MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Neural Fields for Spatial Audio Modeling

Masuyama, Yoshiki

TR2025-171 December 13, 2025

Abstract

• Overview of Spatial Audio and Neural Fields • Neural Fields for Head-Related Transfer Functions • Neural Fields for Room Impulse Responses • Physics-Informed Neural Networks for Room Impulse Responses

Speech and Audio in the Northeast (SANE) 2025

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Neural Fields for Spatial Audio Modeling

Yoshiki Masuyama

November 7, 2025

MITSUBISHI ELECTRIC RESEARCH LABORATORIES (MERL)
Cambridge, Massachusetts, USA
http://www.merl.com

Collaborators on today's topics

Jonathan
Le Roux

Gordon
Wichern

Chiori
Hori

Christoph
Boeddeker

Takahiro
Edo

Current MERL
SA team

François
Germain

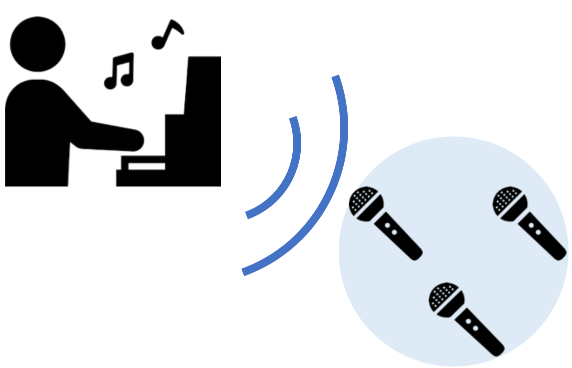
Christopher
Ick

We are hiring interns for next year!

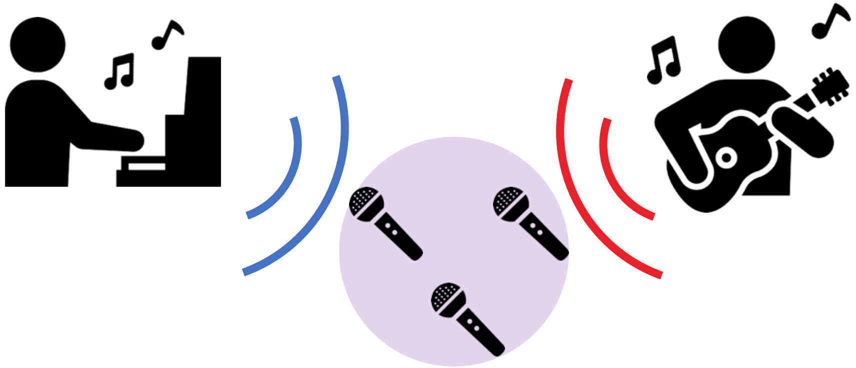
Agenda

- **<u>Overview of Spatial Audio and Neural Fields</u>**
- Neural Fields for Head-Related Transfer Functions
- Neural Fields for Room Impulse Responses
- Physics-Informed Neural Networks for Room Impulse Responses

Spatial Audio Technologies and Applications

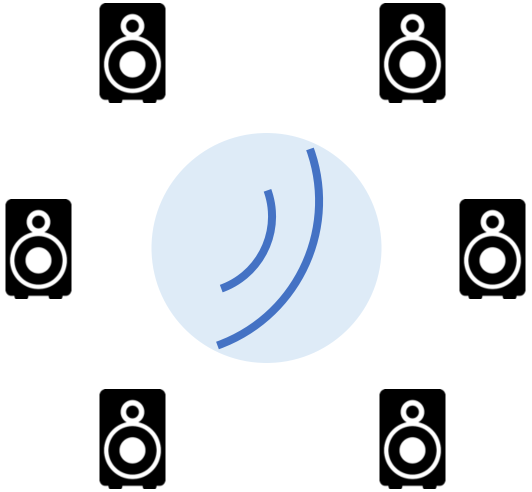


Sound field estimation

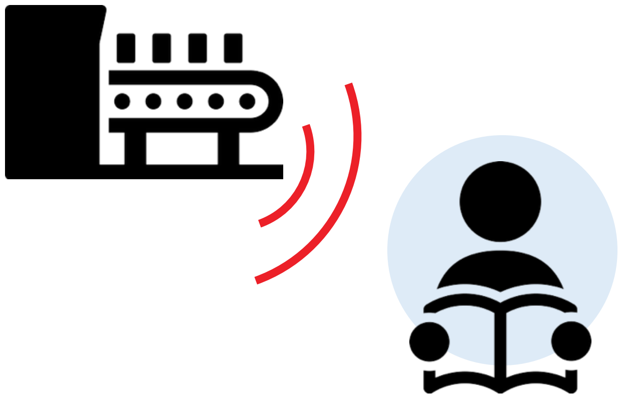


Sound field decomposition

Binaural rendering



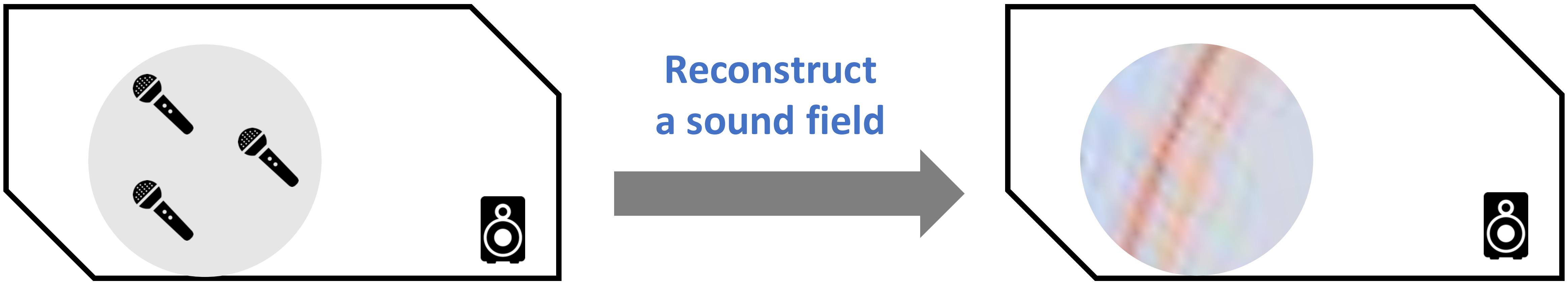
Sound field synthesis



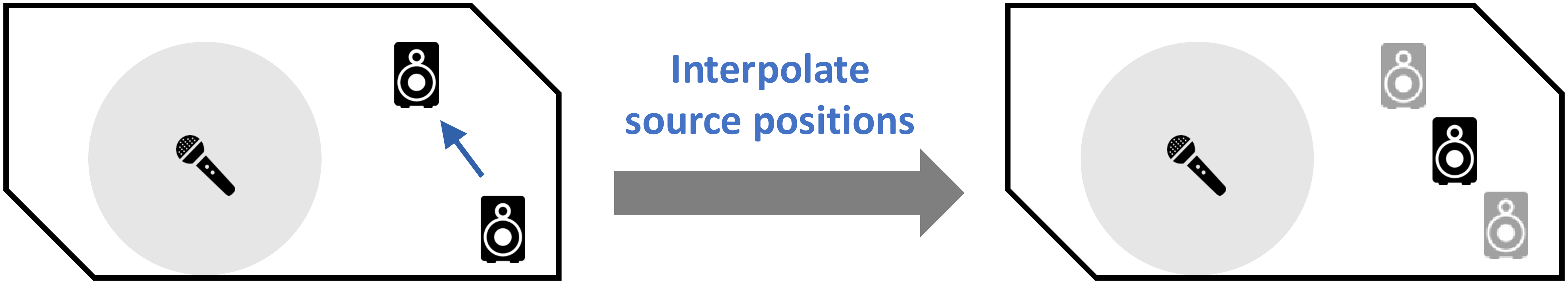
Active noise control

Sound Field Estimation and Beyond

- Sound field estimation predicts continuous pressure from a handful of measurements.

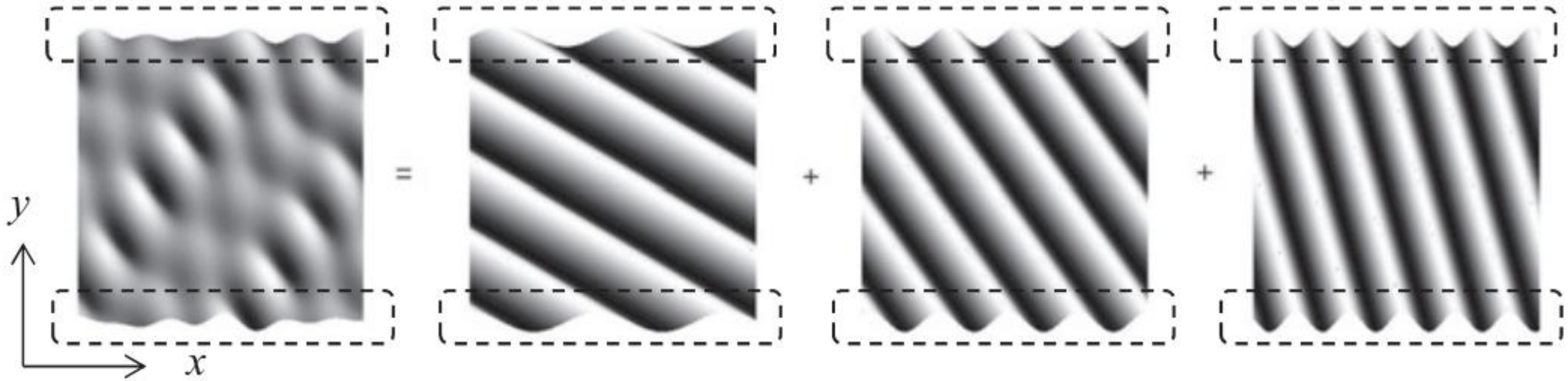


- We are also interested in interpolating sound source positions.

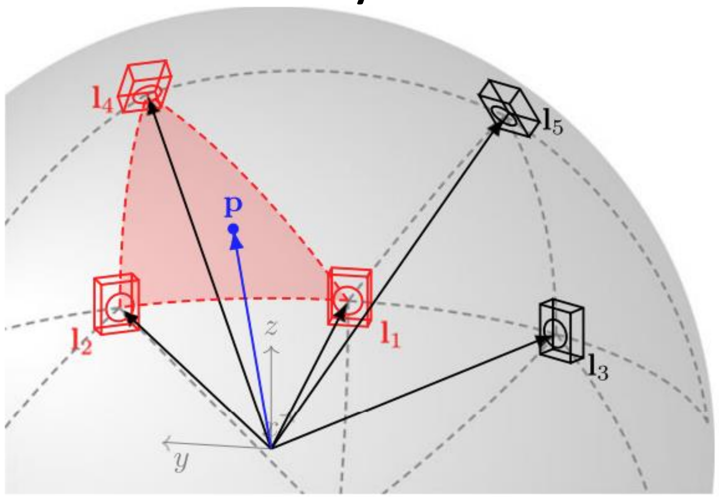


Signal-Processing-Based Approach for Sound Field Estimation

- Sound field estimation is typically realized by decomposing the field into basis functions.
  - Spherical harmonic expansion [Abhayapala+2002], **Plane-wave decomposition** [Rafaely+2004]



- Panning has been widely used to interpolate source positions [Pulkki+1997].

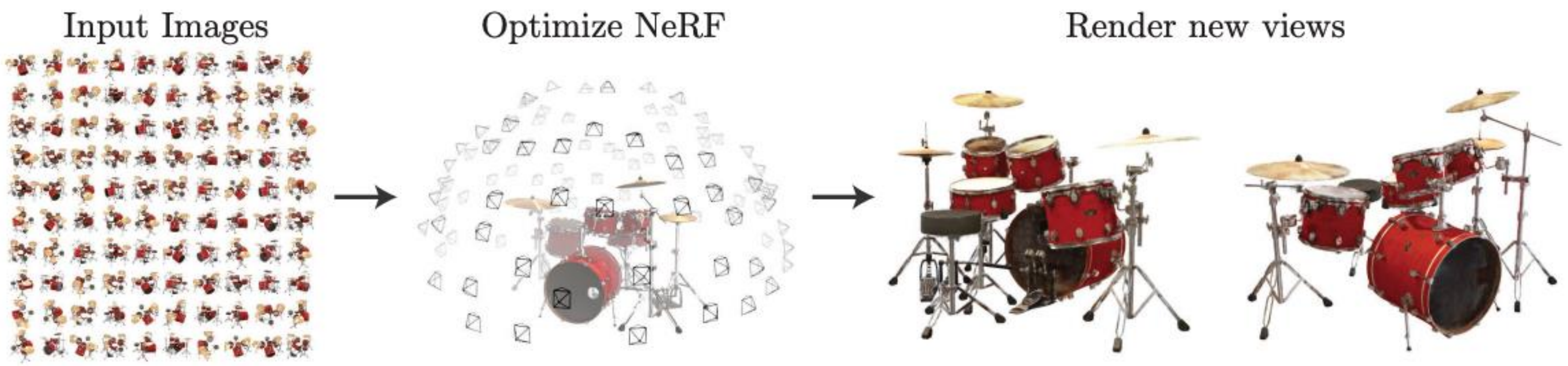


These methods typically require dense measurements.

Figures are from [Haneda2018] and [Franck+2017].

Evolution of Neural Fields (NFs) and Their Application to Spatial Audio

- A neural field is a quantity on coordinates parameterized by a neural network.
  - NeRF predicts color and density from a given 3D camera position and a 2D direction.



