arXiv:2601.11995v1 [cs.MM] 17 Jan 2026

利用推断潜在交互图学习音频-视觉嵌入

Donghuo Zeng*[0000−0002−6425−6270]、Hao Niu、Yanan Wang 和 Masato Taya

KDDI Research, Inc.，日本埼玉{xdo-zen, ha-niu, wa-yanan, ma-taya}@kddi.com

摘要。 要学习稳健的音频-视觉嵌入，需要将真正相关的音频和视觉信号联系起来，同时过滤掉偶然共现的内容——背景噪声、无关元素或未标注事件。大多数对比和三元组损失方法使用每段片段中稀疏的标注标签，并将任何共现都视为语义相似。例如，标注为“火车”的视频也可能包含摩托车音频和视觉内容，因为“摩托车”并非选定的标注；标准方法会将这些共现视为与其他位置的真实摩托车锚点相对的负例，从而造成假阴性，并遗漏真实的跨模态依赖关系。我们提出一个利用软标签预测和推断潜在交互来解决这些问题的框架：(1) *Audio–Visual Semantic Alignment Loss (AV-SAL)* 训练教师网络，生成跨模态对齐的软标签分布，为共现但未标注的事件赋予非零概率，从而丰富监督信号。(2) *Inferred Latent Interaction Graph (ILI)* 将 GRaSP算法应用于教师模型的软标签，以推断类别之间稀疏有向依赖图。该图突显了方向性依赖关系（例如，“火车（视觉）” → “摩托车（音频）”），揭示类别之间可能存在的语义或条件关系；我们将其解释为估计的依赖模式。(3) *Latent Interaction Reg-ularizer (LIR)*：学生网络通过度量损失和由 ILI 图指导的正则化器进行训练，按照软标签概率，将依赖关系相连但未标注的配对的嵌入拉近。在 AVE 和 VEGAS 基准测试上的实验表明，平均精度均值（MAP） consistently 提升，证明将推断潜在交互融入嵌入学习能够提高稳健性和语义连贯性。

关键词： 音频-视觉 · 潜在交互图 · 跨模态检索 · 软标签

1 引言

学习音频和视觉模态之间的对齐表征，对于跨模态检索 [27, 37, 41] 和音频-视觉事件

* 通讯作者

定位等任务至关重要。标准方法通过度量学习（例如对比损失或三元组损失）将音频和视觉输入投影到共享嵌入空间中，使标注的正样本对彼此接近，并将负样本推远。

然而，现实场景中的音视数据远比通常提供的稀疏标注丰富得多。数据集通常只为每个片段指定一个音频标签和一个视觉标签，即使其中同时发生了多个事件。例如，一段短视频片段可能在音频轨道和视觉轨道上都被标注为“火车”（音频内容是火车汽笛或发动机声，视觉内容是火车的外观），但同一片段还可能包含摩托车发动机的轰鸣声以及驶过汽车的车头灯。这些未标注的音频和视觉元素与已标注内容共存，共同构成场景的真实语义——即使它们在标注过程中被认为“不重要”。遗憾的是，当模型使用传统损失函数在此类数据上训练时，会将这些缺少匹配标签的未标注共现视为负样本。具体而言，如果“火车”是唯一的标注标签，那么未标注的“摩托车”音频和“摩托车”视觉虽然共同出现，却会被视为彼此无关，从而使它们的嵌入被推远。这会引发两个根本问题：**(I) 由标注不完整导致的假负样本**。仅依赖标注标签时，标准度量学习方法不可避免地会将语义上有意义但未标注的共现音视对视为负样本 [7,8,11,39,42]。因此，模型会学着排斥现实场景中实际上彼此相关的嵌入。其结果是，同一“火车”片段中同时出现的“摩托车”音频和“摩托车”视觉被错误地视为不相似，从而降低了对真正相关配对的检索效果。**(II) 由偶然共现产生的伪相关**。除了简单地忽略未标注事件之外，标准方法 [17,18,21] 还无法区分真正的语义共现与偶然或背景共现 [2,12]。在嘈杂或复杂场景中，多个重叠元素十分常见；例如，火车场景可能包含交通声、风声或施工声。如果没有机制过滤掉这类偶然信号，仅使用已标注数据训练的模型可能会由于频繁共现，错误地将背景音频（例如交通嗡鸣声）与目标事件（例如火车）联系起来。这会导致损害泛化能力的伪相关：在测试时，模型可能仅因听到交通声，就检索出“火车（视觉）”。

总而言之，这些问题意味着，现有的度量学习 [38]方法所生成的嵌入，既无法捕捉真实世界音视频共现的丰富性，也无法有效滤除背景噪声。换句话说，两个确实具有潜在关系的片段，例如同时包含“火车”和“摩托车”的片段，可能会被错误地表示为毫不相关；而两个仅仅共享背景上下文的片段，例如交通噪声和火车驶过的声音，则可能被错误地视为关系密切。我们的核心洞见是，相关性的重要程度各不相同：有些反映了有意义的语义依赖关系，有些则只是偶然出现。

为解决这一问题，我们提出一个利用预测的软标签分布和依赖关系推断来改进嵌入学习的框架：（1）**AV-**

SAL通过语义对齐损失训练一个教师模型 网络，生成涵盖所有类别的软标签，从而捕捉稀疏标注之外的潜在事件。（2）**ILI图**：我们将 GRaSP [16] 算法应用于教师模型的软标签，以推断类别之间的稀疏有向依赖图。该图编码了软标签/表征空间中的非对称预测依赖关系；虽然在标准结构假设下，这些模式可能具有可解释性，但我们将其视为根据观测模型输出估计得出的依赖结构，而非确凿的因果证明。

(3) **LIR**：我们训练一个学生 网络，并使用源自 ILI 图的潜在交互正则化器，确保已标注对和由依赖关系关联的未标注对都能强化表征，同时降低偶然配对的权重。

我们在 AVE 和 VEGAS 基准测试上验证了该方法，这些数据集中多事件共现十分常见。我们的方法始终优于最先进的基线方法，平均精度均值（MAP）约提升 1.5%，表明推断潜在交互能够显著增强音频-视觉嵌入的鲁棒性和语义连贯性。

2 相关工作

2.1 音频-视觉嵌入学习

利用线性、非线性和深度网络方法学习音频和视觉模态之间的对齐表征，已得到广泛研究。早期工作采用典型相关分析（CCA） [6] 及其扩展方法（聚类典型相关分析 [22],K-CCA [15]），这些方法能够找到相关投影，但难以处理非线性关系，且扩展性较差。深度典型相关分析变体（$DCCA$ [3], C-$DCCA$ [34], VAE-CCA [44]）利用神经网络对非线性和类内相关性建模，但仍受稀疏标注的限制，并忽略未标注共现。基于深度网络的方法从监督分类器 [5,23,29,30,46]或自监督对应目标 [4,14], 中提取嵌入，以捕捉丰富的共现模式。尽管这些方法效果显著，却没有明确区分真正的语义依赖关系与偶然共现或背景共现。双子空间方法 [40]试图分离共享特征和模态特定特征，但构建子空间仍依赖已标注数据。

