Crab$^+$：一种具有显式协作能力的可扩展统一视听场景理解模型

Dongnuan Cai[1, 2, *], **Henghui Du**[1, 3, *], **Chang Zhou**[3], **Xi Chen**[3], **Dan Guo**[4], **Hongyuan Zhang**[5], **李学龙**[2], **Di Hu**[1, ✉]

[1]中国人民大学高瓴人工智能学院， [2]中国电信人工智能研究院（TeleAI）， [3]腾讯PCG在线视频业务部人工智能技术中心， [4]合肥工业大学， [5]香港大学

*共同贡献， ✉ 通讯作者

arXiv:2603.04128v1 [cs.CV] 4 Mar 2026

摘要

构建面向统一场景理解的音视大语言模型（AV-LLM），是多模态智能中的一项关键追求。尽管指令微调能使预训练模型具备多任务能力，但我们观察到，传统的多任务统一方法往往遭受严重的负迁移，近55%的任务相较于单任务训练出现性能下降。我们将这一现象归因于音视任务异质性，其表现为任务粒度差异显著、能力需求各不相同，从而在联合训练下导致负向干扰。为此，我们提出**Crab$^+$**，一种可扩展的统一视听场景理解模型，通过数据和模型两个视角的显式协作来应对任务异质性。从数据视角出发，我们引入AV-UIE v2，一个全面的具有显式推理过程的视听统一指令微调数据集。它包含约222K个样本，涵盖17个数据集和7项任务，使模型能够发现任务特定的交互，并在不同粒度层面上捕捉跨任务关系。从模型视角出发，我们设计了统一输入输出接口，以对齐异构的任务表述。为解决差异化的能力需求，我们提出交互感知 LoRA（I-LoRA），通过动态路由显式建模任务间关系，以协调不同的音视交互模式，从而缓解参数干扰。大量实验表明，Crab$^+$ 比现有统一模型覆盖更广泛的任务范围，同时在多项基准测试上优于专用模型。我们成功扭转了负迁移趋势，实现了正迁移，使多任务学习在近88%的任务上超越单任务基线。这些结果在多种音视频大语言模型范式中均得到一致观察，并通过深入可视化得到验证，表明 Crab$^+$ 是迈向整体视听场景理解的坚实一步。

邮箱： Dongnuan Cai，联系地址： dongnuancai@ruc.edu.com，Di Hu，联系地址：dihu@ruc.edu.com**代码：** https://github.com/GeWu-Lab/Crab_Plus

1 引言

人类天生具备将视觉和听觉信息整合起来，以统一的方式完成日常活动的能力。例如，人类可以整合视觉和听觉线索来识别动作，融合这些模态来确定事件在时空中的位置，并执行对复杂音频-视觉信息进行推理，以回答特定问题。从认知科学的角度来看，这种统一能力有助于相关任务之间的迁移，从而促进任务间协作 [1]。因此，赋予机器类似的音视频数据处理能力，对于发展多模态智能至关重要。

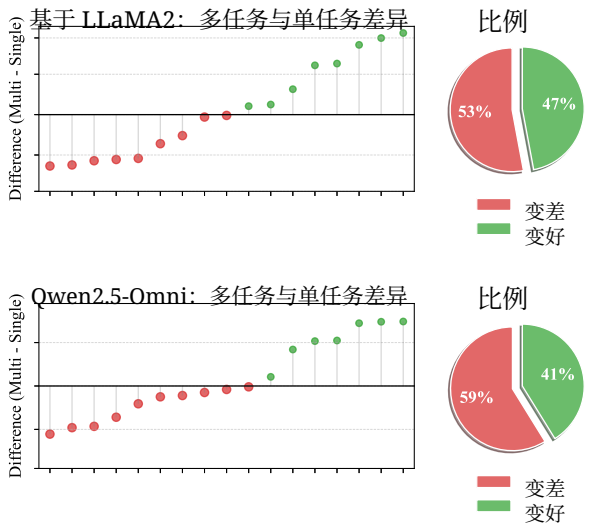


图1 定制化 LLaMA2 和 Qwen2.5-Omni 上的多任务与单任务 LoRA 对比。绿色和红色单元格分别突出显示多任务学习下性能提升和下降的任务。

基础模型的最新进展使范式从特定任务系统转向通用架构。在自然语言处理（NLP）领域，大语言模型（LLM）通过单一架构框架统一处理多种文本任务。受现实世界多模态特性的启发，音视大语言模型（AV-LLM）将大语言模型框架扩展至统一架构中，以执行广泛而复杂的音频-视觉场景理解任务。构建此类模型通常包括两阶段训练策略：首先进行预训练，使多模态特征与大语言模型的语义空间对齐；随后进行指令微调，赋予模型多任务能力。在第二阶段，大多数现有方法 [2, 3] 采用直接指令微调，使音视大语言模型能够执行各种视听任务。然而，鉴于视听任务固有的复杂性，这种直接的指令微调是否足以发挥统一学习的潜力，仍不明确。

为探究这一问题，我们遵循以往工作 [2, 3]，通过在标准指令–响应监督上进行 LoRA微调，对两种具有代表性的音视大语言模型（一个基于 LLaMA2 的模型 [4] 和 Qwen2.5-Omni [5]）进行基准测试。此类监督由独立标注的指令–输出对构成，涵盖17个数据集上的7种多样化视听任务。尽管多任务学习旨在利用任务间的协同效应，我们的实证结果却揭示了一种有悖直觉的现象：在近56%的评估设置中，朴素的任务统一效果不如单任务训练（图1）。我们将这种性能退化归因于任务固有异质性带来的两个根本瓶颈。首先，视听任务粒度各异，阻碍了统一学习下的有效协作。这些任务处于不同的任务粒度层级。例如，低层级定位任务涉及在时间和空间维度上对齐并建模音频-视觉信息。相反，高层级推理任务则需要基于时空输入，对语义、情感和动作进行因果理解。若没有中间表示，直接进行指令微调便无法弥合这些差异。因此，这一差距阻碍了任务间的有效协作。其次，视听能力需求的差异可能导致优化过程中的参数干扰。不同任务对模型提出了不同要求。时间定位任务需要模型具备随时间进行感知与对齐的能力。空间定位任务则要求在区域和像素级别上建模音频信号与视觉信息之间的对应关系。此外，时空推理任务要求模型能够在复杂语境中进行语义推理。现有的参数高效微调方法 [6]通常依赖静态且共享的适配方案，缺乏灵活性，无法适应此类差异化的交互模式，因此在联合优化时往往会导致负迁移。

为解决这些局限性，我们提出 **Crab**$^{+}$，这是一种可扩展的统一视听场景理解模型，通过从数据和模型两个角度开展显式协作，解决任务异质性问题。从**数据角度**来看，我们构建了 AV-UIEv2，这是一个大规模的**音频-视觉**统一**指令**微调**数**据集，包含**显**式推理过程，规模扩大至约 22.2 万个样本，涵盖 7 项任务和 17 个数据集。我们没有简单地聚合数据，而是将显式推理过程用作中间监督级表示，以减少任务粒度差异所导致的语义不一致。从**模型角度**来看，我们通过将所有任务目标表述为序列，建立统一的输入输出接口，将不同的特定任务输出映射为统一的序列表示。为缓解不同能力需求引发的参数干扰，我们进一步提出交互感知 LoRA（I-LoRA）。该模块通过采用动态路由机制，显式建模复杂的任务间关系，从而扩展了传统的静态适配。I-LoRA 使用可学习路由器，将输入词元自适应地路由至适当的 LoRA 头，以此解耦相互冲突的视听交互模式，从而满足不同的任务需求，实现有效的统一学习。如图2所示，Crab^{+} 支持在单一统一模型中执行多种视听场景理解任务，同时能够有效捕捉

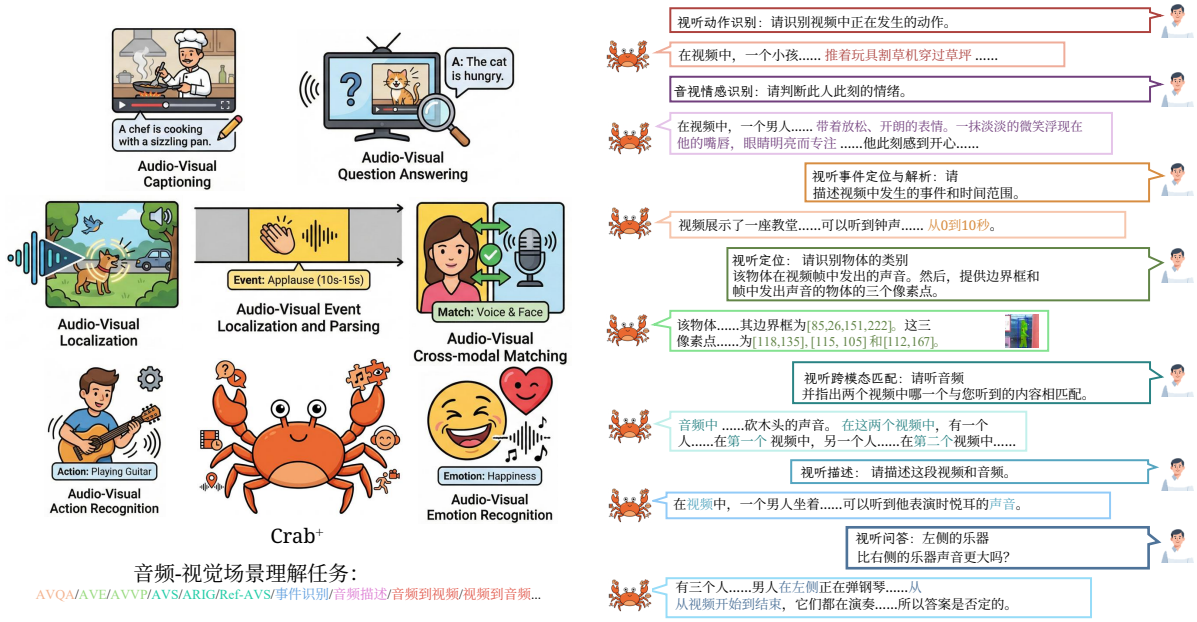


图2 Crab$^+$概览。（a）可扩展的统一架构，能够处理多样化的音视场景理解任务。（b）显式协作，捕捉复杂的跨任务依赖关系，从而提升总体性能。

促进协作的任务间依赖关系。

总体而言，与我们初步的会议版本相比，本研究在数据构建、模型架构和实验分析方面均有显著拓展。[7] 从数据角度来看，我们通过将 AV-UIE 升级为 AV-UIE v2，大幅扩展了视听场景理解的覆盖范围。此次升级将源数据集数量从 8 个增至 17 个，并将任务范围从 4 项扩展至 7 项，纳入了音视情感识别、动作识别和跨模态匹配等新任务。这些新增内容使模型能够适应范围更广、更加复杂的视听场景。从模型角度来看，我们改进了统一输入输出接口，在语义理解之外纳入像素级感知。这一设计无需再对不同模块进行分离优化，从而建立起简洁统一的单阶段训练流程。在实验分析方面，我们不仅对 Crab$^+$进行了性能基准测试，还将我们的方法应用于三种具有代表性的音视频大语言模型构建范式，以评估其泛化能力：配备视觉和音频分支的大语言模型（LLM+V+A）、具备音频能力的视觉大语言模型（V-LLM+A）以及原生音视大语言模型。大量可视化结果和实证分析进一步证实了我们的方法能够促进任务间协同。我们的主要贡献总结如下：

