迈向实时跨语言沟通：实时语音模型的挑战、技术和未来

张绍磊

中国科学院 计算技术研究所

2025-01-16

中国科学院计算技术研究所
Institute of Computing Technology,Chinese Academy of Sciences

传统跨语言沟通

- **传统交互范式：离线交互（一问一答、完整语音）**
- **关键技术：**
  - **机器翻译：**源语言文本→目标语言文本
  - **语音翻译：**源语言语音→目标语言文本（语音）
  - 语音识别、语音合成…

需要等待完整的语音输入

机器翻译

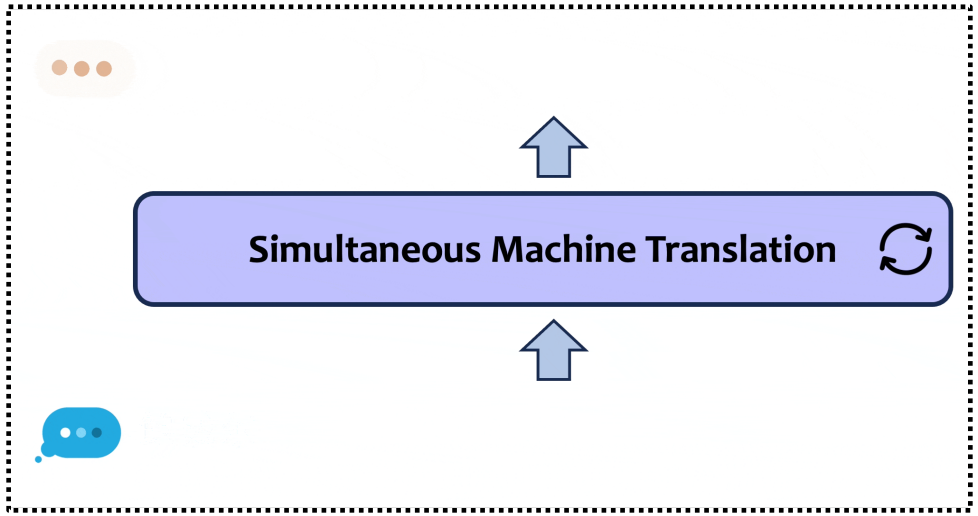
语音翻译

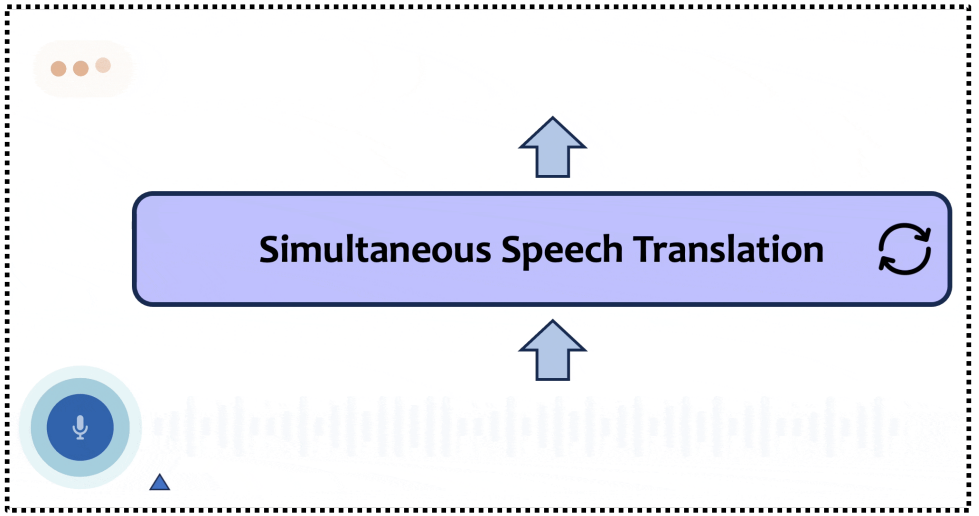
语音识别、合成等技术

实时跨语言沟通

- 实时交互：在接收输入的同时，模型实时生成回复
- 关键技术：
  - 实时机器翻译：源语言文本→目标语言文本
  - 实时语音翻译：源语言语音→目标语言文本（语音）
- 广泛应用于实时场景：国际会议、在线直播、同声传译…

低延时+高生成质量

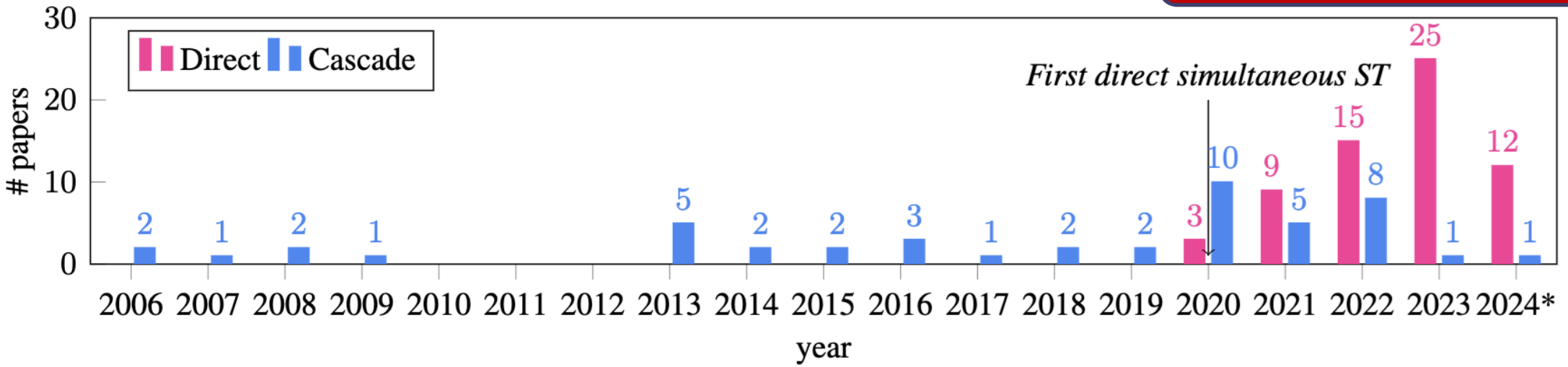




实时语音翻译发展历程

- 逐渐成为机器翻译、语音翻译领域前沿课题
- 级联系统 → 端到端系统
  - 级联：实时语音识别 + 实时机器翻译 + 实时语音合成
  - 端到端：实时语音到文本翻译、实时语音到语音翻译

实时翻译系统数量



图源：How "Real" is Your Real-Time Simultaneous Speech-to-Text Translation System?

实时语音模型研究进展

在说话者讲话的同时
流式生成
（识别/翻译/合成）

级联：实时文本翻译	
未来信息指导的知识蒸馏	AAAI 2021
基于MoE 模型的策略	EMNLP 2021
基于自监督学习的策略	ACL 2022
基于高斯对齐的策略	ACL 2022
基于信息量化的策略	EMNLP 2022
基于概率图模型的策略	ICLR 2023

如何在流式输入中确定生成时机？

端到端：实时语音翻译	
生成模型中的位置偏差	ACL 2022
跨模态信息运输	EMNLP 2022
无监督语音切分/对齐	ACL 2023

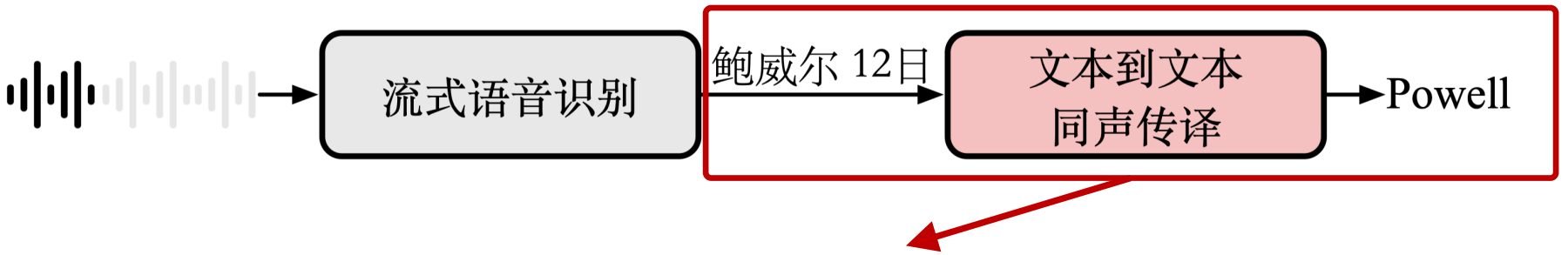
如何建模跨模态跨语言生成？

All in One 流式语音模型	
通用多模态片段对齐	NeurIPS 2023
一体化语音模型	ACL 2024
基于Agent的流式模型	arXiv 2024

如何构建通用模型
完成多种任务？

级联：实时文本翻译

- **级联式实时文本翻译：流式语音识别 → 文本到文本同声传译**



实时策略：决策何时开始生成(读/写)

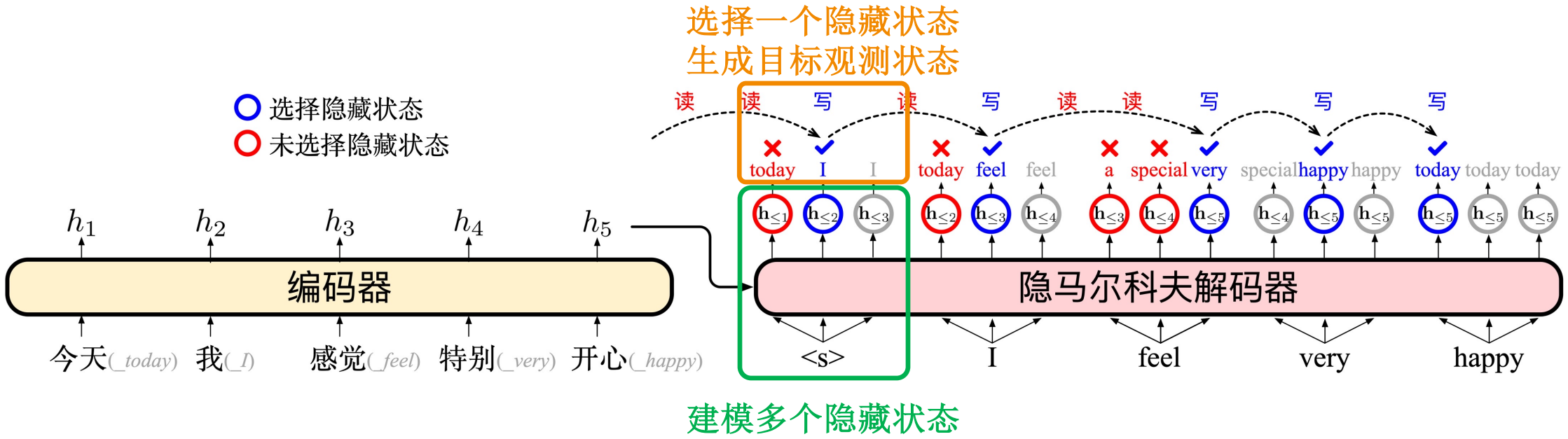
文本输入：	鲍威尔		12日	与	沙龙		举行	了	会谈		
实时策略：	READ	WRITE		READ		WRITE			READ		WRITE
实时文本翻译：	—等待→	Powell held		—等待→	talks with Sharon				—等待→	on 12	

权衡
生成质量和延时

实时策略（生成过程中的隐藏状态）缺乏显式监督

基于概率图模型的实时策略

- 利用概率图模型建模实时生成过程
  - 将策略建模为隐藏状态，生成词建模为观测状态
  - 在所有可能的隐藏状态上，优化观测状态的边际似然 → 学习最佳隐状态，即实时策略



Shaolei Zhang, Yang Feng. Hidden Markov Transformer for Simultaneous Machine Translation. ICLR 2023 Spotlight.

