arXiv:2512.18099v1 [eess.AS] 19 Dec 2025

SAM Audio：音频中的分割一切

Bowen Shi*，**Andros Tjandra***，**John Hoffman***，**Helin Wang***，**Yi-Chiao Wu***，**Luya Gao***，**Julius Richter**†，**Matt Le**†，**Apoorv Vyas**†，**Sanyuan Chen**†，**Christoph Feichtenhofer**‡，**Piotr Dollár**‡，**Wei-Ning Hsu**‡，**Ann Lee**‡

Meta 超级智能实验室

*核心贡献者（从第二作者起随机排序），†贡献者（随机排序），‡项目负责人（随机排序）

通用音频源分离是多模态人工智能系统感知声音并对其进行推理的一项关键能力。尽管近年来取得了长足进展，现有分离模型仍要么面向特定领域，专为语音或音乐等固定类别设计；要么可控性有限，只支持文本等单一提示模态。在这项工作中，我们推出 SAM Audio，这是一种通用音频分离基础模型，可在单一框架中整合文本、视觉和时间跨度提示。SAM Audio 基于扩散Transformer架构构建，采用流匹配方法在涵盖语音、音乐和一般声音的大规模音频数据上训练，并能灵活分离由语言、视觉掩码或时间跨度描述的目标声源。该模型在一系列多样化基准测试中达到最先进性能，涵盖真实场景和专业制作音频中的通用声音、语音、音乐及乐器分离，显著优于以往的通用型和专用系统。此外，我们还推出了一项全新的真实世界分离基准测试，其中包含人工标注的多模态提示，并提出一种无参考评估模型，该模型的评估结果与人类判断高度相关。

日期：2025年12月23日
联系邮箱：Bowen Shi bshi@meta.com，Andros Tjandraandrostj@meta.com
演示：https://aidemos.meta.com/segment-anything/editor/segment-audio**代码：**https://github.com/facebookresearch/sam-audio**网站：**https://ai.meta.com/samaudio/

Meta

1 引言

音频源分离旨在将复杂的声音混合物分解为与不同声音事件相对应的独立源音轨。分离系统在广泛的现实应用中发挥着重要作用。在制作和创意领域，它们能够帮助声音工程师分离并重新混音各个音轨、修复档案录音，或去除不需要的背景噪声。在教育和无障碍领域，分离技术可以为学习者或助听设备突出关键声音或语音。从研究角度来看，它也是研究多模态人工智能系统中音频理解的宝贵试验平台，因为有效的分离需要识别各个声源，而这正是人工智能模型理解音频的关键要素。

音频分离已得到广泛研究，现有方法大致可分为无提示方法和有提示方法。无提示系统旨在将混合音频分解为一组固定的预定义声源，并在语音增强、说话人分离和音乐分轨等专门任务中展现出强劲性能（Mitsufuji 等人，2022年；Fabbro 等人，2024年；Zhao 等，2024；Kong 等，2023年）。然而，这些方法假设输出配置固定，并依赖预定义的声音类别分类体系。因此，当面对开放域混合音频或用户定义的声音类型时，它们难以进行适应，因为声音类别之间的边界模糊且高度依赖上下文。近期进展推动该领域转向有提示分离，即通过外部信号指定目标声源。文本提示（Liu 等，2024b；Yuan 等，2025年；Ma 等，2024；Wang 等，2025a；Hai 等，2024）允许用户描述任意声音事件（例如“狗叫声”“女性语音”），从而不再需要固定的分类体系。视觉提示（Zhao 等人，2018年；Huang 等人，2024年；Li 等，2024b；Ephrat 等人，2018年；Dong 等人，2023年）通过提供实例级定位来补充文本，使多个相似声源

出现在同一场景中时能够进行消歧。

有提示分离范式极大拓展了分离的适用范围，但仍面临若干重要挑战。首先，当前基于文本提示的系统主要在通用音效上进行基准测试，在音乐或语音等专业领域往往表现不佳。例如，基于文本的乐器分离明显落后于 Demucs 等领域专用方法（Rouard 等，2023年；Défossez，2021年）。尽管这些专用模型在各自领域内表现出色，却无法泛化到固定源类型之外。其次，视觉提示系统相对缺乏研究，通常只在小型或合成数据集上进行测试。真实世界视频包含屏幕内外的多种声源，以及模糊的视听对应关系，因此，仅选择一个视觉区域是否足以实现可靠分离尚不明确。最后，现有模型难以区分细微但在感知上十分重要的相似声音差异，例如电影音效的不同变体，而仅靠文本很难准确描述这些差异。这些挑战表明，我们需要一种更加灵活、性能普遍出色的提示框架，能够整合各种模态的线索来引导分离。

与此同时，音频分离研究的进展也受到了缺乏统一基准测试 和可靠评估指标的拖累。现有的文本提示模型通常在彼此不同的测试集上进行评估，且往往仅关注环境声音分离，对语音和音乐的覆盖有限。相比之下，MUSDB（Rafii 等人，2017年）等传统乐器分离基准测试仅涵盖少数预定义音轨，且无法与提示方式兼容。此外，广泛使用的客观指标（如信号失真比（SDR））依赖于使用干净参考音轨的合成混合音频，而在真实录音中，这类数据十分稀缺，甚至完全不存在。无参考指标（如 CLAP 相似度（Wu 等人，2023年））虽能在一定程度上反映音频与文本的对齐情况，但与人类对分离质量的判断相关性较差（见第7.8节）。这些局限使得模型难以进行公平比较，也难以评估其在开放域场景中的真实性能。

在本研究中，我们推出音频万物分割模型（SAM Audio），这是一种通用音频分离基础模型，可在统一框架中整合文本、视觉和时间提示。SAM Audio 允许用户通过文本描述指定要分离的内容，通过视觉掩码（使用正向/负向点击）指定分离的位置，并通过时间跨度提示指定分离的时间。这些模态既可单独使用，也可联合使用，从而实现灵活交互——例如，用户可以在音乐视频中描述“钢琴演奏”，并高亮相应区域，以便精准地将其分离出来。SAM Audio 的核心是一个采用流匹配训练的扩散 Transformer，训练数据为大规模音频混合，涵盖语音、音乐和通用声音事件。推理时，模型会同时生成目标音轨和残余音轨，后者包含其余所有音频内容。

贡献。 我们的主要贡献有三点。

- 我们提出了 SAM Audio，这是首个支持多模态提示（文本、视觉、时间跨度）的基础模型，可单独或组合使用，用于开放领域音频分离，并在通用和专业领域（如语音、音乐）的真实场景及专业音频上取得了最先进的成果。
- 我们提出了跨度提示，这是一种新颖的时间条件信息形式，可为音频分离提供精确的帧级控制。
- 我们开发了统一且全面的分离基准测试 SAM Audio-Bench，涵盖主要音频领域（语音、音乐和音效），包含人工标注的多模态提示，并提供一种与人类感知高度相关的无参考自动评估模型。

2 相关工作

2.1 音频分离模型

说话人分离。传统声源分离旨在将混合音频分解为单个声源。在语音处理中，这通常涉及从混合音频中分离出多个说话人的话语。主流研究将其表述为判别式回归问题：模型为每个声源预测一个潜在域掩码，再通过应用这些掩码重建干净信号。早期

深度学习模型依赖频谱图掩蔽（Takahashi 和 Mitsufuji，2017年；Luo 和 Mesgarani，2018年），而 Conv-TasNet（Luo 和 Mesgarani，2019年）等时域方法则证明，直接处理波形能带来更出色的感知质量。随后，DPRNN（Luo 等人，2020年）和 TF-Locoformer（Saijo 等人，2024年）等双路径架构通过结合块内与块间处理，提升了效率和长程上下文建模能力。基于Transformer的模型进一步拓展了这些思路：SepFormer（Subakan 等人，2021年）采用双路径自注意力来捕捉长期依赖关系，而 MossFormer2（Zhao 等人，2024年）则融合多尺度注意力和掩码细化，以提升实际部署中的鲁棒性并降低延迟。

另一条研究路线将分离视为生成建模问题。生成模型不再预测确定性掩码，而是学习以混合音频为条件的源信号底层分布。早期研究包括 GAN方法（Subakan 和 Smaragdis，2018年；Pascual 等人，2017年）以及归一化流模型（Zhu 等人，2022年）。近期方法（Mariani 等人，2023年；Scheibler 等人，2025年；Dong 等人，2025年）利用扩散和流匹配（Lipman 等人，2023 年）处理复杂的语音重叠。SEP-Diff（Chen 等人，2023年）在梅尔频谱图域中采用去噪扩散过程进行语音分离，而 DiffSep（Scheibler 等人，2023年）则在波形空间中构建混合信号条件评分模型。Diffiner（Sawata 等人，2022年）和 Fast-GeCo（Wang 等人，2024年）等后处理模型则通过扩散优化基于回归的分离器。

除了盲分离之外，目标或可提示的说话人提取利用额外信息来识别所需的说话人。例如，SpeakerBeam（Delcroix 等人，2020年）、VoiceFilter（Wang 等人，2019年）、SpEx+ （Ge 等人，2020年）和 SoloSpeech（Wang 等人，2025b）以注册语句为条件，从特定说话人中提取语音，实现了实例级消歧，但需要干净的参考样本。针对文本提示的说话人分离，相关研究仍然很少；在这类方法中，目标说话人由语义属性指定（例如“女性说话人”或“儿童声音”）。这种方法具有实际价值，因为它无需参考音频即可实现灵活且可解释的控制。

音频增强。音频增强是另一个广泛应用分离方法的领域，旨在去除加性噪声或干扰，同时保留底层信号。尤其是语音增强，长期以来一直被用作自动语音识别和助听器系统的前端（Loizou，2013年；Wang 和 Chen，2018年）。与说话人分离类似，生成建模已被广泛用于语音增强，在各种声学条件下提升感知质量——从早期基于 GAN 的方法（Pascual 等人，2017年；Fu 等人，2019年）到近期基于扩散的模型（Lu 等人，2022年；Richter 等人，2023年）。音乐增强（Schaffer 等人，2022年；Kandpal 等人，2022年）也已得到研究，用于恢复含噪或退化的录音，或作为乐器分离系统的后处理步骤；其中的技术通常改编自语音增强和声源分离。尽管这些增强模型通常侧重于恢复，但我们的目标有所不同。在本文中，我们关注保真分离，即在不改变回声等内在录音属性的情况下，对目标声源进行内容保留式提取，得到分离音轨。

乐器分离。将完整混音音频分离为单个乐器音轨，是分离领域文献中广受研究的另一项任务，也借鉴了许多通用声源分离技术。例如，Demucs（Défossez 等人，2019年）提出了一种采用步幅卷积和学习式合成的时域 U-Net 架构，之后又扩展出混合时频路径和 Transformer 模块，以更好地捕捉长程谐波结构和瞬态（Défossez，2021年；Rouard 等人，2023年）。KUIELAB-MDX-NET (Kim 等人，2021年）采用时频域双流架构，并融合两路输出，在准确率与计算成本之间实现了良好的权衡。最近，Lu 等人 (2024年）提出将输入的复数频谱图划分为多个子频带，再应用分层 Transformer 层来建模频带内和频带间依赖关系。

然而，大多数音乐乐器分离方法仍受限于固定本体。常用的基准测试MUSDB18 (Rafii 等人，2017年）将输出分轨定义为人声、鼓、贝斯和其他。因此，表现出色的模型（Rouard 等人，2023年；Lu 等人，2024年）通常最多只支持 5 条音轨。这一限制削弱了其开放域可用性，也凸显了对更灵活的可提示式分离框架的需求。

目标声音分离。以往大多数声源分离研究都聚焦于语音或音乐等专业领域，而致力于将分离能力推广至开放域声音事件的研究相对较少。

通用声音分离（Kong 等，2023）探索了大规模的本体驱动分离，利用弱标签和层次化条件控制，分离从 AudioSet（Gemmeke 等人，2017年）中提取的预定义声音类别。然而，它依赖固定类别本体，因此灵活性和可控性受限。为克服这一局限，近期方法引入自然语言提示，实现开放词汇分离。AudioSep（Liu 等，2022年）利用 CLAP 对齐的音频-文本嵌入和大规模真实场景混合音频，扩展了声音分离能力。CLIPSep（Dong 等，2023）利用无标签视频训练文本提示分离模型，并以 CLIP（Radford 等人，2021年）嵌入为条件，使音频与文本描述对齐。与这些判别式方法并行，生成式方法也日益受到关注。FlowSep（Yuan 等，2025年）在潜在空间中采用整流流匹配，并以 FLAN-T5 文本嵌入（Raffel 等人，2020年）为条件；SoloAudio（Wang 等，2025a）则采用扩散变换器架构，并通过合成音频增强训练。尽管有提示分离大幅提升了灵活性，但在专业音乐解混或多说话人场景等专业领域中仍表现欠佳，而特定领域系统（Rouard 等，2023）依然更胜一筹。此外，在许多实际应用中，仅靠文本提示不足以区分目标声音，例如区分细微的音效或共现事件。以往研究（Wen 等，2025年；Ge 等，2020年）曾探索将音频作为替代条件模态，而我们则引入跨度提示这一时间条件控制机制，使模型无需外部参考音频即可分离声音事件。

视觉提示分离。 除了文本提示之外，还有一些研究探索将视觉输入作为分离线索，使模型能够识别并提取与特定视觉对象或区域相对应的声音。 Zhao 等人(2018年) 的早期研究表明，通过将视觉区域与相应的声音成分关联起来，空间位置可以引导分离，从而有效分离音乐视频中的乐器。在语音领域，Ephrat 等人(2018年) 展示了如何通过以说话人的面部区域为条件信息来分离目标说话人音频。近期研究（如 IIANet （ Li 等人，2024b）通过模态内与模态间注意力机制整合多层次视听线索，从而改善说话人分离中的跨模态融合。在通用声音与音乐领域，DAVIS（Huang 等人，2025年）将视觉提示分离表述为条件生成任务，并采用基于扩散的框架，以混合音频和视频输入为共同条件合成目标声音。尽管取得了这些进展，视觉提示分离在真实世界开放领域场景中的研究仍不充分。以往的大多数系统都针对特定领域，重点局限于语音或乐器分离，而且通常只在合成数据或小规模基准测试上进行评估。

2.2 音频分离基准与评估

分离基准测试 许多音频源分离研究依赖带有可用音轨的合成混合音频，以便采用基于参考的指标（信号失真比/尺度不变信号失真比）进行评估。尽管这些数据集推动了研究进展（例如用于语音的 WSJ0-2mix 和 WHAM!/WHAMR!）Hershey 等人（2016年）；Wichern 等人（2019年）；Maciejewski 等人（2020年），但它们在真实度和领域多样性方面存在局限。其他数据集套件拓展了条件范围（LibriMix）Cosentino 等人（2020年），或侧重于通过说话人错误率/词错误率进行无参考会话评估（AMI、CHiME-5/6、DIHARD、LibriCSS）Carletta 等人（2006年）；Barker 等人（2018年）；Watanabe 等人（2020年）；Ryant 等人（2019年）；Chen 等人（2020b），但提示模态覆盖范围不足。在音乐领域，MUSDB18/HQ 和 SiSEC/MDX 聚焦于分轨分离（人声、鼓、贝斯、其他）Rafii 等人（2017年）；Mitsufuji 等人（2022年）；Fabbro 等人（2024年），并涌现出 HT-Demucs Rouard 等人（2023年）等性能强劲的系统；Slakh2100/MedleyDB/URMP 则以合成或较小规模的多轨音频提供丰富的乐器种类 Manilow 等人（2019年）；Bittner 等人（2014年）；Li 等人（2018年）。通用声音基准测试（FUSS、AudioSep）旨在实现通用分离，但其数据主要为合成数据（Wisdom 等人，2021年；Liu 等，2022）。

多模态视听方法利用视频（AVSpeech、LRS2/LRS3、VoxCeleb；MUSIC/URMP）Ephrat 等人(2018年）；Afouras 等人(2018年）；Chung 等人(2018年）；Zhao 等人(2018年）；Li 等人(2018年），而 AVS类基准（AVSBench、LU-AVS）则加入了时间声音开关和空间掩码Zhou 等人(2022年）；Liu 等人(2024a）。生成式系统（扩散/流）进一步推动了视听分离的发展Huang 等人(2023年）；Yuan 等人(2025年）。然而，大多数基准测试（i）聚焦于狭窄分类体系，（ii）依赖合成数据或录音室分轨，（iii）仅评估一种提示模态。

分离评估 音频分离系统的传统评估依赖于基于失真的指标，例如信号失真比（SDR）、尺度不变信号失真比（SI-SDR）、信号干扰比（SIR）和信号伪影比（SAR）（Vincent 等人，2006年；Le Roux 等人，2019年）。这些指标量化能量

分离输出与参考信号之间的差异，并已广泛应用于 WSJ0-2mix（Hershey 等人，2016年）和 MUSDB18（Stöter 等人，2019年）等基准测试数据集。然而，这些指标往往无法反映感知质量：两个 SDR 值相近的输出听起来可能截然不同（Le Roux 等人，2019年），而且它们与人类平均意见分（MOS）的相关性较弱，这一点已为人所知（Cartwright 等人，2018年；Cano 等人，2016年）。

主观听音测试仍是评估的黄金标准（Sector，1996年），但成本高、耗时长且难以大规模开展。这导致了一个长期存在的差距：易于计算的失真误差与具有感知意义的评估之间存在脱节。为缩小这一差距，最初为语音编码和传输设计的感知指标（如 POLQA）（Beerends 等人，2013年）试图对听觉机械装置进行建模。尽管这些指标在其预定领域内表现有效，但难以推广至现代分离系统产生的多种伪影（Delgado 和 Herre，2024年）。

近年来，数据驱动的质量预测器日益受到关注，研究表明它们与人类判断的一致性更高（Huang 等人，2022年；Chinen 等人，2020年；Mittag 等人，2021年；Manocha 等人，2021年；Reddy 等人，2021b）。然而，这些研究大多聚焦于语音合成、增强或音频生成。在声源分离领域，评估仍主要依赖类似 SDR 的指标，而听音研究（Jaffe 和 Burgoyne，2025年）证实这些指标与感知判断的相关性较差。这一差距凸显了评估框架的必要性：它们应超越失真误差，更忠实地反映人类的听觉体验。

3 方法

SAM Audio 是一种生成式分离模型，可在文本、视觉和时间跨度提示的条件下，从混合音频中提取目标与残差信号音轨（见图 1）。其核心是，SAM Audio采用基于扩散 Transformer（Peebles 和 Xie，2023年）构建的流匹配模型，并在 DAC-VAE（Polyak 等人，2024年）潜在空间中运行，以联合生成目标音频和残余音频。

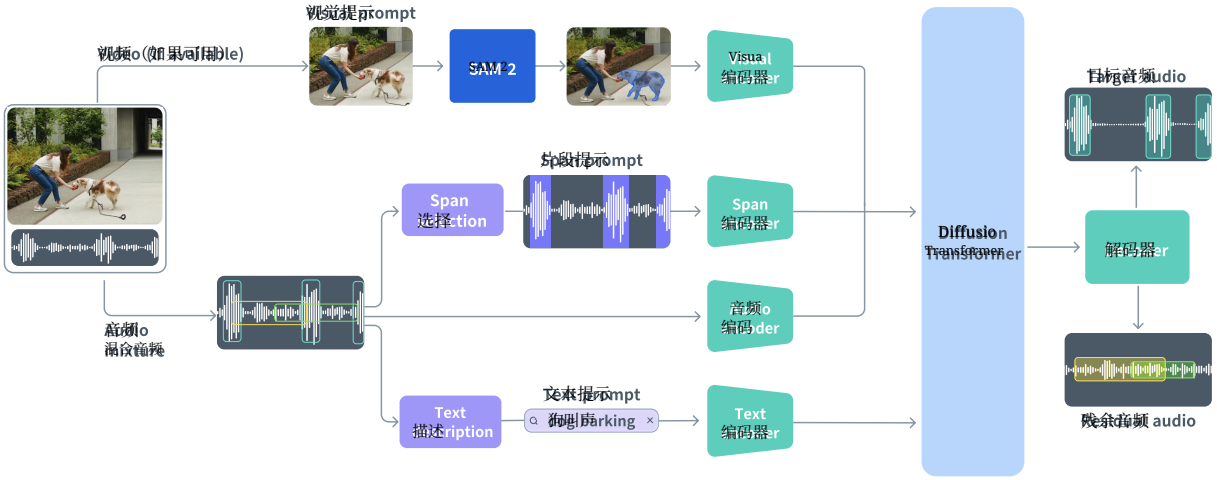

图 1 SAM Audio 概览。给定混合音频，SAM Audio 会根据文本描述（文本提示）、视觉掩码（视觉提示）和时间区间（片段提示）的任意组合，将其分离为目标与残差信号音轨。

3.1 模型架构

3.1.1 使用 DiffusionTransformer 进行流匹配

传统上，音频分离被视为一个判别式学习问题，模型通过在音频频谱图上预测掩码，从混合音频中提取目标信号（Liu 等人，2024b；

Dong 等人，2023年）。然而，隔离目标声音的任务本质上是一对多的——给定一个音频混合，通常会有多个合理的声音，尤其是在嘈杂或重叠的声学场景中。

在 SAM Audio中，我们采用生成建模方法。模型不采用掩码预测，而是学习以多模态提示为条件的目标声音分布，从而更好地捕捉声源的内在多样性。具体而言，SAM Audio 遵循流匹配范式（Lipman 等人，2023年），与近期文本到音频生成（Le 等人，2023年；Vyas 等人，2023年）和视频到音频生成（Polyak 等人，2024年）领域的近期工作类似。

我们的模型学习一个连续向量场，使高斯先验样本 $x_0 \sim \mathcal{N}(0, I)$ 随时间变量 $t \in [0, 1]$ 演化，并将其传输至潜在空间中的数据分布。在每一步 t，模型都会预测瞬时速度场 $u(x_t, c, t; \theta)$，并对其积分以获得时刻 $t = 1$ 的样本 x_1；其中，c 和 θ 分别表示条件和模型参数。正如近期音频生成领域的研究所示，与扩散模型相比，流匹配具有更高的效率和性能（Le 等人，2023 年；Mehta 等人，2024 年；Vyas 等人，2023 年；Prajwal 等人，2024 年；Polyak 等人，2024 年）。

我们采用基于扩散变换器（DiT）的架构（Peebles 和 Xie，2023年），其中每个变换器模块都通过应用于归一化层和残差连接的缩放与平移操作，由流时间嵌入进行调制。一个共享的多层感知机（MLP）将标量 t 映射为六个调制参数（四个缩放参数、两个偏置参数），供所有变换器模块使用；此外，还会添加特定于层的偏置，以捕捉随深度变化的效应。这种参数共享在不影响模型性能的前提下缩减了模型规模。

我们使用单独训练的 DAC-VAE（Polyak 等人，2024年）将音频表示为紧凑的潜在特征序列。该模型采用类似 DAC 的（Kumar 等人，2023年）自编码器架构，但将残差向量量化器（RVQ）替换为变分自编码器（VAE）（Kingma 和 Welling，2013年）瓶颈层。每个音频片段都会被编码为一个 $T \times C$ 序列，采样率为 25 Hz，具有 $C = 128$；与常用的 Encodec 特征（Défossez 等人，2022 年）相比，这种表示具有更高的重建质量和更低的帧率。移除残差向量量化器后，模型以 VAE（Kingma 和 Welling，2013年）的形式进行训练，从而生成平滑的高斯潜变量，非常适合在潜在空间中进行连续流匹配。

对于 SAM Audio，我们通过沿通道维度拼接目标声音和残余声音的 DAC-VAE 特征x_{tgt} 和 x_{res}，形成 $x = [x_{\text{tgt}}, x_{\text{res}}] \in \mathbb{R}^{T \times 2C}$，对二者进行联合建模。随后，模型逐步对这一联合表示进行去噪，从而在单次推理中同时预测这两个组成部分。

3.1.2 提示类型

SAM Audio 支持三种提示类型：文本、时间跨度和视觉。这些提示既可以单独使用，也可以联合使用，以指定要提取的目标声音。

文本提示。 文本提示是对目标声音的自由格式自然语言描述。对于音频分离，我们发现，简洁的名词-动词短语（NP/VP）描述（例如“女性说话”）相比完整句子（例如“一名女性正在发表演讲”），对用户而言更加有效且自然。这种格式也更符合音频编辑中的常见用法。

视觉提示。 视频是多模态音频数据的常见来源，指定视觉区域为分离声源提供了一种直观方式。例如，要从视频中提取狗叫声，用户只需在视觉帧中指示狗所在的区域（见图 1）。我们采用 SAM 2（Ravi 等人，2024年）来获取目标视觉掩码。给定视频 $V \in \mathbb{R}^{T \times H \times W \times 3}$，用户通过点击或边界框与 SAM2 交互，提供二值掩码 $M \in \{0, 1\}^{T \times H \times W \times 3}$。SAM Audio将（$V$，$M$）作为输入。为简便起见，它仅处理经过掩码的视觉帧 $V \odot M$。

片段提示。 仅用文本描述任意音频事件可能颇具挑战，尤其是电影配乐等复杂声景。为此，我们提出了跨度提示，这是一种时间条件控制形式，用于指定目标声音出现的时间区间。其核心直觉是，时间戳可以消除重叠声音事件带来的歧义。例如，在一个包含 0–6 秒女性语音和 1–2 秒狗叫声的混合音频中，仅凭时间区间就足以分离出狗叫声。形式上，片段提示定义为一组时间区间 $S \in \mathbb{R}^{k \times 2}$，标记目标事件的开始和结束时间。需要注意的是，当多个声音事件同时出现时，跨度提示本身可能存在歧义

。实际应用中，它尤其适用于前景声音，或与文本提示、视觉提示结合使用（参见第 7.3节）。

3.1.3 音频混合与提示编码

为生成目标音频，SAM Audio 同时以输入混合音频和三种提示作为条件。视觉提示和片段提示首先被编码为帧对齐特征序列，再与带噪潜变量拼接，然后输入 DiT 主干网络。相比之下，文本提示被编码为全局文本嵌入，并通过 DiT 主干网络中的交叉注意力层融入模型。帧级和全局语义信息相结合，有助于捕捉用于分离的异质线索。

音频编码器。 输入混合音频由单独训练的 DAC-VAE（Polyak 等人，2024年）编码，生成潜在特征序列 $x_{\text{mix}} \in \mathbb{R}^{T \times C}$，频率为 25 Hz。我们直接在 DAC-VAE 空间中进行处理，以最大限度地减少音频中的信息损失，从而确保分离后的输出忠实保留原始音频。

文本编码器。 我们使用预训练的 T5-base 编码器（Raffel 等人，2020年）对文本进行编码，生成特征序列 $c_{\text{text}} \in \mathbb{R}^{N_{\text{txt}} \times d_{\text{txt}}}$，其中 $d_{\text{txt}} = 768$。这些特征通过交叉注意力层注入 DiT 主干网络，使模型能够关注与描述相符的内容。由于提示简短且含义明确，因此无需对文本编码器进行微调。