- 我们提出了Crab$^+$，一种可扩展且统一的音频-视觉场景理解模型，通过显式协作处理任务异质性，从而促进异质任务之间的正迁移。

•我们提出 AV-UIE v2，这是一个大规模数据集，包含近 222K 个样本，涵盖 7 项任务和 17个数据集。它将显式推理过程作为中间监督层级表示，以减少由任务粒度差异引起的语义不一致。

•我们设计了统一输入输出接口以规范任务执行，并结合交互感知 LoRA（I-LoRA）来缓解参数干扰。借助动态路由，I-LoRA 将相互冲突的音视交互模式解耦，以满足不同任务的需求。

2 相关工作

2.1 视听场景理解

音频-视觉场景理解通过整合听觉与视觉模态，实现对环境的全面感知 [8–10]。该领域涵盖多种任务，包括音频-视觉动作识别等基础识别任务 [11], 、时间定位等细粒度感知任务 [12, 13] 和分割任务 [14, 15], ，以及复杂推理，例如

视听问答 [16, 17]。传统研究主要依赖特定任务的架构。音频引导注意力机制等专用模块 [12] 以及层次时空网络 [17], 旨在针对特定任务目标对跨模态相关性进行建模。尽管这些模型在特定任务范围内表现有效，但跨任务泛化能力有限。近年来，AV-LLM 的出现 [3, 18] 标志着研究从专用专家模型向统一通用模型的重要转变。通过利用大语言模型丰富的语义先验，并将多模态特征对齐到统一的语义空间，这些模型能够在单一框架内实现更通用的音频-视觉理解。然而，通过指令微调赋予 AV-LLM 多任务能力会带来重大挑战，尤其是音频-视觉任务异质性导致的负迁移瓶颈。不同于忽视此类干扰的现有方法，我们的工作探究任务异质性的根本原因，并从数据和模型两个角度通过显式协作缓解这一瓶颈。

2.2 统一学习范式

对通用人工智能（AGI）的追求推动机器学习从专用模型转向统一框架。近期研究不再针对特定任务定制架构，而是日益重视具备广泛泛化能力的基础模型。本节将从架构统一和多模态整合两个角度，回顾这一范式的演进。

统一架构。过去，神经网络架构往往带有很强的归纳偏置，以适应特定的数据结构：卷积神经网络用于网格结构数据，循环神经网络用于处理序列依赖关系。Transformer [19] 的出现，通过引入共享的注意力机制，推动了架构统一。该架构将不同输入表示为词元序列，从而在很大程度上使网络设计摆脱了特定任务的限制。在此基础上，近期研究采用这一框架（通常以大语言模型为初始化模型）作为通用骨干网络 [20]。这种通用性有助于以可扩展的方式构建通用模型，使单一架构能够适应各种下游任务。

模态统一。与架构融合并行，机器感知已从彼此孤立的单模态分析，发展到集成式多模态理解。视觉—语言预训练是其中一项关键进展，对比学习框架（如 CLIP [21]和 ALIGN [22] ）将视觉与文本表征对齐到共享的语义空间。在此基础础上，多模态大语言模型（MLLMs） [23,24] 进一步将视觉信息整合进大语言模型，赋予模型全面的多模态理解能力。当前研究正突破视觉—语言的边界，纳入听觉信号。通过同时整合音频、视觉与语言模态，这些方法致力于构建统一接口，以无缝感知和分析复杂环境。

尽管统一学习范式为通用智能提供了有前景的基础，但将其应用于视听场景时，会受到任务异质性所引发的负向干扰的挑战。为此，我们致力于在统一框架内缓解此类干扰，通过在数据和模型层面显式处理异质性，从而实现更稳健的视听场景理解。

3 AV-UIEv2 数据集

正如第 1 节所分析的那样，不同视听任务的粒度存在显著差异。现有视听数据集大多缺乏能够衔接这些任务的合适中间表示，这限制了它们在统一学习中有效促进任务间协同的能力。为缓解这一问题，我们推出 AV-UIE v2，这是一个包含显式推理过程的音视频统一指令微调数据集。AV-UIE 是一个指令微调数据集，涵盖多种视听场景理解任务与场景。通过将显式推理过程用作中间表示，AV-UIE 在不同任务之间建立了合理的关联，从而促进统一学习。该数据集的初步版本已发表于我们的会议论文 [7]。在这项工作中，我们提出 AV-UIE v2，通过将源数据集数量从 8 个增加到 17 个，并将任务范围从 4 项扩展到 7 项，大幅提升了数据集的规模和多样性。新增任务包括情绪识别、动作识别，以及从 VGGSound [25] 派生的跨模态匹配子集。接下来的章节将详细介绍整体构建流程并提供统计分析。

3.1 数据集构建

我们通过整合多种现有音频-视觉来源来构建数据集。为解决这些数据集之间任务粒度不一致的问题，我们将原始标注转换为包含显式推理过程的详细文本描述。这为不同任务之间建立联系提供了统一的表征。如图3(a)所示，每个数据样本都包含原始视频和音频、真实标签，以及一个

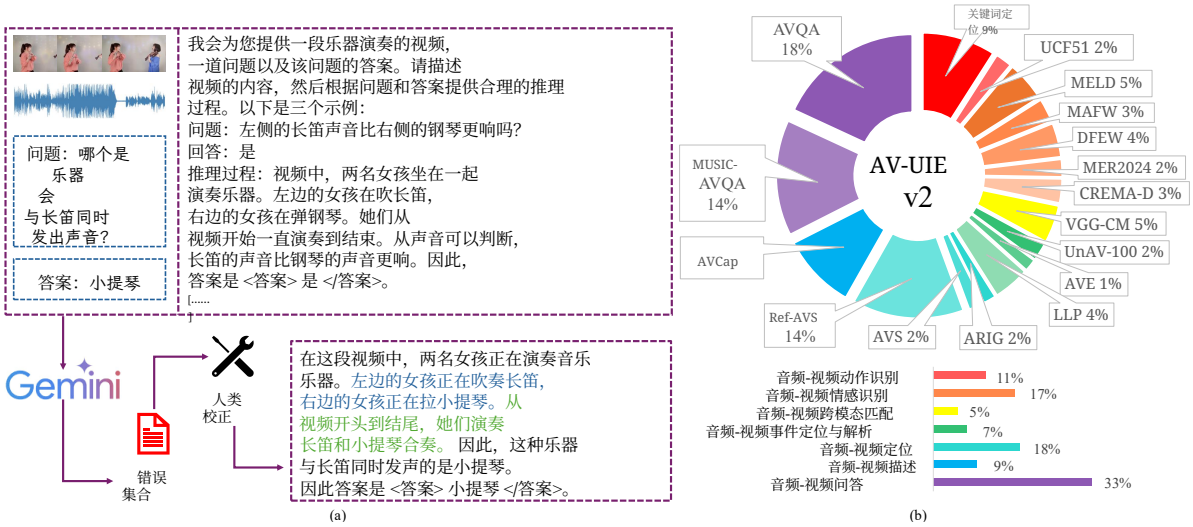


图3 AV-UIE v2 数据集概览。(a) 构建流程。我们利用多模态大语言模型，通过显式推理过程增强原始数据和简单标签。(b) 统计分布。该数据集整合了来自17个数据集的丰富源数据，涵盖7种不同的场景理解任务。

表1 AV-UIE v2 数据集的数据构成，按照任务复杂度从基础识别到复杂推理进行排列。

视听任务	源数据集	领域	样本（K）
动作识别	Kinetics-Sounds [26]	人体动作	20.1
	UCF51 [27]	人体动作	4.5
情感识别	MELD [28]	对话	11.0
	MAFW [29]	电影与电视	6.4
	DFEW [30]	电影与电视	9.3
	MER2024 [31]	电影与电视	4.8
	CREMA-D [32]	录音棚语音	6.6
跨模态匹配	VGG-CM [25]	开放域	10.3
事件定位与解析	UnAV-100 [33]	开放域	4.9
	AVE [12]	开放域	3.3
	LLP [13]	开放域	9.9
定位	ARIG [14]	开放域	3.4
	AVS [14]	开放域	4.0
	Ref-AVS [15]	开放域	30.7
图像描述	AVCap [34]	开放域	20.7
问答	MUSIC-AVQA [17]	音乐	31.9
	AVQA [16]	开放域	40.2

特定任务的提示。我们利用 Gemini 1.5 Pro [35] 通过上下文学习生成推理过程。为确保数据质量，我们实施了两步验证流程。首先，我们要求生成的推理在逻辑上得出原始真实标签。未通过此一致性检查的样本会自动丢弃。其次，我们对随机抽取的一部分有效样本进行人工审核。例如，如图3 (a)所示，一个仅标注为“提琴”的样本（回答“哪种乐器与长笛同时发声？”这一问题）被扩展为一段详细描述，阐述了两种乐器的声学特征和时间同步情况。

