arXiv:2603.07263v1 [cs.SD] 7 Mar 2026

看见上下文：通过多模态推理实现丰富视觉上下文感知语音识别

Wenjie Tian[1], *Mingchen Shao*[1], *Bingshen Mu*[1],*Xuelong Geng*[1],*Chengyou Wang*[1],*Yujie Liao*[1], *Zhixian Zhao*[1], *Ziyu Zhang*[1],*Jingbin Hu*[1],*Mengqi Wei*[1],*Lei Xie*[1], **

[1] 中国西安，西北工业大学twj@mail.nwpu.edu.cn, lxie@nwpu.edu.cn

摘要

视听语音识别（AVSR）是自动语音识别（ASR）的扩展，融合了视觉信号。目前的 AVSR 方法主要关注唇部运动，很大程度上忽略了视频中丰富的上下文信息，例如说话场景和屏幕文字。为应对这类 CAVSR（包含丰富视觉上下文的 AVSR）任务，我们提出了 VASR，旨在“看见”并推理视觉上下文，以提升语音识别效果。具体而言，我们构建了音视频思维链（AV-CoT），明确要求模型在声学信号与视觉证据之间进行中间跨模态对齐。这种由证据驱动的推理缓解了“单模态主导”问题，即模型要么过度依赖视觉上下文，要么无法有效利用视觉上下文。此外，为解决数据匮乏问题，我们构建并发布了相应的数据流水线和测试集。实验表明，AV-CoT 能有效缓解单模态主导问题，并在 CAVSR 任务上取得最先进的性能。该项目已开源。

索引词：视听语音识别、多模态大语言模型、多模态推理、数据匮乏

1. 引言

近年来，自动语音识别（ASR） [1, 2, 3] 取得了显著进展，这在很大程度上得益于训练数据和模型容量的扩大。然而，这类纯音频系统在需要结合上下文进行消歧的场景中仍表现不佳，例如识别同音词、命名实体和领域专有术语。这一局限主要源于系统仅依赖声学信号，而声学信号所提供的上下文证据不足。尽管已有一些视听语音识别（AVSR）研究 [4, 5, 6, 7, 8, 9] 尝试纳入更多上下文，但其中大多数主要局限于唇部运动，这要求说话者面部正对镜头且清晰可见。这些研究忽略了丰富的环境视觉上下文，例如现代多媒体中普遍存在的特定场景、物体和字幕。丰富的视觉上下文对于解决转写歧义极有帮助。对于例如，如图 1 所示，在没有额外上下文线索的情况下，准确转录语音中提及的某个人名或领域专有术语可能存在很大歧义。这引出了一个关键问题：我们能否利用如此丰富的视觉上下文来提高语音识别准确率？我们将这项任务称为上下文音视频语音识别（CAVSR）。

模型和数据这两个关键方面使 CAVSR 面临挑战。从建模角度来看，近期的多模态大语言模型（MLLMs） [10, 11, 12, 13,14] 能够对完整的视频和音频进行编码，因此为 CAVSR 奠定了良好的基础，



图 1：不同识别任务的比较：自动语音识别（*ASR*）、侧重唇读的音视频语音识别（*AVSR*），以及侧重丰富视觉上下文的 *CAVSR*。该示例展示了 *CAVSR*任务如何利用丰富的视觉上下文来提高识别准确率。

然而，我们发现，将 MLLM 直接应用于 CAVSR 会因“单模态主导”而导致系统性失效。具体而言，当屏幕字幕与实际声学证据相矛盾时，模型往往会依据视觉文本产生幻觉，从而忽视音频传达的主要语音事实。反过来，MLLM 也可能忽略有用的视觉线索，仅依赖含糊不清的音频。从数据角度来看，现有视听语音识别数据集不足以用于 CAVSR 研究。如表1所示，大多数视听语音识别数据集主要关注唇部运动，视频中的人脸往往受限，或背景经常模糊，因此几乎无法为自动语音识别提供有意义的视觉上下文。其他工作 [15, 16],提供了与演示文稿幻灯片配对的视频数据，但这些数据集仅限于高度受约束的场景。现有与 CAVSR 相关的工作 [17, 18, 19, 20, 21] 未能应对同音词丰富的语言（如中文）带来的关键挑战，也未探索如何有效利用视觉上下文来消除语音歧义。

