GHENT UNIVERSITY
GLOBAL CAMPUS

MuSe 2023

The Multimodal Sentiment Analysis Challenge: Humor, Emotional Reactions, and Stress

MuSe-Personalization 2023: Feature Engineering, Hyperparameter Optimization, and Transformer-Encoder Re-discovery

Presenters: Ho-min Park

Authors: Ho-min Park, Ganghyun Kim, Arnout Van Messem, Wesley De Neve

Date: October 29, 2023

Location: Ottawa, Canada

GHENT UNIVERSITY GLOBAL CAMPUS

Center for Biosystems & Biotech Data Science

LIÈGE université

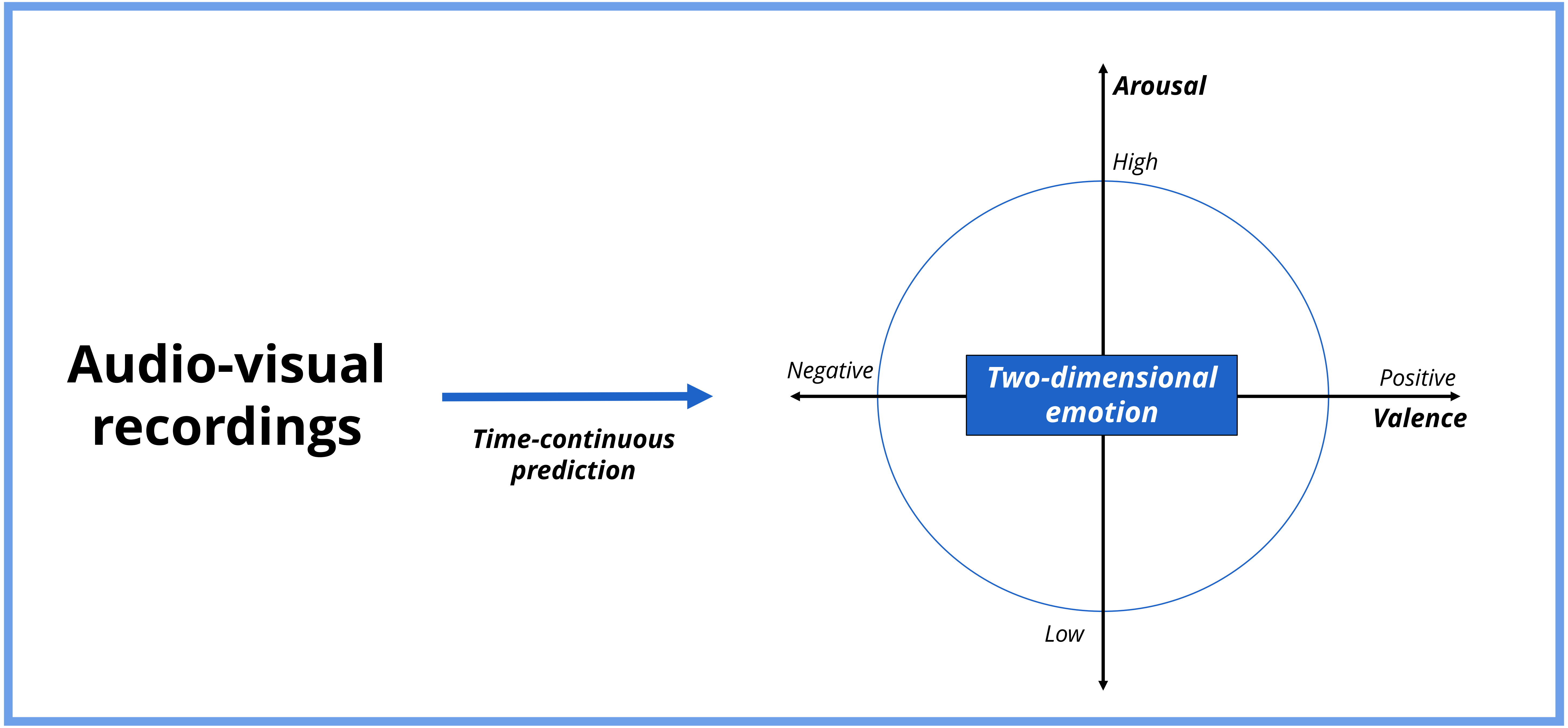
OUTLINE

1. INTRODUCTION

1. INTRODUCTION

01
Introduction

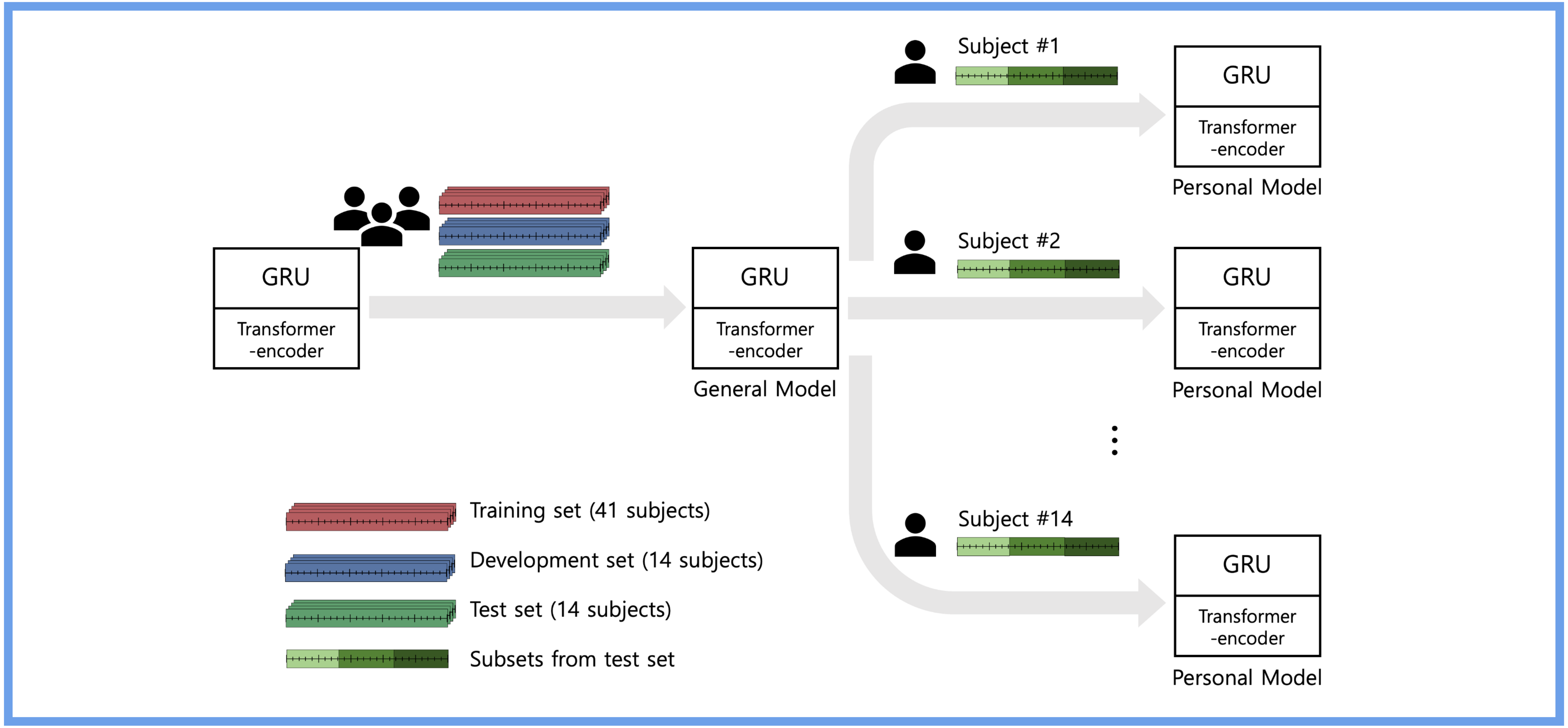
[MuSe-Stress]
Multimodal Emotional Stress Sub-challenge



