arXiv:2611.29411v1 [cs.SD] 22 Nov 2026

谁在演奏什么：从视频中分离相同乐器的音乐

Abhinaba Roy[1], Ambuj Mehrish[2],Sebastiano Vascon[2],Dorien Herremans[1]
新加坡科技与设计大学 [2]威尼斯大学

演示页面： https://amaai-lab.github.io/who_演奏_什么_演示/**代码：** https://github.com/AMAAI-Lab/who_演奏_什么**数据集：** https://huggingface.co/datasets/amaai-lab/who_演奏_什么



图1. **点击 t　　o进行分离。** 点击某位演奏者，只听他们演奏的部分；他们的弓部动作能够区分乐器完全相同的演奏者。

摘要

当合奏团中的两名音乐家演奏相同乐器时，他们的声音相似，乐器外观也相似。纯音频分离器，以及依赖外观的视听分离器，都无法区分他们。两人唯一不同之处在于各自如何运动，以及何时运动。我们研究同乐器的逐演奏者分离：用户在视频中点击一名音乐家，即可单独获得该音乐家的声音，即使视频中的其他演奏者也演奏相同乐器。我们的模型跟踪被点击的演奏者，并利用高层视频特征以及身体和手部关键点的速度来描述他们。预训练音乐分离器通过交叉注意力处理这些标记。早期基于运动的分离器仅在独奏视频的合成混合音频上进行评分。我们首次在真实合奏录音上对同乐器分离进行逐演奏者评估。在 *URMP* 中的 *58*名玩家里，每人都至少与乐曲中的另一名演奏者演奏相同乐器，我们的模型达到 *5.3*分贝的 *SDR* 中位数，而四种已发布的分离模型（纯音频和视听模型）的结果均不超过 *0.3* 分贝。在同乐器二重奏中，差距更大；我们的模型在 *MUSIC*基准测试上也优于视听基线模型。测试表明，性能提升源于动作：将视频替换为静止帧后，提升便不复存在；将视频相对于音频偏移几分之一秒，也会产生同样的结果。

1. 引言

拍摄一段弦乐四重奏的视频，假设你只想听第二小提琴。音乐分离器 [11, 23]会为你分离出一条“小提琴”音轨，但这条音轨包含两把小提琴的演奏。视听分离器可以查看视频，但这类模型大多会将乐器的外观与声音联系起来 [3, 6, 28], ，而两把小提琴看起来一模一样。当

两位演奏者共用一种乐器时，无论音色还是外观，都无法判断哪些音符属于谁。

两位演奏者之间仍有一点不同：他们各自做什么，以及何时做。音符开始时，琴弓会改变方向。音高变化时，左手手指会移动。长笛演奏者会在乐句开始前吸气。这些动作在时间上与该演奏者的音符相对应，与其他人无关。Sound of Motions [28] 和 Music Gesture [5]利用了这一线索，通过运动轨迹以及身体和手指关键点，分离相同乐器的二重奏。然而，两者报告的信号失真比都仅针对由独奏视频生成的合成混合音频，而且我们没有找到任何一方的公开代码。据我们所知，还没有分离器在真实合奏录音上按演奏者逐一进行评估，而这些录音中的两位演奏者演奏的是同一种乐器。

我们以一种用户可以直接使用的形式研究这一问题（图1）。用户点击视频中的一位音乐家，即可单独听到该音乐家的声音。点击的作用类似于可提示分割 [16]：它指出的是目标演奏者，而不是他们演奏的乐器类型。潜在用途包括练习和教学，学生可以借此听到自己的声部；也包括音乐会视频的混音，分离出的声部之后可以恢复或重新制作 [12]。按演奏者分离出的音轨还可用于音乐分析，例如用于训练相似度模型，而这类模型已经使用分离音轨进行训练 [18]。

我们的模型由预训练组件构成。一个可由提示引导的跟踪器会在视频中跟随点击的演奏者。我们从该演奏者的裁剪画面中提取两种描述其动作的信息：冻结的自监督视频编码器输出的令牌，以及身体和手部关键点的帧间速度。随后，预训练音乐分离器通过交叉注意力层关注这些令牌；每个时间层之后都添加了一个交叉注意力层。分离器已经掌握了各种乐器的音色，因此视频只需告诉它要保留哪位演奏者。我们只使用独奏演奏视频的合成混合音频进行训练，从不使用合奏录音。四种训练混合音频中，有三种加入了第二位演奏相同乐器的演奏者。最难的一种中，第二位演奏者是同一视频中另一个时刻的同一表演者，因此外观无法区分两者。

接下来，我们首次针对真实合奏录音中的相同乐器分离进行逐演奏者评估。我们在 URMP 中点击 58 名演奏者 [9], , 每名演奏者所在的乐曲中都有另一名演奏相同乐器的演奏者，并将输出与被点击演奏者自己的音轨进行比较。我们的模型达到 5.30 分贝的 SDR 中位数。四种已发布的分离器中，有三种是视听分离器，其表现均不超过 0.30 分贝，低于一种简单的参照方法（仅调低录音音量，达到 0.82 分贝）。这些分离器输出的配对中另一名演奏者的声音，与被点击演奏者的声音一样响；而我们的模型将该演奏者的声音压低了 23.65 分贝。在纯二重奏中，两名演奏相同乐器的演奏者独自演奏，我们的模型达到 18.73 分贝，而基线方法最高仅为 1.75 分贝。基线方法的这一失效并非设置错误：在它们中有两种接受训练的 MUSIC基准测试上，它们重现了已发表的分数，而我们的模型比其中表现较好的模型高出 6.9 分贝。

良好的得分并不能说明模型为何有效，因此我们直接检验动作的作用。输出会随点击而变化：在58名演奏者中，有51名演奏者的正确点击得分高于点击另一位演奏者，而音视频基线模型的表现接近随机水平。将视频替换为一帧重复画面，保留演奏者的外观、去除其动作后，性能提升便消失了。将视频相对于音频偏移0.16秒，保留外观和动作、仅破坏两者的时序后，对于所有演奏者而言，性能提升同样消失了。最后，一个仅使用关键点速度而不使用像素的变体，在所有演奏者上的得分达到4.15分贝，在纯二重奏上达到12.18分贝；相比之下，不含视频的模型分别为0.52分贝和2.95分贝。仅凭动作就能区分演奏相同乐器的两位演奏者，而对于所有演奏者而言，性能提升取决于动作的时序。

I总之，我们可 ke th三类相关tributions.

• **任务与基准测试**。据我们所知，这是首次在真实合奏录音上对同乐器分离进行逐演奏者评估（第4节）。

• **模型**。一种预训练音乐分离器，以点击的演奏者的视频为条件，仅使用独奏视频的合成混合音频进行训练（第3节）。

• **动作发挥作用的证据**。错误点击、静止帧、时间偏移和仅关键点测试（第5.4节）。

2. 相关工作

纯音频音乐分离。 BS- RoFormer [11] 和 MVSep Mega [23] 等音乐分离器会将录音拆分为不同乐器类别。给定一段四重奏录音，它们会输出一条包含两把小提琴的音轨，因为输入中没有任何信息指出需要哪一把小提琴。采用置换不变训练的盲分离可以拆分音色相同的两个声部 [19],，但输出顺序并不固定，因此无法响应指定某位演奏者的请求。

基于外观的视听分离。 Sound of Pix- els [27] 通过混合两段视频的音频，并要求模型借助帧还原混合音频，来训练分离器。后续模型沿用了这一方案。iQuery [3] 为每种乐器使用一个查询，DAVIS-Flow [6]是一种生成式分离器，而 SAM Audio [21] 是一种通用分离器，可接受视觉掩码作为提示。它们的训练混合音频大多由不同乐器配对而成，因此只需根据乐器外观就足以选出正确的声音。两位演奏相同乐器的演奏者会让这些模型两次获得相同的视觉证据。

基于动作的视听分离。在深度学习之前，Parekh 等人。 [14, 15] 将弦乐演奏者的弓速与矩阵分解的激活值联系起来，而 Li 等人。 [8, 10] 通过将视频中的演奏者与乐谱中的声部进行匹配

，并将运弓和颤音与音符事件关联起来，从而匹配已知乐谱。在学习型模型中，Sound of Motions [28] 将运动轨迹加入分离器，而 Music Gesture [5]加入身体和手指关键点，并报告了钢琴、长笛和小号二重奏的同乐器分离结果。身体部位运动同样是识别视频中人体动作的标准线索 [17]。两者都在由独奏视频构成的合成混合音频上报告了 SDR 分数。Slizovskaia 等人。 [22] 确实在真实的 URMP 录音上进行了测试，但他们将两名使用相同乐器的演奏者视为一个声源。我们的方法在测试时只需点击一次，无需乐谱。

可提示模型与冻结特征。 我们的点击在可提示分割中充当点提示，而我们的模型建立在四个冻结模型之上。SAM 2.1 [16]将点击转换为掩码，并在视频中跟踪该掩码。V-JEPA 2 [1], 是一种自监督视频编码器，RTMW [7],是一个全身关键点估计器，用于描述被跟踪的演奏者。SAM 3 [2] 还接受文本，我们仅使用它来测试以文本提示替代点击的效果。

3. 方法

图2 概述了我们的方法。设 $x \in \mathbb{R}^T$ 为一段包含 N 名演奏者的音频录音，其中

$$x = \sum_{n=1}^{N} s_n, \qquad \hat{s}_k = f_\theta(x, Z_k), \tag{1}$$

其中， s_n 是演奏者 n 的声音， Z_k 表示在第一帧视频中通过点击选中的演奏者。我们的目标是恢复 s_k，即使另一名演奏者 $j \neq k$演奏相同乐器也是如此，从而无法仅凭乐器身份选定目标。

3.1. 演奏者表示

冻结的 SAM 2.1 [16]（采用 Hiera-L 骨干网络）会分割点击的演奏者，并在整个视频中跟踪其掩码。我们在每一帧中围绕掩码截取一个方形裁剪区域，并使用两个冻结的模型对裁剪区域进行编码：自监督视频编码器 V-JEPA 2 [1] 用于提取视觉特征，全身关键点估计器 RTMW [7] 用于提取身体和手部运动。

V-JEPA 2 ViT-L 每两帧生成一个包含 1024 个维度的标记网格 4×4 ，每秒 12.5 个时间步中包含 16 个空间令牌。这些特征保留了有关演奏者、乐器及可见演奏表现的信息。

