从炒作到洞察：重新思考大型语言模型在视觉语音识别中的集成

Rishabh Jain, Naomi Harte

Sigmedia Group, School of Engineering
Trinity College Dublin, Ireland
{rijain, nharte}@tcd.ie

ABSTRACT

自我监督编码器的进步提高了视觉语音识别（VSR）的性能。最近将这些编码器与大型语言模型解码器集成的方法改进了转录准确性；然而，尚不清楚这些改进是源自视觉理解还是更强大的语言建模能力。在本研究中，我们系统地评估了解码器，通过冻结或选择性更新视觉编码器、扩展解码器规模、比较适应策略和架构以及跨 LRS2、LRS3 及其组合的数据集变化来进行训练。在 LRS2、LRS3 和 WildVSR 上的评估显示，扩展和适应带来了有限的改进，而合并数据集增强了泛化能力。语义分析表明，这些增益主要源自词汇处理而非语义处理。我们基于结合数据集训练的 Llama-2-13B 模型在 LRS3 上实现了 24.7%的 WER，在 WildVSR 上实现 47.0%，确立了无额外监督训练模型中的最先进水平。我们的研究结果表明，大型语言模型解码器改进了上下文推理而非视觉特征提取，强调了需要更强的视觉编码器来推动有意义的进步。

Index Terms— 视觉语音识别，大型语言模型，AV-HuBERT，LRS 数据集，Llama

1. 介绍

视觉语音识别（VSR），或唇读，与音频-视频语音识别（AVSR）密切相关，这两项任务采用了类似的底层架构，仅输入模态不同。这些架构从早期的 CNN-RNN 模型 [1] 发展到目前基于序列到序列（S2S）变换器架构的 [2]。自我监督学习（SSL）方法 [3,4] 的引入是一个重要的转折点，特别是在 AV-HuBERT [3] 中，该模型利用大量未标记的音频-视频数据来学习视觉唇部运动与相应音频信号之间的联合表示。在此基础上，研究人员开发了包括 RAVEn [4] 和 BRAVEn [5] 在内的变体，探索了不同的自我监督学习范式和架构修改。这些方法表明，大量的未标记预训练可以建立强大的性能基线，为后续以解码器为中心的改进奠定了基础。半监督方法通过模型如 Auto-AVSR [6] 获得了显著的地位，该模型展示了从预先训练好的 ASR 模型中生成伪标签可以在数千小时的未标记数据上实现有竞争力的表现。这种方法突显了超越传统监督学习范式的数据扩展策略的潜在价值。

最近的研究越来越多地关注将预训练的视觉编码器与大型语言模型（LLMs）解码器 [7,8] 集成，以利用 LLM 的语言知识解决纯粹的视觉处理无法解决的歧义问题 [9]。框架如 VSP-LLM [10] 和 Llama-AVSR [11] 反映了这一趋势，通过轻量级投影层和参数高效的微调（PeFT）方法，例如低秩自适应（LoRA）[12] 或量化 LoRA（Q-LoRA）[13]，将预训练的视觉编码器与冻结的 LLM 连接起来。这些方法旨在将视觉语音信息传递给 LLM，同时受益于其现有的语言上下文。尽管单词错误率（WER）的改进一直被报道，但在 LRS3 [14] 上训练的模型表现出性能紧密聚集的现象（表 1）。这表明观察到的收益可能主要由更强的语言建模驱动，而不是更有效地利用视觉模式来提取口唇运动特征。尚不清楚通过投影层将视觉特征映射到 LLM 是否会产生新的视觉表示。

本文探讨了改进是源自编码器中的新视觉特征、投影层学习还是解码器设置的其他方面。为此，我们在固定视觉编码器的基础上进行了实验，系统地改变了解码器架构、模型规模、适应方法以及训练数据策

略，在领域内和跨领域的基准测试上进行了测试。这种设置隔离了解码器的作用，并阐明了大语言模型整合如何影响视觉语音表示的使用。我们的研究发现对未来的 VSR 研究工作具有重要意义。

