arXiv:2304.08490v1 [cs.CV] 17 Apr 2023

通过拟音类比从视频条件生成音频

Yuexi Du[1,2] Ziyang Chen[1] Justin Salamon[3] Bryan Russell[3] Andrew Owens[1]

[1]密歇根大学 [2]耶鲁大学 [3]Adobe Research

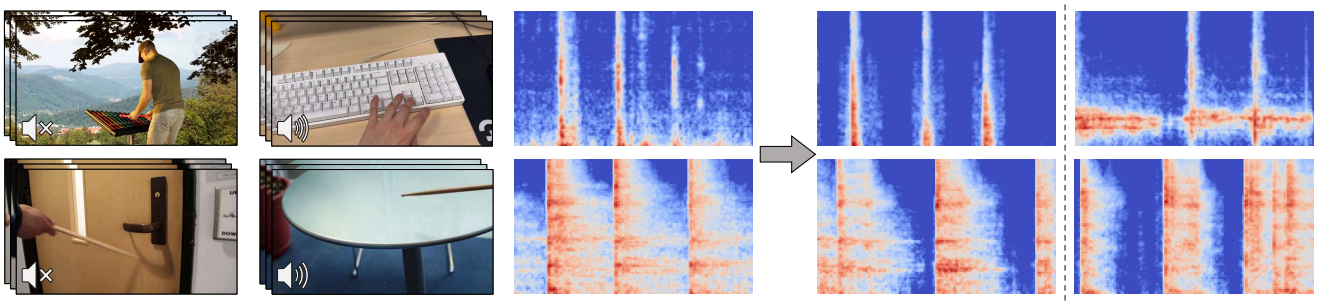


图 1. **通过类比进行条件 Foley 生成**。给定用户提供的条件示例（该示例指定音频应听起来“像什么”），我们为静音输入视频生成音轨。在第一个示例中，我们让木琴敲击声听起来像机械键盘的按键声。在第二个示例中，我们为一段视频生成音轨：视频中鼓槌敲击木门，听起来却像是敲击金属。请注意，生成音频中的声音事件形态（例如，上方示例中的细条纹）与条件音频相符，其起始时刻则与输入示例音频相符。作为参考，我们在右侧提供了输入视频（留出）的声音。**我们欢迎读者前往我们的项目网页**观看并聆听效果。

摘要

设计师为视频添加音效，旨在营造特定的艺术效果，因此音效可能与场景中的真实声音大相径庭。受到这一挑战的启发——即为视频创作音轨，使其不同于真实声音，却仍与屏幕上发生的动作相匹配——我们提出了条件 Foley这一问题。为解决这一问题，我们提出以下贡献。首先，我们提出一种代理任务，用于训练模型：使用从同一源视频中另一时刻采样的条件音视频片段，为输入视频片段预测声音。其次，我们提出一种模型，可根据用户提供的示例为静音输入视频生成音轨，该示例指定视频应听起来“像什么”。人类研究和自动评估指标表明，我们的模型能够成功地根据视频生成声音，同时根据所提供示例的内容调整输出。项目网站：*https://xypb.github.io/CondFoleyGen*。

1. 引言

为视频创作音效时，艺术家常常会从其他来源“借用”声音，然后对其进行处理以配合屏幕上的动作。这些艺术家的目的未必是传达场景中真实的声音，而是营造所需的艺术效果。因此，椰子壳碰撞发出的闷响会变成小跑的马蹄声，煎培根的滋滋声则会变成雨声[1]。

为视频创作音效的问题称为拟音， [1], 通常被表述为预测视频中同时出现的声音， [29,42,68]。然而，声音艺术家所从事的任务略有不同。他们为视频创作的音轨并非视频中的真实声音，但仍能与屏幕上发生的事件合理匹配。此外，这些先前的系统大多没有提供对输出声音的控制能力。

为了在帮助拟音师的同时赋予他们艺术创作的控制权，我们提出了一种条件拟音问题，其灵感来自图像类比领域的经典研究 [27]。我们的任务是根据用户提供的条件视听示例，为输入静音视频生成音轨。该示例指定了输入视频应该“听起来像什么”。生成的音轨应以与所提供示例类似的方式关联输入视频（图1）。这种表述自然地将选择范例声音的问题（这在某种程度上需要艺术家的判断）与操控该声音的问题区分开来

[1]我们鼓励你观看并聆听声音艺术家的工作方式：
https://www.youtube.com/watch?v=UO3N_PRIgX0

，使其与视频相匹配，例如精确调整其时序和音色。

这项拟议任务颇具挑战，因为系统必须学会调整范例（条件）声音，使其与静音视频的视觉内容在时序上相匹配，同时保留范例声音的音色。尽管以往的方法能够预测视频的声音 [29, 42, 68],，但无法融入艺术家的范例条件声音。此外，尽管视觉到声音方法可以将问题表述为根据图像预测视频音轨，但如何获得条件示例的监督信号仍不明确。

为应对这些挑战，我们提出了一项用于学习条件 Foley 的自监督代理任务，以及一个用于解决该任务的模型。我们的代理任务利用了这样一个事实：自然视频往往包含重复事件，这些事件会产生密切相关的声音。为了训练模型，我们从一个视频中随机抽取两组视听片段，并将其中一组用作另一组的条件示例。

我们的模型学会从条件示例中推断场景中的动作类型，并生成相似的声音来匹配输入示例。在测试时，我们的模型能够泛化到来自其他视频的条件声音。为解决该任务，我们训练了一个 Transformer [58]，使其在给定视听示例的条件下，自回归地预测频谱图 VQGAN [13]，的一系列音频编码。我们通过生成大量音轨，并使用视听同步模型 [8, 30, 41] 从中选出与视频时序对齐程度最高的声音，提升模型在测试时的表现。

我们在*Greatest Hits* 数据集 [42],上评估了模型，该数据集包含需要理解材料属性和物理交互的视频；此外，我们还通过高度多样化的*CountixAV* 数据集 [66] 中的定性示例进行评估。通过感知研究和定量评估，我们表明，模型生成的音轨既能传达条件示例中的物理属性，又能反映屏幕上动作的时序和运动。

2. 相关工作

根据图像和视频预测声音。在早期研究中，Van Den Doel 等人 [56] 为物理模拟生成声音。较新的例子包括：为有人用鼓槌敲击物体的视频预测音轨 [42], 根据钢琴演奏生成音乐 [35],根据身体运动 [50], 或舞蹈视频生成音乐 [14, 51], 以及根据唇部动作生成语音 [12, 44]。其他研究则使用自回归声码器 [68], 时间关系网络 [20], 以及视觉引导的生成对抗网络 [21]来预测自然声音（通常是环境声）。Iashin 和 Rahtu [29] 最近使用一种基于梅尔频谱图运行的 VQGAN [13, 57] 来生成声音。我们采用这一架构来执行条件式声音生成。相比之下

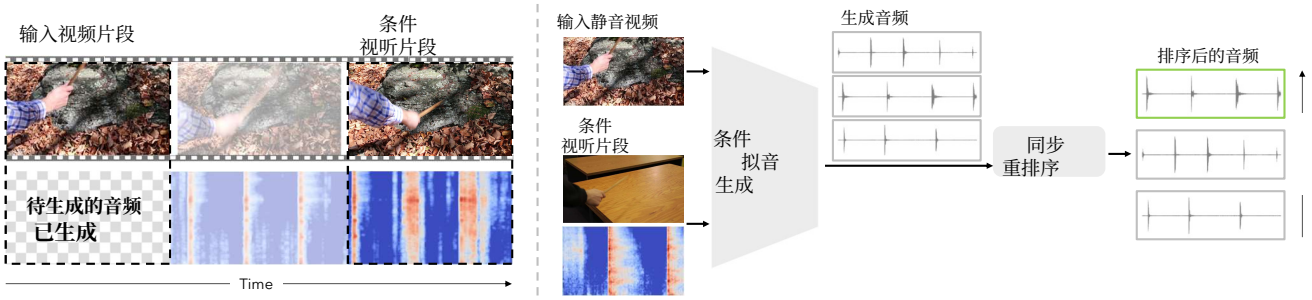
以往方法的目标并非只是根据静音视频估计声音，而是利用用户提供的示例，使声音与场景中的动作相匹配。

声音设计。 电影中出现的声音通常并非在现场录制，而是由艺术家后期加入。声音设计师会进行多个步骤，包括标记需要声音的视觉事件、选择或录制合适的声音，以及使用编辑软件 [1]处理选中的声音。我们的工作针对的是最后这一步处理。其他研究则尝试采用计算方法，在声音设计任务中重新定位视觉信号或使其与音频匹配。Davis 和 Agrawala [10] 通过按估计的节拍对齐视频和音轨，使视频重新定位以匹配音轨。Langlois 和 James [37] 研究了同步视觉以匹配声音的任务。还有研究学习如何为视频匹配相关音乐 [52]。然而，这些研究并未解决本文提出的条件化拟音任务。

交互式风格化。 我们从 Hertzmann 的经典作品中汲取灵感等人。 [27],该作品通过单个由用户提供的图像及其风格化版本示例，学习对输入图像进行重新风格化。与这项工作类似，我们旨在从一组配对数据（一段视频及其声音）推广到另一组配对数据。

Zhang等人 [65] 通过自监督训练流程，利用模拟的用户提示学习为图像着色。与此不同，我们的模型接收用户提供的、来自其他视频的视听示例提示，而不是来自输入内容的标注。已有多种方法被提出，用用户提供的条件对图像进行风格化。Li等人 [39]根据用户提供的声音对输入图像进行风格化。Lee等人 [38] 利用声音和预训练语言模型生成图像。Chen等人 [3]学习改变语音的声学特性，使其与视觉场景相匹配。近年来，许多方法已将风格迁移应用于音频。 [55]其中包括遵循 Gatys等人 [19], 的方法，利用特征统计量分离风格与内容；以及基于 Cycle- GAN [69]迁移音乐音色 [28] 的方法。与此不同，拟音生成需要在没有真实音频的情况下生成声音，因为在拟音师的工作流程中，通常不存在已有的录制声音 [1] 。Nistal等人 [40] 提出使用生成对抗网络 [22] 合成鼓声，其条件是音频感知特征，而非完整的视听条件片段。

自监督与少样本视听学习。我们的工作旨在从数据中学习，而无需人工标注。近年来，利用带有配套音轨的视频学习强大的视听表征，并将其用于其他下游应用，已取得显著进展 [2, 6, 9, 11, 15–17, 24, 41, 54, 63, 64, 67]。应用示例包括利用视听数据进行声源分离 [11, 15, 24, 41, 54, 67]，以及将单声道声音转换为立体声 [16, 17, 63, 64]。我们的工作关注的是另一种



(a) 用于训练的代理任务　　(b) 推理

图 2. **通过拟音类比进行条件视频到音频合成**。(a) 对于我们的代理任务，我们从较长的视频中提取两个片段，并训练模型根据其中一个片段的视听信息预测另一个片段的音轨。通过这一过程，模型学会根据其他视频来生成音轨预测。(b) 测试时，我们向模型提供一个静音输入视频和一个视听片段（取自另一段视频）。我们可以使用视听同步模型对生成的音轨重新排序，并选择时间对齐效果最佳的音轨。

