

ImageBind: One Embedding Space To Bind Them All

Rohit Girdhar* Alaaeldin El-Nouby* Zhuang Liu Mannat Singh
Kalyan Vasudev Alwala Armand Joulin Ishan Misra*
FAIR, Meta AI

<https://facebookresearch.github.io/ImageBind>

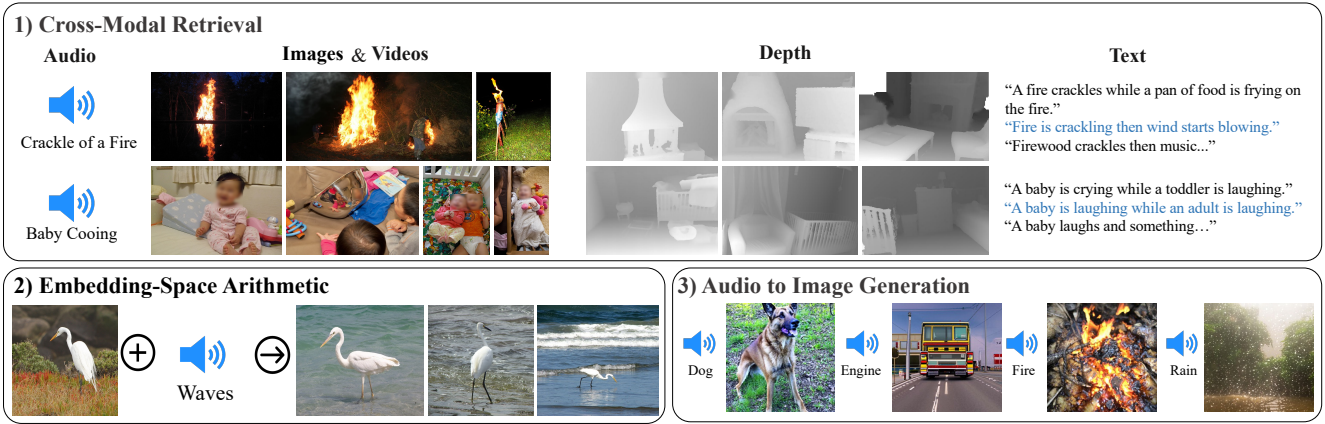


图 1. ImageBind 的联合嵌入空间实现了新颖的多模态能力。通过将六个模态的嵌入对齐到一个共同的空间中，IMAGEBIND 实现了以下功能：1) 跨模态检索，展示了不同模态（如音频、深度或文本）的 *emergent* 对齐，即使它们不常一起观察到。2) 不同模态的嵌入相加自然地组合了它们的语义信息。而且 3) 通过将我们的音频嵌入与预训练的 DALLE-2 [60] 解码器结合使用，实现了音频到图像的生成。该解码器设计用于处理 CLIP 文本嵌入。

Abstract

我们提出了 *IMAGEBIND*，一种学习六种不同模态（图像、文本、音频、深度、热像和 *IMU* 数据）联合嵌入的方法。我们表明，并不需要训练所有配对数据的组合来学习这种联合嵌入，只有图像配对数据就足以将这些模态绑定在一起。*IMAGEBIND* 可以利用最近大规模视觉-语言模型，并通过使用自然图像配对将其零样本能力扩展到新的模态。它能够直接使用“开箱即用”方式实现新颖的应用，包括跨模态检索、通过算术组合模态、跨模态检测和生成。这些新颖的功能随着图像编码器的强度而提升，我们在跨模态的新颖零样本识别任务上取得了新的最优结果，优于专业的监督模型。最后，我们展示了优于先前工作的强少样本识别结果，并证明 *IMAGEBIND* 可作为一种评估视觉模型在视觉

和非视觉任务中的新方法。

1. Introduction

一张图片可以将许多经历联系在一起——海滩的图片可以让我们想起波浪的声音、沙子的质地、微风，甚至能够激发一首诗歌的灵感。

图片的这种“联系”特性为学习视觉特征提供了许多监督来源，通过将它们与与图片相关的任何感官经历进行对齐。理想情况下，对于一个单一的联合嵌入空间，应该通过与所有这些感官对齐来学习视觉特征。然而，这需要获取所有类型和组合的配对数据，这对于相同的一组图片来说是不可行的。

最近，许多方法学习了与文本 [1, 30, 45, 59, 63, 80, 81]、音频 [3, 4, 49, 54, 55, 68] 等对齐的图像特征。

这些方法使用单一配对的模态或者最多几个视觉模态。然而，最终的嵌入只局限于训练所使用的模态对。因此，视频-音频嵌入不能直接用于图像-文本任务，反之亦然。学习真正的联合嵌入的一个主要障碍是缺乏所有模态同时存在的大量多模态数据。

在本文中，我们提出了 **ImageBind**，它通过利用多种类型的图像配对数据学习一个单一的共享表示空间。它不需要包含所有模态同时出现的数据集。相反，我们利用图像的“联系”特性，并且我们展示了仅仅将每个模态的嵌入与图像嵌入对齐就能够在所有模态之间形成出现的对齐。在实践中，IMAGEBIND 利用了规模庞大的（图像、文本）配对数据，并将其与自然配对的数据（如（视频、音频），（图像、深度）等）相结合，以学习一个单一的联合嵌入空间。这使得 IMAGEBIND 能够将文本嵌入与其他模态（如音频、深度等）进行隐式对齐，从而实现了对该模态的零样本识别能力，而无需明确的语义或文本配对。此外，我们展示了它可以以大规模的视觉-语言模型（如 CLIP [59]）进行初始化，从而利用这些模型的丰富的图像和文本表示。因此，IMAGEBIND 可以应用于各种不同的模态和任务，并且只需要很少的训练。

我们使用大规模的图像-文本配对数据以及自然配对的“自监督”数据，包括四种新的模态：音频、深度、热像和惯性测量单元（IMU）读数，并展示了在每个模态的任务上出色的零样本分类和检索性能。这些出现的特性随着底层图像表示的增强而改善。在音频分类和检索基准测试中，IMAGEBIND 的出现零样本分类能力与直接使用音频-文本监督训练的专门模型在 ESC、Clotho、AudioCaps 等基准测试中相当或优于其性能。在少样本评估基准测试中，IMAGEBIND 的表示还优于专门的监督模型。

最后，我们展示了 IMAGEBIND 的联合嵌入可以用于各种组合任务，如图中所示，包括跨模态检索、通过算术组合嵌入、检测图像中的音频源以及根据音频输入生成图像。

Regenerate response

2. Related Work

IMAGEBIND 基于视觉-语言、多模态和自监督研究的几个进展构建而成。

语言图像预训练。将图像与词语或句子等语言信号

一起训练已被证明是一种有效的方法，用于零样本、开放词汇的识别和文本到图像的检索 [13, 17, 37, 66]。语言作为监督信号还可用于学习强大的视频表示 [2, 46, 47]。Joulin 等人 [33] 展示了使用带有嘈杂字幕的大规模图像数据集可以产生强大的视觉特征。最近，CLIP [59]、ALIGN [30] 和 Florence [81] 等模型收集了大量的图像和文本配对数据，并使用对比学习的方式训练模型，将图像和语言输入嵌入到一个共享空间中，展现了令人印象深刻的零样本性能。CoCa [80] 在对比损失之上添加了图像字幕生成的目标，以改善性能。Flamingo [1] 处理任意交错的图像和文本，并在许多样本学习基准测试中取得了最先进的结果。LiT [82] 采用对比训练进行微调，并观察到冻结图像编码器效果最佳。这些先前的工作主要考虑了图像和文本，而我们的工作使得在多个模态上实现了零样本识别的能力。