度量学习方法 [7, 8, 38, 39, 42] 通过对比损失或三元组损失直接约束嵌入距离。基于蒸馏的方法 [31, 36] 则进一步利用教师网络和软标签分布来丰富监督信号。尽管这些方法效果显著，但仍会将未标注却有意义的共现视为负样本，从而导致假阴性和伪相关。

总体而言，现有音视嵌入方法受限于稀疏标注，缺乏捕捉潜在事件的机制，也无法动态区分偶然共现与具有语义意义的依赖关系。这促使研究者探索将软标签监督与依赖关系推断相结合的框架，以更好地反映现实场景的丰富性。

2.2 潜在交互与依赖关系学习

近期研究不再仅仅依赖基于相关性的嵌入，而是探索学习潜在交互结构，以捕捉模态之间有意义的依赖关系 [19]。这些方法旨在从观测数据中推断潜在的依赖模式 [10, 19, 32]。在高层次的视觉与语言任务中（例如，视觉问答 [28]，图像描述 [33]，视觉对话 [20]），研究表明，对潜在交互进行建模能够减少伪相关、分离混杂因素并提升泛化能力。同样，在音视学习中，捕捉此类依赖关系可以减轻缺失标注和偶然共现带来的影响。除了通用依赖关系推断之外，结构化图也被用于显式编码类别之间的关系或嵌入之间的关系。例如，GRaSP [16] 通过结合置换搜索与BIC评分，根据软标签预测构建稀疏有向依赖图。这类图能够突出一致的交互模式，无需真实因果关系标注；它们还可以通过强化语义关联对、降低偶然共现对的权重来指导嵌入学习，使模型能够在标注稀疏或含噪的情况下动态发现潜在语义结构。

通过整合软标签监督与潜在交互推理，模型能够超越静态的基于相关性的嵌入，揭示更深层的语义依赖关系，并生成更具可解释性和迁移性的音频-视觉表示 [36, 45]。

3 所提方法

3.1 符号与概述

设 $\mathcal{D} = \{x_n\}_{n=1}^{N}$ 为包含 N 个片段的数据集。每个片段 x_n 都有一个音频-视觉对 $(x_{a,\,n},\ x_{v,\,n})$。教师模型生成软标签分布 $s_n \in [0,1]^C$（每个类别对应独立的 Sigmoid 分类头，C 为类别数）；将这些分布汇集为 $S \in \mathbb{R}^{N\times C}$。学生嵌入网络为 $z_\theta(x) \in \mathbb{R}^d$，而 $D(z,\ z')$ 表示欧氏距离。

我们的流程遵循两阶段流程（教师模型 $\rightarrow$ 学生）。将教师模型训练至第 E_{teacher} 轮后，我们收集 S 并推断出一个稀疏的有向推断潜在交互（ILI）邻接矩阵 $\widehat{A} = \text{GRaSP}(S) \in \mathbb{R}_{\geq 0}^{2C\times 2C}$。$\widehat{A}_{ij}$ 量化类别 i 到类别 j 的发现的依赖关系（$i,\ j \in \{1, 2, ..., 2C\}$）。在学生训练期间，我们通过潜在交互正则化器（LIR）扩充度量目标，并以 A 加权。

3.2 教师模型：音频–视觉语义对齐（AV-SAL）

教师模型经过训练，能够生成校准软标签，从而发现稀疏标注之外的潜在共现。我们在模态软标签 s_n 与可用的独热标签 $Y_n \in \{0,1\}^C$之间使用对齐损失:

$$\mathcal{L}_{\text{AV-SAL}} \;=\; \frac{1}{N}\sum_{n=1}^{N}\Big(\|s_{a,n} - Y_n\|_2^2 + \|s_{v,n} - Y_n\|_2^2\Big), \tag{1}$$

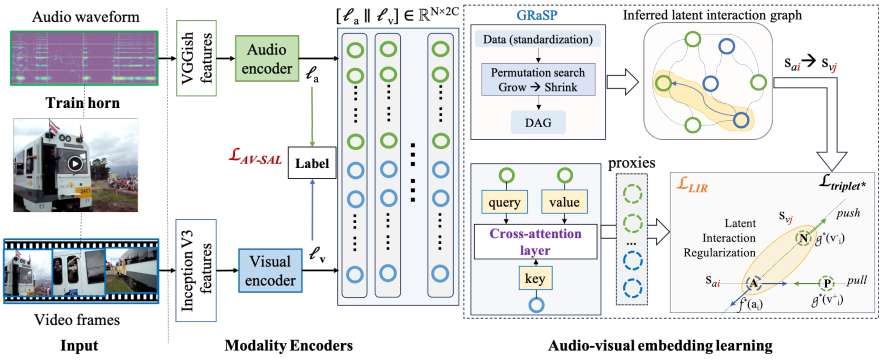


图1：**我们架构的概览**。使用 VGGish 和 Inception V3 提取音频和视觉特征。通过结合代理三元组损失（包含跨模态注意力）的 AV-SAL 优化音频和视觉编码器的 logits，从而生成经过校准的软标签分布。这些软标签$\hat{}$作为输入传递给 GRaSP，以推断有向潜在交互图 A，其中每条边 $s_{a,\,i} \to s_{v,\,j}$ 表示音频标签 i 与视觉标签 j之间的依赖关系。最后，根据发现的依赖关系对嵌入进行加权和归一化，并通过 LIR 项进一步优化。

其中，$s_{a,\,n}$，$s_{v,\,n} \in [0,1]^C$ 是教师模型对片段 n的各模态 Sigmoid输出。教师模型还通过代理-三元组 [38]损失进行训练，使软标签同时反映标注和嵌入：

$$\mathcal{L}^{(\text{teacher})} = \mathcal{L}_{\text{AV-SAL}} + \mathcal{L}_{\text{triplet}*}. \tag{2}$$

给定从交叉注意力层 [38], 提取的代理嵌入 $\ell_{a*}^{(i)}$ 和 $\ell_{v*}^{(i)}$，代理三元组损失定义为

$$L_{\text{triplet}^*} = \frac{1}{N}\sum_{i=1}^{N}\Big[\|\ell_{a*}^{(i)} - \ell_{v*}^{(i+)}\|_2^2 - \|\ell_{a*}^{(i)} - \ell_{v*}^{(i-)})\|_2^2 + \alpha\Big]_+, \tag{3}$$

