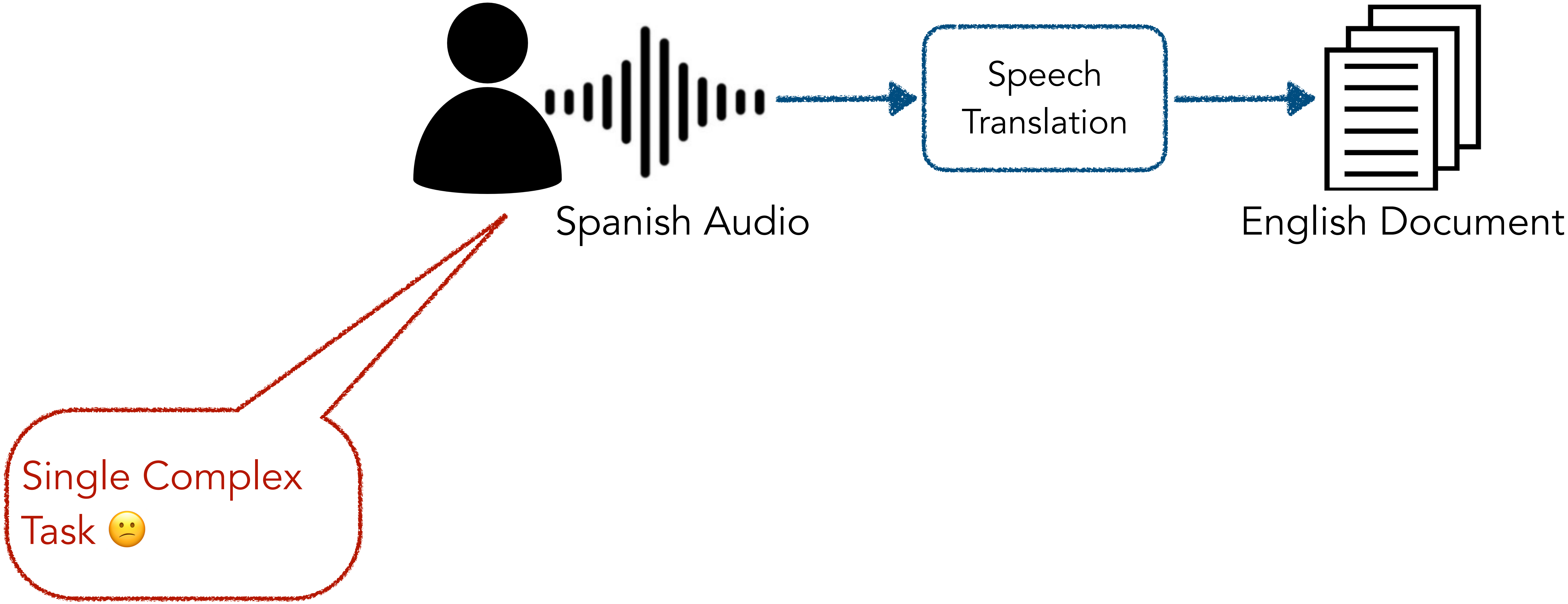
Searchable Hidden Intermediates for End-to-End Models of Decomposable Sequence Tasks

