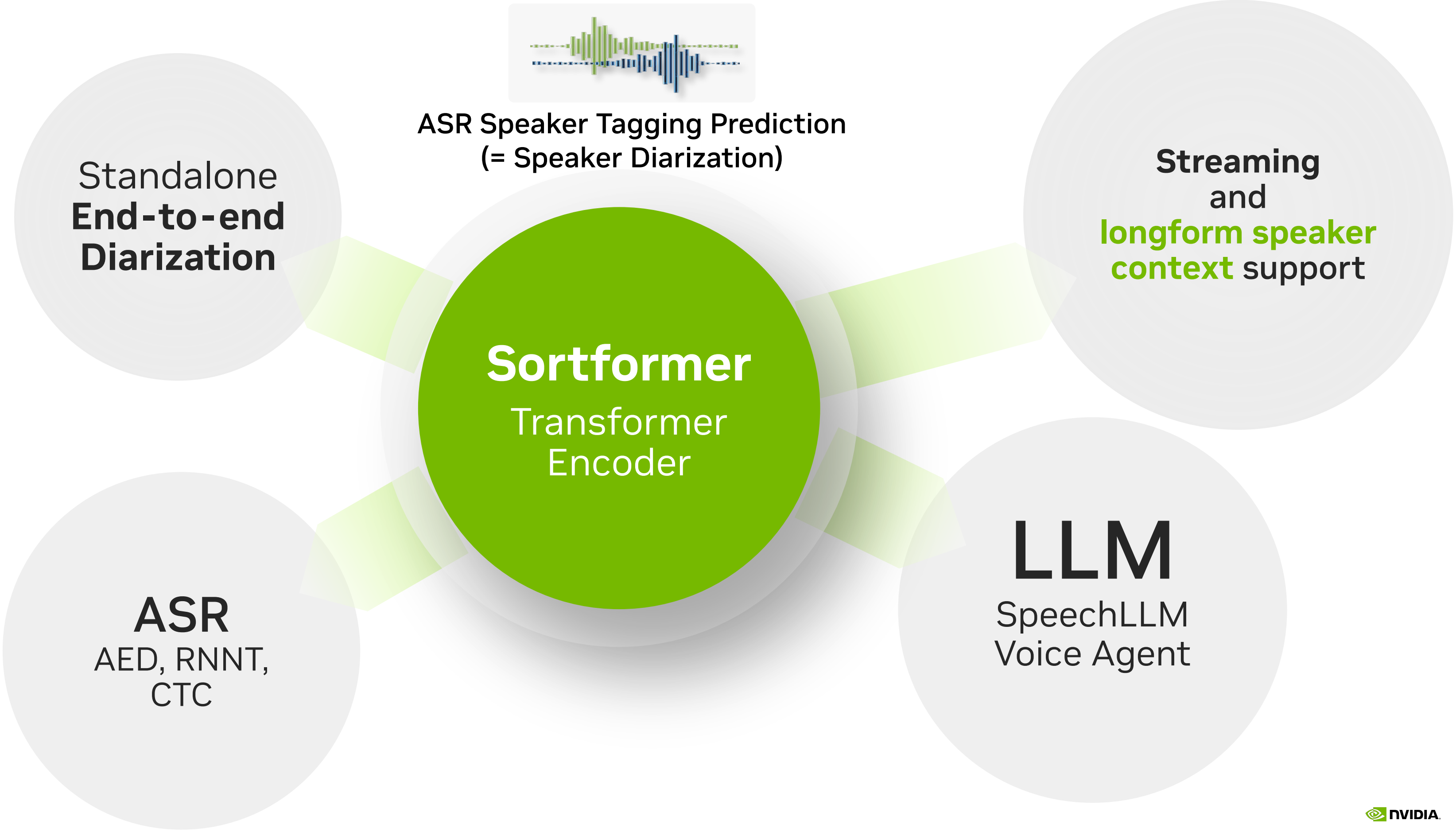
# Sortformer: A Novel Approach for Permutation-Resolved Speaker Supervision in Speech-to-Text Systems

Taejin Park
Ivan Medennikov
Kunal Dhawan
Weiqing Wang
He Huang
Nithin Rao Koluguri
Krishna C. Puvvada
Jagadeesh Balam
Boris Ginsburg

**NVIDIA, Santa Clara, CA, USA**

# Sortformer’s Mission

Power ASR models and LLMs with Speaker Information



ASR Speaker Tagging Prediction
(= Speaker Diarization)

