流式模型研究进展

姓 名：张绍磊

导 师：冯洋 研究员

中国科学院计算技术研究所

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

2024.06.13

流式模型

- 流式模型：输出和输入同时进行
  - 低响应延时、高质量

- 主要内容
  - **Encoder-Decoder 架构：** StreamSpeech: Simultaneous Speech-to-Speech Translation with Multi-task Learning
  - **Decoder-only 架构：** Decoder-only Streaming Transformer for Simultaneous Translation
  - **Offline LLM 架构：** Agent-SiMT: Agent-assisted Simultaneous Machine Translation with Large Language Models

StreamSpeech

StreamSpeech: Simultaneous Speech-to-Speech Translation with Multi-task Learning

Shaolei Zhang[1,3], **Qingkai Fang**[1,3], **Shoutao Guo**[1,3], **Zhengrui Ma**[1,3], **Min Zhang**[4], **Yang Feng**[1,2,3]*

[1]Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS)
[2]Key Laboratory of AI Safety, Chinese Academy of Sciences
[3]University of Chinese Academy of Sciences, Beijing, China
[4]School of Future Science and Engineering, Soochow University


张绍磊


房庆凯


郭守涛


马铮睿


张民


冯洋

StreamSpeech

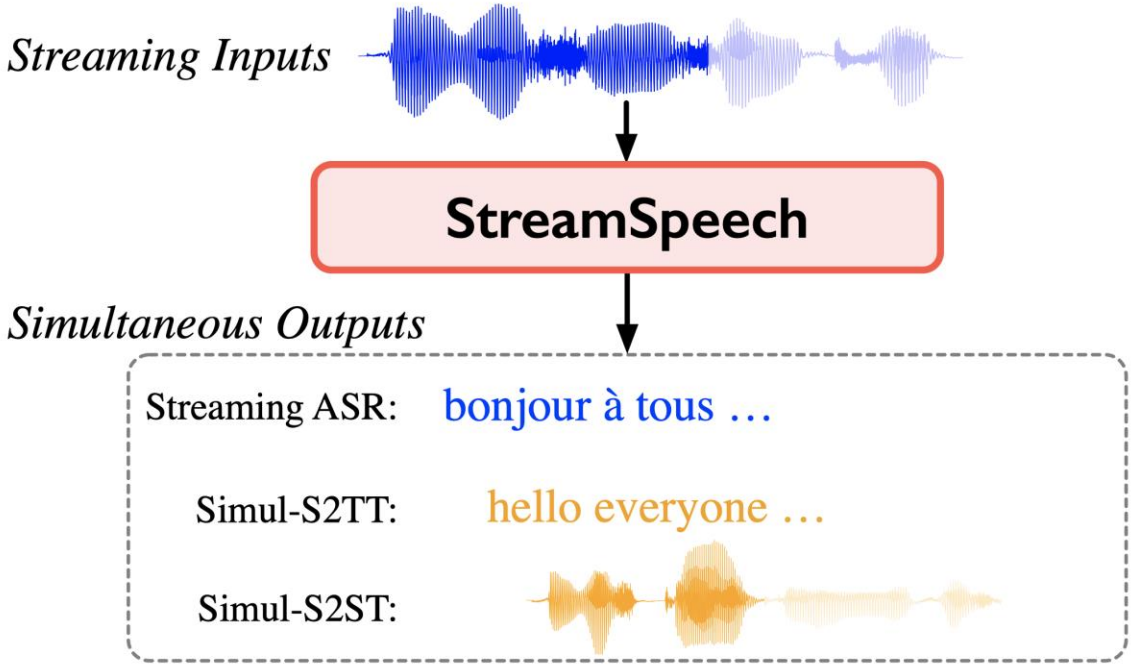
- 能否以端到端的方式完成 流式语音到语音翻译？

- 挑战：
  - 翻译：语音模态的表达形式更加多样化
  - 策略：语音是连续的，且每个词持续时间不确定

- 解决方案：
  - 引入文本模态为翻译和策略提供指导

StreamSpeech

- “All in One”：同时完成语音识别、翻译、合成等任务
- 在翻译过程中提供中间ASR、翻译结果作为额外参考

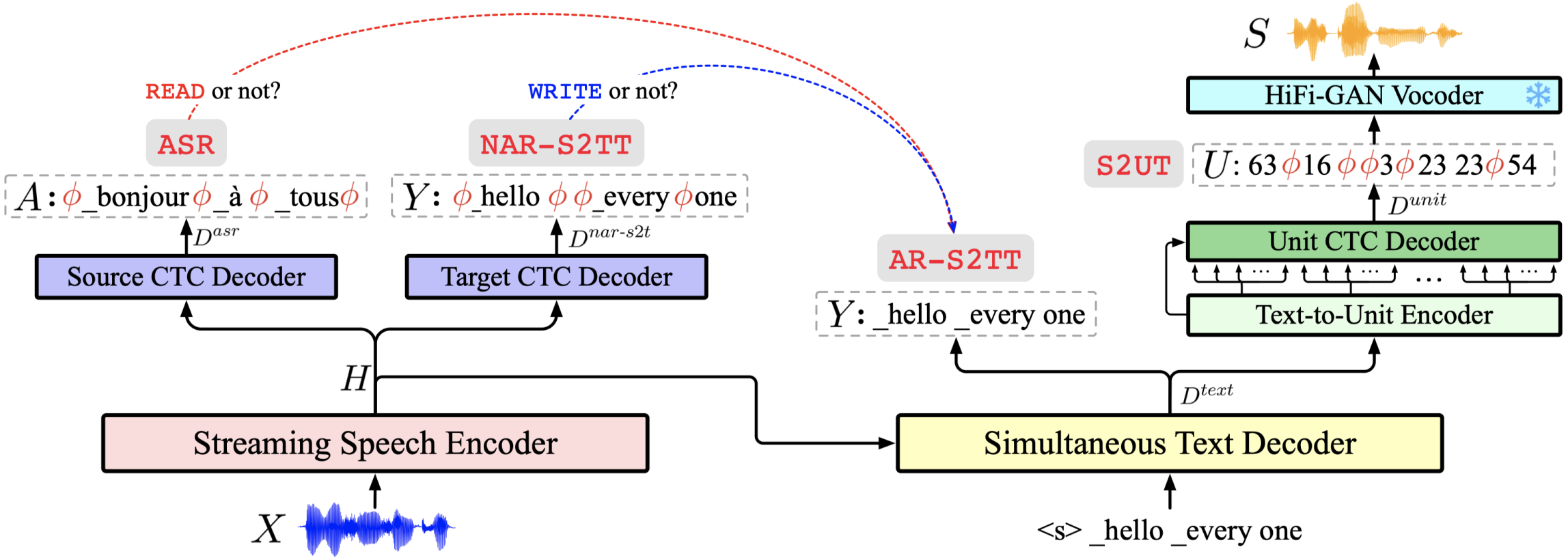


“**All in One**” seamless model for offline and simultaneous ASR / Translation / Synthesis