Siddharth Dalmia, Brian Yan, Vikas Raunak, Florian Metze, Shinji Watanabe

Carnegie Mellon University
Language Technologies Institute

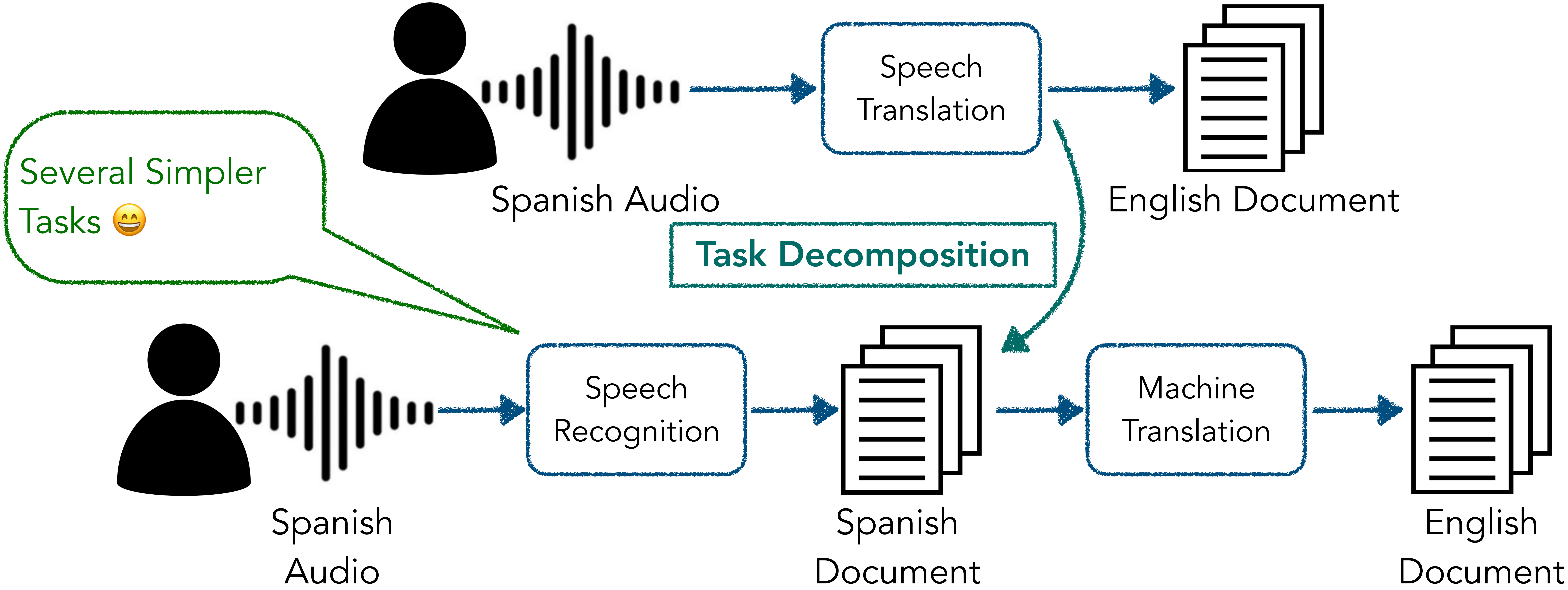
What is Compositionality?

- Compositionality is the principle behind building complex systems by composing together simpler sub-systems.



What is Compositionality?

- Compositionality is the principle behind building complex systems by composing together simpler sub-systems.



Traditional Cascaded Models

- Traditional Cascaded Models exploited the task compositionality to give many interesting properties that facilitate practicality of these models.

    1. The strong search capabilities to compose the final task output from individual system predictions.

    2. The ability to incorporate external models to re-score each individual system.

    3. The ability to easily adapt individual components towards out-of-domain data

    4. The ability to monitor performance of the individual systems towards the decomposed sub-task.

Dalmia et. al., 2021

Can we bring these properties into End-to-End Models?

