DeepASA

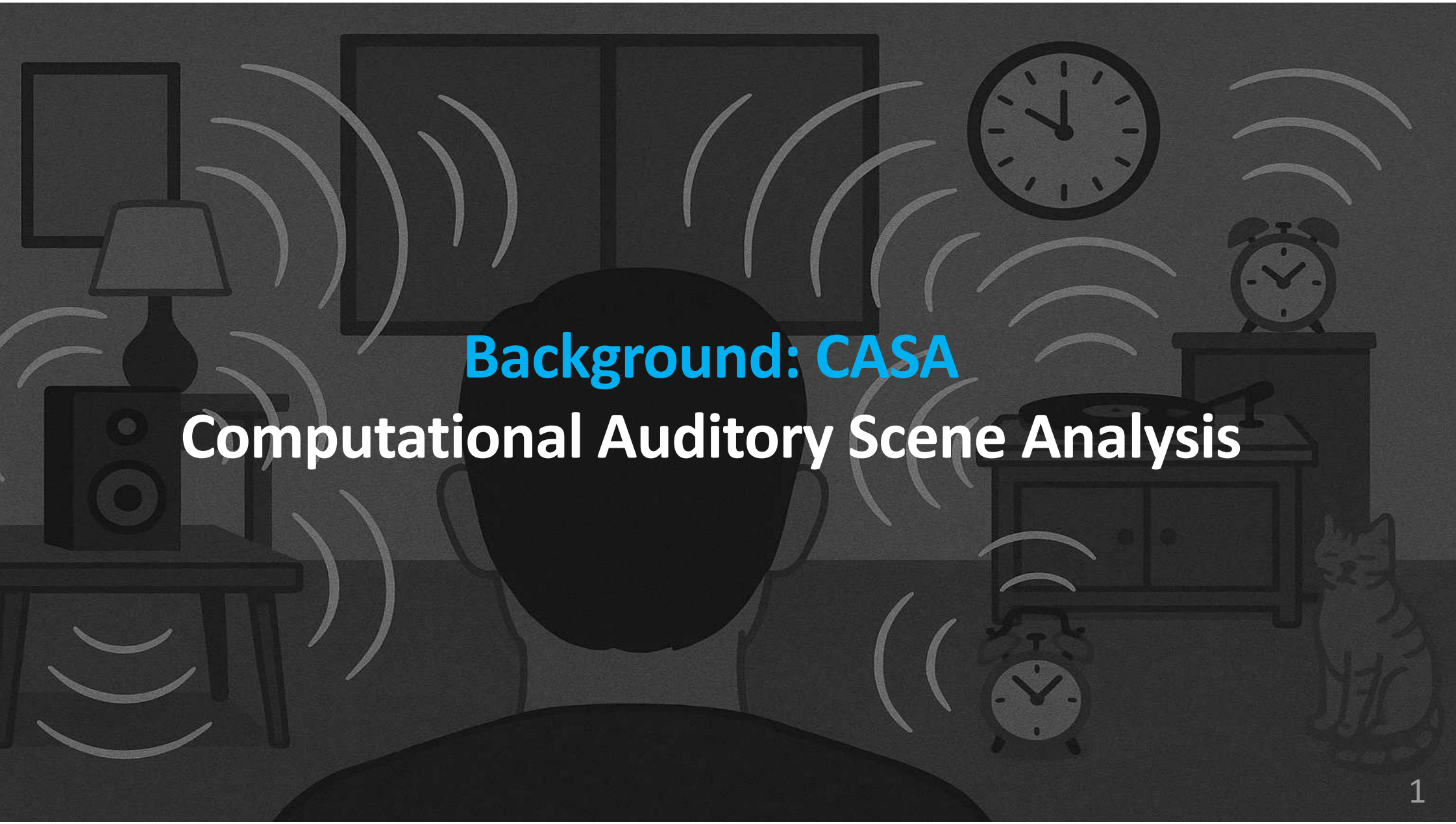
An Object-Oriented Multi-Purpose Network for Auditory Scene Analysis

Dongheon Lee, Younghoo Kwon, Jung-Woo Choi†

School of Electrical Engineering
Korea Advanced Institute of Science & Technology

NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS Conference San Diego , Thu. 4th December 2025 (16:30 – 19:30)



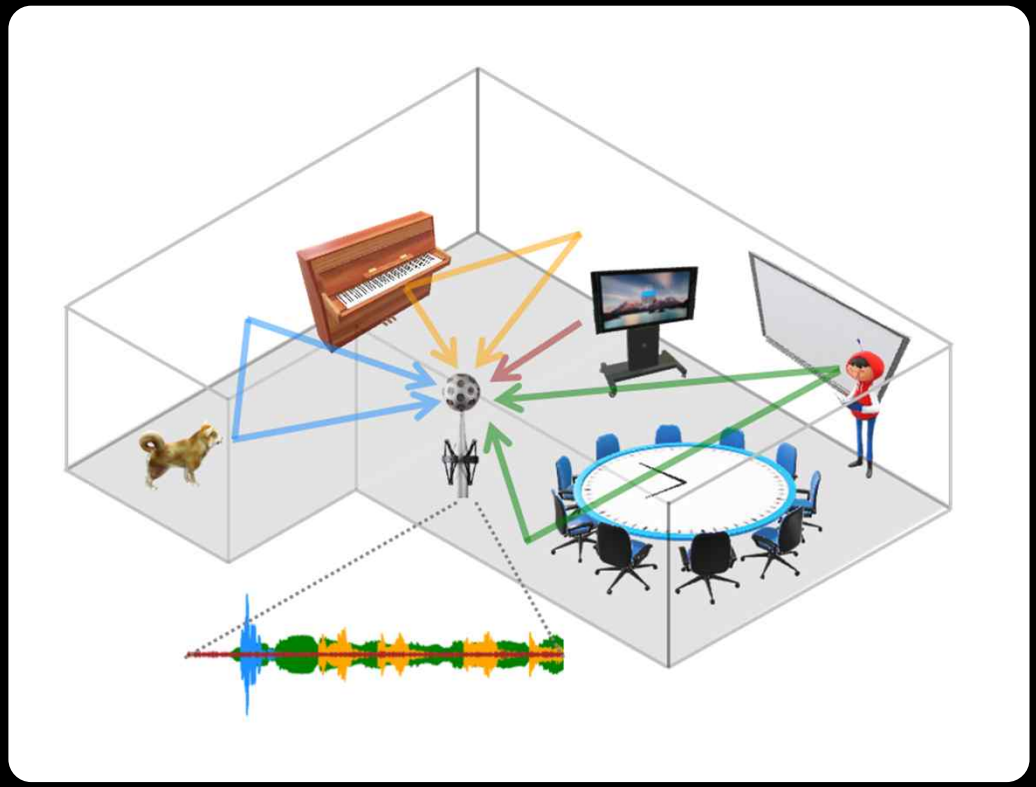
Background: CASA

Computational Auditory Scene Analysis

Auditory Scene Analysis (ASA)

KAIST EE

- **Complex auditory scenes**
  - Simultaneously activated multiple sources
  - Reflections & reverberations
  - Sources from diverse or the same classes
  - Multichannel audio recording



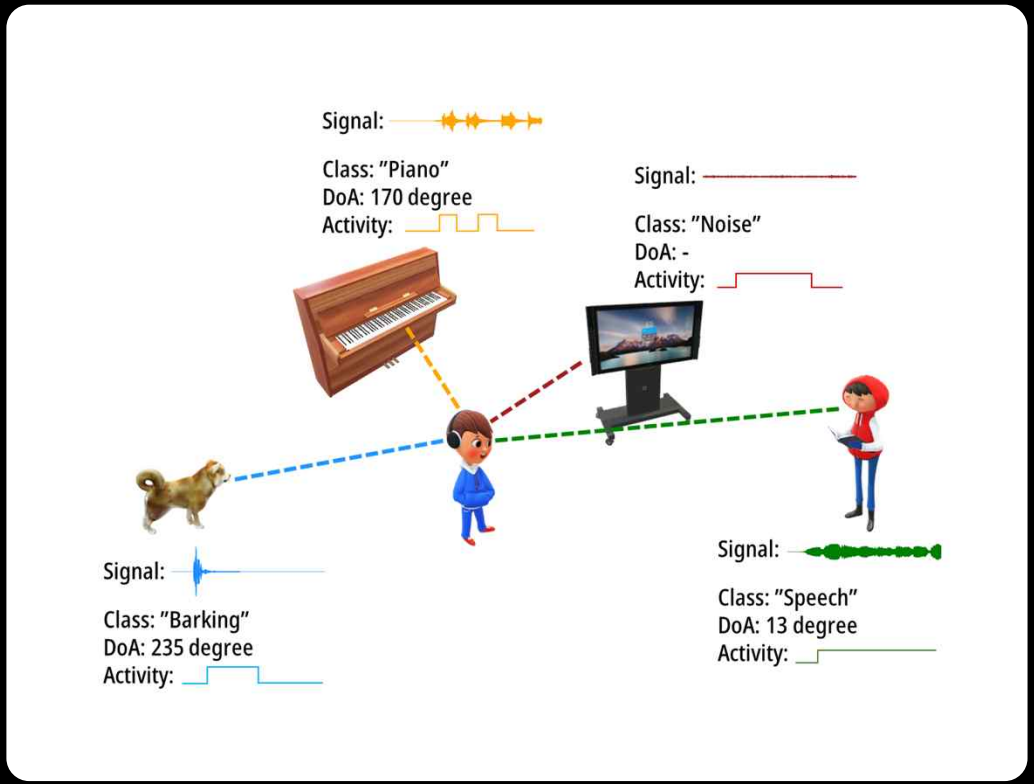
Polyphony

In-class Polyphony

Auditory Scene Analysis (ASA)