应用，但与能够通过少量带标签样例进行自适应的方法相关，例如音频事件检测 [61, 62], 说话人头部生成 [5], 和说话人分离 [7]。

3. 方法

我们的目标是根据用户提供的条件视听示例，为静音输入视频生成音轨，该示例说明视频应当“听起来像什么”。我们学习一个由 θ 参数化的函数 $\mathcal{F}_\theta(\mathbf{v}_q, \mathbf{v}_c, \mathbf{a}_c)$，在给定条件视频$\mathbf{v}_q$和条件音频$\mathbf{a}_c$的情况下，根据输入视频生成音轨。下面介绍如何利用无标注数据训练 $\mathcal{F}_\theta$ 的**代理任务**，以及我们的条件视觉到声音模型。

3.1. 条件预测的代理任务

我们希望设计一种代理任务，使模型能够从每个来源中获取必要信息。具体而言，我们希望输入视频能够指定动作的类型（例如，敲击物体还是刮擦物体）及其时序，而条件视听示例则应指定生成声音的音色（例如，发生交互的材料类型）。

我们将任务定义为视频到音频的预测问题，其中提供来自同一视频的另一个片段作为条件示例（图2a）。训练期间，我们从较长的视频中采样两个片段，其中心时间分别为 t 和 $t+\Delta t$；其中一个用作条件示例，另一个用作输入视频。模型的任务是将条件视频片段作为额外输入，根据静音输入视频预测声音。

根据这一代理任务，我们可以定义一个损失函数 $\mathcal{L}$，用于衡量音频目标 $\mathbf{a}_g$ 与预测结果 $\mathcal{F}_\theta(\mathbf{v}_q, \mathbf{v}_c, \mathbf{a}_c)$；给定对应的输入视频$\mathbf{v}_q$ 和条件音视频片段 $(\mathbf{v}_c, \mathbf{a}_c)$：

$$\mathcal{L}(\mathbf{a}_g, \mathcal{F}_\theta(\mathbf{v}_q, \mathbf{v}_c, \mathbf{a}_c)) \quad (1)$$

这种表述利用了这样一个事实：视频中的动作往往彼此密切相关 [66]（或称“自相似” [49]），例如同一个动作被反复执行时。因此，随机抽取的视频片段对通常包含相关动作。出现这种情况时，模型可以利用条件声音来改进预测。然而，模型不能仅靠“复制粘贴”条件声音来完成任务，因为它必须考虑输入视频中动作的时序内容（以及运动类型）。由于模型经过训练，会假定条件示例能够为输入提供信息，我们通过实证发现，它学会了依据条件声音进行预测。这一发现使我们能够在测试时，替换为从完全不同的视频中抽取的条件声音（图2b）。

3.2. 条件声音预测架构

我们介绍我们的架构：$\mathcal{F}_\theta$ 首先，通过 VQGAN 获取目标声音的代码表示 [13, 29]其次，为给定的输入视频和条件视听对预测输出声音。

矢量量化音频表示。我们遵循 Iashin 和 Rahtu 的方法 [29]，使用 VQGAN 将预测声音表示为离散代码序列 [13, 57]，该 VQGAN 处理梅尔频谱图[2]。我们通过训练自编码器来重建数据集中的声音，并将代码用作其潜在向量，从而学习这种代码。训练完成后，预测代码序列便可转换为波形。

给定波形$\mathbf{a}$ 及其梅尔频谱图 MSTFT(a) $\in \mathbb{R}^{T\times F}$，我们计算嵌入 $\hat{\mathbf{z}}E$ $\in \mathbb{R}^{T'\times F'\times d}$ T' F' (MSTFT(a))，其中 和 定义了一个较低分辨率的时频网格，d 是每个图块处嵌入的维度，E 是一个卷积神经网络。每个

[2]除非另有说明，我们使用对数梅尔频谱图。

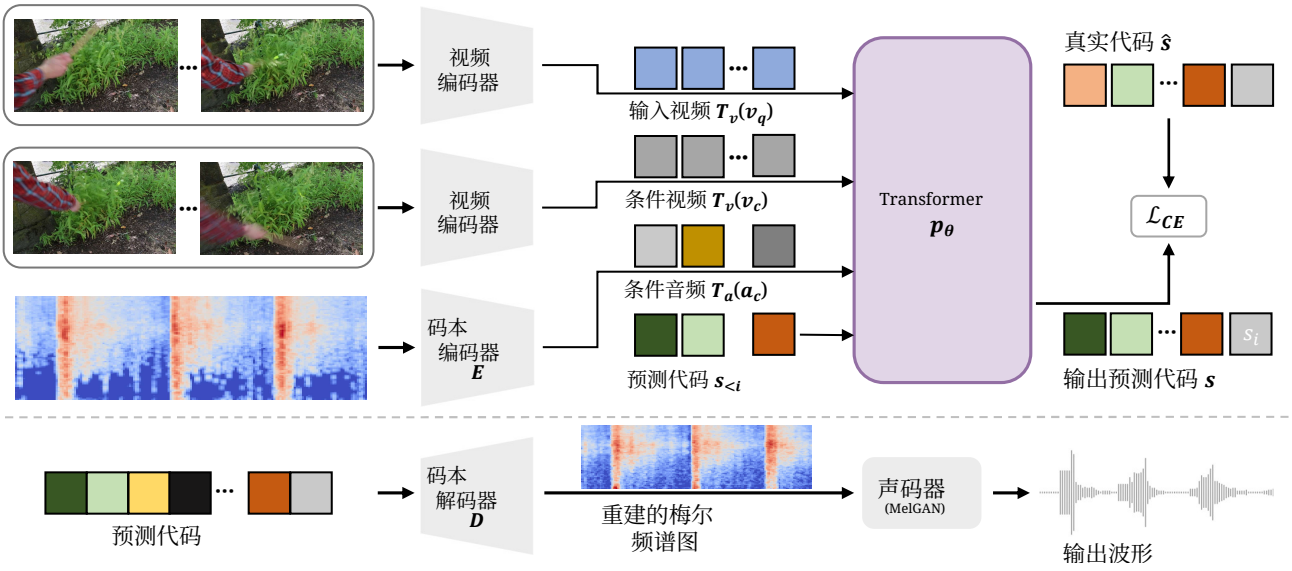


图 3. **条件 Foley 生成**。（上）我们以从同一视频中采样的视听对为条件，为静音视频预测音轨。我们对视频和音频信号进行编码和词元化，并将其输入 Transformer。该 Transformer 自回归地预测一个来自 VQGAN 的代码 [13, 29], , 用于表示输入示例的声音。（下）我们将代码转换为梅尔频谱图，再使用 MelGAN [36] 声码器将其转换为波形，从而生成波形。此处， $\mathcal{L}_{CE}$ 表示交叉熵损失。

嵌入向量随后被替换为码本 $\{\mathbf{c}_k\}_{k=1}^{K}$中最近的条目：

$$\mathbf{z}_{tf} = q(\hat{\mathbf{z}}_{tf}) = \underset{\mathbf{c}_k}{\operatorname{argmin}} ||\hat{\mathbf{z}}_{tf} - \mathbf{c}_k||, \qquad (2)$$

其中 $\hat{}\mathbf{z}^{tf}$ 是时频索引 (t, f)处的嵌入。$\hat{}$我们训练模型重建输入声音 $\mathbf{S}$ $=D(q(E(\text{MSTFT}(\mathbf{a}))))$，其中 D 是一个基于CNN的解码器，且 q 应用于每个嵌入。我们采用 [29], 中的损失函数，该函数将 VQGAN 损失 [13] 调整为适用于频谱图，并联合最小化均方误差重建损失 [57], 、感知损失 [32], 和基于图块的判别器损失 [31]。详情见附录 A.1。

最后，我们可以通过将 声音 $s \in \{0, 1, \cdots, K-1\}^{T' \times F'}$的 VQGAN 码本中每个量化向量替换为其最近码本条目的索引，来获得声音的 $\mathbf{z}$ 代码，即，s_{tf} 是方程（2）中所选码本条目的索引。

自回归声音预测。 有了预测代码 s，现在我们可以表示从静音输入视频和条件示例生成代码 s 的似然。我们按照光栅扫描顺序排列 s的索引 [13]，并自回归地预测这些索引：

$$p_\theta(s|\mathbf{v}_q, \mathbf{v}_c, \mathbf{a}_c) = \prod_i p_\theta(s_i|s_{<i}, \mathbf{v}_q, \mathbf{v}_c, \mathbf{a}_c), \qquad (3)$$

其中 $s_{<i}$ 是序列中之前的索引。根据这些概率，我们将 $\mathcal{L}$ (公式（1）) 定义为预测词元 s_i 与真实标记$\hat{}s_i$之间的交叉熵损失函数。

定义声音的基于代码的表示后，我们介绍用于 $\mathcal{F}_\theta$ 条件声音预测的架构（图3）。按照 [13, 29], , 我们使用基于 GPT-2 [45]的仅解码器Transformer [59]来预测代码序列（公式（3））。该 Transformer 的输入是$\mathbf{v}_q$、$\mathbf{v}_c$和$\mathbf{a}_c$的词元化版本。下面介绍如何将这些信号转换为词元。

输入表示。 我们使用 ResNet（2+1）D-18 [53] 表示每个视频信号。为保留细粒度的时间信息，我们移除所有时间步幅，使最终卷积层的时间采样率与输入视频相同。我们沿空间维度进行平均池化，从而为每一帧生成一个嵌入向量。每个这样的向量都成为一个词元。我们将此分词操作记为 T_v（$\mathbf{v}$）。

我们使用矢量量化嵌入来表示条件音频信号。具体而言，我们计算 $\mathbf{z}^{(c)}$ $=q(E(\text{MSTFT}(\mathbf{a}_c))) \in \mathbb{R}^{T' \times F' \times d}$(式(2))，并按光栅扫描顺序提取其 d 维嵌入向量 $\mathbf{z}_1^{(c)}$、$\mathbf{z}_2^{(c)}$，$\cdots$...、$\mathbf{z}_N^{(c)}$。我们将此标记化操作记为 $T_a(\mathbf{a}_c)$。

我们将这些词元合并为一个序列：$\mathcal{S}$ $=$ Concat(T_v ($\mathbf{v}_c$) $T_v(\mathbf{v}_q)$ $T_a(\mathbf{a}_c)$)。因此，我们对$p_\theta(s|\mathbf{v}_q,$ $\mathbf{v}_c,$ $\mathbf{a}_c)$ $=$ $p_\theta(s|\mathcal{S})$。按照标准做法， [13, 57], 我们自回归地生成音频代码，并将先前生成的代码以其矢量量化表示的形式反馈给模型。

