arXiv:2304.05600v2 [cs.SD] 8 Jun 2024

看起来相似，听起来不同：利用反事实跨模态配对进行视听表征学习

Nikhil Singh[1*] Chih-Wei Wu[2] Iroro Orife[2] MahdiKalayeh[2]1麻省理工学院 [2]Netflix[1]nsingh1@mit.edu [2] [chihweiw, iorife, mkalayeh]@netflix.com

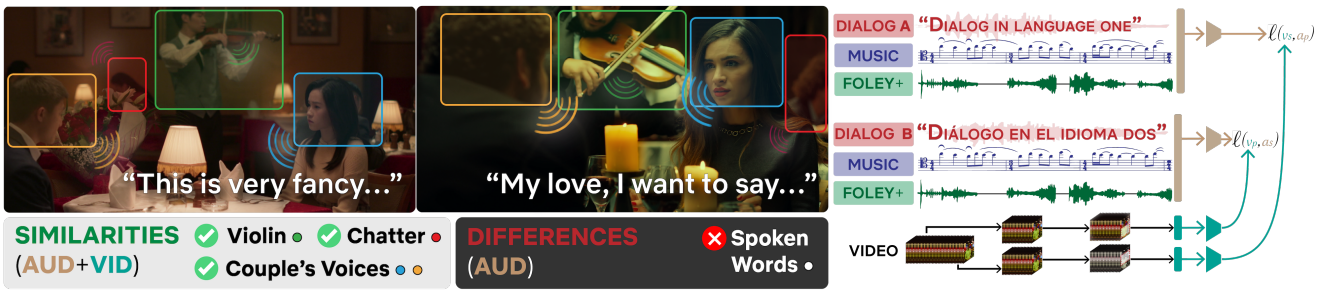

图 1.（左）视听场景在感知上可能十分相似，即使其中说出的词语有所不同，这可能会给自监督视听表征学习带来挑战。（右）我们提出在训练过程中利用电影配音，并证明这能提升所学表征在广泛任务上的质量。

摘要

视听表征学习通常依赖视觉与声音之间的对应关系。然而，同一个视觉场景往往可能对应多条音轨。例如，想象同一条拥挤街道上发生的不同对话。这类反事实配对对视听表征学习的影响此前尚未得到探讨。为研究这一问题，我们使用电影和电视节目的配音版本来增强跨模态对比学习。我们的方法学习将仅在语音上有所不同的替代音轨表征为与同一视频相似。通过一系列全面的实验，考察不同的训练策略后，我们的结果表明，这种通用方法能够提升多种下游听觉及视听任务的表现，同时总体上不会显著影响语言任务的表现。这些发现凸显了在学习场景级视听对应关系时考虑语音变化的重要性，并表明配音音频可作为一种有用的增强技术，帮助训练视听模型，使其在各种下游任务中表现得更加稳健。

1. 引言

两个视频看起来相似，听起来却不同，这可能吗？请看图1左侧的两个场景。它们来自不同来源，但有一些共同元素，例如背景中的小提琴手、远处的其他桌子，以及高档餐厅环境中一对情侣的说话声；但他们在说什么呢？即使其他方面没有变化，这两个场景中的对话内容也可能大不相同。通用的自监督视听表征通常侧重于非语音应用，这一点从现有训练数据集和常见的下游评估任务中可见一斑。仅在音频领域，语义语音处理之外也有无数应用，因此近期出现了一些基准测试，用于评估不同任务类型之间的泛化能力及其权衡 [87, 92]。那么，我们该如何着重从混有语音的视听内容中学习稳健的表征呢？重要的是，除了语义层面之外，还有许多值得关注的语音处理任务，也就是副语言任务，因为语音远不止可听见的文本。这些任务同样需要发现词语之外的其他相似性。

想象一个电影讨论场景，如图2所示。许多视听元素同时存在：背景交谈声、玻璃杯碰撞声、音乐、脚步声和角色的声音，但是先验地，在不改变基本场景属性的情况下，除了局部细节之外，这个场景可以包含许多不同的对话。

*作者在 Netflix 实习期间完成的大部分工作。

唇部动作等特征，这表明存在一种明确的反事实结构。请注意，还有其他与不同问题相关的反事实跨模态结构，例如同一段音乐对应的多个舞蹈视频。口语词语的差异是其中一种特定情况，也是我们探讨的内容。

在这项工作中，我们提出假设：这种看起来相似、听起来却不同的问题可能会出现在真实世界的视听数据分布中，并可能抑制自监督视听表征学习器的性能。既有方法（如跨模态对比学习）让模型学习区分真实视听配对与错误配对，因此可能会受到影响；在这种情况下，语言不同但其他方面相似的音视频配对可能成为混杂因素。然而，即使真实世界的视听场景分布总体上呈现这一趋势，通常也无法获得场景完全相同、仅对白不同的反事实版本。

我们提出利用一种自然呈现出这种反事实式结构的数据来源作为替代方案：**配音**。配音是电影音轨的另一种版本，其中语音被替换为第二语言改编，而其余声音通常保持不变。近期研究表明，利用电影场景进行训练能够取得出色的性能 [14, 40], , 因为与那些经过整理、专注于特定物体或动作的热门视听数据集相比，电影包含多样的视听混合内容。尽管这种分布可能有助于学习侧重于整体场景属性而非对白语义的表征（这正是我们的目标），但对电影中对齐的音频和视频进行对比学习训练，并未明确考虑因语言差异而导致的、看起来相似但听起来不同的场景。我们通过利用电影的多语言配音版本[1]改进了这一策略。具体而言，我们创建了一个电影和电视节目数据集，每部作品最多包含七条音轨：英语（**英语**）、西班牙语（**西班牙语**）、法语（**法语**）、日语（**JA**）、德语（**DE**）、意大利语（**IT**）和韩语（**KO**）。我们将训练策略应用于一种成熟的自监督对比学习框架，即 SimCLR [15], , 并证明这种方法能够提升多模态和单模态设置下的性能。总体而言，本研究的贡献包括：

- 一种使用**配音**（电影的次要音频语言版本）来改进自监督视听表征学习的方法。
- 大量实验表明，该方法不仅提升了多种听觉和视听任务的性能，还在多个基准测试中取得了新的最先进水平。
- 用于研究潜在权衡的其他实验。

[1]预训练数据还包括电视节目剧集。为避免冗长，除非有必要明确区分，否则我们将所有长篇内容统称为电影。



图2。考虑图中的场景。这些对白示-例中，哪一个更有可能？两者在该场景中都合理，但它们的语音声学特征会使配乐产生差异。

这些结果表明，我们可以在不显著影响语言识别和语义语音任务性能的情况下取得提升。

- 用于在各种语言中生成反事实配对的示例流程；我们将该工作流应用于LVU [94] 数据集，并展示了创建替代音轨的可能性，这些音轨有望赋能研究社群进一步研究口语词语对视听表征学习的影响。

2. 相关工作

自监督与多模态学习自监督学习依赖基于数据结构、由人工设计的监督信号构成的代理任务，而非人工标签，来学习有用的表征 [4, 22, 33, 35, 36, 56, 63, 101]。我们聚焦于对比学习，该方法通过最大化同一实例不同视图之间的互信息，展现出优异的性能 [4, 15, 29, 36, 85, 85, 86]。随后可以通过微调，或添加简单（通常是线性）模型，将这些表征适配到新任务；这两种方法对特定任务监督信号的需求都较少。跨模态对比学习尤其利用图像和文本等多模态数据 [71], , 或者像我们的情况一样，利用视频和音频 [1–3, 25, 37, 45, 52, 60, 61, 65, 67, 68, 91, 97, 100]。

视听学习 视听学习利用跨模态对应关系执行动作 [42, 45]和说话人 [17, 59] 识别、声源分离 [9, 73,89], 、媒体合成 [28, 34, 66, 83], 、音频空间化 [30, 57, 98], 、声学模拟 [12, 54, 79], 等任务。许多研究采用对比方法，认识到音频和视频可被视为同一潜在现象的两种互补感官视图，并着重学习协调一致的 [5]表征。以往研究发现，跨模态训练能够取得优于模态内训练的结果 [58], , 因此我们以跨模态设置为基础构建框架。在这项工作中，我们使用来自长篇内容（如电影和电视节目）的多语言音频配音和视频。电影包含丰富的视听对应关系，能够模拟现实世界中的体验，

而且比用户生成的视频更加多样、新颖，同时数量充足且易于扩展 [14, 38, 84]。

通用音频表征学习与评估 声音具有异质性，语音、音乐和环境声音的特征迥然不同。即使在语音领域内部，不同任务之间也存在显著差异，例如语音识别 [16, 49] 和语音情感识别 [78]。这促使研究者开发通用音频表征 [62, 75] 和 HARES [92] 、HEAR [87]等基准测试。我们主要使用 HEAR [87]评估音频，因为它提供了一致的 API。我们的核心假设是，如果跨模态设置中的配音增强训练能够提升表征的通用性，那么模型在各种任务上的表现就应当有所提高，同时避免在语言相关任务上出现显著的性能权衡。

多语言音频 多语言语音处理通过在多样化数据 [18, 32]上进行预训练，推动了语音识别 [7]等领域的发展。近年来，语音到语音翻译也已成为可能 [48]。视听媒体中的语音翻译通常称为**配音**。这是一种视听翻译 [11]，即将媒体作品中的语音内容（例如 电影）重新录制成另一种语言。在许多文化中，配音比字幕更为普遍 [10]。这提供了大规模的自然多语种数据，也为我们关于视听一致性的假设提供了一个具体案例：配音音轨与原版的区别仅在于口语语言。我们希望利用配音中平行的主要与次要音频，它们仅在语音上有所不同，从而学习更稳健的视听表征。我们还构建了一套合成流程来创建反事实配对，以展示反事实跨模态配对这一概念，并为研究社群今后的探索与验证提供支持。

3. 预训练数据集

我们的数据集由 ~20K 部电影和 ~33K 部电视剧集组成，共计 ~59K 个视频小时。为了尽量减少所学习表征中潜在的隐性偏差，我们特别关注预训练数据集中所使用标题的多样性，并且仅使用目录中的一小部分来研究这一问题。图3 详细展示了数据集中所含标题的类型和原始语言分布[2]。每个标题都包含一个视频轨道，以及最多七条音轨：英语（**英语**）、西班牙语（**西班牙语**）、法语（**法语**）、日语（**JA**）、德语（**DE**）、意大利语（**IT**）和韩语（**KO**）。大多数标题只有一条音轨，且几乎总是其原始语言；约四分之一的数据集为多语言数据集，其中每个标题平均提供 2.8 条音轨。这样的数据集使我们能够探究音频中口语词语对于