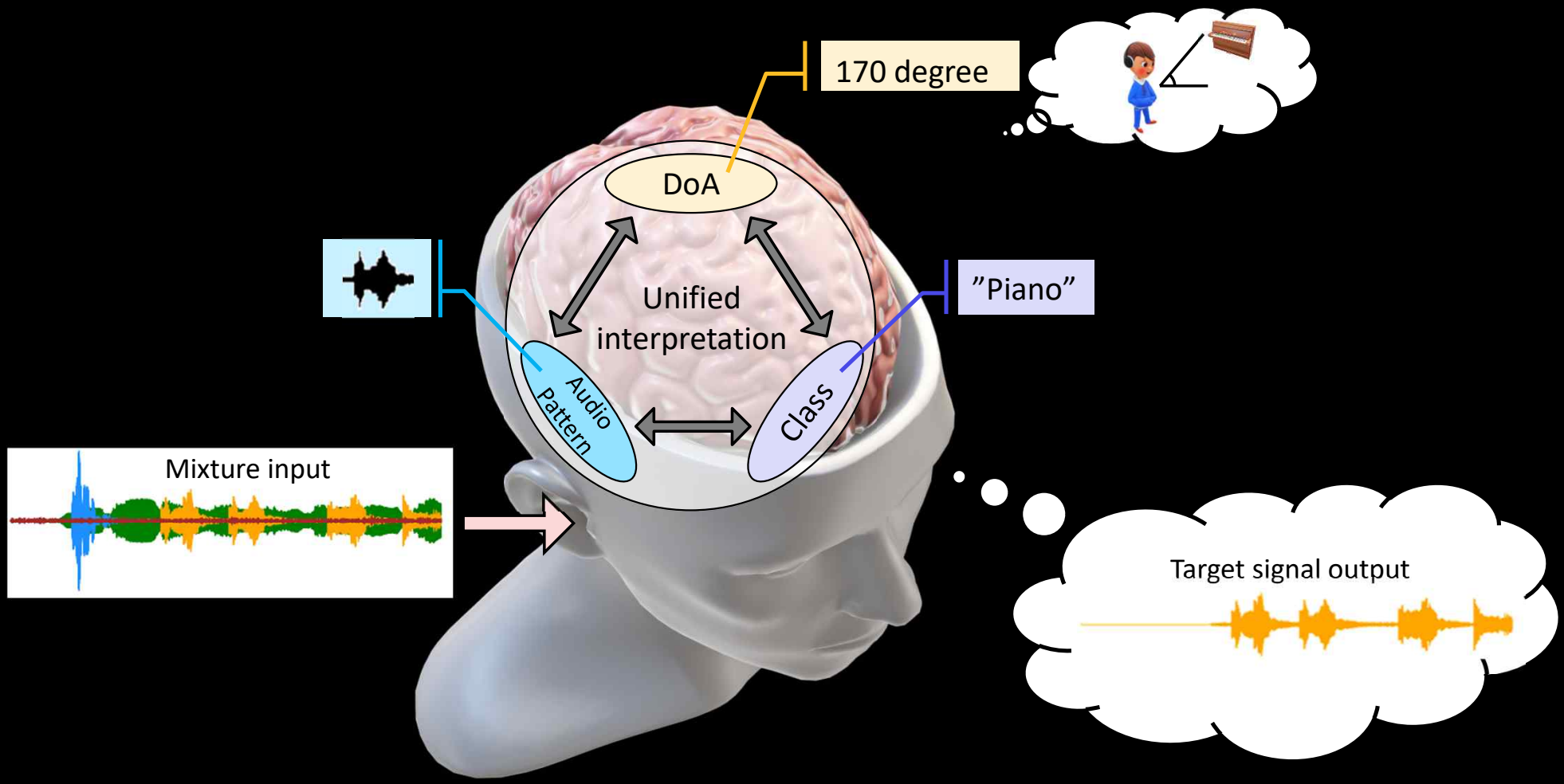
KAIST EE

- **Analysis**
  - Spatial, Temporal, and Spectral relations
- **Tasks**
  - **Information retrieval**
    - Direction-of-arrival estimation (DoAE)
    - Sound event detection (SED)
    - Source Classification (SC)
  - **Source separation**
    - Enhancement (noise suppression)
    - Universal source separation (USS)
    - Target signal extraction (TSE)



Attention-driven auditory streaming

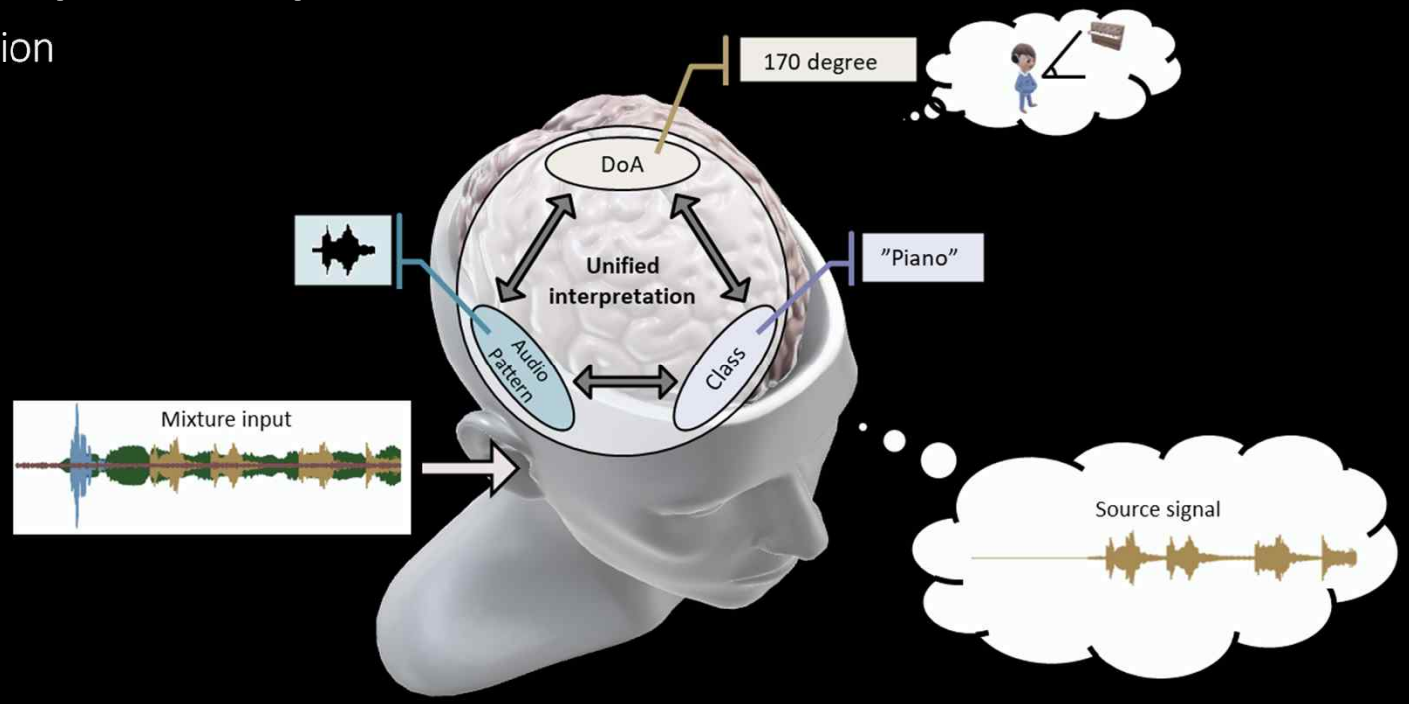
KAIST EE



Challenges remain

KAIST EE

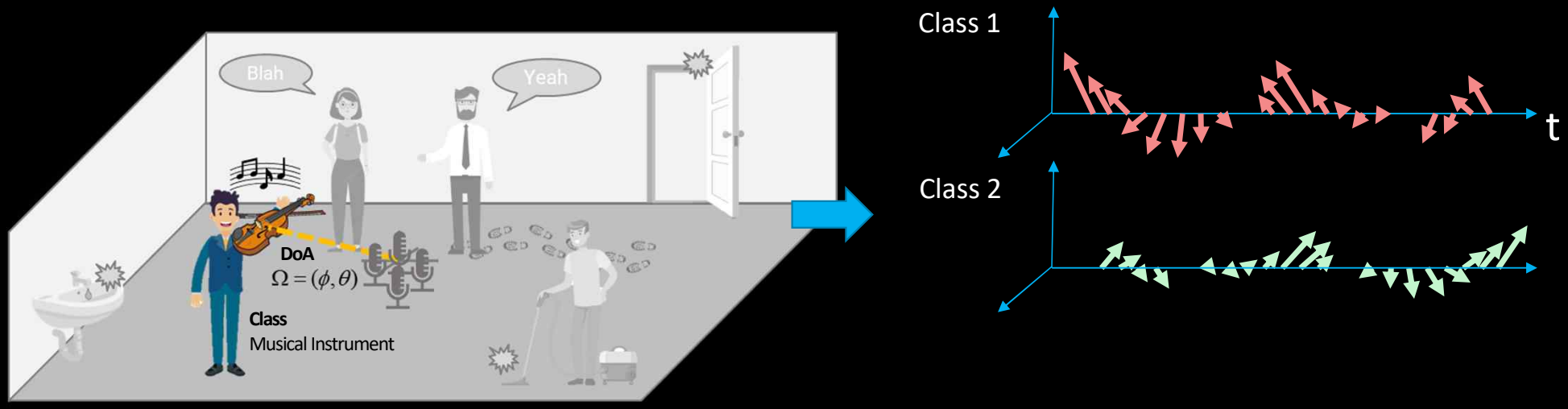
- **Unified ASA model for spatial audio**
  - **A foundation model for various auditory analysis tasks**
  - **A unified model** for **source separation + parameter estimation**
    - Multichannel source separation
    - Noise suppression
    - DoA estimation
    - Event classification



Unified model for ASA

KAIST EE

- **SELD (Sound Event Localization and Detection)**
  - Acoustic parameter estimation networks (DCASE Challenge, Task 3)
  - Maximum polyphonies: 5



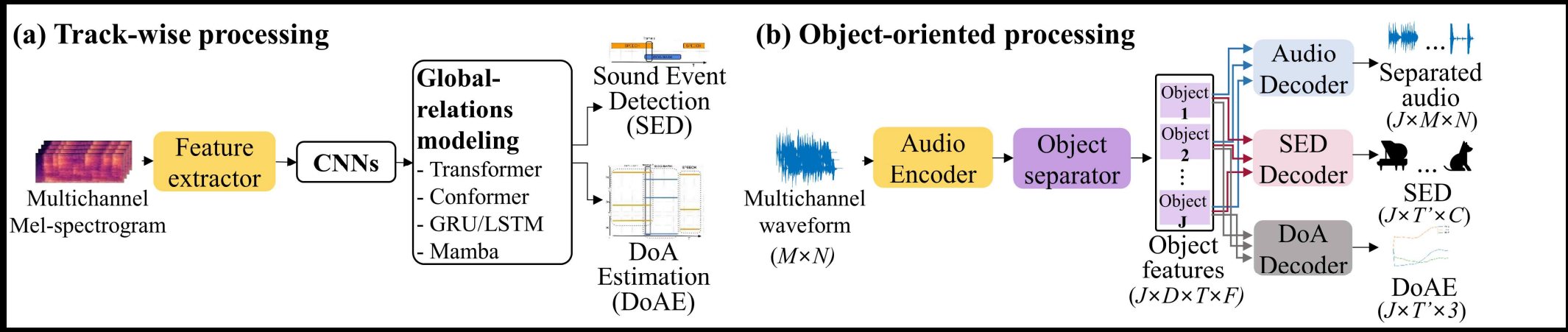
Sound Event Localization and Detection (SELD)[1]

[1] S. Adavanne, et al., "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," IEEE JSTSP, vol. 13, no. 1, pp. 34–48, 2019.