为应对CAVSR任务中的这些挑战，我们提出**VASR**（**视觉感知语音识别**），这是一个旨在“观察”视觉上下文、以解决CAVSR中语言歧义的框架。VASR基于音视频思维链（AV-CoT）机制构建，将CAVSR定义为一个多步骤推理过程：**感知**（提取视觉上下文和音素序列）、**推理**（跨模态消歧）以及**转录**（生成）。通过强制引入这一中间推理结构，AV-CoT能够减轻单一模态主导的问题，并促进音频与视觉信息的证据驱动融合。此外，为克服这一新兴领域中的数据匮乏，我们开发了可扩展的数据流水线，并整理了VASR测试集，这是一个高质量、经人工

**表示通讯作者。

表1： 视听语音识别数据集及其对应的数据类型

数据集	语言	数据类型
LRS 系列 [22]	英语	唇读视频
CMLR [23]	中文	唇读视频
CN-CVS [24]	中文	唇读视频
Voxceleb2 [25]	多语言	唇读视频
SlideSpeech [15]	英语	演示文稿幻灯片
Chinese-LiPS [16]	中文	演示文稿幻灯片
如何2 [17]	英语	通用视频
VisSpeech [18]	英语	通用视频
SlideAVSR [19]	英语	通用视频
AVNS [20]	英语	通用视频
VASR（我们的方法）	中文	通用视频

核验的数据集，旨在极端语言歧义条件下测试模型。我们已在https://github.com/wjtian-wonderful/ContextAVSR/tree/main开源所有数据集、训练代码和模型权重。我们的贡献总结如下：

• 我们提出了 VASR，这是一种专注于 CAVSR 的多模态大语言模型框架，将研究重点从局部唇读转向丰富的视觉感知推理。

• 我们提出了 AV-CoT，这是一种新颖的多模态推理流程，能够明确引导多模态大语言模型执行跨模态消歧，从而缓解“单模态主导”问题。

• 我们推出并发布了 VASR测试集和可扩展的数据流水线，为评估 CAVSR 提供了首个全面的测试集。

• 通过大量实验，我们证明 VASR 的表现显著优于现有的强大多模态大语言模型。

2. 框架

2.1. 整体架构

VASR旨在通过结构化的多模态推理，在多样化的真实场景中实现CAVSR。借助AV-CoT，VASR将CAVSR任务重新表述为一个结构化的感知–推理–转录流程。如图2所示，对于由视频 V 和音频信号 A组成的多模态输入，该框架首先生成一条推理路径，其中包括多模态感知和跨模态消歧推理，随后生成转录文本$\hat{Y}$。

2.2. 多模态编码

基于Qwen2.5-Omni的 [10] 多模态编码器，VASR首先将原始媒体 V 和 A 编码到连续的特征空间中，再按时间顺序交错排列，形成融合多模态流 $\mathcal{M}$。这一统一表示 $\mathcal{M}$ 为后续推理提供了感知基础。

2.3. AV-CoT机制

我们将CAVSR任务表述为结构化的音视频思维链（AV-CoT）过程。给定多模态输入$\mathcal{M}$，转录文本通过因果自回归路径生成，其中感知与推理依次为最终转录文本的生成提供条件。