[2]更多详情请参见附录。

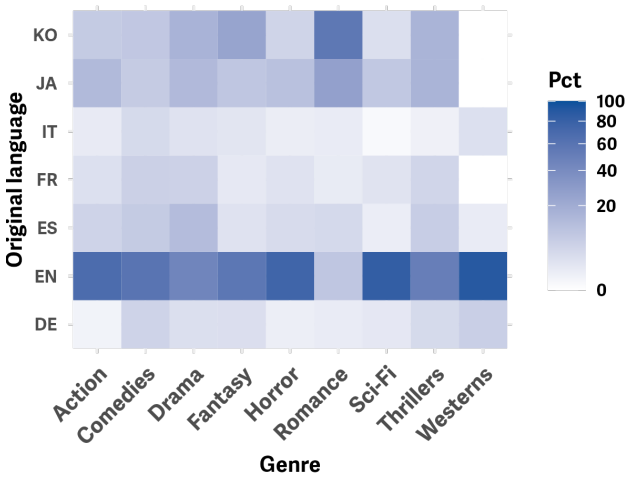

图3. 我们的预训练数据集所收录的电影和电视剧集，选自原始语言和类型各异的多元集合。我们的目标是尽可能减少可能影响自监督模型的潜在内容偏差和故事偏差。请注意，除了用于整理数据集外，我们不会将这些元数据用于表征学习。为便于可视化，我们对每列数据进行了归一化。

自监督视听表征学习的影响。拥有多种配音选项使我们能够研究次要语言之间的权衡，以及“多语言”模型是否可能进一步提升下游性能。

我们认识到，这类数据有可能对研究产生重大裨益。我们正在积极研究必要的法律步骤，以期发布该数据的一个非商业用途版本。图 4 展示了我们数据集中的一些样本，但我们鼓励读者访问项目页面查看更多示例[3]。

4. 方法

4.1. 方法

我们的预训练数据集记为 $\mathcal{X} = \{\mathcal{X}_n | n \in [1 \cdots N]\}$，其中 $\mathcal{X}_n = \{x_{n,m} | m \in [1 \cdots M_n]\}$ 包含M_n 个不重叠的片段，这些片段是根据数据集中 n^{th} 标题的时长按时间分段得到的。$\mathcal{Q}$ 是一个函数类，我们使用它来创建四元组训练实例 $(v_p, a_p, v_s, a_s) \sim \mathcal{Q}(x_{n,m})$ [4]，其中 v_p 和v_s 通过对$x_{n,m}$中的视频模态进行时空增强获得。类似地，a_p 和 a_s 也是针对音频模态的，但与视频不同，我们还可以进一步对音频实例添加配音增强。当有多种语言可用时，这将确保a_p 和 a_s 相似，唯一的区别在于它们所使用的口语语言。

[3]nikhilsinghmus.github.io/lssd[4]下标分别表示主要和次要

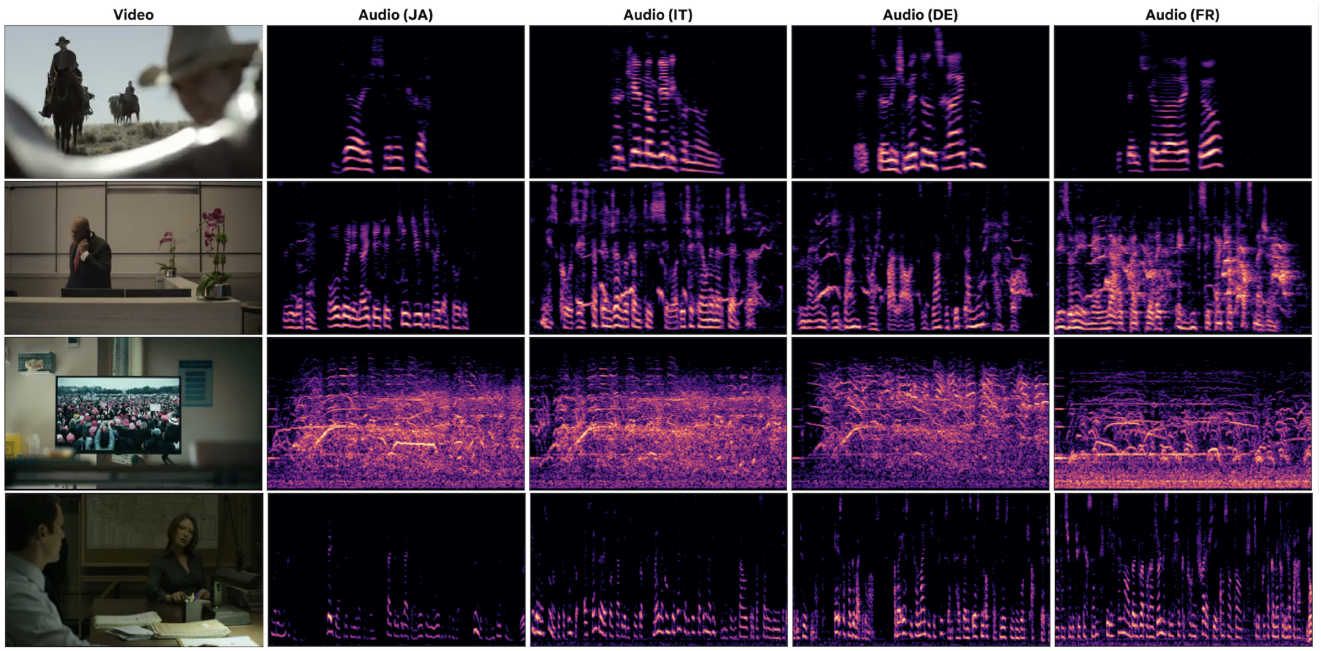

图 4. 我们的预训练数据集中的示例片段，展示了每条音轨的视频静帧和梅尔频谱图。

随机采样负样本，这种在度量学习和对比学习中的传统方法，已被观察到效果不佳 [52, 77]。近期有许多工作开发了所谓的困难负样本挖掘方法，目标是用具有挑战性的示例填充负样本集 [64, 74]。在我们的情况下，数据具有层次结构；片段自然地嵌套在源长篇标题中，并且来自同一标题的片段共享若干共同属性，包括角色、地点、物体、声音和美学特征。因此，遵循先前工作 [40], 创建一个小批次 $\mathcal{B} = \{x_i | i \in [1 \cdots B]\}$的方法，我们首先均匀采样一个标题， $n \sim \mathbb{U}(1, N)$，然后从 $\mathcal{X}_n$ 中抽取多个不同的片段。这确保了对于$\mathcal{B}$中的每个实例，始终有足够数量的来自同一标题的样本充当困难负样本。这一点很重要，因为 $B \ll N$，因此对于 $n \sim \mathbb{U}(1, N)$ 和 $m \neq m'$，$\mathrm{P}(x_{n,\,m} \in \mathcal{B} \wedge x_{n,\,m'} \in \mathcal{B}) \rightarrow 0$。换句话说， $x_i \sim \bigcup_{n=1}^{N} \mathcal{X}_n$ 的朴素随机采样策略主要会导致容易区分的跨标题负样本。

现在我们可以制定训练目标。对于跨模态设置，$\mathcal{B} = \{(v_i, a_i) | i \in [1 \cdots B]\}$表示一个大小为 B的小批量，其中第 i^{th} 个实例的视频和音频模态分别表示为 v_i 和 a_i。我们用 z_v^i 和 z_a^i 表示它们各自的嵌入。对于小批量中的第 i^{th} 个元素，(z_v^i, z_a^i)构成正样本对；同时，假设两种模态都存在负样本对，$\mathcal{N}_i = \{(z_v^i, z_a^j)、(z_v^j, z_a^i) | j \in [1 \cdots B]、 i \neq j\}$ 构成负样本对集合。基于此，公式 1 展示了跨模态归一化温度缩放交叉熵目标 [15]与 i^{th} 个实例相关。由于 $(v, a) \in \{(v_p, a_s)、(v_s, a_p)\}$，实际中我们优化汇总所有可用实例的公式2。

$$\ell_i(v, a) = -\log\left(\frac{e^{((z_v^i)^\intercal (z_a^i))/\tau}}{e^{((z_v^i)^\intercal (z_a^i))/\tau} + \sum\limits_{(z_v', z_a') \in \mathcal{N}_i} e^{((z_v')^\intercal (z_a'))/\tau}}\right) \tag{1}$$

$$\mathcal{L} = \frac{1}{2} \sum_{i=1}^{B} \Big(\ell_i(v_p, a_s) + \ell_i(v_s, a_p)\Big) \tag{2}$$

$$\mathcal{L}_v = \sum_{i=1}^{B} \ell_i(v_p, v_s), \quad \mathcal{L}_a = \sum_{i=1}^{B} \ell_i(a_p, a_s) \tag{3}$$

公式 3 展示了视频和音频模态的损失函数的模态内变体。除非另有明确说明，我们的模型均从头开始训练，并采用跨模态，即如公式2所示，计算模态间的对比损失。我们这样做的依据是早期实验中的观察：从头开始训练且不调整额外的缩放参数时，相比之下，模态内对比任务过于简单，导致相应项提前收敛。这种方法也得到了先前文献的支持 [58]。尽管没有直接优化模态内项，我们仍会在自监督预训练期间跟踪 $\mathcal{L}_v$ 和 $\mathcal{L}_a$，并观察到它们会随着最小化 $\mathcal{L}$而作为副产品降低。在我们的建模方案中，有些变体会将 $\mathcal{L}_v$ 和

$\mathcal{L}_a$ 纳入总损失函数（例如 $\mathcal{L}+\lambda_v\mathcal{L}_v+\lambda_a\mathcal{L}_a$），我们将在第5.2节稍后讨论。

4.2. 架构

为了验证我们的数据和训练方法的效果，我们采用标准的骨干架构。我们的视频模型是一种多尺度视觉变换器 [24],，具体来说是 MViT-S；我们的音频模型采用类似的架构，但略作修改，以便能够处理作为输入的音频频谱图。请注意，我们从头开始，在第3节详述的预训练数据集上训练所有模型。我们使用单个（权重共享的）音频骨干网络处理所有音频频谱图，不论其语言如何。按照对比学习中的常见做法，我们为每种模态各使用一个多层感知机（MLP）投影头，以在训练过程中、计算对比损失之前，进一步降低表征的维度。这些附加层会在预训练后丢弃。

5. 实验

5.1. 下游任务

音频任务与基准 我们在一系列多样化的听觉任务上进行评估，以探究所学习表征的质量，这些任务选自 HEAR [87]挑战基准测试。我们筛选出与假设相关的任务，并聚焦于使用池化表征（而非时间密集型表征）的任务。