各任务间互相指导

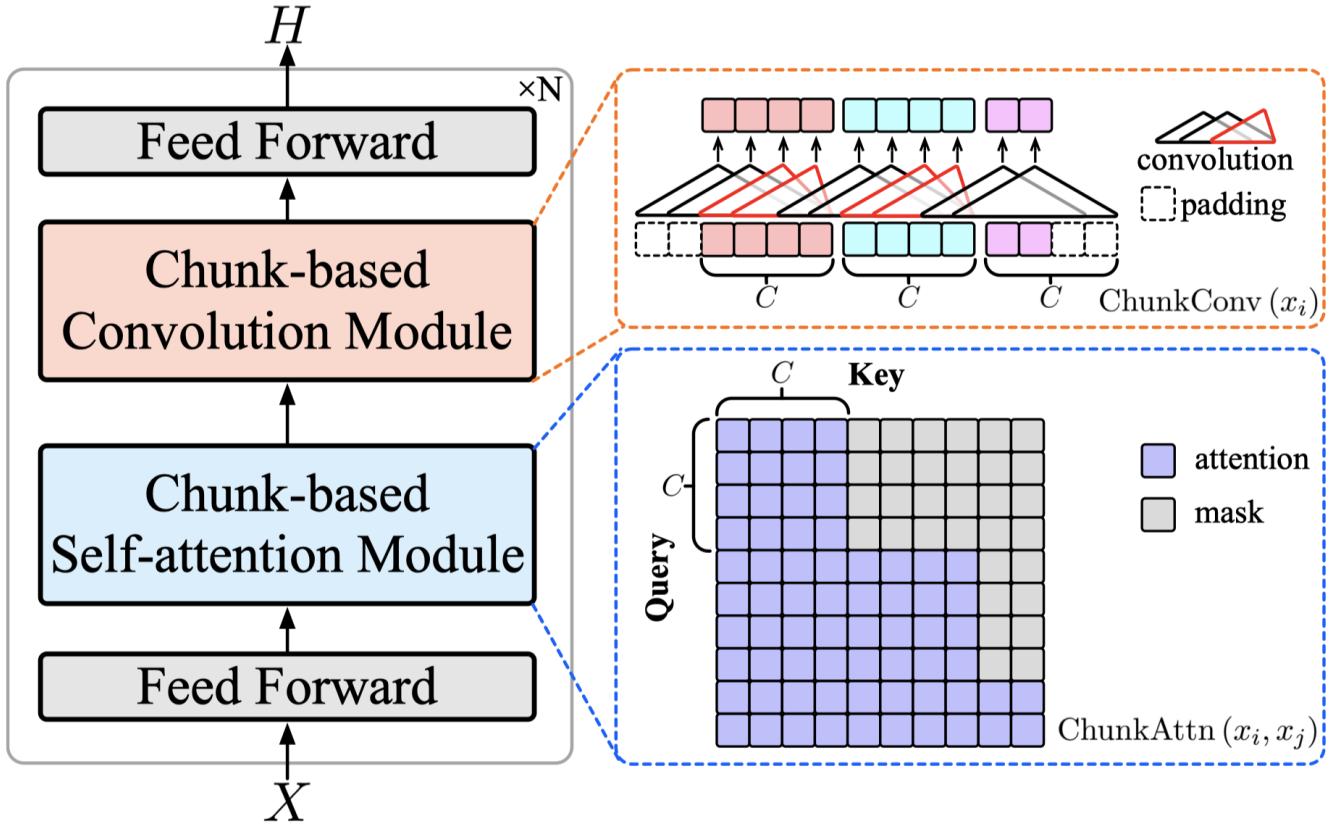
StreamSpeech

- 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



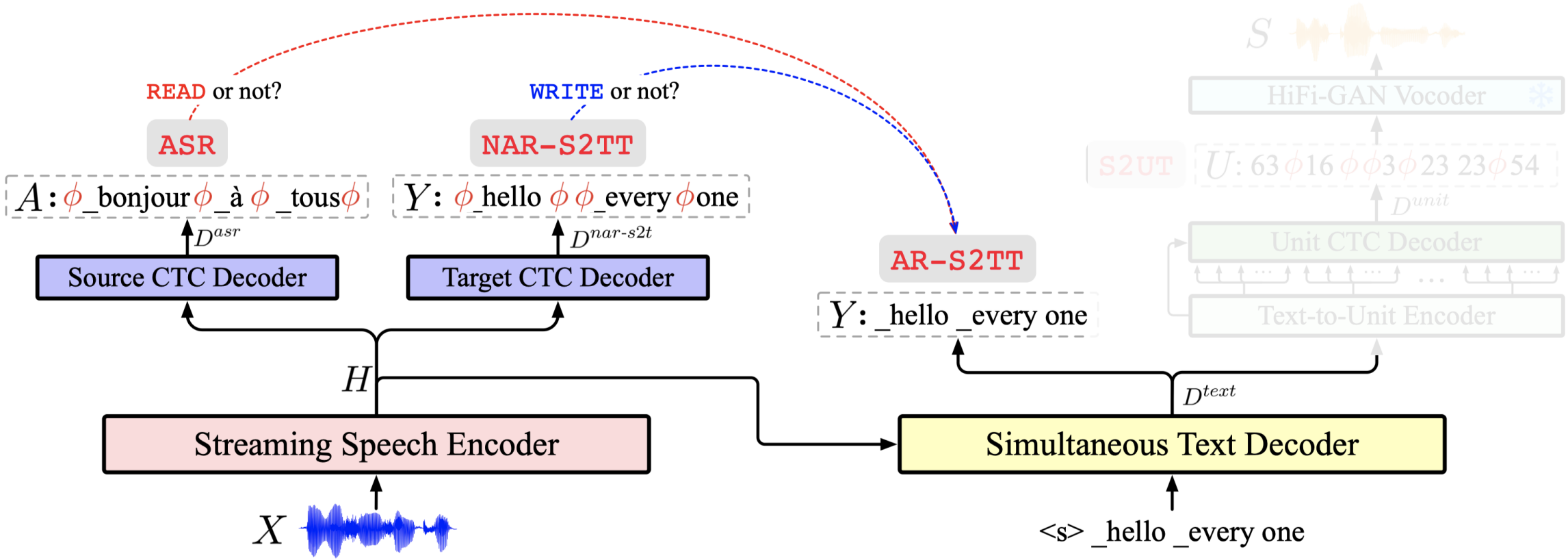
StreamSpeech

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



StreamSpeech

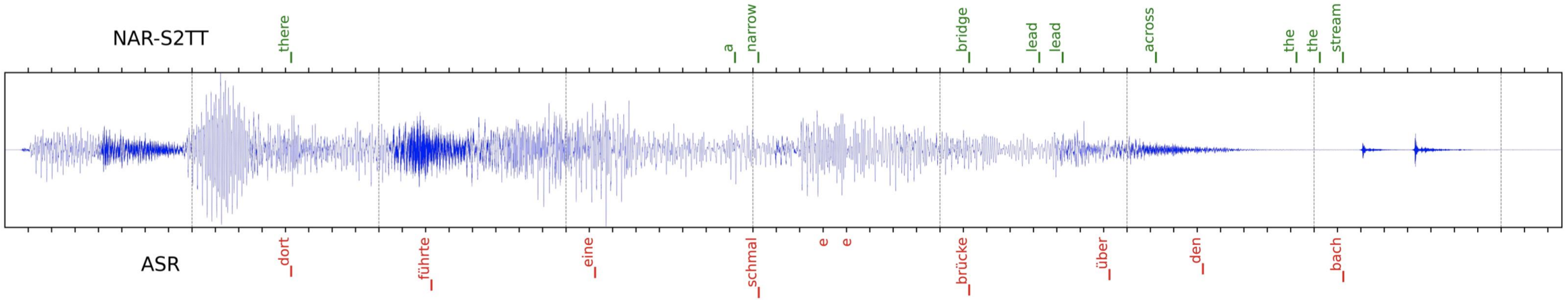
- 实时文本解码器：引入CTC decoder来学习源语音中包含多少文本



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

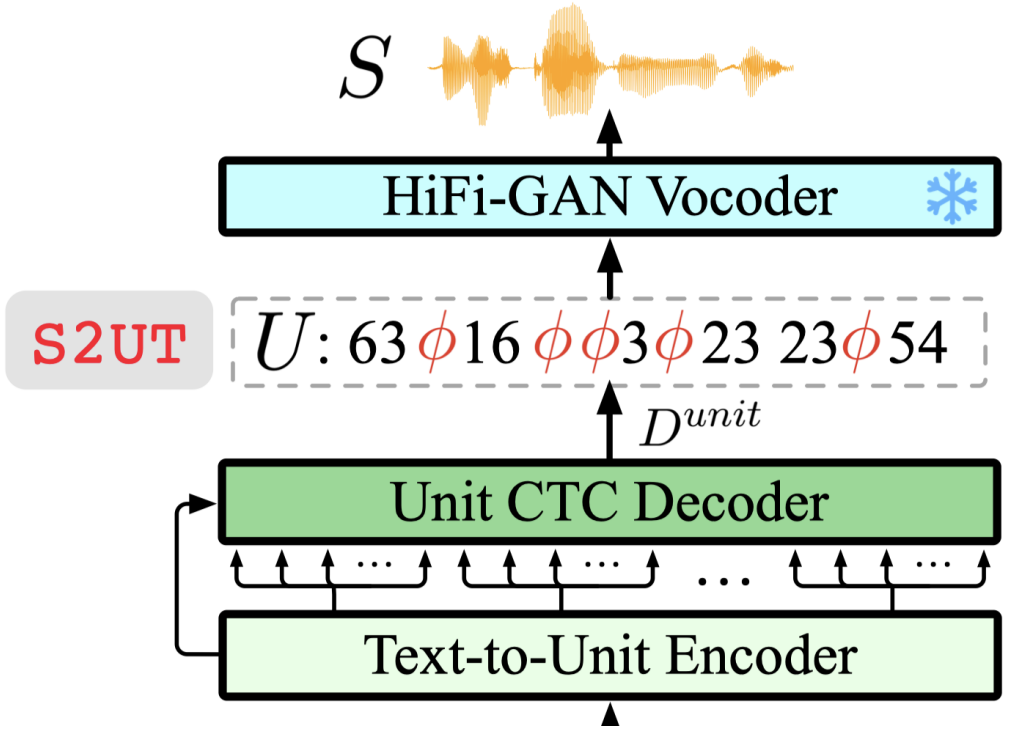
- READ：识别到新的源端词之后开始翻译
- WRITE：根据对应的target词数控制写的数量



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



离线语音到语音翻译

- 超越当前最佳的UnitY
- 平均加速4倍

StreamSpeech 加速比

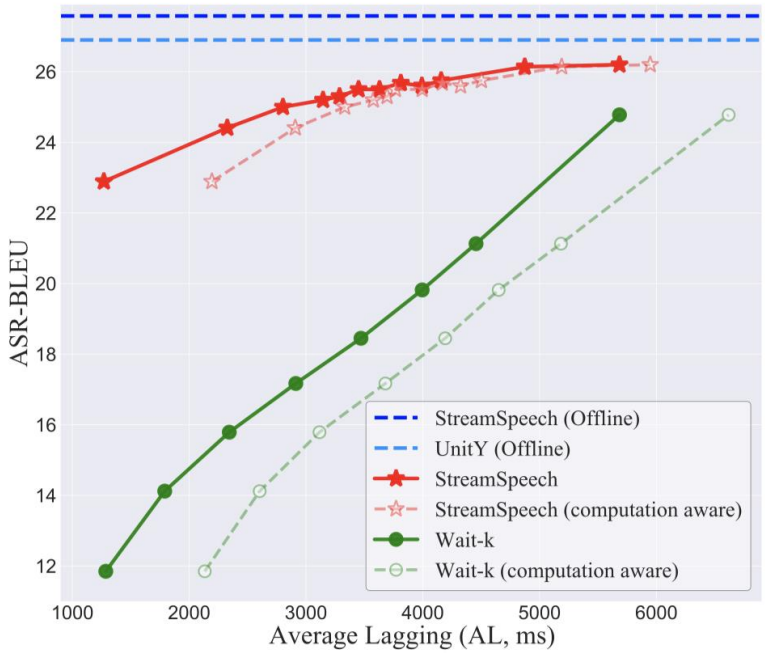