多模态学习。我们的工作将多个模态的表示绑定在一个联合嵌入空间中。之前的研究在监督 [20, 41] 或自监督的情境中探索了多个模态的联合训练 [3, 19, 49, 68, 72]。图像和语言预训练方法（如 CLIP）的成功启发了一些方法，通过将其他模态与语言输入进行匹配，重新学习深层语义表示。各种方法将 CLIP 进行了改进，以提取语义强大的视频表示 [14, 42, 44, 77]。与我们的方法最相关的是，Nagrani 等人 [50] 创建了一个弱标注的视频-音频和字幕配对数据集，可以训练多模态的视频-音频编码器来匹配文本特征，从而实现强大的音频和视频检索以及字幕生成性能。AudioCLIP [26] 在 CLIP 框架中添加了音频作为额外的模态，实现了零样本音频分类。相比之下，IMAGEBIND 不需要所有模态之间的显式配对数据，而是利用图像作为自然的弱监督来统一多个模态。

特征对齐由于其强大的视觉表示能力，预训练的 CLIP 模型已被用作监督其他模型的教师模型 [43, 57, 73]。此外，CLIP 联合图像和文本嵌入空间也被用于各种零样本任务，如检测 [23, 86]、分割 [40]、网格动画 [79] 等，展示了联合嵌入空间的强大能力。PointCLIP [83] 发现预训练的 CLIP 编码器可以通过将点云投影到多个 2D 深度图视图，并使用 CLIP 视觉编码器对其进行编码，用于 3D 识别。在多语言神经机器翻译中，通常观察到并利用了与 IMAGEBIND 的出现行为类似的现象：如果通过学习到的隐式连接将语言训练在相同的潜在空间中，可以在没有提供配对数据的语言对之间进

行翻译 [32, 39]。

3. Method

我们的目标是通过使用图像将所有模态绑定在一起，学习一个单一的联合嵌入空间。我们将每个模态的嵌入与图像嵌入进行对齐，例如使用网络数据将文本对齐到图像，使用带有 IMU 的第一人称摄像头拍摄的视频数据将 IMU 对齐到视频。我们展示了所得到的嵌入空间具有强大的出现零样本行为，能够自动关联模态对，而无需看到特定模态对的任何训练数据。我们在 Figure 2 中说明了我们的方法。

3.1. 初步

对齐特定模态对。对比学习 [27] 是一种通过使用相关样本对（正样本）和不相关样本对（负样本）来学习嵌入空间的通用技术。使用对齐的观测对，对比学习可以对齐诸如（图像，文本）[59]、（音频，文本）[26]、（图像，深度）[68]、（视频，音频）[49] 等模态对。然而，在每种情况下，联合嵌入是使用相同的模态对进行训练和评估的。因此，（视频，音频）嵌入不能直接应用于基于文本的任务，而（图像，文本）嵌入也不能应用于音频任务。

使用文本提示进行零样本图像分类。CLIP [59] 推广了一种基于对齐的（图像，文本）嵌入空间的“零样本”分类任务。这涉及构建一个描述数据集中类别的文本描述列表。根据输入图像与嵌入空间中的文本描述的相似度对图像进行分类。要对其他模态解锁这种零样本分类，需要特定地使用配对的文本数据进行训练，例如（音频，文本）[26] 或（点云，文本）[83]。相比之下，IMAGEBIND 可以在没有配对文本数据的情况下实现模态的零样本分类。

3.2. 将模态与图像绑定

IMAGEBIND 使用模态对 $(\mathcal{I}, \mathcal{M})$ ，其中 \mathcal{I} 表示图像， \mathcal{M} 表示另一个模态，来学习一个单一的联合嵌入。我们使用大规模的网络数据集，其中包含（图像，文本）配对，涵盖了广泛的语义概念。此外，我们利用其他模态（音频、深度、热像和惯性测量单元（IMU））与图像的自然自监督配对。

考虑具有对齐观测的模态对 $(\mathcal{I}, \mathcal{M})$ 。给定图像 \mathbf{I}_i 及其在另一个模态中的相应观测 \mathbf{M}_i ，我们将它们编码

为归一化的嵌入： $\mathbf{q}_i = f(\mathbf{I}_i)$ 和 $\mathbf{k}_i = g(\mathbf{M}_i)$ ，其中 f, g 是深度网络。嵌入和编码器使用 InfoNCE [53] 损失进行优化：

$$L_{\mathcal{I}, \mathcal{M}} = -\log \frac{\exp(\mathbf{q}_i^\top \mathbf{k}_i / \tau)}{\exp(\mathbf{q}_i^\top \mathbf{k}_i / \tau) + \sum_{j \neq i} \exp(\mathbf{q}_i^\top \mathbf{k}_j / \tau)}, \quad (1)$$

其中， τ 是一个标量温度，控制 softmax 分布的平滑程度， j 表示不相关的观测，也称为“负样本”。我们遵循 [74] 的方法，将每个 $j \neq i$ 的样本都视为负样本。该损失函数使得嵌入 \mathbf{q}_i 和 \mathbf{k}_i 在联合嵌入空间中更加接近，从而对齐了 \mathcal{I} 和 \mathcal{M} 。实际上，我们使用对称损失 $L_{\mathcal{I}, \mathcal{M}} + L_{\mathcal{M}, \mathcal{I}}$ 。

未见过的模态对的出现对齐。IMAGEBIND 使用与图像配对的模态，即形如 $(\mathcal{I}, \mathcal{M})$ 的模态对，将每个模态 \mathcal{M} 的嵌入对齐到图像的嵌入。我们观察到在嵌入空间中出现了一种新的行为，即使我们只使用了 $(\mathcal{I}, \mathcal{M}_1)$ 和 $(\mathcal{I}, \mathcal{M}_2)$ 这两对模态进行训练，也能够对齐两对模态 $(\mathcal{M}_1, \mathcal{M}_2)$ 。这种行为使我们能够在训练过程中实现各种零样本和跨模态检索任务，而无需专门针对这些任务进行训练。我们在不观察任何配对的（音频，文本）样本的情况下，实现了最先进的零样本文本-音频分类结果。

3.3. 实现细节

IMAGEBIND 的概念很简单，可以以多种不同的方式实现。我们有意选择了一种灵活且易于研究和采用的基本实现方式。在 § 5 中，我们提供了对于良好的出现“绑定”所关键的设计决策。

编码模态。我们对所有模态编码器使用 Transformer 架构 [71]。对于图像，我们使用 Vision Transformer (ViT) [12]。我们按照 [19] 的方法，使用相同的编码器来处理图像和视频。我们对 ViT 的图像块投影层进行时间膨胀 [7]，并从 2 秒的视频中采样出 2 帧的视频剪辑。对于音频，我们按照 [21] 的方法进行编码，将采样频率为 16kHz 的 2 秒音频转换为包含 128 个梅尔频谱图的梅尔频谱图。由于梅尔频谱图也是一种类似图像的二维信号，我们使用具有大小为 16 和步长为 10 的块的 ViT 进行编码。我们将热像和深度图像视为单通道图像，并使用 ViT 进行编码。我们按照 [20] 的方法将深度转换为视差图以实现尺度不变性。我们提取由加速度计和陀螺仪测量值组成的 IMU 信号，涵盖了 X、Y 和 Z 轴。我们使用 5 秒的片段，得到 2K 个时间步的