声音与场景分类：这些任务完全属于非语言类任务，我们推测，训练时降低对语言的重视应有助于提升此类任务的性能。我们纳入了ESC-50 [69], FSD50K [27], 和人声模仿（VI） [43]。VI 是一项基于发声查询（QBV）的任务，不过由于它基于 AudioSet [31] 本体中的声音事件，因此我们将其归入此类别。非语义语音：语音或副语言学属性中的许多非语义特征可能在不同语言之间共享，而此类信号对于一系列任务都十分重要。我们在此纳入用于情感识别的 CREMA-D [8] 、用于音乐/语音区分的 GTZAN [88]，以及用于说话人数估计的 LibriCount [82] 。我们推测，如果我们的方法提高了对非语言语音属性的关注度，性能应会得到提升。语义语音：为探究一种潜在的权衡关系，我们在语义语音任务上进行评估。我们将关键词理解视为使用池化表征的语音识别的代理任务。为此，我们采用 HEAR [87]中实现的 Speech Commands [93] 的完整版本。语言：衡量可能存在的权衡关系的另一种方式，是评估模型在基于音频的语言识别任务上的表现，以了解对此任务有用的特征是否保留在所学习的表征中。出于这一原因，我们纳入了 VoxLingua107 Top10 [90]。

	# 数据	初始化。	(λ_v, λ_a)	原始语言	avg. 配音	dub 增强
A.1	4.6M	rand.	(0,0)	**ESF**	2.8	✗
A.2	4.6M	随机。	(0,0)	**ESF**	2.8	✓
A.3	4.6M	A.1	(0,0.2)	**ESF**	2.8	✗
A.4	4.6M	A.2	(0,0.2)	**ESF**	2.8	✓
B.1	1180万	随机	(0,0)	**EN**	1.0	✗
B.2	9.8M	随机	(0,0)	$\mathbb{U}\backslash$**EN**	0.2	✗
B.3	19.4M	随机	(0,0)	$\mathbb{U}$	0.6	✗
B.4	5.1M	B.3	(0,0)	$\mathbb{U}$	2.8	✓
B.5	5.1M	B.3	(0.2,0.2)	$\mathbb{U}$	2.8	✓
C.1	19.4M	随机。	(0,0)	$\mathbb{U}$	0.6	✓
C.2	5.1M	C.1	(0.1,0.1)	$\mathbb{U}$	2.8	✓

表1. 不同预训练模型变体的详细信息。这里，**ESF** := {**英语**，**西班牙语**，**法语**} 表示三种语言的并集。$\mathbb{U}$ 表示包含全部七种语言的全集。

视觉与视听任务 我们还会独立评估视觉表征，以及视觉表征与听觉表征协同工作的效果。继近期从长篇内容中进行表征学习的研究之后 [14], 我们将LVU [94] 基准测试纳入评估套件，该基准测试涵盖长视频理解的各个方面。LVU [94] 包含涵盖长篇视频广泛方面的小规模任务，包括内容理解（关系、 说话风格、 场景/地点），以及电影元数据预测（导演、 类型、 编剧、电影上映年份）。在 LVU 任务中，我们同时使用视觉和听觉表征，探究其优势与潜在权衡。总体而言，我们预计性能会有所提升，但说话风格除外，因为目前尚不先验地 明确，在预训练期间降低对口语词汇的重视是否会损害此类下游任务。

评估 自监督预训练完成后，我们舍弃投影头，并使用骨干网络架构从音频和视频资产中提取特征。除非另有说明，我们对输出张量执行时空平均池化，以便为下游任务中的每个数据实例获得一个 d维向量嵌入。然后，我们按照相关基准测试中规定的方法，在这些表征上训练多层感知机或线性探测器。更多实现细节请参见附录。

5.2. 模型

总计，我们训练了**11** 个模型变体，详见表1，并在**15** 个涵盖音频和视频模态的不同任务上对它们进行评估。

第一组（A）模型变体采用小规模多语言预训练方案，用于首次研究配音增强的影响。我们采样英语（**英语**）、

	HEAR							LVU						
	ESC	LibCnt	CREMA	VI	FSD	语音	VoxLng	导演	类型	关系	场景	说话	编剧	Year
A.1	77.20	67.29	59.52	10.37	44.52	74.83	27.16	44.86	54.42	36.59	**45.12**	42.86	**38.10**	41.84
A.2	75.95	67.94	59.76	11.14	44.23	73.80	23.87	47.66	56.63	36.59	41.46	40.74	33.33	41.84
A.3	82.00	67.87	**62.69**	11.39	48.90	**79.47**	**28.70**	**49.53**	57.65	43.90	39.02	43.92	33.93	46.10
A.4	**83.05**	**68.65**	61.95	**12.57**	**49.42**	74.38	26.55	44.86	**59.01**	**46.34**	**45.12**	**48.15**	29.17	**47.52**
B.1	84.15	67.12	61.00	**13.05**	50.29	82.31	24.69	47.66	57.14	51.22	41.46	42.33	32.14	**45.39**
B.2	82.00	67.10	61.98	11.86	49.07	82.90	28.09	42.99	55.95	48.78	42.68	47.62	30.36	44.68
B.3	**85.60**	66.31	62.79	11.55	**53.69**	**83.82**	**30.35**	50.47	**60.20**	46.34	42.68	48.68	37.50	**45.39**
B.4	83.75	68.88	63.18	10.82	51.61	77.12	28.19	**51.40**	59.69	**56.10**	46.34	**49.21**	**38.10**	44.68
B.5	85.25	**69.16**	**63.27**	11.38	52.48	76.99	27.98	**51.40**	58.33	51.22	**52.44**	48.68	36.31	**45.39**
C.1	84.10	67.57	63.70	**12.12**	51.96	**81.88**	29.42	42.99	**58.84**	48.78	46.34	41.27	38.69	41.13
C.2	**85.50**	**68.90**	**64.28**	11.90	**52.55**	77.14	**29.94**	**48.60**	57.65	48.78	**51.22**	**50.79**	**39.88**	**49.65**

表 2. **使用音频的消融实验结果**。 除 FSD50K [27] 和人声模仿 [43]（平均精度均值）外，所有指标均为 top-1准确率。我们遵循 HEAR [87] 基准测试规定的评估策略：在下游任务的冻结嵌入上训练一个多层感知机。对于 LVU [94], , 我们使用官方数据划分，并训练一个线性探测器。结果展示的是测试集上的表现，其中报告的最佳训练轮次根据验证集上的相同指标选取。所有模型变体在 GTZAN 上均取得了 100.0 的 top-1准确率，因此此处未纳入该任务。我们用 **粗体**表示每个消融组中的最佳表现。HEAR [87] 任务从左到右依次为 ESC-50、LibriCount、CREMA-D、人声模仿、FSD-50k、SpeechCommands（完整数据集）和 VoxLingua107 Top10。

	HEAR								LVU							
	ESC	LibCnt	CREMA	VI	FSD	语音	VoxLng	GTZAN		导演	类型	关系	场景	说话	编剧	年份
基准测试 [87]	96.65	78.53	75.21	22.69	65.48	97.79	72.02	99.23	Obj Tr [94]	58.90	56.10	54.70	60.00	40.30	35.10	40.60
基准测试(SSL)	80.50	78.53	75.21	18.48	50.88	96.87	71.40	96.86	M2S [14]	70.90	55.90	71.20	68.20	42.20	53.70	57.80
GURA [95]	74.35	68.34	75.21	18.48	41.32	94.68	71.40	93.59	ViS4mer [39]	62.61	54.71	57.14	67.44	40.79	48.80	44.75
PaSST [46]	94.75	66.01	61.04	18.20	64.09	63.87	25.93	97.69	SCALE [76]	49.09	58.97	76.47	74.02	42.27	62.76	39.23
CLAP [23]	96.70	77.83	64.36	–	58.59	96.83	–	100.0	STCA [20]	66.70	56.62	59.25	69.15	41.62	52.93	53.30
Ours																
B.3 (A)	85.60	66.31	62.79	11.55	53.69	83.82	30.35	100.0	B.3 (V)	69.16	60.88	60.98	63.41	46.03	48.81	52.48
B.4 (A)	83.75	68.88	63.18	10.82	51.61	77.12	28.19	100.0	B.4 (V)	67.29	61.73	60.98	65.85	47.62	41.67	55.32
B.5 (A)	85.25	69.16	63.27	11.38	52.48	76.99	27.98	100.0	B.5 (V)	69.16	64.29	58.54	64.63	46.03	41.07	52.48

表 3. HEAR 各项任务的最先进结果 [87] （新增 GTZAN 音乐/语音）和 LVU [94]。在 HEAR 上，我们将结果与以下模型进行比较：(1) HEAR 排行榜上各项任务的最佳结果；(2) 与 (1) 相同，但仅考虑自监督模型；(3) GURA Fuse HuBERT [95], , 其平均表现最佳；(4) CP-JKU PaSST 2lvl+mel [46], , 在 GURA 模型之后平均表现最强；(5) 最新的 CLAP 模型 [23]。在 LVU 上，我们将结果与原始 LVU 论文中的 Object Transformer [94], 以及近期进展进行比较：ViS4mer [39], 、SVT SCALE 模型 [76], 、STCA [20], 和 Movies2Scenes [14]。Movies2Scenes 使用电影元数据，因此引入了特定任务监督。报告结果时，（A）表示仅使用音频表征，（V）表示仅使用视频表征。

西班牙语（**西班牙语**）或法语（**法语**）标题，这些标题至少有一种可用配音，从而能够系统地研究配音增强的影响。对于每个标题，我们从全部七种语言中采样配音。**A.3** 和 **A.4** 变体引入了一个显式的模态内项，即 $\mathcal{L}_a$。我们假设，在配音增强的作用下，$\lambda_a > 0$ 在语言和语言识别任务上可能产生更大的差距。这是因为该优化过程明确最大化了仅在口语语言上存在差异的音频嵌入之间的相似性，而不是仅通过 $\mathcal{L}$隐式地实现这一点。需要强调的是，**A.3** 和 **A.4** 的预训练步数总量相同；同样，比较**A.1** 和**A.2** 时也是如此。

第二（B）组模型变体旨在了解数据规模和语言多样性的影响。我们

与 A 组的实验相比，预训练样本数量大约翻了一倍，**A** 并研究这是否会带来质量更高的表征。这一点很重要，因为自监督预训练的计算开销很大，而且事先并不清楚 规模更大、更多样化的预训练数据是否必然会带来更好的模型。**B.3** 在包括所有语言在内的全部预训练样本上进行训练，以检验多语言预训练的极限，不使用双重增强。通过比较**B.4** 和**B.5**，我们希望深入了解后者所采用的模态内目标函数的行为。