[Mildenhall+2020]

- Initial spatial audio works used **an NF to represent the sound pressure of a single scene.**

$\mathbf{r} = [x, y, z]^\mathsf{T}, t \longrightarrow$ **Neural Field** $\longrightarrow \circledast \longrightarrow p(\mathbf{r}, t)$ Reverberant signal

Microphone position, time

$s(t)$ Source signal

Deep impulse response (Deep IR)

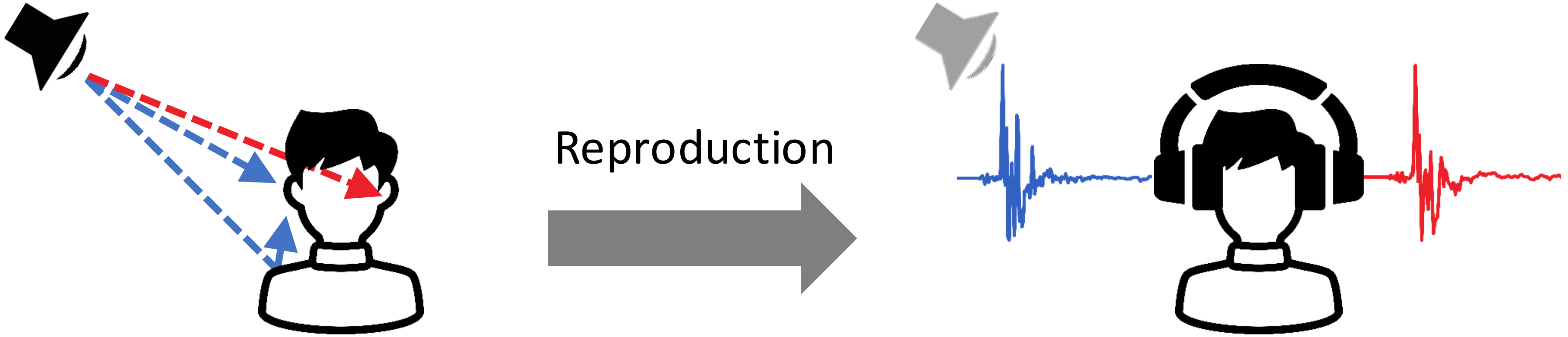
[Richard+2022]

Agenda

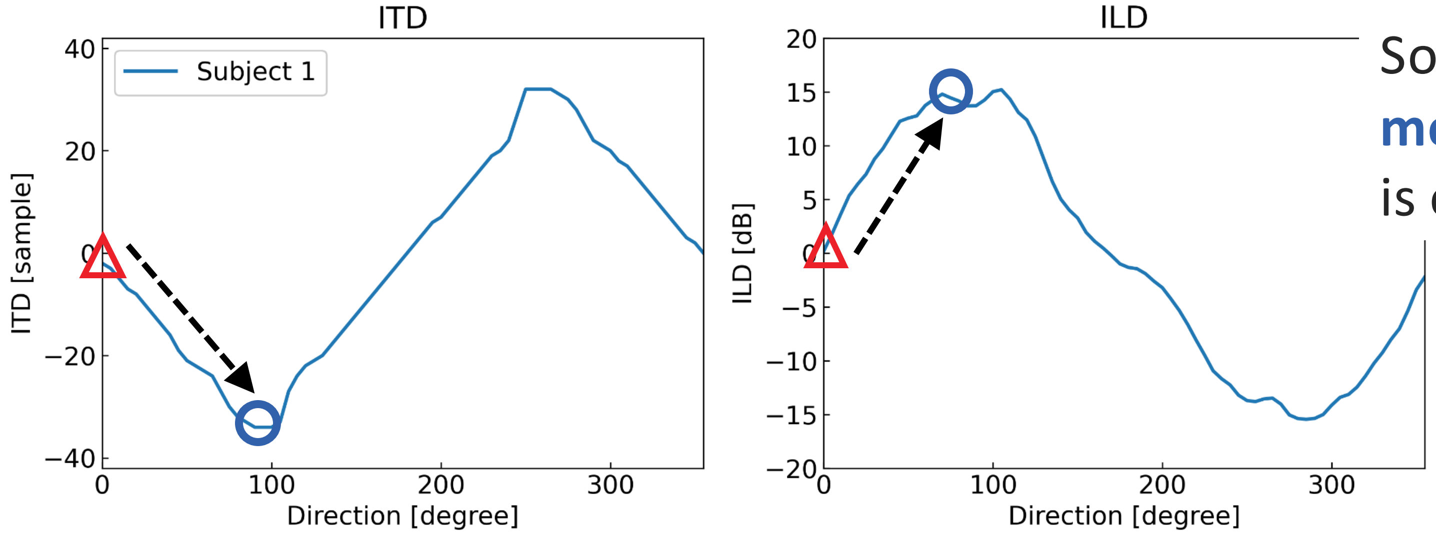
- Overview of Spatial Audio and Neural Fields
- **<u>Neural Fields for Head-Related Transfer Functions</u>**
- Neural Fields for Room Impulse Responses
- Physics-Informed Neural Networks for Room Impulse Responses

Head-Related Transfer Function (HRTF) for Immersive Audio

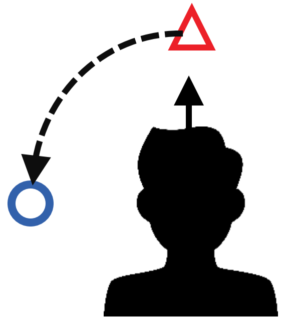
- HRTF represents the acoustic transfer function from the sound source to each ear.



- Azimuth localization relies on interaural time difference and interaural level difference.



Sound reaches the left ear **faster** with **more energy** when the sound source is on the left.



- Elevation localization primarily relies on prominent spectral peaks and notches.

HRTF Spatial Upsampling and Personalization

- Measuring individual HRTFS with dense spatial grids is ideal but **time-consuming.**

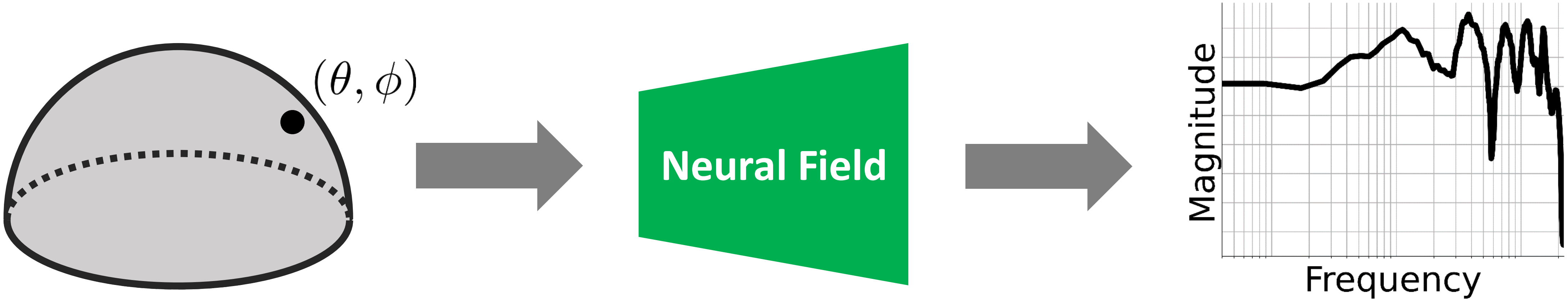
We need to measure HRTFs (impulse responses) several hundreds to thousands of times.

- Our objective is to estimate the HRTF at an unseen direction from sparse measurements.

Figure is from https://www.sonicom.eu/simplifying-hrtf-measurements-using-machine-learning/

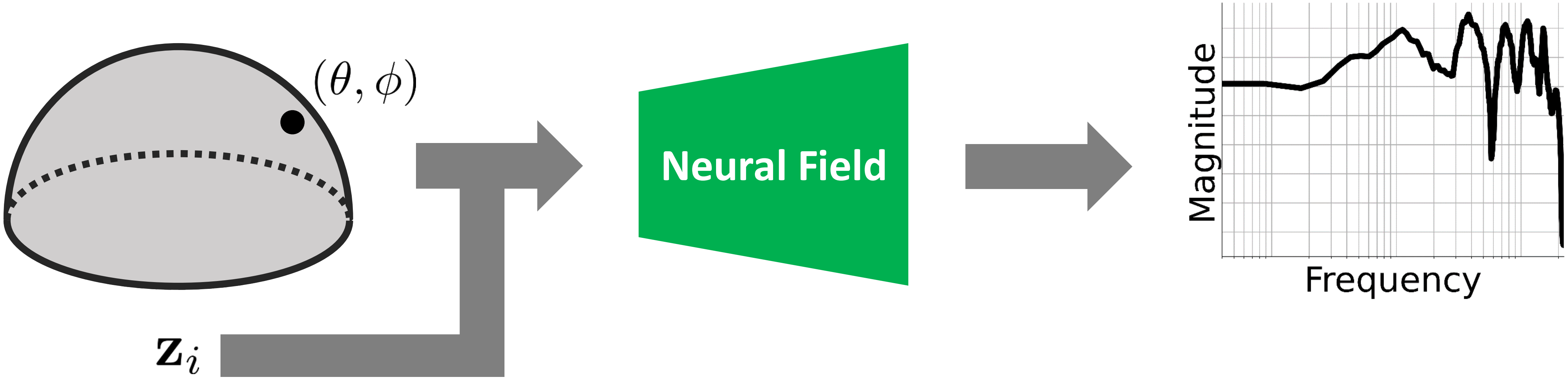
HRTF Field [Zhang+2023]

- **HRTF field models a map from the sound source direction to HRTF magnitude.**
  - Vanilla NF is trained for each subject to reconstruct the measured HRTFs from the direction.



HRTF Field [Zhang+2023]

- **HRTF field models a map from the sound source direction to HRTF magnitude.**
  - Vanilla NF is trained for each subject to reconstruct the measured HRTFs from the direction.



- The authors share the NF across subjects and condition it by **subject-specific parameters.**
  - Subject-specific vector z_i is concatenated to the input like a prompt.

- **The NF is pre-trained on HRTFs for many subjects** and then adapted to the target subject.

Parameter-Efficient Fine-Tuning (PEFT) for HRTF Field [Masuyama+2024]

- We explore other conditioning approaches, e.g., FiLM.
  - Subject-specific modulation is computed from the latent vector $\boldsymbol{z}_i$ by another small network.

$$\mathtt{FiLM}(\mathbf{x}_l \mid \underline{i}) = \boldsymbol{\sigma}_{l,i} \odot \mathtt{Act}(\mathbf{A}_l \mathbf{x}_l + \mathbf{b}_l) + \boldsymbol{\mu}_{l,i}$$

Layer index Subject index Subject-specific

- We also take inspiration from **parameter-efficient fine-tuning (PEFT)** for LLMs.
  - BitFit incorporates subject-specific biases for multiple fully-connected layers.

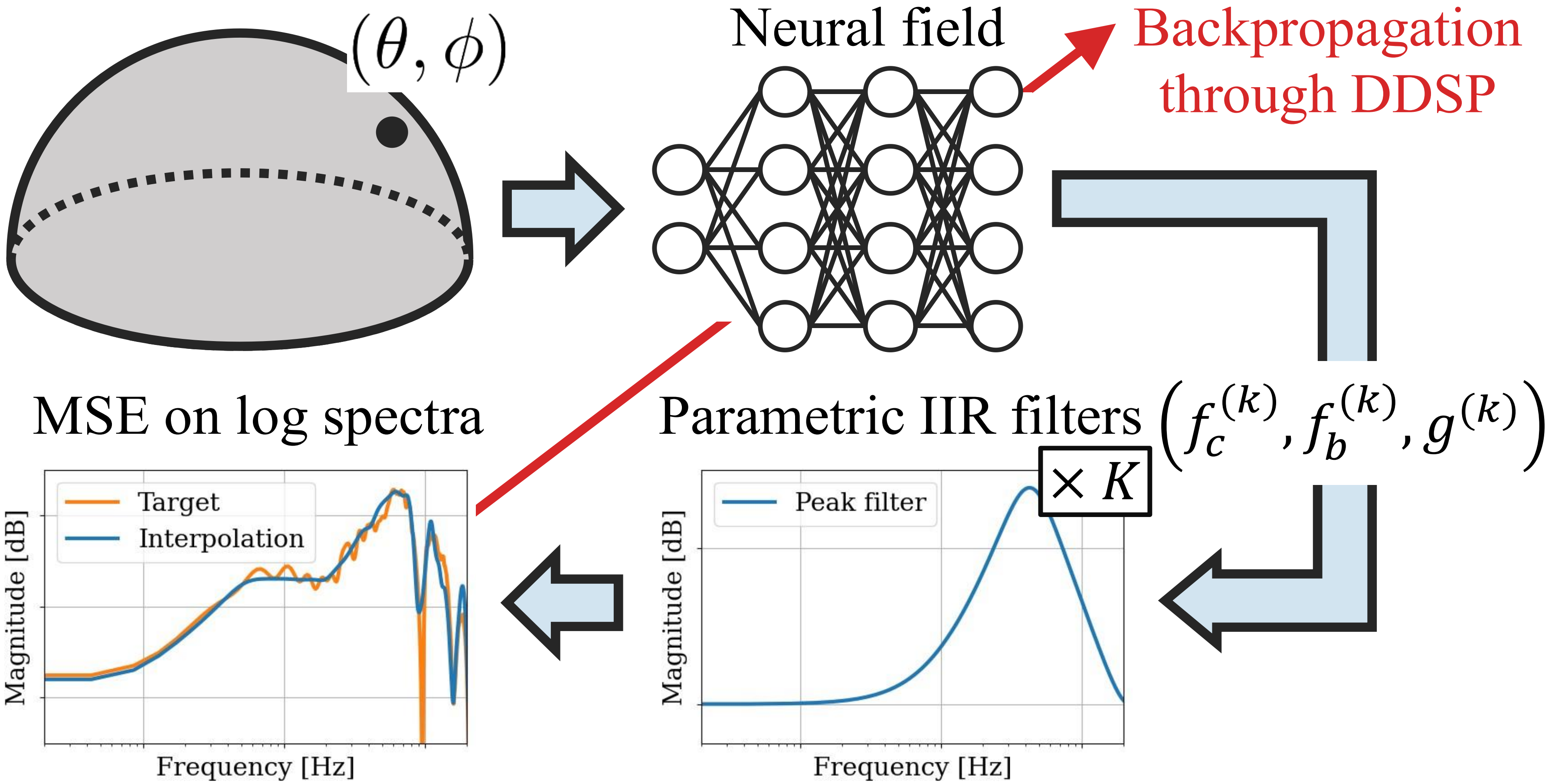
$$\mathtt{BitFit}(\mathbf{x}_l \mid i) = (\mathbf{A}_l \mathbf{x}_l + \underbrace{\mathbf{b}_{l,i}}_{\text{Subject-specific}})$$

