arXiv:2611.33153v1 [cs.CV] 24 Nov 2026

DiaCap：面向视听对话字幕生成，联合学习说话人身份与发言时间

Binfei Ji[1†‡] Zhiwen Cao[2†] Anqi Yang[2†] Prem Seetharaman[2] Shuo Yang[2]
Ratheesh Kalarot[2] Yingyan (Celine) Lin[1] Mingze Xu[21]佐治亚理工学院 [2]Adobe Research
{zhiwenc, anqiy, mingzex}@adobe.com

摘要

我们提出了 *DiaCap*，这是一种用于全面对话描述的视听大语言模型，可为视频中的每段话语生成转录文本、时间戳及说话人的视觉描述。鉴于话语时间对齐与视听说话人分离紧密相关，*DiaCap* 将这两项任务整合到单个模型中，并通过两个互补阶段进行训练。首先，我们使用合成描述文本在 *22.5* 万个精选片段上训练 *DiaCap*，对每段话语进行细致且密集的标注。其次，我们在 *2,900* 个经人工校正的片段上通过强化学习进一步优化模型，利用细粒度反馈直接评估话语时间对齐、说话人分组和说话人匹配。在 *DiaDemBench* 上，我们扩展了该基准，加入话语级时间戳和字幕遮挡；无论使用现成转录文本的 *DiaCap*，还是其端到端变体 *DiaCap-E2E*，均取得了优异结果。具体而言，*DiaCap* 的视听说话人分离性能与 *Gemini 4 Argon* 相当，同时时间戳准确率明显更高（*76.59* 对比 *42.76 TIME*）。它在多语言视频和长达一分钟的视频中也保持了较强的说话人分离能力，并可迁移至视听问答任务。

1. 引言

全面对话描述不仅需要准确的语音转写：模型还必须定位每段话语发生的时间，并识别谁说出了这些话（图1）。不同于传统的说话人分离 [4, 21, 43],即使使用视觉线索，也只会分配匿名标签 [13, 49, 83]，视听说话人分离则通过详细的视觉外观来描述每位说话人。然而，通用视听大语言模型 [25, 26, 64, 85] 和现有的对话字幕生成模型 [8, 72] 仍无法可靠地共同确定何时以及谁，尤其是在多人场景中，或说话人不断进出画面时。

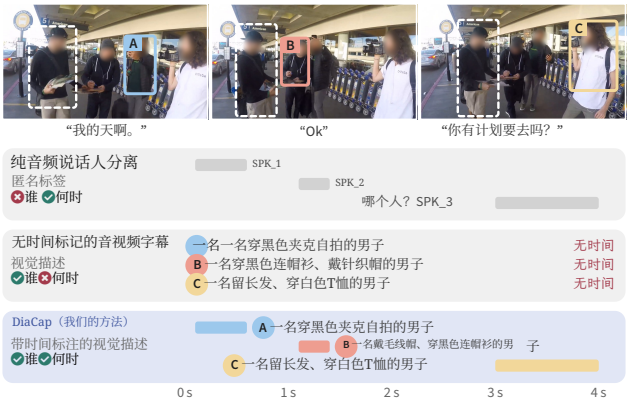


图 1. **全面对话描述**。 DiaCap 在单个描述文本中描述谁说了什么，以及何时说的。纯音频说话人分离仅为每位说话人分配一个匿名标识，而 DiaCap 则通过详细描述识别每位说话人。

我们认为，这一挑战源于对话语学习的困难。

分别进行时间对齐和说话人归属，尽管二者紧密耦合。例如，了解谁在说话有助于定位话语边界，而准确的边界则是将每段话语绑定到正确说话人的必要条件。我们提出 **DiaCap**，一种用于全面对话描述的视听大语言模型（图 2），通过从两种互补信号中学习，在单个模型中统一这两项任务。首先，密集监督提供完整字幕，为每段话语标注其转录文本、时间戳和说话人，并通过视觉外观描述每位说话人。其次，作为更直接的信号，细粒度反馈评估每段话语的时间对齐准确度，以及每位说话人是否被正确识别。

为在大规模数据上提供密集监督，我们首先使用监督微调（SFT）在 225K 个片段上训练 Dia-Cap。这些片段从 1.25M 个原始片段中筛选而来，以确保视听对应关系和数据分布的平衡。由于在这一规模上进行人工标注并不现实，我们使用一种结合两种模型优势的合成标注流水线为这些片段添加标签。视听大语言模型 [26, 64]

†共同贡献。
‡工作于 Adobe 实习期间完成。

能够生成准确的转录文本并进行说话人分离，但只能提供较粗粒度的时间戳，而语音对齐器 [41]能够精确定位语音，却无法识别说话人。我们保留视听大语言模型生成的带说话人归属的话语，并使用对齐器细化其时间戳。由于该处理流程串联了多个模型，因此运行复杂且成本高昂，我们仅将其用作离线数据引擎。

SFT 的下一词元目标逐词元模仿参考字幕，却不会直接衡量预测时间戳和说话人分配的准确性。因此，我们进一步通过强化学习（RL）优化 DiaCap， [91], 并使用细粒度反馈作为奖励。强化学习数据包含约 2.9K 个经过人工校正的片段，这些片段挖掘自视觉上具有挑战性的场景，例如多人场景、面孔较小或模糊的情况，以及说话人延迟出现的情况。我们的奖励并非采用混合不同错误类型的单一分数，而是分解为三项。时间戳奖励 根据每个预测话语边界与参考边界之间的距离进行评分。对于说话人分离，分组奖励 不依赖描述，而是检查每个人的话语是否都对应同一个预测说话人，并对将一个人的话语拆分开以及将多个人合并在一起这两种情况进行惩罚。随后，绑定奖励会询问多模态评判器，每条预测描述是否指向正确的人，并对所有说话人取平均，使次要说话人和主要说话人获得同等重视。

鉴于转录很少依赖视觉线索 [18, 53, 85], ，DiaCap 默认使用现成的转录文本，专注于说话人分离和时间戳标注。我们还训练了 DiaCap-E2E，这是一种端到端变体，可直接输出完整的对话字幕。我们在 DiaDemBench [8], 上进行评估，并为其补充了话语级时间戳和字幕删节，以进行更严格的评估。实验结果表明，两种模型的表现都很出色：DiaCap 使用现成识别器生成的转录文本，而 DiaCap-E2E 则自行预测转录文本。具体而言，在视听说话人分离方面，DiaCap 的表现与 Gemini 4 Argon [25] 相当，而时间戳准确率则显著优于后者（76.59 对比 42.76 时间）。除 DiaDemBench 外，DiaCap 在多语言视频上也保持了出色的说话人分离表现，包括训练数据中代表性不足的语言，以及长达一分钟的视频 [72]。此外，它在迁移至视听问答 [57]方面也展现出良好前景。

2. 相关工作

说话人分离预测每个词或话语是谁说的。纯音频方法仅根据语音特征对语音进行说话人分段 [1, 60], ，方式包括嵌入聚类 [43, 69, 76, 92], 、端到端预测 [21,28, 62], 、混合系统 [4, 36], ，或联合识别与说话人分离 [17, 34, 35, 55, 59, 61, 77, 90]。我们关注视听说话人分离，它还必须将每个话语与一位可见人物关联起来。视听方法通过同步性 [12, 16, 33, 58]和主动说话人检测 [38, 45, 46, 54, 66, 74, 78] 将语音与人脸关联起来，以进行说话人分离 [5, 9, 13,24, 31, 49, 83], ， 但输出的仍是匿名聚类或人脸轨迹。需要为说话人命名的方法必须预先获得身份信息，例如演员图库 [19, 32, 39, 56] 或预定义的外观描述 [2]。而 DiaCap 会根据视频生成每位说话人的视觉描述，并将每个话语关联到相应人物。

时间定位用于确定每个单词或话语的发声时间。时间戳通常通过将转录文本与音频对齐来获取 [3, 42, 52], ， 包括使用基于大语言模型的强制对齐器 [68], ，或由音频语言模型生成带时间戳的描述文本 [41]，或为给定的转录文本生成词级标注 [37]。在视频领域，片段检索 [22, 27, 44] 和密集事件描述 [40, 86] 催生了能够预测时间边界的视频大语言模型和视听大语言模型 [11, 30, 65, 73, 79], ， 但它们定位的是查询片段或场景事件，而非单个话语。只有少数视听大语言模型能够输出话语级时间戳 [25, 64, 72], ，且精度仍较粗略。DiaCap 针对这一话语级任务，同时预测每个话语的边界及其说话人。

视听对话字幕生成。 包括 Qwen-Omni [64, 84, 85], Gem- ini [14, 25, 26], 等在内的通用视听大语言模型 [10, 20, 70, 82, 89, 93], 能够根据同步视频和音频生成转录文本、说话人描述和时间戳。细致的视听描述模型能够提升内容覆盖度和视听对应关系 [6, 7, 23, 48, 51, 67, 71, 80, 81], ，而结构化脚本生成器则会将场景、角色、事件和对话整理成带时间戳的输出 [63, 75, 88]。与我们的工作最相关的是 DiaDem [8] 和 D-ORCA [72], 等对话字幕生成模型，但它们的说话人归属能力仍逊于强大的通用视听大语言模型。此外，DiaDem 不输出时间戳，而 D-ORCA 仅输出精确到整秒的时间戳。DiaCap 填补了这一空白：其说话人归属能力可与最强的通用视听大语言模型相媲美，同时还能生成最准确的语句时间戳。

3. 对话字幕生成

给定一段视频 V 及其音轨 A，对话字幕生成旨在生成一段描述文本 Y，其中包含一系列话语和一个说话人列表（图2）。每个话语都是一名单人说话者发出的一段连续语音片段，并标注其转录文本、开始和结束时间戳以及说话人标签。该标签指向说话人列表中的一项，其中通过视觉外观描述每位说话人（例如，“一名穿蓝色衬衫的男子”），以便将其与视频中的人物相匹配。本文将预测话语级时间戳称为话语定时，将构建说话人列表并将每个话语分配给一名

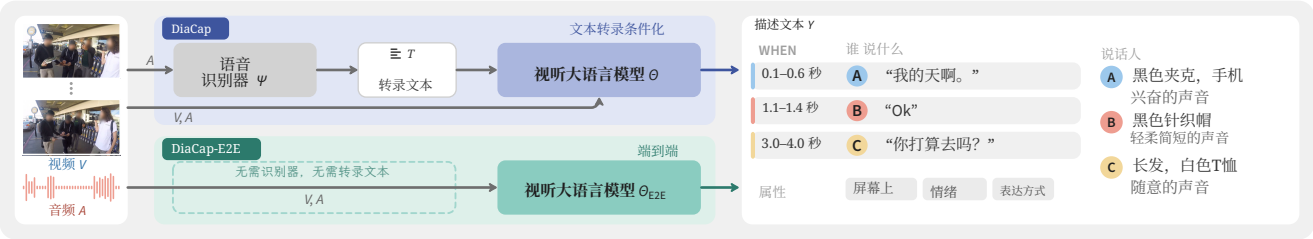


图2。**DiaCap 概览**。语音识别器 Ψ 将音频转录为转录文本，视听大语言模型 Θ 结合视频和音频以及转录文本 T，预测每个话语的时间戳和说话人。DiaCap-E2E 无需识别器即可直接 Y 进行预测。

说话人称为视听说话人分离。

3.1. 流程概述

文本转录条件化流程。强大的现成语音识别器已经能够很好地完成转录任务 [18, 25,53, 85]，而时间戳和说话人分离仍是当前对话字幕生成的主要瓶颈。因此，我们将转录与任务的其余部分分离开来，从而能够复用这些识别器，并让模型专注于预测视听说话人分离结果和精确时间戳。

语音识别器 Ψ生成转录文本，视听大语言模型 Θ 预测时间戳和说话人分离结果：

$$T = \Psi(A) \quad \text{and} \quad Y = \Theta(V, A, T). \tag{1}$$

我们将这一处理流程称为**DiaCap** （图2中的上方图示）。转录文本 T 是一段涵盖片段中全部语音的单一文本，不包含话语边界、时间戳或说话人标签。给定 T后，模型 Θ 将其拆分为多个话语，预测这些话语的时间戳，并对其进行说话人分离。由于 Θ只需要来自 Ψ的文本，任何语音识别器或现有的人类转录文本都可以提供 T，而且无需重新训练，DiaCap 就能受益于更好的转录文本 Θ（第4.4节）。

