LabelBuddy：一款使用人工智能辅助的开源音乐与音频语言标注工具

Ioannis Prokopiou [1,2], Ioannis Sina[3], Agisilaos Kounelis[3],Pantelis Vikatos[2],Themos Stafylakis[1,4]

[1]雅典经济与商业大学，[2]Orfium，[3]帕特雷大学，[4]Archimedes/Athena研究中心，gian.prokopiou@aueb.gr, sinaioannis@gmail.com agis@ceid.upatras.gr, pantelis@orfium.com,tstafylakis@aueb.gr

arXiv:2603.04293v1 [cs.SD] 4 Mar 2026

摘要

机器学习（ML）、大型音频语言模型（LALMs）以及自主人工智能代理在音乐信息检索（MIR）领域的进展，要求从静态标注转向丰富的人类对齐表示学习。然而，能够捕捉音频标注主观细微差别的开源基础设施仍然稀缺，这构成了关键瓶颈。本文介绍**LabelBuddy**，一款开源的协作式自动标注音频标注工具，旨在弥合人类意图与机器理解之间的差距。与静态工具不同，它通过容器化后端将界面与推理解耦，使用户能够接入自定义模型，以进行AI 辅助预标注。我们描述了该系统架构，其支持多用户共识、容器化模型隔离，以及扩展代理与大型音频语言模型的路线图。代码见 https://github.com/GiannisProkopiou/gsoc2022-Label-buddy。

1 引言

用于训练 AI 模型的数据集质量，是影响准确性、可靠性和泛化能力的重要因素(Picard 等人，2020年)。尽管 Zenodo 等公共存储库[1]和数据湖 (Espinal 等人，2022年)，以及 DCASE[2] 和 MIREX[3],等基准测试推动了标准化工作，但现有资源往往针对特定任务。此外，创建专用数据集可能颇具挑战，过程复杂、耗时，通常还需要大量工作并由人工审核 (Voigtlaender 等人，2021年)。延续跨学科影响的传统，自然语言处理 (NLP) 与音乐和音频的交汇开辟了一片新领域，在这里，语言与声音相互融合。大多数音频内容都具有内在的语言维度，但创建能够捕捉这些多模态协同的数据集，仍然是一个耗费大量人力的瓶颈。

音乐信息检索（MIR）领域目前正经历一场转型：从以静态标签分类为特征的判别式范式，转向生成式和基于推理的方法。MusicFlamingo（Ghosh 等人，2025年）、Qwen-Audio（Chu 等人，2023年）以及 Audio Flamingo 3（Goel 等人，2025年）等大型音频语言模型的兴起，带来了思维链推理和对话式音频理解等新能力。然而，这些模型的效能在很大程度上取决于其与人类意图对齐的质量。近期的综述指出，客观指标往往无法捕捉审美细微差异，因此有必要转向基于人类反馈的强化学习（RLHF）和严格的主观评估方法（Kader 和 Karmaker，2025年）。

当前工作流往往彼此割裂，用户不得不采用互不连贯的工作流，将数据整理与关键的人工主观评估阶段（例如 MUSHRA、GoListen 或成对偏好测试）分开进行（Schoeffler 等人，2018年；Barry 等人，2021年）。用户使用基于波形的工具进行分段（Grover 等人，2020年），使用独立平台处理文本，并采用专门软件进行主观评估（例如 WebMUSHRA（Schoeffler 等人，2018年））。这种割裂阻碍了高效的人在回路（HITL）流程的发展；在这类流程中，模型输出的不确定性应当指导数据采集。此外，生成式音乐领域的“指标危机”——客观评分（如 FAD）无法与人类感知相关联（Kader 和 Karmaker，2025年）——要求开发能够无缝衔接标注与主观偏好排序的工具。

为了解决这一问题，我们推出**LabelBuddy**，这是一款开源、协作式自动标注音频工具，配备人工智能辅助功能，包括：

[1]https://zenodo.org/ [2]https://dcase.community/ [3]https://www.music-ir.org/mirex

1. **解耦式人工智能辅助：** 通过隔离的容器化架构，使用声明式 YAML 文件注入模型预测结果。我们提供 YOHO（Venkatesh 等人，2022年）、musicnn（Pons 和 Serra，2019年）、PANNs（Kong 等人，2020年）等预训练模型，以及 Music Flamingo（Ghosh 等人，2025年）等大型音频语言模型，用于人工智能辅助的预标注标签，将用户精力从创建转向核验；经批准的标签还可用于微调模型。