其中，α是间隔超参数，用于确保锚点与正样本的距离至少比其与负样本的距离近 α。此处，($\ell_{a*}^{(i)}$，$\ell_{v*}^{(i+)}$)构成一个正样本对（标签相同），而 $\ell_{v*}^{(i-)}$则从不同类别中采样。

3.3 推断 ILI图

教师阶段（训练轮次 E_{teacher}）结束后，我们收集所有片段的教师模型 $logits$，并构建 logits 数据矩阵：

$$\mathcal{L} = \big[\ell^{(1)}, \ldots, \ell^{(N)}\big]^\top \in \mathbb{R}^{N\times 2C},$$

其中每一行 $\ell^{(i)} = [\ell_a^{(i)}, \ell_v^{(i)}]$ 将片段 i的 C 音频和 C 视觉 logits 拼接在一起。由于不同类别的 logits 可能具有不同的尺度，我们按列对其进行标准化 $\mathcal{L}$，得到 $\widetilde{\mathcal{L}}$（每列均为零均值和单位方差）。我们应用 GRaSP [16] 来 $\widetilde{\mathcal{L}}$。GRaSP

通过贪心 tuck/score 过程搜索排列（变量排序）空间，并针对每一种候选排序 π，使用 Grow–Shrink 程序为每个节点确定父节点集合 $P_j(\pi)$。GRaSP 的标准局部评分是高斯线性模型的贝叶斯信息准则；GRaSP 选择总评分最低的排序 $\pi^\star$，并返回由此导出的稀疏有向无环图。

作为一种面向鲁棒性、针对特定应用的调整，对于每个具有前驱节点 $P_j(\pi)$ 的节点 j，我们可以选择在标准化后的 logits 上拟合逐节点 ℓ_1 正则化线性模型：

$$\widehat{\beta}^{(j)}(P_j(\pi)) \;=\; \arg\min_{\beta}\; \frac{1}{2n}\big\|\widetilde{\mathcal{L}}_{:,j} - \widetilde{\mathcal{L}}_{:,P_j(\pi)}\beta\big\|_2^2 \;+\; \lambda_{\mathrm{reg}}\|\beta\|_1, \tag{4}$$

其中，$\lambda_{\mathrm{reg}} \geq 0$ 是一个较小的正则化参数（设定 $\lambda_{\mathrm{reg}} = 0$ 后，式（4）将简化为普通最小二乘法）。我们取初步的有向权重

$$\widetilde{A}_{i\to j} \;=\; \begin{cases} \big|\widehat{\beta}_i^{(j)}\big|, & i \in P_j(\pi^\star), \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

也就是说，对于非父节点，系数为零。GRaSP 通过其贪心 tuck/score 过程选择最佳排序 $\pi^\star$ 及相应的父节点集合 $P_j(\pi^\star)$；上述逐节点拟合仅用于为选定边提供基于系数的权重。为构建 LIR 使用的最终邻接矩阵 $\widehat{\ }$，我们 (i) 将 A 的对角线元素设为零，并 (ii) 用其 L_1 范数对矩阵进行归一化：

$$\widetilde{A}_{ii} := 0 \quad \forall i, \qquad \widehat{A} \;=\; \frac{\widetilde{A}}{\|\widetilde{A}\|_1}. \tag{6}$$

由此得到的 $\widehat{A} \in \mathbb{R}^{2C\times 2C}_{\geq 0}$ 是一个稀疏有向依赖矩阵（即 ILI 图），用于对 LIR（式（7））赋权。作为实践说明，应将 $\widehat{A}$ 理解为从教师 logits 推导出的估计依赖图，而非确凿的因果证据；因此，我们会报告边稳定性（通过自助法/不同随机种子评估），并在实验中将 LIR 限制于权重较高且稳定的边。式（6）中的 ℓ_1 正则化逐节点回归以及对角线处理和归一化，都是为增强鲁棒性和可解释性而采用的实现选择。

3.4 潜在交互正则化器（LIR）

$\widehat{\ }$LIR 使用 A，将预测为会共同激活依赖类别的片段的学生嵌入拉近，从而纠正由稀疏标注造成的假阴性。定义采样分布 $\mathcal{S}_{ij}$，从中抽取音频/视觉片段对 $(n,\ m)$，其中类别 i（分别为 j）在教师软标签中处于激活状态（例如 $s_n[i] > \tau$）。LIR 为

$$\mathcal{L}_{\mathrm{LIR}}(\theta;\widehat{A}) \;=\; \sum_{i=1}^{C}\sum_{j=1}^{C} W((m,i),(n,j))\widehat{A}_{ij}\; \mathbb{E}_{(n,m)\sim\mathcal{S}_{ij}}\big[\, D\big(z_\theta(x_n),\, z_\theta(x_m)\big)\,\big]. \tag{7}$$

$\widehat{\ }$在实践中，我们通过小批量采样近似期望，并对 A 进行归一化，使其满足 $\sum_{i,\ j}\widehat{A}_{ij} = 1$。我们将 $W(a,\ a) = W(v,\ v)$ 设为 0.1，并将 $W(a,\ v) = W(v,\ a)$ 设为 0.4，以强调跨模态潜在交互连接。

总体训练目标 教师网络和学生网络共用相同的模态编码器和嵌入架构。我们采用转换轮次为 $M \in \{300, 400, 500, \ldots, 900\}$ 的两阶段流程。教师阶段（第 1 至 M 轮）的目标是学习经过良好校准的软标签，以进行潜在发现。在学生阶段（第 M 轮至第 1000 轮），推断出潜在交互图后，我们引入 LIR，以因果结构指导嵌入学习：

$$\mathcal{L}^{(student)} = \mathcal{L}_{\text{AV-SAL}} + \mathcal{L}_{\text{triplet*}} + \gamma \mathcal{L}_{\text{LIR}}, \tag{8}$$

其中，$\gamma = 0.005$ 是通过经验选择的。AV-SAL 生成稳健的软标签；三元组损失保留标注关系；LIR 则通过强制施加学到的潜在交互链接来纠正假阴性。我们通过随机梯度下降 [24] 和反向传播优化 $\mathcal{L}$。

4 实验

4.1 数据集与评估

我们在视听跨模态检索（AV-CMR）任务上评估模型，使用 VEGAS [47] 和 AVE [29] 数据集，并采用标准划分和预处理方式 [40]，以确保各类别的表征比例均衡。VEGAS 包含 28,103 个 YouTube 片段（2–10 秒），涵盖 10 个事件类别，其中 22,482 个用于训练，5,621 个用于测试。AVE 包含 1,955 个片段，涵盖 15 个类别，其中 1,564 个用于训练，391 个用于测试。音频特征通过对每个片段的片段级 VGGish [9] 嵌入进行提取，并对 1 秒片段取平均，得到 128 维嵌入。视觉特征是每个片段的 1024 维描述符，通过对每隔 1 秒采样的帧级 InceptionV3 [1] 特征进行平均池化获得。