端到端流程。 我们还训练了一个单一模型 Θ_{E2E}，使其直接根据视频和音频预测转录文本、时间戳和说话人分离结果：

$$Y = \Theta_{\text{E2E}}(V, A). \tag{2}$$

我们将其称为 **DiaCap-E2E** （图2下方的面板）。它与 DiaCap 使用相同的骨干网络、训练数据和输出格式，但不将 T 作为输入。这表明，即使不使用外部识别器，我们的框架也同样适用。

3.2. 监督微调

监督微调需要大量且多样的视频，并且这些视频都具有完整的对话字幕（即，转录文本、时间戳和视听说话人分离标注）。据我们所知，目前没有公开数据集能大规模提供此类标注，也没有任何单一现有模型能够可靠地生成这些标注。因此，我们通过数据整理和合成标注，从原始视频构建监督微调语料库（图 3）。

数据整理。 我们通过四个步骤整理出一组在音视频上对齐且均衡的片段（图3，步骤 1）。

标注。 我们从一个内部收集的约 125 万个、时长为 4 至 30 秒的片段集合开始，其中大多数片段截取自更长的视频。对于音频，pyannote.audio [4, 62] 对语音进行说话人分离，并提供说话人的数量以及重叠语音的时长；语言则取自该集合的元数据。对于音视频对应关系，Synchformer [33] 预测相关性得分r，该得分表示声音是否由可见的视觉内容产生，以及音频与视频之间的偏移量 δ。对于视觉内容，我们跟踪并聚类人脸，以统计可见人数；LR-ASD [47] 检测其中哪些人在说话。

视听过滤。 只有当 $r > 0.3$ 和$|\delta| < 0.5$s 时，我们才保留片段；这样可以剔除音频与画面无关的片段（例如包含画外音的片段），以及音画不同步、因而无法将唇部运动与语音对应起来的片段。该过滤器保留了约 80% 的数据池。我们不要求画面中出现可见或正在说话的人脸，因为画外说话者也是任务的一部分，仅凭声音即可对其进行描述。

去重与重采样。 筛选后的数据池极不平衡（例如，英语和普通话约占片段总数的 70%，而说话人数达到三人或以上的片段仅占 16%）。我们首先移除来自同一源视频的重复窗口，并限制每个源视频所含片段的数量，以免少数长视频占据主导。随后，我们抽取分层样本，按优先级平衡 18 种以上目标语言、说话人数、片段时长、重叠语音的数量，以及可见人脸和说话人脸的数量。最后两项因素有助于保留包含重叠语音或多个可见人物的稀有片段，而这些片段最难进行说话人分离。

无语音片段。 零说话人层提供不含语音的片段，例如仅包含音乐、环境音或静音的片段。这些片段无需标注，并分配空目标，使模型学会不臆造对话。最终语料库包含约 22.5 万个片段（其中 18.9 万个含语音，3.6 万个不含语音），总时长为 1,175 小时。附录 A展示了说话人数和语言的分布。

合成标注。在这种规模下，人工标注成本高、耗时长，尤其是说话人分离。因此，我们构建了一条合成标注流水线，其中结合了两种互补的模型——

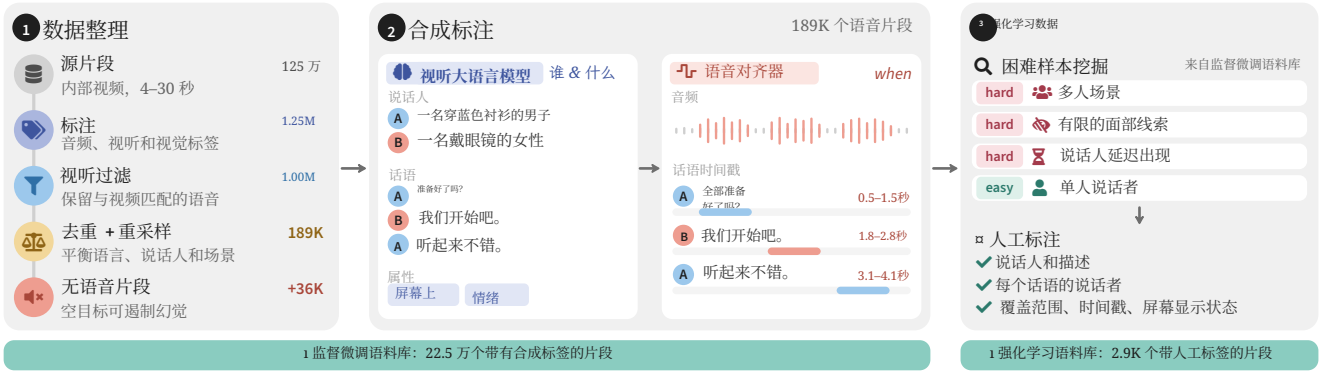


图 3. **训练数据筛选与标注**。(1) 音频、视听和视觉专家为每个片段添加标签。我们移除包含无关或不同步音频的片段，对其余片段去重并重新采样，以平衡语言、说话人数和视觉场景，并添加空目标的无语音片段。(2) 视听大语言模型标注谁说了什么，语音对齐器则为每段话语添加时间戳。(3) 人工标注员修正从监督微调语料库中挖掘的困难片段，以及数量较少的单说话人片段，以构建强化学习语料库。

——（图3，步骤 2）。视听大语言模型（例如Gemini 3.1 Pro [26]）能够生成准确的转录文本并进行说话人分离，但即使经过仔细提示，其时间戳仍较为粗略。语音对齐器 TAC [41],可以精确定位语音，却无法识别说话人。视听大语言模型首先标注转录文本、话语分段、说话人列表、说话人分配和属性。随后，TAC 在保留其余标注内容的同时，细化每段话语的时间戳，并丢弃未能通过对齐的片段。与单独使用 Gemini 3.1 Pro 相比，这种组合在保留其转录文本和说话人分离结果的同时，将时间准确率从 23.54 大幅提升至 72.42（表1）。我们将这条处理流程用作离线数据引擎，但串联多个模型和一个专有大语言模型，成本过高，不适合日常使用。

学习目标。我们基于预训练的音频-视觉语言模型（例如，Qwen3-Omni [85]），并使用标准的下一词元预测对其进行训练。给定描述文本 Y作为词元序列 $y_1, \ldots, y_L$，损失为

$$\mathcal{L}_{\text{SFT}} = -\sum_{j=1}^{L} \log P_{\Theta}\left(y_j \mid y_{<j}, V, A, T\right), \qquad (3)$$

DiaCap-E2E 使用相同的损失进行训练，但不包含 T。如果基准模型在回答前会生成思考轨迹，我们会在每个目标前添加一个空的思考块（<think></think>），使模型直接输出描述文本。

3.3. 强化学习

我们进一步通过强化学习优化 Θ 和 Θ_{E2E}，为时间戳和音视频说话人分离提供直接训练信号。**困难样本挖掘与标注**。我们重点关注三类视觉上难以处理的场景：(1) 多人场景，其中有多张脸始终可见，但只有一人说话；(2) 有限视觉证据，由于分辨率低、画面模糊、遮挡或不利的观看角度，难以观察到说话人的嘴部动作；以及 (3) 说话人延迟出现，即语音开始时屏幕上显示的是未说话的人，而真正的说话人稍后才出现。我们根据视频标签和合成标注（例如，人脸数量和说话人数）中的规则，从监督微调语料库中筛选出各类片段，并加入一小部分单说话人片段，以避免策略偏向于识别额外说话人，最终得到约 2.9K 个片段（图 3，步骤 3）。为获取这些片段上可靠的参考标注 Y^*，人工标注员会对照视频校正每个片段的合成标注，包括修正说话人列表、重新分配话语、补全缺失的语音并调整时间戳。

奖励设计。采样字幕 Y 和参考文本Y^* 可能会将语音划分为不同的话语，因此我们首先在词级进行匹配。由于两份描述文本包含的词语几乎相同，我们通过编辑距离对齐它们的词序列；对于中日韩语言，则使用角色而非词语进行对齐。这样，每个参考词都会映射到包含该词的预测话语，从而获得其预测说话人和时间戳。

时间戳奖励。对于每个参考话语的开始和结束时间，我们取相应的预测边界：当误差在 0.3 秒以内时，给予 1 分；误差达到 1 秒时线性递减至 0 分。预测结果可能会合并两个连续的参考话语，从而遗漏它们之间的边界。对于此类遗漏的边界，仅当其标志着说话人变化或至少 0.5 秒的停顿时，才给予 0 分；因此，合并同一说话人的连续语音不会受到惩罚。参考话语内部多出的预测边界不参与评分。R_{time} 是所有参与评分的边界的平均得分。

说话人分离奖励。我们通过两种奖励评估说话人分离：分组奖励用于检查每个人的话语是否归入同一个预测说话人之下，而不考虑描述；绑定奖励则用于检查每个预测说话人是否被描述为正确的人——

物。令 $\mathcal{G}$ 和 $\mathcal{P}$ 分别表示参考说话人和预测说话人，令 n_g 表示参考说话人g的话语数量，其中 $N = \sum_g n_g$。我们在 C_{pg} 中统计g 被分配给预测说话人 p的话语数。若某段话语被拆分到多个预测说话人，则根据其对齐词语按比例计数；同时由多人共同说出的内容不纳入统计。对于分组奖励，我们通过一对一映射 π 将每个参考说话人匹配到一个不同的预测说话人，使其涵盖的话语数量最多：

$$R_{\text{grp}} = \frac{1}{N} \max_{\pi} \sum_{g \in \mathcal{G}} C_{\pi(g)\, g}. \tag{4}$$

将一个人拆分成多个说话人，或将多个人合并为一个说话人，都会导致部分话语无法匹配，从而降低奖励。对于绑定关系，多模态评判器会观看片段，并在描述 p和 g 指的是同一个人时设定 $J_{pg} = 1$，否则设定 $J_{pg} = 0$:

$$R_{\text{bind}} = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \frac{1}{n_g} \sum_{p \in \mathcal{P}} C_{pg}\, J_{pg}. \tag{5}$$

它对参考说话人取平均值，因此，话语较少的说话人与主要说话人的权重相同，这与 DiaDem 的按话语加权词错误率指标 REF 不同 [8]。

最终奖励为

$$R = \lambda_{\text{time}} R_{\text{time}} + \lambda_{\text{grp}} R_{\text{grp}} + \lambda_{\text{bind}} R_{\text{bind}} + \lambda_{\text{asr}} R_{\text{asr}}, \tag{6}$$

其中，$R_{\text{asr}} = 1 - \min(\text{WER},1)$用于评估预测转录文本与参考文本之间的差异。DiaCap 将转录文本作为输入，因此我们设定 $\lambda_{\text{asr}} = 0$、$\lambda_{\text{time}} = 0.2$和 $\lambda_{\text{grp}} = \lambda_{\text{bind}} = 0.4$。DiaCap-E2E 会自行生成转录文本，因此我们添加 $\lambda_{\text{asr}} = 0.2$，并将 λ_{grp} 降低至 0。2，同时保留 $\lambda_{\text{time}} = 0.2$ 和 $\lambda_{\text{bind}} = 0.4$。无法解析的描述文本奖励为零。

策略优化。我们使用由监督微调模型初始化的 DAPO [91]，对 Θ 和 Θ_{E2E} 进行优化。对于每个片段，我们采样一组包含八条描述文本的样本，并将每条描述文本的优势计算为其奖励减去组内平均值，从而提高奖励高于组平均水平的描述文本的似然。遵循 Dr. GRPO [50]，，我们不使用组内标准差对优势进行归一化，因为这种归一化会过度提高那些描述文本奖励差异很小的片段的权重。与 DAPO 一样，我们丢弃奖励完全相同的样本组，因为这些样本组不提供学习信号，并对过长的响应施加软惩罚。

4. 实验

4.1. 设置

基准测试。我们在 DiaDemBench [8]，上进行评估，该基准包含 1,039 个视频片段，每段中的话语都标注了转录文本以及对说话人的自然语言描述。由于 DiaDem 未提供时间戳，我们手动标注了每段话语的开始和结束时间；下文中的 *DiaDemBench* 均指带有这些时间戳的基准测试。为防止模型利用内嵌字幕进行转录，我们进一步使用 PP-OCRv5 [15]检测每一帧中的字幕，并将检测到的区域模糊处理，同时保持音频和标注不变；这影响了 1,039 个片段中的 598 个。我们将这个去除字幕的版本称为 *DiaDemBench-NoSub*，除非另有说明，所有结果均在该版本上报告。

评估指标。我们采用 DiaDemBench [8]的 REF 和自动语音识别指标，并引入 TIME 来评估时间戳。*REF* 用于衡量每条话语是否归属于正确的说话人，评判依据为 Gemini-2.5-Flash。*ASR*用于衡量匹配话语的转录准确率。*TIME* 用于衡量每条匹配话语的开始时间和结束时间是否都在参考文本对应时间的 0.3 秒范围内。