基于概率图模型的实时策略

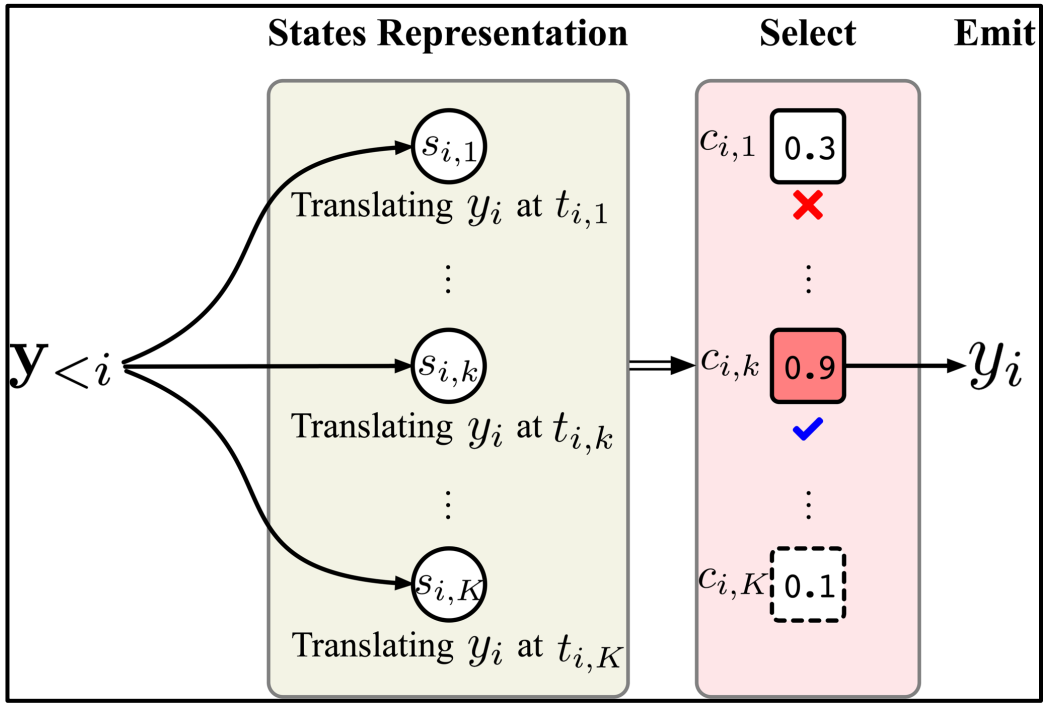
- 流式生成过程
  - 生成多个状态，对应不同翻译时刻
  - 根据状态的置信度选择一个状态
  - 从选择的状态发射目标词
- 训练
  - 优化目标序列在所有状态上的的边际似然：

$$p(\mathbf{y} \mid \mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) \times p(\mathbf{z})$$

$$\mathcal{L}_{hmm} = -\log p(\mathbf{y} \mid \mathbf{x})$$

  - 最小化期望延时：

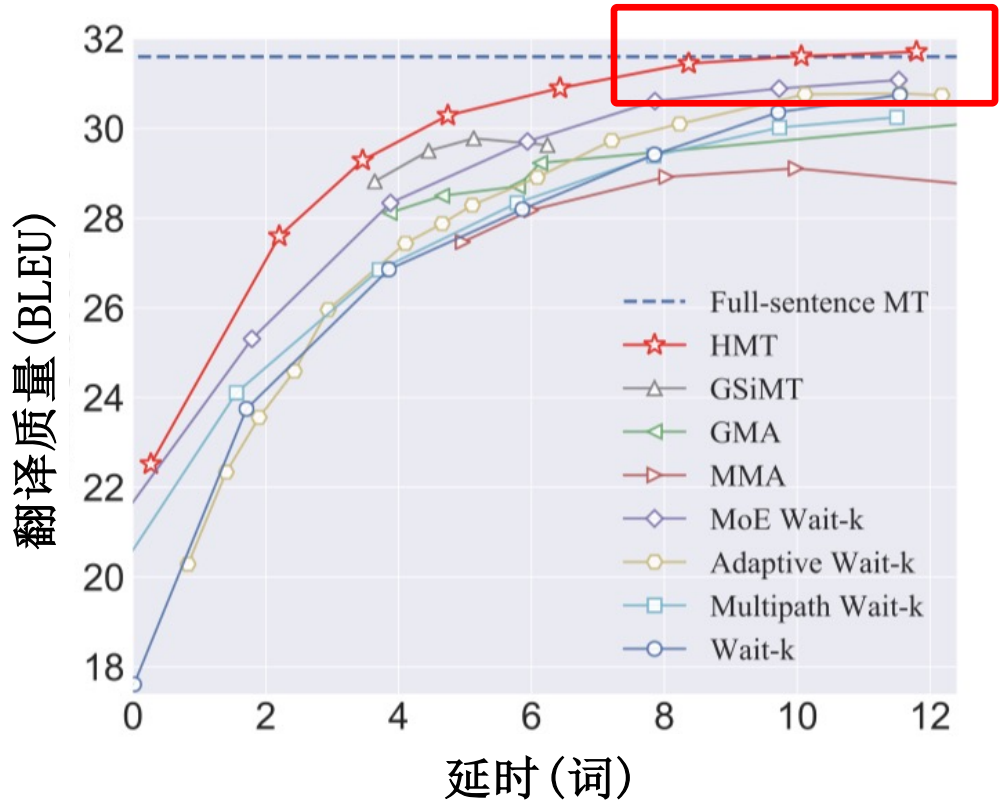
$$\mathcal{L}_{latency} = \sum_{\mathbf{z}} p(\mathbf{z}) \times \mathcal{C}(\mathbf{z})$$



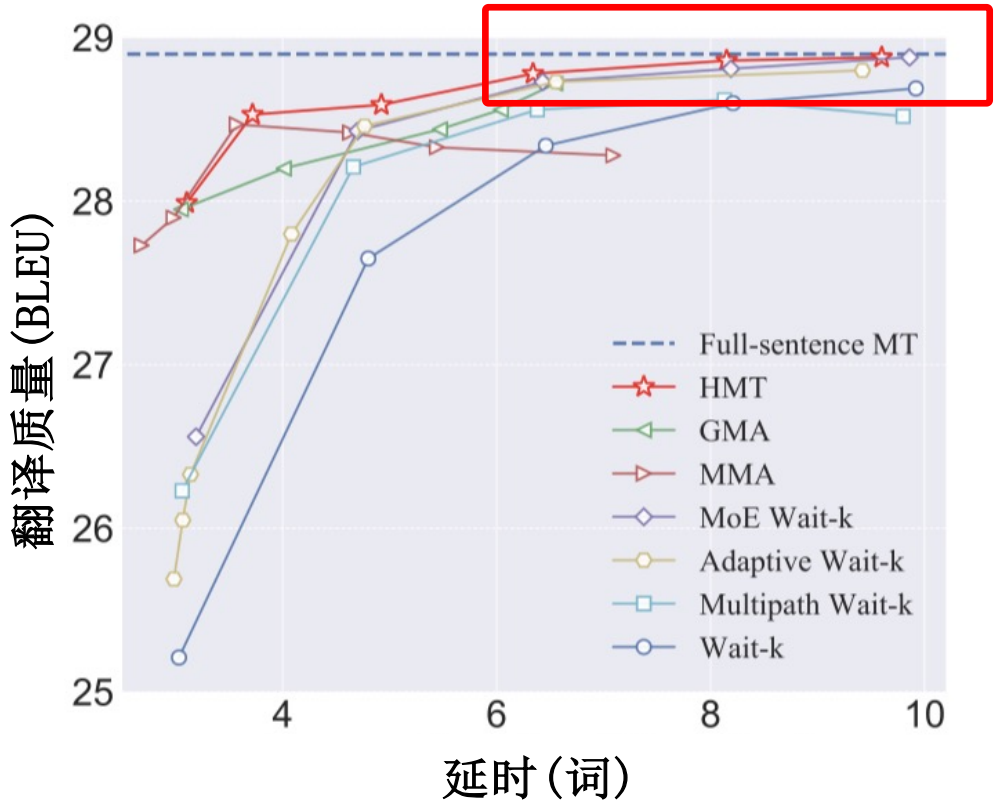
Shaolei Zhang, Yang Feng. Hidden Markov Transformer for Simultaneous Machine Translation. ICLR 2023 Spotlight.

级联式实时文本翻译性能

- 目前最佳的实时文本翻译方法：滞后8个词，媲美离线机器翻译



（a）德语→英语

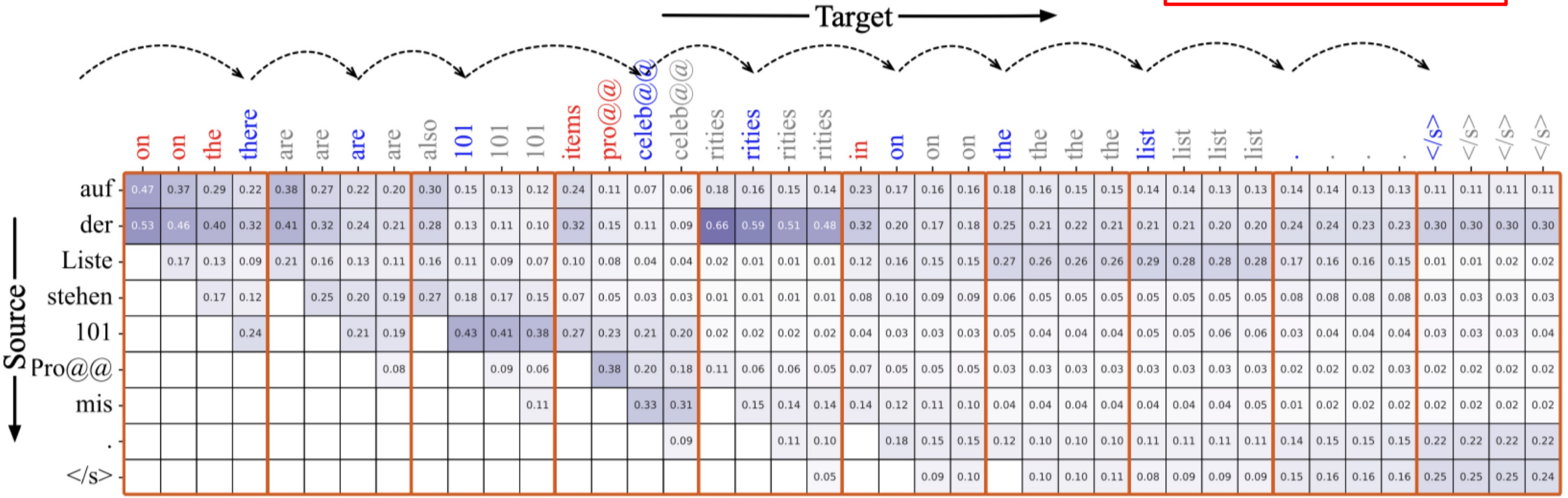


（b）英语→越南语

级联式实时文本翻译性能

- 在候选状态间跳转、选择、发射

红色：被选择状态
蓝色：未选择状态

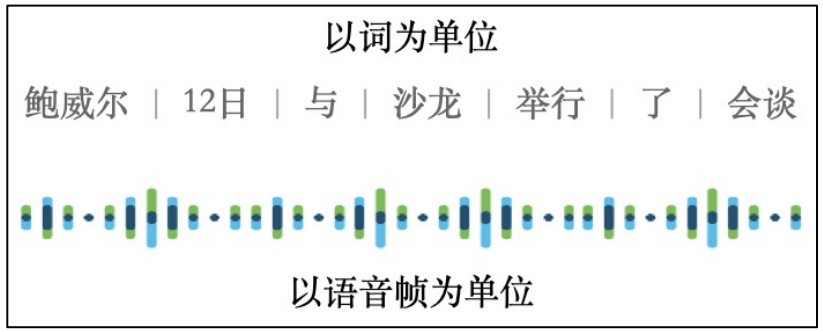


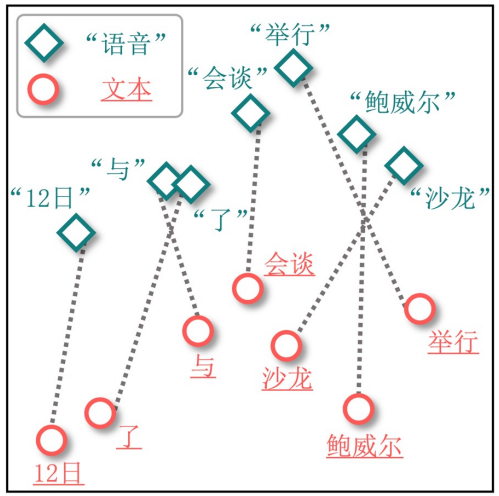
端到端：实时语音翻译

- **跨模态跨语言流式生成：**端到端流式语音翻译

语音输入：

流式语音翻译：—等待→ Powell held —等待→ talks with Sharon —等待→ on 12



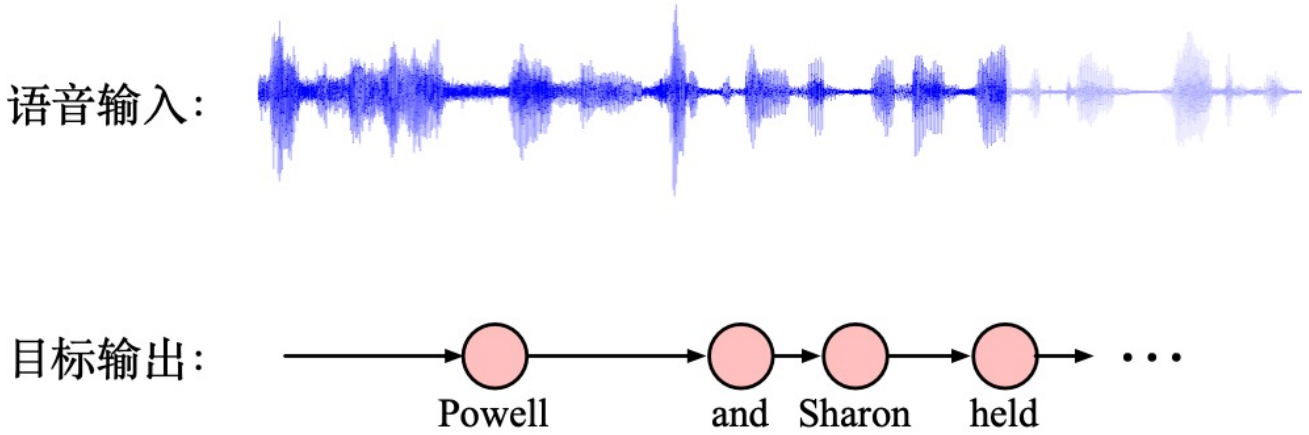


结构差异
源序列长度远大于目标序列

表示差异
语音表示和文本表示无法对齐

如何解决语音和文本模态差异?