评估时，我们使用嵌入之间的余弦相似度进行音频到视觉和视觉到音频检索，并据此为检索候选项排序。按照先前的研究 [39, 40]，分别计算两个跨模态检索方向（音频到视觉和视觉到音频）的平均精度均值（MAP），然后取其平均值。

4.2 实现设置

每个模态分支由三个包含 1,024 个单元的全连接层构成，使用 Tanh 激活函数和 0.15 的丢弃法，随后通过线性投影映射到嵌入空间，其维度与事件类别数量一致（AVE 为 15，VEGAS 为 10）。我们对每个数据集分别进行训练，批量大小为 400，训练 1,000 个轮次，采用三元组损失（间隔 $= 1$ 2 [13] 1×10^{-4} $\beta_1 = 0.9$ $\beta_2 =$.）和 0.999 **1** Adam优化器（学习率 , , ）。训练采用两阶段流程 [38]: **) 预训练**（教师模型，训练轮次 1–400）：在原始嵌入上优化 AV-SAL 和基于代理的三元组损失；使用 GRaSP 从软标签输出中提取潜在交互邻接矩阵。**2) 潜在交互细化**（学生模型，训练轮次 401–1000）：继续优化三元组损失和 AV-SAL，并基于 GRaSP 图进行潜在交互正则化。

实验使用 PyTorch 1.12 在 Ubuntu 22.04 上实现，并采用 NVIDIA RTX 3080 GPU（10 GB），使用固定随机种子 42 以确保可复现性。代码见 [1]。

[1] https://github.com/ZenzenDatabase/causalCMR

表1：**与最先进方法的平均精度均值比较**。 最佳分数以粗体标示，次佳分数加下划线。

模型	AVE 数据集			VEGAS 数据集		
	一个 → V V →A		Avg.	A → V V →A		Avg.
随机情况	0.127	0.124	0.126	0.110	0.109	0.109
典型相关分析 [6]	0.190	0.189	0.190	0.332	0.327	0.330
核典型相关分析 [15]	0.133	0.135	0.134	0.288	0.273	0.281
深度典型相关分析 [3]	0.221	0.223	0.222	0.478	0.457	0.468
C-典型相关分析 [22]	0.228	0.226	0.227	0.711	0.704	0.708
C-DCCA [34]	0.230	0.227	0.229	0.722	0.716	0.719
TNN-C-CCA [42]	0.253	0.258	0.256	0.751	0.738	0.745
VAE-CCA [44]	0.328	0.302	0.315	0.821	0.824	0.822
CCTL [39]	0.328	0.267	0.298	0.766	0.765	0.766
VideoAdviser [31]	-	-	-	0.825	0.819	0.822
EICS [40]	0.337	0.279	0.308	0.797	0.779	0.788
TLCA [35]	0.410	0.451	0.431	0.822	0.838	0.830
MSNSCA [43]	0.323	0.343	0.333	0.866	0.865	0.866
AADML [38]	0.890	0.883	0.887	0.901	0.891	0.896
我们的方法	**0.906**	**0.899**	**0.903**(+1.62%)	**0.913**	**0.907**	**0.910**(+1.45%)

4.3 性能比较

我们将所提方法与具有代表性的视听跨模态检索（AV-CMR）方法进行比较，其中既包括经典的基于相关性的技术，也包括近期的深度学习模型，并将它们分为三类：（1） *CCA-based methods* 包括典型相关分析（CCA） [6],K-CCA [15], 聚类典型相关分析（Cluster-CCA） [22],深度典型相关分析（DCCA） [3], C-DCCA [34], 结合聚类典型相关分析的三元组方法（TNN-C-CCA） [42], 和 VAE-CCA [44]。这些方法通过学习投影（线性或非线性）来最大化模态间相关性。（2）采用 Transformer 或蒸馏的表征学习方法，例如 MSNSCA [43] ，它使用带有共享交叉注意力的多尺度网络实现细粒度音视对齐。VideoAdviser [31] 方法利用基于 CLIP 的教师-学生蒸馏，获得稳健的音频-视觉嵌入。（3 *Metric learning*）包括 *methods* CCTL [39], EICS [40], 和 AADML [38]。这些方法通过对比或三元组目标学习判别性嵌入向量。

如表1所示，我们的模型在两个数据集上均取得了最高的平均精度均值，分别在 AVE 和 VEGAS 上以 1.62% 和 1.45% 的优势超越了最强基线。在跨模态检索中，鉴于现有基线的性能已十分强劲且接近饱和，这些提升尤为显著，彰显了我们方法的有效性。

4.4 消融研究

潜在交互图分析 我们通过在七个检查点（训练轮次=300、400、...、900）上对软标签输出应用 GRaSP，跟踪跨模态潜在交互的演变。我们使用频率矩阵 $\mathbf{F} \in [0,1]^{2C\times 2C}$总结时间一致性：

$$F_{ij} \;=\; \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{\widehat{A}_{ij}^{(t)} > \varepsilon\}, \qquad T=7,\ \varepsilon \geq 0, \tag{9}$$