基线方法。我们将 DiaCap 与最先进的专有模型和开放权重模型进行比较，并使用各模型的官方实现进行评估。通用视听大语言模型（例如，Gemini 3.1 Pro）使用我们提供的提示，其中要求填写描述文本的每个字段。对话字幕生成模型（例如，DiaDem）使用其官方提示，而一个纯文本提取器 [14] 则将它们的自由格式描述文本转换为对话。我们还报告了用于生成 DiaCap 合成标签的标注流水线（第3.2节）。

评估协议。我们在两种设置下报告结果。在预测转录文本设置中，不提供真实文本，因此基线方法自行转录语音，而 DiaCap 则使用现成语音识别器生成的转录文本。在真实转录文本设置中，模型还会收到未分段的真实转录文本，从而消除转录错误，并单独评估说话人分离和时间戳。

4.2. 实现细节

模型架构。我们基于 Qwen3-Omni-30B-A3B-Thinking [85]构建 DiaCap 和 DiaCap- E2E。我们对大语言模型的注意力投影层应用 LoRA [29]，秩设为 512，并使用 $\alpha = 1024$；同时冻结专家模块、路由器以及音频和视觉编码器。我们将每段视频缩放至短边不超过 360 像素，同时保持宽高比，并以每秒 4 帧的速率采样。

监督微调训练。我们使用合成数据流程标注的 225K 个片段训练 Θ 和 Θ_{E2E}，该流程使用 Gem- ini 3.1 Pro [26] 作为视听大语言模型。数据格式相应地调整为符合两个模型所需的输入输出格式。我们使用 AdamW，参数为 $\beta_1 = 0.9$和 $\beta_2 = 0.95$，峰值学习率为 6×10^{-5}，先进行 100 个预热步数，再在 2 个 epoch 内采用余弦衰减。训练在 16 个 NVIDIA B200 GPU 上进行，全局批量大小为 128；我们选取训练 1.5 个 epoch 后的检查点。

强化学习训练。在监督微调后，我们进一步使用 DAPO [91] 对 2.9K 个人工校正片段训练 Θ 和 Θ_{E2E}。我们使用 32 个片段的全局批量大小，每个片段有 8 条采样轨迹。我们以 10^{-5} 的学习率训练 100 步，用于 Θ

方法 / 字幕	预测转录文本 REF↑ ✗	✓	ASR↑ ✗	✓	时间↑ ✗	✓	真实转录文本 REF↑ ✗	✓	时间↑ ✗	✓
专有										
标注流水线	69.18	71.31	76.57	80.56	72.42	74.49	-	-	-	-
Qwen3.5-Omni-Plus	66.78	66.51	77.04	78.15	22.17	24.87	71.32	71.87	28.29	31.26
Gemini 3.1 Pro	69.18	71.31	76.57	80.56	23.54	23.81	80.08	79.82	32.27	32.11
Gemini 4 Argon [25]	70.93	73.56	77.41	81.85	42.76	44.80	80.87	80.67	48.08	48.83
开放权重										
ARC-Qwen-Video-Narrator [23]	33.90	31.91	49.75	46.48	-	-	-	-	-	-
HumanOmniV2 [87]	15.37	14.97	23.03	22.14	-	-	-	-	-	-
MiMo-V2.6-Flash [82]	58.04	-	68.67	-	17.96	-	72.62	-	18.42	-
Qwen2.5-Omni [84]	31.50	30.92	55.10	54.11	4.65	4.71	40.05	40.11	9.31	8.82
UGC-VideoCaptioner [81]	24.93	24.57	37.11	37.54	-	-	34.11	34.49	5.23	4.99
AVoCaDO [7]	42.44	40.11	53.80	51.85	-	-	-	-	-	-
D-ORCA [72]	50.97	51.29	63.09	66.91	11.24	11.84	-	-	-	-
Qwen3-Omni-Thinking [85]	52.76	49.88	74.29	71.50	22.52	18.40	55.28	55.14	18.81	17.41
DiaDem [8]	63.79	65.87	75.62	**79.46**	-	-	-	-	-	-
DiaCap（监督微调）	68.58	69.44	78.38	78.50	75.69	75.55	79.11	79.15	86.42	86.06
DiaCap（监督微调 +强化学习）	**71.70**	**71.87**	78.99	78.97	76.59	**76.56**	**81.05**	**81.01**	**88.76**	**89.31**
DiaCap-E2E（监督微调）	70.60	71.72	78.98	81.76	74.86	75.44	-	-	-	-
DiaCap-E2E（监督微调 +强化学习）	71.59	**73.02**	**79.55**	**82.13**	**76.64**	**77.83**	-	-	-	-

表1。 **与最先进模型的比较**。 黑色列（✗）是在 DiaDemBench-NoSub 上评估的，该基准会遮去内嵌字幕；而 灰色列（✓）则使用相同的标注，在带字幕的原始视频上进行评估。 **粗体** 和下划线分别标出表现最佳的开放权重模型结果和专有模型结果。

以及用于 Θ_{E2E} 的 60 步。绑定奖励使用 Gemini-2.5-Flash [14] 作为评判器。

推理。 在预测转录设置下，除非另有说明，DiaCap 使用 Qwen3-Omni-Captioner [85] 作为语音识别器 Ψ；而 DiaCap-E2E 会自行预测转录文本。在真实转录文本设置下，DiaCap 则直接使用真实转录文本。

4.3. 主要结果

表 1 比较了 DiaCap 与最先进模型的表现。我们讨论 DiaDemBench-NoSub（✗ 列）上的结果。**与开放权重模型的比较**。 表现最佳的开放权重模型 DiaDem 不预测时间戳，而 DiaCap 的表现优于它 +7.91 REF，DiaCap-E2E 紧随其后。在会预测时间戳的模型中，最佳 TIME 仅为 22.52（Qwen3-Omni-Thinking），而我们的两个模型均超过 76。

与专有模型的比较。在预测转录文本设置下，我们的两个模型在全部三项指标上都优于 Gemini 4 Argon，其中 TIME 指标的优势尤为显著，超过 +33。DiaCap 的表现也优于生成其监督微调标签的标注流水线，REF、自动语音识别和 TIME 指标分别高出 +2.52、+2.42 和 +4.17；DiaCap-E2E 的优势幅度也相近。

真实转录文本设置。 由于 DiaCap-E2E 不接受转录文本作为输入，因此我们仅评估 DiaCap。在 REF 方面，DiaCap 与 Gemini 4 Argon 相当（81.05 *vs* 80.87），TIME 领先 +40.68，并且 REF 比其基准模型高出 +25.77。在使用真实转录文本时，所有模型的 REF 都有所提升，这表明转录质量会显著影响说话人分离。

方法	转录文本	REF↑	自动语音识别↑	时间↑
DiaCap-E2E	自行预测	71.59	79.55	76.64
DiaCap	Gemini 4 Argon	70.76	78.23	74.68
	Scribe v2	71.11	78.37	76.25
	Qwen3-Omni-Captioner*			76.59
	MAI-Transcribe-2	**72.48**	**81.23**	**78.93**

表2. **文本转录条件化与端到端预测**。DiaCap-E2E 直接预测转录文本；DiaCap 使用语音识别器生成的转录文本，无需重新训练。两者均采用监督微调 + 强化学习进行训练。 *默认识别器。

强化学习的影响。强化学习使 DiaCap 的预测转录 REF 提升 +3.12，从低于 Gemini 3.1 Pro 提升至所有模型中的最高水平。DiaCap-E2E 从更强的监督微调模型起步，因此 REF 的增幅较小，但强化学习提升了其全部三个指标，并在开放权重自动语音识别和 TIME 方面取得最佳表现。

定性结果。图4 展示了一个包含三位说话人的示例，其中两人戴着相似的军用头盔。Gem- ini 4 Argon 错误归属了话语，还凭空生成了一条额外话语；DiaDem 只描述了三位说话人中的两位。DiaCap 能区分这两位外观相似的说话人，并正确归属和对齐全部四条话语（片段级 TIME：100，而Gemini 4 Argon 为 44）。更多定性示例见附录E。

4.4. 消融研究

除非另有说明，所有消融实验均在 DiaDe mBench-NoSub 上采用 Predicted-transcript 设置。

字幕的影响。 我们还在

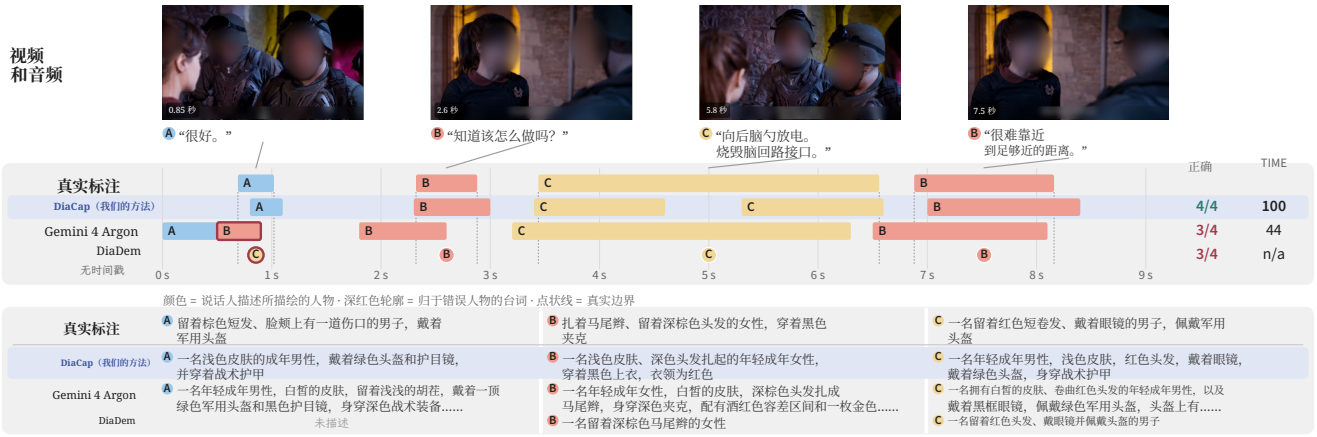


图4. **与外观相似说话者的定性比较**。每条轨迹展示一个模型随时间预测的话语，并根据其描述所对应的真实说话者着色；错误说话者以深红色描边标出。

原始 DiaDemBench 视频（带有内嵌字幕）（✓表 1中的列）上评估了所有模型。在预测转录设置下，许多基线方法通过从字幕中读取对话（例如， +4.44 的自动语音识别和 +2.63 的 REF（Gemini 4 Argon））而取得了显著提升；DiaCap-E2E 也会自行转录语音。使用真实转录文本时，这些提升便会消失（例如，Gemini 4 Argon 的 REF 为 80.87 对比 80.67），这证实字幕主要有助于转录。相比之下，DiaCap 在所有指标上的变化均小于 0.2，因为它采用的是仅音频的现成转录文本。因此，我们推荐使用 DiaDemBench-NoSub，这是一种更严格但更公平的设置，要求模型转录语音，而不是利用字幕。

转录文本条件化。我们将 DiaCap 与 DiaCap-E2E 进行比较，两者均采用监督微调 + 和强化学习训练，以测试现成转录文本对描述文本的影响（表2）。使用默认的 Qwen3-Omni-Captioner 时，两个模型的表现相当：DiaCap-E2E 的自动语音识别得分更高（+0.56），而 REF 和 TIME 的差异不到 0.11。由于 DiaCap 将转录文本作为输入，因此无需重新训练即可使用其他语音识别器 [18, 25, 53, 85]。随着识别器的自动语音识别得分从 78.23 升至 81.23，其 TIME 也从 74.68 升至 78.93；使用表现最强的识别器 MAI-Transcribe-2 时，DiaCap 在表2中取得了最佳结果。

强化学习数据。 我们测试困难样本挖掘和人工审核能否改善 DiaCap 的强化学习数据，并将其与随机采样的片段和自动标注进行比较（表 3）。在使用相同自动标注的情况下，挖掘片段比随机片段高出 +0.43 REF 和 +0.79 时间。对挖掘标签进行人工审核后，REF 进一步提高 +1.38，时间提高 +0.44；与使用自动标注的随机片段相比，总体提升 +1.81 REF 和 +1.23 时间。

强化学习奖励设计。 我们在 DiaCap 上将分解式奖励与一个基线进行比较，该基线用 REF 替代分组和绑定奖-励 [8]（表 4）。尽管基线直接优化 REF，我们的奖励仍优于它，提升了+1.30 REF 和 +0.67 TIME，且 REF 优于监督微调