生成波形。我们的完整模型 $\mathcal{F}_\theta$ 首先使用 Transformer 生成代码，再通过解码器 D将其转换为梅尔频谱图，然后将

模型	Task							
	材料			动作			起始时刻	
	匹配 准确率（%）	不匹配 准确率（%）	总体 准确率（%）	匹配 准确率（%）	不匹配 准确率（%）	总体 准确率（%）	# 起始时刻 准确率 (%)	起始同步 平均精度 (%)
风格迁移* [18, 55]	30.0	33.5	32.3	20.8	36.6	31.3	19.1	46.9
起始点迁移	**54.8**	**51.4**	**52.6**	**69.0**	**44.7**	**52.9**	**24.8**	**71.9**
随机水平	5.9	5.9	5.9	50.0	50.0	50.0	–	–
SpecVQGAN [29]	25.4	26.8	26.1	52.3	43.1	46.2	11.3	51.0
SpecVQGAN - 微调版 [29]	29.9	25.7	27.2	70.6	58.4	62.5	25.8	59.3
我们的方法 - 无条件	21.3	24.9	23.7	61.4	55.1	57.2	24.6	59.3
我们的方法 - 基线	41.1	41.6	41.4	67.5	59.2	62.0	**26.5**	**60.0**
我们的方法 - 使用重排序	**43.4**	**45.2**	**44.0**	**78.2**	**61.3**	**66.7**	25.3	54.3

表1. **自动评估指标**。我们衡量生成的声音具备条件示例的材料属性、输入示例的动作，以及生成声音相对于原始声音的起始时刻数量和时序的比例。此外，我们还根据条件示例和输入示例的动作及材料是否**匹配**（或**不匹配**），进一步细分自动化指标。起始时刻数量通过判断生成的声音是否与原始声音具有相同数量的起始时刻来衡量。我们计算距真实值 0.1 秒以内的起始时刻预测的平均精度，以评估生成声音起始时刻的时序。 $*$ 表示该模型是一个“预言机”，可以访问输入示例的声音。

梅尔频谱图转换为波形。为了完成这最后一步，我们遵循 [29] 并使用预训练的 MelGAN声码器 [36]。我们发现，与标准的 Griffin-Lim 相比，这种方法生成的结果显著更好 [23]。

基于视听同步的重排序。受其他跨模态生成研究的启发，我们使用重排序来改进模型的预测结果 [47]。我们生成大量声音，然后由一个独立分类器评判并选出最佳声音。通常，这类方法会使用一个模型来判断输入与输出之间的多模态一致性。然而在我们的场景中，这样的分类器理应将条件生成的声音视为与输入视频和条件视频都不匹配。我们转而提出使用视听同步模型 [8, 30, 41]，测量预测声音与输入视频之间的时间对齐情况。这些模型会预测一个时间偏移，使视觉和音频数据达到最佳对齐。

如图 2b 所示，我们使用现成的同步模型 [30] 来估计音频与视频之间的偏移量 t 以及预测的置信度。我们从所有输出中找出绝对偏移量的最小值 $\min|t|$。然后，移除绝对偏移量大于 $\min|t|+\tau$ 的输出，其中 τ 是偏移容差。最后，我们选择置信度最高的声音。

4. 实验

为了评估我们的方法，我们结合使用自动化评估指标和人工感知研究。

4.1. 实验设置

数据集。我们使用视频片段数据集训练条件式拟音生成模型：*Greatest Hits* [42]，该数据集由鼓槌与不同物体互动的视频组成这些物体位于不同场景中；以及*CountixAV* [66]，其中包含来自野外 YouTube 视频的重复动作视频，涵盖多达 23 个不同类别。这些数据集对音频生成颇具挑战，因为它们要求时序精确，并根据细粒度的视觉属性细微调整声音。具体而言，*Greatest Hits* 任务要求理解鼓槌的运动以及物体的材料属性。由于该数据集中的动作和材料易于分析，我们将其用于定量评估；而对于 *CountixAV*，我们提供定性结果，并获准使用类似的野外视频进行评估。有关数据集和实现细节的更多信息，请参阅附录A.2 和附录A.3。

其他模型。我们选取了多种其他模型进行比较。首先，我们使用 **SpecVQGAN** 模型，该模型由 Iashin 和 Rahtu 提出 [29]，，它是一种基于双流视觉网络的先进视觉到声音预测方法。我们使用公开发布的实现和权重。为进行公平比较，我们还在 *Greatest Hits* [42] 数据集上对模型进行了微调，以用于自动化指标评估。

我们还对模型进行了几项消融实验：**无条件示例：**从模型中移除所有条件信息。该模型是一种视觉到声音预测方法，在控制与我们模型之间的架构和数据差异后，类似于 SpecVQGAN [29]。**无条件视频：**该模型获得条件音频 $\mathbf{a}_c$，但没有视频 $\mathbf{v}_c$，因此无法观察条件视频中音频与视觉事件之间的关联。**无增强：**使用未经过音频增强的模型。**随机条件示例：**模型使用与输入视频无关、因而无法提供有效信息的条件视听片段进行训练。我们从数据集中的其他视频里随机选取这些片段。**重新排序的示例：**我们为每个对象生成 100 个输出

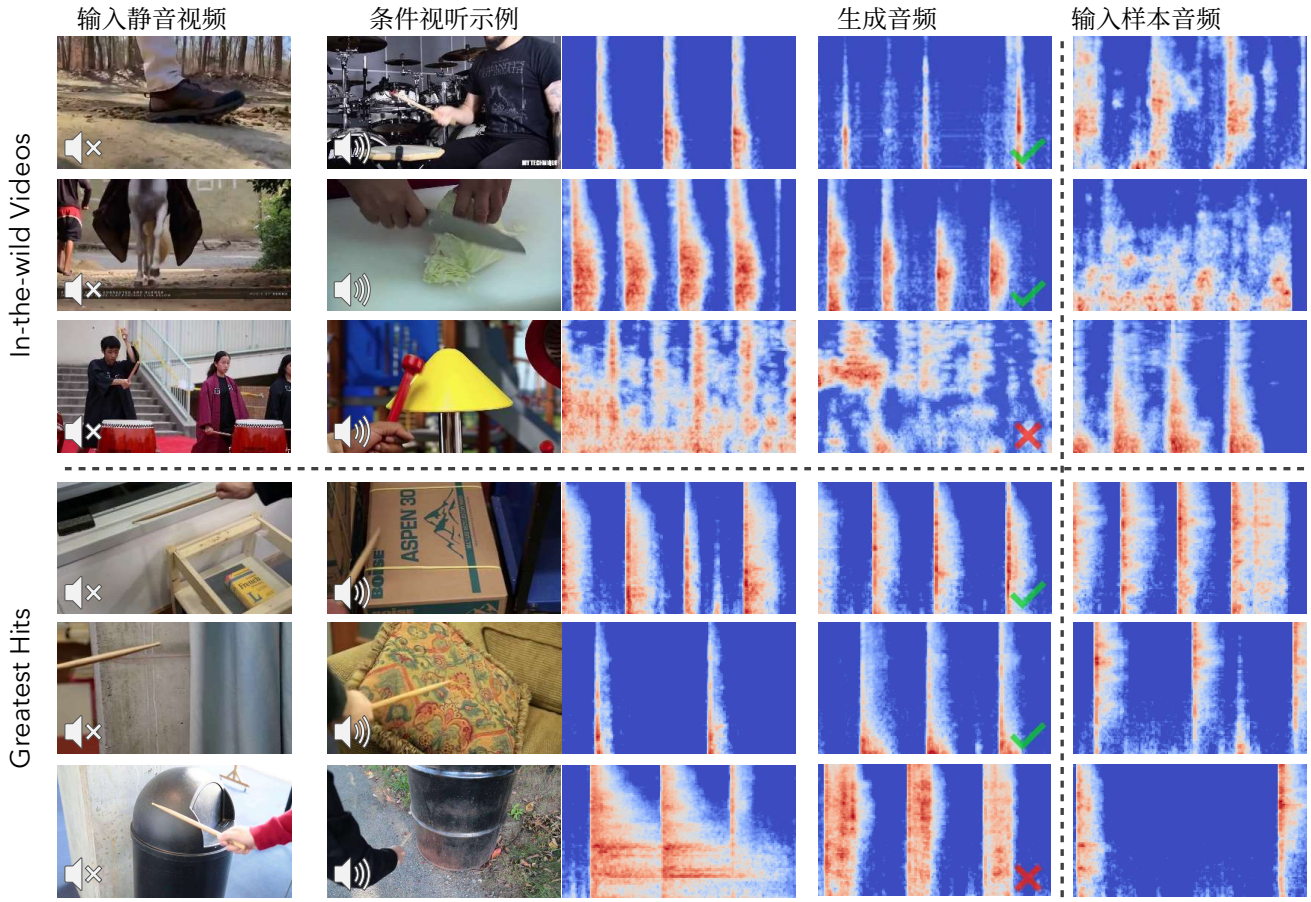


图4。 **定性结果**。我们展示了模型的结果。我们展示了模型在 *CountixAV* 数据集 [66] （第1–3行）和 *Greatest Hits* 数据集 [42] （第4–6行）上的互联网视频结果。第3行和第6行展示了失败案例，生成音频上标有红色叉号。生成音频中的击打时序和次数与输入视频对应的（留出）音频大体一致，这表明生成音频与输入视频中的动作相匹配。生成音频的频率与条件示例大致相符，说明其音色相似。 **如需聆听声音并查看更多示例，请访问我们的 项目网页**。

输入与条件配对，然后对它们重新排序。

为了更好地了解模型的行为，我们将其与两种“非生成式”方法进行比较。首先，我们提出一种名为**起始点迁移**的方法。该方法不生成声音，而是采用手工设计的方法，从条件示例中迁移声音。我们训练了一个 ResNet（2+1）-D [53] 模型，用于检测条件视频和输入视频中的音频起音，然后从条件示例的随机起始时刻提取声音并进行迁移（详情见附录A.4 ）。其次，我们评估了一种音频**风格迁移**方法。我们使用 Ulyanov 的模型 [55], ，该模型将 Gatys等人提出的风格化方法 [18] 应用于频谱图。需要指出的是，该方法需要将音频作为输入，并非专为拟音生成而设计。为此，我们向模型提供输入视频的真实音频$\mathbf{a}_q$，从而赋予它预言机信息。

4.2. 自动化音色评估

一种成功的预测方法应准确传达输入视频中的动作，同时体现条件示例中的材料属性。为评估是否满足这一点，我们使用*Greatest Hits* 数据集 [42]中的标签，训练分类器来识别动作（敲击与刮擦）和材料（一个 17 类分类问题）。随后，我们用该分类器对预测声音、条件声音和输入声音进行分类，并比较估计出的标签。

声音分类器。 我们对预训练的 VGGish 分类器进行微调 [25,26]，使其在 *GreatestHits* [42]数据集上根据梅尔频谱图识别动作或材料。为避免歧义，我们仅使用包含单一材料或动作类型的视频片段。更多详情见附录 A.5。