我们遵循这一流程来处理多个现有的数据集。这些数据集包括 MUSIC-AVQA [17],AVQA [16], AVE [12], UnAV-100 [33], LLP [13], MELD [28], CREMA-D [32], MAFW [29],DFEW [30], MER2024 [31], Kinetics-Sounds（KS） [26, 36], UCF51 [27], , 以及我们提出的 VGGSound 跨模态匹配（VGG-CM）数据集。为解决原始 LLP 训练数据缺少特定时间区间的问题，我们利用多模态大语言模型的时间推理能力生成区间标注。对于 AVS- Bench [14] 和 Ref-AVS [15], , 我们对输出格式进行调整，以支持空间定位和分割任务。对于定位任务，我们从对象掩码中提取边界框，如 $[x_{left}, y_{top}, x_{right}, y_{bottom}]$所示。对于分割任务，我们在此基础上通过预处理额外采样掩码内的三个点，以引导像素级定位。此外，我们还引入了 VALOR [34]提出的视听感知数据集，以丰富我们的格式。通过严格的质量控制协议，这一过程构建了 AV-UIE v2 数据集。

3.2 数据集分析

AV-UIE v2 数据集约包含 22.2 万个样本，每个样本都标注了显式推理过程，以建立不同任务之间的联系。如表 1所示，并在图3 (b)中可视化，数据分布遵循层级结构，旨在评估模型在不同粒度任务上的能力。在基础层面，基本识别任务占据了数据集的很大一部分，包括动作识别（11%）和情感识别（17%）。这些任务旨在从

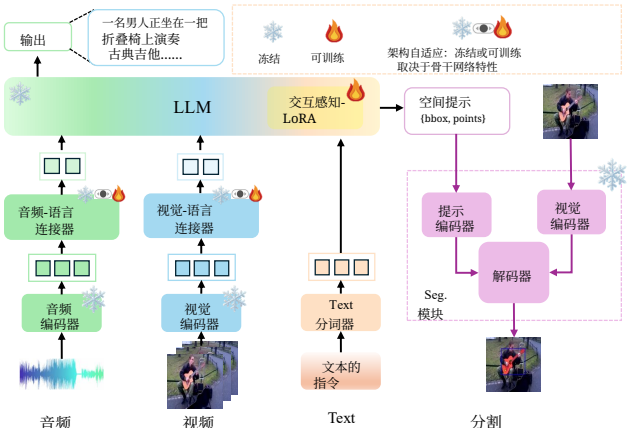


图4 我们的统一输入输出接口由视觉分支、音频分支、大语言模型骨干网络和分割模块组成，并使用交互感知LoRA在AV-UIE v2数据集上进行训练。

多模态输入（例如 Kinetics-Sounds 和 MELD）中识别人类活动和情感状态。在此基础上，数据集还纳入了需要精确跨模态对齐的细粒度时空感知任务，包括跨模态匹配（5%）、事件定位与解析（7%）以及定位（18%）。在这些任务中，模型需要将音频线索对应到相应的视觉空间或时间语境中（例如 Ref-AVS）。在最高层级，复杂推理任务在数据集中占比最大，包括图像描述（9%）和问答（33%）。这类任务要求整合多模态信息，生成语义描述或回答具有挑战性的问题（例如 MUSIC-AVQA 和 AVQA）。

4 方法

为解决第1节所讨论的局限性，我们提出了一种统一模型，通过一致的输入输出接口规范任务表述，并集成交互感知 LoRA（I-LoRA）以减轻任务间干扰。该方法的设计不依赖特定架构。我们主要在原生音视频大语言模型范式（Crab$^+$）上实现该方法，以检验性能上限，并利用其联合预训练实现的强跨模态对齐。为进一步展示我们方法的泛化能力，我们还将其扩展到复合范式（即 LLM+视觉+音频和视觉大语言模型+音频）。以下小节将详细介绍统一接口、架构实例化以及 I-LoRA 机制。

4.1 统一输入输出接口

为弥合不同任务粒度之间的差距，我们设计了一种统一接口，将所有任务目标表示为序列，如图4所示，从而实现精简的单阶段训练流程，并避免分离优化。

视觉分支。 视觉分支由预训练视觉编码器 $\mathcal{E}_V(\cdot)$和连接器 $\psi_V(\cdot)$组成，后者将视觉特征投影到大语言模型的嵌入空间中。给定长度为 T_v 的视频帧序列 $V = \{I_i \in \mathbb{R}^{H\times W\times C}\}_{i=1}^{T_v}$，其中 H、W和 C分别表示高度、宽度和通道数，每一帧被划分为 $N_v = \frac{H\times W}{P^2}$ 个互不重叠的图像块。这些图像块被线性映射到 D_v维嵌入空间，从而得到第 i帧的图像块级特征 $f_v^i \in \mathbb{R}^{N_v\times D_v}$。聚合后的视觉表示记为 $F_v = \{f_v^i\}_{i=1}^{T_v}$。随后，连接器 $\psi_V(\cdot)$对这些特征进行变换并展平，以匹配大语言模型的输入维度，生成最终的视觉令牌 $H_v \in \mathbb{R}^{L_v\times h}$，其中 L_v 为视觉令牌数量，h 为大语言模型的隐藏维度。

音频分支。 类似地，音频分支采用预训练音频编码器 $\mathcal{E}_A(\cdot)$ 和连接器 $\psi_A(\cdot)$，将音频特征与语言嵌入对齐。给定由 T_a 个音频片段 $A = \{S_i \in \mathbb{R}^{L_s}\}_{i=1}^{T_a}$ 组成的序列，其中 L_s 表示每个片段的长度，每个片段都会被映射到 D_a 维嵌入空间中，从而得到第 i 个片段的片段级特征 $f_a^i \in \mathbb{R}^{N_a\times D_a}$。聚合后的音频表示记为 $F_a = \{f_a^i\}_{i=1}^{T_a}$。随后，连接器 $\psi_A(\cdot)$ 对这些特征进行变换和展平，使其匹配 LLM 的输入维度，生成最终的音频令牌 $H_a \in \mathbb{R}^{L_a\times h}$，其中 L_a 为音频令牌的数量，h 为 LLM 的隐藏维度。

TokenFusion 与大语言模型骨干网络。 完成特征编码后，多模态令牌会与由分词器生成的文本令牌拼接。该统一序列被输入大语言模型骨干网络，以进行联合自回归生成，从而在单一语言建模接口内实现音频、视觉和文本模态之间的综合推理。

分割模块。 以往方法[37, 38] 通常通过专用标记将分割整合进多模态大语言模型中，往往需要多阶段训练。受 Seg-Zero 启发[39], 我们采用解耦的推理与分割方法，将 SAM2[40]作为冻结的分割模块，如图4所示。多模态大语言模型预测空间提示（例如边界框和点），随后利用这些提示向 SAM2 查询以生成掩码。该设计有效地将多模态定位与分割解耦，从而实现统一的

单阶段架构。形式化而言，给定训练帧 $I \in \mathbb{R}^{H\times W\times 3}$ 及其真实二值掩码$M \in \{0,1\}^{H\times W}$，我们将监督目标定义为$\mathcal{Y}_{\text{seg}}$。该目标由最小外接边界框 b 以及通过基于几何的采样策略得到的一组点组成。我们选取掩码区域内最大的三个内切圆的圆心，允许圆之间部分重叠，但要求任意两圆的交并比（IoU）不超过最大阈值。这些圆心作为点提示 $\{p_1, p_2, p_3\}$，构成监督集合：

$$\mathcal{Y}_{\text{seg}} = \{b, p_1, p_2, p_3\}. \tag{1}$$

在推理时，给定图像 I、音频 A和指令 X_{text}，模型直接预测以下空间参数：

$$(\hat{b}, \hat{p}_1, \hat{p}_2, \hat{p}_3) = \mathcal{F}_{\text{MLLM}}(I, A, X_{\text{text}}), \tag{2}$$

其中， $\mathcal{F}_{\text{MLLM}}(\cdot)$表示多模态大语言模型。随后，预测得到的空间提示由 SAM2 的提示编码器 $\mathcal{E}_{\text{prompt}}(\cdot)$ 编码，并通过掩码解码器 $\mathcal{D}_{\text{mask}}(\cdot)$与来自 SAM2 视觉编码器 $\mathcal{E}_{\text{vision}}(\cdot)$的图像特征结合，生成最终掩码：

$$\hat{M} = \mathcal{D}_{\text{mask}}\big(\mathcal{E}_{\text{prompt}}(\hat{b}, \hat{p}_1, \hat{p}_2, \hat{p}_3), \mathcal{E}_{\text{vision}}(I)\big). \tag{3}$$

基于上述分支和模块，我们的方法构建了统一的输入输出接口，可接收视听输入及特定任务指令，并生成相应的面向任务的输出。

4.2 架构实例化

原生音视频大语言模型实现。 Crab$^+$ 的主要实现采用原生音视频大语言模型范式，并在大规模音视频数据上进行预训练。我们基于 Qwen2.5-Omni [5]构建。视觉分支采用 Qwen2.5-VL [41] 骨干网络，这是一种采用窗口注意力以实现高效空间建模的视觉Transformer，后接一个用于词元压缩的合并器。音频分支采用 Qwen2-Audio [42]架构，使用一个通过线性投影层连接的 Whisper-Large-v3 [43] 编码器。

扩展至复合架构。 为评估我们的方法在多种架构上的鲁棒性，我们将其应用于通过轻量级连接器桥接各模态的复合范式。对于大语言模型+视觉+音频设置，我们采用 LLaMA2-7B-Chat [4] 和 LLaMA3-8B-Instruct [20] 作为文本骨干网络，并集成独立的视觉（CLIP-ViT-L/14 [21]）与音频（BEATs [44]）编码器。每个模态通过一个