强化学习数据 标签	REF↑	自动语音识别↑	时间↑
随机 自动	69.89	78.53	75.36
挖掘的 自动的	70.32	78.90	76.15
挖掘的 人工审核的	**71.70**		**76.59**

表3。 **强化学习数据对 DiaCap 的影响**。 所有变体均从相同的监督微调检查点开始，并使用公式（6）中的奖励。

方法	REF↑	自动语音识别↑	时间↑
SFT	68.58	78.38	75.69
+采用 0.8 REF+0.2R_{time} 的强化学习	70.40	78.86	75.92
+强化学习 w/ 0.4R_{grp}+0.4R_{bind}+0.2R_{time}	**71.70**		**76.59**

表4. **强化学习奖励对 DiaCap 的影响**。 两种奖励均在相同的人工审核的挖掘片段上训练，并在强化学习第100步进行评估；阴影行是我们的奖励（公式（6））。

提升了 +3.12，而非 +1.82。我们认为这主要归功于绑定奖励：它对每位说话人赋予同等权重，而 REF 受常见说话人主导，对正确识别少数说话人的贡献认可甚少。

4.5. 泛化

多语言视频。 由于 DiaDemBench 主要涵盖英语、中文、日语和韩语，我们整理了一个多语言基准，用于评估 DiaCap 在更多语言上的表现。该基准包含14种语言的140个片段，并按照 DiaDemBench-NoSub 的方式删除了烧录字幕。其中10种语言在我们的 SFT 和强化学习数据中具有充分代表性，而越南语、泰语、波斯语和缅甸语的占比不足（在 SFT 数据中最多仅占0.25%）。对于真实标注，我们保留 Gemini、MAI-Transcribe-2 和 Scribe v2 对转录文本意见一致的片段，并人工修正说话人描述、说话人分配和时间戳。我们在真实转录文本设置下进行评估。如表5所示，DiaCap 达到71.67的 REF 和91.45的 TIME，显著领先于 MiMo-V2.6-Flash 和 Qwen3-Omni-思考模式，并完成了每个片段，而 MiMo-V2.6-

方法	#已完成	REF↑	时间↑
MiMo-V2.6-Flash	102/140	64.33	21.70
Qwen3-Omni-Thinking	138/140	39.95	20.29
DiaCap	140/140	**71.67**	**91.45**

表5. **14种语言的对话字幕生成**（真实转录文本设置）。#已完成不包括模型陷入重复循环的片段，这表明模型对该语言的处理能力较差。REF 和 TIME 根据已完成的片段计算。

方法	准确率↑	词错误率↓	交并比↑	REF↑	时间↑
专有					
Gemini 3.1 Pro	86.96	15.96	49.56	69.82	56.87
Gemini 4 Argon	88.82	16.34	62.86	71.58	65.36
开放权重					
DiaDem	80.46	23.32	-	62.50	-
D-ORCA	80.00	19.07	56.80	62.45	**63.28**
Qwen3-Omni-Thinking	76.86	20.62	42.14	58.33	49.65
DiaCap	**87.05**				

表 6. **DVDBench 上的长视频评测 [72]**。准确率、词错误率和交并比是 DVDBench 的原始指标，时间采用整秒时间戳，并使用 1 秒容差。**粗体** 和下划线分别标示表现最佳的开放权重和专有模型结果。

方法	直接		先生成字幕	
	说话人↑	音频↑	说话人↑	音频↑
仅转录文本	-	-	37.2	33.7
Gemma 4 E4B	31.7	38.2	-	-
Qwen3-Omni-Thinking	54.0	56.5	-	-
video-SALMONN 2+	49.4	53.6	-	-
DiaDem	52.0	52.7	59.1	56.9
D-ORCA	55.5	55.1	**62.7**	58.8
DiaCap	**58.2**	**62.0**	59.0	**67.1**

表7。**AV-SpeakerBench 上的视听问答（视听问答）[57]**。以说话人和音频为中心的问题的准确率。直接：每个模型根据视频、音频和问题作答。先生成字幕：每个模型先生成描述文本，再由 Qwen3-Omni-Thinking 根据描述文本回答问题。

Flash 在140个片段中的38个上陷入重复循环。这表明，DiaCap 的说话人分离和时间对齐能够很好地跨语言泛化。

长视频。DVDBench [72] 包含 964 个英语片段，平均时长为 84 秒，最长可达 5 分钟，并带有整秒时间戳。我们将每个视频切分为时长 50–60 秒、重叠 20 秒的窗口，分别生成描述文本，再通过关联相邻窗口中共有的台词来连接说话人并合并窗口，无需使用额外模型。基线方法则对每个片段进行单次处理。与开放权重模型相比，DiaCap 的准确率、词错误率、交并比和 REF 均表现最佳，在时间方面仅落后于 D-ORCA（表6）。其准确率与 Gemini 3.1 Pro 相当（87.05对比86.96），但词错误率和 REF 均不及两种 Gemini 模型，这可能是因为每个窗口生成描述文本时都没有利用视频的其余部分。

视听问答（AV-QA）。我们测试 DiaCap 的能力是否能够迁移到视听问答任务中



图 5。**失败案例**。上：较小且较远的人脸会导致分段和归属错误。下：未能识别重叠语音。

（表7）。AV-SpeakerBench [57] 包含 3,212 道四选一问题（例如，说话人识别和语音的时间定位），这些问题无法仅凭单一模态回答。尽管 DiaCap 仅针对对话字幕进行了微调，但在直接设置下，其表现优于基准模型 Qwen3-Omni-Thinking，分别高出 +4.2 个说话人准确率和 +5.5 个音频准确率。在先描述设置下，其生成的描述文本取得了最佳音频准确率（67.1，相比 D-ORCA 的 58.8）。我们还将其与仅转录文本进行比较；在该设置中，仅向 Qwen3-Omni-Thinking 提供 DiaCap 以其为输入的语音转录文本。DiaCap 在说话人准确率和音频准确率上分别高出 +21.8 和 +33.4，表明这些增益来自模型本身，而非转录文本。

5. 局限性

首先，DiaCap 在具有挑战性的场景中需要更强的感知能力。例如，较小且较远的人脸只能提供很少的唇部运动证据，导致 DiaCap 将同一话语拆分到图 5 中的两位说话人之间。其次，DiaCap 若具备更强的推理能力将从中受益：当视觉证据不可靠时，模型可以根据对话语境推断说话人，正如图 5 所示。第三，转录仍然是说话人分离的瓶颈：对于给定真实标注转录文本的 Di-aCap，REF 提高 +9.35（表 1）。采用强化学习的 DiaCap-E2E 已经在没有外部识别器的情况下达到与转录文本条件模型相当的水平，而端到端模型中更强的转录能力还可以进一步缩小这一差距。

6. 结论

我们提出了 DiaCap，这是一种用于全面对话字幕生成的视听大语言模型，能够联合学习话语定时和视听说话人分离。该模型通过 225K 个合成标注片段进行密集监督训练，并利用 2.9K 个人工校正片段上的强化学习获得细粒度反馈。实验结果表明，DiaCap 实现了出色的时间戳准确率和强大的视听说话人分离能力；此外，无需外部识别器，DiaCap-E2E 也展现出了同样出色的性能。

A. SFT语料库统计

由于筛选后的数据池不平衡，我们对其进行重采样，以构建监督微调语料库（第3.2节）。

在语言方面，我们保留了所有来自16种代表性不足的语言的片段。为避免少数长视频占据过多比例，对于英语、普通话和印地语，我们每个视频最多保留一个片段。代表性不足的语言不受此限制，以确保其中的片段不会被遗漏。

在说话人数量方面，我们保留所有可获取的、包含 3+ 位说话人的片段。如果长视频中有多个候选片段，我们会选择包含 3+ 位说话人的片段。随后加入双说话人片段，直到包含 2+ 位说话人的片段占语料库的64%。单说话人片段的比例不能大幅低于21%，因为我们完整保留的代表性不足的语言片段大多只有一位说话人。我们还加入约16%的无语音片段，这些片段仅包含环境声音、音乐、音效或静音。

最终语料库包含225K个片段，图6 展示了详细的分布情况。

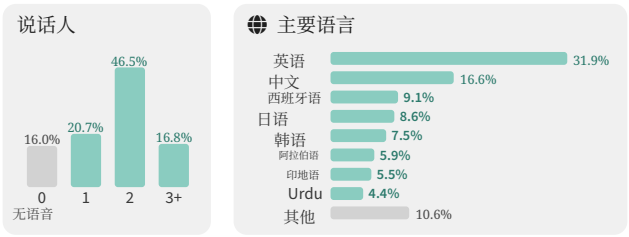


图6. **监督微调语料库统计**。

B. 原始视频上的结果

表8 至 10 重复表2 至4 在使用原始视频的 DiaDemBench 上，其中内嵌字幕未被遮盖，如表✓ 的列1所示。

方法	转录文本	REF↑	自动语音识别↑	时间↑
DiaCap-E2E	自行预测	73.02	82.13	77.83
DiaCap	Scribe v2	71.42	78.24	76.08
	Qwen3-Omni-Captioner*			76.56
	MAI-Transcribe-2	72.96	81.20	**78.73**
	Gemini 4 Argon	**73.30**	**82.25**	77.34

表 8. **文本转录条件化** 与**端到端预测** 在采用原始视频的 DiaDemBench 上的表现。DiaCap 和 DiaCap-E2E 均采用监督微调和强化学习进行训练，其他设置遵循表 2。 *默认识别器。 **粗体**表示最佳结果。

转录文本条件化。 表 8重复了表 2 中对原始视频的比较。删减视频中得到的两项发现依然成立。首先，DiaCap-E2E表现出色。使用默认的Qwen3-Omni-Captioner时，它如今在全部三项指标上都优于DiaCap，分别高出 +1.15 REF、 +3.16 的自动语音识别，以及 +1.27 的时间。其次，DiaCap无需重新训练即可使用不同的识别器，而且更强的识别器能生成更好的描述文本。使用 Gemini 4 Argon 时，DiaCap取得了最高的 REF（73.30）和自动语音识别（82.25）分数，均高于 DiaCap-E2E。使用 MAI-Transcribe-2 时，它取得了最高的时间分数（78.73）。

这些结果也支持像主论文中那样在删减视频上进行评估。当内嵌字幕清晰可见时，能够看到视频的模型可以直接读取屏幕上的文字，而不必从音频中识别文字。这会改变哪一个识别器表现更好。Gemini 4 Ar- gon 是表 8 中唯一能够看到视频的识别器，借助它，DiaCap 目前取得了比使用仅音频 MAI-Transcribe-2 更高的自动语音识别（ASR）得分（82.25 对比 81.20），这一结果与删减视频中的结果相反。DiaCap-E2E 同样能够看到视频，也以相同方式获益。与使用仅音频 Qwen3-Omni-Captioner 的 DiaCap 相比，其自动语音识别（ASR）领先幅度从删减视频中的 0.56 增长到此处的 3.16。因此，原始视频中的部分得分来自读取字幕，这有利于能够看到视频的模型，但对于模型将语音与说话人联系起来的能力几乎没有说明。删减字幕会消除这一捷径，因此每个模型都必须从音频中转录——

强化学习数据	标签	REF↑	自动语音识别↑	时间↑
随机	自动	70.88	78.66	75.85
挖掘的	自动	71.34	78.92	76.31
挖掘的	人工审核的	**71.87**		**76.56**

表 9。**强化学习数据的影响**在使用原始视频的 DiaDemBench 上。所有变体均从相同的监督微调检查点开始，并使用我们的奖励。**粗体**标示最佳结果。

方法	REF↑	自动语音识别↑	时间↑
SFT	69.44	78.50	75.55
+强化学习，使用 0.8 REF+0.2R_{time}	70.71	78.85	75.93
+强化学习，使用 0.4R_{grp}+0.4R_{bind}+0.2R_{time}	**71.87**		**76.56**

表10。**强化学习奖励的影响**在使用原始视频的 DiaDe mBench 上。两种奖励均在相同的经人工审核的挖掘片段上进行训练，阴影行是我们的奖励。**粗体**表示最佳结果。

方法	准确率↑	词错误率↓	交并比↑	REF↑	时间↑
专有					
Gemini 3.1 Pro	86.96	15.96	49.56	69.82	56.87
Gemini 4 Argon	88.82	16.34	62.86	71.58	65.36
开放权重					
DiaDem	80.46	23.32	-	62.50	-
D-ORCA	80.00	19.07	56.80	62.45	63.28
Qwen3-Omni-Thinking	76.86	20.62	42.14	58.33	49.65
DiaCap	**87.05**	16.68	**61.18**	69.34	62.03
DiaCap-E2E	86.72	**15.89**	60.60		**63.81**