Object-oriented processing (OOP)

KAIST EE

- Comparison to Track (class)-based processing



DeFT-Mamba

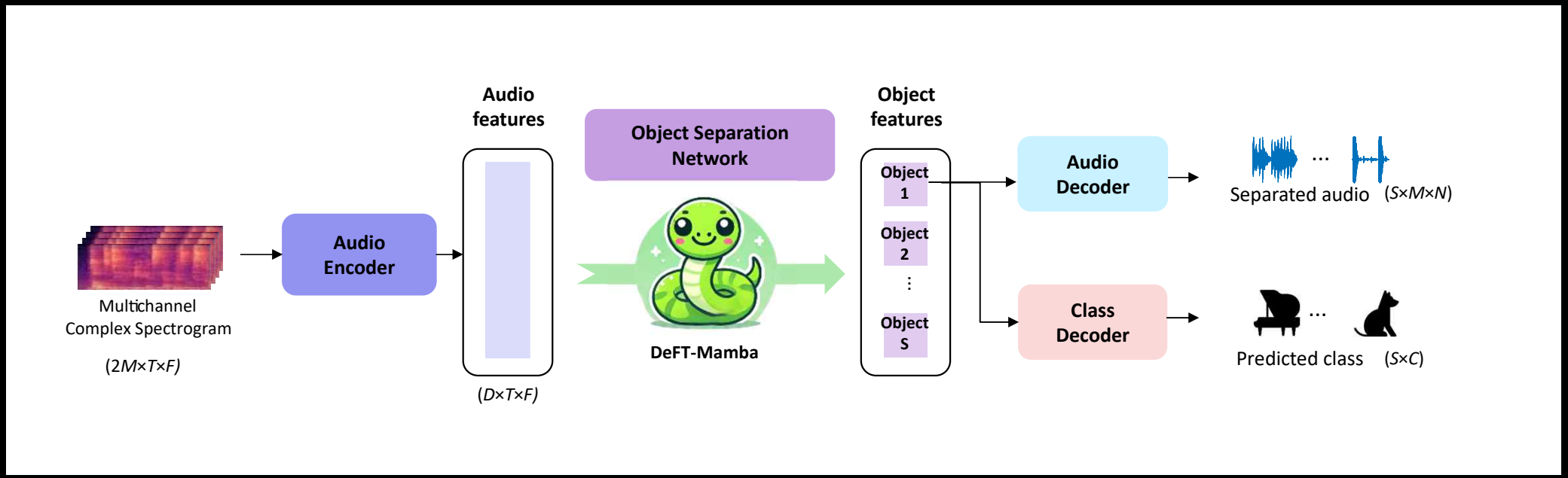
Chicken or Egg ?

DeepASA
An Object-Oriented Multi-Purpose Network
for Auditory Scene Analysis

Universal DeFT-Mamba

KAIST EE

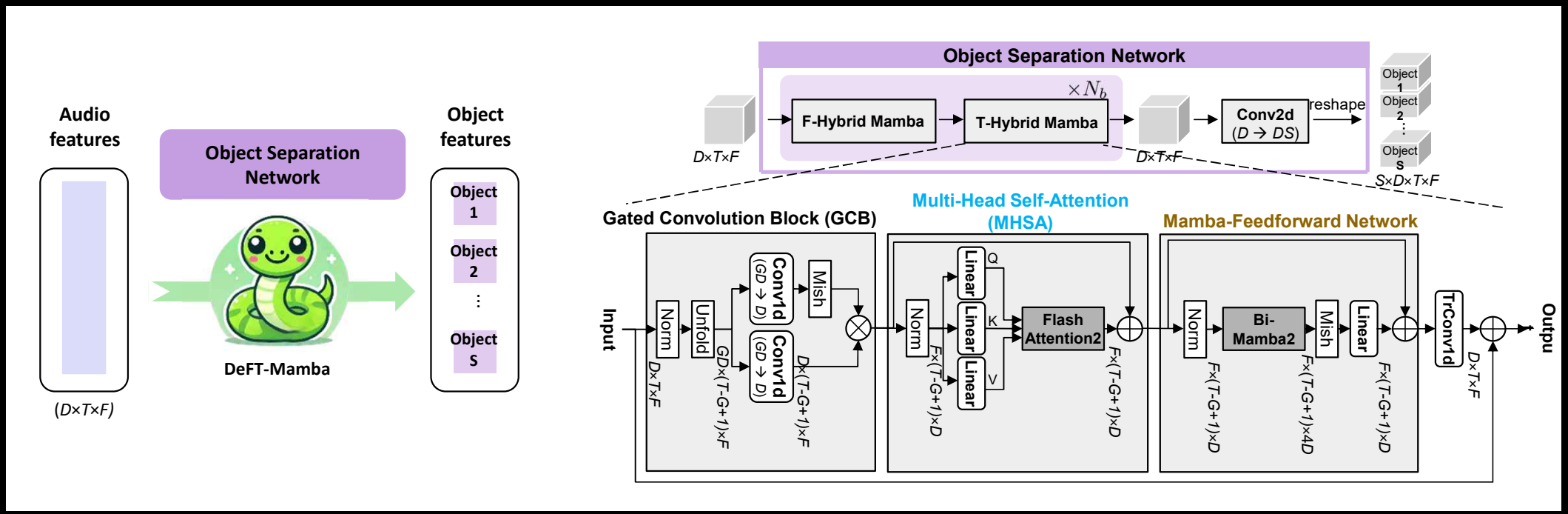
- **Object-oriented model with Multihead decoder**
  - **DeFT-Mamba:** Object feature separation through Transformer + Mamba architecture
  - **Multihead decoder:** Simultaneous estimation of audio signals and SELD parameters



DeFT-Mamba (2025)

KAIST EE

- Leveraging **Transformer** and **Mamba** architectures
  - **Transformer** captures **global relations** independent of sequence order
  - **Bi-Mamba2** identifies **sequential relation**, functioning as a **position-aware feedforward network**



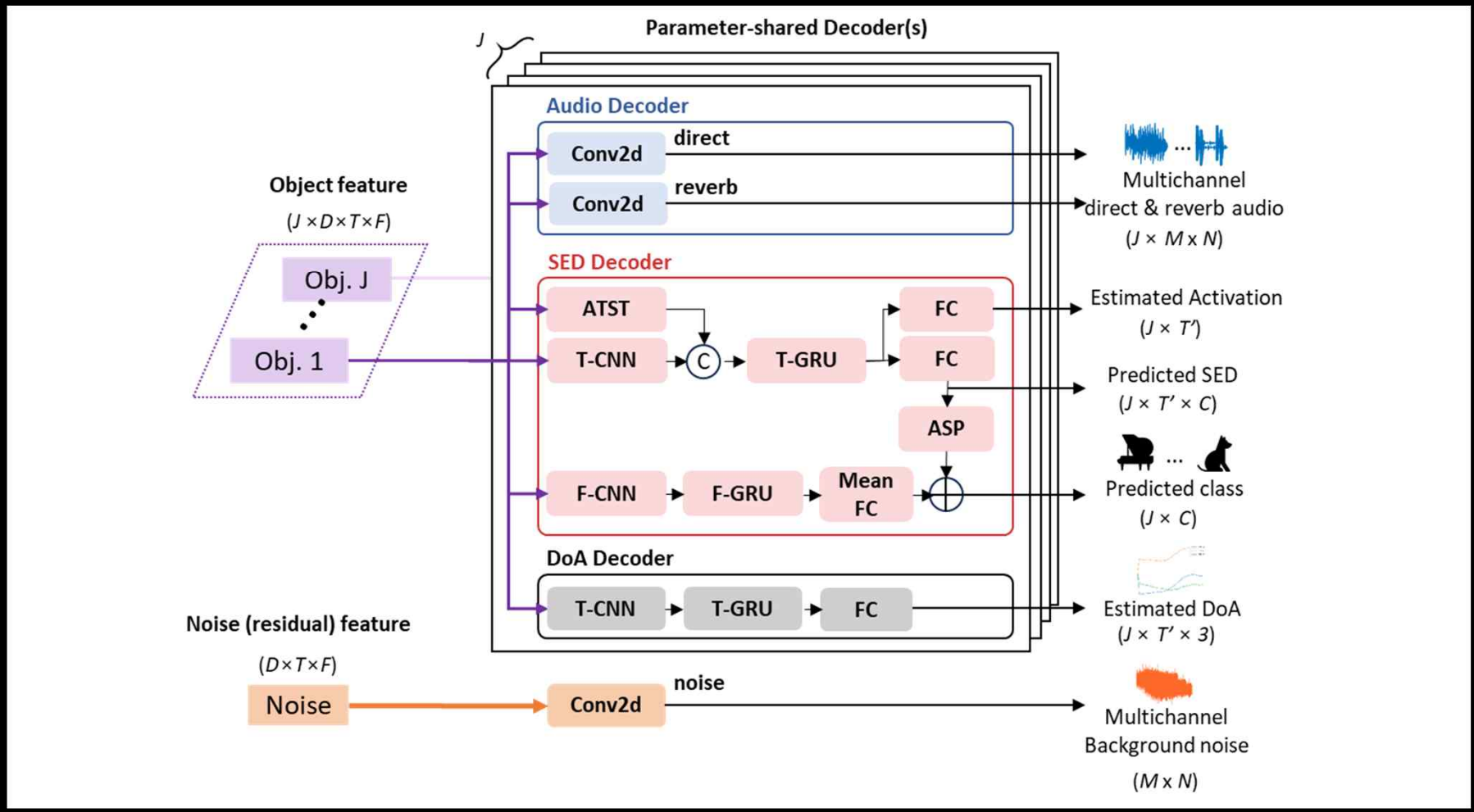
One-for-all model for SELD & SEP tasks

KAIST EE

1. **Object-oriented Processing (OOP)**
   - Inheriting the OOP structure of DeFT-Mamba
   - Less permutation ambiguity
2. **Multichannel-to-multichannel Separation**
   - Separation / dereverberation in multichannel format for spatial audio rendering
3. **SELD + Separation**
   - SEP, Dereverberation, DOAE, SED, SC in a single model