RTMW 每帧检测 48 个身体和手部关键点。对于关键点 i 的位置 $p_{t,\,i} \in \mathbb{R}^2$ 和置信度$c_{t,\,i}$ ，我们计算边界框归一化速度，并将速度与置信度拼接起来：

$$\begin{aligned} u_{t,i} &= \frac{p_{t,i} - p_{t-1,i}}{b_t}, \\ m_t &= [u_{t,1}, \ldots, u_{t,48}, c_{t,1}, \ldots, c_{t,48}] \in \mathbb{R}^{144}, \end{aligned} \tag{2}$$

其中， b_t 是演奏者在帧 t 中边界框的大小。我们以 25 Hz 的频率为每帧生成一个运动向量。与绝对位置相比，使用速度更能突出动作，并消除对演奏者位置的直接依赖。

视频标记与关键点标记共同构成 Z_k。由于这些标记与分离器的音频标记具有不同的采样率，我们将它们的时间戳映射到共享的 10 毫秒网格上。旋转位置嵌入 [25] 对这些时间戳进行编码，使交叉注意力能够将演奏者的动作与相应的声音联系起来。

3.2. 演奏者条件分离器

我们使用 MVSep Mega [23],a BS-RoFormer [11] 预训练音频骨干网络进行初始化，该网络经过预训练，可分离 53 个乐器声部。模型采用 2048 个采样点的窗口和 512 个采样点的帧移计算短时傅里叶变换，将频谱划分为多个频带，并通过 12 个宽度为 256 的模块处理频带特征。模块 ℓ 先应用时间变换器 Time_ℓ，再应用逐频带变换器 Band_ℓ。我们保留预训练的频带分割模块和全部 12 个模块，共计 2630 万个参数。

我们在每个 Time_ℓ 后添加一个交叉注意力层，以音频特征 H 作为查询，以演奏者令牌 Z_k 作为键和值：

$$\begin{aligned} H &\leftarrow \text{Time}_\ell(H), \\ H &\leftarrow H + W_\ell \,\text{CrossAttn}_\ell(H, Z_k), \\ H &\leftarrow \text{Band}_\ell(H). \end{aligned} \tag{3}$$

这样，演奏者专属信息便会在分离器的时间处理过程中发挥作用，且早于信息在频带之间交换。输出投影 W_ℓ 初始化为零，从而在初始化时保留预训练骨干网络的行为。

我们将 Mega 的 53 个乐器专属输出头替换为一个单独的输出头，用于预测所选演奏者的复数掩码 M，使得 $\hat{s}_k = \text{iSTFT}(M \odot X)$ 与 X =STFT(x)。我们训练骨干网络、交叉注意力层和掩码头，参数总数为 4390 万。

3.3. 训练

我们使用 Solos [13] 和 MUSIC-21 [27, 28] 中的独奏演奏视频构建合成混合音频，并以此进行训练。数据涵盖 26 种乐器，包含 1,088 条训练录音、51 条验证录音和 78 条测试录音。我们排除了属于 MUSIC 和 MUSIC-21 基准测试集的 158 条录音，且不使用 URMP 进行训练。

每个示例都将一个四秒的目标片段与干扰音频相结合。给定目标音频 s_a、其演奏者标记Z_a和干扰信号 s_b，模型预测

$$x = s_a + s_b, \qquad \hat{s}_a = f_\theta(x, Z_a). \tag{4}$$

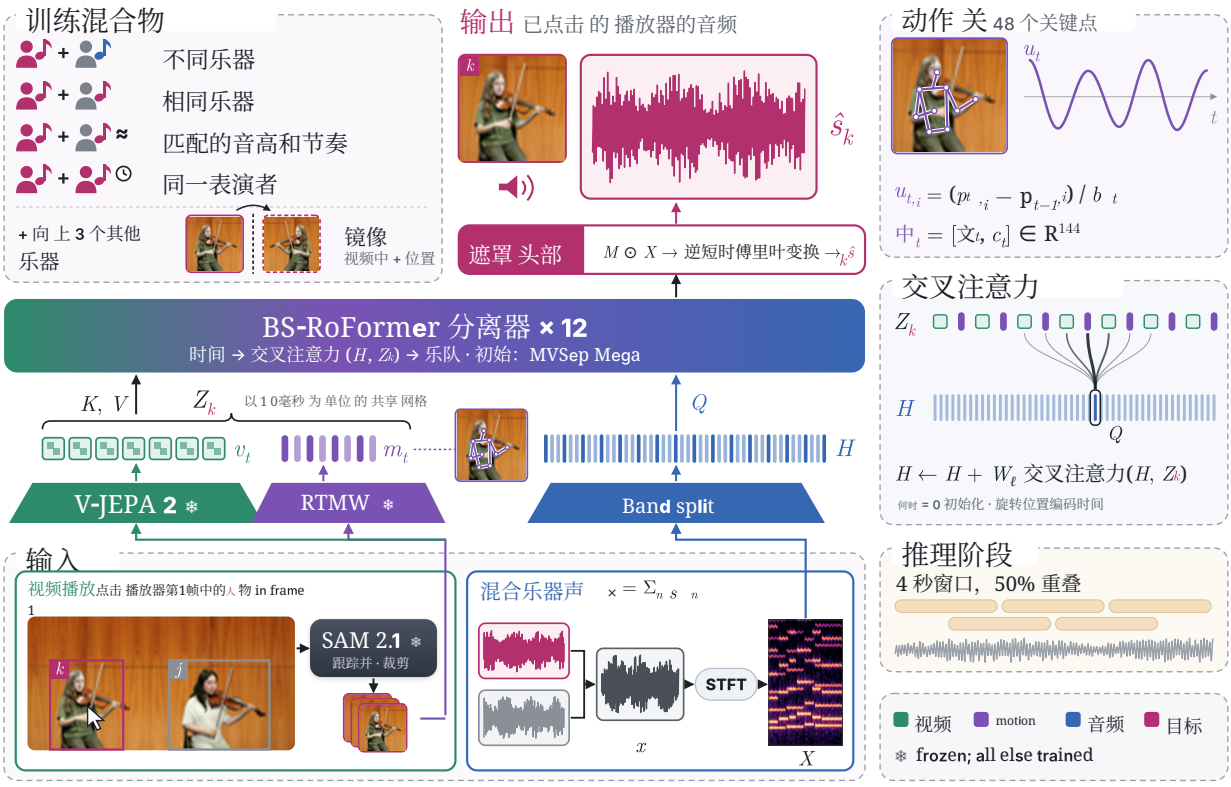

图2. **方法概览**。点击目标演奏者。跟踪裁剪区域提供视频特征和关键点速度，这些信息通过交叉注意力对预训练音乐分离器进行条件化。掩码头重建所选演奏者的声音。

仅提供目标视频的令牌，因此分离器必须利用这些令牌来识别混合音频中的对应声源。

干扰音频中包含第二位演奏者，其呈现方式有四种。第二位演奏者演奏不同的乐器（概率为0.20）、在另一个视频中演奏相同乐器（0.30）、在另一个视频中演奏与目标片段音高和速度相匹配的相同乐器（0.25），或者是目标视频中不同时间出现的同一表演者，演奏相同乐器且录音环境相同（0.25）。这四种情况从第一个训练步骤开始就进行混合，我们还会加入最多三个其他乐器的片段。相同乐器和同一表演者的情况使得乐器身份和外观无法成为可靠的源选择捷径，从而促使模型依据演奏者动作的时序进行判断。

为考虑观看方向，我们以 0.5 的概率将演奏者视频水平镜像。随后，V-JEPA 2 会处理镜像后的帧；对于关键点，我们则将水平速度取反，并交换左、右。

我们优化 ℓ_1 波形损失函数 $\|\hat{s}_a - s_a\|_1$，并结合复数频谱 ℓ_1 损失函数 $\|\text{短时傅里叶变换}_r(\hat{s}_a) - \text{短时傅里叶变换}_r(s_a)\|_1$，采用五种分辨率 r。

训练使用 AdamW，批次大小为 32 个四秒混合音频。预训练骨干网络的学习率设为新层所用学习率的六分之一。我们选择验证集目标源信噪比最高的检查点。推理时，录音以四秒窗口处理，窗口之间有 50% 重叠。补充材料介绍了损失函数加权方式、分辨率及更多优化细节。

4. 实验设置

URMP 同乐器测试集。URMP [9] 包含带视频的室内乐曲目，并为每位演奏者提供单独的音频轨道。各声部由演奏者逐一录制，随后合成为一段录音和一段视频。我们选取了同一乐曲中乐器相同的 29 对演奏者，共有 58名玩家：26 名小提琴演奏者、2 名中提琴演奏者、12 名长笛演奏者、16 名小号演奏者和 2 名萨克斯管演奏者。每首乐曲有两至五名演奏者。58 名演奏者中有八名属于纯二重奏，即四首由这对演奏者单独演奏的曲目；其余 50 名演奏者则与其他乐器一同演奏。每次点击都对应演奏者身体上的三个点（鼻子、肩膀、髋部），这些点由行人检测器在第一帧标注，并经人工目视检查。第 5.6节测试其他提示。

指标和统计数据。我们使用 museval [24] 的信号失真比（SDR）为每个输出评分，并将其与点击的演奏者自己的音轨进行比较，采样率为 URMP 的 48 千赫，所有输出均重采样至该采样率。

		URMP，真实合奏录音						留出混合音频	
		全部（58）		弦乐器	管乐器与	Pure	与其他	All	Pure 纯二重奏 ‡
方法	输入	中位数 [95% 区间]	我们的方法更好	(28)	铜管乐器 (30)	二重奏 (8)	乐器 (50)	(500)	仅 (134)
未改变的录音	none	−6.29 [−6.79, −5.61]	58	−6.71	−5.77	0.00	−6.58	−4.46	−0.26
缩放录音， x/N	none	0.82 [0.66, 1.00]	46	0.73	1.02	2.91	0.80	1.09	2.65
SAM Audio	A +V	−3.92 [−4.58, −3.16]	57	−3.97	−3.73	−0.20	−4.05	−2.47	−0.07
iQuery	A +V	−0.45 [−0.78, −0.10]	51	−0.62	−0.12	0.29	−0.49	−0.66*	0.46*
DAVIS-Flow	A +V	0.01 [−0.09、0.01]	47	0.01	0.01	0.02	0.00	−0.18	0.29
MVSep Mega	A	0.30 [0.19, 0.78]	42	0.43	0.30	1.75	0.28	0.26†	0.14†
Ours	A +V	**5.30** [1.86, 10.31]		**6.20**	**3.62**	**18.73**	**3.02**	**9.56**	**14.36**