Models	Fr→En		Es→En		De→En	
	ASR-BLEU	Speedup	ASR-BLEU	Speedup	ASR-BLEU	Speedup
UnitY	27.77	1.0×	24.95	1.0×	18.74	1.0×
StreamSpeech	**28.45**	**3.6×**	**27.25**	**4.5×**	**20.93**	**4.5×**

StreamSpeech 离线性能

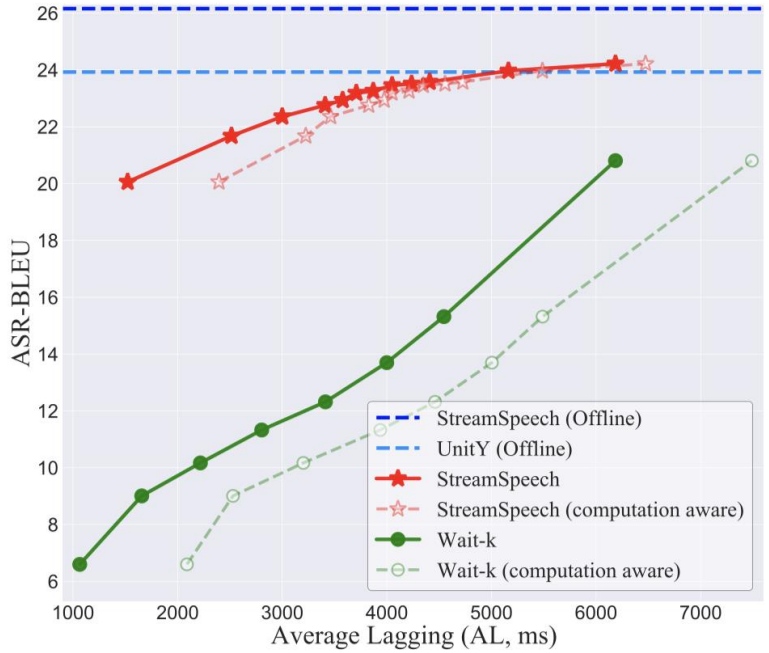
Models	#Param.	Fr→En		Es→En		De→En		Average	
		greedy	beam10	greedy	beam10	greedy	beam10	greedy	beam10
Ground Truth	-	84.52		88.54		75.53		82.86	
S2UT	73M	20.91	22.23	16.94	18.53	2.46	2.99	13.44	14.58
Translatotron	79M	16.96	/	8.72	/	1.97	/	9.22	/
Translatotron 2	87M	25.49	26.07	22.35	22.93	16.24	16.91	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