视频编码器。为提供视觉 grounding，我们以感知编码器（PE）提取的帧级视频特征为条件（Bolya 等人，2025年）。感知编码器是一种先进的视觉编码器，采用大规模对比式视觉语言学习进行训练，在图像与视频基准测试上取得了领先的零样本准确率。与 CLIP（Radford 等人，2021年）或 MetaCLIP（Xu 等人，2023年）相比，后两者广泛用于以往的视频条件音频生成（Polyak 等人，2024年）或分离研究（Dong 等人，2023年），PE 能生成更稳健、语义更丰富的表征，尤其是在动作和场景上下文方面（Bolya 等人，2025年）。我们提取帧级 PE 特征，并对其重采样，使其与音频帧率匹配。随后，门控线性投影层将这些特征映射到 DiT 维度，所得序列再按帧与音频表征拼接。

片段编码器。与文本转语音（TTS）中使用的音素序列类似，我们将原本以一组时间区间 $S \in \mathbb{R}^{k \times 2}$ 表示的片段提示转换为帧同步词元序列 $S'_{1:T}$。每个词元 $s'_t \in \{\text{<sil>}, +\}$ 表示目标事件在帧 t 是静音还是处于活动状态。我们使用可学习的嵌入表对词元序列进行嵌入，并沿通道维度与音频特征拼接。这样便提供了明确的时间先验，引导模型聚焦于需要提取的相关区域。

并非所有训练样本都包含全部三种提示模态。在这种情况下，我们会提供虚拟条件：文本使用空字符串，视频使用全零向量序列，时间跨度使用特殊空值标记序列（<null>）。为提升鲁棒性，我们还会在训练期间以概率p_{video}、p_{text}、p_{span} 随机丢弃每个可用提示，并将其替换为相应的虚拟条件。这有助于模型在推理时稳健地处理缺失的提示。

3.2 训练目标

流匹配目标　　标。 在每个流步骤 t，模型都会接收带噪声的潜变量 x_t 和条件集合

$$c = \{x_{\text{mix}}, c_{\text{text}}, c_{\text{vid}}, c_{\text{span}}\},$$

其中，c_{text} 表示文本特征，c_{vid} 表示掩码视频特征，c_{span} 表示目标音频片段嵌入。模型预测速度场 $u(x_t, c, t; \theta)$，用于更新 x_t，使其趋向干净目标残差表示

$$\mathcal{L}_{FM} = \|u(x_t, t, c; \theta) - (x_1 - (1 - \sigma_{min})x_0)\|, \tag{1}$$

其中，x_1 和 $x_0 \sim \mathcal{N}(0, 1)$分别是目标音频和随机噪声。

音频表示对齐。为了根据用户提示分离出正确的声音事件，模型必须先推断要提取什么，以及它在混合音频中的发生时间。为鼓励这种行为，我们

在Yu 等人(2025年)的方法基础上，引入辅助多层感知器层，将混合音频与输入提示进行条件化后形成的中间联合表示投影到外部音频事件检测（AED）模型的嵌入空间中。随后，我们应用辅助对齐损失，迫使这些投影后的表示与真实目标音频的 AED 嵌入相匹配。通过明确引导隐藏状态与事件级语义和时间结构对齐，这一目标补充了流匹配损失，并促使模型在训练过程中内化“提取什么以及从何处提取”。

对于对齐目标，我们首先提取目标音频嵌入 $a_{tgt} = \text{AED}(x_{tgt}) \in \mathbb{R}^{T \times F}$，其中 F是 AED 模块的特征维度。然后，令 $h_t \in \mathbb{R}^{T \times D}$ 为 DiT 隐藏表示，其中 T 是输入音频长度， D 是 Transformer 的通道维度，我们将其投影为$\hat{a}_t = \phi(h_t) \in \mathbb{R}^{T \times F}$，其中 $\phi(\cdot)$ 是一个小型 MLP。然后，我们通过最小化下式来最大化 a_{tgt} 与 $\hat{a}_t$之间的余弦相似度

$$\mathcal{L}_{aux} = \mathbb{E}_t \left[1 - \text{sim}(\hat{a}_t, a_{tgt}) \right], \tag{2}$$

其中 $\text{sim}(A, B) = \frac{A \cdot B}{\|A\| \|B\|}$。

我们将 FM 损失和辅助损失合并为:

$$\mathcal{L} = \mathcal{L}_{FM} + \lambda * \mathcal{L}_{aux}$$

其中 λ 是用于控制对齐损失影响的超参数。我们使用 (Kong 等人，2020年) 中的 AED 模型来提取目标音频表示 a_{tgt}。对于辅助 MLP aux_ϕ，我们采用了一个隐藏维度为 2048 的三层多层感知器，并使用 GeLU 激活函数和层归一化。

3.3 通过跨度预测增强文本提示

文本提示为指定目标声源提供了最便捷的接口，因为用户只需描述所需的声音即可。尽管操作简单，跨度标注能在目标声音处于活动状态时提供帧级控制，从而实现更准确的分离。在经验上，联合文本与跨度条件信息在通用声音事件分离方面优于仅使用文本的条件信息（见第7.3节）。然而，获取真实片段需要手动边界标注，这可能非常耗时。为弥合这一差距，我们提出了一种简单的方法，用于在推理时近似跨度标注。为此，我们使用一个名为 $\text{PE}_{\text{A-Frame}}$ 的辅助模型来预测片段。

具体而言，$\text{PE}_{\text{A-Frame}}$ （Vyas 等人，2025年）是一种语言查询声音事件检测模型，能够学习检测音频信号中由自由文本描述的声音处于活动状态的精确时间帧。给定一个音频片段和一段文本描述（例如“狗叫声”），该模型会预测帧级概率分数，以指示该事件发生的时间。该模型通过引入帧级损失函数扩展了 CLAP 风格音频文本嵌入，从而实现对连续音频中语言信息的细粒度时间定位。

给定音频混合 x_{mix} 和文本提示 c_{text}，我们使用 $\text{PE}_{\text{A-Frame}}$ (Vyas 等人，2025年)来估计文本描述所指声音事件逐帧的活动情况。通过对逐帧活动情况进行阈值处理，我们得到近似的时间跨度序列$\hat{c}_{\text{span}}$。与 Wang 等人（2022a）类似，随后我们将预测时间段$\hat{c}_{\text{span}}$ 和原始文本提示 c_{text}都作为分离模型的条件信息。具体而言，给定 SAM 音频模型$\mathcal{M}$ 和文本条件信息 c_{text}，我们通过 $\mathcal{M}(x_{\text{mix}}, c_{\text{text}}, \text{PE}_{\text{A-Frame}}(x_{\text{mix}}, c_{\text{text}}))$进行文本提示，而不是使用 $\mathcal{M}(x_{\text{mix}}, c_{\text{text}})$。

3.4 使用多重扩散进行长音频分离

受 GPU 内存限制，无法在单次前向传递中处理任意长度的音频。简单的分块处理方法是将混合音频划分为互不重叠的片段并独立应用 SAM Audio，但这会引入边界伪影和不连续性，对于持续或缓慢变化的声音尤其如此。为确保时间连贯性，我们采用多重扩散方法（Bar-Tal 等人，2023年；Polyak 等人，2024年），并将其应用于音频分离。

具体来说，我们将混合音频划分为重叠窗口。对于窗口 j，我们构建特定于窗口的条件信息 $c^{(j)} = \{x^{(j)}_{\text{mix}}\ c^{(j)}_{\text{vid}}\ c^{(j)}_{\text{span}},\ c_{\text{text}}\}$。在每个流匹配步骤 t_i 中，我们使用共享的时间计划 $\{t_i\}$ 并行求解所有窗口的常微分方程：$\tilde{x}^{(j)}_{t_{i+1}} = x^{(j)}_{t_i} + \Delta t_i\, u\left(x^{(j)}_{t_i}, t_i, c^{(j)}\right)$。

随后，我们使用归一化软掩码将局部预测合并到全局潜变量中 :

$$x_{t_{i+1}} = \sum_j \mathrm{pad}\Big(m^{(j)} \odot \tilde{x}^{(j)}_{t_{i+1}},\, j\Big), \qquad \sum_j \mathrm{pad}\Big(m^{(j)},\, j\Big) = \mathbf{1},$$

其中 $m^{(j)}$ i 是应用于片段 j 的三角窗，pad 则将其补零至全局长度 h.

这一迭代过程使信息能够在每个扩散步骤中跨重叠区域传播，从而实现平滑且全局连贯的长音频分离，避免边界伪影。

4 数据

SAM Audio 的训练数据由以下形式的配对组成：(x_{mix}， x_{tgt}， x_{res}， c)，其中 x_{mix} 是音频混合， x_{tgt} 是目标音轨， x_{res} 是残余音轨，而 $c \in$ {文本、视频、时间跨度} 则是一个或多个条件提示。本节介绍了构建这类训练元组的策略。我们首先介绍用于构成 (x_{mix}， x_{tgt}， x_{res}) 的音频数据，然后说明如何为 c 生成提示。由于 SAM Audio 涵盖所有音频子模态，包括语音、音乐和通用音效，因此我们针对每种音频子模态制定了数据策略。

4.1 音频

类别	输入混合音频 x_{mix}	目标音频 x_{tgt}	残余音频 x_{res}
全真实三元组	✓	✓	✓
合成混合音频	✗	✓	✓
伪标注音轨	✓	✗	✗

表1 SAM Audio 中使用的训练数据类型概览。每行表示混合音频、目标音频和残差信号来自真实音频（√）还是合成数据/伪标注数据（）。

总体而言，我们的训练数据来自（1）一个大规模、中等质量的音视频语料库（约 ∼1M 小时），以及（2）一组小到中等规模的高质量音频数据集（见表 2）。只有少部分数据包含可直接用于模型训练的真实音轨。其余数据则根据我们是合成混合音频（x_{mix}），还是合成目标/残差信号（$x_{\mathrm{tgt}}, x_{\mathrm{res}}$），用于构建合成混合音频或伪标注数据。

表 1 总结了 SAM Audio 中采用的三种数据构建范式。各范式的区别在于，混合音频、目标音频和残差信号来自真实录音，还是通过合成或伪标注获得。

数据源	模态	含分轨	质量	声音	音乐	语音	样本数（百万）	#小时 (K)
通用视频	音频、视频	✗	中等	✓	✓	✓	$\mathcal{O}$(100)	$\mathcal{O}$(1000)
通用音频	音频	✗	中等	✓	✓	✓	$\mathcal{O}(1)$	$\mathcal{O}(1)$
语音对话	音频	✓	High	✗	✗	✓	$\mathcal{O}$(10)	$\mathcal{O}$(10)
高品质音乐	音频	✗	High	✗	✓	✗	$\mathcal{O}$(10)	$\mathcal{O}$(10)
多轨音乐	音频	✓	High	✗	✓	✗	$\mathcal{O}$(0。1)	$\mathcal{O}(1)$
高质量音效	音频	✗	High	✓	✓	✓	$\mathcal{O}$(10)	$\mathcal{O}$(10)
高质量视频	音频、视频	✗	High	✓	✓	✓	$\mathcal{O}$(0.1)	$\mathcal{O}$(0.1)

表2 SAM Audio 的训练数据来源。“带音轨”表示是否提供真实目标／残差信号轨。

4.1.1 全真实三元组

理想的训练数据形式（x_{mix}， x_{tgt}， x_{res}）由真实且完全分离的音轨组成，这些音轨叠加后形成自然混合信号：$x_{\mathrm{mix}} = x_{\mathrm{tgt}} + x_{\mathrm{res}}$。这类数据提供了尽可能干净的监督信号，因为模型可以直接学习如何从混合音频映射到其组成声源，而不会受到合成伪影的影响。

音乐和语音领域都有全真实三元组。

完全真实的音乐。 我们使用了内部的高质量音乐数据，其中包含清晰的乐器音轨（即 多轨音乐 ，见表 2），用于引入乐器提取能力。该数据集共包含 10,610 首独特的音乐作品，总时长为 536 小时。每首音乐作品由多种乐器（例如鼓、贝斯、吉他）和人声轨组成。我们通过混合 x_{mix}、 x_{tgt}和 x_{res} 同一作品中的这些乐器，构造三元组。假设 $x_{\text{mix}} = \sum_{i=1}^{N} x_i$ ，其中 x_i 表示每条乐器音轨的音频，那么我们会创建 N 个目标不同的三元组；对于每段乐器音频 x_i，定义 $x_{\text{tgt}} = x_i$ 和残差信号$x_{\text{res}} = \sum_{j=1,\ j\neq i}^{N} x_j$。为了提高鲁棒性，我们不只是简单地将各条音轨相加，而是根据相对于目标音频的信噪比 ± 5 ，对每条乐器音轨重新缩放。

完全真实的语音。 我们使用了一个对话语音语料库（即 语音对话 ，见表 2），其中包含 21,910 小时的音频。每段对话都提供两条干净的说话人分轨，分别记作 x_{spk1} 和 x_{spk2}。我们将混合音频构造为 $x_{\text{mix}} = x_{\text{spk1}} + x_{\text{spk2}}$ ，并构造两个训练三元组： $(x_{\text{mix}}, x_{\text{spk1}}, x_{\text{spk2}})$ 和 $(x_{\text{mix}}, x_{\text{spk2}}, x_{\text{spk1}})$。为提高鲁棒性，我们将残余说话人的音频重新缩放，使其相对于目标说话人的信噪比偏移量从 ± 15分贝中随机采样。

4.1.2 合成音频混合

尽管真实分轨数据极具价值，但相对稀缺，且通常仅适用于特定领域（例如，多说话人语音数据或乐器数据集）。合成混合音频通过随机混合两段音频生成，是以往研究中常用的训练数据策略（Liu 等人，2022年；Yuan 等人，2025年；Ma 等人，2024年），常用于通用音效。由于SAM Audio涵盖特殊领域和音频领域，我们针对各个领域设计了相应的噪声混合策略。

合成噪音乐。我们使用了内部数据，其中包含 $\sim 20K$ 小时的纯净音乐（即高品质音乐，见表2）。为了创建合成混合音频，我们将该数据集与通用声音数据中的非音乐音频混合，信噪比为 ± 15。为提升人声与背景音乐分离效果，我们使用音频事件检测（Koutini 等人，2022年）将音乐数据集划分为有人声和无人声两类，并相应设置文本提示（见第4.2.1节）。

合成噪声语音。对于语音提取，我们使用第4.1.1节所述的同一对话语音语料库。为了构建合成混合音频，我们将每段对话中的两条说话人分轨直接相加，形成单通道混合音频 $x_{\text{speech}} = x_{\text{spk1}} + x_{\text{spk2}}$，其中自然包含单说话人和多说话人片段。对于噪声成分 x_{noise}，我们从通用声音池中随机采样一个非语音音频片段。最终的混合音频构造为 $x_{\text{mix}} = x_{\text{speech}} + x_{\text{noise}}$。

合成通用声音。我们将音效数据分为两种主要类型：真实场景声音（即通用视频和通用音频，见表2）以及高质量声音（即高质量音效，见表2）。真实场景音频由在非受控环境中录制的内容组成，通常包含多个重叠的声音事件以及环境声。相比之下，专业声音录音是在受控设置下制作的，作为专业音频制作的源素材；它们通常只包含单个声音事件，没有背景噪声。在进行音频混合时，我们通过将真实场景片段与其他真实场景片段混合，以及将专业声音片段与其他专业声音片段或真实场景片段混合，来生成训练样本。

4.1.3 伪标签数据引擎

上述随机混合策略的一项局限是，它经常生成不真实、无法充分反映真实世界音频混合的混合音频。例如，将体育场中的人群欢呼声与在森林中录制的鸟鸣声混在一起，会产生一种在自然环境中很少出现的不自然组合。使用这类混合音频进行训练，可能会妨碍模型学习有意义的分离线索。

为了解决这一问题，我们使用 SAM Audio 的一个中间检查点来合成更真实的训练元组。该检查点采用相同的训练方案进行训练，但未使用伪标注数据。我们使用这个早期版本的 SAM Audio，从自然音频录音中生成目标与残差信号，本质上是从未标注混合音频中自助生成新的训练数据。

数据合成流程由两个阶段组成：使用 SAM Audio 进行分离 ，以及使用专用模型筛选分离音频的各种特征。

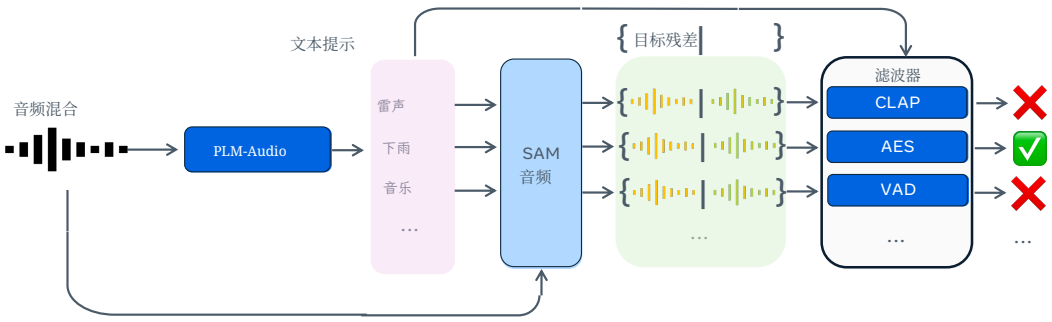

图2 我们的伪标签数据合成流程示意图。PLM-Audio 从混合音频中生成文本提示，用于引导 SAM Audio 生成目标/残差信号。过滤阶段仅保留高质量的伪标注音轨。

分离。 在分离阶段，我们将中间版 SAM Audio 检查点作为数据引擎，从大规模未标注音频录音中生成目标音轨和残差音轨，形成一个伪标签训练候选样本池。

鉴于文本提示分离具有良好的可扩展性，我们主要关注基于文本的合成。首先，我们使用音频描述模型 PLM-Audio Vyas 等人(2025年)（见第 4.2.1节），从 100 万小时的 真实场景通用声音数据中提取文本描述。PLM-Audio 的输出是一系列音频事件描述，我们将每条描述用作 SAM Audio 伪标注的文本提示。

筛选。 实际上，这个初步检查点的分离质量存在显著差异。为避免降低最终模型的性能，我们采用严格的筛选策略，根据涵盖多个质量方面的一组标准剔除低质量候选样本。我们使用 CLAP（Wu 等人，2023年）衡量文本-音频对齐，通过 Audiobox-aesthetic 模型的制作复杂度（PC）维度（Tjandra 等人，2025年）评估音频洁净度，并使用语音活动检测器（pydub）检测静音过多的输出。为加强视觉提示训练，我们进一步筛选出视听对应关系较强的子集。对于分离时使用的同一文本提示，我们使用内部开发的文本提示视频分割模型获取视觉掩码，并计算视听对齐分数和掩码覆盖率。对于前一项指标，我们使用 ImageBind（Girdhar 等人，2023年）这一大规模音频-视频-文本对比模型，计算音轨与掩码视觉帧的嵌入之间的余弦相似度。只有满足表3 中所有条件的伪标签样本才会被保留。

标准	阈值
文本-音频筛选（必须全部通过）	
CLAP（文本，目标音频）	> 0.35
CLAP（文本，残余音频）	< 0.0
美学PC分数（目标音频）	< 2.5
静音比例（目标音频）	$< 95\%$
额外视听筛选	
掩码覆盖率（掩码区域）	> 0.02
ImageBind（目标音频，掩码区域）	> 0.2

表3 合成训练数据的过滤标准。只有满足所有文本-音频标准的样本才会保留；视觉提示样本还需满足额外的视觉相关约束。

伪标注应用于我们训练数据中的通用视频部分（表 2）。我们仅对包含多个声音事件的混合音频进行伪标注。经过多阶段过滤后，得到的伪标记数据集规模远小于用于合成混合音频的高质量音频。表 4 总结了用于 SAM Audio 训练的最终伪标记数据集。

数据源	模态	合成音轨	声音	音乐	语音	#样本 (M)	#小时 (K)
PL-Audio	音频	✓	✓	✓	✓	$\mathcal{O}(1)$	$\mathcal{O}(1)$
PL-Video	音频、视频	✓	✓	✓	✓	$\mathcal{O}(0.1)$	$\mathcal{O}(0.1)$

表 4 伪标签 SAM Audio 中使用的训练数据。纯音频输入和音频-视频输入均经过 b 处理 y
中间版 SAM 用于生成合成音轨的音频检查点。

4.2 提示生成

给定一个音频元组（x_{mix}， x_{tgt}， x_{res}），我们会生成相应的提示 c，用于指定目标音频x_{tgt}。本节将介绍如何生成文本、视觉和基于片段的提示。

4.2.1 文本

为目标音频生成文本提示，相当于对其内容进行简洁且忠实于语义的描述。在分离任务中，我们发现，采用名词短语／动词短语形式的提示比完整的自然语言句子更直观、更有效。例如，相较于*"a dog is barking"*，更倾向于使用*"dog barking"*，因为它能更直接地指定要提取的声音事件。

为大规模获取高质量提示，我们训练了内部开发的音频语言模型*PLM-Audio*。该模型是一种具备视觉感知能力、经微调用于音频描述生成的音频大语言模型。只要有原始元数据，我们就会将其与 PLM 输出合并，并舍弃音频文本对应关系较低的字幕。

PLM-Audio。 PLM-Audio 是感知语言模型（PLM）（Cho 等人，2025年）的视听版本。它使用视听编码器 PE-AV（Vyas 等人，2025年），而不是视觉感知编码器（PE）（Bolya 等人（2025年）。PE-AV 提取帧级视听表征，并使用拥有 8B 参数的 LLaMA 解码器（Dubey 等人，2024年）将其解码为语言输出。

我们采用三个阶段训练 PLM-Audio，训练方案与 PLM（Cho 等人，2025年）的训练方案非常相似：

1.预热阶段。我们冻结骨干编码器，仅微调一个轻量级投影多层感知机，将 PE-Audio 嵌入映射到大语言模型嵌入空间。这使特征空间对齐，同时保留预训练表征。2. 中期训练。我们解冻整个模型，并使用由 PE-Audio 生成的大规模合成描述语料进行微调，其中包含纯音频样本和视听样本。我们优先选择音视频对齐分数较高的话语（例如 ImageBind 相似度较高的话语），确保训练聚焦于有充分依据的样本。3. 后期训练。我们使用精选任务组合进行微调，这些任务鼓励以名词短语／动词短语形式生成提示，并提升对下游分离相关属性的覆盖度。

在可用时纳入元数据。我们的部分训练数据来自专业录制的环境音与音效数据集。这些数据集通常包含标题、描述和关键词标签等元数据。然而，这些元数据可能存在噪声，包含无关或过于宽泛的信息。与此同时，PLM-Audio 通常会生成相关的音频描述，但偶尔也会产生假阳性声音事件。为获得可靠的名词短语／动词短语形式提示，我们采用三步改写流程：

1. 对每个音频片段运行 PLM-Audio，以获得初始描述。2. 使用大语言模型（Dubey 等人，2024年）将所有可用元数据（标题、关键词、描述）与 PLM-Audio 生成的描述合并，生成经过清理的详细描述，同时剔除无关内容。3. 指示同一个大语言模型仅从详细描述中提取*NP/VP* 风格的声音事件短语。

CLAP过滤。 由于我们的字幕大多是自动生成而非人工标注的，因此我们会执行筛选步骤，以去除质量较差的字幕。为此，我们使用 CLAP 计算文本与音频的相似度（Wu 等人，2023年），并丢弃相似度低于 0.28的样本；在初步实验中，这一阈值约对应第 25 百分位数。我们评估了多个百分位数阈值

（$p = \{0.05, 0.10, 0.25, 0.50\}$），并发现，在 $p = 0.25$ 进行筛选，可以在数据规模和字幕质量之间取得最佳平衡。

请注意，我们的训练数据中有一小部分属于这样的领域：无需使用上述流程即可获得准真实文本描述。对于全真实音乐子集（第4.1.1节），我们将简单模板应用于乐器标签（例如钢琴 → 钢琴演奏），以生成文本提示。对于全真实语音子集（第4.1.1节），我们使用预训练性别分类器（Huh，2024年）识别说话者性别，然后将其映射为模板化描述（例如女性 →女性说话）。

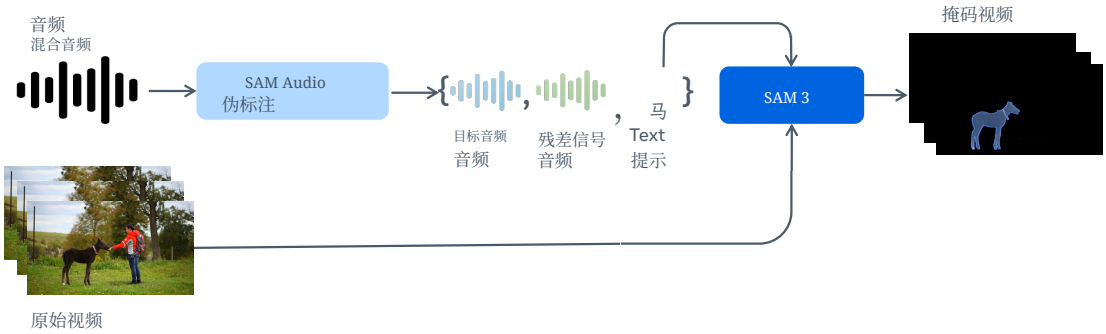

图3 伪标注视觉数据示意图。伪标注流程会生成目标音频的文本字幕，并用其提示 SAM3，以获取视觉掩码。

4.2.2 视频

在推理过程中，视觉提示允许用户通过在视频上提供视觉掩码来指定目标声源，模型则需要分离出相应的声音。视觉掩码使用 SAM 2（Ravi 等，2024年）生成，覆盖所有帧中的单个物体或感兴趣区域。模型训练的主要目标是促使模型将视觉接地声源与音频关联起来，从而让用户在测试时只需选择视频中的一个物体，即可进行实例级分离。

为任意音频合成视频颇具挑战，因为这需要生成时空连贯的帧。因此，我们完全依赖带有配对音轨的自然视频。我们的视觉训练数据构建方案由两个互补部分组成：(a) 完整视频，(b) 视觉掩码。

完整视频。 对于大部分训练数据，我们直接使用完整的视频片段，不依赖显式的分割掩码，使模型能够学习视觉内容与音频事件之间的弱监督关联。然而，许多真实世界视频包含视觉上无法对应的声音，例如后期制作时添加的背景音乐或画外旁白。使用此类样本进行训练可能会产生负面影响，因为这会促使模型在视觉帧与无关音频之间建立虚假的关联。为缓解这一问题，我们计算每个视频的 ImageBind 分数（Girdhar 等人，2023年），并仅保留分数高于阈值的视频，使数据集倾向于包含源自场景、且在视觉上有依据的声源。我们将此筛选应用于 ∼1M 小时通用声音语料库。

视觉掩码。 对于伪标注音视频数据，我们会显式构建与目标声音对应的视觉掩码。与整段视频条件信息不同，伪标注片段通常包含多个发声物体，其中只有一部分与目标音频相匹配。因此，我们使用从 PLM-Audio 获取的目标声音文本字幕提示 SAM3（Carion 等人，2025年）（见图 3），以此生成掩码。实际中，我们发现掩码质量存在显著差异，原因包括音视频不匹配（例如，字幕描述的声音在画面中并不存在）以及 SAM3 预测错误。因此，第 4.1.3 节所述的基于 ImageBind 的筛选流程至关重要，它能确保最终训练集中只保留视听对应关系可靠的样本。