  - LoRA employs subject-specific low-rank weights for each fully-connected layer.

$$\mathtt{LoRA}(\mathbf{x}_l \mid i) = (\mathbf{A}_l \mathbf{x}_l + \underbrace{\mathbf{u}_{l,i} \mathbf{v}_{l,i}^{\mathsf{T}}}_{\text{Subject-specific}} \mathbf{x}_l + \mathbf{b}_l)$$

Neural IIR Filter Field (NIIRF) [Masuyama+2024]

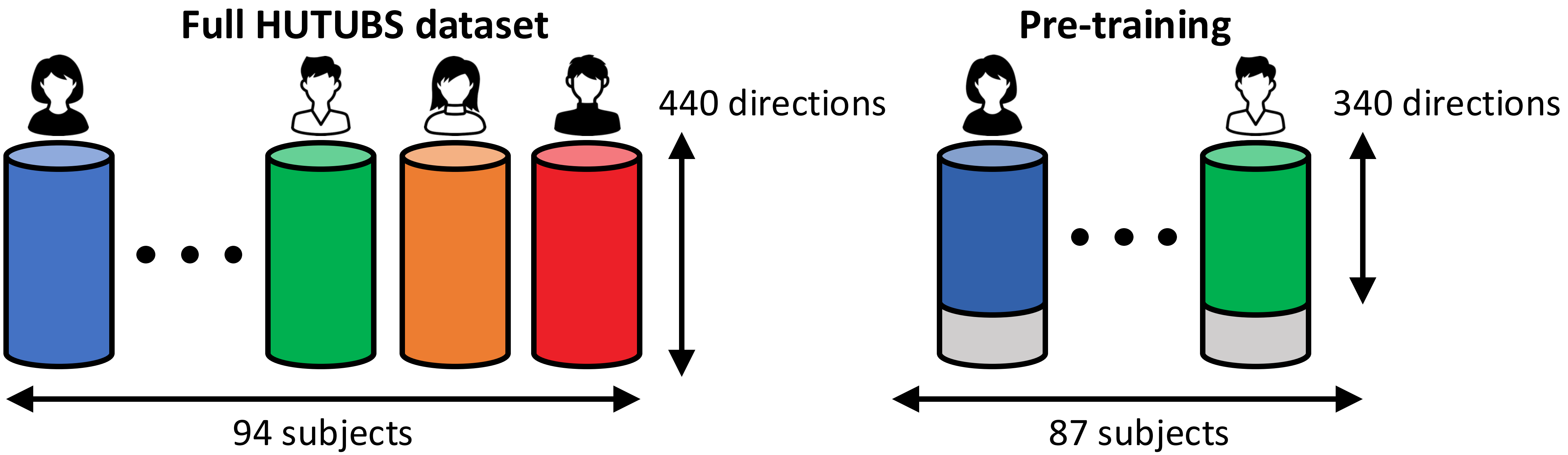
- We propose to integrate **a neural field** and **cascaded parametric IIR filters [Ramos+2013].**
- IIR filters can approximate HRTFs with **fewer coefficients** and **reduce memory footprint.**
- We can **use backpropagation thanks to differentiable DSP (DDSP)** for the IIR filters.



Experiments for Different Adaptation Techniques

- We used the HUTUBS dataset [Brinkmann+2019] to analyze the adaptation techniques.
  - It consists of HRTFs for 94 subjects, where 440 directions are measured for each subject.
  - We used 87 subjects for pre-training and the remaining 7 subjects for evaluation.

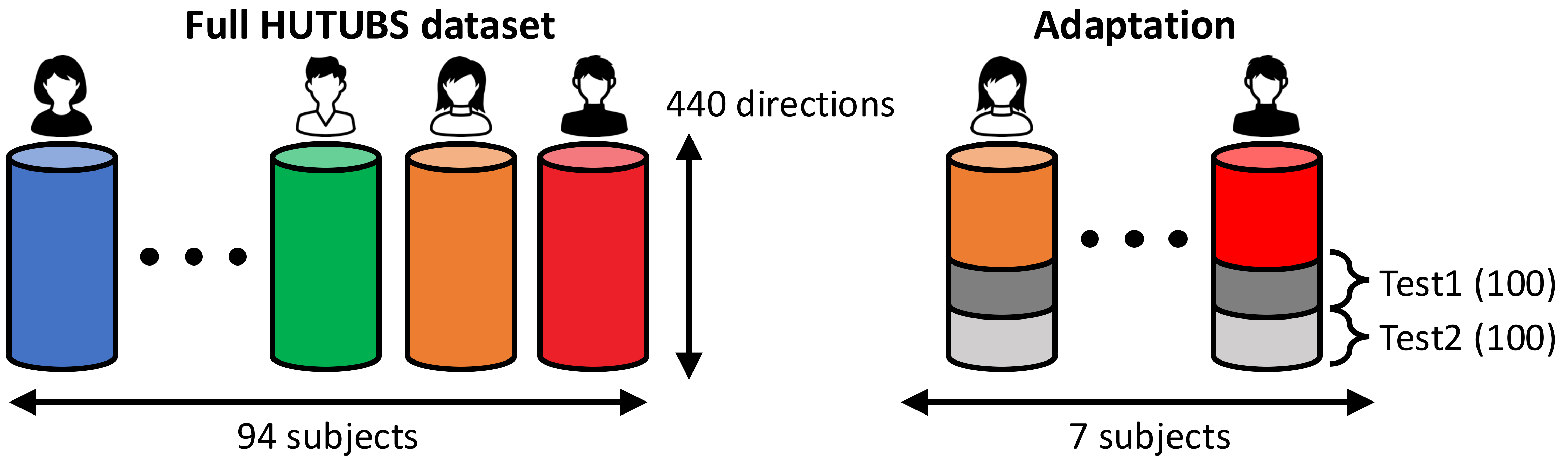
- The main NF and subject-specific parameters were pre-trained on HRTFs of the 87 subjects.



Experiments for Different Adaptation Techniques

- We used the HUTUBS dataset [Brinkmann+2019] to analyze the adaptation techniques.
  - It consists of HRTFs for 94 subjects, where 440 directions are measured for each subject.
  - We used 87 subjects for pre-training and the remaining 7 subjects for evaluation.

- The main NF and subject-specific parameters were pre-trained on HRTFs of the 87 subjects.
- The subject-specific parameters were then adapted to each target subject.



Experiments for Different Adaptation Techniques

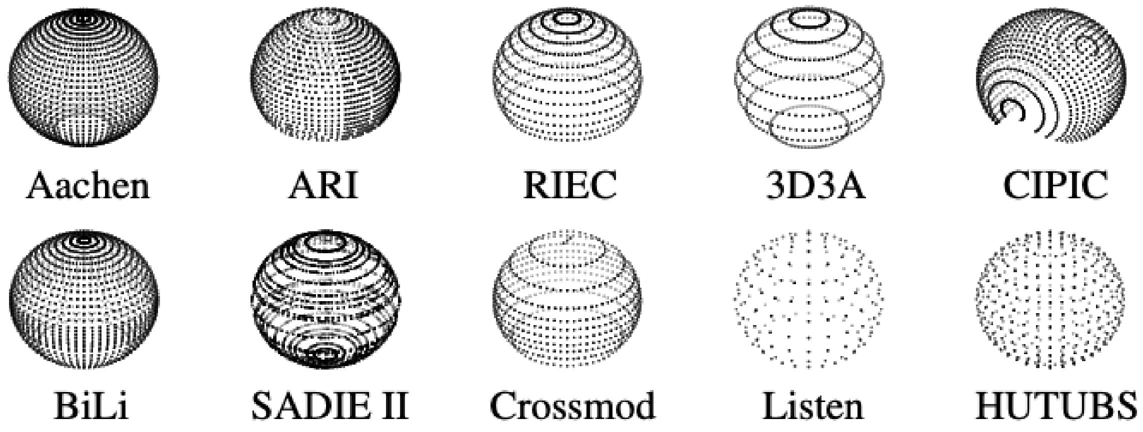
- FiLM outperforms the original conditioning-by-concatenation (i.e., prompting).
- Mag. NF and NIIRF are comparable at the directions included in the pre-training set (Test1).
- **NIIRF with LoRA performs best at unseen directions (Test2).**

RMSE on log-magnitude spectra

			Seen positions (Test1)					*Unseen positions (Test2)*				
			Number of measurements					Number of measurements				
Method	Adaptation	Q	10	20	30	50	100	10	20	30	50	100
Mag. NF	Conditioning by concatenation	32	4.8	4.7	4.6	4.6	4.5	4.9	4.8	4.8	4.8	4.7
	FiLM	32	**4.3**	4.2	4.2	4.1	4.1	**4.5**	**4.4**	4.4	4.4	4.3
	BitFit	2562	**4.3**	**4.0**	3.9	3.7	**3.5**	5.0	4.8	4.6	4.4	4.4
	LoRA	5122	**4.3**	**4.0**	**3.8**	**3.6**	**3.5**	5.2	5.0	4.9	4.8	4.6
NIIRF Proposed	Conditioning by concatenation	32	4.8	4.7	4.7	4.6	4.6	5.0	4.9	4.9	4.8	4.7
	FiLM	32	4.3	4.2	4.2	4.2	4.2	**4.5**	4.5	4.4	4.4	4.4
	BitFit	2248	**4.3**	**4.0**	3.9	3.7	**3.5**	4.8	4.5	4.4	4.2	4.1
	LoRA	4808	**4.3**	**4.0**	3.9	3.7	**3.5**	4.7	**4.4**	**4.2**	**4.1**	**4.0**

Challenges Toward Large-Scale Pre-training

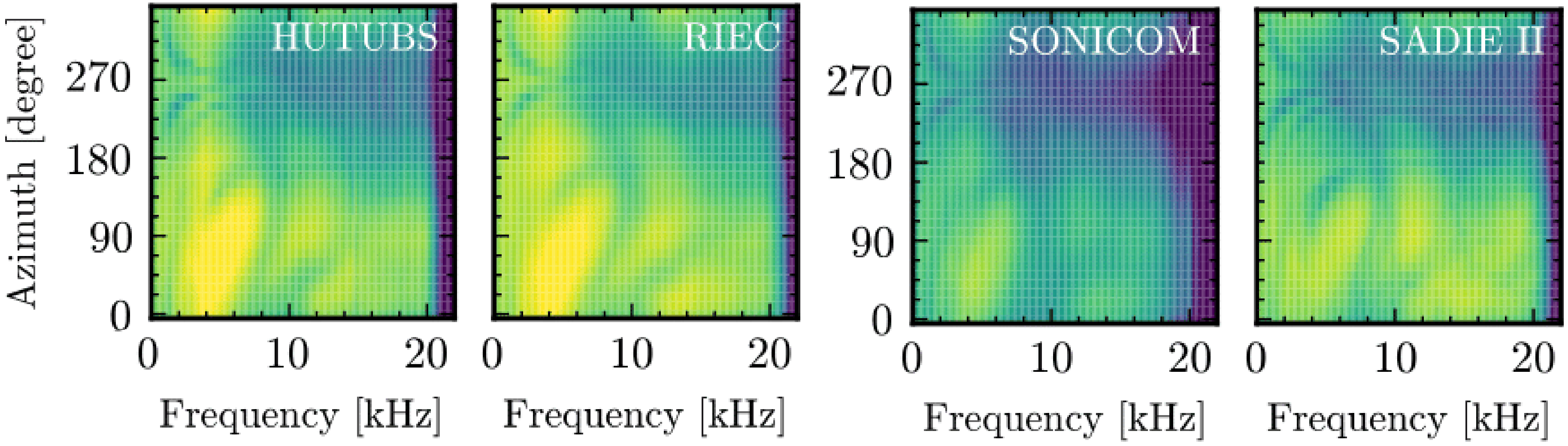
- We would like to train a generic NF with a large amount of training data.



We can train an NF on a combined dataset regardless of different spatial grids.