2. **协作共识：** 原生支持多用户角色（管理员、标注者、复核者），以确保真实标注可靠性。

3. **混合工作流支持：** 该架构旨在同时支持基于区域的标注和主观偏好聚合。

2 相关工作

本节回顾标注平台、面向大型音频语言模型的 HITL 工作流，以及用于主观评估和基于人类反馈的强化学习的基础设施。

标注平台与领域特异性。数据整理的方式因模态而异。对于文本，BRAT（Stene- torp 等人，2012年）、Paladin（Nghiem 等人，2021年）和 PubAnnotation（Kim 和 Wang，2012年）可用于语言标注，而 CAT（Bartalesi Lenzi 等人，2012年）和 MDSWriter（Meyer 等人，2016年）等专用框架则分别用于提升语义效率和生成摘要。文本标注领域也探索了主动学习策略（例如 ActiveAnno（Wiechmann 等人，2021年）、APLenty（Nghiem 和 Ananiadou，2018年））。在视觉领域，VIA（Dutta 和 Zisser- man，2019年）和 Annotation Web（Smistad 等人，2021年）等工具展现了领域专用界面的必要性。

在音频领域，Audino（Grover 等，2020年）和 BAT（Meléndez Catalán 等，2017年）在声音事件检测（SED）和显著性等时间任务方面表现出色，而基于库的解决方案（如 Aubio（Brossier 等，2019年）则便于特征提取。其他工具，如 Gecko（Levy 等，2019年），则专注于语音分割。然而，这些工具通常缺乏现代推理模型使用所需的解耦式 AI 架构。相比之下，Label Studio（Tkachenko 等，2020-2022）和 Prodigy（Montani 和 Honnibal，2018年）等通用 HITL 平台提供了强大的后端，但通常会将审阅者角色和共识指标等协作功能限制在付费企业层级中

表1：LabelBuddy 与现有工具的比较。

Tool	音频特定	解耦 AI 辅助	Open 来源	协作实验室 Consensus
Audino	✓	-	✓	-
BAT	✓	-	✓	-
Aubio	✓	-	✓	-
Gecko	✓	-	✓	-
Prodigy	-	✓	-	-
Label Studio（社区版）	-	✓	✓	-
LabelBuddy	✓	✓	✓	✓

此外，它们缺乏对专业音频工具（Cartwright 等，2017年）中常见的音乐结构（例如小节、节拍）的原生支持。LabelBuddy 与其他现有工具的比较见表1。

大型音频语言模型与人在回路工作流。当前最先进技术已从固定词汇表自动标注器转向大型音频语言模型（大型音频语言模型），例如 Audio Flamingo 3（Goel 等人，2025年）和 Qwen-Audio（Chu 等人，2023年）。这些模型采用统一编码器进行“思维链”推理。为使其与人类判断保持一致，我们采用“单次迭代”人在回路方法。近期关于视频标注的研究（Gutiérrez 等人，2025年）表明，简单的模型辅助预标注能够缩短任务耗时，同时不降低质量；我们将这一理念同样应用于音频，方法类似于 NEAL（Gibbons 等人，2023年）。

主观评估与基于人类反馈的强化学习。生成式音乐面临的一个关键瓶颈是“指标危机”：弗雷歇音频距离（FAD）等分数无法与人类感知相关联（Kader 和 Karmaker，2025年；Gui 等人，2023年）。因此，该领域正转向基于人类反馈的强化学习（Cideron 等人，2024年；Liu 等人，2025年）。目前，评估与标注彼此分离，依赖 Web-MUSHRA（Schoeffler 等人，2018年）等独立工具。LabelBuddy 旨在通过整合成对偏好聚合方法（如贝叶斯Bradley-Terry（BBQ）（Aczel 等人，2025年）），将两者统一起来。

3 系统架构

LabelBuddy 通过模块化的容器化架构，解决了标注工具中的“耦合问题”：界面与特定模型后端之间存在硬编码关联。如图1所示，该系统将轻量级用户交互层（Django）与计算密集型推理层（Docker）解耦。