第三组（C）实验探究了更深层架构的影响，即 MViT-B [24]（相较于作为默认架构的 MViT-S [24]）。我们保持数据规模和多样性

	导演	类型	关系	场景语音		编剧	Year
A.1	53.27	54.59	43.90	52.44	34.39	36.90	42.55
A.2	53.27	55.44	41.46	50.00	**41.27**	35.12	42.55
A.3	57.01	**57.48**	**46.34**	**57.32**	39.68	**38.69**	46.10
A.4	**63.55**	**57.48**	36.59	53.66	36.51	33.93	**47.52**
B.1	60.75	55.78	**48.78**	53.66	38.10	35.71	42.55
B.2	54.21	57.65	46.34	51.22	37.04	**38.69**	44.68
B.3	**65.42**	57.48	41.46	53.66	39.68	38.10	45.39
B.4	62.62	**58.50**	36.59	**59.76**	**43.39**	35.12	46.81
B.5	62.62	58.16	43.90	**59.76**	39.15	37.50	**49.65**
C.1	**63.55**	55.10	43.90	57.32	**40.74**	**39.29**	**45.39**
C.2	61.68	**56.63**	**46.34**	**60.98**	40.21	36.90	43.97

表4。**采用视频的消融实验结果**。所有指标均为 top-1 准确率。我们遵循 LVU 基准测试规定的数据划分，并在下游任务的冻结**视频** 嵌入上训练线性探测器。结果展示在测试集上，报告的最佳轮次根据验证集选出。我们用**粗体**标出每组消融实验中的最佳性能。

与B.3 、B.4 和B.5 变体相同。与这些变体类似，我们首先使用全部数据进行训练，然后从C.1 的最终检查点继续微调，仅使用包含多个音轨的标题子集。这样可以确保在C.2 的每个优化步骤中都进行配音增强。

借助涵盖音频和视频模态的多种预训练模型与下游任务，我们现在可以全面研究多语言和多模态表示学习的各个方面。

5.3. 消融实验

配音增强预训练有帮助吗? 为了解决这个问题，我们首先考察表2 中的第一（A）组模型变体。2我们假设，配音增强应能提升声音/场景分类和非语义语音任务的性能。在 HEAR [87]基准测试中，除 CREMA-D [8]，之外，我们的定量结果证实了这一点。LVU [94] 任务也被视为非语言任务，表2 显示，在大多数任务中，配音增强变体相较于其基线对应版本都带来了显著的性能提升。我们的第二个假设是，配音增强应会影响语言任务和语言识别任务，因为其目标是减弱口语词语在音频表征中的影响。事实上，我们可以看到A.4 使用配音增强的方法表现不佳A.3 在 Speech Commands 和 VoxLingua 上。表2 还表明，配音增强可从模态内目标即 $\mathcal{L}_a$中获益；并且要使这种方法有效，我们实际上需要——正如预期的那样——在预训练期间拥有足够数量的包含替代音轨的样本。

配音增强模型仍然能够识别语言并执行语言学任务吗? 表中所示结果2 在 VoxLingua上表明，在小规模（A 变体）和大规模（B 变体）设置中强制采用配音增强明显影响语言识别性能。我们通过比较 A.2 对比 A.1 ，或 B.4 对比 B.3 来衡量这一点。对于 Speech Commands，我们也观察到类似的行为 [93]，，将我们的语言学性能代理指标实现为关键词检测。然而，在这两种情况下，性能下降都不足以阻止经过配音增强的模型识别语言或执行语言学任务。我们假设存在这种建模权衡， 即 尽管性能可能下降，但其影响将是有限的。

视频表征的质量是否受到影响? 为回答这一问题，我们查看表 4 ，其中对 LVU [94] 任务进行评估，使用冻结的视频嵌入上的线性探测器。在小规模预训练设置中，我们观察到一种混合模式：配音增强变体 即 A.2 和 A.4 在 3 项任务（“导演”、“说话方式”和“年份”）上优于其对应版本，而在其余任务上则表现更差或持平。在大规模预训练设置中，我们看到更明确的趋势：B.4 和 B.5 在 7 项 LVU 任务中的 5 项上表现优于 B.3 ，这表明在多样化的评估集上，配音增强预训练总体上也有助于仅视频任务。

语言多样性如何影响预训练? 要妥善回答这一研究问题，需要进一步考察 B.1 、B.2和 B.3。值得再次强调的是，尽管预训练样本数量不同（见表1），为了进行公平比较，我们训练这三种模型变体时所采用的梯度优化步数大致相同。总体而言，在音频（参见表2）和视频（参见表4）任务中，我们都观察到，最大限度地增加语言多样性（参见 B.3 ）能够提升性能。然而，在预训练期间纳入英语（**英语**）标题似乎至关重要，因为英语是占比最高的原始语言（见图 3）。表 2 展示了 VoxLingua [90] 和 Speech Commands [93]，中的明显规律：预训练期间语言多样性越高，性能提升越显著，例如，VoxLingua 的绝对提升达 5.6% [90]。

更深的架构更好吗? 对于表 2 和 4中的每项任务，我们可以比较最强的B 模型变体与 C 变体。除少数例外情况外，我们的定量结果并未表明，使用 MViT-B [24] 并配备~40% 更多参数，能够相较于较小的 MViT-S [24] 带来足够显著的提升，从而证明预训练期间显著增加的计算量是合理的。我们承认，如果下游任务是通过微调（而不是线性/多层感知机探测）进行评估的，尤其是 HEAR 中的大规模任务，这一结论可能并不成立 [87]。

补充实验在附录中，我们在一个小型配音视听数据集上提供了更多结果，并匹配了较小的骨干架构；在该数据集中，四种语言之间完全平衡（超过 700 个**英语**标题，其中有

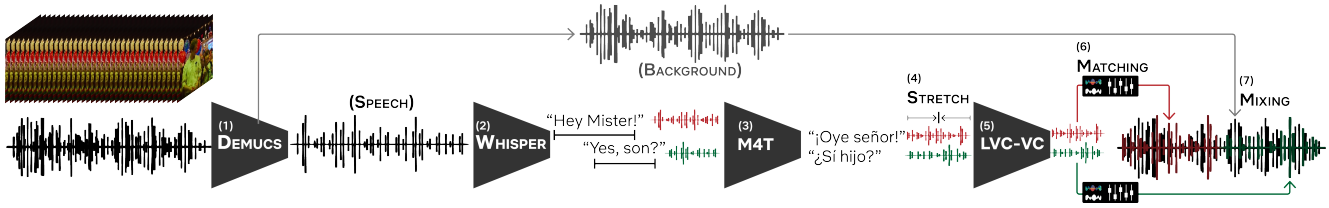

图 5。生成合成反事实对的流程。

西班牙语、**法语**和**日语** 均可用）。我们还将其与语音移除策略进行比较：我们对完整数据集进行源分离，并移除语音部分，作为另一种降低语音重要性的策略。由于语言保持平衡，我们还评估了使用特定配音增强组合（例如 **英语+西班牙语**）的“双语”模型。这些结果系统地表明，即使在规模较小的设置中，配音增强训练也有益；其效果优于语音移除策略；而多语言模型（使用多个配音，并像本文主要结果中那样随机采样）还能进一步提升鲁棒性。

5.4. 与最先进水平的比较

HEAR 表 3 将我们的结果与 HEAR 上的多个优秀结果进行了比较 [87] 任务。在 ESC-50、FSD50K 和 GTZAN 音乐/语音任务上，我们的结果优于 HEAR 排行榜上表现最佳的自监督结果，且至少优于另一项结果。在大多数任务上（人声模仿除外），我们的表现至少优于其中一个模型，展现了在这些不同任务上的稳健性。

LVU 同样在表 3中，我们将最强模型与 7 项 LVU 基准测试上的最先进结果进行比较 [94] 任务。我们的模型在类型和 说话 任务上取得了新的最先进表现，较以往结果有了显著提升。不考虑使用电影元数据的 Movies2Scenes [14],时，我们在导演 和 年份 任务上也取得了最先进结果（总计 7 项中的 4 项）。在其余任务中，我们的结果也极具竞争力。这表明，在我们采用配音增强的数据集上预训练的模型，在多种视频理解基准测试上的表现可媲美甚至超过现有最佳模型。总体而言，这些结果凸显了我们方法的有效性。

6. 合成反事实对

为推动对音视频表征学习中反事实配对的研究，我们提出了一套模块化流程，如图 5 所示，用于在任意目标片段上模拟类似配音的反事实配对，使其与我们预训练数据中的一对多音视频分布相似。尽管该流程生成的内容质量有限，但它可作为一种简单工具，用于……

缓解研究社群开展类似研究时的数据限制。

步骤为：(1) 使用 Demucs 从背景声音中分离语音 [19], (2) 使用 Whisper 转录并分割语音 [72], 生成带时间戳的片段 (3) 使用 SeamlessM4T 将语音（或可选地，将文本）翻译成目标语 言 [6](4) 使用拉伸将翻译与原始片段对齐 (5) 使用 LVC-VC 转换人声，使其匹配原始演员的声音 [41] (6) 使用 pyloudnorm 对输出进行响度 归一化和均衡匹配，使其与原始音频匹配 [81]并使用 matchering[5] (7) 将片段重新放回其原始位置，与背景音频重新混音，并与原始视频复用。该流程还实现了其他中间步骤，例如重采样，以衔接主要步骤。

作为面向社区的概念验证资源，我们使用这一流程制作了多语言版本的 LVU [94]。LVU-M 展示了大规模生成反事实数据的可行性。我们将开源这一流程，以便创建此类“看起来相似，听起来不同”的数据集。我们也希望未来的进展能够提升此类数据结构的质量，并推动更深入的研究。

7. 结论

在这项工作中，我们提出了看起来相似，听起来却不同 这一问题：感知上相似的场景可能包含不同的语音内容。我们证明，可以利用具有相似结构的反事实数据来源——配音电影——在成熟的跨模态对比学习方案中提升视听表征学习效果。我们在一个包含大量电影和电视节目的预训练数据集上开展实验，结果表明，这种方法能够提升多种听觉和视听任务的表现。因此，配音增强训练是一种可扩展且有效的方法，无需监督即可学习更稳健的视听表征。

参考文献

[1] Humam Alwassel、Dhruv Mahajan、Bruno Korbar、Lorenzo Torresani、Bernard Ghanem 和 Du Tran。通过跨模态音视频聚类进行自监督