2. 当前的 VSR 方法

通过关注在类似监督设置下训练的系统，我们旨在评估近期的视频超分辨率(VSR)进展是否真正反映了架构上的改进，而不是来自额外监督或适应方法的好处。为此，我们分析了先前在 LRS3 [14] 和 WildVSR [15] 数据集上的结果，考虑使用至少 433 小时的 LRS3 训练的模型以确保公平比较。我们排除了采用附加监督技术（如自训练、伪标签或语言模型重新评分）的具体结果。除了 BRAVEn [5](30h) 展示数据扩展在第 4.5 节外，低资源微调模型也被排除。正如表 1 所总结的，在 LRS3 上的词错误率（WER）紧密集中在 24% 到 28%之间，表明解码器设计的变化或轻微训练差异影响有限。WildVSR 数据集也显示出类似的趋势。较大的改进主要出现在预训练中或当标记数据显著增加时，这表明当前的视觉编码器可能接近其性能极限。将 AV-HuBERT 编码器 [3] 与 LLM 解码器（如 VSP-LLM [10] 和 Llama-AVSR [11]）或 Whisper 解码器 [16] 配对的模型仅显示出比 S2S 基准线略有改进。

3. 实验设置

VSR 中的一个关键问题是哪些因素驱动性能：解码器大小（1B 到 13B）、调整策略（如 LoRA 和 QLoRA）、LLM 架构，还是训练数据组成。为了分离这些影响，我们进行了控制实验来评估它们对识别和推理的影响。第 4 节呈现了结果，而本节概述了数据集、指标和方法。

3.1. 数据集和评估指标

我们使用三个标准基准来训练和评估我们的模型，这些基准共同涵盖了广泛的 VSR 场景。LRS2 [19] 包含 144,482 个片段，其中培训时间为 224 小时，验证和测试时间各自少于 1 小时。LRS3 [14] 由来自 TED 和 TEDx 演讲的 151,819 个片段组成，总共有 433 小时用于训练，0.9 小时用于测试。**野外超分辨率** [15] 是一个利用 LRS3 管道构建的开放领域测试集，但数据来源是未受限制的 YouTube 视频，这些视频在说话者、录音条件和视觉质量方面有更大的变异性，使其成为一个具有挑战性的测试集。所有输入视频被重新采样为 16 fps；嘴部区域通过 RetinaFace [20] 检测，裁剪成 96×96 灰度补丁，并在训练期间使用随机裁剪进行增强。

Table 1. 之前在 LRS3（L3）和 WildVSR（WV）数据集上报告的超分辨率（VSR）结果。

Model	PT	FT	Decoder	L3 ↓	WV ↓
AV-HuBERT [3]	1759	433	S2S	28.6	51.7
AVH+Whisper [16]	1759	433	Whisper	24.3	-
Auto-AVSR [6]	–	818	S2S	33.0	-
Auto-AVSR [6]	–	3448	S2S	19.1	38.6
RAVEn [4]	1759	433	S2S	27.8	52.2
BRAVEn [5]	1729	433	S2S	26.6	-
BRAVEn [5]	3052	30	S2S	24.8	-
BRAVEn [5]	2649	433	S2S	23.6	-
VSP-LLM [10]	1759	433	L-7B	26.7	55.1
VSP-LLM (E) [10]	1759	433	L-7B	25.4	51.6
Llama-AVSR [11]	1759	433	L-8B	26.9	-
Llama-AVSR (E) [11]	1759	433	L-8B	25.3	-
Llama-AVSR (E) [11]	1759	1756	L-8B	24.0	-

↓：越低越好；AVH= AV-HuBERT；PT= 预训练小时数（未标记）；**金融技术**= 微调小时数（已标记）；E 表示更新了 AV-HuBERT 编码器而不是冻结或使用 LoRA；**序列到序列**= Transformer 序列到序列解码器；L-7B= Llama2-7B [17]；L-8B= Llama3.1-8B [18]。

我们的主要评估指标是 WER [21]；然而，对于某些实验，我们还报告字符错误率（CER）[21] 和语义度量，如 sWER [22]、语义相似性 [23]、BERTScore [24] 和 METEOR [25]，详见第 4.4 节。

3.2. 方法论

我们利用两种不同的模型来研究解码器设计在 VSR 中的影响。第一种使用标准的 AV-**胡伯特** [3] 视觉编码器与 S2S 变压器解码器，作为强大的基线，在 LRS3 上实现了 28.6%的 WER（见表 1）。AV-HuBERT 是一个自监督框架，通过混合 ResNet-Transformer 架构学习视听语音表示。第二种模型基于 VSP-LLM [10] 框架，该框架将视觉语音处理与大语言模型结合以增