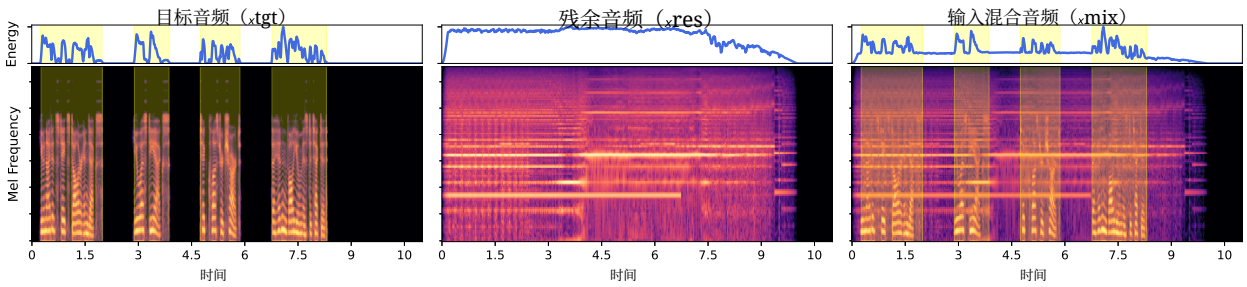

图4 时间跨度生成示意图。目标音频 x_{tgt}、残差信号 x_{res} 和混合音频 x_{mix} 的均方根能量（上）与梅尔频谱图（下）。黄色区间表示检测到的、对应于活跃声音事件的片段。

4.2.3 时间跨度

并非所有音频混合都适合进行跨度提示。尤其是以绵长、连续的环境音（如下雨声、交通声）为主的混合音频，时间结构较弱，因此时间跨度信息难以发挥作用。我们的关键发现是，跨度提示对于突发性离散声音事件（例如关门声 或狗吠声）最为有效，因为这些事件的时间位置能为分离提供有力线索。

为此，我们仅使用满足这一属性的音频来构造音频混合。对于通用声音，目标音频取自高质量音效声音数据（第4.1.2节），其中主要包含干净、孤立的声音事件；噪声音频则取自较长时长的片段（>10 秒），这些片段通常包含环境背景或持续背景。对于音乐和语音，我们还纳入了全真实音乐数据集和全真实语音数据集，以确保广泛的领域覆盖度。

给定目标音频后，我们使用静音阈值为pydub的语音活动检测（VAD），阈值为 -40dBFS，并将最短发声时长设为250 毫秒，以获得一个二值帧级掩码，用于标示发声区域和静音区域。随后，将连续的发声片段转换为时间区间，并将其作为提供给SAM Audio的片段提示。图4展示了时间跨度提示音频数据的生成概览。

5 评估

5.1 SAM Audio-Bench

5.1.1 SAM Audio-Bench：统一模态、领域与真实感

我们发现现有分离基准测试中存在三个长期未解决的问题：(i) 真实感与参考音轨之间的矛盾（真实场景声学环境与音轨可用性），(ii) 文本/视觉/时间提示模态覆盖有限，以及 (iii) 在统一协议下跨领域的广度不足（语音、音乐、乐器、一般声音）。

我们推出一个真实场景、多模态的分离基准测试，以解决这些问题 s:

1. **生态有效性**：所有条目均取自真实场景音频/视频或制作级视频：**AudioSet**（Gemmeke 等人，2017年）、**VGGSound**（Chen 等人，2020a）、**MUSIC** （Zhao 等人，2018年）、**MUSIC-AVQA** （Li 等人，2022年）、**AVSpeech** （Ephrat 等人，2018年）和 **CondensedMovies** （Bain 等人，2020年）。2. **多模态提示**：每个 10 秒测试实例都包含人工标注的视觉 **SAM 掩码片段**（ Kirillov 等人，2023 年）（当发声物体出现在屏幕上时）、时间**正/负片段**和语言**文本描述**。3. **分类覆盖范围**：音轨分类体系以**AudioSet** 本体（Gemmeke 等人，2017年）为基础，并支持标注员扩展类别。专门任务包括：**语音清理、说话人分离、音乐清理/移除、乐器音轨**（37 个类别）和**一般声音**。

标注协议。 标注员会枚举声音事件并提供文本描述（随后我们会将其规范化为简洁的名词或动词短语），为可见的发声源绘制视觉掩码片段——这些掩码片段

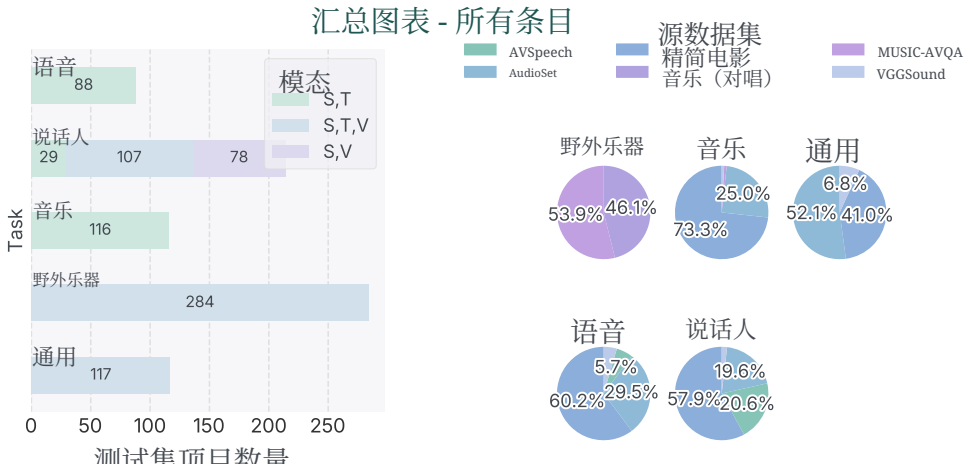

图5 SAM Audio-Bench 中任务、模态和数据集覆盖范围的总结。模态缩写如下：“T”表示条目可使用纯文本提示（例如，对于说话人分离，这意味着文本描述可以明确对应到单个说话人）；“V”表示目标声音出现在屏幕上，并且提供了 SAM 掩码片段；“S”表示目标声音具有事件边界。

适用于视频中的每一帧，且每段视频都统一为每秒24帧——并标注时间上的存在/缺失片段。每个（目标声音，视频）配对都会生成可互换的文本、视觉和时间提示，从而支持跨模态组合进行受控消融实验。

5.1.2 与现有评估集的总结比较

表 5 总结了各代表性基准测试与我们的SAM Audio-Bench在模态覆盖范围、真实度、领域范围和参考可用性方面的情况。

如表 5 所示，SAM Audio-Bench 独特地结合了：(i) 真实场景中的真实音频/视频；(ii)在同一批条目上使用多模态提示（文本、视觉掩码片段、时间跨度）；*(iii)*跨领域覆盖（语音、音乐、乐器、一般声音）；以及 *(iv)*无需参考的人工评估。

图 5 展示了已发布的整个 SAM Audio-Bench 数据集的汇总统计数据，包括每项任务可用的测试集项目数量，以及每项任务支持哪些提示模态；我们还展示了 SAM Audio-Bench 视频的来源数据集明细。

中的其他图　　附录 B 展示了每项任务的其他细分数据，涵盖整个数据集以及所
以获得更均衡的子集，该子集用于人工评估。
子集，该子集用于人工评估。

5.2 SAM Audio JudgeModel

为开发 SAM 音频评判器，我们系统地研究了与感知相契合的音频分离评估。具体而言，我们旨在设计一个评估框架，使其：(i) 无需参考信号即可运行，因此适用于现实应用；(ii) 能够对分离音频进行细粒度的感知评估；(iii) 与人类听觉判断具有更强的相关性。

5.2.1 数据收集

现有的音频分离评估指南通常较为简单，只关注一些粗略标准，例如分离音频与提示之间的相关性，或输出的总体音频质量Liu 等人 (2022年)。这种高层次且定义宽泛的评估目标会使评估结果含糊不清。在这些设置下收集的评分往往混杂了多个感知方面，并且可能因评分者的个人解读而偏向某些标准，导致结果难以保持一致地解读。

表5 比较 代表性音频分离基准测试与 SAM Audio-Bench 的比较。真实：Y/N/M（混合 d 真实+合成数据）。提示 词：T=文本、V=视觉、S=时间跨度。

		任务覆盖范围						
基准测试	**真实**	**语音清洁**	**说话人分离**	**通用音乐**	**清洁乐器音轨**		**提示词**	**源信号/混合信号**
WSJ0-2mix / WHAM! / WHAMR![a]	N	–	✓	–	–	–	–	合成双说话人混合语音 来自 WSJ0； WHAM/WHAMR 添加真实 噪声
LibriMix[b]	N	–	✓	–	–	–	–	合成语音混合 来自 LibriSpeech +噪声
DNS / VoiceBank+DEMAND[c]	M	✓	–	–	–	–	–	噪声语音来自 VoiceBank、DNS 挑战赛中的噪声音频 （合成数据 +一些真实数据）
FUSS[d]	N	–	–	✓	–	–	–	来自 多种声音事件音轨的合成混合音频
MUSDB18 / MDX[e]	M	–	–	–	–	✓	–	录音室多轨音乐； 人声/鼓/贝斯/其他 音轨
Slakh2100[f]	N	–	–	–	–	✓	–	完全合成 MIDI 渲染的多轨 音乐
MedleyDB / URMP[g]	M	–	–	–	–	✓	–	真实录音室多轨 合奏 ，乐器音轨
AudioSep（AudioSet / AudioCaps / Clotho 混合音频）[h]	N	✓	–	✓	✓	✓	T	合成混合音频来自 AudioSet、AudioCaps， Clotho 等。
DCASE 语言查询源 分离。[i]	M	–	–	✓	–	–	T	合成混合音频来自 带字幕/标注的声音 事件（例如， FSD50K/AudioSet 风格）
基于 CLAP 的文本查询 分隔符。[j]	N	–	–	✓	–	–	T	来自 AudioSet / FSD50K 风格的合成混合音频 数据集
基于 AVSpeech 的视听语音 分离。[k]	M	✓	✓	–	–	–	V	合成双说话人混合语音， 来源于 AVSpeech、 LRS2/3、VoxCeleb
音乐[l]	M	–	–	–	–	✓	V	真实在线演出 视频；合成数据或精选的 多乐器混合音频
AVSBench / LU-AVS[m]	M	–	–	✓	–	✓	V,S	真实的类 YouTube 视频 带有物体/乐器 声音、音视频掩码/片段
SAM Audio-Bench（我们的）	**Y**	✓	✓	✓	✓	✓	**T,V,S**	**来自真实野外音视频** **AudioSet ，VGGSound ，音乐 ，AVSpeech、CondensedMovies**

[a]WSJ0-2mix；WHAM!/WHAMR!（Hershey 等人，2016年；Wichern 等人，2019年；Maciejewski 等人，2020年）。b LibriMix（Cosentino 等人，2020年）。c VoiceBank+DEMAND；DNS 挑战赛数据集（Valentini-Botinhao，2017年；Reddy 等人，2021a）。d FUSS（Wisdom 等人，2021年）。e MUSDB18/HQ；SiSEC/MDX（Rafii 等人，2017年；Mitsufuji 等人，2022年；Fabbro 等人，2024年）。f Slakh2100（Manilow 等人，2019年）。g MedleyDB；URMP（Bittner 等人，2014年；Li 等人，2018年）。[h]基于 AudioSet、AudioCaps、Clotho 等混合音频构建的 AudioSep 基准（Liu 等，2022；Kim 等人，2019年；Drossos 等人，2020年；Gemmeke 等人，2017年）。i DCASE 语言查询源分离基准（例如，DCASE 2024 Task 8）。[j]基于 CLAP 的文本查询分离基准（Ma 等，2024）。[k]在 AVSpeech、LRS2/LRS3、VoxCeleb 上进行的音视频语音分离（Ephrat 等人，2018年；Afouras 等人，2018年；Chung 等人，2018年）。l MUSIC 是典型的音视频乐器分离数据集，Zhao 等人（2018年）曾使用该数据集。m AVSBench；LU-AVS（Zhou 等人，2022年；Liu 等人，2024a）。缩写：真实 Y/N/M；语音清理；说话人分离；通用开放域／通用声音分离；音乐清理；乐器音轨分离；提示：T文本、V视觉、S时间跨度。

此外，现有研究很少探讨音频分离任务本身的难度。以往大多数研究仅关注评估模型输出，却未系统分析重叠声源的数量、响度不平衡或声音之间的声学相似度等内在因素如何影响人类对任务难度的感知。了解分离难度至关重要，因为这有助于深入认识当前模型的局限性和鲁棒性，促进训练与评估的课程设计，并支持真实应用中的难度感知基准测试和自适应模型选择。

为更好地表征音频分离模型的性能和音频分离任务的内在难度，我们提出了一套新的人工标注指南：**SAM 音频评判器（SAJ）**，其中定义了九个感知维度。

SAJ 的性能维度用于评估模型分离目标声音的效果：

- **召回率**：提取音频是否包含提示中指定的所有目标声音？
- **精确率**：模型从提取音频中去除非目标声音的效果如何？
- **忠实度**：对于提取音频中包含的目标声音，它们听起来与原始混合音频中的对应声音有多相似？
- **总体质量**：模型输出的总体感知质量如何？

此外，SAJ 难度维度用于评估分离任务本身的复杂程度：

- **计数**：源音频中有多少个非目标声音？
- **重叠程度**：目标声音与非目标声音的重叠程度如何？
- **响度**：目标声音相对于非目标声音有多响？
- **混淆度**：非目标声音有多容易被误认为是目标声音？
- **总体难度**：综合考虑以上所有因素，从混合音频中提取目标声音有多困难？

5.2.2 数据标注

基于上述定义，我们设计了一项数据标注任务，以收集 SAJ 数据，由人工评分员使用五点李克特量表（1–5）对九个维度进行评估。我们重点关注**文本提示**SAJ，这是最常见的场景。我们提供了全面的标注指南，对各个维度作出详细说明，并明确评估时需要考虑的细粒度方面。附录 C.1展示了标注界面的示例。为帮助评分者校准判断，指南还配有大量音频示例和评分参考，说明各维度上高分或低分的判定标准。我们还设计了评分员资格认证计划，以确保选出高质量的标注员。详情见附录C.2。

划分	模态	时长	样本
训练集	语音	59.31 小时	13,149
	音乐	133.64 小时	26,101
	声音	117.52 小时	37,444
测试集	语音	6.38 小时	2,311
	音乐	9.32 小时	3,367
	声音	31.72 小时	11,476

表 6 SAM 音频评判器数据集中的音频样本及其时长

配对数据准备。我们收集了一套涵盖音乐、语音和音效的全面数据集。为缓解真实数据与模拟数据之间的不匹配，我们使用真实混合音频和合成混合音频作为输入音频。表6 展示了详细的数据统计信息。对于每个数据源，我们采用原始声音标注（例如狗叫声、男性说话等），或采用 PLM-Audio 生成的声音类型预测作为文本提示，具体说明见第4.2.1节。我们采用与 SAM Audio 模型相同的文本提示设置

，包括语音提取、说话人提取、音乐提取、乐器分离和通用声音事件提取。遵循（Hai 等人，2024年；Wang 等人，2025a），文本提示采用单一声音描述的形式，可以指代特定声源（例如狗叫声、吉他），也可以指代涵盖多个声源的通用声音类别（例如狗、播放的音乐）。对于每种模态，我们收集了多种公开可用的音频分离模型以及中间版 SAM Audio 检查点的输出，相关内容列于表7。为建立一个适用于不同音频模态、经过校准的统一 SAJ 分数，我们在标注过程中打乱了所有模态的音频样本。我们还进行了响度归一化，以消除音频音量变化可能带来的混杂影响。最后，我们为每个音频样本收集了三份独立评分，以减少方差并提高可靠性。

模态	音频分离模型
语音	SAM Audio、MossFormer2、Tiger、FastGeCo
音乐	SAM Audio、AudioSep、FlowSep、ClapSep、SoloAudio、Demucs、Spleeter
声音	SAM Audio、AudioSep、FlowSep、ClapSep、SoloAudio

表 7 SAM Audio Judge 数据集所用的模型输出

5.2.3 SAJ 模型训练

我们的主要 SAJ 模型旨在预测分离模型性能，因此我们仅使用性能维度上的标注。如图6所示，SAJ 模型接收三种输入：输入音频、输出音频和文本提示。我们使用预训练音频编码器提取音频特征，并使用文本编码器提取文本特征。这两个编码器均采用自 PE-AV（Vyas 等人，2025年），该模型通过大规模视频-音频-文本对比学习进行训练。

文本特征在时间上进行对齐，以匹配音频特征的长度，然后输入 Transformer 提取联合多模态表示。随后应用多个线性层来预测 SAJ 分数，包括召回率、精确率、忠实度、总体评分等。

此外，我们发现，引入一项代理任务来预测输出音频是否符合文本提示（Wang 等人，2022年b，a），能够显著提升模型性能。为此，我们使用一个大规模模拟数据集对整个模型进行预训练，以完成这项文本—音频对齐检测任务；该数据集提供了混合音频中各分离音轨的数据。我们交替将目标声音和一段

从同一混合音频中随机选取非目标声音，用以表示目标声音是否存在。我们使用一个额外的线性层，预测文本提示所描述的目标声音是否存在。完成这一预训练阶段后，我们对 SAJ 模型进行微调，以预测最终的 SAJ 分数。

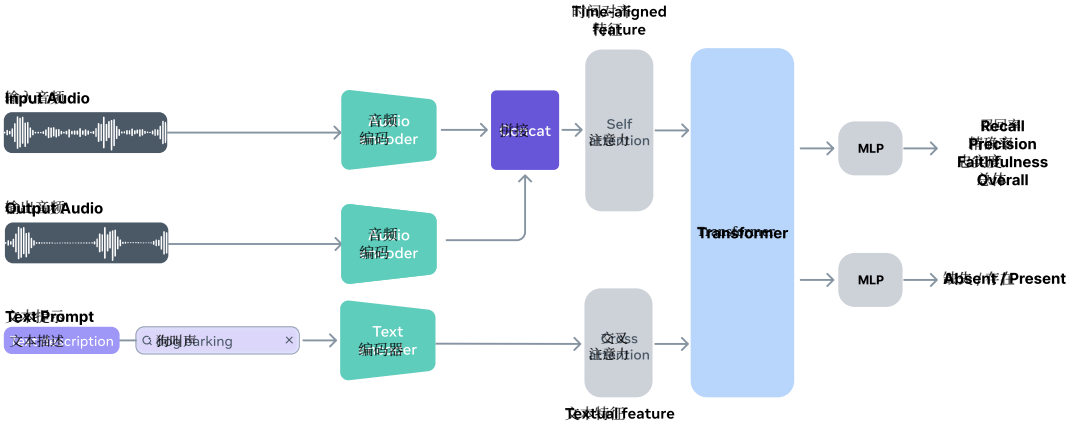

图 6 SAM 音频评判器模型示意图 el

5.3 主观评估协议

鉴于客观指标在真实场景内容上的局限性，且目前缺乏可靠、通用的分离质量无参考度量，我们的基准测试采用人工评估作为主要方法。我们使用**并排绝对类别评分（ACR）**协议，并始终设置偏好决胜项。这种混合方法既能提供绝对质量信号，也能进行稳健的相对比较，同时减少评估者常见的偏差。

简而言之，标注员会看到源音频（以及适用时的视频）、用户提示和两个模型的提取输出，并需要判断每个输出在多大程度上呈现了所请求的目标声音。该协议首先确认目标声音是否确实出现在源音频中，随后评估提取音频中包含了多少目标内容、是否有部分内容缺失，以及提取出的目标声音与原声有多相似。标注员还会评估非目标声音是否存在及其程度，包括判断这些声音是源自源音频，还是模型引入的伪影。回答完这些结构化问题后，评分者会给出 1–5 分的总体得分，以反映输出对提示的忠实度和声学保真度；如果模型得分相同，则需在两个输出之间强制选择其偏好。此流程一致适用于文本提示、视觉提示和片段提示条件。

从实证结果来看，与其他评估协议相比，这种并列 ACR 框架结合始终启用的偏好决胜项，具有明显优势。偏好决胜项提升了标注者间一致性，并产生区分度更高、更具判别力且与表达出的偏好相符的 ACR 差值。并列呈现也降低了评分差异的不确定性：与单刺激评分相比，置信区间最多缩窄 20%，这意味着在 A/B 敏感度相当的情况下，成本约可节省 30%。尽管与仅进行成对偏好评估相比，该协议提高了每个项目的处理时间，但它能在一次评估中同时得出绝对判断和相对判断。最后，与先前研究中报告的锚定效应一致，绝对 ACR 评分仍取决于上下文。

对于每一对系统，主观评估都会针对总体质量、覆盖度、正确性和忠实度这四个维度，给出系统 A 相对于系统 B 的绝对评分（OVR）和净胜率（NWR）。在主要结果中，我们主要报告总体 OVR 和 NWR。

主观评估协议的完整细节见附录 A.1 和 A.2。

5.4 客观指标

文本提示。 对于文本提示分离，我们使用 SAJ 和 CLAP（Wu 等人，2023年）评分[1]来评估模型性能。SAJ 衡量多个维度上的分离质量，为简洁起见，我们报告其总体得分。CLAP 评估分离音频与输入文本提示之间的语义对齐情况，作为感知正确性的替代指标。

跨度提示。 对于时间段提示分离，我们复用与目标区域相关联的真实文本描述，从而能够使用相同的 SAJ 和 CLAP 评分。此外，我们引入 **SpanIoU** 指标，通过计算预测片段与参考片段之间的交并比（IoU）来量化时间对齐情况。预测时间段通过对分离波形应用简单的语音活动检测（VAD）获得，使用的是`pydub` 库中的 pydub。

视觉提示。 对于视觉提示分离，CLAP 等基于文本的指标效果较差，因为文本描述通常不足以区分视觉定位区域。相反，我们使用 ImageBind（Girdhar 等人，2023年）评分来衡量视听一致性，该评分反映了分离音频与对应视觉掩码区域之间的一致性。

6 实验设置

6.1 训练和推理配置

我们训练了三种SAM AUDIO 变体，其参数预算分别为 500M、1B 和 3B（表 8）。参数量不包括 PE 视觉编码器、T5 文本编码器和 DAC-VAE 音频编解码器。辅助

[1]我们使用以下检查点：https://huggingface.co/lukewys/laion_clap/blob/main/630k-best.pt。

对于 500M、1B 和 3B 模型，辅助对齐损失分别注入第 2、4 和 6 个 Transformer 模块。

所有模型均按照Polyak 等人 (2024年）提出的两阶段方案进行优化：预训练阶段之后，接着是在精选高质量数据上进行微调阶段。模型在表2所列的通用视频数据生成的合成音频混合物上进行预训练。其余高质量数据以及所有伪标注数据仅用于微调。由于许多训练音频超过了常见的上下文长度，所有片段都限制为 30 秒，较长的片段则会随机分块。训练采用全分片数据并行，以适配模型规模。

模型	总参数量	层	注意力维度	前馈网络维度
SAM Audio-Small	500M	12	1,536	6,144
SAM Audio-Base	1B	16	2,048	8,192
SAM Audio-Large	3B	22	2,816	11,264

表 8 SAM Audio 模型配置。

预训练。 我们采用有效批次大小为 1,024 个序列，每个序列截断或填充至 30 秒（= 750 音频令牌）。模型经过 500K 次更新训练，使用固定学习率 1×10^{-4}，并在此之前进行 5K 步线性预热。整个训练过程中均使用 AdamW，权重衰减为 0.1，精度为 bf16。预训练期间启用辅助对齐损失，权重设为 1。此外，我们采用条件丢弃：音频混合、视频和文本提示分别以 0.3 的概率独立丢弃。

微调。 由于微调数据的长度变化较大，我们采用可变长度批处理，并根据模型规模分别将每批令牌预算设为 {96K, 120K, 144K} 个令牌，对应 {500M, 1B, 3B} 模型。微调进行 300K 步，先以 5K 步预热至峰值学习率 1×10^{-4}，之后保持恒定。我们维护指数移动平均（EMA）检查点（衰减率为 0.999），并将其用于推理。微调期间禁用辅助损失（权重为 0），因为我们观察到，在干净输出上训练并无收益。微调期间，我们也禁用条件丢弃。

推理。我们采用 16 步中点 ODE 求解器，不使用无分类器引导，因为 CFG 在我们的设置中没有带来改进。此外，我们采用束宽为 8 的候选项重排序：对于文本提示，使用 SAM 音频评判器和 CLAP 评分的线性组合（权重分别为 1 和 5）；对于跨度提示，使用片段 IoU；对于视觉提示，使用音频与掩码区域之间的 ImageBind 相似度。默认情况下，我们为仅文本分离启用跨度预测，因为这能提升总体性能（见第7.4节）。具体而言，我们使用 PE-A 帧模型（Vyas 等人，2025年）根据文本提示估计与目标声音对应的时间跨度。然后，将预测出的时间跨度与原始文本提示结合，并用作 SAM Audio 的条件信息。我们将帧概率阈值设为 0。3，与 Vyas 等人 （ 2025 年）中的阈值设置一致。

由于 SAM Audio-Large 在三个变体中表现最佳，因此我们在所有关键比较和分析中均采用 SAM Audio-Large；除非另有说明，本文后续将其简称为 SAM Audio。

6.2 任务

我们对SAM Audio 进行基准测试，涵盖多种分离任务，这些任务反映了第5节中提到的常见现实应用场景。每项任务都用于评估模型单独或联合利用文本、视频和片段提示的能力。

文本提示任务 。我们在广泛的文本提示源分离任务上评估 SAM Audio s:

1. **通用声音事件提取：**根据文本查询（例如“狗叫声”或“玻璃破碎”）提取任意目标声音事件。此任务用于测试模型处理通用音频事件提取的能力。

2. **语音提取：**从噪声音频混合中分离出所有语音。说话人数不受限制，目标是不论说话人数或背景噪声如何，都能恢复完整的语音轨道。3. **说话人提取：**根据描述说话人性别或年龄等属性的文本提示（例如“女性说话”），分离出特定说话人的语音。此任务用于评估细粒度的说话人条件控制。我们主要关注基于性别的分离。4. **音乐提取：**从混合音频中分离音乐，包括音乐伴有语音或音效的情况。5. **真实场景乐器分离：**从完整音乐音频中提取单个乐器音轨（例如“钢琴”“鼓”）。此基准测试涵盖干净的录音室录音和真实场景中的噪声音乐混合音频。6. **专业乐器分离：**从专业录制的音乐（例如录音室音轨）中分离音轨。与通用乐器分离基准测试不同，此任务侧重于高质量的多轨录音，并采用 MUSDB18（Rafii 等人，2017年）作为标准基准测试。