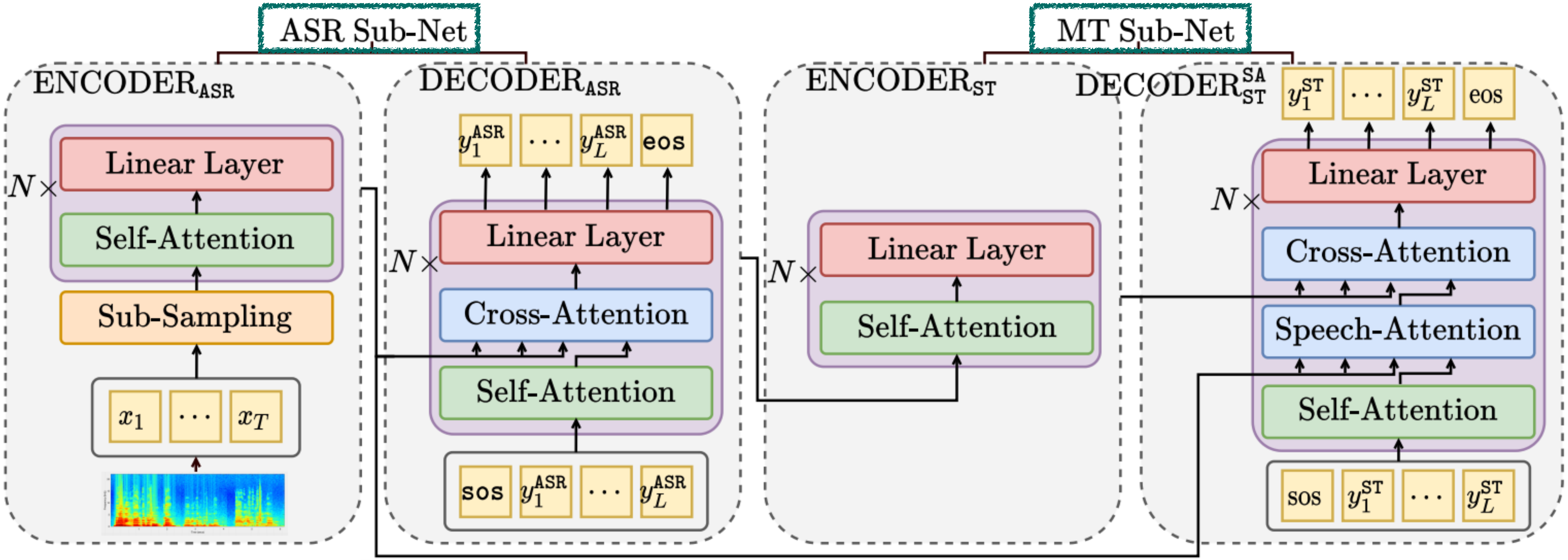
Searchable Hidden Intermediates Framework

- General end-to-end framework to exploit natural decomposition in sequence tasks.
  - A sequence task, A → C is decomposable, if there is an intermediate sequence B for which A → B sequence transduction followed by B → C prediction achieves the original task.
    - For instance, Speech Translation using ASR intermediates
- Learn $P(\mathrm{C} \mid \mathrm{A})$ through $\max_{\mathrm{B}}(P(\mathrm{C} \mid \mathrm{A}, \mathrm{B})P(\mathrm{B} \mid \mathrm{A}))$, approximated using Viterbi search.

Dalmia et. al., 2021

Multi-Decoder Model with Searchable Intermediates

(Completed Work)