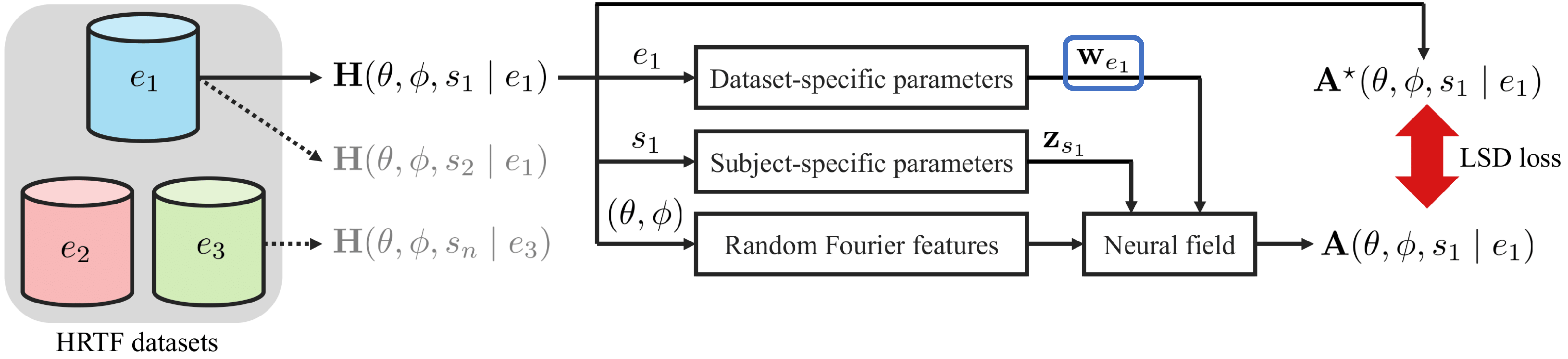
[Zhang+2023]

- **Differences in recording setups** also affect HRTFs from different datasets [Pauwels+2023].



Subject- and Dataset-Aware NF (SuDaField) [Masuyama+2025a]

- **We introduce dataset-specific parameters** in addition to the subject-specific parameters.
  - These parameters are shared across subjects within the same dataset.
  - We expect **they capture the effects from measurement setups specific to each dataset.**

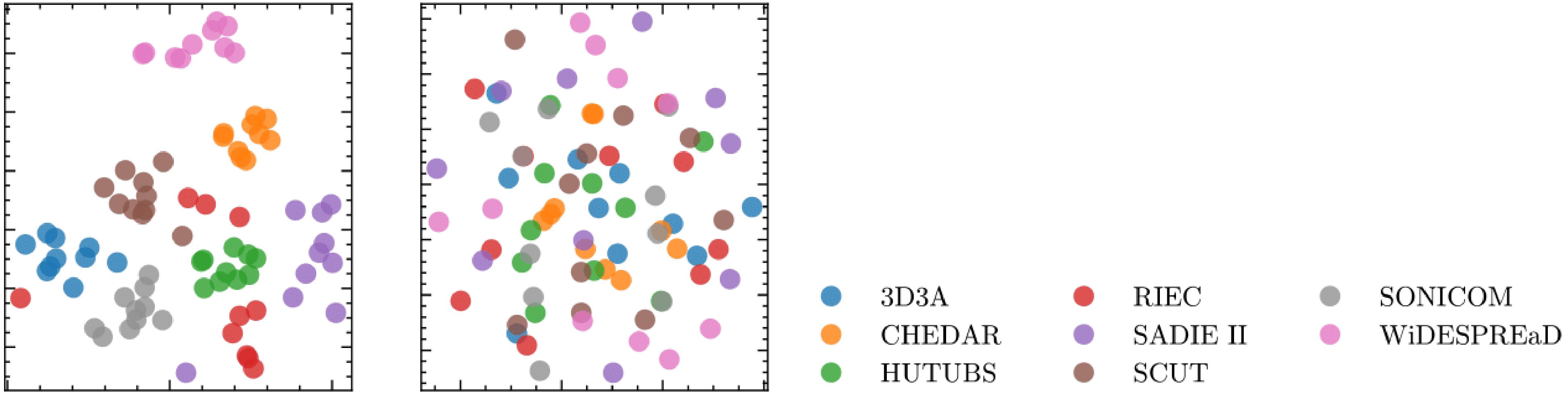


- We switch the bias at each layer depending on the subject or dataset index.

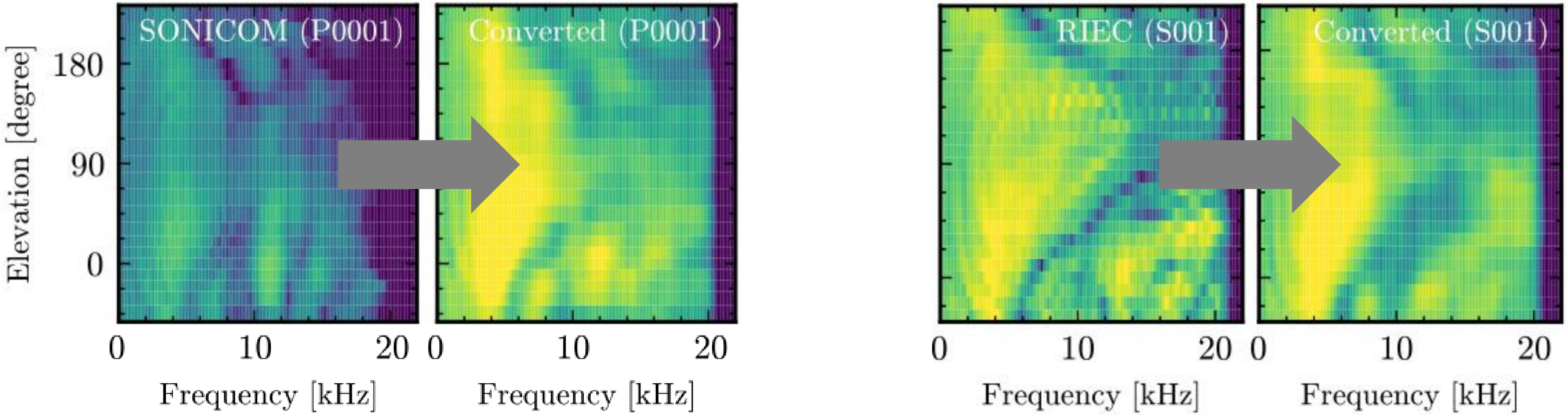
$$\mathbf{x}_l = \begin{cases} \mathtt{GeLU}(\mathbf{P}_l \mathbf{x}_{l-1} + \mathbf{q}_l) & \text{Generic (w/o BitFit)} \\ \mathtt{GeLU}(\mathbf{P}_l \mathbf{x}_{l-1} + \mathbf{q}_l + \mathbf{z}_{s,l}) & \text{Subject-specific} \\ \mathtt{GeLU}(\mathbf{P}_l \mathbf{x}_{l-1} + \mathbf{q}_l + \mathbf{w}_{e,l}) & \text{Dataset-specific} \end{cases}$$

SuDaField for HRTF Conversion

- Decoupled parameters disentangle the subject- and dataset-specific factors.

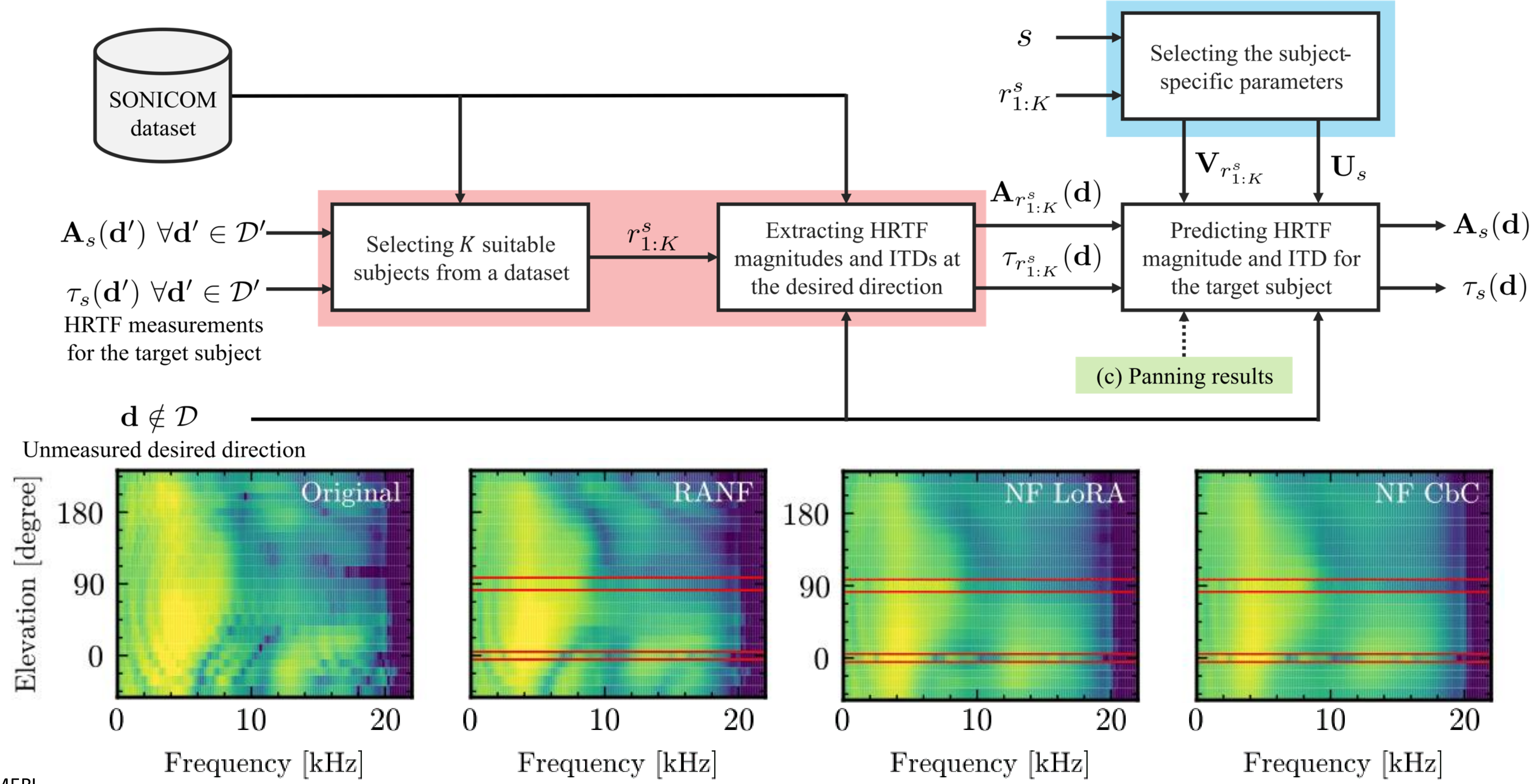


- We can convert HRTFs from one dataset to another by swapping the parameters.



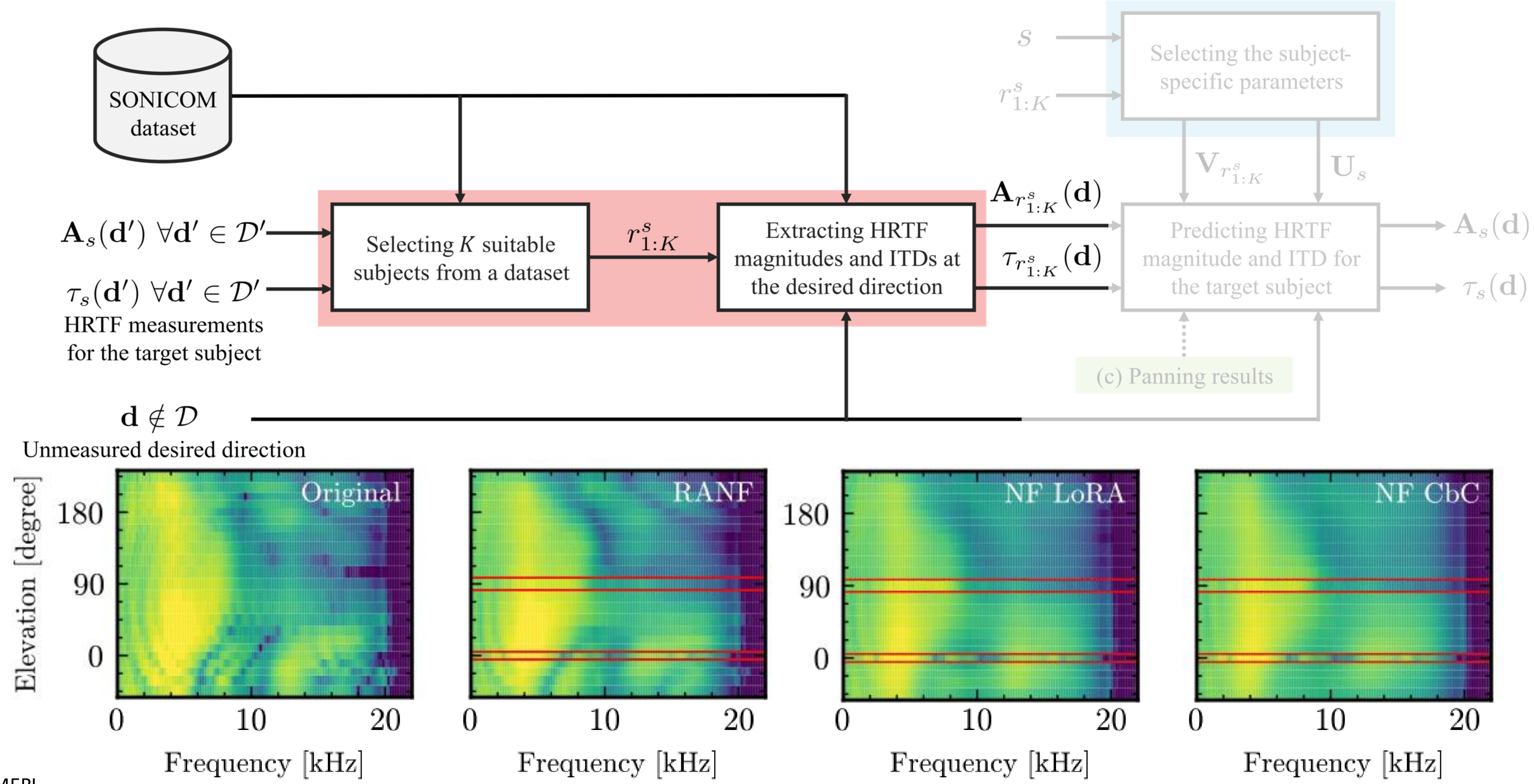
Retrieval Augmented Generation for HRTF Field [Masuyama+2025b,c]