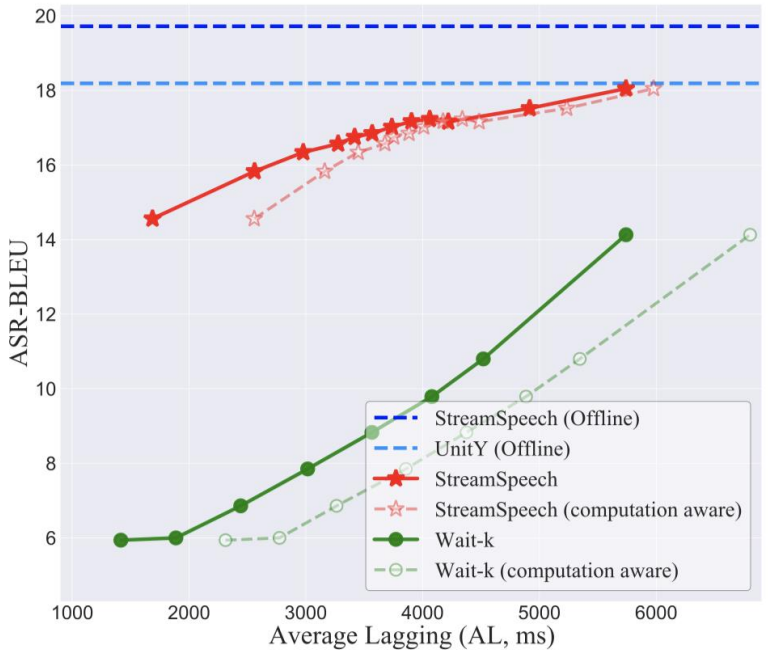
- 滞后2s，取得不错的翻译性能



(a) Fr→En



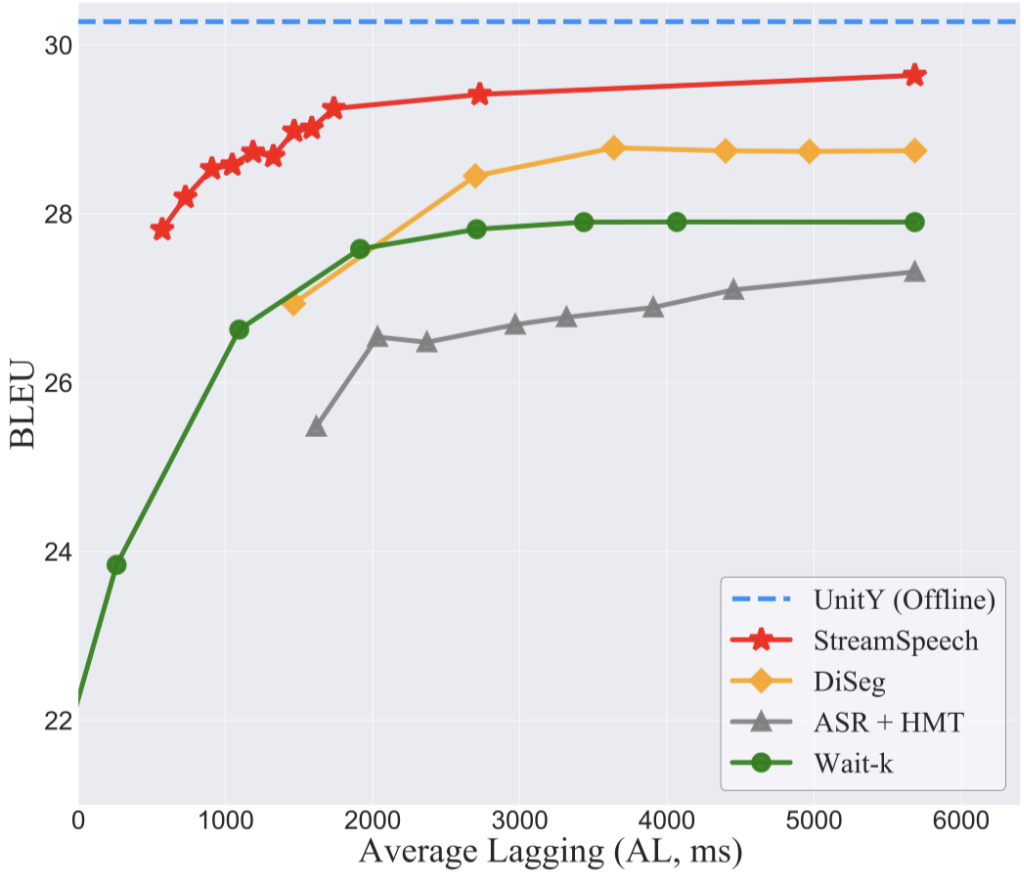
(b) Es→En



(c) De→En

实时语音到文本翻译

- 超过之前的端到端、级联方案



流式语音识别

- 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

StreamSpeech

- 实时语音到语音翻译Demo，同步输出中间结果

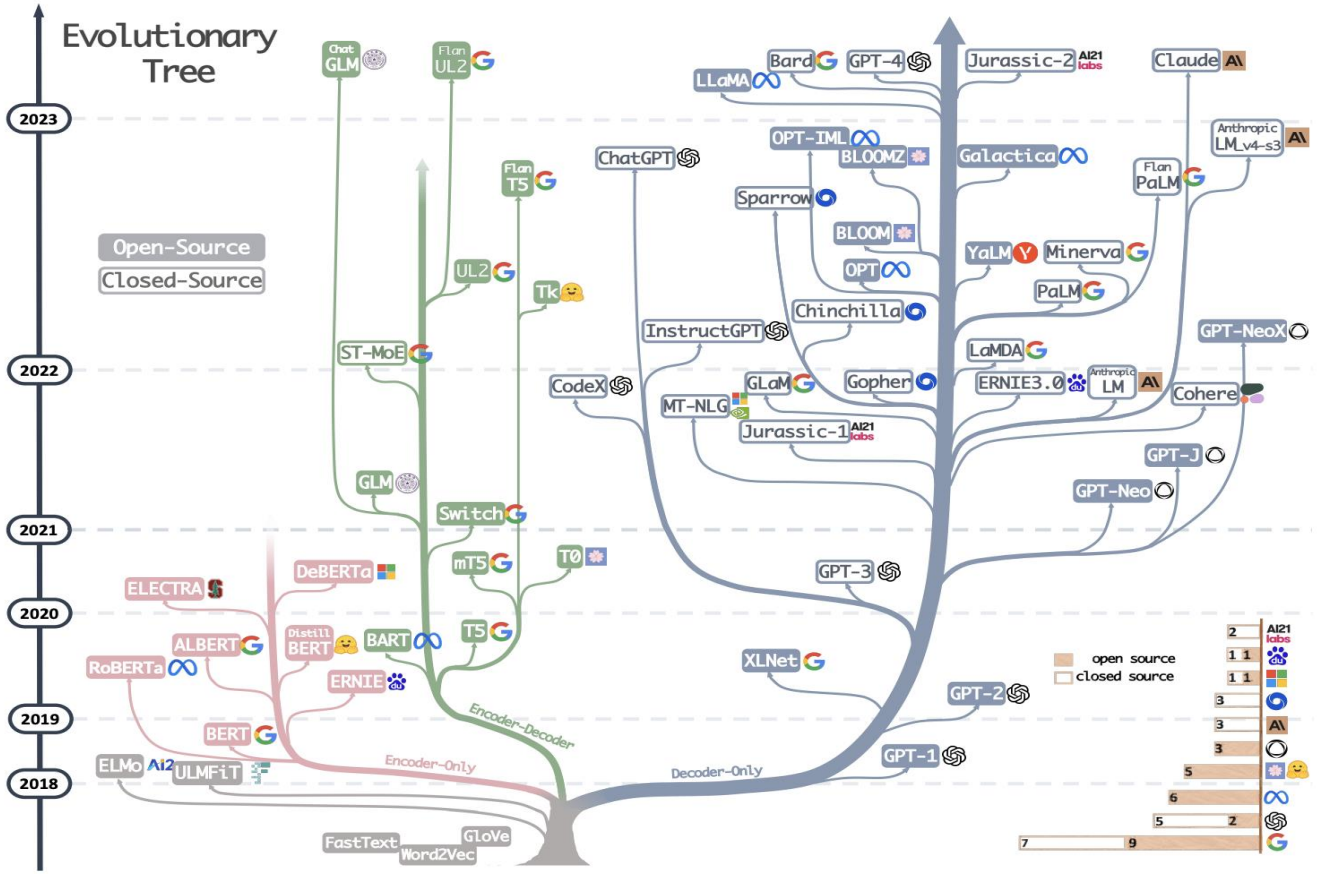
Twitter 10w+观看
收获600+stars

StreamSpeech

- 总结：
  - “All in One”端到端模型：处理离线/实时语音识别、翻译、合成任务
  - 无需设计复杂策略
  - 可以呈现中间结果，提供更全面的体验

