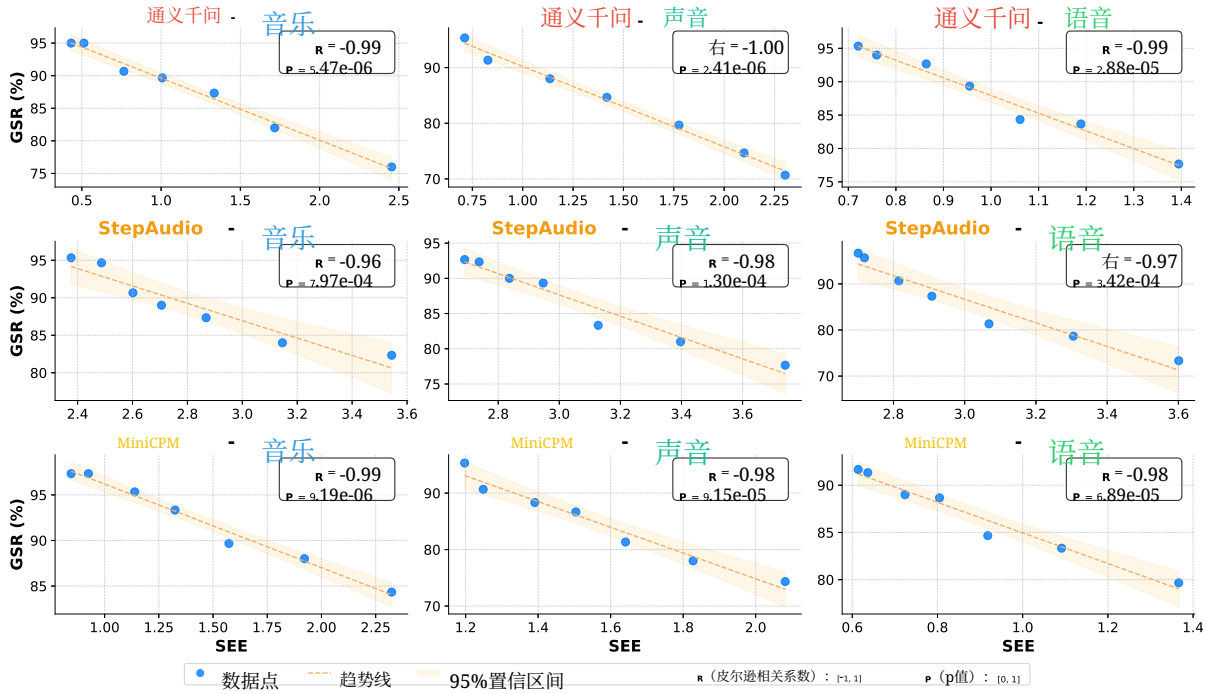

图 3: SEE 与 GSR 呈负相关，并且在不同任务类型中均观察到一致的趋势。

最后，我们对保留层进行聚合，为输入计算一个单一分数：

$$\mathrm{SEE}(x) = \frac{1}{|\mathcal{L}^\star|} \sum_{\ell \in \mathcal{L}^\star} \mathrm{SEE}_\ell(x), \tag{10}$$

SEE(x) 衡量与噪声子空间对齐的嵌入能量，因此可直接量化模型内部的噪声干扰。

3.3 信号嵌入能量中和（SEEN）的构建

SEE 用于量化输入激活噪声方向的程度，而 SEEN 则利用相同的噪声子空间，从中间激活中移除相应分量，旨在减少噪声偏差，同时不改变模型参数，也不需要额外的训练数据。

给定保留的层集合 $\mathcal{L}^\star$ 和噪声矩阵 $\{\mathbf{Q}_\ell\}_{\ell \in \mathcal{L}^\star}$，我们对音频嵌入中的激活进行处理。对于输入 x 和层 $\ell \in \mathcal{L}^\star$，我们首先通过将 $\mathbf{A}_\ell(x) \in \mathbb{R}^{T(x) \times d_\ell}$ 投影到噪声子空间，重构原始隐藏空间中的噪声分量：

$$\mathbf{C}_\ell(x) = \mathbf{A}_\ell(x)\mathbf{Q}_\ell\mathbf{Q}_\ell^\top \in \mathbb{R}^{T(x) \times d_\ell}. \tag{11}$$

SEEN 随后通过减去投影分量来中和激活：

$$\widetilde{\mathbf{A}}_\ell(x) = \mathrm{SEEN}(x) = \mathbf{A}_\ell(x) - \lambda \mathbf{C}_\ell(x), \quad \lambda \in [0, 1],\ \ell \in \mathcal{L}^\star. \tag{12}$$

这里，λ 控制中和强度（默认设置为 $\lambda = 1$）。

随后继续使用 $\widetilde{\mathbf{A}}_\ell(x)$ 进行前向传播。如此一来，SEEN 提供了一种有效且轻量的方法，使噪声缓解与 SEE 保持一致，从而减少内部扰动，同时保留生成所需的其余语义分量。

4 实验

4.1 设置

模型。我们在 3 个模型上进行了实验，包括通义千问（`Qwen-2.5-omni-7b`）（Chu 等人，2024a年）、MiniCPM（`Minicpm-o-2.6`）（Yao 等人，2024年）以及 StepAudio（`Step-Audio-2-mini`）（Wu 等人，2025）。更多实验细节见附录 A

数据集。在实验中，我们主要使用 MMAU（Kumar 等人，2025）和 LibriSpeech（Panayotov 等人，2015年）数据集评估我们的方法。根据底层音频模态，任务分为四类：语音转文本（STT）、语音、声音和音乐。STT 评估词级识别。语音侧重于世界知识问答。声音评估环境声音感知，而音乐则考察多元文化音乐推理。

在噪声设置中，我们采用随机生成的白高斯噪声（高斯）（Ko 等，

噪声	高斯	人群	交通	机器	动物	淋浴	Wind
干净	1.212	1.212	1.212	1.212	1.212	1.212	1.212
10db	1.767	1.709	1.633	2.106	1.462	1.503	1.402
0db	2.052	2.287	2.324	2.808	1.544	1.536	1.551
-10分贝	2.436	3.101	3.481	3.306	1.797	1.586	1.921

表 1：使用混合噪声类型生成的 SEE 的效果。带下划线的项目是已知噪声类型。

数据集	方法	高斯	人群	机器	交通
音乐	干净	0.35	1.32	1.30	1.17
	噪声	2.45	5.97	11.15	5.84
声音	干净	0.67	2.12	2.47	1.85
	噪声	2.72	5.71	10.87	5.30
语音	干净	0.66	3.12	3.54	3.28
	噪声	1.63	4.79	6.74	4.14

表 2：SEE 在不同噪声类型下的区分度。

2015年），以及 PNL 数据集中的噪声类别（Hu 和 Wang，2010年），包括人群噪声（人群）、机械噪声（机器）和车辆噪声（交通）。

基线。 我们将模型在正常（干净）和噪声条件下输出的差异作为指标，而非真实准确率。

对于传统音频降噪，我们选取两种基于频率的方法和两种基于模型的方法作为基线。具体而言，STFT（Ephraim 和 Malah，2003年）和 WT（Donoho，2002年）代表基于频率的方法，而 Segan（Pascual 等人，2017年）和 DFL（Purwins 等人，2019年）则作为基于模型的基线。

指标。 使用信噪比（SNR）对噪声强度进行标准化；对于未标注实验，信噪比设为 -10 dB。我们采用生成成功率（GSR）作为主要指标，用于衡量含噪样本相对于干净输入的准确率。我们通过计算 SEE 指数与 GSR 的皮尔逊相关系数（Benesty 等人，2009）及其相关 p 值，评估 SEE 指数的有效性。

4.2 使用 SEE 量化噪声强度

我们在三个 LALM 框架中评估信号嵌入能量（SEE），以考察它们对噪声的鲁棒性。如图3所示，随着 SEE 增加，GSR 持续下降，且皮尔逊相关系数较高（$R \in [-0.96, -1.00]$，$P \ll 0.001$），这表明 SEE 与损坏条件下的性能下降密切相关。这证实 SEE 是一种以模型为中心的指标，能够捕捉嵌入空间干扰，从而预测生成失败。附录B中提供了更多实证证据。