表11. **DVDBench 长视频评测**。 准确率、词错误率和交并比是 DVDBench 的原始指标，而时间采用 1 秒容差来评估整秒时间戳。**粗体** 和下划线分别标示表现最佳的开放权重模型和专有模型结果。

方法	以音频为中心							以说话人为中心			
	数量↑ (288)	时长↑ (236)	强度↑音高↑ (206)	(206)	速率↑ (209)	ASR↑ (201)	All↑ (1,346)	数量↑ (402)	检测↑ (427)	识别↑ (422)	All↑ (1,251)
直接：模型根据视频、音频和问题作答											
Gemma 4 E4B	34.0	39.0	34.5	42.2	37.3	43.8	38.2	29.1	34.2	31.8	31.7
Qwen3-Omni-Thinking	**40.3**	58.5	67.5	57.8	52.2	69.7	56.5	**45.8**	56.7	59.2	54.0
video-SALMONN 2+	29.5	59.7	54.4	57.8	56.5	73.1	53.6	39.1	55.7	52.8	49.4
DiaDem	29.5	58.9	59.7	56.8	49.3	70.6	52.7	43.5	56.9	55.2	52.0
D-ORCA	38.5	62.3	54.4	49.0	57.9	74.6	55.1	42.5	**63.9**	59.2	55.5
DiaCap	37.2	**68.6**	70.9	68.4	**61.7**	74.1	**62.0**	44.3	63.5	**66.1**	**58.2**
DiaCap-E2E	34.4	65.7	**74.3**	**69.9**	61.2	**75.6**	61.7	41.0	63.7	64.0	56.5
先生成字幕：模型撰写描述文本，*Qwen3-Omni-Thinking* 根据该描述文本作答											
仅转录文本	36.5	25.4	23.3	35.0	31.1	51.7	33.7	31.3	49.6	30.3	37.2
DiaDem	52.8	54.2	59.7	51.0	62.2	63.7	56.9	56.5	64.9	55.7	59.1
D-ORCA	53.8	63.6	57.3	54.9	52.2	**72.6**	58.8	**57.7**	**68.1**	**61.8**	**62.7**
DiaCap	59.7	67.8	62.6	**66.5**	**77.0**	71.6	67.1	51.5	63.7	61.4	59.0
DiaCap-E2E	**63.9**	**70.8**	**63.1**	**66.5**	69.9	70.1	**67.2**	51.5	**68.1**	61.1	60.4

表12。 **AV-SpeakerBench 上的视听问答 [57] 按任务划分。** 音频和以说话人为中心的任务的准确率（%），括号内为问题数量。 全部 汇总每个类别的问题。 直接：每个模型根据视频、音频和问题作答。 先生成字幕：每个模型先撰写不依赖问题的字幕，再由 Qwen3-Omni-Thinking 根据字幕作答。 **粗体** 表示各区块中的最佳结果。

字幕，从而消除这一捷径，因此每个模型都必须仅从音频中转录。

强化学习数据。 表 3中的趋势在原始视频（表 9）上同样成立。使用相同的自动标注时，挖掘片段比随机片段高出 +0.46 REF 和 +0.46 TIME。对挖掘标签进行人工审核后，REF 进一步提高 +0.53，TIME 进一步提高 +0.25，总体而言，相较于带有自动标注的随机片段，REF 提高 +0.99，TIME 提高 +0.71。

强化学习奖励设计。表4 中的趋势在原始视频上也同样成立（表10）。我们的奖励优于基于 REF 的基线，分别高出 +1.16 REF 和 +0.63 TIME。与监督微调相比，它使 REF 提高了 +2.43，而基线仅提高 +1.27，尽管基线直接优化 REF。

C. 更多泛化结果

第4.5节评估了 DiaCap 在 DiaDemBench 之外处理长视频和视听问答的表现。本节则在相同的两个基准测试上评估 DiaCap-E2E。

C.1. 长视频

表 11将 DiaCap-E2E 加入表6中的 DVDBench。两个变体均采用第 4.5节的重叠窗口和跨窗口说话人关联，唯一差异在于，DiaCap-E2E 自行转录每个窗口，而不是采用语音识别器生成的转录文本。DiaCap-E2E 在 REF（69.46对比 69.34）上与 DiaCap 相当，并使 TIME 提高了+1.78（63.81对比 62.03），取得最佳开放权重模型结果，领先于 D-ORCA（63.28）。其 WER 为 15.89，是所有模型中最低的，包括 Gemini 3.1 Pro（15.96）和 Gemini 4 Ar- gon（16.34）；而 DiaCap 在准确率（87.05对比 86.72）和交并比（61.18对比 60.60）上保持小幅领先。与 DiaDemBench（表1）上的情况一样，DiaCap-E2E 的表现与 DiaCap 一样强，在这里处理的是最长五分钟的视频。

C.2. 按任务划分的视听问答

表 12 报告了 AV-SpeakerBench [57]在表7所示两种设置下的各任务结果。在直接设置中，DiaCap 和 DiaCap-E2E 在以音频为中心的（+5.5 和 +5.2）以及以说话人为中心的平均分（+4.2 和 +2.5）上均优于其基准模型 Qwen3-Omni-Thinking。它们在以音频为中心的（+6.9 和+6.6）以及以说话人为中心的（+2.7 和 +1.0）任务上也均优于 D-ORCA。在先生成字幕设置中，它们取得了最佳的以音频为中心的准确率（67.1 和 67.2 对比 D-ORCA 的 58.8）。

DiaCap-E2E。DiaCap-E2E 在两种设置下的表现均与 DiaCap 相当。直接作答时，它在以音频为中心的任务和以说话人为中心的任务上分别与 DiaCap 相差 0.3 和 1.7 分（61.7 对 62.0，56.5 对 58.2）。作为字幕生成器，它在以音频为中心的任务上与 DiaCap 表现相当（67.2 对 67.1），在以说话人为中心的任务上则高出+1.4 分（60.4 对 59.0）。

以音频为中心的任务。在直接设置下，DiaCap 和 DiaCap-E2E 在六项以音频为中心的任务中的五项（时长、强度、音高、速率和语音识别）上优于基准模型；在音高任务上的提升最高，分别为 +10.6 和 +12.1。这些任务询问每个人如何以及何时说话，而我们的描述文本会提供每段话语的时间戳、文字内容和声音描述，因此尽管这两个模型仅经过微调以生成描述文本，这项能力也能迁移到问答任务中。DiaCap-E2E 还学习了转录语音，因此是语音识别任务上的最佳模型（75.6）。例外的是语音计数任务，这两个模型的表现都不如基准模型（37.2 和 34.4 相比 40.3）。这是因为相关问题询问某个短语说了多少次，且往往要求统计某人在视觉事件标记的时间段内说出该短语的次数；直接作答的模型必须在不写下任何内容的情况下完成统计。当模型先写出每段话语及其说话人（先生成字幕）时，准确率分别升至 59.7



图7。**失败案例：画外说话者与屏幕上的说话者被合并**。在一段手持手机视频中，女孩（A）问她的母亲要去哪里，母亲（B）在画外回答。DiaCap 将其描述为一名单人说话者，把母亲的回答（“我想去哪儿就去哪儿”）归给女孩，并省略了母亲被截断的最后几句话。由于画面中始终只有女孩，模型没有推断出还有第二个人在说话。

和 63.9，DiaCap-E2E 在这一组中表现最佳。

以说话人为中心的任务。 在直接设置下，说话人识别的提升最大，DiaCap 比基准模型提高了 +6.9（66.1 对比 59.2），并且是表现最佳的模型。这些问题询问某句台词由谁说出，或某个事件发生前后由谁说话；这要求将每段话语与画面中说出该话语的人物关联起来，而这正是我们的 SFT 和强化学习所训练的视听说话人分离核心。在说话者计数方面，DiaCap 比基准模型低 −1.5（44.3对比 45.8）。每个答案都要给出说话人和画内人物的数量，而 DiaCap 的错误主要出现在后者上，因为它会漏数部分可见的人物；这是一种不属于我们训练范围的视觉能力。在说话人检测方面，该任务询问某人在事件发生前或后是否说话，DiaCap 比基准模型提高了 +6.8，但比 D-ORCA 低 −0.4（63.5 对比 63.9）。与其基准模型一样，它倾向于回答“是”，在答案为“否”的 185 个问题中正确回答了 78 个（对比 D-ORCA 的 96 个）。

D. 其他失败案例

图7 展示了另一个失败案例，其中一名画外说话者被合并到了屏幕上唯一的人物身上。与图 5一样，每个 DiaCap 条形标记都显示了预测说话人和文本，而下方没有条形标记的真实话语则未被预测出来。处理重叠语音并纠正识别错误是未来的研究方向。

E. 其他定性结果

图8–13 展示了更多采用图4布局的示例。每幅图都为每条真实话语展示一个关键帧，并标注该时刻说出的台词。时间线将真实标注与 DiaCap、DiaCap- E2E、Gemini 4 Argon [25], 以及 DiaDem [8]进行比较。每个条形表示一条话语，放置在其预测时间戳处，并根据模型所描述的真实说话者进行着色（X 表示从未说话的人），且深红色轮廓表示错误的说话人。DiaDem 不预测时间戳，因此其话语会以点的形式绘制在与之匹配的真实话语下方。对于每个模型，正确数量表示归因于正确说话人的真实话语数量，TIME 是片段级 TIME 得分。DiaCap 采用预测转录文本设置，而基线方法采用端到端方式。每条时间线下方的面板列出了各模型说话人描述中的视觉部分。

不可见的说话人。 在对话场景中，摄像机通常拍摄听者而非说话人（图 8 和 9），或者说话人背对摄像机（图 10）。基线方法往往会将这些台词分配给可见的人物，而 DiaCap 和 DiaCap-E2E 则能跨镜头将语音与正确的人物对应起来。

快速交谈中的短轮次。 较长话轮之间的单词回答或简短插话很容易被吸收到相邻的话轮中（图11 和 12）。DiaCap 和 DiaCap-E2E 能以准确的边界将这些话轮分开，而时间戳不精确或没有时间戳的基线方法会将它们合并，从而遗漏该话轮及其说话人。

几位外貌相似的说话人。 当一个场景中有三位或更多说话人时，描述必须能够区分每个人（图9 和 13）。DiaDem 的通用描述，例如“穿实验服的男子”，适用于多个人，会导致说话人身份互换或合并；而 DiaCap 的描述始终足够具体，能够识别每位说话人。

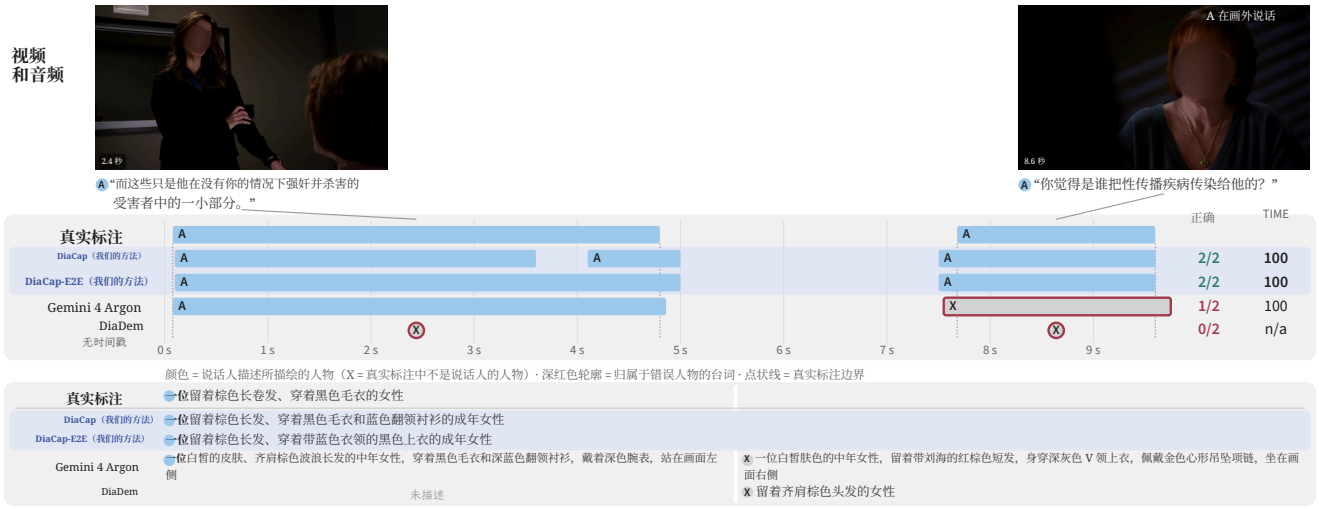


图8。**一名走出画面的说话人**。在这个典型的正反打对话中，同一名女性说出了两句台词。第二句台词期间，摄像机切换到听者身上，而说话人位于画面外。DiaCap 和 DiaCap-E2E 将两句台词都归因于同一名说话人。Gemini 4 Argon 将第二句台词归给了画面中可见的听者（留着短红棕色头发）。DiaDem 的描述（“齐肩棕色头发的女性”）过于笼统，无法识别说话人，因此它的两句台词都未被归属。