评估指标。 我们采用了两项评估指标来衡量我们的两项标准：动作准确率，即预测声音中估计动作类别与（留出的）输入声音相同的比例

；以及材质准确率，即预测声音中估计材质类别与条件示例相同的比例。

结果。我们发现（表1），在材质指标上，我们的模型表现显著优于 SpecVQGAN，也优于不使用条件示例的变体；这两者都是无条件的视觉到声音预测方法。正如预期，起始音迁移模型取得了近乎最优的表现，因为它只是从条件声音中迁移声音，而这些声音（显而易见地）很可能具有相同的估计类别。另一方面，这一起始点迁移基线在动作指标上的表现很差，因为它没有机制根据视频中的动作调整迁移的声音（例如，将敲击声转换为刮擦声）。相比之下，我们的模型在这一指标上表现出色。

为了进一步了解性能差异的来源，我们根据条件示例的属性是否与输入声音的属性匹配，对结果进行了细分（表 1）。正如预期，起始时刻迁移方法在材质指标上表现出色，因为“复制粘贴”条件声音是一种简单直接的解决方案。另一方面，当动作类型不匹配时，我们的生成式方法显著优于该方法，因为它们能够调整声音以匹配动作。

此外，我们发现，同步重排序显著提升了材料和动作任务上的性能（表 1）。重新排序后的模型在所有三个动作相关指标上均优于起始时刻迁移方法。在材料相关指标方面，它也缩小了与起始时刻迁移方法之间的差距。这证明了重排序方法的有效性，也体现了将我们的方法设定为生成模型的优势。

4.3. 自动起始音评估

为评估生成音色的质量，我们考察条件式拟音生成模型生成的声音，其起始时刻是否与输入视频中（留出的）声音的起始时刻相匹配。我们评估两项标准：生成的声音是否包含正确数量的起始时刻，以及这些起始时刻的时序是否与输入视频中的相匹配。

评估指标。我们测量视频片段中起始时刻数量与真实音频相同的比例。按照 Owens 等人的做法， [43], 我们报告正确起始时刻检测的平均精度，其中相对波幅用作每个起始时刻的置信度。若检测结果与真实值的偏差在 0.1 秒以内，则视为正确。

结果。我们注意到（表 1），在两项指标上，我们的模型都优于所有生成式基线和风格迁移方法。起始时刻迁移方法在起始音同步任务上表现最佳并不令人意外，因为它是专门针对一个密切相关的任务进行训练的。与起始音迁移模型相比，我们的模型能够成功生成正确数量的起始时刻，这可以归因于条件中提供了有关音视频关系的额外信息，从而帮助模型更好地理解视频中的动作。如表 1所示，移除条件信息后，捕捉正确起始时刻数量的准确率下降至与起始时刻迁移方法相同的水平。有趣的是，我们发现（表 1），重新排序后的模型性能有所下降。这可能是因为模型从其训练所用的 *VGG-Sound*数据集 [4] 转移到了同步网络 [30] 所处的领域，导致它难以推断声音的精确时序（而不是仅仅整体判断两路信号是否同步）。

模型	变化	Task	
		材料 选中(%) ↑	同步 选中(%) ↑
风格迁移* [18, 55] – 起始点迁移	–	9.9 (±2.3) **64.7** (±3.8)	10.6 (±2.4) **57.3** (±3.9)
SpecVQGAN [29]	–	16.3 (±2.9)	18.0 (±3.0)
Ours	base	50.0 (±0.0)	50.0 (±0.0)
	- 条件	35.3 (±3.8)	40.1 (±4.0)
	- 条件视频	46.1 (±4.0)	45.0 (±4.0)
	- 增强	51.3 (±3.9)	49.7 (±4.0)
	使用随机条件	45.5 (±4.0)	47.5 (±4.0)
	+ 重新排序	**54.3** (±3.4)	**53.8** (±3.4)

表2。**感知研究结果**。我们报告了在研究中的两个问题上，参与者选择某一方法的结果而非我们的方法（基线）的比例。作为参考，我们还列出了我们的基线方法在该研究中应获得的比例（50%）。结果以95%置信区间表示。

4.4. 感知研究

我们还通过一项使用 Amazon Mechanical Turk 开展的感知研究对模型进行了评估。我们向参与者提供条件音视频片段，以及两个由不同模型生成声音的输入视频（我们的基础模型加上一个随机选取的替代模型）。

我们要求参与者根据两个与自动化指标相似的标准对生成的声音进行判断。参与者需要选择：1）哪一个结果中的音频与视频中的动作同步得更好；2）哪一个结果中的声音最像条件示例中物体或材料发出的声音。研究共有376名参与者，他们观看了21组视频，这些视频从评估集中随机抽取，其中前5组用于练习，不计入统计。

与我们模型的其他变体进行比较。我们通过感知研究中选择相应方法的情况，评估模型不同变体的影响（表2）。经过重新排序的模型整体表现最佳——

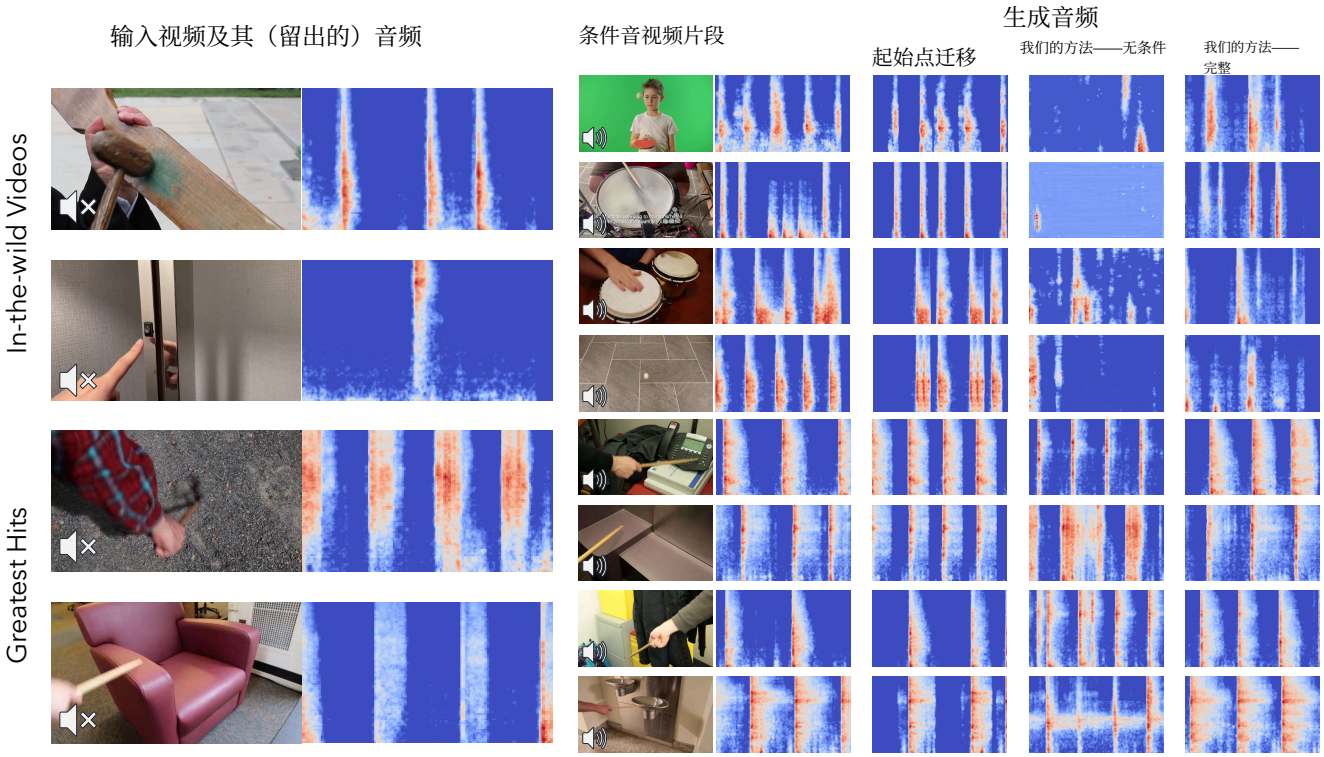


图 5. **模型比较**。我们展示了多个模型的条件 Foley 生成结果，并使用了多种条件声音。我们展示了公开来源的演示视频（第 1–2 行）以及 *Greatest Hits* 数据集 [42]（第 3–4 行）的结果。每个输入视频都与两个不同的条件视频配对。对于不使用条件示例的模型变体，我们提供了 2 个样本。

而未经重排序的基础模型在两项指标上的表现也优于大多数基线方法。不使用条件示例的模型在材质指标上的表现较差，但同步性能的下降幅度相对较小。这是可以理解的，因为该模型没有使用条件示例，但仍被要求与视频保持同步。未使用条件视频的模型在材质指标上略有下降，而同步性能的下降幅度相对较大，这可能是因为它无法观察条件示例中图像与声音之间的关系。

与其他方法的比较。我们将模型与其他方法进行比较。总体而言，不含条件示例的变体（这是一种视觉到声音的方法）优于 SpecVQGAN [29]。风格迁移 [18, 55] 模型表现不佳（从定性结果来看，往往还会包含伪影）。有趣的是，起始音迁移模型在感知研究中的表现相当出色，优于我们的基线模型和经过重排序的模型，尽管它并未针对场景中的动作调整输出（表 1）。这不难理解，因为它会在与撞击同步的选定时刻，从条件示例中“复制并粘贴”声音。因此，用户很可能会发现它与条件声音完全匹配，并且（显而易见地）传达相同的属性。然而，这种情况只会在音频事件在时间上清晰分离时出现；我们预计，当声音不易划分为离散起始音，或起始时刻模糊不清时，该模型会失效。

4.5. 定性结果

我们在图 4中展示了测试集预测声音的可视化结果。通过这些可视化结果可以看出，我们的模型能够成功生成与条件声音频谱特性相似的声音，同时匹配输入视频中（留出的）声音的时序和动作。我们还展示了使用在 *CountixAV*数据集 [66]上训练的模型，在野外视频上应用我们的方法所得的定性结果（图 *4* [42]）。我们采用了与 *GreatestHits* [42]相同的训练和生成方案。模型再次生成了条件频谱特征与输入动作同步相匹配的声音。

在图 5中，我们展示了结果如何随条件声音变化，并在两个数据集上将我们的方法与基线方法进行比较。对于每个输入视频，我们展示了不同模型的预测结果。可以看到，我们的方法会根据条件声音调整输出（例如，根据条件示例是塑料还是金属来改变输出）。我们还看到，起始时刻迁移方法

会在正确的时间“复制并粘贴”条件示例中的声音。我们还观察到，在更真实的场景中，起始点迁移基线表现不佳（见图 5 Video in the wild 部分），因为其中的动作和声音更加丰富、复杂。我们展示了模型在无条件输入下随机生成的两个样本。可以看到，预测结果通常与输入声音相匹配，而非条件声音，并且生成音频的音色存在很大差异。请访问我们的 项目网页 收听输出结果，并查看更多定性结果。