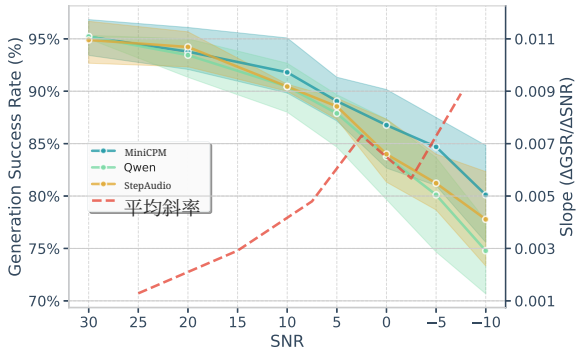

图 4：生成成功率（GSR）随信噪比（SNR）的变化，显示当信噪比低于 $= 10$ 时，生成成功率急剧下降。

上述分析表明，SEE 可作为一种内部指标；我们还评估了其相对于标准物理信噪比尺度的鲁棒性，以验证所选的 GSR 区间。图 4 展示了鲁棒性与信噪比的关系。对于所有大型音频语言模型，随着信噪比降低，GSR 会单调下降，并在低于 ≈ 10dB 时下降得更为显著，这标志着一个信噪比敏感区间。斜率曲线通过显示低信噪比下更大的边际损失，证实了这一转变。这个物理转折点与我们先前分析中 SEE 开始升高时对应的 GSR 范围一致，证实所选区间能够有效捕捉噪声导致的失效。

4.3 SEE 在不同噪声类型下的有效性

我们评估了 SEE 在不同噪声类型下的表现，以验证其泛化能力。通义千问上的结果（表2）显示，干净条件与噪声条件之间存在明显差异。具体而言，在不同请求类型下，干净输入与含噪输入的 SEE 范围互不重叠，且所有噪声类别下的 SEE 均有所升高。

此外，我们评估了 SEE 在不同噪声类别中的适用性，并针对四种具有代表性的噪声类型（高斯、人群、机器和交通）对 SEE 进行了优化。表 1 显示，在相同信噪比下，这些类别的 SEE值保持一致，这表明 SEE 主要受噪声强度驱动，而非特定的声学特征。

数据集	方法	MiniCPM		Qwen		StepAudio	
		噪声	干净	噪声	干净	噪声	干净
音乐	STFT	84.33%	93.33%	71.67%	90.33%	79.00%	92.67%
	WT	82.00%	94.67%	67.67%	91.00%	77.67%	94.00%
	Segan	82.67%	89.67%	71.00%	86.33%	78.00%	87.33%
	DFL	82.00%	88.33%	67.33%	76.67%	75.67%	85.33%
	SEEN	**85.00%**	**99.00%**	**76.99%**	**97.67%**	**82.67%**	**98.00%**
声音	STFT	72.00%	86.67%	64.33%	88.33%	76.00%	87.67%
	WT	70.33%	92.33%	63.00%	91.67%	71.33%	92.33%
	Segan	73.00%	87.00%	58.67%	83.33%	73.67%	87.00%
	DFL	73.67%	81.33%	60.00%	79.33%	72.33%	83.00%
	SEEN	**75.27%**	**98.67%**	**72.00%**	**99.67%**	**78.33%**	**99.00%**
语音	STFT	75.33%	94.00%	68.00%	97.00%	66.33%	93.33%
	WT	70.67%	93.00%	68.00%	97.00%	55.67%	94.67%
	Segan	69.00%	93.00%	64.67%	91.33%	60.67%	90.00%
	DFL	72.67%	93.00%	65.33%	92.00%	62.00%	92.33%
	SEEN	**80.00%**	**98.00%**	**78.33%**	**99.00%**	**74.00%**	**96.33%**

表3：采用不同去噪策略时噪声条件下的生成成功率。

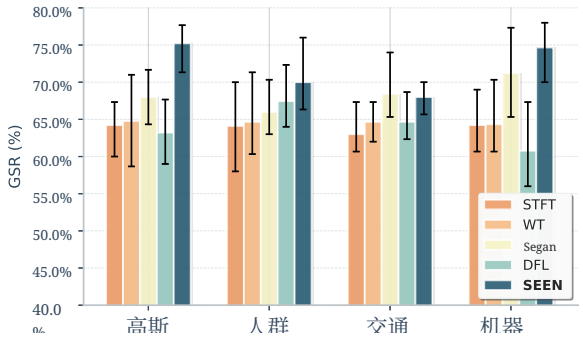

图5：SEEN提升了不同噪声类型下的生成质量，且未出现一致性的性能下降。

强度，而非特定的声学特征。

为进一步考察泛化能力，我们引入了三种未包含在优化过程中的额外噪声类别。表1 显示，与直接优化的噪声类型相比，其SEE值略低，这可能是由于频谱和时间结构存在差异。然而，随着信噪比下降，SEE仍呈单调递增趋势。这种保持不变的单调性证实，即使在未见过的条件下，SEE仍然对噪声强度保持敏感。更详细的结果见附录E

4.4 传统降噪方法中的SEE

我们使用 SEE 分析预处理样本。如图6所示，去噪输入的 SEE 通常高于原始输入。这与波形级证据形成对比：波形级证据表明，这些方法能够降低声学噪声，凸显出去噪目标与 LALM 语义编码之间的不匹配。

模型	数据集	早期	中期	ALL	SEEN
MiniCPM	干净	0.002	0.004	0.004	0.763
	噪声	0.126	0.270	0.456	1.777
Qwen	干净	0.001	0.000	0.015	0.633
	噪声	0.024	0.005	1.751	1.972
StepAudio	干净	0.005	0.022	0.092	2.489
	噪声	0.312	0.661	5.233	3.878

表4：SEE 的层选择消融实验。

定量来看，去噪音频的 SEE 与信噪比额外提高 $= 0$ 的输入相当，这表明去噪伪影的作用类似于模型内部表征中的额外噪声。

此外，去噪会导致波形与干净嵌入之间出现独特的错位。尽管 SEE 能捕捉残余干扰，但这些失真已超出原始噪声的范畴。这表明编码空间中增加了变化性，可能会妨碍鲁棒性导向训练。更多结果见附录 C 和 F

4.5 使用 SEEN 抑制噪声

表3 评估了 SEEN 对生成质量的影响。在干净或低噪声环境下，SEEN 对性能的影响微乎其微，并能保持语义保真度，这证实了抑制与噪声相关的方向不会损害正常请求。在有噪声的情况下，SEEN 在所有数据集和模型上均能稳定提升 GSR，平均表现优于传统和学习型去噪器

模型	MiniCPM	Qwen	StepAudio
None	84.33%	76.00%	82.33%
0.25	84.67%	76.33%	82.67%
0.5	84.33%	76.33%	82.67%
0.75	85.00%	76.67%	82.67%
1	**85.00%**	**77.00%**	**82.67%**
1.2	85.00%	76.67%	82.67%

表5：SEEN 对中和强度 λ 的敏感性。

模型	MiniCPM	Qwen	StepAudio
None	84.33%	76.00%	82.33%
0.1	**85.00%**	**77.00%**	**82.67%**
0.3	84.33%	77.00%	82.33%
0.5	83.33%	77.33%	82.33%
0.7	84.33%	77.33%	82.33%
1	83.67%	77.33%	82.33%

表 6：SEEN 对余弦阈值的敏感性 δ。

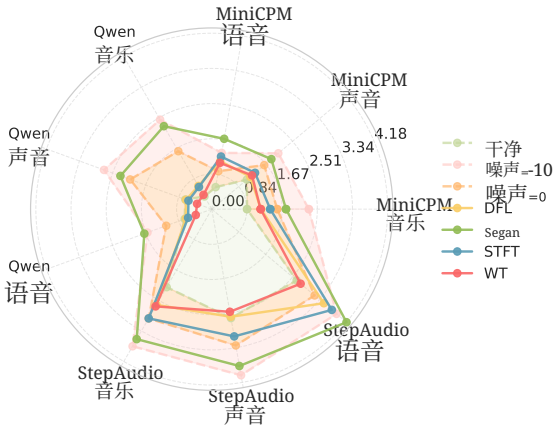

图6：通过 SEE 测量的波形去噪效果。许多传统去噪器无法降低 SEE，甚至可能使其升高。较低的 SEE 表示噪声对齐的干扰更少。

6.7%。这些结果表明，嵌入空间缓解比声学抑制更有效，因为它在针对噪声引发的语义偏差时不会引入新的失真。附录H提供了更多分析。

我们还评估了 SEEN 在三个 MMAU数据集上面对不同噪声类型时的鲁棒性。如表5所示，SEEN 在所有测试条件下均能有效抑制噪声。附录D中报告了 SEEN 方法在不同信噪比设置下的更多结果。

4.6 消融分析

噪声监测的层选择。我们首先对 SEE 中采用的层选择策略进行消融。为评估其必要性，我们构建了三种替代方案：(i) 将该方法应用于前三分之一的层（早期），(ii) 仅应用于中间三分之一的层（中期），以及 (iii) 应用于全部层（全部）。

表 4 显示，将 SEE 限制在早期层或中期层并不能产生显著结果。这是因为在这些层进行筛选时得到的有效方向很少，导致检测结果不稳定。将该方法应用于全部层是可行的。然而，在不同模型中，变化幅度会变得不稳定，并可能受到早期层和中期层干扰的影响。我们的方法与提取出的噪声子空间更加一致，也能更好地保持跨模态嵌入一致性。