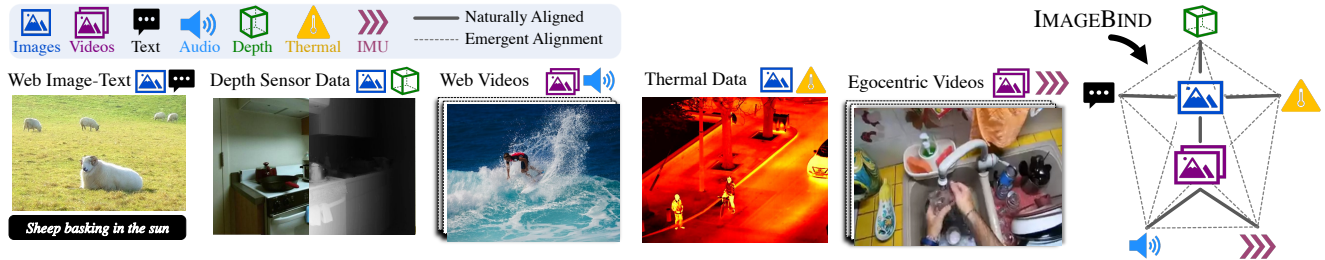


图 2. ImageBind overview. 不同的模态在不同的数据源中自然地对齐，例如图像 + 文本和视频 + 音频在网络数据中，深度或热像与图像配对，以及使用第一人称摄像头拍摄的视频中的 IMU 数据等等。IMAGEBIND 将所有这些模态链接在一个共同的嵌入空间中，实现了新的出现对齐和能力。

Dataset	Task	#cls	Metric	#test
Audioset Audio-only (AS-A) [18]	Audio cls.	527	mAP	19048
ESC 5-folds (ESC) [58]	Audio cls.	50	Acc	400
Clotho (Clotho) [16]	Retrieval	-	Recall	1045
AudioCaps (AudioCaps) [36]	Retrieval	-	Recall	796
VGGSound (VGGs) [8]	Audio cls.	309	Acc	14073
SUN Depth-only (SUN-D) [67]	Scene cls.	19	Acc	4660
NYU-v2 Depth-only (NYU-D) [64]	Scene cls.	10	Acc	653
LLVIP (LLVIP) [31]	Person cls.	2	Acc	15809
Ego4D (Ego4D) [22]	Scenario cls.	108	Acc	68865

表 1. Emergent zero-shot classification datasets for audio, depth, thermal, and Inertial Measurement Unit (IMU) modalities. 我们评估了 IMAGEBIND 在这些任务上的零样本分类能力，且没有对这些模态进行任何训练，也没有使用配对的文本数据进行训练。对于每个数据集，我们报告了任务（分类或检索），类别数目（#cls），评估指标（准确率或均值平均精度），以及测试样本数目（#test）。

IMU 读数，使用卷积核大小为 8 的一维卷积进行投影。得到的序列使用 Transformer 进行编码。最后，我们遵循 CLIP [59] 的文本编码器设计。

我们为图像、文本、音频、热像、深度图像和 IMU 使用单独的编码器。我们在每个编码器上添加一个模态特定的线性投影头，以获得固定大小的 d 维嵌入，对其进行归一化，并在 Eq 1 中的 InfoNCE 损失中使用。除了便于学习外，这种设置还使我们能够使用预训练模型初始化编码器的子集，例如使用 CLIP [59] 或 OpenCLIP [29] 初始化图像和文本编码器。

4. Experiments

首先，我们描述了主要的实验设置，并在补充材料中提供了完整的细节。

自然配对的模态和数据集。我们在六个模态上使用 IMAGEBIND- 图像/视频、文本、音频、深度、热像和 IMU。如 § 3.3 所述，我们将视频视为 2 帧图像，并对其进行与图像相同的处理。对于自然配对的数据，我们使用 Audioset 数据集 [18] 中的 (视频, 音频) 配对、SUN RGB-D 数据集 [67] 中的 (图像, 深度) 配对、LLVIP 数据集 [31] 中的 (图像, 热像) 配对以及 Ego4D 数据集 [22] 中的 (视频, IMU) 配对。对于这些模态的配对，我们不使用任何额外的监督信息，如类别标签、文本等。由于 SUN RGB-D 和 LLVIP 相对较小，我们按照 [20] 的方法，将它们复制 50 倍进行训练。

大规模图像-文本配对。我们利用来自大规模网络数据 [59] 的图像-文本配对信息进行训练。为了方便实验，我们使用在数十亿个图像-文本配对上训练的预训练模型。具体来说，我们在实验中使用了 OpenCLIP [29] 的预训练视觉 (ViT-H, 630M 个参数) 和文本编码器 (302M 个参数)。

每个模态的编码器。我们将音频转换为 2D 梅尔频谱图 [21]，将热像和深度模态转换为 1 通道图像，并分别使用 ViT-B、ViT-S 编码器进行编码。图像和文本编码器在 IMAGEBIND 训练过程中保持冻结，而音频、深度、热像和 IMU 编码器则进行更新。

出现的零样本 vs. 零样本。方法 (例如 CLIP [59]、AudioCLIP [26] 等) 使用模态配对 (图像, 文本) 和 (音频, 文本) 来展示使用文本提示进行零样本分类的能力。相比之下，IMAGEBIND 使用仅图像配对数据将模态绑定在一起。因此，仅通过在 (图像, 文本) 和 (图像, 音频) 上进行训练，IMAGEBIND 就能执行使用文本提示的音频零样本分类。由于我们不直接训练这种能力，我们将其称为出现的零样本分类，以区别于专门







											
	IN1K	P365	K400	MSR-VTT	NYU-D	SUN-D	AS-A	VGGS	ESC	LLVIP	Ego4D
Random	0.1	0.27	0.25	0.1	10.0	5.26	0.62	0.32	2.75	50.0	0.9
IMAGEBIND	77.7	45.4	50.0	36.1	54.0	35.1	17.6	27.8	66.9	63.4	25.0
Text Paired	-	-	-	-	41.9*	25.4*	28.4 [†] [26]	-	68.6 [†] [26]	-	-
Absolute SOTA	91.0 [80]	60.7 [65]	89.9 [78]	57.7 [77]	76.7 [20]	64.9 [20]	49.6 [38]	52.5 [35]	97.0 [9]	-	-

表 2. Emergent zero-shot classification of IMAGEBIND using text prompts **highlighted in blue**. IMAGEBIND 将图像与文本、深度、音频、热图和 IMU 等模态进行了对齐。由此产生的嵌入空间可以将文本嵌入与非图像模态相关联，并导致强大的零样本分类能力。我们在音频和 IMU 等非视觉模态上表现出色。我们尽可能地与“Text Paired”基准进行比较，该基准使用与该模态配对的文本数据进行训练。* 我们使用 OpenCLIP ViT-H [29] 处理的灰度深度图像。[†][26] 在训练期间使用 AS 类别名称作为监督，因此不是“零样本”。总体而言，IMAGEBIND 在零样本分类任务上表现出强大的性能，甚至与这些上限相比也不逊色。我们还报告了每个数据集的绝对最新结果（SOTA）供参考，这些结果通常使用了额外的监督、模型集成等。我们报告了除了 MSR-VTT（Recall@1）和 Audioset Audio-only（mAP）以外的所有数据集的 top-1 分类准确率。

使用配对的文本监督训练的方法。

下游任务的评估。我们在许多不同的下游任务上对 IMAGEBIND 进行了全面评估，并使用不同的协议进行评估。我们在 Table 1 中总结了评估中使用的主要数据集。

4.1. 出现的零样本分类

我们在出现的零样本分类任务上评估 IMAGEBIND，并使用 [59] 中的文本提示模板（完整细节见 Appendix B）。我们在 Table 2 中报告了结果。每个任务衡量了 IMAGEBIND 在没有同时观察到这些模态的情况下，将文本嵌入与其他模态关联的能力。

鉴于我们问题设置的新颖性，没有与 IMAGEBIND 进行直接比较的“公平”的基线。尽管如此，我们将其与使用某些模态（如音频 [26, 50]）配对的先前工作进行了比较，对于某些“类似视觉”的模态，如深度和热像，我们直接使用了 CLIP 模型。我们还报告了每个基准任务的最佳已报告的有监督上界。