- We propose RANF, **a retrieval augmented HRTF spatial upsampling method.**



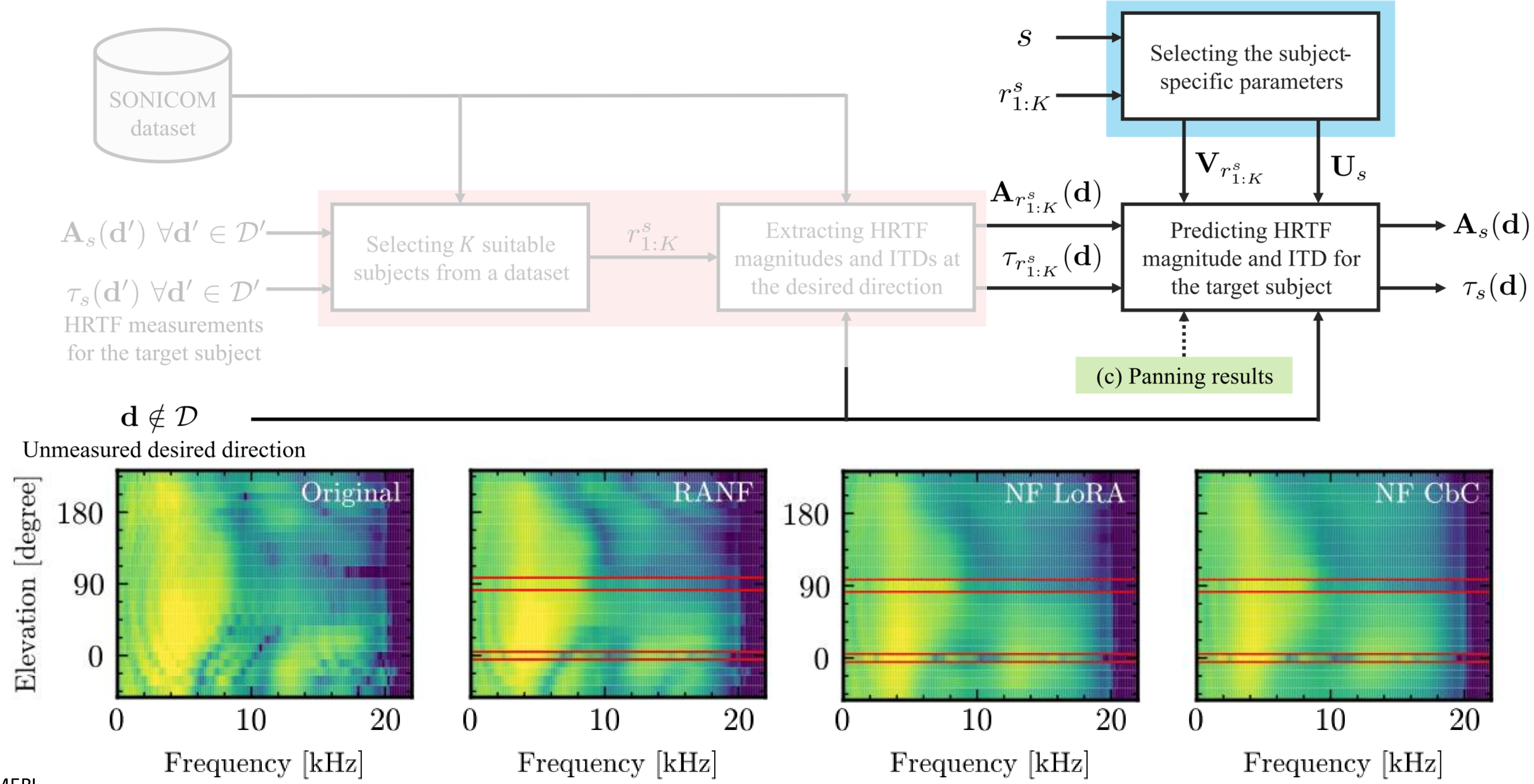
Retrieval Augmented Generation for HRTF Field [Masuyama+2025b,c]

- We propose RANF, **a retrieval augmented HRTF spatial upsampling method.**



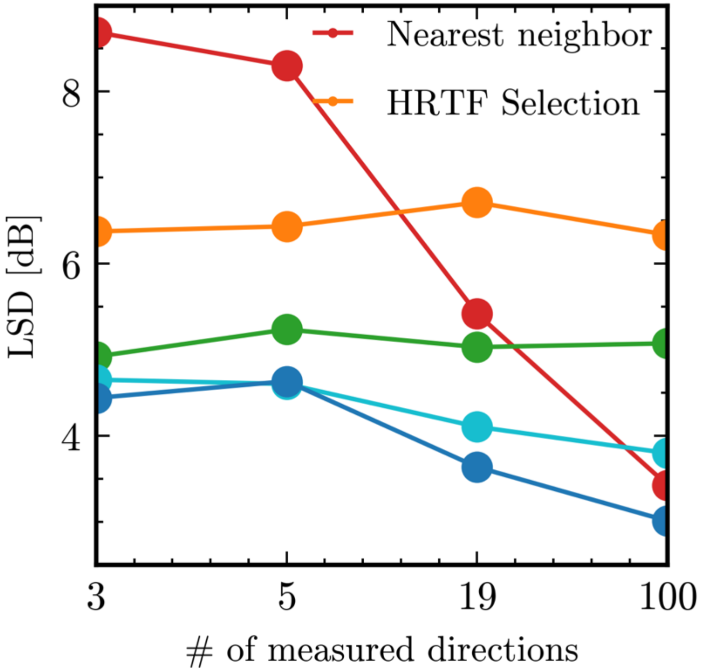
Retrieval Augmented Generation for HRTF Field [Masuyama+2025b,c]

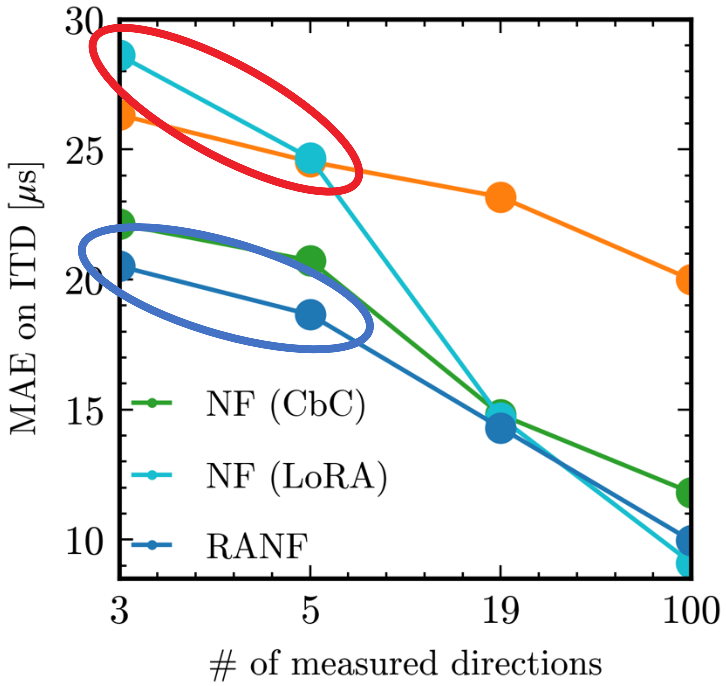
- We propose RANF, **a retrieval augmented HRTF spatial upsampling method.**



Experimental Comparison of RANF with Vanilla NFs

- **RANF substantially improves performance from NFs and just selecting the best subject.**

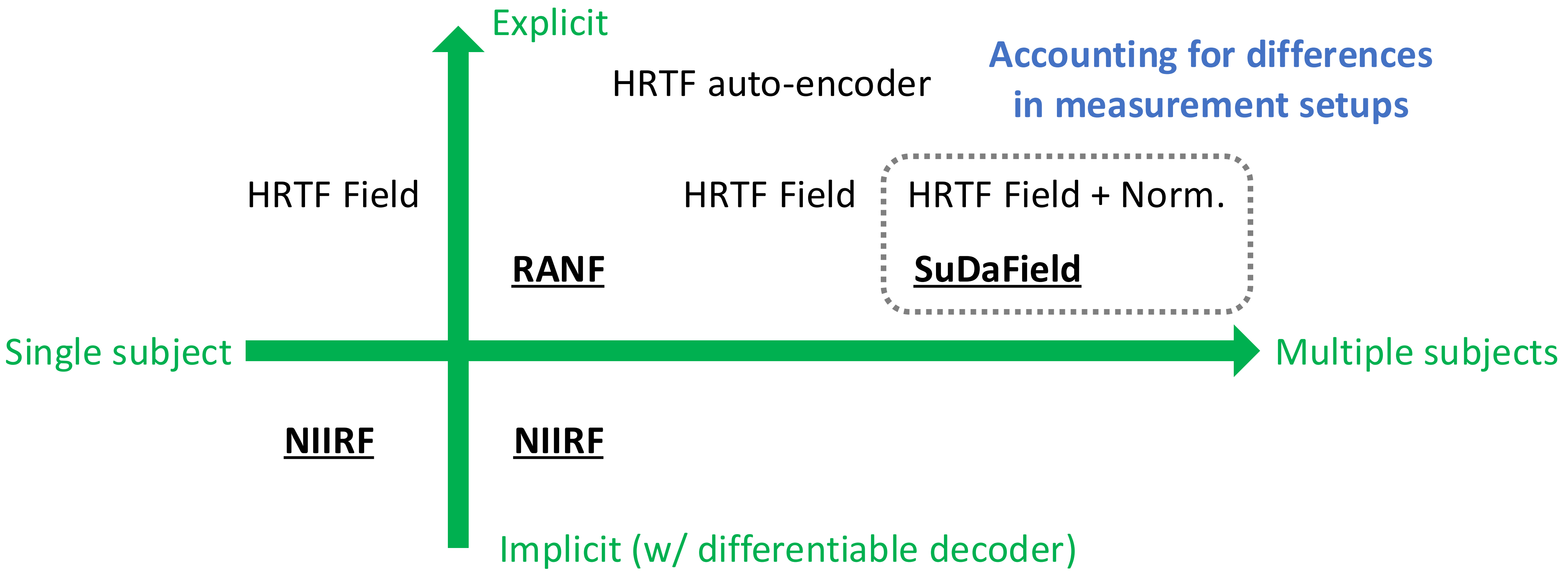




RANF yields the lowest ITD error under sparse scenarios, whereas NF (LoRA) results in high ITD error.

NFs for HRTF Upsampling and Personalization

- Many methods **represent HRTFs for multiple subjects using a small number of subject-specific parameters.**
- NIIRF implicitly represents HRTFs, predicting parameters of biquad filters.

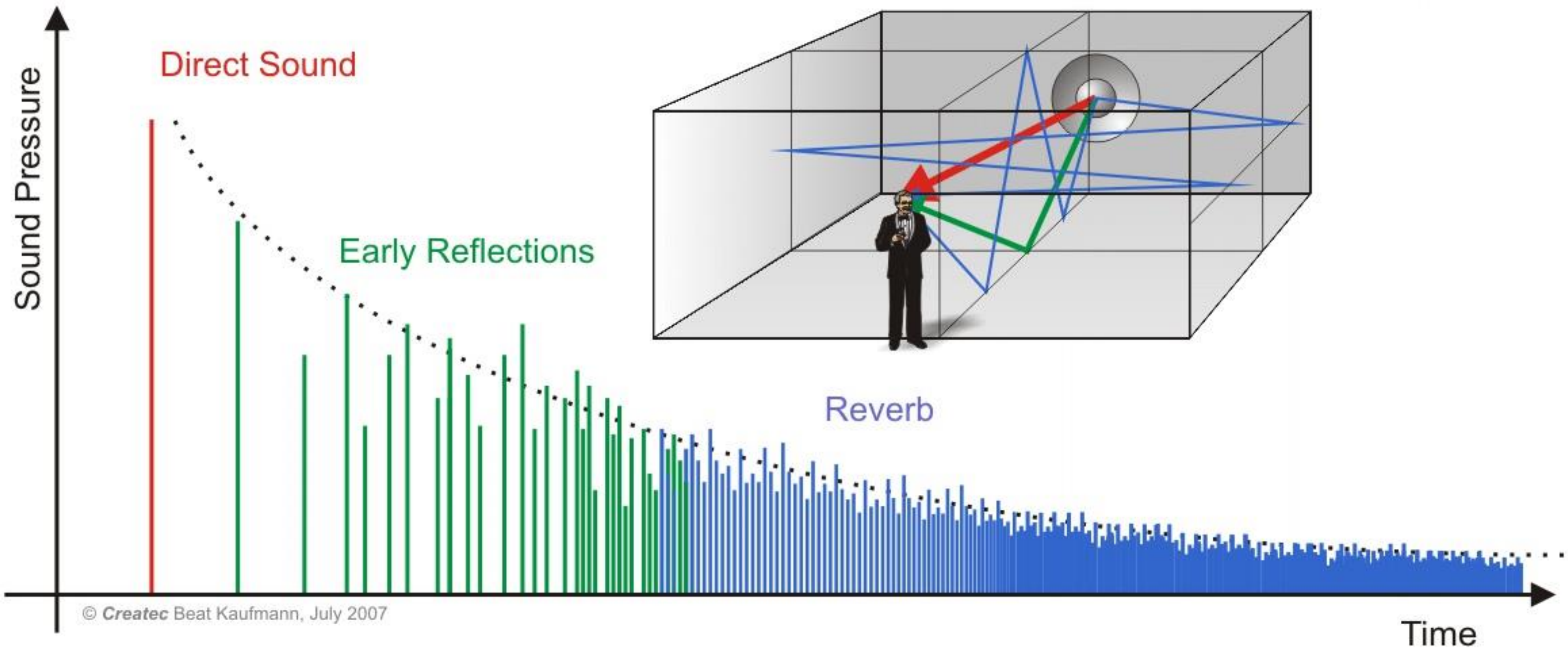


Agenda

- Overview of Spatial Audio and Neural Fields
- Neural Fields for Head-Related Transfer Functions
- **Neural Fields for Room Impulse Responses**
- Physics-Informed Neural Networks for Room Impulse Responses