视觉提示任务。 视觉提示由一对视频掩码和原始视频片段组成。这些任务用于评估模型在视觉信息的条件下执行实例级声源分离的能力：

1. **通用声音事件分离：** 提取与高亮感兴趣区域相关联的声音（例如，分离汽车鸣笛声）。2. **乐器分离：** 在音乐视频中，提取与高亮掩码对应的乐器音频，无论混合音频是否嘈杂。3. **说话人分离：** 在多说话人视频中，分离高亮说话人的语音。

时间跨度提示任务。 时间跨度提示任务与文本提示任务相对应。模型不再以文本描述为条件，而是以指定目标声源活跃时间的时间跨度为条件。

在每种模态中，我们会过滤掉提示存在歧义、可能对应多个合理声源的样本（例如，包含多个视觉上无法区分的乐器的视频，或时间上无法分辨的重叠事件）。评估集中的每个样本均为 $\sim 10s$，以便进行主观评估。由于 MUSDB 中的原始音乐长达数分钟，我们会随机截取 10 秒的原始完整混合音频片段，用于专业乐器分离。

6.3 基线方法

我们将 SAM Audio 与涵盖所有提示方式的多种基线方法进行比较。为确保公平比较，我们针对每个模型所设计的任务进行评估，并在可用时采用其原生输入格式。

6.3.1 文本提示基线方法

现有关于文本引导的音频分离的研究，大多聚焦于通用音频事件，并将语音和音乐等宽泛类别视为单一类别。我们选取了四种具有代表性且较新的开放领域基线方法进行比较：AudioSep（Liu 等人，2022年）、FlowSep（Yuan 等人，2025年）、SoloAudio（Wang 等人，2025a）和 CLAPSep（Ma 等人，2024年）。这些模型旨在处理通用音频，并基于大规模音频文本数据集进行训练。尽管它们并非针对特定领域专门设计，但其预训练数据通常包含语音和音乐样本，因此可以在通用和专业任务上与SAM Audio 进行比较。

除这些开放域系统外，我们还纳入了一系列针对特定音频领域的专业基线。与通用模型不同，这些系统不支持自由文本提示；相反，它们会将混合音频分解为由固定本体定义的音轨（例如人声、鼓和贝斯）。为确保公平比较，我们提取与目标事件类型对应的分离音轨，并将其与 SAM Audio 进行评估。以下介绍这些专业基线。

乐器分离我们将其与 Demucs（Rouard 等人，2023年）、Spleeter（Hennequin 等人，2020年）、AudioShake（AudioShake，2025年）、MoisesAI（Moises.AI，2025年）、LalalAI（Lalal.AI，2025年）和 FADR（FADR，2025年）进行评估。Demucs 和 Spleeter 是公开可用的模型，其他则是通过公共 API 使用的专有系统。所有基线模型支持的音轨种类都很有限（见表9）。在专业乐器分离基准测试中，我们采用 MUSDB（Rafii 等人，2017年），该基准仅包含人声、鼓和贝斯分离。相比之下，在真实场景的乐器分离任务中，我们不采用 Demucs 和 Spleeter，因为它们支持的音轨词汇仅涵盖所需乐器中的一小部分。对于其余基线，我们仅在目标乐器属于其支持列表的示例上进行比较。

模型	支持的乐器
Demucs（Rouard 等人，2023年）	人声、鼓、贝斯
Spleeter（Hennequin 等人，2020年）	人声、鼓、贝斯、钢琴
AudioShake（Audioshake，2025年）	人声、鼓、吉他、贝斯、风声、钢琴
MoisesAI（Moises.AI，2025年）	人声、吉他、贝斯、鼓、钢琴、风声、弦乐器
LalalAI（Lalal.AI，2025年）	人声、鼓、贝斯、吉他、钢琴、合成器、弦乐器、风声
FADR（FADR，2025年）	人声、鼓、钢琴、吉他、弦乐器、风声
SAM Audio	开放词汇

表 9 各基线方法支持的乐器类型。

语音分离 语音分离旨在去除语音录音中的背景噪声。我们将SAM Audio 与 LalaAI（Lalal.AI，2025年）、ElevenLabs（ElevenLabs，2025年）、Auphonic（Auphonic，2025年）和 AudioShake（Audioshake，2025年）进行比较。为确保比较公平，我们禁用基线方法中的后处理或增强模块，因为我们的评估关注的是分离保真度——即模型分离语音内容的能力——而非感知增强或混响抑制。

音乐分离 对于音乐分离，我们以 MoisesAI (Moises.AI，2025年) 和 AudioShake (Au-dioshake，2025年) 为基准进行评测。这两者都能隔离或移除背景音乐。与语音分离类似，我们禁用任何额外的增强模块。我们分别报告音乐提取 (隔离音乐) 和音乐移除 (移除音乐，同时保留其他声音)的结果。

说话人分离 说话人分离旨在隔离特定个体的语音。大多数专用系统依赖固定音轨分解，且原生不支持通过提示来指定任意说话人。为便于在提示设置下进行比较，我们选择 CLAP 相似度分数最高的分离音轨 (Wu 等人，2023年)。我们评测 Mossformer2 (Zhao 等人，2024年)、Tiger (Xu 等人，2024年) 和 FastGeCo (Wang 等人，2024年)，以及专有的 AudioShake 模型 (Audioshake，2025年)。

6.3.2 视觉提示基线

与文本引导的分离相比，视觉提示音频分离的研究要少得多。据我们所知，目前没有可用的商业模型，因此我们重点关注公开的研究模型。我们在通用声音、说话人和乐器分离任务上评估SAM Audio与两个通用视觉分离模型 DAVIS-Flow（Huang 等人，2025年）和 CLIPSep（Dong 等人，2023年）的表现。此外，对于视觉提示乐器分离基准测试，我们还纳入了 DAVIS-Flow 的乐器专用检查点[2] （记为 *DAVIS-Flow(Music)*）。

在视觉分离任务中，说话人分离得到了更广泛的研究（Wu 等人，2019年；Pan 等人，2025年；Li 等人，2024a，b）。因此，我们将其与两个强大的基线方法 IIANet（Li 等人，2024b）和 AV-Mossformer2（Zhao 等人，2025年）进行比较；这两个模型在近期的视觉提示语音分离基准测试（ClearVoice，2025年）上取得了最先进结果。这些模型依赖包含人脸检测和唇部运动提取的预处理流程；我们遵循其官方预处理设置，并相应输入掩码视频。实际上，我们的评估视频中约有 20% 在 AV-Mossformer2 预处理阶段失败，因此这些样本不纳入与SAM Audio的比较。

[2]我们使用 DAVIS-F 分别在 AVE 和 MUSIC 上训练、用于通用分离和乐器特定分离的低检查点

6.3.3 跨度提示基线

据我们所知，目前没有支持片段提示的公开分离模型。因此，我们将自己的文本提示模型用作基线，将仅片段条件信息和文本+片段联合条件信息与仅文本分离进行比较，以量化片段提示带来的益处。

7 SAM Audio 结果

本节将介绍SAM Audio的主要结果，涵盖文本、视觉和跨度提示、声音移除、模型延迟以及长音频分离。详细的消融研究将在附录D中介绍。

7.1 文本提示分离

			通用音效			语音			说话人			音乐			野外乐器			专业乐器		
模型	可提示的	开放源代码软件	SAJ	CLA		SAJ	CLAP	O	J	CLAP	OVR	AP	OVR	SAJ	OVR	SAJ	CLA			
MossFormer2（Zhao 等人，2024年）	✓	✗	-	-	-	-	-	-	2.43	0.14	2.54	-	-	-	-	-	-	-	-	-
Tiger（Xu 等人，2024年）	✓	✗	-	-	-	-	-	-	2.47	0.15	2.50	-	-	-	-	-	-	-	-	-
Fast-GeCo（Wang 等人，2024年）	✓	✗	-	-	-	-	-	-	2.66	0.16	2.71	-	-	-	-	-	-	-	-	-
Demucs（Rouard 等人，2023年）	✓	✗	-	-	-	-	-	-	-	-	-	-	-	-	-	-		4.48	0.15	4.26
Spleeter（Hennequin 等人，2020年）	✓	✗	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4.26	0.11	3.90
FlowSep（Yuan 等人，2025年）	✓	✓	2.36	0.21	2.65	2.18	0.20	2.14	1.85	0.09	2.13	2.73	0.18	2.90	2.37	0.10	2.69	2.13	-0.01	2.02
AudioSep(Liu 等人，2022年)	✓	✓	2.63	0.25	2.88	2.93	0.28	2.85	2.50	0.17	2.79	3.47	0.27	3.51	2.16	0.13	2.59	2.34	0.04	2.45
CLAPSep(Ma 等人,2024年)	✓	✓	2.68	0.23	2.92	2.30	0.22	2.47	2.80	0.17	2.79	2.48	0.04	2.97	2.47	0.14	2.81	2.48	0.04	2.56
SoloAudio（Wang 等人，2025a）	✓	✓	3.29	0.25	2.97	3.45	0.30	3.32	2.26	0.19	2.45	2.68	0.21	2.47	2.92	0.13	2.71	2.65	0.01	2.30
AudioShake（AudioShake，2025年）	✗	✗	-	-	-	3.90	0.28	3.95	3.28	0.14	3.51	3.22	0.29	3.37	3.37	0.29	3.43	3.87	0.29	4.28
MoisesAI（Moises.AI，2025年）	✗	✗	-	-		-	-		-	-	-	3.79	0.27	3.90	3.03	0.29	3.12	3.78	0.28	4.22
FADR（FADR，2025年）	✗	✗	-	-	-	-	-	-	-	-	-	-	-	-	2.44	0.19	2.45	3.63	0.25	3.92
LalalAI（Lalal.AI，2025年）	✗	✗	-	-	-	3.77	0.33	3.92	-	-	-	-	-	-	3.07	0.25	3.03	3.83	0.27	4.18
Auphonic（Auphonic，2025年）	✗	✗	-	-	-	4.32	0.27	4.08	-	-	-	-	-	-	-	-	-	-	-	-
ElevenLabs（ElevenLabs，2025年）	✗	✗	-	-	-	3.79	0.25	3.72	-	-	-	-	-	-	-	-	-	-	-	-
SAM Audio	✓	✓	**4.35**	**0.31**	**3.59**	**4.67**	**0.35**	**4.29**	**4.51**	**0.18**	**4.15**	**4.45**	**0.26**	**4.05**	**4.32**	**0.31**	**4.00**	**4.82**	**0.28**	**4.45**

表10 C　　与文本提示基线的比较。–：不适用。总体主观评分：总体主观评分　　e.

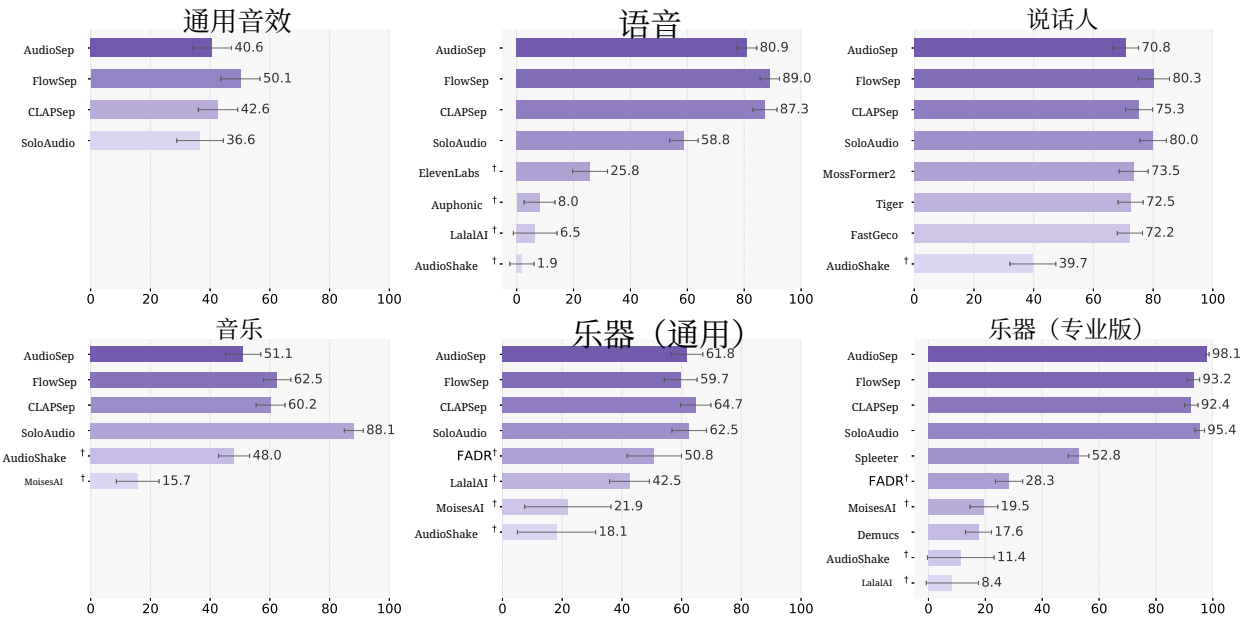

图7 SAM Audio 在文本提示任务中相较于最先进分离模型的净胜率（%）。 †：专有模型

表 10展示了 SAM Audio 与多种公开文本提示分离模型之间的定量比较。基线方法分为两大类：旨在处理各种分离任务的通用模型（如 AudioSep、FlowSep、SoloAudio 和 CLAPSep），以及专门针对特定领域

优化的模型，例如语音或音乐领域（如 MossFormer2、Demucs、Spleeter）。我们还在图 7中展示了相对于基线方法的净胜率。在表 10中，除了客观指标外，我们还展示了总体主观评分（即 OVR）。由于我们的评估协议采用成对比较设置，最终的 OVR 分数通过对所有涉及 SAM AUDIO 的成对比较的总体偏好得分取平均值计算得出。

总体而言，SAM Audio 在几乎所有类别中都持续大幅超越此前的所有模型。在通用声音事件分离方面，SAM Audio 相较于表现最佳的公开通用声音分离模型之一 SoloAudio（Wang 等，2025a），净胜率约为 $\sim 36\%$。在乐器分离或说话人分离等专业领域，我们发现 FlowSep（Yuan 等，2025）或 AudioSep（Liu 等，2022）等通用模型无法达到有竞争力的质量。它们的平均评判分数仍低于 3，明显落后于 Demucs（Rouard 等，2023）等专用系统。然而，即便在这些领域，SAM Audio 也优于专用模型（例如，SAM Audio 相较于 Demucs 的净胜率为 $= 17.6\%$）。虽然专有系统通常优于 OSS 对应系统，但 SAM Audio 在大多数设置下仍表现更出色。在面向专业音频的乐器分离任务（MUSDB（Rafii 等，2017））中，SAM Audio 的总体主观评分为 4.45，高于 AudioShake（Audioshake，2025）的 4.28（相对提升 $\sim 4\%$）；在说话人分离任务中，SAM Audio 的评分为 4.15，而 AudioShake 为 3.51，对应的净胜率提升为 $\sim 39\%$。通过统一训练，SAM Audio 能够跨领域泛化，并实现最先进性能。

7.2 视觉提示分离

		通用音效		说话人		指令（野外）	
模型	**通用**	IB	OVR	IB	OVR	IB	OVR
AV-MossFormer2（Zhao 等人，2025年）	✗	-	-	0.20	2.62	-	-
IIANet（Li 等人，2024b）	✗	-	-	0.16	2.41	-	-
ClipSep（Dong 等人，2023年）	✓	0.16	1.53	0.14	1.47	0.15	1.12
DAVIS-Flow（Huang 等人，2025年）	✓	0.14	1.96	0.13	1.97	0.13	2.08
DAVIS-Flow（音乐）（Huang 等人，2025年）	✗	-	-	-	-	0.13	2.40
SAM Audio	✓	**0.25**	**2.61**	**0.24**	**3.07**	**0.24**	**2.56**

表 11 C　与视觉提示基线的比较。–：不适用。OVR：总体主观评分　re.

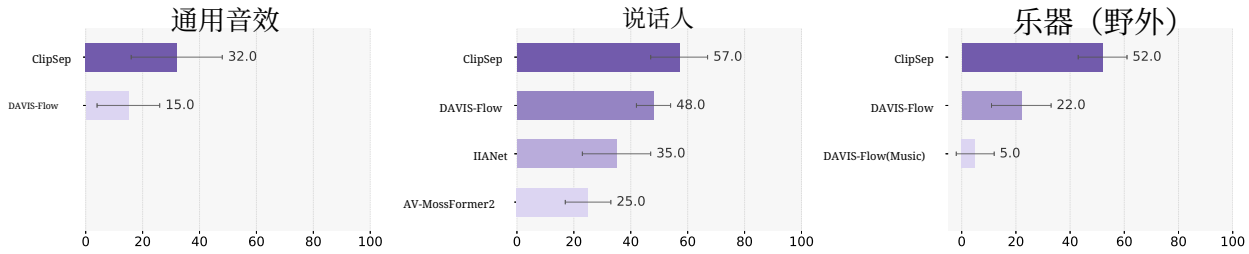

图 8 SAM Audio 在视觉提示任务中相较于最先进分离模型的净胜率（%）

我们在表 SAM AUDIO 与现有视觉提示分离模型之间进行比较，结果见表 11。与文本提示分离相比，目前公开可用的通用视觉分离系统要少得多。

在所有设置下，SAM Audio 相较于以往工作都取得了更大的提升。平均而言，SAM Audio 大幅优于 DAVIS；根据分离任务的不同，其净胜率为 5% 至 48%。与文本提示分离中观察到的趋势相似，我们发现，通用视觉模型在乐器分离和说话人分离等专业领域表现不佳。相比之下，SAM Audio 不仅保持领先，还以约 25% 的优势超越说话人分离领域最优秀的专业基线，并以约 5% 的优势超越乐器分离领域最优秀的专业基线。

此外，我们注意到，与文本提示相比，视觉提示的总体主观评分明显较低。文本提示的优势主要源于其拥有规模大得多的高质量文本训练数据，而基于视频的监督通常噪声更大，这是因为

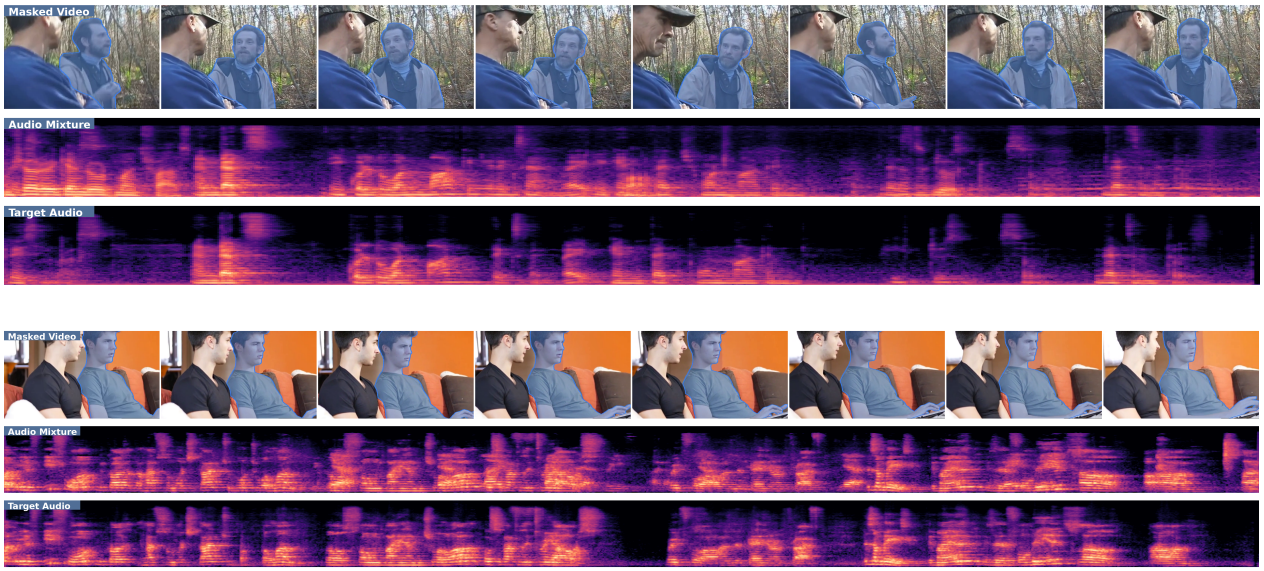

图9 文本描述存在歧义时的视觉提示样本。每个示例展示了（**顶部**）掩码视频、（**中部**）输入混合频谱图，以及（**底部**）分离后的目标频谱图。每段视频中均突出显示目标说话人。

训练中的视觉掩码存在错误，且频繁出现画外声音。除了规模优势外，文本通常也能为目标声源提供更具体的线索。例如，“男人拖着脚步走”这样的提示直接指向一个独特的声音事件，而人物的视觉掩码本质上具有歧义，因为它可能对应多种声音。

尽管如此，在仅靠文本不足以应对的场景中，视觉提示仍能发挥互补作用。一个常见的例子是对话场景：多个同性别的人同时说话。诸如“男性语音”这样的文本提示无法区分两位男性说话人，但视觉掩码能够定位目标说话人，从而有效地实现分离。在这类情况下，视觉线索能够提供实例级定位，而仅靠文本很难做到这一点（见图9）。

提示模态	通用			语音			说话人			音乐			野外乐器		
	SAJ CLAP	总体主		SAJ CLAP	总体主		AJ CLAP	总体主观		CLAP	总体主观评		LAP	总体主观评分	
text	4.11	0.31	3.32	4.59	0.33	4.18	4.08	0.17	3.63	4.30	0.28	4.09	**4.45**	0.30	3.78
span	3.27	0.30	3.54	3.37	0.28	3.65	3.26	0.26	4.08	2.18	-0.04	2.57	1.81	0.01	2.20
文本 + 时间跨度	**4.25**	**0.31**	**4.04**	**4.66**	**0.35**	**4.33**	**4.51**	**0.18**	**4.22**	**4.38**	**0.27**	**4.19**	4.42	**0.32**	**3.88**

表12 跨 度提示与文本提示在 SAM Audio 中的比较。总体主观评分：总体主观评分。对于以下所有指标，数值越高越好。

7.3 时间段提示分离

表 12 和图 11 比较了以文本、时间段以及文本和时间段+作为输入的条件模型。为了更好地消融真实时间段的影响，文本提示中不使用预测时间段。单独使用时间段提示并不能稳定提升性能，因为目标音频和干扰声音可能在同一时间区域内同时出现。这种影响在语音和音乐等时长较长或属于环境声音的音频中尤为明显，仅使用时间段提示的模型性能大幅下降（NWR 为 -16%至 -49.6%，相较于仅使用文本的基线）。相反，对于时长较短且定位明确的声音，时间线索的区分度更高，仅使用时间段提示的条件输入可带来明显提升（声音的 NWR 比仅使用文本高 $+26\%$）。

尽管存在这些波动，结合文本与时间段输入仍能持续提升各领域的性能，NWR 达到 $+12.9\%$ 至 $+39.0\%$。这些结果表明，片段提示提供的时间定位能够补充文本中的语义信息；同时使用两者可以实现更精准的分离。

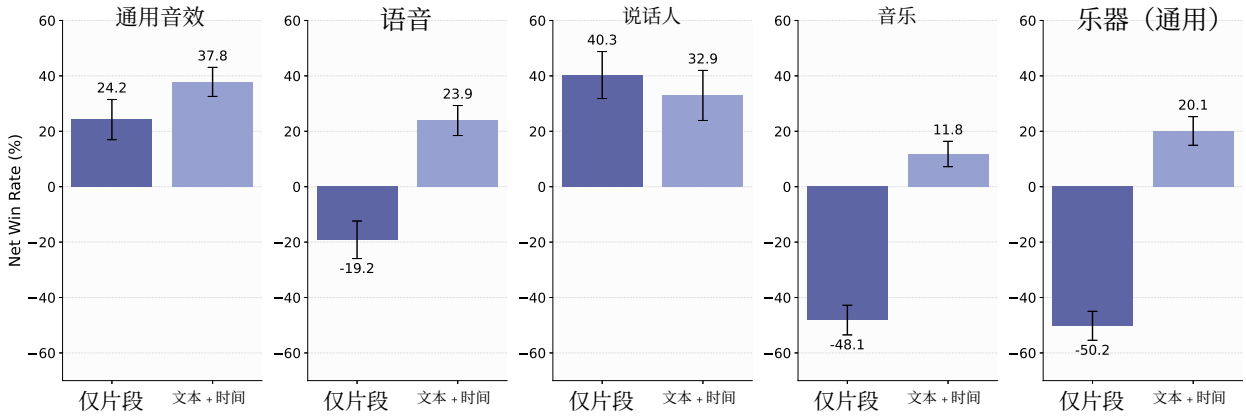

图 10 SAM Audio 使用文本与时间跨度 / 仅时间跨度作为输入时，相较于纯文本模型的净胜率（%）

7.4 跨度预测提升文本提示分离效果

在文本提示中，我们默认将预测时间段作为额外输入。这里，我们比较了在所有分离任务中使用与不使用跨度预测的效果。比较结果见表 13。请注意，文本 + 预测时间段 的 OVR 分数与表10中的不同，因为在我们的人工评估协议下，每次成对评估的绝对评分取决于所使用的具体基线。

引入预测时间段可提升大多数领域的性能，包括通用音效、语音、说话人和音乐；在这些领域中，时间线索对于消除目标声音的歧义至关重要。我们发现，专业乐器分离的性能略有下降，这可能是因为 MUSDB 中许多乐器音轨覆盖整个片段，时间线索能够带来的额外收益有限。使用预测片段与真实片段之间的差距很小（见表 12），这表明SAM Audio 对片段估计中的时间误差具有很强的鲁棒性。重要的是，跨度预测无需耗费成本进行人工标注，就能提升分离质量，使SAM Audio 能够大规模实现精准分离。

Task	带时间跨度预测	SAJ	CLAP	OVR
通用音效	✗	4.11	**0.31**	3.36
	✓	**4.35**	**0.31**	**3.89**
语音	✗	4.59	0.33	4.17
	✓	**4.67**	**0.35**	**4.22**
说话人	✗	4.08	0.17	3.62
	✓	**4.51**	**0.18**	**4.01**
音乐	✗	4.30	**0.28**	**4.16**
	✓	**4.45**	0.26	4.12
野外乐器	✗	**4.45**	0.30	3.70
	✓	4.32	**0.31**	**3.88**
专业乐器	✗	**4.83**	**0.28**	**4.16**
	✓	4.82	**0.28**	4.12

表 13 文本提示中使用与不使用预测时间段的对比。总体主观评分：总体主观评分。对于以下所有指标，数值越高越好。

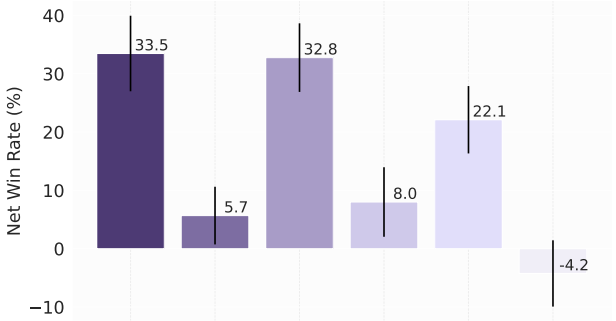

图11 SAM Audio 使用预测时间段与不使用预测时间段进行文本提示时的净胜率（%）。

7.5 声音移除

SAM Audio 会同时输出<拆分 id='1'>目标音频 </拆分>和<拆分 id='3'>残差信号 </拆分>音频。前面的章节重点评估目标音频的质量，现在我们评估残差信号，对应于<拆分 id='5'>移除 </拆分>任务——从混合音频中移除提示指定的声音。我们以文本提示音乐移除作为具有代表性的移除任务。