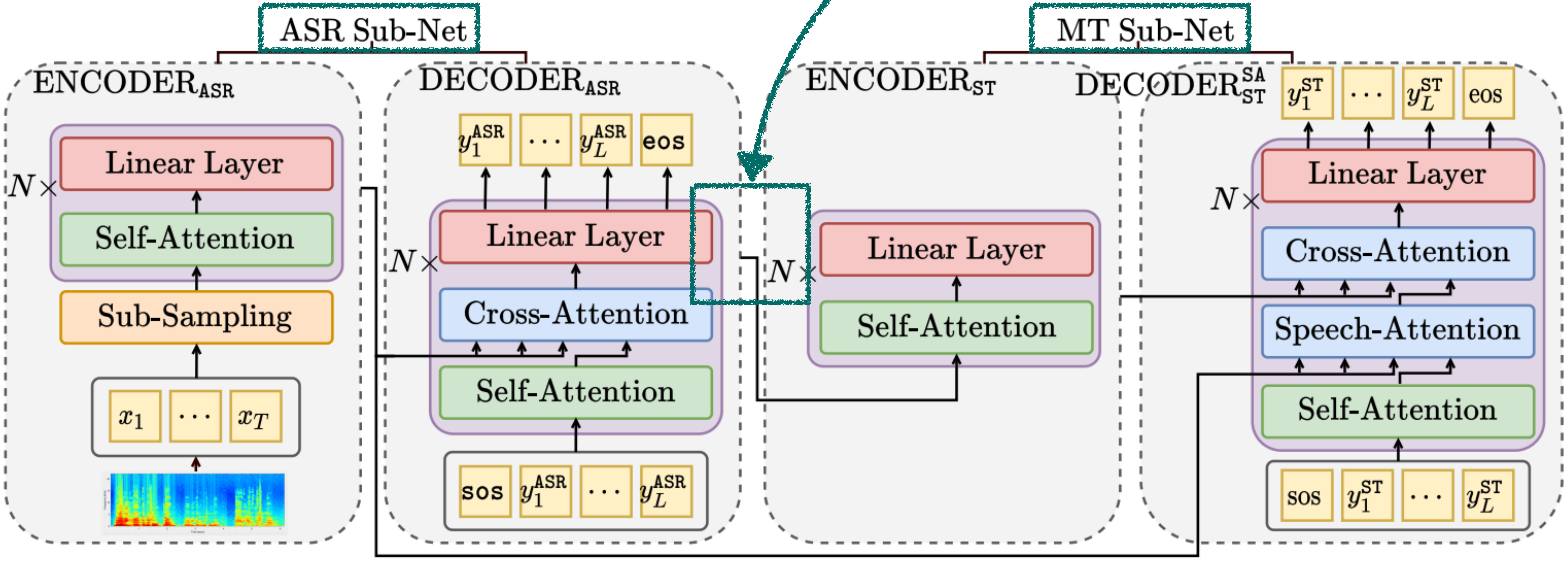


Multi-Decoder Model with Searchable Intermediates

(Completed Work)

Pass Decoder Hidden Representations:

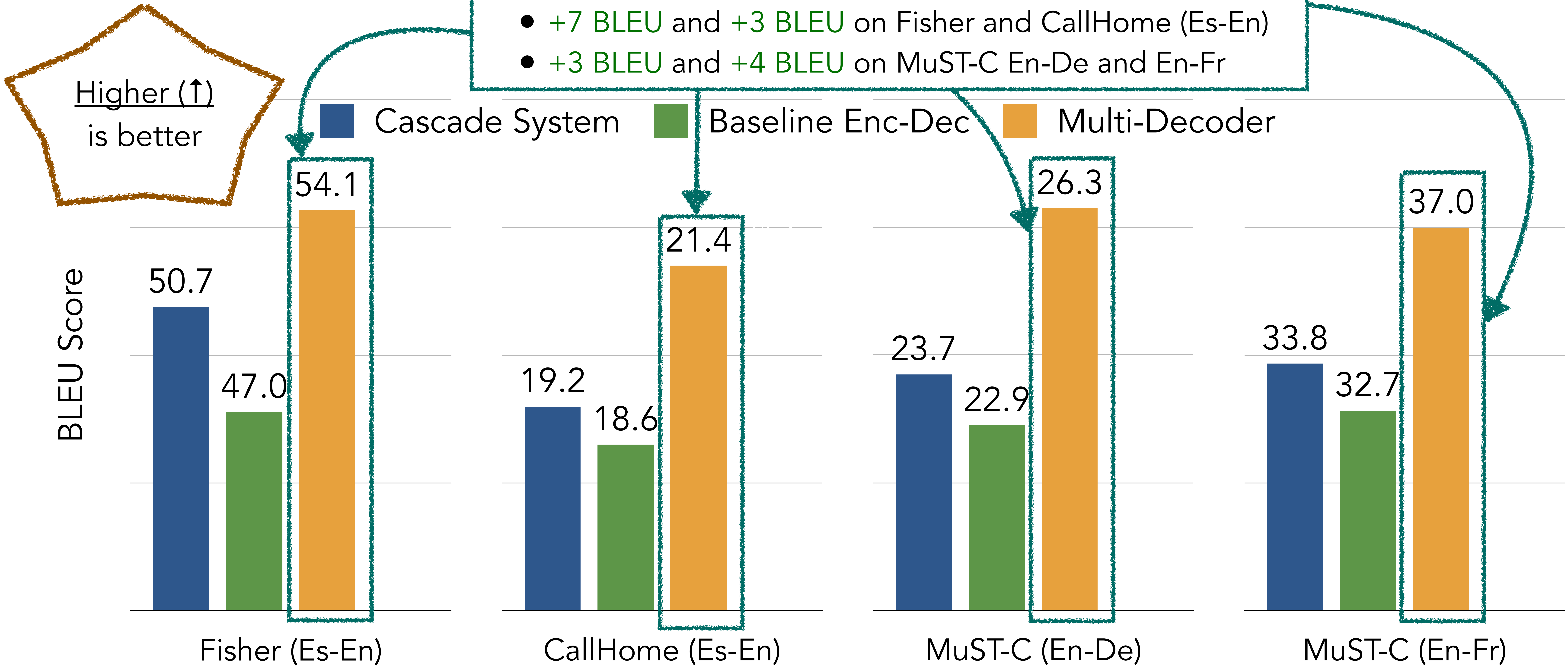
- ASR Sub-Net maps input to sequence of decoder hidden representations $\mathbf{h}^{D_B}$
- MT Sub-Net maps $\mathbf{h}^{D_B}$ to final ST output
- During inference, approximate $\mathbf{h}^{D_B}$ with $\mathbf{h}^{D_B}_{\text{Beam}}$