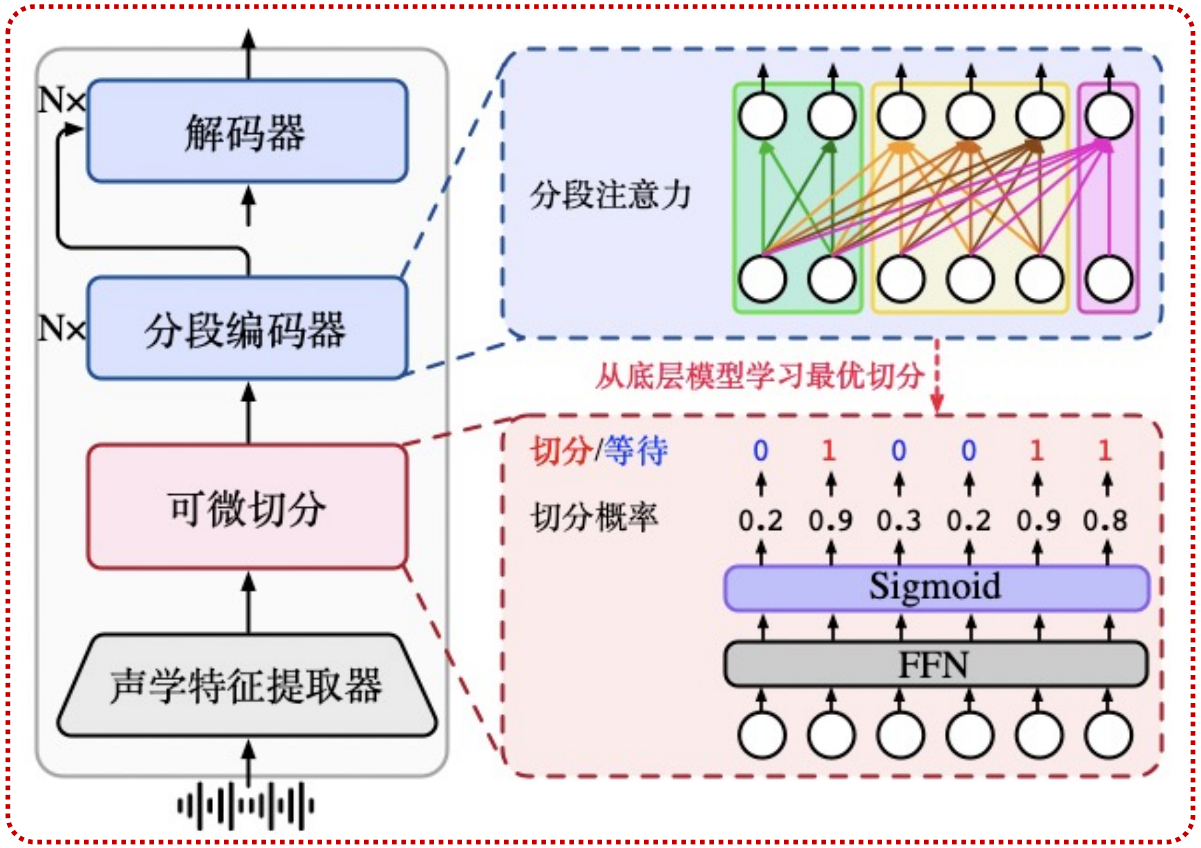
- 语音和文本模态存在模态鸿沟，翻译策略难以学习
  - 结构差异：语音序列是连续信号，序列长度长、缺少显式切分
  - 表示差异：语音模态和文本模态在表示空间的分布存在差异



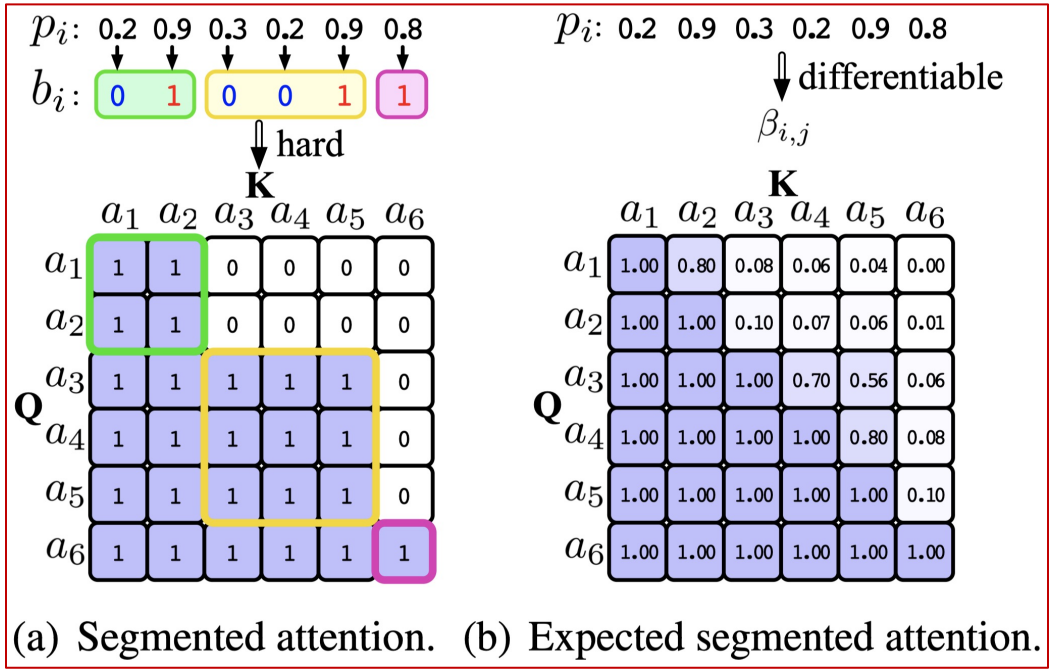
- 解决方案：
  - 1. 自适应语音分段：解决跨模态结构差异
  - 2. 建模词级别对齐：解决跨模态表示差异

基于可微语音分段的实时语音翻译

- 无监督地切分语音序列为语音片段



根据切分概率，计算期望语音表示
与底层模型联合训练

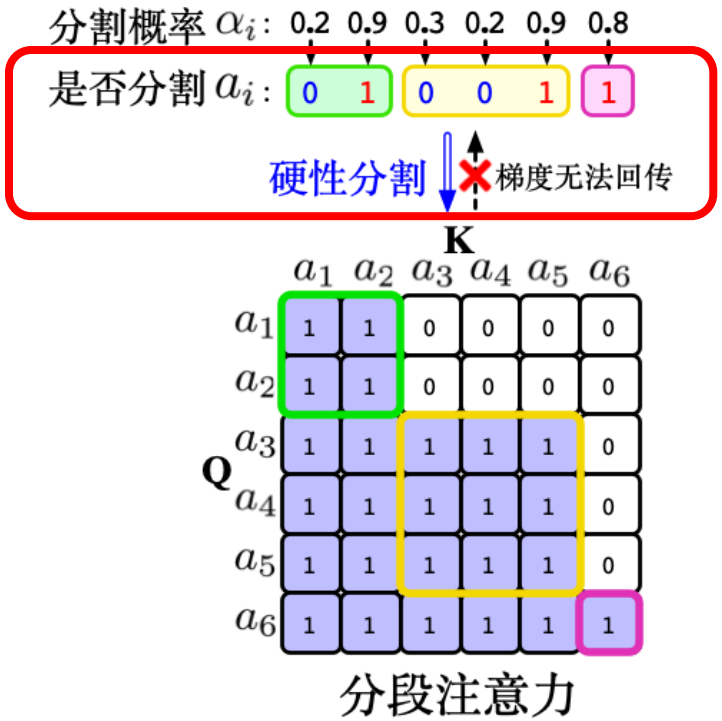


无监督语音分段

Shaolei Zhang, Yang Feng. End-to-End Simultaneous Speech Translation with Differentiable Segmentation. Findings of ACL 2023.

基于可微语音分段的实时语音翻译

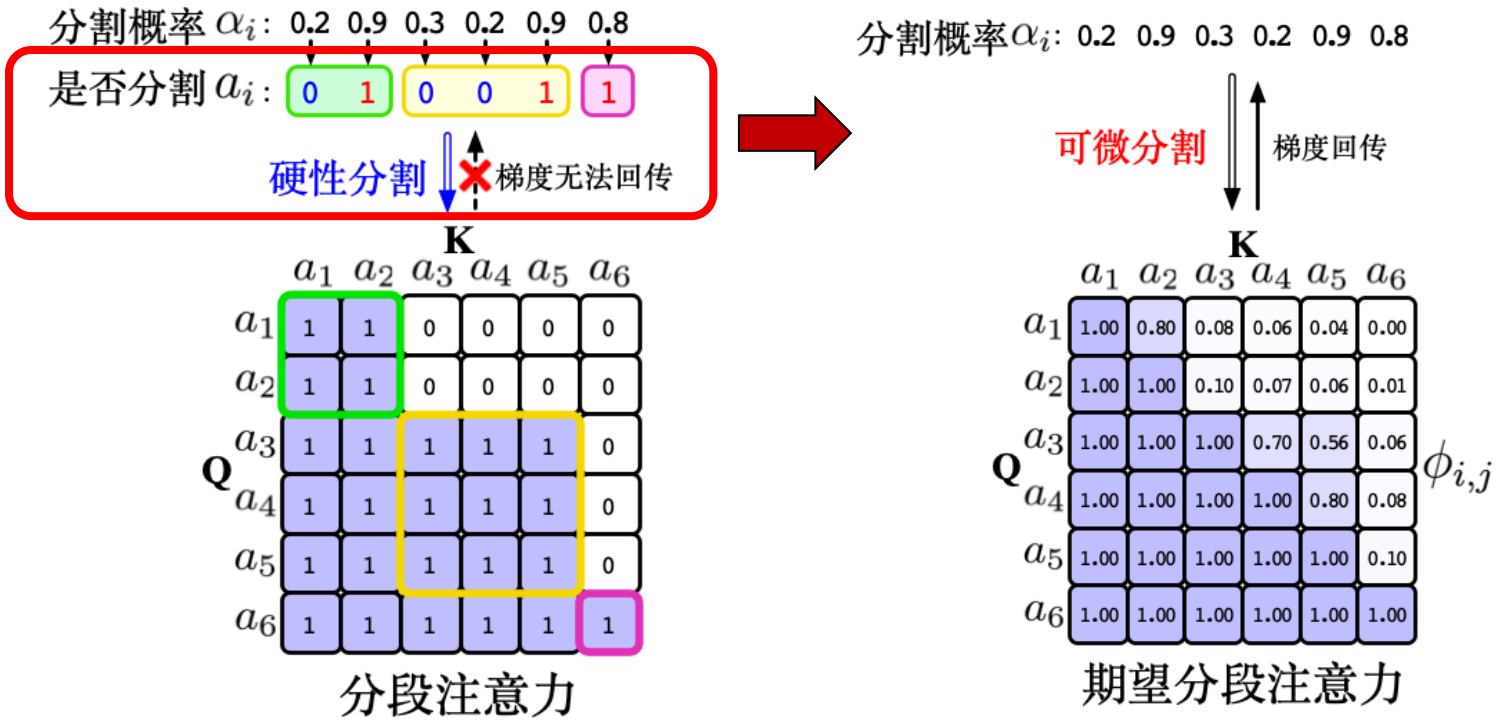
- **自适应语音分段**：将语音序列分割成若干完整的语音片段
  - 核心问题：用0/1硬性分割不可导 ⇒ 梯度无法回传，无法学习



Shaolei Zhang, Yang Feng. End-to-End Simultaneous Speech Translation with Differentiable Segmentation. Findings of ACL 2023.