<拆分 id='1'>在现有基线方法中，只有 MoisesAI（</拆分><拆分 id='3'>Moises.AI</拆分><拆分 id='5'>，</拆分><拆分 id='7'>2025年</拆分><拆分 id='9'>）和 AudioShake（</拆分><拆分 id='11'>Audioshake</拆分><拆分 id='13'>，</拆分><拆分 id='15'>2025年</拆分><拆分 id='17'>）支持显式声音移除。因此，我们将</拆分><拆分 id='19'>SAM Audio</拆分><拆分 id='21'>与这两个系统进行比较。如表</拆分><拆分 id='23'>14</拆分><拆分 id='25'>和图</拆分><拆分 id='27'>12</拆分><拆分 id='29'>所示，</拆分><拆分 id='31'>SAM Audio </拆分><拆分 id='33'>的表现优于这两个系统。其趋势与图</拆分><拆分 id='35'>7</拆分><拆分 id='37'>中的音乐提取结果相呼应，表明提取与移除模式具有很高的相关性。<拆分 id='39'>SAM Audio </拆分><拆分 id='41'>取得的高 OVR 分数进一步展现了该模型干净抑制目标声源的能力。</拆分>

模型	OVR
AudioShake（Audioshake，2025年）	3.75
MoisesAI(Moises.AI，2025年)	4.00
SAM Audio	**4.05**

表14 音乐移除与基线方法的比较

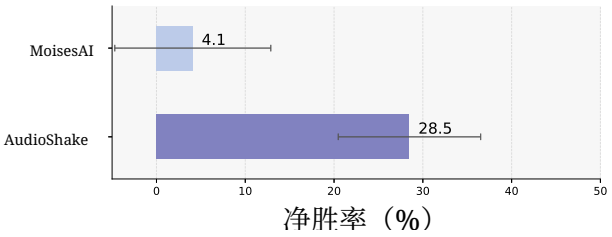

Figure12 SAM Audio 在音乐移除方面相较于基线方法的净胜率（%）

7.6 延迟

我们的默认推理配置采用 16 个常微分方程步数，并使用中点求解器。对于文本提示，在单个 A100 GPU 上，SAM Audio-Large 对 10 秒输入进行一次分离约需 7.3 秒，其中模型前向计算耗时 6.5 秒，片段提示耗时 0.1 秒，评判器重排序耗时 0.5 秒。

我们进一步通过调整常微分方程步数，研究推理成本与输出质量之间的权衡。如图 13 所示，正如预期，增加常微分方程步数通常能提升各项任务的性能。尽管如此，即使只有两步常微分方程步数，模型也能取得出乎意料的竞争力结果（例如语音分离）。从定性角度来看，说话人分离或乐器分离在 16 步与 2 步常微分方程步数下的性能差距更大。相比之下，对于通常较短且稀疏的通用音效，即使神经函数评估次数较少，仍能获得可接受的感知质量。我们推测，输入混合音频提供了强有力的条件信号，因此与完全生成式任务相比，分离所需的细化步骤更少例如文本转语音。总体而言，在许多实际的 SAM Audio 分离场景中，较少的神经函数评估次数能够在速度和质量之间实现良好的权衡。

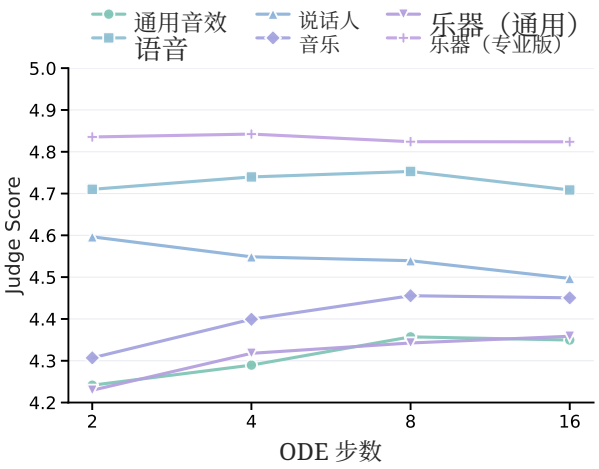

图 13 中点求解器下改变常微分方程步数的影响。减少步数可降低计算量，但会适度牺牲质量。

7.7 长篇音频分离

评估集中的大多数样本时长约为 10 秒。为进一步评估 SAM Audio 处理较长输入时的性能，我们采用第 3 节所述的多重扩散方法，评估其长音频分离性能。具体而言，我们采用 20 秒的窗口，并设置 5 秒的上下文重叠。作为比较，我们考虑两种基线方法：(a) **分块分离**，将音频划分为 20 秒的片段，分别处理后再拼接；(b) **单次分离**，在一次前向传递中处理整段音频。

我们整理了一个内部测试集，其中包含 50 个时长为 1 分钟的音频样本，用于评估长时程分离质量。如表 15 所示，单次处理模型在长录音上的性能明显下降，这与预期相符，因为大多数训练样本的时长都短于 30 秒。分块处理基线缓解了这一问题，但在片段边界处引入了可听见的不连续性。相比之下，所提出的多重扩散策略在整个序列中保持了较高的感知质量，并取得了最高的评判分数。

方法	SAJ	CLAP
单次处理	3.48	0.26
分块处理	3.57	0.24
多重扩散	**3.67**	**0.27**

表15 长音频分离方法对比。

7.8 SAM Audio 评估结果

7.8.1 主观评分相关性

我们将提出的 SAM 音频评判模型与几种涵盖不同评估范式的代表性基线系统进行比较：

- CLAP（Wu 等人，2023年）：一种大规模对比式音频文本模型，使用数百万个音频字幕对进行训练。我们将其文本提示与输出音频嵌入之间的余弦相似度作为感知对齐的替代指标。这代表了当前音频语言对应评估的标准方法。

- SDR Estimator（Dang 等人，2023年）：一种回归模型，无需真实参考即可估算信号失真比。它体现了传统基于失真的指标的性能，这类指标关注信号保真度，而非感知判断。我们使用与 SAJ 模型相同的架构训练 SDR 估算器，但将信号失真比值设为训练目标。我们使用均衡训练数据集，涵盖目标音频电平从 -25 分贝到 25 分贝，共计 496 小时，包含语音、音乐和音效。该 SDR 估算器在语音、音乐和音效上的 PPC 分数分别为 0.923、0.681 和 0.665。

- Gemini-2.5-pro（Comanici 等人，2025年）：一种能够对文本和音频输入进行推理的大型多模态大语言模型。我们提示它根据 SAJ 所采用的相同评估维度，对分离音频质量进行评分，以此作为基于语言模型的感知评估基线。我们还提供了多个具有不同评分的示例，以实现少样本学习。我们的提示见附录 C.3。

这些基线涵盖了基于信号处理的分离度量和通用基于多模态大语言模型的指标这两种主要范式，从而能够与我们面向特定任务的 SAJ 模型进行全面比较。我们报告自动指标与人工评分之间的皮尔逊相关系数（PCC）和斯皮尔曼秩相关系数（SRCC）。

	语音				音乐				声音			
模型	总体	召回率	精确率	忠实度	总体	召回率	精确率	忠实度	总体	召回率	精确率	忠实度
皮尔逊相关系数（PCC）												
CLAP	0.490	0.431	0.283	0.477	0.487	0.416	0.385	0.432	0.367	0.431	0.283	0.418
SDR Estimator	0.336	0.004	0.403	0.055	0.369	0.157	0.388	0.182	0.181	0.040	0.222	0.055
Gemini-2.5-pro	0.487	0.498	0.169	0.430	0.351	0.287	0.115	0.303	0.462	0.493	0.192	0.369
SAM 音频评判器	**0.883**	**0.943**	**0.841**	**0.891**	**0.815**	**0.858**	**0.766**	**0.791**	**0.815**	**0.837**	**0.775**	**0.818**
斯皮尔曼等级相关系数（SRCC）												
CLAP	0.380	0.291	0.325	0.273	0.285	0.293	0.199	0.296	0.493	0.376	0.388	0.406
SDR Estimator	0.338	0.000	0.395	0.079	0.390	0.203	0.375	0.210	0.173	0.053	0.220	0.073
Gemini-2.5-pro	0.495	0.361	-0.015	0.117	0.338	0.232	0.010	0.008	0.390	0.324	-0.006	0.180
SAM 音频评判器	**0.817**	**0.573**	**0.774**	**0.573**	**0.714**	**0.569**	**0.658**	**0.476**	**0.781**	**0.660**	**0.734**	**0.607**

表16 SAM 音频评判模型与基线方法的比较

如表 16所示，所提出的 SAM 音频评判模型在三种模态——即 语音、音乐和音效——上，在 PCC 和 SRCC 两项指标中均持续优于所有基线方法。SAJ 与人工评分的相关性显著更高：语音、音乐和声音的 PCC 分别达到 0.883、0.815 和 0.815，SRCC 分别达到 0.817、0.714 和 0.781。相比之下，CLAP 等基线模型和

Gemini-2.5-pro 的相关性中等，而基于失真的指标（SDR Estimator）无法捕捉感知质量，相关性往往较低甚至为负。SAJ 在各种模态上的稳定表现凸显了其将能力从语音泛化到音乐、环境声等更复杂声学领域的优势。这些结果证实，所提出的 SAJ 模型通过利用音频-文本联合表征和文本条件预训练，能够有效捕捉人类感知判断，提供比现有基线方法更可靠、细粒度更高的评估框架。附录C.4 介绍了 SAJ 支持的自动化细粒度性能分析。

重排序器 A	重排序器 B	语音	说话人	音乐	野外乐器	通用
SAJ	CLAP	0.17	0.19	0.09	0.01	0.15
CLAP 与 SAJ	CLAP	0.18	0.18	0.14	0.03	0.06
CLAP 与 SAJ	SAJ	0.00	0.02	0.10	0.15	0.07

表17 重排序器之间的净胜率比较（重排序器 A 与重排序器 B）

7.8.2 将评判器用作重排序器

为评估不同重排序器的影响，我们将它们应用于 SAM Audio 模型。该模型会针对每个输入混合音频生成八个候选分离输出。每种重排序器都根据自身的评分机制选出最佳候选项，随后在测试集上评估所选输出（参见表 6）。我们比较三种配置：评判器、CLAP，以及二者的组合（CLAP与评判器组合）。组合重排序器按照 5:1 的权重比例（CLAP：评判器）对两种评分进行线性组合。

表 17 比较了五项分离任务（语音、说话人、音乐、乐器和通用声音分离）在不同重排序配置下的 NWR。总体而言，我们观察到，相较于仅使用 CLAP，将 SAM 音频评判模型集成作为重排序器能够持续提升或稳定性能。

在比较评判器与 CLAP 时，评判器重排序器在大多数类别中取得了更高的 NWR（例如，语音中的 0.17 对 0.19，以及声音分离中的 0.15），表明其更擅长使选择决策与感知质量保持一致。在结合两种信号后（CLAP 与评判器组合 vs. CLAP），混合重排序器进一步提升了音乐分离（0.14）和乐器分离（0.03）中的 NWR，说明 CLAP 评分可以作为 SAJ 的补充。最后，当将 CLAP 与评判器组合直接与评判器进行评估比较时，相较于与 CLAP 的评估结果，NWR 总体有所下降（例如，语音中为 0.00、说话人中为 0.02，以及音乐/乐器中为 0.10–0.15），这表明 SAJ 分数已经捕获了大部分重排序信号。

总之，结果表明，基于评判器的重排序器在各类音频领域中都非常有效。结合评判器与 CLAP 可带来最大的提升，尤其是在语音和通用声音分离方面；仅使用 CLAP 进行重排序往往表现不佳。

8 结论

我们介绍了 SAM Audio，这是一种通用音频分离模型，支持多模态提示，并在各类音频分离任务中实现了最先进的性能。SAM Audio 通过利用流匹配扩展数据规模和模型容量，推动了通用音频分离的发展。为缓解真实音轨匮乏的问题，我们开发了可扩展的数据构建流程，包括领域感知合成混音，以及大规模伪标注流程；后者利用中间版 SAM Audio 检查点，从自然录音中自助生成音轨。这些策略无需人工进行音轨标注，即可实现广泛覆盖。我们还引入了视觉提示和片段提示，作为文本提示的补充模态。值得注意的是，片段提示显著提升了基于文本的分离效果，并支持实用的迭代优化。

为支持未来研究，我们还发布了 SAM Audio-Bench 和 SAM Audio-Judge。SAM Audio-Bench 提供了一个经过精心平衡的基准测试，其中包含由人工标注的文本、视觉和片段提示。

SAM Audio-Judge 提供了一种无参考指标，其与人类感知的相关性显著高于现有替代方案；它还可用作后处理模块，以提升分离质量。

尽管性能表现出色，我们的结果也揭示了若干局限性。具体而言，视觉提示的效果明显不如文本提示，而与语音等专业领域相比，通用音效仍然更具挑战性。要弥补这些不足，可能需要更强的视听基础对齐能力，以及对复杂多源声学场景进行更好的建模，这些问题留待未来工作解决。

致谢

作者谨此感谢 Peng-Jen Chen、Dangna Li、Robin San Roman、Leying Zhang、Carleigh Wood、Andrew Westbury、George Orlin、Anushka Sagar、Vivian Lee、Cynthia Gao、Ida Cheng、Roman Rädle、Victor Loh、Alex He、Dex Honsa、Eric Gan、Kei Koyama、Kevin Ngo、Meng Wang、Michelle Chan、Phillip Thomas、Andrew Huang、Robbie Adkins、Jason Holland、Josh Terry、Ben Samples、Karla Martucci、Bruno Nakano、Yoko Kristiansen、Ashley Gabriel、Athyu Eleti、AndyBass、HelenKlein、Emma LeibmanBaker、Britt Montalvo、Nikhila Ravi 和 Manohar Paluri，在本研究过程中给予了启发性的讨论和及时的支持。

参考文献

Triantafyllos Afouras、Joon Son Chung 和 Andrew Zisserman。用于视觉语音识别的大规模数据集。*arXiv预印本 arXiv:1809.00496*，2018年。

Audioshake。Audioshake，2025年。https://www.audioshake.ai/。

Auphonic。Auphonic，2025年。https://auphonic.com。

Max Bain、Arsha Nagrani、Gül Varol 和 Andrew Zisserman。精简版：基于故事的检索与时间对齐视频字幕。*arXiv预印本 arXiv:2107.11850*，2021年。

Omer Bar-Tal、Lior Yariv、Yaron Lipman 和 Tali Dekel。多重扩散：融合扩散路径以实现可控图像生成。*arXiv预印本 arXiv:2302.08113*，2023年。

Jon Barker、Shinji Watanabe、Emmanuel Vincent 和 Jan Trmal。第五届“chime”语音分离与识别挑战赛：数据集、任务和基线方法。载于 *Interspeech 2018*论文集，第1561–1565页，2018年。

John G Beerends、Christian Schmidmer、Jens Berger、Matthias Obermann、Raphael Ullmann、Joachim Pomy 和 Michael Keyhl。感知客观听音质量评估（POLQA）：用于端到端语音质量测量的第三代国际电信联盟电信标准化部门标准，第一部分——时间对齐。《音频工程学会期刊》，61(6)：366–384，2013年。

Rachel M Bittner、Justin Salamon、Mike Tierney、Matthias Mauch、Chris Cannam 和 Juan Pablo Bello。《MedleyDB：面向注释密集型音乐信息检索研究的多轨数据集》。载于*ISMIR*，第 14 卷，第 155–160 页，2014年。

Daniel Bolya、Po-Yao Huang、Peize Sun、Jang Hyun Cho、Andrea Madotto、Chen Wei、Tengyu Ma、Jiale Zhi、Jathushan Rajasegaran、Hanoona Rasheed、Junke Wang、Marco Monteiro、Hu Xu、Shiyu Dong、Nikhila Ravi、Daniel Li、Piotr Dollár 和 Christoph Feichtenhofer。《感知编码器：最佳视觉嵌入并非位于网络的输出端》。*arXiv:2504.13181*，2025年。

Estefania Cano、Derry FitzGerald、Antoine Liutkus、Mark D Plumbley 和 Fabian-Robert Stöter。音源分离算法的质量评估：主观评分与客观指标。载于*2016*年第*24*届欧洲信号处理会议（*EUSIPCO*），第 1758–1762 页。IEEE，2016年。

Nicolas Carion、Laura Gustafson、Yuan-Ting Hu、Shoubhik Debnath、Ronghang Hu、Didac Suris、Chaitanya Ryali、Kalyan Vasudev Alwala、Haitham Khedr、Andrew Huang、Jie Lei、Tengyu Ma、Baishan Guo、Arpit Kalla、Markus Marks、Joseph Greer、Meng Wang、Peize Sun、Roman Rädle、Triantafyllos Afouras、Effrosyni Mavroudi、Katherine Xu、Tsung-Han Wu、Yu Zhou、Liliane Momeni、Rishi Hazra、Shuangrui Ding、Sagar Vaze、Francois Porcher、Feng Li、Siyuan Li、Aishwarya Kamath、Ho Kei Cheng、Piotr Dollár、Nikhila Ravi、Kate Saenko、Pengchuan Zhang 和 Christoph Feichtenhofer。《SAM 3：基于概念分割万物》，2025年。https://arxiv.org/abs/2511.16719。

Jean Carletta、Simone Ashby、Sebastien Bourban、Mike Flynn、Mael Guillemot、Thomas Hain、Jaroslav Kadlec、Vasilis Karaiskos、Wessel Kraaij、Melissa Kronenthal、Guillaume Lathoud、Mike Lincoln、Agnes Lisowska、Iain McCowan、Wilfried Post、Dennis Reidsma 和 Pierre Wellner。《AMI会议语料库：预告》。载于 Steve Renals 和 Samy Bengio（编），多模态交互机器学习，第 28–39 页，柏林、海德堡，2006年。Springer Berlin Heidelberg。国际标准书号 978-3-540-32550-5。

Mark Cartwright、Bryan Pardo 和 Gautham J Mysore。《用于声源分离评估的众包成对比较》。载于*2018*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*），第 606–610 页。IEEE，2018年。

Bo Chen、Chao Han、Haibin Wu、Chun-Yi Kuo、Ke Hu、Yi-Chiao Wu、Yanmin Qian、Hung-yi Lee、Xuankai Chang、Dong Yu、Ehsan Hosseini-Asl、Kunal Dhawan、Jia Qi、Wenze Ren、Yu Tsao、Sabato Marco Siniscalchi、Eunsol Choi、Ann Lee、Shinji Watanabe、Carlos Aguirre、Ruchao Fan、Roshan Sharma、Xuanru Zhou、Shukjae Choi、Lucas Bandarkar、Benjamin Peloquin、Haoyu Gao、Edson Cataluna、Samuel Pessoa、Samuel Guimarães、Jinchuan Tian、Yuchen Hu、Ivan Kukanov、Yiming Gao、Yifan Peng、Yihan Wu、Xuankai Chang、Zhaoheng Ni、Vanshika Chhaparia、Kumud Tripathi、Yiwen Shao、Jiatong Shi、Anshul Kumar、Jiawei Ge、Miao Liu、Yuxun Tang、Andrew Wang、Jionghao Han、Dunjie Fang、Jinyuan Li、Hexin Liu、Lingxuan Ye、Ruihan Xu、Kanghyun Choi、Isaac Chung、Avihu Dekel、Anuj Diwan、Neelay Shah、Haoqin Tu、Jagadeesh Balam、Tanel Alumäe、Zhaolin Li、Stefan Constantin、Erica Cooper、Leshem Choshen、Hagai Aronowitz、Yifan Peng、Ziyang Ma、Siddhant Arora、Rohan Kadekodi、Fabian Ritter-Gutierrez、Shuai Wang、Jionghao Han、Saurabhchand Bhati、Yosuke Kashiwagi、Yiwen Zhao、Keer Xu、Yanning Zhang、Chao-Han Huck Yang、Sid Arora、Hayato Futami、Jee-weon Jung、Ruchira Sharma、Suwon Shon、Daniel Mas Montserrat、Darshana Fernandez、Siddhant Arora、Vrunda N Sukhadia、Julia Mendelev、Rohita Mocharla、Wenxi Chen、Omar Abdelwahab、Rehana Mahfuz、Yifan Yang、Jiachen Lian、Xuanru Zhou、Junteng Jia、Abrham Kiflu Gebrekidan、Kunal Dhawan、Ankita Pasad、Arunkumar A、Bhuvana Ramabhadran、Joseph Marvin Imperial、Matthew C. Hunt、Parvathy Krishnaswamy、Vaibhav Visweswaran、Meganathan Ramanathan、Rong Fu、Haset Gebremedhin、Pooja Hiregoudar、Anjali Jayakrishnan、Priyanka Shekar、Anjali Sarvaiya、Yash Patel、Ifeoluwa Wuraola、Mehdi Hosseini Kivanani、Minghao Yang、Joshua Jacob、Meena Saravanan、Kaavya Chaparala、Mohammad Anas Jawed、Emiru Tsunoo、Andrew Rouditchenko、Muhammad Shakeel、Darsh Agrawal、Dhavalkumar Patel、Matthew Wiesner、Satoshi Nakamura、Vishal Sunder、Saurabh Kataria、Osama Hanna、Samuel Thomas、Gene-Ping Yang、Rishika Agarwal、Xiaofei Wang、Rajesh Gupta、Xiaodong Cui、Hong-Kwang Kuo、Zvi Kons、Aditya Patwardhan、Shammur Absar Chowdhury、Shinnosuke Takamichi、Ryuichiro Hataya、Jun Shimamura、Tomohiro Tanaka、Takanori Ashihara、Wataru Ikeda、Wangyou Zhang、Nabarun Goswami、Sadhana Kumaravel、Sathvik Udupa、Yooncheol Ju、Shota Horiguchi、Dancheng Liu、Sharon Goldman、Eunjung Yeo、Alexander Polok、Matthew Maciejewski、Vimal Manohar、Piotr Żelasko、Lucas Ondel Yang、Alexander Richard、Leda Sari、Prashanth Gurunath Shivakumar、Arnab Ghoshal、Kohei Matsuura、Takanori Ashihara、Moris Simko、Ivan Medennikov、Taejin Park、Virender Kadyan、Dominik Wagner、Yuan Tseng、Layne Berry、Xuenan Xu、Ashi Garg、Chutong Meng、Kalvin Chang、Amir Hussein、Dongji Gao、Baochen Yang、Brian Yan、Manthan Thakker、Anfeng Xu、Yuanyuan Zhang、Xinzhao Wang、Hexin Liu、Akshaj Gupta、Chen Feng、Duygu Altinok、Shuaishuai Ye、Toshko Mihaylov、Christophe Veaux、Rohit Paturi、Alkis Koudounas、Marco Sini、Katrin Tomanek、Liangyou Li、Xiaosong Ma、Bashar Awwad Shiekh Hasan、Saidul Islam、Omar Mohamed Al Ghadi、St Shalev Ron、Hsin-Min Wang、Bin Ma、Chng Eng Siong、Ruobing Xie、Shuai Fang、Kangdi Mei、Chao Zhang、Shuai Wang、Haizhou Li、Yang Xiang、Xiaoxi Mao、Prabhat Pandey、Ujan Deb、Kunal Sinha、Amrit Kumar Muduli、Abhayjeet Singh、Murali Karthick Baskar、Luciana Ferrer、Erik Zarrella、Marco Gaido、Sara Papi、Maike Züfle、Iuliia Nigmatulina、Jörg Tiedemann、Lasse Hansen、Marcely Zanon Boito、Oleg Rybakov、Minsik Cho、Weiyi Sun、Xianrui Zheng、Kunal Dhawan、Vitaly Lavrukhin、Nithin Rao Koluguri、Taejin Park、Agoston Weller、Boris Ginsburg、Shinji Watanabe 和 Graham Neubig。ESPnet-SLU 进展、理解与基准测试。载于*2025*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*），第 1–5 页。IEEE，2023年。

Honglie Chen、Weidi Xie、Andrea Vedaldi 和 Andrew Zisserman。Vggsound：一个大规模视听数据集。发表于 *ICASSP 2020-2020 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，第 721–725 页。IEEE，2020a。Zhuo Chen、Takuya Yoshioka、Liang Lu、Tianyan Zhou、Zhong Meng、Yi Luo、Jian Wu、Xiong Xiao 和 Jinyu Li。连续语音分离：数据集与分析。发表于 *ICASSP 2020-2020 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，第 7284–7288 页。IEEE，2020b。

Michael Chinen、Felicia SC Lim、Jan Skoglund、Nikita Gureev、Feargus O Gorman 和 Andrew Hines。Visqol v3：一种开源、可用于生产环境的感知语音与音频指标。发表于*2020*年第*12*届多媒体体验质量国际会议 *(QoMEX)*，第 1–6 页。IEEE，2020 年。

Jang Hyun Cho、Andrea Madotto、Effrosyni Mavroudi、Triantafyllos Afouras、Tushar Nagarajan、Muhammad Maaz、Yale Song、Tengyu Ma、Shuming Hu、Hanoona Rasheed、Peize Sun、Po-Yao Huang、Daniel Bolya、Suyog Jain、Miguel Martin、Huiyu Wang、Nikhila Ravi、Shashank Jain、Temmy Stark、Shane Moon、Babak Damavandi、Vivian Lee、Andrew Westbury、Salman Khan、Philipp Krähenbühl、Piotr Dollár、Lorenzo Torresani、Kristen Grauman 和 Christoph Feichtenhofer。Perceptionlm：用于精细视觉理解的开放获取数据和模型。*arXiv:2504.13180*，2025年。Joon Son Chung、Arsha Nagrani 和 Andrew Zisserman。Voxceleb2：深度说话人识别。发表于《*Interspeech 2018* 论文集》，第 1086–1090 页，2018年。ClearVoice。Clearvoice，2025年。https://github.com/modelscope/ClearerVoice-Studio/tree/main/train/target_ 说话人_提取#4-基于人脸或唇部影像进行条件控制的视听说话人提取。Gheorghe Comanici、Eric Bieber、Mike Schaekermann、Ice Pasupat、Noveen Sachdeva、Inderjit Dhillon、Marcel Blistein、Ori Ram、Dan Zhang、Evan Rosen 等。Gemini 2.5：借助高级推理、多模态、长上下文和新一代智能体能力，拓展前沿。*arXiv*预印本 *arXiv:2507.06261*，2025年。

Erica Cooper 和 Junichi Yamagishi。探究合成语音平均意见分评分中的范围均衡偏差。载于*Proc. Interspeech 2023*，第 1104–1108 页，2023年。