其中，$\mathbf{F}_{ij}$表示边（$i \to j$）出现的检查点所占比例。

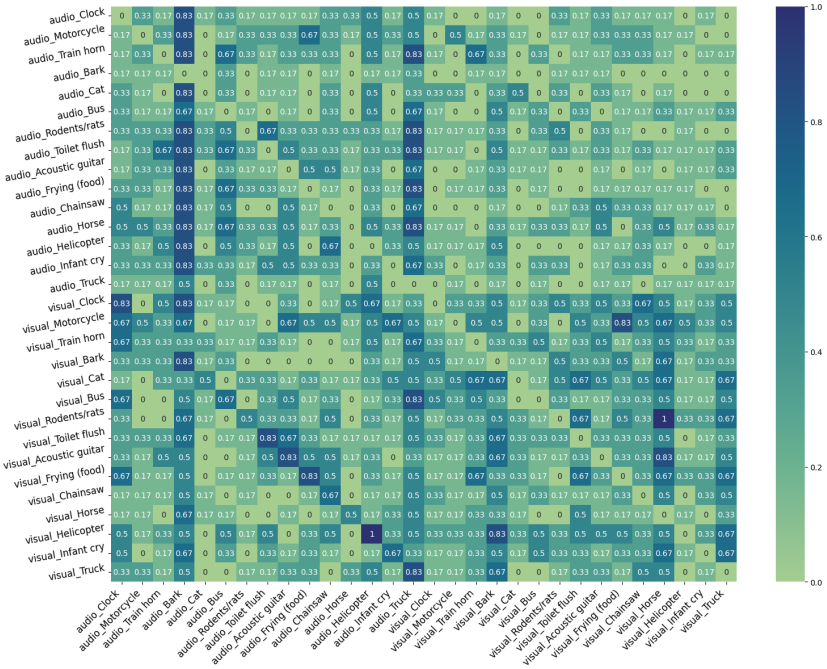


图2：**潜在交互边频率热图**（在 AVE 上）。每个单元格显示在七个检查点中，类别 i 被推断为影响类别 j 的频率。

图 2 将$\mathbf{F}$可视化。高频跨模态链接（例如，“直升机（视觉）$\rightarrow$ 直升机（音频）”为 0.99；“卡车（视觉） $\rightarrow$ 卡车（音频）”为 0.83；“煎炸声（视觉）$\rightarrow$ 煎炸声（音频）”为 0.83）表明，在训练过程中学到的关联具有持久性和稳健性。共有 81 个类别对的频率超过 0.714（即至少出现在 7 个检查点中的 5 个），这表明某些潜在交互链接能够被持续恢复——这可能反映了较强的时间/上下文耦合，而非偶然共现。这些稳定模式凸显了利用推断出的潜在交互来揭示高置信度跨模态关系的价值。

演变中的潜在交互子图案例分析。图 3 展示了训练轮次中的三个代表性子图（上：以“摩托车（音频）”为源节点；中：以“火车汽笛（视觉）”为源节点；下：以“公交车（视觉）”为目标节点）。**(i) “摩托车（音频）”（源节点）**。该节点的高激活始终与音频类别（如“吠叫”“链锯”）和视觉类别（如“摩托车”“火车汽笛”“直升机”）的模型激活增强相关。与机械相关事件之间的稳定连接表明，引擎类声音能够可靠地指示户外/机械场景，而部分连接（如与“直升机”或“冲马桶声”的连接）会随训练轮次而变化。**(ii) “火车汽笛（视觉）”（源节点）**。火车汽笛的视觉线索反复预测音频事件（如“时钟”“摩托车”“直升机”）及邻近的视觉类别（如“卡车”“冲马桶声”），反映出声学信息丰富的交通环境。**(iii) “公交车（视觉）”（目标节点）**。作为目标节点，“公交车（视觉）”接收来自多种源节点（“婴儿哭声”“煎炸声（食物）”“啮齿动物/老鼠”“火车汽笛”）的有向输入，表明公交车场景出现在多样的城市、家庭和交通环境中。

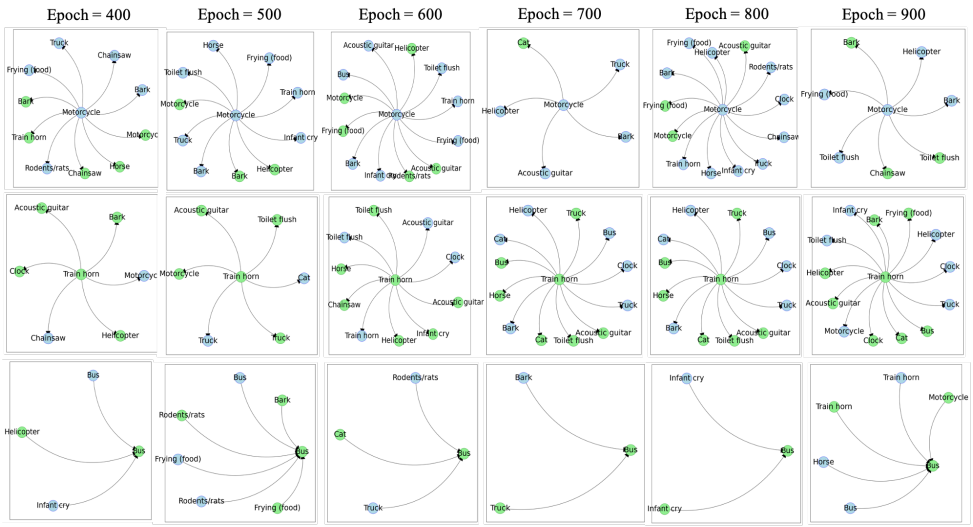


图 3：AVE 数据集在不同训练轮次（400–900）下的潜在交互子图。顶行显示以“摩托车（音频）”为原因的子图，中间行显示以“火车汽笛（视觉）”为原因的子图，底行显示以“公交车（视觉）”为结果的子图。蓝色节点表示音频类别，绿色节点表示视觉类别。

总的来说，这些示例表明，推断出的有向依赖关系既具有可解释性，又是动态变化的：某些关联能够被稳定地识别出来（高置信度信号），而另一些则会在训练过程中演变——这支持利用模型推断的依赖信息来改进跨模态嵌入学习。

推断潜在交互的影响 我们通过改变插入轮次 $M \in \{300, 400, 500, 600, 700, 800, 900\}$ 4 （基线：不插入），研究推断潜在交互图注入时机的敏感性。图中绘出了这些设置下 AVE 的训练损失和测试平均精度均值。在 M引入推断交互会立即扰动损失，随后迅速趋于稳定，表明优化器能够适应额外的正则化。与基线相比

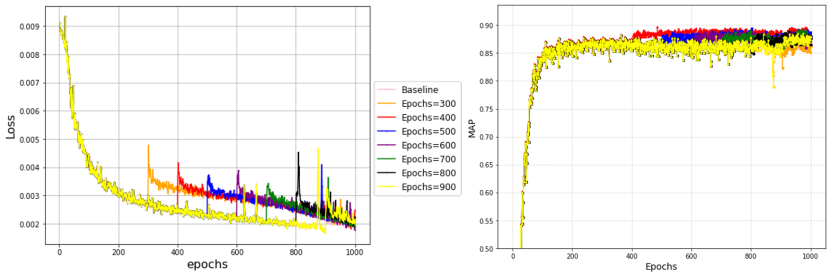


图 4：不同潜在交互插入轮次下 AVE 的训练损失（左）和测试平均精度均值（右）$M \in \{300, 400, \ldots, 900\}$（基线 =不插入）。