基于可微语音分段的实时语音翻译

- 基于可微分段的自适应语音切分（Differentiable Segmentation）
  - 引入分割概率，以期望形式通过梯度回传实现无监督分段



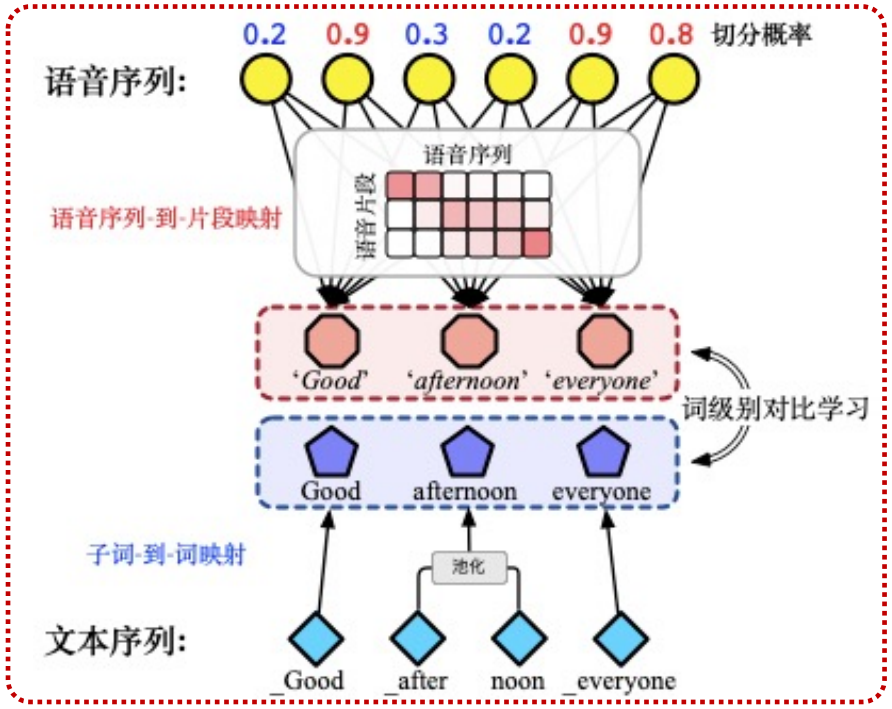
语音帧x_j处于x_j的同一片段或之前片段的概率

$$\phi_{i,j} = \begin{cases} \prod_{l=i}^{j-1}(1-\alpha_l), & \text{if } i < j \\ 1, & \text{if } i \geq j \end{cases}$$

Shaolei Zhang, Yang Feng. End-to-End Simultaneous Speech Translation with Differentiable Segmentation. Findings of ACL 2023.

基于可微语音分段的实时语音翻译

- 无监督地切分语音序列为语音片段
- 通过词级别对比学习进行跨模态对齐



文本表示：

$$f_k^t = \frac{1}{r_k - l_k + 1} \sum_{i=l_k}^{r_k} e_i$$

语音片段期望表示：

$$p(a_i \in \mathrm{Seg}_k) = p(a_{i-1} \in \mathrm{Seg}_{k-1}) \times p_{i-1} + p(a_{i-1} \in \mathrm{Seg}_k) \times (1 - p_{i-1}).$$

$$f_k^s = \sum_{i=1}^{|\mathbf{a}|} p(a_i \in \mathrm{Seg}_k) \times a_i.$$

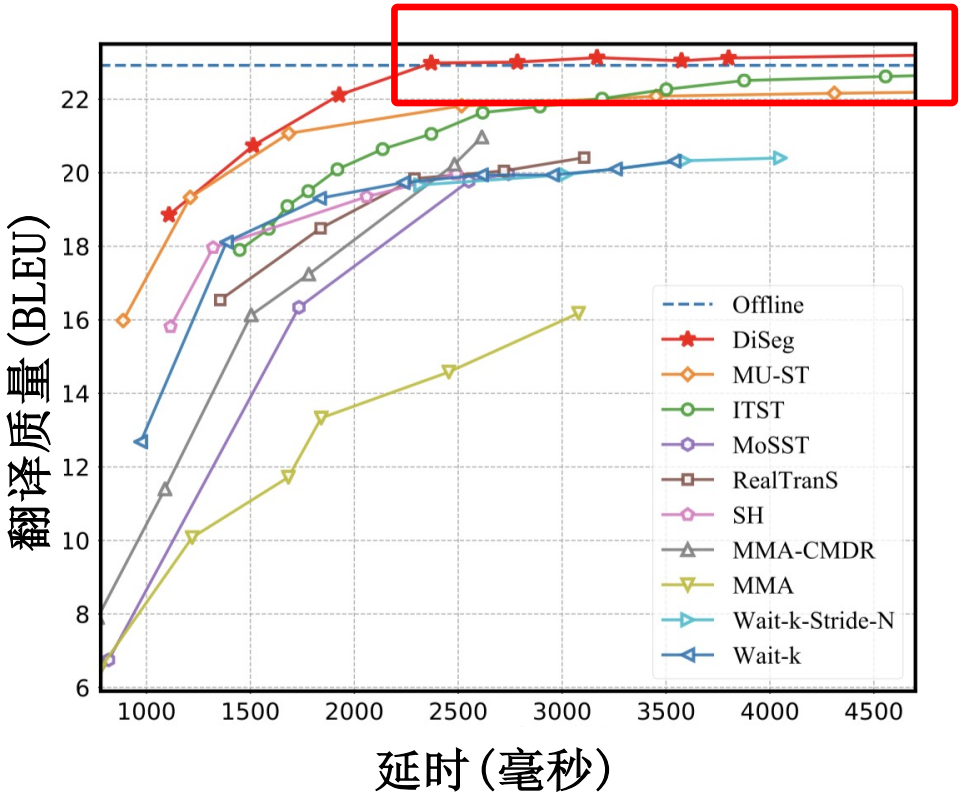
词级别对比学习：

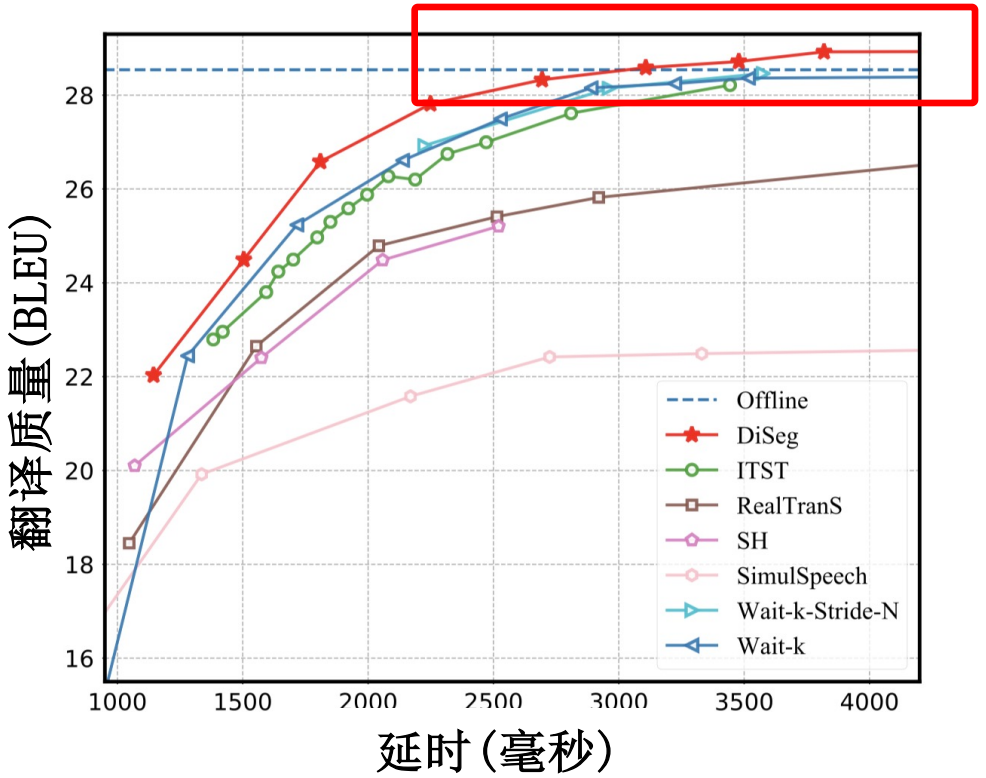
$$\mathcal{L}_{ctr} = -\sum_{\mathbf{f}^s, \mathbf{f}^t} \log \frac{\exp\left(sim\left(f_k^s, f_k^t\right)/\tau\right)}{\sum_{n=1}^{K} \exp\left(sim\left(f_k^s, f_n^t\right)/\tau\right)}.$$

Shaolei Zhang, Yang Feng. End-to-End Simultaneous Speech Translation with Differentiable Segmentation. Findings of ACL 2023.

端到端实时语音翻译性能

- 目前最佳的实时语音翻译方法：滞后2.5秒时，性能媲美离线语音翻译





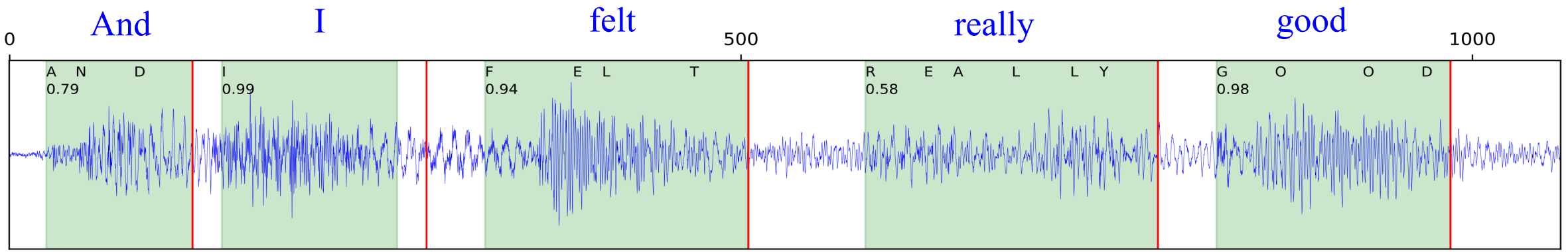
Shaolei Zhang, Yang Feng. End-to-End Simultaneous Speech Translation with Differentiable Segmentation. Findings of ACL 2023.

自适应语音分段准确率

- 语音分段准确率
  - 可微分割（DiSeg）取得较高分段质量
  - 保持语音完整性

表：Buckeye数据集上的分段准确率

Methods	P(↑)	R(↑)	F1(↑)	OS(0)	R-val(↑)
ES K-Means	30.7	18.0	22.7	-41.2	39.7
BES GMM	31.7	13.8	19.2	-56.6	37.9
VQ-CPC	18.2	54.1	27.3	196.4	-86.5
VQ-VAE	16.4	56.8	25.5	245.2	-126.5
SCPC	35.0	29.6	32.1	-15.4	44.5
DSegKNN	30.9	32.0	31.5	3.5	40.7
Fixed(280ms)	28.1	16.3	20.7	-42.0	38.4
DiSeg	34.9	32.3	**33.5**	-7.4	**44.6**



Shaolei Zhang, Yang Feng. End-to-End Simultaneous Speech Translation with Differentiable Segmentation. Findings of ACL 2023.

“All in One” 实时语音模型

- 级联式、端到端语音翻译往往完成单一任务
- 国际会议、AR眼镜等实时场景
  - 用户在看到翻译结果的同时，希望能听到翻译语音
  - 更全面的跨语言沟通体验

“All in One” 实时语音模型

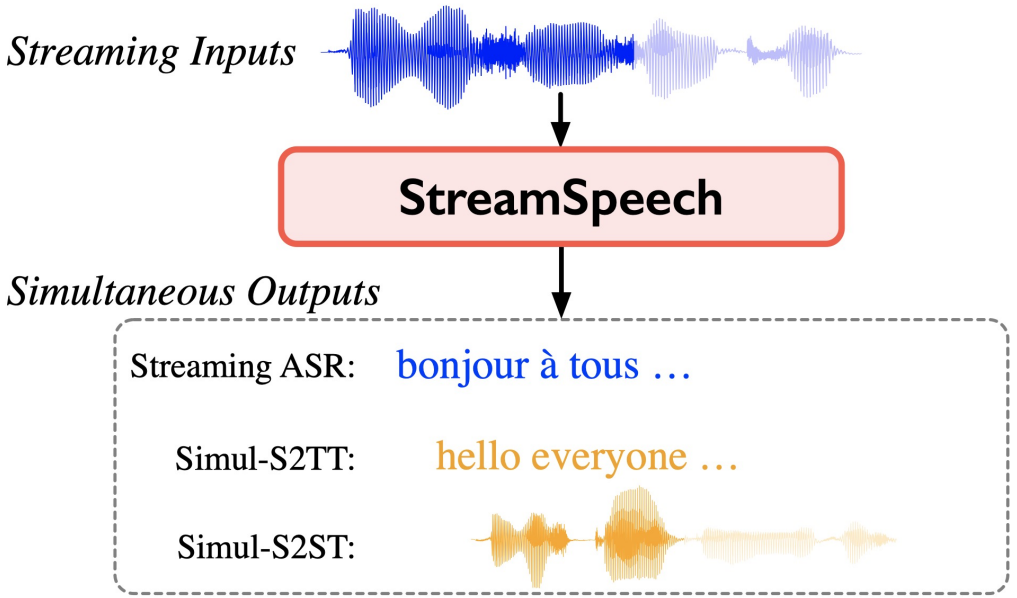
- 级联式、端到端语音翻译往往完成单一任务
- 国际会议、AR眼镜等实时场景
  - 用户在看到翻译结果的同时，希望能听到翻译语音
  - 更全面的跨语言沟通体验

如何利用端到端模型
做到边听边看？

“All in One” 实时语音模型

- 以端到端的无缝方式完成离线/任意延时下的语音识别、语音翻译、语音合成
- 仅需一个通用模型：StreamSpeech

支持8种语音任务
320毫秒延迟



“**All in One**” seamless model for offline and simultaneous ASR / Translation / Synthesis

