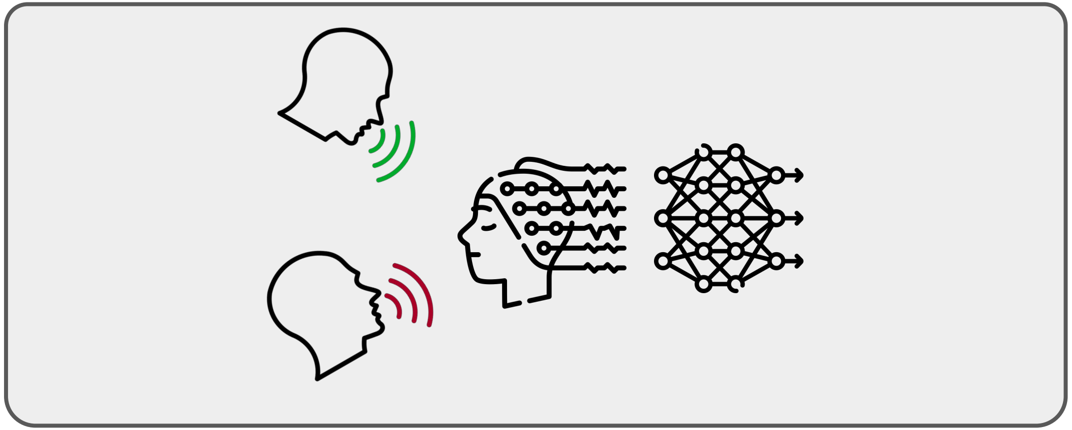
Exploring Foundation Models for Auditory Attention Decoding

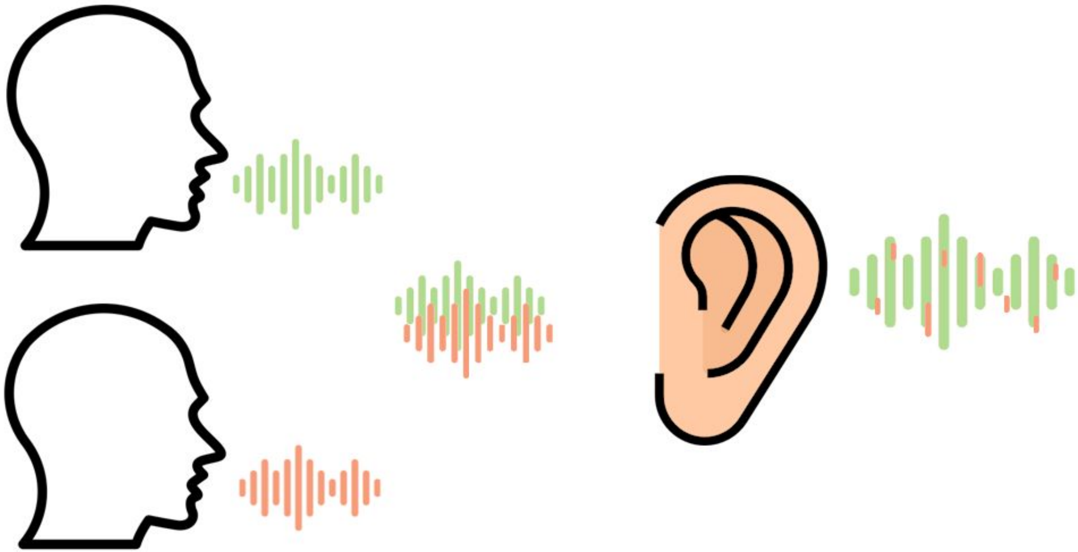


Rasmus Steen Mikkelsen (s204135)
Victor Tolsager Olesen (s204141)

Introduction

Introduction

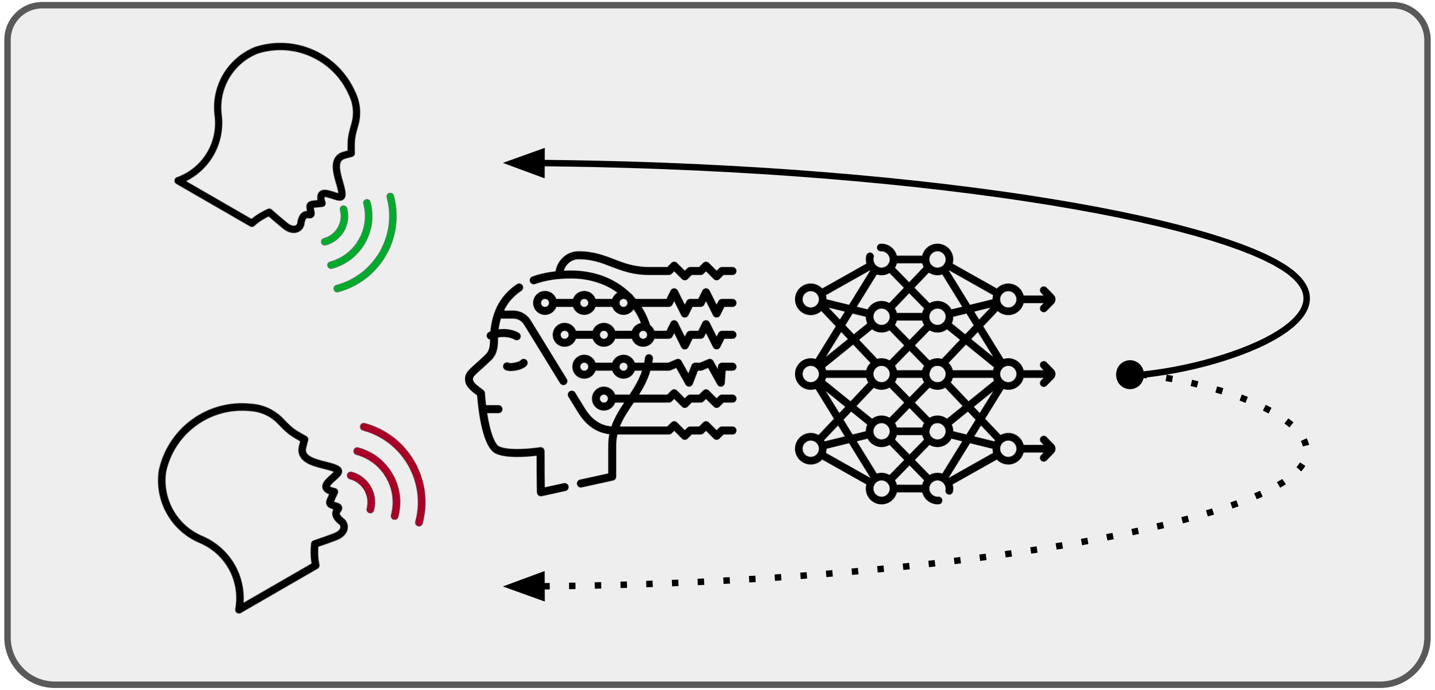
- Cocktail party effect
- Hearing aid users



Introduction

Auditory Attention Decoding

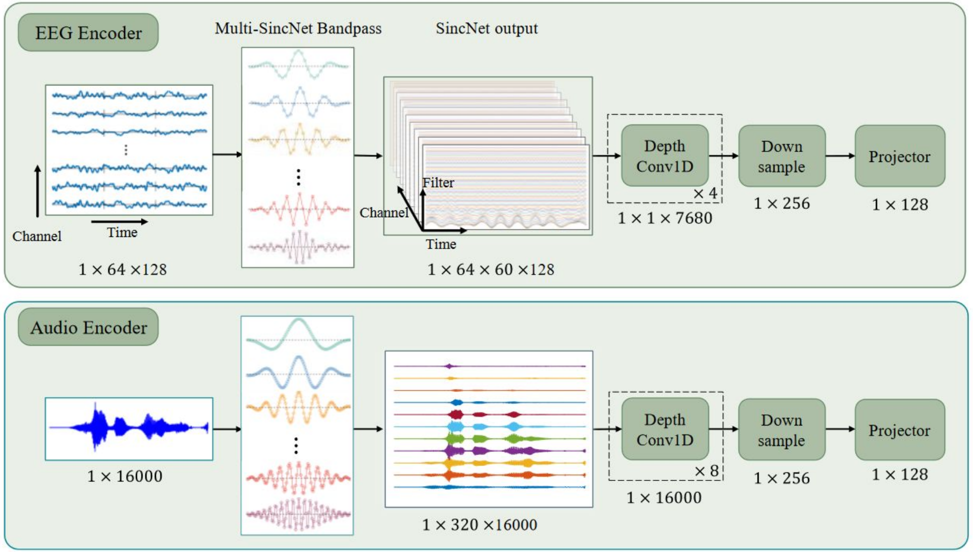
- AAD: Audio+EEG → Attention
- Decision window: Time segment used to predict



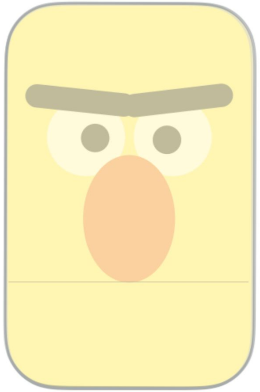
Introduction

Foundation models

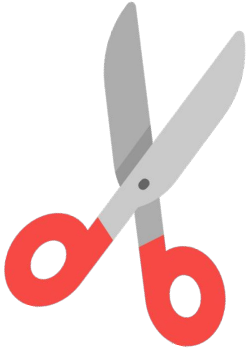
- Foundation Models
- SOTA AAD Models



NLP
BERT

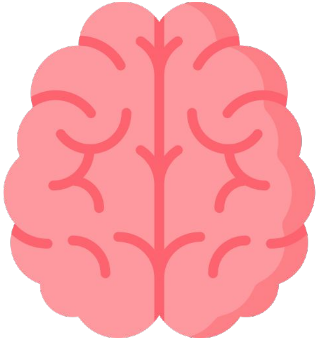


Vision+Text
CLIP



Audio+Text
CLAP

EEG
LaBraM



Introduction

Research questions

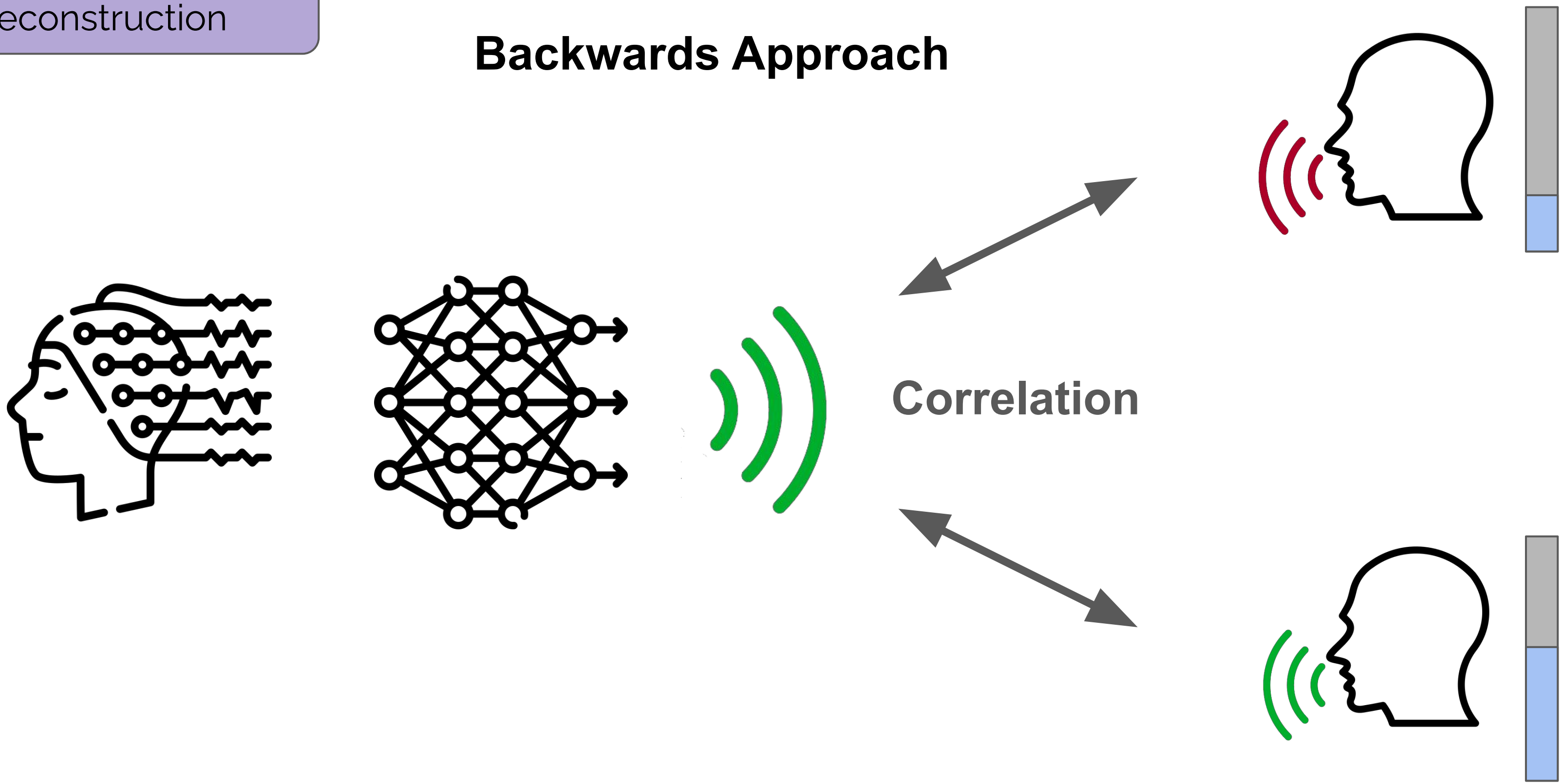
RQ1: How do CLAP and LaBraM perform as pretrained feature extractors for auditory attention decoding?

RQ2: How does contrastive learning compare to supervised classification for training robust AAD models using CLAP and LaBraM?

RQ3: How does the length of decision windows affect performance?