表2：不同度量学习方法在 AVE 和 VEGAS 数据集上的各组件对最终目标损失的影响。

模型	AVE 数据集			VEGAS 数据集		
	A→V	V→A	平均值	A→V	V→A	平均值
AV-SAL+代理（三元组）+LIR	**0.906**	**0.899**	**0.903**	**0.913**	**0.907**	**0.910**
AV-SAL+代理（困难三元组）+LIR	0.893	0.889	0.891	0.873	0.867	0.870
AV-SAL+代理（对比）+LIR	0.822	0.858	0.840	0.795	0.814	0.805
AV-SAL+代理（N对）+LIR	0.835	0.846	0.841	0.658	0.611	0.634
AV-SAL+代理（三元组）	0.890	0.883	0.887	0.901	0.890	0.896
AV-SAL+代理（困难三元组）	0.883	0.885	0.884	0.865	0.851	0.858
AV-SAL+代理（对比）	0.798	0.854	0.826	0.783	0.807	0.795
AV-SAL+代理（N对）	0.828	0.833	0.830	0.647	0.612	0.629
AV-SAL+三元组+LIR	0.408	0.451	0.430	0.763	0.727	0.745
AV-SAL+困难三元组+LIR	0.407	0.439	0.423	0.774	0.759	0.767
AV-SAL+对比+LIR	0.376	0.381	0.379	0.712	0.714	0.713
AV-SAL+N对+LIR	0.326	0.340	0.333	0.568	0.550	0.559
代理（三元组 [25]）+LIR	0.886	0.880	0.883	0.890	0.883	0.887
代理（困难三元组 [25]）+LIR	0.877	0.880	0.879	0.857	0.846	0.852
代理（对比 [7]）+LIR	0.790	0.844	0.817	0.776	0.794	0.785
代理（N对 [26]）+LIR	0.822	0.828	0.825	0.642	0.603	0.623

（无交互），所有交互感知设置均提升了平均精度均值，说明推断出的依赖关系提供了有用的监督。 $M = 400$取得了最佳权衡：损失平稳收敛，测试平均精度均值最高。过早引入交互可能会干扰初始表征学习，而过晚引入则会削弱其塑造嵌入的能力； $M = 400$ 在图推断的标签可靠性与留出充足训练时间以便图引导表征优化之间取得了平衡。

损失组件的影响 表2报告了 AVE 和 VEGAS 上不同组合的平均精度均值（MAP）分数，这些组合涉及 AV-SAL、基于代理的度量学习（三元组、困难三元组、对比、N对）和 LIR。表格中的各区块分别展示了各组件的作用：(1) AV-SAL +代理 (+损失) + LIR（完整模型），(2) 不含 LIR 的 AV-SAL +代理 (+损失)，(3) AV-SAL + 损失 +LIR（无代理），以及 (4) 代理 (+损失) +LIR（无 AV-SAL）。完整模型表现最佳：AV-SAL + 代理（三元组） + LIR 取得了最高的 MAP（**AVE 上为 0.903**，**VEGAS 上为 0.910**），分别比表现最接近的 AV-SAL+代理基线高出 $+0.016$ 和 $+0.014$。

这表明，LIR 能够强化跨模态对齐，尤其是与 AV-SAL 和 Proxy 搭配时。在基于代理的损失中，三元组监督效果最强，带来了最稳健的提升，并生成了判别性最强的嵌入。总体而言，消融实验证实， $AV\text{-}SAL + Proxy\ (Triplet) + LIR$ 是最优配置，其中 AV-SAL 和 LIR 发挥了互补作用。

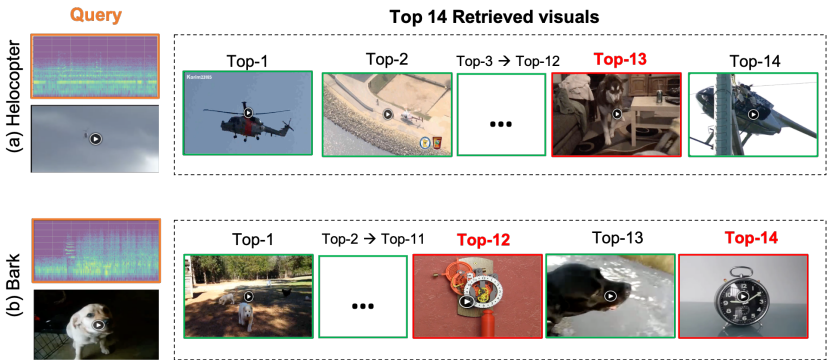


图 5：AVE 上音频到视觉检索的定性结果。图中展示了两个标签不同的音频查询对应的排名前 14 的检索视觉样本。正确匹配以绿色突出显示，错误匹配以红色突出显示。

4.5 定性结果

我们展示了 AVE 数据集上的音频到视觉检索定性示例，以评估所学跨模态嵌入的语义对齐效果（图 5）。少数检索错误可由推断出的依赖图中学到的交互强度解释（图 2）。对于直升机音频查询，模型将音频–直升机与视觉–直升机强烈关联（边频率为 $= 1.0$），因此检索结果大多正确。排名第 13 位唯一错误的狗叫视觉样本，对应较弱但非零的依赖关系（音频–直升机 $\rightarrow$ 视觉–狗叫，频率为 $= 0.5$），这表明模型存在部分混淆，可能是由于声学特征相似（例如高能量的户外声音）。相比之下，对于狗叫音频查询，音频–狗叫与视觉–时钟之间不存在推断出的依赖关系（频率为 $= 0$），也没有指向视觉–狗叫的强方向性依赖。因此，排名较低位置出现的少数错误时钟样本反映了交互图未能强化的残余歧义。总体而言，正确检索与较强的自依赖关系相符，而错误则出现在依赖信号较弱或缺失的低排名位置，这表明图正则化会影响嵌入结构，但不会对其作出僵化的规定。

5 结论

我们提出了一种音频-视觉嵌入学习框架，旨在应对稀疏标注和偶然共现带来的挑战。我们的方法首先在不同模态之间对齐教师生成的软标签（AV-SAL），然后应用 GRaSP 算法推断有向潜在交互图，最后通过潜在交互正则化（LIR）强化这些联系，从而有效区分有意义的语义依赖关系与表面相关性。在 AVE 和 VE-GAS 上进行的实验表明，与最先进的基线方法相比，平均精度均值（MAP）稳定提升约 1.5%，证明了所学嵌入更加稳健且语义上更加连贯。未来，我们计划探索更丰富的潜在交互建模方法——或许能捕捉类似因果关系的规律，并通过纳入自监督策略来减少对人工标签的依赖。

6 讨论与澄清审稿意见

ILI图——这并非因果关系的论断。我们并未声称推断出的 ILI图构成确定的因果证据；相反，我们将 GRaSP 应用于教师 logits，从观测模型输出中恢复一个估计的、稀疏有向依赖矩阵。因此，有向边表示教师逻辑值/表征空间中的非对称预测依赖关系（即类别 i 的激活如何帮助预测类别 j），而非经证实的干预或因果机制。为减少伪相关连接，我们采用对角线置零和 $L1$ 归一化，在多个检查点之间进行阈值筛选和边稳定性聚合，并将 LIR 限制于权重较高且稳定的边；这些做法都有助于使图偏向紧凑、可复现的依赖关系。重要的是，该图仅作为软正则项（LIR），与 AV-SAL 和三元组监督共同使用，因此它会推动嵌入形成一致的共同激活，而不会强制施加可能有误的硬性关系。最后，我们承认，有些边反映的可能是声学相似性或频繁共现，而非语义因果关系；要确立真正的因果关系，需要进行干预或实验验证，这超出了本观察性研究的范围。