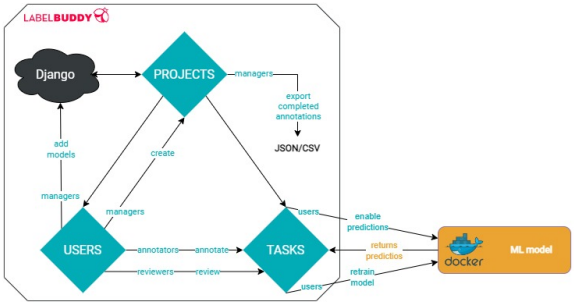


图1：**系统架构概览**：该架构将 Django Web 服务器与 Docker 化的机器学习推理解耦。

3.1 后端与数据模型

核心应用基于 Django 构建，并使用关系型数据库管理三个主要实体：项目、用户和任务。

• **基于角色的访问控制与隐私：** 为防止数据泄露，系统实施基于角色的访问控制（RBAC）。管理者拥有全面监督权限，而标注员和审核员仅能访问分配给自己的任务队列。

• **数据序列化：** 标注以 JSON 对象的形式存储，其中包含时间边界和本体标签。管理者可以将共识数据导出为 CSV/JSON格式，以便直接集成到机器学习训练流水线中。

3.2 容器化推理引擎

管理者通过 YAML 配置文件定义模型，指定 Docker镜像、输入/输出架构和资源约束。当请求 AI 辅助时，后端通过 RESTful Flask API 与模型容器通信。该设计确保了沙箱隔离（模型在独立环境中运行）和可扩展性（推理可以部署到 AWS/Azure 等远程云节点）。

4 工作流与界面

该平台支持完整的“人在回路中”（HITL）生命周期。

4.1 项目设置与任务管理

工作流从**仪表板**开始，管理者可在此创建项目并分配用户角色。

• **模型集成：** 在**模型页面**中，管理人员上传 YAML 配置文件以附加推理容器。此界面提供高级控制功能：监控训练损失/准确率、下载权重文件，以及使用项目经过验证的数据触发微调任务（指定训练轮次和学习率）。

•**任务导入：** 管理者通过**项目页面**上传音频（WAV/MP3），然后通过共享池或互斥分配策略将任务分发给标注员。

4.2 标注循环

核心标注工作流程利用 wavesurfer.js 实现响应迅速的波形可视化。

1.**AI 辅助预标注：** 标注员触发“按需预测”，将音频序列化后发送至当前 Docker 容器。系统会将返回的预测结果呈现为可编辑区域（图 2），使人工任务从创建转向验证。

2. **审核与 Consensus：** 已完成的任务会进入**审核界面**，审核员可以回放特定区域，并通过反馈批准或拒绝标注。这一质量保证（QA）流程对于创建用于生成式对齐的高保真数据集至关重要。

5 案例研究：NLP 音乐标注

为展示 LabelBuddy 的实用性，我们介绍一种用于创建音乐描述数据集的参考工作流程。此任务要求将音频信号与丰富的自然语言描述对齐。

模型集成 项目经理定义一个封装多模态模型（例如 Music Flamingo 检查点）的 Docker 容器。YAML 配置将模型的文本输出映射到 LabelBuddy 标注区域：

图像：“my-repo/music-flamingo:v1” 输入_架构：{ "audio": "wav" } 输出_架构：- { "type": "text", "label": "说明文字" } 资源：{ "gpu": " true" }

“人工验证”循环。 标注员会看到一个原始音轨队列。与其从头开始撰写描述（这在认知上负担很重），他们不妨触发**预标注**功能。后端容器会处理音频，并返回候选文本描述：“一段带有慢节奏和黑胶唱片噼啪声的低保真嘻哈音轨。”