表1。**主要结果**。以分贝为单位的 SDR 中位数，数值越高越好。A：音频；V：视频。我们的方法更优：我们的模型优于该行所示演奏者组合。纯二重奏：仅包含该配对中两位演奏者的乐曲。 *仅使用 iQuery 的 11 个类别（220 个混合音频，其中 63 个包含两位演奏者）。 †仅使用 Mega 有音轨的 408 个混合音频（114 个）。 ‡ 两位演奏者演奏相同乐器。

演奏者的得分是以 1 秒窗口计算的中位数，我们报告演奏者之间的中位数。对于全部 58名玩家的中位数，我们基于 10,000 次重采样给出 95%自助法区间；由于一对演奏者来自同一首乐曲，因此在重采样时将这两名演奏者作为一个整体。

两个简单基准。以较低音量返回录音，即使不进行任何分离，也会提高 SDR。因此，我们报告未改变的录音 x 和缩放录音 x/N，其中 N 为演奏者人数。只有优于缩放录音的方法才算有效。

留出合成混合音频。 作为第二项测试，我们从 Solos 和 MUSIC-21 划分中留出的 78 段录音（第 3.3 节）构建了 500 个时长为 4 秒的混合音频。每个混合音频都包含同一种乐器的两位演奏者，其中 366 个还包含一到三种其他乐器。第二位演奏者来自另一个视频（170 个混合音频）、来自另一个音高和速度匹配的视频（149 个），或来自同一视频的另一个时刻（181 个）。这些混合音频未用于任何设计决策或检查点选择。由于它们的构建方式与我们的训练混合音频类似，因此更有利于我们的模型，而非基线方法。

基线方法。 我们比较了四种已发布的模型，均按照作者发布时的设置运行，且未在我们的数据上重新训练。**SAMAudio** [21] 接收点击的演奏者的 SAM 2.1 掩码作为视觉提示，**iQuery** [3]接收演奏者的裁剪画面，并以乐器名称作为查询，**DAVIS-Flow** [6] 接收演奏者的裁剪画面。**MVSep Mega** [23] 仅处理音频。给定乐器后，它会返回该乐器的音轨，其中包含这一对演奏者。iQuery 和 DAVIS-Flow 的音频输出采样率为 11 千赫，因此在我们的 48 千赫评分中，5.5 千赫以上的所有内容都会丢失。我们没有找到 Sound of Motions 和 Music Gesture 的任何公开代码。

消融模型。 我们使用训练数据进行消融实验。 课程学习模型 仅改变训练混合音频的顺序（第 3.3 节）：依次在三种情形下各训练 4k 步（不同乐器、来自另一个视频的相同乐器、同一表演者在另一时刻），然后使用完整混合音频。其他所有消融模型都像我们的模型一样混合所有情形。无视频模型 没有演奏者标记，也没有交叉注意力。仅关键点模型 去掉了 V-JEPA 2 标记，因此它能看到演奏者如何移动，却看不到任何像素。无镜像模型 在没有镜像增强的情况下训练；另一个模型则从随机权重开始初始化无镜像模型的骨干网络，而不是使用 Mega 的权重。

5. 结果

5.1. 与已发布的分离模型比较

在表 1 的 URMP 各列中，没有任何已发布的基线方法优于缩放录音。它们中表现最好的是纯音频 Mega，在所有演奏者上的提升为 0.30 分贝；三种视听基线模型的表现均不超过 0.01 分贝。我们的模型在全部 58名玩家 上达到 5.30 分贝，其 95% 区间为 1.86 至 10.31 分贝，高于每个基线方法和两个简单基准。该区间较宽，是因为该方法在某些乐曲上表现出色，但在其他乐曲上则会失败（第5.7节）。在纯二重奏中的 8 名演奏者上，该模型达到 18.73 分贝，并在全部 8 名演奏者上领先于每个基线方法。留出合成混合音频呈现出相同的结果，且差距更大：我们的模型达到 9.56 分贝，而基线方法最高仅为 0.46 分贝。

为了区分视频和训练数据各自的影响，我们在相同数据上训练了同一个不使用视频的模型（表 5，第 c 行）。该模型在所有演奏者上的得分为 0.52 分贝，在纯二重奏上的得分为 2.95 分贝，与缩放录音的结果接近，这正如一个无法知道目标演奏者是谁的模型所预期的那样。增益来自视频，而非训练数据。

基线方法留下了什么。信号失真比将所有不需要的声音合并计算。为了了解具体残留了哪些声音（表。2），我们将每个输出拟合为三种真实信号的加权和：点击的演奏者、配对中的另一名演奏者（我们称之为搭档），以及其他乐器。这是一种简单的增益拟合，

方法	搭档		其他乐器
	全部（58）	纯二重奏（8）	(50)
SAM Audio	−0.61	0.14	−17.39
iQuery	0.08	−0.76	2.32
DAVIS-Flow	0.26	1.07	4.78
MVSep Mega	0.00	0.00	18.09
无视频模型‡	0.00	0.00	−0.82
课程学习模型‡	21.14	32.81	14.23
Ours	**23.65**	**39.01**	**21.23**

表2。**输出中剩下的内容**。点击的演奏者下方每种不需要的声音的级别，单位为分贝；数值越高，泄漏越少。搭档：同乐器配对中的另一名演奏者。 ‡消融实验。

方法	平均信号失真比 [95% 区间]	在其论文中
混合音频本身	0.16	
iQuery	11.61 [10.88、12.36]	11.17
DAVIS-Flow	12.21 [11.40、13.03]	12.01
Ours	**19.08** [18.22, 19.97]	

表3。**基线方法自身的基准测试**。来自 MUSIC 和 MUSIC-21 测试录音的 200 组由两种不同乐器构成的混合音频，按照基线方法的评估方案进行评分。

并非 museval 的信号干扰比（补充材料）。所有基线方法，包括我们的无视频模型，都会让搭档的音量与点击的演奏者大致相当，差值介于 −0.76 和 1.07 分贝之间，因此没有一种方法能将演奏同一种乐器的两名演奏者分开。Mega 能很好地去除其他乐器（18.09 分贝），却无法去除搭档的声音（0.00 分贝）：它分离的是乐器，而非演奏者。我们的模型则将搭档的声音压低到比点击的演奏者低 23.65 分贝；在纯二重奏中低 39.01 分贝；其他乐器的声音则低 21.23 分贝。

基线方法自身的基准测试。我们还在这两个基线方法训练和测试所用的 MUSIC 基准测试上评估了 iQuery、DAVIS-Flow 和我们的模型（表3）。我们从 MUSIC 和 MUSIC-21 的 77 段测试录音中构建了 200 组由两种不同乐器构成的混合音频，这些录音均未用于训练我们的模型或选择其检查点。两个基线方法复现其已发表的分数时，误差均在 0.5 分贝以内，因此它们在 URMP 上表现不佳不太可能是设置错误所致。我们的模型比表现较好的基线方法高 6.9 分贝，并且在至少 183 组混合音频上表现更好。这里的每组混合音频都包含两种不同乐器，因此该基准测试评估的是常见任务，而非同乐器情形。我们的模型还以 Mega 为起点；Mega 能分离 53 种音轨，因此接触过的乐器种类远多于 MUSIC 的 11 种。

5.2. 听力测试

信号失真比会比较波形，并且取决于响度，因此我们还会在 MUSHRA测试中询问听众 [20]。在每次试验中，听众都会看到一帧视频画面，其中点击的演奏者已被标记，并听到该演奏者的真实音轨作为参考音频。随后，听众会对七个未标注的音频片段进行 0 到 100 分的评分

方法	所有试验（10）	纯二重奏（3）
隐藏参考音频	95.3（91.7，98.1）	95.5（91.4，98.6）
SAM Audio	30.5（16.8，45.2）	38.5 (25.1，52.4)
iQuery	22.1 (14.2, 30.3)	19.7 (11.4, 29.6)
DAVIS-Flow	18.2 (10.7, 26.8)	11.4（1.4，27.2）
MVSep Mega	30.2（21.1，40.6）	29.8 (18.9, 40.5)
Ours	**69.0** (50.0, 85.2)	**88.0** （73.8, 98.1）
锚点（未改变的录音）	26.0（16.3，36.6）	31.3 (19.0, 45.4)

表 4. **聆听测试（MUSHRA）**。16 名听众的平均分数（范围为 0 至 100），分数越高越好；括号内为基于听众和试验计算的 95% 自助法置信区间。

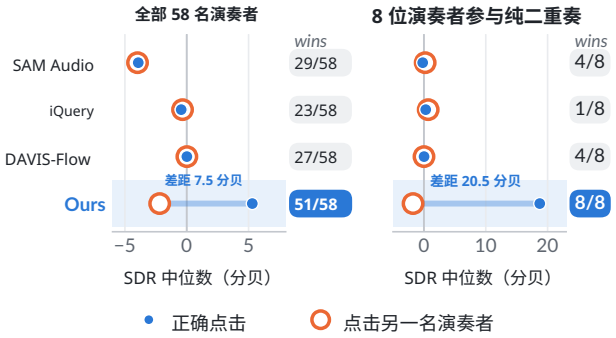

图 3. **输出是否跟随点击?** 将点击分别放在配对中某位演奏者和另一名演奏者的音轨上，并据此对该演奏者的音轨评分。胜场：正确点击得分更高的演奏者。在所有演奏者中，基线方法获胜的比例约为一半。

以评估每个片段与参考音频的接近程度：隐藏的参考音频、我们的模型、四种基线方法，以及作为锚点的未改变的录音。

共有 10 次试验，每次试验来自一首不同的乐曲，且乐曲随机选取：3 首二重奏、2 首三重奏、3 首四重奏和 2 首五重奏。10 秒的节选按固定规则选取，且在任何人听取输出之前就已确定；所有音频片段均归一化至 −23LUFS，因此响度不会影响结果。共有 16 名听众参与。我们的模型在所有试验中的得分为 69.0，在纯二重奏中的得分为 88.0（表4）。表现最佳的基线方法在所有试验中的得分分别为 30.5（SAM Audio）和 30.2（MVSep Mega），而 SAM Audio 在纯二重奏中的得分为 38.5。16 名听众无一例外地给我们的模型打出了高于所有基线方法的平均评分（$p = 1.2 \times 10^{-4}$；对听众的平均评分进行精确Wilcoxon符号秩检验，并经霍姆校正）。基线方法相较于未改变的录音几乎没有提升：在所有试验中，iQuery 和 DAVIS-Flow 的得分低于未改变的录音，而 SAM Audio 和 MVSep Mega 的得分仅高出不到 5 分。

5.3. 输出是否跟随点击?