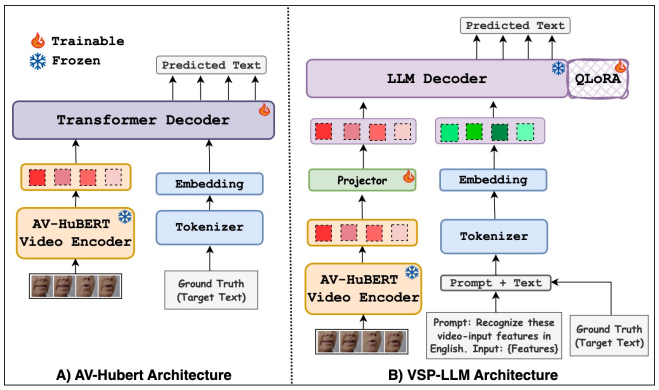

Fig. 1. 基线 AV-HuBERT 模型与本工作中使用的修改后的 VSP-LLM 架构的比较。(A) AV-HuBERT [3] (B) VSP-LLM [10]

强上下文理解和多任务能力，在 LRS3 上实现了 26.7% 的 WER（见表 1）。VSP-LLM 通过一个视觉到文本投影层将视觉特征映射到大语言模型输入空间，并使用 4 位 QLoRA [13] 以实现高效低精度的大语言模型适应。为了确保公平比较，我们排除了原始 VSP-LLM 中使用的去重复步骤。两个模型均使用相同的预训练 AV-HuBERT 编码器（在 LRS3+VoxCeleb2 上进行过训练），确保性能差异源自解码器设计而非优化设置。图 1 描述了两个模型的架构。在大多数实验中，视觉编码器是冻结的，当我们在一定数量的更新后解冻它时，我们会明确报告这一点。

4. 结果与讨论

4.1. 从 1B 到 13B：解码器大小和数据量如何驱动 AV-HuBERT 和 VSP-LLM 的性能

表 2 显示，在 LRS2 数据集上，AV-HuBERT 和 VSP-LLM（8B）表现相同（WER 为 24.4%），表明在这种设置下解码器集成没有带来任何好处。扩展到 13B 在 LRS2 上增加了域内误差（WER 为 28.4%），这表明过度拟合，但在 LRS3 上提供了适度的改进，而 WildVSR 的性能仍低于基线。在对 LRS3 数据集进行微调时，VSP-LLM（13B）实现了 WER 25.7%，超过了 8B 模型和已发布的 AV-HuBERT* 基线，并且甚至推广到了具有挑战性的 WildVSR 测试集。在结合的 657 小时语料库上进行训练产生了最大的收益：

Table 2. 报告了 AV-HuBERT 和 VSP-LLM 在 LRS2、LRS3 和 WildVSR 上的 WER 比较。

Model	LRS2 ↓	LRS3 ↓	WildVSR ↓
Finetuned on LRS2 (224 h)			
AV-HuBERT	24.4	38.1	52.9
VSP-LLM (8B)	24.4	36.4	56.4
VSP-LLM (13B)	28.4	36.5	53.6
Finetuned on LRS3 (433 h)			
AV-HuBERT*	38.0	28.7	51.7
VSP-LLM (8B)	37.9	27.8	53.8
VSP-LLM (13B)	38.2	25.7	49.8
Finetuned on LRS2 + LRS3 (657 h)			
AV-HuBERT	26.6	28.3	47.9
VSP-LLM (1B)	23.9	26.1	48.5
VSP-LLM (13B)	23.1	24.7	47.0

对所有训练，编码器 AV-HuBERT 都是从 22K→45K 更新的，除了用 433 小时 LRS3 训练的基础模型 AV-HuBERT*（由 AV-HuBERT 作者提供 [3]）；所有 VSP-LLM 模型都是以 4 位精度使用 QLoRA 进行训练的；1B：Llama-3.2-1B；8B：Llama3.1-8B；13B：Llama-2-13B-hf；↓：越低越好。

即使是 1B 解码器也优于 AV-HuBERT，而 13B 模型以 LRS2 上的 WER 为 23.1%，LRS3 上为 24.7%，WildVSR 上为 47.0% 设定了新的 SOTA。这些结果突显了两个关键见解：(i) 扩展解码器容量仅在与丰富且多样的数据配对时才能带来有意义的改进，以及 (ii) LLM 解码器主要增强了上下文推理而非视觉表示。值得注意的是，我们在结合数据集上训练的 13B 模型实现了迄今为止使用类似标注资源模型中报告的最低 WER，并排除了那些依赖额外监督如自训练、伪标签或 LM 重新评分的方法。