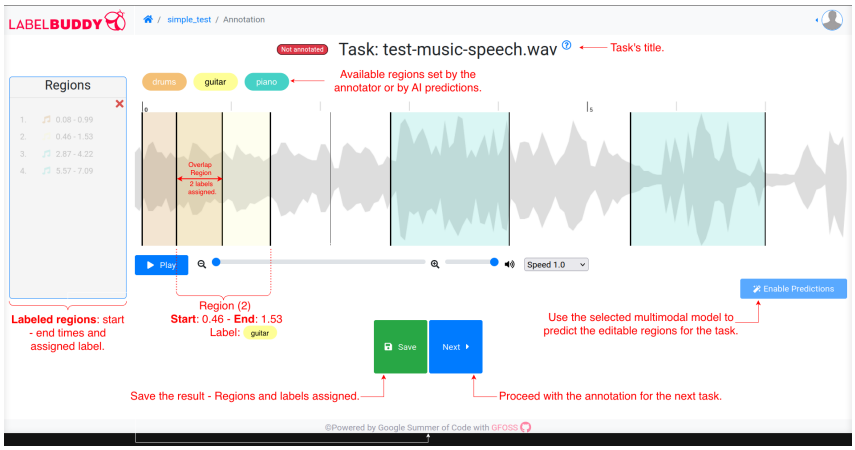


图 2：**标注界面**，以可编辑的波形区域显示 AI生成预测。

修正与 Consensus。 标注员会修正特定的幻觉（例如，将“黑胶唱片噼啪声”改为“雨声”），并调整时间戳边界。如果多名标注员处理同一音轨，审阅界面会突出显示文本描述中的语义分歧，从而能够创建基于共识的可靠真实标签。

多模态导出。 最终数据集将导出为 JSONL 或 CSV 文件，其中包含对齐的（音频_路径、文本_描述）配对，可立即用于下游音频到文本生成模型的微调。

6 讨论与未来路线图

LabelBuddy 当前版本解决了迫在眉睫的基础设施挑战：将标注前端与快速演进的模型后端解耦。随着该领域迈向大型音频语言模型和自主智能体，我们未来的路线图将致力于弥合人类意图与机器表征之间的差距：

智能体推理。 传统主动学习依赖不确定性采样，而大型音频语言模型的兴起要求我们转向对话式协助。我们正在扩展后端API，以支持 Qwen-Audio 等推理能力更强的模型（Chu 等人，2023年）。未来版本将允许标注员向模型提问并获取“思维链”解释，从而将工作流从简单的标签核验转变为协作推理。这与近期研究发现一致：交互式推理能够减少复杂标注任务中的幻觉。

集成式主观评估（基于人类反馈的强化学习）。 承认“指标危机”：FAD 分数无法捕捉美学质量（Kader 和 Karmaker，2025年），LabelBuddy 旨在发展成为基于人类反馈的强化学习工作台。我们的目标是实现原生的“成对偏好”界面（数据集 A 与数据集 B），并将其直接集成到审核流程中。为处理嘈杂的人类评分者，后端将集成贝叶斯 Bradley–Terry（BBQ）模型（Aczel 等人，2025年），提供稳健的偏好聚合，以使生成模型保持一致（Cideron 等人，2024年）。

提升感知有效性。 为纠正模型倾向于依赖文本先验而非音频内容的问题——RULis-tening 基准测试指出了这一缺陷（Zang 等人，2025年），我们计划引入需时间戳问答模板。这些模板将促使模型和人工标注员都以特定频谱区域为依据，支撑每一项语义判断，确保未来的数据集能够推动真正的听觉感知，而非纯文本推理。

评估计划。 为验证其效用，我们建议使用 DCASE 2024 数据开展一项试点研究，测量：（a）相较于从头标注所节省的时间；（b）标注者间一致性（Fleiss Kappa）；以及（c）使用 LabelBuddy 整理的数据训练的基线声音事件检测模型在 PSDS 上的提升。

7 结论

LabelBuddy 是探索语言与音频之间多模态协同的重要基础设施。通过将界面与推理过程解耦，它使社区能够整理丰富的语言基础数据集，为现代 NLP驱动的音乐理解提供所需的数据。无论是用于常规标注，还是满足新兴的基于人类反馈的强化学习（RLHF）需求，LabelBuddy 都提供了开放、可扩展的工作台，以加深人类对音频的感知与机器表征之间的联系。

8 伦理声明

AI辅助标注工具的发展引发了人们对劳动力替代和偏见的担忧。LabelBuddy 旨在增强而非取代人类专业知识，让人在回路中对关键判断进行把关。此外，我们希望通过促进开放数据集的创建，推动高质量训练数据的获取民主化，避免资源集中于大型科技公司。我们确保所有集成模型的使用均符合其研究许可。

致谢

本工作得到谷歌编程之夏（GSoC）和开放技术联盟（GFOSS）的支持，并由欧盟地平线欧洲研究与创新计划在AIXPERT项目（资助协议编号101214389）下资助，该项目旨在开发一个由生成式人工智能驱动的智能体式多层框架，用于创建可解释、可问责且透明的人工智能系统。

参考文献

Till Aczel、Lucas Theis和Wattenhofer Roger。2025年。基于噪声成对比较的高效贝叶斯推理。*arXiv*预印本 *arXiv:2510.09333*。