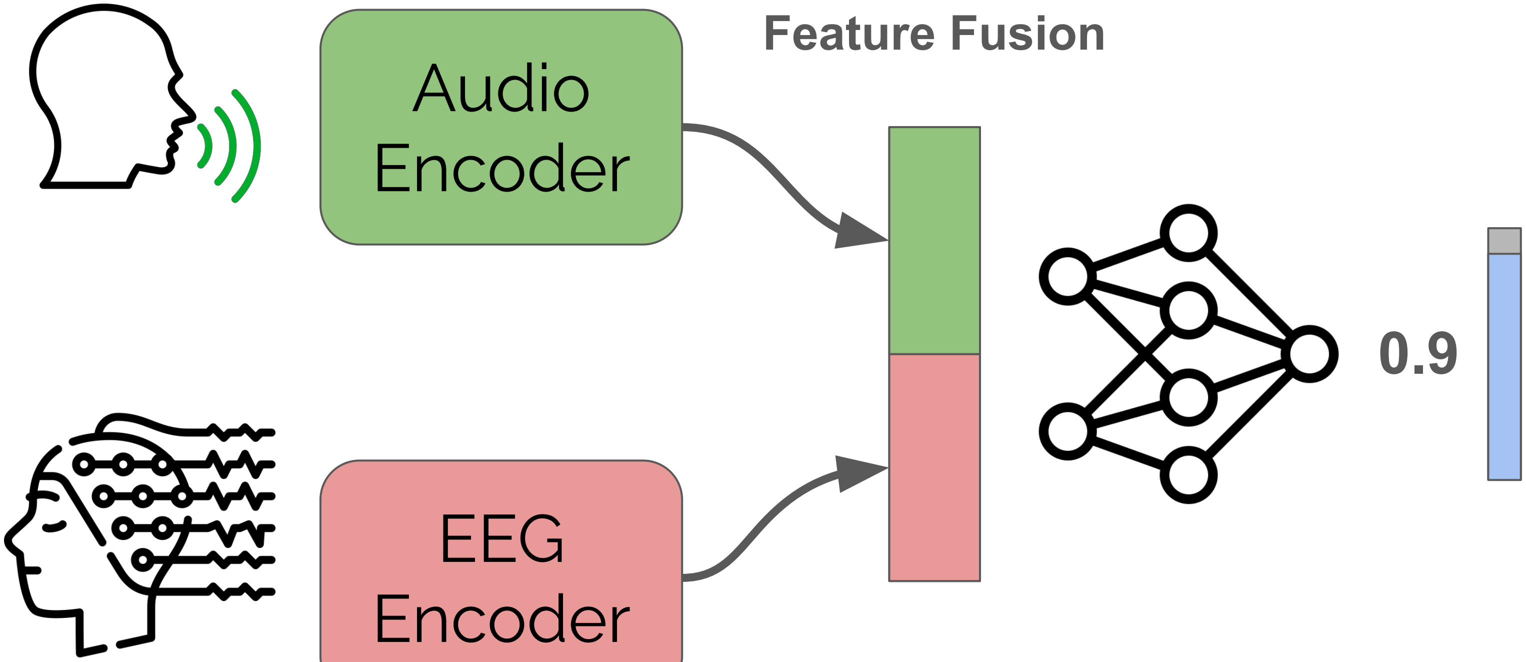
Literature Review

Backwards Approach

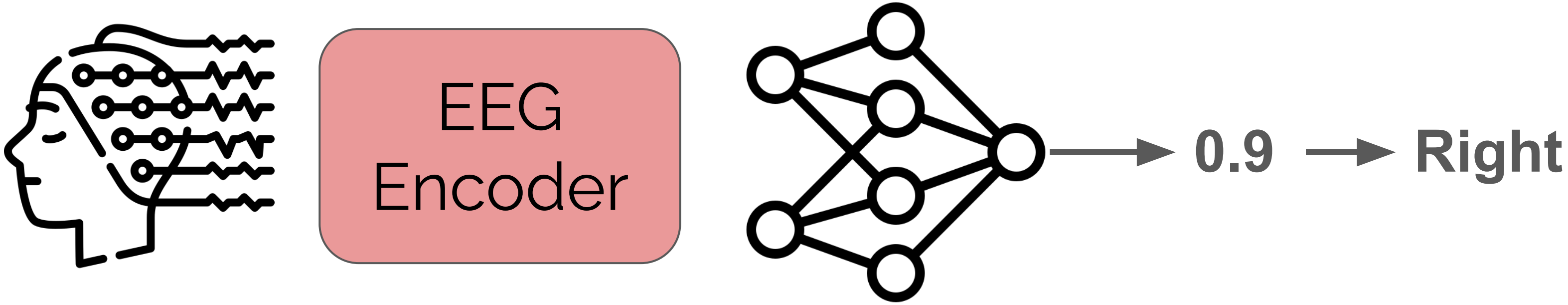


Literature Review

Direct Classification



Auditory Spatial Attention Decoding



Literature Review

Why Direct Classification?

[..] the process of stimulus reconstruction [..] is not optimized to effectively detect attention. [...] the compression of multichannel EEG signals into a single waveform through stimulus reconstruction reduces the available information for analysis[1]

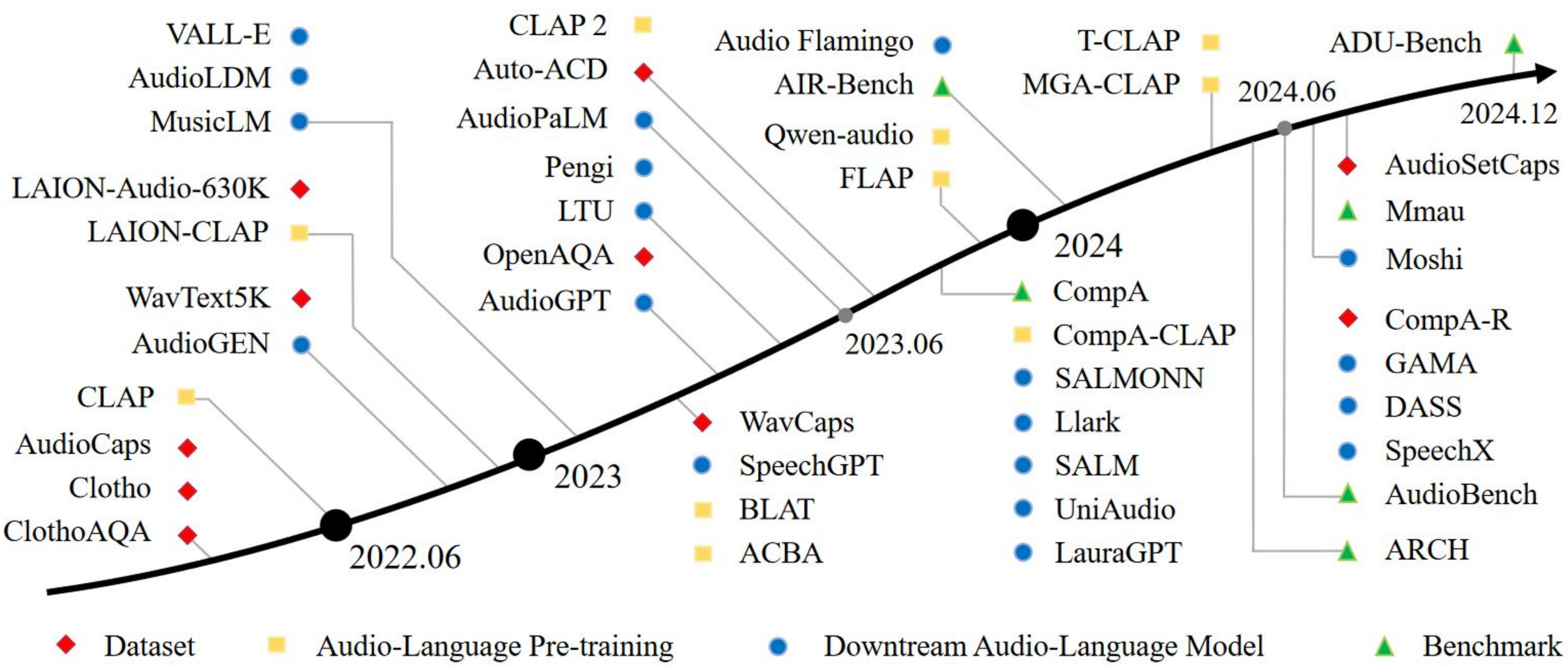
[...] correlation between the reconstructed and the attended speech envelopes is generally weak[2]

[1]: Siqi Cai et al. "EEG-based Auditory Attention Detection in Cocktail Party Environment."
[2]: Enze Su et al. "STAnet: A Spatiotemporal Attention Network for Decoding Auditory Spatial Attention From EEG."

Literature Review

Audio Foundation Models

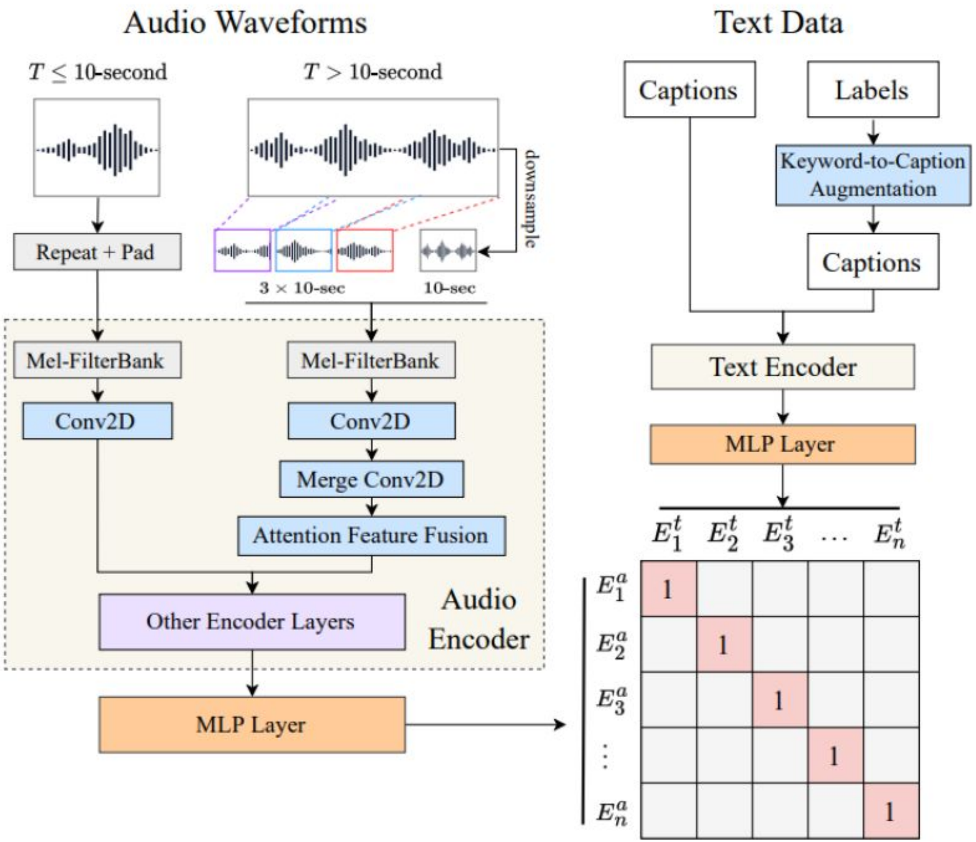
- Larger models
- Our model: LAION-CLAP



Literature Review

LAION-CLAP

- Contrastive Language Audio Pretraining (CLAP)
- Trained on multiple datasets

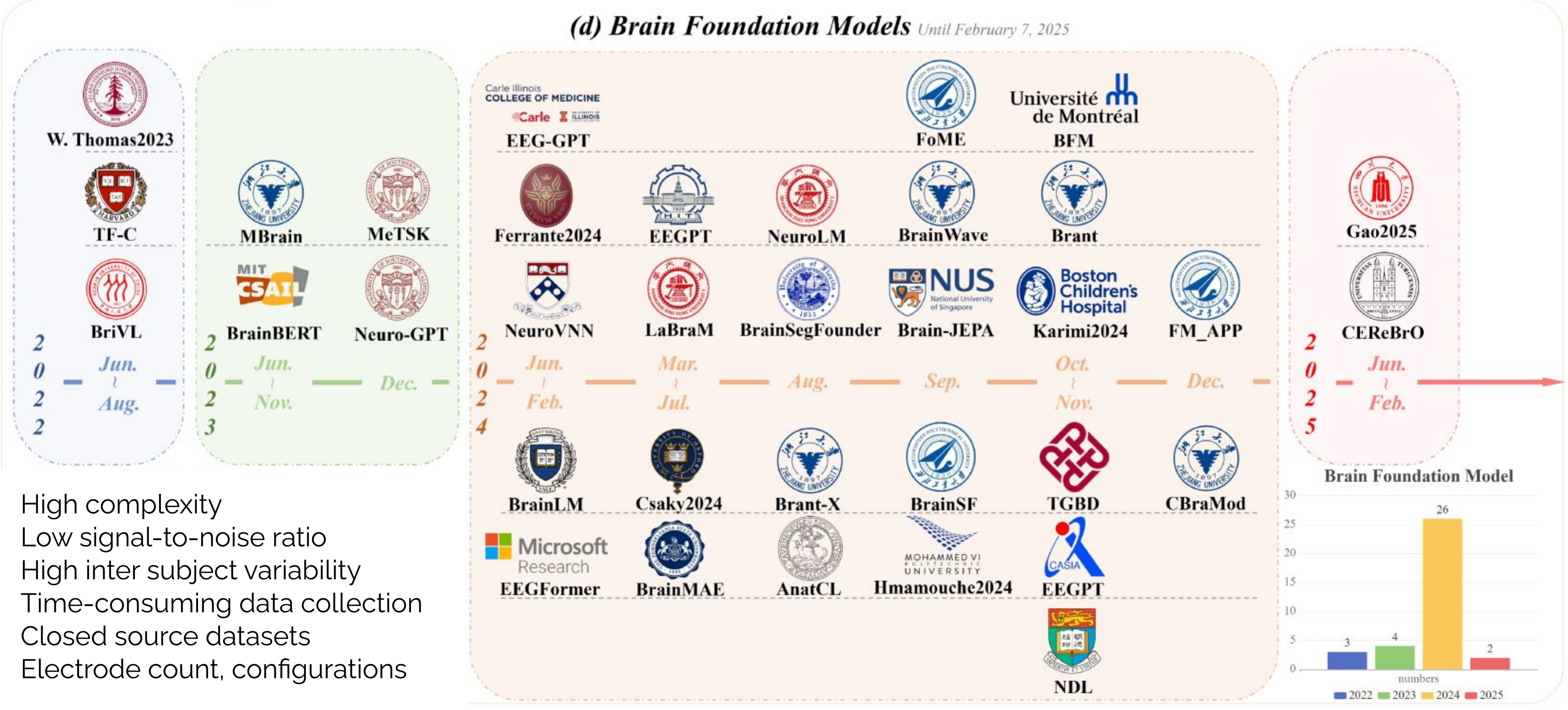


Traffic_Light.wav

A group of people standing on the street near a busy freeway.