5. 讨论

本文提出了条件 *Foley* 任务，并提出了一种通过自监督学习解决这一问题的方法。我们在*Greatest Hits* 数据集上评估了该方法，通过感知研究和自动化指标发现，我们的模型成功学会了从条件声音中迁移相关信息，同时与静音输入视频中的事件相匹配。我们还展示了该模型在更复杂、更真实的公开来源视频数据上的有效性，该模型是在*CountixAV*数据集上训练的。

我们认为，这项工作有望开辟多个研究方向。我们的工作解决了声音设计流程中的一个环节——调整声音以匹配视频。我们将此视为迈向更广泛目标的一步，即实现半自动化的“用户参与”声音设计。我们也将这项工作视为迈向能够通过类比进行学习的合成方法的一步，这延续了图像类比等经典工作的传统 [27]。我们将在项目网站上发布代码和模型。

局限性与更广泛的影响。 尽管配乐生成可用于电影制作等创意应用，但也可能被用于制作有潜在用途来传播虚假信息的视频，并造成负面后果。图像和音频取证领域有助于减轻这一影响。

致谢。感谢 Jon Gillick、Daniel Geng 和 Chao Feng 提供的宝贵讨论。本研究部分由 DARPA Semafor 和 Cisco Systems 资助，并获得 Adobe 的捐赠。文中表达的观点、意见和/或研究结果仅代表作者，不应被解读为美国国防部或美国政府的官方观点或政策。

参考文献

[1] Vanessa Theme Ament。《拟音圣经：电影、游戏和动画拟音艺术》。Routledge，2014年。1，2[2] Relja Arandjelovic 和 Andrew Zisserman。《会发声的物体》。载于欧洲计算机视觉会议论文集（*ECCV*），第 435–451 页，2018年。2

[3] Changan Chen、Ruohan Gao、Paul Calamia 和 Kristen Grauman。视觉声学匹配。arXiv 预印本 arXiv:2202.06875，2022年。2[4] Honglie Chen、Weidi Xie、Andrea Vedaldi 和 Andrew Zisserman。Vggsound：一个大规模视听数据集。载于《ICASSP 2020-2020 IEEE国际声学、语音与信号处理会议（ICASSP）》，第721–725页。IEEE，2020年。7[5] Lele Chen、Guofeng Cui、Celong Liu、Zhong Li、Ziyi Kou、Yi Xu 和 Chenliang Xu。基于节奏性头部运动的说话人头部生成。载于《欧洲计算机视觉会议》，第35–51页。Springer，2020年。3[6] Ziyang Chen、Xixi Hu 和 Andrew Owens。从静默中提取结构：从环境声中学习场景结构。载于第五届机器人学习年会，2021年。2[7] Joon Son Chung、Bong-Jin Lee 和 Icksang Han。是谁说的？真实会议中的视听说话人分离。arXiv 预印本 arXiv:1906.10042，2019年。3[8] Joon Son Chung 和 Andrew Zisserman。时过境迁：自然场景中的自动唇形同步。载于《亚洲计算机视觉会议》，第251–263页。Springer，2016年。2，5[9] Chenye Cui、Yi Ren、Jinglin Liu、Rongjie Huang 和 Zhou Zhao。Varietysound：通过无监督信息解耦实现音色可控的视频到声音生成。arXiv 预印本 arXiv:2211.10666，2022年。2[10] Abe Davis 和 Maneesh Agrawala。视觉节奏与节拍。载于《IEEE计算机视觉与模式识别会议研讨会论文集》，第2532–2535页，2018年。2[11] Ariel Ephrat、Inbar Mosseri、Oran Lang、Tali Dekel、Kevin Wilson、Avinatan Hassidim、William T Freeman 和 Michael Rubinstein。鸡尾酒会场景中的聆听：一种与说话人无关的视听语音分离模型。SIGGRAPH，2018年。2[12] Ariel Ephrat 和 Shmuel Peleg。Vid2speech：从静音视频重建语音。载于《2017年IEEE国际声学、语音与信号处理会议（ICASSP）》，第5095–5099页。IEEE，2017年。2[13] Patrick Esser、Robin Rombach 和 Bjorn Ommer。驯化用于高分辨率图像合成的变换器。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第12873–12883页，2021年。2，3，4，12[14] Chuang Gan、Deng Huang、Peihao Chen、Joshua B Tenenbaum 和 Antonio Torralba。拟音音乐：学习从视频生成音乐。载于《欧洲计算机视觉会议》，第758–775页。Springer，2020年。2[15] Ruohan Gao、Rogerio Feris 和 Kristen Grauman。通过观看未标注视频学习分离物体声音。载于《欧洲计算机视觉会议论文集（ECCV）》，第35–53页，2018年。2[16] Ruohan Gao 和 Kristen Grauman。2.5D视觉声音。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第324–333页，2019年。2[17] Rishabh Garg、Ruohan Gao 和 Kristen Grauman。面向视频双耳音频生成的几何感知多任务学习。arXiv 预印本 arXiv:2111.10882，2021年。2

[18] Leon A Gatys、Alexander S Ecker 和 Matthias Bethge。艺术风格的神经算法。arXiv 预印本 arXiv:1508.06576，2015年。5，6，7，8，13[19] Leon A Gatys、Alexander S Ecker 和 Matthias Bethge。使用卷积神经网络进行图像风格迁移。载于《IEEE计算机视觉与模式识别会议论文集》，第2414–2423页，2016年。2[20] Sanchita Ghose 和 John Jeffrey Prevost。Autofoley：利用深度学习为无声视频合成同步音轨。IEEE多媒体汇刊，23：1895–1907，2020年。2[21]Sanchita Ghose 和 John J Prevost。Foleygan：在无声视频中实现视觉引导的生成对抗网络式同步声音生成。arXiv 预印本 arXiv:2107.09262，2021年。2[22] Ian Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、Bing Xu、David Warde-Farley、Sherjil Ozair、Aaron Courville 和 Yoshua Bengio。生成对抗网络。载于《神经信息处理系统进展》，第2672–2680页，2014年。2 – .[23] Daniel Griffin 和 Jae Lim。基于改进短时傅里叶变换的信号估计。IEEE声学、语音与信号处理汇刊，32(2)：236–243，1984年。5[24] John Hershey 和 Michael Casey。基于隐马尔可夫模型的视听声音分离。《神经信息处理系统进展》，14，2001年。2[25] Shawn Hershey、Sourish Chaudhuri、Daniel PW Ellis、Jort F Gemmeke、Aren Jansen、R Channing Moore、Manoj Plakal、Devin Platt、Rif A Saurous、Bryan Seybold 等。用于大规模音频分类的卷积神经网络架构。arXiv 预印本 arXiv:1609.09430，2016年。6，13[26] Shawn Hershey、Sourish Chaudhuri、Daniel P. W. Ellis、Jort F. Gemmeke、Aren Jansen、Channing Moore、Manoj Plakal、Devin Platt、Rif A. Saurous、Bryan Seybold、Malcolm Slaney、Ron Weiss 和 Kevin Wilson。用于大规模音频分类的卷积神经网络架构。载于《国际声学、语音与信号处理会议（ICASSP）》，2017年。6，13[27] Aaron Hertzmann、Charles E Jacobs、Nuria Oliver、Brian Curless 和 David H Salesin。图像类比。载于《第28届计算机图形学与交互技术年会论文集》，第327–340页，2001年。1，2，9[28] Sicong Huang、Qiyang Li、Cem Anil、Xuchan Bao、Sageev Oore 和 Roger B Grosse。Timbretron：用于音乐音色迁移的 WaveNet（CycleGAN（CQT（音频））流程。arXiv 预印本 arXiv:1811.09620，2018年。2[29] Vladimir Iashin 和 Esa Rahtu。驯化视觉引导的声音生成。arXiv 预印本 arXiv:2110.08791，2021 年。1，2，3，4，5，7，8，12，13，14[30] Vladimir Iashin、Weidi Xie、Esa Rahtu 和 Andrew Zisserman。时空稀疏：利用可训练选择器实现视听同步。arXiv 预印本 arXiv:2210 07055，2022年。2，5，7 . .[31] Phillip Isola、Jun-Yan Zhu、Tinghui Zhou 和 Alexei A Efros。通过条件生成对抗网络进行图像到图像转换。载于《IEEE计算机视觉与模式识别会议论文集》，第1125–1134页，2017年。4，12

[32] Justin Johnson、Alexandre Alahi 和李飞飞。《用于实时风格迁移和超分辨率的感知损失》，2016年。4, 12[33] Diederik Kingma 和 Jimmy Ba。《Adam：一种随机优化方法》。2015年国际学习表征会议。12, 13[34] Diederik P Kingma 和 Max Welling。《自编码变分贝叶斯》。arXiv预印本 arXiv:1312.6114，2013年。12[35] A Sophia Koepke、Olivia Wiles、Yael Moses 和 Andrew Zisserman。《视觉到声音：一种端到端的视觉钢琴转录方法》。载于 ICASSP 2020—2020 IEEE声学、语音与信号处理国际会议（ICASSP），第1838–1842页。IEEE，2020年。2[36] Kundan Kumar、Rithesh Kumar、Thibault de Boissiere、Lucas Gestin、Wei Zhen Teoh、Jose Sotelo、Alexandre de Br´ebisson、Yoshua Bengio 和 Aaron C Courville。《Melgan：用于条件波形合成的生成对抗网络》。神经信息处理系统进展，第32卷，2019年。4, 5[37] Timothy R Langlois 和 Doug L James。《逆向拟音动画：将刚体运动与声音同步》。ACM图形学汇刊（TOG），33(4):41，2014年。2[38] Seung Hyun Lee、Wonseok Roh、Wonmin Byeon、Sang Ho Yoon、Chan Young Kim、Jinkyu Kim 和 Sangpil Kim。《声音引导的语义图像操控》。arXiv预印本 arXiv:2112.00007，2021年。2[39] Tingle Li、Yichen Liu、Andrew Owens 和 Hang Zhao。《从视听关联中学习视觉风格》。arXiv，2022年。2[40] Javier Nistal、Stefan Lattner 和 Gael Richard。《Drumgan：利用生成对抗网络，通过音色特征条件控制合成鼓声》。arXiv预印本 arXiv:2008.12073，2020年。2[41] Andrew Owens 和 Alexei A Efros。《利用自监督多感官特征进行视听场景分析》。欧洲计算机视觉会议（ECCV），2018年。2, 5[42] Andrew Owens、Phillip Isola、Josh McDermott、Antonio Torralba、Edward H Adelson 和 William T Freeman。《视觉指示的声音》。CVPR，2016年。1, 2, 5, 6, 8, 12, 13, 14[43] Andrew Owens、Phillip Isola、Josh McDermott、Antonio Torralba、Edward H Adelson 和 William T Freeman。《视觉指示的声音》。载于计算机视觉与模式识别（CVPR），2016年。7[44] KR Prajwal、Rudrabha Mukhopadhyay、Vinay P Namboodiri 和 CV Jawahar。《学习个人说话风格，以实现准确的唇动到语音合成》。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第13796–13805页，2020年。2[45] Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei、Ilya Sutskever 等。《语言模型是无监督的多任务学习者》，2019年。4[46] Colin Raffel、Brian McFee、Eric J Humphrey、Justin Salamon、Oriol Nieto、Dawen Liang、Daniel PW Ellis 和 C Colin Raffel。《MIR Eval：常用音乐信息检索指标的透明实现》。载于 ISMIR，第367–372页，2014年。14[47] Aditya Ramesh、Mikhail Pavlov、Gabriel Goh、Scott Gray、Chelsea Voss、Alec Radford、Mark Chen 和 Ilya Sutskever。