[5] https://github.com/sergree/matchering

学习。载于《神经信息处理系统进展》，第9758–9770页，2020年。2[2] Relja Arandjelovic 和 Andrew Zisserman。看、听与学习。载于《IEEE国际计算机视觉会议论文集》，第609–617页，2017年。[3] Yusuf Aytar、Carl Vondrick 和 Antonio Torralba。Soundnet：从无标签视频中学习声音表征。载于《神经信息处理系统进展》，第892–900页，2016年。2[4] Philip Bachman、R Devon Hjelm 和 William Buchwalter。通过最大化跨视图互信息学习表征。载于《神经信息处理系统进展》，第15535–15545页，2019年。2[5] Tadas Baltruˇsaitis、Chaitanya Ahuja 和 Louis-Philippe Morency。多模态机器学习：综述与分类。IEEE模式分析与机器智能汇刊，41(2)：423–443，2018年。2[6]Loıc Barrault、Yu-An Chung、Mariano Cora Meglioli、David Dale、Ning Dong、Paul-Ambroise Duquenne、Hady Elsahar、Hongyu Gong、Kevin Heffernan、John Hoffman 等。SeamlessM4T——大规模多语言与多模态机器翻译。arXiv预印本，arXiv:2308 11596，2023年 8 。 。[7] William Byrne、Peter Beyerlein、Juan M Huerta、Sanjeev Khudanpur、Bhaskara Marthi、John Morgan、Nino Peterek、Joe Picone、Dimitra Vergyri 和 T Wang。迈向语言无关的声学建模。载于《2000 I EEE国际声学、语音与信号处理会议。会议论文集》（目录编号：00CH37100），第II1029–II1032页。IEEE，2000年。3[8] Houwei Cao、David G Cooper、Michael K Keutmann、Ruben C Gur、Ani Nenkova 和 Ragini Verma。CREMA-D：众包情感多模态演员数据集。IEEE情感计算汇刊，5(4)：377–390，2014年。5，7，3[9] Moitreya Chatterjee、Jonathan Le Roux、Narendra Ahuja 和 Anoop Cherian。用于音频源分离的视觉场景图。载于《IEEE/CVF国际计算机视觉会议论文集》，第1204–1213页，2021年。2[10] Frederic Chaume。视听翻译的转向：新受众与新技术。Translation Spaces，2(1)：105–123，2013年。3[11] Frederic Chaume。视听翻译：配音。Routledge，2020年。3[12] Changan Chen、Ruohan Gao、Paul Calamia 和 Kristen Grauman。视觉声学匹配。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第18858–18868页，2022年。2[13] Honglie Chen、Weidi Xie、Andrea Vedaldi 和 Andrew Zisserman。VGGSound：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。1，2，3[14] Shixing Chen、Chun-Hao Liu、Xiang Hao、Xiaohan Nie、Maxim Arap 和 Raffay Hamid。Movies2Scenes：使用电影元数据，用于学习场景表征。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 6535–6544 页，2023年。2, 3, 5, 6, 8[15] Ting Chen、Simon Kornblith、Mohammad Norouzi 和 Geoffrey Hinton。用于视觉表征对比学习的简单框架。arXiv 预印本 arXiv:2002 05709，2020年 2, 4 . .[16] Po-Han Chi、Pei-Hung Chung、Tsung-Han Wu、Chun-Cheng Hsieh、Yen-Hao Chen、Shang-Wen Li 和 Hung-yi Lee。音频 ALBERT：用于音频表征自监督学习的轻量级 BERT。载于 2021年 IEEE口语技术研讨会（SLT），第 344–350 页。IEEE，2021年。3[17] Joon Son Chung、Arsha Nagrani 和 Andrew Zisserman。VoxCeleb2：深度说话人识别。arXiv 预印本 arXiv:1806.05622，2018年。2[18] Alexis Conneau、Alexei Baevski、Ronan Collobert、Abdelrahman Mohamed 和 Michael Auli。用于语音识别的无监督跨语言表征学习。arXiv 预印本 arXiv:2006.13979，2020年。3[19] Alexandre D´efossez、Nicolas Usunier、L´eon Bottou 和 Francis Bach。Demucs：利用额外的未标注混音数据进行音乐源深度提取。arXiv 预印本 arXiv:1909.01174，2019年。8, 2[20]Ali Diba、Vivek Sharma、Mohammad Arzani、Luc Van Gool 等。时空卷积注意力视频网络。载于 IEEE/CVF国际计算机视觉会议论文集，第 859 869 页，2023年 6 – 。[21] Shuangrui Ding、Maomao Li、Tianyu Yang、Rui Qian、Hao-hang Xu、Qingyi Chen、Jue Wang 和 Hongkai Xiong。通过前景-背景合并实现运动感知对比视频表征学习。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第9716–9726 页，2022年。2[22] Carl Doersch、Abhinav Gupta 和 Alexei A Efros。通过上下文预测进行无监督视觉表征学习。载于 IEEE国际计算机视觉会议论文集，第1422–1430 页，2015年。2[23] Benjamin Elizalde、Soham Deshmukh、Mahmoud Al Ismail 和 Huaming Wang。CLAP：通过自然语言监督学习音频概念。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。6[24] Haoqi Fan、Bo Xiong、Karttikeya Mangalam、Yanghao Li、Zhicheng Yan、Jitendra Malik 和 Christoph Feichtenhofer。多尺度视觉变换器。载于 IEEE/CVF国际计算机视觉会议论文集，第 6824–6835 页，2021年。5, 6, 7[25] Haytham M Fayek 和 Anurag Kumar。利用弱标注数据进行大规模音视频声音学习。arXiv 预印本 arXiv:2006.01595，2020年。2[26] Christoph Feichtenhofer。X3D：扩展架构以实现高效视频识别。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 203–213 页，2020年。2

[27] Eduardo Fonseca、Xavier Favory、Jordi Pons、Frederic Font 和 Xavier Serra。FSD50K：一个由人工标注声音事件的开放数据集。《IEEE/ACM音频、语音与语言处理汇刊》，30：829–852，2021年。5，6，3[28] 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。施普林格，2020年。2[29] Tianyu Gao、Xingcheng Yao 和 Danqi Chen。SimCSE：句子嵌入的简单对比学习。载于EMNLP（1），2021年。2[30]Rishabh Garg、Ruohan Gao 和 Kristen Grauman。用于从视频生成双耳音频的几何感知多任务学习。arXiv预印本 arXiv:2111 10882，2021年 2 . .[31] 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。5[32] Arnab Ghoshal、Pawel Swietojanski 和 Steve Renals。深度神经网络的多语言训练。载于《2013年 IEEE国际声学、语音与信号处理会议》，第7319–7323页。IEEE，2013年。3[33] Spyros Gidaris、Praveer Singh 和 Nikos Komodakis。通过预测图像旋转进行无监督表征学习。载于《国际学习表征会议》，2018年。2[34] Yudong Guo、Keyu Chen、Sen Liang、Yong-Jin Liu、Hujun Bao 和 Juyong Zhang。AD-NeRF：用于说话头部合成的音频驱动神经辐射场。载于《IEEE/CVF国际计算机视觉会议论文集》，第5784–5794页，2021年。2[35] Raia Hadsell、Sumit Chopra 和 Yann LeCun。通过学习不变映射进行降维。载于《2006年 IEEE计算机学会计算机视觉与模式识别会议（CVPR'06）》，第1735–1742页。IEEE，2006年。2[36] Kaiming He、Haoqi Fan、Yuxin Wu、Saining Xie 和 Ross Girshick。用于无监督视觉表征学习的动量对比。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第9729–9738页，2020年。2[37] Di Hu、Feiping Nie 和 Xuelong Li。用于无监督视听学习的深度多模态聚类。载于《IEEE计算机视觉与模式识别会议论文集》，第9248–9257页，2019年。2[38] Qingqiu Huang、Yu Xiong、Anyi Rao、Jiaze Wang 和 Dahua Lin。MovieNet：用于电影理解的整体性数据集。载于《欧洲计算机视觉会议》，第709–727页。施普林格，2020年。3[39] Md Mohaiminul Islam 和 Gedas Bertasius。利用状态空间视频模型进行长电影片段分类。载于《欧洲计算机视觉会议》，第87–104页。施普林格，2022年。6。

[40] Mahdi M Kalayeh、Shervin Ardeshir、Lingyi Liu、Nagendra Kamath 和 Ashok Chandrashekar。电影中的视听对比学习负采样研究。arXiv预印本 arXiv:2205.00073，2022年。2，4[41] Wonjune Kang、Mark Hasegawa-Johnson 和 Deb Roy。采用位置可变卷积的端到端零样本语音转换。载于《INTERSPEECH 2023论文集》，第2303–2307页，2023年。8[42] Evangelos Kazakos、Arsha Nagrani、Andrew Zisserman 和 Dima Damen。Epic-fusion：用于第一人称动作识别的视听时间绑定。载于《IEEE/CVF国际计算机视觉会议论文集》，第5492–5501页，2019年。2[43] Bongjun Kim、Madhav Ghei、Bryan Pardo 和 Zhiyao Duan。人声模仿数据集：使用AudioSet本体对声音事件进行人声模仿的数据集。载于DCASE，第148–152页，2018年。5，6，3[44] Diederik P Kingma 和 Jimmy Ba。Adam：一种随机优化方法。arXiv预印本 arXiv:1412 6980，2014年 3 。.[45] Bruno Korbar、Du Tran 和 Lorenzo Torresani。通过自监督同步协同学习音频和视频模型。载于《神经信息处理系统进展》，第7763–7774页，2018年。2[46] Khaled Koutini、Jan Schl¨uter、Hamid Eghbal-zadeh 和 Gerhard Widmer。使用Patchout高效训练音频Transformer。arXiv预印本 arXiv:2110.05069，2021年。6[47] Hildegard Kuehne、Hueihan Jhuang、Est´ıbaliz Garrote、Tomaso Poggio 和 Thomas Serre。HMDB：用于人体动作识别的大型视频数据库。载于2011年国际计算机视觉会议，第2556–2563页。IEEE，2011年。1，2，3，5[48] Ann Lee、Hongyu Gong、Paul-Ambroise Duquenne、Holger Schwenk、Peng-Jen Chen、Changhan Wang、Sravya Pop- uri、Juan Pino、Jiatao Gu 和 Wei-Ning Hsu。真实数据上的无文本语音到语音翻译。arXiv预印本 arXiv:2112.08352，2021年。3[49] Andy T Liu、Shu-wen Yang、Po-Han Chi、Po-chun Hsu 和 Hung-yi Lee。Mockingjay：使用深度双向Transformer编码器进行无监督语音表征学习。载于ICASSP 2020—2020 IEEE国际声学、语音与信号处理会议（ICASSP），第6419–6423页。IEEE，2020年。3[50] Ilya Loshchilov 和 Frank Hutter。解耦权重衰减正则化。arXiv预印本 arXiv:1711.05101，2017年。1，2[51]Loren Lugosch、Mirco Ravanelli、Patrick Ignoto、Vikrant Singh Tomar 和 Yoshua Bengio。面向端到端口语语言理解的语音模型预训练。arXiv预印本 arXiv:1904.03670，2019年。3，4，5[52] Shuang Ma、Zhaoyang Zeng、Daniel McDuff 和 Yale Song。音频-视频表征的主动对比学习。arXiv预印本 arXiv:2009.09805，2020年。2，4[53] Ken MacLean。Voxforge。Ken MacLean。*[在线]*。可访问：http://www. voxforge. org/home。