4.2. 从比特到大脑：量化和适应如何塑造解码

VSR-LLM(4 位 QLoRA，Llama-2-7B)和 Llama-AVSR（全精度 LoRA，Llama-3.1-8B）在适应性和精度上有所不同，但两者报告的 WER 几乎相同：26.7 对比 26.9（表 1）。这引发了对这些解码器方面的适应实际上如何影响性能的问题。为了进一步研究这个问题，我们设计了一组量化实验，使用 Llama-3.2-1B 模型。该模型较小，适合我们的 GPU 限制，并支持全精度训练、LoRA 和 QLoRA。如表 3 所述的这种设置允许对适应性和精度策略进行有控制的比较。所有模型都在 LRS3 上最多进行了 30K 次更新的微调，以确保配

Table 3. 报告了在 LRS2、LRS3 和 WildVSR（WV）上使用 VSP-LLM 进行量化实验的 WER。

Adaptation Approach	LRS2 ↓	LRS3 ↓	WV ↓
QLoRA (4-bit) [13]	39.80	28.19	51.60
LoRA (16-bit) [12]	39.35	28.59	51.93
Full Training/ No LoRA (16-bit)	47.33	37.31	65.43

AV-HuBERT **编码器** 的更新来自所有实验的 22K→30K ； 所有模型都是在 LRS3 数据集上使用 llama-3.2-1B 进行训练的；**位**的数量表示模型权重的数值**精度**;↓ 越低越好。

置的一致性。

结果表明，无论是 4 位 QLoRA 还是 16 位 LoRA 配置，在 LRS2、LRS3 和 WildVSR 上都实现了比全精度（无 LoRA 设置）更低的 WER。在全精度情况下，损失在训练结束时（30K 次更新后）仍在下降，这表明延长训练时间可能进一步提高性能。这些发现表明，量化水平和适应方法对准确性的影响力较小，编码器架构限制仍然是主要的制约因素。

4.3. 从 LLM 到 LLM：理解解码器架构如何影响性能

为了评估解码器设计的影响，我们将四个大小相似的 LLM（约 13B）[17,26–28] 进行比较，这些模型都在 LRS3 数据集上进行了相同的微调。尽管它们的核心管道和训练设置相似（详见表 4），但这些模型在解码器块、位置编码、注意力变体以及预训练数据方面有所不同。此分析旨在确定架构和数据驱动的差异是否转化为性能提升，或者不同设计下的性能保持相似。

性能差异在所有测试架构中（表 4）的 1.5%WER 范围内变化。Llama-2-13B [17] 表现最佳，略微优于 Phi-4 [26] 设计。Vicuna-13B [27] 和 Qwen2.5-14B [28] 略有不足，这可能与其优化对话和多语言任务而非通用解码有关。这些结果表明，在固定编码器的情况下，解码器中的架构变化和预训练数据对 VSR 准确性的影响很小。

4.4. 从指标到意义：理解解码器在词汇和语义维度上的限制

为了评估 LLM 解码器是否在行为上与 AV-HuBERT 有质的不同，我们使用多个互补指标进行性能评估，而不仅仅是 WER。虽然 WER 捕捉到了词级别的

Table 4. 大语言模型解码器架构和预训练数据对 LRS3（L3）上 VSR 性能的影响。

Model	Architecture & Data Details	L3 ↓
Llama-2-13b [17]	RoPE; MHA; SwiGLU; RMSNorm; 4k ctx; pretrained on 2T tokens of multilingual (web,code,dialogue); instruction-tuned	25.7
phi-4 [26]	MHA; GELU; LayerNorm; 16k ctx; English web and code; synthetic reasoning; SFT + DPO	26.2
vicuna-13b-v1.5 [27]	Extention of Llama-2 13B; RoPE; MHA; SwiGLU; RMSNorm; 4k ctx; multilingual pre-training; ShareGPT SFT (125k conversations)	26.5
Qwen2.5-14B [28]	RoPE + QKV bias; GQA; SwiGLU; RMSNorm; 131k ctx; trained on 40+ lang, web, code, and structured data	27.2