Joris Cosentino、Manuel Pariente、Samuele Cornell、Antoine Deleforge 和 Emmanuel Vincent。LibriMix：一个用于通用语音分离的开放源代码数据集。*arXiv*预印本 *arXiv:2005.11262*，2020年。

Shaoxiang Dang、Tetsuya Matsumoto、Yoshinori Sakaguchi、Ryoichi Takashima 和 Tetsuji Ogawa。使用半监督学习结合基于自监督学习的 SNR 估计器进行单声道语音分离。载于*Proc. INTERSPEECH*，第 2023 卷，第 3759–3763 页，2023年。

Alexandre Défossez。发表于*ISMIR 2021*音乐源分离研讨会论文集，2021年。

Alexandre Défossez、Nicolas Usunier、Léon Bottou 和 Francis Bach。波形域中的音乐源分离。*arXiv*预印本 *arXiv:1911.13254*，2019年。

Alexandre Défossez、Jade Copet、Gabriel Synnaeve 和 Yossi Adi。高保真神经音频压缩。机器学习研究汇刊，2022年。

Marc Delcroix、Tsubasa Ochiai、Katerina Zmolikova、Keisuke Kinoshita、Naohiro Tawara、Tomohiro Nakatani 和 Shoko Araki。利用时域 SpeakerBeam 提升目标语音提取中的说话人区分能力。载于*ICASSP 2020-2020 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，第 691–695 页。IEEE，2020年。

Pablo M Delgado 和 Jürgen Herre。迈向更优的客观感知音频质量评估——第 1 部分：一种新颖的数据驱动认知模型。*IEEE/ACM*音频、语音与语言处理汇刊，2024年。

Hao-Wen Dong、Naoya Takahashi、Yuki Mitsufuji、Julian McAuley 和 Taylor Berg-Kirkpatrick。Clipsep：利用含噪无标签视频学习文本查询式声音分离。载于国际学习表征会议论文集 *(ICLR)*，2023年。

Jinwei Dong、Xinsheng Wang 和 Qirong Mao。Edsep：一种有效的基于扩散的语音源分离方法。载于 *ICASSP 2025-2025 IEEE*国际声学、语音与信号处理会议 *(ICASSP)*，第 1–5 页。IEEE，2025年。

Konstantinos Drossos、Stylianos Ioannis Mimilakis、Dmitriy Serdyuk、Gerald Schuller、Tuomas Virtanen 和 Yoshua Bengio。Maestro 数据集。载于*ICASSP 2020-2020 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，第 7 卷，第 3777–3781 页。IEEE，2020年。

Abhimanyu Dubey、Abhinav Jauhri、Abhinav Pandey、Abhishek Kadian、Ahmad Al-Dahle、Aiesha Letman、Akhil Mathur、Alan Schelten、Amy Yang、Angela Fan、Anirudh Goyal、Anthony Hartshorn、Aobo Yang、Archi Mitra、Archie Sravankumar、Artem Korenev、Arthur Hinsvark、Arun Rao、Aston Zhang、Aurélien Rodriguez、Austen Gregerson、Ava Spataru、Baptiste Rozière、Bethany Biron、Binh Tang、Bobbie Chern、Charlotte Caucheteux、Chaya Nayak、Chloe Bi、Chris Marra、Chris McConnell、Christian Keller、Christophe Touret、Chunyang Wu、Corinne Wong、Cristian Canton Ferrer、Cyrus Nikolaidis、Damien Allonsius、Daniel Song、Danielle Pintz、Danny Livshits、David Esiobu、Dhruv Choudhary、Dhruv Mahajan、Diego Garcia-Olano、Diego Perino、Dieuwke Hupkes、Egor Lakomkin、Ehab AlBadawy、ElinaLobanova、Emily Dinan、EricMichael Smith、FilipRadenovic、FrankZhang、GabrielSynnaeve、Gabrielle Lee、Georgia Lewis Anderson、Graeme Nail、Grégoire Mialon、Guan Pang、Guillem Cucurell、Hailey Nguyen、Hannah Korevaar、Hu Xu、Hugo Touvron、Iliyan Zarov、Imanol Arrieta Ibarra、Isabel M. Kloumann,

Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone，以及其他作者。Llama 3 模型群。*CoRR*，abs/2407.21783，2024年。doi: 10.48550/ARXIV.2407.21783。https://doi.org/10.48550/arXiv.2407.21783。

ElevenLabs。ElevenLabs，2025年。https://elevenlabs.io/voice-isolator。

Ariel Ephrat、Inbar Mosseri、Oran Lang、Tali Dekel、Kevin Wilson、Avinatan Hassidim、William T Freeman 和 Michael Rubinstein。《聆听鸡尾酒会：用于语音分离的说话人无关音视模型》。*ACM* 图形学汇刊（*TOG*），37(4):1–11，2018年。

Giorgio Fabbro、Stefan Uhlich、Chieh-Hsin Lai、Woosung Choi、Marco Martínez-Ramírez、Weihsiang Liao、Igor Gadelha、Geraldo Ramos、Eddie Hsu、Hugo Rodrigues等。《2023 声音解混挑战赛——音乐解混赛道》。国际音乐信息检索学会汇刊，7(1)，2024年。

FADR。Fadr，2025年。https://fadr.com/。

Szu-Wei Fu、Chien-Feng Liao、Yu Tsao 和 Shou-De Lin。《MetricGAN：基于生成对抗网络的语音增强黑盒指标分数优化》。载于国际机器学习会议，第 2031–2041 页。PmLR，2019年。

Meng Ge、Chenglin Xu、Longbiao Wang、Eng Siong Chng、Jianwu Dang 和 Haizhou Li。SpEx+：一种完全时域说话人提取网络。载于*Interspeech 2020*，第1406–1410页，2020年。

Jort F Gemmeke、Daniel PW Ellis、Dylan Freedman、Aren Jansen、Wade Lawrence、R Channing Moore、Manoj Plakal 和 Marvin Ritter。《Audio set：用于音频事件的本体和人工标注数据集》。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。

Rohit Girdhar、Alaaeldin El-Nouby、Zhuang Liu、Mannat Singh、Kalyan Vasudev Alwala、Armand Joulin 和 Ishan Misra。ImageBind：一个能够绑定所有内容的嵌入空间。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 15180–15190 页，2023年。

Kilem L Gwet。《评定者间信度手册：测量评分者间一致程度的权威指南》。Advanced Analytics, LLC，2014年。

Jiarui Hai、Helin Wang、Dongchao Yang、Karan Thakkar、Najim Dehak 和 Mounya Elhilali。Dpm-tse：一种用于目标声音提取的扩散概率模型。载于*ICASSP 2024 - 2024*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*），第 1196–1200 页，2024年。

Romain Hennequin、Anis Khlif、Felix Voituret 和 Manuel Moussallam。Spleeter：一种采用预训练模型的快速高效音乐源分离工具。*Journal of Open Source Software*，5(50)：2154，2020年。

John R Hershey、Zhuo Chen、Jonathan Le Roux 和 Shinji Watanabe。深度聚类：用于分割和分离的判别式嵌入。载于*2016*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*），第31–35页。IEEE，2016年。

Chao Huang、Susan Liang、Yapeng Tian、Anurag Kumar 和 Chenliang Xu。DAVIS：利用生成式扩散模型实现高质量视听分离。*arXiv e-prints*，第arXiv–2308页，2023年。

Chao Huang、Susan Liang、Yapeng Tian、Anurag Kumar 和 Chenliang Xu。来自多样类别的高质量视觉引导声音分离。载于亚洲计算机视觉会议（*ACCV*）论文集，第35–49页，2024年12月。

Chao Huang、Susan Liang、Yapeng Tian、Anurag Kumar 和 Chenliang Xu。通过视觉引导的生成建模实现跨多样类别的高质量声音分离。*arXiv*预印本 *arXiv:2509.22063*，2025年。

Wen-Chin Huang、Erica Cooper、Yu Tsao、Hsin-Min Wang、Tomoki Toda 和 Junichi Yamagishi。VoiceMOS 挑战赛 2022。*arXiv*预印本 *arXiv:2203.11389*，2022年。

Jaesung Huh。语音性别分类器。Huggingface 模型集合。https://huggingface.co/JaesungHuh/语音性别分类器，2024年。

Noah Jaffe 和 John Ashley Burgoyne。《音乐源分离对决：比较客观指标与人类感知》。*arXiv*预印本 *arXiv:2507.06917*，2025年。

Nikhil Kandpal、Oriol Nieto和Zeyu Jin。通过图像转换处理会议(ICASSP)，第3124–3128 页。IEEE，2022年。

Chris Donahue、Zhuo Chen和Kim Kuhn。Wavegrad：用于音频生成的梯度。载于2019年计算语言学协会北美分会年会：人类语言技术论文集，第 *1* 卷（长篇和短篇论文），第 119–132 页，2019年。

Minseok Kim、Woosung Choi、Jaehwa Chung、Daewon Lee 和 Soonyoung Jung。Kuielab-mdx-net：一种用于音乐分轨的双流神经网络。载于 *ISMIR*，2021年。

Diederik P Kingma 和 Max Welling。自动编码变分贝叶斯。*arXiv*预印本 *arXiv:1312.6114*，2013.

Alexander Kirillov、Eric Mintun、Nikhila Ravi、Hanzi Mao、Chloe Rolland、Laura Gustafson、Tete Xiao、Spencer Whitehead、Alexander C Berg、Wan-Yen Lo 等。Segment Anything。发表于*IEEE/CVF* 国际计算机视觉会议论文集，第 4015–4026 页，2023年。

Qiuqiang Kong、Yin Cao、Turab Iqbal、Yuxuan Wang、Wenwu Wang 和 Mark D Plumbley。Panns：用于音频模式识别的大规模预训练音频神经网络。*IEEE/ACM*音频、语音与语言处理汇刊，28：2880–2894，2020年。

Qiuqiang Kong、Ke Chen、Haohe Liu、Xingjian Du、Taylor Berg-Kirkpatrick、Shlomo Dubnov 和 Mark D Plumbley。利用弱标注数据进行通用声源分离。*arXiv*预印本 *arXiv:2305.07447*，2023年。

Khaled Koutini、Jan Schlüter、Hamid Eghbal-zadeh 和 Gerhard Widmer。使用 Patchout 高效训练音频变换器。收录于*Interspeech 2022*，国际语音通信协会第*23*届年会，韩国仁川，*2022*年*9*月*18*日至*22*日，第 2753–2757 页。国际语音通信协会，2022年。doi: 10.21437/Interspeech.2022-227。https: //doi.org/10.21437/Interspeech.2022-227。

Rithesh Kumar、Prem Seetharaman、Alejandro Luebs、Ishaan Kumar 和 Kundan Kumar。采用改进的 PRVQGAN 进行高保真音频压缩。神经信息处理系统进展，36:27980–27993，2023年。

Lalal.AI. LalalAI，2025年. https://www.lalal.ai.

Matt Le、Apoorv Vyas、Bowen Shi、Brian Karrer、Leda Sari、Rashel Moritz、Mary Williamson、Vimal Manohar、Yossi Adi、Jay Mahadeokar 和 Wei-Ning Hsu。《Voicebox：大规模文本引导的多语言通用语音生成》。*arXiv*，abs/2306.15687，2023年。

Jonathan Le Roux、Scott Wisdom、Hakan Erdogan 和 John R Hershey。SDR——半生不熟还是恰到好处？载于*ICASSP 2019*，第626–630页，2019年。

Bochen Li、Xinzhao Liu、Karthik Dinesh、Zhiyao Duan 和 Gaurav Sharma。创建用于多模态音乐分析的多轨古典音乐演奏数据集：挑战、见解与应用。*IEEE*多媒体汇刊，21(2)：522–535，2018年。

Guangyao Li、Yake Wei、Yapeng Tian、Chenliang Xu、Ji-Rong Wen 和 Di Hu。学习在动态音视场景中回答问题。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。

Kai Li、Fenghua Xie、Hang Chen、Kexin Yuan 和 Xiaolin Hu。受皮层-丘脑-皮层回路启发的音视语音分离模型。*IEEE*模式分析与机器智能汇刊，46(10)：6637–6651，2024a。

Kai Li、Runxuan Yang、Fuchun Sun 和 Xiaolin Hu。IIANet：一种用于音视语音分离的模态内与模态间注意力网络。载于 Ruslan Salakhutdinov、Zico Kolter、Katherine Heller、Adrian Weller、Nuria Oliver、Jonathan Scarlett 和 Felix Berkenkamp（编），第*41*届国际机器学习会议论文集，机器学习研究论文集第235卷，第29181–29200页。PMLR，2024b。

Yaron Lipman、Ricky Q. Chen、Heli Ben-Hamu、Maxim Nickel 和 Matthew Le。用于生成建模的流匹配。载于国际学习表征会议，2023年。

Chen Liu、Peng Li、Qingtao Yu、Hongwei Sheng 和 Dezhong Wang。Lipsch-Net：一种用于音视语音分离的模态内与模态间注意力网络。载于 Ruslan Salakhutdinov、Zico Kolter、Katherine Heller、Adrian Weller、Nuria Oliver、Jonathan Scarlett 和 Felix Berkenkamp（编），第*41*届国际机器学习会议论文集，机器学习研究论文集第235卷，第29181–29200页。PMLR，2024b。

Yaron Lipman、Ricky Q. Chen、Heli Ben-Hamu、Maxim Nickel 和 Matthew Le。用于生成建模的流匹配。载于国际学习表征会议，2023年。

Chen Liu、Peng Li、Qingtao Yu、Hongwei Sheng 和 Dezhong Wang。Lipsch-Net：一种用于音视语音分离的模态内与模态间注意力网络。载于 Ruslan Salakhutdinov、Zico Kolter、Katherine Heller、Adrian Weller、Nuria Oliver、Jonathan Scarlett 和 Felix Berkenkamp（编），第*41*届国际机器学习会议论文集，机器学习研究论文集第235卷，第29181–29200页。PMLR，2024b。

Xubo Liu、Haohe Liu、Qiuqiang Kong、Xinhao Mei、Jinzheng Zhao、Qiushi Huang、Mark D Plumbley 和 Wenwu Wang。《分离你所描述的内容：语言查询音频源分离》。载于*Interspeech 2022* 论文集，第 1801–1805 页，2022年。

Xubo Liu、Qiuqiang Kong、Yan Zhao、Haohe Liu、Yi Yuan、Yuzhuo Liu、Rui Xia、Yuxuan Wang、Mark D Plumbley 和 Wenwu Wang。《分离你描述的任何对象》。*IEEE/ACM*音频、语音与语言处理汇刊，2024b。Philipos C Loizou。《语音增强：理论与实践》。CRC出版社，2013年。Wei-Tsung Lu、Ju-Chiang Wang、Qiuqiang Kong 和 Yun-Ning Hung。《使用频带分割旋转位置编码变换器进行音乐源分离》。载于 *2024* 年 *IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 481–485 页。IEEE，2024年。Yen-Ju Lu、Zhong-Qiu Wang、Shinji Watanabe、Alexander Richard、Cheng Yu 和 Yu Tsao。《用于语音增强的条件扩散概率模型》。载于*ICASSP 2022-2022 IEEE*国际声学、语音与信号处理会议（*ICASSP*），第 7402–7406 页。IEEE，2022年。Yi Luo 和 Nima Mesgarani。《TasNet：用于实时单通道语音分离的时域音频分离网络》。载于*2018*年*IEEE*国际声学、语音与信号处理会议，第 696–700 页。IEEE，2018年。Yi Luo 和 Nima Mesgarani。《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):1256–1266，2019年。Yi Luo、Zhuo Chen 和 Takuya Yoshioka。《双路径循环神经网络：用于时域单通道语音分离的高效长序列建模》。载于*ICASSP 2020-2020 IEEE*声学、语音与信号处理国际会议，第 46–50 页。IEEE，2020年。Hao Ma、Zhiyuan Peng、Xu Li、Mingjie Shao、Xixin Wu 和 Ju Liu。《CLAPSep：利用对比预训练模型进行多模态查询条件目标声音提取》。*IEEE/ACM*音频、语音与语言处理汇刊，2024年。Matthew Maciejewski、Gordon Wichern、Emmett McQuinn 和 Jonathan Le Roux。《WHAMR!：含噪混响单通道语音分离》。载于*ICASSP 2020-2020 IEEE*声学、语音与信号处理国际会议，第 696–700 页。IEEE，2020年。

Ethan Manilow、Gordon Wichern、Prem Seetharaman 和 Jonathan Le Roux。《Cutting music source separation some Slakh: A study in a controlled dataset and its influence on training quality and quantity》。载于*2019*年*IEEE*音频与声学信号处理应用研讨会（*WASPAA*），第45–49页。IEEE，2019年。

Pranay Manocha、Zeyu Jin、Richard Zhang 和 Adam Finkelstein。CDPAM：用于感知音频相似度的对比学习。载于*ICASSP 2021-2021*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*），第196–200页。IEEE，2021年。

Giorgio Mariani、Irene Tallini、Emilian Postolache、Michele Mancusi、Luca Cosmo 和 Emanuele Rodolà。用于同时进行音乐生成与分离的多源扩散模型。*arXiv*预印本 *arXiv:2302.02257*，2023年。Shivam Mehta、Ruibo Tu、Jonas Beskow、Éva Székely 和 Gustav Eje Henter。Matcha-TTS：采用条件流匹配的快速文本转语音架构。载于论文集 *ICASSP*，2024年。

Yuki Mitsufuji、Giorgio Fabbro、Stefan Uhlich、Fabian-Robert Stöter、Alexandre Défossez、Minseok Kim、Woosung Choi、Chin-Yun Yu 和 Kin-Wai Cheuk。2021年音乐分轨挑战赛，808395，2022年 1。《Frontiers in Signal Processing》，：。

Gabriel Mittag、Babak Naderi、Assmaa Chehadi 和 Sebastian Möller。NISQA：一种基于深度卷积神经网络自注意力模型的多维语音质量预测方法，使用众包数据集。发表于*Proc. Interspeech 2021*，2131，2021年 2127页 –。Moises.AI。Moises.AI，2025年。https://moises.ai/。Viet-An Nguyen、Peibei Shi、Jagdish Ramakrishnan、Udi Weinsberg、Henry C Lin、Steve Metz、Neil Chandra、Jane Jing 和 Dimitris Kalimeris。CLARA：标签和评审者置信度。发表于 第*26*届 *ACM SIGKDD* 知识发现与数据挖掘国际会议论文集，第2542–2552页，2020年。

Zexu Pan、Gordon Wichern、François G Germain、Kohei Saijo 和 Jonathan Le Roux。用于多模态测听语言人提取的用插即用共现上下注意力 *arXiv*预印本 *arXiv:2504.12734*，2025年。

Santiago Pascual、Antonio Bonafonte 和 Joan Serrà。《SEGAN：语音增强生成对抗网络》。载于《*Interspeech 2017*》论文集，第 3642–3646 页，2017年。

William Peebles 和 Saining Xie。《基于变换器的可扩展扩散模型》。载于《IEEE/CVF 国际计算机视觉会议论文集》，第 4195–4205 页，2023年。Adam Polyak、Amit Zohar、Andrew Brown、Andros Tjandra、Animesh Sinha、Ann Lee、Apoorv Vyas、Bowen Shi、Chih- Yao Ma、Ching-Yao Chuang 等。《Movie Gen：一系列媒体基础模型》。arXiv预印本 arXiv:2410 13720，2024年。。KR Prajwal、Bowen Shi、Matthew Le、Apoorv Vyas、Andros Tjandra、Mahi Luthra、Baishan Guo、Huiyu Wang、Triantafyllos Afouras、David Kant 等。《MusicFlow：用于文本引导音乐生成的级联流匹配》。载于《国际学习表征会议》，2024年。pydub。pydub。https://github.com/jiaaro/pydub。Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。《通过自然语言监督学习可迁移视觉模型》。载于《国际机器学习会议》，第 8748–8763 页。PmLR，2021年。Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J Liu。《探索统一文本到文本变换器的迁移学习能力极限》。机器学习研究期刊，21(140)：1–67，2020年。Zafar Rafii、Antoine Liutkus、Fabian-Robert Stöter、Stylianos Ioannis Mimilakis 和 Rachel Bittner。《用于音乐分离的 MUSDB18 语料库》，2017年12月。https://doi.org/10.5281/zenodo.1117372。Nikhila Ravi、Valentin Gabeur、Yuan-Ting Hu、Ronghang Hu、Chaitanya Ryali、Tengyu Ma、Haitham Khedr、Roman Rädle、Chloe Rolland、Laura Gustafson、Eric Mintun、Junting Pan、Kalyan Vasudev Alwala、Nicolas Carion、Chao-Yuan Wu、Ross Girshick、Piotr Dollár 和 Christoph Feichtenhofer。《SAM 2：分割图像和视频中的任意对象》。arXiv预印本 arXiv:2408.00714，2024年。https://arxiv.org/abs/2408.00714。Chandan KA Reddy、Harishchandra Dubey、Vishak Gopal、Ross Cutler、Sebastian Braun、Hannes Gamper、Robert Aichner 和 Sriram Srinivasan。《ICASSP 2021深度噪声抑制挑战赛》。载于《ICASSP 2021-2021 IEEE国际声学、语音与信号处理会议（ICASSP）》，第 6623–6627 页。IEEE，2021a。Chandan KA Reddy、Vishak Gopal 和 Ross Cutler。《DNSMOS：用于评估噪声抑制器的非侵入式感知客观语音质量指标》。载于《ICASSP 2021-2021 IEEE国际声学、语音与信号处理会议（ICASSP）》，第 6493–6497 页。IEEE，2021b。Julius Richter、Simon Welker、Jean-Marie Lemercier、Bunlong Lay 和 Timo Gerkmann。《使用基于扩散的生成模型进行语音增强和去混响》。IEEE/ACM音频、语音与语言处理汇刊，31：2351–2364，2023年。Simon Rouard、Francisco Massa 和 Alexandre Défossez。《用于音乐声源分离的混合变换器》。载于 ICASSP 23，2023年。Neville Ryant、Kenneth Church、Christopher Cieri、Alejandrina Cristia、Jun Du、Sriram Ganapathy 和 Mark Liberman。《第二届DIHARD说话人日志挑战赛：数据集、任务和基线方法》。载于《Interspeech 2019》论文集，第 978–982 页，2019年。Kohei Saijo、Gordon Wichern、François G Germain、Zexu Pan 和 Jonathan Le Roux。《TF-LoCoformer：通过卷积进行局部建模、用于语音分离和增强的变换器》。载于《2024年第18届国际声学信号增强研讨会（IWAENC）》，第 205–209 页。IEEE，2024年。Ryosuke Sawata、Naoki Murata、Yuhta Takida、Toshimitsu Uesaka、Takashi Shibuya、Shusuke Takahashi 和 Yuki Mitsufuji。《Diffiner：用于语音增强的多功能基于扩散的生成式精炼器》。arXiv预印本 arXiv:2210.17287，2022年。Noah Schaffer、Boaz Cogan、Ethan Manilow、Max Morrison、Prem Seetharaman 和 Bryan Pardo。《通过生成建模增强音乐分离》。载于《ISMIR 2022混合会议》，2022年。Robin Scheibler、Youna Ji、Soo-Whan Chung、Jaeuk Byun、Soyeon Choe 和 Min-Seok Choi。《基于扩散的生成式语音源分离》。载于《ICASSP 2023-2023 IEEE国际声学、语音与信号处理会议（ICASSP）》，第 1–5 页。IEEE，2023年。Robin Scheibler、John R Hershey、Arnaud Doucet 和 Henry Li。《通过流匹配进行声源分离》。arXiv预印本 arXiv:2505.16119，2025年。国际电信联盟。电信标准化部门。《传输质量主观测定方法》。国际电信联盟，1996年。

国际电信联盟无线电通信全会，2，2014年。Fabian-Robert Stöter、Antoine Liutkus 和 Nobutaka Ito。2018年信号分离评测竞赛。载于潜变量分析与信号分离国际会议，第293–305页。施普林格，2018年。Fabian-Robert Stöter、Stefan Uhlich、Antoine Liutkus 和 Yuki Mitsufuji。Open-Unmix——音乐源分离的参考实现。*Journal of Open Source Software*，4(41):1667，2019年。Cem Subakan、Mirco Ravanelli、Samuele Cornell、Marco Bronzi 和 Jianyuan Zhong。注意力就是语音分离所需的一切。载于*ICASSP*，2021年。Y Cem Subakan 和 Paris Smaragdis。生成对抗式源分离。载于*2018*年*IEEE*国际声学、语音与信号处理会议 *(ICASSP)*，第26–30页。IEEE，2018年。Naoya Takahashi 和 Yuki Mitsufuji。用于音频源分离的多尺度多频带密集网络。载于*2017*年*IEEE*音频与声学信号处理应用研讨会 *(WASPAA)*，第21–25页。IEEE，2017年。Andros Tjandra、Yi-Chiao Wu、Baishan Guo、John Hoffman、Brian Ellis、Apoorv Vyas、Bowen Shi、Sanyuan Chen、Matt Le、Nick Zacharov 等。Meta Audiobox Aesthetics：面向语音、音乐和声音的统一自动质量评估。*arXiv*预印本 *arXiv:2502.05139*，2025年。Amos Tversky 和 Daniel Kahneman。不确定情境下的判断：启发式与偏差：判断中的偏差揭示了不确定情境下思维的一些启发式。科学，185(4157):1124–1131，1974年。Cassia Valentini-Botinhao。用于训练语音增强算法和文本转语音模型的含噪混响语音数据库。*(无标题)*，2017年。Praveen Srinivasa Varadhan、Amogh Gulati、Ashwin Sankar、Srija Anand、Anirudh Gupta、Anirudh Mukherjee、Shiva Kumar Marepally、Ankur Bhatia、Saloni Jaju、Suvrat Bhooshan 等。重新思考MUSHRA：应对文本转语音评估的现代挑战。*arXiv*预印本 *arXiv:2411.12719*，2024年。Emmanuel Vincent、Rémi Gribonval 和 Cédric Févotte。盲音频源分离的性能测量。《*IEEE*音频、语音与语言处理汇刊》，14(4):1462–1469，2006年。Apoorv Vyas、Bowen Shi、Matthew Le、Andros Tjandra、Yi-Chiao Wu、Baishan Guo、Jiemin Zhang、Xinyue Zhang、Robert Adkins、William Ngan 等。Audiobox：通过自然语言提示实现统一音频生成。*arXiv*预印本 *arXiv:2312.15821*，2023年。Apoorv Vyas、Heng-Jui Chang、Cheng-Fu Yang、Po-Yao Huang、Luya Gao、Julius Richter、Sanyuan Chen、Matthew Le、Piotr Dollár、Christoph Feichtenhofer、Ann Lee 和 Wei-Ning Hsu。借助大规模多模态对应学习，推动视听感知的前沿发展。2025年。DeLiang Wang 和 Jitong Chen。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):1702–1726，2018年。Helin Wang、Dongchao Yang、Chao Weng、Jianwei Yu 和 Yuexian Zou。利用时间戳信息改进目标声音提取。载于*INTERSPEECH*论文集，第1526–1530页，2022a。Helin Wang、Dongchao Yang、Yuexian Zou、Fan Cui 和 Yujun Wang。检测你想要的声音：目标声音检测。载于*DCASE*，2022b。Helin Wang、Jesús Villalba、Laureano Moro-Velazquez、Jiarui Hai、Thomas Thebaud 和 Najim Dehak。采用快速生成式校正的抗噪语音分离。载于*INTERSPEECH 2024*论文集，第2165–2169页，2024年。Helin Wang、Jiarui Hai、Yen-Ju Lu、Karan Thakkar、Mounya Elhilali 和 Najim Dehak。SoloAudio：语言导向音频扩散变换器实现的目标声音提取。载于*ICASSP 2025-2025*年*IEEE*国际声学、语音与信号处理会议 *(ICASSP)*，第1–5页。IEEE，2025a。Helin Wang、Jiarui Hai、Dongchao Yang、Chen Chen、Kai Li、Junyi Peng、Thomas Thebaud、Laureano Moro Velazquez、Jesus Villalba 和 Najim Dehak。SoloSpeech：通过级联生成流程提升目标语音提取的可懂度和质量，2025b。https://arxiv.org/abs/2505.19314。Quan Wang、Hannah Muckenhirn、Kevin Wilson、Prashant Sridhar、Zelin Wu、John R Hershey、Rif A Saurous、Ron J Weiss、Ye Jia 和 Ignacio Lopez Moreno。VoiceFilter：通过说话人条件频谱掩蔽实现定向语音分离。载于*INTERSPEECH 2019*论文集，第2728–2732页，2019年。