Standalone
**End-to-end Diarization**

**Sortformer**
Transformer Encoder

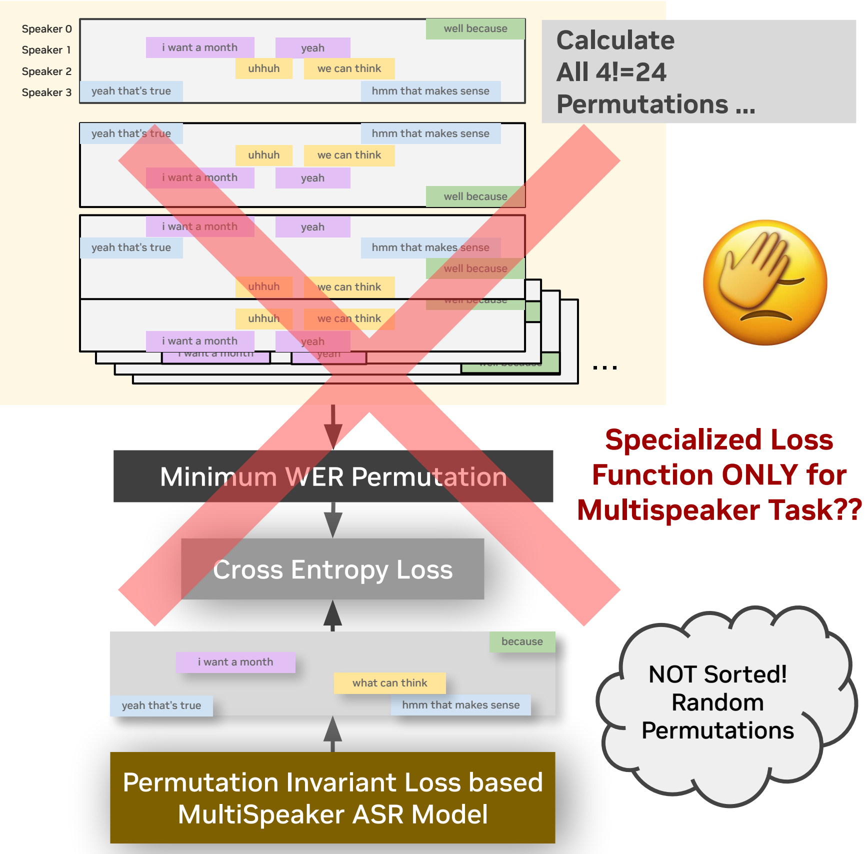
**Streaming** and **longform speaker context** support

**ASR**
AED, RNNT, CTC

**LLM**
SpeechLLM
Voice Agent

# Train Multispeaker ASR Just Like Normal ASR

Resolve Permutation with Sorting.



Calculate All 4!=24 Permutations ...

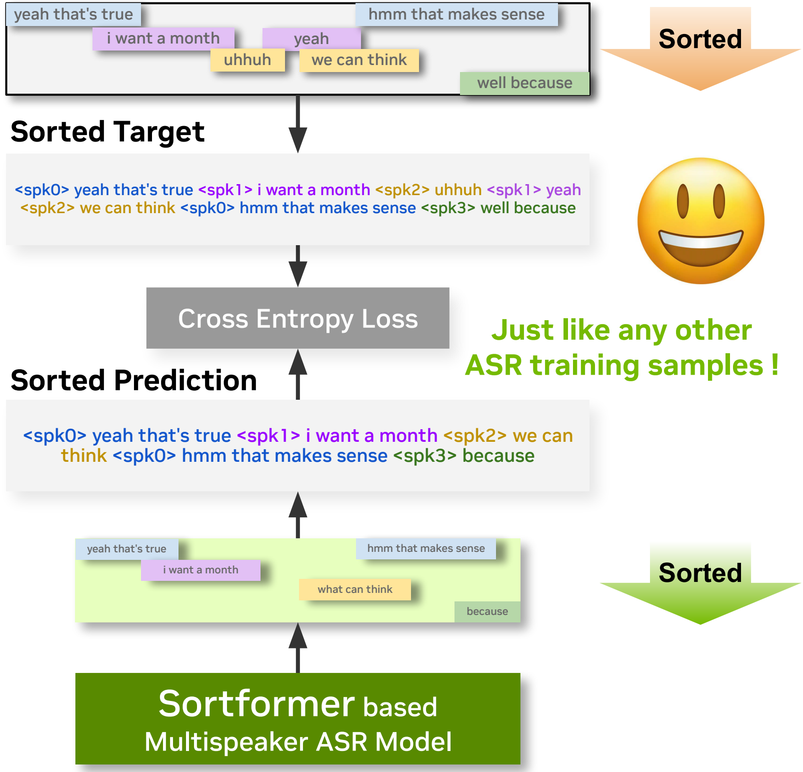
Minimum WER Permutation

Cross Entropy Loss

Permutation Invariant Loss based MultiSpeaker ASR Model

Specialized Loss Function ONLY for Multispeaker Task??

NOT Sorted! Random Permutations



Sorted Target

<spk0> yeah that's true <spk1> i want a month <spk2> uhhuh <spk1> yeah <spk2> we can think <spk0> hmm that makes sense <spk3> well because

Cross Entropy Loss

Sorted Prediction

<spk0> yeah that's true <spk1> i want a month <spk2> we can think <spk0> hmm that makes sense <spk3> because

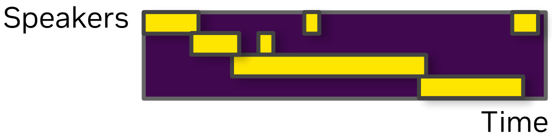
Sortformer based Multispeaker ASR Model

Just like any other ASR training samples !