《零样本文本到图像生成》。载于国际机器学习会议，第8821–8831页。PMLR，2021年。5[48] Tim Sainburg、Marvin Thielk 和 Timothy Q Gentner。《发现、可视化并量化多样动物发声 repertoire 中的潜在结构》。PLoS计算生物学，16(10):e1008228，2020年。12[49] Eli Shechtman 和 Michal Irani。《在图像和视频中匹配局部自相似性》。载于2007年IEEE计算机视觉与模式识别会议，第1–8页。IEEE，2007年。3[50]Kun Su、Xiulong Liu 和 Eli Shlizerman。《多乐器演奏者网络：从身体动作中无监督生成音乐》。arXiv预印本 arXiv:

2012 03478, 2020 2。[51]

Kun Su、Xiulong Liu 和 Eli Shlizerman。《听起来如何？》。神经信息处理系统进展，第34卷，2021年。2[52] D´ıdac Sur´ıs、Carl Vondrick、Bryan Russell 和 Justin Salamon。《是时候探讨音乐与视频中的艺术对应关系了》。CVPR，2022年。2[53] Du Tran、Heng Wang、Lorenzo Torresani、Jamie Ray、Yann LeCun 和 Manohar Paluri。《深入探究用于动作识别的时空卷积》。载于2018年IEEE/CVF计算机视觉与模式识别会议，第6450–6459页，2018年。4, 6, 12[54] Efthymios Tzinis、Scott Wisdom、Aren Jansen、Shawn Hershey、Tal Remez、Daniel PW Ellis 和 John R Hershey。《走进真实世界：利用 Audioscope 无监督分离屏幕内声音》。arXiv预印本 arXiv:2011 01143，2020 2。[55] Dmitry Ulyanov。《音频纹理合成与风格迁移》。https : / / dmitryulyanov . github . io / audio - texture - synthesis - and - style - transfer，2016年。2, 5, 6, 7, 8, 13[56] Kees Van Den Doel、Paul G Kry 和 Dinesh K Pai。《Foleyautomatic：用于交互式模拟和动画的基于物理的音效》。载于第28届计算机图形学与交互技术年度会议论文集，第537–544页。ACM，2001年。2[57] Aaron Van Den Oord、Oriol Vinyals 等。《神经离散表示学习》。神经信息处理系统进展，第30卷，2017年。2, 3, 4, 12[58] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。《注意力就是一切》。载于神经信息处理系统进展，第5998–6008页，2017年。2[59] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Lukasz Kaiser 和 Illia Polosukhin。《注意力就是一切》。载于神经信息处理系统（NIPS），2017年。4[60] Prateek Verma 和 Julius O S mith。《音频频谱图的神经风格迁移》。arXiv预印本 arXiv:1801.01589，2018年。2[61] Yu Wang、Nicholas J Bryan、Justin Salamon、Mark Cartwright 和 Juan Pablo Bello。《谁说了算？重新思考音频的少样本学习》。载于2021年IEEE音频与声学信号处理应用研讨会（WASPAA），第36–40页。IEEE，2021年。3

[62] Yu Wang、Justin Salamon、Nicholas J Bryan 和 Juan Pablo Bello。少样本声音事件检测。载于《ICASSP 2020—2020 IEEE国际声学、语音与信号处理会议》（ICASSP），第81–85页。IEEE，2020年。3[63] Xudong Xu、Hang Zhou、Ziwei Liu、Bo Dai、Xiaogang Wang 和 Dahua Lin。无需双耳音频的视觉引导的双耳音频生成。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第15485–15494页，2021年。2[64] Karren Yang、Bryan Russell 和 Justin Salamon。辨别左右：学习视觉与声音的空间对应关系。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第9932–9941页，2020年。2–。[65] Richard Zhang、Jun-Yan Zhu、Phillip Isola、Xinyang Geng、Angela S Lin、Tianhe Yu 和 Alexei A Efros。利用学习到的深度先验实现实时用户引导的图像着色。arXiv预印本，arXiv:1705.02999，2017年。2[66] Yunhua Zhang、Ling Shao 和 Cees GM Snoek。通过视觉与声音进行重复活动计数。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第14070–14079页，2021年。2、3、5、6、8、12[67] 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）论文集，第570–586页，2018年。2[68] Yipin Zhou、Zhaowen Wang、Chen Fang、Trung Bui 和 Tamara L Berg。视觉转声音：为野外视频生成自然声音。载于《IEEE计算机视觉与模式识别会议论文集》，第3550–3558页，2018年。1、2–。[69] Jun-Yan Zhu、Taesung Park、Phillip Isola 和 Alexei A Efros。使用循环一致对抗网络进行非配对图像到图像转换。2017年。2

A.1. VQGAN 码本损失

我们遵循 Iashin 和 Rahtu 的公式 [29]，该公式将 VQGAN损失 [13, 57] 应用于频谱图。该损失 $\mathcal{L}$ 由三个独立部分组成：重建损失和码本损失 [34, 57] $\mathcal{L}_{VQVAE}$、感知损失 [32] $\mathcal{L}_{perceptual}$以及基于补丁的判别器损失 [31] $\mathcal{L}_{disc}$。我们给出了用于学习码本的最终损失：

$$\mathcal{L}_{VQGAN} = \mathcal{L}_{VQVAE} + \mathcal{L}_{perceptual} + \mathcal{L}_{disc} \quad (4)$$

接下来我们将介绍这些损失函数。

重建损失和码本损失。 给定输入波形$\mathbf{a}$ 及其梅尔频谱图$\mathbf{S} = \text{MSTFT}(\mathbf{a}) \in \mathbb{R}^{T \times F}$，其中 T和 F分别表示时间维度和频率维度，编码器 E 会将$\mathbf{S}$ 编码为嵌入 $\hat{}\mathbf{z}^{tf}$ 及其对应的量化码 $q(\hat{}\mathbf{z}^{tf}) = \mathbf{z}^{tf}$。随后，解码器 D 会将$\mathbf{z}^{tf}$ 重建为梅尔频谱图 $\hat{}\mathbf{S} = D(q(E(\text{MSTFT}(\mathbf{a}))))$。按照 VQVAE 的做法，我们最小化$\mathbf{S}$ 与 $\mathbf{S}$ 之间的重建损失，以及 $\hat{}\mathbf{z}^{tf}$ 与$\mathbf{z}^{tf}$之间的码本损失：

$$\mathcal{L}_{VQVAE} = \underbrace{||\mathbf{S} - \hat{\mathbf{S}}||}_{\text{recon. loss}} + \underbrace{||\hat{\mathbf{z}}_{tf} - \text{sg}[\mathbf{z}_{tf}]||_2^2 + ||\text{sg}[\hat{\mathbf{z}}_{tf}] - \mathbf{z}_{tf}||_2^2}_{\text{codebook loss}}. \quad (5)$$

其中，sg 表示停止梯度操作。

感知损失。 我们使用现成的类 VGGish 模型 [29]，（该模型是以 VGG-16 为骨干网络、在 VGGSound 数据集上训练的 VGGish 模型变体）来评估感知损失。对于原始梅尔频谱图和重建梅尔频谱图的第 i^{th} 层特征，$\mathbf{S}^i$ 和 $\hat{}\mathbf{S}^i$，相应的感知损失定义为二者的平方 L2 距离：

$$\mathcal{L}_{perceptual} = \sum_{i=1}^{N} \frac{1}{F^i T^i} ||\mathbf{S}^i - \hat{\mathbf{S}}^i||_2^2, \quad (6)$$

其中，N是我们选用来计算感知损失的层数。我们的模型使用了 $N = 5$层，选择方式如 [29]所示。

基于图块的判别器损失。 我们使用 Isola等人提出的判别器损失。 [31]:

$$\mathcal{L}_{disc} = \log \mathcal{D}(\mathbf{S}) + \log(1 - \mathcal{D}(\hat{\mathbf{S}})), \quad (7)$$

其中，$\mathcal{D}$ 应用于多个尺度的全卷积网络，并用作判别器。

A.2. 数据集

所有数据均来自*Greatest Hits* 数据集（由 Owens等人在实验室采集的数据集） [42], 以及*CountixAV* [66] 数据集。*Greatest Hits*数据集由 977 段视频（11 小时）组成，内容是鼓槌在不同场景中与各种物体互动。*CountixAV* [66]数据集由 1483 段视频（4.1 小时）组成，内容是在真实场景中重复进行的动作，包括弹跳球、跳绳及其他动作。

为了评估我们的条件 Foley 生成任务，我们从 *GreatestHits* [42] 数据集中随机抽取时长为 2 秒的测试视频片段，作为输入视频。我们将每个静音输入视频与来自不同测试视频的 3 个条件音视频片段随机配对。由此得到条件 Foley 评估集，共包含 582 个输入—条件配对，涵盖 17 种不同材料和两种动作类型。

我们还在 *CountixAV* [66] 数据集上评估模型，随机从视频中选取时长为 2 秒的片段。我们采用降噪算法 [48]，在训练和评估前提升声音质量。考虑到缺少 *CountixAV* [66] 数据集中视频的使用许可，我们在获得适当授权的公开视频上定性展示结果。附录A.11中列出了这些视频的出处。

A.3. 实现细节

为了训练完整模型，我们先训练 VQGAN，再使用其学习到的代码训练音频预测器。我们使用批量大小 32、学习率为 1.44×10^{-4}，并采用 Adam [33]，训练 VQGAN 约 400 个训练轮次。我们使用批量大小 8、学习率为 1.6×10^{-4}，并使用 4 块 NVIDIA A40 图形处理器训练 Transformer 50 个训练轮次。VQGAN 的训练约需 5 天，Transformer 的训练约需 20 小时。

我们的模型对条件视频和输入视频均使用时长为 2 秒的视频片段，视频采样率为 15 Hz，音频采样率为 22.05 kHz。VQGAN 码本编码器从梅尔频谱图中生成 12×5 时频网格；该梅尔频谱图由波形转换而来，使用了 80 个梅尔频带和 1,024 个傅里叶滤波器。我们将 $d = 256$ 用作码本嵌入维度。随机裁剪 2 秒视频（30 帧）并将其调整为 112×112，然后将其表示为 30 个 1024 维特征向量，这些向量通过对 ResNet 特征图执行 1×1 卷积获得。在训练期间，我们在提取片段之前，对频谱图应用频率掩蔽和时间掩蔽形式的数据增强。模型以相同方式在 *Greatest Hits* [42] 数据集和 *CountixAV* [66]数据集上进行训练。