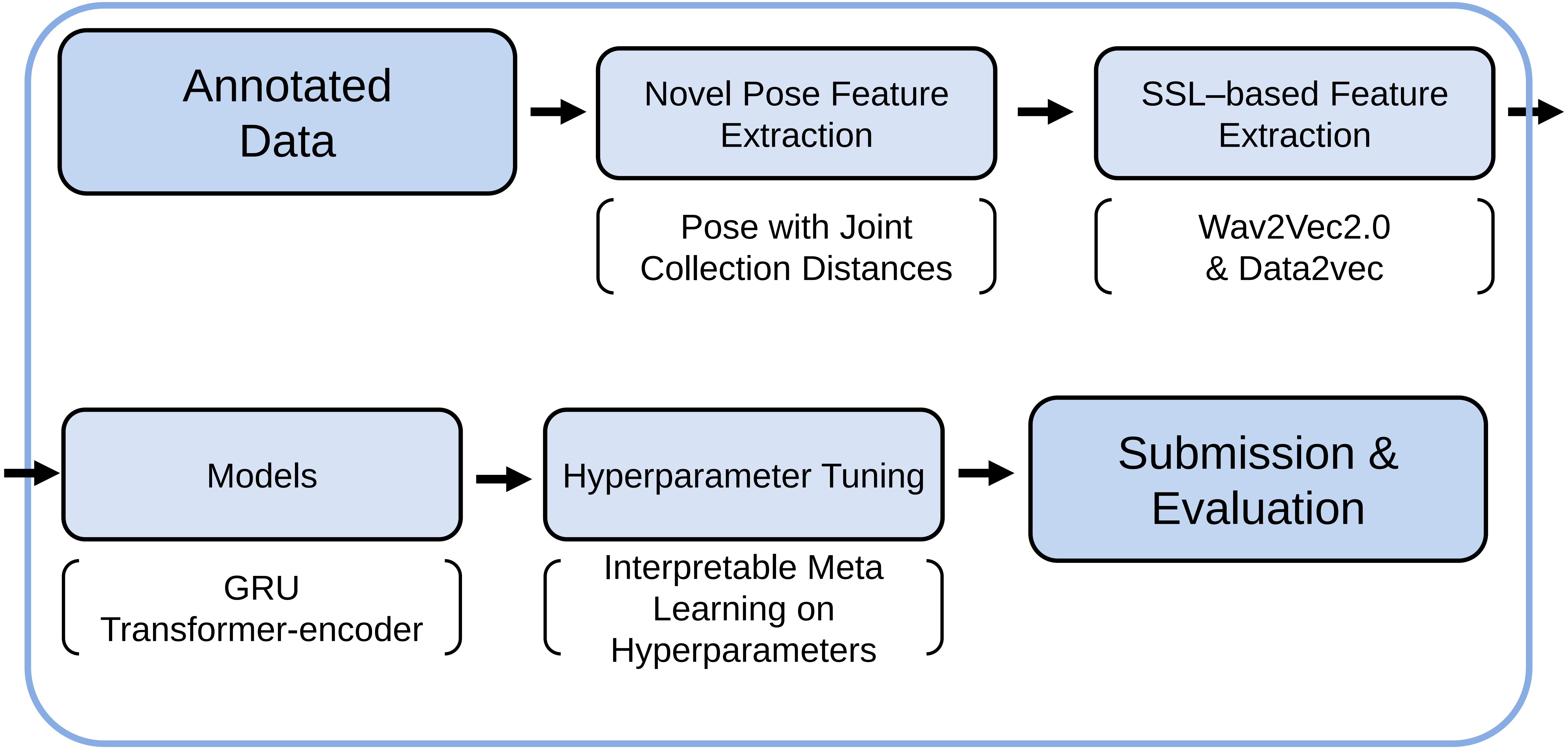
1. INTRODUCTION

01
Introduction

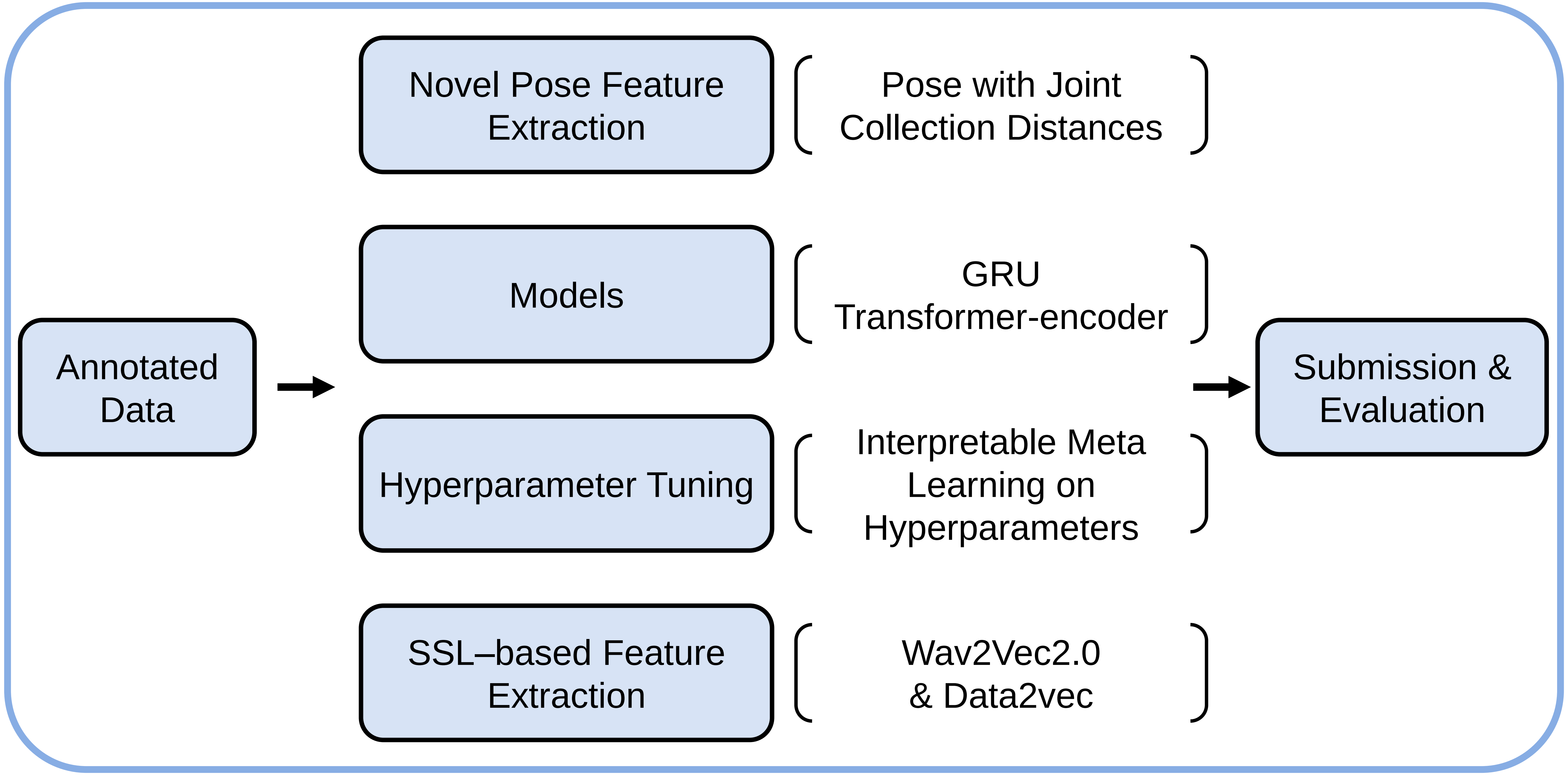
[MuSe-Personalisation]
Multimodal Emotional Stress Sub-challenge