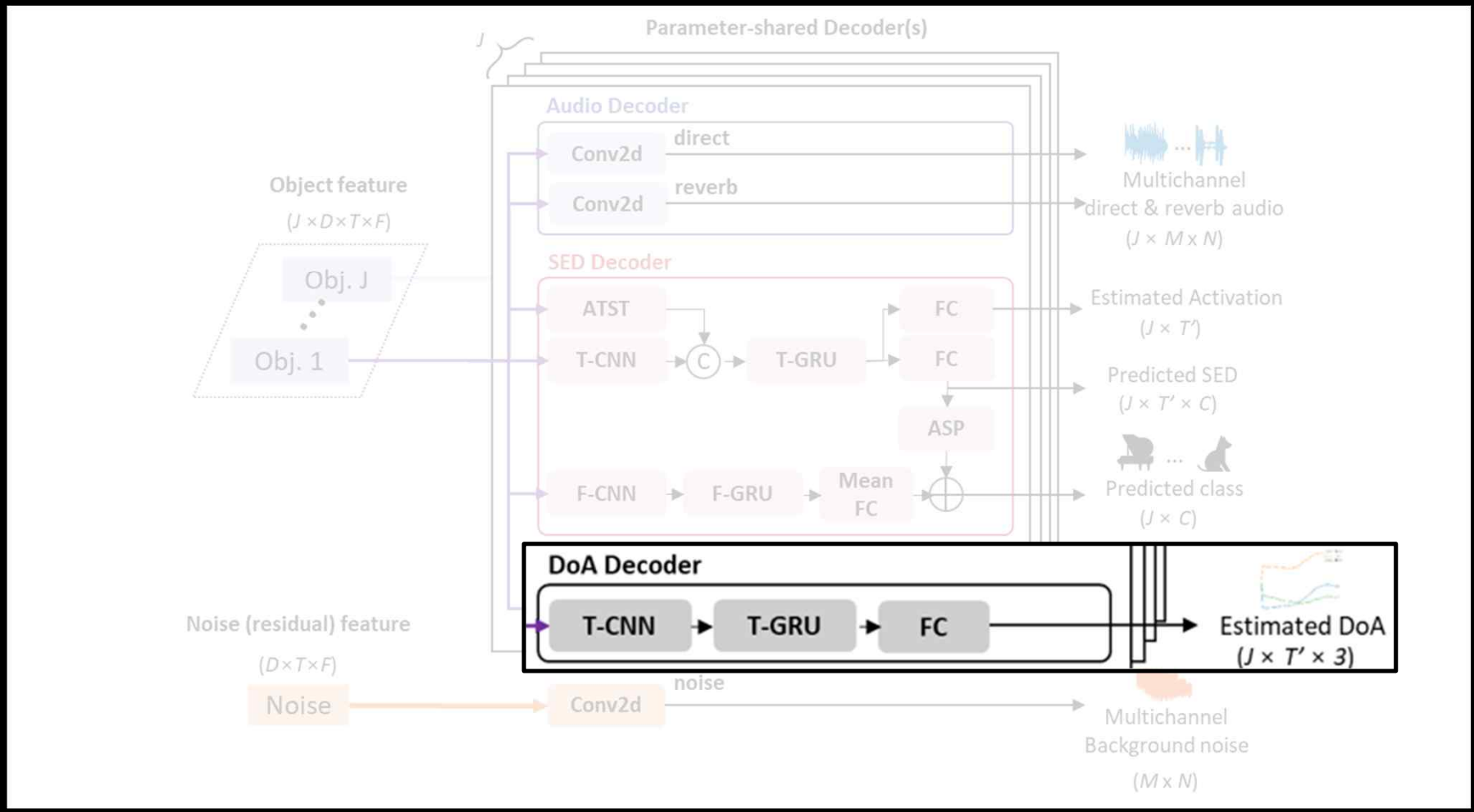
Parameter-shared Decoders

KAIST EE



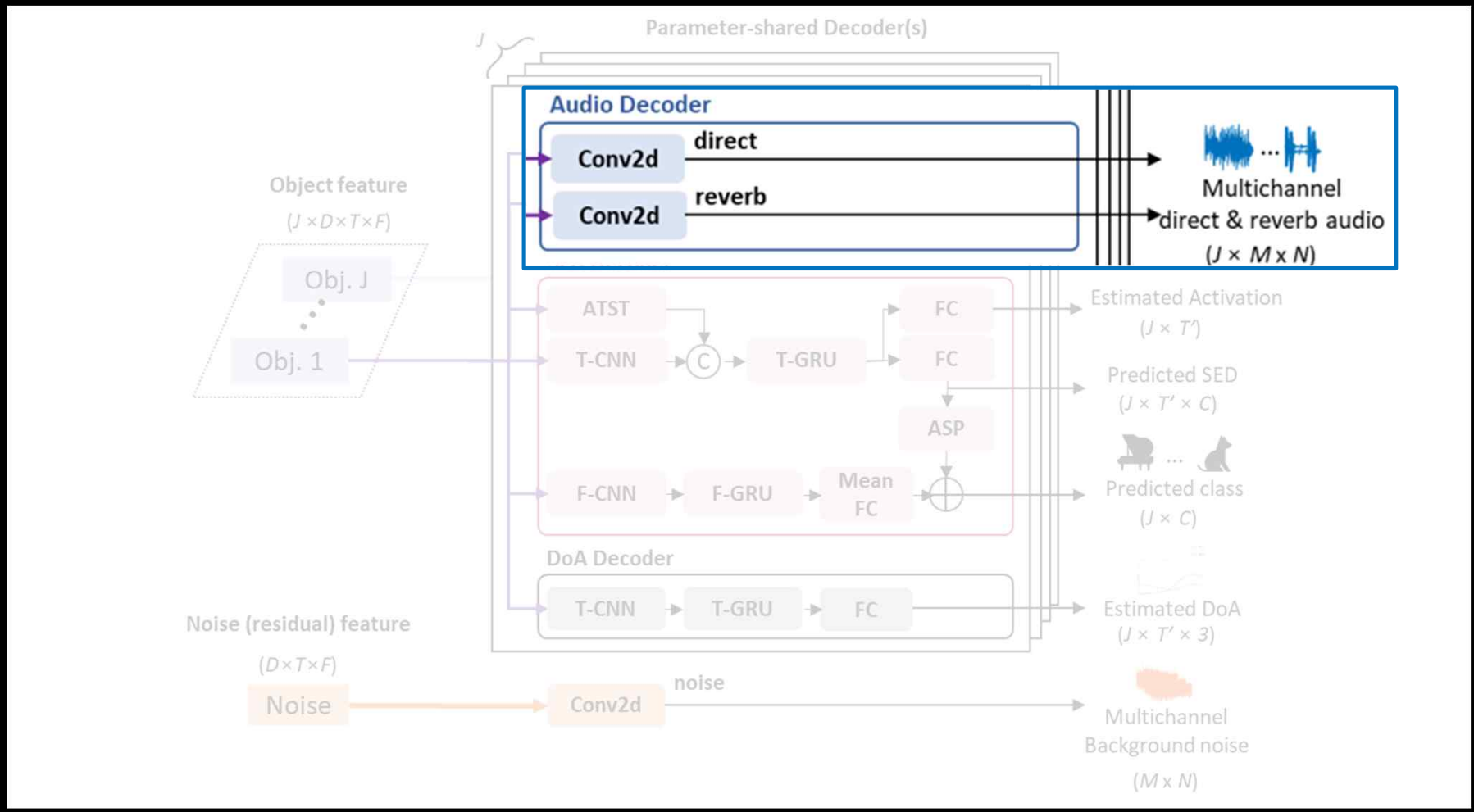
Parameter-shared Decoders – DoA Decoder

KAIST EE



Parameter-shared Decoders – Audio Decoder

KAIST EE



Effect of direct and reverb decoder on DoAE

KAIST EE

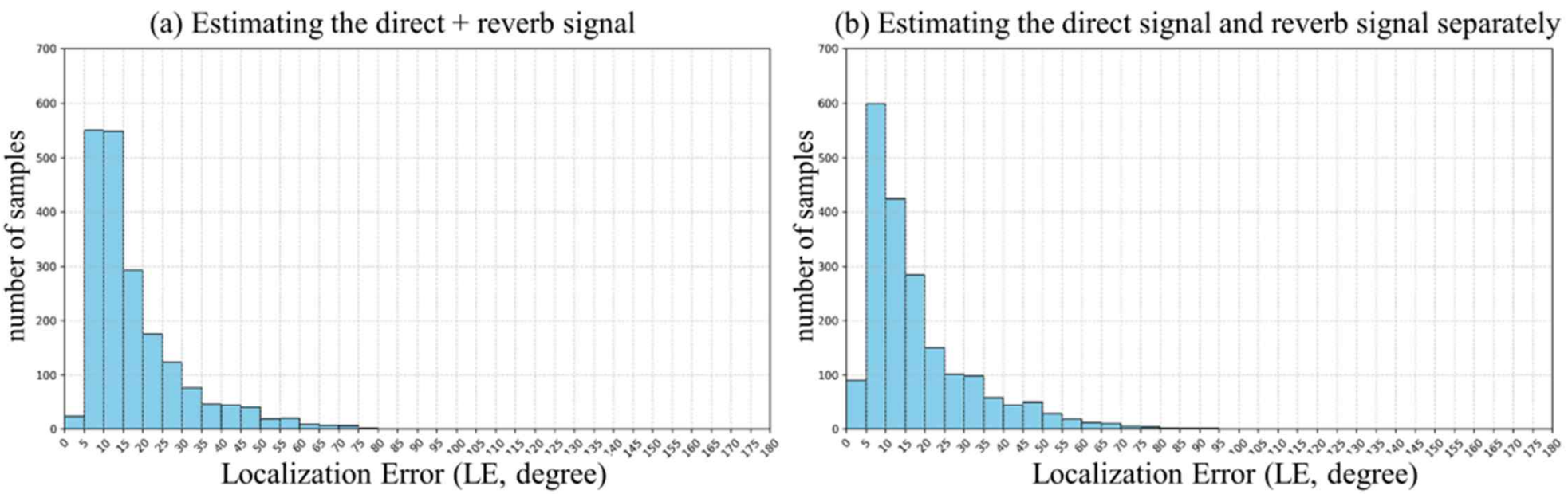


Figure 8: Comparison of LE histograms when estimating (a) direct + reverb signals together and (b) direct and reverb separately.