IMAGEBIND 在出现的零样本分类任务中取得了较高的性能。在每个基准任务上，IMAGEBIND 都取得了显著的增益，甚至与针对特定模态和任务进行训练的有监督专用模型相比具有相似的性能。这些结果表明，IMAGEBIND 将模态进行了对齐，并将与图像相关的文本监督隐式传递给其他模态，例如音频。特别是，IMAGEBIND 对于音频和 IMU 等非视觉模态显示出强大的对齐能力，这表明它们与图像的自然配对是一个强大的监督来源。为了完整起见，我们还报告了标准的零样本图像（ImageNet [62] - IN1K，Places-365 [85] -

	Emergent	Clotho		AudioCaps		ESC
		R@1	R@10	R@1	R@10	Top-1
<i>Uses audio and text supervision</i>						
AudioCLIP [26]	✗	—	—	—	—	68.6
<i>Uses audio and text loss</i>						
AVFIC [50]	✗	3.0	17.5	8.7	37.7	—
<i>No audio and text supervision</i>						
IMAGEBIND	✓	6.0	28.4	9.3	42.3	66.9
<i>Supervised</i>						
AVFIC finetuned [50]	✗	8.4	38.6	—	—	—
ARNLQ [52]	✗	12.6	45.4	24.3	72.1	—

表 3. Emergent zero-shot audio retrieval and classification. 我们将 IMAGEBIND 与先前的零样本音频检索和音频分类方法进行比较。在不使用音频特定监督的情况下，IMAGEBIND 在零样本检索任务上表现优于先前的方法，并在分类任务上具有可比较的性能。IMAGEBIND 的零样本表现接近专门的有监督模型。

P365）和视频（Kinetics400 [34] - K400，MSR-VTT 1k-A [76] - MSR-VTT）任务。由于图像和文本编码器是使用 OpenCLIP 初始化（并冻结）的，这些结果与 OpenCLIP 的结果一致。

4.2. 与先前工作的比较

现在我们将 IMAGEBIND 与先前的零样本检索和分类任务的工作进行比较。

零样本文本到音频检索和分类。与 IMAGEBIND 不同，先前的工作使用配对数据进行训练，例如，AudioCLIP [26] 使用（音频，文本）监督进行训练，AVFIC [51] 使用自

动挖掘的（音频，文本）配对。我们将它们在文本到音频检索和分类任务中的零样本性能与 IMAGEBIND 的出现检索和分类性能进行比较，结果见 Table 3。

IMAGEBIND 在音频文本检索基准上明显优于先前的工作。在 Clotho 数据集上，与 AVFIC 相比，IMAGEBIND 的性能提升了一倍，尽管在训练过程中没有使用任何音频文本配对。与有监督的 AudioCLIP 模型相比，IMAGEBIND 在 ESC 上实现了可比的音频分类性能。值得注意的是，AudioCLIP 使用 AudioSet 的类别名称作为音频-文本训练的文本目标，因此被称为“有监督”。IMAGEBIND 在所有三个基准测试中展现出了强大的音频和文本模态对齐能力，验证了使用图像作为桥梁的音频和文本特征之间的对齐。

文本到音频和视频检索。我们使用 MSR-VTT 1k-A 基准来评估文本到音频和视频检索性能，结果见 ??。仅使用音频，IMAGEBIND 在出现的检索任务中表现出强大的性能，与先前的 MIL-NCE 等工作的视频检索性能相比。对于我们的模型，文本到视频性能也很强（见 Table 2 中的 36.1% R@1），因为它使用了 OpenCLIP 的视觉和文本编码器，并且优于许多先前的方法。然而，结合音频和视频模态进一步提升了性能，显示了 IMAGEBIND 特征的实用性，即使在已经强大的检索模型上也是如此。

4.3. 少样本分类

我们通过在音频和深度分类任务上评估 IMAGEBIND 的标签效率来评估其少样本分类能力，结果见 ??。我们使用 IMAGEBIND 的音频和深度编码器，在音频和深度分类上进行评估。对于 ≥ 1 -shot 结果，我们按照 [49, 59] 的方法，在固定特征上训练线性分类器（详细信息见 Appendix B）。

在少样本音频分类任务中（左侧的 ??），我们与（1）在 Audioset 音频上训练的自监督 AudioMAE 模型和（2）在音频分类上微调的有监督 AudioMAE 模型进行比较。这两个基线都使用与 IMAGEBIND 相同容量的 ViT-B 音频编码器。在所有设置上，IMAGEBIND 在 ≤ 4 -shot 分类的 top-1 准确率上相对于 AudioMAE 模型获得了约 40% 的准确率增益。在 ≥ 1 -shot 分类上，IMAGEBIND 的性能与有监督模型相媲美甚至更好。IMAGEBIND 的出现零样本性能超过了 ≤ 2 -shot 的有监督性能。

对于少样本深度分类，我们将其与多模态的 MultiMAE [4] 模型进行比较，该模型使用了图像、深度和语义分割数据进行训练。在所有少样本设置上，IMAGEBIND 显著优于 MultiMAE。总的来说，这些结果表明，通过使用图像对齐训练的 IMAGEBIND 音频和深度特征具有强大的泛化能力。

4.4. 分析和应用

多模态嵌入空间算术。我们研究了 IMAGEBIND 的嵌入是否可以用于跨模态进行信息组合。在 ?? 中，我们展示了通过将图像和音频嵌入相加来获得图像检索结果。这种联合嵌入空间允许我们组合两个嵌入：例如，图像中的水果在桌子上 + 鸟鸣声，并检索包含这两个概念的图像，即树上有鸟的水果。这种出现的组合性，即可以从不同模态中组合语义内容，可能会使一系列组合任务成为可能。

将基于文本的检测器升级为基于音频的检测器。我们使用预训练的基于文本的检测模型 Detic [86]，简单地将其基于 CLIP 的“类别”（文本）嵌入替换为 IMAGEBIND 的音频嵌入。在不进行训练的情况下，这就创建了一个基于音频的检测器，可以根据音频提示检测和分割对象。如 ?? 所示，我们可以使用狗吠声提示检测器，从而定位狗。

将基于文本的扩散模型升级为基于音频的模型。我们使用预训练的 DALL-E-2 [60] 扩散模型（私人重新实现），并将其提示嵌入替换为我们的音频嵌入。如 Figure 1 所示，我们观察到我们可以重新调整扩散模型以使用不同类型的声音生成合理的图像。

总之，这些结果显示了 IMAGEBIND 在多模态分析和应用方面的强大能力，允许语义内容从不同的模态中组合，并能够将基于文本的检测器和扩散模型升级为基于音频的模型。

5. Ablation Study

我们研究了学习不同模态的联合嵌入空间的各种设计选择。由于消融实验设置与 § 4 类似，我们只记录主要差异（完整细节请参见 Appendix C）。我们在 ESC 的第一折上报告了消融研究的结果。默认情况下，我们使用 ViT-B 编码器对图像、音频、深度和热像模态进行训练，并在其中训练 16 个时期（与 § 4 的 32 个时期相比）。对于 IMU，我们使用一个轻量级的 6 层编

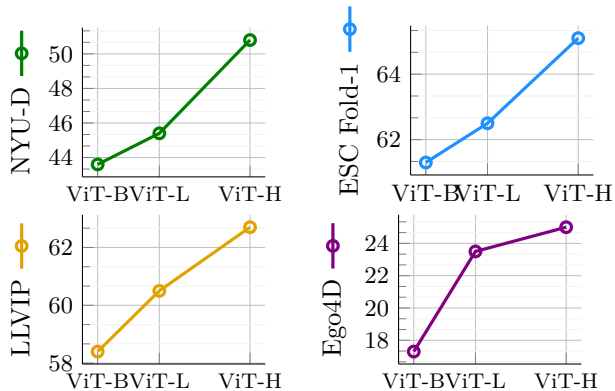


图 3. 调整图像编码器的规模，同时保持其他模态编码器的规模不变。我们评估了深度、音频、热图和 IMU 模态的新兴零样本分类性能。调整图像编码器的规模显著提高了零样本分类结果，表明更强大的视觉表示改善了各模态之间的“绑定”。