1. INTRODUCTION

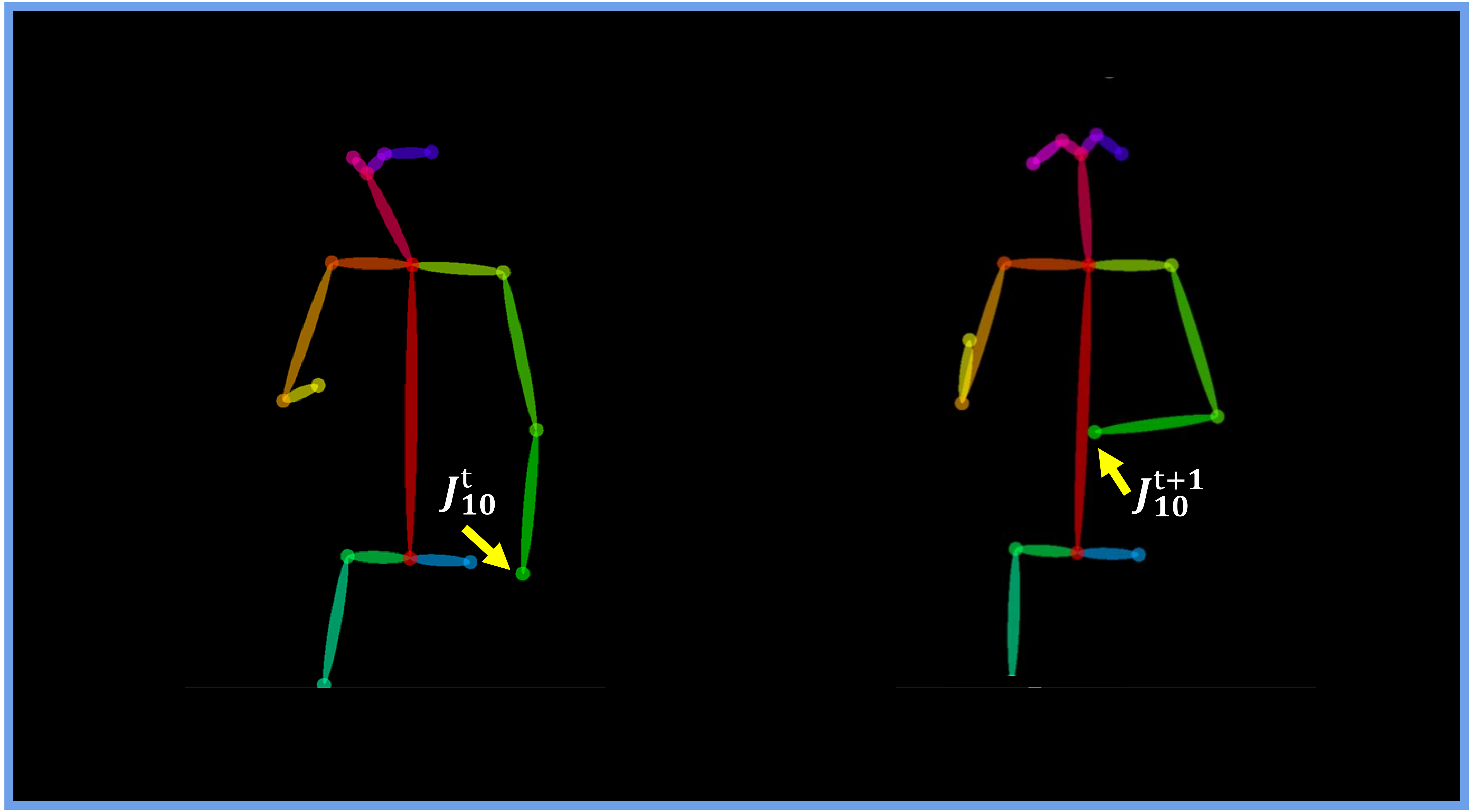


1. INTRODUCTION



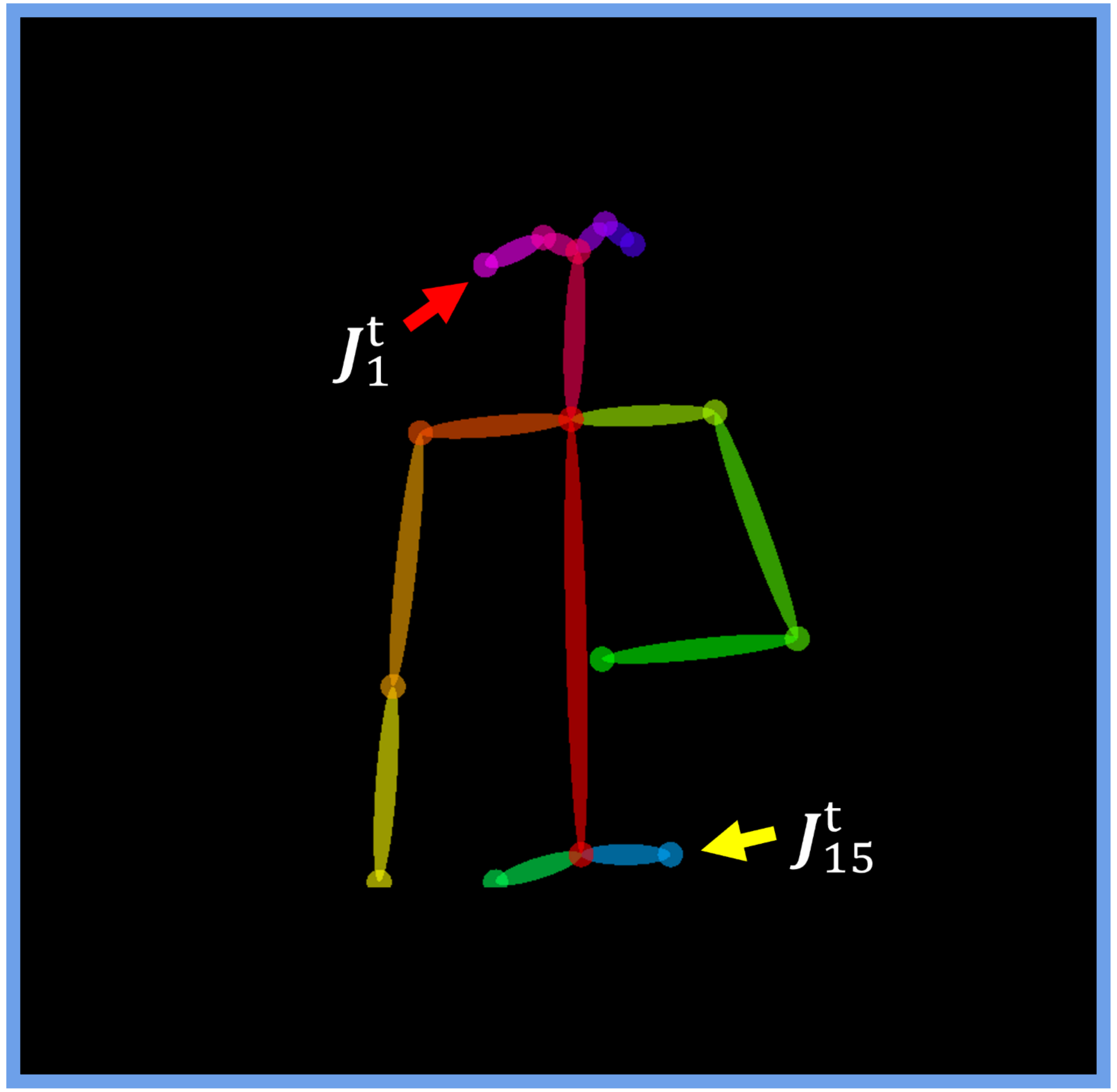
2. METHODS

2. METHODS



Last year's pose feature: simple Euclidean distance over time

2. METHODS



Pose and joint

	J_1^t	J_2^t	$\cdots$	J_{15}^t
J_1^t	0	30	$\cdots$	320
J_2^t	30	0	$\vdots$	230
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
J_{15}^t	320	230	$\cdots$	0