[2012], 访问日

2018 3, 4, 5
期：。

[54] Sagnik Majumder、Changan Chen、Ziad Al-Halah 和 Kristen Grauman。环境声学的少样本视听学习。arXiv预印本 arXiv:2206.04006，2022年。2[55] Annamaria Mesaros、Toni Heittola 和 Tuomas Virtanen。用于城市声学场景分类的多设备数据集。arXiv预印本 arXiv:1807.09840，2018年。3[56] Ishan Misra 和 Laurens van der Maaten。自监督学习预任务不变表征。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第6707–6717页，2020年。2[57] Pedro Morgado、Nuno Nvasconcelos、Timothy Langlois 和 Oliver Wang。面向360度视频的空间音频自监督生成。《神经信息处理系统进展》，第31卷，2018年。2[58]Pedro Morgado、Nuno Vasconcelos 和 Ishan Misra。基于跨模态一致性的视听实例判别。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第12475–12486页，2021年。2，4[59] Arsha Nagrani、Joon Son Chung、Weidi Xie 和 Andrew Zisserman。Voxceleb：真实环境中的大规模说话人验证。Computer Speech & Language，60：101027，2020年。2[60] Arsha Nagrani、Chen Sun、David Ross、Rahul Sukthankar、Cordelia Schmid 和 Andrew Zisserman。Speech2action：用于动作识别的跨模态监督。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第10317–10326页，2020年。2[61] Arsha Nagrani、Shan Yang、Anurag Arnab、Aren Jansen、Cordelia Schmid 和 Chen Sun。多模态融合的注意力瓶颈。《神经信息处理系统进展》，第34卷：14200–14213，2021年。2[62] Daisuke Niizumi、Daiki Takeuchi、Yasunori Ohishi、Noboru Harada 和 Kunio Kashino。面向音频的BYOL：通用音频表征的自监督学习。载于《2021年国际联合神经网络会议（IJCNN）》，第 1 8 2021 3 –页。IEEE，。[63] Mehdi Noroozi 和 Paolo Favaro。通过拼图求解进行视觉表征的无监督学习。载于《欧洲计算机视觉会议》，第69–84页。施普林格，2016年。2[64] Hyun Oh Song、Yu Xiang、Stefanie Jegelka 和 Silvio Savarese。通过提升式结构化特征嵌入实现深度度量学习。载于《IEEE计算机视觉与模式识别会议论文集》，第4004–4012页，2016年 4 – 。[65] Andrew Owens 和 Alexei A Efros。利用自监督多感官特征进行视听场景分析。载于《欧洲计算机视觉会议（ECCV）论文集》，第631–648页，2018年。2[66] Andrew Owens、Phillip Isola、Josh McDermott、Antonio Torralba、Edward H Adelson 和 William T Freeman。视觉指示的声音。载于《IEEE计算机视觉与模式识别会议论文集》，第2405–2413页，2016年。2[67] Andrew Owens、Jiajun Wu、Josh H McDermott、William T Freeman 和 Antonio Torralba。环境声音提供用于视觉学习的监督。在欧洲计算机视觉会议论文集，第801–816页。施普林格，2016年。2[68] Mandela Patrick、Yuki M Asano、Polina Kuznetsova、Ruth Fong、Jo˜ao F Henriques、Geoffrey Zweig和Andrea Vedaldi。对比自监督学习中的变换组合。载于IEEE/CVF国际计算机视觉会议论文集，第9577–9587页，2021年。2[69] Karol J Piczak。使用卷积神经网络进行环境声音分类。载于2015年IEEE第25届信号处理机器学习国际研讨会（MLSP），第1–6页。IEEE，2015年。5[70] Karol J Piczak。ESC：用于环境声音分类的数据集。载于第23届ACM多媒体国际会议论文集，第1015–1018页，2015年。3[71] 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论文集，第8748–8763页。PMLR，2021年。2[72] Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey和Ilya Sutskever。通过大规模弱监督实现稳健的语音识别。载于国际机器学习会议论文集，第28492–28518页。PMLR，2023年。8[73] Bertrand Rivet、Wenwu Wang、Syed Mohsen Naqvi和Jonathon A Chambers。视听语音源分离：关键方法概述。《IEEE信号处理杂志》，31(3)：125–134，2014年。2[74] Joshua Robinson、Ching-Yao Chuang、Suvrit Sra和Stefanie Jegelka。使用困难负样本进行对比学习。arXiv预印本 arXiv:2010.04592，2020年。4[75] Aaqib Saeed、David Grangier和Neil Zeghidour。通用音频表征的对比学习。载于ICASSP 2021-2021年IEEE国际声学、语音与信号处理会议（ICASSP）论文集，第3875–3879页。IEEE，2021年。3[76] Sepehr Sameni、Simon Jenni和Paolo Favaro。用于视频表征学习的时空裁剪聚合。载于IEEE/CVF国际计算机视觉会议论文集，第5664–5674页，2023年。6[77] Nikunj Saunshi、Orestis Plevrakis、Sanjeev Arora、Mikhail Khodak和Hrishikesh Khandeparkar。对比无监督表征学习的理论分析。载于国际机器学习会议论文集，第5628–5637页。PMLR，2019年。4[78] Ben Shneiderman。语音识别的局限。《ACM通讯》，43(9)：63–65，2000年。3[79] Nikhil Singh、Jeff Mentch、Jerry Ng、Matthew Beveridge和Iddo Drori。Image2reverb：跨模态混响脉冲响应合成。载于IEEE/CVF国际计算机视觉会议论文集，第286 295页，2021年 2 – 。[80] Khurram Soomro、Amir Roshan Zamir和Mubarak Shah。UCF101：一个包含101类野外视频人类动作的数据集。arXiv预印本 arXiv:1212.0402，2012年。1、2、3、5

[81] Christian J Steinmetz和Joshua Reiss。pyloudnorm：一个简单而灵活的Python响度计。载于第150届音频工程学会大会。音频工程学会，2021年。8[82] Fabian-Robert St¨oter、Soumitro Chakrabarty、Emanu¨el Ha-bets和Bernd Edler。Libricount：一个用于说话人数估计的数据集，2018年。5、3[83] Kun Su、Xiulong Liu和Eli Shlizerman。Audeo：为无声表演视频生成音频。神经信息处理系统进展，33 3337，2020年3325 2：–。[84]Makarand Tapaswi、Yukun Zhu、Rainer Stiefelhagen、Antonio Torralba、Raquel Urtasun和Sanja Fidler。MovieQA：通过问答理解电影故事。载于IEEE计算机视觉与模式识别会议论文集，第4631–4640页，2016年。3[85] Yonglong Tian、Dilip Krishnan和Phillip Isola。对比多视图编码。载于欧洲计算机视觉会议论文集，第776–794页。施普林格，2020年。2[86] Michael Tschannen、Josip Djolonga、Paul K Rubenstein、Sylvain Gelly和Mario Lucic。关于表征学习中的互信息最大化。载于国际学习表征会议，2019年。2[87] Joseph Turian、Jordie Shier、Humair Raj Khan、Bhiksha Raj、Bj¨orn W Schuller、Christian J Steinmetz、Colin Malloy、George Tzanetakis、Gissel Velarde、Kirk McNally等。HEAR 2021：音频表征的全面评估。arXiv预印本 arXiv:2203.03022，2022年。1、3、5、6、7、8[88] George Tzanetakis。GTZAN音乐/语音数据集，1999年。5、3[89] Efthymios Tzinis、Scott Wisdom、Aren Jansen、Shawn Hershey、Tal Remez、Daniel PW Ellis和John R Hershey。走进真实场景：使用Audioscope无监督分离视听屏幕内声音。arXiv预印本 arXiv:2011.01143，2020年。2[90] J¨orgen Valk和Tanel Alum¨ae。VoxLingua107：用于口语语言识别的数据集。载于2021年IEEE口语技术研讨会（SLT）论文集，第652–658页，2021年 5、7、4 – 。[91] Luyu Wang、Pauline Luc、Adria Recasens、Jean-Baptiste Alayrac和Aaron van den Oord。通用音频表征的多模态自监督学习。arXiv预印本 arXiv:2104.12807，2021年。2[92] Luyu Wang、Pauline Luc、Yan Wu、Adria Recasens、Lucas Smaira、Andrew Brock、Andrew Jaegle、Jean-Baptiste Alayrac、Sander Dieleman、Joao Carreira等。迈向通用音频表征学习。载于ICASSP 2022-2022年IEEE国际声学、语音与信号处理会议（ICASSP）论文集，第4593–4597页。IEEE，2022年。1、3[93] Pete Warden。Speech Commands：用于有限词汇语音识别的数据集。arXiv预印本 arXiv:1804.03209，2018年。5、7[94] Chao-Yuan Wu和Philipp Kr¨ahenb¨uhl。迈向长时段视频理解。载于计算机视觉与模式识别会议，2021年。2、5、6、7、8、4

[95] Tung-Yu Wu、Tsu-Yuan Hsu、Chen-An Li、Tzu-Han Lin 和 Hung-yi Lee。《自监督语音模型用于音频表征的有效性》。载于 *HEAR*：音频表征的整体评估，第 90–110 页。PMLR，2022年。 6[96] Yusong Wu、Ke Chen、Tianyu Zhang、Yuchen Hui、Taylor Berg-Kirkpatrick 和 Shlomo Dubnov。《通过特征融合和关键词到描述增强进行大规模对比式语言-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。 2[97] Fanyi Xiao、Yong Jae Lee、Kristen Grauman、Jitendra Malik 和 Christoph Feichtenhofer。《用于视频识别的视听慢快网络》。*arXiv* 预印本 *arXiv:2001.08740*，2020 年。 2[98] Xudong Xu、Hang Zhou、Ziwei Liu、Bo Dai、Xiaogang Wang 和 Dahua Lin。《无需双耳音频的视觉信息引导双耳音频生成》。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[99]Xuenan Xu、Zhiling Zhang、Zelin Zhou、Pingyue Zhang、Zeyu Xie、Mengyue Wu 和 Kenny Q Zhu。《BLAT：基于 AudioSet 标签引导合成数据的语言-音频预训练引导方法》。*arXiv* 预印本 *arXiv:2303.07902*，2023 年。 2[100] 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[101] Richard Zhang 、 Phillip Isola 和 Alexei A Efros。《多彩的图像着色》。载于 欧洲计算机视觉会议，第 649–666 页。施普林格，2016年。 2[102] Minghao Zhu、Xiao Lin、Ronghao Dang、Chengju Liu 和 Qijun Chen。《面向对比式视频表征学习的细粒度时空运动对齐》。载于 第 *31* 届 *ACM* 多媒体国际会议论文集，第 4725–4736 页，2023年。 2