对超参数的敏感性。接下来，我们研究 SEEN 对两个关键超参数的敏感性：抑制强度 λ 和余弦相似度阈值 δ。表 5 显示，λ 取值越大，抑制效果越好；增强噪声方向抑制有助于提升生成质量。然而，当抑制比超过 1 时，会引入与噪声方向相反的分量，造成额外的语义扰动。表6 报告了余弦相似度的影响。在大多数情况下，放大 δ 会干扰主要语义方向，最终导致性能下降。

5 结论

我们提出信号嵌入能量（SEE），一种用于量化现实噪声影响的指标。SEE 基于从模型内部表征中导出的结构化激活子空间，相较于干净音频特征，能够更准确地刻画噪声感知。以 SEE 作为探针，我们表明，最先进的 LALM 对噪声高度敏感。更为关键的是，传统语音降噪技术往往无法降低 SEE，这揭示了声学抑制与 LALM 嵌入层级推理之间存在语义错位。受 SEE 启发，我们提出信号嵌入能量中和（SEEN），一种从嵌入激活中移除噪声子空间的免训练方法。SEEN 能够持续降低 SEE，并优于现有降噪方法。总体而言，我们的研究凸显了语义层级标准的必要性，并为未来的大型音频语言模型设计提供了原则性方向。

局限性

我们的方法假设能够获取来自目标部署环境的对齐干净请求和纯噪声录音，以估计稳定的噪声子空间。在实践中，这类“仅噪声”数据集无法在训练期间直接获取；必须从实际应用环境中收集噪声。噪声类别相对单一或较少的下游应用环境更易于收集，也适合进行 SEE 测量。变化高度多样的环境可能会增加获取纯噪声录音的难度。

在方法上，我们当前的 SEE 通过均值池化压缩可变长度序列，这可能会低估受到时间性损坏的线索和副语言线索。虽然该方法在我们测试的场景中效果良好，但过度激进的中和可能会在某些声学条件下去除与任务相关的信息。更好的做法不是消除噪声语义成分，而是增强模型对不完整任务语义的理解。

一个有前景的方向是将 SEE 用作训练鲁棒性信号，而不仅仅是缓解工具。这可以通过在噪声增强下将 SEE 纳入正则化项，明确惩罚嵌入中的噪声对齐能量来实现；也可以利用 SEE 选择或学习嵌入空间增强模块，在不单纯依赖声学保真度的情况下减少语义干扰。我们将这一方向留待未来工作探索。

参考文献

Jacob Benesty、Jingdong Chen、Yiteng Huang 和 Israel Cohen。2009。皮尔逊相关系数。载于语音处理中的噪声抑制，第 1–4 页。施普林格。

Steven Boll. 2003年. 使用谱减法抑制语音中的声学噪声。《IEEE Transactions on Acoustics, Speech, and Signal Processing》，27(2):113–120。

Albrecht Böttcher 和 David Wenzel. 2008年. Frobenius 范数与交换子。《线性代数及其应用》，429(8-9):1864–1885。

Sanyuan Chen、Yu Wu、Chengyi Wang、Shujie Liu、Daniel Tompkins、Zhuo Chen、Wanxiang Che、Xi- angzhan Yu 和 Furu Wei。2023年。BEATs：使用声学分词器进行音频预训练。发表于国际机器学习大会，第 5178–5193 页。PMLR。

Yunfei Chu、Jin Xu、Qian Yang、Haojie Wei、Xipin Wei、Zhifang Guo、Yichong Leng、Yuanjun Lv、Jinzheng He、Junyang Lin、Chang Zhou 和 Jingren Zhou。2024a年。Qwen2-Audio 技术报告。*arXiv* 预印本 *arXiv:2407.10759*。

Yunfei Chu、Jin Xu、Qian Yang、Haojie Wei、Xipin Wei、Zhifang Guo、Yichong Leng、Yuanjun Lv、Jinzheng He、Junyang Lin 等人. 2024b年. Qwen2-Audio 技术报告。*arXiv* 预印本 *arXiv:2407.10759*。

YunfeiChu、JinXu、Xiaohuan Zhou、Qian Yang、Shil- iang Zhang、Zhijie Yan、Chang Zhou 和 Jingren Zhou. 2023年. 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*。

Dimitar D Deliyski、Heather S Shaw 和 Maegan K Evans。2005年。环境噪声对声学嗓音质量测量的不利影响。*Journal ofVoice*，19(1):15–28。David L Donoho。2002年。通过软阈值去噪。*IEEE* 信息论汇刊，41(3):613–627。Yariv Ephraim 和 David Malah。2003年。使用最小均方误差短时谱幅度估计器进行语音增强。*IEEE* 声学、语音与信号处理汇刊，32(6):1109–1121。David Gelbart 和 Nelson Morgan。2002年。双重难题：处理远场自动语音识别中的噪声和混响。发表于*Interspeech*，第2185–2188页。Karan Goel、Nazneen Rajani、Jesse Vig、Zachary Taschdjian、Mohit Bansal 和 Christopher Ré。2021年。Robustness Gym：统一自然语言处理评估领域。*NAACL-HLT 2021*，第42页。Yuan Gong、Hongyin Luo、Alexander H Liu、Leonid Karlinsky 和 Ja mes Glass。2024年。Listen, Think, and Understand。发表于国际学习表征会议。Shreyas Gopal、Ashutosh Anshul、Haoyang Li、Yue Heng Yeo、Hexin Liu 和 Eng Siong Chng。2025。面向噪声鲁棒自动语音识别的离散语音表征可解释解耦。*arXiv* 预印本 *arXiv:2510.25150*。Guanyu Hou、Jiaming He、Yinhang Zhou、Ji Guo、Yi-tong Qiao、Rui Zhang 和 Wenbo Jiang。2025。评估大型音频语言模型对音频注入的鲁棒性：一项实证研究。*arXiv* 预印本 *arXiv:2505.19598*。Guoning Hu 和 DeLiang Wang。2010年。一种用于基频估计和有声语音分离的串联算法。*IEEE* 音频、语音与语言处理汇刊，18(8):2067–2079。

Yuchen Hu、Chen Chen、Chao-han Huck Yang、Ruizhe Li、Chao Zhang、Pin-Yu Chen 和 Ensiong Chng。2024年。大语言模型是噪声鲁棒语音识别的高效学习器。发表于国际学习表征会议。

Tom Ko、Vijayaditya Peddinti、Daniel Povey 和 San- jeev Khudanpur。2015年。用于语音识别的音频增强。载于*Proc.Interspeech2015*，第 3586–3589 页。

Pankaj Kumar 和 Subhankar Mishra。2025。大语言模型的鲁棒性：缓解策略与评估指标综述。*arXiv preprintarXiv:2505.18658*。

Sonal Kumar、Šimon Sedlá˘cek、Vaibhavi Lokegaonkar、Fernando López、Wenyi Yu、Nishit Anand、Hyeong- gon Ryu、Lichang Chen、Maxim Pli˘cka、Miroslav Hlavá˘cek 等。2025。MMAU-Pro：用于全面评估音频通用智能的一项具有挑战性的综合基准测试。*arXiv* 预印本 *arXiv:2508.13992*。Kushal Lakhotia、Eugene Kharitonov、Wei-Ning Hsu、Yossi Adi、Adam Polyak、Benjamin Bolte、Tu-Anh Nguyen、Jade Copet、Alexei Baevski、Abdelrahman Mohamed 等。2021。基于原始音频的生成式口语语言建模。计算语言学协会会刊，9:1336–1354。Bohan Li、Wenbin Huang、Yuhang Qiu、Yiwei Guo、Han- kun Wang、Zhihan Li、Jing Peng、Ziyang Ma、Xie Chen 和 Kai Yu。2025a。ISA-Bench：大型音频语言模型指令敏感性的基准测试。*arXiv* 预印本 *arXiv:2510.23558*。Chen-An Li、Tzu-Han Lin 和 Hung-yi Lee。2025b。静音何时重要：无关音频对大型音频语言模型中文本推理的影响。*arXiv* 预印本 *arXiv:2510.00626*。Jing Li、Jingyuan Li、Guo Yang、Lie Yang、Haozhuang Chi 和 Lichao Yang。2025c。大型语言模型和多模态大型模型在自动驾驶中的应用：综述。《*Drones*》，9(4):238。Jinyu Li、Li Deng、Reinhold Haeb-Umbach 和 Yifan Gong。2015。鲁棒自动语音识别：通往实际应用的桥梁。学术出版社。Jae Lim。1978年。退化语音的全极点建模。*IEEE Transactions onAcoustics,Speech,andSignalProcessing*，26(3):197–210。Liang Lin、Miao Yu、Kaiwen Luo、Yibo Zhang、Lilan Peng、Dexian Wang、Xuehai Tang、Yuanhe Zhang、Xikang Yang、Zhenhong Zhou 等。2025。藏于噪声之中：通过潜在声学模式触发器揭示音频大语言模型对齐中的后门。*arXiv* 预印本 *arXiv:2508.02175*。