# **Sortformer:** bridging between timestamps and tokens

We want to handle inference and training with token labels.

Timestamps



**Inference (Speaker Supervision)**

<spk0> oh <spk1> great <spk0> did you grow up there
<spk2> yeah <spk3> so you <spk2> born
<spk3> have <spk2> and <spk3> to <spk2> raised

**Gradient/Loss (Train)**

## A *bridge* between **timestamps** and **Tokens!**

Timestamps to Tokens?

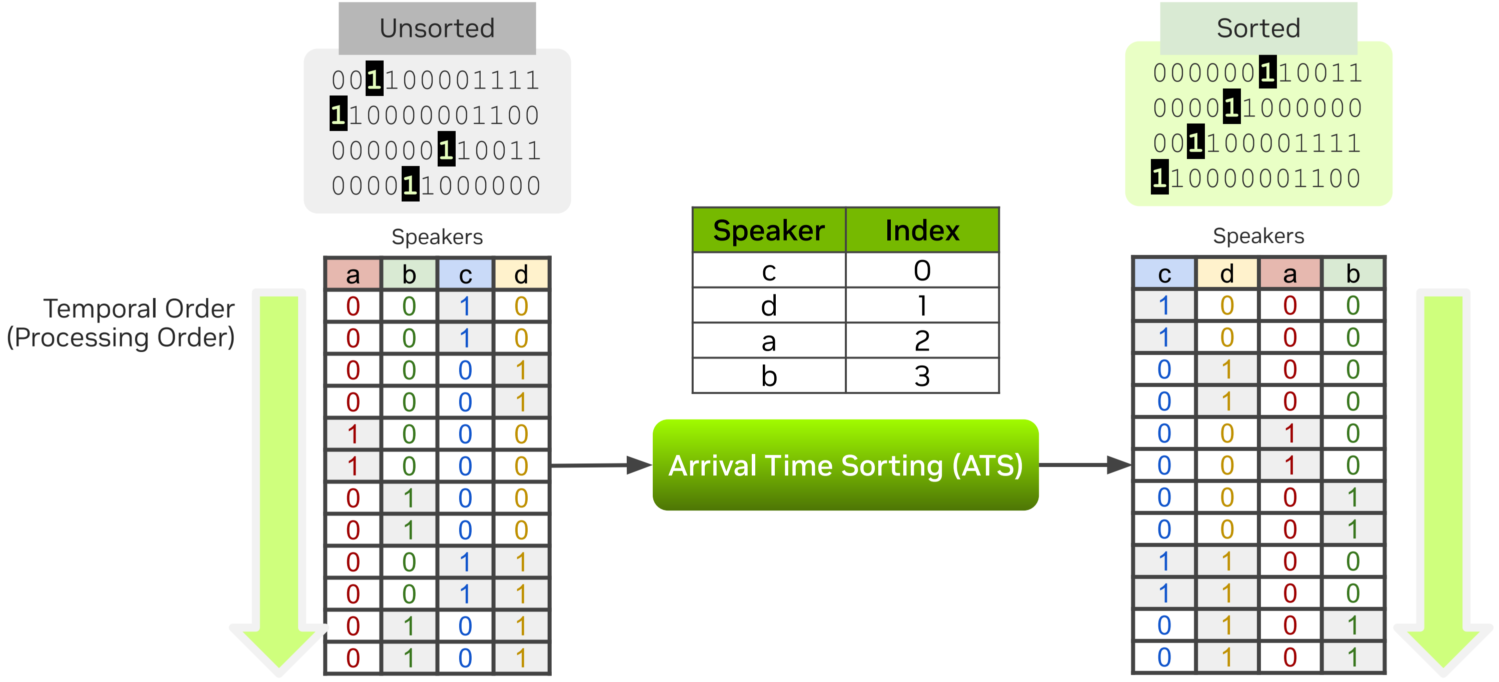
Word Alignments - Done by RNNT, CTC and Transformer AED

Speaker Permutations - Done by Sortformer

# Sortformer End-to-End Diarizer Models
# Permutation can be also resolved by Sorting!

## Sortformer Overview

**Sortformer's Speaker Diarization: Who** spoke **When + and who spoke first ?**



Unsorted

001100001111
110000001100
000000110011
000011000000

Speakers

| a | b | c | d |
|---|---|---|---|
| 0 | 0 | 1 | 0 |
| 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 1 |
| 1 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 |
| 0 | 1 | 0 | 0 |
| 0 | 0 | 1 | 1 |
| 0 | 0 | 1 | 1 |
| 0 | 1 | 0 | 1 |
| 0 | 1 | 0 | 1 |