Effect of direct and reverb decoder on DoAE

KAIST EE

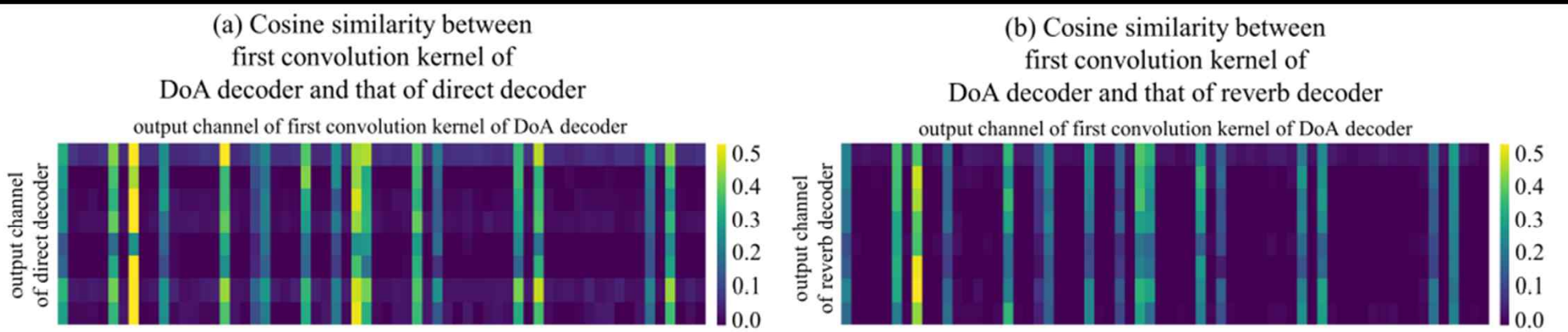
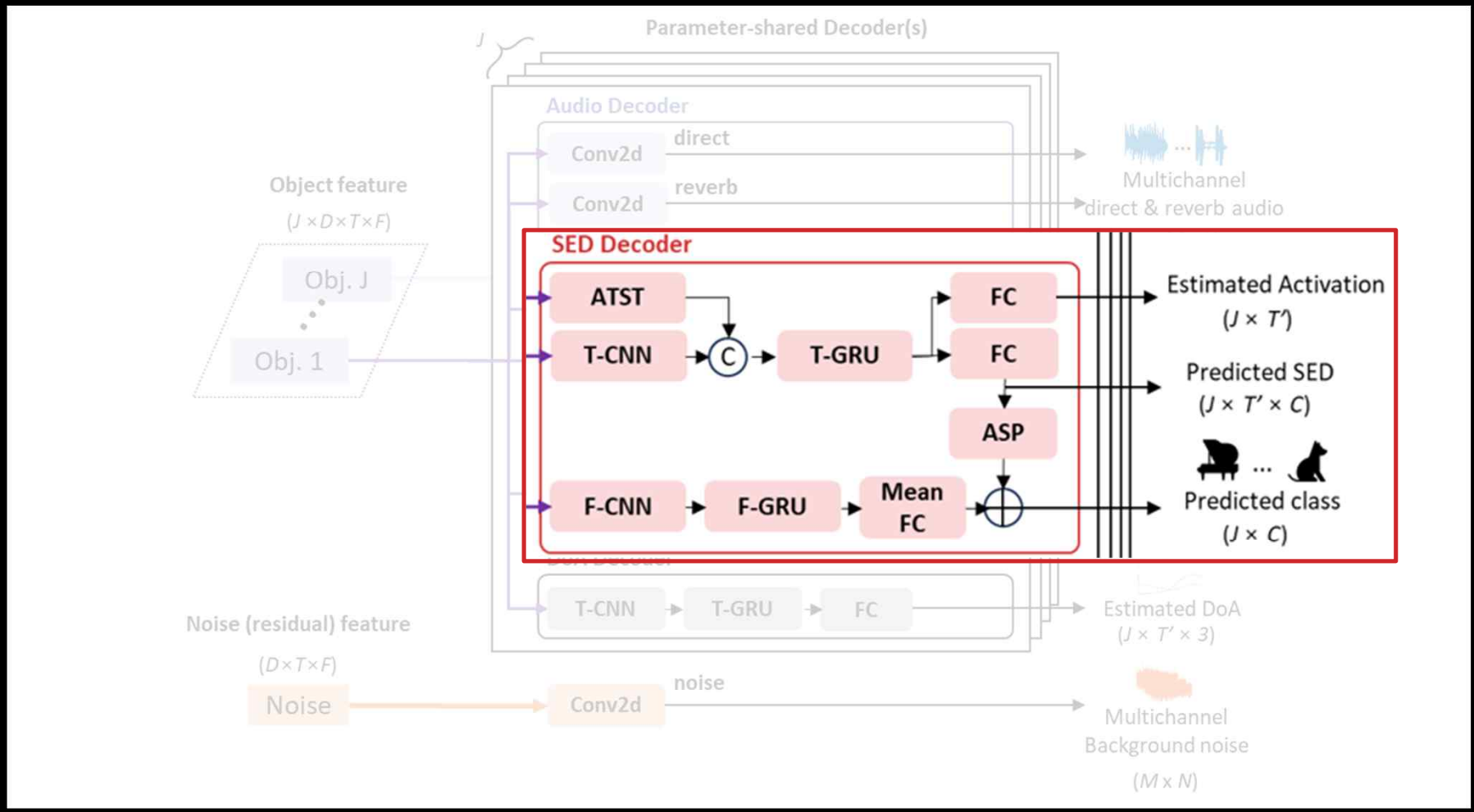


Figure 9: Comparison of cosine similarity between the weight of the first convolution kernel of the DoA decoder and the convolution kernel weights of (a) direct decoder and (b) reverb decoder