码器，宽度为 512 维，8 个头，并训练了 8 个时期。文本编码器遵循 [59] 的设计，是一个具有 512 维度宽度的十二层 Transformer。我们使用 CLIP 模型 [59] 初始化图像和文本编码器。

5.1. 扩展图像编码器

IMAGEBIND 的核心思想是将所有模态的嵌入与图像嵌入对齐。因此，图像嵌入在未见模态的紧密对齐和出现式零样本分类中起着核心作用，我们研究了图像编码器规模对出现式零样本性能的影响。我们改变图像编码器的大小，并训练一个与图像表示匹配的深度、音频等模态的编码器。为了独立评估图像表示的影响，我们固定了其他模态编码器的大小。我们使用预训练的 CLIP (ViT-B 和 ViT-L) 和 OpenCLIP (ViT-H) 图像和文本编码器进行此实验。我们在 Figure 3 中的结果显示，IMAGEBIND 在所有模态上的出现式零样本性能随着视觉特征的改进而提高。对于深度和音频分类，比较强大的 ViT-H 相对于 ViT-B 图像编码器分别提供了 7% 和 4% 的增益。因此，更强大的视觉特征可以改善甚至非视觉模态的识别性能。

5.2. 训练损失和架构

我们研究了训练设计选择对出现式零样本分类的影响。我们重点研究了两种具有不同特征的模态——视觉和空间特性的深度模态，以及非视觉和有时间特性的

音频模态。我们发现研究这些不同模态有助于获得稳健且可转移的设计决策。

对比损失的温度。我们在 Table 4a 中研究了温度 τ (Eq 1) 对深度和音频分类的影响。我们尝试了可学习温度初始化为 0.07 (在对数尺度上参数化)，与使用固定温度的不同数值进行比较。与 [59] 不同，我们观察到固定温度对深度、热像和 IMU 分类效果更好。此外，我们发现较高的温度对于深度、热像和 IMU 编码器的训练效果更好，而较低的温度对于音频模态效果更好。

投影头。我们将每个编码器使用的投影头从线性层变为具有 768 个隐藏维度的 MLP。在 Table 4b 中的结果显示，线性投影头对于两种模态都表现更好。这与标准的自监督方法 (如 SimCLR [10]) 不同，它们的性能会随着 MLP 投影头的改进而提高。

训练时期。我们改变训练的时期数，并在 Table 4c 中报告了分类性能。较长的训练时期一致地提高了两种模态在所有数据集上的出现式零样本性能。

配对图像的数据增强。在 IMAGEBIND 训练期间，我们对图像进行了基本的数据增强 (裁剪、颜色抖动) 或更强的数据增强，其中还包括 RandAugment [11] 和 RandErase [84]。我们在 Appendix C 中指定了数据增强的参数。较强的数据增强有助于处理来自 SUN RGB-D 数据集中少量的 (图像，深度) 配对数据，提高了深度分类的性能。然而，对于音频来说，强烈的数据增强使任务过于困难，导致 ESC 上的性能显著下降了 34%。

深度特定设计选择。我们在 Table 4e 中改变了用于训练的空间裁剪类型。根据 CMC [68] 的方法，我们使用来自相应图像和深度配对的两个不对齐的随机裁剪，与我们默认选择的空间对齐的随机裁剪进行比较。与 CMC 不同的是，我们观察到随机裁剪严重降低了性能，对 SUN-D 的影响超过 10%。与传统的自监督学习不同，我们从图像文本中学到的图像表示更具语义性，因此空间上不对齐的裁剪会对性能产生负面影响。在 Table 4f 中我们观察到，对深度分类使用 RandomErase [84] 可以提升性能。

音频特定设计选择。我们在训练时使用时间上对齐的样本或不对齐的样本进行视频-音频对齐，并在 Table 4g 中衡量最终性能。与深度分类的观察类似，时间上对齐的样本表现出更好的性能。在 Table 4h 中显示，对音频使用频率掩蔽增强也可以略微提高性能。

音频和深度编码器的容量及其对分类性能的影响在 Ta-

Temp →	Learn 0.05 0.07 0.2 1.0	Proj head →	Linear MLP	Epochs →	16 32 64	Data aug →	Basic Strong
SUN-D	24.1 27.0 27.3 26.7 28.0	SUN-D	26.7 26.5	SUN-D	26.7 27.9 29.9	SUN-D	25.4 26.7
ESC	54.8 56.7 52.4 45.4 24.3	ESC	56.7 51.0	ESC	56.7 61.3 62.9	ESC	56.7 22.6
(a) Temperature for loss.		(b) Projection Head.		(c) Training epochs.		(d) Data aug for image.	
Spatial align →	None Aligned	Data aug →	None RandErase	Temporal align →	None Aligned	Data aug →	Basic +Freq mask
SUN-D	16.0 26.7	SUN-D	24.2 26.7	ESC	55.7 56.7	ESC	56.5 56.7
(e) Spatial alignment of depth.		(f) Depth data aug.		(g) Temporal alignment of audio.		(h) Audio data aug.	

表 4. 训练损失和架构设计决策及其对紧急零样本分类的影响。在 § 4 中，灰色突出显示了结果的设置。(a) 对于所有模态，固定温度的对比损失优于可学习的温度。(b) 用于计算深度或音频嵌入的线性投影头比多层感知机投影头效果更好。(c) 较长的训练时间提高了零样本分类性能，对于两种模态都是如此。(d) 更强的图像增强提高了深度分类的性能，而基本增强显著提高了音频分类的性能。(e, f) 在训练 IMAGEBIND 时使用空间对齐的图像和深度裁剪显著提高了性能。同样，RandErase 增强对于深度的良好零样本分类至关重要。(g, h) 时域对齐的音频和视频匹配提供了改进的性能，而对音频进行频率增强略微提高了性能。

Image Encoder	Audio Encoder (ESC)		Depth Encoder (SUN)	
	ViT-S	ViT-B	ViT-S	ViT-B
ViT-B	52.8	56.7	30.7	26.7
ViT-H	54.8	60.3	33.3	29.5

表 5. 音频编码器和深度编码器的容量及其对性能的影响。较强的图像编码器提高了音频和深度任务的性能。由于（图像，深度）对的数量较少，较小的编码器提高了深度的性能。对于音频分类，较大的编码器效果更好。

Batch size →	512	1k	2k	4k
NYU-D	47.3	46.5	43.0	39.9
ESC	39.4	53.9	56.7	53.9

表 6. Effect of scaling batch size. We found the optimal batch size for contrastive loss varied by the modality. For image-depth task, a smaller batch size was better, likely due to the small size and limited diversity of the original dataset. For audio-video task where we have a lot more positive and negative audio-video pairs, using a large batch size lead to better results.

ble 5 中报告。较小的深度编码器可以改善性能，这可能是由于（图像，深度）数据集相对较小的缘故。相反，我们观察到更大的音频编码器可以提高性能，尤其是与高容量的图像编码器配对时。

批量大小的影响。在 Table 6 中，我们评估了批量大小对所学特征表示的影响。如表所示，批量大小可以根据相应的预训练数据集的大小和复杂性在不同模态

	IN1K	VGGs ESC	SUN-D NYU-D
DINO [6]	64.4	17.2 44.7	26.8 48.8
DeiT [70]	74.4 [†]	9.6 25.0	25.2 48.0

表 7. ImageBind 作为评估工具。我们使用不同的方法初始化（并固定）图像编码器，并对其他模态进行对齐。IMAGEBIND 评估了视觉特征对多模态任务的影响。[†] 使用 IN1K 监督进行训练。