| Speaker | Index |
|---|---|
| c | 0 |
| d | 1 |
| a | 2 |
| b | 3 |

Sorted

000000110011
000011000000
001100001111
110000001100

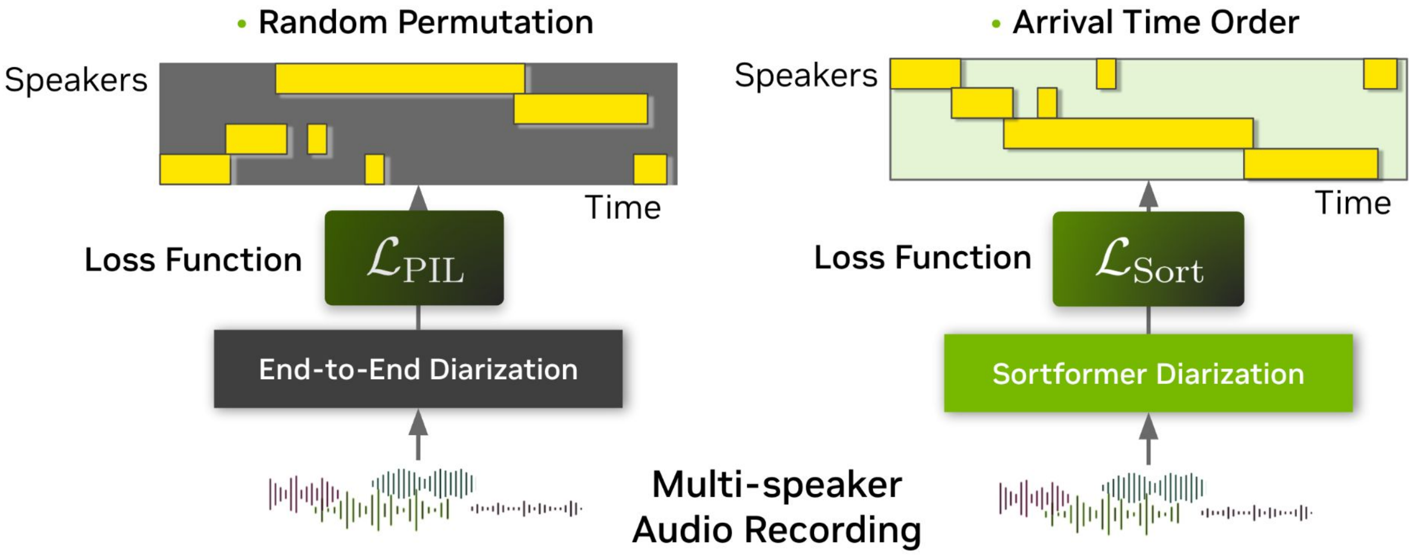
Speakers

| c | d | a | b |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 |
| 0 | 1 | 0 | 0 |
| 0 | 0 | 1 | 0 |
| 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 |
| 1 | 1 | 0 | 0 |
| 0 | 1 | 0 | 1 |
| 0 | 1 | 0 | 1 |