多模态感知。 AV-CoT 从感知 on 阶段开始，提取视觉上下文和声学音素线索。给定多模态输入 $\mathcal{M}$，模型首先提取可观察到的视觉线索，如场景、屏幕文字和主题，以形成视觉上下文 C_v。随后，它从音频中提取音素序列 P_a。对于普通话，我们使用拼音作为基本单位。形式上，这一阶段对应于给定多模态输入 $\mathcal{M}$时生成感知状态 $S_p = \{C_v, P_a\}$ 的条件概率。由于生成过程是自回归的，因此可分解为：

$$P(S_p|\mathcal{M}) = \prod_{i=1}^{|C_v|} P(c_i|\mathcal{M}, c_{<i}) \prod_{j=1}^{|P_a|} P(p_j|\mathcal{M}, C_v, p_{<j}), \quad (1)$$

其中， c_i 和 p_j 分别表示 C_v 和 P_a中的词元。这为后续的跨模态消歧推理奠定了基础。

跨模态消歧。 推理阶段执行跨模态消歧。模型不会直接将 P_a 映射为文本，而是生成推理轨迹 R，将歧义语音片段与视觉语义 C_v对齐。形式上，以 $\mathcal{M}$和感知状态为条件，自回归地生成 R：

$$P(R|\mathcal{M}, C_v, P_a) = \prod_{k=1}^{|R|} P(r_k|\mathcal{M}, C_v, P_a, r_{<k}). \quad (2)$$

为说明该推理轨迹中的核心消歧机制，考虑一个歧义语音片段 P_a p_{sub} =（例如音节 he yı）。传统的纯音频系统会选择 y =arg max $P($ $y|p_{sub})$，通常默认选择最常见的同音词。相比之下，AV-CoT会以视觉上下文 C_v为条件，从候选集 $\mathcal{H}(p_{sub})$中进行选择：

$$y^* = \arg\max P(y \mid p_{sub}, C_v). \quad (3)$$

通过整合 C_v，模型能够有效地重新调整每个候选项的可能性权重。推理轨迹 R 会明确阐述这一逻辑筛选过程，剔除语法上不合适的选项和上下文关联较弱的匹配项。这一过程确保所选的语义表示 y^* 既符合语音特征，又有视觉依据。

转写生成。 最后的转录阶段将推断出的感知和推理结果转换为转录文本。由于视觉接地 C_v、语音提取 P_a 和消歧 R 已经建模，解码便成为一个约束生成过程。转录文本 $\hat{Y}$ 通过以整个前序推理链为条件进行自回归生成。形式上，生成长度为 T 的最终转录序列的概率定义如下：

$$P(\hat{Y}|\mathcal{M}, C_v, P_a, R) = \prod_{t=1}^{T} P(\hat{y}_t|\mathcal{M}, C_v, P_a, R, \hat{y}_{<t}), \quad (4)$$

其中, $\hat{y}_t$ 是最终转录文本的第 t 个字符。这里， R 作为解码的引导约束。

最后，通过将这三个阶段串联起来，VASR框架的总体目标得以优雅地统一。在给定多模态输入 $\mathcal{M}$ 的条件下，生成感知状态$S_p = \{C_v, P_a\}$、推理轨迹 R以及最终转录 Y 的联合概率，按照我们提出的因果依赖链进行分解：

$$P(\hat{Y}, R, S_p|\mathcal{M}) = P(S_p|\mathcal{M}) \cdot P(R|\mathcal{M}, S_p) \cdot P(\hat{Y}|\mathcal{M}, S_p, R). \quad (5)$$