可提示式分离器必须在点击不同演奏者时输出不同的声音。我们通过点击配对中的另一位演奏者来测试每位演奏者，并根据输出进行评分

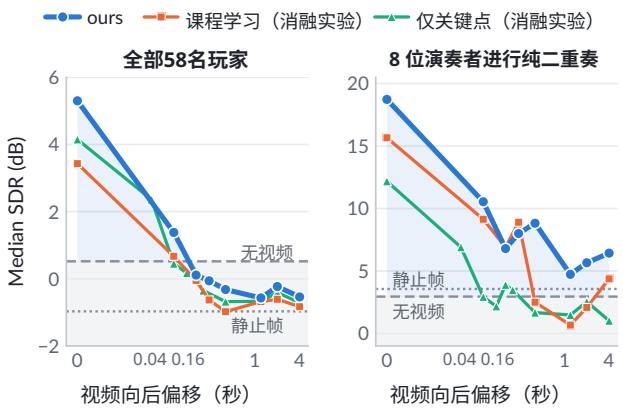

图4。**动作随时间变化，促成了分离**。视频相对于音频延迟时的 SDR 中位数（0.04 秒以上为对数坐标轴）。在所有演奏者中，0.08 至 0.16 秒的偏移会使所有模型达到无视频模型的水平（虚线）。在纯二重奏中，我们的模型仍保留部分优势（蓝色区域）。点线：使用静止帧的我们的模型。

并与第一位演奏者的音轨进行比较。如果某种方法能够跟随点击，正确点击的得分会更高。

对于我们的模型，58名演奏者中有51名以及纯二重奏中的全部8名演奏者，其正确点击的得分更高（图3）。错误点击会使所有演奏者的中位数从5.30降至 −2.19分贝，并使纯二重奏中的中位数从18.73降至 −1.76分贝。三个视听基线模型在58名演奏者中有23至29名上胜出。例如，无论点击哪位演奏者，SAM Audio 大多都会返回同一位更突出的演奏者。

5.4. 模型是否使用动作?

点击测试无法说明模型是否能识别演奏者 是谁——是通过裁剪画面中的位置或服装识别，还是依据演奏者的动作来判断。我们通过三种方式进行测试，从最弱到最强（图4）。

静止帧。 我们将每位演奏者的视频替换为其中间帧，并重复播放，同时将关键点速度设为零。模型仍能看到演奏者是谁以及他们演奏什么，但看不到他们如何运动。所有演奏者的中位数从 5.30 分贝降至 −0.97 分贝，纯二重奏中的中位数则从 18.73 降至 3.56 分贝。现在，正确点击仅在 58 名演奏者中的 26 名身上胜出，这相当于随机猜测；在二重奏中，模型大多会漏掉两位演奏者中的一位。有一点需要注意：模型在训练时从未见过静止不动的演奏者，因此静止帧也是一种不常见的输入，这或许是它的得分低于无视频模型（ 0.52 分贝）的原因。

时间偏移。 我们保留真实视频，并将其相对于音频向后偏移，超出乐曲末尾的部分则绕回开头。演奏者的外观和动作都与之前相同，只有动作与声音之间的时序发生变化，因此，只识别演奏者的模型仍会得到原来的分数。我们在不重新训练的情况下测试了我们的模型和两个消融实验：课程学习模型和仅关键点模型。使用 V-JEPA 2 令牌的模型可以按 0.08 秒的步长偏移，即每个令牌覆盖的时长；仅关键点模型则可以按单个视频帧偏移，即 0.04 秒。

综合所有演奏者来看，要获得增益，动作与声音的时序需要对齐到约 0.1 秒以内。偏移 0.08 秒，模型的表现就会从 5.30 分贝降至 1.38 分贝。偏移 0.16 秒时，三个模型的表现都低于无视频模型；即使偏移增大到 4 秒，增益也不会恢复。对于仅关键点模型，仅一帧的偏移就会产生影响：偏移 0.04 秒，其二重奏乐谱分数便从 12.18 分贝降至 6.90 分贝。在纯二重奏中，无论偏移多少，我们的模型都能保留部分增益，介于 4.73 至 10.54 分贝之间；同步时为 18.73 分贝，无视频时为 2.95 分贝。在这种情况下，模型能够部分依据不受时序影响的特征（例如演奏者的身份）来区分两位演奏者。仅关键点模型看不到外观，从偏移 0.4 秒起，其表现便降至无视频模型的水平左右。因此，在二重奏中，时序是模型所利用因素中的主要部分，但并非全部。

仅关键点。 仅关键点模型从未看到任何像素：它只接收 48 个关键点的速度和检测置信度，不包含帧中的绝对位置。由于模型是以这种方式训练的，针对静止帧的异议并不适用。在纯二重奏上，它达到 12.18 分贝；在全部 58 名演奏者上达到 4.15 分贝，而无视频模型对应的结果为 2.95 和 0.52 分贝（表 5）。它对 58 名演奏者中的 51 名取得了正确点击获胜，与我们的模型相同。仅凭关键点运动，就能在真实录音中区分演奏相同乐器的两位演奏者。三项测试结果一致。在所有演奏者中，去掉动作或其时序都会消除这一优势，而只保留动作则能保留大部分优势。

同一表演者，另一个时刻。 留出混合音频也表明了同样的趋势。其中 181 个混合音频里的第二位演奏者，是同一视频中同一表演者在另一个时刻的画面，因此两者看起来完全相同。无视频模型达到 1.39 分贝，仅关键点模型达到 3.86 分贝，我们的模型达到 6.31分贝。

时序，而非数量。动作更多的演奏者并不会因此更容易被分离出来。与无视频模型相比，我们的模型所取得的提升与演奏者的关键点平均速度并不相关（手臂的斯皮尔曼相关系数为 0.07，手的为 0.03，$p \geq 0.6$）。平均速度衡量的是演奏者动作的幅度，而非动作是否与音符对齐。

5.5. 消融实验

在表 5 中，(b) 至 (e) 行分别改变了模型中的一个因素：训练计划、视频输入或镜像增强。没有视频（c 行）时，模型的表现会降至缩放录音的水平，而且完全无法跟随点击。

课程学习。从第一步开始混合第3.3 节中的四种情况（表5第 a 行），比逐一引入这些情况（第 b 行，第4节）在 URMP 上效果更好：所有演奏者的结果为 5.30 分贝，而非 3.43 分贝；58 名演奏者中有 44 名的结果更高，纯二重奏上的结果为 18.73 分贝，而非 15.66 分贝。对于留出混合音频，两种方法的差异不大（9.56 分贝和 9.90 分贝）。

模型	URMP All	URMP Str.	URMP W&B	URMP 二重奏	持有-out
(a) **我们的方法**	**5.30**	**6.20**	3.62	**18.73**	9.56
(b) 课程学习	3.43	3.44	2.73	15.66	**9.90**
(c) (a)，无视频	0.52	0.61	0.38	2.95	1.24
(d) (a)，仅关键点	4.15	5.04	**3.83**	12.18	6.46
(e) (a)，不使用镜像增强。	3.29	5.19	0.66	11.04	9.54
(f) (e)，测试视频为 960 px	0.44	4.34	−0.24	6.75	9.54
(g) (f)，骨干网络未经过预训练	0.40	0.71	0.36	0.51	2.08

表5。**消融实验**。 SDR 中位数，单位为分贝。第 (b) 至 (e) 行各自对 (a) 行的一项内容进行调整；第 (f) 和 (g) 行则各自对其所指行的一项内容进行调整。Str.：弦乐器。W&B：木管乐器与铜管乐器。

补充材料按组、提示以及时间偏移对它们进行了比较。

视频令牌。 第 (d) 行移除了 V-JEPA 2 令牌，保留了关键点。加入视频令牌后，所有演奏者的中位数从 4.15 分贝升至 5.30 分贝，纯二重奏的中位数则从 12.18 分贝升至 18.73 分贝。视频令牌对同一演奏者的混合音频也有帮助，因为外观在此无法发挥作用（从 3.86 升至 6.31 分贝）；因此，它们携带了关键点未能捕捉到的演奏者动作信息。对于木管乐器与铜管乐器，两种模型的表现相近（3.83 对 3.62 分贝）。我们尚未训练同时使用视频令牌且不使用关键点的模型。

镜像处理与裁剪尺寸。 镜像增强在 URMP 的每一列中都有帮助（第 e 行和第 a 行）：木管乐器与铜管乐器的得分从 0.66 升至 3.62 分贝，纯二重奏从 11.04 升至 18.73 分贝；58 名演奏者中，正确点击获胜的人数从 47 人增至 51 人。留出集得分没有变化（9.54 和 9.56 分贝）。视频令牌对演奏者在画面中的大小也很敏感。我们的训练裁剪宽约 360 像素，而 URMP 的裁剪宽约 480 像素。将 URMP 视频缩小至 960 像素后，裁剪宽度缩至约 240 像素，无镜像模型在所有演奏者上的得分便从 3.29 降至 0.44 分贝。

预训练分离器。 如果没有 Mega 的权重（第 g 行与第 f 行相比），纯二重奏的乐谱得分会从 6.75 分贝降至 0.51 分贝，留出集得分则从 9.54 分贝降至 2.08 分贝。在全分辨率下进行评分时，它也几乎失去了全部增益（第 e 行在所有演奏者上的得分为 0.37 分贝，而原本为 3.29 分贝）。它仍会跟随点击（在全分辨率下，58 名演奏者中有 48 名），但分离效果很差。预训练骨干网络提供了分离能力，而视频提供了选择演奏者的依据。

5.6. 其他提示

除了默认的三点（第 4 节）之外，我们还尝试了仅使用鼻子、在第一帧上绘制框，以及每次都将三个点随机移动，并让它们完整经过从 SAM 2.1 跟踪到分离器的整个流程。它们全部都有效（表 6）：58 名演奏者中有 48 或 49 名与默认提示的差距保持在 1 分贝以内，而纯二重奏保持在 18.94