可扩展性与图稀疏性。GRaSP 的搜索无法轻易扩展到超大规模的类别词汇表，这促使我们在较小的标签集（10、15 个类别）上开展实验。在实践中，GRaSP 是一个离线且不频繁执行的训练阶段步骤，而 GRaSP 的稀疏化目标函数能够生成紧凑图，并且只需进行适度且固定的 $lamba$（我们不会对 $lamba$进行过度调优）。对于更大的标签集，可以采用一些实用的调整方案，例如在聚类后的元类别上推断依赖关系，将推理限制在每个样本的前 K 个/高置信度标签上，或采用近似/贪心/基于采样的搜索或参数化依赖关系估计器，以少量近似误差换取可处理性。重要的是，这些调整方案保留了部署简便性：推理时只使用学生模型。

对教师质量的依赖：我们承认传播教师错误的风险。为缓解这一问题，教师使用硬标签进行训练，并结合渐进式自蒸馏，使其输出具有依据且校准效果更佳；ILI图根据聚合的软标签分布进行推断，并通过阈值处理生成稀疏图，以丢弃较弱的连接；此外，LIR 作为软正则项与监督损失共同使用，而不是作为硬监督。这些设计选择能够抑制虚假的教师信号，而不是强制采用这些信号。不同检查点之间推断边的实证稳定性进一步降低了模型对偶发教师噪声的敏感性。

流程复杂度：我们的方法采用协调一致的两阶段流程，而非两种不同的模型架构。教师模型和学生模型使用相同的模态编码器与嵌入架构。教师阶段从第 1 个训练轮次持续到转换轮次 M（例如 $M \in \{300,400,\ldots,900\}$），以学习经过校准的软标签；在$M$，我们仅推断一次稀疏 ILI 图；随后，学生阶段从 M 继续训练至 1000，并在损失函数中加入潜在交互正则化器（LIR）。因此，额外的复杂度来自训练阶段调度的选择，而非多模型部署成本：推理时仅使用紧凑的学生模型（不使用教师模型或图），因此运行时延迟、内存占用和工程集成都与标准的单模型基线相当。我们的消融实验（第 4.4 节，图 4）表明，在适当时机仅插入一次图即可带来稳定的性能提升，同时将额外训练工作量控制在有限范围内。

参考文献

1. Sami Abu-El-Haija、Nisarg Kothari、Joonseok Lee、Paul Natsev、George Toderici、Balakrishnan Varadarajan 和 Sudheendra Vijayanarasimhan。Youtube-8m：大规模视频分类基准。 *arXiv preprint arXiv:1609.08675*，2016年。2. Vedika Agarwal、Rakshith Shetty 和 Mario Fritz。迈向因果视觉问答：通过不变与协变语义编辑揭示并减少伪相关。载于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 9690 9698, 2020，页码 –。3. Galen Andrew、Raman Arora、Jeff Bilmes 和 Karen Livescu。深度典型相关分析。载于 *Proceedings of the 30th International Conference on Machine Learning* 28 *Proceedings of Machine Learning Research*，第 Vol. 卷, of，页码 pp.1247–1255，亚特兰大，佐治亚州，美国，2013 年 6 月 17–19 日。PMLR。4. Relja Arandjelovic 和 Andrew Zisserman。看、听、学。载于 *Proceedings of the IEEE international conference on computer vision* 609 617, 2017，页码 –。5. Ying Cheng、Ruize Wang、Zhihao Pan、Rui Feng 和 Yuejie Zhang。看、听与注意：用于自监督音视频表征学习的协同注意力网络。载于 *Proceedings of the 28th ACM International Conference on Multimedia*，MM '20，第 3884–3892 页，纽约，纽约州，美国，2020年。计算机协会。6. Hardoon David、R.、Szedm´ak S´andor 和 Shawe-Taylor John。典型相关分析：概述及其在学习方法中的应用。 *Neural Computation.* 16 12 2639 2664, 2004，第 Vol. 卷（no. ）：pp. –。7. Raia Hadsell、Sumit Chopra 和 Yann LeCun。通过学习不变映射实现降维。载于 *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)* 2, 1735 1742 2006，volume pages –。IEEE，。8. Alexander Hermans、Lucas Beyer 和 Bastian Leibe。为行人重识别中的三元组损失辩护。 *arXiv preprint arXiv:1703.07737*，2017年。9. 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。载于 *2017 ieeeinternational conference on acoustics, speech and signal processing (icassp)*，页码 131–135。IEEE，2017年。10. Joseph W Hogan。统计学中的因果推断入门：Judea Pearl、Maria Glymour 和 Nicholas Jewell 著，约翰·威立父子出版公司有限公司，英国奇切斯特，2019年。11. Prannay Khosla、Piotr Teterwak、Chen Wang、Aaron Sarna、Yonglong Tian、Phillip Isola、Aaron Maschinot、Ce Liu 和 Dilip Krishnan。监督式对比学习。 *Advances in neural information processing systems*，33：18661–18673，2020年。12. Jae Myung Kim、A Koepke、Cordelia Schmid 和 Zeynep Akata。揭示并缓解跨模态检索中的伪相关。载于 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*，页码 2585–2595，2023年。13. Diederik P Kingma。Adam：一种随机优化方法。 *arXiv preprint arXiv:1412.6980* 2014，。14. 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。 *Advances in Neural Information Processing Systems* 31 2018，,。15. Pei Ling Lai 和 Colin Fyfe。核典型相关分析与非线性典型相关分析。*Int. J. Neural Syst.*，第10卷（第5期）：pp.365–377，2000年。