Shinji Watanabe、Michael Mandel、Jon Barker、Emmanuel Vincent、Ashish Arora、Xuankai Chang、Sanjeev Khudanpur、Vimal Manohar、Daniel Povey、Desh Raj 等。CHiME-6挑战赛：应对未分段录音中的多说话人语音识别。载于*CHiME 2020*论文集，第1–7页，2020年。Yutong Wen、Ke Chen、Prem Seetharaman、Oriol Nieto、Jiaqi Su、Rithesh Kumar、Minje Kim、Paris Smaragdis、Zeyu Jin 和 Justin Salamon。Promptsep：通过多模态提示实现生成式音频分离。*arXiv*预印本 *arXiv:2511.04623*，2025年。Gordon Wichern、Joe Antognini、Michael Flynn、Licheng Richard Zhu、Emmett McQuinn、Dwight Crow、Ethan Manilow 和 Jonathan Le Roux。WHAM!：将语音分离扩展至噪声环境。载于*INTERSPEECH 2019*论文集，第1368–1372页，2019年。Scott Wisdom、Hakan Erdogan、Daniel PW Ellis、Romain Serizel、Nicolas Turpault、Eduardo Fonseca、Justin Salamon、Prem Seetharaman 和 John R Hershey。免费通用声音分离数据有何特别之处？载于*ICASSP 2021-2021*年*IEEE*国际声学、语音与信号处理会议 *(ICASSP)*，第186–190页。IEEE，2021年。Jian Wu、Yong Xu、Shi-Xiong Zhang、Lian-Wu Chen、Meng Yu、Lei Xie 和 Dong Yu。时域音视语音分离。载于*2019*年*IEEE*自动语音识别与理解研讨会 *(ASRU)*，第667–673页。IEEE，2019年。

Yusong Wu、Ke Chen、Tianyu Zhang、Yuchen Hui、Taylor Berg-Kirkpatrick 和 Shlomo Dubnov。采用特征融合和关键词到字幕增强的大规模对比语言-音频预训练。载于*ICASSP 2023-2023 IEEE*国际声学、语音与信号处理会议 *(ICASSP)*，第 1–5 页。IEEE，2023年。

Hu Xu、Saining Xie、Xiaoqing Ellen Tan、Po-Yao Huang、Russell Howes、Vasu Sharma、Shang-Wen Li、Gargi Ghosh、Luke Zettlemoyer 和 Christoph Feichtenhofer。揭开 CLIP 数据的奥秘。*arXiv*预印本 *arXiv:2309.16671*，2023年。

Mohan Xu Kai 用于高效语音分离的时频交错增益提取与重建。载于第十三届国际学习表征会议，2024年。

Sihyun Yu、Sangkyung Kwak、Huiwon Jang、Jongheon Jeong、Jonathan Huang、Jinwoo Shin 和 Saining Xie . 用于生成的表征对齐：训练扩散变换器比你想象的更容易。发表于国际学习表征会议，2025年。

Yi Yuan、Xubo Liu、Haohe Liu、Mark D Plumbley 和 Wenwu Wang。Flowsep：采用整流流匹配的语言查询式声音分离。发表于 *ICASSP 2025-2025 IEEE*国际声学、语音与信号处理会议（*ICASSP*），第 1–5 页。IEEE，2025年。

Hang Zhao、Chuang Gan、Andrew Rouditchenko、Carl Vondrick、Josh McDermott 和 Antonio Torralba。像素之声。发表于欧洲计算机视觉会议论文集（*ECCV*），第 570–586 页，2018年。

Shengkui Zhao、Yukun Ma、Chongjia Ni、Chong Zhang、Hao Wang、Trung Hieu Nguyen、Kun Zhou、Jia Qi Yip、Dianwen Ng 和 Bin Ma。Mossformer2：结合 Transformer 和无RNN循环网络，以增强时域单声道语音分离。发表于 *ICASSP*，第 10356–10360 页。IEEE，2024年。国际标准书号 979-8-3503-4485-1。 http://dblp.uni-trier.de/db/conf/icassp/icassp2024.html#ZhaoMNZ000YN024。

Shengkui Zhao、Zexu Pan 和 Bin Ma。Clearervoice-studio：连接先进的语音处理研究与实际部署。*arXiv*预印本 *arXiv:2506.19398*，2025年。Jinxing Zhou、Jianyuan Wang、Jiayi Zhang、Weixuan Sun、Jing Zhang、Stan Birchfield、Dan Guo、Lingpeng Kong、Meng Wang 和 Yiran Zhong。音频-视觉分割。发表于欧洲计算机视觉会议，2022年。

Ge Zhu、Jordan Darefsky、Fei Jiang、Anton Selitskiy 和 Zhiyao Duan . 使用生成式流进行音乐源分离 . *IEEE*信号处理快报，29：2288–2292，2022年。

附录

A 主观评估

A.1 设计

我们的协议旨在满足四项要求：(i) 在复杂的真实音频/视频中具备生态有效性；(ii) 对消融实验中的微小差异具有敏感性；(iii) 在不同提示模态（文本、视觉掩码片段、时间跨度）和领域（语音、音乐、乐器、一般声音）之间具有可比性；(iv) 考虑到测试项的规模和多样性，具备操作可行性。这些标准促使我们在此任务中弃用经典的单刺激平均意见分/比较平均意见分和 MUSHRA 式多刺激测试，转而采用并列绝对类别评分混合方案。

A.1.1 为什么不用 MUSHRA、平均意见分或 CMOS?

MUSHRA（ITU-R BS.1534）适用于每个条目都有高质量参考和锚点的情况，可通过隐藏参考和低质量锚点进行多刺激评分系列(2014年）。然而，在我们的设置中，(a) 我们往往无法从真实混合音频中获得任何可信的参考提取结果；(b) 对于异质目标（语音、乐器、事件），锚点选择并非易事，且可能导致锚点偏差和参考匹配偏差——质量较高的参考可能拉低原本表现良好的输出的评分；评分可能偏向“与参考的相似度”，而非感知质量Varadhan 等人(2024年）；(c) 多刺激测试环节会加剧范围均衡和会话漂移效应Cooper 和 Yamagishi(2023年）。再加上显著的成本和时间开销，MUSHRA 并不适用于我们的真实场景、多领域分离基准测试。

MOS/ACR（ITU-T P.800）采用五点评分量表，对单个刺激进行简单的绝对评分Sector （1996年）。尽管单刺激 MOS 在操作上很简单，但在我们的应用场景中存在一些众所周知的缺点：（i）在消融研究中，对细微模型差异的敏感度有限；（ii）容易受到锚定效应和范围偏差的影响——在一次评测过程中，绝对分数会随上下文变化而漂移 Tversky 和 Kahneman （1974年）；Cooper 和 Yamagishi （2023年）；以及（iii）若不大幅增加样本量，就无法直接比较模型的相对排名。

CMOS/CCR（成对比较）通过要求评估者表达对两个刺激的偏好（有时还需说明偏好强度），来提高对相对差异的敏感度 Sector （1996年）。然而，纯偏好设计无法提供关于绝对性能阈值（例如，是否“足以部署”）的见解；当不同错误模式产生相似的整体印象时，这种设计也可能不够可靠。

A.1.2 我们的混合方案：并列绝对类别评分与偏好决胜项

因此，我们采用**并列 ACR** 协议，并始终提出**偏好问题**：

- **呈现方式**：针对同一项目的两个模型输出（A、B）会并列呈现，同时提供目标规格（文本提示、SAM 掩码片段和/或时间跨度）以及原始混合音频作为参考背景。

- **绝对评分**：评估者分别对 A 和 B 在与分离目标相对应的三个维度上进行独立的 5 分制 ACR 评分：召回率（提取出多少目标音频）、精确率（混入多少非目标音频）和忠实度（与目标声音在上下文中的听感相似程度）。我们还会对两个模型输出进行总体评分。

- **偏好决胜项**：即使 ACR 评分不同，评估者也始终回答一个直接偏好问题（“您更喜欢哪段提取音频，A 还是 B？”）。这个“强制”决胜项能够提高标注者间一致性（IAA），并使增量估计更加准确。只有在评估者给 A 和 B 的 ACR 评分相同时才有条件地提出该问题，可能会促使评估者更频繁地给出不同评分，以避免额外的条件问题（不过，这一假设并未经过直接检验）。

- **模态覆盖范围**：同一协议适用于各种提示模态——文本、视觉（SAM 掩码片段）和时间跨度，从而能够针对相同的真实场景项目进行受控消融（例如，仅文本与文本+视觉，以及仅视觉）。

这种混合方式兼具两全其美：为部署就绪度提供绝对质量信号，并为消融实验和模型选择提供相对信号。

A.1.3 实证观察

在多项研究中，我们观察到：

- **标注者间一致性有所提升**。加入偏好决胜项后，偏好一致性和绝对差值的一致性均有所提高，相较于不含决胜项的 ACR 和仅成对偏好协议，提升情况通过 Gwet's AC2 衡量（Gwet，2014年）。

- **差值更显著**：加入决胜项后，评估者给出的平均ACR评分差值更大（每个项目约增加 $+0.15$ 分），使绝对分数与表达出的偏好相符，并提高了对近似平局比较的敏感度。

- **差值的不确定性更低**：与单刺激评分相比，并排判断可将模型间差异的置信区间（CI）宽度最多缩小 $\sim$20 %，这意味着在 A/B测试中，达到相同敏感度所需的成本可降低 $\sim$30%；CI 宽度通过自助法衡量。

- **处理时间权衡**：与仅成对方式相比，并列绝对类别评分的单项处理时间大致翻倍（每项问题更多；CMOS 的处理时间为 2 分钟，而仅成对方式为 1 分钟），但一次分配即可同时获得绝对和相对信号。

- **锚定效应持续存在**：绝对分数取决于比较语境——将某个输出与质量更高的对应输出并列呈现，会压低其 ACR评分，这与更广泛的锚定效应和参考匹配文献一致Tversky 和 Kahneman (1974年)；Varadhan 等人 (2024年)；Cooper 和 Yamagishi(2023年)。因此，我们更强调比较统计量（胜率、ACR差值），而不是原始绝对均值。

A.1.4 操作细节

指令与训练。 附录 A.2 包含完整的评估者指令和示例。评估者接受的训练包括：(i) 关注所提供的提示（文本/掩码片段/片段），(ii) 将"精确率"（非目标泄漏）与"召回率"（遗漏目标）区分开来，(iii) 根据混合音频中目标音频的音色/时间特征来评判"忠实度"（而不是与任何外部参考的相似度）。

随机化与盲化。 模型身份经过盲化；每个条目的 A/B顺序均随机排列；条目会被分批处理，以避免同一领域/模态的条目连续出现过多。

质量控制。 我们采用多种工具进行质量控制：（i）供应商侧质量保证流程，由一组经过培训且值得信赖的核心评估者检查部分标注员的评估结果；（ii）对评分者质量进行贝叶斯建模：我们使用 CLARA（Nguyen 等人，2020年）估计评分者混淆矩阵的后验分布，并剔除那些相对于真实值的偏差汇总指标异常偏高的评分者。

考虑到锚定效应的风险，我们更重视比较性统计数据，而非原始绝对均值，并提醒不要过度解读不同实验间绝对水平的变化。

A.1.5 与标准和文献的定位

我们的协议继承了 ACR/MOS Sector（1996年）的绝对评分可解释性，同时保留了类似 CMOS/CCR Sector（1996年）的相对敏感度，但避免了 MUSHRA Series（2014年）和 Varadhan 等（2024年）中对参考音频/锚点的依赖以及多刺激偏差。该协议与近期大规模音频分离评测的发现相符：基于参考的指标（SDR/SI-SDR/SAR）与人类对各音轨的感知之间存在脱节（Stöter 等（2018年））；也与有关评估者行为（锚定效应、范围均衡、会话漂移）的更广泛研究结论相符（Tversky 和 Kahneman（1974年）；Cooper 和 Yamagishi（2023年））。

局限性。并列绝对类别评分仍会受到上下文效应的影响；因此，对绝对分数应谨慎解读。开发与人类判断相关性更高的无参考客观指标——尤其是针对人声和复杂混合音频的指标——仍是一个有待探索的领域 Stöter 等人(2018年)。

A.2 主观评估协议

下面将介绍评估音频分离模型时向标注员提供的说明内容。

音频分离

你将评估执行音频分离的模型。

什么是音频分离？使用音频分离模型时，用户会向模型提供音轨或带声音的视频，并描述他们希望模型提取声音中的哪些部分（例如“吉他”或“狗叫声”）。随后，模型会提取用户指定的音频部分，不包含其他任何声音。

关键术语

提示用户对目标音频的描述（例如“狗叫声”或视频中视觉突出显示的狗）。

源音频提供给模型的原始视频或音频。

目标声音源音频中用户要求提取的部分。

非目标声音源音频中用户未要求提取的所有其他声音 r.

提取 的音频模型根据源音频和提示生成的音频。

什么是“失真”？

当我们说某个声音发生了“失真”时，是指该声音与其在源音频中的原本声音相比发生了变化。

失真 失真是指源自源音频中可辨识声源的声音修改。这些修改并非由与源音频中存在的其他声音进行混音所导致。如果你无法判断某个声音是失真，还是仅仅是与源音频中的其他声音混音所产生的结果，可以将其视为失真。

失真的示例

- 不同程度的“贝斯”或“高音”：目标声音听起来可能更偏“贝斯”或不那么偏“贝斯”，或者高音调声音更多。
- 音高差异：如果声音的音高高于或低于源音频中的音高。
- 源音频中无法辨别来源的静电噪声 / 伪影 / 其他噪声：例如源音频中不存在或难以辨别的噼啪声、静电噪声、爆裂声等。
- 失真：例如，语音在源音频中听起来正常，但失真了。

概述

您将看到的模型输入和输出 在本次评估中，您将看到：

- 源音频（以及可能提供的视频）。
- 提示将采用本文档后面介绍的三种形式之一。

- 由两个独立模型提取的音频。

首先，系统会询问你几个问题，了解各个模型的提取音频与用户在提示中要求的目标声音匹配得有多好。回答完这些问题后，你还需要根据以下方面为提取音频给出总体得分：

1. 对提示的忠实度：模型在多大程度上遵循了用户关于要提取哪些声音的指令？

2. 对目标音频声音的忠实度：提取出的目标声音听起来是否确实像源音频中的目标声音？

你可能会看到的提示类型

提示类型	示例	描述
基于文本	—	用户希望模型从源音频中提取的目标声音的文本描述 从源音频中。
基于视频	—	源视频中高亮显示的部分，该部分正在产生目标声音 用户希望模型从源音频中提取内容。
基于片段	—	高亮显示 黄色 音频中包含用户希望模型提取的目标声音的片段 希望模型提取**声音**所在的位置；高亮显示 红色 音频片段 目标声音**不存在**的位置；没有片段标记的音频区域可能 包含也可能不包含目标声音。

您将要进行标注的内容

针对每个模型提取音频的调查问题 我们会要求您针对一组特定输入，提供有关每个模型性能各个方面的具体信息。首先，我们会提出一个问题来

了解提示中所请求的目标声音是否确实存在于源音频中。第一个问题与模型输出完全无关！

问题 1。 源音频是否包含提示中所请求的全部目标声音 ？

- 是
- No
- 可能，我无法判断

If t如果上述问题的答案是“是”，接下来将会问你以下问题 ：

Q2. 提取音频中包含了多少目标声音？无论其中是否存在其他非目标声音或失真，都请考虑目标声音的（1）时长和（2）数量。

- 全部
- 大部分（提取音频中包含了大部分目标声音，但仍有部分缺失）
- 一部分（提取音频中仅包含少部分目标声音，大部分缺失）
- 没有

如果答案是 Q2 不是“全部”，你将被询问目标声音缺少哪些方面 ：

Q2a. 选择一个或多个选项，说明提取音频中可能缺少目标声音的哪些方面：

- 提示中要求的一个或多个目标声音完全不存在。
- 一个或多个目标声音在其部分时长内缺失。

- 当一个或多个复杂目标声音出现在提取音频中时，仅有部分内容被提取出来（不考虑失真） [例如，如果“音乐”是目标声音，而提取音频中缺少一种乐器]。
- 其他（请注明）。

如果您对 Q2 的回答不是“无”，我们还会询问相似度：

Q3. 对于提取音频中存在的目标声音，它们听起来与源音频中的声音有多相似？

- 完全相同
- 有一些失真 / 伪影
- 有中度失真 / 伪影
- 听起来完全不同，几乎认不出来

之后，您将回答以下问题：

问题 4. 提取音频中是否存在可辨别的非目标声音？

- 是
- No

问题 5. 多　　个提取音频中的非目标声音也存在于源音频中　　？

- 全部
- 部分
- 没有

Q6. 你觉得模型从提取音频中去除非目标声音的效果如何？　　？

- 几乎完美：源音频中存在的绝大多数非目标声音都已从提取音频中滤除，且没有添加额外的非目标声音。
- 效果尚可：源音频中存在的大多数非目标声音都已从提取音频中滤除。
- 较差：进行了一些筛选，但源音频中的大多数非目标声音也大多保留在提取音频中。
- 非常差：没有筛除任何非目标声音，而且可能还出现了额外的非目标声音。

Q7. 源音频中是否有任何非目标声音在提取音频中完全未经过滤？

- 是，有多个
- 是，有一个
- 否 [提取音频中的任何非目标音频都已从源音频中至少部分筛除]

各模型提取音频的总体得分

回答完上述问题后，我们会请您按 1 到 5 分的尺度，为每个模型的表现打出总体得分。理想情况下，评分应综合并符合您在上述问题中提供的所有信息，但同样重要的是，我们也希望您运用自己的判断。

评分	标准
5：完美	所有目标声音都完整保留，听起来与其在源音频中的声音完全一致。提取音频中不存在非目标声音。
4：良好	提取音频仅存在轻微问题（例如，部分目标音频可能缺失或略有失真，可能包含极少量非目标声音）。
3：良好	提取音频存在一些严重问题：目标音频可能仅保留约一半，和/或非目标声音约保留一半。
2：较差	提取音频存在许多严重问题：提取音频中包含一些目标音频，但严重失真或缺失，并且存在多个非目标声音。
1：糟糕	提取音频完全不正确；提取音频中没有任何目标音频，或者，即使存在目标音频，也没有过滤掉任何非目标音频。

您更喜欢哪个模型的提取音频？如果您给每个模型输出的评分相同，我们还会请您根据自己的判断选出表现更好的模型。请尽可能凭自己的判断选择您更喜欢的输出，即使两个模型的表现都不理想。

- 第 1 组
- 第 2 组
- 无法决定

常见问题

不同类型的声音：你可能需要提供不同目标声音或非目标声音数量方面的输入。有时这很简单（音频中有狗叫声和猫叫声——这里有两种不同的声音），但在许多情况下，这项任务可能更难界定。例如，在一段有两个人交谈、背景中还播放着一首摇滚歌曲的音频里，算作三种声音吗：每个人的说话声各算一种，音乐算一种？还是需要将构成摇滚歌曲的每种乐器都算作一种不同的声音？

- 如果提示要求指定某种乐器或歌曲组成部分，你可以将源音频中的每种乐器视为单独的目标声音或非目标声音。
- 背景噪声：对于难以清晰辨别的环境声（如街道声音、鸣笛声、汽车驶过等，或咖啡馆中的声音），可以将其视为单一的非目标“背景”声音。

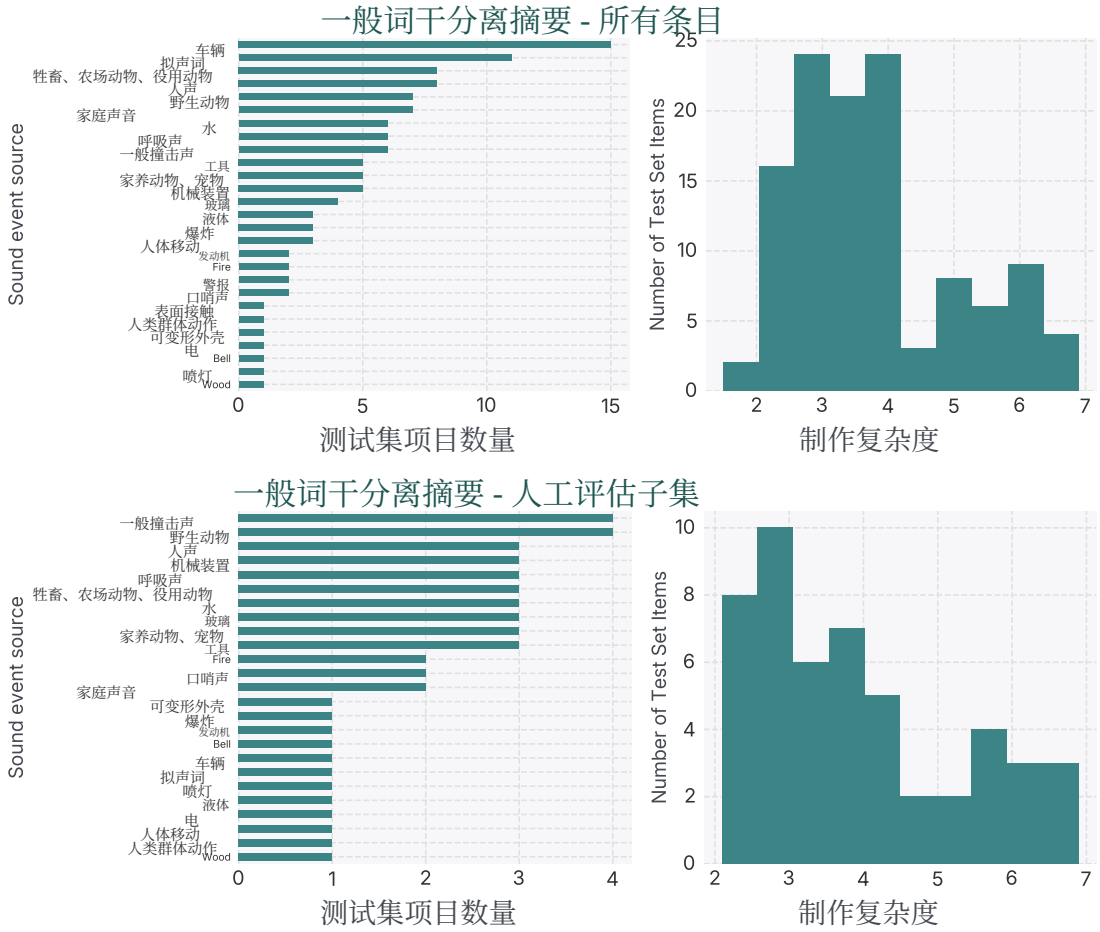

图 14 通用声音分离任务的统计数据。**左侧**：使用 AudioSet 层级结构的第二层对目标声音进行分类。**右侧**：每个基准测试条目的音频制作复杂度分布。**顶部**：SAM Audio-Bench 中所有样本的摘要；**底部**：用于人工评估的子集样本摘要。

B SAM Audio-Bench

B.1 特征描述与统计

图 14 和 16 概述了 SAM Audio-Bench 中乐器分离任务和说话人分离任务的数据分布情况。乐器方面，基准测试中的数据主要集中在少数几种常见声源上，同时也包含长尾的低频乐器；大多数音频仅包含两三种活跃乐器。说话人方面，数据集涵盖多种说话人类型和提示方式，目标说话人与干扰说话人之间的重叠程度也存在显著差异。非语音干扰声种类繁多，从环境声到背景音乐都有，反映了真实世界音频混合的复杂性。

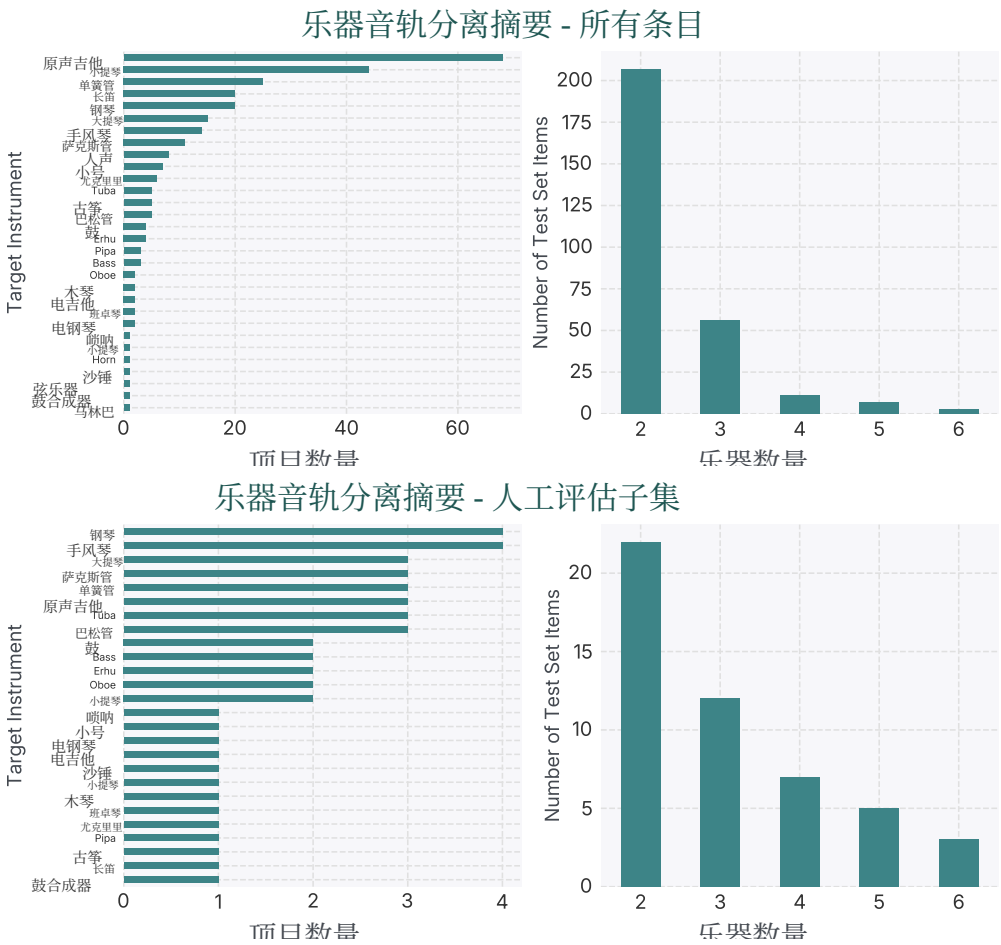

图15 SAM Audio-Bench 中乐器音轨分离任务的统计数据。**左侧**：目标乐器的直方图。**右侧**：测试集中每个项目的视频中演奏的乐器数量直方图。**顶部**：SAM Audio-Bench 中所有样本的摘要；**底部**：用于人工评估的子集样本摘要。