Literature Review

Brain Foundation Models



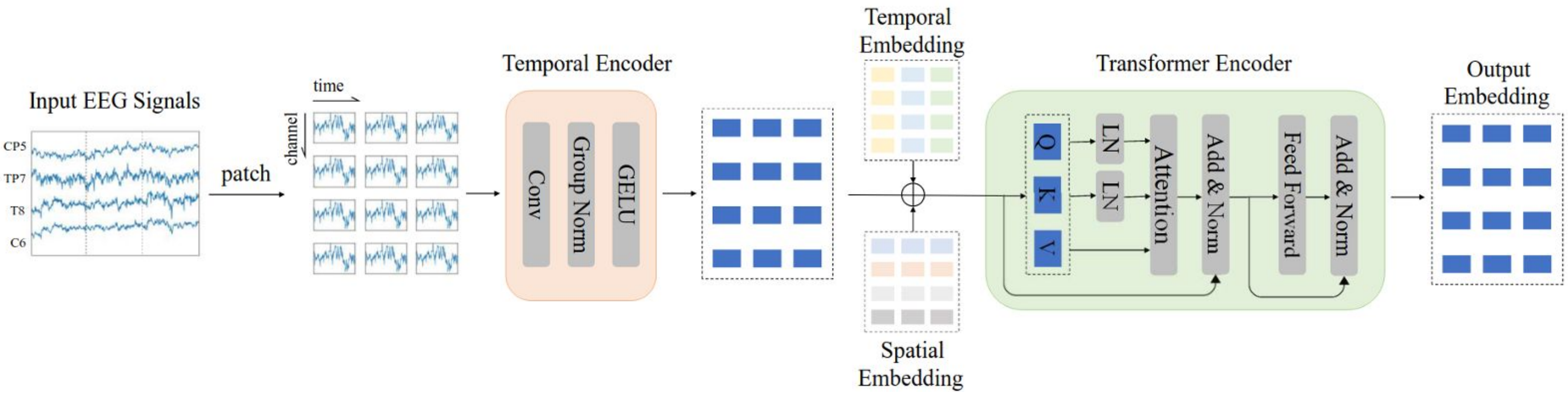
- High complexity
- Low signal-to-noise ratio
- High inter subject variability
- Time-consuming data collection
- Closed source datasets
- Electrode count, configurations

Literature Review

LaBraM

- Large Brain Model (LaBraM)

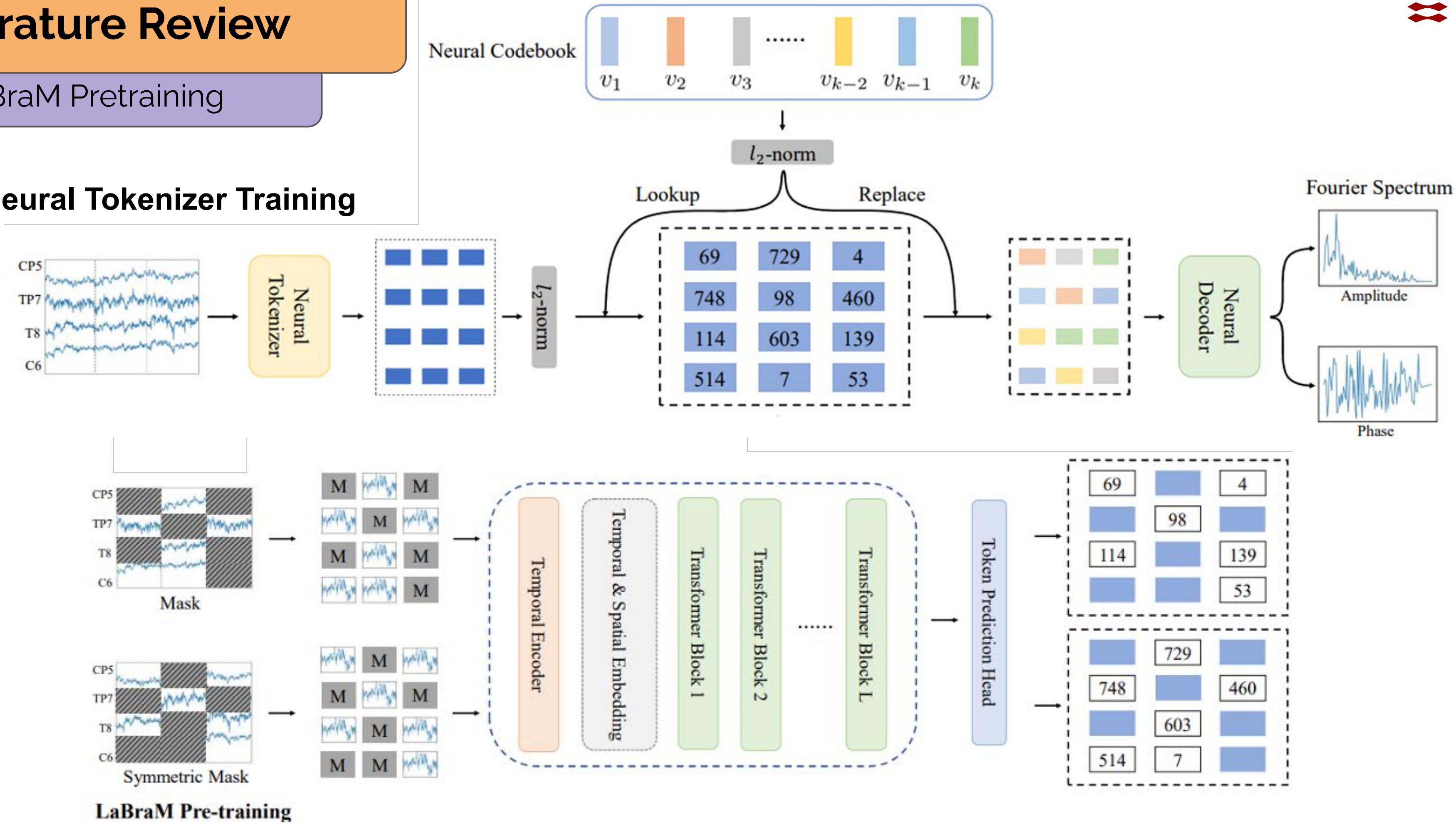
Neural Transformer



Literature Review

LaBraM Pretraining

Neural Tokenizer Training



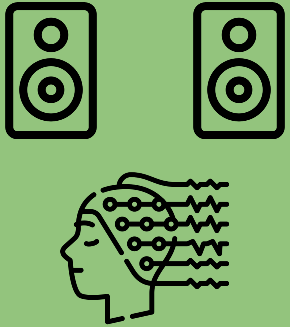
Data

Data

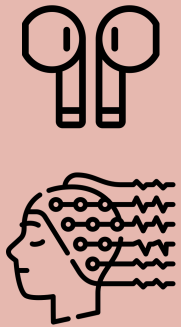
Overview

- 26 subjects
- Five conditions
- Male audio clips: 200, Female audio clips: 165
- Trial length: 1 minute

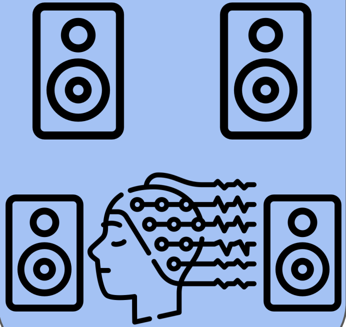
Freefield

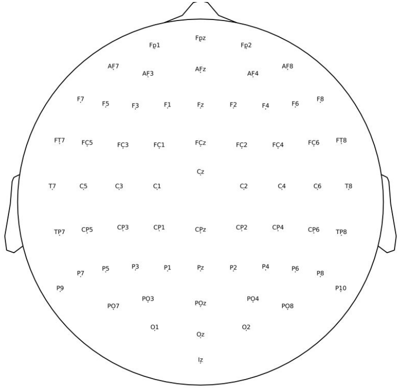


Insertphones



-1,-4,-7dB





Data

Missing data

- 3 subjects missing, left with 23 subjects
- 3364 trials

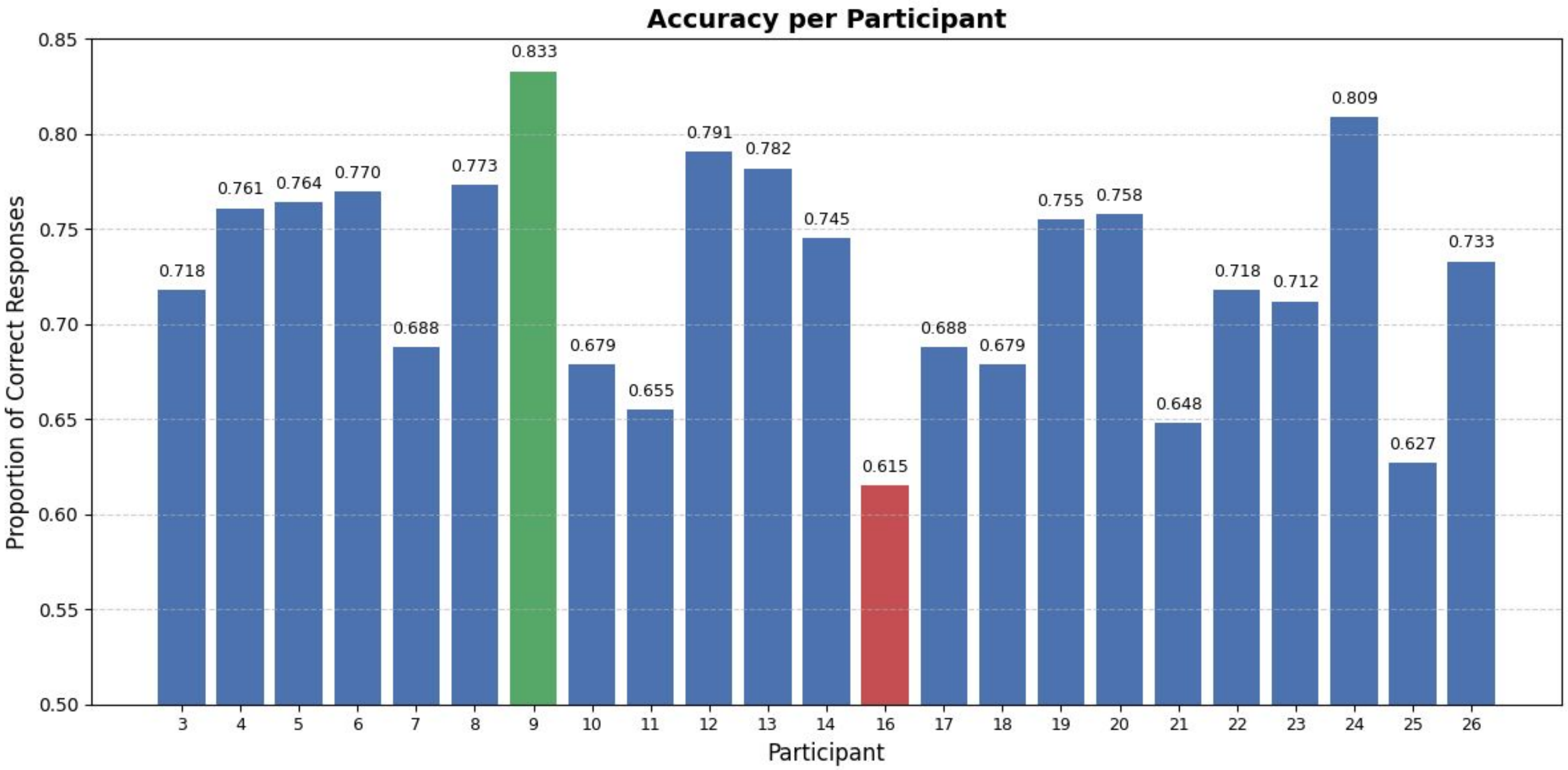
Subject	1	2	4	5	8	14	15	16	23	25
Insert	×	×	✓	✓	×	✓	×	×	✓	✓
Free	×	×	✓	×	✓	✓	×	×	✓	×
-1dB	×	×	✓	✓	✓	×	×	×	×	✓
-4dB	×	×	×	✓	✓	✓	×	✓	✓	✓
-7dB	×	×	✓	✓	✓	✓	×	✓	×	✓

Subject	Condition	# Missing Trials
10	Insert	16
20	-7dB	11
26	Insert	16
26	-4dB	15

Data

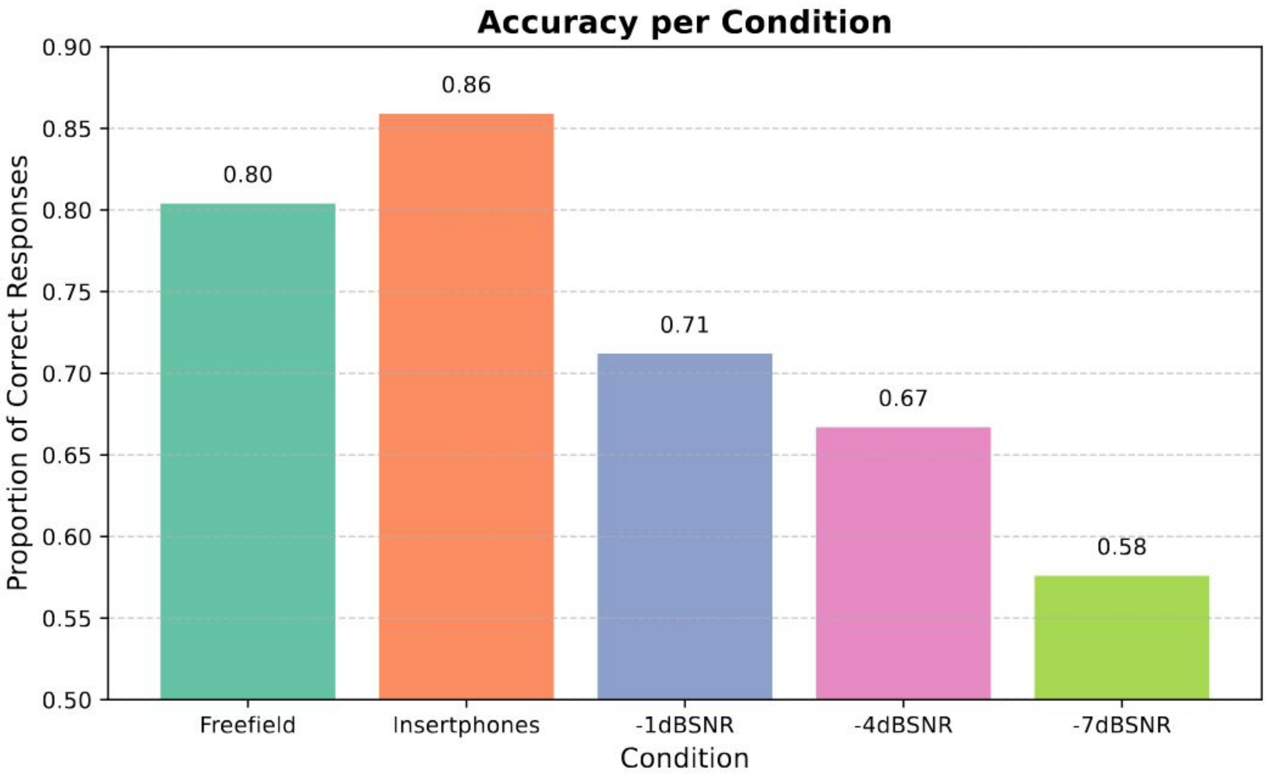
Response accuracy

2 yes/no questions per trial



Data

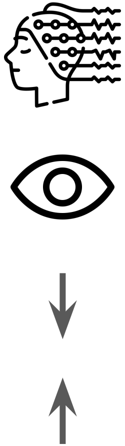
Response accuracy



Data

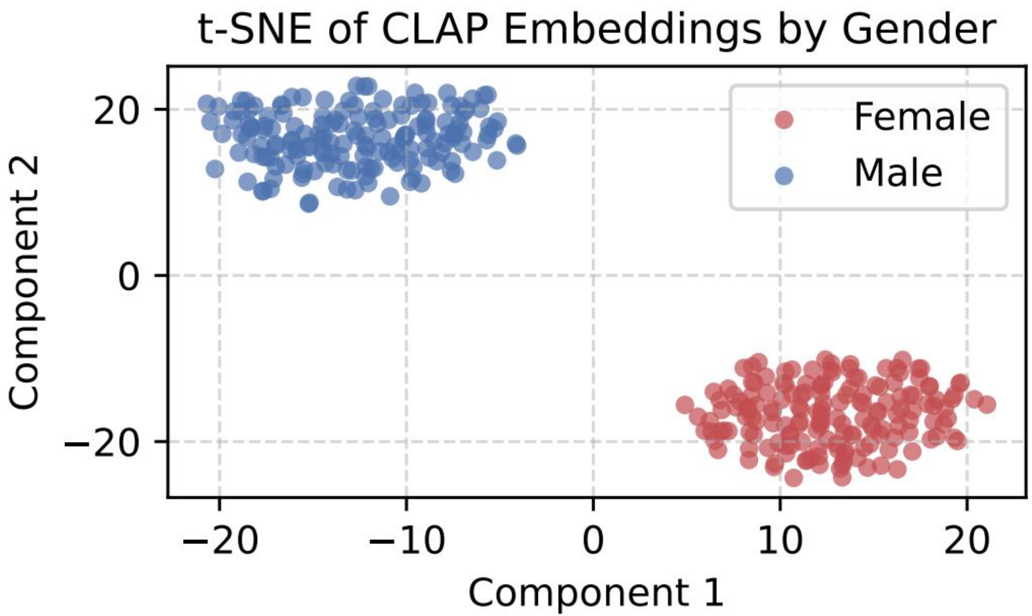
Preprocessing

1. EEG was bandpass filtered between 0.5-30Hz
2. Independent Component Analysis (ICA) to remove EEG artifacts
3. EEG downsampled from 8192Hz → 200Hz
4. Audio upsampled from 44100Hz → 48000Hz



