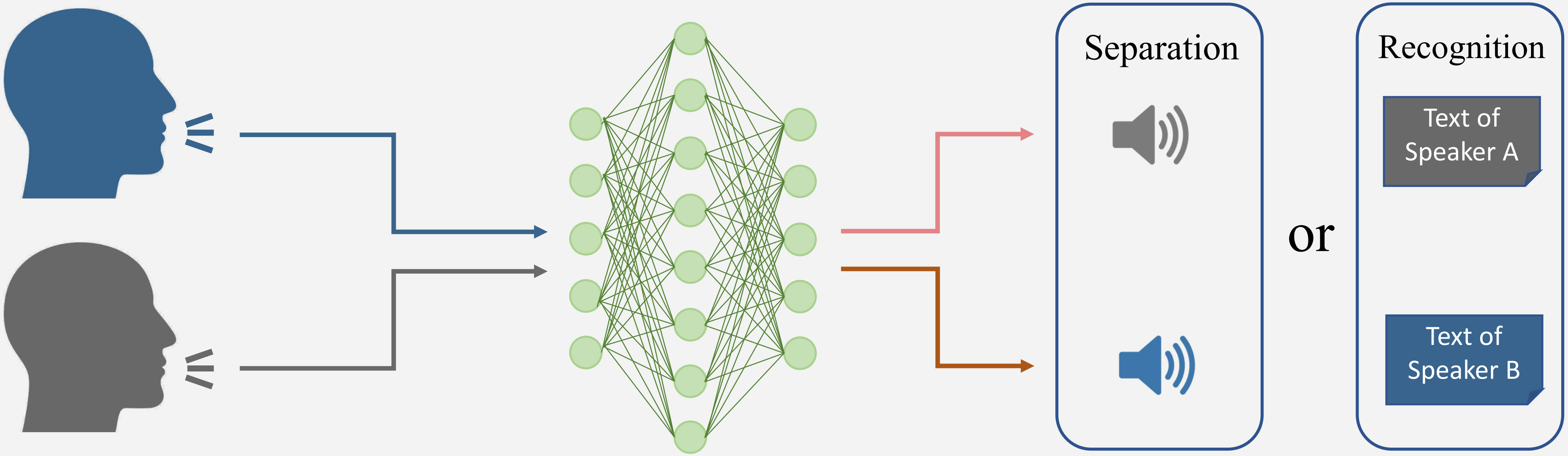
End-to-End Multi-speaker Speech Recognition with Transformer

Xuankai Chang[1], Wangyou Zhang[2],
Yanmin Qian[2], Jonathan LeRoux[3], Shinji Watanabe[1]
[1]Center for Language and Speech Processing, Johns Hopkins University, USA
[2]SpeechLab, Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
[3]Mitsubishi Electric Research Laboratories (MERL), USA

Background

- Multi-speaker speech processing (Cocktail party problem)



End-to-End is attractive

✓ **No need** for parallel clean audios

End-to-End is attractive

- ✓ **No need** for parallel clean audios
- ✓ Simplifying the complicated model-building