16. Wai-Yin Lam、Bryan Andrews 和 Joseph Ramsey。最稀疏置换算法的贪心松弛。在 *Uncertainty in Artificial Intelligence*，第 1052–1062 页。PMLR，2022年。17. Liunian Harold Li、Mark Yatskar、Da Yin、Cho-Jui Hsieh 和 Kai-Wei Chang。Vi-sualbert：用于视觉与语言的一种简单且高性能的基线。*arXiv preprint arXiv:1908.03557* 2019，。18. Jiasen Lu、Dhruv Batra、Devi Parikh 和 Stefan Lee。Vilbert：为视觉与语言任务预训练与任务无关的视觉语言表征。*neural information processing systems* 32 2019，。19. Judea Pearl 和 Dana Mackenzie。*The book of why: the new science of cause and effect* 2018。Basic Books，。20. Jiaxin Qi、Yulei Niu、Jianqiang Huang 和 Hanwang Zhang。改进视觉对话的两项因果原则。在 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 10860 10869, 2020，第 – 页。21. Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sand-hini 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。在 *International conference on machine learning* 8748 8763 2021，第 – 页。PMLR，。22. Nikhil Rasiwasia、Dhruv Mahajan、Vijay Mahadevan 和 Gaurav Aggarwal。聚类典型相关分析。在 *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics* 823 831，，第 pp. – 页，冰岛雷克雅未克，2014年。JMLR.org。23. Joseph Roth、Sourish Chaudhuri、Ondrej Klejch、Radhika Marvin、Andrew Gal-lagher、Liat Kaver、Sharadh Ramaswamy、Arkadiusz Stopczynski、Cordelia Schmid、Zhonghua Xi 等。AVA主动说话人：用于主动说话人检测的音频-视觉数据集。在 *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 4492 4496 2020，第 – 页。IEEE，。24. Sebastian Ruder。梯度下降优化算法概述。*arXiv preprint arXiv:1609.04747* 2016，。25. Florian Schroff、Dmitry Kalenichenko 和 James Philbin。FaceNet：用于人脸识别和聚类的统一嵌入。在 *Proceedings of the IEEE conference on computer vision and pattern recognition* 815 823, 2015，第 – 页。26. Kihyuk Sohn。采用多类N对损失目标改进深度度量学习。*Advances in neural information processing systems*，29，2016年。27. Didac Suris、Amanda Duarte、Amaia Salvador、Jordi Torres 和 Xavier Giro-i Nieto。用于视频与音频检索的跨模态嵌入。在 *Proceedingsof the European Conference on Computer Vision (ECCV) Workshops*，2018年9月。28. Damien Teney、Ehsan Abbasnejad 和 Anton van den Hengel。通过数据反乱序提升视觉问答的泛化能力。在 *Proceedings of theIEEE/CVF International Conference on Computer Vision (ICCV)*，第 1417–1427 页，2021年10月。29. Yapeng Tian、Jing Shi、Bochen Li、Zhiyao Duan 和 Chenliang Xu。非受限视频中的音频-视觉事件定位。在 *Proceedings of the European conference on computer vision (ECCV)* 247 263, 2018，第 – 页。30. Amirsina Torfi、Seyed Mehdi Iranmanesh、Nasser Nasrabadi 和 Jeremy Dawson。用于跨音频-视觉匹配识别的3 *IEEE Access* 5 22081 22091, 2017 d卷积神经网络。，：–。31. Yanan Wang、Donghuo Zeng、Shinya Wada 和 Satoshi Kurihara。Videoad-visor：用于多模态迁移学习的视频知识蒸馏。 *IEEE Access*，11 51229: –51240，2023年。

32. Yixin Wang 和 David M Blei。多重原因的益处。 *Journal of the American Statistical Association* 114 528 1574 1596, 2019 ，( ): –。33. Xu Yang、Hanwang Zhang 和 Jianfei Cai。去混淆图像描述：因果回顾。 *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45 11 12996 13010, 2021 ，( ): –。34. Yi Yu、Suhua Tang、Kiyoharu Aizawa 和 Akiko Aizawa。基于类别的深度CCA：利用多模态数据进行细粒度场所发现。*neural networks and learning systems* 30 4 1250 1258, 2019 *IEEE transactions on* ，Vol. (no.):pp. –。35. Donghuo Zeng 和 Kazushi Ikeda。采用课程增强的两阶段三元组损失训练，用于视听检索。 *arXiv preprint arXiv:2310.13451*，2023年。36. Donghuo Zeng 和 Kazushi Ikeda。采用渐进式自蒸馏的度量学习，用于视听嵌入学习。在 *ICASSP 2025 - 2025 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)* 1 5, 2025 ，，第 – 页。37. Donghuo Zeng 和 Keizo Oyama。学习用于跨模态检索的联合嵌入。在 *International Conference on Data Mining Workshops (ICDMW)*，第 pp.1070–1071 页，中国北京，2019年。IEEE。38. Donghuo Zeng、Yanan Wang、Kazushi Ikeda 和 Yi Yu。用于视听检索的锚点感知深度度量学习。在 *Proceedings of the 2024 International Conference on Multimedia Retrieval* 211 219, 2024 ，第 – 页。39. Donghuo Zeng、Yanan Wang、Jianming Wu 和 Kazushi Ikeda。标签空间中的完整跨三元组损失，用于视听跨模态检索。在 *2022 IEEE International Symposium on Multimedia (ISM)* 1 9 2022 ，第 – 页。IEEE，。40. Donghuo Zeng、Jianming Wu、Gen Hattori、Rong Xu 和 Yi Yu。学习用于视听跨模态检索的显式和隐式双重公共子空间。*ACM Transactions on Multimedia Computing, Communications and Applications*，vol.19(no.2s):pp.1–23，2023年。41. Donghuo Zeng、Yi Yu 和 Keizo Oyama。通过监督式深度CCA学习用于跨模态音乐视频检索的视听嵌入。在 *2018 IEEE International Symposium on Multimedia (ISM)* 143 150 2018 ，第 – 页。IEEE，。42. Donghuo Zeng、Yi Yu 和 Keizo Oyama。采用聚类CCA的深度三元组神经网络，用于视听跨模态检索。 *ACM Transactions on Multimedia Com-puting, Communications and Applications*，vol.16(no.3):pp.1–23，2020年。43. Jiwei Zhang、Yi Yu、Suhua Tang、Wei Li 和 Jianming Wu。采用共享交叉注意力的多尺度网络，用于视听相关性学习。 *Neural Computing and Applications* 1 15, 2023 ，第 – 页。44. Jiwei Zhang、Yi Yu、Suhua Tang、Jianming Wu 和 Wei Li。采用CCA的变分自编码器，用于视听跨模态检索。 *ACM Trans. MultimediaComput. Commun. Appl.*，19(3s)，2023年2月。45. Kun Zhang、Shaoan Xie、Ignavier Ng 和 Yujia Zheng。从多个分布中进行因果表征学习：一般设定。 *arXiv preprint arXiv:2402.05052* 2024 ，。46. Lecheng Zheng、Yu Cheng、Hongxia Yang、Nan Cao 和 Jingrui He。用于多视图子空间学习的深度共同注意力网络。在 *conference 2021* 1528 1539, 2021 *Proceedings of the web* ，第 – 页。47. Yipin Zhou、Zhaowen Wang、Chen Fang、Trung Bui 和 Tamara L Berg。视觉到声音：为野外视频生成自然声音。在 *conference on computer vision and pattern recognition* 3550 3558, 2018 *Proceedings of the IEEE* ，第 – 页。