Data

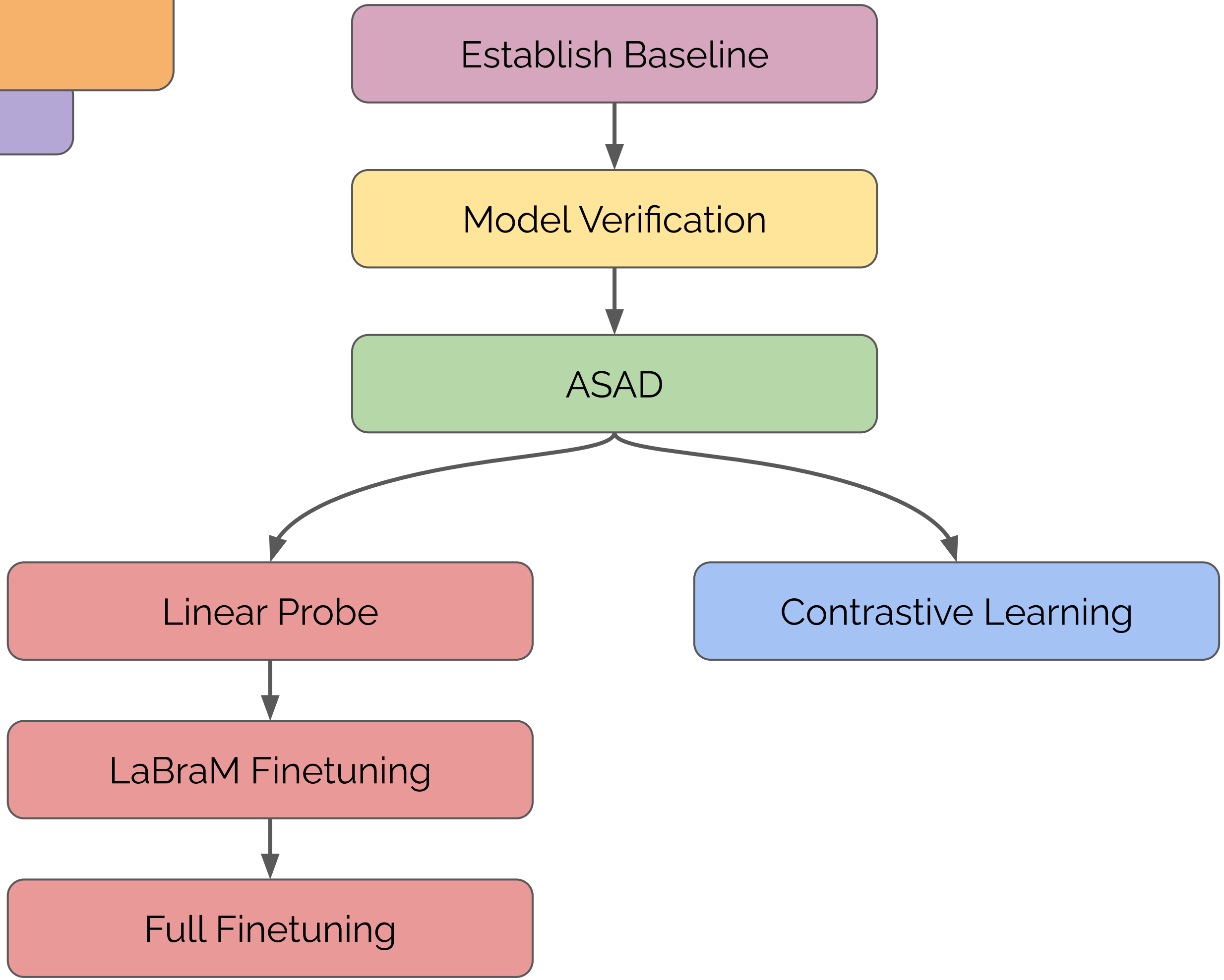
Data visualization



Methodology

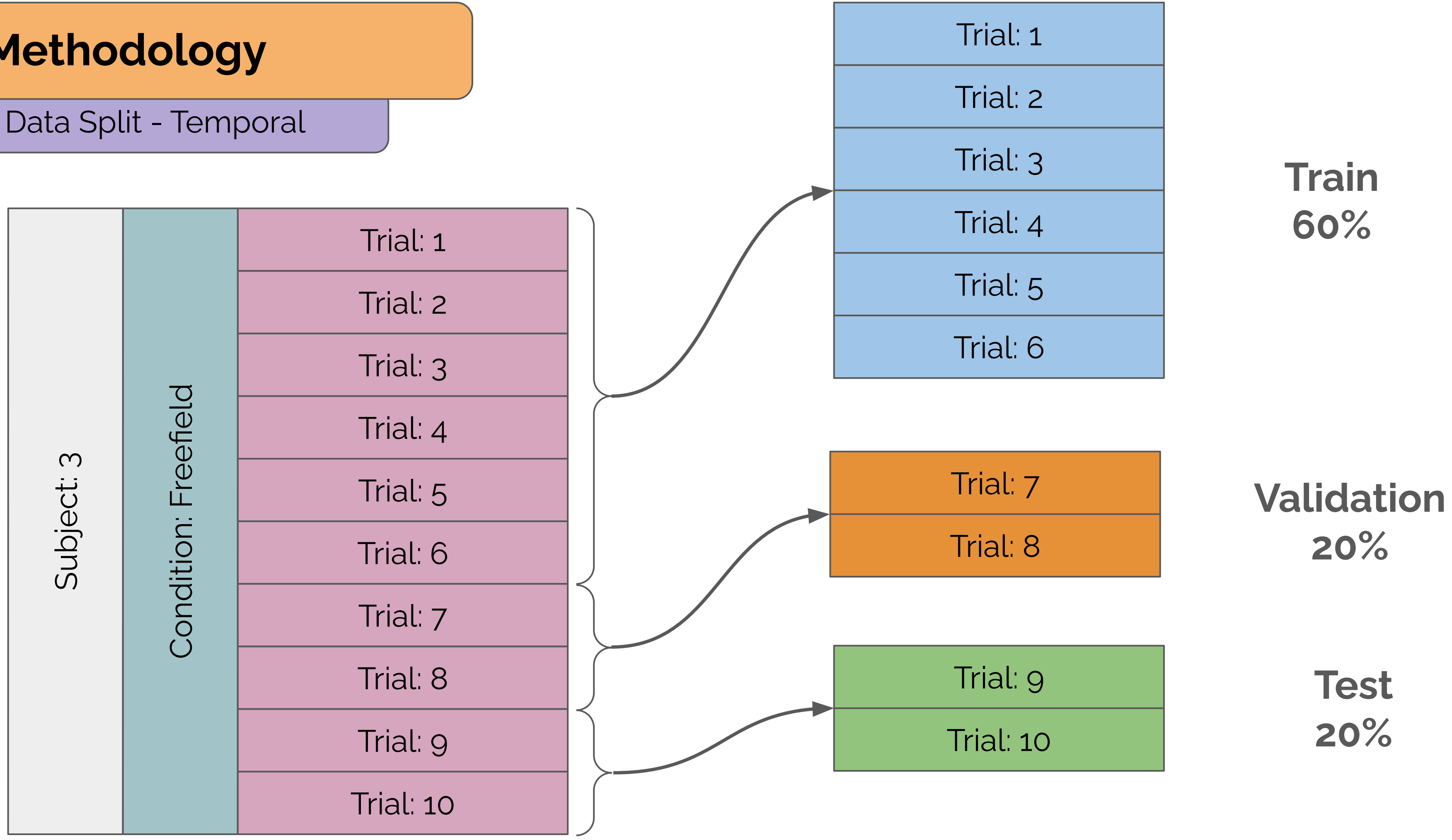
Methodology

Process



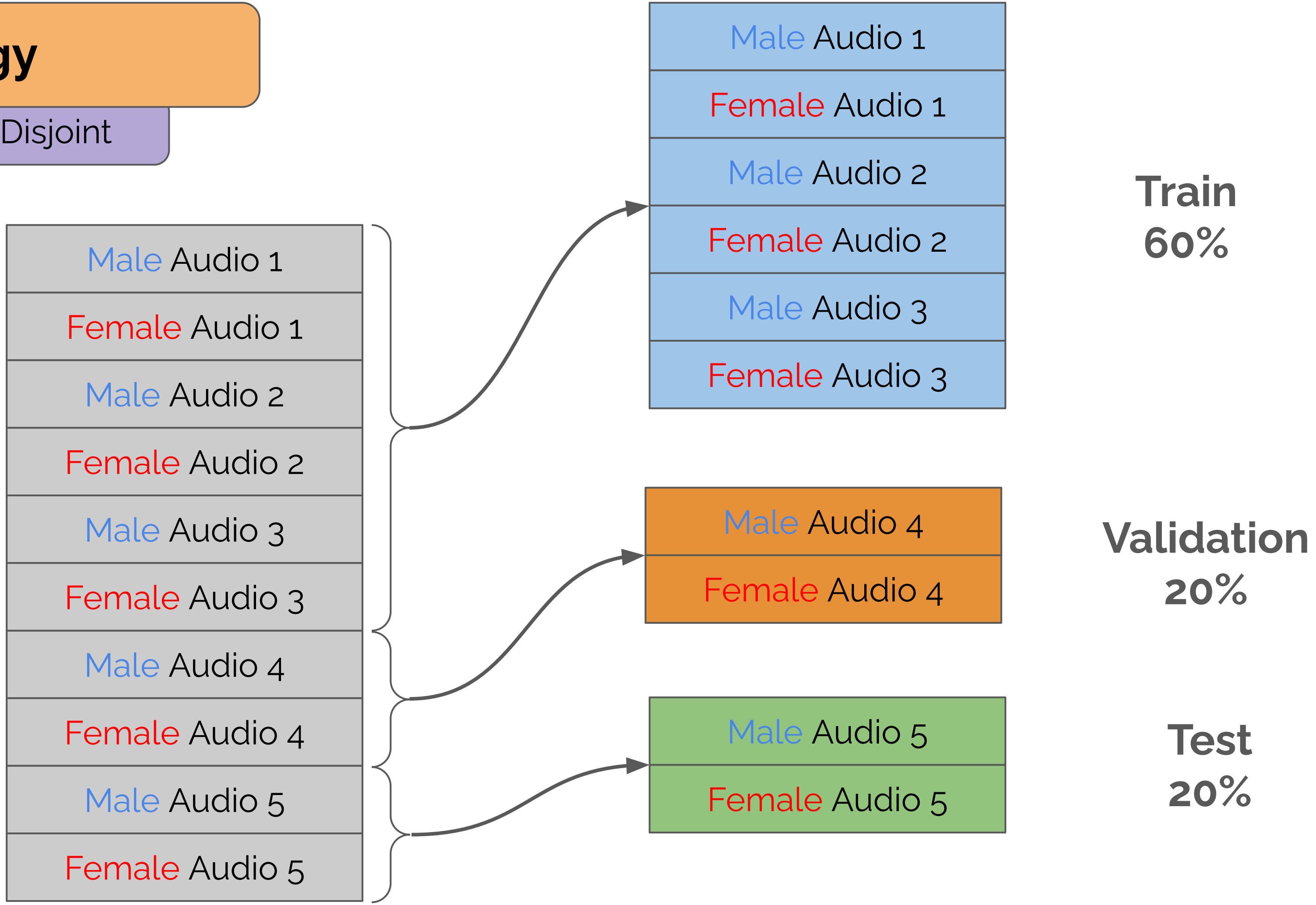
Methodology

Data Split - Temporal



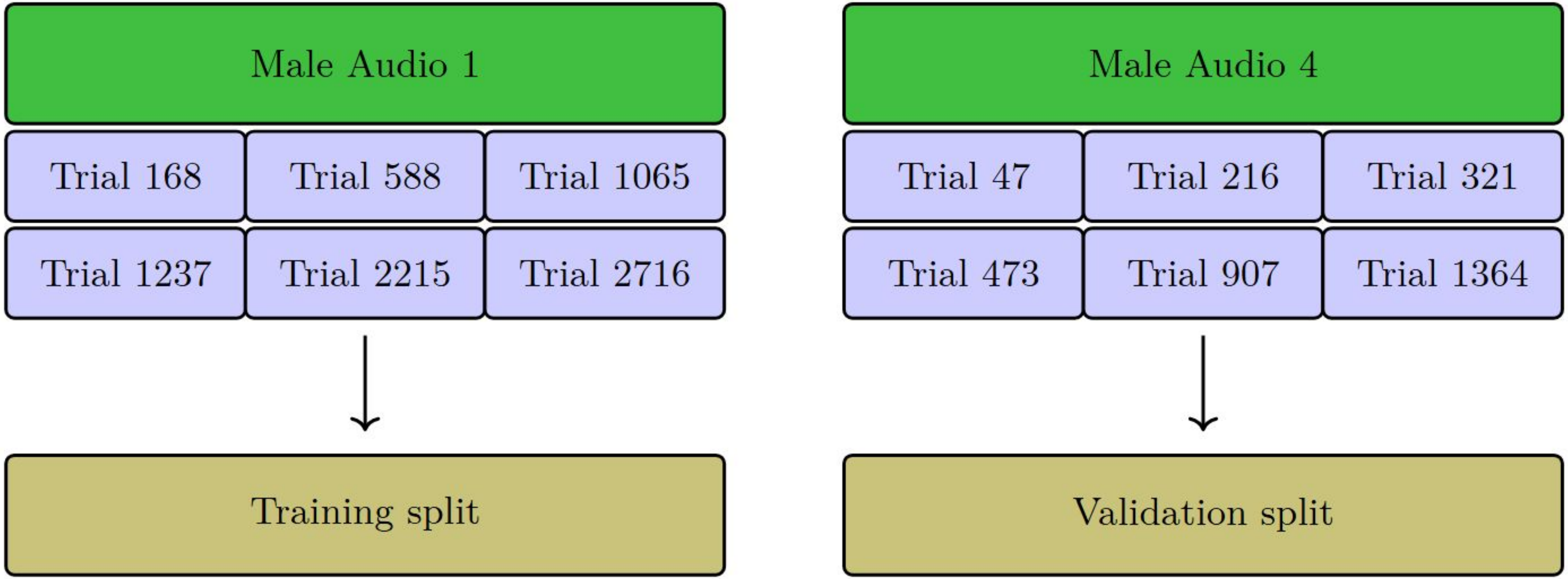
Methodology

Data Split - Audio Disjoint



Methodology

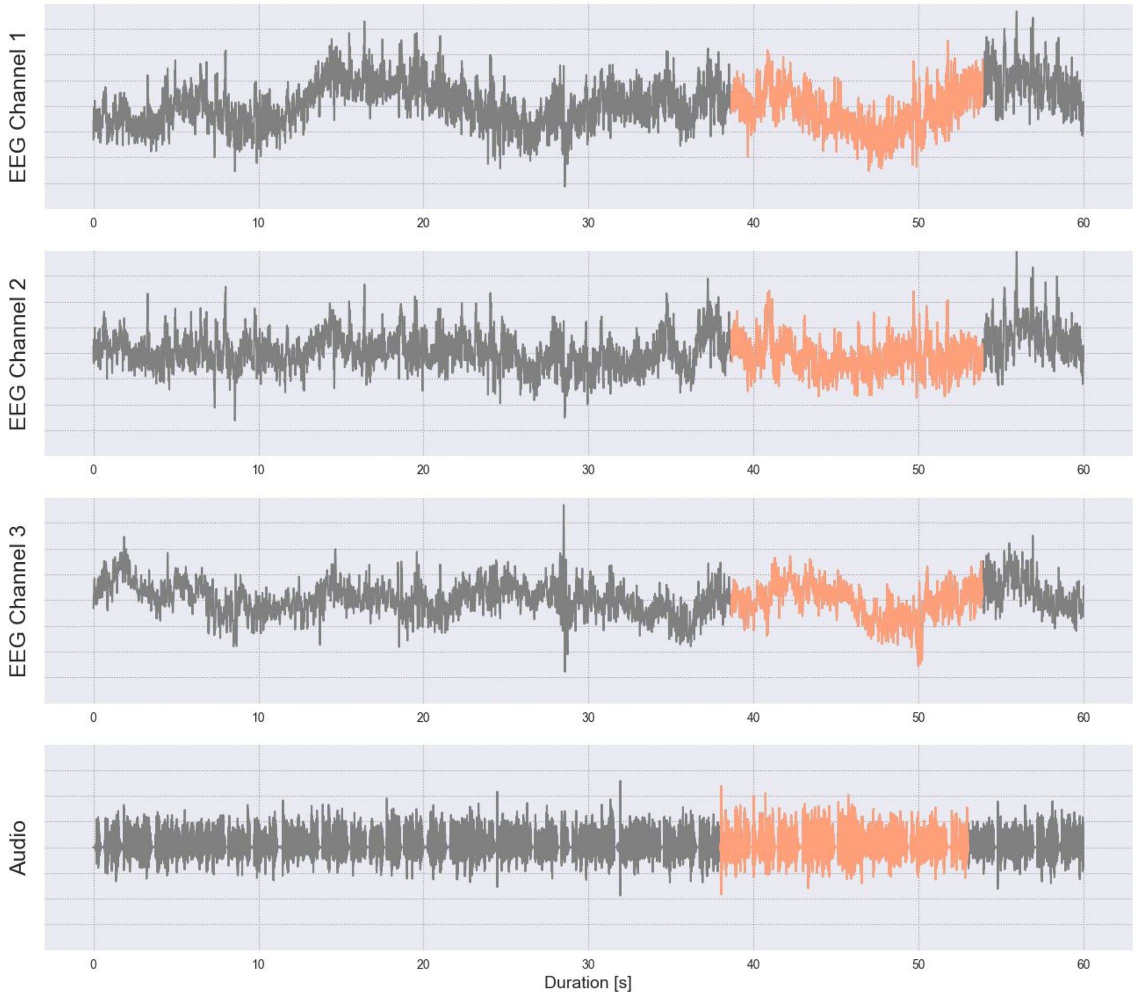
Data Split - Audio Disjoint



Methodology

Trial sampling

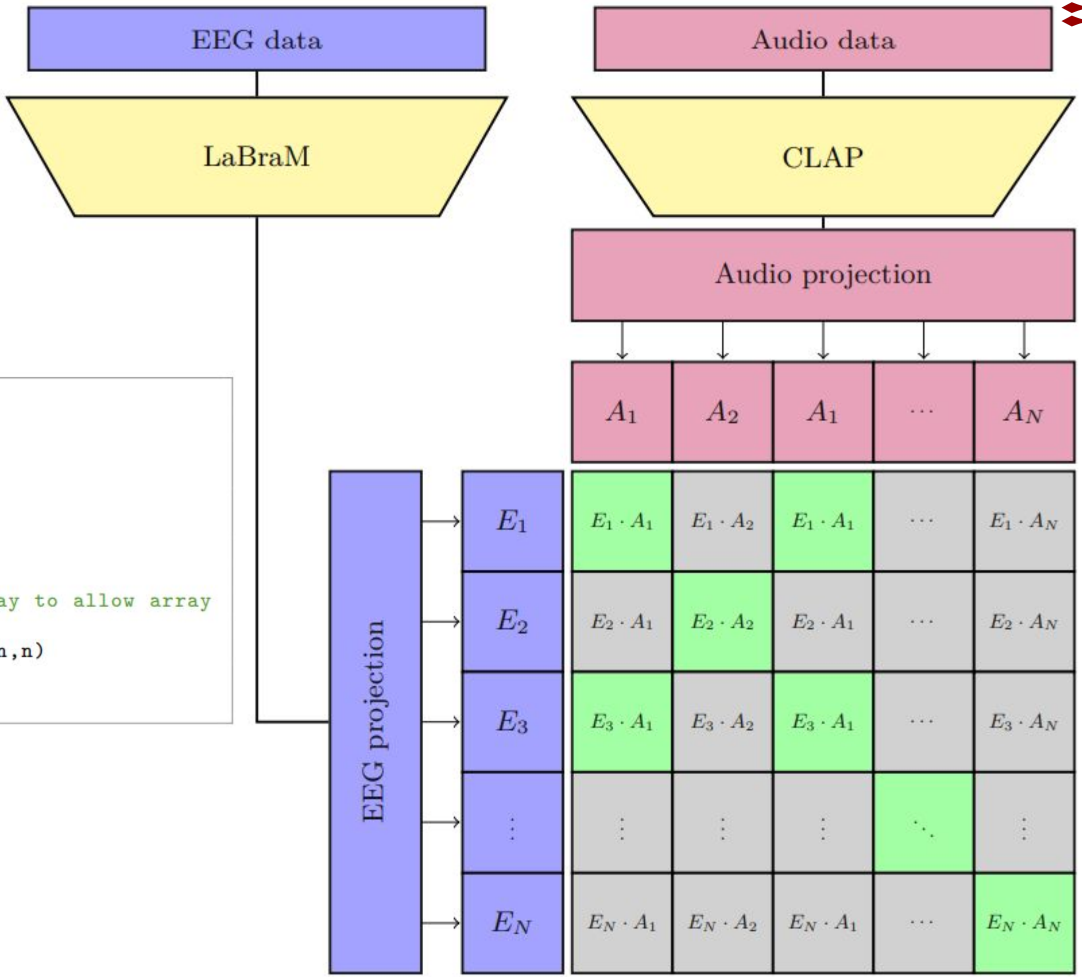
- Randomized trial segments
- Fixed validation segments
- Three augmentations:
  - Channel dropout
  - FT Surrogate
  - Time Reverse



Methodology

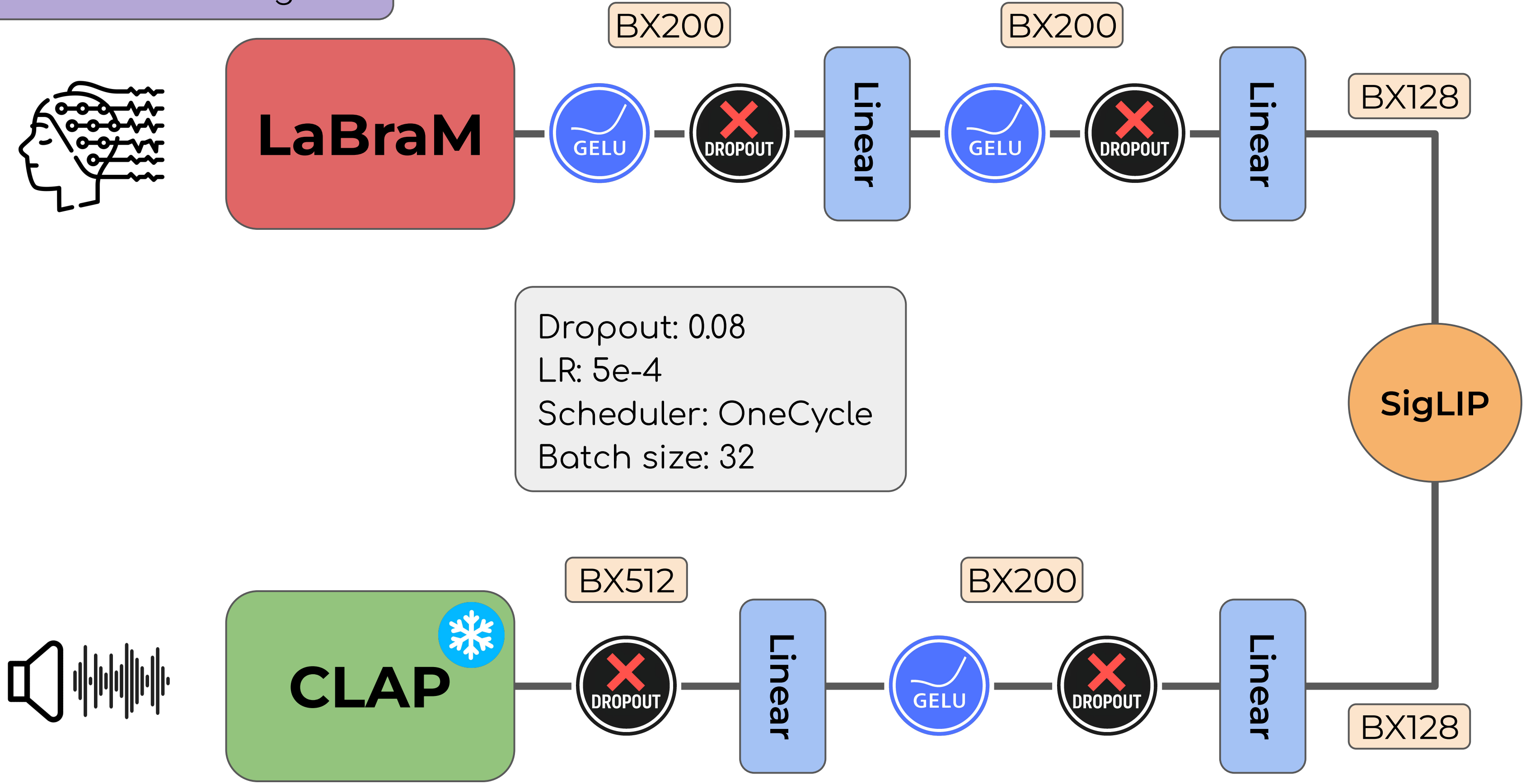
Contrastive learning

```