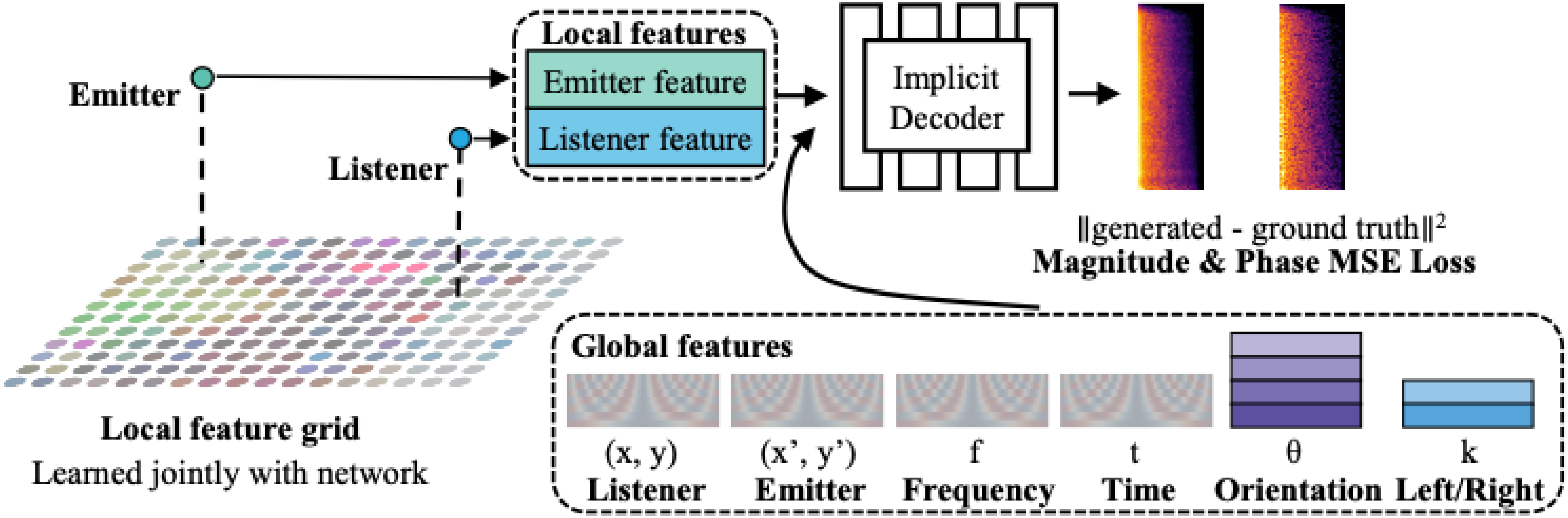
Room Impulse Responses (RIRs)

- RIRs capture sound propagation from the source position to a microphone.
  - We typically assume the sound propagation is linear and time-invariant.
  - RIR depends not only on the source/microphone positions but also on room settings.



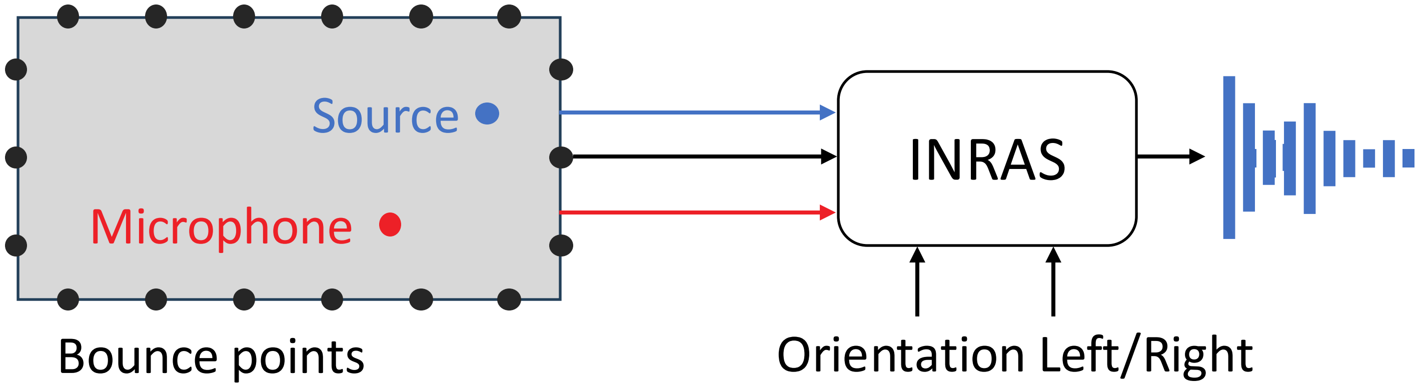
NFs for RIRs

- Neural Acoustic Fields (NAFs) continuously represent sound fields.
  - The original NAF predicts RIR from the speaker and listener positions [Luo+2022].
  - **A scene-specific local feature is optimized together with the decoder for each scene.**



Implicit Neural Representation for Audio Scenes (INRAS) [Su+2022]

- **INRAS [Su+2022] leverages room geometry** as scene-specific context.
  - Bounce points $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_K] \in \mathbb{R}^{3 \times K}$ are sampled from the given room mesh.



  - INRAS is trained using time-domain and STFT-domain loss functions at each channel.

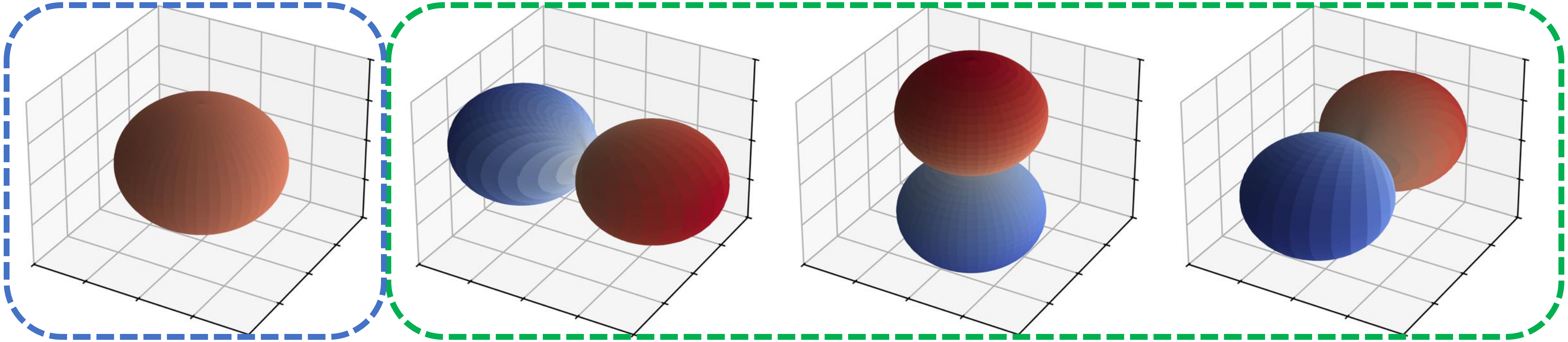
$$\mathcal{L}_{\mathrm{MSE}}(\mathbf{h}, \hat{\mathbf{h}}) = \|\hat{\mathbf{h}} - \mathbf{h}\|_2^2$$

$$\mathcal{L}_{\mathrm{STFT}}(\mathbf{h}, \hat{\mathbf{h}}) = \||\hat{\mathbf{H}}| - |\mathbf{H}|\|_1 + \|\angle\hat{\mathbf{H}} - \angle\mathbf{H}\|_2 + \cdots$$

- **The training of INRAS and its variants has been based on single-channel criteria.**

Direction-Aware Neural Acoustic Field (DANF) [Ick+2025]

- **DANF predicts first-order Ambisonics (FOA)-RIR** to provide 3D directional information.
  - **FOA contains three directional components along each Cartesian axes (X, Y, Z)** in addition to the omnidirectional component (W).



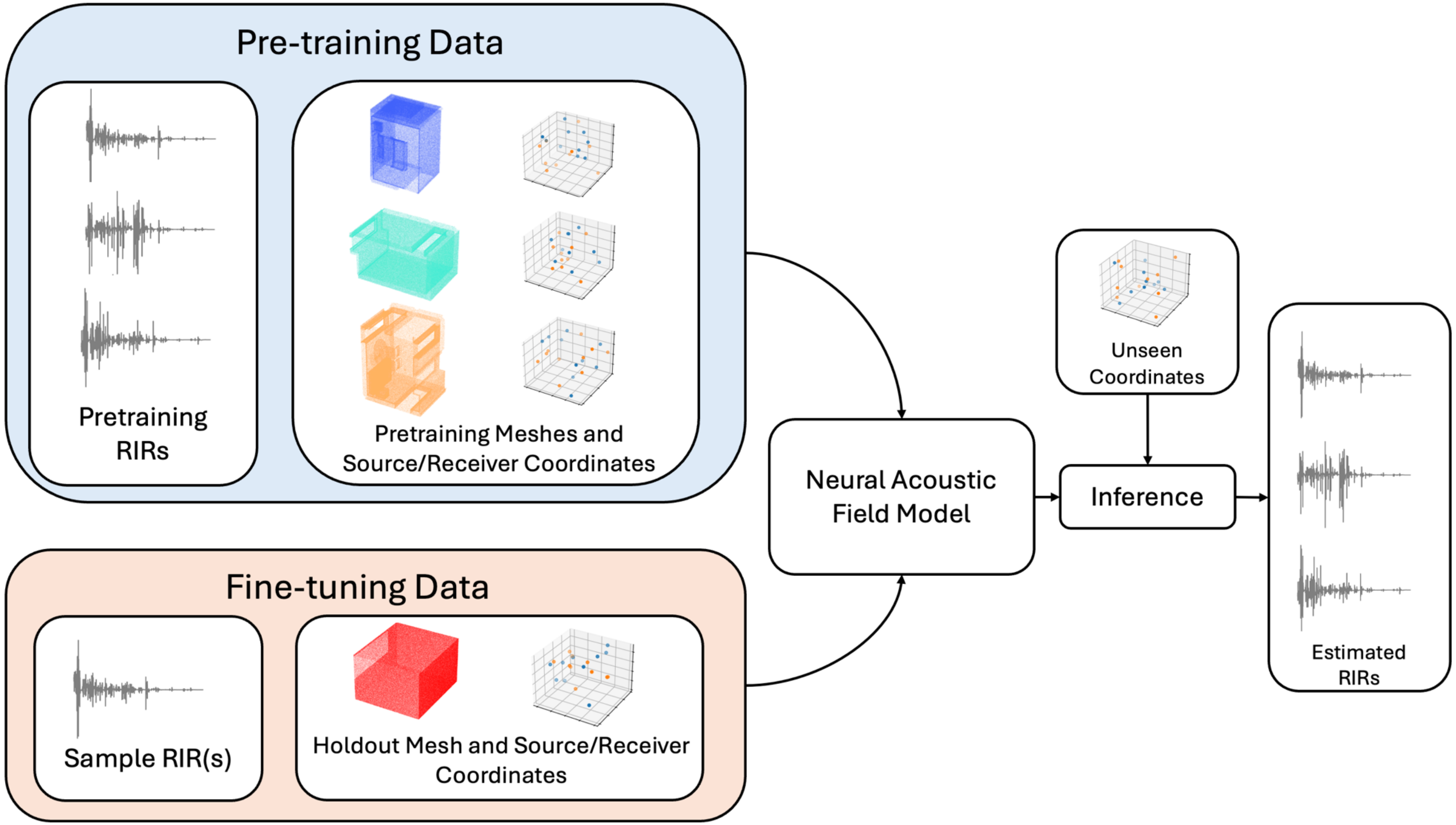
Basis for W-channel Basis for Y-channel Basis for Z-channel Basis for X-channel

  - **We derive a direction-aware loss function** based on the intensity vector.

$$\mathcal{L}_{\mathrm{IV}}(\hat{\mathbf{h}}, \mathbf{h}) = \frac{1}{2}\left(1 - \frac{\langle \mathbf{I}(\hat{\mathbf{h}}); \mathbf{I}(\mathbf{h})\rangle}{\|\mathbf{I}(\hat{\mathbf{h}})\|\|\mathbf{I}(\mathbf{h})\|}\right)$$

Setup for Few-Shot Adaptation of DANF

- We adapt a pre-trained DANF to new rooms with a limited number of RIRs.
  - DANF was pre-trained on FOA-RIRs from 90 rooms.
  - The pre-trained DANF was fine-tuned for each of 10 new rooms.



Results for Few-Shot Adaptation

- Zero-shot (the pre-trained DANF without fine-tuning) does not perform well.
  - **The bounce points seem to be insufficient to capture room characteristics.**
- **Fine-tuning is beneficial when the available FOA-RIRs for the target room are limited.**
  - The warm-start consistently outperforms the cold-start, i.e., DANF trained from scratch.
  - LoRA is an efficient solution since the required parameters for each scene is much less.