和 19.18分贝。差异源于略有不同的

提示	All	Str.	W&B	Pure 二重奏	在 1 分贝以内 默认的
三点（默认）	5.30	6.20	3.62	18.73	
一点（鼻子）	4.25	4.25	4.24	19.18	第 48 项，共 58 项
Box	4.64	5.39	4.58	18.94	第 49 项，共 58 项
三点，已移动	5.63	5.97	3.58	19.12	第 49 项，共 58 项
文本，通过 SAM 3	−0.77	−1.55	−0.77	18.73	

表6。**其他提示**。 我们的模型，SDR 中位数（分贝），未重新训练。移动后：每个点随机移动，距离最多为演奏者身高的 10%。

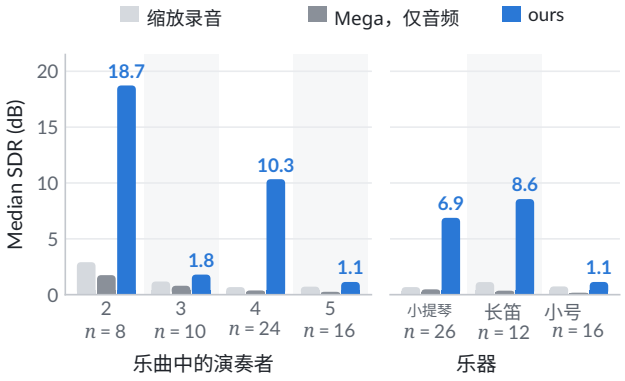

图5。**方法的适用范围**。 按每首乐曲的演奏者数量（左）和乐器（右）统计的 SDR 中位数； n：点击的演奏者。中提琴和萨克斯管各只有一组配对，未绘出。

SAM 2.1 掩码，而非来自不同演奏者的掩码。

文本提示也可以替代点击。SAM 3 [2]会将“左侧小提琴”之类的短语转换为掩码，而分离器保持不变。对于 58名玩家，SAM 3 能为其中 33 名选出目标演奏者，其中包括纯二重奏中的全部 8 名，所得结果与点击相同。它会为 11 名选出另一名演奏者，例如把小提琴选成中提琴；对于 14 名则找不到演奏者，其中大多数是长笛，此时不会分离出任何声音。

5.7. 适用场景与失效场景

我们的模型在图5中的每一组都优于所有基线。它在纯二重奏（18.73 分贝）和四人乐曲（10.34 分贝）中表现最强，在长笛和小提琴上也是如此。三组表现较弱：三人乐曲（1.80 分贝）、五人乐曲（1.14 分贝）和小号（1.14 分贝），每组都只略高于缩放录音。小号在与圆号、长号或大号同时演奏时表现不佳，而小号二重奏本身表现良好（19.73 和 20.58 分贝）。五人乐曲中的两对小提琴也表现不佳，尽管跟踪准确，且每对的两个声部演奏的音符并不相同；我们未能找到原因。总体而言，在 58 名演奏者中，有 12 名使用我们的模型时效果不如缩放录音：模型对大多数演奏者帮助很大，但对少数演奏者毫无帮助。

6. 结论与局限性

我们研究了当另一名演奏者使用相同乐器时，如何分离被点击的音乐家。仅使用合成混合音频进行训练后，我们的视频条件分离器在 URMP 的 58 名演奏者中取得了 5.30 分贝的 SDR 中位数，在同乐器二重奏中则达到 18.73 分贝。动作控制结果表明，性能提升取决于视听对齐。在一些规模较大的合奏团和混合铜管乐曲中，分离效果仍不够理想。评估使用的是分开录制的声部，部分设计选择参考了 URMP，并且每种设置只训练了一次。该模型使用 26 种乐器进行训练。结果可能不适用于其他乐器或其他音乐风格。现场合奏团以及无效点击的拒绝处理仍有待进一步研究，同时也无法排除与 MVSep Mega 未公开训练数据存在重叠的可能性。

大语言模型使用声明

我们使用了 Anthropic 的大语言模型 Claude 作为写作与编辑助手。**文本：** 该模型对作者撰写的论文草稿和图注进行了编辑，以提升清晰度、简洁性，并确保术语和符号的一致性。**图示：** 作者使用 draw.io 手动绘制示意图，之后在模型的协助下对其进行了完善。该模型根据我们的数据编写了结果图的绘图代码。**代码**：作者使用 Claude Code，让模型编写了部分训练、评估和消融实验代码。作者审阅并核实了所有经模型编辑的文本、图示和代码，并对论文内容承担全部责任。

参考文献

[1] Mido Assran、Adrien Bardes、David Fan、Quentin Garrido、Russell Howes、Matthew Muckley、Ammar Rizvi、Claire Roberts、Koustuv Sinha、Artem Zholus 等。V-JEPA 2：自监督视频模型助力理解、预测与规划。*arXiv* 预印本 *arXiv:2506.09985*，2025年。3[2] Nicolas Carion、Laura Gustafson、Yuan-Ting Hu、Shoubhik Debnath、Ronghang Hu、Didac Suris Coll-Vinent、Chaitanya Ryali、Kalyan Vasudev Alwala、Haitham Khedr、Andrew Huang 等。SAM 3：基于概念进行任意对象分割。发表于国际学习表征会议（*ICLR*），2026年。3，8[3] Jiaben Chen、Renrui Zhang、Dongze Lian、Jiaqi Yang、Ziyao Zeng 和 Jianbo Shi。i Query：将乐器作为查询，用于视听声音分离。发表于*2023 IEEE/CVF* 计算机视觉与模式识别会议（*CVPR*），第 14675–14686 页。IEEE，2023年。1，2，5[4] Xize Cheng、Siqi Zheng、Zehan Wang、Minghui Fang、Ziang Zhang、Rongjie Huang、Shengpeng Ji、Jialong Zuo、Tao Jin 和 Zhou Zhao。OmniSep：通过查询混合实现统一的全模态声音分离。发表于*ICLR*，2025年。12[5] Chuang Gan、Deng Huang、Hang Zhao、Joshua B Tenenbaum，以及 Antonio Torralba。Music Gesture：用于视觉声音分离。载于2020年IEEE/CVF计算机视觉与模式识别会议（CVPR），第10475–10484页。IEEE，2020年。2，3，12[6] Chao Huang、Susan Liang、Yapeng Tian、Anurag Kumar 和 Chenliang Xu。通过视觉引导的生成式建模实现多样类别中的高质量声音分离。《国际计算机视觉杂志》，134(3):104，2026年。1，2，5[7] Tao Jiang、Xinchen Xie 和 Yining Li。RTMW：实时多人二维和三维全身姿态估计。arXiv预印本 arXiv:2407.08634，2024年。3[8] Bochen Li、Karthik Dinesh、Zhiyao Duan 和 Gaurav Sharma。观看与聆听：基于乐谱的音轨与演奏者关联，应用于室内乐演出视频。载于2017年IEEE国际声学、语音与信号处理会议（ICASSP），第2906–2910页。IEEE，2017年。2[9] 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出数据集：挑战、见解与应用。《IEEE多媒体汇刊》，21(2):522–535，2018年。2，4[10] Bochen Li、Karthik Dinesh、Chenliang Xu、Gaurav Sharma 和 Zhiyao Duan。室内乐演出中的在线视听源关联。《国际音乐信息检索学会汇刊》，2 42，2019 1 29 2 ( ): – 。[11] Wei-Tsung Lu、Ju-Chiang Wang、Qiuqiang Kong 和 Yun-Ning Hung。使用频带分割 RoPE 变换器进行音乐源分离。载于ICASSP 2024-2024 IEEE国际声学、语音与信号处理会议（ICASSP），第481–485页。IEEE，2024年。1，2，3[12] Jan Melechovsky、Ambuj Mehrish、Abhinaba Roy 和 Dorien Herremans。Sonicmaster：迈向可控的一体化音乐修复与母带处理。arXiv预印本 arXiv:2508.03448，2025年。2[13] Juan F Montesinos、Olga Slizovskaia 和 Gloria Haro。Solos：用于视听音乐分析的数据集。载于2020年第22届IEEE多媒体信号处理国际研讨会（MMSP），第1–6页。IEEE，2020年。3[14] Sanjeel Parekh、Slim Essid、Alexey Ozerov、Ngoc QK Duong、Patrick P´erez 和 Ga¨el Richard。利用视频对象信息引导音频源分离。载于2017年IEEE音频与声学信号处理应用研讨会（WASPAA），第61–65页。IEEE，2017年。2[15] Sanjeel Parekh、Slim Essid、Alexey Ozerov、Ngoc QK Duong、Patrick P´erez 和 Ga¨el Richard。运动信息引导的音频源分离。载于2017年IEEE国际声学、语音与信号处理会议（ICASSP），第6–10页。IEEE，2017年。2[16] Nikhila Ravi、Valentin Gabeur、Yuan-Ting Hu、Ronghang Hu、Chaitanya Ryali、Tengyu Ma、Haitham Khedr、Roman R¨adle、Chloe Rolland、Laura Gustafson 等。SAM 2：图像与视频中的任意对象分割。载于国际学习表征会议（ICLR），2025年。2，3[17] Abhinaba Roy、Biplab Banerjee 和 Vittorio Murino。用于中层动作的判别式身体部位交互挖掘