目前的大模型

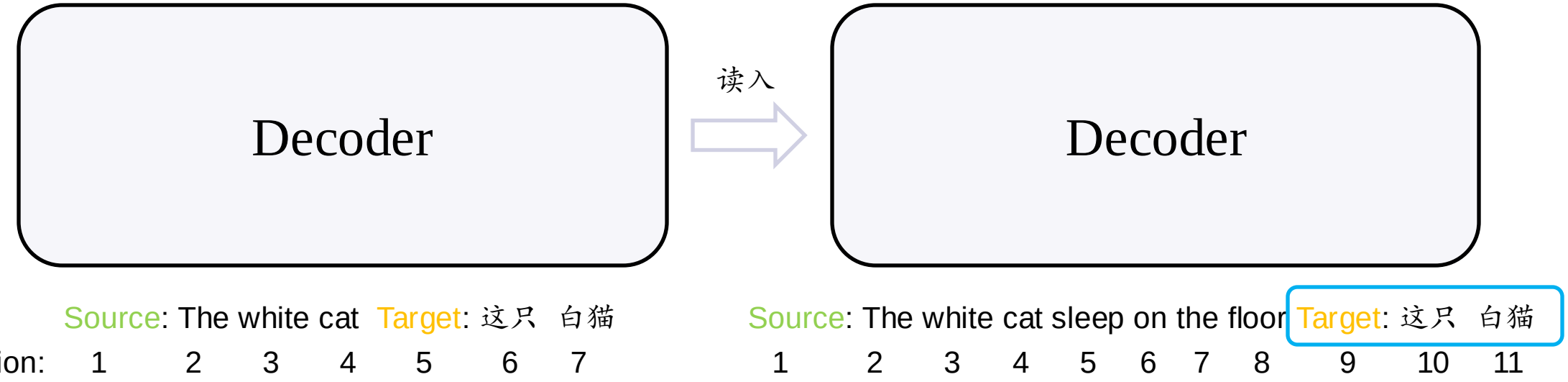
- Decoder-only架构逐渐成为主流架构



Yang et al. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. In arXiv:2304.13712.

当Decoder-only模型遇上流式输入

- 挑战
  - 推理代价：因流式输入而导致已生成内容的重复编码
  - 训练代价：策略学习的指数级复杂度



Decoder-only Streaming Transformer (DST)

Decoder-only Streaming Transformer for Simultaneous Translation

Shoutao Guo [1,3], **Shaolei Zhang** [1,3], **Yang Feng** [1,2,3]*
[1]Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS)
[2] Key Laboratory of AI Safety, Chinese Academy of Sciences
[3] University of Chinese Academy of Sciences, Beijing, China

郭守涛

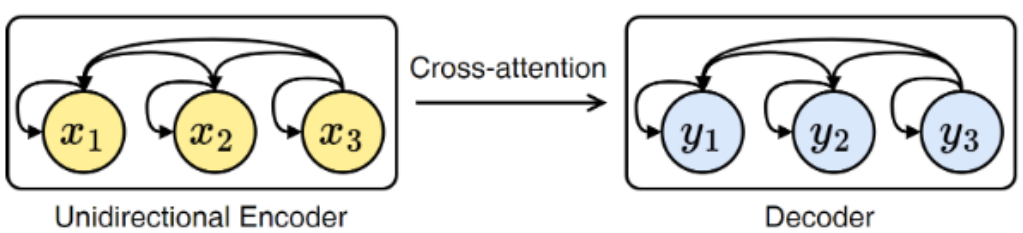
张绍磊

冯洋

19

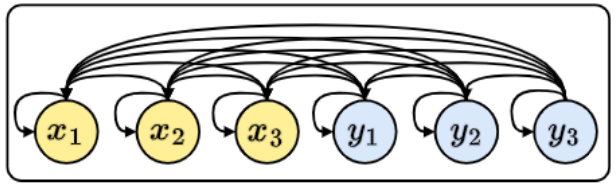
Decoder-only Streaming Transformer (DST)