Yen-Ju Lu、Zhong-Qiu Wang、Shinji Watanabe、Alexan- der Richard、Cheng Yu 和 Yu Tsao。2022年。用于语音增强的条件扩散概率模型。收录于 *ICASSP 2022-2022 IEEE* 声学、语音与信号处理国际会议（*ICASSP*），第7402–7406页。IEEE。

Yi Luo 和 Nima Mesgarani。2019年。Conv-tasnet：超越用于语音分离的理想时频幅度掩蔽。*IEEE/ACM* 音频、语音与语言处理汇刊，27(8):1256–1266。

Vassil Panayotov、Guoguo Chen、Daniel Povey 和 Sanjeev Khudanpur。2015年。LibriSpeech：基于公版有声读物的自动语音识别语料库。载于*Acous-tics,Speech andSignal Processing (ICASSP), 2015IEEE International Conference on*，第 5206–5210 页。IEEE。

Daniel S Park、William Chan、Yu Zhang、Chung-Cheng Chiu、Barret Zoph、Ekin D Cubuk 和 Quoc V Le。2019年。SpecAugment：一种用于自动语音识别的简单数据增强方法。载于 *INTERSPEECH*，第 2613–2617 页。Santiago Pascual、Antonio Bonafonte 和 Joan Serrà。2017年。Segan：语音增强生成对抗网络。载于 *2017*年 *INTERSPEECH* 会议论文集，第 3642–3646 页。Jing Peng、Yucheng Wang、Yangui Fang、Yu Xi、Xu Li、Xizhuo Zhang 和 Kai Yu。2024年。语音大语言模型综述。*arXiv* 预印本 *arXiv:2410.18908*。Hendrik Purwins、Bo Li、Tuomas Virtanen、Jan Schlüter、Shuo-Yiin Chang 和 Tara Sainath。2019年。音频信号处理中的深度学习。*IEEE* 信号处理精选主题期刊，13(2):206–219。Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey 和 Ilya Sutskever。2023年。通过大规模弱监督实现鲁棒语音识别。载于国际机器学习大会，第 28492–28518 页。PMLR。Paul K Rubenstein、Chulayuth Asawaroengchai、Duc Dung Nguyen、Ankur Bapna、Zalán Borsos、Félix de Chaumont Quitry、Peter Chen、Dalia El Badawy、Wei Han、Eugene Kharitonov 及其他 1 人。2023年。AudioPaLM：一种能够说话和聆听的大语言模型。*arXiv* 预印本 *arXiv:2306.12925*。Yongliang Shen、Kaitao Song、Xu Tan、Wenqi Zhang、Kan Ren、Siyu Yuan、Weiming Lu、Dongsheng Li 和 Yueting Zhuang。2024年。TaskBench：用于任务自动化的大语言模型基准测试。神经信息处理系统进展，37:4540–4574。Gilbert W Stewart。1993年。奇异值分解的早期历史。*SIAM* 综述，35(4):551–566。

Changli Tang、Wenyi Yu、Guangzhi Sun、Xianzhao Chen、Tian Tan、Wei Li、Lu Lu、Zejun Ma 和 Chao Zhang。2023年。Salmonn：迈向大语言模型的通用听觉能力。*arXiv* 预印本 *arXiv:2310.13289*。

Bin Wang、Xunlong Zou、Geyu Lin、Shuo Sun、Zhuo- han Liu、Wenyu Zhang、Zhengyuan Liu、AiTi Aw 和 Nancy Chen。2025。Audiobench：音频大语言模型的通用基准测试。载于*2025* 年美国语言学会计算语言学分会会议论文集：人类语言技术（第 *1* 卷：长篇论文），第 4297–4316 页。

DeLiang Wang 和 Jitong Chen。2018。基于深度学习的监督式语音分离：综述。*IEEE/ACM* 音频、语音与语言处理汇刊，26(10)：1702–1726。

Shinji Watanabe、Takaaki Hori、Shigeki Karita、Tomoki Hayashi、Jiro Nishitoba、Yuya Unno、Nelson En- rique Yalta Soplin、Jahn Heymann、Matthew Wiesner、Nanxin Chen 等。2018。Espnet：端到端语音处理工具包。载于国际语音通信协会年会论文集，*INTERSPEECH*，第 2018 卷，第 2207–2211 页。

Boyong Wu、Chao Yan、Chen Hu、Cheng Yi、Chengli Feng、Fei Tian、Feiyu Shen、Gang Yu、Haoyang Zhang、Jingbei Li、Mingrui Chen、Peng Liu、Wang You、Xiangyu Tony Zhang、Xingyuan Li、Xuerui Yang、YayueDeng、Yechang Huang、Yuxin Li 等 90 人。2025。Step-audio 2 技术报告。预印本，arXiv:2507.16632。

Zhen Xiong、Yujun Cai、Zhecheng Li、J unsong Yuan 和 Yiwei Wang。2025。用声音思考：音频思维链使大型音频语言模型能够进行多模态推理。*arXiv* 预印本 *arXiv:2509.21749*。

Yong Xu、Jun Du、Li-Rong Dai 和 Chin-Hui Lee。2014。一种基于深度神经网络的语音增强回归方法。*IEEE/ACM* 音频、语音与语言处理汇刊，23(1):7–19。

Qian Yang、Jin Xu、Wenrui Liu、Yunfei Chu、Ziyue Jiang、Xiaohuan Zhou、Yichong Leng、Yuanjun Lv、Zhou Zhao、Chang Zhou及另外1位作者。2024年。Air-bench：通过生成式理解对大型音频语言模型进行基准测试。载于计算语言学协会第*62*届年会论文集（第*1*卷：长篇论文），第1979–1998页。

Yuan Yao、Tianyu Yu、Ao Zhang、ChongyiWang、Junbo Cui、Hongji Zhu、Tianchi Cai、Haoyu Li、Weilin Zhao、Zhihui He及另外1位作者。2024年。MiniCPM-V：手机上的GPT-4V级多模态大语言模型。*arXiv*预印本 *arXiv:2408.01800*。

Dong Zhang、Shimin Li、Xin Zhang、Jun Zhan、PengyuWang、Yaqian Zhou和Xipeng Qiu。2023年。SpeechGPT：赋能大语言模型内在的跨模态对话能力。载于计算语言学协会论文集：*EMNLP 2023*年，第15757–15773页。

A 更多实验设置

本节全面介绍了本研究中使用的实验配置，涵盖分类和转录任务。

A.1 模型架构与音频处理

所有实验均使用以 bfloat16 精度加载的全模态大型语言模型（LLM）。我们采用 sdpa（缩放点积注意力）实现，以提高计算效率。对于声学输入，所有原始音频信号均使用 librosa 库重采样至 16,000 Hz，并转换为单声道。所有实验均在配备 NVIDIA RTX 5090 GPU 的服务器上进行。

A.2 任务定义与提示

我们在两种不同类型的任务上评估我们的方法：

1.**多项选择题问答（MCQA）**：应用于音乐、声音和语音数据集。模型会收到提示，要求回答多项选择题，并受到严格限制：“回答多项选择题时，只输出一个字母。”我们使用正则表达式模式r’([A-Za-z])’解析最终答案。

2. **语音转文本（STT）**：应用于 *Lib-riSpeech* 数据集。模型使用以下提示进行转录：“请仔细聆听音频片段并转录其内容。”

A.3 噪声合成与数据集规模

为模拟现实世界中的干扰，我们在七个信噪比水平下将各种类型的环境噪声注入干净音频样本中：$\{-10, -5, 0, 5, 10, 20, 30\}$。样本量设置为 MCQA 任务的 $N = 300$ 和 STT 任务的 $N = 100$，以确保结果具有统计显著性。

A.4 SEE 干预与超参数

SEE 通过在目标层上注册前向钩子来实现。干预所使用的目标层集合通过 算法 1 中所述的诊断程序确定。

•**目标层**：该抑制应用于模型的编码器层。对于通义千问，目标层从第 23 层开始一直到最终层（model.thinker.audio_tower.layers[23:]）。对于 MiniCPM，目标层从第 18 层一直到最终层（model.apm.layers[18:]）。对于 StepAudio，目标层从第 27 层一直到最终层（model.llm.encoder.blocks[27:]）

•**抑制因子**：干预强度 $\alpha = 1.0$。余弦相似度阈 $\delta = 0.1$。用于确定保留奇异值数量的能量阈值比 $\rho = 0.95$。

• **噪声子空间（V_{noise}）**：噪声奇异向量是在离线校准阶段推导得到的，使用从 Speech 数据集中采样的 50 个具有代表性的纯噪声片段，并将强度标准化为相当于 0 dB 信噪比的参考水平。