# eeg_embed - EEG model embedding [n, d]
# audio_embed - Audio model embedding [n, d]
# target_ids - ids of audio segments [n]
# b, t_prime - learnable bias and temperature
# n - mini-batch size
eeg_embed_z = l2_normalize(eeg_embed)
audio_embed_z = l2_normalize(audio_embed)
t = exp(t_prime)
# ~ is used as a short hand for adding a new axis to an array to allow array
    broadcasting
labels = 2 * (target_ids[:, ~] == target_ids[~, :]) - ones(n,n)
logits = dot(eeg_embed_z, audio_embed_z.T) * t + b
loss = -sum(log_sigmoid(labels * logits)) / n
```



Methodology

Contrastive learning



Dropout: 0.08
LR: 5e-4
Scheduler: OneCycle
Batch size: 32

Results & Discussion

Results & Discussion

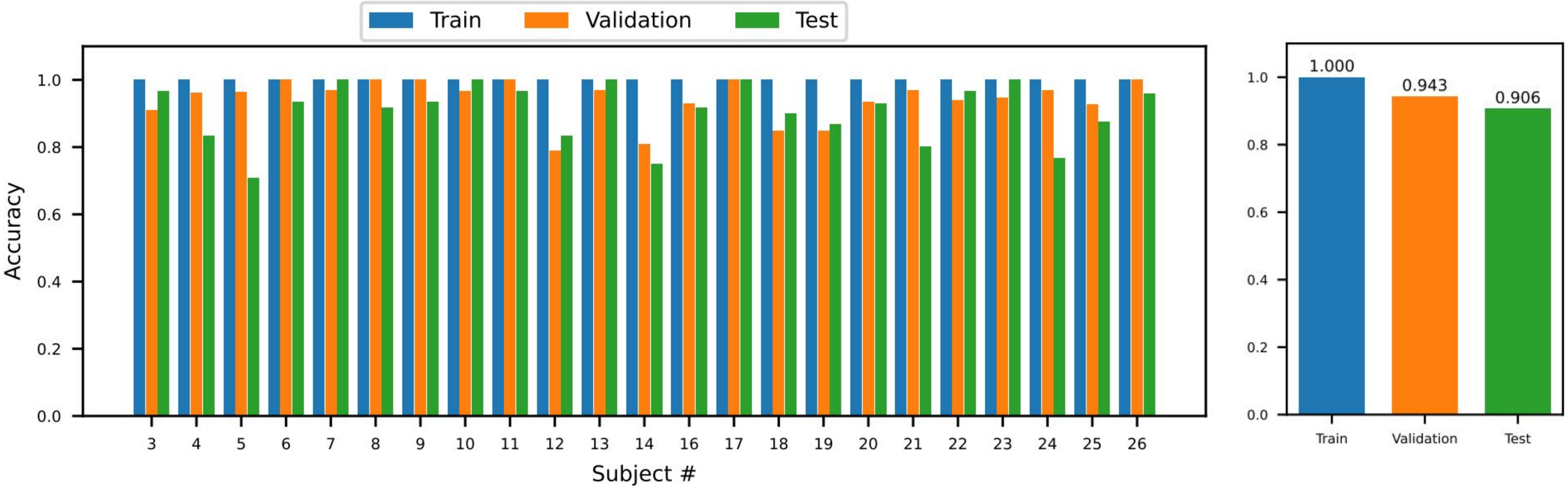
Baseline

- Each experiment used a 15 second decision window
- Only ran experiments with a single seed
- Backwards model