之间变化。

使用 ImageBind 评估预训练视觉模型，见 Table 7。我们使用预训练模型来初始化视觉编码器，并将其保持不变。我们使用图像配对数据来对齐和训练文本、音频和深度编码器（详细信息请参见附录 Appendix B）。与有监督的 DeiT 模型相比，自监督的 DINO 模型在深度和音频模态上的紧急零-shot 分类效果更好。此外，紧急零-shot 性能与 ImageNet 上的纯视觉性能并不相关，这表明这些任务衡量的是不同的特性。IMAGEBIND 可以作为评估视觉模型在多模态应用中强度的有价值工具。

6. Discussion and Limitations

IMAGEBIND 是一种简单而实用的方法，只使用图像对齐来训练联合嵌入空间。我们的方法实现了所有模态的紧急对齐，可以通过交叉模态检索和基于文本的零-shot 任务进行衡量。我们实现了跨不同模态的丰富的组合多模态任务，展示了一种评估预训练视觉模型在非视

觉任务上的方法，并可以对模型如 Detic 和 DALLÉ-2 进行音频上的”升级”。进一步改进 IMAGEBIND 的方法有多种途径。我们可以通过使用其他对齐数据来丰富我们的图像对齐损失，例如与其他模态（例如音频与 IMU）配对的文本，或者模态之间的配对。我们的嵌入是在没有特定下游任务的情况下训练的，因此性能落后于专门的模型。进一步研究如何将通用嵌入适应每个任务，包括结构化预测任务（如检测）将是有益的。最后，新的基准数据集（例如我们的紧急零-shot 任务）可以用于衡量多模态模型的紧急能力，从而帮助创建令人兴奋的新应用。我们的模型是一个研究原型，不能直接用于实际应用（请参见附录 Appendix F）。

致谢：作者们感谢 Uriel Singer、Adam Polyak 和 Naman Goyal 对 DALLÉ-2 实验的帮助，以及整个 Meta AI 团队的许多有益的讨论。

References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 1, 2
- [2] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *NeurIPS*, 2020. 2
- [3] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, 2017. 1, 2
- [4] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multima: Multi-modal multi-task masked autoencoders. *arXiv preprint arXiv:2204.01678*, 2022. 1, 6
- [5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. *arXiv preprint arXiv:2104.00650*, 2021.
- [6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 8
- [7] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, 2017. 3
- [8] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, 2020. 4, 14
- [9] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP*, 2022. 5
- [10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 7, 16
- [11] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR*, 2020. 7
- [12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3
- [13] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017. 2
- [14] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 2
- [15] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. In *NeurIPS*, 2022. 15
- [16] Frederic Font, Gerard Roma, and Xavier Serra. Freesound technical demo. In *ACM international conference on Multimedia*, 2013. 4, 14
- [17] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc’Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013. 2
- [18] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 2017. 4, 14

- [19] Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Omnimae: Single model masked pretraining on images and videos. *arXiv preprint arXiv:2206.08356*, 2022. 2, 3
- [20] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A Single Model for Many Visual Modalities. In *CVPR*, 2022. 2, 3, 4, 5, 14
- [21] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Interspeech*, 2021. 3, 4, 15
- [22] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of ego-centric video. In *CVPR*, 2022. 4, 15
- [23] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 2
- [24] Saurabh Gupta, Pablo Arbelaez, and Jitendra Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *CVPR*, 2013. 15
- [25] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from rgb-d images for object detection and segmentation. In *ECCV*, 2014. 15
- [26] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio, 2021. 2, 3, 4, 5
- [27] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006. 3
- [28] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *ECCV*, 2016. 16
- [29] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 4, 5
- [30] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 1, 2
- [31] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *ICCV*, 2021. 4, 14, 15
- [32] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google’s multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017. 3
- [33] Armand Joulin, Laurens van der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from large weakly supervised data. In *European Conference on Computer Vision*, pages 67–84. Springer, 2016. 2
- [34] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, AMustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 5
- [35] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Slow-fast auditory streams for audio recognition. In *ICASSP*, 2021. 5
- [36] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *NAACL*, 2019. 4, 14
- [37] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 2
- [38] Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. In *Interspeech*, 2022. 5
- [39] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*, 2017. 3
- [40] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. 2
- [41] Valerii Likhoshesterov, Anurag Arnab, Krzysztof Choromanski, Mario Lucic, Yi Tay, Adrian Weller, and

- Mostafa Dehghani. Polyvit: Co-training vision transformers on images, videos and audio. *arXiv preprint arXiv:2111.12993*, 2021. 2
- [42] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. In *ECCV*, 2022. 2
- [43] Xingbin Liu, Jinghao Zhou, Tao Kong, Xianming Lin, and Rongrong Ji. Exploring target representations for masked autoencoders. *arXiv preprint arXiv:2209.03917*, 2022. 2
- [44] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. corr abs/2104.08860 (2021). *arXiv preprint arXiv:2104.08860*, 2021. 2
- [45] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018. 1
- [46] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncured instructional videos. In *CVPR*, 2020. 2
- [47] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. *ICCV*, 2019. 2
- [48] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019.
- [49] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *CVPR*, 2021. 1, 2, 3, 6
- [50] Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. Learning audio-video modalities from image captions. In *ECCV*, 2022. 2, 5
- [51] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *NeurIPS*, 2021. 5
- [52] Andreea-Maria Oncescu, A Koepke, Joao F Henriques, Zeynep Akata, and Samuel Albanie. Audio retrieval with natural language queries. *arXiv preprint arXiv:2105.02192*, 2021. 5, 14
- [53] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. In *NeurIPS*, 2018. 3
- [54] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018. 1
- [55] Mandela Patrick, Yuki M Asano, Ruth Fong, João F Henriques, Geoffrey Zweig, and Andrea Vedaldi. Multi-modal self-supervision from generalized data transformations. In *ICCV*, 2021. 1
- [56] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, Joao Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824*, 2020.
- [57] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. 2
- [58] Karol J Piczak. Esc: Dataset for environmental sound classification. In *ACM MM*, 2015. 4, 14
- [59] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 3, 4, 5, 6, 7, 15, 16, 17
- [60] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 6, 15
- [61] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 2020. 15
- [62] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015. 5

- [63] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. [1](#)
- [64] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgb-d images. In *ECCV*, 2012. [4](#), [14](#), [15](#)
- [65] Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens van der Maaten. Revisiting weakly supervised pre-training of visual perception models. In *CVPR*, 2022. [5](#)
- [66] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2014. [2](#)
- [67] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 2015. [4](#), [14](#)
- [68] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019. [1](#), [2](#), [3](#), [7](#)
- [69] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022. [15](#)
- [70] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. [8](#)
- [71] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. [3](#)
- [72] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. BEVT: Bert pretraining of video transformers. In *CVPR*, 2022. [2](#)
- [73] Yixuan Wei, Han Hu, Zhenda Xie, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *arXiv preprint arXiv:2205.14141*, 2022. [2](#)
- [74] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. [3](#)
- [75] Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. In *NeurIPS*, 2022.
- [76] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. [5](#)
- [77] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*, 2022. [2](#), [5](#)
- [78] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *CVPR*, 2022. [5](#)
- [79] Kim Youwang, Kim Ji-Yeon, and Tae-Hyun Oh. Clip-actor: Text-driven recommendation and stylization for animating human meshes. In *ECCV*, 2022. [2](#)
- [80] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. [1](#), [2](#), [5](#)
- [81] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. [1](#), [2](#)
- [82] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, 2022. [2](#)
- [83] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *CVPR*, 2022. [2](#), [3](#)
- [84] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, 2020. [7](#)
- [85] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for

scene recognition using places database. In *NeurIPS*, 2014. 5

- [86] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 2, 6, 16

A. Datasets and Metrics

AudioSet (AS) [18]. 这个数据集用于训练和评估。它包含来自 YouTube 的 10 秒视频, 被注释为 527 个类别。它包含了 3 个预定义的划分, 平衡划分约有 20,000 个视频, 测试划分约有 18,000 个视频, 非平衡训练划分约有 2,000,000 个视频。在训练中, 我们使用了 2,000,000 个非平衡数据集, 但没有任何标签, 并且只用于音频-视频匹配。在零-shot 评估中, 我们使用了测试集, 并使用文本类别名称以及后面描述的模板计算每个类别的逻辑值 (logits), 如 Appendix B.3 中所述。所使用的度量标准是 top-1 准确率。