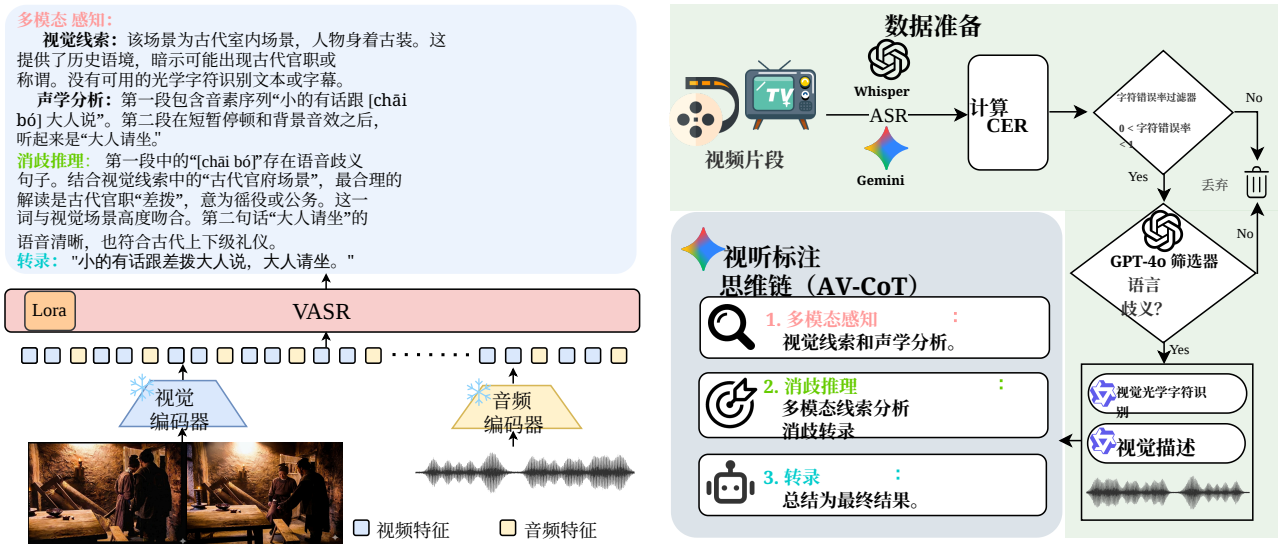


图2： *VASR*框架概览。（左）该模型架构实现了 *CAVSR*。一个定性示例展示了模型如何通过多模态推理解决语音歧义。（右） *AV-CoT* 的数据处理流程及其构建过程。

通过在训练过程中显式最大化这一联合概率，AV-CoT有效地迫使多模态大语言模型在判定其含义之前先澄清所见与所闻，从根本上弥合环境视觉感知与精确语言解码之间的差距。

3. 数据集

3.1. 数据流水线

为克服面向CAVSR的开源数据集稀缺问题，我们设计了一条自动化、可扩展的数据整理流水线。其目标是识别在语言歧义方面具有挑战性且视觉信息丰富的数据，并生成高质量的多模态标注。如图2右侧所示，我们将数据处理分为两个阶段：数据准备和标注。

数据准备。 为提高数据效率，我们首先采用基于声学歧义和语言复杂性的过滤机制。我们使用两个最先进水平模型进行初始自动语音识别：Gemini2.5Pro [14] 和 Whisper-Large-V3 [26]。我们计算两个模型所生成转录文本之间的字符错误率（CER）。为筛选出具有挑战性的样本，我们对数据集进行严格过滤，仅保留 $0 < \text{字符错误率} < 1$的片段。字符错误率 $= 0$ 的样本被丢弃，因为我们认为它们过于简单，仅靠音频模态便已足够。字符错误率 ≥ 1 的样本则被移除，以排除严重噪声数据。由于非零字符错误率并不能保证存在语言歧义，我们利用GPT4o分析转录文本，并判断其是否适合CAVSR。

因此，为避免标注过程中出现单模态主导，我们会明确提取与语音消歧相关的关键视觉元素，而不是直接使用原始视频帧。我们使用 Qwen2.5-VL [27] 进行光学字符识别（OCR）。为区分语音文本与环境背景，Qwen2.5-VL 会将屏幕文字分为两类：语音字幕，以及水印、品牌标志或印在物体上的文字等背景文字。同时，我们使用 Qwen2.5-VL 生成全面的视觉场景视频描述，捕捉环境、视频中出现的物体和动作。