A.5 关于基线方法

为确保比较公平且严谨，本研究中包含的所有传统基线方法均严格按照其原始论文中规定的参数设置和实验流程实现与配置。除遵循各自作者提供的建议外，未对这些基线进行额外的超参数调优，以维护比较分析的完整性。

B SEE 的补充分析

SEE 在干净音频上保持稳定。 我们在三个 LALM 框架上评估了信号嵌入能量（SEE）。图 7 显示，对于干净音频，SEE 在每个模型内部均保持稳定，在不同类别之间也仅表现出轻微变化，这表明当输入干净时，该指标在很大程度上不受内容差异影响。干净音频的 SEE 嵌入集中在一个相对紧凑的区域内，在噪声对齐子空间上的激活也十分有限。

SEE 跟踪噪声导致的性能下降。随着信噪比 (SNR) 降低，注入噪声变得更强，SEE 也单调上升，反映出向噪声方向的投影逐渐增大。图 4总结了对生成的相应影响：当 SNR 低于 10 dB 时，任务性能下降超过 10%；在严重损坏情况下，SNR $= -10$ 准确率最高下降 25%。这种性能下降与 SEE 的变化密切一致；在相同的 SNR

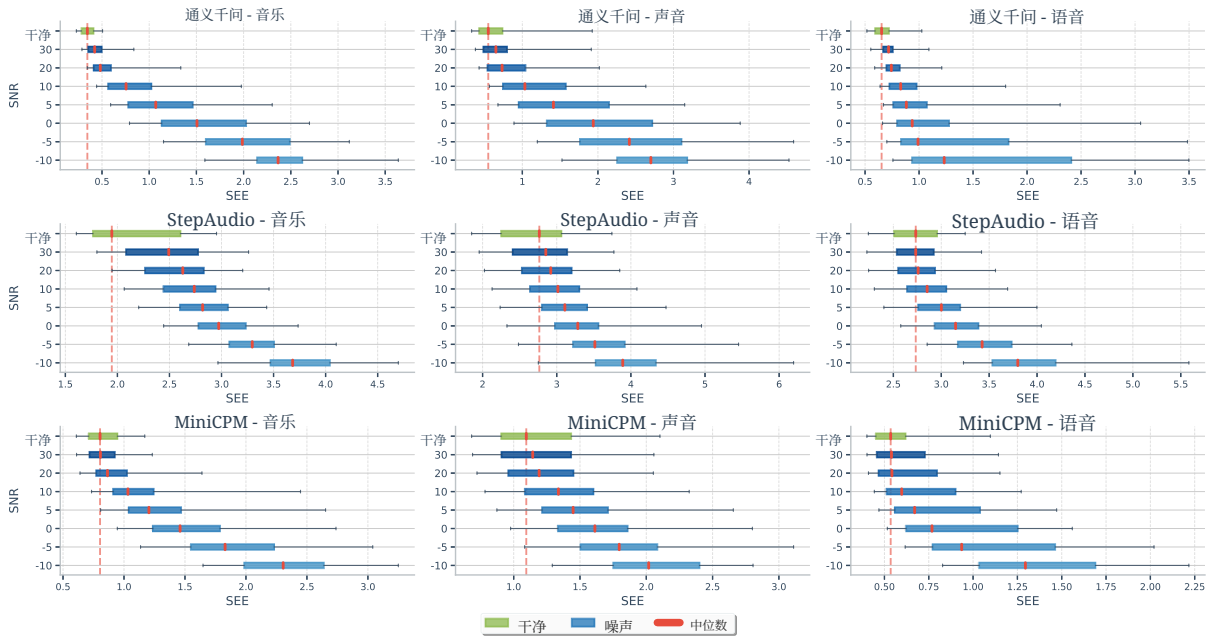

图 7：三种大型音频语言模型在不同噪声强度下的 SEE。干净输入产生的 SEE 稳定，而 SEE 会随着信噪比降低而单调增加。箱线图展示了四分位距和中位数。

方法	MiniCPM				Qwen				StepAudio			
	STT	音乐	声音	语音	STT	音乐	声音	语音	STT	音乐	声音	语音
BASE	0.481	0.834	1.176	0.561	0.854	0.354	0.669	0.657	2.409	2.137	2.676	2.735
噪声	1.378	2.306	2.064	1.357	1.082	2.455	2.725	1.625	3.872	3.763	4.007	3.871
STFT	1.416	1.399	1.347	1.318	0.607	0.604	0.599	0.600	4.442	3.049	3.025	3.658
WT	1.188	1.213	1.255	1.153	0.442	0.391	0.369	0.404	3.001	2.667	2.476	2.756
Segan	1.906	1.816	1.910	1.682	2.051	2.330	2.270	1.675	5.078	3.829	3.698	4.014
DFL	1.237	1.194	1.206	1.171	0.604	0.614	0.647	0.683	3.628	2.703	2.714	3.406
SEEN	**0.008**	**0.017**	**0.016**	**0.009**	**0.007**	**0.015**	**0.017**	**0.010**	**0.040**	**0.044**	**0.047**	**0.039**

表7：SE E 在不同设置下的得分。SEEN 始终会降低 SEE，而大多数波形级去噪 rs
会使 SE E 在不同程度上升高（已加下划线）。

范围内，SEE 急剧上升，并与干净音频基线明显区分开来。总体而言，这些结果表明，SEE 能可靠地跟踪噪声强度，并捕捉声学损坏加剧如何转化为干扰。

C 通过 SEE 评估 SEEN

我们将 SEEN 用作嵌入空间去噪模块，并在表7中报告其对 SEE 的影响。与波形级去噪器不同，SEEN 能显著降低 SEE。由于 SEEN 和 SEE 具有相似的计算结构，SEEN 主要抑制 SEE 可直接捕获的成分，这限制了直接评估的可解释性。

D SEEN 在不同信噪比下的性能总结

与原始噪声损坏相比，SEEN 在多个数据集上提升了任务性能，但提升幅度有限，如表 8和表 9所示。对于这一现象，我们有以下推测。SEEN 通过抑制噪声对齐子空间中的激活来发挥作用，因此只能从表征中去除与噪声相关的成分。然而，当噪声损坏输入本身的语义内容时，由此产生的任务信息会变得不完整或模糊，仅靠表征级抑制无法恢复这些信息。因此，如果模型对不完整或退化的输入缺乏鲁棒性，单独应用 SEEN 不太可能带来显著的性

数据集	方法	MiniCPM		Qwen		StepAudio	
		噪声	干净	噪声	干净	噪声	干净
音乐	None	84.33%	100.00%	76.00%	100.00%	82.33%	100.00%
	STFT	84.33%	93.33%	71.67%	90.33%	79.00%	92.67%
	WT	82.00%	94.67%	67.67%	91.00%	77.67%	94.00%
	Segan	82.67%	89.67%	71.00%	86.33%	78.00%	87.33%
	DFL	82.00%	88.33%	67.33%	76.67%	75.67%	85.33%
	SEEN	**85.00%**	**99.00%**	**76.99%**	**97.67%**	**82.67%**	**98.00%**
声音	None	74.33%	100.00%	70.67%	100.00%	77.67%	100.00%
	STFT	72.00%	86.67%	64.33%	88.33%	76.00%	87.67%
	WT	70.33%	92.33%	63.00%	91.67%	71.33%	92.33%
	Segan	73.00%	87.00%	58.67%	83.33%	73.67%	87.00%
	DFL	73.67%	81.33%	60.00%	79.33%	72.33%	83.00%
	SEEN	**75.27%**	**98.67%**	**72.00%**	**99.67%**	**78.33%**	**99.00%**
语音	None	79.67%	100.00%	77.67%	100.00%	73.33%	100.00%
	STFT	75.33%	94.00%	68.00%	97.00%	66.33%	93.33%
	WT	70.67%	93.00%	68.00%	97.00%	55.67%	94.67%
	Segan	69.00%	93.00%	64.67%	91.33%	60.67%	90.00%
	DFL	72.67%	93.00%	65.33%	92.00%	62.00%	92.33%
	SEEN	**80.00%**	**98.00%**	**78.33%**	**99.00%**	**74.00%**	**96.33%**

表8：Ge 在噪声（信噪比=-10）和干净条件下，采用不同降噪策略的生成成功率 s.