Conference

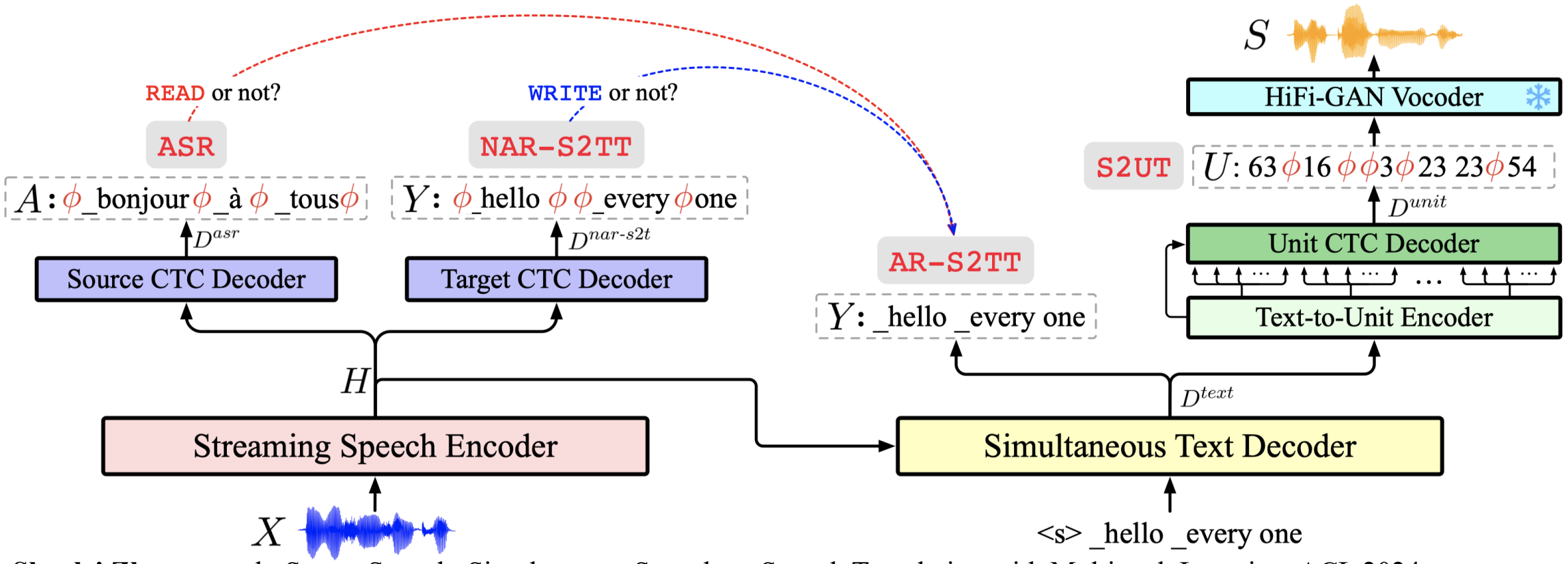
live broadcasts

AR Glasses

Shaolei Zhang, et, al. StreamSpeech: Simultaneous Speech-to-Speech Translation with Multi-task Learning. ACL 2024.

StreamSpeech

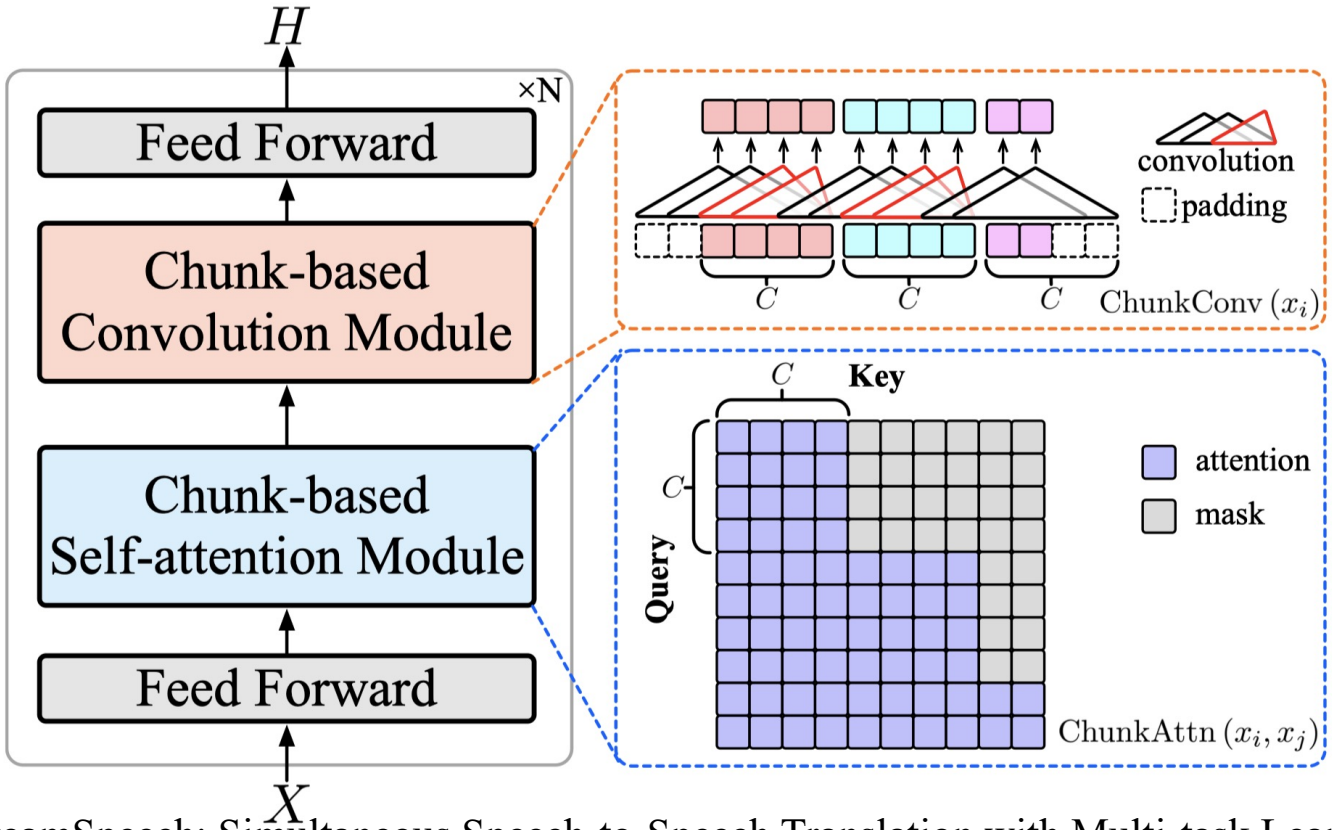
- 流式语音编码器 + 实时文本解码器 + 同步语音合成模块
- CTC Decoder 通过辅助任务学习序列间的对齐



Shaolei Zhang, et, al. StreamSpeech: Simultaneous Speech-to-Speech Translation with Multi-task Learning. ACL 2024.

StreamSpeech

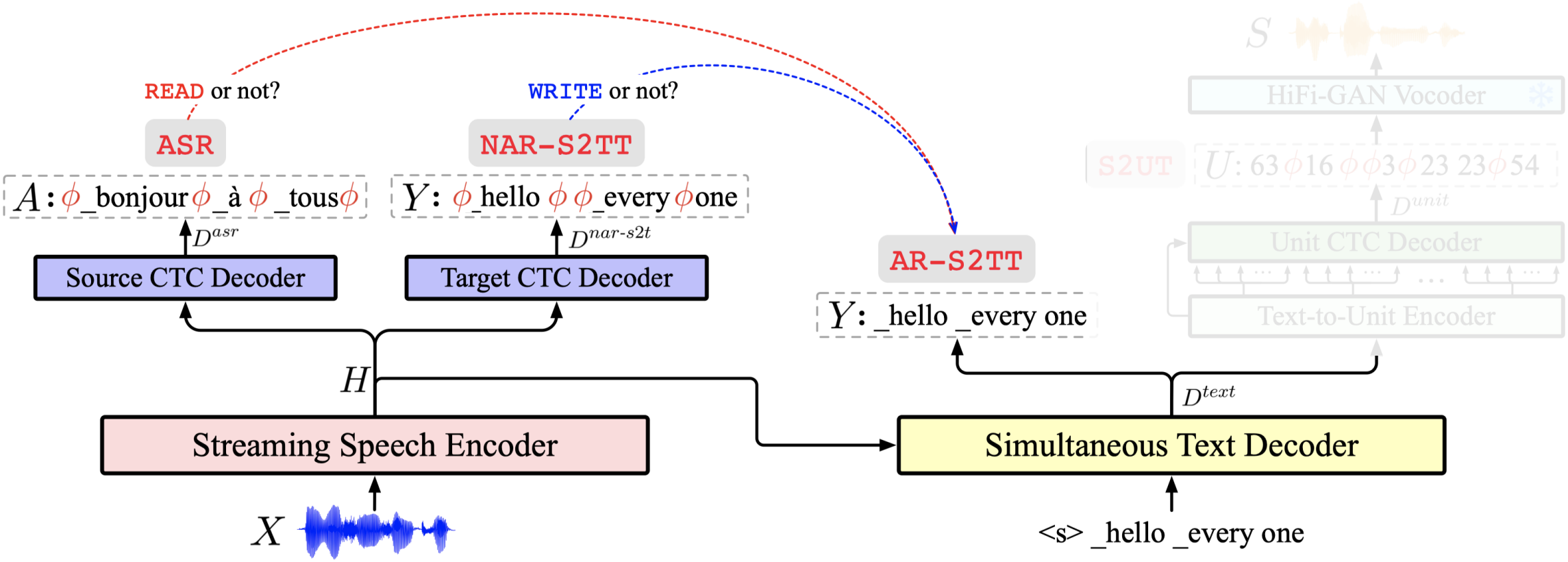
- 流式语音编码器：chunk-based Conformer
- 按chunk处理流式语音输入



Shaolei Zhang, et, al. StreamSpeech: Simultaneous Speech-to-Speech Translation with Multi-task Learning. ACL 2024.

StreamSpeech

- 实时文本解码器：引入CTC decoder来学习源语音中包含多少文本

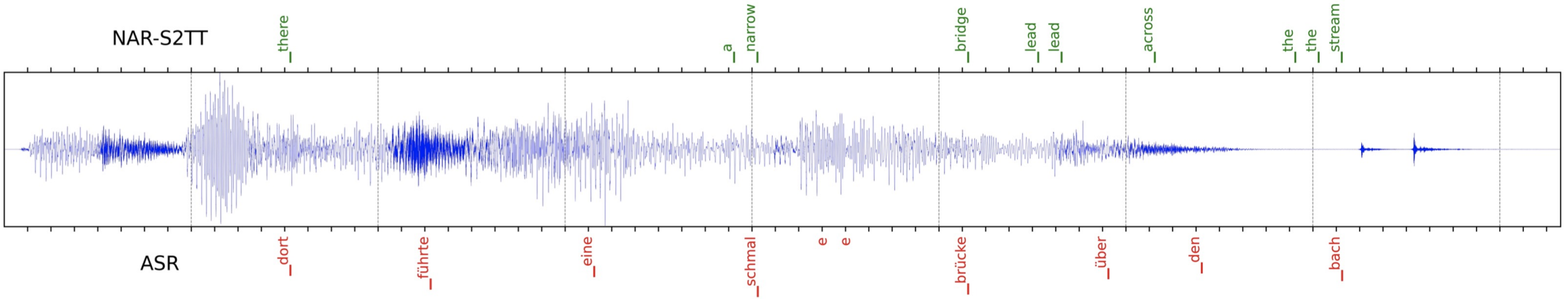


Shaolei Zhang, et, al. StreamSpeech: Simultaneous Speech-to-Speech Translation with Multi-task Learning. ACL 2024.

StreamSpeech

- 实时文本解码器：引入CTC decoder来学习源语音中包含多少文本

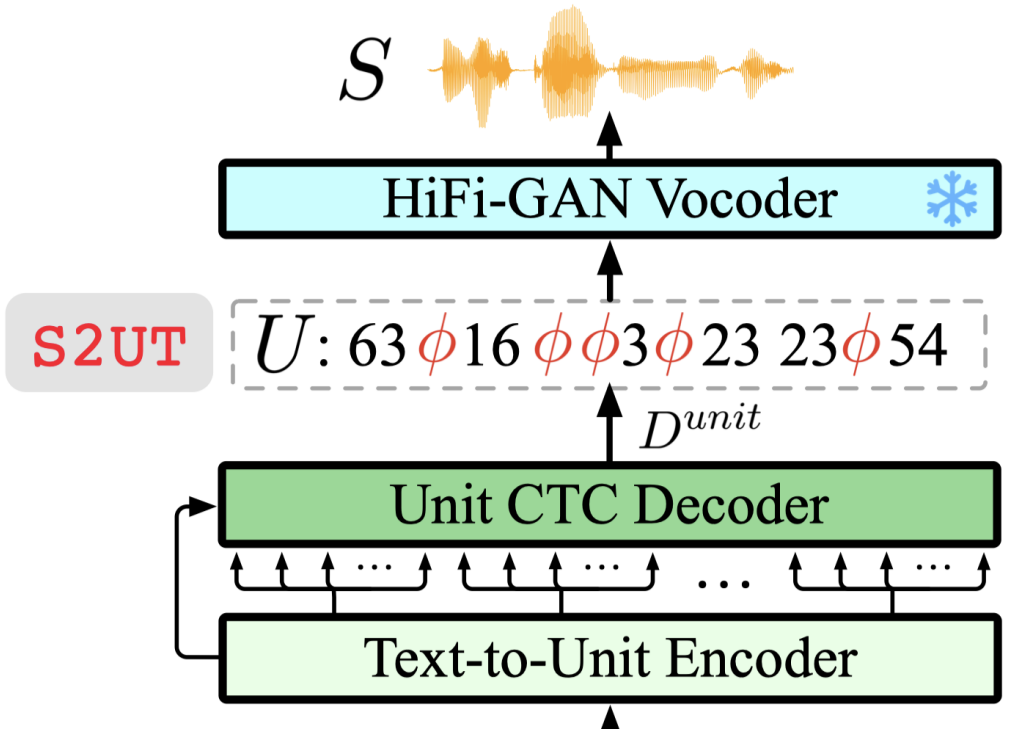
- READ：识别到新的源端词之后开始翻译
- WRITE：根据对应的target词数控制写的数量



Shaolei Zhang, et, al. StreamSpeech: Simultaneous Speech-to-Speech Translation with Multi-task Learning. ACL 2024.