Dan Barry、Qijian Zhang、Pheobe Wenyi Sun和Andrew Hines。2021年。Go Listen：一个端到端在线听力测试平台。《开放研究软件期刊》，9(1)。

Valentina Bartalesi Lenzi、Giovanni Moretti 和 Rachele Sprugnoli。2012年。Cat：CELCT 标注工具。收录于第八届语言资源与评估国际会议论文集（*LREC'12*），第 333–338 页。LREC。

PaulBrossier、Tintamar、Eduard Müller、NilsPhilippsen、Tres Seaver、Hannes Fritz、cyclopsian、Sam Alexan- der、Jon Williams、James Cowgill 和 Ancor Cruz。2019年。aubio/aubio：0.4.9。

Mark Cartwright、Ayanna Seals、Justin Salamon、Alex Williams、Stefanie Mikloska、Duncan MacConnell、Edith Law、Juan P Bello 和 Oded Nov。2017年。聆听声音：探究可视化效果和复杂度对众包音频标注的影响。《*ACM* 人机交互汇刊》，1（CSCW）：1–21。

Yunfei Chu、JinXu、XiaohuanZhou、Qian Yang、Shil- iang Zhang、Zhijie Yan、Chang Zhou 和 Jingren Zhou。2023年。Qwen-Audio：推动通用 audio understanding via unified l$_{\text{arge-sca}}$ le audi o- 语言模型。*arXiv* 预印本 *arXiv:2311.07919*。

Geoffrey Cideron、Sertan Girgin、Mauro Verzetti、Damien Vincent、Matej Kastelic、Zalán Borsos、Brian McWilliams、Victor Ungureanu、Olivier Bachem、Olivier Pietquin 等。2024年。MusicRL：使音乐生成与人类偏好对齐。*arXiv* 预印本 *arXiv:2402.04229*。Abhishek Dutta 和 Andrew Zisserman。2019年。用于图像、音频和视频的 VIA 标注软件。发表于第 *27* 届 *ACM* 多媒体国际会议论文集，第 2276–2279 页。

Xavier Espinal、Maria Giuffrida、Marieke Willems 和 Rita Meneses。2022 年。让大型科学实验数据触手可及，方便研究人员获取。

Sreyan Ghosh、Arushi Goel、Lasha Koroshinadze、Sang- gil Lee、Zhifeng Kong、Joao Felipe Santos、Ramani Duraiswami、Dinesh Manocha、Wei Ping、Moham- mad Shoeybi 等。2025年。Music Flamingo：扩展音频语言模型中的音乐理解能力。*arXiv* 预印本 *arXiv:2511.10289*。

Anthony Gibbons、Ian Donohue、Courtney Gorman、Emma King 和 Andrew Parnell。2023年。Neal：一款开源音频标注工具。*PeerJ*，11:e15913。

Arushi Goel、Sreyan Ghosh、Jaehyeon Kim、Sonal Ku- mar、Zhifeng Kong、Sang-gil Lee、Chao-Han Huck Yang、Ramani Duraiswami、Dinesh Manocha、Rafael Valle 等。2025年。Audio Flamingo 3：借助完全开放的大型音频语言模型推进音频智能。*arXivpreprint arXiv:2507.08128*。

Manraj Singh Grover、Pakhi Bamdev、Yaman Kumar、Mika Hama 和 Rajiv Ratn Shah。2020年。Audino：一款面向音频和语音的现代标注工具。*arXivpreprint arXiv:2006.05236*。

Azalea Gui、Hannes Gamper、Sebastian Braun 和 Dimitra Emmanouilidou。2023年。调整 Fréchet 音频距离以用于生成音乐评估。*arXivpreprint arXiv:2311.01616*。

Juan Gutiérrez、Ángel Mora、Pablo Regodón、Silvia Rodríguez 和 José Luis Blanco。2025年。AI 增强视频标注：评估流程改进。*arXiv* 预印本 *arXiv:2510.21798*。

Faria Binte Kader 和 Santu Karmaker。2025年。音乐生成评估指标综述。*arXiv* 预印本 *arXiv:2509.00051*。

Jin-Dong Kim 和 Yue Wang。2012年。PubAnnotation——一个持久且可共享的语料库与标注存储库。载于*BioNLP*：*2012*年生物医学自然语言处理研讨会论文集，第 202–205 页。

Qiuqiang Kong、Yin Cao、Turab Iqbal、Yuxuan Wang、Wenwu Wang 和 Mark D Plumbley。2020年。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、语音和语言处理汇刊，28:2880–2894。