AV-CoT标注。 获得视频标注后，我们将原始音频和视觉标注输入 Gemini2.5Pro，以生成推理路径。首先，我们进行多模态感知。对于视觉模态，我们将光学字符识别结果和视频描述总结为视觉线索。在声学分析中，我们提取语音序列，并利用发音序列（即中文拼音）识别歧义片段。提供更多语言和情境语境，有助于模型推断含糊或发音有歧义的语音片段中最可能的词语。随后，我们利用生成的视觉线索消解发音序列中的歧义片段，从而得到最终转录文本。

3.2. VASR测试集

我们从所用数据集中随机抽取 2,000 条话语，人工核验相应的转录文本，并构建最终的 CAVSR 测试集。同时，我们人工检查每个样本是否适用于 CAVSR任务，并剔除过于简单的样本。在人工筛选过程中，GPT 生成的理由供人工参考。最终测试集包含 1,981 条话语。

4. 实验

4.1. 数据集

我们使用了多个开源中文数据集，包括 MEIJU [28]和 MER 25 [29], ch-sims-v2 [30], 以及 Chinese-LiPS [16] 训练集。所有开源数据均使用我们提出的数据处理流程进行统一处理，总计约 230 小时。我们使用 Chinese-LiPS 测试集作为开源测试集。对于通用场景，我们还纳入了所提出的 VASR测试集，作为一个更全面、更具挑战性的数据集。

4.2. 实现细节

我们在所有实验中均采用 Qwen2.5-Omni-7B 作为基础模型。VASR 的视觉编码器和音频编码器在整个训练阶段均保持冻结。所有训练实验

表2：对比模型性能比较．最佳和次佳结果分别以粗体和下划线标示 *ly.*

模型	参数	输入模态	Task	字符错误率(%) ↓	
				Chinese-LiPS	VASR测试
Gemini2.5Pro [14]	-	A/V	CAVSR	4.59	<u>11.81</u>
Intern-S1 [12]	9B	A/V	CAVSR	71.94	17.6
MiniCPM-o2.6 [13]	8B	A/V	CAVSR	73.92	19.49
Qwen3Omni-Instruct-30B-A3B [11]	30B	A/V	CAVSR	<u>4.41</u>	11.97
Qwen3Omni-Thinking-30B-A3B [11]	30B	A/V	CAVSR	7.56	12.39
Qwen2.5Omni-7B [10]	7B	A/V	CAVSR	22.45	12.21
Qwen2.5Omni-3B [10]	3B	A/V	CAVSR	27.89	13.06
豆包语音识别	-	A	ASR	3.47	17.17
Qwen2.5Omni-7B	7B	A	ASR	11.71	19.25
VASR	7B	A/V	CAVSR	**1.80**	**11.02**

表3：用于验证 *AV-CoT* 贡献的消融实验。*"w/o AV-CoT"*：不使用 *AV-CoT* 进行直接监督微调；*"w/ Black Video"*：推理时使用黑色视频（全零帧）；*"w/ Random Video"*：推理时使用随机视频。

模型/指标	字符错误率(%) ↓	
	Chinese-LiPS	VASR测试
VASR	1.80	11.02
不使用 AV-COT	2.65	12.66
使用黑色视频	4.19	16.25
使用随机视频	5.29	18.94

均在启用混合精度的 16 张 NVIDIA A800 GPU 上执行。我们使用 LoRA 对思考模块进行微调，共训练三个轮次，批量大小总计为 96，学习率为 1×10^{-4}。LoRA 配置采用 8 的秩、32 的 alpha 值和 0.05 的丢弃率。在所有实验中，我们使用余弦学习率调度器，并设置 3,000 个预热步数。所有输入均使用 Qwen2.5-Omni 的处理器进行预处理，以确保与基础模型兼容。

4.3. 对比模型与评估指标。

由于若干现有的无约束音视频语音识别模型尚未开源，我们只能将基线模型 Qwen2.5-Omni-7B [10] 与几种强大的 Omni 模型进行比较：Qwen3-Omni-30B-A3B-Instruct [11], Qwen3-Omni-30B-A3B-Thinking [11], Intern-S1-9B [12], 和 MiniCPM-o2.6-8B [13]。我们还纳入了两个强大的商业基线：Gemini2.5Pro [14] 和豆包语音识别 [1]。我们采用字符错误率（CER）来衡量中文识别的字符错误率。