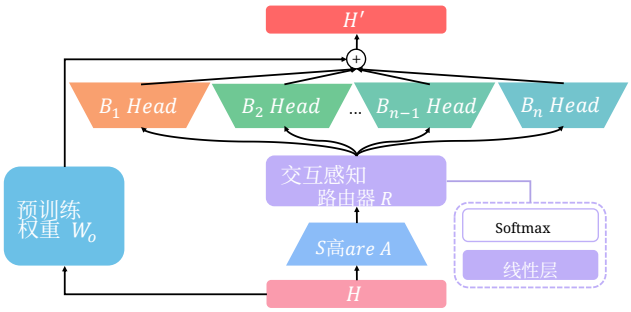


图 5 交互感知LoRA由一个共享矩阵 A和多个 B 矩阵组成，输入会动态路由到特定的 B 矩阵。

Q-Former [23]，随后接一个两层多层感知机。该设置用于检验该方法协调松散耦合模态的能力。相反，视觉大语言模型+音频设置在现有视觉语言模型 Qwen2.5-VL [41],的基础上，增加一条由 BEATs、窗口级Q-Former [45], 以及随后的两层多层感知机组成的音频分支。一个关键差异在于初始化：与受益于已具备先验跨模态对齐检查点的原生模型不同，这些复合架构组合的是独立编码器。因此，它们在指令微调之前需要一个初始的连接器对齐阶段。通过在这些变体以及原生 Crab$^+$上验证我们的方法，我们确保该方法在不同的模型构建范式下均具有鲁棒性。

4.3 交互感知LoRA

LoRA [6] 通过将权重更新分解为低秩矩阵，实现了较高的参数效率。然而，在多任务学习中，为不同任务强制采用共享的静态参数会限制灵活性。异构的任务需求往往会引发参数干扰，导致负迁移。尽管 Multi-LoRA [46] 通过为各个任务分配彼此独立的适配器来解决这一问题，但这种参数隔离会阻止模型利用音视频数据固有的共享表示。为此，我们提出交互感知 LoRA（I-LoRA），这是一种动态适配模块，能够在共享知识与任务专属化之间取得平衡。

如图5所示，I-LoRA 将权重更新分解为共享低秩矩阵 A、一组专用 LoRA 头 $\{B_i\}_{i=1}^{n}$和一个交互感知路由器 R。这一设计基于这样一个洞见：不同的视听任务共享共同的感知基础。因此，我们将这种通用能力显式编码在共享矩阵 A中，以确保参数效率和知识共享。为适应不同任务所需的不同交互模式，路由器R 和多个 B 头协同运作，以解耦

任务特定的适配与共享表示。不同于僵化的任务专属分配，I-LoRA 采用令牌级软路由策略，动态合成权重以满足每个令牌的语义需求，同时最大限度地减少统一模型中的参数干扰。

形式上，输入 H 通过沿序列维度拼接视觉令牌 $H_v \in \mathbb{R}^{L_v \times h}$、音频令牌$H_a \in \mathbb{R}^{L_a \times h}$和文本令牌 $H_t \in \mathbb{R}^{L_t \times h}$ 构造而成：

$$H = [H_v; H_a; H_t] \in \mathbb{R}^{L \times h}, \tag{4}$$

其中 $L = L_v + L_a + L_t$ 是总序列长度。在大语言模型骨干网络中，I-LoRA 应用于冻结线性变换 $W_0 \in \mathbb{R}^{d \times h}$。该架构采用共享低秩矩阵 $A \in \mathbb{R}^{r \times h}$ 进行初始下投影。交互感知路由器由 $W_r \in \mathbb{R}^{n \times r}$ 参数化，并利用投影特征 $HA^\top$ 计算路由分数：

$$S = \text{softmax}((HA^\top)W_r^\top) \in \mathbb{R}^{L \times n}, \tag{5}$$

其中 $S_{t,\ i}$ 表示第 i个头中第 t个令牌的路由权重。更新量 ΔH 对 n 个头 $\{B_i \in \mathbb{R}^{d \times r}\}_{i=1}^{n}$的输出进行聚合，并按照各自的路由分数加权：

$$\Delta H = \sum_{i=1}^{n} \text{diag}(S_{:,i})(HA^\top B_i^\top), \tag{6}$$

其中 diag($S_{:,\ i}$) $\in \mathbb{R}^{L \times L}$ 实现逐令牌门控。最后，经过 I-LoRA 增强的层的输出为：

$$H' = HW_0^\top + \Delta H, \tag{7}$$

其中， $H' \in \mathbb{R}^{L \times d}$ 是更新后的隐藏表示。

我们提出了一种统一接口，将三种音视频大语言模型范式中的任务标准化为单阶段流水线，并在大语言模型的标准下一令牌预测目标下进行训练。借助 I-LoRA 模块，我们有效地平衡了共享参数与特定参数之间的权衡。该方法调和了任务异质性所导致的冲突，为复杂环境中的统一学习释放了协同效应。

5 实验

本节对我们的方法进行了全面评估。我们首先介绍实验设置，包括任务、数据集和指标。随后，我们详细说明了各种音视频大语言模型架构的训练策略及其实现。我们的分析首先从将该方法与具有代表性的基线进行比较，涵盖统一的音视大语言模型和专用的非大语言模型。随后，我们比较单任务与多任务学习，并对 I-LoRA 进行可视化，以验证我们的设计依据。我们还开展了广泛的消融研究，以考察 I-LoRA 的有效性、通过移除头来评估头的作用，以及对 B-头数量的敏感性。最后，我们通过定性案例研究，展示该方法处理复杂视听语境的能力。

5.1 任务、数据集与评估指标

音视动作识别。该任务旨在从视听流中识别人类动作。我们在 Kinetics-Sounds（KS）[26],（涵盖 31 个动作类别）和 UCF51（UCF101[27] 的 51 类子集）上进行评估。对于 KS，我们采用官方数据集划分；对于 UCF51，则采用第一个预定义划分，并报告 Top-1 准确率。

音视情感识别。该任务根据视听输入预测所表达的情绪。我们在 CREMA-D [32], 、MAFW [29], 、DFEW [30],、MER2024 [31], 和 MELD [28] 上进行评估。CREMA-D 包含表演式情感语音，而其他数据集则由来自电视或电影的自然表情组成。对于 CREMA-D，我们遵循 [59–61]的划分协议；对于 MAFW 和 DFEW，采用第一个官方划分；对于 MELD，采用官方划分。MER2024 在其“训练与验证集”上进行训练。遵循既有基准测试，我们报告 CREMA-D 的准确率，MELD 的准确率和加权 F1（WAF），以及其余数据集的非加权平均召回率/加权平均召回率（UAR/WAR）。

跨模态匹配。这项二元分类任务旨在判断音频片段与视频是否对应同一事件。我们评估了提出的 VGGSound 跨模态匹配（VGG-CM）基准，并报告准确率。

视听事件定位与解析。事件定位用于确定视听事件的时间边界，而解析则将这些事件分类为单模态或多模态。我们采用 AVE [12] 和 UnAV-100 [33] 进行定位，并采用 LLP [13] 进行解析。AVE数据集包含 28 个类别的 4,143 个 10 秒片段，而 UnAV-100 涵盖未剪辑视频中来自 100 个类别的 30,059 个标注事件。LLP 包含 11,849 个精选 YouTube 视频。按照标准协议，我们报告 AVE 上的定位准确率，以及 LLP 上片段级和事件级（交并比 0.5）的 F1 分数。为降低 UnAV-100 长序列的计算成本，我们使用时间动态稳定的抽样视频片段子集进行训练，从而在减少输入数据的同时保持事件连贯性。

表2 与统一音频-视觉大语言模型的性能比较。

方法	MUSIC ACC	AVQA ACC	AVE ACC	AVVP 片段级	AVVP 事件级	ARIG 交并比	ARIG 曲线下面积	AVS S4 $\mathcal{J}$	AVS S4 $\mathcal{F}$	AVS MS3 $\mathcal{J}$	AVS MS3 $\mathcal{F}$	Ref-AVS Seen $\mathcal{J}$	Ref-AVS Seen $\mathcal{F}$	Ref-AVS 未见 $\mathcal{J}$	Ref-AVS 未见 $\mathcal{F}$	CRE ACC	DFEW 加权平均召回率	DFEW 非加权平均召回率	MAFW 加权平均召回率	MAFW 非加权平均召回率	MELD 加权平均 F 值	MELD 非加	KS ACC	UCF51 ACC	VGG-CM A→V ACC	VGG-CM V→A ACC
TimeChat [47]	–	–	–	51.28	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
猫鼬 [48]	–	–	–	54.96	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
GroundingGPT [49]	–	–	–	–	–	44.02	45.00	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
AnyRef [50]	–	–	–	–	–	–	–	**82.80**	**90.80**	55.60	**66.30**	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
SimToken [51]	–	–	–	–	–	–	–	–	–	–	–	**72.00**	**81.30**	**69.80**	**79.10**	–	–	–	–	–	–	–	–	–	–	–
PandaGPT [3]	33.70	79.80	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
MacawLLM [52]	31.80	78.70	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Video LLaMA [18]	36.60	81.00	–	–	–	–	–	–	–	–	–	–	–	–	–	–	35.75†	26.77†	–	–	–	–	–	–	–	–
Video LLaMA 2 [53]	73.60	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
CAT+ [54]	**86.40**	**92.60**	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
OMCAT [55]	73.80	90.20	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
UIO-2 [56]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	89.30	–	–	–
VITA [57]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	32.07†	21.36†	33.38†	14.05†	–	–	–	–	–	–
Emotion-LLaMA [58]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**77.06**†	**64.21**†	–	–	–	–	–	–	–	–
Crab [7]	78.94	–	80.15	59.00	54.44	41.78	42.00	73.25	86.81	**58.21**	66.24	40.54	58.00	45.55	63.00	–	–	–	–	–	–	–	–	–	–	–
Crab+	81.09	92.16	–	**59.47**	**55.79**	–	–	–	–	–	–	–	–	–	–	–	–	55.83	–	–	–	–	–	–	–	–

注：符号 – 表示该任务不适用，或未报告结果。符号 † 标识包含多个标准划分的数据集；我们仅报告第一个划分上的结果，以 适应多任务学习的计算复杂度。

视听定位。该任务旨在分割并定位发声物体。我们使用 AVSBench [14]，它为单声源（S4）和多声源（MS3）设置提供像素级掩码。为评估参考定位，我们还纳入了包含已见和未见子集的 Ref-AVS。与以往工作一致，我们对 S4 使用第一帧，对 MS3 使用均匀抽样的帧。此外，S4 的边界框用作音频指代图像定位（ARIG）的标注。对于 Ref-AVS，我们从每个 10 秒片段中抽取代表性帧。我们使用杰卡德指数 $\mathcal{J}$ 和 F分数 $\mathcal{F}$ 报告性能。对于 ARIG，我们采用交并比（IoU）和曲线下面积（AUC）作为评估指标。