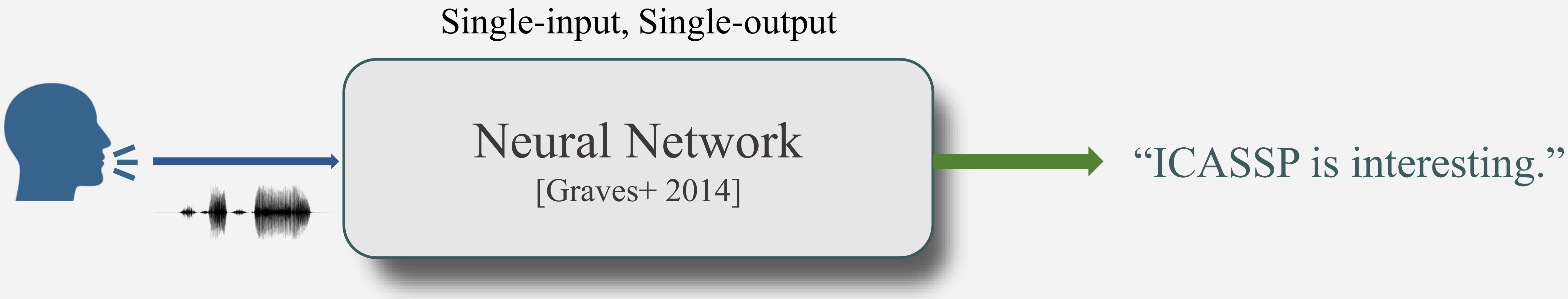
End-to-End is attractive

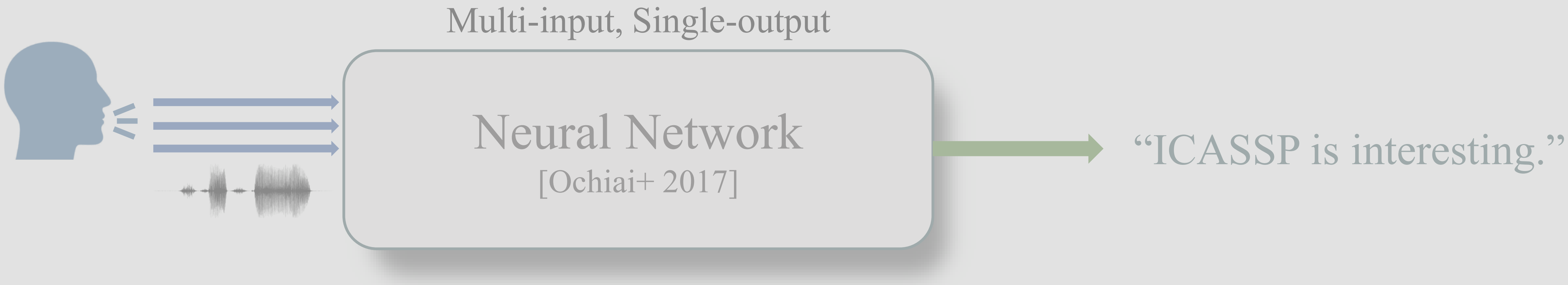
✓ **No need** for parallel clean audios

✓ Simplifying the complicated model-building

✓ Natural incorporation with Linguistic Information

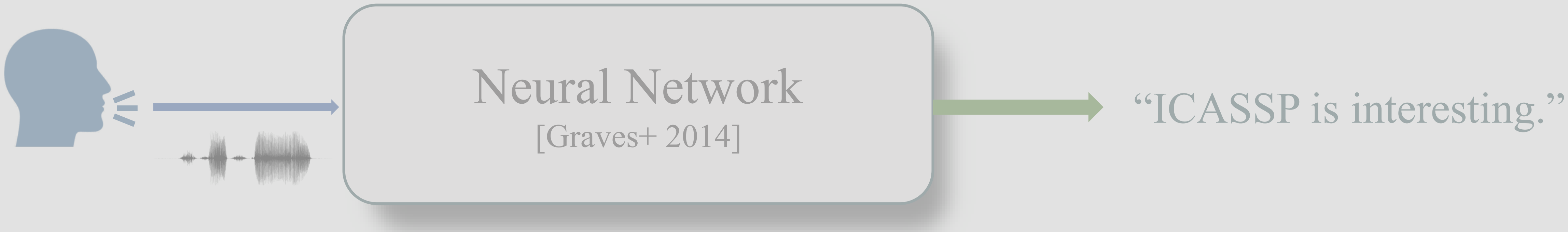
End-to-End speech recognition





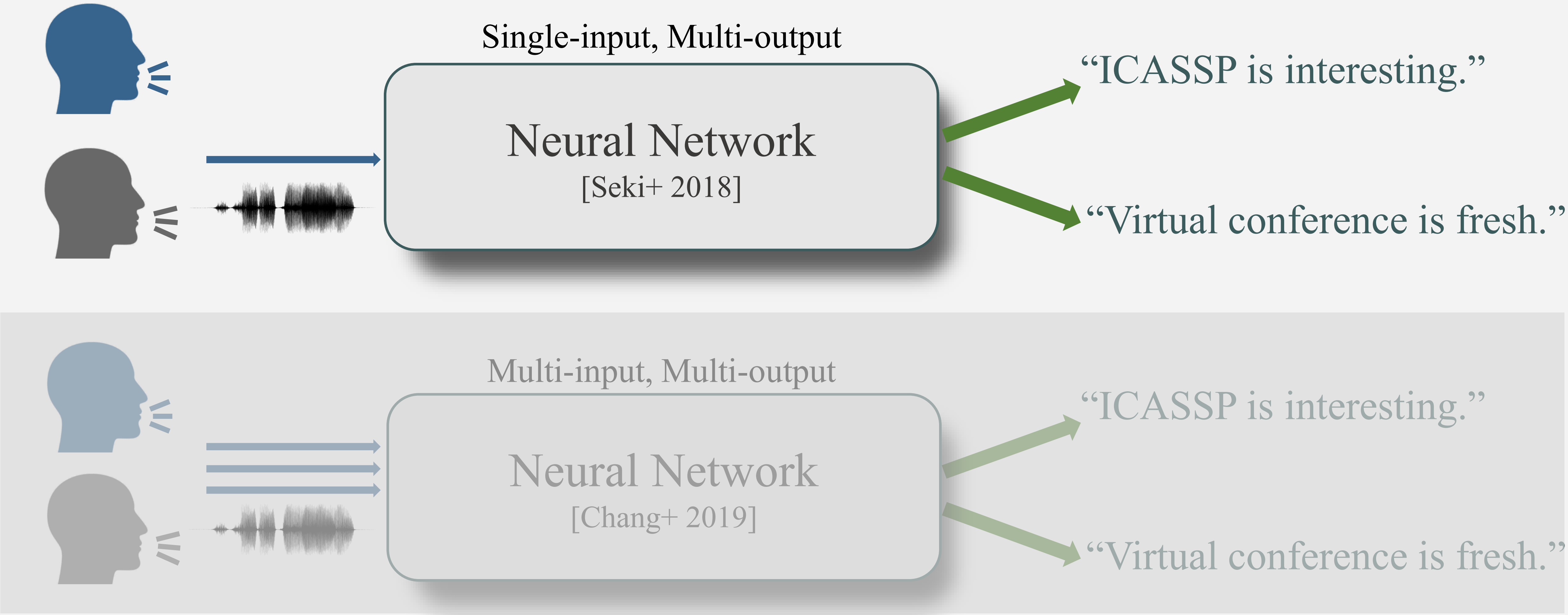
End-to-End speech recognition

Single-input, Single-output

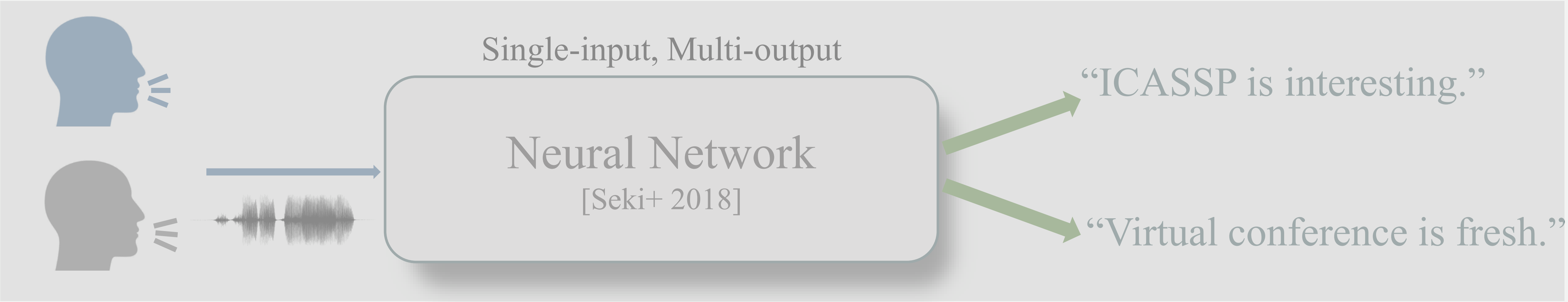


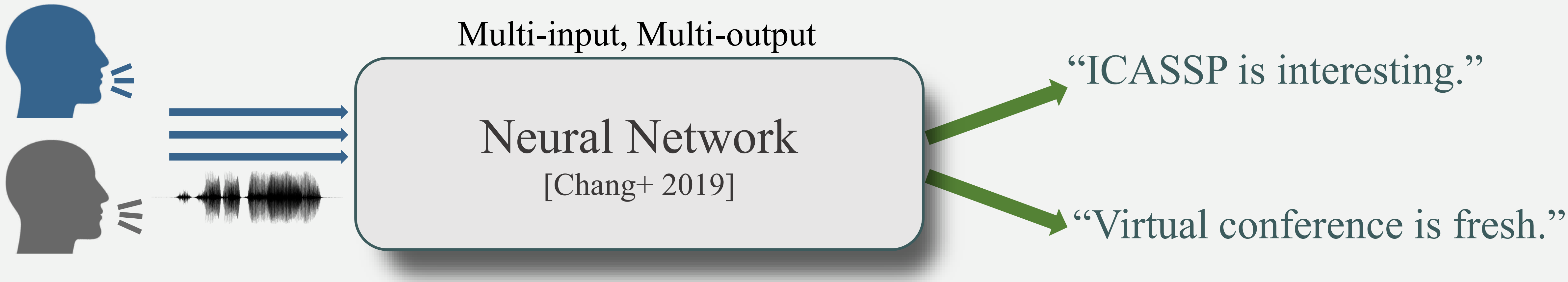
Multi-input, Single-output

End-to-End speech recognition

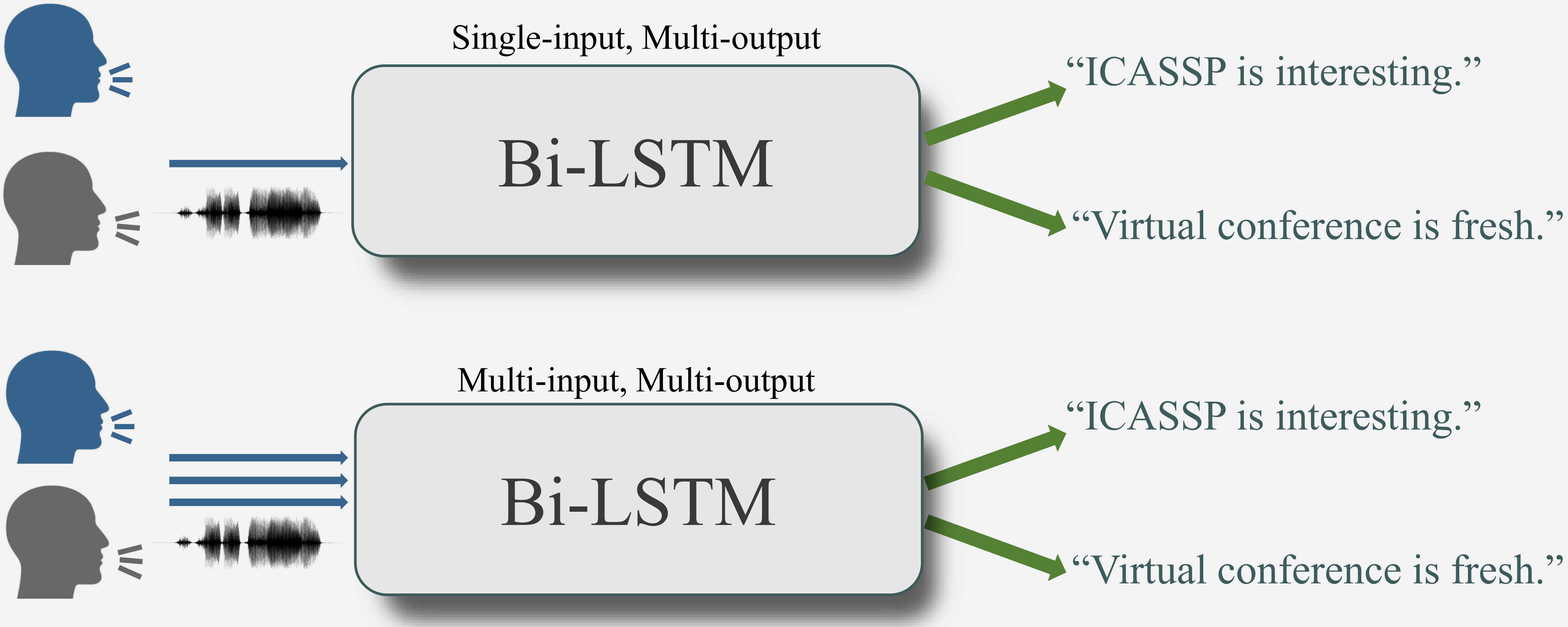


End-to-End speech recognition

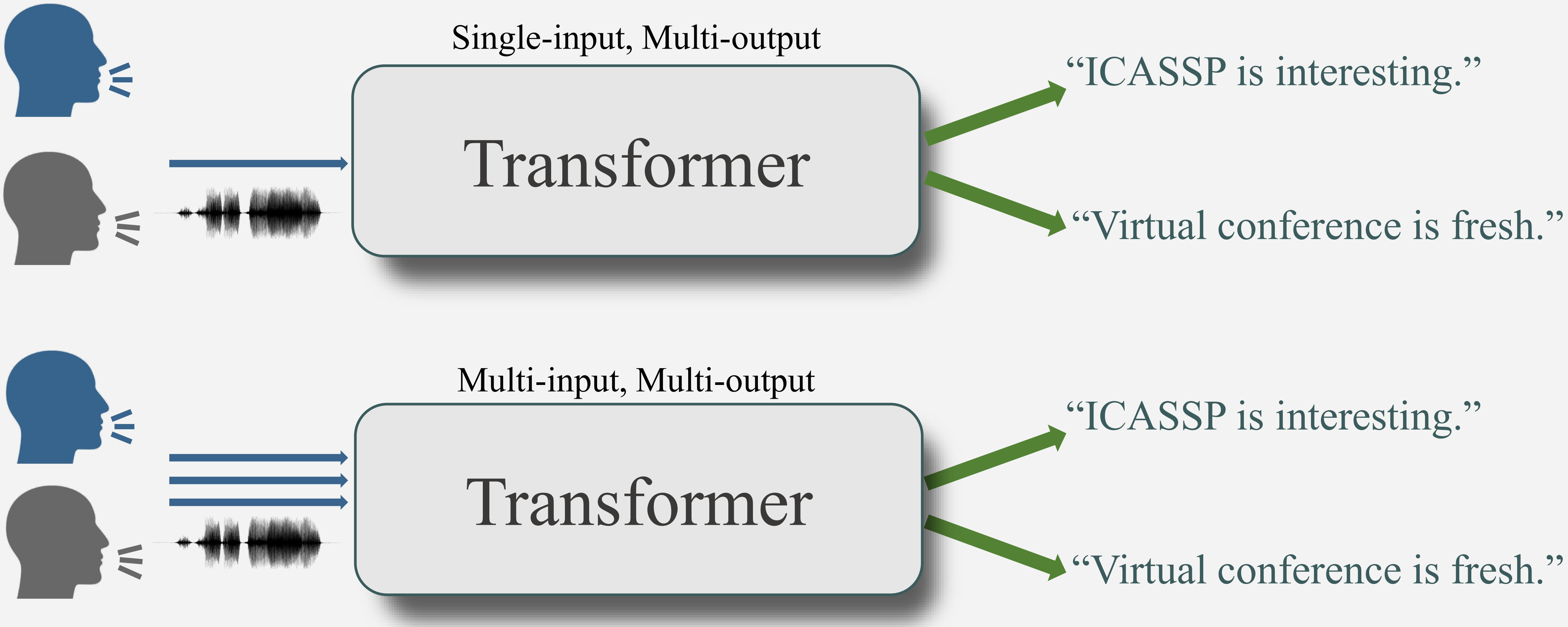




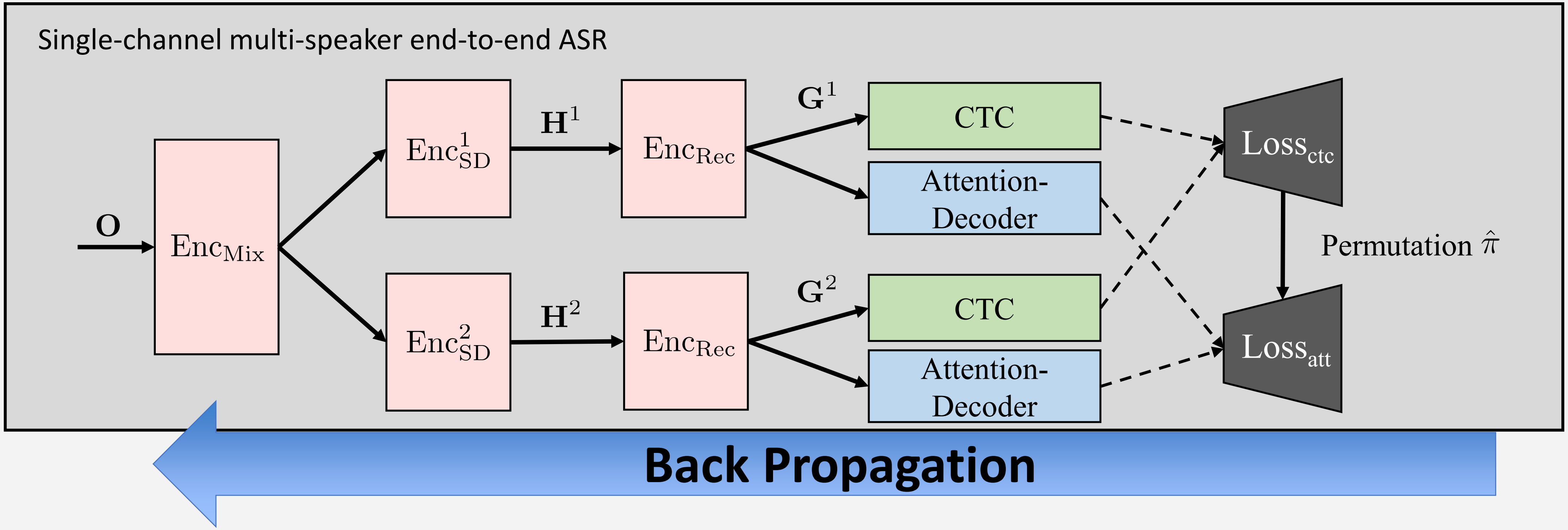
End-to-End speech recognition



End-to-End speech recognition



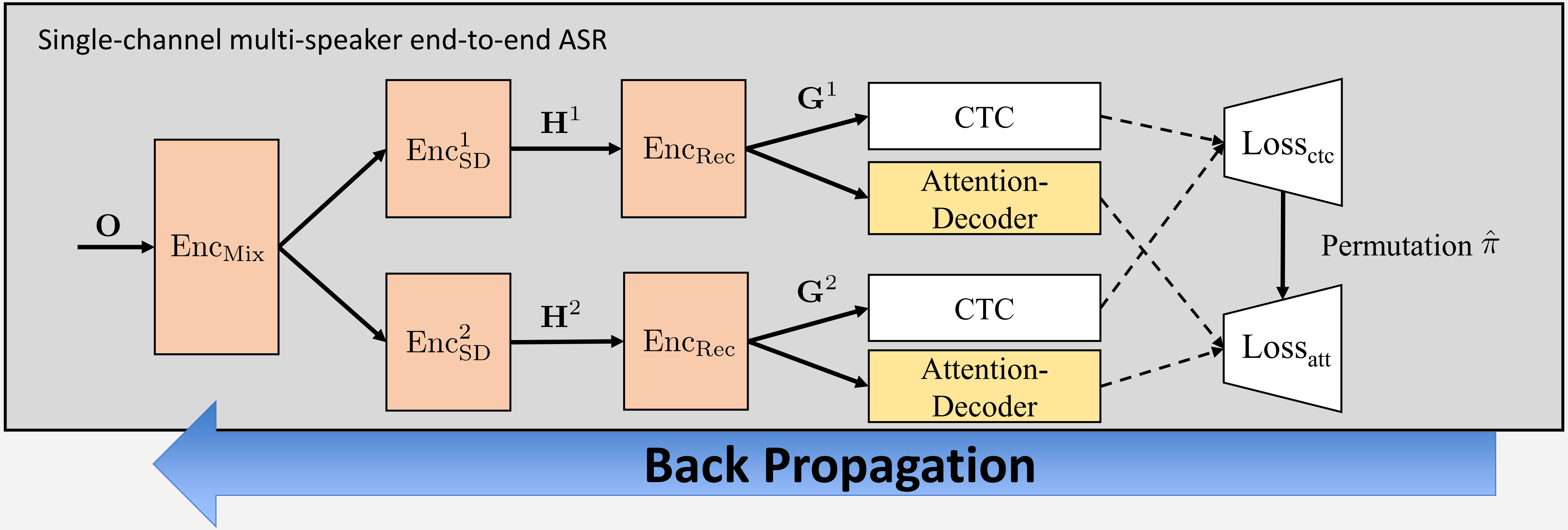
Single-Channel Multi-Speaker E2E ASR



1. **Encoder**: separating and encoding as high dimensional representation