- Encoder-Decoder 架构 vs Decoder-only 架构



Encoder-Decoder 架构

VS



Decoder-only 架构

- 解决方案

推理代价：因流式输入而导致已生成内容的**重复编码**。

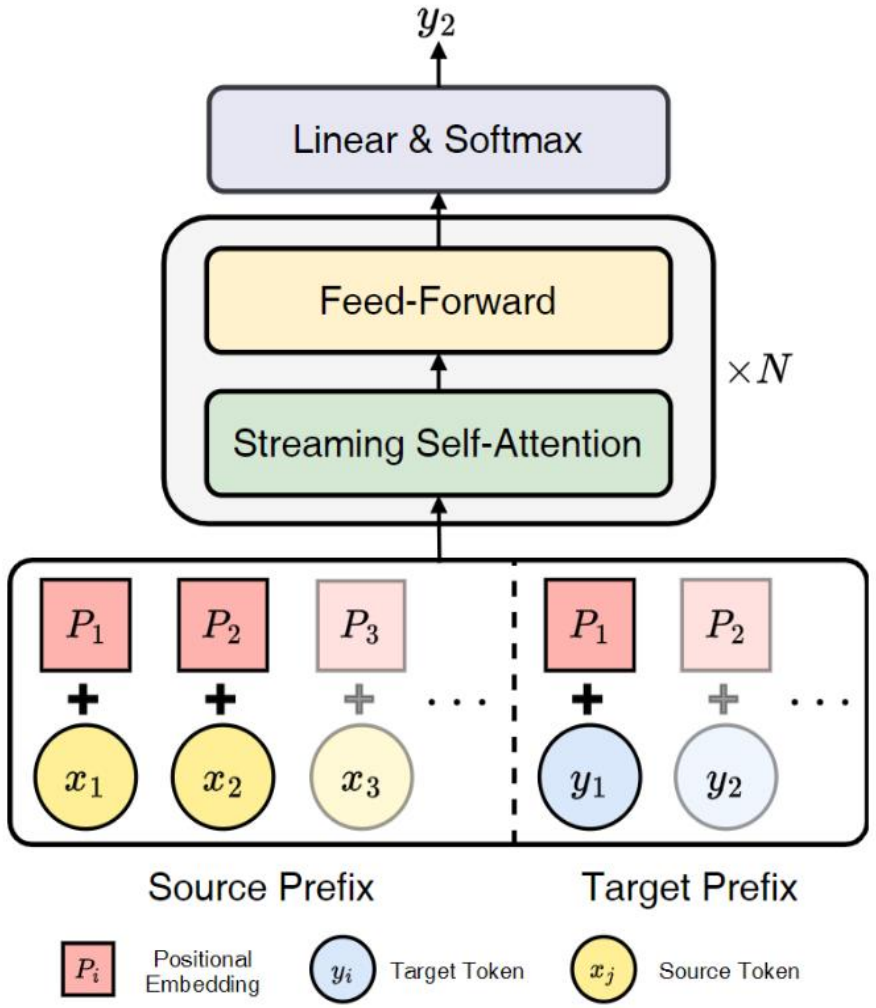
解决方案：源端和目标端信息的独立位置编码。

训练代价：策略学习的指数级复杂度。

解决方案：期望注意力机制让目标端单词适应不同长度源端前缀，降低训练复杂度为平方级别。

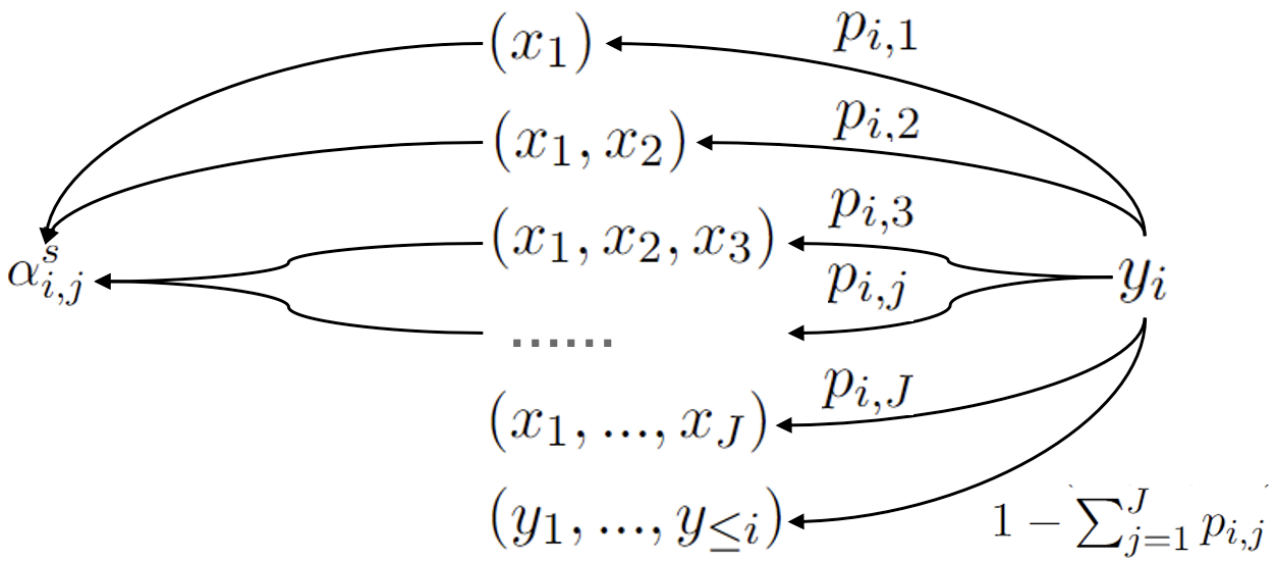
Decoder-only Streaming Transformer (DST)

- 模型架构
  - 独立位置编码
    - 防止已生成目标端前缀的重复编码
  - 流式自注意力 (Streaming Self-Attention)
    - 替代了掩码自注意力
    - 决定翻译策略
    - 利用动态规划学习翻译策略



Streaming Self-Attention

- **训练：**源端和目标端采用不同注意力机制
  - 源端单词内部仍执行掩码自注意力机制
  - 目标端单词采用二级注意力分配
    - 首先分配注意力给不同长度的源端前缀
    - 随后计算源端单词从不同前缀中获取的期望注意力



Streaming Self-Attention

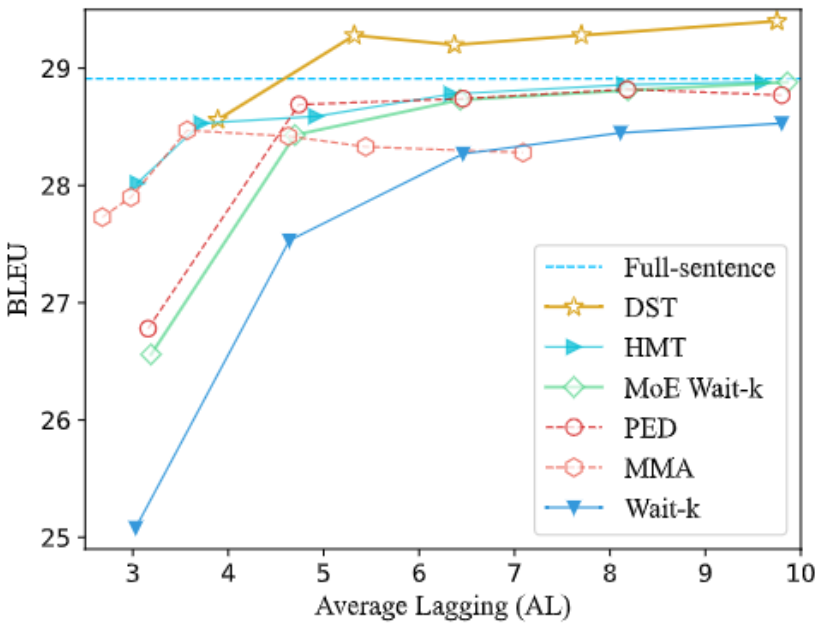
- **推理**：通过累积目标端给源端前缀的注意力，并据此判断策略

$$\sum_{j=1}^{m} p_{i,j} > \delta_{infer}$$

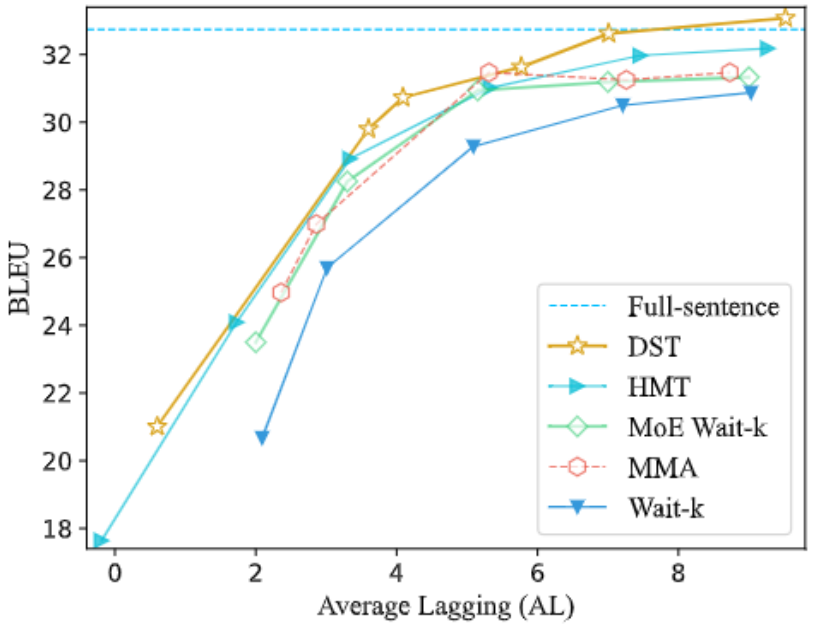
  - δ_{infer} ：决策阈值
  - 注意力足够，则执行 WRITE 动作
  - 否则，执行 READ 动作

主实验结果

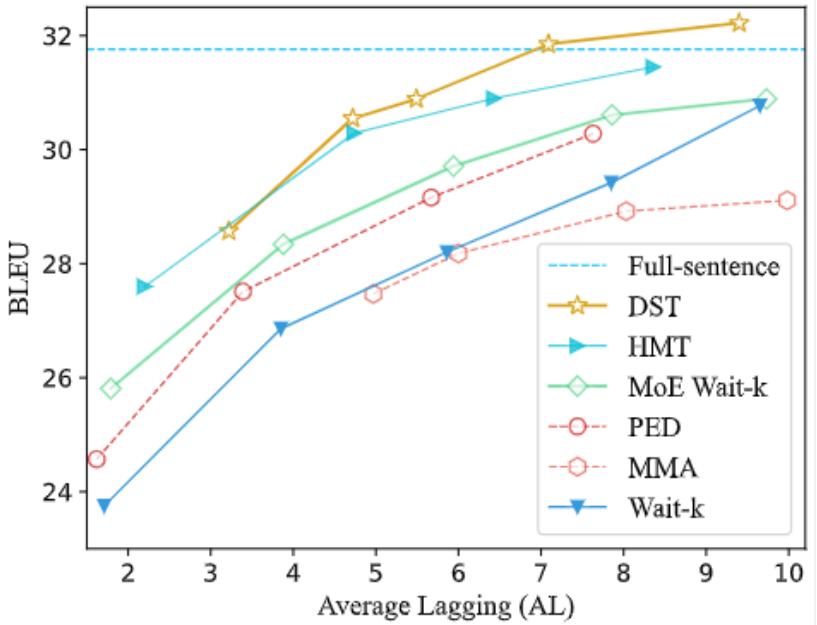
- AL: 平均延时, BLEU: 翻译质量
- 16 layer Decoder-only 结构