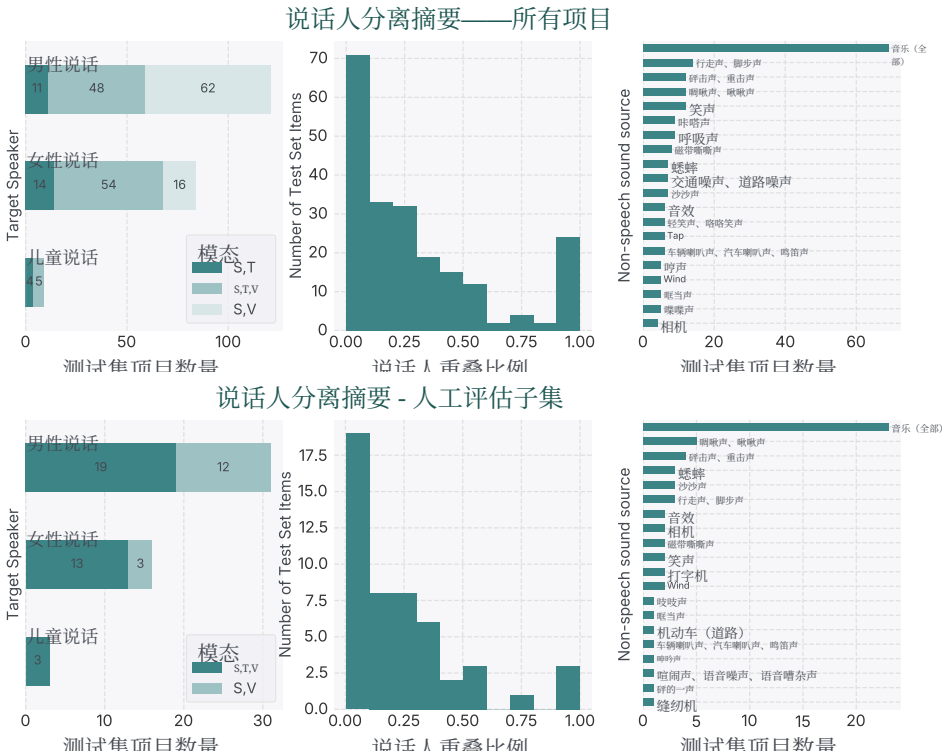

图16 SAM Audio-Bench 说话人分离任务的统计数据。**左侧**：目标说话人类型及可用于提示的模态分布（部分条目可能没有可用的视觉或文本提示）。**中间**：“说话人重叠”的分布——目标说话人时长中同时有其他语音出现的比例。**顶部**：SAM Audio-Bench 所有样本的摘要，**底部**：用于人工评估的子集样本摘要。

C SAM 音频评判器

C.1 数据标注协议

下面展示的是标注员在评审 SAM 音频评审分数时收到的说明内容。

音频分离

你将评估执行音频分离的模型。

什么是音频分离？使用音频分离模型时，用户会向模型提供一段音轨或一段带声音的视频，并描述希望模型提取声音中的哪些部分（例如“吉他”或“狗叫声”）。随后，模型会提取用户指定的音频部分，不含任何其他声音。

关键术语

提示 用户对目标音频的描述（例如“狗叫声”）。

源音频 提供给模型的原始视频或音频。

目标声音 用户请求提取的源音频片段。

非目标声音 源音频中用户未请求的所有其他声音。

提取音频 模型根据源音频和提示生成的音频。

什么是“失真”？

当我们说某个声音发生了“失真”时，是指该声音与其在源音频中的原本听感相比发生了变化。

失真 失真是指源自源音频中某个可辨识声源的声音发生了变化。这些变化并非由与源音频中其他声音混音造成。如果你无法判断某个声音是失真，还是仅仅因为与源音频中的其他声音混音而产生的结果，也可以将其视为失真。

失真的示例

- 不同程度的“贝斯”或“高音”：目标声音听起来可能更“有贝斯感”或更缺乏“贝斯感”，或者高音调声音更多。
- 音高差异：声音的音高高于或低于源音频中的音高。
- 源音频中没有明显来源的静电噪声、伪影或其他噪声：例如源音频中不存在或难以判断是否存在的噼啪声、静电噪声、爆裂声等。
- 失真：例如，源音频中的语音并未失真，但听起来却失真。

概述

模型输入和　　你将看到的输出 在此评估中，你将看到：

- 源音频（以及可能提供的视频）。
- 提示是对用户希望模型从源音频中提取的目标声音的文本描述。
- 两个独立模型提取出的音频。

首先，你需要回答几个问题，评估每个模型提取出的音频与用户在提示中要求的目标声音匹配得有多好。回答完这些问题后，你还需要根据以下方面为提取音频给出总体评分：

1. 对提示的忠实度：模型在多大程度上遵循了用户指令，提取指定的声音？2. 对目标音频声音的忠实度：提取的目标声音听起来是否确实像源音频中的目标声音？

你将要标注的内容

首先，我们会提出三个问题（Q1-Q3），评估提示中指定的目标声音是否确实出现在源音频中，以及是否还包含任何非目标声音。

如果 Q1 和 Q3 的答案都是“是”，接下来会请你回答四个问题（Q4-Q8），以评估音频分离任务的难度。注意：*Q1-Q8 与模型输出无关！*

之后，我们会请你回答几个问题（Q9-Q12），以评估模型提取音频的质量。注意：*Q9-Q12 需要模型输出！*

Q1. 你是否清楚提示所指的是源音频中的哪些声音？

- 是
- No

Q2. （如果回答“否”，请回答此后续问题）

- 源音频中没有符合提示描述的声音。
- 我不确定提示所描述的声音应该是什么样的。
- 提示含义不明确：源音频中有多个声音都可能是提示所指的声音，因此无法确定哪个是目标声音。

Q3. 源音频中是否包含提示未要求的任何非目标声音？

- 是
- No
- 可能，我无法判断

如果 Q3 的答案是“是”，接下来会向你提出以下问题：

Q4. 源音频中有多少个非目标声音？

说明：非目标声音是指提示中未要求、但清晰可听到的任何声音。同一类型声音的多个实例（例如反复咳嗽）计为一个。无法辨认、但明确来自不同声源的声音（例如语音与机械声）应分别计数。完全无法辨认或含糊不清的背景声音应合并计为一个非目标声音。

- 1
- 2
- 3
- 4
- 5+

Q5. 目标声音与非目标声音有多大程度的重叠？

- 目标声音可听见时，不会同时听见非目标声音。

- 目标声音与非目标声音有少量重叠（< 25%）。
- 目标声音在约一半的时长内与一个或多个非目标声音重叠。
- 目标声音在大部分时长内都与一个或多个目标声音重叠（~75%）
- 所有目标声音都与一个或多个非目标声音完全重叠。
- 目标声音含糊不清，所以我无法判断。

Q6. 目标声音与非目标声音相比有多响?

指南：评估整个音频中目标声音相对于非目标声音的总体感知响度。如果相对响度随时间变化，请重点关注目标声音的主要趋势或最突出的部分。

- 比非目标声音响得多。
- 比非目标声音稍响。
- 响度大致相同。
- 比非目标声音稍轻。
- 比非目标声音轻得多。
- 目标声音不明确，因此无法判断。

Q7. 非目标声音与目标声音有多容易混淆?

指南：当非目标声音与目标声音具有相似的声学特征（例如音高、音色、时序），或出现在相近的时刻时，可能会造成混淆。例如，如果提示是"婴儿哭声"，而背景中有一声猫叫短暂地听起来像婴儿哭声，那么根据听起来有多相似，可能会造成轻微或中度混淆。

- 不混淆（非目标声音与目标声音截然不同。）
- 轻微混淆（非目标声音大多可以区分，但有时可能短暂地听起来像目标声音。）
- 中度混淆（非目标声音的某些部分听起来像目标声音，偶尔会造成混淆。）
- 严重混淆（非目标声音常常难以与目标声音区分。）
- 极易混淆（非目标声音与目标声音几乎无法区分。）

Q8. 综合考虑 Q4-Q7 中列出的所有因素，你认为从源音频中提取目标声音有多困难?

指南：当非目标声音与目标声音具有相似的声学特征（例如音高、音色、时序），或出现在相似的时刻时，可能会产生混淆。例如，如果提示是"婴儿哭声"，而背景中有一只猫叫声短暂地听起来像婴儿哭声，那么根据声音相似的程度，可能会造成轻微或中等程度的混淆。

- 1: 非常容易（目标声音响亮、清晰，与非目标声音区别明显且几乎没有重叠。非目标声音非常少。）
- 2: 容易
- 3: 中等
- 4: 困难
- 5: 非常困难（目标声音很轻，与非目标声音相似且大量重叠。非目标声音很多）

以下　　这些问题需要模型输出！接下来，你将被问到以下问题　　:

Q9. 无论是否存在其他非目标声音或失真，提取音频中包含了多少目标声音？请从以下两个方面考虑：**(1)** 时长；**(2)** 目标声音的数量。

- 全部
- 大部分（提取音频中包含大部分目标声音，但有一部分缺失）
- 一部分（提取音频中包含一半目标声音，另一半缺失）
- 少部分（提取音频中包含少部分目标声音，但大部分缺失）
- 没有

如果你对“提取音频中包含多少目标声音”的回答不是“没有”，我们还会询问提取音频中的目标声音听起来与源音频中的声音有多相似。

Q10. 对于提取音频中存在的目标声音，它们听起来与在源音频中的声音有多相似?

- 完全相同
- 轻微失真/伪影
- 中度失真/伪影
- 严重失真/伪影
- 听起来完全不同，几乎无法辨认

之后，你将被要求

Q11. 你认为模型从提取音频中移除非目标声音的效果如何?

- 非常好：提取音频中没有任何非目标声音。
- 近乎完美：源音频中的大部分非目标声音都已被筛选掉，只剩下极少量非目标声音。
- 尚可：源音频中的部分非目标声音已被移除，但提取音频中仍残留明显一部分。
- 较差：筛选效果有限；源音频中的大部分非目标声音仍存在于提取音频中。
- 极差：没有移除任何非目标声音，提取过程中还可能引入额外的非目标声音。

Q12. 总体得分

回答完以上问题后，我们会请您按 1 到 5 分的标准，对模型的表现进行总体评分。理想情况下，评分应综合考虑并与您以上提供的所有信息保持一致；但同样重要的是，我们希望您运用自己的判断。

常见问题

不同类型的声音：系统可能会要求你提供不同目标声音或非目标声音数量的信息。有时这很简单（音频中有狗叫声和猫叫声——这里有两种不同的声音），但在许多情况下，这项任务可能更难以明确判断。例如，在一段有两个人说话、背景中还播放着一首摇滚歌曲的音频里——这算三种声音吗：每个人的说话声各算一种，音乐算一种？还是需要把构成这首摇滚歌曲的每种乐器都分别算作一种不同的声音?

评分	标准
5：完美	所有目标声音均完整呈现，且与其在源音频中的声音完全一致。提取音频中不存在非目标声音。
4：良好	提取音频仅存在轻微问题（例如，部分目标音频可能缺失或略有失真，也可能包含极少量非目标声音）。
3：合格	提取音频存在一些严重问题：目标音频可能只保留了一半，和/或非目标声音大约保留了一半。
2：较差	提取音频存在许多严重问题：提取音频中包含一些目标音频，但声音严重失真、缺失，而且还包含几种非目标声音。
1：糟糕	提取音频完全不正确；提取音频中没有任何目标音频，或者，即使有，也没有过滤掉任何非目标音频。

- 如果提示要求识别特定乐器或歌曲组成部分，你可以将源音频中的每种乐器视为单独的目标声音或非目标声音。
- 背景噪声：对于难以辨别的环境声音（街道声音、鸣笛声、汽车驶过等；或咖啡馆的声音），可以将其视为一种非目标“背景”声音。

语音提示：你可能会看到一些提示，例如“其中一位男性在说话”“其中一位女性在说话”或“其中一人在说话”。在这些情况下，请在整段音频中始终关注同一位说话人。

- 例如，如果提示是“其中一位男性在说话”，而提取音频中有两位男性在说话，那么由于与提示不一致，总体得分应评为较低。
- 由于可能有多个候选样本，你可以在 Q5 和 Q6 中选择“目标声音含糊不清，因此我无法判断”。

音量：也是影响评分的重要因素。

- 对于 Q10，改变目标声音的音量应视为失真。
- 对于 Q11，降低非目标声音的音量应视为一种有效的移除方式 al.

失真：如果提取音频的质量听起来比源音频更好，那么在回答 Q10 时也应将其视为一种失真。

C.2 SAM 音频评判器中的高质量评分者筛选

我们设计了一套评分员资格认证计划，以确保选出高质量的标注员。具体而言，我们整理了一个包含 40 个样本的黄金集，其评分由专家标注并作为真实值。随后，我们邀请外包供应商对该黄金集进行标注，并使用贝叶斯评分员建模方法对其标注员进行资格认证。该模型会估计每位标注员混淆矩阵的后验分布，从而反映他们在不同评分值上的可靠性和偏差。基于此后验分布，我们推导出每位标注员评分的预期平均绝对偏差（MAD），这一指标能够自然地考虑未完成的任务、项目评分的不确定性以及分数分布的不均衡。我们选取总体得分预期 MAD 最低的标注员作为合格评分员。虽然我们不会直接将评分与专家标注进行比较，但会将专家标注用于抽查，以确保选出的标注员具有合理的共识度和一致的评分行为。最终，我们成功招募了 128 名合格评分员，代表了普通公众中多元的主观视角。

C.3 Gemini-2.5-pro 评审评估提示

你将评估执行音频分离的模型。
使用音频分离模型时，用户会向模型提供一段音轨，并描述希望模型提取声音中的哪些部分（例如“吉他”或“狗叫声”）。随后，模型会从音频中提取用户指定的部分，不包含其他声音。

用户对目标音频的描述（例如“狗叫声”或视频中视觉突出显示的狗）称为提示。提供给模型的原始视频或音频称为源视频/音频。用户希望从源音频中提取的部分称为目标声音。源音频中用户未指定的其他声音称为非目标声音。模型根据源音频和提示生成的音频称为提取音频。

任务：

- 用户提供：1. 一条提示，即对用户希望模型从源音频中提取的目标声音的文本描述；2. 两个音频文件：– 源音频（包含目标声音和其他声音）– 音频分离模型输出的提取音频

请从以下四个维度，结合混合音频和目标描述评估提取音频：

- **召回率**：目标声音中有多少出现在提取音频中，无论其中是否存在其他非目标声音或失真？请从以下两方面考虑其占比：(1) 时长；(2) 目标声音的数量。
- **精确率**：对于提取音频中出现的目标声音，它们听起来与源音频中的声音有多相似？
- **忠实度**：你认为模型从提取音频中去除非目标声音的效果如何？
- **总体**：请为模型的性能提供总体得分。

请为每个维度提供 1（**非常差**）到 5（优秀）的评分：

- **1: 糟糕**：提取音频完全不正确；提取音频中不含任何目标音频，或者即使包含，也没有过滤掉任何非目标声音。
- **2: 较差**：提取音频存在许多严重问题：其中包含部分目标音频，但失真严重、内容缺失，而且存在许多非目标声音。
- **3: 尚可**：提取音频存在一些严重问题：目标音频可能只保留了一半，和/或非目标声音约有一半仍然存在。● **4: 良好**：提取音频仅存在一些小问题（例如，可能缺少少量目标音频，或有轻微失真；也可能存在极少量非目标声音）。
- **5: 完美**：所有目标声音均完整保留，且听起来与源音频中的声音完全一致。提取音频中不存在非目标声音。

我们还提供了 5 个参考示例，说明评分 1、2、3、4、5 分别代表什么。

C.4 用于自动化和细粒度性能分析的 SAJ

我们可以使用 SAJ，针对任意目标音频概念列表，自动抽取不同难度级别的评估样例，从而实现完全自动化的细粒度性能分析流程。为此，我们训练了一个文本提示式 SAJ 模型，用于预测分离任务的内在难度。由于人工标注显示，4 级和 5 级的样例较少，因此我们将它们合并为一个四级量表（1–4），并使数据集在各级别间保持均衡。

与标准 SAJ 模型不同，后者使用混合音频和输出音频来预测分离性能；难度模型只将混合音频和文本提示作为输入，因此能够在运行任何分离系统之前估计任务难度。

我们将预测的难度级别应用于 SAJ 测试集，并对相应的 SAM Audio 输出进行人工评分。如表 18 所示，人工标注性能随着难度增加而单调下降：1 级案例最简单，而 4 级则表现出明显退化。这使得我们能够自动筛选出不同难度级别的评估子集，从而支持细粒度的概念特定鲁棒性分析。

难度级别	总体	召回率	精确率	忠实度
1	3.716	4.084	3.984	3.963
2	3.327	3.995	3.537	3.798
3	3.272	4.022	3.326	3.807
4	2.894	3.791	3.024	3.528

表18 不同任务难度级别下人工评定的分离质量。

D SAM Audio 消融实验

D.1 模型规模的影响

表 19 和20，以及图 17，对比了三种规模的模型——拥有 5 亿、10 亿和 30 亿参数的模型。在规模对比中，我们禁用了跨度预测。总体而言，30亿参数模型在大多数任务中性能最佳，不过在某些情况下（如通用音效分离），其表现与较小模型持平或略逊一筹。

扩大模型容量在专业领域带来的收益最大。例如，在乐器分离任务中，SAM Audio-Large 相较于较小的变体取得了显著提升：在真实场景乐器分离任务中，其 NWR 比 SAM Audio-Base 高 23%，比 SAM Audio-Small 高 20%。这些结果表明，更大的模型能够更好地捕捉结构化领域（如乐器）中实现高保真分离所需的细粒度声学线索，而较小的模型在更广泛的声音类别上仍具竞争力。

	通用音效			语音			说话人			音乐			野外乐器			专业乐器		
模型	SAJ CLAP	总体主		SAJ CLAP	总体主		AJ CLAP	总体主观		CLAP	总体主观评		LAP	总体主观评分		AP	总体主观评分	
SAMAUDIO-SMALL	**4.25**	0.30	**3.62**	4.55	**0.35**	3.99	3.89	**0.17**	3.12	4.32	**0.28**	4.11	4.27	0.27	3.56	4.78	**0.30**	4.24
SAM Audio-Base	4.23	0.28	3.28	**4.61**	0.33	**4.25**	3.94	0.15	3.57	4.26	0.27	3.87	4.33	0.29	**3.66**	4.78	**0.30**	4.27
SAM Audio-Large	4.11	**0.31**	3.50	4.59	0.33	4.03	**4.08**	**0.17**	**3.60**	**4.30**	**0.28**	**4.22**	**4.45**	**0.30**	**3.66**	**4.83**	0.28	**4.49**

表19 不同规模 SAM Audio 在文本提示下的比较。–：不适用。总体主观评分：总体主观评分。

	通用音效		说话人		指令（野外）	
模型	IB	OVR	IB	OVR	IB	OVR
SAM Audio-Small	0.24	2.62	0.23	2.79	0.21	2.25
SAM Audio-Base	**0.25**	**2.63**	0.24	**3.25**	0.22	**2.76**
SAM Audio-Large	**0.25**	2.61	**0.24**	2.95	**0.24**	2.58

表 20 视觉提示中不同规模 SAM Audio 的比较。–：不适用。OVR：总体主观评分。

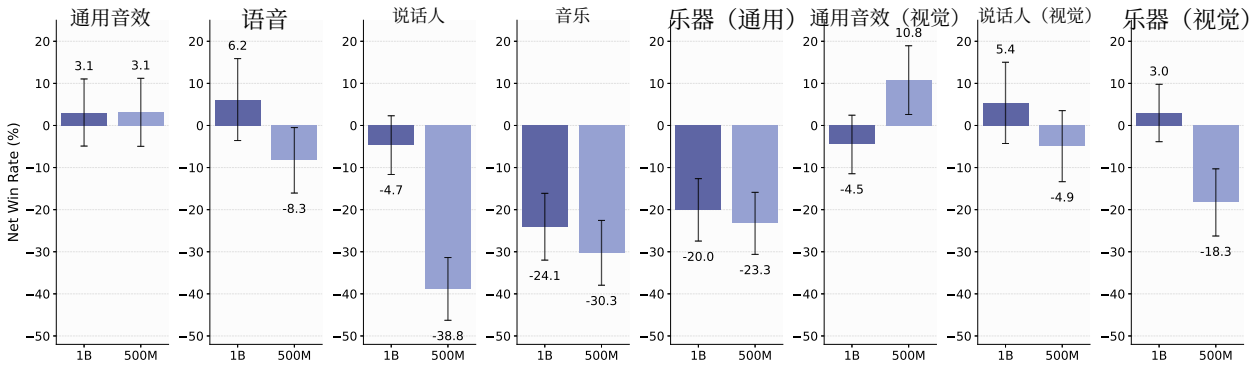

图 17 净值　　SAM Audio-Base 和 SAM Audio-Small 相对于 SAM Audio-Large 的胜率（%）

D.2 辅助损失的影响

表 21 报告了在 SAM AUDIO 预训练期间加入表示对齐损失的效果。我们比较了两个 30亿参数预训练检查点：(1) 不使用辅助损失的基线模型；(2) 使用辅助损失并通过 $\lambda = 1.0$进行训练的模型。我们在通用音效测试集上评估了两个检查点在文本提示和视觉提示分离任务中的表现，因为预训练模型单独用于语音和音乐等专业领域时性能较差。

	通用音效（文本）		通用音效（视觉）
辅助目标	SAJ	CLAP	IB
None	2.48	0.24	0.17
目标AED	**3.18**	**0.29**	**0.18**

表21 基于AED的表示对齐损失对预训练的影响

		Text						视觉		
分离任务	**阶段**	通用音效	语音	说话人	音乐	野外乐器	专业乐器	通用音效	说话人	野外乐器
SAJ (↑)	PT	3.93	2.90	3.28	4.14	3.17	3.36	-	-	-
	FT	**4.14**	**4.55**	**4.07**	**4.38**	**4.48**	**4.82**	-	-	-
CLAP (↑)	PT	**0.31**	0.28	**0.23**	**0.31**	0.21	0.15	-	-	-
	FT	0.30	**0.36**	0.21	**0.31**	**0.30**	**0.29**	-	-	-
指令型基线 (↑)	PT	-	-	-	-	-	-	**0.24**	0.22	0.20
	FT	-	-	-	-	-	-	**0.24**	**0.24**	**0.22**
AES-PC (↓)	PT	2.57	2.91	2.98	4.53	3.54	4.72	3.09	3.49	3.69
	FT	**2.08**	**1.89**	**1.94**	**4.44**	**3.16**	**3.24**	**2.22**	**2.20**	**3.26**

表22 音频分离任务中预训练与微调结果的比较。–：不适用。

如表 21所示，加入基于 AED 的对齐目标后，两种设置下的性能均得到稳定提升；与视觉提示相比，文本提示情况下的增益更大（文本对齐相对提升超过 20%）（∼5% 相对提升）。该消融实验仅在预训练期间进行，因为我们在微调阶段不应用辅助损失。我们推测，这一收益源于预训练语料库中更嘈杂的目标音频：对齐损失有助于模型学习对分离任务有益的中间语义表示。

D.3 微调的效果

表22比较了仅经过预训练的30亿参数模型与在精选分离数据上进一步微调的同一模型。除标准指标外，我们还报告了 Audiobox-Aesthetics 的 PC 分数（Tjandra 等人，2025年），该分数可作为音频清洁度的替代指标。

总体而言，微调带来了稳定的提升，不过提升幅度因任务而异。对于文本提示分离，乐器提取、语音提取和说话人分离的提升最为显著。这些任务受益于高质量数据集的可用性；这些数据集包含专业录制的音轨，能够提供干净可靠的监督信号。相比之下，通用声音事件分离和音乐分离的提升较小。大规模预训练已经涵盖了广泛分布的声音事件和混合音频，其中包括许多带有背景音乐的混合音频，因此微调能够带来额外提升的空间较小。

对于视觉提示分离，通用 SFX 的视觉对齐分数（指令型基线）保持相对稳定，这与文本提示分离中观察到的趋势一致。与通用声音事件相比，微调数据对音乐和语音视频的覆盖度更广。大规模音频视觉预训练已经在视觉区域与音频之间建立了强关联，这解释了 SFX 所获得的增量收益较小。

最后，根据 PC 指标，我们观察到各项任务的音频清洁度均有显著提升。经过微调的模型始终能生成更干净、伪影更少的分离音频，这得益于微调过程中使用的干净音频目标。

D.4 使用伪标注音频源数据的效果

表 23比较了在微调阶段使用与不使用伪标注音频数据的效果。伪标注数据的详细信息汇总于表 4。基线模型仅使用完全真实的三元组和合成混合音频进行训练。如表 23所示，加入伪标注数据后，结果始终

文本提示和视觉提示均有所提升。值得注意的是，通用声音的 AES-PC 提升最为显著，这表明伪标注音频有助于模型生成更干净的分离音轨。

		Text						视觉		
分离任务	**设置**	通用音效	语音	说话人	音乐	野外乐器	专业乐器	通用音效	说话人	野外乐器
SAJ (↑)	无伪标签	4.02	4.50	4.06	4.34	4.34	4.80	-	-	-
	PL	**4.14**	**4.55**	**4.07**	**4.38**	**4.48**	**4.82**	-	-	-
CLAP (↑)	无伪标签	**0.30**	0.34	**0.21**	**0.31**	0.29	0.28	-	-	-
	PL	**0.30**	**0.36**	**0.21**	**0.31**	**0.30**	**0.29**	-	-	-
指令型基线 (↑)	不使用伪标签	-	-	-	-	-	-	0.23	0.21	**0.22**
	PL	-	-	-	-	-	-	**0.24**	**0.24**	**0.22**
AES-PC (↓)	不使用伪标签	2.36	1.95	2.02	**4.38**	3.19	**3.24**	2.28	2.29	**3.25**
	PL	**2.08**	**1.89**	**1.94**	4.44	**3.16**	**3.24**	**2.22**	**2.20**	3.26

表 23 使用伪标注音频声部分数据的效果