2. **Decoder**: generating the output token sequence
3. **CTC** : determining the permutation of reference sequences

Hiroshi Seki, et al. "A purely end-to-end system for multi-speaker speech recognition", ACL, 2018

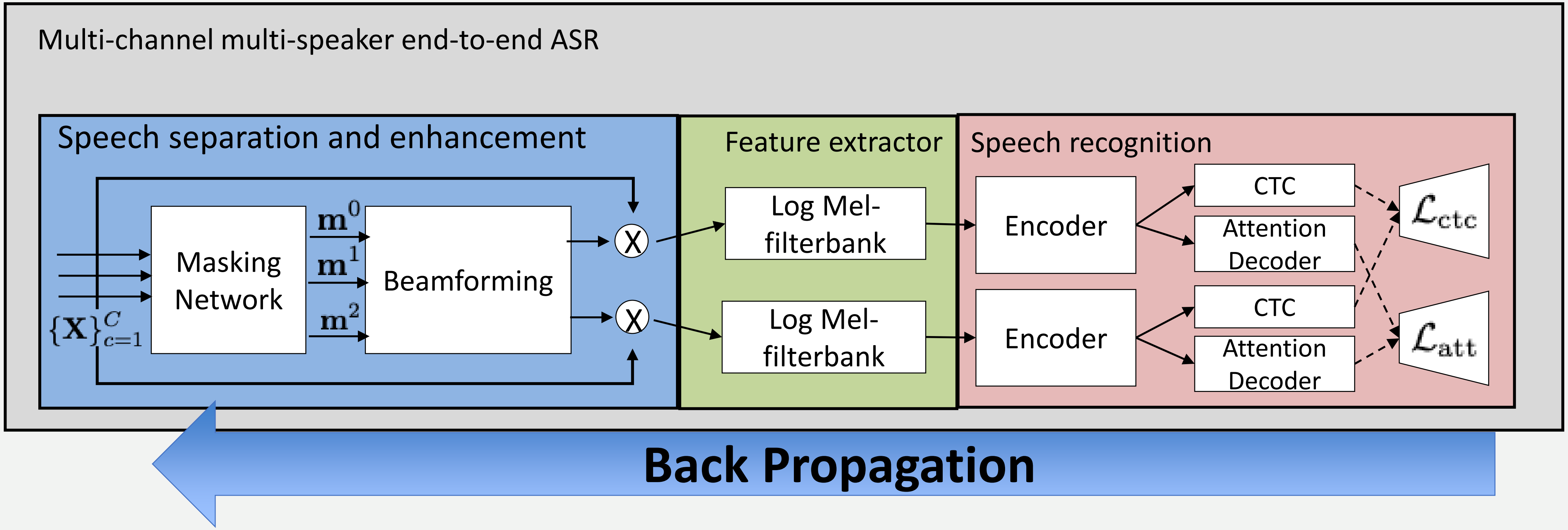
Single-Channel Multi-Speaker E2E ASR



1. **Encoder**: separating and encoding as high dimensional representation
2. **Decoder**: generating the output token sequence
3. **CTC** : determining the permutation of reference sequences

Hiroshi Seki, et al. "A purely end-to-end system for multi-speaker speech recognition", ACL, 2018

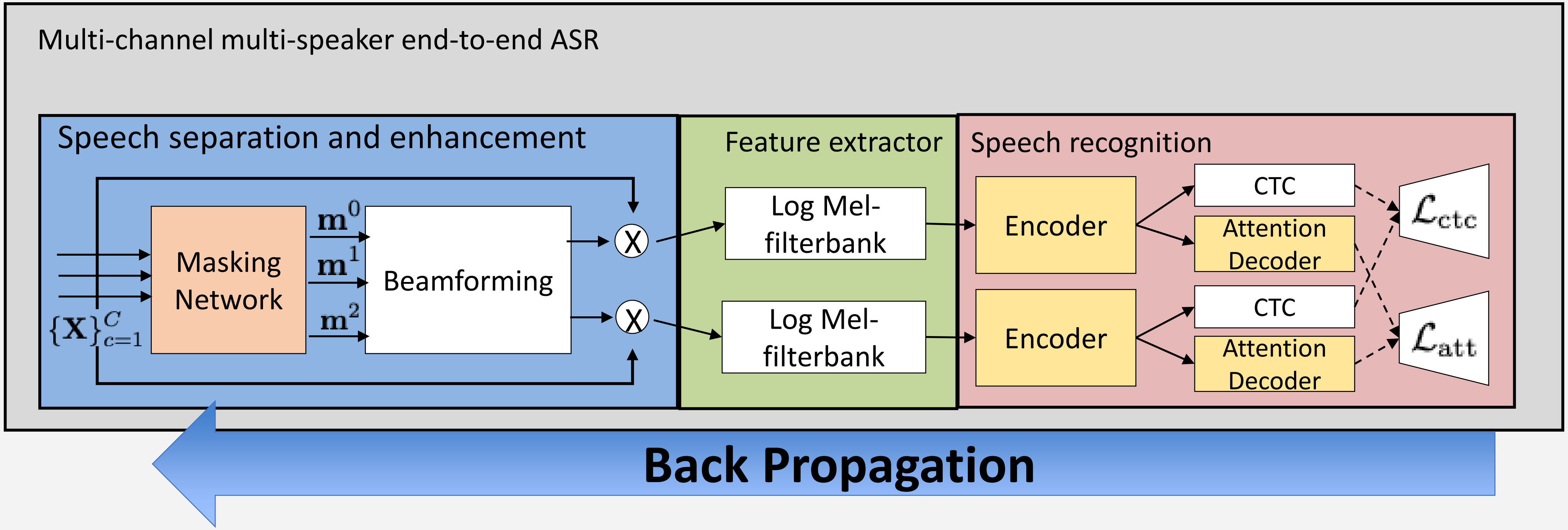
Multi-Channel Multi-Speaker ASR



1. **Speech separation: Multi-source mask-based neural beamformer**
2. **Feature extraction: STFT → Log Mel-filterbank**
3. **Speech recognition: Joint CTC/attention-based encoder-decoder**

Xuankai Chang, et al. "MIMO-Speech: End-to-end multi-channel multi-speaker speech recognition", ASRU, 2019