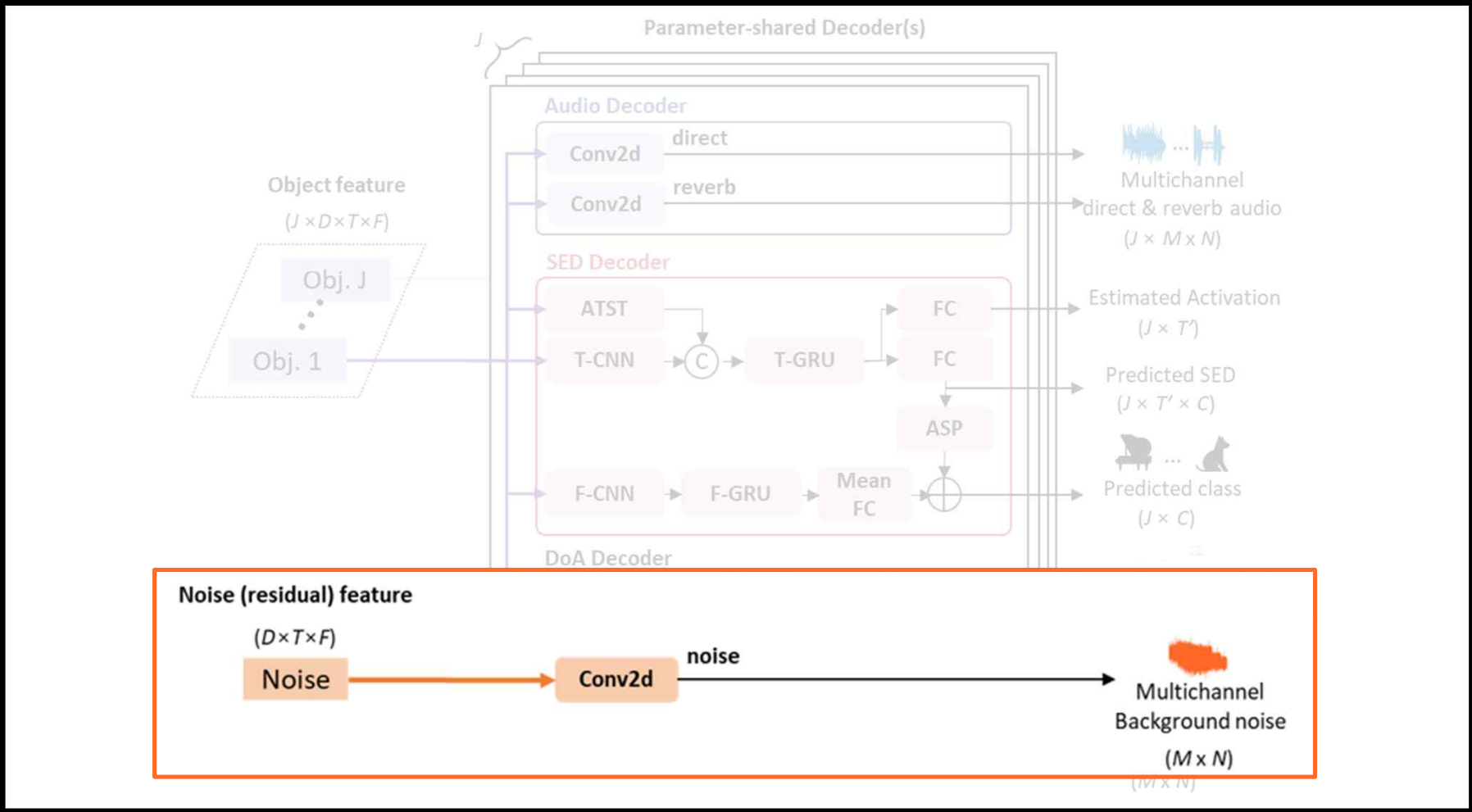
Parameter-shared Decoders – SED Decoder

KAIST EE



Parameter-shared Decoders – Noise Decoder

KAIST EE



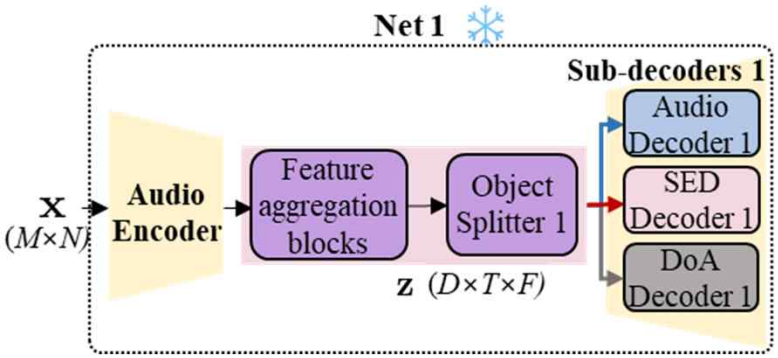
One-for-all model for SELD & SEP tasks

KAIST EE

1. **Object-oriented Processing (OOP)**
   - Inheriting the OOP structure of DeFT-Mamba
   - Less permutation ambiguity
2. **Multichannel-to-multichannel Separation**
   - Separation / dereverberation in multichannel format for spatial audio rendering
3. **SELD + Separation**
   - SEP, Dereverberation, DOAE, SED, SC in a single model
4. **Chain-of-Inference (CoI)**
   - Iterative refinement from the previous erroneous estimation

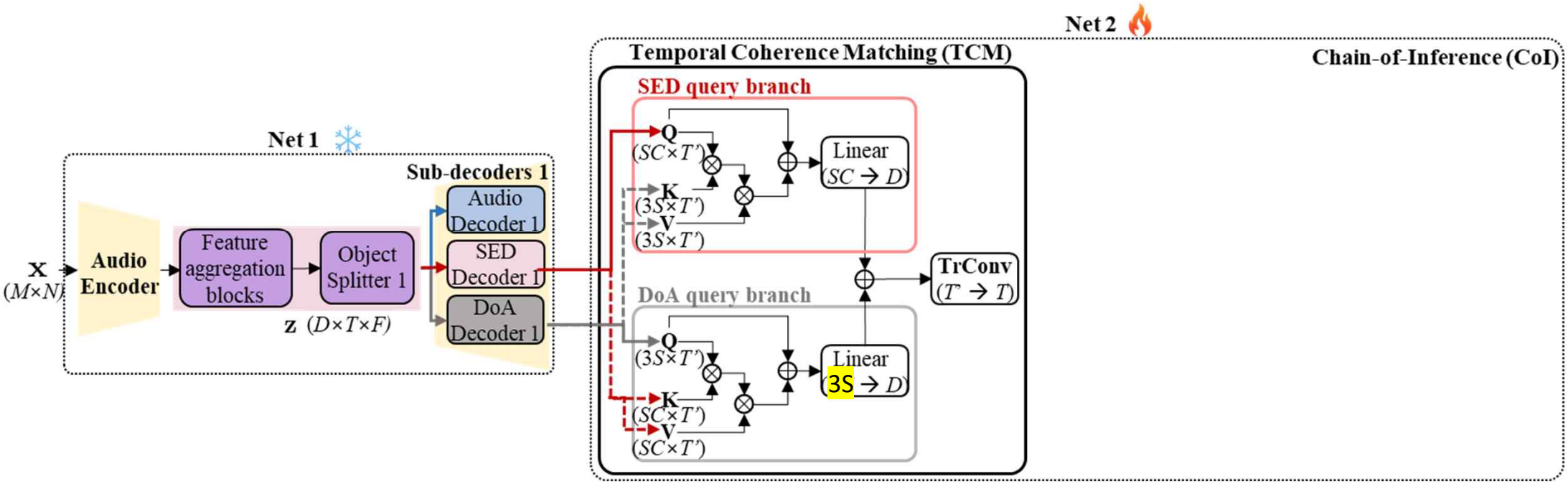
Chain-of-Inference (CoI)

KAIST EE



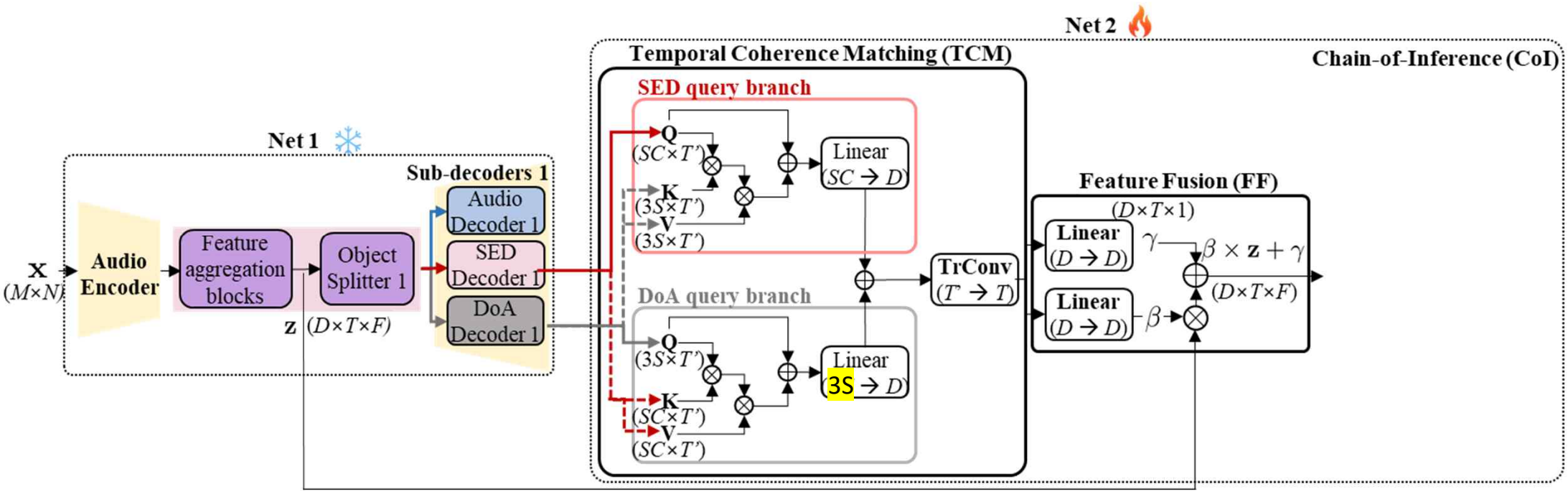
Chain-of-Inference (CoI)

KAIST EE



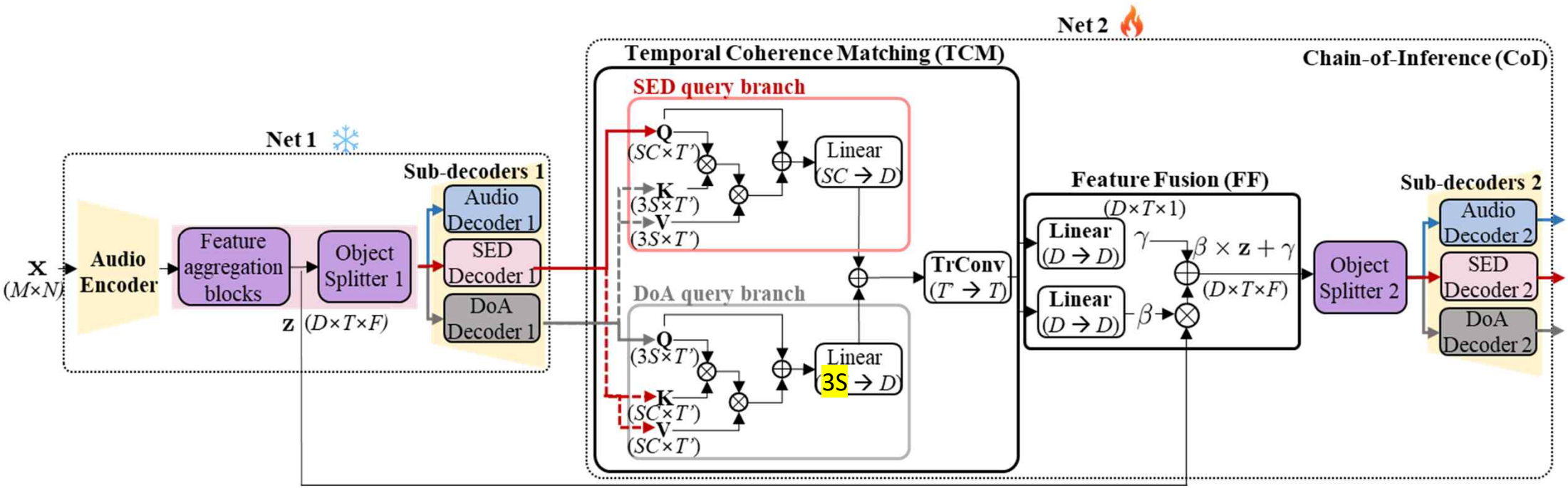
Chain-of-Inference (CoI)

KAIST EE



Chain-of-Inference (CoI)

KAIST EE



Chain-of-Inference (CoI)

KAIST EE

Table 8: Performance comparison by repeating chain-of-inference

Stage	USS		SED		DoAE		SELD ↓	Complexities	
	SI-SDRi ↑	SDRi ↑	ER ↓	F1 ↑	LE ↓	LR ↑		Param.	MAC/s
without CoI	11.0	11.7	28.8	70.2	18.5	76.9	0.230	8.2 (+96.8) M	99.1 G
CoI (1st stage)	**11.2**	**12.0**	**25.0**	**74.1**	**17.0**	**78.1**	**0.206**	12.1 (+96.8) M	104.0 G
CoI (2nd stage)	11.1	11.9	26.8	72.3	17.2	77.5	0.216	16.0 (+96.8 M)	118.9 G