# Sortformer End-to-End Diarizer Models
# Permutation can be also resolved by Sorting!

## Sortformer Overview

**Sortformer's Speaker Diarization: Who** spoke **When + and who spoke first ?**



- Random Permutation

Speakers

Time

Loss Function

$\mathcal{L}_{\text{PIL}}$

End-to-End Diarization

- Arrival Time Order

Speakers

Time

Loss Function

$\mathcal{L}_{\text{Sort}}$

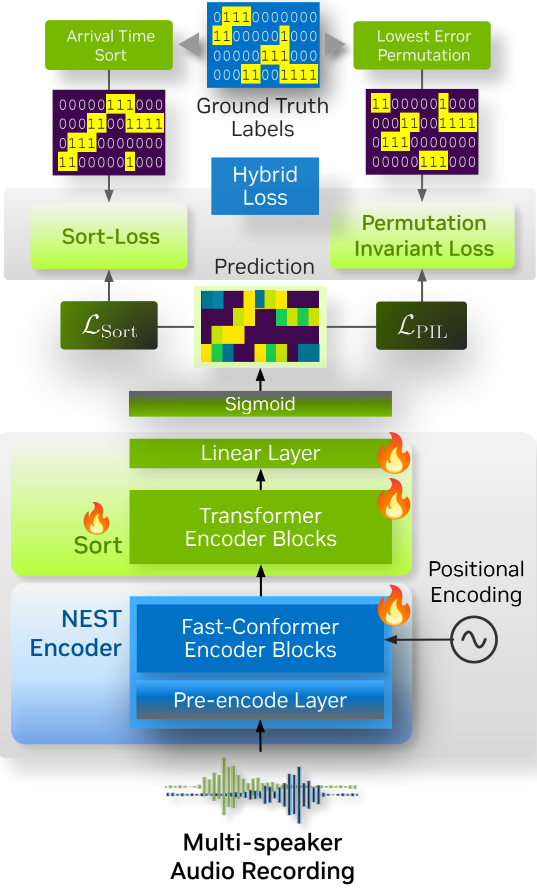
Sortformer Diarization

Multi-speaker
Audio Recording

# Sortformer End-to-End Diarizer Models

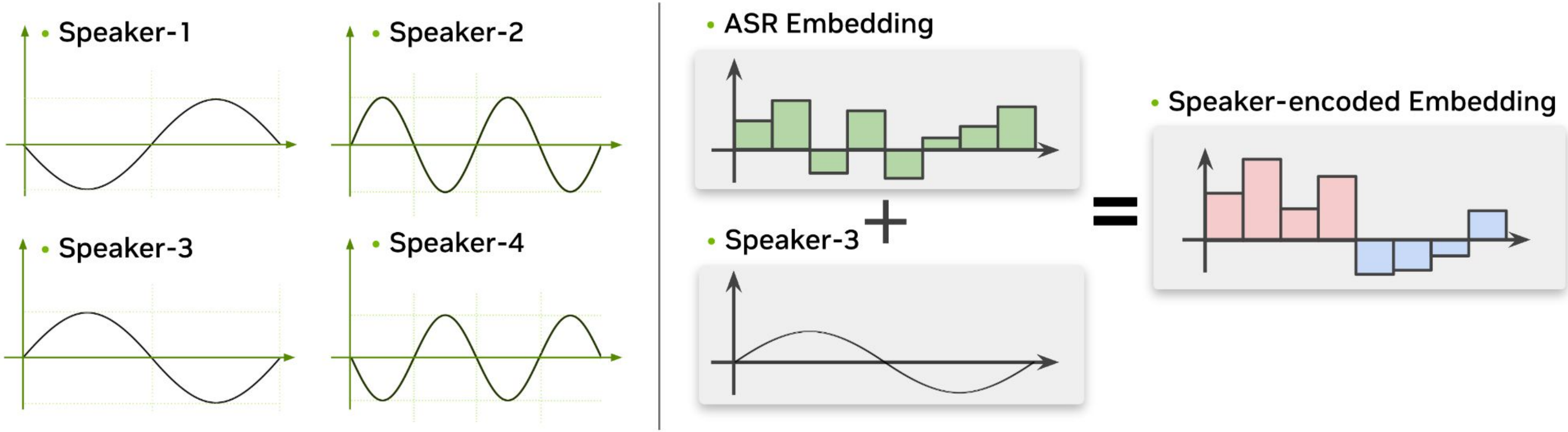
## Sortformer Overview

- **Sortformer Diarizer**
  - **Hybrid Loss**
    - We use both Sort-Loss and permutation invariant loss.
    - Still the model output is sorted
  - Based on **NEST SSL based pretrained Encoder**
    - Fast-Conformer Encoder
  - Very simple architecture (**Encoder-only**)
    - No encoder-decoderno attractors (unlike EEND-EDAAED-EEND)
    - Not an autoregressive wayone-pass Transformer-based encoding
  - To be used as: Either **Stand-alone** or **Diarization Encoder with ASR**
    - Designed to provide “speaker encoding” to ASR/speechLM models



# Sortformer is an Encoder Type Model

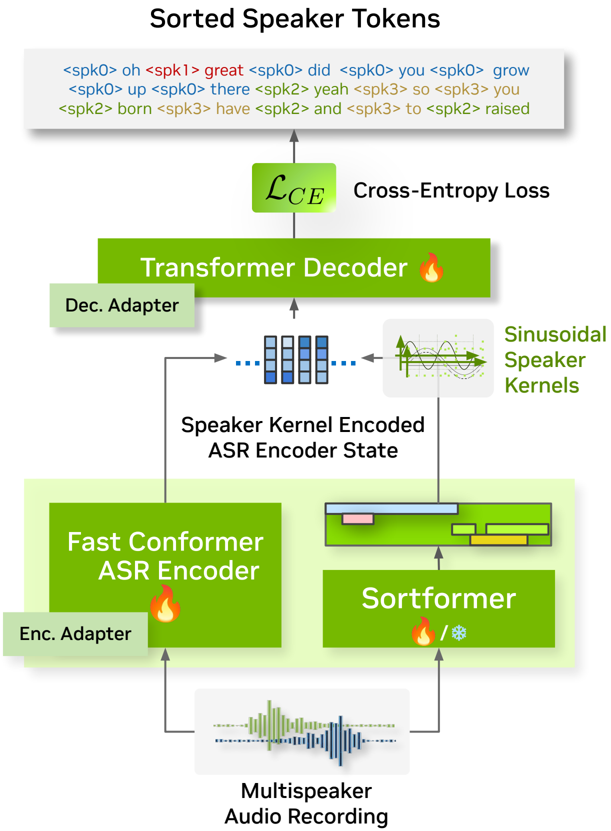
## Speaker Encoding with Differentiable Kernel Functions



- **Speaker Kernels** for adding speaker information into encoder states:
  - Inspired by positional encoding but conveying speaker information.
  - Differentiable functions and can propagate gradient to diarization module.