Multi-Channel Multi-Speaker ASR

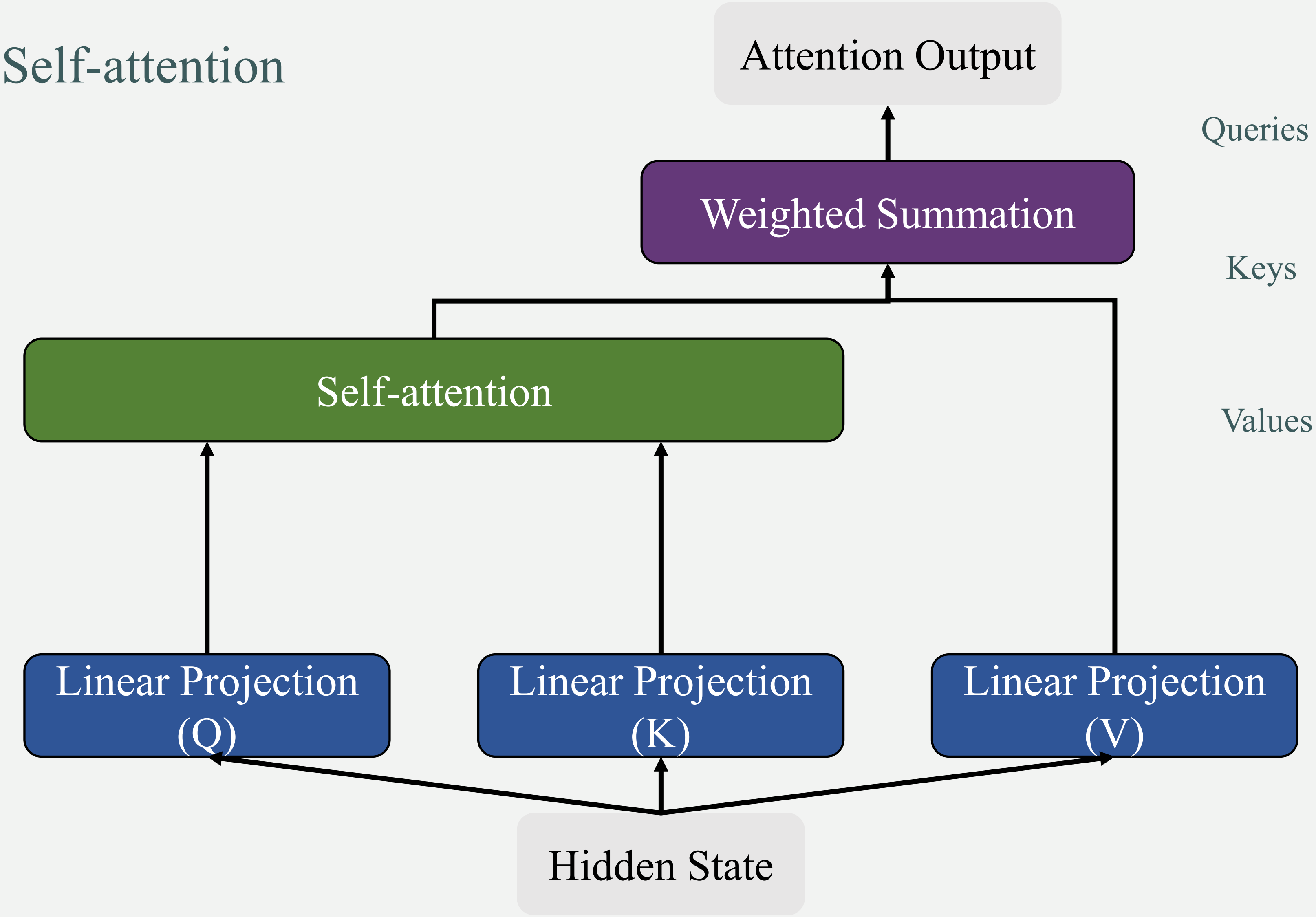


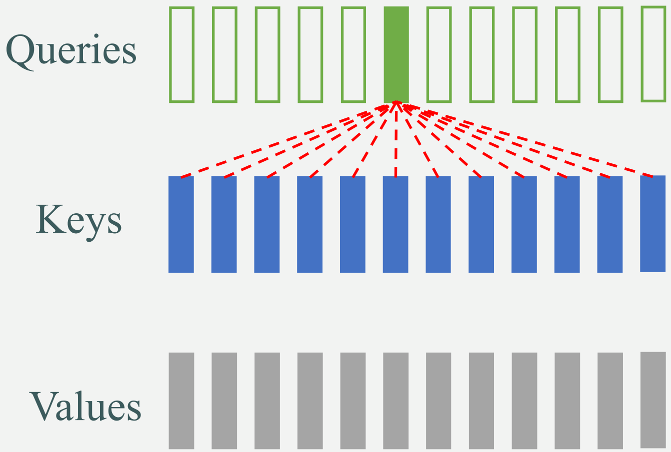
1. **Speech separation: Multi-source mask-based neural beamformer**
2. **Feature extraction: STFT → Log Mel-filterbank**
3. **Speech recognition: Joint CTC/attention-based encoder-decoder**

Xuankai Chang, et al. "MIMO-Speech: End-to-end multi-channel multi-speaker speech recognition", ASRU, 2019

Transformer (Self-attention)

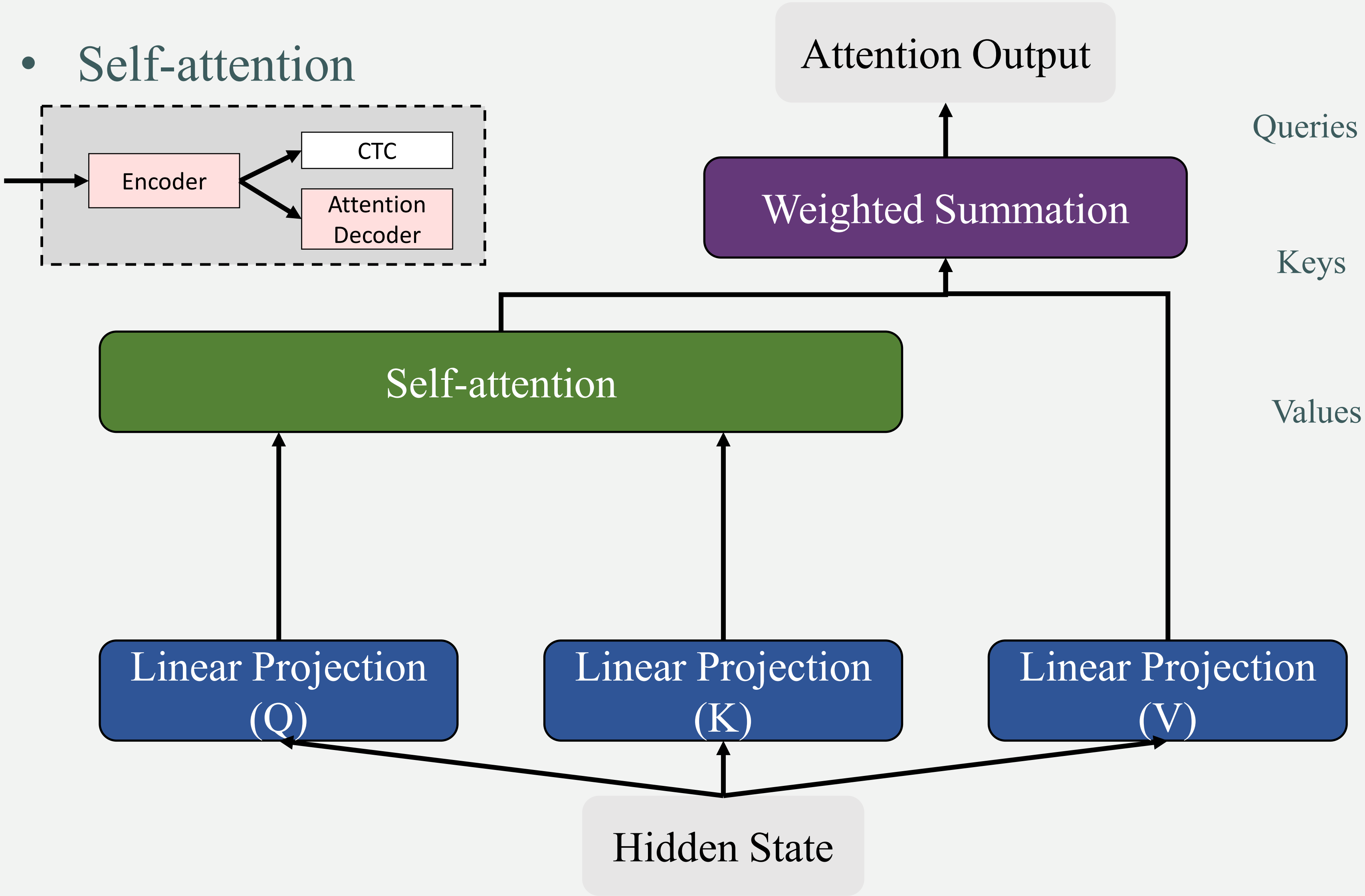
- Self-attention

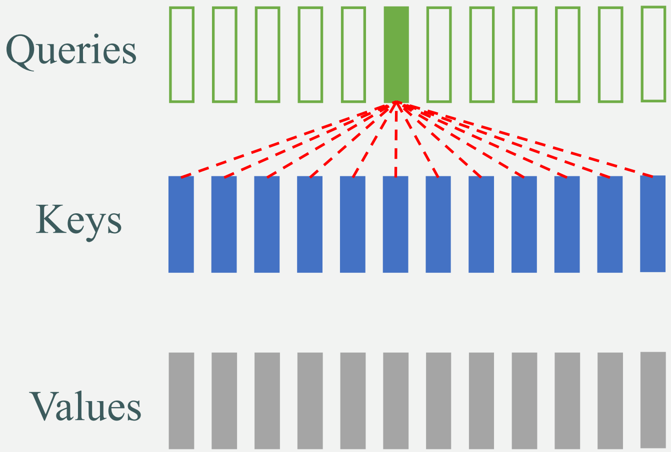




Transformer (Self-attention)