One-for-all model for SELD & SEP tasks

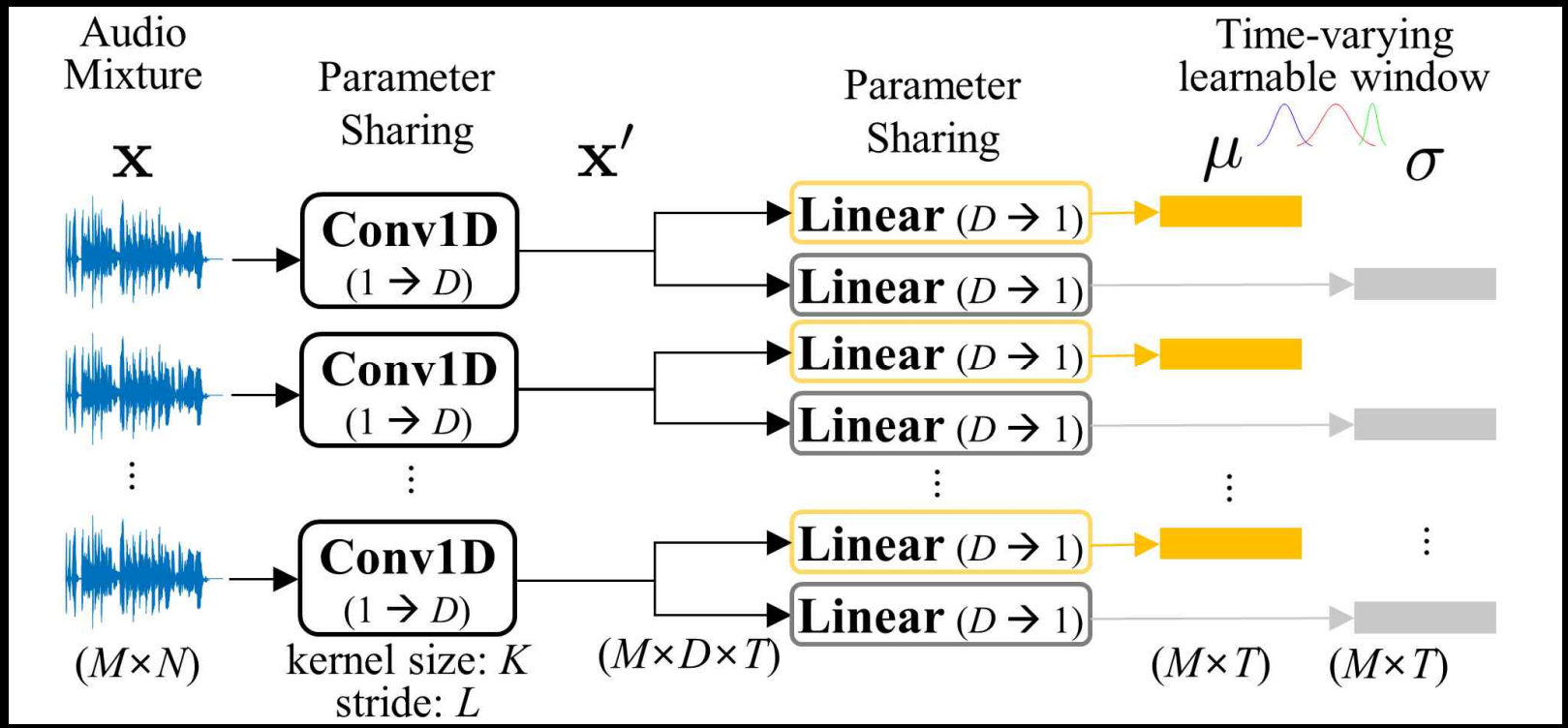
KAIST EE

1. **Object-oriented Processing (OOP)**
   - Inheriting the OOP structure of DeFT-Mamba
   - Less permutation ambiguity
2. **Multichannel-to-multichannel Separation**
   - Separation / dereverberation in multichannel format for spatial audio rendering
3. **SELD + Separation**
   - SEP, Dereverberation, DOAE, SED, SC in a single model
4. **Chain-of-Inference (CoI)**
   - Iterative refinement from the previous erroneous estimation
5. **Adaptive STFT**
   - Partially learnable STFT adapting to input waveform

Adaptive STFT

KAIST EE

- Variable Gaussian time window applied before STFT



Adaptive STFT

KAIST EE

- Variable Gaussian time window applied before STFT

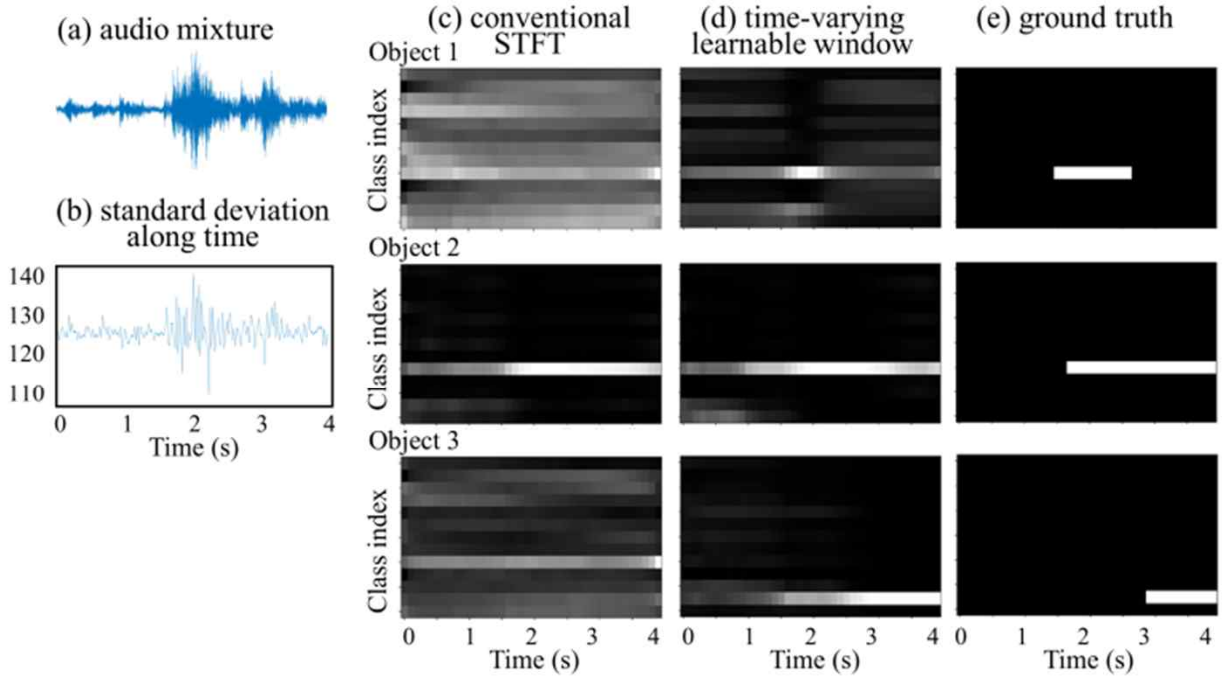


Figure 11: (a) Time-domain waveform of the audio mixture, (b) window length along the time frame, SED results for (c) conventional STFT, (d) time-varying learnable window, (e) ground truth.

Adaptive STFT

KAIST EE

- Variable Gaussian time window applied before STFT

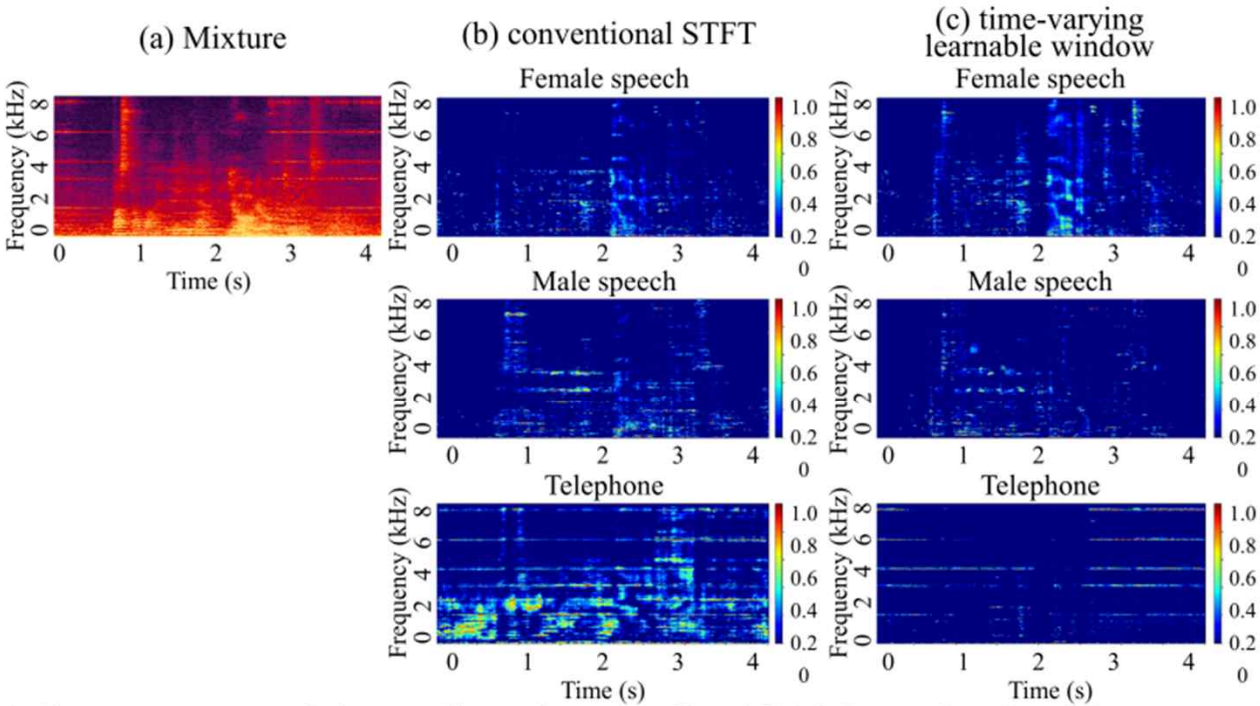


Figure 12: (a) Spectrogram of the audio mixture, GradCAM results for (b) conventional STFT, (c) time-varying learnable window.

Ablation studies (ASA2 dataset)

KAIST EE

Table 1: Ablation study on the proposed model using the ASA2 dataset. The models with (+) or (-) indicate the addition or removal of the corresponding blocks from the baseline. The (+) in the parameter size indicates the parameter size of the pre-trained ATST.