推理时，我们将重排序容差设为$\tau = 0.2$。

A.4. 起始点迁移基线方法训练细节

为了根据视频预测起始时刻，我们使用与我们的主模型相同变体的 ResNet（2+1）D-18 [53] 视觉网络

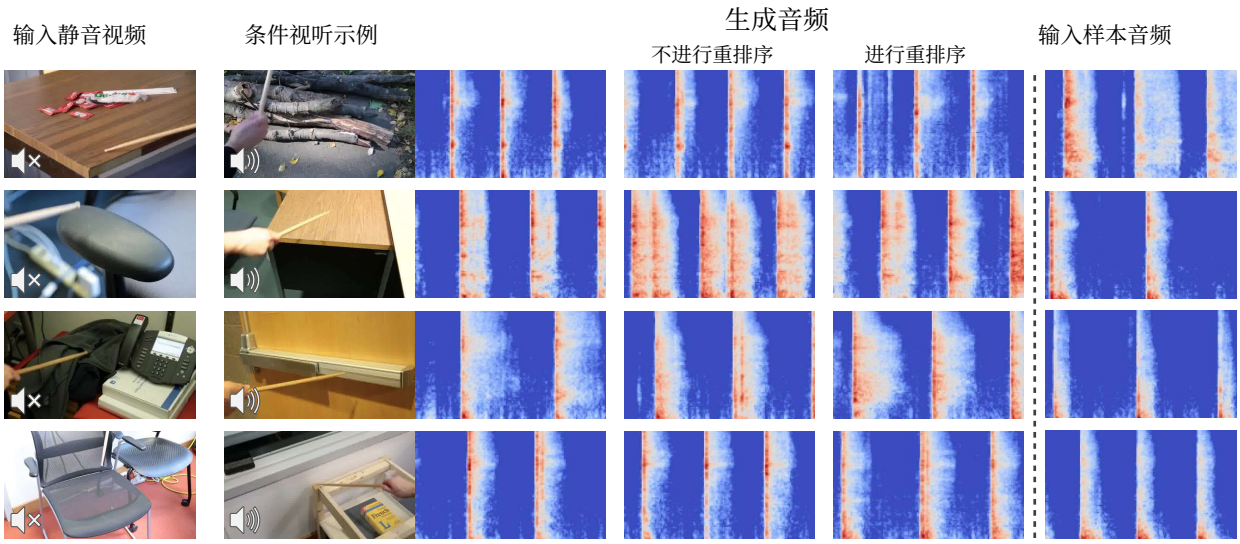


图 6. **重排序后的比较结果**。我们展示了 4 对结果，以定性比较重排序的效果。

（即移除时间步幅后的模型）。我们的模型在池化后使用一个全连接层输出一个预测向量（每个输入帧对应一个起始时刻预测）。我们根据*GreatestHits* [42]数据集提供的时间标签，将其对齐到最近的帧，以此获得真实起始时刻。我们使用二元交叉熵损失，对发生在错误时刻的起始时刻预测进行惩罚。由于每个输入视频可能包含多个声音事件，因此我们根据输入视频中的起始时刻总数，平均分配损失中各起始时刻的权重，使权重之和为一。视频和音频的配置与我们的模型相同：帧率为 15 Hz，音频采样率为 22.05 kHz。我们训练模型 100 个训练轮次，批量大小为 12 个、时长为 2 秒的视频片段，学习率为 1×10^{-4}，并使用 Adam [33]进行优化。模型在单个 NVIDIA RTX-2080 GPU 上训练。

该模型用于检测输入视频和条件视频中的起始时刻。随后，我们随机将条件视频中的声音复制并粘贴到起始时刻。我们加入起始点迁移方法，以帮助分析 Greatest Hits 上的生成模型 [40]。该方法只需检测起始时刻，并从条件示例中复制粘贴声音。按照设计，当只有一种材料时，它能在材料指标上轻松取得近乎完美的表现，在起始时刻指标上也表现良好，因为它直接使用起始时刻标签进行训练。我们强调，这种基线方法并非生成式方法；事实上，我们展示了当输入与条件配对中的动作不同时，该方法会失效（表 1 第 6 列）。从定性结果来看，我们发现该方法在 CountixAV 上经常彻底失效 [60]，因为那里没有清晰的起始时刻。起始时刻的数量往往不准确，而且该方法会受到背景噪声的影响而失效（例如，图5 中第 1 行）。

模型	窗口大小		
	0.1 秒 平均精度 (%)	0.2 秒 平均精度(%)	Avg. 平均精度(%)
风格迁移* [18, 55]	46.9	46.7	58.3
起始点迁移	**71.9**	**78.4**	**76.5**
SpecVQGAN [29]	59.3	65.5	64.1
我们的方法 - 无条件	59.3	73.2	69.8
我们的方法 - 基线	**60.0**	**74.0**	**70.0**

表3. **起始点同步窗口大小评估**。我们测量起始时刻预测在不同窗口大小（0.1 秒和 0.2 秒）内的平均精度。我们还测量了不同窗口大小（从 0.10 秒到 0.25 秒，步长为 0.05 秒）下的平均精度分数。

A.5. 定量实验的声音分类器训练细节

我们对两个预训练的 VGGish 分类器进行微调 [25, 26]，根据 *Greatest Hits* [42] 数据集提供的标签，预测视频中呈现的动作或材料。在相同的输入视频设置下，我们使用 Adam 训练这两个模型，并采用提前停止准则 [33]。模型在单个 NVIDIA RTX-2080 GPU 上训练，学习率为 10^{-4}，批量大小为 32 个时长为 2 秒的视频片段。训练后的模型在材料任务上的验证准确率为 75。6%，在动作预测任务上的准确率为92。0%。

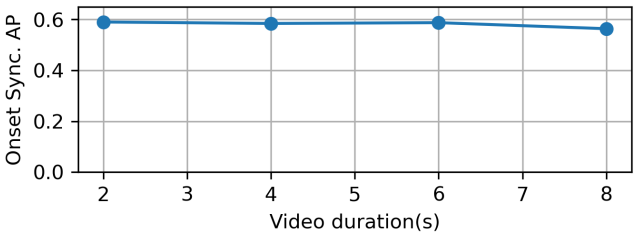


图 7. **不同持续时间下的自动起音同步评估**。我们评估了起始点同步在不同持续时间视频上的表现。

A.6. 不同窗口大小下的起音检测实验

在起音检测实验中，我们选择 0.1 秒作为检测窗口大小，这一数值遵循 MIR Eval MIR Eval [46]以及 *GreatestHits* [42]中音频起音检测所采用的标准值。除此之外，我们还尝试了使用其他窗口大小的不同方法。我们发现（表3），在这项补充实验中，我们的方法优于所有其他生成式基线方法。同时，当窗口大小增大或对不同窗口大小的结果取平均时，我们的方法与起始时刻迁移基线方法之间的差距会缩小。

A.7. 生成较长音频

遵循 Iashin 等人的做法 [29], ，我们使用 2 秒的滑动窗口生成较长的音频片段。在此过程中，我们始终将待生成的词元置于所有可见词元的中心，然后再将这些词元输入变换器。我们在 *Greatest Hits* 数据集 [40] 中的 454 对视频上评估了起始同步性能。由图 7 可见，对于时长最长达 6 秒的视频，我们的模型仍能取得出色表现。我们在项目网页上提供了生成 4 秒、6 秒和 8 秒音频的示例。

A.8. 定性结果的重排序

从图 6的定性结果中，我们注意到重排序提高了同步性能。请注意，经过重排序生成的声音在图 6的第 2 行和第 3 行中表现出明显更好的同步性能。

A.9. 人类研究详情

在人类研究中，我们总共从亚马逊土耳其机器人招募了 609 名参与者。根据参与者对哨兵示例的回答，我们筛除了 233 条反馈。在该示例中，其中一个视频配上了明显错误的声音，既与视频中的动作不匹配，音色也与选中的条件视频不相近。最终，我们获得了 376 条有效反馈。

我们向他们提供了以下说明：

- 测试时请使用耳机。您可能会听到一些刺耳或不协和声音，因此请务必在测试前调节好设备音量。
- 该任务大约需要 15 分钟完成。
- 您将参与一项涉及视觉与听觉感知的实验。要完成此任务，您需要观看并聆听 21 组视频，并为每组视频回答两个问题。每组包含以下视频（所有视频均带有音频）：
  - 一个输入视频：视频 #1– 两个输出视频：视频 #2 和视频 #3
- 您的任务是回答底部的两个问题：
  - 一个输入视频：视频 #1。输出视频（#2 或 #3）中，哪个视频的音频与视频中的动作最同步？ – 根据输出视频中的动作，哪个输出视频（#2 或 #3）的音频听起来最符合视频 #1 中的物体或材料？
- 在开始主要任务之前，你将完成一项简短的练习，内容为 5 组视频（约 3 分钟），以便熟悉界面。

我们还提供了参与者在测试过程中将看到的说明页面和主测试页面的屏幕截图，见图 8 和图9。

A.10. 随机选取的结果

我们在图 10 中提供了随机选取的结果（见本补充材料）。另请观看所提供的视频。值得注意的是，在大多数情况下，生成音频的音色与条件视频非常接近。同时，模型通常也能生成与输入视频中动作时序相匹配的声音。

A.11. 视频鸣谢

我们已获得许可，可以使用和编辑*Greatest Hits*数据集中的视频，也自行录制了其中一些视频。此处列出论文及补充材料中使用的公开来源且已获授权的视频的鸣谢信息。

1. fotorealis- https://stock.adobe.com/video/bambino- suona - bongo- percussioni- per - musicoterapia/536371955- Adobe Stock扩展许可。2. FreeStockFootageClub- https://www.youtube.com/watch?v=zI1_fAtpHQc - YouTube 知识共享署名许可。3. Ignat Gorazd - https://www.youtube.com/watch?v=1_ckbCU5aQs - YouTube 知识共享署名许可。