T：兆；**上下文**: 上下文窗口；MHA: 多头注意力；**软烤**: 监督微调；DPO: 直接偏好优化；GQA: 分组查询注意力；**旋转位置嵌入**: 旋转位置嵌入；QKV: 查询/键/值；模型训练直到 50K 更新；编码器从 22K→50K 更新;↓: 数值越低越好。

准确性，CER [21] 强调了字符级别的错误。语义-WER [22] 仅计算那些改变含义的替换，并且语义相似度 [23] 测量嵌入空间中句子级别的接近程度。BERTScore [24] 对齐上下文化嵌入以捕捉语义等价性，而 METEOR [25] 考虑了精度、召回率、语言变化（同义词、词干提取、释义）和词序。AV-HuBERT 和 VSP-LLM 都在 LRS3 上进行了 30K 次更新的微调，编码器被冻结，确保差异反映了解码器的贡献而非视觉表示的变化。

如表 5 所示，在 LRS2 和 LRS3 测试集上评估时，两个模型在所有指标上的性能都紧密匹配，差异通常在 1 个百分点之内。VSP-LLM 偶尔会实现稍低的 WER 或更高的语义分数，但改进幅度小且不一致，这表明 LLM 解码器并未显著改变识别行为。这突出显示了视觉贡献来自于编码器，而解码器的变化仅带来微小的语言收益。

4.5. 从瓶颈到突破：重新思考视觉前端

VSR 性能越来越受到当前视觉编码器限制的约束。大多数 SOTA 系统依赖于 AV-HuBERT 或其变体 [10, 11, 29]，在 LRS3 上导致 WER 紧密聚类，并揭示了无论解码器架构还是训练策略如何，视觉特征提取都存在一个共同的上限。BRAVEn [5] 的扩展实验（表 1）显示，当总训练时长保持不变（预训练+微

Table 5. 扩展评估指标对比 AV-HuBERT 与 VSP-LLM 在 □ LRS2（L2）和 □ 长期记忆检索 3 (L3) 上的表现。

Model	CER↓	WER↓	sWER↓	SS↑	BS ↑	MET↑
AVH (L2)	24.7	38.0	31.2	0.64	0.93	0.63
VLM (L2)	26.2	38.3	31.8	0.64	0.93	0.61
AVH (L3)	18.7	28.7	22.7	0.71	0.94	0.70
VLM (L3)	19.4	28.5	22.7	0.72	0.95	0.71

AVH = AV-HuBERT; **多模态学习** = VSP-LLM; **置信错误率** = 字符错误率; WER = 单词错误率; sWER = 语义 WER; **样本大小** = 语义相似度; BS = BERTScore; **金属** = METEOR. 所有 VLM 使用 llama-2-7b 解码器;↓越低越好;↑越高越好。

调）时，将标注数据从 30 小时增加到 433 小时，只带来适度的 WER 改进（24.8%到 23.6%），突显了编码器作为主要瓶颈的作用。我们的研究揭示了这些约束条件，指出了推进 VSR 领域的有前途的方向。一个前景广阔的方法是开发大规模自监督视觉编码器，在成千上万小时未标注视频上进行训练，以学习更丰富的唇部运动表示，并克服当前的性能上限。向 Vision Transformers（ViTs）[30, 31] 的转变也很有希望，因为它们比 CNN 更好地捕捉全局时空模式，并可能解决在视-音映射 [32] 中的关键限制。此外，Auto-AVSR 的 [6] 编码器，在 3,448 小时伪标注数据上训练时，可以作为特征提取器使用，当与基于 LLM 的解码器配对时，提供监督下的视觉表示。这些发现具有重要意义，表明在解码器优化方面的进一步投资可能会因性能主要受到编码器架构限制而产生递减回报。作为未来工作的一部分，我们旨在探索其中一些方向以开发更有效的 VSR 系统。

5. 结论

我们对 LLM 在 VSR 中的整合分析显示，通过投影层将视觉特征映射到 LLMs 中仅带来了微小的 WER 改进，这主要得益于预训练的语言知识而非学习新的视觉表示。尽管这些改进可以测量，但它们并不代表近期文献中经常提到的重大突破。近乎普遍依赖于 AV-HuBERT 编码器已经塑造了一个研究领域，在这个领域中，尽管解码器设计或训练方法有所变化，核心的视觉表征学习仍然基本没有改变。我们的发现表明，VSR 中的实质性进展需要在视觉编码器架构上取得进步，仅专注于解码器的战略对于改善 VSR 性能潜力有限。

6. 致谢

本出版物源自由 Taighde Éireann – Research Ireland 资助的研究，资助编号为 22/FFP-A/11059。

7. REFERENCES

[1] I. Fung et al., "End-to-end low-resource lip-reading with maxout cnn and lstm," in ICASSP, 2018, pp. 2511–2515.