ESC-50 (ESC) [58]. 我们使用这个数据集以零-shot 方式评估学习到的表示。该任务是“环境音分类”(ESC)。它包含了 2000 个 5 秒的音频片段, 分为 50 个类别。它有预定义的 5 折评估, 每个折包含 400 个测试音频片段。在这项工作中, 我们对每个折的评估集进行零-shot 预测, 并报告 5 折平均性能。对于消融实验, 我们仅使用第一个折以便于计算。所使用的度量标准是 top-1 准确率。

Clotho (Clotho) [16]. 这是一个来自 Freesound 平台的带有文本描述的音频数据集。它包含了一个开发集和一个测试集, 分别包含 2893 个和 1045 个音频片段, 每个片段都关联着 5 个描述。我们考虑文本 \rightarrow 音频检索任务, 并将每个关联的描述视为一个单独的测试查询, 从音频片段集合中进行检索。所使用的度量标准是 $\text{recall}@K$, 其中如果在前 K 个检索到的音频片段中包含了与真实音频相对应的结果, 则认为给定的测试查询被正确解决。

AudioCaps (AudioCaps) [36]. 这是一个包含音频视觉片段和文本描述的数据集, 来自 YouTube。它包含了 AudioSet 数据集中的片段, 如前面所述。我们使用 [52] 提供的划分,

它删除了与 VGGSound 数据集重叠的片段。最终我们得到 48198 个训练样本, 418 个验证样本和 796 个测试样本。我们只在测试集上对我们的模型进行零-shot 评估。任务是文本 \rightarrow 音频检索, 评估使用 $\text{recall}@K$ 。

VGGSound (VGS) [8]. 该数据集包含大约 200,000 个 10 秒长度的视频片段, 注释了 309 个声音类别, 包

括人类动作、发声物体和人物-物体交互等。我们只使用测试集中的音频 (共有 14,073 个片段) 进行零-shot 分类。评估使用的度量标准是 top-1 准确率。

SUN RGB-D (SUN). 我们使用 SUN RGB-D [67] 数据集中提供的注册 RGB 和深度图像进行训练, 这些图像来自 `train` 集合, 大约有 5,000 对图像和深度图像。我们遵循 [20] 对深度图像进行两步后处理 - 1) 我们使用填充的深度值, 2) 将它们转换为视差进行尺度归一化。该数据集仅用于训练, 因此我们不使用任何元数据或类标签。

SUN Depth-only (SUN-D). 我们只使用 SUN RGB-D [67] 数据集的 `val` 划分中的大约 5,000 个深度图像, 并将其表示为 SUN Depth-only。该数据集仅用于评估, 我们不使用 RGB 图像。我们对深度图像进行类似 SUN RGB-D 的处理 (填充深度, 转换为视差)。我们使用数据集中的 19 个场景类别, 并使用它们的类名构建零-shot 分类模板。

NYU-v2 Depth-only (NYU-D). 我们仅使用 NYU-v2 Depth-only [64] 数据集的 794 个 `val` 集深度图像进行评估。我们对深度图像进行类似 SUN Depth-only 的后处理。数据集中包含 10 个场景类别的名称。第 10 个场景类别称为“other”, 对应于 18 个不同的语义类别 - ['basement', 'cafe', 'computer lab', 'conference room', 'dinetette', 'exercise room', 'foyer', 'furniture store', 'home storage', 'indoor balcony', 'laundry room', 'office kitchen', 'playroom', 'printer room', 'reception room', 'student lounge', 'study', 'study room']。对于零-shot 评估, 我们将第 10 类别的余弦相似度定义为这 18 个类别名称之间的最大余弦相似度。

LLVIP (LLVIP). LLVIP 数据集 [31] 包含 RGB 图像和热成像 (红外低光) 图像对。该数据集在室外环境中使用固定摄像机观察街景, 包含在低光条件下拍摄的 RGB 图像与红外图像 (8~14 μm 频率)。RGB 热图对在数据集发布时进行了注册。我们使用包含 12025 个 RGB 图像和热图像对的 `train` 集进行训练。对于评估, 我们使用包含 3463 对 RGB 图像和热图像的 `val` 集。由于原始数据集是为检测任务设计的, 我们将其后处理为一个二分类任务。我们裁剪出行人边界框和随机边界框 (与行人具有相同的长宽比和大小), 以创建一个平衡

https://www.robots.ox.ac.uk/~vgg/research/audio-retrieval/resources/benchmark-files/AudioCaps_retrieval_dataset.tar.gz

的边界框集合，总共有 15809 个边界框（其中 7931 个是“person”边界框）。对于零-shot 分类，我们使用以下类别名称来表示“person”类 - ['person', 'man', 'woman', 'people'], 并使用 ['street', 'road', 'car', 'light', 'tree'] 表示背景类别。

Ego4D (Ego4D) [22]. 对于 Ego4D 数据集，我们考虑场景分类任务。在 Ego4D 数据集的 9,645 个视频中，存在 108 个唯一的场景。我们筛选掉所有标注了多个场景的视频，得到了 7,485 个仅有一个场景标注的视频。对于每个视频，我们选择包含同步 IMU 信号和对齐叙述的所有时间戳，并在每个时间戳周围采样 5 秒的片段。数据集随机划分为训练集（510,142 个片段）和测试集（68,865 个片段）。在训练过程中，我们只使用视频帧和相应的 IMU 信号。我们使用测试集来评估零-shot 场景分类性能，其中每个 IMU 信号片段被赋予视频级别的场景标签作为其真实标签。

A.1. 数据表示

我们使用标准的 RGB 和 RGBT 表示来处理图像和视频。对于视频，我们使用 2 帧的片段，受到 ViT 风格视频架构的最新研究的启发 [15, 69]，其中一个视频补丁的大小为 $2 \times 16 \times 16$ ($T \times H \times W$)。我们调整视觉编码器的权重以适应时空补丁，并在推理时在多个 2 帧的片段上聚合特征。因此，我们可以直接将在图像-文本数据上训练的模型用于视频。

对于热成像数据，由于当前红外热传感器返回的数据是单通道的 [31]，我们只使用了单通道图像。对于单视角深度，我们尝试了不同的编码方式-绝对深度 [64]（由 Kinect 等传感器返回）、逆深度 [61]、视差 [61] 和 HHA [24, 25]。总体而言，我们发现视差表示（一个单通道图像）效果最好。对于音频，我们将原始波形处理成 mel 频谱图 [21]，如主文中所述。对于 IMU，我们使用一个 $6 \times T$ 张量表示随时间变化的 IMU 传感器读数序列。

B. 评估细节

我们现在描述了本文中使用的评估设置。

B.1. 推理实现细节

IMU: 对于 IMU，我们采样固定长度的 5 秒片段，中心对齐于与叙述对齐的时间戳。对于每个片段，我们获

得一个 6×2000 维的输入，并使用每个片段作为独立的测试样本进行场景分类的零-shot 性能测量。

B.2. 少样本评估细节

音频: 对于使用 ESC 进行音频少样本训练，我们的模型和基线模型使用 AdamW 进行训练，学习率为 1.6×10^{-3} ，权重衰减为 0.05，共训练 50 个 epoch。
深度: 对于使用 SUN 进行深度少样本训练，我们的模型和基线模型使用 AdamW 进行训练，学习率为 10^{-2} ，无权重衰减，共训练 60 个 epoch。