附录

摘要

在本附录中，我们介绍预训练流程的相关信息以及补充实验的结果，并展示合成数据示例。本文档的结构如下所示。

目录

A. 预训练细节 *1*

B. 额外实验*B.1.*动作识别结果.*B.2. VGGSound* 结果. *1*
B.3. 受控数据集和模型.*B.3.1* 评估.*B.3.2* 结果.*B.3.3* 探索权衡. 1
1
2
3
3
4

C. 合成反事实对示例 *4*

A. 预训练详情

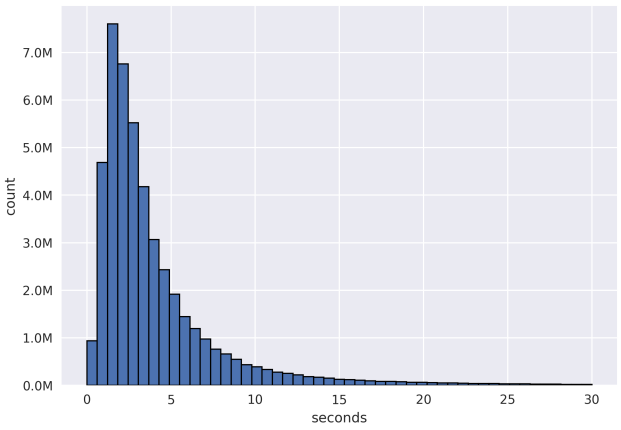

图 A.1。我们数据集中观测到的镜头长度分布。

数据预处理 我们将长篇内容按时间切分为镜头（摄像机切换）。图 A.1 展示了镜头时长的分布。我们忽略短于 3 秒和长于 12 秒的镜头。前一项限制是为了确保片段足够长，适用于我们的模型；后一项限制则是为了提高训练吞吐量。每种预训练设置下的镜头总数见表 1 中的 #data 列。在预训练期间创建小批次时，我们确保每个批次中有1 8 的样本来自同一个长篇内容来源（例如，同一部电影），以创建困难负样本。生成四元组训练实例（v_p, a_p, v_s, a_s）的过程如下：

1.给定一个标题，随机选择一个镜头。2.时间抖动：随机选择两个3 秒的时间窗口。这两个源自同一镜头的片段是我们的主要实例和次要实例。对于次要实例，如果有备用音频轨道（即配音），则其音频的语言与主要实例中的语言不同。

3. 对于每一对音频和视频：**视频**：重采样至 25 fps，均匀采样 16帧，将视频较短边随机缩放至 256-320范围内，然后随机裁剪为224x224。**音频**：重采样至 48kHz，将音频转换为梅尔频谱图（nfft=1024、hoplength=501、num mels=96），转换为分贝刻度，并应用时间和频率遮罩，-其最大值为相应轴的 50%。

模型和预训练超参数 多层感知机投影头的输出维度为 512。潜在嵌入（z）在计算损失之前进行 L2 归一化。目标函数中的温度因子 τ 设为 0.07。我们使用 AdamW 优化器 [50]，学习率为 3e-4，权重衰减为 5e-2。我们在 32 个 NVIDIA A100GPU 上训练12 个 epoch，每个GPU 的批量大小为 64，并采用半余弦学习率退火策略，该策略在 2 个预热 epoch 后启动。

B. 其他实验

B.1. 动作识别结果

我们报告了在 **UCF101** [80]和 **HMDB51** [47],这两个知名基准测试上的结果，以评估我们模型的视频专属性能，如表 B.1所示。我们不同模型变体的性能相当，表明配音增强训练不一定会降低视频专属性能。此外，我们还与近期最先进结果进行了比较；与我们一样，这些结果也未使用微调。请注意，这些结果使用线性探测器，而我们使用的多层感知机探测器则是通过对探测策略进行网格搜索得出的。尽管如此，我们在不进行微调的情况下显著优于这些结果（>12% 绝对值），这证明了我们所学习表征的价值。

B.2. VGGSound 结果

我们报告了在 **VGGSound** [13]，这一视听基准测试上的结果，其中我们重点关注音频结果，如表 B.2所示。再次说明，我们的模型表现

模型	UCF101 [80]	HMDB51 [47]
B.3	88.90	69.35
B.4	88.20	68.91
B.5	87.99	69.43
FIMA [102]	76.40	47.30
FAME [21]	72.20	42.20

表 B.1。视频模型在 UCF101 [80] 和 HMDB51 [47] 数据集上的性能，与近期结果进行比较，这些结果不 涉及微调。

模型	VGGSound
B.3	43.49
B.4	41.95
B.5	42.96
LAION-CLAP [96]	46.20
BLAT [99]	42.90

表 B.2. 音频模型在 VGGSound [13]数据集上的性能，与近期未在下游数据集上进行微调的结果进行比较。
LAION-CLAP 报告的结果使用了关键词到描述增强。

各变体具有可比性，并且我们的结果与近期不使用微调的最先进结果相比具有竞争力。

B.3. 受控数据集与模型

本节讨论一组规模较小、控制更严格的实验所采用的方法及其结果。预训练数据集由 748 部电影组成，内容约为 1300 小时的视频。每部电影都包含一条视频轨道和四条音轨：英语（**英语**）为主要语言，另有三个配音版本：西班牙语（**西班牙语**）、法语（**法语**）和日语（**JA**）；这些语言的配音都相对常见。提供多种配音选项，使我们能够研究次要语言之间的权衡，以及“多语言”模型是否能进一步提升性能。

视频模型是中型 X3D [26],，这是一种高效的基于 ResNet 的模型。我们的音频模型是 Acoustic ResNet 50 [97],，以音频频谱图作为输入。两个模型都会为每个片段输出 1024 维表征。对于使用主要语言和次要语言（配音语言）的音频变体，我们共享骨干网络权重（即 Acoustic ResNet50）。对于主要语言音频和次要语言音频，我们不共享多层感知机权重，以获得更大的灵活性。与我们的主要实验一样，我们主要以跨模态方式训练这些模型，即计算模态之间的对比代价。

我们使用 4 块 A100 GPU 训练这些模型 10 个 epoch，每块 GPU 的批量大小为 26。我们使用负采样参数 k（从与正样本片段相同的电影中抽取样本），并将其设为每块 GPU 12 个。我们使用 AdamW 优化器 [50]，其 β= 为 (0.9, 0.999)，学习率为 0.001，权重衰减为 0.05，并采用带有半个 epoch 预热的余弦学习率调度。

总的来说，在这些规模较小、控制更严格的实验中，我们比较以下模型变体：

1. 单语言 （**英语**）：在这一基线中，我们将使用两种不同增强方式训练的主要语言（英语）音频分别视为“主要”和“次要”（a_p=**英语**； a_s=**英语**）音频。这样做是为了考虑每个已见样本进行两次增强所可能产生的影响，这种情况会出现在配音增强案例中，尽管它不会改变数据分布。这是一个基于 SimCLR 的设置，其中两条音频路径分别与视频进行对比。

2. 双语 **(西班牙语、法语、JA)**：我们每次引入一种辅助音频，以探索配音增强训练假设（a_p=**英语**； a_s= **西班牙语** 或**法语** 或 **JA**）。

3. 多语言 **(+EFJ)**：在这里，我们针对每个批次从给定列表（西班牙语、法语和日语）中随机选择一种辅助音频（a_p= **英语**； $a_s \in_R$ {**西班牙语**、**法语**、**JA**}）。样本顺序是随机的，因此实际上我们只是按轮询方式依次遍历列表。我们旨在探究使用多种辅助音频是否会带来额外益处或弊端。

4. 无语音 **(SEP)**：我们建立了另一个基线，在该设置下，语音经过分离后，我们仅使用视频 + 非语音音频进行训练。这使我们能够考察，在侧重非语音的任务中，单纯移除语音是否足以提升性能。我们使用预训练的 Hybrid Demucs v3 模型 [19] 将人声与其余部分分离，再将其他音轨混合在一起。这里没有辅助音频（a_p=**英语** SEP）。请注意，该变体使用 44.1千赫兹音频进行训练，因为这是 Demucs 模型的输入和输出采样率。虽然 Demucs 是针对音乐分离任务训练的，但我们发现，在实际应用中，它在我们的数据集上处理语音的效果很好。我们使用默认的预训练模型（mdx extra q）。

5.纯音频 (单语言： **音频**，以及多语言：**音频**$_{+EFJ}$）：最后，我们考察两种纯音频模型。其数据与单语言和多语言设置类似，但不包含视频。此时的目标函数是模态内的，即在两个音频片段之间进行。单语言版本采用两个增强副本进行标准的音频对比训练。这些模型无法用于视觉或视听任务，但在此我们旨在评估在没有视频的情况下，配音增强训练是否以及在多大程度上能够带来改进。

	Task	M	基线 (SimCLR) AV	SEP	A	配音增强 ES	FR	JA	EFJ	A_{+EFJ}
Snd/Scn	ESC-50 [70]	A	.527±.012	.570±.028	◦ 220±.027	◦ 580±.019	◦ 575±.031	◦ **590±.036**	◦ 587±.009	◦ 550±.026
	FSD50K [27]	A	.296	.307	.109	**.317**	.313	.311	.313	.277
	TUT18 [55]	A	.853	.857	.682	**.884**	.881	.849	.867	.801
	声音模仿 [43]	A	.042	.051	.022	.045	.047	.045	.050±.006	**.055**
	VGGSound [13]	AV	.303	.287	—	**.323**	.314	.314	.311	—
NonSem	CREMA-D [8]	A	.514±.012	.489±.009	.354±.022	.528±.009	.540	.520±.010	**.548±.012**	.530±.011
	GTZAN 音乐/语音 [88]	A	.954±.054	.931±.099	.866±.119	.946±.082	.891±.142	.931±.092	**.969±.054**	.954±.054
	LibriCount [82]	A	.654±.026	◦ 608±.016	◦ 505±.014	◦ 671±.025	◦ **706±.021**	◦ 681±.016	◦ 676±.013	.678±.022

表 B.3. **对照实验评估结果。** 除 FSD50K 外，所有指标均为 top-1准确率 [27] 和声音模仿 [43] (平均精度均值)。 **粗体** 表示最高分， 灰色 表示最低分。任务类型包括 **声音/场景** = 分类和**非语义** = 非语义语音。

B.3.1 评估

评估任务 HEAR 之外 [87] 除了主要实验中使用的任务外，我们还将其他音频任务的结果纳入这一受控设置，以便更全面地了解受控设置。首先，我们加入 HARES [92]中的音频任务；具体而言，包括用于声学场景识别的 TUT 18 [55] 、用于语音命令识别的 Fluent Speech Commands [51] ，以及用于语言识别的 VoxForge [53] ，以补充现有的 HEAR 任务。与主要结果的附录一样，我们还纳入了仅视频动作识别任务 HMDB51 [47] 和 UCF101 [80]。最后，我们加入一项视听任务（ VG-GSound [13]），以便更好地与**SEP**进行比较，因为该基线完全不涉及语音。我们推测，**SEP** 在某些情况下会表现出色，但总体而言，配音增强模型将更胜一筹，因为它们保留了视觉上出现并发出声音的语音动作之间的视听关系。