[2] O. Chang et al., "Conformer is all you need for visual speech recognition," in ICASSP, 2024, pp. 10136–10140.

[3] B. Shi et al., "Learning audio-visual speech representation by masked multimodal cluster prediction," in ICLR, 2022.

[4] A. Haliassos et al., "Jointly learning visual and auditory speech representations from raw data," ICLR, 2023.

[5] A. Haliassos et al., "Braven: Improving self-supervised pre-training for visual and auditory speech recognition," in ICASSP, 2024, pp. 11431–11435.

[6] P. Ma et al., "Auto-avsr: Audio-visual speech recognition with automatic labels," in ICASSP, 2023, pp. 1–5.

[7] J. Yeo et al., "MMS-LLaMA: Efficient LLM-based audio-visual speech recognition with minimal multimodal speech tokens," in ACL Findings, 2025, pp. 20724–20735.

[8] D. Zhang et al., "Mm-llms: Recent advances in multimodal large language models," in ACL Findings, 2024.

[9] W. Wang et al., "Visionllm: Large language model is also an open-ended decoder for vision-centric tasks," in NeurIPS, 2023, vol. 36, pp. 61501–61513.

[10] J. Yeo et al., "Where visual speech meets language: VSP-LLM framework for efficient and context-aware visual speech processing," in EMNLP, 2024, pp. 11391–11406.

[11] U. Cappellazzo et al., "Large language models are strong audio-visual speech recognition learners," in ICASSP, 2025, pp. 1–5.

[12] E. J Hu et al., "Lora: Low-rank adaptation of large language models.," ICLR, 2022.

[13] T. Dettmers et al., "Qlora: Efficient finetuning of quantized llms," NeurIPS, 2023.

[14] T. Afouras et al., "Lrs3-ted: a large-scale dataset for visual speech recognition," in arXiv preprint arXiv:1809.00496, 2018.

[15] Y. Djilali et al., "Do vsr models generalize beyond lrs3?," in WACV, 2024, pp. 6635–6644.

[16] K R Prajwal et al., "Speech Recognition Models are Strong Lip-readers," in Interspeech, 2024, pp. 2425–2429.

[17] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," in in arXiv preprint arXiv:2307.09288, 2023.

[18] A. Grattafiori et al., "The llama 3 herd of models," in in arXiv preprint arXiv:2407.21783, 2024.

[19] T. Afouras et al., "Deep audio-visual speech recognition," in TPAMI, 2022, p. 13 – 18.

[20] J. Deng et al., "Retinaface: Single-shot multi-level face localisation in the wild," in CVPR, 2020, pp. 5202–5211.

[21] Thennal D K et al., "Advocating character error rate for multilingual ASR evaluation," in NAACL, 2025, pp. 4926–4935.

[22] C. Spiccia et al., "Semantic word error rate for sentence similarity," in ICSC, 2016, pp. 266–269.

[23] V. Mayil et al., "Pretrained sentence embedding and semantic sentence similarity language model for text classification in nlp," in AISP, 2023, pp. 1–5.

[24] T. Zhang et al., "Bertscore: Evaluating text generation with bert," in in arXiv preprint arXiv:1904.09675, 2020.

[25] S. Banerjee et al., "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in ACL Workshop, 2005.

[26] M. Abdin et al., "Phi-4 technical report," in in arXiv preprint arXiv:2412.08905, 2024.

[27] W. Chiang et al., "Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality," 2023.

[28] A. Yang et al., "Qwen2.5 technical report," in in arXiv preprint arXiv:2412.15115, 2025.

[29] A. Rouditchenko et al., "Whisper-Flamingo: Integrating Visual Features into Whisper for Audio-Visual Speech Recognition and Translation," in Interspeech 2024, 2024, pp. 2420–2424.

[30] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," ICLR, 2021.

[31] A. Arnab et al., "Vivit: A video vision transformer," in ICCV, 2021, pp. 6836–6846.

[32] M. Thomas et al., "Vallr: Visual asr language model for lip reading," in arXiv preprint arXiv:2503.21408, 2025.