4.4. 实验结果

表2展示了VASR与最先进水平的商用和开源模型，以及强劲的纯音频基线之间的性能对比。我们提出的VASR在各种场景中始终取得最先进水平的结果，涵盖从密集文本视频到视觉信息丰富的视频。它以显著优势超越所有基线，包括Gem- ini2.5Pro和Qwen3Omni系列等大型模型。值得注意的是，VASR以Qwen2.5omni-7B为骨干网络，并通过LoRA进行微调使用的数据仅有数百小时。这凸显了AV-CoT方法在解决CAVSR任务时的简洁性和显著效果。尽管Qwen2.5Omni-7B的自动语音识别性能不及豆包语音识别，但我们提出的VASR模型有效利用视觉上下文，进一步提升了识别任务的性能上限。通过比较Qwen2.5Omni-7B的自动语音识别和CVASR结果，我们发现，若不解决单模态主导问题，丰富的视觉上下文往往会降低而非提升模型的识别性能。

在 Chinese-LiPS 数据集上，Intern-S1 和 MiniCPM-o2.6 出现了显著异常，其字符错误率分别高达 71.94% 和 73.92%。我们认为，这是因为它们受到幻灯片上密集文字的严重干扰。这种失效模式凸显了应对单模态主导这一关键局限的难度。

4.5. 消融研究

为验证 AV-CoT 的作用，我们在不包含 AV-CoT 的数据上对模型进行微调。如表3所示，这会导致模型在两个基准测试上的字符错误率均有所上升，表明模型未能充分利用视觉上下文。

为进一步验证 AV-CoT 确实利用视觉上下文来提升识别准确率，我们保持推理音频不变：一个变体使用随机选取的视频（使用随机视频），另一个变体使用所有帧均为零的视频（使用黑色视频）。黑色视频的结果表明，移除视觉上下文会导致性能下降，这说明模型会主动利用视觉线索，而非忽略视觉模态。与此同时，正如预期的那样，这两个变体的性能都下降到与豆包语音识别相当的水平，但仍优于大多数多模态基线。这表明模型不会过度依赖可能产生误导的视觉信号，并且即使视觉信息受到破坏，也能保持稳定的性能，说明 AV-CoT 有助于缓解单模态主导问题。

5. 结论与局限性

在这项工作中，我们提出了 CAVSR任务，将 AVSR 从唇读拓展到利用丰富的视觉上下文来消解语言歧义。我们提出了 VASR，它利用 AV-CoT 对转写过程进行显式建模，以应对单模态主导问题。此外，我们开发并发布了一套可扩展的数据构建流水线，并推出了 VASR

[1] https://www.volcengine.com/docs/6561/1354869?lang=en

测试集，以推动对 CAVSR任务开展系统性研究。大量实验表明，VASR 达到了最先进的性能，显著优于其他强大的多模态大语言模型。然而，预训练模型 Qwen2.5-Omni 中视觉编码器的帧率过低，因而我们无法将唇读任务纳入本项工作。

6. 生成式人工智能使用披露

本研究使用生成式人工智能工具进行文字润色和数据生成。此类工具不参与方法设计、实验、结果解读或科学成果的产出。作者对本文的所有内容承担全部学术责任。全体作者均同意提交本文。