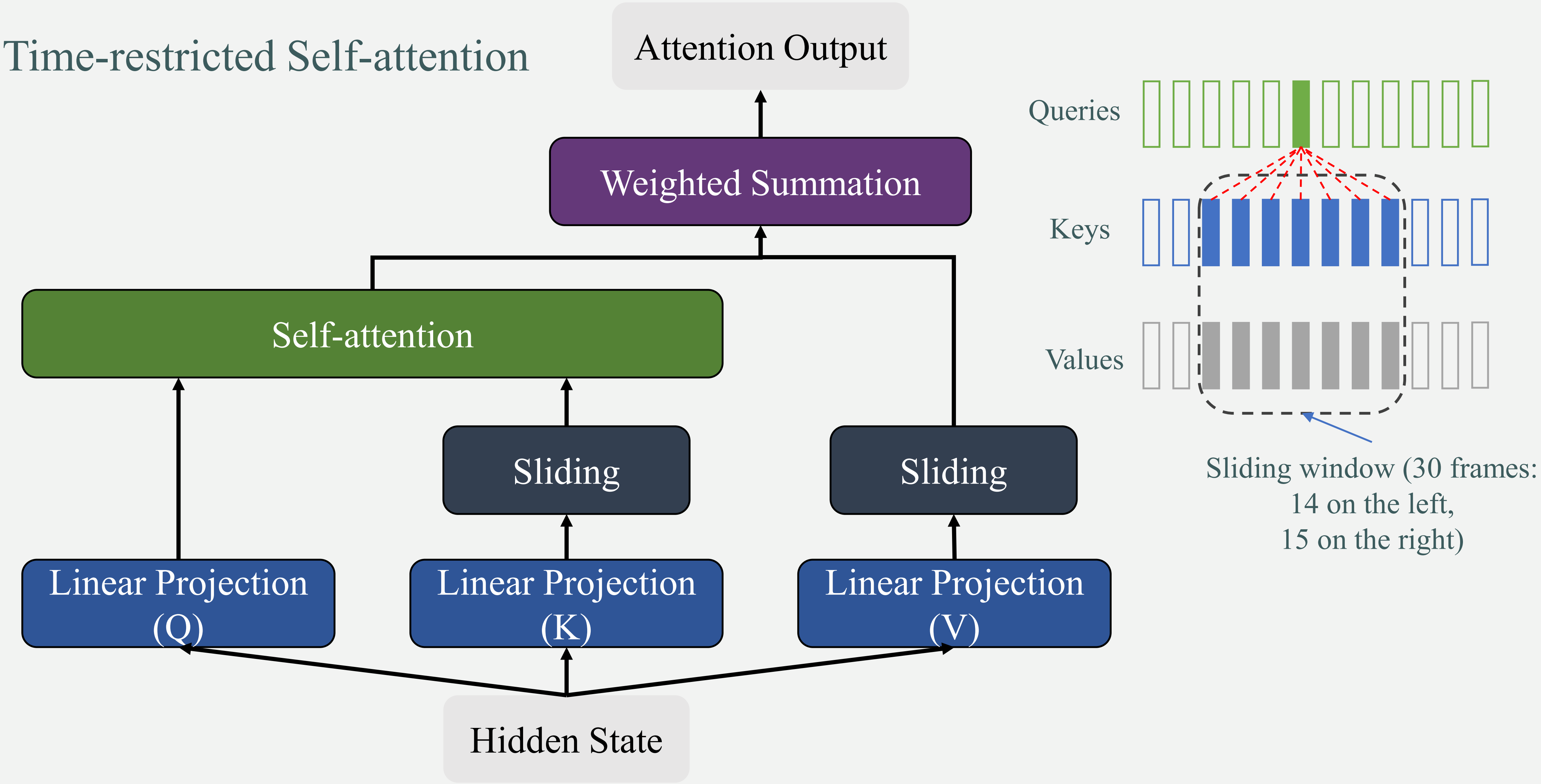
- Self-attention





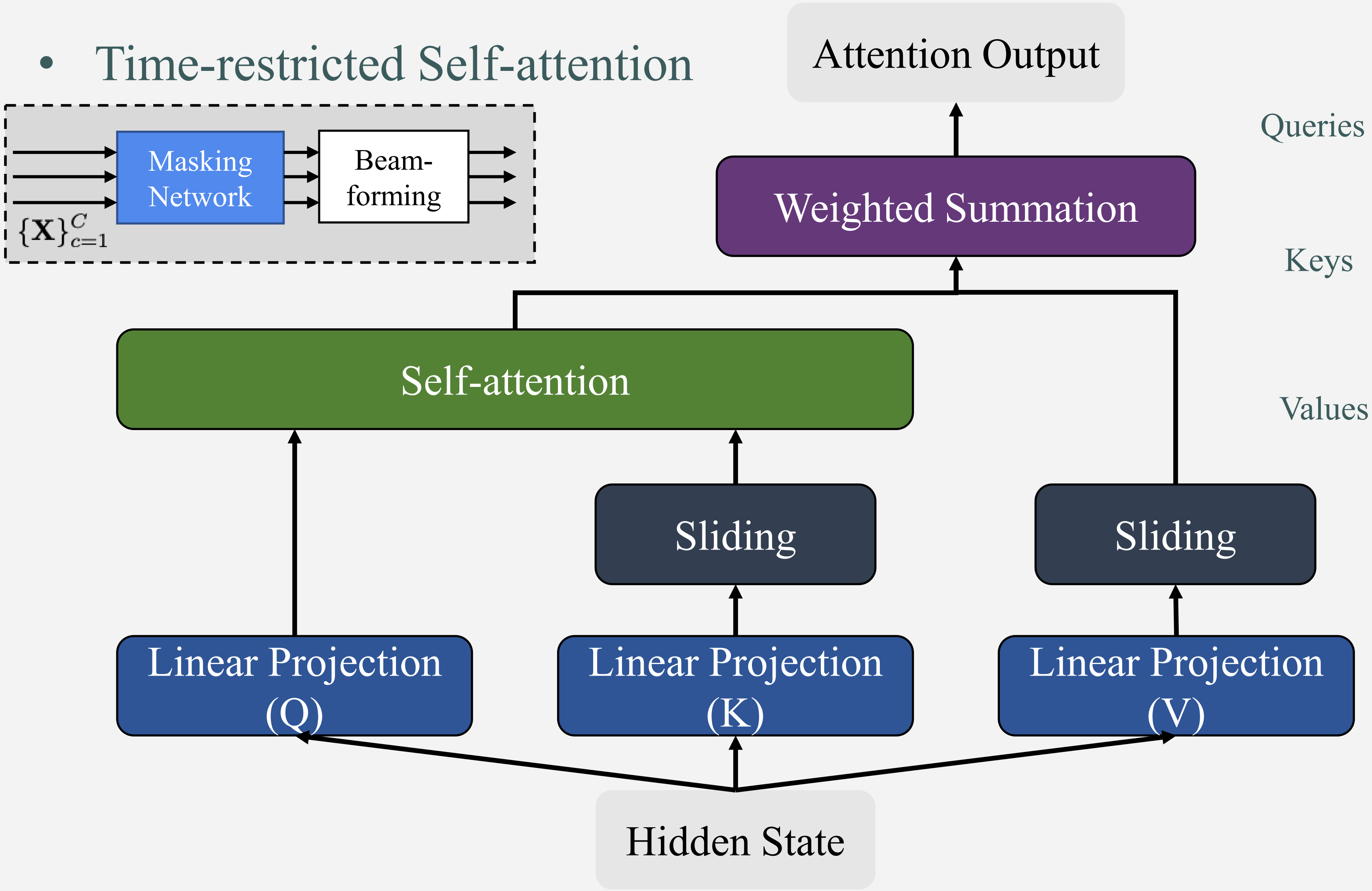
Transformer (Self-attention)

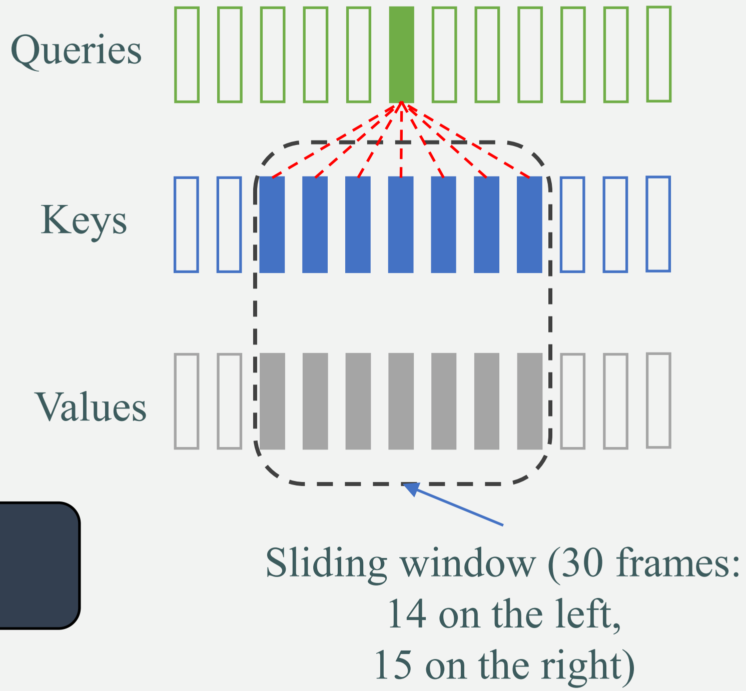
- Time-restricted Self-attention



Transformer (Self-attention)

- Time-restricted Self-attention





Experiment – Data

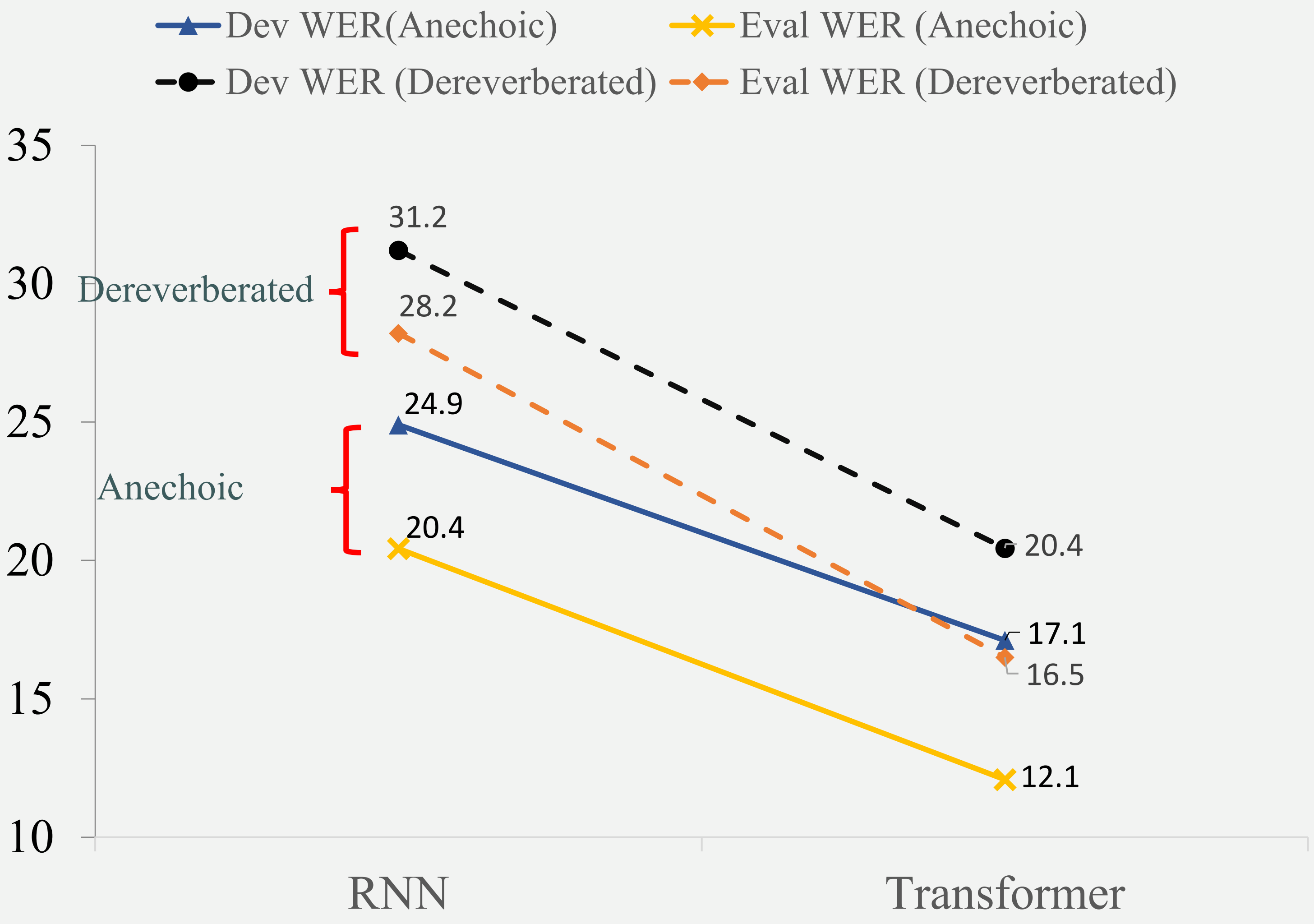
Data	Name	Note
Single-channel single-speaker	WSJ	-
Single-channel **multi**-speaker	wsj1-2mix [1]	-
Multi-channel **multi**-speaker	Spatialized wsj1-2mix[1] 2 versions: • Anechoic • Reverberant	Train: 98.5 hr Dev: 1.3 hr Eval: 0.8 hr