表示与分类。《视觉传播与图像表示杂志》，55:829–840，2018年。3[18] Abhinaba Roy、Junyi Liang 和 Dorien Herremans。Merit：学习用于音频相似度的解耦音乐表示。arXiv预印本 arXiv:2605.27346，2026年。2[19] Saurjya Sarkar、Emmanouil Benetos、Mark Sandler 等。EnsembleSet：用于室内乐团分离的全新高质量合成数据集。载于ISMIR，第625 632页，2022 2 – 。[20] Michael Schoeffler、Fabian-Robert St¨oter、Bernd Edler 和 J¨urgen Herre。迈向新一代基于网络的实验：依据ITU-R BS.1534建议（MUSHRA）评估基本音频质量的案例研究。载于第1届Web音频会议（WAC）论文集，第1–6页，2015年。6，15[21] Bowen Shi、Andros Tjandra、John Hoffman、Helin Wang、Yi-Chiao Wu、Luya Gao、Julius Richter、Matt Le、Apoorv Vyas、Sanyuan Chen 等。SAM Audio：音频中的任意对象分割。arXiv预印本 arXiv:2512.18099，2025年。2，5[22] Olga Slizovskaia、Gloria Haro 和 Emilia G´omez。面向乐器演奏的条件式源分离。IEEE/ACM音频、语音与语言处理汇刊，29:2083–2095，2021年。3[23] Roman Solovyev、Ilya Kiselev、Alexander Stempkovskiy 和 Tatiana Gabruseva。Music-Source-Separation-Training（MSST）：用于训练和评估音乐解混模型的统一框架，2026年。1，2，3，5，11[24] Fabian-Robert St¨oter、Antoine Liutkus 和 Nobutaka Ito。2018年信号分离评测竞赛。载于国际潜变量分析与信号分离会议，第293–305页。Springer，2018年。4[25] Jianlin Su、Murtadha Ahmed、Yu Lu、Shengfeng Pan、Wen Bo 和 Yunfeng Liu。RoFormer：采用旋转位置嵌入的增强型Transformer。《神经计算》，568 127063，2024 3：。[26] Akira Takahashi、Shusuke Takahashi 和 Yuki Mitsufuji。MMAudioSep：驾驭视频到音频生成模型，实现视频/文本查询声音分离。载于ICASSP，2026年。12[27] Hang Zhao、Chuang Gan、Andrew Rouditchenko、Carl Vondrick、Josh McDermott 和 Antonio Torralba。像素之声。载于欧洲计算机视觉会议（ECCV），第587–604页。Springer，2018年。2，3，12[28] Hang Zhao、Chuang Gan、Wei-Chiu Ma 和 Antonio Torralba。运动之声。载于2019年IEEE/CVF国际计算机视觉会议（ICCV），第1735–1744页。IEEE，2019年。1，2，3，12

W谁演奏什么：从视频中分离同种乐器的音乐 o

补充材料

本文档提供论文中提及的设置、完整表格和评估方案。除非另有说明，所有得分均为分贝制中位数信号失真比，与论文中的定义相同。我们的方法是论文中的模型，从训练第一步起便使用各种混合音频进行训练。课程学习、无视频、仅关键点和无镜像模型，以及未使用 Mega 权重的模型，均为消融实验（论文，第4节）。

A. 训练和推理设置

损失函数。论文中的损失函数（第3.3节）由一个 ℓ_1 波形项和五个复数频谱 ℓ_1 项组成，权重均为1：

$$\mathcal{L} = \mathcal{L}_{\text{wav}} + \sum_{r=1}^{5} \mathcal{L}_{\text{spec}}^{(r)}, \tag{S1}$$

$$\mathcal{L}_{\text{wav}} = \|\hat{s}_a - s_a\|_1, \tag{S2}$$

$$\mathcal{L}_{\text{spec}}^{(r)} = \|\text{STFT}_r(\hat{s}_a) - \text{STFT}_r(s_a)\|_1, \tag{S3}$$

其中， s_a 是训练混合音频中的目标音频，$\hat{s}_a$ 是模型对其的估计值（论文，公式4）。这里， $\|\cdot\|_1$ 表示所有元素的平均绝对值；对于频谱项，它涵盖实部和虚部。短时傅里叶变换 $_r$ 采用长度为4096、2048、1024、512或256个采样点的汉宁窗，帧移为147个采样点，且不进行归一化。这些是 MVSep Mega 训练配置中的损失函数设置， [23],我们的骨干网络由此开始。

优化。 我们使用 AdamW 进行训练（学习率 3×10^{-4}, $\beta =$ (0.9,0.95)，权重衰减为 0.01），预热 1,000 步，并在 52k 步内采用余弦调度。预训练骨干网络的学习率为新层的六分之一。每个批次包含 32 个、时长为 4 秒的混合音频，在四个图形处理器上各处理八个。对于每段混合音频，第二位演奏者会被选为不同乐器（概率为 0.20）、来自另一个视频的相同乐器（0.30）、来自另一个视频且音高和速度匹配的相同乐器（0.25），或同一演奏者在另一时刻的演奏（0.25）；此外，还会加入零至三个其他乐器的音频片段。我们在固定的验证集混合音频上测量目标演奏者的信噪比；未经处理的混合音频得分为 −3.97 分贝，并保留信噪比最高的检查点。我们的模型在第 37k 步达到最佳验证信噪比 10.27 分贝。训练在四块 RTX PRO 6000 GPU 上耗时 13.7 小时。镜像视频的 V-JEPA 2 令牌在训练前计算一次。

模型大小。 我们的模型有 43.9M 个训练参数，其中 26.3 M 是预训练骨干网络。仅使用关键点的模型有 43.6M 个训练参数。SAM 2.1 (Hiera-L)、V-JEPA 2 (ViT-L) 和 RTMW 均被冻结，且未计入。

推理。 我们以 4 秒为窗口处理一首乐曲，窗口之间采用 50% 重叠，并使用汉宁窗对窗口进行融合。预训练骨干网络需要立体声输入，因此我们将单声道录音同时馈入两个声道，并对两个输出取平均。

B. 基线方法的运行方式

我们没有训练或调优任何基线。

SAM Audio。 我们使用已发布的大型检查点 sam-audio-large-tv。视觉提示是所点击演奏者的 SAM 2.1 掩码。我们关闭排序器，并以 10 秒为块处理乐曲，块之间重叠 1 秒。

iQuery。 我们使用作者提供的代码和在 MUSIC 上训练的权重。模型接收所点击演奏者的裁剪图像，以及该演奏者所用乐器的名称作为查询。

DAVIS-Flow。 我们使用作者提供的代码和在 MUSIC 上训练的权重。模型接收所点击演奏者的裁剪图像。由于模型的训练目标是从两个声源的平均值中还原出一个声源，因此我们将其输出缩放为原来的一半。不进行此缩放时，该模型在所有演奏者上的得分为 −2.40 分贝，在纯二重奏上的得分为 0.01 分贝；进行缩放后，得分分别为 0.01 和 0.02 分贝。

MVSep Mega。 模型会获知点击的演奏者所演奏的乐器，并返回该乐器的音轨，其中包含所有演奏该乐器的演奏者。

采样率。 iQuery 和 DAVIS-Flow 的音频输出采样率为 11 千赫。所有输出都会重采样至 48 千赫，并与全频带音轨进行评分，因此这两个模型会丢失 5.5 千赫以上的内容。

留出混合音频。 iQuery 在乐器属于其 11 个类别之一的 220 个混合音频上进行评分，Mega 则在其拥有音轨的 408 个混合音频上进行评分。独奏视频中只出现一个人，因此 SAM Audio 接收不带掩码的片段，而 iQuery 和 DAVIS-Flow 接收完整帧。500 个混合音频中有 255 个使用的视频出现在 iQuery 或 DAVIS-Flow 的 MUSIC 训练列表中。不使用这些混合音频时，我们的模型在所有混合音频上的得分为 9.71 分贝，在仅两名演奏者的情况下为 13.54 分贝；所有基线的得分均不超过 0.57 分贝。

MUSIC基准测试。 为确认我们正确运行了 iQuery 和 DAVIS-Flow，我们按照它们各自论文中的基准测试（论文，表 3）对其进行评分。我们从 MUSIC 的 77 段测试录音中构建了 200 个由两种不同乐器组成的混合音频

Row	中位数 [95% 区间]	我们的方法更优
未改变的录音	−6.29 [−6.79, −5.61]	58 项中的 58 项
缩放录音	0.82 [0.66, 1.00]	58 页中的第 46 页
SAM Audio	−3.92 [−4.58, −3.16]	58 页中的第 57 页
iQuery	−0.45 [−0.78, −0.10]	58 个中的 51 个
DAVIS-Flow	0.01 [−0.09, 0.01]	58 个中的 47 个
MVSep Mega	0.30 [0.19, 0.78]	58 个中的 42 个
课程学习模型	3.43 [1.01, 6.79]	58 个中的 44 个
我们的方法，静态帧	−0.97 [−1.69, −0.47]	58名中的 50名
Ours	5.30 [1.86, 10.31]	

表 S1. **全部58名演奏者的区间和计数**。 我们的方法更优：得分高于该行方法的演奏者数量。

以及 MUSIC-21 的录音，这些录音涵盖了 11 种 MUSIC 乐器。这些录音均未用于训练我们的模型或选择其检查点。我们遵循两种基线方法的协议。窗口长度为 5.94 秒（65,535 个采样点）。每个模型接收混合音频和完整帧，具体方式与各自代码完全一致：iQuery 和我们的模型接收两个声源之和，而 DAVIS-Flow 接收它们的平均值。得分为窗口内的 BSS Eval v3 SDR，对两个声源和所有混合音频取平均，采样率为 11,025 Hz，我们的输出也重采样至该采样率。两种基线方法的结果与其已发表的分数相差不到 0.5 分贝：iQuery 达到 11.61 分贝（其论文中为 11.17），DAVIS-Flow 达到 12.21 分贝（其论文中为 12.01）。我们的模型达到 19.08 分贝，课程学习模型达到 18.89 分贝。与 DAVIS-Flow 相比，我们的模型在 200 个混合音频中的 183 个上表现更好；与 iQuery 相比，则在 187 个上表现更好。

无法运行的方法。 我们没有找到 Sound of Motions 的公开代码 [28] 和 Music Gesture 的公开代码 [5], 而且我们无法获得 Sound of Pixels 的权重 [27], MMAu-dioSep [26]和 OmniSep [4]。

C. 区间和计数

表 S1 列出了每一行的区间和计数，包括本文的课程学习模型和静态帧测试。中位数的区间来自10,000次自助法重采样；每次重采样时，配对中的两名演奏者会一同抽取。

纯二重奏。 对于四首纯二重奏中的八名演奏者，我们模型的区间为 11.74 至 20.16 分贝（中位数为 18.73）。对于全部八名演奏者，我们的模型均优于所有已发布的基线，也优于缩放录音（2.91分贝）。这些区间基于四首乐曲，结果较为粗略。

D. 按合奏规模和乐器划分的结果

表 S2 列出了论文图5中每种方法背后的数据。我们的模型在每个组中都优于缩放录音；优势最小的分别是三人乐曲（1.80 对 1.18 分贝）、五人乐曲（1.14 对 0.72 分贝）和小号（1.14 对 0.74分贝）。