7. 参考文献

[1] X. Shi、X. Wang、Z. Guo、Y. Wang、P. Zhang、X. Zhang、Z. Guo、H. Hao、Y. Xi、B. Yang、J. Xu、J. Zhou 和 J. Lin，“Qwen3-ASR 技术报告”，*CoRR*，第 abs/2601.21337 卷，2026 年。[2] K. Xu、F. Xie、X. Tang 和 Y. Hu，“FireRedASR：从编码器-解码器到大语言模型集成的开源工业级普通话语音识别模型”，*CoRR*，第 abs/2501.14350 卷，2025 年。[3] Z. Gao、S. Zhang、I. McLoughlin 和 Z. Yan，“Paraformer：用于非自回归端到端语音识别的快速且准确的并行变换器”，载于 *Proc.Interspeech*，2022 年，第 2063–2067 页。[4] U. Cappellazzo、X. Liu、P. Ma、S. Petridis 和 M. Pantic，“Omni-AVSR：借助大语言模型实现统一的多模态语音识别”，*CoRR*，第 abs/2511.07253 卷，2025 年。[5] J. Xue、X. Liu、X. Wu、X. Yin、D. Huang 和 F. Yu，“AD-AVSR：面向鲁棒视听语音识别的非对称双流增强”，*CoRR*，第 abs/2508.07608 卷，2025 年。[6] H. Wang、P. Guo、P. Zhou 和 L. Xie，“MLCA-AVSR：基于多层交叉注意力融合的视听语音识别”，载于 *Proc.ICASSP*，2024 年，第 8150–8154 页。[7] U. Cappellazzo、M. Kim 和 S. Petridis，“通过基于套娃结构的多模态大语言模型实现自适应视听语音识别”，*CoRR*，第 abs/2503.06362 卷，2025 年。[8] J. H. Yeo、H. Rha、S. J. Park 和 Y. M. Ro，“MMS-LLaMA：仅使用少量多模态语音词元的高效基于大语言模型的视听语音识别”，载于 *ACL Findings*，ACL 2025，奥地利维也纳，2025年7月27日至8月1日，Findings of ACL 系列，第 ACL 2025 卷，2025 年，第 20724–20735 页。[9] J. H. Yeo、M. Kim、C. W. Kim、S. Petridis 和 Y. M. Ro，“Zero-AVSR：通过学习语言无关的语音表征，利用大语言模型实现零样本视听语音识别”，*CoRR*，第 abs/2503.06273 卷，2025 年。[10] J. Xu、Z. Guo、J. He、H. Hu、T. He、S. Bai、K. Chen、J. Wang、Y. Fan、K. Dang、B. Zhang、X. Wang、Y. Chu 和 J. Lin，“Qwen2.5-Omni 技术报告”，*CoRR*，第 abs/2503. . 卷，2025 年。[11] J. Xu、Z. Guo、H. Hu、Y. Chu、X. Wang、J. He *et al.*，“Qwen3-Omni 技术报告”，*CoRR*，第 abs/2509.17765 卷，2025 年。[12] L. Bai、Z. Cai、Y. Cao、M. Cao、W. Cao、C. Chen *et al.*，“Intern-S1：科学多模态基础模型”，*CoRR*，第 abs/2508.15763 卷，2025 年。[13] Y. Yao、T. Yu、A. Zhang、C. Wang、J. Cui、H. Zhu、T. Cai、H. Li、W. Zhao、Z. He、Q. Chen、H. Zhou、Z. Zou、H. Zhang、S. Hu、Z. Zheng、J. Zhou、J. Cai、X. Han、G. Zeng、D. Li、Z. Liu 和 M. Sun，“MiniCPM-V：手机上的 GPT-4V 级多模态大语言模型”，*CoRR*，第 abs/2408.01800 卷，2024 年。[14] G. Comanici、E. Bieber、M. Schaekermann、I. Pasupat、N. Sachdeva *et al.*，“Gemini 2.5：凭借先进推理、多模态、长上下文和新一代智能体能力推进前沿”，*CoRR*，第 abs/2507.06261 卷，2025 年。[15] H. Wang、F. Yu、X. Shi、Y. Wang、S. Zhang 和 M. Li，“SlideSpeech：大规模幻灯片增强型音视频语料库”，载于 *Proc.ICASSP*，2024 年，第 11076–11080 页。[16] J. Zhao、Y. Jia、S. Wang、J. Zhou、H. Wang 和 Y. Qin，“Chinese-Lips：包含唇读和演示文稿幻灯片的中文音视频语音识别数据集”，载于 *Proc.ICME*，2025 年，第 1–6 页。