# Sortformer is an Encoder Type Model

## Benefits of Sortformer



### 1. Why not one Transformer? Why multiple sub-models ?

a. **Data Scarcity**

b. **Long Context Problem** in Diarization.

c. **Streaming and Long-form inference** of speaker information

### 2. Make multi-speaker ASR + diarization Easy

a. Train the whole system only using token objectives (i.e.ASR/LLM training objective).

   i. No need for timestamp annotation

   ii. No need for considering permutation-invariant loss

b. **One model** performs diarization + ASR (Multi-speaker ASR)

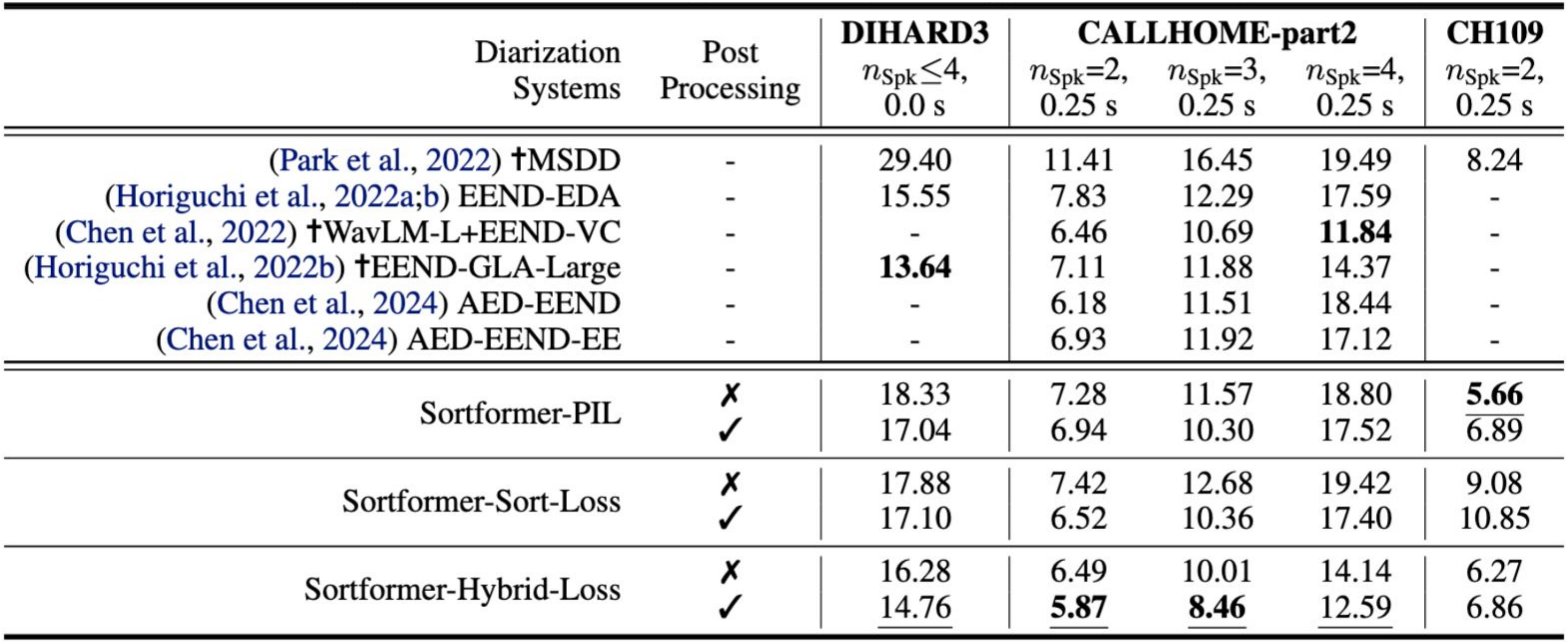
   i. Easier domain optimization than cascaded systems