[1] Hiroshi Seki, et al. "A purely end-to-end system for multi-speaker speech recognition", ACL, 2018

[1] The spatialization toolkit is available at http://www.merl.com/demos/deep-clustering/spatialize_wsj0-mix.zip

Results – Single-channel multi-speaker

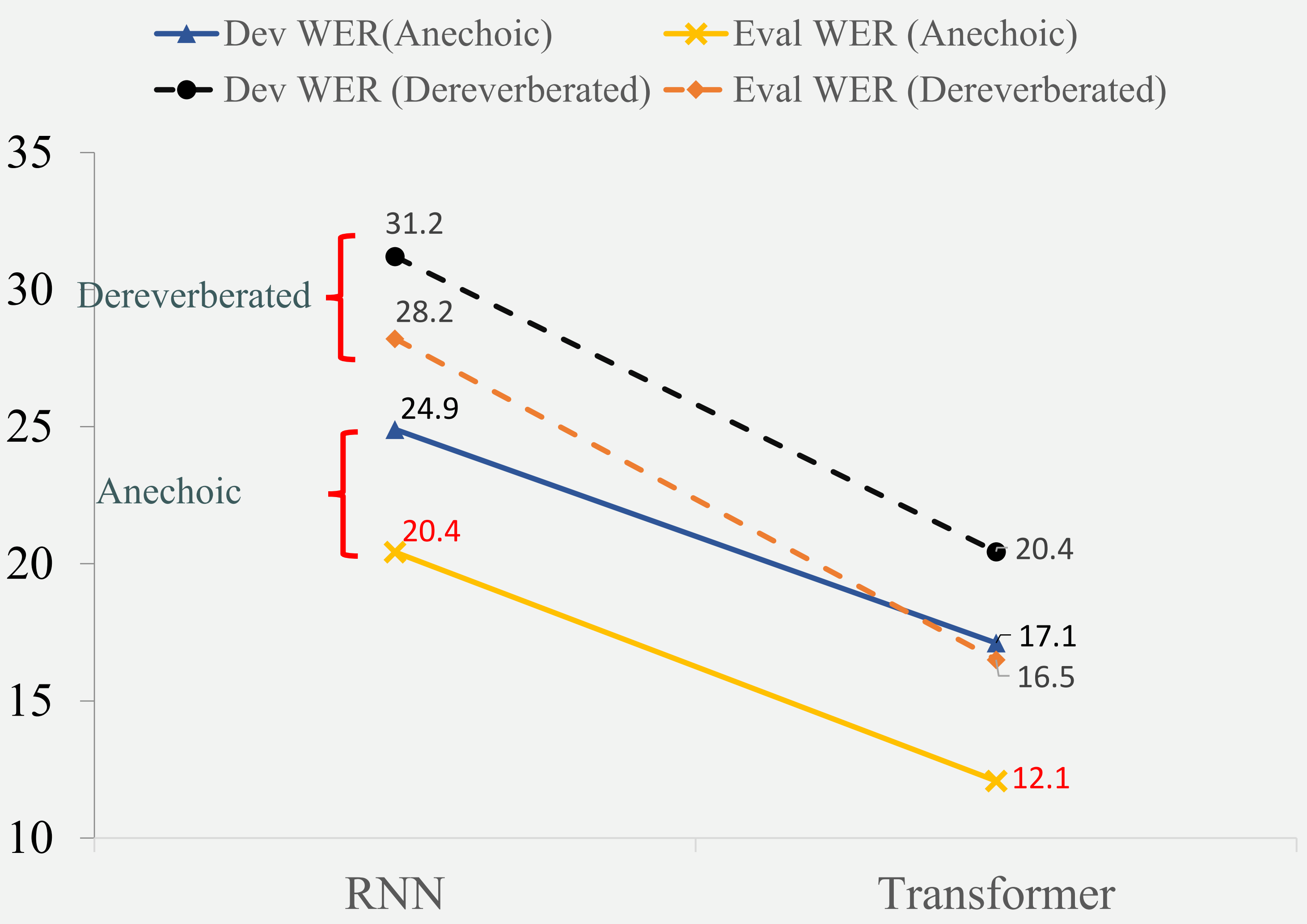
- **Anechoic**
  - **1st Channel**
- **Reverberant**
  - **Nara-WPE preprocessing**
  - **1st Channel**



Results – Single-channel multi-speaker

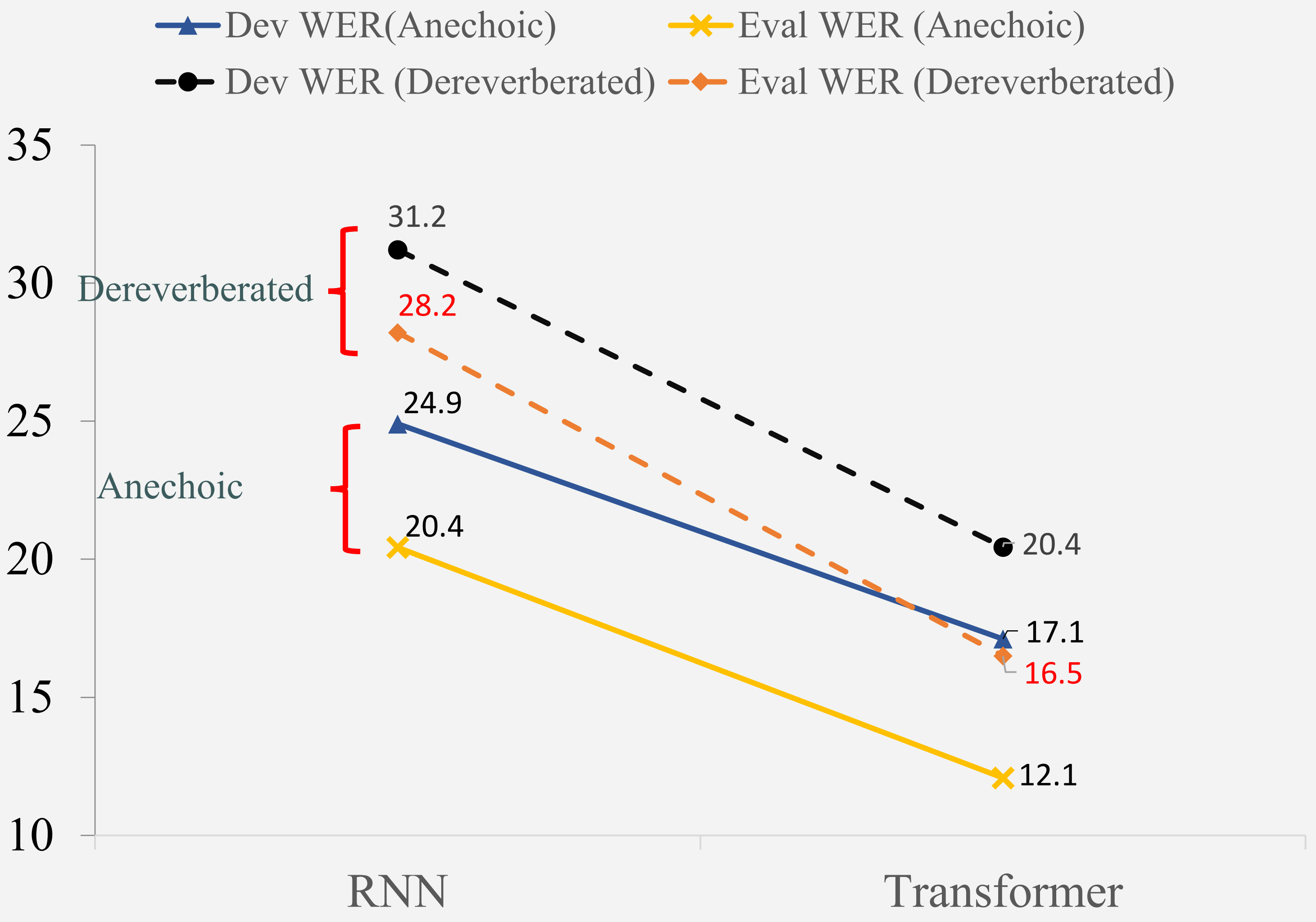
- **Anechoic**
  - **1st Channel**

- **Reverberant**
  - **Nara-WPE preprocessing**
  - **1st Channel**



Results – Single-channel multi-speaker

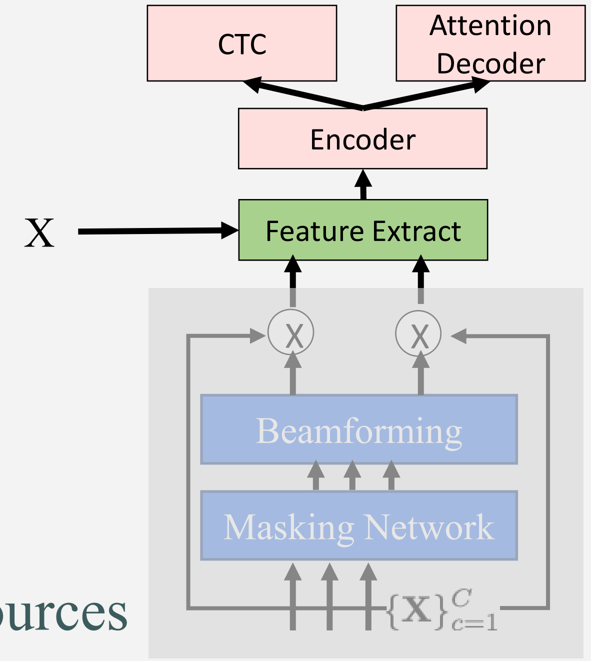
- **Anechoic**
  - **1st Channel**
- **Reverberant**
  - **Nara-WPE preprocessing**
  - **1st Channel**