Distance matrix between joints

This year's pose feature: applying joint collection distances

2. METHODS

Sampling rate: 2Hz

Raw Video → OpenPose → JCD

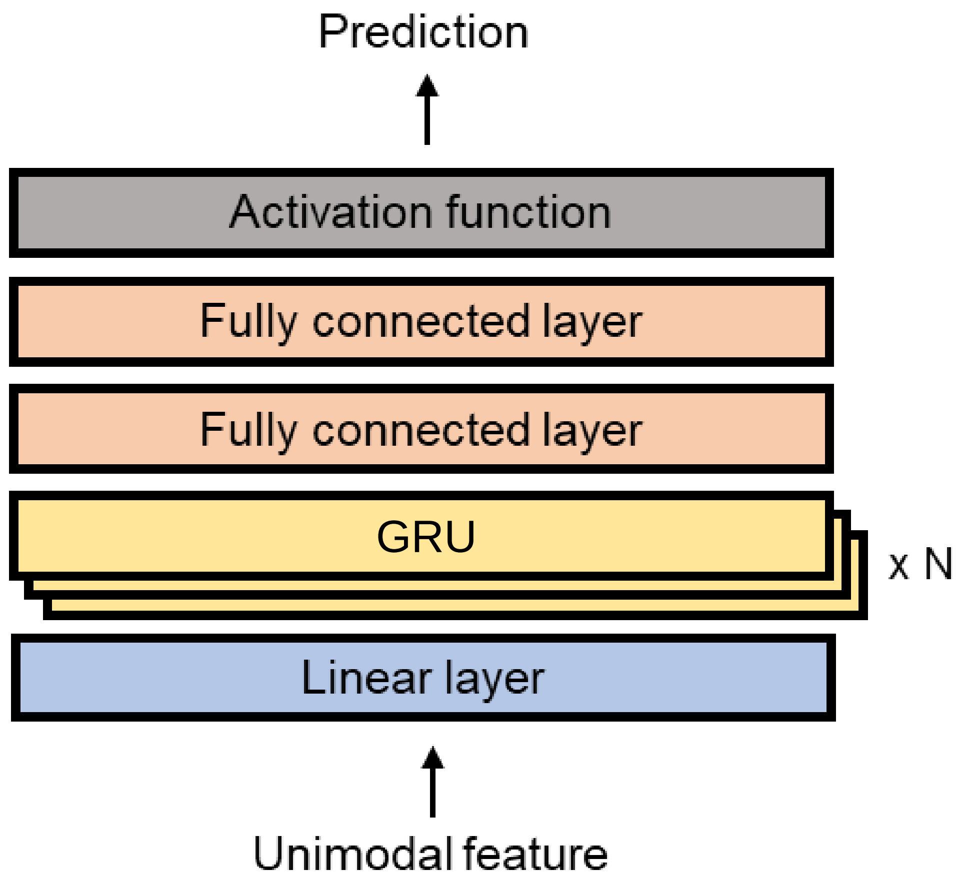
JCD → JCD-feature