   ii. Easy to deploy

# Sortformer End-to-End Diarizer Models

## Experimental Results

- **Sortformer Speaker Diarization**

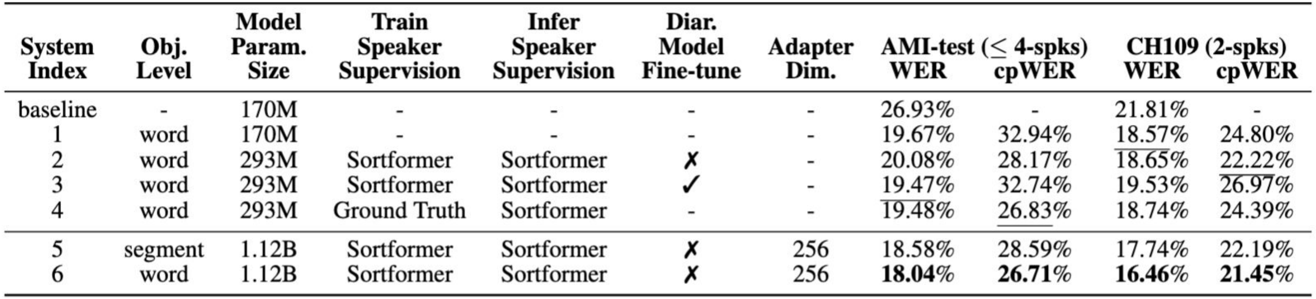


| Diarization Systems | Post Processing | DIHARD3 | CALLHOME-part2 | | | CH109 |
|---|---|---|---|---|---|---|
| | | $n_{\text{Spk}} \leq 4$, 0.0 s | $n_{\text{Spk}}=2$, 0.25 s | $n_{\text{Spk}}=3$, 0.25 s | $n_{\text{Spk}}=4$, 0.25 s | $n_{\text{Spk}}=2$, 0.25 s |
| (Park et al., 2022) †MSDD | - | 29.40 | 11.41 | 16.45 | 19.49 | 8.24 |
| (Horiguchi et al., 2022a;b) EEND-EDA | - | 15.55 | 7.83 | 12.29 | 17.59 | - |
| (Chen et al., 2022) †WavLM-L+EEND-VC | - | - | 6.46 | 10.69 | **11.84** | - |
| (Horiguchi et al., 2022b) †EEND-GLA-Large | - | **13.64** | 7.11 | 11.88 | 14.37 | - |
| (Chen et al., 2024) AED-EEND | - | - | 6.18 | 11.51 | 18.44 | - |
| (Chen et al., 2024) AED-EEND-EE | - | - | 6.93 | 11.92 | 17.12 | - |
| Sortformer-PIL | ✗ | 18.33 | 7.28 | 11.57 | 18.80 | **5.66** |
| | ✓ | 17.04 | 6.94 | 10.30 | 17.52 | 6.89 |
| Sortformer-Sort-Loss | ✗ | 17.88 | 7.42 | 12.68 | 19.42 | 9.08 |
| | ✓ | 17.10 | 6.52 | 10.36 | 17.40 | 10.85 |
| Sortformer-Hybrid-Loss | ✗ | 16.28 | 6.49 | 10.01 | 14.14 | 6.27 |
| | ✓ | 14.76 | **5.87** | **8.46** | 12.59 | 6.86 |

# Sortformer End-to-End Diarizer Models

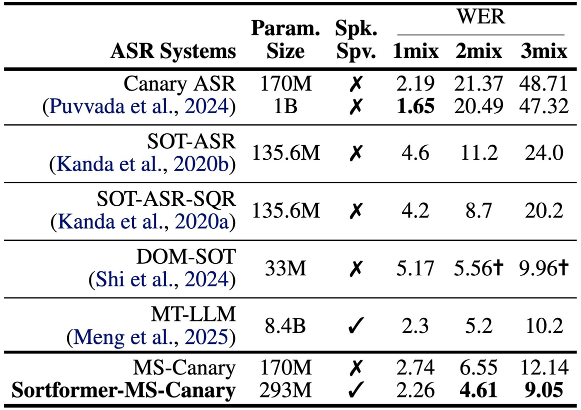
## Experimental Results

- **Multi-speaker Canary: Multi-speaker ASR with Canary supervised by Sortformer Diarizer (Ablation Study)**



| System Index | Obj. Level | Model Param. Size | Train Speaker Supervision | Infer Speaker Supervision | Diar. Model Fine-tune | Adapter Dim. | AMI-test ($\leq$ 4-spks) WER | AMI-test ($\leq$ 4-spks) cpWER | CH109 (2-spks) WER | CH109 (2-spks) cpWER |
|---|---|---|---|---|---|---|---|---|---|---|
| baseline | - | 170M | - | - | - | - | 26.93% | - | 21.81% | - |
| 1 | word | 170M | - | - | - | - | 19.67% | 32.94% | 18.57% | 24.80% |
| 2 | word | 293M | Sortformer | Sortformer | ✗ | - | 20.08% | 28.17% | 18.65% | 22.22% |
| 3 | word | 293M | Sortformer | Sortformer | ✓ | - | 19.47% | 32.74% | 19.53% | 26.97% |
| 4 | word | 293M | Ground Truth | Sortformer | - | - | 19.48% | 26.83% | 18.74% | 24.39% |
| 5 | segment | 1.12B | Sortformer | Sortformer | ✗ | 256 | 18.58% | 28.59% | 17.74% | 22.19% |
| 6 | word | 1.12B | Sortformer | Sortformer | ✗ | 256 | **18.04%** | **26.71%** | **16.46%** | **21.45%** |

- **Multi-speaker Canary: Overlapped Speech Evaluations**
  - **LibriSpeechMix:** 0.5~10 sec of overlaps, up to 3 speakers