对于视觉和视听任务，我们使用随机梯度下降训练线性探测器 200 个周期，并按照余弦调度将学习率设为 0.2。我们使用 2 块 A10 GPU 进行训练，总批量大小为 1024。对于 HEAR 任务，我们采用所提供 API 的策略和 48kHz 数据。对于 HARES 任务，我们遵循作者的规 [92] 定：通常训练 400K 步，学习率调度包括 5K 步线性预热，之后余弦衰减（最大学习率为 0.0002，使用 Adam [44] 优化器）。我们使用 2 块 GPU 进行训练，总批量大小为 64。在所有相关情况下，我们都将单声道音频复制到第二个声道，以构造伪立体声输入，使其与模型架构相匹配。

B.3.2 结果

我们总共训练了 8 种不同的模型变体，并在 15 项不同任务上对它们进行了评估。表B.3 展示了我们假设性能会有所提升的主要任务（N=8），按模态和任务类型分组。

配音增强预训练有帮助吗? 对于表B.3中的所有任务，一个或多个配音增强模型的表现都优于单语言**英语**模型。在 8 项任务中的 6 项里，所有配音增强变体的表现都优于**英语**，唯独两个最简单的任务（TUT18 和 GTZAN）例外。我们对声音与场景分类任务以及非语义语音任务做出了这一结果的假设；在这些任务中，我们始终观察到显著提升。这与论文正文中的结果一致。

性能提升是否仅仅源于降低语音的权重? 我们考察了源分离版本来回答这个问题，因为它代表了语音被完全（尽可能）移除这一极端情况。尽管源分离变体在声音与场景分类任务上提供了强有力的基线，但其表现仍不及一个或多个配音增强模型。我们认为这是因为模型重新聚焦于非语音元素所致。不过，尽管该变体在这些任务中表现出色，但也存在缺点。首先，在 VGGSound（视听分类）和两项视觉任务上的表现低于所有其他模型（见表B.4中的权衡结果）。我们怀疑这是因为源分离版本中的听觉通道与视觉通道之间存在明显差异，即语音。当人物说话时，如果视觉内容所对应的听觉流中几乎没有或完全没有语音内容，这可能会干扰两种表征的协调。请注意，人物是 VGGSound 中的一个大类[6]。

其次，**SEP** 在非语义语音任务以及（表B.4）语言识别任务上的表现显著较差，GTZAN 除外；我们发现，GTZAN 通常是一项更简单的任务。从直觉上看，这很合理：实际上，这个变体并未接触语音，其表现也低于单语言变体。这些结果表明

[6]www.robots.ox.ac.uk/ vgg/数据/vggsound

这是一种权衡：将源分离作为预处理方法，除了计算成本极高、会削弱自监督假设（因为依赖第三方监督模型）之外，还会导致副语言任务表现不佳；这类任务需要关注语言之外的语音特征。

语言越多越好吗？ 鉴于配音增强训练的优势，我们想知道在训练中加入更多语言是否能进一步提升性能。结果并未表明情况如此，但请注意，在表B.3中，**EFJ** 模型最少成为性能最低的配音增强变体（8 项任务中仅有 1 项）。此外，多语言变体在 3 项非语义语音任务中的 2 项上表现良好。尽管副语言特征可能因语言而异，但其中仍存在一些共性，或许具有实用价值；许多实际场景也能从多样化示例中获益。多语言模型的稳健性表明，在对下游任务了解甚少的情况下，它可以作为合理的默认选择；因此，我们在论文正文中规模更大的实验中采用了类似的多语言方法。

即使没有视频，配音增强是否仍有益？ 该A+EFJ 变体始终优于**AUD** 模型（包括我们稍后为权衡分析而考察的所有音频任务，结果见表B.4）。**AUD** 在所有相关任务上的表现都是最弱的，这表明跨模态训练具有优势。此外，在某些任务上，多语言变体的表现接近甚至超过跨模态变体（例如在人声模仿任务上）。当然，这一变体无法用于视觉或多模态任务，而且整体表现仍远逊于多模态配音增强模型，但它表明，即使是单模态配音增强训练也具有显著价值。

B.3.3 探索权衡

表格中的 7 项任务结果B.4 有助于我们评估在较小规模且受控的设置下可能存在的权衡，以补充此前的结果。

配音增强模型仍然能够识别语言吗？ 配音增强变体在 VoxLingua 上的表现通常相近或略差 [90] 但在 VoxForge上的表现似乎更好 [53], 这两项语言识别任务都是如此。后者是一个大规模的用户提交数据集，因此其听觉特征可能与前者不同。综合这些结果，我们认为，配音增强模型能够在其多层感知机前的特征中保留对语言识别有用的信息。更通用的听觉特征可能不会编码语音语义，在这些任务中仍然具有判别性。

它们能否区分口语词语？ 与论文主体部分的结果一致，我们没有观察到语言学任务上的明显性能下降。这表明，经过配音增强的模型所学习到的特征保留了与语音相关的信息，可用于识别词语或命令等任务。不过，源分离模型的特征似乎也有助于完成这些任务，这说明非语音特征以及更通用的声音信号表征可能也有帮助。我们将在下文进一步探讨这一点；结果表明，其中一个数据集（Fluent Speech Commands [51]）中的背景噪声可能为提升性能提供有用信号。

仅视频任务的性能是否受到影响？ 在视觉动作识别任务上，经过配音增强的变体所取得的结果似乎与基线相近。基线在 HMDB51 上略好，在 UCF101 上则略差。这表明，配音增强预训练可能不会显著影响模型在仅视频任务上的整体表现，这与我们的主要模型变体在表格B.1 中呈现的结果相似。

C. 合成反事实对示例

图 C.2展示了从合成生成的 LVU 数据集版本中截取的片段 [94], 我们将其称为 **LVU-M**，正如正文中所述。与图4类似，频谱图展示了同一片段的替代音轨之间的变化和共性。这些示例为任意选取，既展现了与视觉内容的一致性（例如人声、总体时间安排等），也展现了由于伪影、缺乏完整的声学上下文（例如混响）以及当前所提出流程的其他局限性而产生的偏离。我们仅展示这些片段中间的 10 秒，以便于检查。

	Task	M	基线 (SimCLR) AV	SEP	A	配音增强 ES	FR	JA	EFJ	A_{+EFJ}
SemSp	FlSpComm [51]	A	.379	.400	.263	.391	**.410**	.402	.373	.368
	SpComm5h [93]	A	.298	**.372**	.144	.362	.344	.325	.300	.231
	SpCommFull [93]	A	.471	.489	.162	.477	**.537**	.530	.491	.298
Lang	VoxForge [53]	A	.546	.516	.504	.580	.584	**.592**	.571	.543
	VoxLingua10 [90]	A	**.251**$_{\pm.045}$	.226$_{\pm.033}$	.111$_{\pm.012}$	.229$_{\pm.016}$	.237$_{\pm.050}$	.246$_{\pm.032}$	.227$_{\pm.043}$	.201$_{\pm.009}$
Act	HMDB51 [47]	V	**.341**	.319	–	.330	.324	.322	.333	–
	UCF101 [80]	V	.531	.496	–	.540	.523	.538	**.542**	–

表 B.4. **对照实验中的潜在权衡：配音增强是否会对语言或纯视觉任务的表现产生负面影响？**本表中的任务包括**语义语音**（FlSpComm [51], SpComm5h [93], 和 SpCommFull [93]）和**语言识别**（VoxForge [53]和 VoxLingua10 [90]），以及 2 项**动作识别**仅视频任务（HMDB51 [47] 和 UCF101 [80]）。结果各不相同，且往往显示差异相对较小，方向也不尽相同；总体而言，这表明语言相关任务和纯视觉任务的表现并未受到显著影响。

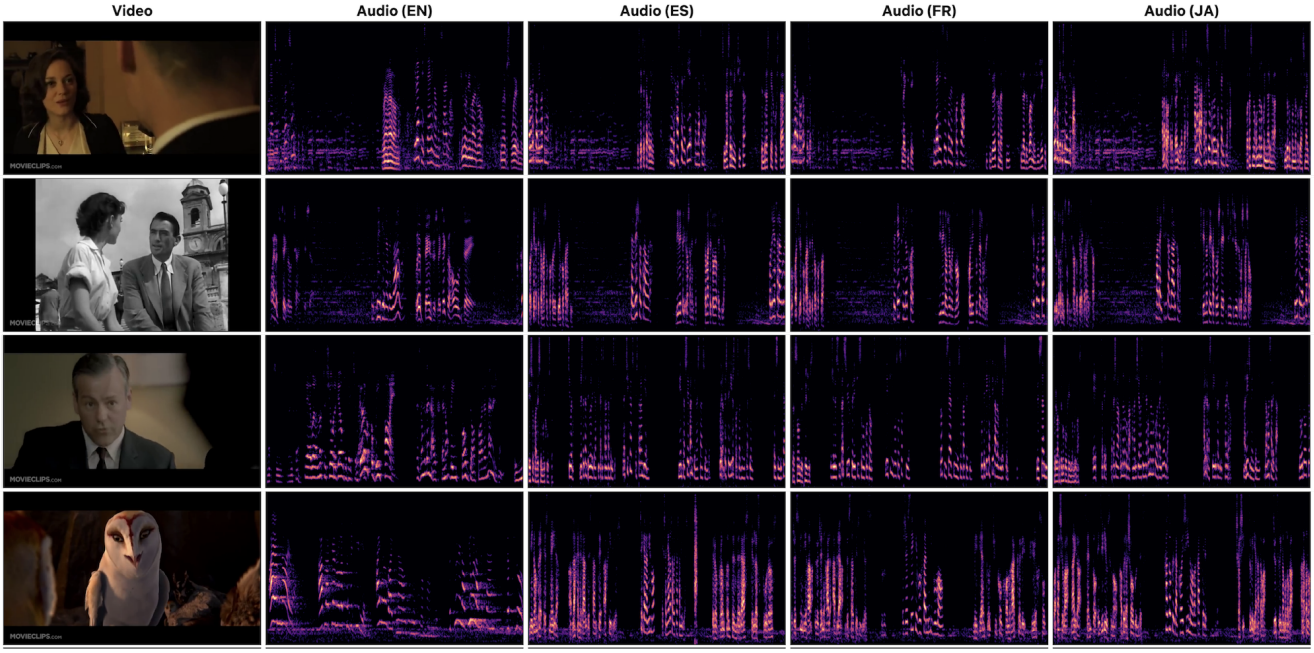

图 C.2. **LVU-M**中的片段示例。