方法	高斯	人群	交通	机器
None	74.78%	69.67%	67.89%	73.45%
STFT	68.00%	66.00%	68.44%	71.22%
WT	63.22%	67.44%	64.67%	60.78%
Segan	64.78%	64.66%	64.67%	64.33%
DFL	64.22%	64.11%	63.00%	64.22%
SEEN	**75.22%**	**70.00%**	**68.00%**	**74.67%**

表9：SEEN 在各种噪声类型下提升了生成质量。

能提升。尽管存在这一限制，SEEN 仍能持续带来小幅提升，这表明即使信息损失占主导地位，减少噪声诱导偏差仍能带来可衡量的益处。

方法	STFT	WT	Segan	DFL
时间（秒）	23.11	2.21	4.67	1156.15
复杂度	$O(N \log N)$	$O(N)$	$O(N \cdot L \cdot C)$	$O(N \cdot L \cdot K^2)$

表 10：效率分析：100 个音频样本的执行时间和时间复杂度

E SEE 指标的可迁移性

为评估所提出的 SEE 指标的可迁移性和普适性，我们在三种不同的音频大语言模型上进行了广泛的跨模型评估：MiniCPM、通义千问和 StepAudio。实验方案包括使用每个噪声类别仅含 50 个样本的有限校准集提取方向向量。随后，将这些向量应用于一个更大的、此前未见过的测试集，其中每种噪声类型包含 300 个样本。

实验结果汇总于表11，一致表明，在全部三种模型和各种声学环境（高斯、人群、交通和机器）中，SEE 指标与信噪比（SNR）呈现稳定的单调关系。具体而言，当信噪比从 10dB 降至 -10dB 时，SEE 值逐步升高。这一趋势在不同模型架构和噪声条件下始终一致，凸显了 SEE 指标在表征音频退化方面具有很强的可迁移性和可靠性。

F 不同增强方法间的表征相似度比较

表格12–18 展示了在不同噪声级别和语音增强方法下，干净语音与其带噪语音或增强语音之间的余弦相似度。具体而言，干净_与_带噪用于衡量干净语音与带噪语音之间的相似度，而干净_与_增强 用于衡量干净语音与经不同语音增强方法处理后的语音之间的相似度。

数据集	信噪比	MiniCPM				Qwen				StepAudio			
		高斯	人群交通		机器	高斯	人群交通		机器	高斯	人群交通		机器
音乐	干净	3.136	3.136	3.136	3.136	0.891	0.891	0.891	0.891	5.430	5.430	5.430	5.430
	10	3.016	3.290	3.140	3.323	1.578	1.498	1.414	1.868	7.623	6.780	6.884	7.132
	0	3.228	3.579	3.369	3.440	1.895	2.217	2.283	2.643	8.163	8.458	8.310	8.370
	-10	3.567	3.569	3.562	3.455	2.453	3.159	3.616	3.27	8.612	9.654	9.291	9.770
声音	干净	3.145	3.145	3.145	3.145	1.533	1.533	1.533	1.533	7.884	7.884	7.884	7.884
	10	3.060	3.161	3.056	3.235	1.956	1.919	1.853	2.344	8.416	8.337	8.426	8.363
	0	3.223	3.377	3.279	3.308	2.209	2.358	2.365	2.972	8.449	8.603	8.516	9.025
	-10	3.549	3.545	3.634	3.350	2.420	3.044	3.346	3.342	8.168	8.967	8.351	9.906
语音	干净	2.461	2.461	2.461	2.461	1.839	1.839	1.839	1.839	7.322	7.322	7.322	7.322
	10	2.764	2.747	2.641	2.871	2.007	2.026	1.945	2.028	7.749	8.014	7.859	7.834
	0	3.085	3.067	2.985	3.089	1.798	2.034	1.904	2.093	8.453	9.004	8.309	8.194
	-10	3.703	3.706	3.862	3.391	1.768	2.353	2.681	2.354	8.583	9.703	8.890	8.559

表 11：SEE 指标在不同噪声条件下的跨噪声迁移能力

数据集	方法	MiniCPM		Qwen		StepAudio	
		干净_vs_噪声	干净_与_增强后	干净_与_含噪	干净_与_增强后	干净_与_含噪	干净_与_增强后
音乐	STFT	0.7138	0.6121	0.7549	0.7048	0.7121	0.6897
	WT	0.7138	0.6108	0.7549	0.6335	0.7121	0.6543
	Segan	0.7138	0.6273	0.7549	0.6304	0.7121	0.6481
	DFL	0.7138	0.6775	0.7549	0.6944	0.7121	0.6736
	SEEN	**0.7138**	**0.7341**	**0.7549**	**0.7628**	**0.7121**	**0.7142**
声音	STFT	0.7446	0.6339	0.7754	0.6938	0.8627	0.7924
	WT	0.7446	0.6423	0.7754	0.6410	0.8627	0.7263
	Segan	0.7446	0.6880	0.7754	0.7098	0.8627	0.8015
	DFL	0.7446	0.7284	0.7754	0.6880	0.8627	0.7804
	SEEN	**0.7446**	**0.7858**	**0.7754**	**0.7827**	**0.8627**	**0.8631**
语音	STFT	0.5837	0.4708	0.7271	0.6880	0.8399	0.8119
	WT	0.5837	0.3683	0.7271	0.5345	0.8399	0.6943
	Segan	0.5837	0.3997	0.7271	0.6303	0.8399	0.7831
	DFL	0.5837	0.5439	0.7271	0.6588	0.8399	0.7986
	SEEN	**0.5837**	**0.6579**	**0.7271**	**0.7345**	**0.8399**	**0.8476**

表 12：不同增强方法和噪声条件下干净语音与噪声或增强语音之间的余弦相似度（信噪比=-5）

如结果所示，传统语音增强方法往往无法缩小干净语音与带噪语音之间的表征差距。在许多情况下，这些方法甚至会扩大这一差距，导致余弦相似度低于原始带噪输入。相比之下，我们提出的方法始终能够缩小干净语音与带噪语音之间的相似度差距，这表明该方法在噪声扰动下保留语义和声学一致性的能力更强。这项辅助实验进一步证明了我们提出的方法优于传统方法。

G 关于基线方法的效率分析

为评估计算效率，我们使用 Speech 数据集中的 100 个音频样本进行了基准测试，信噪比为=0 dB。如表10所示，外部降噪方法不可避免地会带来额外的处理时间：WT 和短时傅里叶变换方法分别需要 2.21 秒和 23.11 秒，而 DFL 模型则需要 1156.15 秒。由于 DFL 实现支持的 CUDA 版本有限，我们在 CPU 上进行了评估。对音频大语言模型应用传统降噪或外部深度学习降噪通常会造成这种“瓶颈”延迟，而我们的方法不会增加任何处理时间，能够更顺畅地集成到实时应用中。

推理阶段所评估方法的时间复杂度也汇总于表10中。此处，N 表示信号中的音频样本总数。对于深度学习模型（Segan 和 DFL），L 表示隐藏层的数量，C 表示特征通道的数量，K 表示卷积核的大小。

数据集	方法	MiniCPM		Qwen		StepAudio	
		干净_对比_噪声	干净_对比_增强	干净_对比_噪声	干净_对比_增强	干净_对比_噪声	干净_对比_增强
音乐	**STFT**	0.6523	0.5965	0.6701	0.6551	0.7031	0.6912
	WT	0.6523	0.5752	0.6701	0.6390	0.7031	0.6566
	Segan	0.6523	0.6078	0.6701	0.6012	0.7031	0.6459
	DFL	0.6523	0.6525	0.6701	0.6648	0.7031	0.6656
	SEEN	**0.6523**	**0.6509**	**0.6701**	**0.6823**	**0.7031**	**0.7002**
声音	**STFT**	0.7129	0.6188	0.7034	0.6460	0.8291	0.7641
	WT	0.7129	0.6031	0.7034	0.6297	0.8291	0.6921
	Segan	0.7129	0.6693	0.7034	0.6557	0.8291	0.7713
	DFL	0.7129	0.6951	0.7034	0.6512	0.8291	0.7510
	SEEN	**0.7129**	**0.7124**	**0.7034**	**0.7125**	**0.8291**	**0.8114**
语音	**STFT**	0.5314	0.4141	0.6452	0.6058	0.8060	0.7692
	WT	0.5314	0.3365	0.6452	0.4967	0.8069	0.6932
	Segan	0.5314	0.3760	0.6452	0.5359	0.8060	0.7410
	DFL	0.5314	0.4517	0.6452	0.5867	0.8060	0.7560
	SEEN	**0.5314**	**0.5292**	**0.6452**	**0.6474**	**0.8069**	**0.7868**