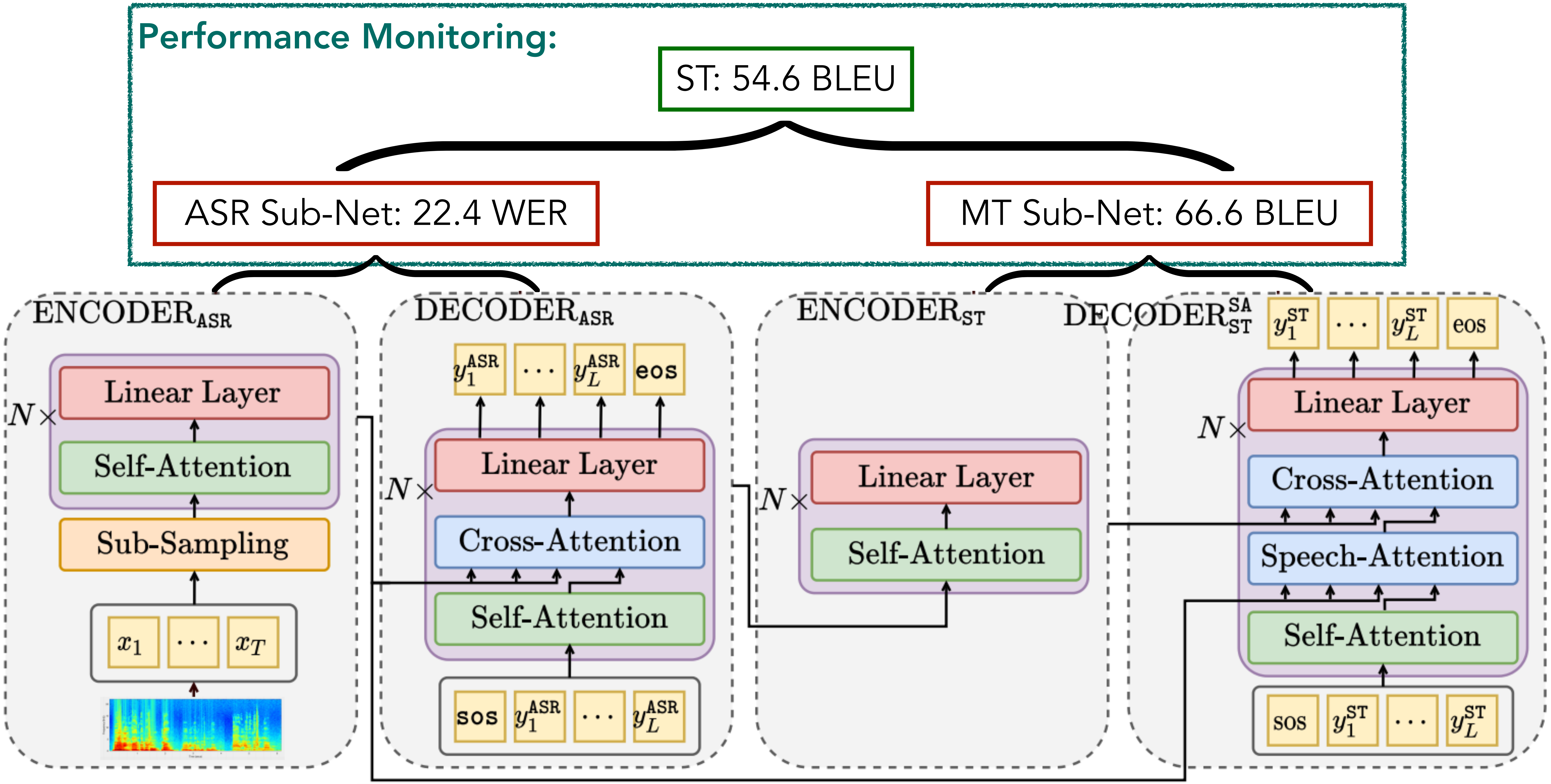
Comparison with Encoder-Decoder

Outperforms both Encoder-Decoder and Cascaded Models -

- +7 BLEU and +3 BLEU on Fisher and CallHome (Es-En)
- +3 BLEU and +4 BLEU on MuST-C En-De and En-Fr