StreamSpeech

- 同步语音合成：非自回归text-to-unit
- 语音合成是单调对齐的
- Unit序列过长：非自回归加速



Shaolei Zhang, et, al. StreamSpeech: Simultaneous Speech-to-Speech Translation with Multi-task Learning. ACL 2024.

离线语音到语音翻译

- 超越当前最佳的UnitY（Meta发布）
- 推理平均加速4倍

StreamSpeech 推理加速比

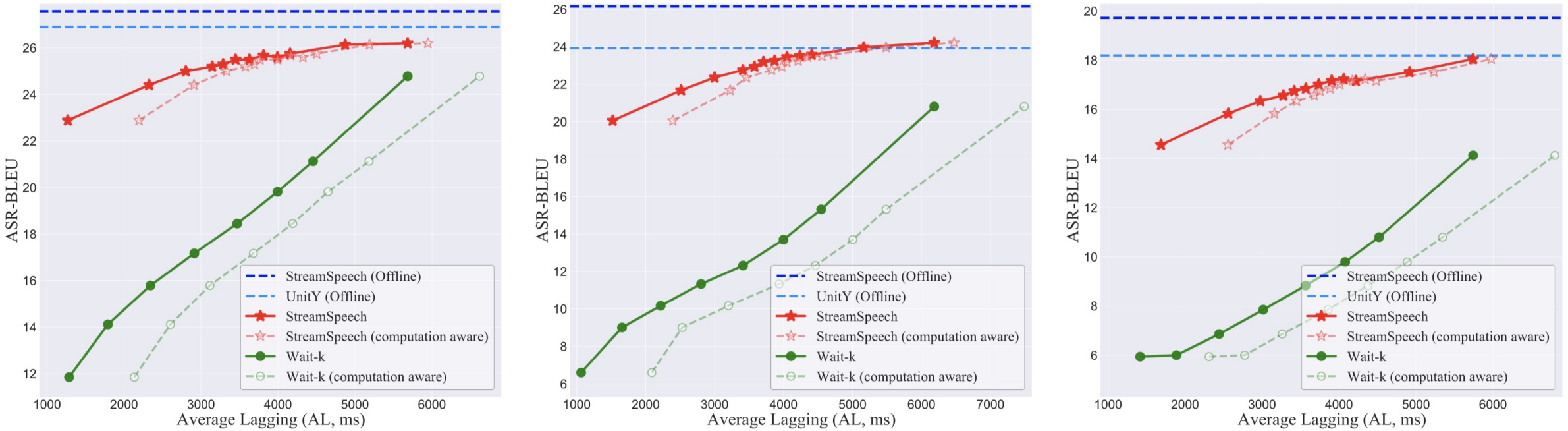
Models	Fr→En		Es→En		De→En	
	ASR-BLEU	Speedup	ASR-BLEU	Speedup	ASR-BLEU	Speedup
UnitY	27.77	1.0×	24.95	1.0×	18.74	1.0×
StreamSpeech	**28.45**	**3.6×**	**27.25**	**4.5×**	**20.93**	**4.5×**

StreamSpeech 离线语音到语音翻译性能

Models	#Param.	Fr→En		Es→En		De→En		Average	
		greedy	beam10	greedy	beam10	greedy	beam10	greedy	beam10
Ground Truth	-	84.52		88.54		75.53		82.86	
S2UT	73M	20.91	22.23	16.94	18.53	2.46	2.99	13.44	14.58
Translatotron	79M	16.96	/	8.72	/	1.97	/	9.22	/
Translatotron 2	87M	25.49	26.07	22.35	22.93	16.24	16.91	21.36	21.97
DASpeech	93M	25.03	/	21.37	/	16.14	/	20.85	/
UnitY	67M	26.90	27.77	23.93	24.95	18.19	18.74	23.01	23.82
StreamSpeech	70M	**27.58****	**28.45****	**26.16****	**27.25****	**19.72****	**20.93****	**24.49**	**25.54**

实时语音到语音翻译

- 滞后2s，取得不错的翻译性能



(a) Fr→En　(b) Es→En　(c) De→En

Shaolei Zhang, et, al. StreamSpeech: Simultaneous Speech-to-Speech Translation with Multi-task Learning. ACL 2024.

实时语音识别性能

- 噪音等极端场景
- 滞后100毫秒，超过Wav2Vec2-large、Whisper-base

Models	#Parm.	AL (ms)↓	WER↓
Wav2Vec2-large	315M	5684.38	26.17
Whisper-base	74M	5684.38	38.04
StreamSpeech	70M (33M used)	109.127	25.46
		267.891	25.54
		431.652	25.20
		757.989	24.67

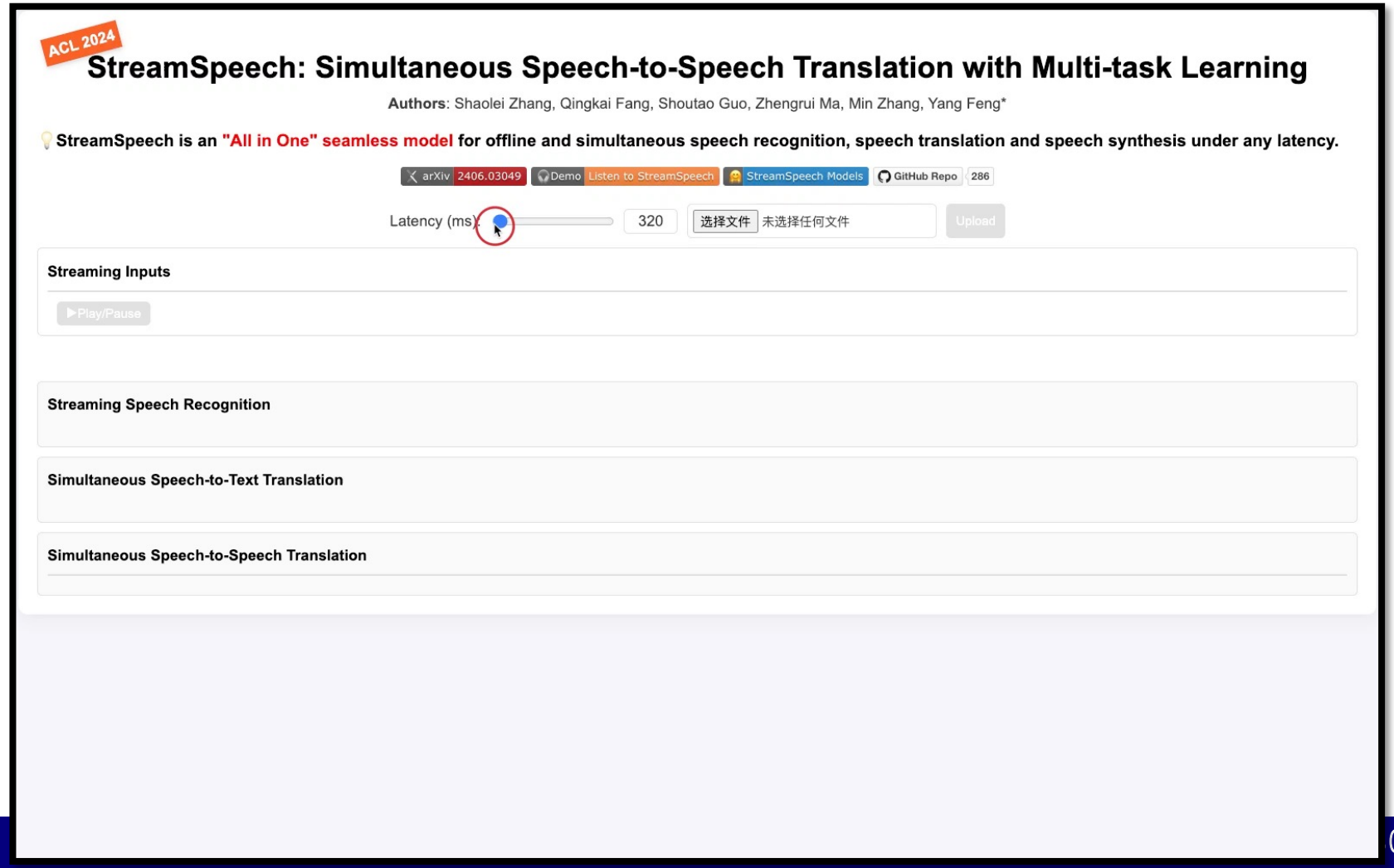
Shaolei Zhang, et, al. StreamSpeech: Simultaneous Speech-to-Speech Translation with Multi-task Learning. ACL 2024.

StreamSpeech

- Web Demo：实时语音到语音翻译，同步输出中间结果

首个一体化离线/流式语音模型
Twitter 300K+观看
开源模型1000 stars

StreamSpeech GitHub



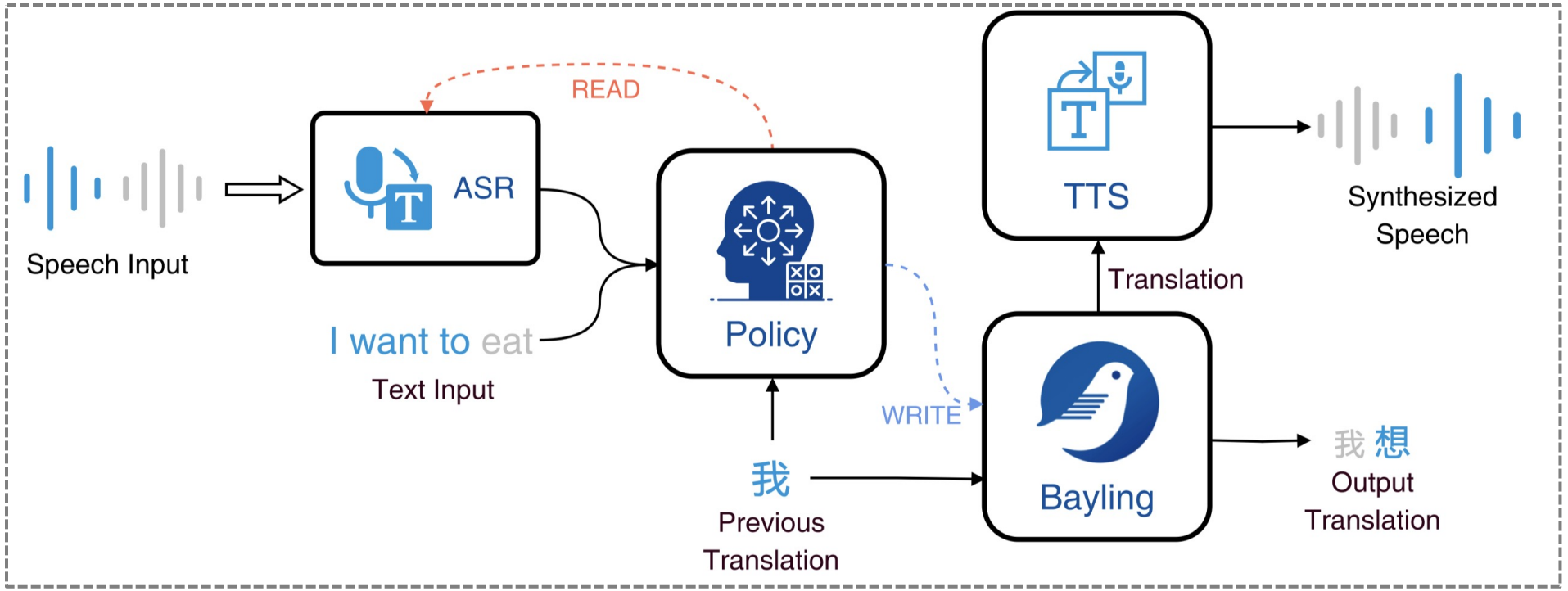
Agent-SiMT：基于智能体的实时模型

- 现有方法普遍采用单一Transformer来共同完成生成+策略
  - 策略决策和翻译生成两项子任务强行耦合
  - 未能利用LLMs的优势、模型翻译性能不佳