		1 training example				80 training examples				800 training examples			
	N_p	**T60**	**C50**	**EDT**	**DoA**	**T60**	**C50**	**EDT**	**DoA**	**T60**	**C50**	**EDT**	**DoA**
Zero-Shot	0	4.78	14.07	474.24	111.10	-	-	-	-	-	-	-	-
Cold-Start	3.5×10^6	22.06	26.79	945.64	67.87	22.72	26.80	944.19	59.99	0.46	2.89	10.21	32.86
Warm-Start	3.5×10^6	2.68	4.68	29.34	52.56	1.22	2.73	18.29	31.94	0.49	2.39	9.28	27.08
LoRA(3)	2.9×10^4	2.34	6.36	39.88	67.32	1.44	3.75	21.51	33.68	1.40	3.29	20.52	34.12
LoRA(1)	9.6×10^3	4.82	7.07	88.15	55.43	1.82	3.76	27.22	42.32	1.31	3.86	22.66	40.67

Here, EDT is in milliseconds.

Results for Few-Shot Adaptation

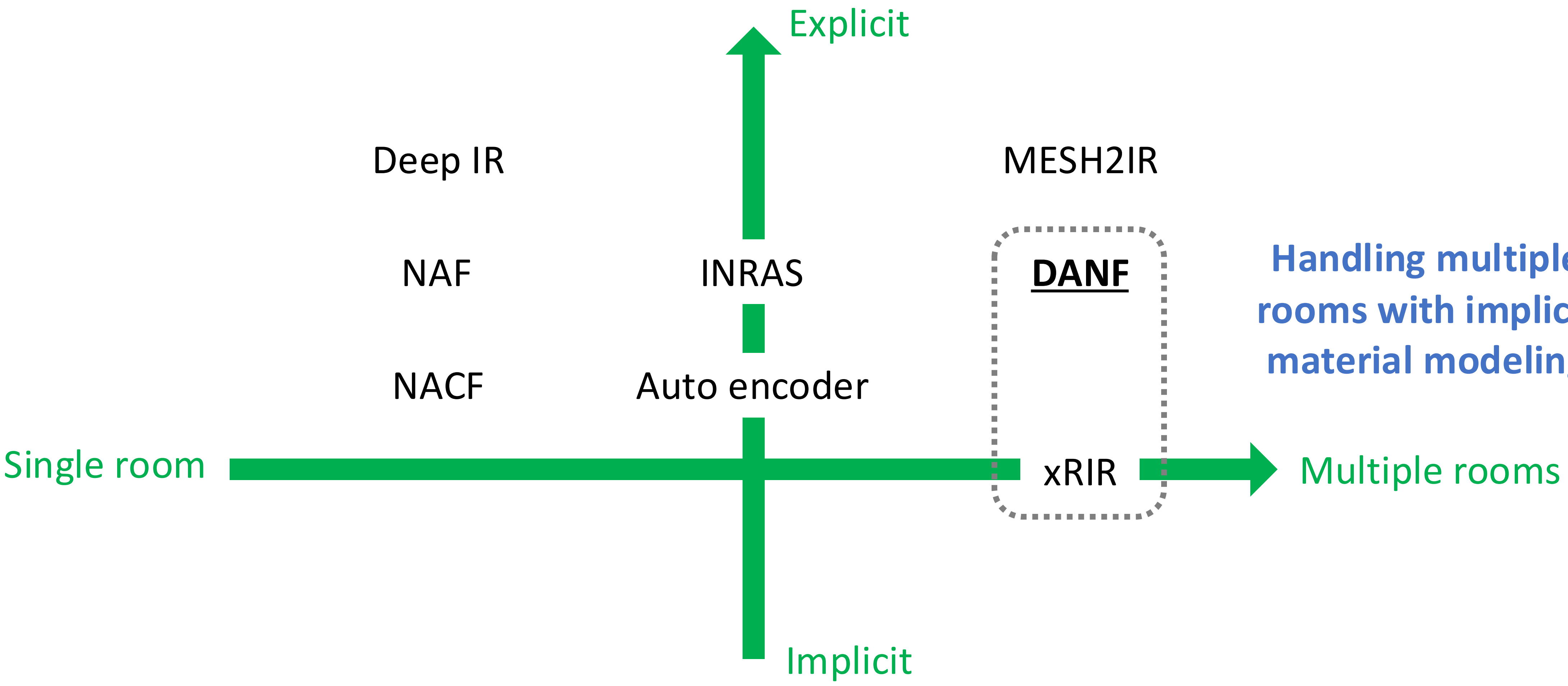
- Zero-shot (the pre-trained DANF without fine-tuning) does not perform well.
  - **The bounce points seem to be insufficient to capture room characteristics.**

- **Fine-tuning is beneficial when the available FOA-RIRs for the target room are limited.**
  - The warm-start consistently outperforms the cold-start, i.e., DANF trained from scratch.
  - **LoRA is an efficient solution** since the required parameters for each scene is much less.

		1 training example				80 training examples				800 training examples			
	N_p	**T60**	**C50**	**EDT**	**DoA**	**T60**	**C50**	**EDT**	**DoA**	**T60**	**C50**	**EDT**	**DoA**
Zero-Shot	0	4.78	14.07	474.24	111.10	-	-	-	-	-	-	-	-
Cold-Start	3.5×10^6	22.06	26.79	945.64	67.87	22.72	26.80	944.19	59.99	0.46	2.89	10.21	32.86
Warm-Start	3.5×10^6	2.68	4.68	29.34	52.56	1.22	2.73	18.29	31.94	0.49	2.39	9.28	27.08
LoRA(3)	2.9×10^4	2.34	6.36	39.88	67.32	1.44	3.75	21.51	33.68	1.40	3.29	20.52	34.12
LoRA(1)	9.6×10^3	4.82	7.07	88.15	55.43	1.82	3.76	27.22	42.32	1.31	3.86	22.66	40.67

NFs for RIR Spatial Interpolation

- Early works optimize an NF for each room from scratch.
- More recent methods apply a single NF to multiple rooms.
  - **Cross-room generalization seems more challenging than the case of HRTFs.**



Handling multiple rooms with implicit material modeling

Agenda

- Overview of Spatial Audio and Neural Fields
- Neural Fields for Head-Related Transfer Functions
- Neural Fields for Room Impulse Responses
- **Physics-Informed Neural Networks for Room Impulse Responses**

Formulation of NFs for Sound Field Reconstruction

- NFs predict a sound pressure from a given position and time.
  - They leverage powerful modeling capability and flexibility of neural networks.

$$p(\mathbf{r}, t) \approx \hat{p}(\mathbf{r}, t) = \mathrm{NF}_\theta(\mathbf{r}, t)$$

Euclidean coordinate of microphone position and time

- NFs are typically trained to reconstruct given D measurements.

$$\mathcal{L}_{\text{data}} = \frac{1}{DL} \sum_{d=0}^{D-1} \sum_{l=0}^{L-1} |\hat{p}(\mathbf{r}_d, t_l) - p(\mathbf{r}_d, t_l)|$$

Index of measured positions

Index of discrete time

- **We can predict RIRs in a grid-less manner**, but **NFs do not incorporate physical principles.**

Physics-Informed Neural Networks (PINNs)

- PINNs encode partial differential equations (PDEs) governing sound propagation into NFs.
  - **Deviation from the governing PDEs, e.g., the wave equation, is used as a soft penalty term.**
  - We can compute penalty terms at arbitrary microphone position and time.

$$\mathcal{L}_{\text{wave}} = \underbrace{\mathbb{E}_{\mathbf{r}\in\Omega}\mathbb{E}_{t\in[0,T]}}_{\text{Sampling the position and time}} \underbrace{\left| \Delta\hat{p}(\mathbf{r},t) - \frac{1}{c_0^2}\frac{\partial^2 \hat{p}(\mathbf{r},t)}{\partial t^2} \right|}_{\text{Calculated by using automatic differentiation}}$$

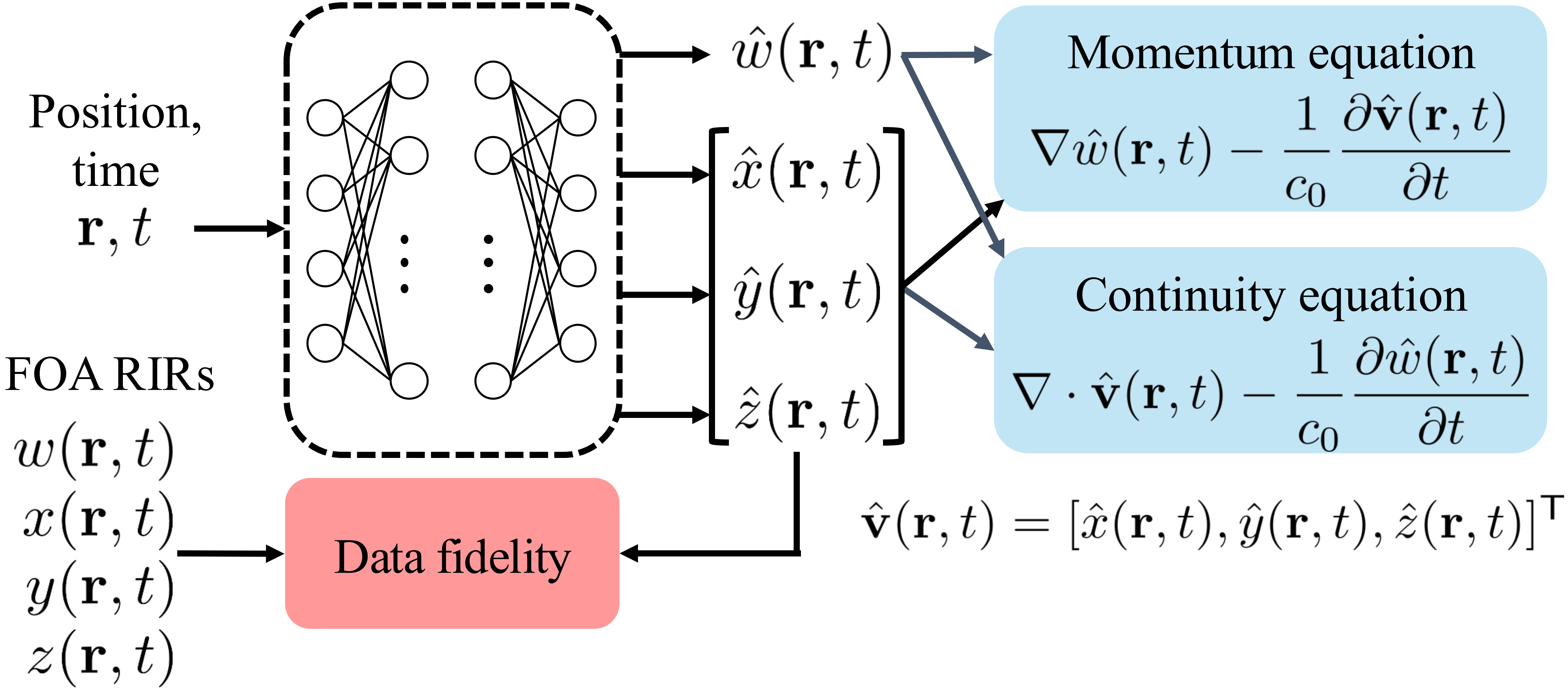
- PINNs are trained to minimize a sum of the data-fidelity and regularization terms.

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{data}} + \lambda\mathcal{L}_{\text{wave}}$$

- **PINNs have been successfully applied to RIR interpolation [Pezzoli+2023, Karakonstantis+2024].**

Physics-Informed DANF (PI-DANF) [Masuyama+2025d]

- We propose a physics-informed extension of DANF (PI-DANF).
  - The outputs of PI-DANF are **regularized by two physical principles of sound propagation.**
  - The penalty terms capture the relationship between the zeroth- and first-order components.



Physics-Informed Priors for FOA

- The sound pressure and particle velocities satisfy the following two equations.

$$\nabla p(\mathbf{r},t) + \rho_0 \frac{\partial \mathbf{u}(\mathbf{r},t)}{\partial t} = \mathbf{0} \quad \textbf{Linearized momentum equation}$$

$$\rho_0 \nabla \cdot \mathbf{u}(\mathbf{r},t) + \frac{1}{c_0^2}\frac{\partial p(\mathbf{r},t)}{\partial t} = 0 \quad \textbf{Continuity equation}$$

- We derive two penalty terms from these equations and the properties of FOA.
  - The predicted FOA RIRs are enforced to satisfy the relationships at any position and time.

$$\mathcal{L}_{\mathrm{momentum}} = \mathbb{E}_{\mathbf{r}\in\Omega}\mathbb{E}_{t\in[0,T]}\left\| \nabla \hat{w}(\mathbf{r},t) - \frac{1}{c_0}\frac{\partial \hat{\mathbf{v}}(\mathbf{r},t)}{\partial t}\right\|_1$$