视听描述。此任务生成能够反映音视语义的自然语言描述。我们使用 VALOR [34] 的子集来增强模型的基础音视感知能力。

视听问答。此任务需要利用联合的声学和视觉线索推断答案。我们在 MUSIC-AVQA [17], 和 AVQA [16], 上评估方法；前者包含涵盖 22 种乐器类别的音乐场景，后者包含 57.3K 个问答对，涉及 9 种问题类型。我们根据官方划分报告准确率。

5.2 训练策略

我们针对每种模型的特定架构定制训练流程。由于 LLM+V+A 和 V-LLM+A 缺乏固有的音视先验，我们分两个阶段训练它们：(1) 特定模态连接器预训练，随后进行 (2) 音视多任务指令微调。原生音视频大语言模型已具备基础的音视感知能力，因此直接进行单阶段指令微调。

阶段1：模态专属连接器预训练。对于大语言模型+视觉+音频，视觉连接器在一个来自 VideoLLaVA [62], 的数据集上进行预训练，同时冻结所有其他参数，从而建立基本的视觉感知。对于大语言模型+视觉+音频和视觉大语言模型+音频中的音频处理，音频连接器在 AudioSet [63]上进行预训练。由于 AudioSet 的标签较为粗略，我们使用 Qwen2Audio [42] 生成字幕作为监督信号。仅更新音频连接器参数。

阶段2：音频-视觉多任务指令微调。三种范式均在 AV-UIE v2 数据集上进行微调。默认情况下，仅优化交互感知LoRA。对于大语言模型+视觉+音频，我们额外微调音频与视觉连接器，以提升感知对齐。对于视觉大语言模型+音频和原生音视频大语言模型，连接器保持冻结，以维持其预训练编码器的跨模态一致性。

5.3 实现细节

在音视情感识别任务中，鉴于视频片段时长较短，我们对每个视频采样 224×224帧；对于所有其他任务，则使用 10 帧。所有视觉输入均调整为 K。原始音频波形重采样至 16 kHz。视觉和音频 Q-Former 的可学习 token 数量 [40]均设为 32。对于窗口级Q-Former，每个 0.33 秒的音频片段映射为一个可学习 token，因此 30 秒的片段共有 88 个 token。分割模型使用预训练的 SAM2.1-Large 权重 +进行初始化。我们根据实验设置调整交互感知LoRA模块的超参数。为最大限度地发挥 Crab 中原生音视频大语言模型 $r = 128$的能力，我们将秩 $\alpha = 256$、缩放因子以及dropout $r = 16$率设为 0.10。为与第 1节中的分析实验（验证多任务有效性）保持一致，我们使用 B 、 $\alpha = 32$，并将 dropout 率设为 0.05。LoRA 矩阵采用 3 个头。

表3 识别任务上的方法比较。

方法	CREMA-D	MELD	DFEW	MAFW	KS	UCF51
OGM-GE [59]	64.34	–	–	–	66.35	–
MMPareto [64]	70.19	–	–	–	69.13	–
CAL [65]	79.30	–	–	–	74.82	–
DialogueRNN [66]	–	56.12	–	–	–	–
DQ-Former [67]	–	64.88	–	–	–	–
ACoSDN [68]	–	**68.04**	–	–	–	–
DFER-CLIP [69]	–	–	71.25†	52.59†	–	–
VAEmo [70]	–	–	75.78†	58.91†	–	–
AVF-MAE++ [71]	–	–	**77.45**†	**60.24**†	–	–
AttenClus [72]	–	–	–	–	73.91	84.79†
DMRN [12]	–	–	–	–	77.50	82.93†
MAFnet [73]	–	–	–	–	83.94	86.72†
Crab+	**84.41**	52.08	64.34	45.60	**91.12**	**94.04**

注：† 表示所有划分的平均结果。由于多任务学习的计算复杂度较高，我们使用第一个划分进行评估。

表4 时间定位任务上的方法比较。

方法	AVVP		AVE
	片段级	事件级	
AVT [74]	–	–	75.80
MM-Pyramid [75]	59.20	53.04	77.80
CMBS [76]	55.00	48.48	–
DHHN [77]	**60.32**	55.06	–
MoLT [78]	–	–	83.50
Crab+	59.47	**55.79**	**83.58**

表5 空间定位任务上的方法比较。

方法	ARIG	S4	MS3	已见	未见
LVS [79]	23.69	–	–	–	–
FNAC [80]	27.15	– – –			–
AVSBench [14]	–	78.70	54.00	23.20	32.36
AVSegFormer [81]	–	**83.06**	**61.33**	33.47	36.05
EEMC [15]	–	–	–	34.20	49.54
SAM2-LOVE [82]	–	–	–	43.50	66.50
Crab+	**79.62**	75.75	52.64	**65.45**	**68.98**

5.4 结果与分析

与音视统一模型的比较。表2对我们的方法与现有音视大语言模型进行了比较分析。与以往仅适用于有限任务子集的模型不同，我们的模型涵盖了范围更广的视听场景。值得注意的是，Crab+ 相较于 Crab 实现了显著进步，其任务执行能力覆盖范围更广，并在大多数基准测试中取得了更优的性能。与其他音视统一模型相比，我们的方法在 AVE（83.58 对 80.15）、AVVP（59.47 对 59.00）和关键词定位（91.12 对 89.30）等任务上表现更优。此外，在 AVS、Ref-AVS 和 DFEW 等任务上也取得了有竞争力的结果。这表明，我们提出的方法有效地平衡了任务多样性与特定任务上的专业能力。

与专用音视模型的比较。尽管我们的统一模型旨在实现通用视听场景理解，但与特定任务的专家模型相比，它仍然表现出色。在表中报告的识别

表6 AVQA任务上的方法比较。

方法	MUSIC-AVQA				AVQA
	音频	视觉	音频-视觉	总体	
ST-AVQA [17]	73.87	74.40	69.53	71.59	–
PSTP-Net [83]	70.91	77.26	72.57	73.52	90.20
SaSR-Net [84]	73.56	73.28	74.66	74.21	89.90
QA-TIGER [85]	78.58	85.14	73.74	77.62	–
HCRN+HAVF[86]	–	–	–	–	89.00
Crab+	**79.44**	**92.45**	**76.20**	**81.09**	**92.16**

任务结果3凸显了多任务整合本身的难度。尽管在若干情绪识别基准上仍存在轻微的性能差距，我们的模型在 CREMA-D 上表现稳健，得分为 84.41；在动作识别方面则取得更优结果，KS 上达到 91.12，UCF51 上达到 94.04。对于时间定位，如表4所示，我们的方法仍极具竞争力。在 AVE数据集上，其准确率达到 83.58，略优于此前表现最佳的方法。在 AVVP任务上，模型取得了最高的事件级得分 55.79，同时在片段级上保持了相当的性能。我们模型的优势在空间定位任务上尤为突出。表5显示，我们的方法相较于专门设计的基线方法取得了显著提升，比先前方法高出约 +52.47（ARIG）和 +21.95（Ref-AVS（已见）子集）。此外，正如表+ 所示，Crab 在问答任务上也表现出色6。得益于强大的视觉感知能力，其在 MUSIC-AVQA 上的总体准确率达到 81.09，在 AVQA基准上达到 92.16。总而言之，这些结果表明，所提出的统一模型达到了可与特定任务方法相媲美的性能。

单任务与多任务。 在验证了 Crab 的性能+之后，我们进一步研究了统一学习方法在不同 AV-LLM 构建范式（例如，LLM+视觉+音频和视觉大语言模型+音频）中的泛化能力。我们使用 I-LoRA，在 AV-UIE v2 数据集上对单任务和多任务设置进行了比较分析。为确保比较公平，我们保持训练配置一致，仅改变数据组成：单任务模型在各自的数据集上训练，而多任务模型则在涵盖所有任务的综合数据集上训练。如图6所示，在我们的统一学习方法下，不同范式的音视大语言模型在多任务设置中始终优于单任务基线方法。从量化结果来看，我们的方法在 94% 的任务上取得了性能提升，仅在其余任务中观察到轻微的负迁移，净收益高达 88%。这与第 1节的分析形成对比：朴素的多任务学习（采用相互独立的指令遵循，并使用

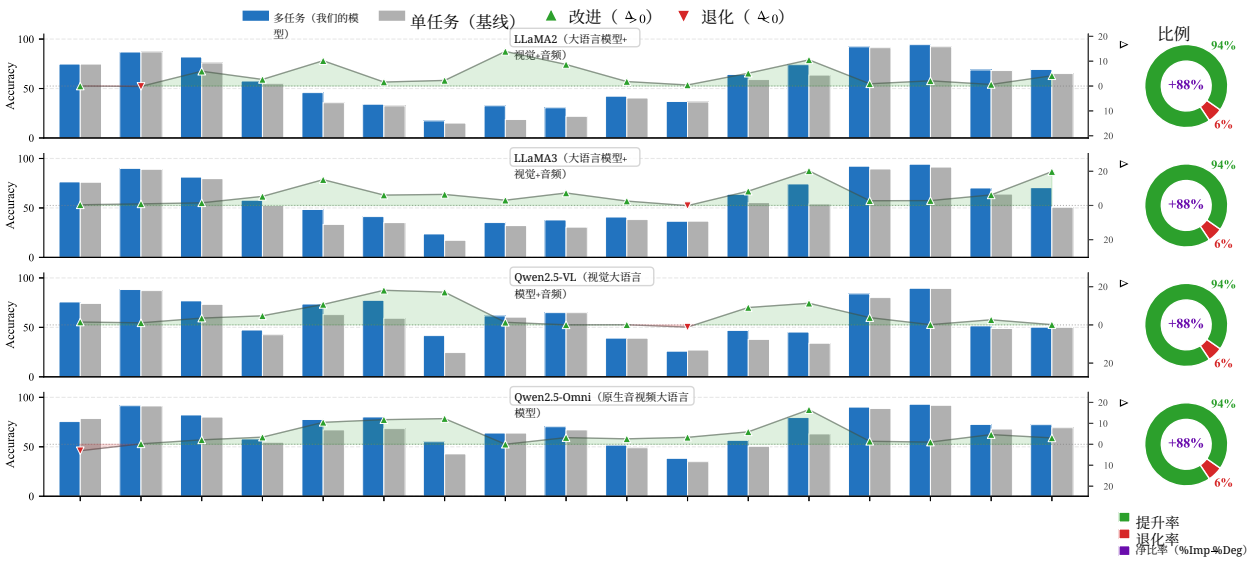


图6 三种代表性音视频大语言模型范式下，多任务与单任务学习的协同效应对比分析。对于每个模型，我们报告表现提升的任务比例和净得分；净得分定义为受益于多任务学习的任务百分比与遭受负迁移的任务百分比之差。结果表明，我们的方法有效缓解了任务异质性，促进了任务间协同。