Model variation		USS		SED		DoAE		SELD ↓	Complexities	
		SI-SDRi ↑	SDRi ↑	ER ↓	F1 ↑	LE ↓	LR ↑		Param.	MAC/s
Framework	DeFT-Mamba-MISO [38]	10.4	11.3	-	-	-	-	-	**3.6 M**	**83.8 G**
	DeFT-Mamba-MIMO	10.0	10.9	-	-	-	-	-	3.6 M	83.8 G
	(+) SELDNet [28]	10.2	11.1	42.0	58.2	28.6	63.2	0.341	5.4 M	86.1 G
	(+) SED, DoA decoder	10.4	11.4	40.0	60.3	22.9	65.7	0.317	7.2 M	88.4 G
Object separator	(-) Unfold	10.3	11.4	39.8	60.5	22.2	66.1	0.314	7.2 M	88.0 G
	(-) Unfold, F-Mamba	10.0	11.1	39.9	60.4	22.0	65.9	0.315	6.3 M	77.1 G
SED decoder	(+) F-CRNN	10.3	11.2	38.0	61.1	21.9	68.8	0.301	8.1 M	90.7 G
	(+) ATST + T-CRNN [45]	10.2	11.2	35.7	66.1	21.2	71.0	0.276	6.3 (+96.8) M	96.5 G
	(+) ATST + T- &F-CRNNs	10.3	11.2	34.1	66.6	21.3	72.8	0.266	8.1 (+96.8) M	98.8 G
Audio decoder	(+) Noise decoder	11.0	11.7	30.3	69.8	21.2	76.0	0.241	8.1 (+96.8) M	98.9 G
	(+) Direct/reverb, noise decoder	10.8	11.5	30.0	69.6	19.1	76.2	0.237	8.1 (+96.8) M	99.0 G
Dynamic STFT	(+) Time-invariant window	10.7	11.4	30.2	69.8	19.3	76.0	0.238	8.1 (+96.8) M	99.0 G
	(+) Time-variant window	11.0	11.7	28.8	70.2	18.5	76.9	0.230	8.2 (+96.8) M	99.1 G
CoI	**(+) Chain-of-inference**	**11.2**	**12.0**	**25.0**	**74.1**	**17.0**	**78.1**	**0.206**	12.1 (+96.8) M	104.0 G

Ablation studies (ASA2 dataset)

KAIST EE

Table 2: Ablation study for the chain-of-inference architecture

Model variation	USS		SED		DoAE		SELD ↓	Complexities	
	SI-SDRi ↑	SDRi ↑	ER ↓	F1 ↑	LE ↓	LR ↑		Param.	MAC/s
without chain-of-inference	11.0	11.7	28.8	70.2	18.5	76.9	0.230	8.2 (+96.8) M	99.1 G
without audio decoder 2 (+) SED branch	11.0	11.8	26.6	71.8	18.2	76.5	0.221	10.3 (+96.8) M	101.6 G
without audio decoder 2 (+) DoA branch	11.0	11.8	28.2	70.6	17.6	76.0	0.228	10.3 (+96.8) M	101.6 G
without audio decoder 2 (+) SED & DoA	11.0	11.7	26.5	73.0	17.3	77.2	0.214	12.1 (+96.8) M	104.0 G
+ Chain-of-inference	**11.2**$_{\pm 0.1}$	**12.0**$_{\pm 0.1}$	**25.0**$_{\pm 0.4}$	**74.1**$_{\pm 0.3}$	**17.0**$_{\pm 0.3}$	**78.1**$_{\pm 0.4}$	**0.206**$_{\pm 0.001}$	12.1 (+96.8) M	104.0 G

SELD performance

KAIST EE

- Pretrain on ASA2 → Finetune on STARSS23

SELD Models (STARSS23 dataset)	ER↓	F1↑	LE↓	LR↑	SELD↓
CST-former	59	42.6	20.5	61.3	0.416
MFF-EINV2	54	42.5	18.7	62.6	0.398
CST-former2	42	59.7	15.6	68.4	0.301
EINV-2 with data augmentation chain (DCASE, 2nd rank)	42	57.5	15.8	72.7	0.301
NERC-SLIP (DCASE, 1st rank)	40	64	13.4	74	0.277
NERC-SLIP (DCASE, 1st rank) +ensemble	38	**66**	12.8	**75**	0.26
DeepASA	**33.7**	63.1	**9.8**	74.6	**0.253**

Source separation performance

KAIST EE

- Pretrain on ASA2 → Finetune on MC-FUSS

USS Model (MC-FUSS dataset)	J=2	J=3	J=4	Total	Param.	MAC/s
ByteDance-uss	14.8	14.4	12.7	14	28.0(+80.7) M	40.1G
MC-BSRNN	15.7	15.2	11.4	14.1	12.2M	**15.3G**
TF-GridNet	17.2	16.1	12.5	15.3	14.7M	462G
DeFTAN-II	17.6	16.3	12.8	15.6	4.1M	66.1G
SpatialNet	17.8	16.5	13.1	15.8	7.3M	71.8G
DeFT-Mamba	18.4	17.1	**13.8**	16.4	4.2M	58.2G
DeepASA	**18.9**	**19.1**	**17.6**	**18.5**	**3.8M**	84.0G

Multi-task performance

KAIST EE

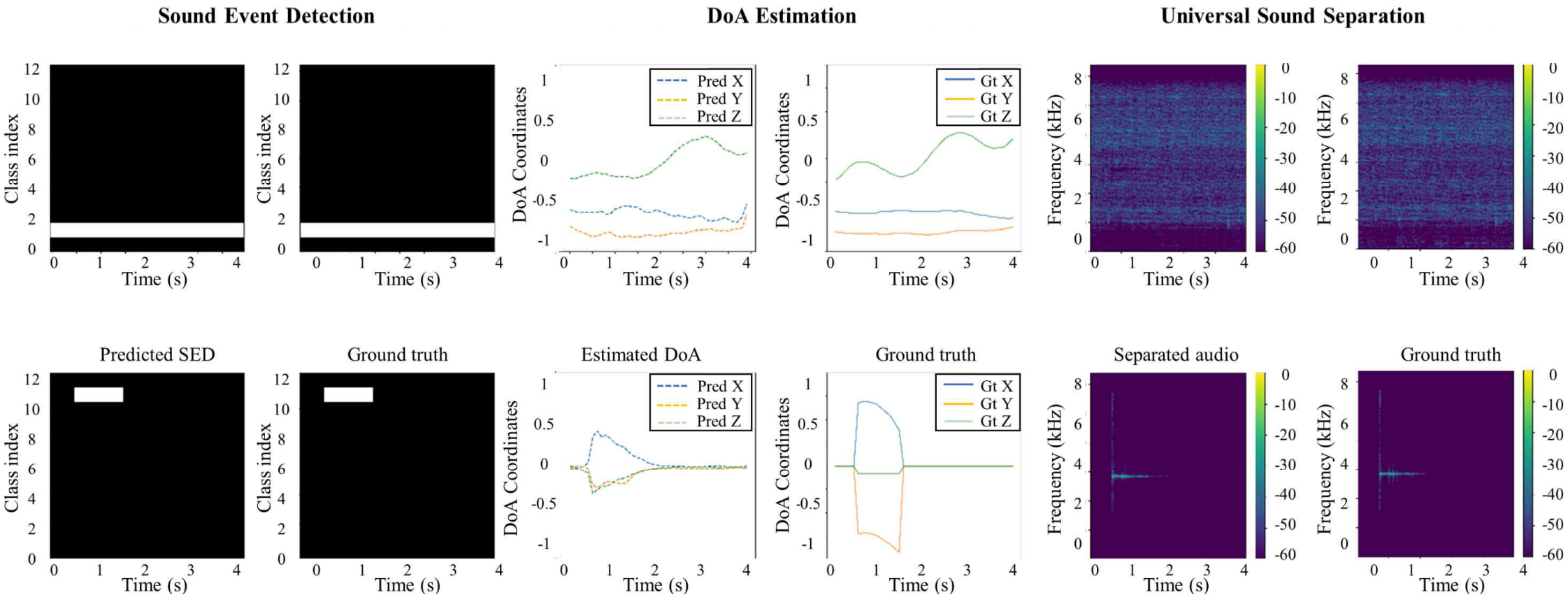
- **Multitask performance on ASA2 dataset**

Table 9: Comparison with SOTA models on ASA2 dataset

model	USS		SED		DoAE		SELD ↓	Complexities	
	SI-SDRi ↑	SDRi ↑	ER ↓	F1 ↑	LE ↓	LR ↑		Param.	MAC/s
ByteDance-uss [7]	5.2	8.3	-	-	-	-	-	28 (+80.7) M	40.1 G
TF-GridNet [47]	8.7	10.7	-	-	-	-	-	14.7 M	462 G
SpatialNet [49]	9.6	10.2	-	-	-	-	-	7.3 M	71.8 G
DeFT-Mamba [38]	10.4	11.3	-	-	-	-	-	3.6 M	83.8 G
EINV2 [53]	-	-	48.5	39.5	27.1	51.5	0.431	51.5 M	6.7 G
ResNet Conformer [59]	-	-	45.7	41.0	26.6	53.7	0.414	13.6 M	7.6 G
SELD-Mamba [55]	-	-	43.5	42.7	25.5	56.7	0.396	75.1 M	4.3 G
MFF-EINV2 [50]	-	-	42.1	43.2	25.8	60.7	0.381	54.8 M	14.3 G
DeepASA (w/o ATST)	11.0	11.7	30.1	69.5	18.5	74.8	0.240	8.2 M	91.0 G
DeepASA	11.0	11.7	28.8	70.2	18.5	76.9	0.230	8.2 (+96.8) M	99.1 G
(+) Chain-of-inference	**11.2**$_{\pm 0.1}$	**12.0**$_{\pm 0.1}$	**25.0**$_{\pm 0.4}$	**74.1**$_{\pm 0.3}$	**17.0**$_{\pm 0.3}$	**78.1**$_{\pm 0.4}$	**0.206**$_{\pm 0.001}$	12.1 (+96.8) M	104.0 G

Classification & Localization

KAIST EE



Hugging Face

NeurIPS

Paper

ASA2 dataset

DeepASA

An Object-Oriented Multi-Purpose Network for Auditory Scene Analysis