Golan Levy、Raquel Sitman、Ido Amir、Eduard Golshtein、Ran Mochary、Eilon Reshef、Roi Reichart 和 Omri Allouche。2019年。Gecko：一种用于高效标注人类对话的工具。载于*INTERSPEECH*，第 3677–3678 页。

Cheng Liu、Hui Wang、Jinghua Zhao、Shiwan Zhao、Hui Bu、Xin Xu、Jiaming Zhou、Haoqin Sun 和 Yong Qin。2025年。Musiceval：一个带有专家评分、用于自动文本到音乐评估的生成式音乐数据集。*arXiv* 预印本 *arXiv:2501.10811*。

Blai Meléndez Catalán、Emilio Molina 和 Emilia Gómez Gutiérrez。2017年。BAT：一款开源、基于 Web 的音频事件标注工具。

Christian M Meyer、Darina Benikova、Margot Mieskes 和 Iryna Gurevych。2016年。MDSWriter：用于创建高质量多文档摘要语料库的标注工具。载于*ACL 2016* 系统演示，第 97–102 页。

Ines Montani 和 Matthew Honnibal。2018年。Prodigy：一种全新的标注工具，可实现极其高效的机器教学。《人工智能》，即将发表。

Minh-Quoc Nghiem 和 Sophia Ananiadou。2018年。APLenty：一款利用主动与前瞻式学习创建高质量数据集的标注工具。载于*EMNLP*（演示），第 108–113 页。

Minh-Quoc Nghiem、Paul Baylis 和 Sophia Anani- adou。2021年。Paladin：一款基于主动与前瞻式学习的标注工具。载于计算语言学协会欧洲分会第 *16* 届会议论文集：系统演示，第 238–243 页。

Sylvaine Picard、Camille Chapdelaine、Cyril Cappi、Lau- rent Gardes、Eric Jenn、Baptiste Lefevre 和 Thomas Soumarmon。2020年。确保机器学习认证的数据集质量。载于*2020*年 *IEEE* 软件可靠性工程国际研讨会（*ISSREW*），第 275–282 页。IEEE。

Jordi Pons 和 Xavier Serra。2019年。musicnn：用于音乐音频标注的预训练卷积神经网络。*arXiv* 预印本 *arXiv:1909.06654*。

Michael Schoeffler、Sarah Bartoschek、Fabian-Robert Stöter、Marlene Roess、Susanne Westphal、Bernd Edler 和 Jürgen Herre。2018年。webmushra——用于在线听音测试的综合框架。《开放研究软件期刊》，6(1)。

Erik Smistad、Andreas Østvik 和 Lasse Løvstakken。2021年。Annotation Web——一款用于超声图像的开源网页版标注工具。载于 *2021*年 *IEEE*国际超声研讨会（*IUS*），第1–4页。IEEE。

Pontus Stenetorp、Sampo Pyysalo、Goran Topi´c、Tomoko Ohta、Sophia Ananiadou 和 Jun'ichi Tsujii。2012年。BRAT：一款用于自然语言处理辅助文本标注的网页版工具。载于第*13*届计算语言学协会欧洲分会会议论文集：演示，第102–107页。

Maxim Tkachenko、Mikhail Malyuk、Andrey Holmanyuk 和 Nikolai Liubimov。2020-2022年。Label Studio：数据标注软件。可从 https://github.com/heartexlabs/label-studio 获取的开源软件。

Satvik Venkatesh、David Moffat 和 Eduardo Reck Miranda。2022年。只听一次：一种用于音频分割和声音事件检测的类YOLO算法。《应用科学》，12(7):3293。

Paul Voigtlaender、Lishu Luo、Chun Yuan、Yong Jiang和Bastian Leibe。2021年。降低视频目标分割数据集的标注工作量。载于*IEEE/CVF*冬季计算机视觉应用会议论文集，第3060–3069页。

Max Wiechmann、Seid Muhie Yimam和Chris Biemann。2021年。ActiveAnno：集成主动学习的通用文档级标注工具。载于 *2021*年计算语言学协会北美分会会议论文集：人类语言技术：演示，第99–105页。

Yongyi Zang、Sean O'Brien、Taylor Berg-Kirkpatrick、Julian McAuley和Zachary Novack。2025年。你真的在听吗？提升音乐问答基准测试中的感知意识。*arXiv*预印本*arXiv:2504.00369*。