[17] R. Sanabria、O. Caglayan、S. Palaskar、D. Elliott、L. Barrault、L. Specia 和 F. Metze，“How2：用于多模态语言理解的大规模数据集”，*CoRR*，第 abs/1811 00347 卷，2018 年。[18] V. Gabeur、P. H. Seo、A. Nagrani、C. Sun、K. Alahari 和 C. Schmid，“AVATAR：无约束视听语音识别”，载于*Proc.Interspeech*，2022 年，第 2818–2822 页。[19] H. Wang、S. Kurita、S. Shimizu 和 D. Kawahara，“SlideAVSR：用于视听语音识别的论文讲解视频数据集”，*CoRR*，第 abs/2401.09759 卷，2024 年。[20] C. Luo、Y. Liu、W. Sun 和 Z. Sun，“由背景视觉场景驱动的多模态语音识别”，载于 *Proc.ICASSP*，2024 年，第 10926–10930 页。[21] P. H. Seo、A. Nagrani 和 C. Schmid，“Avformer：将视觉信息注入冻结的语音模型以实现零样本视听自动语音识别”，载于 *Proc.CVPR*，2023 年，第 22922–22931 页。[22] J. S. Chung、A. W. Senior、O. Vinyals 和 A. Zisserman，“野外场景中的句子唇读”，载于 *Proc.CVPR*，2017 年，第 3444–3453 页。[23] Y. Zhao、R. Xu 和 M. Song，“用于中文普通话唇读的级联序列到序列模型”，载于 *Proc.MMAsia*，2019 年，第 32:1–32:6 页。[24] C. Chen、D. Wang 和 T. F. Zheng，“CN-CVS：用于大词汇量连续视觉语音合成的普通话音视频数据集”，载于 *Proc.ICASSP*，2023 年，第 1–5 页。[25] J. S. Chung、A. Nagrani 和 A. Zisserman，“Voxceleb2：深度说话人识别”，载于 *Proc.Interspeech*，2018 年，第 1086–1090 页。[26] A. Radford、J. W. Kim、T. Xu、G. Brockman、C. McLeavey 和 I. Sutskever，“通过大规模弱监督实现鲁棒语音识别”，载于 *Proc.ICML*，机器学习研究论文集系列，第 202 卷，2023 年，第 28492–28518 页。[27] S. Bai、K. Chen、X. Liu、J. Wang、W. Ge、S. Song、K. Dang、P. Wang、S. Wang、J. Tang、H. Zhong、Y. Zhu、M. Yang、Z. Li、J. Wan、P. Wang、W. Ding、Z. Fu、Y. Xu、J. Ye、X. Zhang、T. Xie、Z. Cheng、H. Zhang、Z. Yang、H. Xu 和 J. Lin，“Qwen2.5-VL 技术报告”，*CoRR*，第 abs/2502.13923 卷，2025 年。[28] R. Liu、X. Xing、Z. Lian、H. Li、B. W. Schuller 和 H. Zuo，“MEIJU——首届多模态情感与意图联合理解挑战赛”，载于 *Proc.ICASSP*，2025 年，第 1–2 页。[29] Z. Lian、R. Liu、K. Xu、B. Liu、X. Liu、Y. Zhang、X. Liu、Y. Li、Z. Cheng、H. Zuo、Z. Ma、X. Peng、X. Chen、Y. Li、E. Cambria、G. Zhao、B. W. Schuller 和 J. Tao，“MER 2025：当情感计算遇上大语言模型”，*CoRR*，第 abs/2504.19423 卷，2025 年。[30] Y. Liu、Z. Yuan、H. Mao、Z. Liang、W. Yang *et al.*，“让声学和视觉线索发挥作用：CH-SIMS v2.0 数据集与视听混合一致性模块”，载于 *Proc.ICMI*，2022 年，第 247–258 页。