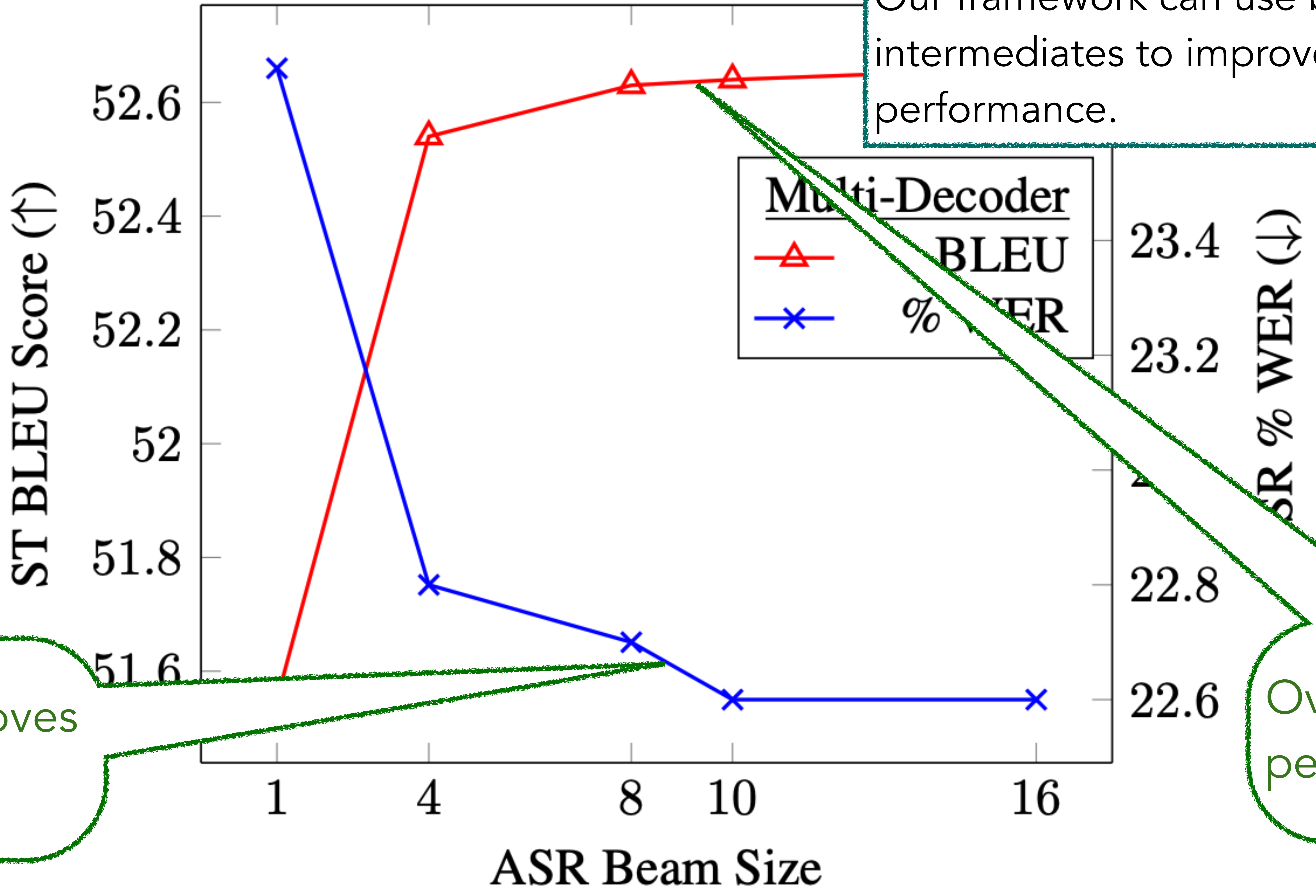


Performance Monitoring



Retrieval with Beam Search

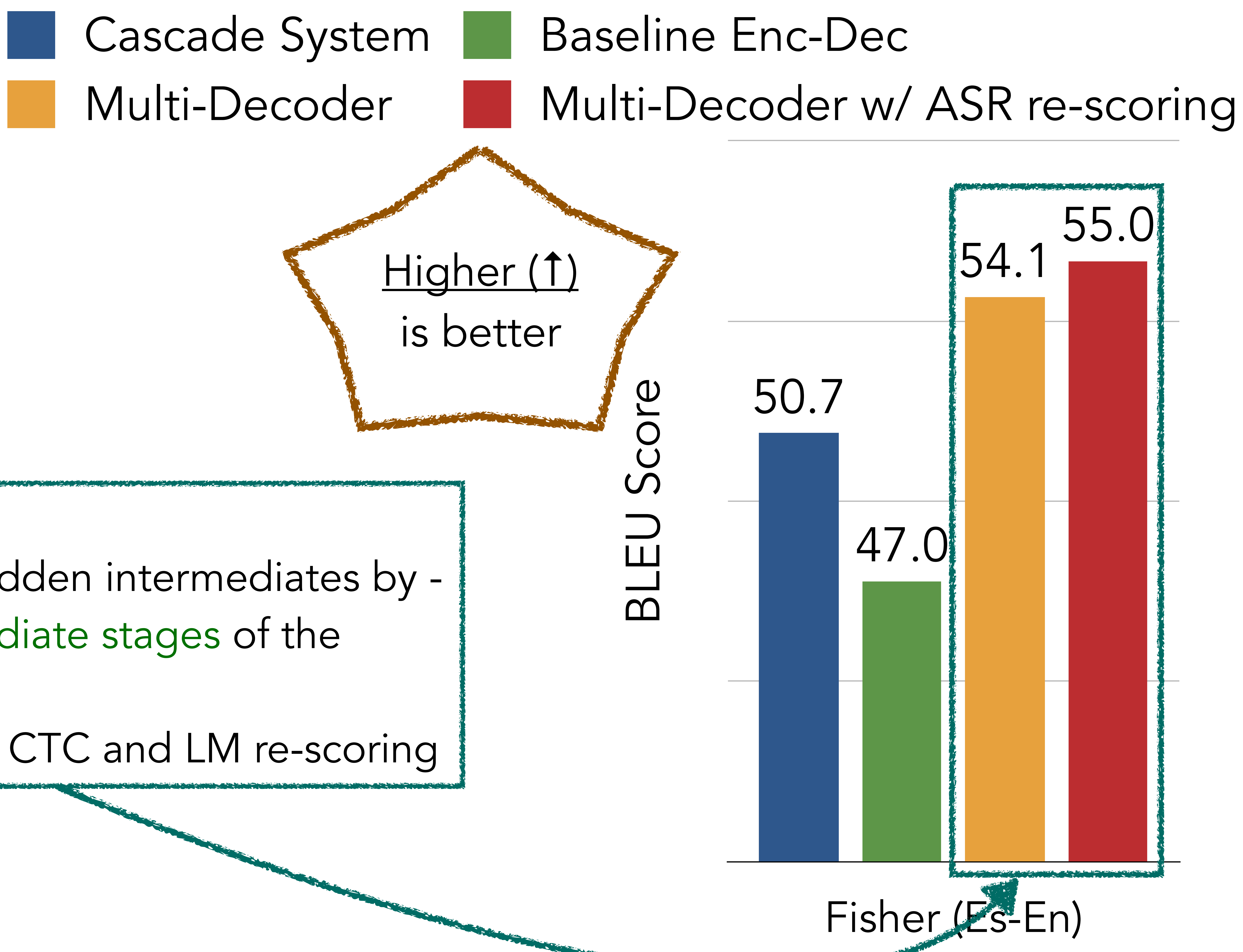
Search and Retrieval:
Our framework can use beam search at ASR intermediates to improve the overall ST performance.



As ASR quality improves with larger beam

Overall ST performance goes up!