LoRA）会导致约 55% 的任务出现性能退化，净收益为 -10%。我们的结果表明，通过将正向收益率从 45% 提高到 94%，并将净收益从 -10% 提高到 +88%，我们的方法有效缓解了任务异质性，从而促进任务之间稳健的正迁移与协同效应。

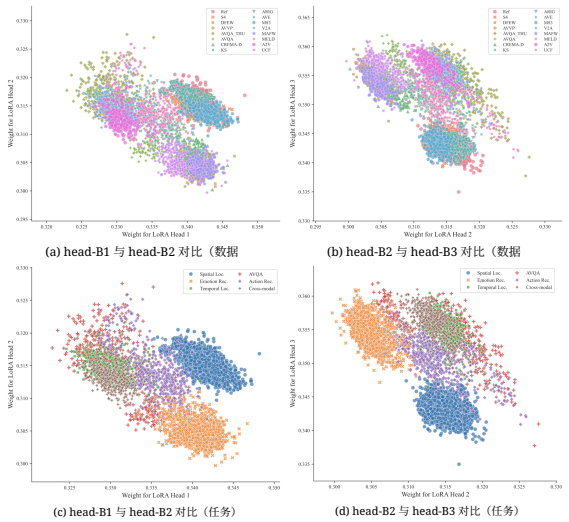


图7 路由器权重的可视化。头对在（a）–（b）中按数据集、在（c）–（d）中按任务进行比较。需求相似的任务形成不同的簇，表明该方法能够厘清复杂的音视频模式。

可视化分析。为了考察内部适应-I-LoRA的能力，我们分析了路由器权重、激活分布和参数相似度（图7–9）。尽管I-LoRA采用令牌级软路由而非任务级硬路由，路由器权重仍呈现出清晰的语义聚类。簇不仅出现在不同数据集之间，也出现在不同任务之间，例如空间定位（如ARIG和Ref-AVS），同一任务的数据集会形成紧凑的组。这表明路由器能够有效捕捉任务特定语义。我们进一步分析了注意力头级激活行为。总体而言，所有头在多样输入上均保持相对均衡的激活，平均激活概率约为0.3。尽管存在这种均衡，不同的头仍表现出不同的功能偏好：head-B1侧重于空间定位和情感识别，head-B2强调时间定位，head-B3专门负责AVQA和跨模态匹配。为验证这种功能分化，我们计算了三个头的展平权重向量之间的成对相似度，并在所有层和LoRA模块上取平均。平均相似度为0.630，表明在角色专门化的同时存在部分参数共享。逐层分析显示，浅层相似度较高，更深层相似度逐渐降低，表明随着网络深度增加，任务特定分化愈发明显。

LoRA消融研究。为验证 I-LoRA 的有效性，我们将其与LoRA基线进行基准测试。我们在 AV-UIE 上采用统一的多任务训练方法，

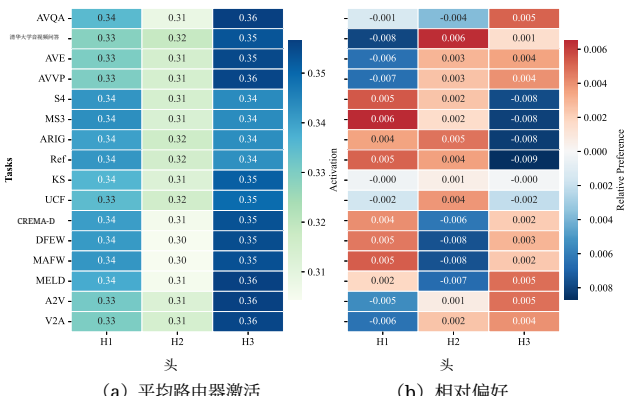


图8 路由器激活分析。(a) 平均激活显示出均衡的贡献（≈ 0.3）。(b) 相对模式揭示了特定任务的头偏好。

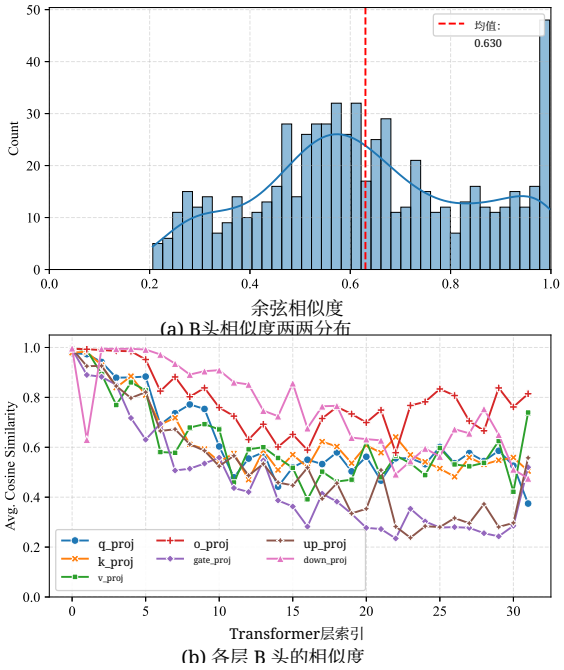


图9 路由矩阵的相似性分析 B。(a) 总体相似性（均值 0.630）。(b) 逐层相似性，显示更深层中的差异。

表7 丢弃注意力头的结果。

方法	CREMA-D	AVQA	MUSIC	AVVP	KS	ARIG 音频→视觉	视觉→音频
丢弃头-1	74.06	86.57	71.11	51.91	91.01	40.22	68.14 66.15
丢弃头部-2	72.37	85.30	71.55	54.65	91.57	43.94	67.01 67.83
丢弃头部-3	71.76	84.34	69.80	39.72	90.76	44.40	67.88 67.14
不丢弃	**74.12**	**86.92**	**74.73**	**59.77**	**92.21**	**45.78**	**68.81 69.28**

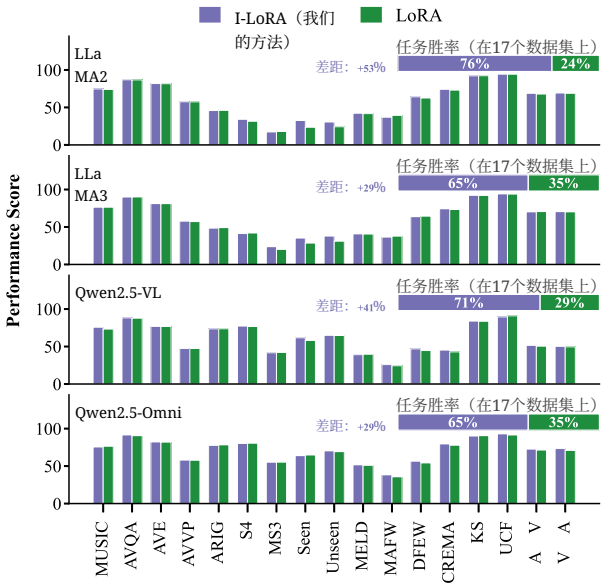


图10 关于 LoRA 的消融结果。虽然 LoRA 在部分任务上有所提升，但 I-LoRA 显著增加了具有协同效应的任务数量。

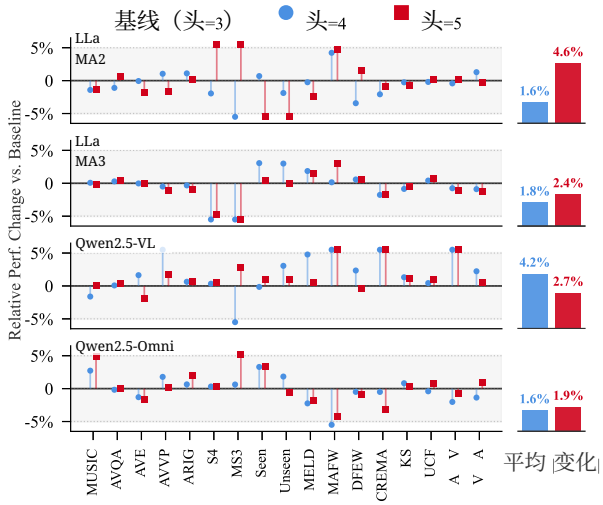


图11 对 B 头数量的消融。我们的方法在该超参数上保持稳定，没有显著波动。

v2数据集，旨在缓解数据中固有的任务异质性。如图 10所示，结果表明，跨任务联合训练本身就能带来性能提升，LoRA 在有限的一部分任务上取得了更好的结果便是明证。这一观察体现了多任务学习的内在潜力。通过任务间交互显式解决参数干扰问题，I-LoRA 进一步放大并稳定了这种正向迁移，从而在大多数基准测试中取得更优表现。在不同的模型架构中，