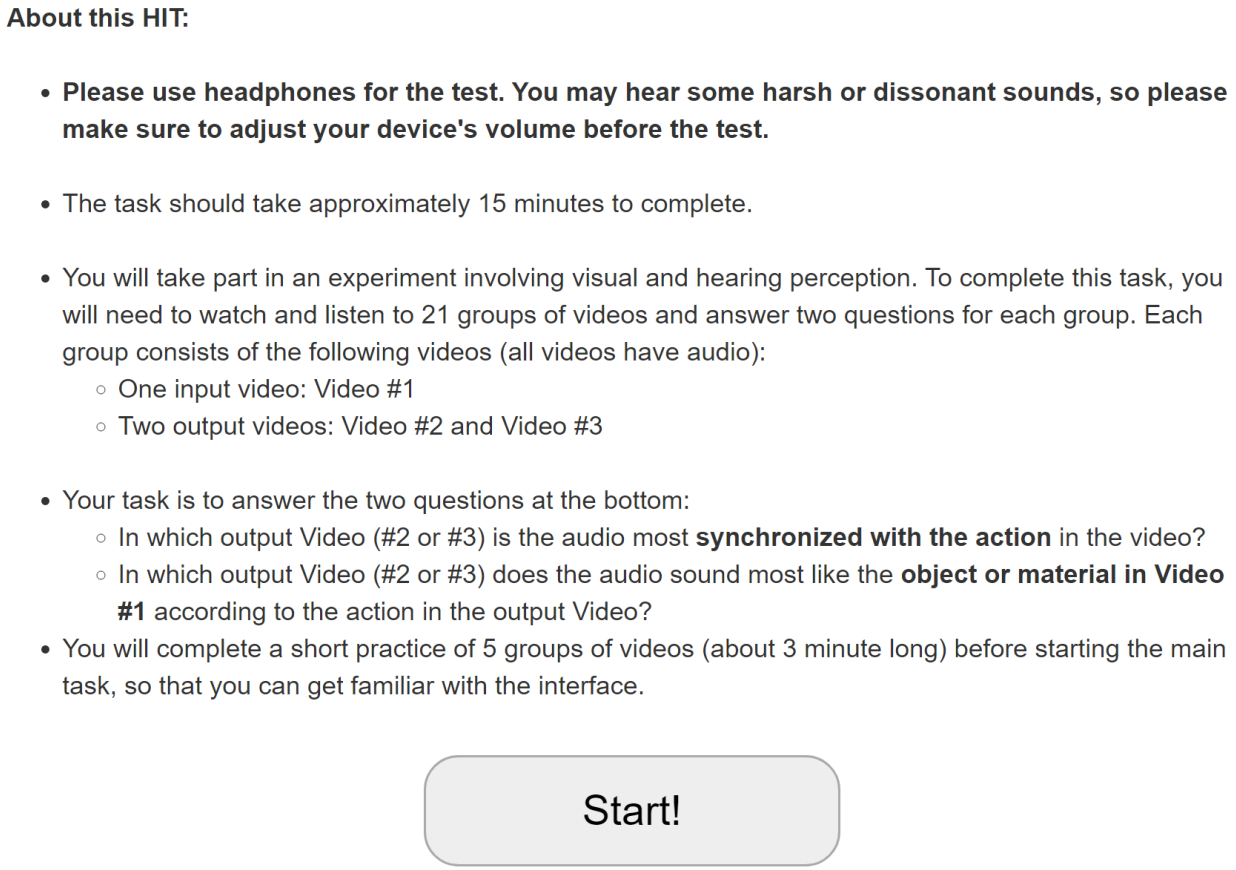


图8。**AMT测试说明页面**。我们展示了参与者在测试开始时将看到的说明页面屏幕截图。图像中的敏感信息已被移除。

4. kriista -https://www.youtube.com/watch?v=6d1YS7fdBK4 - YouTube 知识共享署名许可。5. Over & Out -https://www.youtube.com/watch?v=SExIpBIBj_k - YouTube 知识共享署名许可。6. Over & Out -https://www.youtube.com/watch?v=XxmZxM8AtUc - YouTube 知识共享署名许可。7. Percussion Play - https://www.youtube.com/watch?v=fcjfKvdkJyI - YouTube 知识共享署名许可。8. Percussion Play - https://www.youtube.com/watch?v=xcUyiXt0gjo - YouTube 知识共享署名许可。9. PhotoSerg - https://stock.adobe.com/video/孩子-抛接-乒乓球-绿幕/99378579 - Adobe Stock扩展许可。10. PUSAT E-PEMBELAJARAN UMS - https://www.youtube.com/watch?v=S6TkbV4B4QI - YouTube 知识共享署名许可。11. Suliman Razvan - https://stock.adobe.com/video/Christian-monk-用-一把小木槌-敲击-一块大木头-toaca-以召唤-兄弟们-来祈祷/93378558- Adobe Stock 扩展许可。12. Thomas Cremier - https://www.youtube.com/watch?v=GFmuVBiwz6k - YouTube 知识共享署名许可。

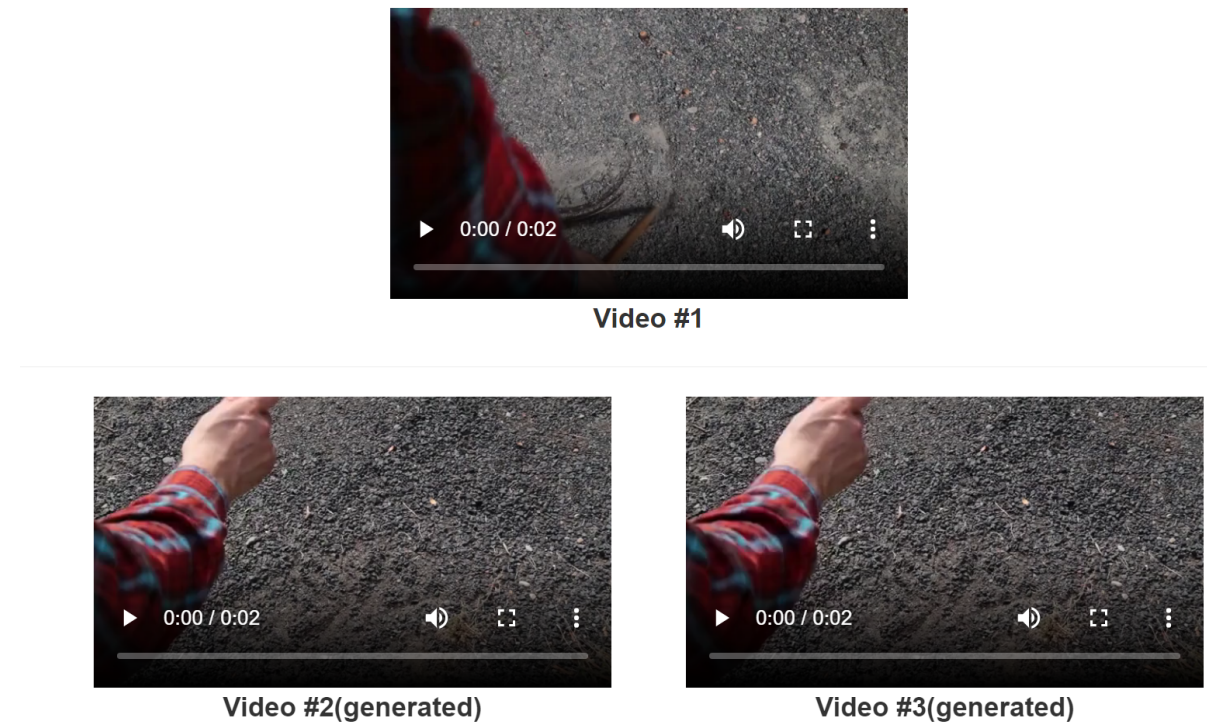


图9。**AMT测试的测试页面**。我们展示了参与者在测试期间会看到的主测试页面屏幕截图。参与者需要回答两个问题，然后才能进入下一组视频。点击“提交”按钮后，参与者将进入下一题。

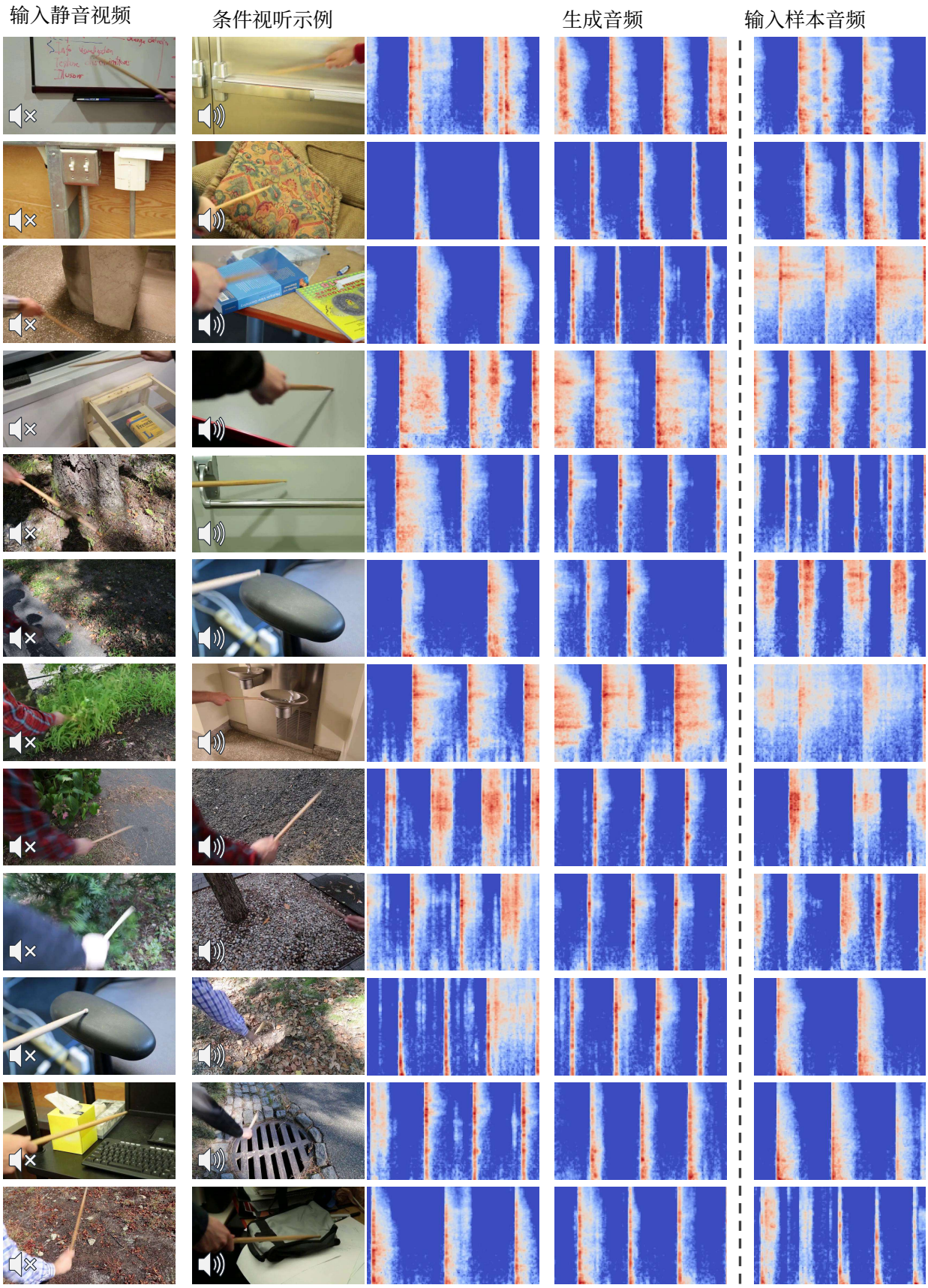


图10。**随机选取的结果**。我们展示了模型生成的12个结果。请参阅视频以聆听这些结果。17