JCD → JCD-time

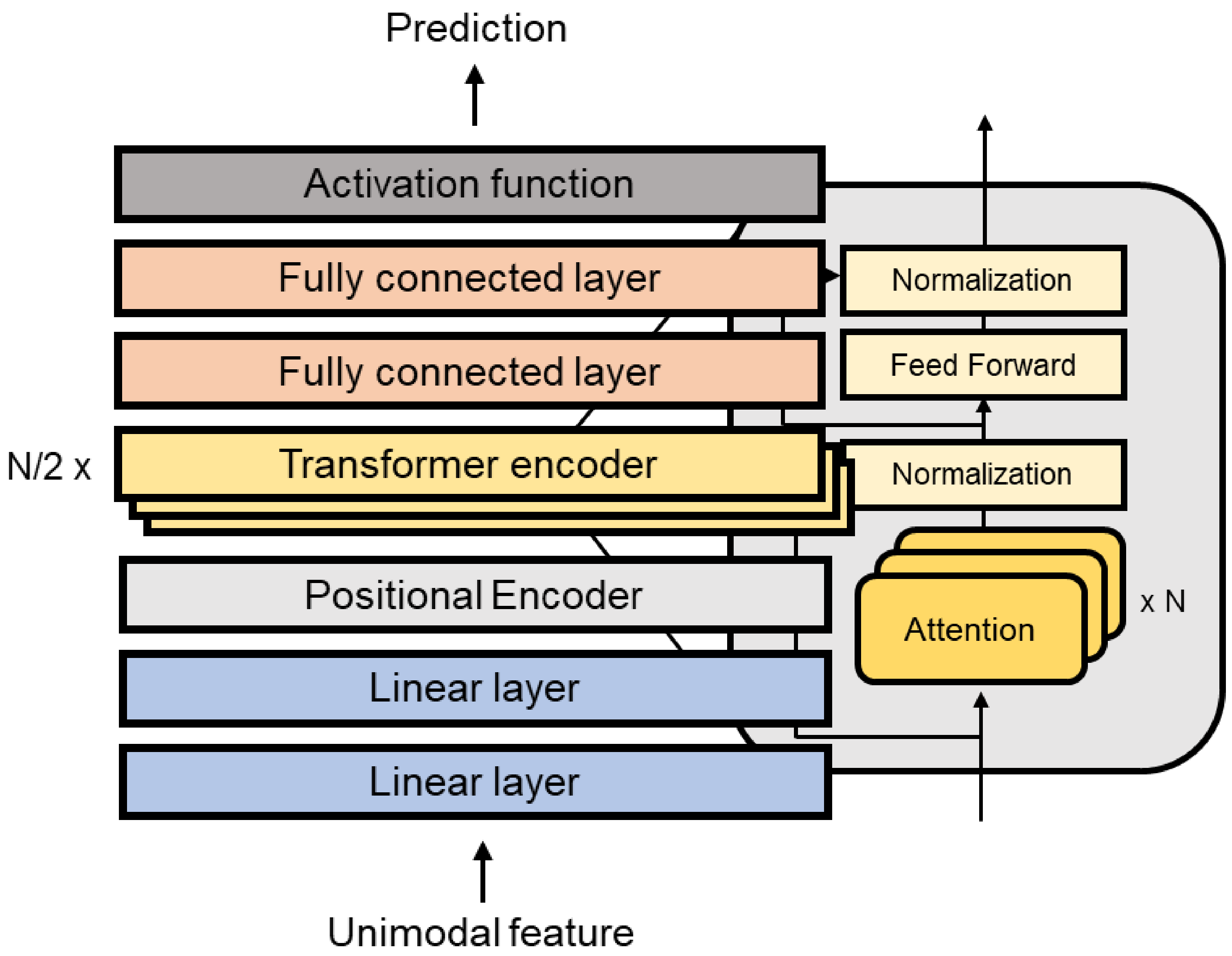
JCD → JCD-both

2. METHODS

GRU

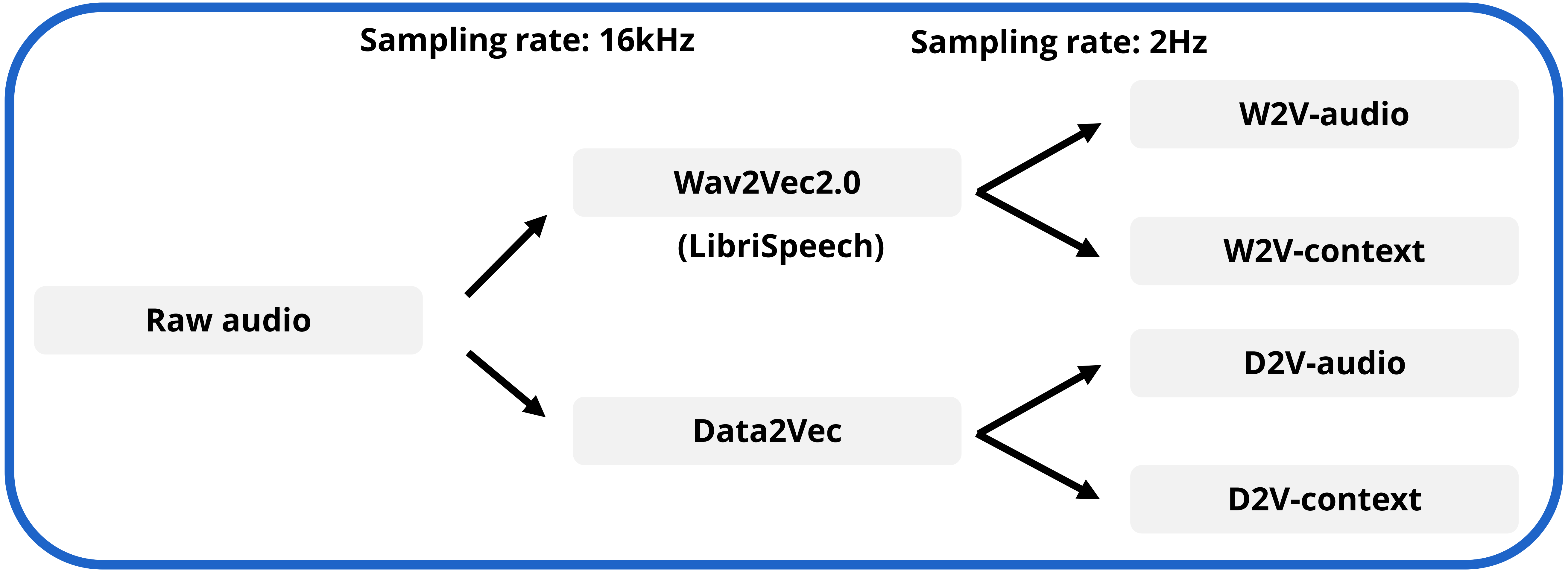


Transformer Encoder



2. METHODS

Baseline



Ours

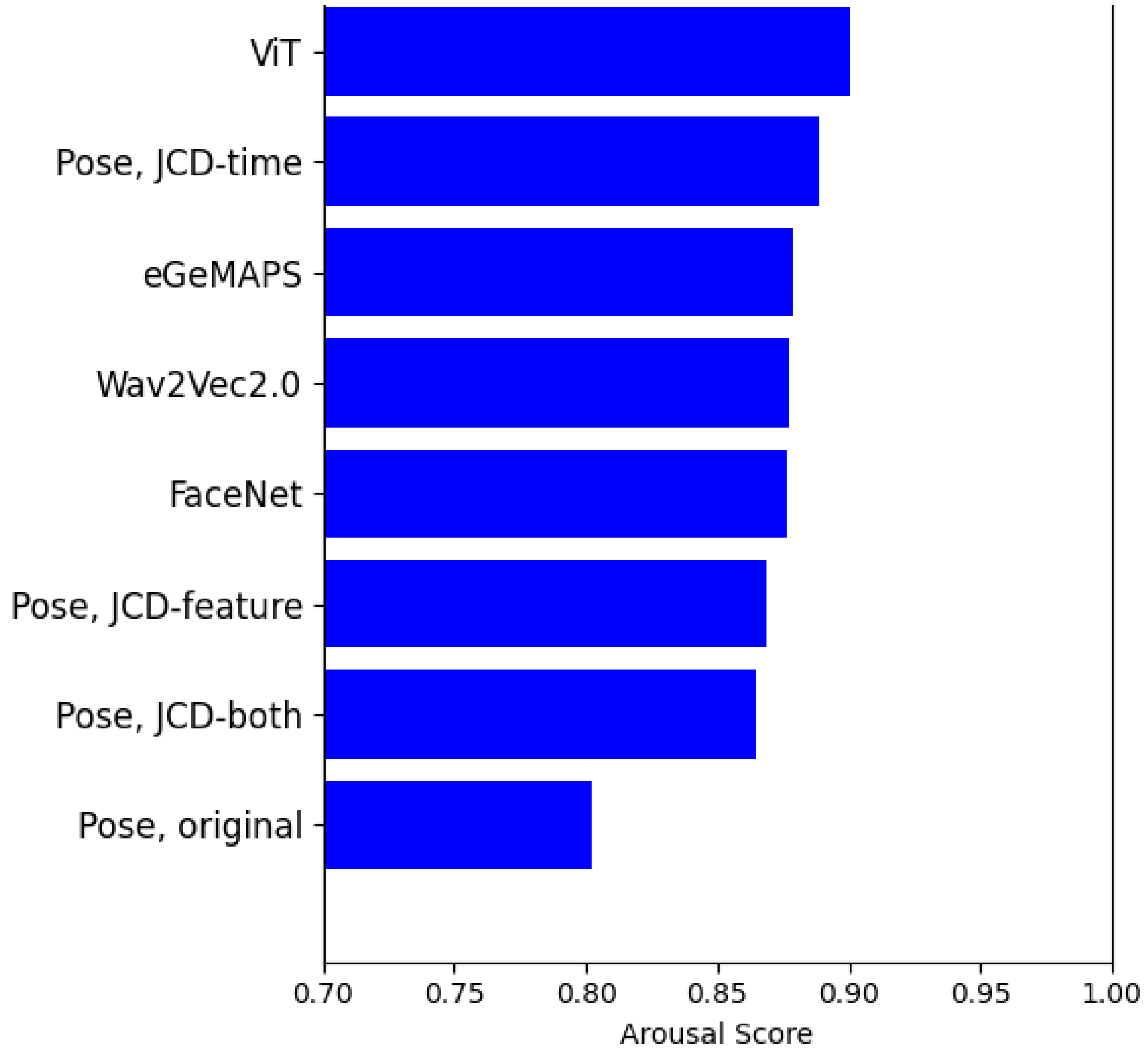