E. 留出混合音频

表 S3 按演奏者人数划分了500个留出混合音频。对于任何演奏者人数，都没有基线方法达到1.1分贝；只有四位演奏者的 Mega（1.08分贝）高于缩放混合物（0.94分贝）。随着演奏者人数增加，每个模型的表现都会下降；我们的模型从两位演奏者时的14.36分贝降至五位演奏者时的6.21分贝。表 S4 按第二位演奏者的选择方式对相同的混合音频进行了划分。对于每种类型，消融实验的排序都相同：无视频、仅关键点，然后是课程学习模型。对于每种类型，我们的模型与课程学习模型的差距都在0.6分贝以内。这些混合音频的构建方式与我们的训练混合音频相同，因此对我们的模型而言属于域内数据，对基线方法而言则属于域外数据。

F. 时间偏移

对于每个被点击的演奏者，我们将演奏者标记相对于音频固定延后一定时间，并在乐曲末尾循环回到开头。对于使用 V-JEPA 2 令牌的模型，这些令牌与关键点标记会一起偏移。所有模型均使用论文中的检查点，不进行重新训练。表S5 列出了我们测试的所有偏移量。

逐个演奏者来看。 偏移 0.08 秒时，在我们的模型中，58 名演奏者里有 45 名的得分低于未偏移时；课程学习模型中有 46 名，仅关键点模型中有 44 名。每位演奏者的得分降幅中位数分别为 1.06、1.04 和 1.64 分贝。对于我们的模型，这一人数在偏移 0.16 秒时增至 46 名，偏移 0.4 秒时增至 48 名，偏移 4 秒时增至 50 名。

G. 运动量

对于每位点击的演奏者，我们取模型接收的关键点速度，并计算乐曲中检测到该关键点的各帧的平均速度。我们对 6 个手臂关键点和 42 个手部关键点分别进行此计算。演奏者的增益是我们的模型得分减去无视频模型得分。表格 S6 显示，无论是对所有演奏者还是在单种乐器范围内，都没有显著相关性。演奏得更响的演奏者也可能动作更多，因此最后一列剔除了录音中各演奏者的响度影响（偏相关）。在 29 组配对中，动作更多的演奏者有 13 组获得了更大的增益。增益确实与演奏者在录音中的响度相关（斯皮尔曼 0.37，$p = 0.004$）。课程学习模型呈现出相同的情况：对所有演奏者而言，三种相关性分别为 0.04、−00.05 和 0.16；在 29 组配对中，动作更多的演奏者有 12 组获得了更大的增益。我们只检验了平均速度，没有尝试其他动作指标。

H. 点击测试

表 S7 给出了所有模型的点击测试结果。对于我们的模型，弦乐演奏者中有 28 人中的 24 人正确点击获胜，木管和铜管乐器演奏者中有 30 人中的 27 人正确点击获胜。

组	n	未改变录音	缩放录音	SAM Audio	iQuery	DAVIS-Flow	MVSep Mega	课程学习模型	关键点only	**Ours**
乐曲中的演奏者										
2	8	0.00	2.91	−0.20	0.29	0.02	1.75	15.66	12.18	18.73
3	10	−4.56	1.18	−2.88	−0.52	0.00	0.80	2.03	1.08	1.80
4	24	−6.70	0.68	−4.32	−0.57	0.01	0.38	6.07	6.03	10.34
5	16	−6.95	0.72	−4.90	−0.13	−0.29	0.26	−0.06	0.26	1.14
所点击玩家的乐器										
小提琴	26		0.68	−4.06	−0.54	0.01	0.48	3.66	5.31	6.88
中提琴	2		0.94	−2.68	−1.80	−1.84	−0.16	1.63	1.74	4.03
长笛	12		1.13	−3.16	0.00	0.02	0.36	7.05	4.23	8.56
小号	16		0.74	−4.40	−0.75	0.00	0.18	0.68	1.08	1.14
萨克斯管	2		2.84	4.30	0.01	0.28	1.76	16.77	15.34	14.78

表 S2. **按合奏规模和乐器划分的 URMP**。 n 是点击的演奏者人数。课程学习模型和仅关键点模型是消融实验。中提琴和萨克斯管各只有一组配对，数量太少，无法对这些乐器得出结论。

混合音频中的演奏者	2	3	4	5	All
n	134	126	121	119	500
未改变的混合音频	−0.26	−4.19	−5.84	−7.45	−4.46
缩放混合物，x/N	2.65	1.28	0.94	0.62	1.09
SAM Audio	−0.07	−2.19	−3.31	−3.85	−2.47
iQuery*	0.46	−1.03	−0.46	−2.25	−0.66
DAVIS-Flow	0.29	−0.04	−0.23	−0.97	−0.18
MVSep Mega†	0.14	0.80	1.08	−0.04	0.26
仅关键点	11.68	8.43	5.34	3.58	6.46
课程学习模型	14.31	10.82	8.93	5.85	9.90
Ours	14.36	10.56	8.42	6.21	9.56

表 S3. **按演奏者数量划分的留出混合音频**。每个混合音频包含两位演奏者演奏的一种乐器，以及零至三种其他乐器。 *仅包括乐器属于 iQuery 的 11 个类别之一的混合音频：共 63、54、59 和 44 个，合计 220 个。 †仅包括 Mega 有对应音轨的混合音频：共 114、107、91 和 96 个，合计 408 个。两位演奏者这一列对应论文表 1 中留出的纯二重奏列。

第二位演奏者	n	No 视频	关键点only	课程学习模型	**Ours**
另一个视频	170	1.22	8.49	11.46	10.87
另一个视频，匹配	149	1.18	8.06	10.47	10.63
同一视频的另一个时刻	181	1.39	3.86	6.51	6.31
所有混合物	500	1.24	6.46	9.90	9.56
仅两名玩家	134	2.44	11.68	14.31	14.36

表 S4。**按第二位演奏者类型划分的留出混合音频**。匹配：另一个视频在音高和速度上相匹配。最困难的情况是同一演奏者在另一个时刻演奏，此时两位演奏者看起来完全相同。仅两名玩家：论文表 1 中的留出仅纯二重奏列。

偏移（秒）	**Ours** All	**Ours** 二重奏	课程学习 All	课程学习 二重奏	仅关键点 All	仅关键点 二重奏
0	5.30	18.73	3.43	15.66	4.15	12.18
0.04					2.36	6.90
0.08	1.38	10.54	0.67	9.14	0.45	2.93
0.12					0.17	2.17
0.16	0.11	6.79	−0.05	6.90	−0.04	3.89
0.20					−0.35	3.47
0.24	−0.06	8.00	−0.63	8.90		
0.4	−0.32	8.82	−0.98	2.50	−0.68	1.67
1.2	−0.57	4.73	−0.68	0.67	−0.67	1.48
2	−0.23	5.66	−0.61	2.08	−0.41	2.54
4	−0.54	6.43	−0.83	4.37	−0.71	1.02
无视频模型	0.52	2.95	0.52	2.95	0.52	2.95

表 S5。**时间偏移**。 所有 58 名演奏者以及纯二重奏中的 8 名演奏者的 SDR 中位数（单位：分贝）； 此处将演奏者的视频相对于音频按给定时长向后偏移。空白单元格表示未进行测试。采用 V-JEPA 2 令牌的模型只能按 0.08 秒的步长偏移，即每个令牌所覆盖的时长。

演奏者	n	手臂速度	手速	手臂速度，已去除响度
All	58	0.07 (0.60)	0.03 (0.80)	0.18 (0.17)
小提琴	26	0.11 (0.61)	0.31 (0.13)	0.24 (0.24)
小号	16	0.19 (0.47)	0.12 (0.66)	0.19 (0.47)
长笛	12	−0.23 (0.47)	−0.43 (0.17)	−0.17 (0.60)

表 S6. **动作更多的演奏者是否获益更多？** 演奏者平均关键点速度与我们的模型相较于无视频模型的增益之间的斯皮尔曼相关系数，括号内为 p 值。

I. 泄漏划分

论文（表 2）将每个输出中的不需要的声音分为搭档和其他乐器。我们将输出y 按 48 千赫切分为 1 秒的窗口，并对每个窗口进行最小二乘拟合

正方形，

$$y \approx a\,s + b\,q + c\,r, \tag{S4}$$

其中， s 是点击的演奏者的真实音轨， q 是搭档的真实音轨，而 r 是减去 s 和 q后的录音，其中包含其他乐器的声音，并且在 Pure 二重奏中为零。两个泄漏

	全部（58）		纯二重奏（8）	
方法	正确 / 其他胜场	正确	/ 其他胜场	
SAM Audio	−3.92 / −3.98	29	−0.20 / 0.20	4
iQuery	−0.45 / −0.34	23	0.29 / 0.69	1
DAVIS-Flow	0.01 / 0.01	27	0.02 / 0.01	4
(b) 课程学习		52		8
(d) 仅关键点		51		8
(e) 不进行镜像增强。		47		8
(g) 未使用预训练		49		5
(g)，全分辨率		48		
Ours, 静止帧		26		
Ours	5.30 / −2.19	51	18.73 / −1.76	8

表 S7。**点击测试**。 点击正确演奏者和配对中另一名演奏者时的 SDR 中位数，以及正确点击得分更高的演奏者人数（胜场）。中间各行是论文表5中的消融实验，字母标记相同。(g) 行按论文中的设置以 960 px 进行评分，也以全分辨率进行评分。空白单元格表示未报告。MVSep Mega 和无视频模型对两种点击返回相同的输出，因此正确点击从未获胜。

窗口的值为

$$10\log_{10}\frac{\|as\|^2}{\|bq\|^2} \quad \text{and} \quad 10\log_{10}\frac{\|as\|^2}{\|cr\|^2}, \tag{S5}$$

针对搭档和其他乐器。跳过被点击的演奏者未演奏的窗口。先对窗口取中位数，再对演奏者取中位数。此拟合对每个信号和窗口使用一个增益，而非 BSS Eval 的 512 抽头滤波器，因此其数值并非 museval 的信号干扰比。