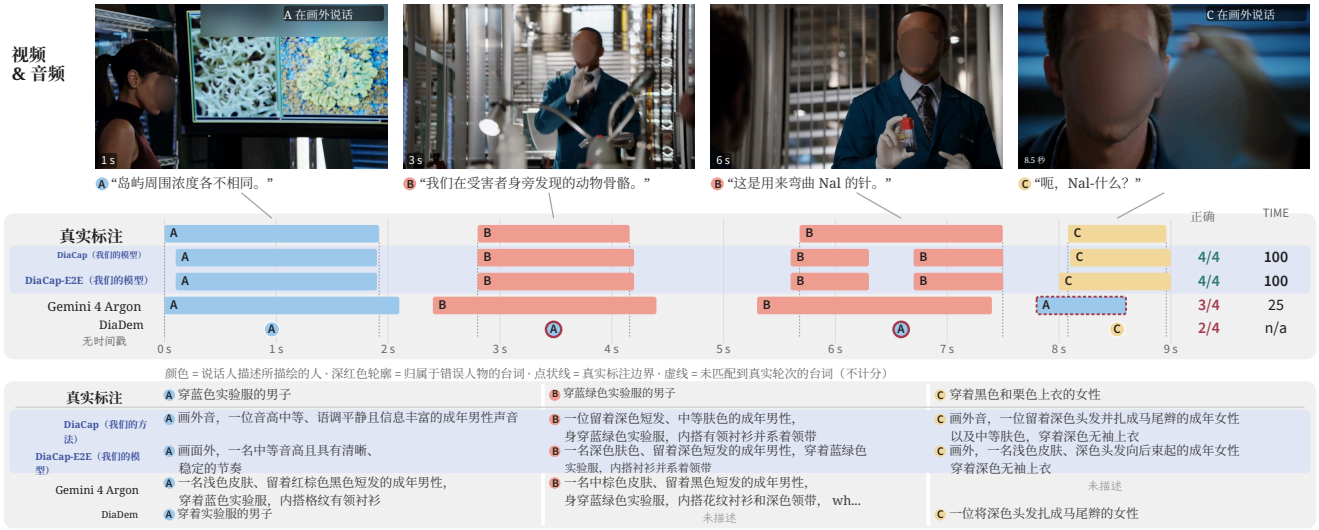


图9。**多镜头场景中的画外说话者**。第一和最后一个话轮是在摄像机拍摄其他人时说出的。DiaCap 和 DiaCap-E2E 正确地将全部四个话轮归属给相应说话人，并将这些说话人描述为画外声音。Gemini 4 Argon 将女性说的“呃，Nal——什么？”归给了穿蓝色实验服的男子，并截断了这句话。DiaDem 仅将两名男子都描述为“穿实验服的男子”，因此第二名男子的两个话轮被归给了第一名男子。

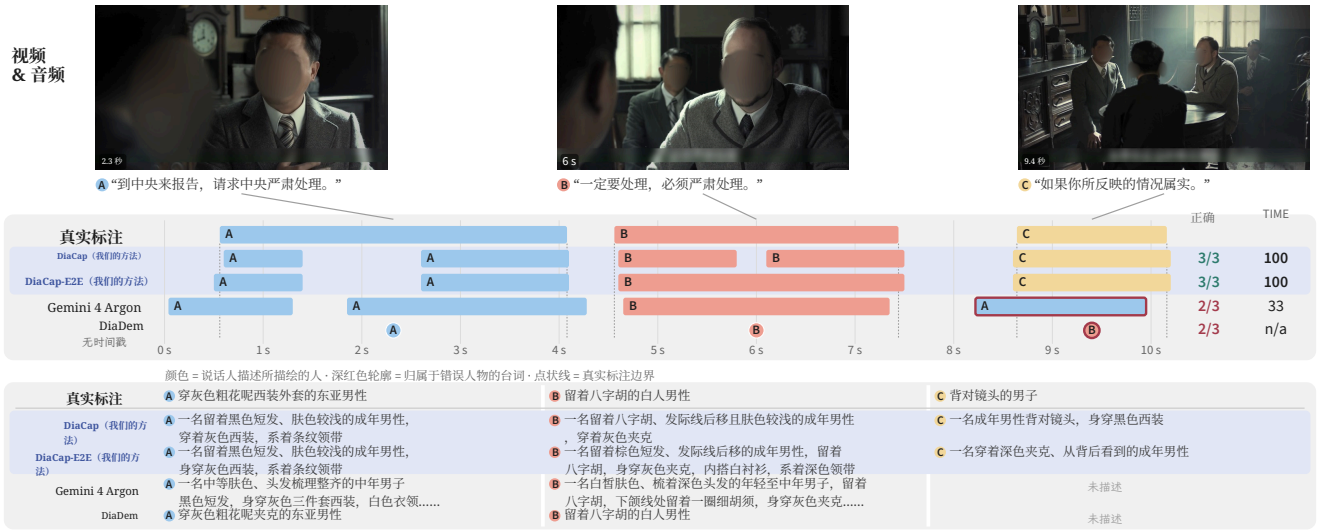


图 10. **一位只能看到背影的说话人**。最后一句台词由背对摄像机的男子说出。DiaCap 和 DiaCap-E2E 将这句台词归给了他。Gemini 4 Argon 将其归给穿灰色西装的男子，DiaDem 则归给留着八字胡的男子；这两人的脸都清晰可见。

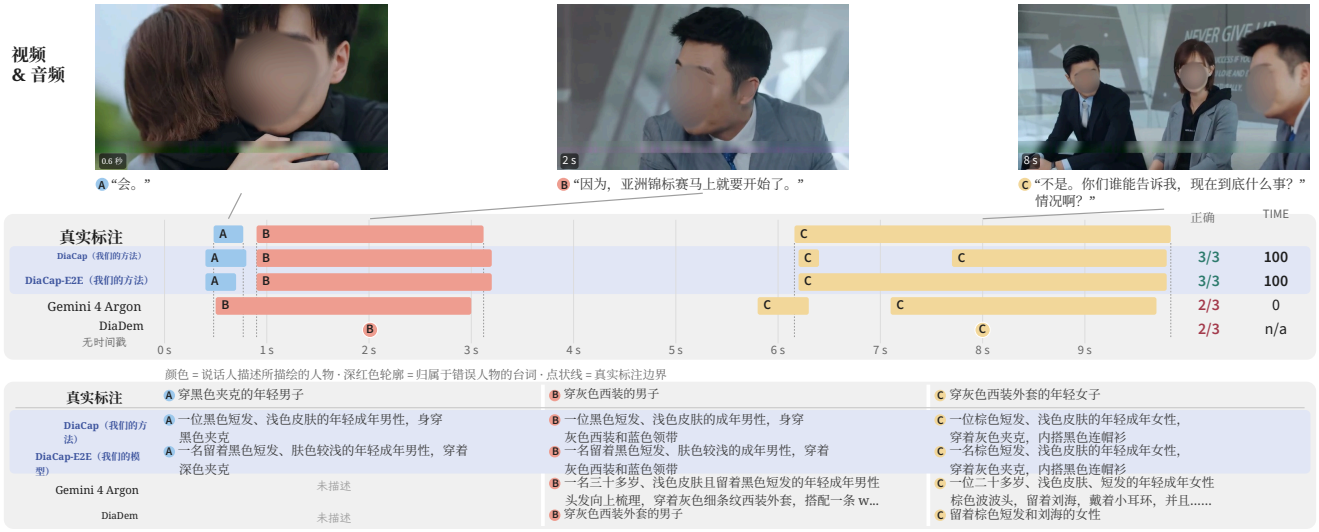


图 11. **另一位说话人发言前的一个词轮次（普通话）**。年轻男子用一个持续 0.3 秒的单音节作答（第一帧），穿灰色西装的男子随即接话（第二帧）。DiaCap 和 DiaCap-E2E 将这段简短发言保留为单独一行，并正确识别了说话人。Gemini 4 Argon 和 DiaDem 则将其并入下一位说话人的台词，因此丢失了轮次 A。Gemini 的时间戳也遗漏了所有边界（时间 0），而 DiaDem 没有预测出任何边界。这个例子表明，准确的时间戳与说话人分离密切相关：缺乏精确轮次边界的模型往往会将简短轮次并入相邻的较长轮次。



图12. **第三位说话者短暂插话。** DiaCap 和 DiaCap-E2E 正确归属了全部四个话轮，并将每个边界都置于容差区间内。Gemini 4 Argon 将女性的两个话轮合并成了一句台词，因此在对 Gemini 评分时，我们将同一说话者连续的真实话语合并，这样真实话轮就剩下三个。Gemini 的时间戳错过了大多数边界（时间 33）。DiaDem 对最后那段插话的说话者仅描述为“男人的声音”，这与场景中的任何人都不相符。

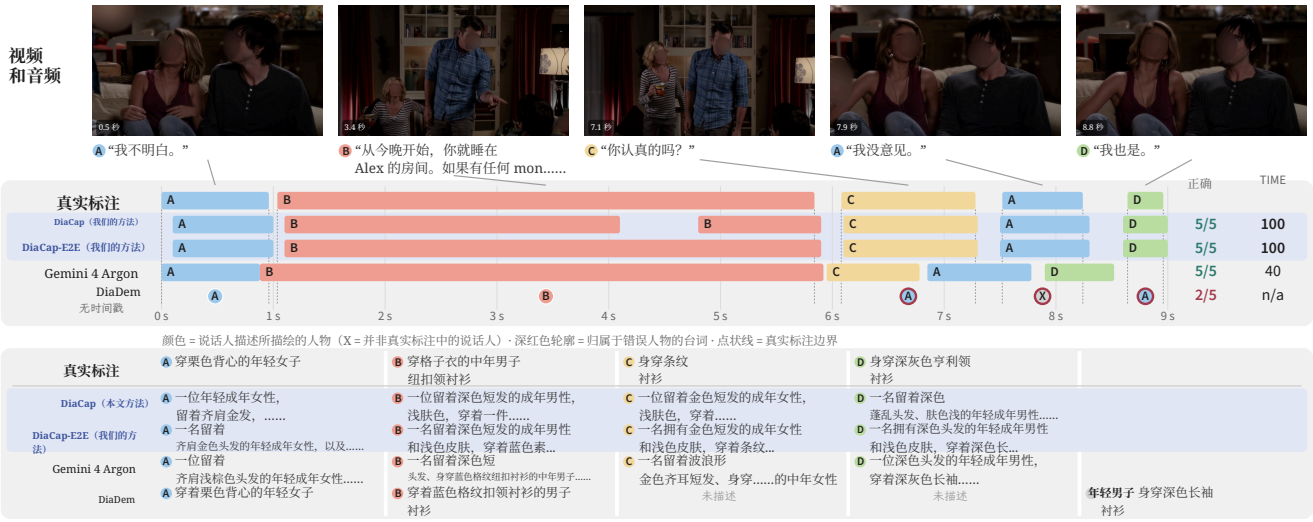


图13. **一个房间里的四位说话者。** DiaCap、DiaCap-E2E 和 Gemini 4 Argon 正确归属了全部五个话轮。DiaDem 错判了五个话轮中的三个。它将“你是认真的吗？”（C）和“我也是。”（D）归给了年轻女子，却把她自己说的“我对此没意见。”归给了一名年轻男子。Gemini 的时间戳在最后三个话轮上偏移了约 0.7 秒（时间 40），而 DiaCap 的两个版本都将每个边界置于容差区间内（时间 100）。