Retrieval with External Models



Search and Retrieval:
Our framework has the ability to retrieve better hidden intermediates by -
- Re-scoring using external models at intermediate stages of the network during inference.
- On Fisher Es-En improves by +1 BLEU using CTC and LM re-scoring

Adapting Individual Components

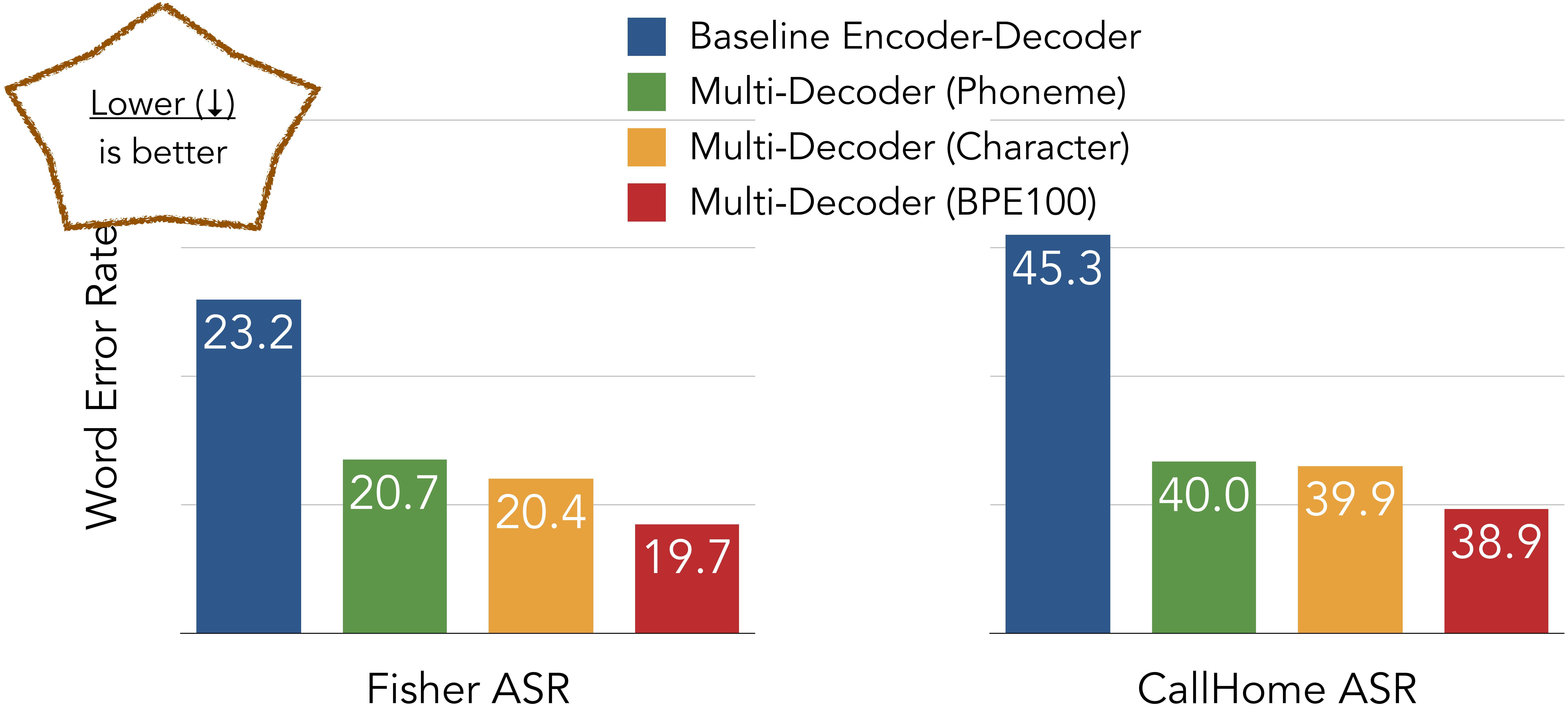
Search and Retrieval:

Our framework has the ability to adapt individual components of the E2E model towards out-of-domain data.

- We can re-score ASR sub-net with in-domain LM.
- Improves ASR by 10% lower WER, improving the overall ST by +2.4 BLEU

Model	Overall ST(↑)	Sub-Net ASR(↓)
IN-DOMAIN ST MODEL		
Baseline (Wang et al., 2020b)	12.0	-
OUT-OF-DOMAIN ST MODEL		
Multi-Decoder	12.6	46.5
+ASR Re-scoring w/ in-domain LM	**15.0**	**36.7**

Decomposing Speech Transcripts



Thank you