表 13：不同增强方法和噪声条件（信噪比=-10）下干净语音与噪声或增强语音之间的余弦相似度

数据集	方法	MiniCPM		Qwen		StepAudio	
		干净_对比_含噪	干净_对比_增强	干净_对比_含噪	干净_与_增强后	干净_与_含噪	干净_与_增强后
音乐	**STFT**	0.7967	0.6714	0.8280	0.7631	0.7435	0.6994
	WT	0.7967	0.6394	0.8280	0.6539	0.7435	0.6553
	Segan	0.7967	0.6861	0.8280	0.6895	0.7435	0.6629
	DFL	0.7967	0.7084	0.8280	0.7275	0.7435	0.6847
	SEEN	**0.7967**	**0.8099**	**0.8280**	**0.8312**	**0.7435**	**0.7371**
声音	**STFT**	0.8129	0.6849	0.8418	0.7420	0.8928	0.8170
	WT	0.8129	0.6762	0.8418	0.6675	0.8928	0.7460
	Segan	0.8129	0.7488	0.8418	0.7669	0.8928	0.8267
	DFL	0.8129	0.7609	0.8418	0.7235	0.8928	0.8068
	SEEN	**0.8129**	**0.8486**	**0.8180**	**0.8462**	**0.8928**	**0.8718**
语音	**STFT**	0.7042	0.6155	0.8097	0.7703	0.8495	0.8446
	WT	0.7042	0.4365	0.8097	0.6083	0.8495	0.7077
	Segan	0.7042	0.5320	0.8097	0.7206	0.8495	0.8186
	DFL	0.7042	0.6295	0.8097	0.7253	0.8495	0.8348
	SEEN	**0.7042**	**0.7257**	**0.8097**	**0.8082**	**0.8495**	**0.8486**

表 14：不同增强方法和噪声条件下干净语音与噪声或增强语音之间的余弦相似度（信噪比=0）

H SEEN 有效性声明

在我们的实验中，随着损坏程度加剧，SEE 始终升高，且与干净输入范围保持明显区分，这表明它捕获了嵌入空间中稳定的噪声对齐成分。然而，应用 SEEN 抑制这一成分后，生成质量仅有小幅但稳定的提升，并未带来显著的鲁棒性恢复。可测量性与可恢复性之间的这一差距表明，SEE 所识别的干扰是导致下游故障的机制中的核心成分之一。

一种合理的解释是，SEE 隔离出的子空间代表噪声会稳定扰动模型内部编码倾向的方向，因此移除这些方向可以减少嵌入偏差并缓解虚假激活。

然而，生成错误并非完全由有偏表示造成；噪声较强时，也可能在源头抹去或扭曲任务相关声学线索，造成无法通过减去投影项来恢复的真实信息损失。在这种情况下，SEEN 可以消除干扰，却无法重建缺失内容，因此可实现的改进幅度取决于可恢复信息的多少，而非干净与含噪性能之间的差距。这种影响在音频输入中可能尤为明显。帧级分词通常会产生较长的令牌序列，且每个令牌的信息密度相对较低；这意味着损坏可能在许多令牌中分散造成细微失真，并逐渐累积为语义退化，难以通过一次子空间抑制来逆转。

这些观察指明了一个自然的改进方向

数据集	方法	MiniCPM		Qwen		StepAudio	
		干净_与_噪声	干净_与_增强	干净_与_噪声	干净_与_增强	干净_与_噪声	干净_与_增强
音乐	**STFT**	0.8566	0.7610	0.8750	0.8083	0.7702	0.7220
	WT	0.8566	0.6815	0.8750	0.6955	0.7702	0.6724
	Segan	0.8566	0.7659	0.8750	0.7599	0.7702	0.6887
	DFL	0.8566	0.7326	0.8750	0.7562	0.7702	0.6880
	SEEN	**0.8566**	**0.8636**	**0.8750**	**0.8749**	**0.7702**	**0.7822**
声音	**STFT**	0.8681	0.7491	0.8909	0.7817	0.9184	0.8390
	WT	0.8681	0.7268	0.8909	0.7104	0.9184	0.7899
	Segan	0.8681	0.8075	0.8909	0.8134	0.9184	0.8455
	DFL	0.8681	0.7864	0.8909	0.7556	0.9184	0.8290
	SEEN	**0.8681**	**0.8948**	**0.8909**	**0.8923**	**0.9184**	**0.9067**
语音	**STFT**	0.7990	0.7449	0.8683	0.8297	0.9002	0.8774
	WT	0.7990	0.5480	0.8683	0.6902	0.9002	0.7549
	Segan	0.7990	0.6618	0.8683	0.8002	0.9002	0.8392
	DFL	0.7990	0.7105	0.8683	0.7909	0.9002	0.8617
	SEEN	**0.7990**	**0.8213**	**0.8683**	**0.8757**	**0.9002**	**0.8944**

表 15：不同增强方法和噪声条件（信噪比=5）下干净语音与噪声或增强语音之间的余弦相似度

数据集	方法	MiniCPM		Qwen		StepAudio	
		干净_与_噪声	干净_与_增强	干净_与_噪声	干净_对比_增强	干净_对比_噪声	干净_对比_增强
音乐	**STFT**	0.8979	0.8319	0.9066	0.8404	0.7984	0.7468
	WT	0.8979	0.7395	0.9066	0.7493	0.7984	0.6942
	Segan	0.8979	0.8209	0.9066	0.8200	0.7984	0.7153
	DFL	0.8979	0.7522	0.9066	0.7782	0.7984	0.6917
	SEEN	**0.8979**	**0.9035**	**0.9066**	**0.9048**	**0.7984**	**0.7891**
声音	**STFT**	0.9076	0.7966	0.9243	0.8095	0.9393	0.8495
	WT	0.9076	0.7864	0.9243	0.7646	0.9393	0.8384
	Segan	0.9076	0.8489	0.9243	0.8442	0.9393	0.8534
	DFL	0.9076	0.8032	0.9243	0.7779	0.9393	0.8445
	SEEN	**0.9076**	**0.9274**	**0.9243**	**0.9234**	**0.9303**	**0.9172**
语音	**STFT**	0.8635	0.8304	0.9018	0.8615	0.9221	0.9010
	WT	0.8635	0.6704	0.9018	0.7695	0.9221	0.8111
	Segan	0.8635	0.7550	0.9018	0.8514	0.9221	0.8613
	DFL	0.8635	0.7745	0.9018	0.8360	0.9221	0.8801
	SEEN	**0.8635**	**0.8893**	**0.9018**	**0.8983**	**0.9221**	**0.9050**

表 16：不同增强方法和噪声条件下干净语音与噪声或增强语音之间的余弦相似度（信噪比=10）

。尽管 SEE 提供了一项有用的标准，可用于诊断和比较噪声引起的干扰，但要实现更强的鲁棒性，可能需要采用训练阶段机制，明确增强模型在声学证据部分缺失或退化时推断语义的能力。例如，可以考虑将 SEE 指导的目标纳入鲁棒性导向训练，或采用强调在帧级损坏情况下保留语义的增强课程。我们认为这些是值得探索的假设，而非已获证实的解决方案，并将系统探索训练侧增强作为一项重要的后续工作。

I 算法流程

在本附录中，我们介绍所提出的 SEEN 框架的正式流程。该过程分为三个主要部分：算法 1：层定位与噪声提取（算法 1）、算法 2：基于能量的评估（算法 2）以及算法 3：最终的稳健推理（算法 3）。这些算法共同构成一个完整流程，用于诊断并消除大型音频语言模型中的噪声对齐激活。

J AI写作辅助披露

我们仅使用人工智能工具润色语言，以提升表达的清晰度和可读性。人工智能工具未参与本研究的科学内容、构思、分析或结论。

数据集	方法	MiniCPM		Qwen		StepAudio	
		干净_与_噪声	干净_与_增强	干净_与_噪声	干净_与_增强	干净_与_噪声	干净_与_增强
音乐	**STFT**	0.9530	0.8985	0.9444	0.8886	0.8480	0.7906
	WT	0.9530	0.8340	0.9444	0.8427	0.8480	0.7551
	Segan	0.9530	0.8854	0.9444	0.8749	0.8480	0.7570
	DFL	0.9530	0.7716	0.9444	0.7988	0.8480	0.7039
	SEEN	**0.9530**	**0.9492**	**0.9444**	**0.9413**	**0.8480**	**0.8361**
声音	**STFT**	0.9687	0.8667	0.9621	0.8483	0.9661	0.8804
	WT	0.9687	0.8819	0.9621	0.8596	0.9661	0.9046
	Segan	0.9687	0.8978	0.9621	0.8719	0.9661	0.8618
	DFL	0.9687	0.8215	0.9621	0.8004	0.9661	0.8588
	SEEN	**0.9687**	**0.9650**	**0.9621**	**0.9590**	**0.9661**	**0.9436**
语音	**STFT**	0.9404	0.9186	0.9345	0.9016	0.9590	0.9292
	WT	0.9404	0.8445	0.9345	0.8928	0.9590	0.8782
	Segan	0.9404	0.8698	0.9345	0.8863	0.9590	0.8784
	DFL	0.9404	0.8391	0.9345	0.8699	0.9590	0.9025
	SEEN	**0.9404**	**0.9379**	**0.9343**	**0.9307**	**0.9590**	**0.9497**