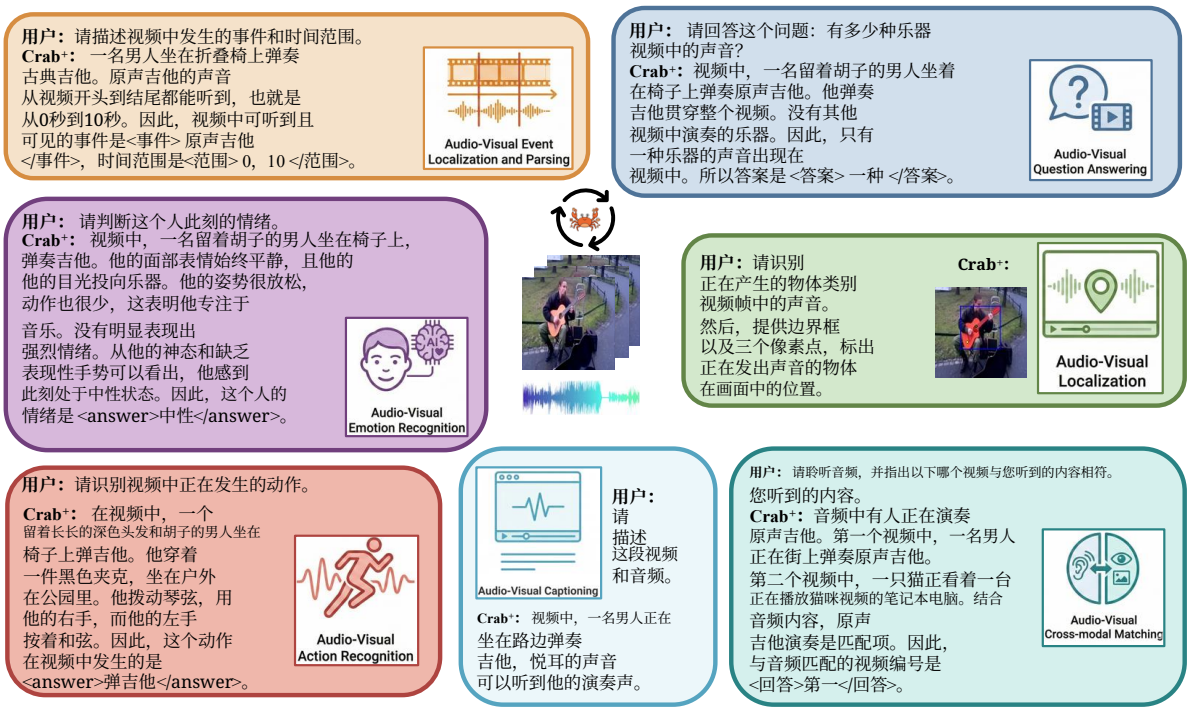


图12 定性结果。Crab$^+$ 能够在单个样本上执行多种不同的音频-视觉场景理解任务。

我们的方法在 65%–76% 的任务中取得了最佳性能；就任务覆盖率而言，相较于 LoRA基线，绝对提升了 29%–53%。

头部消融研究。为了探究每个 LoRA 头的作用，我们采用了“头部剔除”策略，将 LLaMA2 骨干网络中特定头部的参数置零。如表 7 所示，移除任何一个头部都会导致所有任务的性能持续下降，涵盖 AV QA、时间定位（AVVP）、空间定位（ARIG）和跨模态感知（$A \rightarrow V$， $V \rightarrow A$）。这些结果表明，每个头部捕获的都是互补信息，而非高度相似或冗余的表征，从而共同对视听空间进行建模。此外，各任务性能退化程度的差异与我们此前的可视化分析一致。定量结果与定性结果相互印证，进一步支持了我们设计的合理性。总体而言，结果表明，I-LoRA 适用于处理复杂音频-视觉场景理解中的任务异质性，有助于任务间协作，并实现更有效的统一学习。

头数量的消融研究 B**-头**。我们进一步研究 I-LoRA 对B-头数量的敏感性。直观而言，头的数量决定了适应子空间的广度。头的数量越多理论上可为新任务提供更大的容量。然而，这种配置需要权衡。数量过少会限制对复杂任务的建模能力；反之，数量过多则难以建立头与头之间的联系，从而实现有效的统一学习。为评估稳定性，我们将 $N = 3$ 设为基线，并在不同的 AV-LLM 上测试将头的数量增加到 4 和 5 的影响。随后，我们以该基线为参照，分别展示特定任务的偏差和总体平均偏差。如图 11所示，敏感性因骨干架构而异。基于 LLaMA 2 的模型在 $N = 5$时偏差最大，为 4.6%。与此同时，Qwen2.5-Omni 的偏差最小，为 1.6%。尽管存在这些差异，所有 AV-LLM 的性能波动相对于基线均保持在合理的 5% 范围内。这些结果证实了我们方法的稳定性。

定性案例研究。图12 展示了一个具有代表性的视频片段的定性分析：一名男人在路边弹奏原声吉他。该模型在多种视听任务中表现稳定。分析逐步展现了模型的能力。从基础识别开始，模型准确识别出弹奏吉他的动作，并推断出演奏者的情绪状态。在进一步进行细粒度感知时，它能精准定位事件的时间边界，并分割出画面中的发声乐

器。尽管存在干扰项，它也能根据输入音频识别对应的视频片段，成功完成跨模态匹配。最后，在复杂推理方面，模型能够理解并综合音频和视觉内容，正确回答“视频中有几种乐器发声？”这一问题。该案例展现了低层感知与高层推理的无缝整合，验证了其作为统一视听助手的潜力。

6 讨论

我们提出 Crab$^+$，一个可扩展且统一的视听场景理解模型，旨在缓解朴素多任务指令微调中观察到的严重负迁移。Crab$^+$ 并不将任务异质性视为障碍，而是从数据和模型两个视角，通过显式协作来调和这些差异。在数据方面，我们构建 AV-UIE v2，借助显式推理过程弥合输出差异。在模型方面，我们设计统一的输入-输出接口以规范多样的任务结构，并提出交互感知 LoRA（I-LoRA），动态协调相互冲突的能力需求。通过对齐这些异构任务，Crab$^+$ 成功扭转了性能退化的趋势，将负向干扰转化为正协同，并为通用视听场景理解确立了稳健的基线。

局限性。尽管显式推理有益于复杂场景，但对较简单的任务应用冗长的推理链可能引发过度解读，从而引入噪声并干扰直接预测。因此，开发能够使推理粒度与任务复杂度动态对齐的自适应数据策略，仍是一个开放性挑战。

未来工作。展望未来，我们计划扩展统一模型，将生成式任务纳入其中，进一步推动通用音频-视觉场景理解的发展。

参考

[1] N. A. Taatgen, 《认知技能的本质与迁移》。心理学评论，第 120 卷，第 3 期，第 439 页，2013 年。

[2] Y. Tang 等，“通过伪未剪辑视频赋能大语言模型进行视听时序理解”，载于 人工智能促进会会议论文集，第 39 卷，第 7 期，2025年，第 7293–7301 页。

[3] Y. Su 等，“PandaGPT：一个可通过指令完成所有任务的模型”，*arXiv* 预印本 *arXiv:2305.16355*，2023年。

[4] H. Touvron等，“Llama 2：开放基础模型与微调后的聊天模型”，*arXiv* 预印本 *arXiv:2307 09288*，2023年。。

[5] J. Xu等，“3-omni技术报告，”*arXiv* 预印本 *arXiv:2503.20215*，2025年。

[6] E. J. Hu 等，“Lora：大型语言模型的低秩适配”，载于学习表征国际会议论文集，2022年，第1–13页。[7] H. Du 等，“Crab：一种具有显式协作机制的统一音频-视觉场景理解模型”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2025年，第18 804–18 814页。[8] L. Feng、C. Zhang和X.-L. Zhang，“迈向灵活且统一的语音增强架构”，*Vici- nagearth*，第2卷第1期，第14页，2025年。[9] L. Wang 等，“Erf-ba-tfd+：一种用于音频-视觉深度伪造检测的多模态模型”，*Vicinagearth*，第2卷第1期，第1–12页，2025年。[10] Y. Wei、D. Hu、Y. Tian和X. Li，“音频-视觉语境中的学习：综述、分析与新视角”，*arXiv*预印本 *arXiv:2208.09579*，2022年。[11] R. Gao、T.-H. Oh、K. Grauman和L. Torresani，“听声观影：通过预览音频进行动作识别”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2020年，第10 457–10 467页。[12] Y. Tian、J. Shi、B. Li、Z. Duan和C. Xu，“无约束视频中的视听事件定位”，载于欧洲计算机视觉会议论文集，2018年，第247–263页。[13] Y. Tian、D. Li和C. Xu，“统一多感知知觉：弱监督视听视频解析”，载于欧洲计算机视觉会议论文集，2020年，第436–454页。[14] J. Zhou 等，“音频-视觉分割”，载于欧洲计算机视觉会议论文集，2022年，第386–403页。[15] Y. Wang 等，“Ref-avs：在视听场景中指代并分割对象”，载于欧洲计算机视觉会议论文集，2024年，第196–213页。[16] P. Yang 等，“Avqa：用于视频视听问答的数据集”，载于*ACM*国际多媒体会议论文集，2022年，第3480–3491页。

[17] G. Li 等，“学习回答动态视听场景中的问题”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2022年，第19 108–19118页。

[18] H. Zhang, X. Li, and L. Bing, “Video-llama：一种面向视频理解的指令微调视听语言模型，”*arXiv* 预印本*arXiv:2306.02858*，2023年。

[19] A. Vaswani 等，“注意力就是你所需要的一切”，载于神经信息处理系统进展，2017年。[20] A. Grattafiori 等，“Llama 3 模型群”，*arXiv* 预印本 *arXiv:2407.21783*，2024年。

[21] A. Radford 等，“通过自然语言监督学习可迁移视觉模型”，载于机器学习国际会议论文集，2021年，第8748–8763页。

[22] C. Jia 等人，“利用噪声文本监督扩展视觉与视觉-语言表征学习”，载于国际机器学习会议论文集，2021年，第4904–4916页。[23] J. Li、D. Li、S. Savarese和S. Hoi，“BLIP-2：利用冻结图像编码器和大语言模型引导语言-图像预训练”，载于国际机器学习会议论文集，2023年，第19730–19742页。[24] H. Liu、C. Li、Q. Wu和Y. J. Lee，“视觉指令微调”，载于神经信息处理系统进展，2023年，第34892–34916页。[25] H. Chen、W. Xie、A. Vedaldi和A. Zisserman，“VggSound：大规模视听数据集”，载于国际声学、语音与信号处理会议，2020年，第721–725页。[26] R. Arandjelovic和A. Zisserman，“Look, Listen and Learn”，载于*IEEE*国际计算机视觉会议论文集，2017年，第609–617页。[27] K. Soomro、A. R. Zamir和M. Shah，“UCF101：野外视频中101类人体动作数据集”，*arXiv*预印本 *arXiv:1212.0402*，2012年。[28] S. Poria 等人，“MELD：用于对话中情感识别的多模态多方数据集”，*arXiv*预印本 *arXiv:1810.02508*，2018年。[29] Y. Liu 等人，“MAFW：用于野外动态面部表情识别的大规模多模态复合情感数据库”，载于*ACM*国际多媒体会议论文集，2022年，第24–32页。[30] X. Jiang 等人，“DFEW：用于识别野外动态面部表情的大规模数据库”，载于*ACM*国际多媒体会议论文集，2020年，第2881–2889页。[31] Z. Lian等人，“MER 2024：半监督学习、噪声鲁棒性与开放词汇多模态情绪识别”，载于第二届多模态负责任情感计算国际研讨会论文集，2024年，第41–48页。[32] H. Cao 等人，“CREMA-D：众包情绪多模态演员数据集”，*IEEE*情感计算汇刊，第5卷，第4期，第377–390页，2014年。