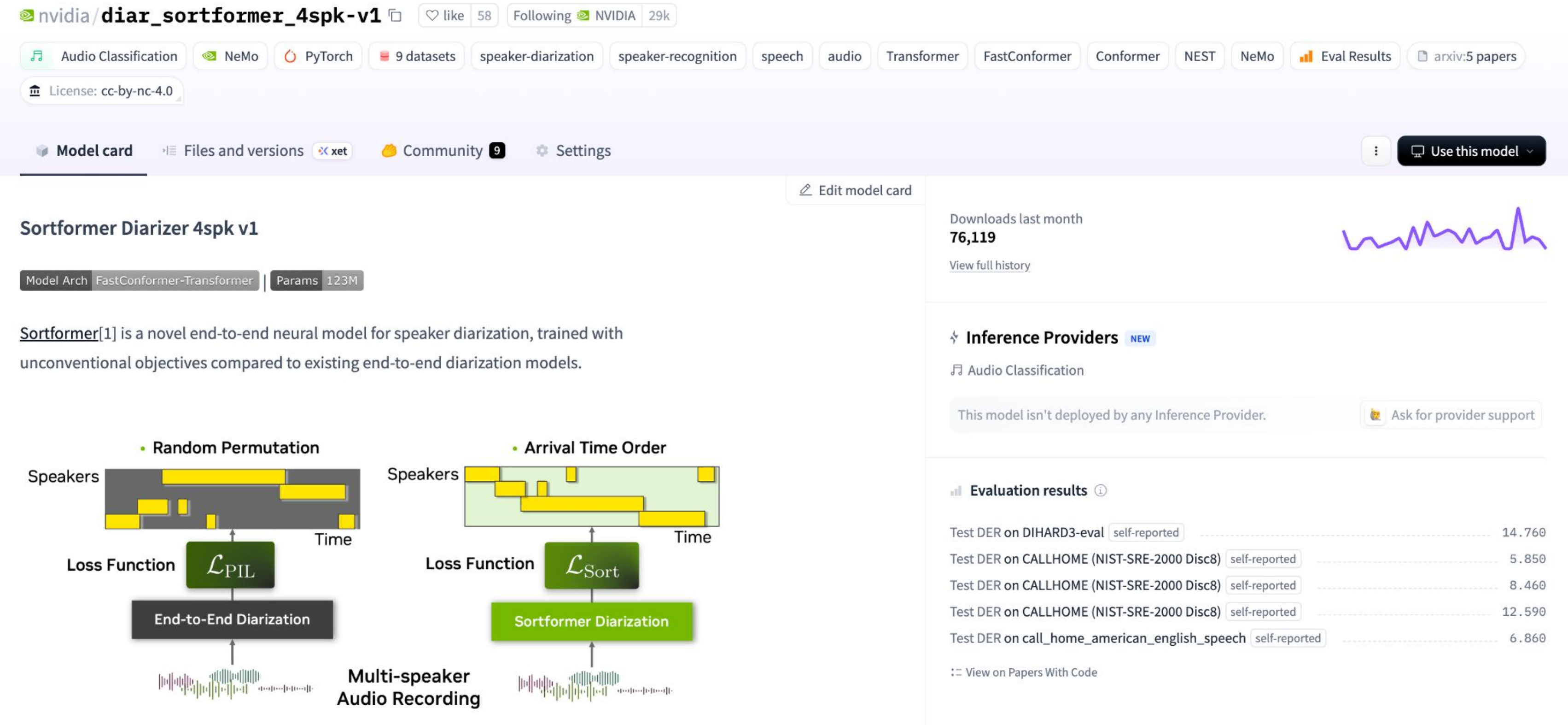


| ASR Systems | Param. Size | Spk. Spv. | WER 1mix | WER 2mix | WER 3mix |
|---|---|---|---|---|---|
| Canary ASR | 170M | ✗ | 2.19 | 21.37 | 48.71 |
| (Puvvada et al., 2024) | 1B | ✗ | **1.65** | 20.49 | 47.32 |
| SOT-ASR<br>(Kanda et al., 2020b) | 135.6M | ✗ | 4.6 | 11.2 | 24.0 |
| SOT-ASR-SQR<br>(Kanda et al., 2020a) | 135.6M | ✗ | 4.2 | 8.7 | 20.2 |
| DOM-SOT<br>(Shi et al., 2024) | 33M | ✗ | 5.17 | 5.56† | 9.96† |
| MT-LLM<br>(Meng et al., 2025) | 8.4B | ✓ | 2.3 | 5.2 | 10.2 |
| MS-Canary | 170M | ✗ | 2.74 | 6.55 | 12.14 |
| **Sortformer-MS-Canary** | 293M | ✓ | 2.26 | **4.61** | **9.05** |

# Sortformer is an Open-Source Model

**[Link]** [Hugging Face: diar_sortformer_4spk_v1]

NVIDIA NeMo Toolkit



nvidia/**diar_sortformer_4spk-v1** ♡ like 58 Following NVIDIA 29k

Audio Classification | NeMo | PyTorch | 9 datasets | speaker-diarization | speaker-recognition | speech | audio | Transformer | FastConformer | Conformer | NEST | NeMo | Eval Results | arxiv:5 papers | License: cc-by-nc-4.0

## Sortformer Diarizer 4spk v1

Model Arch FastConformer-Transformer | Params 123M

Sortformer[1] is a novel end-to-end neural model for speaker diarization, trained with unconventional objectives compared to existing end-to-end diarization models.

Downloads last month
**76,119**
View full history

**Inference Providers** NEW

Audio Classification

This model isn't deployed by any Inference Provider. Ask for provider support

**Evaluation results**

| | | |
|---|---|---|
| Test DER on DIHARD3-eval | self-reported | 14.760 |
| Test DER on CALLHOME (NIST-SRE-2000 Disc8) | self-reported | 5.850 |
| Test DER on CALLHOME (NIST-SRE-2000 Disc8) | self-reported | 8.460 |
| Test DER on CALLHOME (NIST-SRE-2000 Disc8) | self-reported | 12.590 |
| Test DER on call_home_american_english_speech | self-reported | 6.860 |

View on Papers With Code

# Sortformer: A Novel Approach for Permutation-Resolved Speaker Supervision in Speech-to-Text Systems

**Thank you for your Attention !**