B.3. 零-shot 评估细节

查询模板: 对于所有评估，我们使用了 CLIP [59] 中默认模板集合。

请注意，我们对非视觉的模式（如音频和深度）也使用相同的模板，因为我们只使用与图像相关的语义/文本监督。

B.4. 定性评估细节

跨模式最近邻: 我们通过对温度缩放后的嵌入特征进行检索来执行最近邻计算。最近邻是使用余弦距离计算的。

在图中 Figure 1，我们展示了对 ESC 的音频、对 IN1K 和 COCO 的图像进行检索，对 SUN-D 的深度进行检索，以及对 AudioCaps 的文本进行检索。
嵌入空间运算: 对于运算，我们再次使用温度缩放后的嵌入特征。我们对特征进行了 ℓ_2 归一化，然后将它们乘以 0.5 并求和。我们使用组合后的特征进行最近邻检索，使用余弦距离计算，如上所述。在图中 Figure 1，我们展示了来自 IN1K 的图像和 ESC 的音频的组合，并展示了来自 IN1K 的检索结果。

音频 → 图像生成: 为了从音频片段生成图像，我们依赖于重新实现的 DALL-E-2 [60] 模型。在 DALL-E-2 中，为了从文本提示生成图像，图像生成模型依赖于预训练的 CLIP-L/14 文本编码器生成的文本嵌入。由于 ImageBind 自然地将 CLIP 的嵌入空间与论文中提出的其他模式对齐，我们可以通过使用 ImageBind 的音频编码器生成的经过温度缩放的音频嵌入作为 DALL-E-2 中图像生成模型中 CLIP 文本嵌入的代理来实现零-shot 音频到图像的生成。

https://github.com/openai/CLIP/blob/main/notebooks/Prompt_Engineering_for_ImageNet.ipynb

使用音频检测物体：我们从 ESC 的验证集中提取了所有音频描述符，使用 ImageBind 的 ViT-B/32 编码器，总共得到 400 个描述符。我们使用现成的基于 CLIP 的 Detic [86] 模型，并将音频描述符用作 Detic 的分类器，替代 CLIP 基于文本的“类别”嵌入。在图 ?? 中的定性结果中，我们使用了 0.9 的得分阈值。

C. Pretraining details

C.1. Best setup

In Table 8 we detail the hyperparameters used to pre-train each of the models reported in Table 4. Our experiments were done on 32GB V100 or 40GB A100 GPUs.

Config	AS	SUN	LLVIP	Ego4D
Vision encoder	ViT-Huge			
embedding dim.	768	384	768	512
number of heads	12	8	12	8
number of layers	12	12	12	6
Optimizer	AdamW			
Optimizer Momentum	$\beta_1 = 0.9, \beta_2 = 0.95$			
Peak learning rate	1.6e-3	1.6e-3	5e-4	5e-4
Weight decay	0.2	0.2	0.05	0.5
Batch size	2048	512	512	512
Gradient clipping	1.0	1.0	5.0	1.0
Warmup epochs	2			
Sample replication	1.25	50	25	1.0
Total epochs	64	64	64	8
Stoch. Depth [28]	0.1	0.0	0.0	0.7
Temperature	0.05	0.2	0.1	0.2
Augmentations:				
RandomResizedCrop				
size	—	224px		—
interpolation	—	Bilinear	Bilinear	—
RandomHorizontalFlip	—	$p = 0.5$	$p = 0.5$	—
RandomErase	—	$p = 0.25$	$p = 0.25$	—
RandAugment	—	9/0.5	9/0.5	—
Color Jitter	—	0.4	0.4	—
Frequency masking	12	—	—	—

表 8. Pretraining hyperparameters

对比损失批处理大小与模态之间的关系。虽然对比损失确实需要较大的批处理大小，但这个要求并没有随着

模态数量的增加而增加。正如在附录 Appendix B 中提到的，我们的实验 (Table 2) 每次只对一对模态进行小批量采样：(视频, 音频) 采用 2K 的批处理大小，(图像, 深度)、(图像, 热度) 和 (视频, IMU) 采用 512 的批处理大小。这些批处理大小比之前的研究 [10, 59] 中使用的大于 32K 的批处理大小要小。

模态的组合。在 ?? 中，我们展示了组合音频和视频模态的结果。我们通过从每个样本中提取两个模态的嵌入，并对这些嵌入进行线性组合来进行模态的组合。在这种组合中，我们使用了 0.95 的视频权重和 0.05 的音频权重，经过实验发现这样的组合效果最好。

C.2. 消融实验设置

以下是我们在 § 5 中进行评估时使用的设置。与最佳设置不同，所有消融实验都使用 ViT-Base 作为视觉和模态特定编码器。模型训练了 16 个 epoch，除非另有说明。

在 Table 4b 中，线性投影头和 MLP 投影头之间的差异如下所述：在我们的实验中，MLP 投影头并没有改善性能。

Linear	Linear(in_dim, out_dim)
MLP	Linear(in_dim, in_dim), GELU, Linear(in_dim, out_dim)

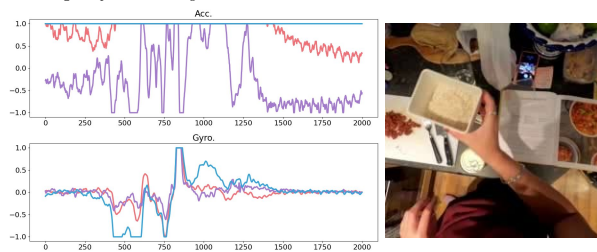
D. Additional Results

Qualitative results. We show additional results (along with audio) in the accompanying video. 异质模态的实际应用。一般来说，共享的嵌入空间可以支持各种跨模态的搜索和检索应用。例如，由于 IMU 传感器的普及（如手机、AR/VR 头盔、健康追踪器），ImageBind 可以允许用户使用文本查询在 IMU 数据库中进行搜索（无需使用 IMU-文本对进行训练）。基于 IMU 的文本搜索在医疗保健和活动搜索等领域具有应用前景。例如，在 Figure 4 中，我们展示了给定文本搜索查询时的 IMU（和相应的视频）检索示例。检索到的 IMU 样本，以 3 通道加速度计 (Acc) 和陀螺仪 (Gyro) 记录的形式呈现，与文本查询相匹配。

E. 其他消融实验

损失设计选择。由于模态特定的编码器是与冻结的图

Text query: "Cooking a meal"



Text query: "A person doing gardening work outdoors"

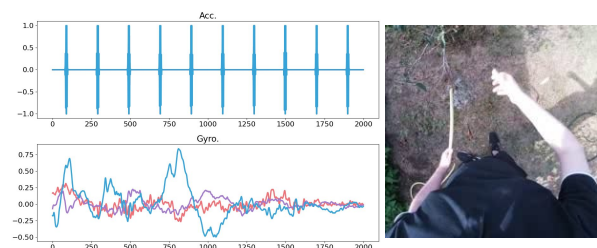


图 4. IMU retrievals. Given a text query, we show some IMU retrievals and corresponding video frames.

像编码器对齐训练的，我们尝试使用 ℓ_2 回归目标。在 ZS SUN 的 top-1 准确率中，我们观察到回归作为唯一目标（25.17

F. 伦理考虑

ImageBind 学习了多个模态的联合嵌入空间。这样的嵌入空间旨在将不同模态中的语义相关概念关联起来。然而，这样的嵌入空间可能也会创建无意识的关联。因此，包括 ImageBind 在内的联合嵌入模型必须通过测量这些关联及其影响的视角进行仔细研究。ImageBind 利用预训练模型在基于 Web 的大规模数据上学习了图像-文本嵌入，这些数据存在各种已记录的偏见 [59]。对于学习其他模态（如音频、热度、深度和 IMU）的联合嵌入，我们利用了 Appendix A 中提到的数据集。因此，这些联合嵌入仅限于数据集中存在的概念。例如，我们使用的热度数据集仅限于室外街景，而深度数据集仅限于室内场景。