[33] T. 耿郑J. “未剪辑视频中视听事件的密集定位：大规模基准和基线”，载于计算机视觉与模式识别会议论文集，2023年，第 22 942–22 951 页。

[34] S. Chen等, “Vast：视觉-音频-语言全方位感知预训练模型与数据集”，*arXiv*预印本 *arXiv:2304.08345*，2023年。

[35] G. Team 等，“Gemini 1.5：解锁跨数百万上下文标记的多模态理解”，*arXiv* 预印本 *arXiv:2403.05530*，2024年。

[36] W. Kay等, “The Kinetics 人类动作视频数据集”，*arXiv* 预印本 *arXiv:1705.06950*，2017年。

[37] X. Lai等, “Lisa：通过大语言模型进行推理分割”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2024年，第9579–9589页。

[38] Z. Ren等, “Pixellm：使用大规模多模态模型进行像素推理”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2024年，第26374–26383页。

[39] Y. Liu等, “Seg-Zero：通过认知强化进行推理链引导的分割”，*arXiv*预印本 *arXiv:2503.06520*，2025年。

[40] N. Ravi等, “Sam 2：在图像和视频中分割任意对象”，*arXiv*预印本 *arXiv:2408.00714*，2024年。

[41] S. Bai等, 《Qwen2-vl 技术报告》，*arXiv* 预印本 *arXiv:2502.13923*，2025年。

[42] Y. Chu等, “Qwen2-audio 技术报告”，*arXiv* 预印本 *arXiv:2407.10759*，2024年。

[43] A. Radford 等，“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，2023年，第 28 492–28 518 页。

[44] S. Chen等, “BEATs：采用声学分词器进行音频预训练”，*arXiv* 预印本 *arXiv:2212.09058*，2022年。

[45] C. Tang 等，“SALMONN：迈向大语言模型的通用听觉能力”，*arXiv* 预印本 *arXiv:2310.13289*，2023年。

[46] Y. Wang, W. Zhou, Z. Jiang, Z. Huang, Y. Ding, J. Liu, Z. Liu, and R. Xia, “MuLoRA：普及化，以实现更好的多任务学习”，*arXiv* 预印本 *arXiv:2311.11501*，2023年。

[47] S. Ren、L. Yao、S. Li、X. Sun 和 L. Hou，“Timechat：一种面向长视频理解的时间敏感型多模态大语言模型”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2024年，第 14 313–14323 页。

[48] S. Chowdhury等, “Meerkat：具备时空定位的音频-视觉大语言模型”，载于欧洲计算机视觉会议论文集，2024年，第52–70页。

[49] Z. Li 等，“GroundingGPT：语言增强的多模态定位模型”，*arXiv* 预印本 *arXiv:2401.06071*，2024年。

[50] J. He等，“具有细粒度视觉感知的多模态指令调优大语言模型”，载于 *IEEE/CVF* 计算机视觉与模式识别会议论文集，2024年，第 13 980–13 990 页。[51] D. Jin 等，“SimToken：用于指代式音视频分割的简单基线”，*arXiv* 预印本 *arXiv:2509.17537*，2025年。[52] C. Lyu 等，“Macaw-llm：集成图像、音频、视频和文本的多模态语言模型”，*arXiv* 预印本 *arXiv:2306.09093*，2023年。[53] Z. Cheng 等，“Videollama 2：推进视频大语言模型中的时空建模与音频理解”，*arXiv* 预印本 *arXiv:2406.07476*，2024年。

[54] Q. Ye等, “Ola：探究并增强大语言模型的音频-视觉理解能力”，发表于*IEEE*模式分析与机器智能汇刊，第47卷，第10期，第8674–8690页，2025年。

[55] A. Goel等人, “Omcat：全上下文感知变换器，”*arXiv*预印本 *arXiv:2410.12109*，2024年。

[56] J. Lu, 等人, “Unified-io 2：利用视觉、语言、音频与动作扩展自回归多模态模型，”载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2024年，第26 439–26 455页。

[57] C. Fu, 等人, “Vita：迈向开源交互式全模态大语言模型”，*arXiv* 预印本 *arXiv:2408.05211*，2024年。

[58] Z. Cheng 等，“Emotion-llama：通过指令微调实现多模态情绪识别与推理”，载于神经信息处理系统进展，2024年，第110805–110853页。

[59] X. Peng, Y. Wei, A. Deng, D. Wang, 和 D. Hu, “通过即时梯度调制实现平衡多模态学习”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2022年，第8238–8247页。

[60] C. Huang、Y. Wei、Z. Yang和D. Hu，“用于平衡多模态信息获取的自适应单模态调节”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2025年，第25854–25863页。[61] Y. Wei、D. Hu、H. Du和J.-R. Wen，“用于平衡多模态学习的即时调制”，*IEEE*模式分析与机器智能汇刊，第47卷，2025年，第485号，第469–页。[62] B. Lin 等，“Video-llava：通过投影前对齐学习统一的视觉表示”，*arXiv*预印本 *arXiv:2311.10122*，2023年。[63] J. F. Gemmeke 等，“Audio set：用于音频事件的本体与人工标注数据集”，载于国际声学、语音与信号处理会议论文集，2017年，第776–780页。

[64] Y. Wei 和 D. Hu, “Mmpareto：通过无害单模态辅助提升多模态学习能力”，*arXiv* 预印本 *arXiv:2405.17730*，2024年。

[65] Z. Xu, Y.Kou, K. Wu 和 H. Liu，《不平衡与噪声条件下鲁棒多模态融合的贡献引导的非对称学习》，

imbalance andnoise," *arXiv* 预印本 *arXiv:2510.26289*，2025年。

[66] N. Majumder等，"Dialoguernn：用于对话情绪检测的注意力循环神经网络"，发表于人工智能促进会会议论文集，第33卷，第01期，2019年，第6818–6825页。

[67] Y. Jing 和 X. Zhao，"DQ-Former：通过动态模态优先级查询Transformer，实现认知对齐的对话多模态情感识别"，载于*ACM* 多媒体国际会议论文集，2024年，第 4795–4804 页。

[68] J. Ding、B. Zhang、D. Li、J. Wang 和 H. Lin，"解耦跨模态交互以增强对话中的多模态情感识别"，载于第 *27* 届多模态交互国际会议论文集，2025年，第 344–353 页。

[69] Z. Zhao 和 I. Patras，"利用提示引导视觉语言模型进行动态面部表情识别"，*arXiv*预印本 *arXiv:2308.13382*，2023年。[70] H. Cheng 等，"VAEmo：通过知识注入实现视听情绪的高效表征学习"，载于*ACM*国际多媒体会议论文集，2025年，第5547–5556页。[71] X. Wu 等，"用于高效情感视频面部分析的可扩展视听掩码自编码器"，*arXiv*预印本 *arXiv:2509.24214*，2025年。[72] X. Long 等，"注意力簇：一种纯注意力机制的局部特征整合方法，用于视频分类"，载于*IEEE*计算机视觉与模式识别会议论文集，2018年，第7834–7843页。[73] M. Brousmiche、J. Rouat 和 S. Dupont，"用于视听事件识别的多模态注意力融合网络"，*Information Fusion*，第85卷，第52–59页，2022年。[74] Y.-B. Lin 和 Y.-C. F. Wang，"结合实例注意力的视听变换器，用于视听事件定位"，载于亚洲计算机视觉会议论文集，2020年。[75] J. Yu、Y. Cheng、R.-W. Zhao、R. Feng 和 Y. Zhang，"MM-Pyramid：用于视听事件定位和视频解析的多模态金字塔注意力网络"，载于*ACM*国际多媒体会议论文集，2022年，第6241–6249页。[76] Y. Xia 和 Z. Zhao，"用于视听事件定位的跨模态背景抑制"，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2022年，第19 989–19 998页。[77] X. Jiang 等，"DHHN：用于弱监督视听视频解析的双层级混合网络"，载于*ACM*国际多媒体会议论文集，2022年，第719–727页。[78] K. Rho、H. Lee、J. W. Cho 和 J. S. Chung，"MoLT：用于高效视听学习的逐层令牌混合"，*arXiv*预印本 *arXiv:2512.00115*，2025年。[79] H. Chen 等，"以艰难方式定位视觉声音"，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2021年，第16 867–16 876页。

[80] W. Sun等，"通过假阴性感知对比学习实现音频-视觉声源定位"，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2023年，第6420–6429页。

[81] S. Gao, Z. Chen, G. Chen, W. Wang, and T. Lu, "Avseg-former：基于Transformer的音频-视觉分割"，发表于人工智能促进协会会议论文集，第38卷，第11期，2024年，第155–163页。

[82] Y. Wang, W. Liu, G. Li, J. Ding, D. Hu, and X. Li, "Prompting segment anything model with domain-adaptive prototype for generalizable video polyp segmentation," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 1777–1786. [Online]. Available: https://doi.org/10.1145/3746027.3754760

[83] G. Li, Y. Wei, Y. Tian, C. Xu, J.-R. Wen, and D. Hu, "Learning to answer questions in dynamic audio-visual scenarios," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 19 108–19 118. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.01852

[84] T. Yang 等，"Sasr-net：用于音频-视觉问答的源感知语义表示网络"，*arXiv*预印本 *arXiv:2411.04933*，2024年。[85] H. Kim 等，"用于音频-视觉问答的问题感知高斯专家"，载于*IEEE/CVF*计算机视觉与模式识别会议，2025年，第 13 681–13 690 页。[86] T. M. Le、V. Le、S. Venkatesh 和 T. Tran，"用于视频问答的层次条件关系网络"，载于*IEEE/CVF*计算机视觉与模式识别会议，2020年，第 9972–9981 页。