Data Scheduling in Multi-channel Training

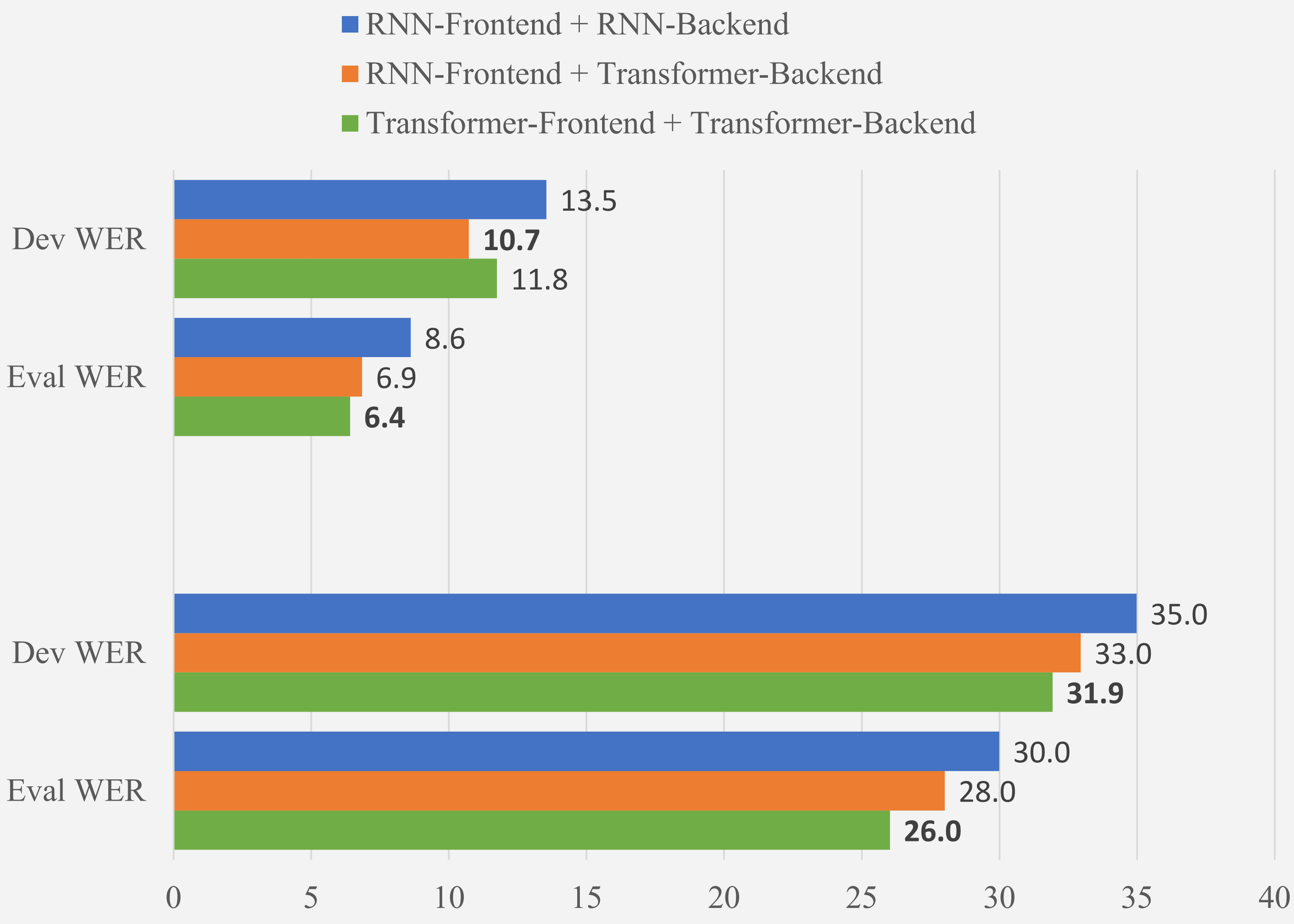
1. Include original WSJ **(single-channel single speaker)**
   - Bypassing the frontend
   - Helps regularize training
     - Improves backend ASR performance
     - Benefits frontend performance

2. Curriculum Learning
   - In the order of **balanced** → **unbalanced** energy between the sources
     1) **balanced** means both streams in the frontend can be trained.
     2) **unbalanced** samples to refine one of the streams.