表 17：不同增强方法和噪声条件（信噪比=20）下干净语音与噪声或增强语音之间的余弦相似度

数据集	方法	MiniCPM		Qwen		StepAudio	
		干净_与_噪声	干净_与_增强	干净_与_噪声	干净_对比_增强	干净_对比_嘈杂	干净_对比_增强
音乐	**STFT**	0.9783	0.9137	0.9671	0.9092	0.8967	0.8196
	WT	0.9783	0.8818	0.9671	0.8892	0.8967	0.7968
	Segan	0.9783	0.9070	0.9671	0.8938	0.8967	0.7832
	DFL	0.9783	0.7840	0.9671	0.8121	0.8967	0.7109
	SEEN	**0.9783**	**0.9743**	**0.9671**	**0.9637**	**0.8967**	**0.8822**
声音	**STFT**	0.9873	0.8835	0.9811	0.8672	0.9809	0.8960
	WT	0.9873	0.9204	0.9811	0.9027	0.9809	0.8912
	Segan	0.9873	0.9092	0.9811	0.8827	0.9809	0.8637
	DFL	0.9873	0.8320	0.9811	0.8123	0.9809	0.8670
	SEEN	**0.9873**	**0.9833**	**0.9811**	**0.9772**	**0.9809**	**0.9582**
语音	**STFT**	0.9605	0.9371	0.9527	**0.9253**	0.9816	0.9418
	WT	0.9605	0.9186	0.9527	0.9403	0.9816	0.9038
	Segan	0.9605	0.8880	0.9527	0.8946	0.9816	0.8833
	DFL	0.9605	0.8726	0.9527	0.8878	0.9816	0.9173
	SEEN	**0.9605**	**0.9580**	**0.9527**	**0.9588**	**0.9816**	**0.9612**

表 18：不同增强方法和噪声条件下干净语音与噪声或增强语音之间的余弦相似度（信噪比=30）

算法 1 SEE 设置：层定位与噪声提取

要求： 语义集 $X^s = \{x_i^s\}_{i=1}^m$、噪声集 $X^n = \{x_i^n\}_{i=1}^m$、阈值 α，δ、稳定项 ε**确保：** 保留层 $\mathcal{L}^\star$、噪声基 $\{\mathbf{Q}_\ell\}_{\ell\in\mathcal{L}^\star}$

1: 对于 $\ell = 1$**到** L**，执行 2: 对于** $i = 1$**到** m**，执行 3: 提取** $\mathbf{A}_\ell(x_i^s) \in \mathbb{R}^{T(x_i^s)\times d_\ell}$**和** $\mathbf{A}_\ell(x_i^n) \in \mathbb{R}^{T(x_i^n)\times d_\ell}$**4:** $\mathbf{a}_\ell(x_i^s) \leftarrow \frac{1}{T(x_i^s)}\sum_{t=1}^{T(x_i^s)} \mathbf{A}_\ell(x_i^s)_{t,:} \in \mathbb{R}^{d_\ell}$**5:** $\mathbf{a}_\ell(x_i^n) \leftarrow \frac{1}{T(x_i^n)}\sum_{t=1}^{T(x_i^n)} \mathbf{A}_\ell(x_i^n)_{t,:} \in \mathbb{R}^{d_\ell}$**6: 结束循环 7: 堆叠** $\{\mathbf{a}_\ell(x_i^s)\}_{i=1}^m$ **以形成** $\mathbf{S}_\ell \in \mathbb{R}^{m\times d_\ell}$**；堆叠** $\{\overline{\mathbf{a}_\ell(x_i^n)}\}_{i=1}^m$ **以形成** $\mathbf{N}_\ell \in \mathbb{R}^{m\times d_\ell}$**8: 幅度：** $\mathbf{M}_\ell \leftarrow \|\mathbf{S}_\ell - \mathbf{N}_\ell\|_F$**9: 方向：** $\mathbf{D}_\ell \leftarrow \frac{|\mathrm{vec}(\mathbf{S}_\ell)^\top \mathrm{vec}(\mathbf{N}_\ell)|}{\|\mathrm{vec}(\mathbf{S}_\ell)\|_2 \|\mathrm{vec}(\mathbf{N}_\ell)\|_2 + \varepsilon}$**10: 结束循环 11:** $\mathbf{M} \leftarrow \frac{1}{L}\sum_{\ell=1}^L \mathbf{M}_\ell$**,** $\mathbf{D} \leftarrow \frac{1}{L}\sum_{\ell=1}^L \mathbf{D}_\ell$**12:** $\ell^\star \leftarrow \mathbf{min}\{\ell \mid \mathbf{M}_\ell > \mathbf{M} \wedge \mathbf{D}_\ell > \mathbf{D}\}$**13:** $\mathcal{L}^\star \leftarrow \{\ell^\star, \ell^\star + 1, \ldots, L\}$**14: 对每个** $\ell \in \mathcal{L}^\star$ **执行 15: 奇异值分解：** $\mathbf{S}_\ell = \mathbf{U}_\ell^s \mathbf{\Sigma}_\ell^s (\mathbf{V}_\ell^s)^\top$**；** $\mathbf{N}_\ell = \mathbf{U}_\ell^n \mathbf{\Sigma}_\ell^n (\mathbf{V}_\ell^n)^\top$**16:** $\mathcal{I}_\ell^s \leftarrow \{j \mid \sigma_{\ell,j}^s > \alpha\}$**；** $\mathcal{I}_\ell^n \leftarrow \{j \mid \sigma_{\ell,j}^n > \alpha\}$**17: 对每个** $j \in \mathcal{I}_\ell^n$**执行 18:** $\mathbf{m}_{\ell,j} \leftarrow \mathbf{max}_{k\in\mathcal{I}_\ell^s} \left|\mathbf{cos}(\mathbf{v}_{\ell,j}^n, \mathbf{v}_{\ell,k}^s)\right|$**19: 结束循环 20:** $\mathcal{J}_\ell \leftarrow \{j \in \mathcal{I}_\ell^n \mid \mathbf{m}_{\ell,j} < \delta\}$**；** $r_\ell \leftarrow |\mathcal{J}_\ell|$**21: 使用** $s_\ell[j] = 1$ **构建掩码** $s_\ell \in \{0,1\}^{d_\ell}$**，当且仅当** $j \in \mathcal{J}_\ell$**时；** $\mathbf{M}_\ell \leftarrow \mathbf{diag}(s_\ell)$ **22:** $\mathbf{Q}_\ell \leftarrow \mathbf{NonZeroCols}(\mathbf{V}_\ell^n \mathbf{M}_\ell) \in \mathbb{R}^{d_\ell \times r_\ell}$ **{equivalently keep columns** $\mathcal{J}_\ell$ **of** $\mathbf{V}_\ell^n$**} 23: 结束循环**

算法 2 SEE 评估：计算输入的信号嵌入能量

要求： 测试输入 x、保留层 $\mathcal{L}^\star$、噪声基 $\{\mathbf{Q}_\ell\}_{\ell\in\mathcal{L}^\star}$、稳定项 ε**确保：** SEE(x)

1：前向传播以获取所有 $\ell \in \mathcal{L}^\star$ 对应的 $\mathrm{A}_\ell(x) \in \mathbb{R}^{T(x)\times d_\ell}$ 2：对每个 $\ell \in \mathcal{L}^\star$ 执行 3：投影：$\mathrm{Z}_\ell(x) \leftarrow \mathrm{A}_\ell(x)\mathrm{Q}_\ell \in \mathbb{R}^{T(x)\times r_\ell}$4：$\mathrm{SEE}_\ell(x) \leftarrow \frac{1}{T(x)}\sum_{t=1}^{T(x)} \mathrm{Z}_{\ell,x,t,:}^2$，$\mathrm{A}_{\ell,x,t,:}$，$\overline{\|\cdot\|_2^2+\varepsilon}$5：结束循环

6：$\mathrm{SEE}(x) \leftarrow \frac{1}{|\mathcal{L}^\star|}\sum_{\ell\in\mathcal{L}^\star} \mathrm{SEE}_\ell(x)$

算法 3SEEN 推理：中和噪声对齐激活

要求：测试输入 x、保留层 $\mathcal{L}^\star$、$\{\ell\}_{\ell\in\mathcal{L}^\star}$、$\{\lambda\}_{\ell\in\mathcal{L}^\star}$、噪声基 Q、中和强度 确保：中和激活 $\{\mathrm{A}_\ell(x)\}$1：前向传播以获取所有 $\ell \in \mathcal{L}^\star$ 对应的 $\mathrm{A}_\ell(x) \in \mathbb{R}^{T(x)\times d_\ell}$ 2：对每个 $\ell \in \mathcal{L}^\star$ 执行 3：重建噪声分量：$\mathrm{C}_\ell(x) \leftarrow \mathrm{A}_\ell(x)\mathrm{Q}_\ell\mathrm{Q}_\ell^\top \in \mathbb{R}^{T(x)\times d_\ell}$4：中和：$\mathrm{A}_\ell(x) \leftarrow \mathrm{A}_\ell(x) - \lambda\mathrm{C}_\ell(x)$ 5：结束循环6：使用 $\mathrm{A}_\ell(x)$ 继续前向传播，以获得最终生成输出