# Conditions	Validation accuracy	Test accuracy
Two conditions	0.643	0.604
Five conditions	0.564	0.568

Results & Discussion

Condition classification

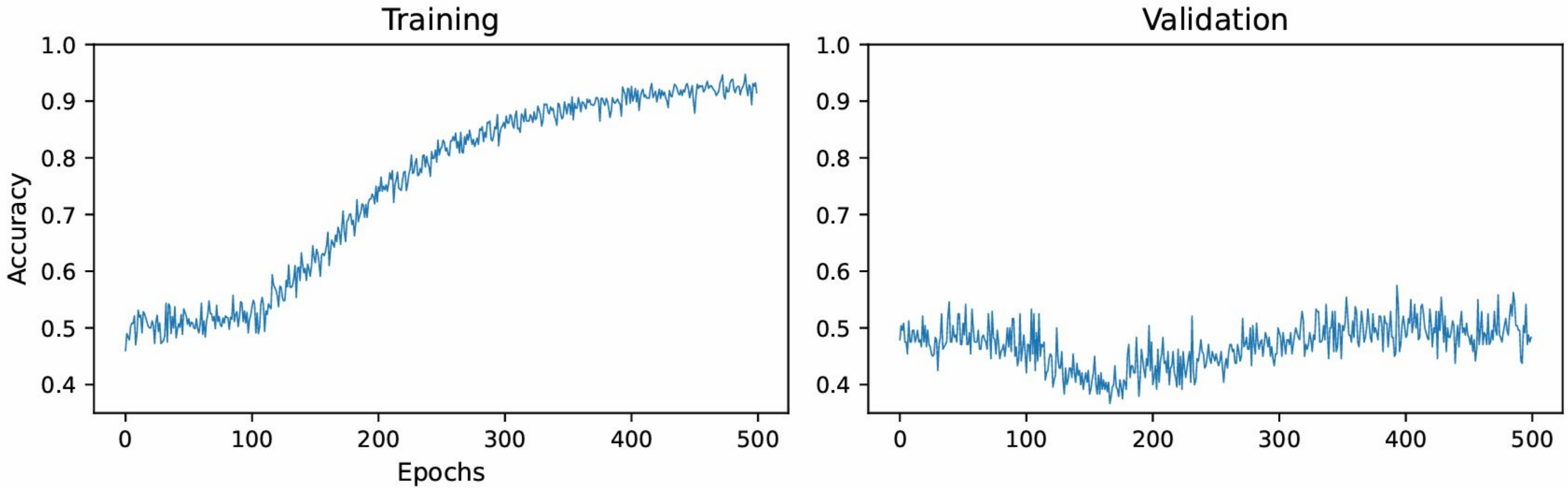


Results & Discussion

Contrastive learning

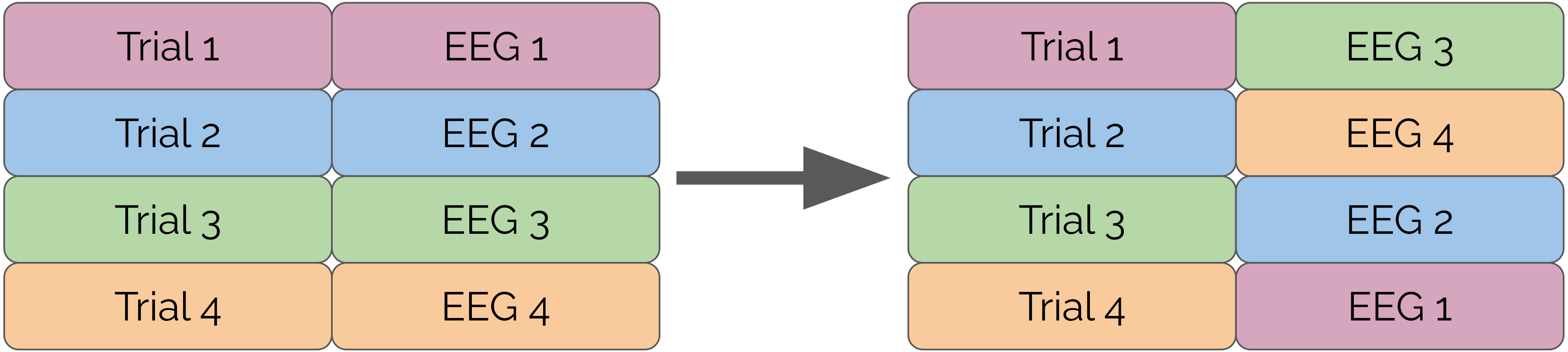
- Overfitting
- Memorization

Temporal Split



Results & Discussion

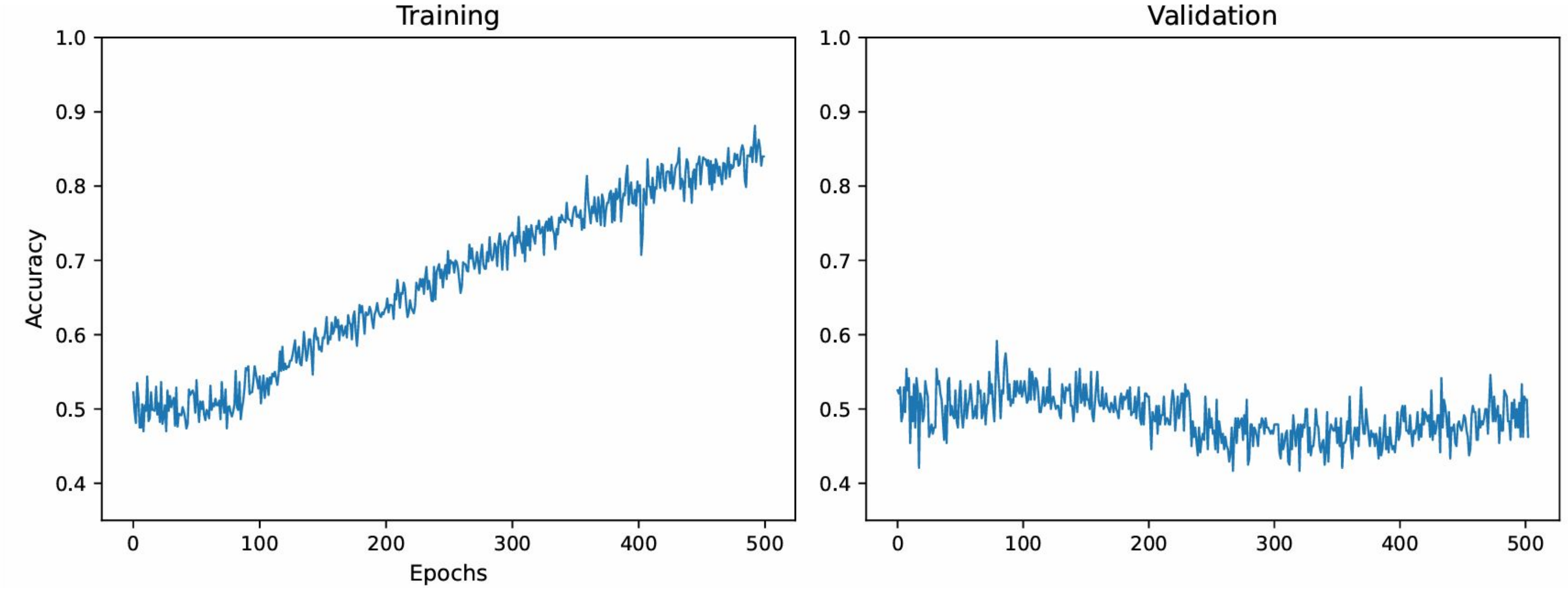
Contrastive learning



Results & Discussion

Contrastive learning

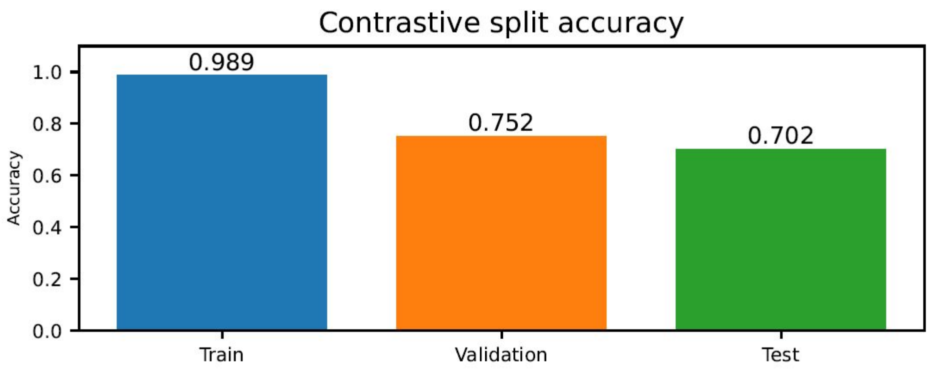
Temporal Split with mismatched EEG



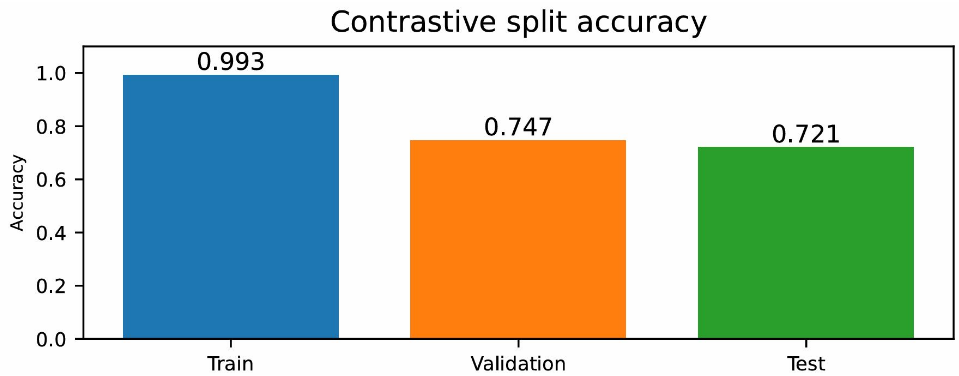
Results & Discussion

Contrastive learning

Noise-free conditions



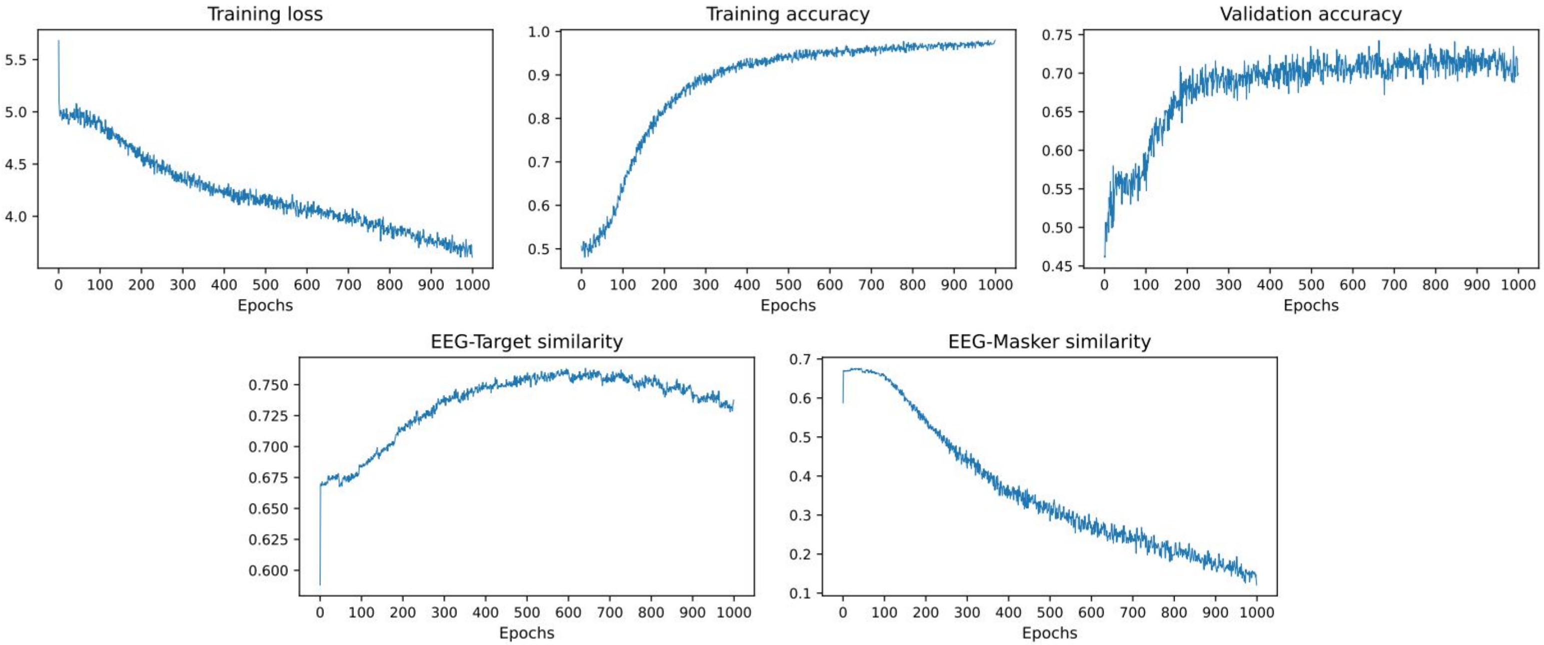
All conditions



Results & Discussion

Contrastive learning

All conditions

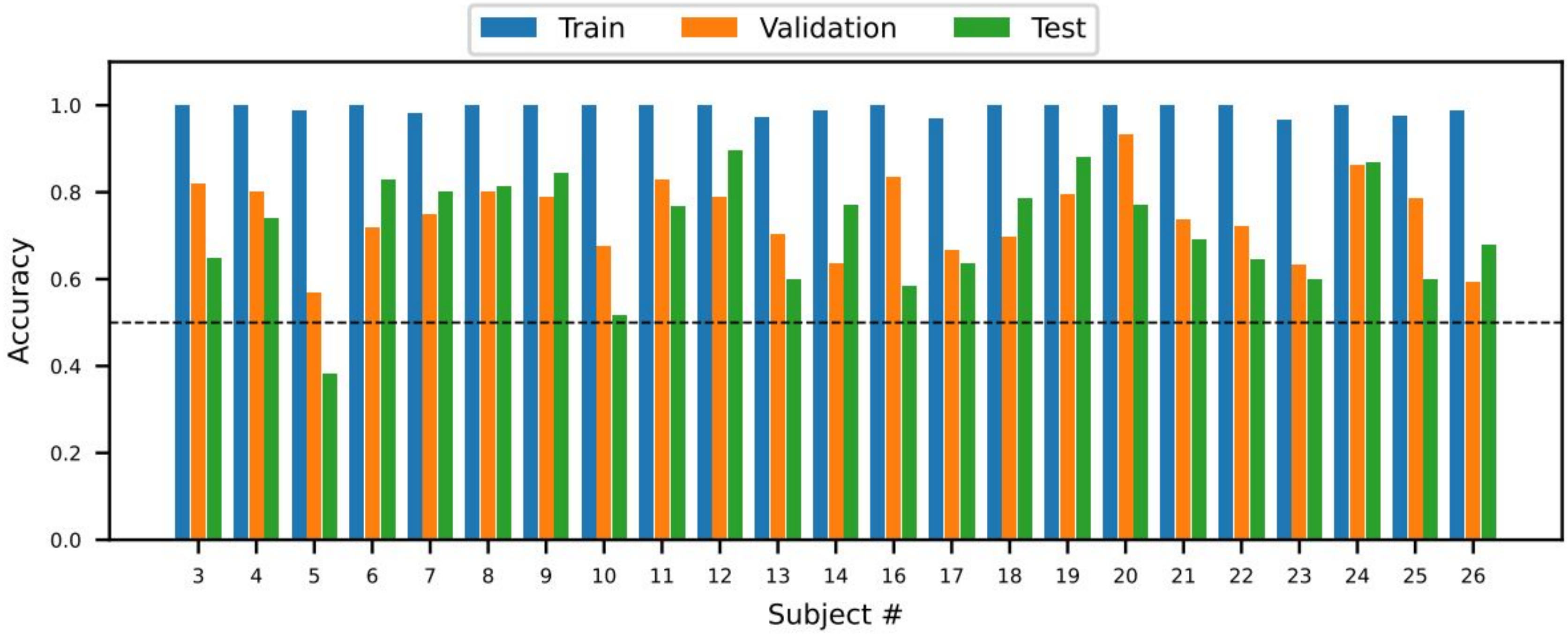


Results & Discussion

Contrastive learning

- Better than random guessing
- High response accuracy + no missing data-> high model accuracy (9, 12, 24)