J. 其他提示

点击方式变化。 除三个默认点（鼻子、肩部、髋部）外，还分别以三种方式为58名演奏者提供提示：仅鼻子、第一帧中演奏者周围的边界框，以及将三个点分别随机移动最多相当于演奏者身高10%的距离。每个提示都依次经过 SAM 2.1 跟踪、关键点提取、全分辨率 V-JEPA 2 特征处理和我们的模型处理，且未进行重新训练。与默认提示相比，得分低于1分贝以上的情况包括：仅使用鼻子的8名演奏者、使用边界框的5名演奏者，以及使用移动点的7名演奏者。得分高于1分贝以上的情况分别有2名、4名和2名演奏者。最差的单个案例是一名小提琴演奏者：使用默认提示时得分为9.20分贝，而仅使用鼻子时为1.13分贝。对于所有演奏者，课程学习模型在仅使用鼻子、使用边界框和使用移动点时分别达到3.22、3.04和2.95分贝；相比之下，我们的模型分别达到4.25、4.64和5.63分贝。

文本提示。 SAM 3 接收诸如“左侧小提琴”这样的短语，并返回一个掩码来替代点击。分离器未改变。SAM 3 为 58 名演奏者中的 33 名以及纯二重奏中的全部 8 名演奏者选中了目标演奏者。它

组	n	Ours	课程学习	我们的方法在以下情况下更高
2名玩家（纯二重奏）	8	**18.73**	15.66	8 个中的 6 个
3名玩家	10	1.80	**2.03**	8/10
4名玩家	24	**10.34**	6.07	20/24
5名玩家	16	**1.14**	−0.06	10/16
小提琴	26	**6.88**	3.66	26 个中的 19 个
中提琴	2	**4.03**	1.63	2 个中的 2 个
长笛	12	**8.56**	7.05	12 个中的 10 个
小号	16	**1.14**	0.68	13/16
萨克斯管	2	14.78	**16.77**	0/2
All	58	**5.30**	3.43	44/58

表 S8。 **在 URMP 上比较常规训练（我们的方法）与课程学习**。SDR 中位数，单位为分贝。

为 11 名演奏者选中了另一名演奏者，例如把中提琴误选为小提琴；随后，错误演奏者的声音得分介于 −4.6 和 −0.6 分贝之间。它没有为 14 名演奏者找到匹配的演奏者，其中大多数演奏长笛。对于这些情况，没有分离出任何声音，我们对未改变的录音进行评分。当 SAM 3 选中目标演奏者时，输出与点击所得相同。

点击没有人的地方。对于 29 首乐曲中的每一首，我们都在第一帧的背景上点击一次：点击位置取自网格点，网格间距为帧宽的 5%，并且该点距所有检测到的人最远。这些点与最近的人相距 240 至 320 px。点击会经过完整流程。相对于录音，输出的中位电平为 −7.5 分贝，范围为 −19.4 至 −4.0 分贝；而真实点击为 −6.1dB。对于包含木管乐器与铜管乐器的乐曲，输出更轻（小号乐曲中为 −10 至 −19dB），而在弦乐乐曲中则接近真实点击的电平。与配对中距离较近的那位演奏者相比，其信号失真比中位数为 −0.07 dB，而真实点击为 5.30 dB，因此它返回的是一段不属于任何一位演奏者的音乐。

K. 课程学习消融

调度。 课程学习模型与我们的方法唯一的不同在于训练混合音频的顺序。它先以不同乐器作为第二位演奏者训练 4k 步，然后以来自另一个视频的相同乐器训练 4k 步，再以同一表演者在另一时刻的演奏训练 4k 步，并从第 12k 步起使用我们模型的完整混合音频。模型、数据、镜像增强、步数预算、学习率调度和检查点选择均相同。三个阶段结束时，其验证信噪比分别为 0.99、3.10 和 4.29 分贝；在第 44k 步的最佳检查点处，验证信噪比达到 9.81 分贝。连续五次验证检查均未见性能提升后，提前停止机制于第 49k 步结束了训练；该训练在四块 RTX PRO 6000 GPU 上耗时 17.6 小时。

结果。 表格。 S8和 S9 比较了这两个模型。我们的模型在 URMP 的所有演奏者、纯二重奏、

Test	**Ours**	课程学习
留出，2名玩家	14.36	14.31
留出，3名玩家	10.56	10.82
留出，4名玩家	8.42	8.93
留出，5名玩家	6.21	5.85
留出 ，全部 500	9.56	9.90
留出 ，另一个视频	10.87	11.46
留出、匹配	10.63	10.47
留出、同一表演者	6.31	6.51
正确点击胜场 ，全部／二重奏	51／8	52／8
MUSIC基准测试	19.08	18.89
鼻子／方框／移动点，全部	4.25／4.64／5.63	3.22 / 3.04 / 2.95
时间偏移 0.08 秒，全部 / 二重奏	1.38 / 10.54	0.67 / 9.14
时间偏移 0.4 秒，全部 / 二重奏	−0.32 / 8.82	−0.98 / 2.50
验证信噪比	10.27	9.81

表 S9. **常规训练（我们的方法）与其他测试中的课程学习进行比较。** 在留出混合音频中，我们的模型在 500 个混合音频中的 271 个上表现更好。

四人和五人合奏曲中的小提琴、中提琴、长笛和小号、验证集混合音频、MUSIC基准测试以及使用其他所有提示时，表现都更好。课程学习模型在三人乐曲中的中位数更高，不过在这 10 位演奏者中，我们的模型在 8 位演奏者上的表现更好；在唯一一组萨克斯管配对中也是如此。在留出混合音频上，这两个模型的差异不大：分别为 9.56 和 9.90 分贝，而我们的模型在 500 个混合音频中的 271 个上表现更好。课程学习模型在三位和四位演奏者参与的情况下，以及面对三种第二位演奏者中的两种时，表现高出 0.2 至 0.6 分贝；而我们的模型在两位和五位演奏者参与的情况下，以及在匹配视频上表现更好。在纯二重奏中，视频偏移 0.4 秒或更长时，课程学习模型的乐谱分数下降更多（表 S5）。每个模型我们只训练了一次，因此不能据此声称课程学习导致了这一结果。

L. 失败案例

对于 58 位演奏者中的 12 位，我们的方法不如缩放录音；对于其中 16 位，我们的方法未能胜过 MVSep Mega（论文，表 1）。

与其他铜管乐器相邻的小号。 五对小号分别与长号、圆号或大号相邻演奏。使用我们的模型时，每对中的两位演奏者得分分别为 1.07 和 −3.22、1.21 和 −0.85、0.37 和 −2.00、2.55 和 −0.50，以及 −3.41 和 −2.06 分贝。小号二重奏的得分为 19.73 和 20.58 分贝。

两对小提琴。 在五人乐曲中，两对小提琴使用我们的模型时未能成功分离。四位演奏者的得分为 −3.32、−2.30、−3.12 和 −2.24 分贝，均低于缩放录音。跟踪结果正确，并且每对中的两个声部并未演奏相同的音符。我们不知道这些配对为何失败。

得分最低的演奏者。 四个最低得分分别为 −3.41 分贝（小号）、−3.32分贝（小提琴）、 −3.22 分贝（小号）和−3.12 分贝（小提琴）。

试验	URMP 乐曲	演奏者	点击的演奏者	开始时间（秒）
1	04	2	长笛	5
2	05	2	小号	29
3	02	2	小提琴	5
4	14	3	长笛	15
5	13	3	小提琴	5
6	25	4	小提琴	29
7	34	4	小号	83
8	29	4	长笛	11
9	40	5	长笛	5
10	44	5	中提琴	26

表 S10。**聆听测试试验。**每首 URMP 乐曲对应一个试验。演奏者：乐曲中的演奏者。开始时间：10 秒摘录片段的起始位置。试验 1 至 3 为纯二重奏。

M. 听音测试方案

本节介绍论文第 5.2 节和表 4 中听音测试的流程。该测试遵循 MUSHRA [20]，并在网页浏览器中进行。

试验。 试验中会显示一帧视频静止画面，并用红框标出点击的演奏者。页面会告知听众，图片仅用于说明点击了哪位演奏者，评判音频片段时应凭听觉判断。听众可以播放参考音频（即点击的演奏者的真实音轨），以及七段未标注的音频片段：隐藏参考音频、我们的模型、MVSep Mega、SAM Audio、iQuery、DAVIS-Flow，以及作为锚点的未改变的录音。听众需按每段音频与参考音频的相似程度打分，范围为 0 到 100，评分等级依次标为差、较差、一般、良好和优秀。说明要求听众听到其他演奏者的声音、失真或音符缺失时，均相应降低该音频片段的评分；同时也会告知听众，其中一段音频片段本身就是参考音频。只有播放过的音频片段才能解锁对应的滑块，且听众可以按需反复播放任何音频片段。缩放录音不纳入测试，因为在响度匹配后，它听起来与锚点相同。

试验与音频片段。共有 10 个试验，每首 URMP 乐曲对应一个（表 S10）：其中包括 3 个纯二重奏、2 个三重奏、3 个四重奏和 2 个五重奏，涉及 4 名弦乐演奏者和 6 名木管或铜管演奏者。这些试验通过固定随机种子随机抽取，在各组之间保持平衡，且每首乐曲只出现一次。每个试验的片段取自乐曲开始 5 秒之后的第一个 10 秒窗口，在此窗口中，配对中的两位演奏者都至少演奏 9 秒。当演奏者的真实音轨高于 −40dBFS 时，即视为正在演奏。这项规则在任何人听取输出之前就已确定。每个音频片段都截取自论文中评分的整首乐曲输出，归一化至 −23 LUFS，并以 44.1 千赫的单声道 WAV 文件格式存储。每位听众的试验顺序以及每个试验中的音频片段顺序都会随机打乱。

听众。共有 16 名听众参与，其中没有人是本文作者。页面要求他们戴上耳机

并在安静的房间里进行。进度会保存在浏览器中，因此听众可以暂停后再继续。没有移除任何听众。

分析。 对于每种方法，我们报告平均分数及其95%自助法区间；自助法区间是对听众和试验进行重采样后计算的，分别涵盖全部10次试验和3次二重唱试验（论文，表4）。为了将我们的模型与每种基线方法进行比较，我们计算每位听众对每种方法的平均评分，并对16位听众应用精确的双侧Wilcoxon符号秩检验，再对四种基线方法进行Holm校正。每位听众给我们模型的平均评分都高于任何一种基线方法，因此每项检验均得到 $p = 2/2^{16} \approx 3.1 \times 10^{-5}$，经校正后为 1.2×10^{-4}。