(a) En→Vi



(b) En→Ro



(c) De→En

Decoder-only Streaming Transformer (DST)

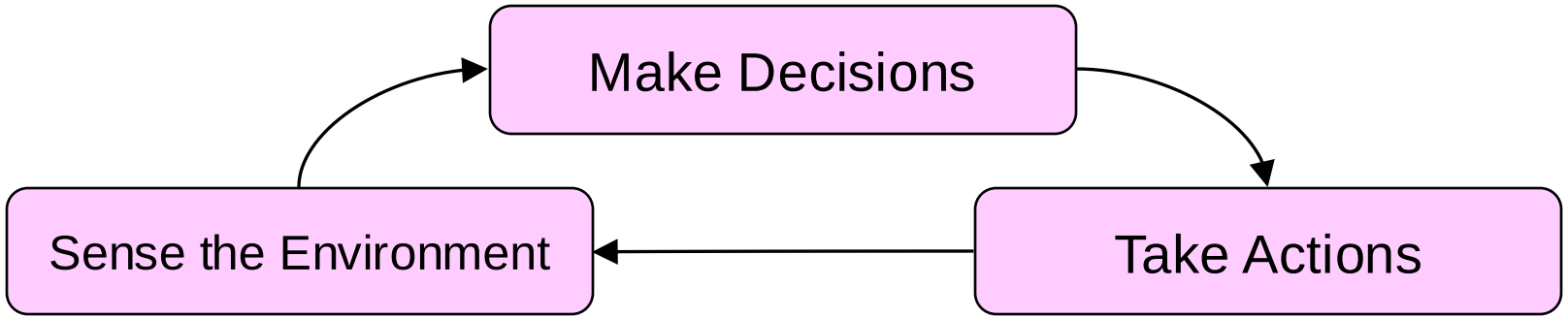
- 结论
  - From Scratch 的同传模型虽擅长决策，但其生成能力仍存在欠缺
  - LLMs虽然具有更好的生成能力，但并不擅长处理流式输入

- 尝试利用Agent工作流的方式结合二者的优势

智能体

An autonomous agent is a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future.

—— **Franklin and Graesser**



Agent-SiMT

Agent-SiMT: Agent-assisted Simultaneous Machine Translation with Large Language Models

Shoutao Guo[1,3], **Shaolei Zhang**[1,3], **Zhengrui Ma**[1,3], **Min Zhang**[4], **Yang Feng**[1,2,3*]
[1]Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS)
[2] Key Laboratory of AI Safety, Chinese Academy of Sciences
[3] University of Chinese Academy of Sciences, Beijing, China
[4] School of Future Science and Engineering, Soochow University