参考文献

[1] Xavier Anguera Miro、Simon Bozonnet、Nicholas Evans、Corinne Fredouille、Gerald Friedland 和 Oriol Vinyals。说话人分离：近期研究综述。IEEE音频、语音与语言处理汇刊，20(2):356–370，2012。2[2] Detao Bai、Xihan Wei 和 Zhiheng Ma。HumanOmni-Speaker：识别谁在何时说了什么。载于《计算机视觉——ECCV 2026》，第547–564页。Springer，2026。2[3] Max Bain、Jaesung Huh、Tengda Han 和 Andrew Zisserman。WhisperX：长音频的精准计时语音转写。载于 Interspeech，第4489–4493页，2023。2[4] Herv´e Bredin。pyannote.audio 2.1说话人分离流水线：原理、基准与实践指南。载于 Interspeech，第1983–1987页，2023。1、2、3[5] Andrew Brown、Vicky Kalogeiton 和 Andrew Zisserman。面部、身体、声音：基于多模态的视频人物聚类。载于 IEEE/CVF国际计算机视觉会议研讨会（ICCVW）论文集，第3177–3187页，2021年。2[6] Sihan Chen、Handong Li、Qunbo Wang、Zijia Zhao、Mingzhen Sun、Xinxin Zhu 和 Jing Liu。VAST：视觉-音频-字幕-文本全模态基础模型与数据集。载于《神经信息处理系统进展》，第72842–72866页，2023。2[7] Xinlong Chen、Yue Ding、Weihong Lin、Jingyun Hua、Linli Yao、Yang Shi、Bozhou Li、Yuanxing Zhang、Qiang Liu、Pengfei Wan、Liang Wang 和 Tieniu Tan。AVoCaDO：由时间编排驱动的视听视频描述器。arXiv预印本 arXiv:2510.10395，2025年。2、6[8] Xinlong Chen、Weihong Lin、Jingyun Hua、Linli Yao、Yue Ding、Bozhou Li、Bohan Zeng、Yang Shi、Qiang Liu、Yuanxing Zhang、Pengfei Wan、Liang Wang 和 Tieniu Tan。Dia-Dem：推进面向多模态大语言模型的视听视频描述中的对话描述。arXiv预印本 arXiv:2601.19267，2026年。1、2、5、6、7、11[9] Luyao Cheng、Hui Wang、Chong Deng、Siqi Zheng、Yafeng Chen、Rongjie Huang、Qinglin Zhang、Qian Chen、Xihao Li 和 Wen Wang。融合音频、视觉与语义信息，以增强多方对话中的多模态说话人分离。载于第63届计算语言学协会年会论文集（第1卷：长篇论文），第19914–19928页，2025年。2[10] Zesen Cheng、Sicong Leng、Hang Zhang、Yifei Xin、Xin Li、Guanzheng Chen、Yongxin Zhu、Wenqi Zhang、Ziyang Luo、Deli Zhao 和 Lidong Bing。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。2[11]Sanjoy Chowdhury、Sayan Nag、Subhrajyoti Dasgupta、Jun Chen、Mohamed Elhoseiny、Ruohan Gao 和 Dinesh Manocha。Meerkat：用于时空定位的视听大语言模型。载于欧洲计算机视觉会议（ECCV），2024年。2[12] Joon Son Chung 和 Andrew Zisserman。Out of time：自然环境中的自动唇形同步。载于亚洲计算机视觉会议（ACCV）研讨会，第251–263页，2016年。2[13] Joon Son Chung、Bong-Jin Lee 和 Icksang Han。《谁在说话？：真实世界会议中的音视频说话人日志》。载于 Interspeech，第 371–375 页，2019年。1, 2[14] Gheorghe Comanici、Eric Bieber、Mike Schaekermann、Ice Pasupat、Noveen Sachdeva、Inderjit Dhillon 等。《Gemini 2.5：以高级推理、多模态、长上下文和新一代智能体能力推动前沿发展》。arXiv 预印本 arXiv:2507.06261，2025年。2, 5, 6[15] Cheng Cui、Yubo Zhang、Ting Sun、Xueqing Wang、Hongen Liu、Manhui Lin、Yue Zhang、Tingquan Gao、Changda Zhou、Jiaxuan Liu 等。《PP-OCRv5：一种专用的500万参数模型，在OCR任务上可媲美十亿参数视觉语言模型》。arXiv 预印本 arXiv:2603.24373，2026年。5[16] Ross Cutler 和 Larry Davis。《看谁在说话：利用视频与音频相关性进行说话人检测》。载于 IEEE 多媒体与博览会国际会议（ICME），第 1589–1592 页，2000年。2[17] Laurent El Shafey、Hagen Soltau 和 Izhak Shafran。《通过序列转换联合进行语音识别和说话人分离》。载于 Interspeech，第 396–400 页，2019年。2[18] ElevenLabs。《Scribe v2》。https://elevenlabs.io/blog/introducing-scribe-v2，2026年。语音转文本模型。2, 3, 7[19] Mark Everingham、Josef Sivic 和 Andrew Zisserman。《“你好！我叫……Buffy”——电视视频中角色的自动命名》。载于英国机器视觉会议（BMVC），2006年。2[20] Chaoyou Fu、Haojia Lin、Zuwei Long、Yunhang Shen、Yuhang Dai、Meng Zhao 等。《VITA：迈向开源交互式全能多模态大语言模型》。arXiv 预印本 arXiv:2408.05211，2024年。2[21] Yusuke Fujita、Naoyuki Kanda、Shota Horiguchi、Kenji Nagamatsu 和 Shinji Watanabe。《采用无排列目标的端到端神经说话人分离》。载于 Interspeech，第 4300–4304 页，2019年。1, 2[22] Jiyang Gao、Chen Sun、Zhenheng Yang 和 Ram Nevatia。《TALL：通过语言查询进行时间活动定位》。载于 IEEE国际计算机视觉会议（ICCV）论文集，第 5277–5285 页，2017年。2[23] Yuying Ge、Yixiao Ge、Chen Li、Teng Wang、Junfu Pu、Yizhuo Li、Lu Qiu、Jin Ma、Lisheng Duan、Xinyu Zuo、Jinwen Luo、Weibo Gu、Zexuan Li、Xiaojing Zhang、Yangyu Tao、Han Hu、Di Wang 和 Ying Shan。《ARC-Hunyuan-Video-7B：真实世界短视频的结构化视频理解》。arXiv 预印本 arXiv:2507.20939，2025年。2, 6[24] Israel D. Gebru、Sil`eye Ba、Xiaofei Li 和 Radu Horaud。《基于时空贝叶斯融合的音视频说话人分离》。IEEE模式分析与机器智能汇刊，40(5):1086–1099，2018年。2[25] 谷歌。《Gemini 4 Argon》。https://aistudio.google.com/docs/models/gemini-4-argon，2026年。访问日期：2026年10月22日。1, 2, 3, 6, 7, 11[26] Google DeepMind。《Gemini 3.1 Pro 模型卡》。https://deepmind.google/models/model-cards/gemini-3-1-pro/，2026年。访问日期：2026年11月9日。1, 2, 4, 5

[27] Lisa Anne Hendricks、Oliver Wang、Eli Shechtman、Josef Sivic、Trevor Darrell 和 Bryan Russell。《利用自然语言定位视频中的片段》。载于 IEEE国际计算机视觉会议（ICCV）论文集，第 5804–5813 页，2017年。2[28] Shota Horiguchi、Yusuke Fujita、Shinji Watanabe、Yawen Xue 和 Kenji Nagamatsu。《针对未知数量说话人、采用基于编码器-解码器吸引子的端到端说话人分离》。载于 Interspeech，第 269–273 页，2020年。2[29] Edward J. Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。《LoRA：大语言模型的低秩适配》。载于国际学习表征会议（ICLR），2022年。5[30] Bin Huang、Xin Wang、Hong Chen、Zihan Song 和 Wenwu Zhu。《VTimeLLM：助力大语言模型理解视频片段》。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第 14271–14280 页，2024年。2[31] Liangbin Huang、Xiaohua Liao、Chaoqun Cui、Shijing Wang、Zhaolong Huang、Yanlong Du 和 Wenji Mao。《CineSRD：利用视觉、声学和语言线索进行开放世界视觉媒体说话人分离》。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第 8836–8845 页，2026年。2[32] Jaesung Huh 和 Andrew Zisserman。《上下文中的角色感知音视频字幕生成》。载于亚洲计算机视觉会议（ACCV）论文集，第 2613–2631 页，2024年。2[33] Vladimir Iashin、Weidi Xie、Esa Rahtu 和 Andrew Zisserman。《Synchformer：利用稀疏线索实现高效同步》。载于 ICASSP 2024——2024年IEEE国际声学、语音与信号处理会议，第 5325–5329 页。IEEE，2024年。2, 3[34] Naoyuki Kanda、Yashesh Gaur、Xiaofei Wang、Zhong Meng 和 Takuya Yoshioka。《用于端到端重叠语音识别的序列化输出训练》。载于 Interspeech，第 2797–2801 页，2020年。2[35] Naoyuki Kanda、Xiong Xiao、Yashesh Gaur、Xiaofei Wang、Zhong Meng、Zhuo Chen 和 Takuya Yoshioka。《先转录后分离：使用端到端说话人归属语音识别处理不限数量说话人的神经说话人分离》。载于 IEEE国际声学、语音与信号处理会议（ICASSP），第 8082–8086 页，2022年。2[36] Keisuke Kinoshita、Marc Delcroix 和 Naohiro Tawara。《融合端到端神经网络与聚类式说话人分离：兼得两者之长》。载于 IEEE国际声学、语音与信号处理会议（ICASSP），第 7198–7202 页，2021年。2[37] Zvi Kons、Avihu Dekel、Hagai Aronowitz、Vishal Sunder 和 Ron Hoory。《使用非自回归大语言模型生成词级时间戳并进行说话人归属》。arXiv 预印本 arXiv:2609.15218，2026年。2[38] Okan K¨op¨ukl¨u、Maja Taseska 和 Gerhard Rigoll。《如何设计用于自然环境中音视频主动说话人检测的三阶段架构》。载于 IEEE/CVF国际计算机视觉会议（ICCV）论文集，第 1173–1183 页，2021年。2[39] Bruno Korbar、Jaesung Huh 和 Andrew Zisserman。《看、听与识别：角色感知的视听字幕生成》。载于2024年IEEE国际声学、语音与信号处理会议（ICASSP）论文集，第2975–2979页。IEEE，2024年。2[40] Ranjay Krishna、Kenji Hata、Frederic Ren、李飞飞和 Juan Carlos Niebles。《视频事件的密集字幕生成》。载于IEEE国际计算机视觉会议（ICCV）论文集，第706–715页，2017年。2[41]Sonal Kumar、Prem Seetharaman、Ke Chen、Oriol Ni- eto、Jiaqi Su、Zhepei Wang、Rithesh Kumar、Dinesh Manocha、Nicholas J. Bryan、Zeyu Jin 和 Justin Salamon。《TAC：带时间戳的音频字幕生成》。arXiv预印本 arXiv:2602.15766，2026年。2, 4[42] Ludwig K¨urzinger、Dominik Winkelbauer、Lujun Li、Tobias Watzel 和 Gerhard Rigoll。《面向德语端到端语音识别的大型语料库CTC分段》。载于Speech and Computer（SPECOM），第267–278页，2020年。2[43] Federico Landini、J´an Profant、Mireia Diez 和 Luk´aˇs Bur- get。《说话人分离中x向量序列的贝叶斯隐马尔可夫模型聚类（VBx）：理论、实现及标准任务分析》。Computer Speech & Language，71：101254，2022年。1, 2[44] Jie Lei、Tamara L. Berg 和 Mohit Bansal。《通过自然语言查询检测视频中的片段和精彩时刻》。载于神经信息处理系统进展，2021年。2[45]Juan Le´on Alc´azar、Fabian Caba、Long Mai、Federico Perazzi、Joon-Young Lee、Pablo Arbel´aez 和 Bernard Ghanem。《上下文中的活跃说话人》。载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第12462–12471页，2020年。2[46] Junhua Liao、Haihan Duan、Kanghui Feng、Wanbing Zhao、Yanbing Yang 和 Liangyin Chen。《用于主动说话人检测的轻量级模型》。载于2023年IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第22932–22941页。IEEE，2023年。2[47] Junhua Liao、Haihan Duan、Kanghui Feng、Wanbing Zhao、Yanbing Yang、Liangyin Chen 和 Yanru Chen。《LR-ASD：用于主动说话人检测的轻量级鲁棒网络》。国际计算机视觉杂志，133(7)：4749–4769，2025年。3[48] Jing Liu、Sihan Chen、Xingjian He、Longteng Guo、Xinxin Zhu、Weining Wang 和 Jinhui Tang。《VALOR：视觉-音频-语言全方位感知预训练模型与数据集》。IEEE模式分析与机器智能汇刊，47(2)：708–724，2025年。2[49] Tao Liu、Shuai Fan、Xu Xiang、Hongbo Song、Shaoxiong Lin、Jiaqi Sun、Tianyuan Han、Siyuan Chen、Binwei Yao、Sen Liu、Yifei Wu、Yanmin Qian 和 Kai Yu。《MSDWild：自然环境中的多模态说话人分离数据集》。载于Interspeech，第1476–1480页，2022年。1, 2[50] Zichen Liu、Changyu Chen、Wenjun Li、Penghui Qi、Tianyu Pang、Chao Du、Wee Sun Lee 和 Min Lin。《理解类R1-Zero训练：批判性视角》。arXiv预印本 arXiv:2503.20783，2025年。5[51] Ziyang Ma、Ruiyang Xu、Zhenghao Xing、Yunfei Chu、Yux- uan Wang、Jinzheng He、Jin Xu、Pheng-Ann Heng、Kai