- 采用智能体工作流方式解决流式问题
  - 解耦两项任务
  - 充分利用LLM和Transformer-Based SiMT的优势

Agent-SiMT：基于智能体的实时模型

- 让LLM处理流式输入：决策何时开始输出
  - 策略模块 (Policy Model)：用于决策READ/WRITE action
  - LLM：较强的生成能力，但无法决策生成的时机
  - 记忆模块: 存储指令、当前上下文状态



Agent-SiMT：基于智能体的实时模型

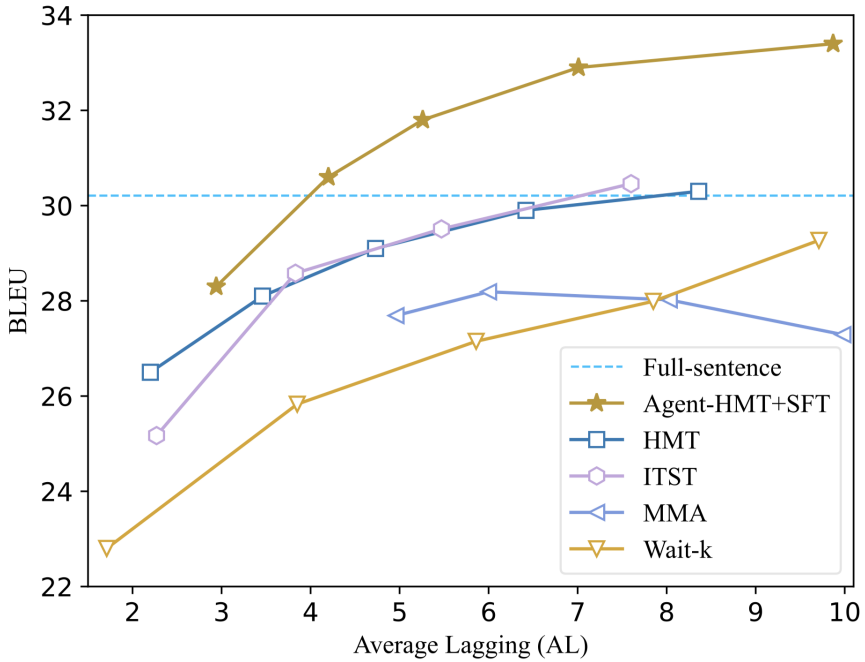
- Llama 2 + HMT、BayLing 2 + HMT
- 超过Meta发布的SeamlessM4T/SeamlessStreaming 模型

Models	S2TT				S2ST			
	Zh⇒En		En⇒Zh		Zh⇒En		En⇒Zh	
	BLEU	COMET	BLEU	COMET	ASR-BLEU	ASR-COMET	ASR-BLEU	ASR-COMET
SeamlessM4T-Large v2	22.77	82.43	28.06	79.76	23.68	82.17	26.26	73.35
BayLing-2-7B-Translate	22.20	83.72	**39.60**	86.09	23.13	83.23	31.34	**81.15**
BayLing-2-13B-Translate	23.07	83.89	39.16	**86.17**	23.70	83.13	**31.60**	80.76
BayLing-3-8B-Translate	**25.40**	**84.76**	38.64	85.99	**26.22**	**84.06**	31.46	80.70

BayLing + HMT, Zh-En离线语音翻译

Models	Zh⇒En			En⇒Zh		
	LAAL	SacreBLEU	COMET	LAAL	SacreBLEU	COMET
SeamlessStreaming	**2560.34**	**20.88**	**81.20**	2064.104	21.93	72.16
BayLing-2-7B-Translate (HMT)	3322.24	19.35	80.61	2112.70	29.62	77.34
BayLing-2-13B-Translate (HMT)	3214.67	19.56	81.96	**2028.47**	**30.19**	**76.97**