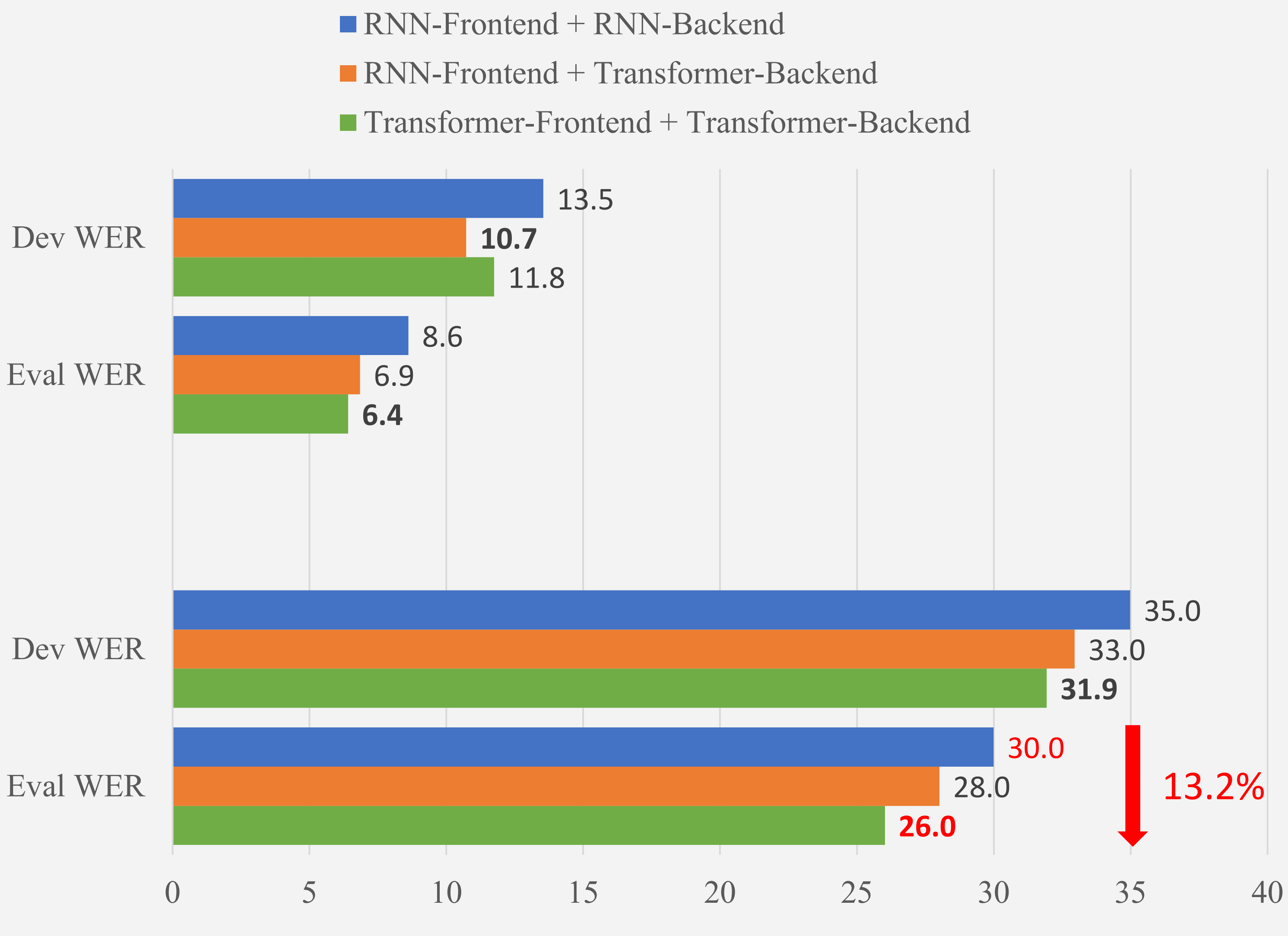
Results – Multi-channel multi-speaker

- **Anechoic**
- **Reverberant**



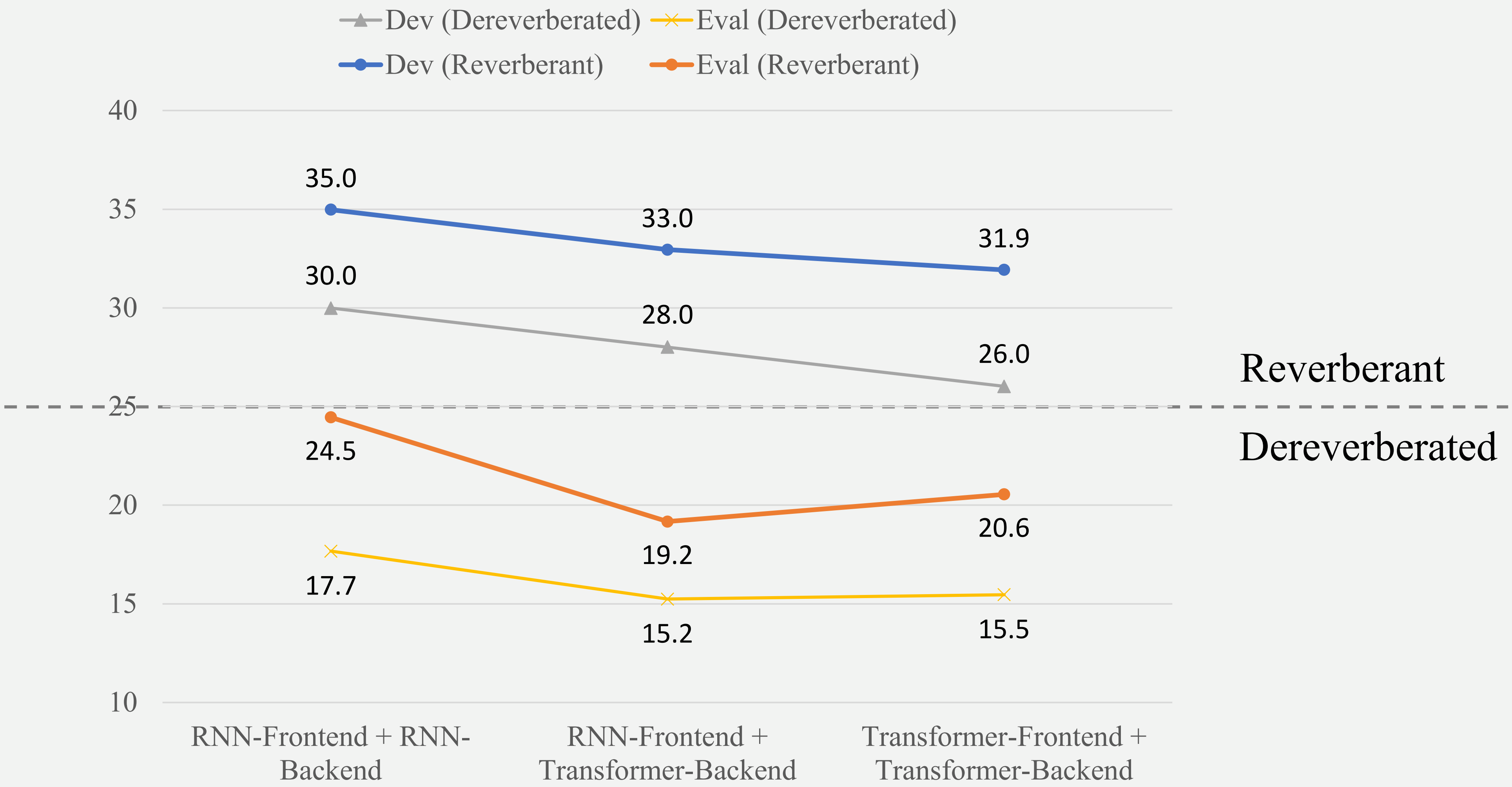
Results – Multi-channel multi-speaker

- **Anechoic**
- **Reverberant**



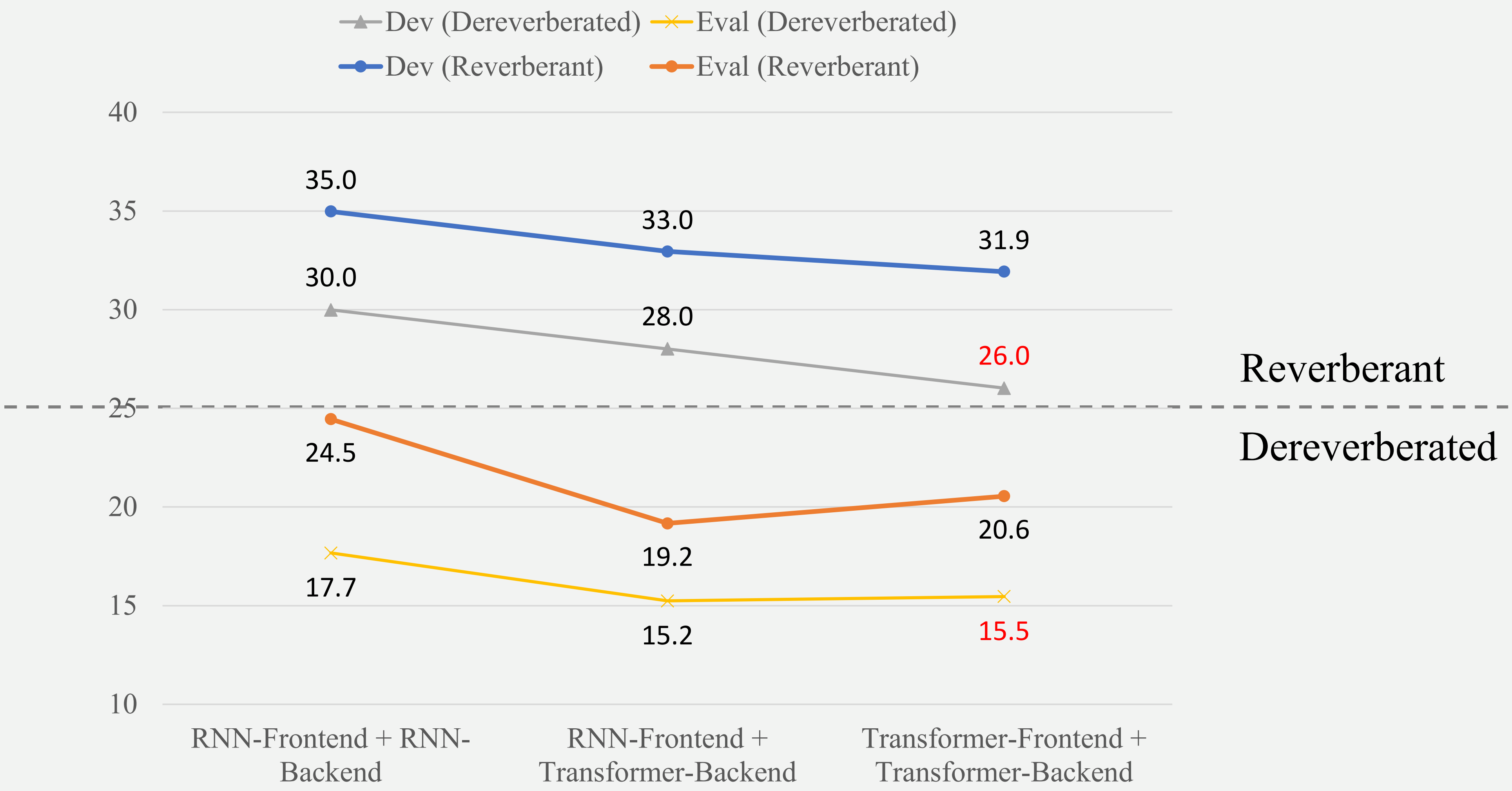
Results – Multi-channel

- ❑ With external dereverberation (WPE)

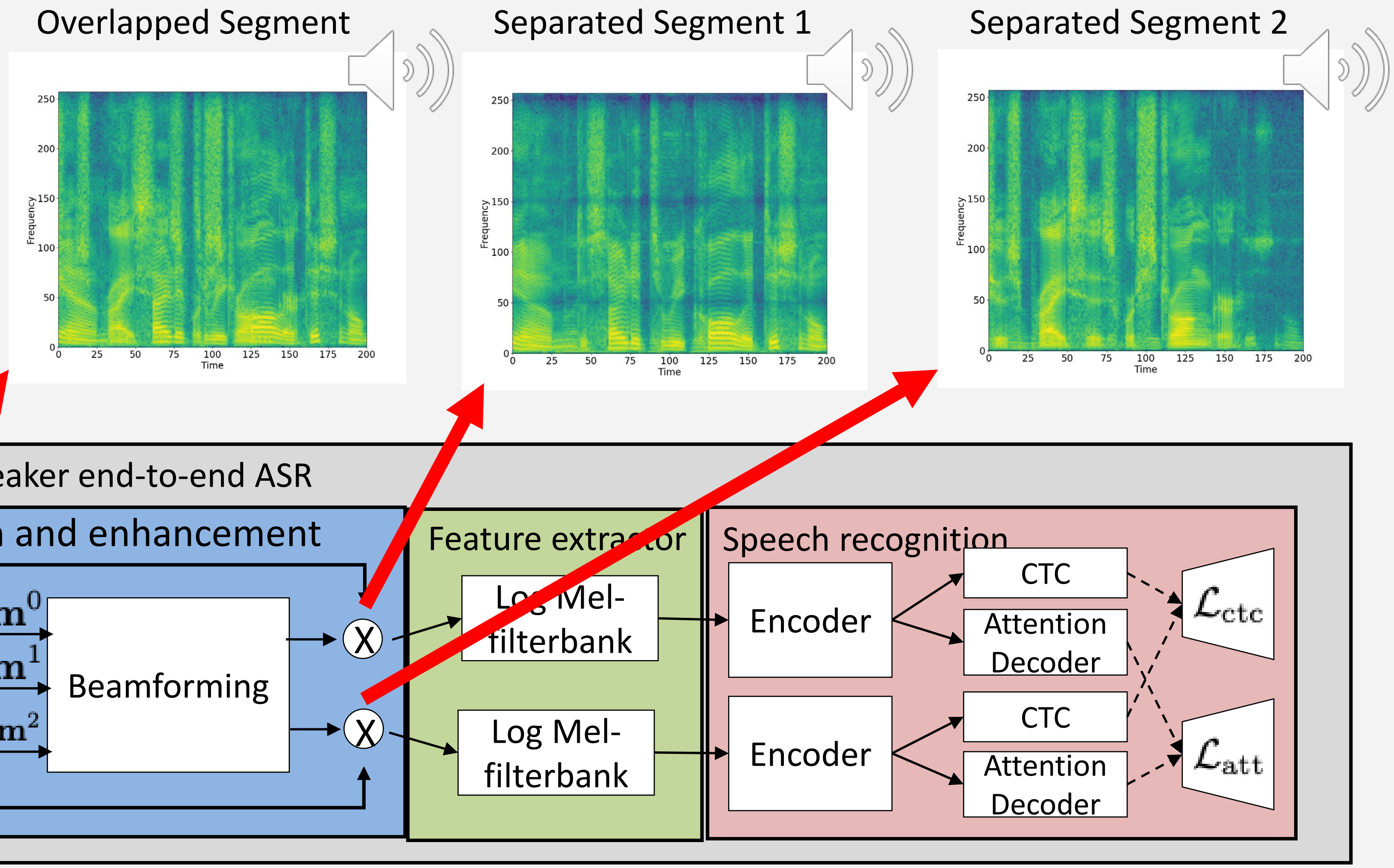


Results – Multi-channel

- With external dereverberation (WPE)



Speech Separation Ability



Conclusion

- Transformer based multi-speaker end-to-end ASR
  - ❑ Single-channel
  - ❑ Multi-channel
    - ❑ Backend ASR: encoder & decoder
    - ❑ Frontend masking network: local self-attention
      - ❑ First to apply **self-attention** in **speech separation**.
- Future work
  - ❑ To improve the performance of the model with Transformer frontend
  - ❑ To integrate dereverberation in the system
  - ❑ To apply the model on real data

Thanks!
Q & A

- Special thanks to my co-authors:

Wangyou Zhang

Yanmin Qian

Jonathan Le Roux

Shinji Watanabe