Yu、Junyang Lin、Eng Siong Chng 和 Xie Chen。《Omni-Captioner：面向全方位细粒度感知的数据流水线、模型与基准》。载于国际学习表征会议，2026年。2[52]Michael McAuliffe、Michaela Socolof、Sarah Mihuc、Michael Wagner 和 Morgan Sonderegger。《Montreal Forced Aligner：使用Kaldi进行可训练的文本-语音对齐》。载于Interspeech，第498–502页，2017年。2[53] Microsoft AI。MAI-Transcribe-2。https : / / microsoft . ai / models / mai - transcribe - 2/，2026年。语音转文本模型。2, 3, 7[54] Kyle Min、Sourya Roy、Subarna Tripathi、Tanaya Guha 和 Somdeb Majumdar。《学习用于主动说话人检测的长期时空图》。载于欧洲计算机视觉会议（ECCV），第

371 387, 2022 2 – 。

[55] MOSI.AI。《MOSS转录与说话人分离技术报告》。arXiv预印本 arXiv:2601.01554，2026年。2[56] Arsha Nagrani 和 Andrew Zisserman。《从Benedict Cumberbatch到Sherlock Holmes：无需剧本识别电视剧集中的角色》。载于英国机器视觉会议（BMVC），2017年。2[57] Le Thien Phuc Nguyen、Zhuoran Yu、Samuel Low Yu Hang、Subin An、Jeongik Lee、Yohan Ban、SeungEun Chung、Thanh-Huy Nguyen、JuWan Maeng、Soochahn Lee 和 Yong Jae Lee。《看、听与理解：多模态大语言模型视听人类语音理解基准测试》。载于IEEE/CVF计算机视觉与模式识别会议（CVPR）成果论文集，第

2272 2283, 2026 2, 8, 10 – 。

[58] Andrew Owens 和 Alexei A. Efros。《利用自监督多感官特征进行视听场景分析》。载于欧洲计算机视觉会议（ECCV），第639–658页，2018年。2[59] Taejin Park、Ivan Medennikov、Kunal Dhawan、Weiqing Wang、He Huang、Nithin Rao Koluguri、Krishna C. Puvvada、Jagadeesh Balam 和 Boris Ginsburg。《Sortformer：一种用于语音转文本系统中置换解析的说话人监督的新方法》。载于国际机器学习会议（ICML），

2025 2 。

[60] Tae Jin Park、Naoyuki Kanda、Dimitrios Dimitriadis、Kyu J. Han、Shinji Watanabe 和 Shrikanth Narayanan。《说话人分离综述：深度学习的最新进展》。Computer Speech & Language，72：101317，2022年。2[61] Zhiliang Peng 等。《VIBEVOICE-ASR技术报告》。arXiv预印本 arXiv:2601.18184，2026年。2[62] Alexis Plaquet 和 Herv´e Bredin。《神经说话人分离的幂集多分类交叉熵损失》。载于Interspeech 2023，第3222–3226页，2023年。2, 3[63] Junfu Pu、Yuxin Chen、Teng Wang 和 Ying Shan。《OmniScript：面向长篇电影视频的视听脚本生成》。arXiv预印本 arXiv:

2604 11102, 2026 2 。

[64] 3 5 Qwen团队。《Qwen . -Omni技术报告》。arXiv预印本 arXiv:2604.15804，2026年。1, 2[65] Shuhuai Ren、Linli Yao、Shicheng Li、Xu Sun 和 Lu Hou。《TimeChat：用于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（CVPR）论文集，第14313–14323页，2024年。2[66]Joseph Roth、Sourish Chaudhuri、Ondrej Klejch、Rad- hika Marvin、Andrew Gallagher、Liat Kaver、Sharadh Ramaswamy、Arkadiusz Stopczynski、Cordelia Schmid、Zhonghua Xi 和 Caroline Pantofaru。AVA Active Speaker：用于主动说话人检测的音视频数据集。载于IEEE国际声学、语音与信号处理会议（ICASSP），第 4492–4496 页，2020年。2[67] Jiapeng Shi、Qiuxia Hou、Jiaqi Leng、Junke Wang、Yanhao Zhang、Haonan Lu、Ziyi Ye 和 Zuxuan Wu。EVA-Cap：通过以事件为中心的对齐优化视听视频描述。载于神经信息处理系统进展，2026年。已录用。2[68]Xian Shi、Xiong Wang、Zhifang Guo、Yongqi Wang、Pei Zhang、Xinyu Zhang、Zishan Guo、Hongkun Hao、Yu Xi、Baosong Yang、Jin Xu、Jingren Zhou 和 Jun- yang Lin。Qwen3-ASR技术报告。arXiv预印本 arXiv:2601.21337，2026年。2[69] David Snyder、Daniel Garcia-Romero、Gregory Sell、Daniel Povey 和 Sanjeev Khudanpur。X向量：用于说话人识别的鲁棒深度神经网络嵌入。载于 IEEE国际声学、语音与信号处理会议（ICASSP），第 5329–5333 页，2018年。2[70] Guangzhi Sun、Wenyi Yu、Changli Tang、Xianzhao Chen、Tian Tan、Wei Li、Lu Lu、Zejun Ma、Yuxuan Wang 和 Chao Zhang。video-SALMONN：语音增强型音视频大语言模型。载于国际机器学习会议（ICML），2024年。2[71] Changli Tang、Yixuan Li、Yudong Yang、Jimin Zhuang、Guangzhi Sun、Wei Li、Zejun Ma 和 Chao Zhang。video-SALMONN 2：描述文本增强型音视频大语言模型。arXiv预印本 arXiv:2506.15220，2025年。2[72] Changli Tang、Tianyi Wang、Fengyun Rao、Jing Lyu 和 Chao Zhang。D-ORCA：以对话为中心的鲁棒音视频描述优化。arXiv预印本 arXiv:2602.07960，2026年。1, 2, 6, 8[73] Yunlong Tang、Daiki Shimada、Jing Bi、Mingqian Feng、Hang Hua 和 Chenliang Xu。利用伪未剪辑视频增强大语言模型的音视频时间理解能力。载于 AAAI人工智能会议论文集，2025年。2[74] Ruijie Tao、Zexu Pan、Rohan Kumar Das、Xinyuan Qian、Mike Zheng Shou 和 Haizhou Li。有人在说话吗？探索用于音视频主动说话人检测的长期时序特征。载于ACM国际多媒体会议，第 3927–3935 页，2021年。2[75] 腾讯混元团队。Script-a-Video：通过因式分解流和关系定位生成深度结构化音视频描述。arXiv预印本 arXiv:2604.11244，2026年。2[76] Quan Wang、Carlton Downey、Li Wan、Philip Andrew Mans- field 和 Ignacio Lopez Moreno。基于长短期记忆网络的说话人分离。载于 IEEE国际声学、语音与信号处理会议（ICASSP），第 5239 5243 页，2018年 2 – 。[77] Quan Wang、Yiling Huang、Guanlong Zhao、Evan Clark、Wei Xia 和 Hank Liao。DiarizationLM：利用大语言模型进行说话人日志后处理。载于 Interspeech，第 3754–3758 页，2024年。2

[78] Xizi Wang、Feng Cheng 和 Gedas Bertasius。LoCo Net：用于主动说话人检测的长短期上下文网络。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第 18462 18472 页，2024年 2 – 。[79] Ye Wang、Ziheng Wang、Boshen Xu、Yang Du、Kejun Lin、Zihan Xiao、Zihao Yue、Jianzhong Ju、Liang Zhang、Dingyi Yang、Xiangnan Fang、Zewen He、Zhenbo Luo、Wenxuan Wang、Junqi Lin、Jian Luan 和 Qin Jin。Time-R1：用于时间视频定位的预训练后大型视觉语言模型。arXiv预印本 arXiv:2503.13377，2025年。2[80] Mingyang Wu、Kaituo Feng、Bohao Li、Kaixiong Gong、Zi- hao Yin 和 Xiangyu Yue。AVCap：通过细节感知奖励强化音视频联合描述。arXiv预印本 arXiv:2608.06930，2026年。2[81] Peiran Wu、Yunze Liu、Zhengdong Zhu、Enmin Zhou 和 Junxiao Shen。UGC-VideoCaptioner：全能型UGC视频细节描述模型与全新基准测试。arXiv预印本 arXiv:2507.11336，2025年。2, 6[82] 小米LLM-Core团队。MiMo-V2.6：推动强化学习迈向自我改进。arXiv预印本 arXiv:2610.11959，2026年。2, 6[83] Eric Zhongcong Xu、Zeyang Song、Satoshi Tsutsui、Chao Feng、Mang Ye 和 Mike Zheng Shou。AVA-AVD：自然环境中的音视频说话人分离。载于 ACM国际多媒体会议，第 3838–3847 页，2022年。1, 2[84]Jin Xu、Zhifang Guo、Jinzheng He、Hangrui Hu、Ting He、Shuai Bai、Keqin Chen、Jialin Wang、Yang Fan、Kai Dang、Bin Zhang、Xiong Wang、Yunfei Chu 和 Jun- yang Lin。Qwen2.5-Omni技术报告。arXiv预印本 arXiv:2503.20215，2025年。2, 6[85] Jin Xu、Zhifang Guo、Hangrui Hu、Yunfei Chu、Xiong Wang、Jinzheng He、Yuxuan Wang、Xian Shi、Ting He、Xinfa Zhu 等。Qwen3-Omni技术报告。arXiv预印本 arXiv:2509.17765，2025年。1, 2, 3, 4, 5, 6, 7[86] Antoine Yang、Arsha Nagrani、Paul Hongsuck Seo、An- toine Miech、Jordi Pont-Tuset、Ivan Laptev、Josef Sivic 和 Cordelia Schmid。Vid2Seq：用于密集视频描述的大规模视觉语言模型预训练。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第 10714–10726 页，2023年。2[87] Qize Yang、Shimin Yao、Weixuan Chen、Shenghao Fu、De- tao Bai、Jiaxing Zhao、Boyuan Sun、Bowen Yin、Xihan Wei 和 Jingren Zhou。HumanOmniV2：从理解到结合上下文的全模态推理。arXiv预印本 arXiv:2506.21277，2025年。6[88] Linli Yao、Yuancheng Wei、Yaojie Zhang、Lei Li、Xinlong Chen、Feifan Song、Ziyue Wang、Kun Ouyang、Yuanxin Liu、Lingpeng Kong、Qi Liu、Pengfei Wan、Kun Gai、Yuanxing Zhang 和 Xu Sun。TimeChat-Captioner：利用时间感知的结构化音视频描述为多场景视频编写脚本。载于第43届国际机器学习会议论文集，第 147673–147693 页。PMLR，2026年。2[89]Hanrong Ye、Chao-Han Huck Yang、Arushi Goel、Wei Huang、Ligeng Zhu、Yuanhang Su、Sean Lin、An-Chieh Cheng、Zhen Wan、Jinchuan Tian 等。OmniVinci：En-

用于全模态理解大语言模型的架构与数据。 *arXiv*预印本 *arXiv:2510.15870*，2025年。2[90] Han Yin、Yafeng Chen、Chong Deng、Luyao Cheng、Hui Wang、Chao-Hong Tan、Qian Chen、Wen Wang 和 Xian- gang Li。SpeakerLM：利用多模态大语言模型实现端到端、多功能的说话人分离与识别。*AAAI* 人工智能会议论文集，40(40):34467–34475，2026 年。2[91] Qiying Yu、Zheng Zhang、Ruofei Zhu 等。DAPO：一个可大规模应用的开源大语言模型强化学习系统。发表于 神经信息处理系统进展，2025年。2，5[92] Aonan Zhang、Quan Wang、Zhenyao Zhu、John Paisley 和 Chong Wang。全监督说话人分离。发表于*IEEE*国际声学、语音与信号处理会议（*ICASSP*），第6301–6305页，2019年。2[93] Hang Zhang、Xin Li 和 Lidong Bing。Video-LLaMA：一种用于视频理解的指令微调音视语言模型。发表于自然语言处理实证方法会议论文集：系统演示，第543–553页，2023年。2