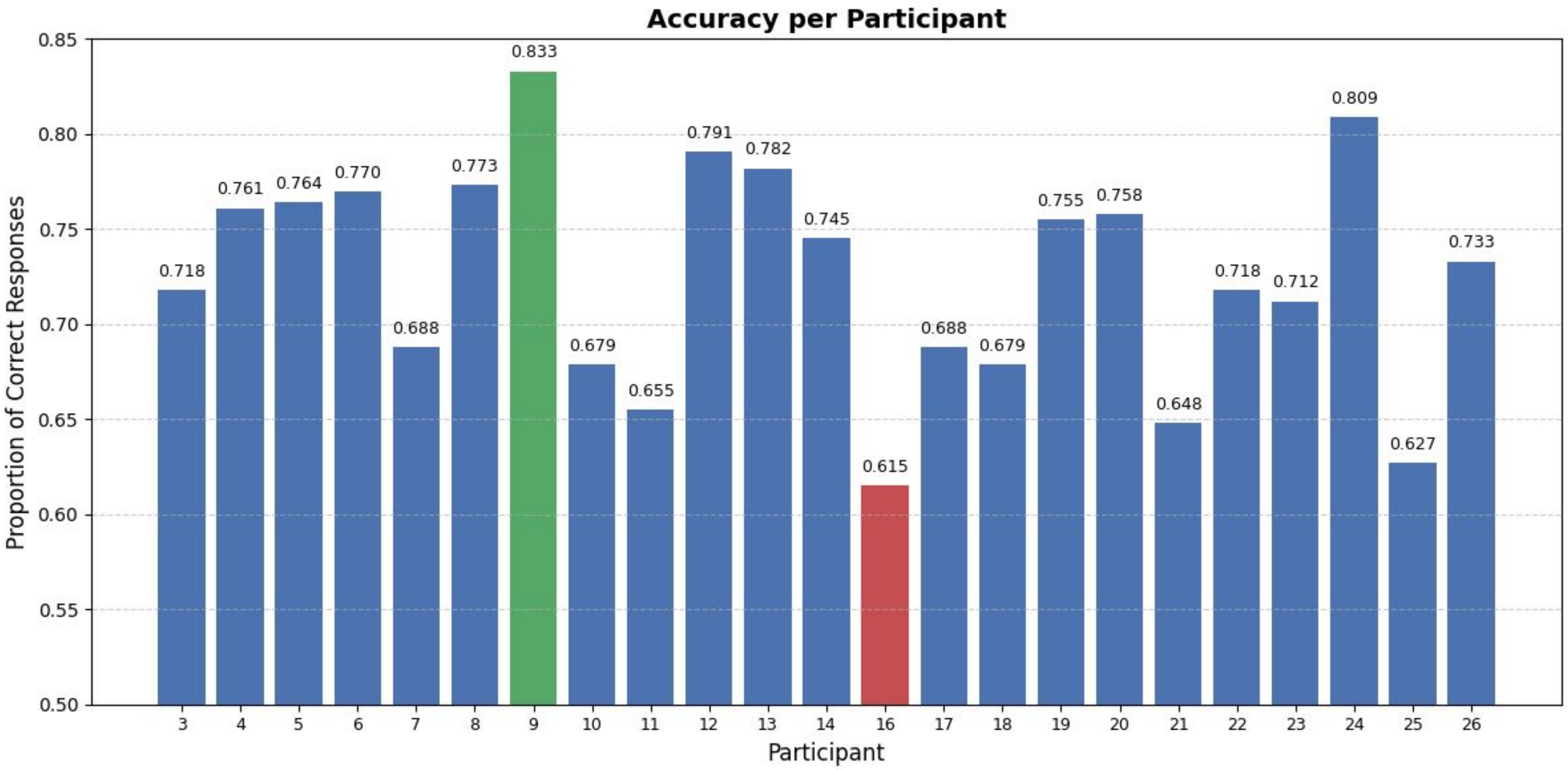
Subject accuracy on all conditions



Data

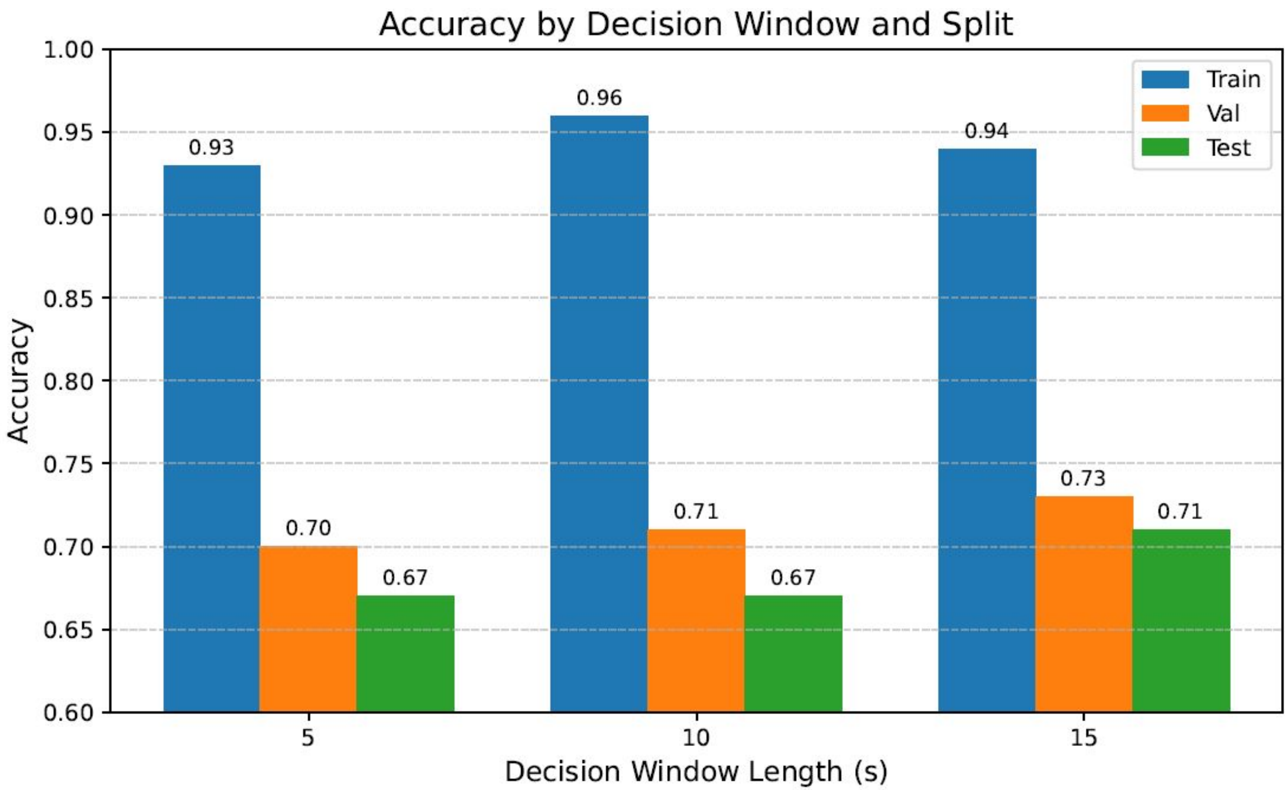
Response accuracy

2 yes/no questions per trial



Results & Discussion

Contrastive learning



Results & Discussion

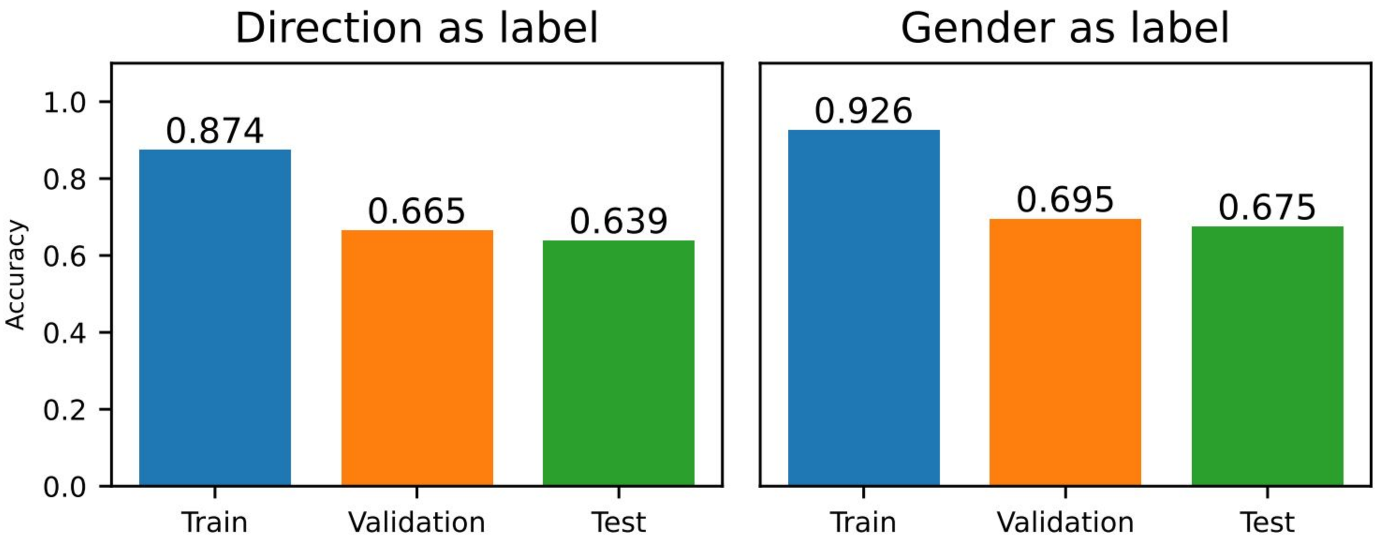
Augmentation results

- TR: Time Reversal
- DR: Channel Dropout
- FTS: Fourier Transform Surrogate

	Train	Val	Test
No aug	**0.989**	**0.752**	0.702
TR	0.987	0.748	**0.725**
DR	0.946	0.711	0.667
FTS	0.905	0.714	0.694

Results & Discussion

ASAD



Results & Discussion

Direct classification

	Train	Validation	Test
Linear probe	0.572	0.521	0.522
LaBraM finetuning	**0.984**	**0.707**	**0.676**
Full finetuning	0.722	0.523	0.492

Conclusion

Conclusion

RQ1

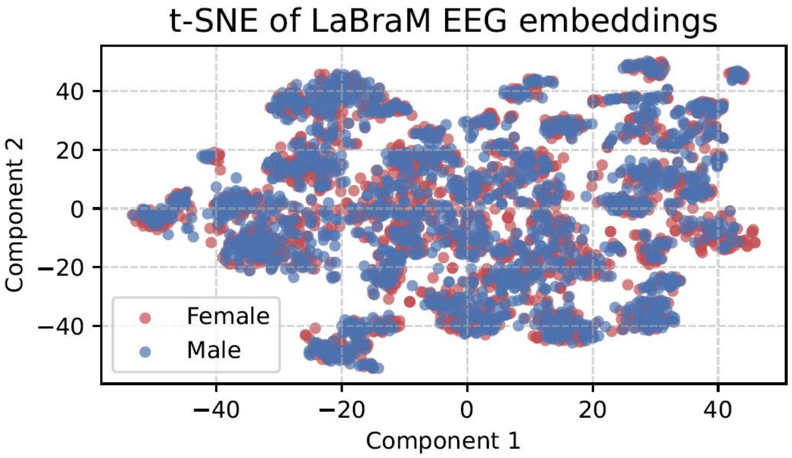
RQ1: How do CLAP and LaBraM perform as pretrained feature extractors for auditory attention decoding?

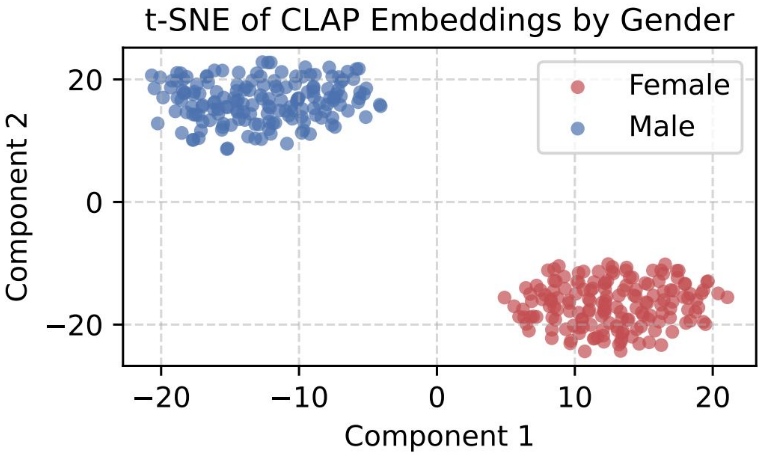
	Train	Validation	Test
Linear probe	0.572	0.521	0.522

Conclusion

RQ1

RQ1: How do CLAP and LaBraM perform as pretrained feature extractors for auditory attention decoding?