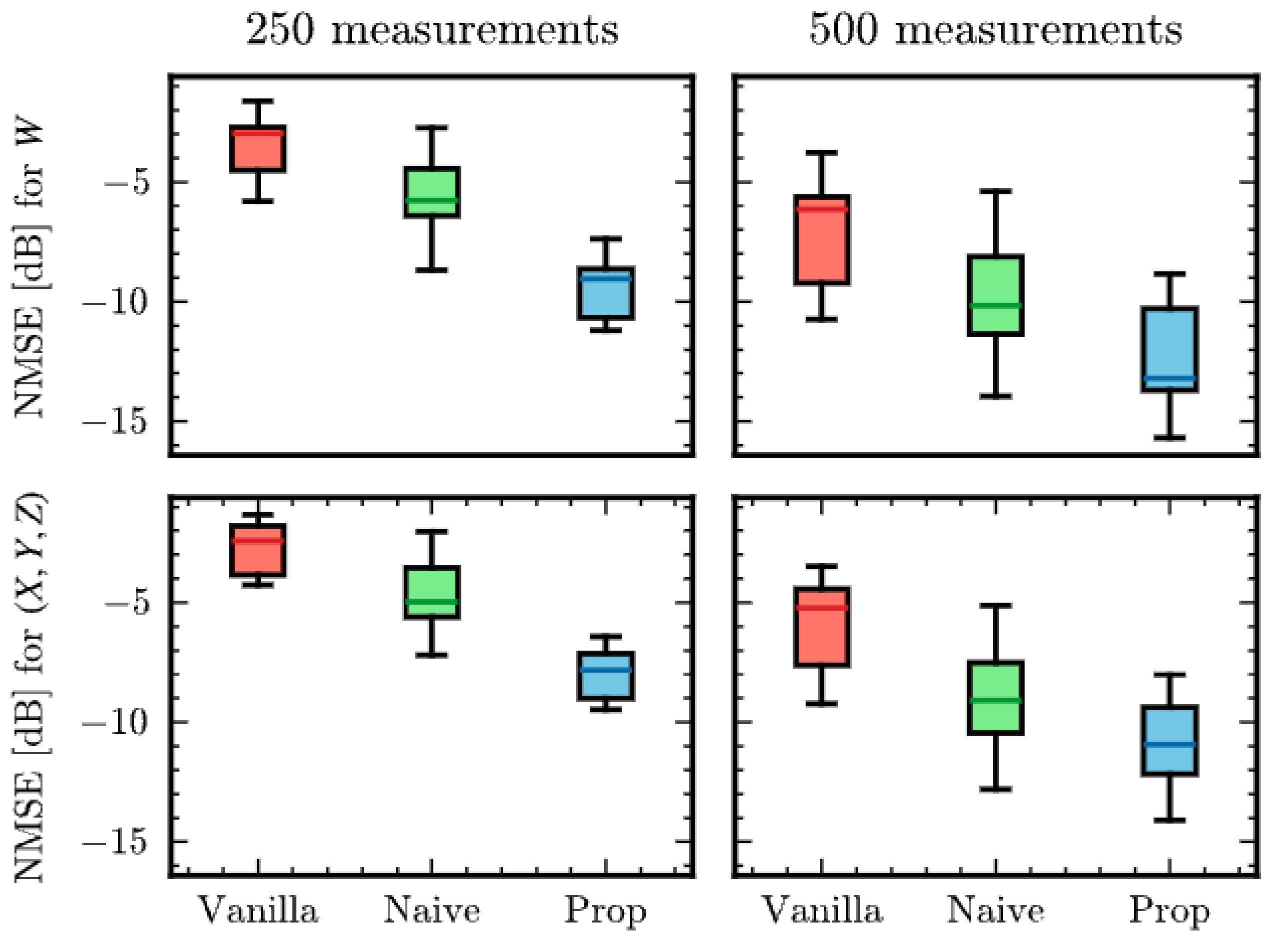
$$\mathcal{L}_{\mathrm{continuity}} = \mathbb{E}_{\mathbf{r}\in\Omega}\mathbb{E}_{t\in[0,T]}\left| \nabla \cdot \hat{\mathbf{v}}(\mathbf{r},t) - \frac{1}{c_0}\frac{\partial \hat{w}(\mathbf{r},t)}{\partial t}\right|$$

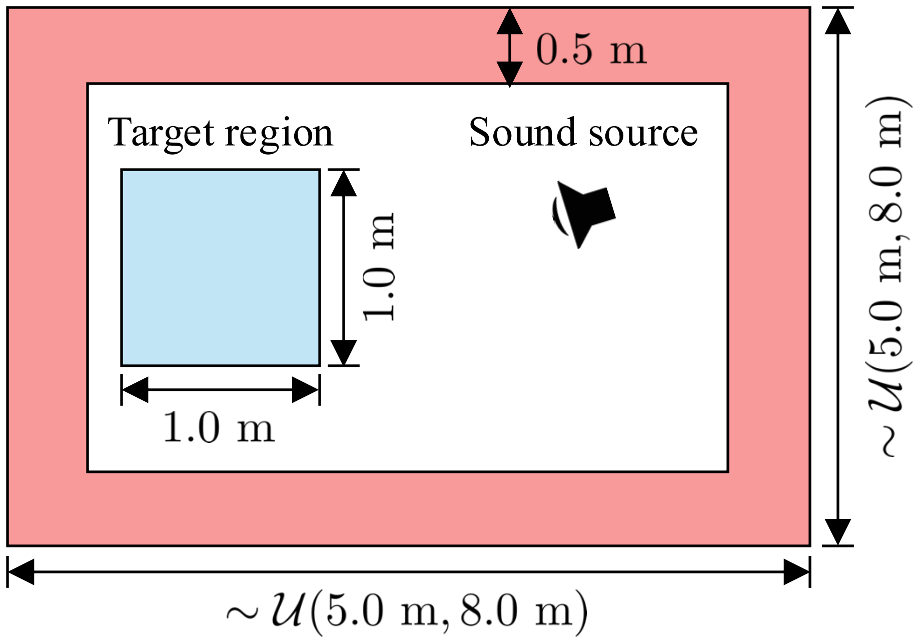
$$\boxed{\mathbf{v}(\mathbf{r},t) = -\rho_0 c_0 \mathbf{u}(\mathbf{r},t)}$$

We consider air sound propagation in a source-free region, assuming air as an inviscid fluid.

Quantitative Results

- The naive physics-informed method outperforms the vanilla NF.
- **The proposed PI-DANF consistently performs best.**
  - The proposed penalty terms are more beneficial than the existing one only for the W-channel.

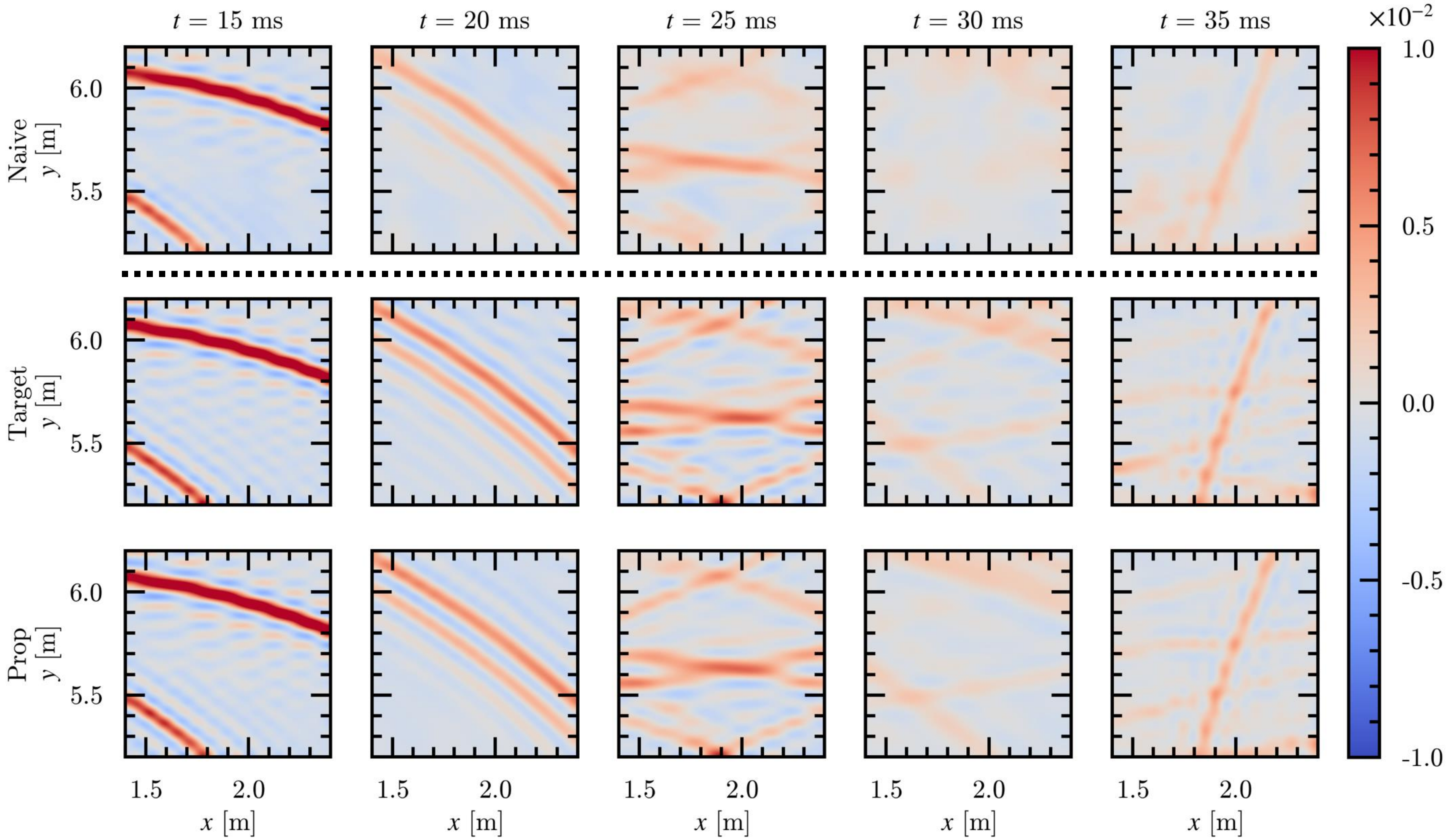




Simulated by HARP [Saini+2024] on 10 shoebox rooms with random geometry

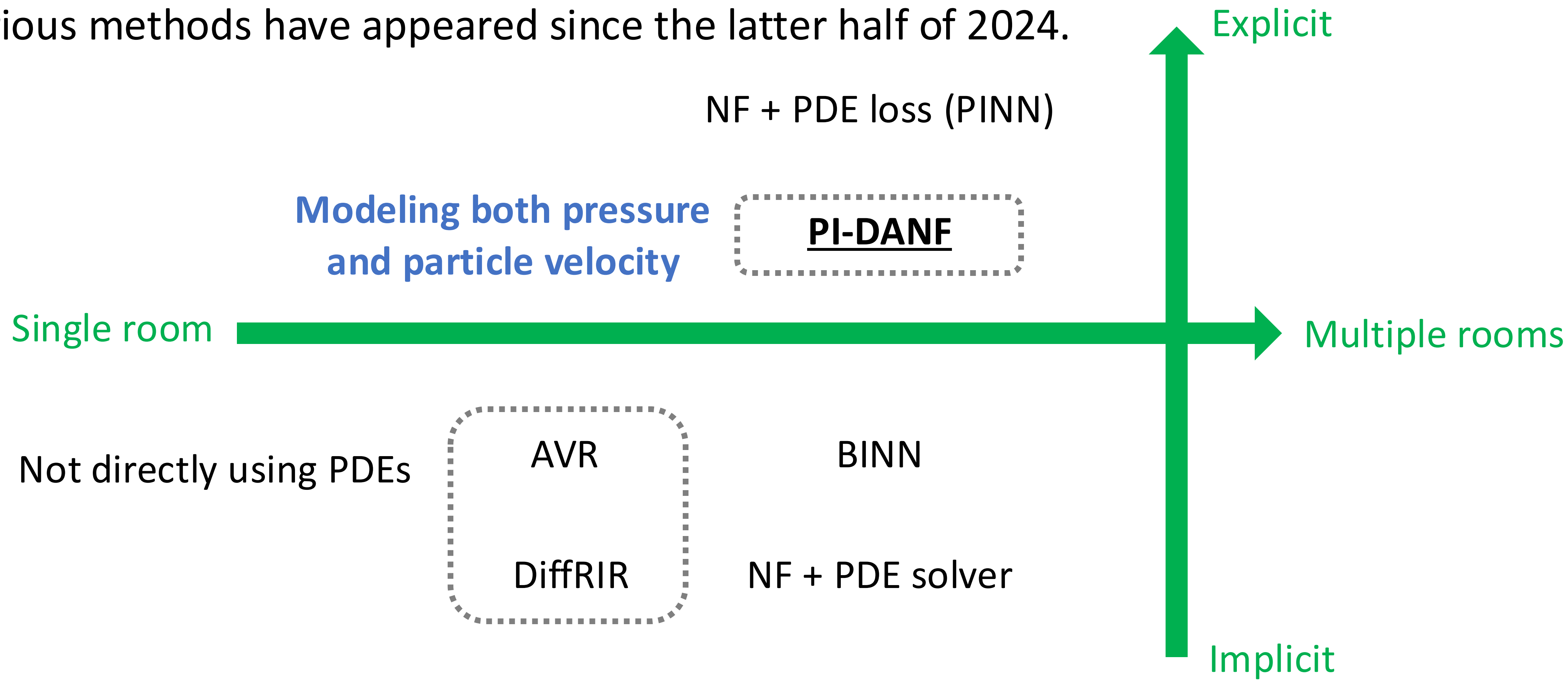
Qualitative Results

- **PI-DANF achieves better reconstruction** than the naive method, especially for $t > 25$ ms.



PINNs for RIR Spatial Interpolation

- Physics-informed methods focus on room-wise modeling.
  - The wave-based priors improve precise waveform-level modeling.
  - These priors may not be beneficial to capture coarse characteristics of RIRs across rooms.
- Various methods have appeared since the latter half of 2024.



Summary

- NFs have been actively applied to HRTF and RIR modeling.
  - **A single NF is shared across multiple subjects/rooms to exploit the learned prior knowledge.**
- Physics-informed methods have been developed, but they focus on room-wise modeling.
  - These methods are suitable for waveform-level precise modeling.
  - **There is a gap between cross-room models focusing on coarse characteristics of RIRs (e.g., envelope, RT60, …).**
- Many challenges remain such as
  - Limited size of datasets
  - Multi-modal integration
  - Training and inference computational costs

https://github.com/merlresearch

MITSUBISHI ELECTRIC
Changes for the Better

MITSUBISHI ELECTRIC
Changes for the Better