郭守涛　张绍磊　马铮睿　张民　冯洋

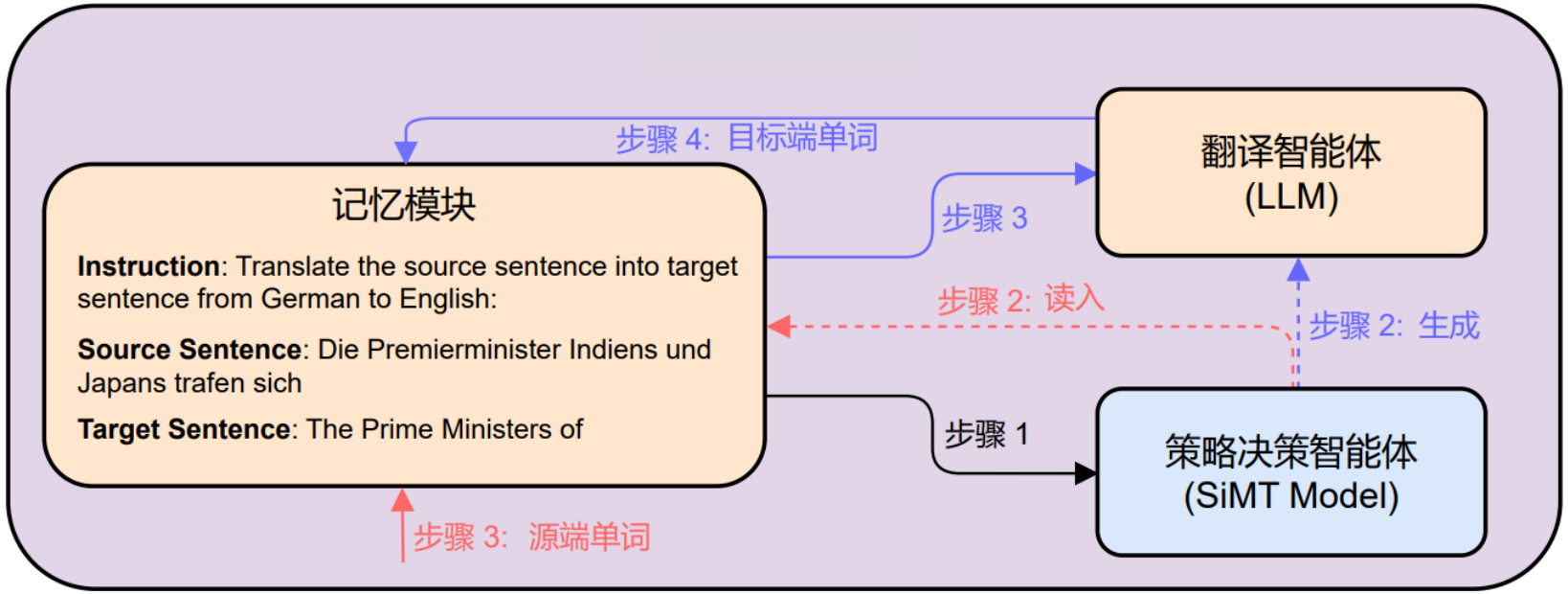
Agent-SiMT

- 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

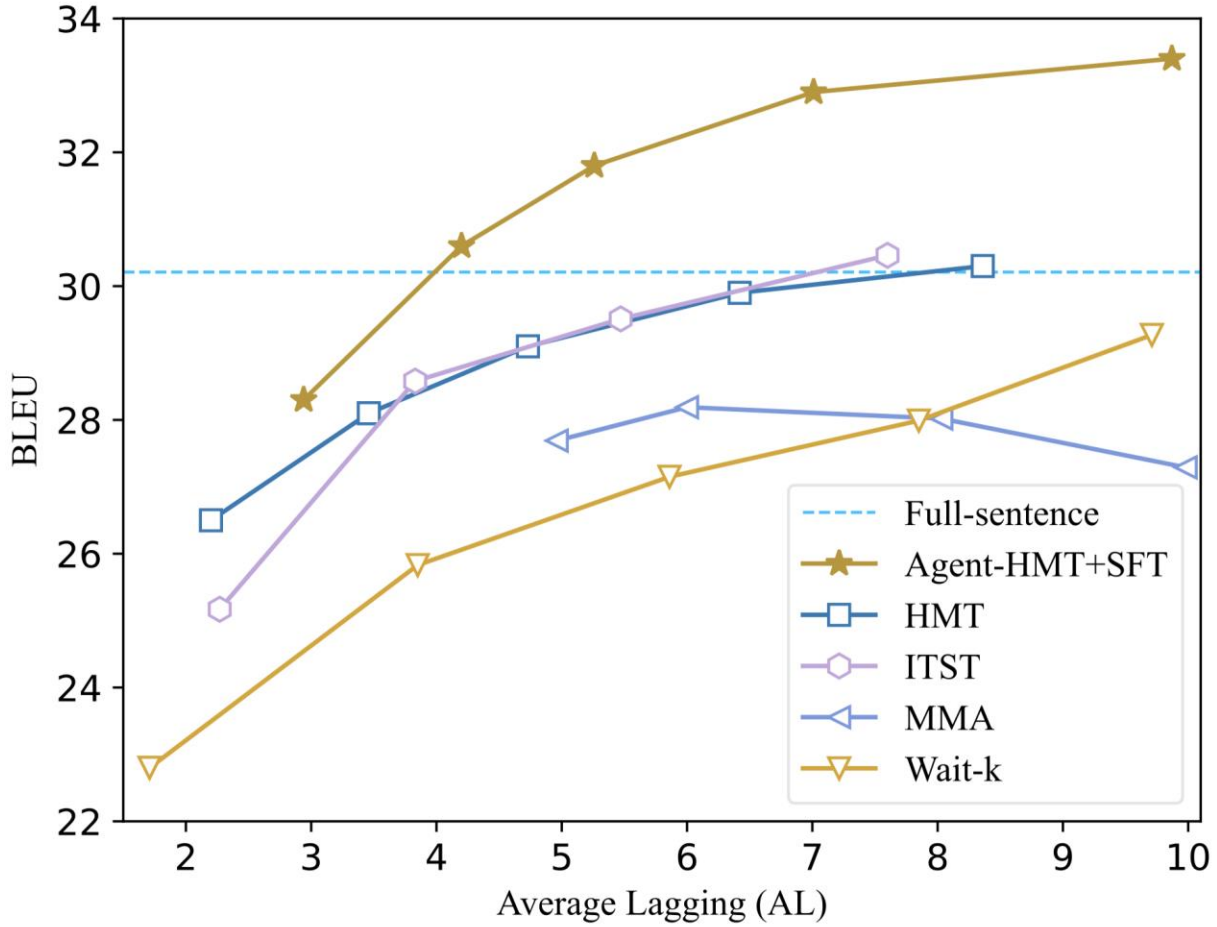
Agent-SiMT

- 策略决策智能体 (SiMT Model)： 擅长决定生成时机，生成能力一般
- 翻译智能体 (LLM): 较强的生成能力，但较难决策生成的时机
- 记忆模块: 存储指令、当前翻译状态



主实验结果

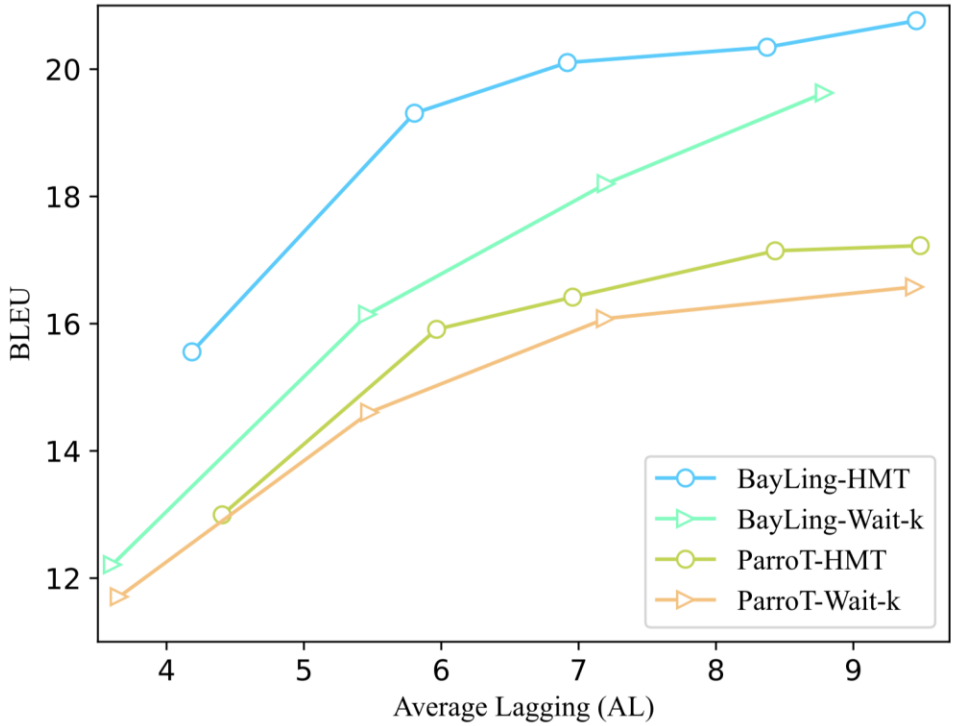
- Llama 2 + HMT
- LLM的生成能力对SIMT翻译质量有明显帮助



不同LLM、SiMT模型

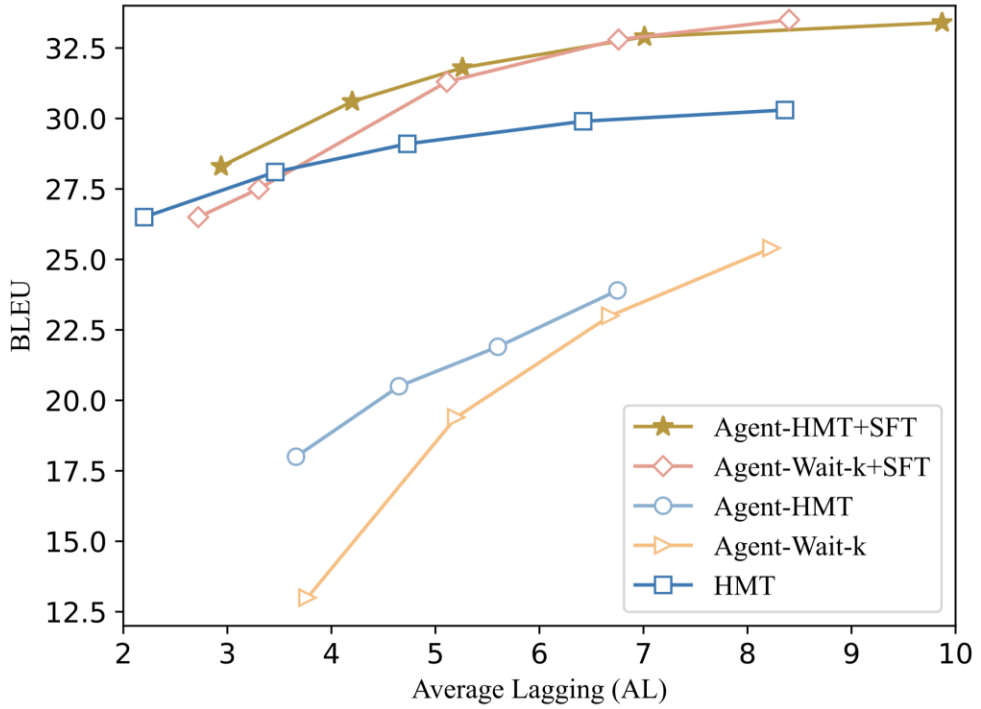
- 充分发挥翻译LLM的能力

翻译LLM：BayLing、ParroT



Zh-En

策略模块：HMT、Wait-k



De-En

总结

- 智能体
  - 智能体大致包含感知、决策和动作模块
  - LLMs的引入使得智能体的决策能力、交互方式都有了较大的提升
  - 合理设计工作流能够充分发挥LLMs的优势，甚至让较弱的LLMs超越更强的LLMs

32

谢 谢 大 家！

张绍磊

中国科学院计算技术研究所

zhangshaolei20z@ict.ac.cn