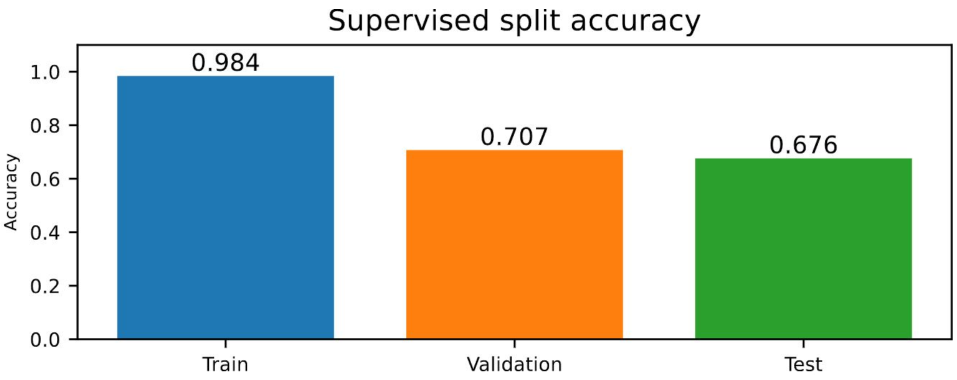


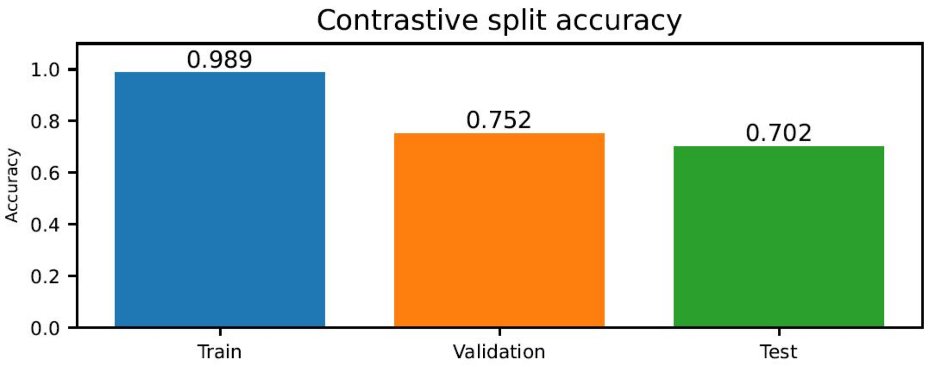


Conclusion

RQ2

RQ2: How does contrastive learning compare to supervised classification for training robust AAD models using CLAP and LaBraM?

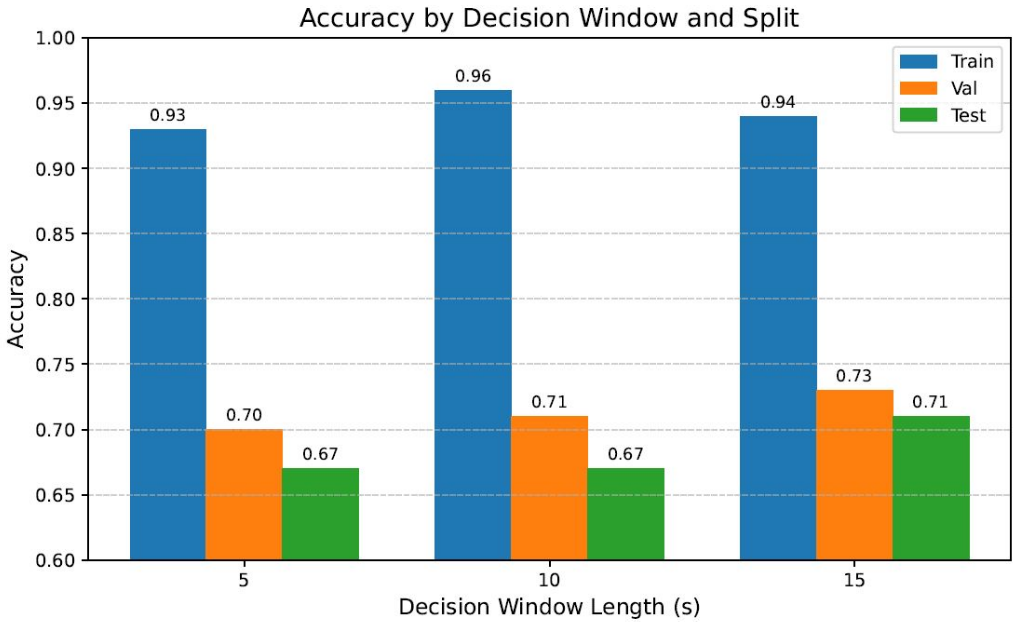




Conclusion

RQ3

RQ3: How does the length of decision windows affect performance?



Thank you for your Attention

Appendix

Results & Discussion

Baseline

- Each experiment used a 15 second decision window
- Only ran experiments with a single seed
- Backwards TRF model

Two condition performance

Split	Validation accuracy	Test accuracy
Temporal	0.588	0.633
Audio-disjoint	0.643	0.604

Five condition performance

Split	Validation accuracy	Test accuracy
Temporal	0.593	0.599
Audio-disjoint	0.564	0.568

Literature Review

Why Direct Classification?

[..] the process of stimulus reconstruction [..] is not optimized to effectively detect attention. [...] the compression of multichannel EEG signals into a single waveform through stimulus reconstruction reduces the available information for analysis[1]

[The neural network] outperforms the baseline linear stimulus reconstruction method, improving decoding accuracy [...] from 59% to 87%[2]

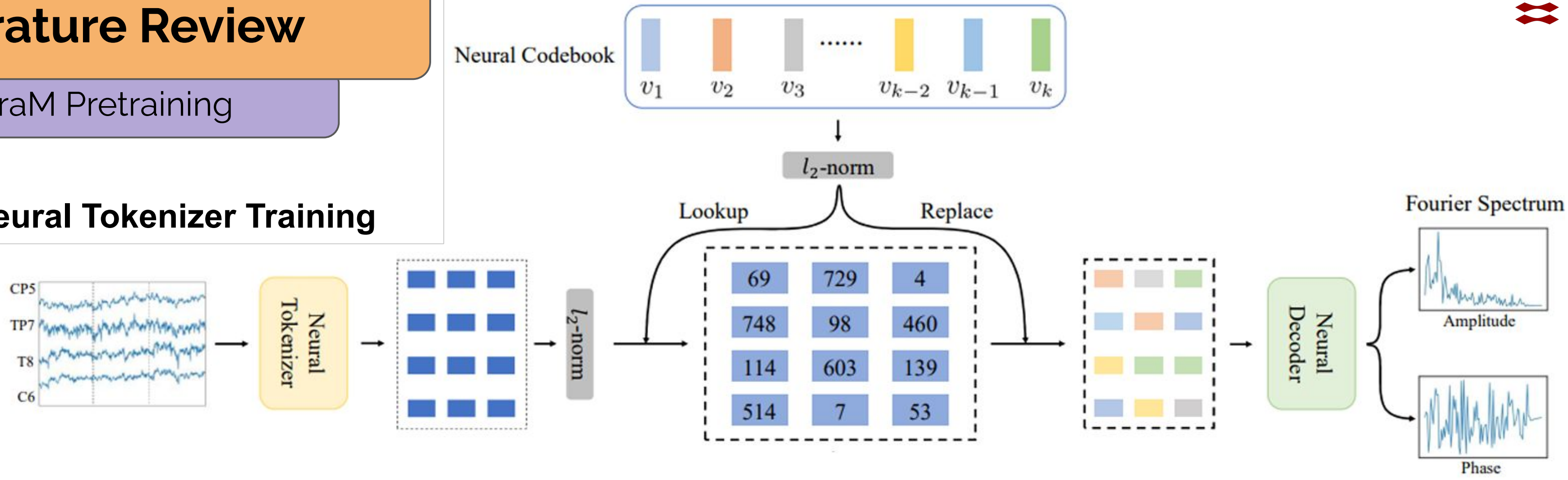
[...] correlation between the reconstructed and the attended speech envelopes is generally weak[3]

[1]: Siqi Cai et al. “EEG-based Auditory Attention Detection in Cocktail Party Environment.”
[2]: Gregory Ciccarelli et al. “Comparison of Two-Talker Attention Decoding from EEG with Nonlinear Neural Networks and Linear Methods.”
[3]: Enze Su et al. “STAnet: A Spatiotemporal Attention Network for Decoding Auditory Spatial Attention From EEG.”

Literature Review

LaBraM Pretraining

Neural Tokenizer Training



$$\mathcal{L}_T = \sum_{x \in \mathcal{D}} \sum_{i=1}^{N} \left\| o_i^A - A_i \right\|_2^2 + \left\| o_i^\phi - \phi_i \right\|_2^2 + \left\| \mathrm{sg}(\ell_2(p_i)) - \ell_2(v_{z_i}) \right\|_2^2 + \left\| \ell_2(p_i) - \mathrm{sg}(\ell_2(v_{z_i})) \right\|_2^2$$

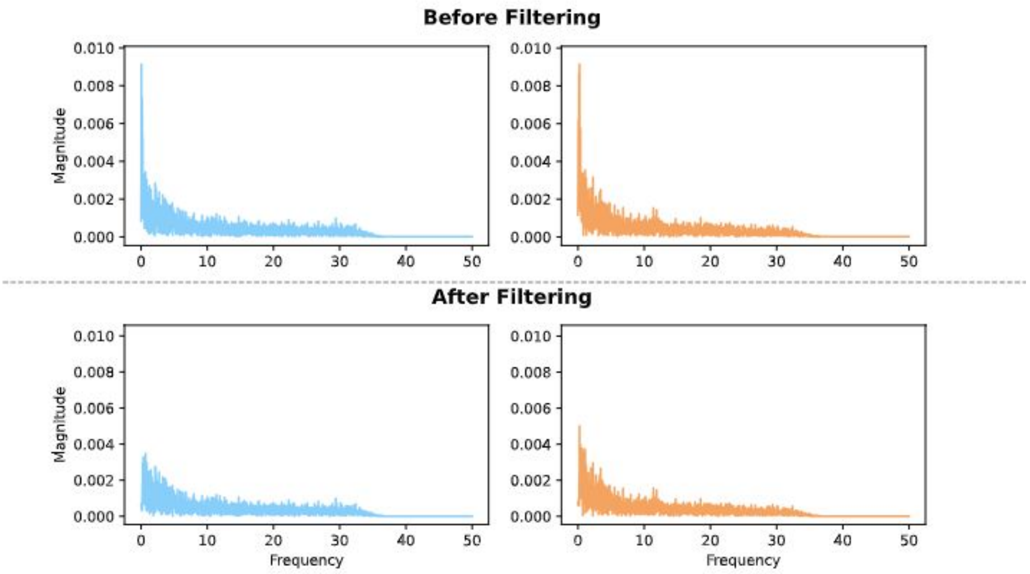
Predicted amplitude
Predicted phase
Tokenizer Vector

Actual amplitude
Actual phase
Codebook Vector

Data

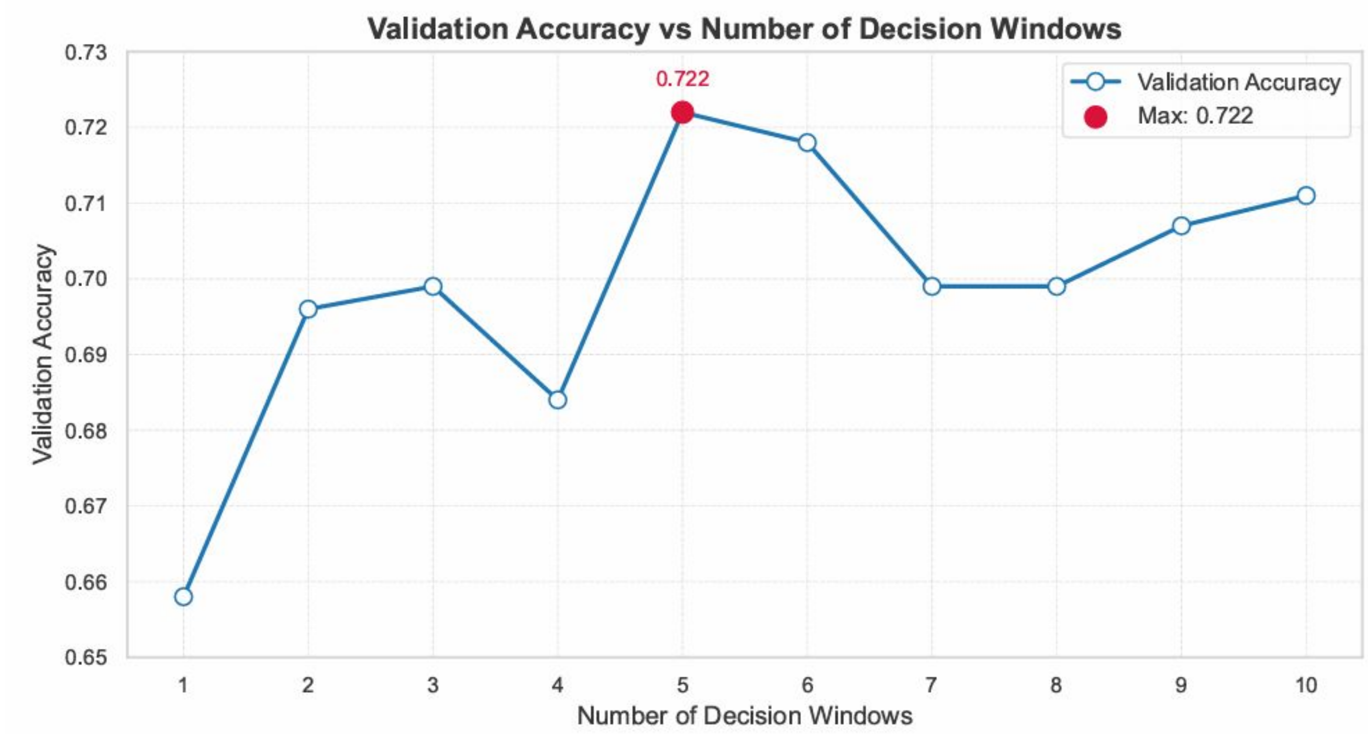
Preprocessing

- EEG was bandpass filtered between 0.5-30Hz
- ICA to remove EEG artifacts
- EEG downsampled from 8192Hz → 200Hz
- Audio upsamples from 44100Hz → 48000Hz



Results & Discussion

Contrastive learning



Results & Discussion

Comparisons

Lund Contrastive

- Hearing impaired subjects
- Unspecified background noise
- CNN + attention
- Subject specific architecture

Lund DCCA

- No added background noise
- Whisper + Deep Canonical-correlation analysis

	Lund Contrastive[1]	Lund DCCA[2]	Our Model
Accuracy	71.5%	67.9%	67.0%

(5 second decision window)

[1] Gautam Sridhar et al. "Improving auditory attention decoding i noisy environments for listeners with hearing impairment through contrastive learning"
[2] Alessandro Celoria et al. "An ASR-based Hybrid Approach for Auditory Attention Decoding"

Results & Discussion

Out-of-sample classification