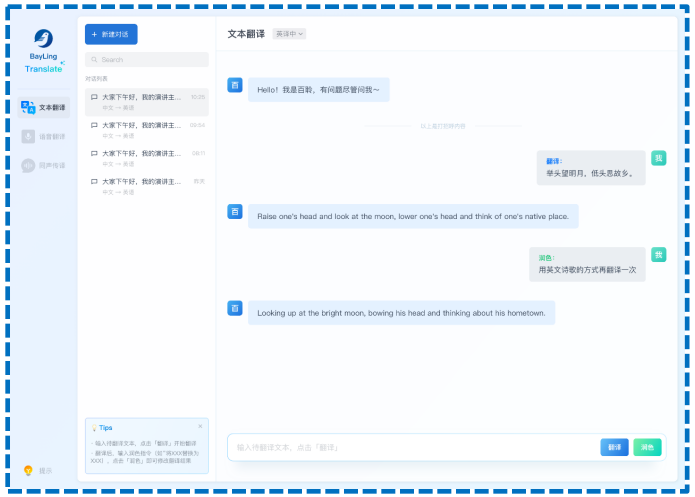
BayLing + HMT, Zh-En实时语音翻译



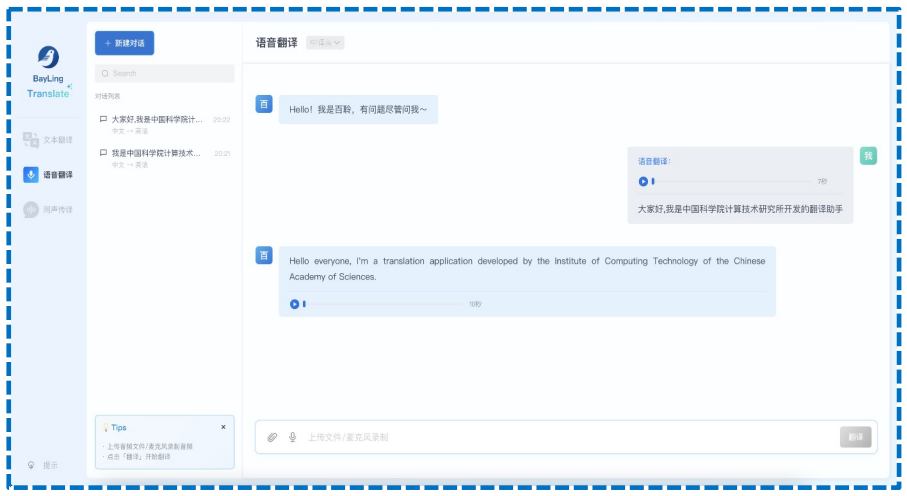
Llama + HMT, 德英实时文本翻译

Agent-SiMT：基于智能体的实时模型

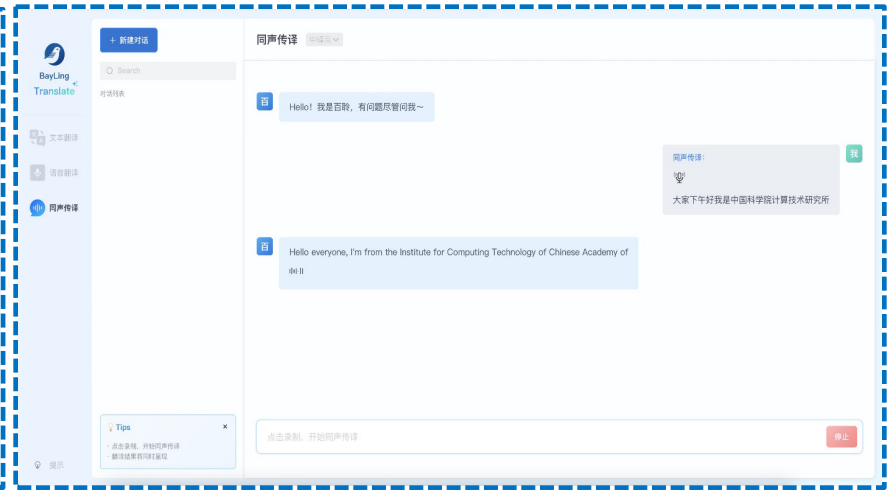
- 在线Demo：开放了翻译部分功能
  - 支持文本翻译、语音翻译、同声传译



文本翻译



语音翻译



同声传译

首个以LLM为核心的实时翻译系统

LLM-based Interaction

- 如何构建能实时沟通的大模型？

- 依赖多模态大模型：语音、视觉…
  - 实时沟通的核心：快速、高效

LLM-based Interaction

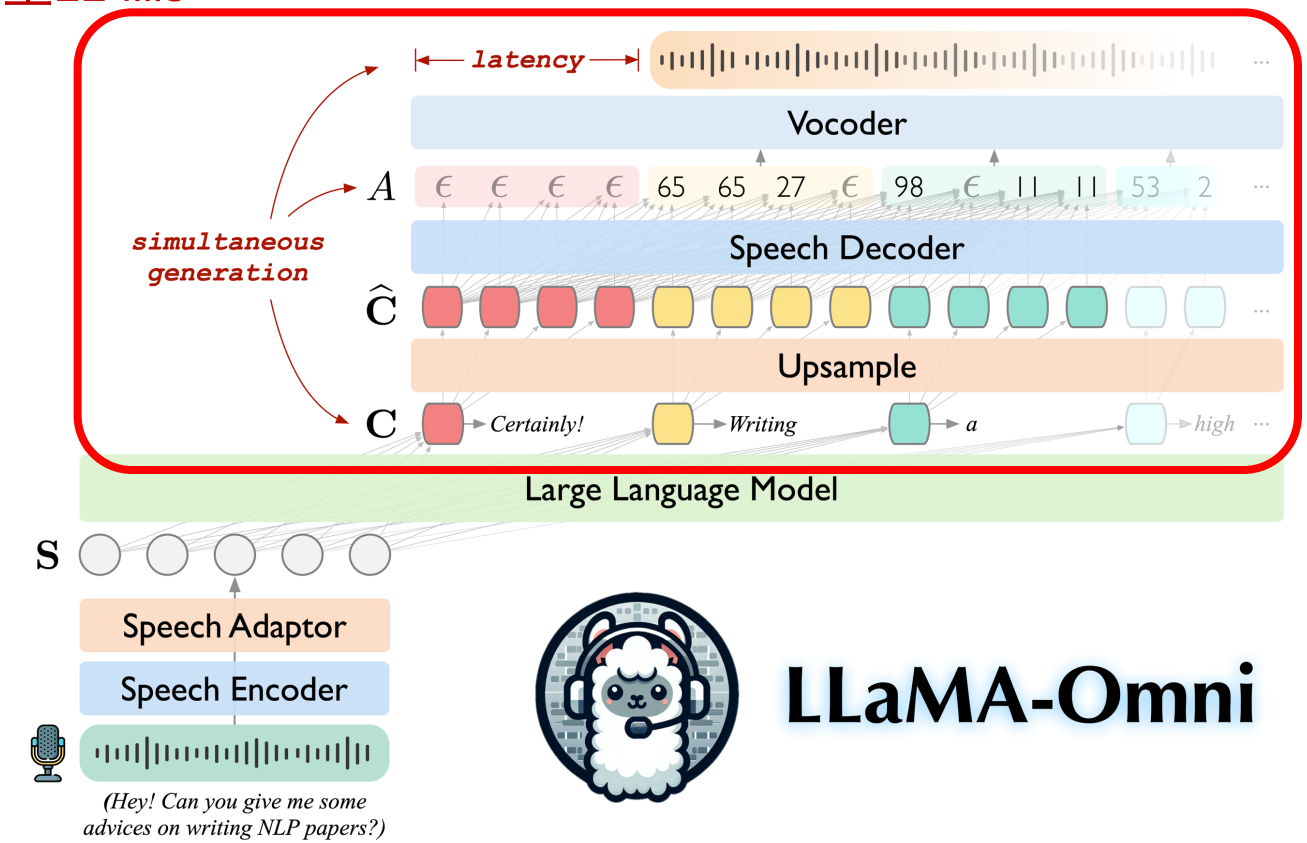
- 如何构建能实时沟通的大模型？

- 依赖多模态大模型：语音、视觉…
  - 实时沟通的核心：快速、高效

构建快速、高效的
语音大模型和视觉大模型

即时交互语音大模型：LLaMA-Omni

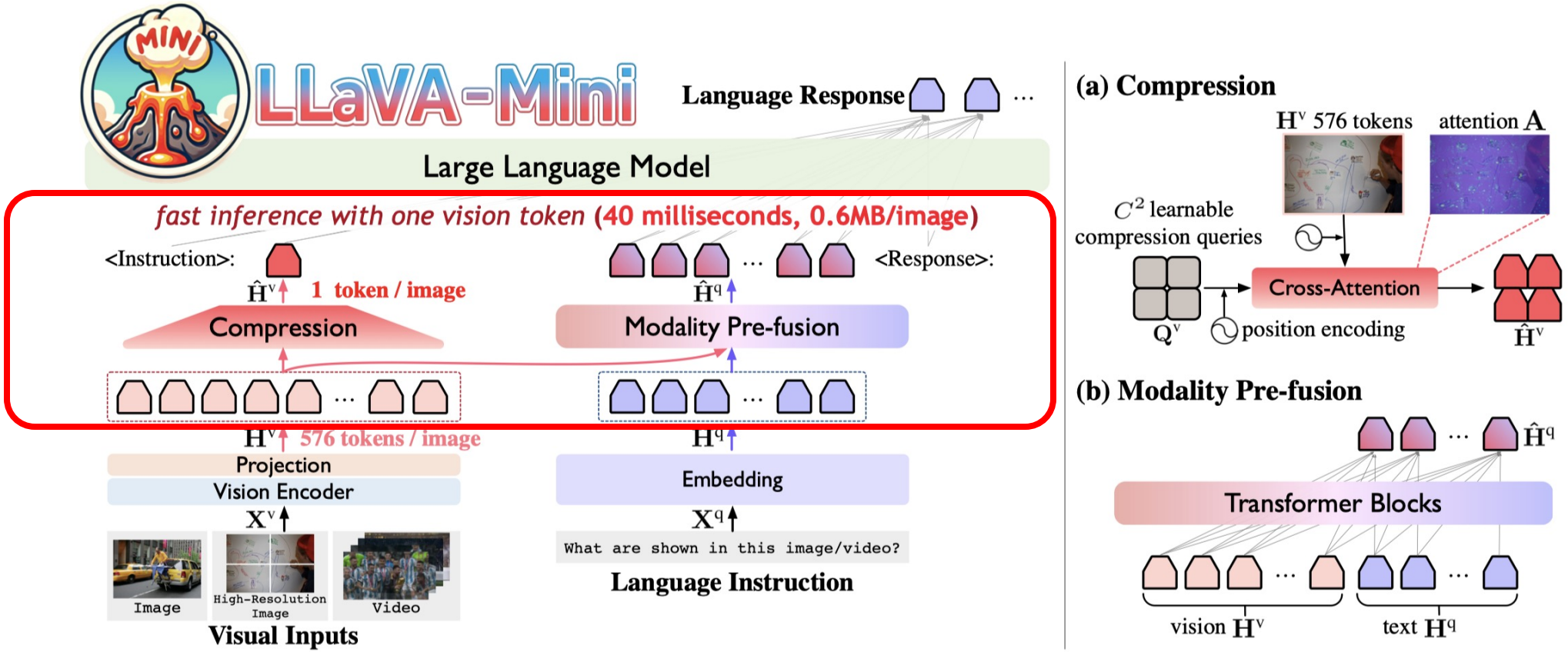
- **LLaMA-Omni：类GPT-4o即时交互大模型**
  - 同步生成文本和语音回复，延时低至224ms
  - 语音编码器：Whisper-large-v3
  - 语音适配器：下采样 + MLP
  - LLM：Llama-3.1-8B-Instruct
  - 语音解码器
    - 模型结构是 2 层单向 Transformer
    - 使用 CTC 建模，非自回归生成



Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, Yang Feng. LLaMA-Omni: Seamless Speech Interaction with Large Language Models. Preprint 2024.

高效视觉大模型：LLaVA-Mini

- **LLaVA-Mini：统一图像、高分辨率图像、视频理解**
  - 模态预融合+压缩：使用1个token表示一张图像，视频按照1fps采样

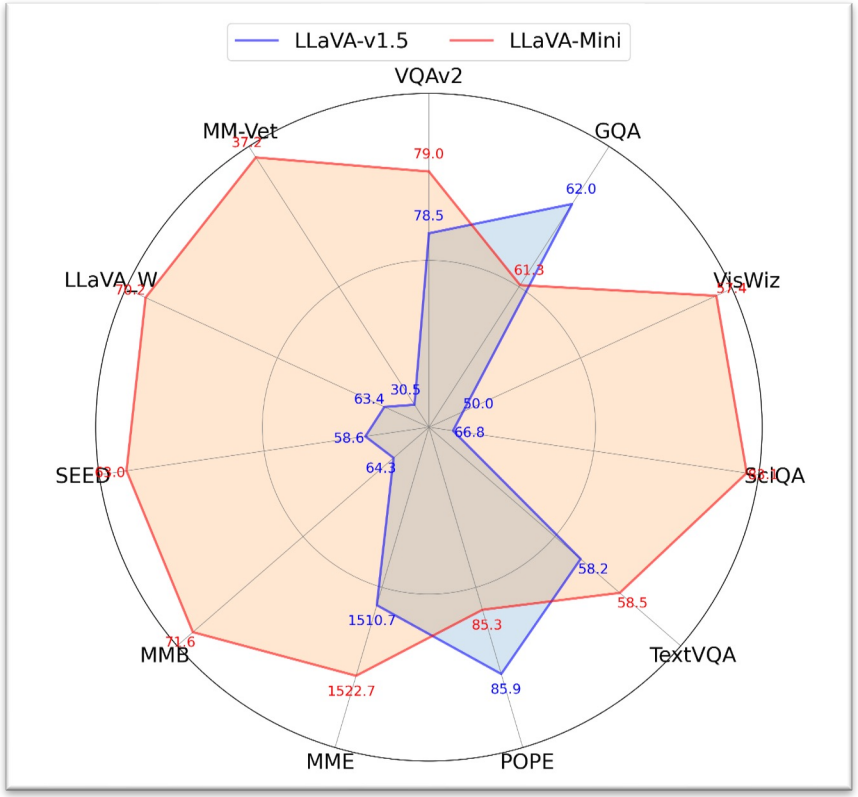


Shaolei Zhang, et,al. LLaVA-Mini: Efficient Image and Video Large Multimodal Model with One Vision Token. Preprint.

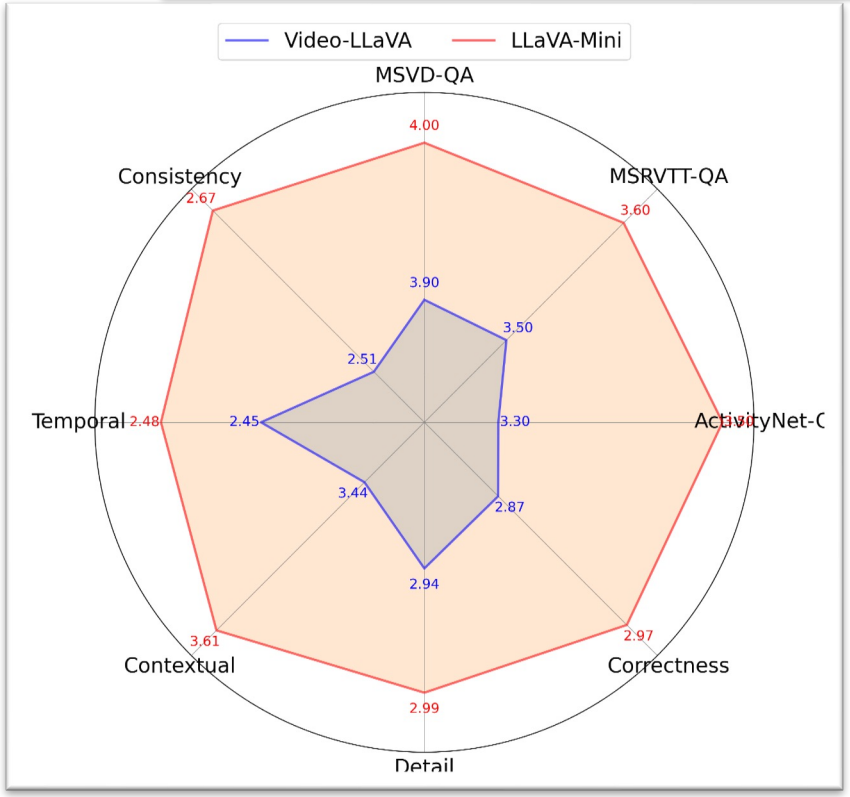
高效视觉大模型：LLaVA-Mini

- **LLaVA-Mini：统一图像、高分辨率图像、视频理解**
  - 使用1个token表示一张图像，视频按照1fps采样

媲美576个token的LLaVA-v1.5



图像理解

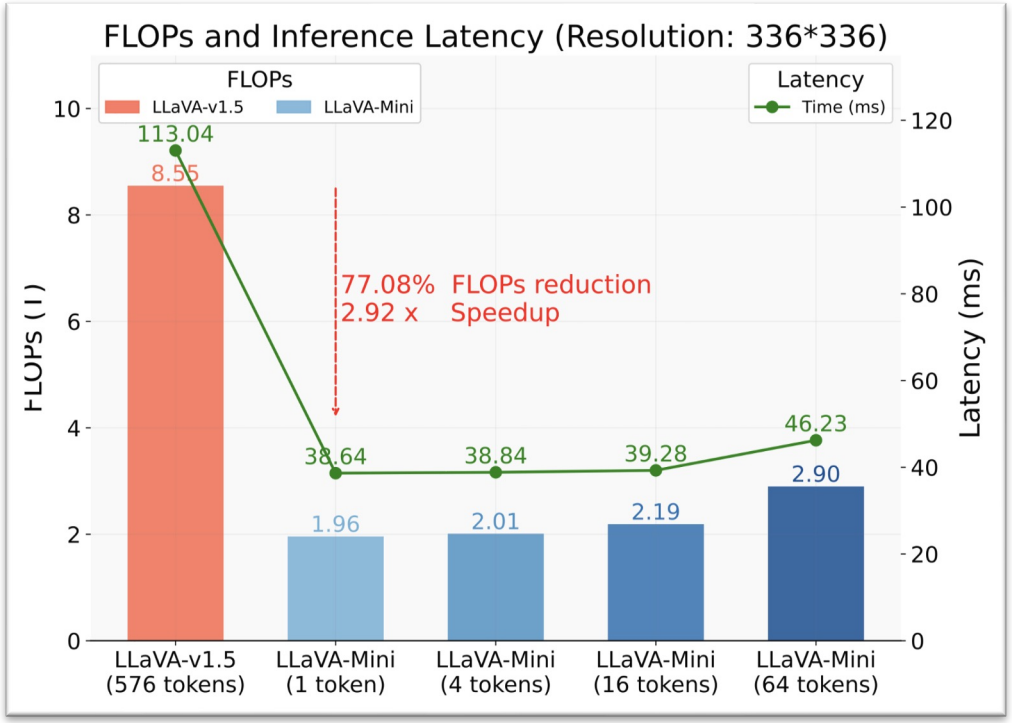


视频理解

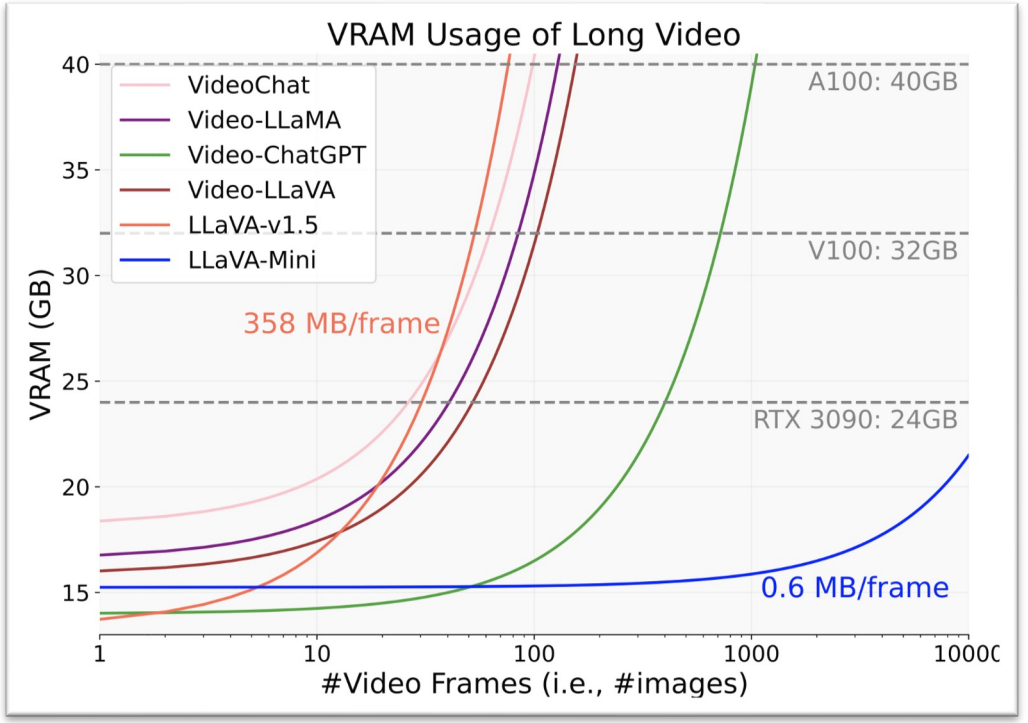
高效视觉大模型：LLaVA-Mini

- **LLaVA-Mini：统一图像、高分辨率图像、视频理解**
  - 使用1个token表示一张图像，视频按照1fps采样

3倍推理加速、500倍显存减少



图像处理延时**40毫秒**



支持**3小时+**视频理解

实时跨语言交互

- 实时语音模型：
  - 级联（HMT）、端到端（DiSeg）、All in One（StreamSpeech）、LLM-based（Agent-SiMT）
- 高效多模态大模型：LLaMA-Omni、LLaVA-Mini
- 未来趋势：
  - 通用性：一个模型支持多项任务
  - LLM-based：充分利用LLM的通用性、鲁棒性
  - 端侧模型：耳机、眼镜的轻量化模型

端侧一体化实时语音模型 StreamSpeech

即时语音交互大模型 LLaMA-Omni

高效图像视频理解大模型 LLaVA-Mini

谢 谢 大 家 !

更多研究/项目介绍

张绍磊

中国科学院计算技术研究所

zhangshaolei20z@ict.ac.cn

https://zhangshaolei1998.github.io/