2. METHODS

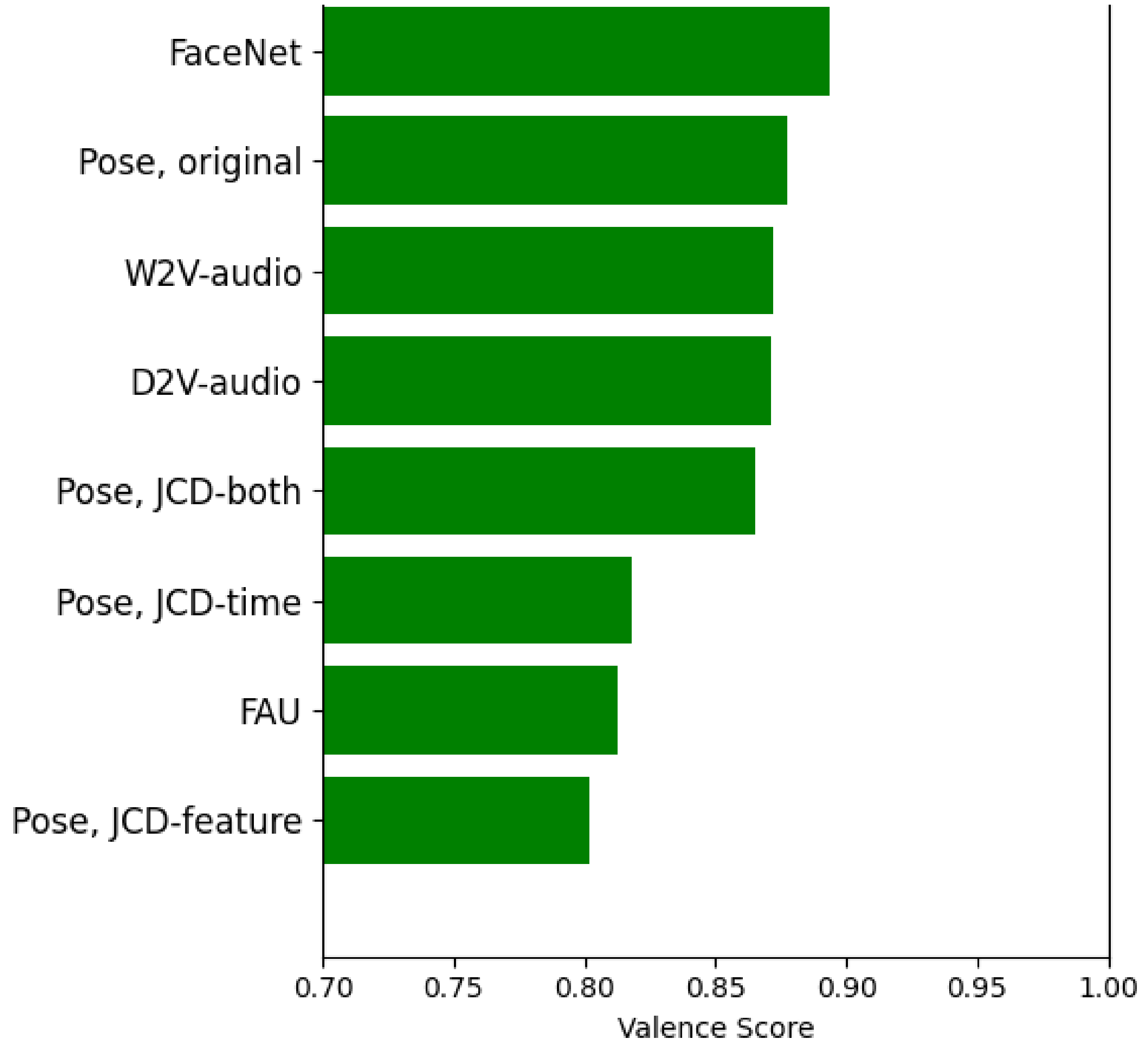
Hyperparameter Setting → Random Forest Regressor → Development CCC Prediction

Model type	Hyperparameter name	Min value	Max value	Number of configurations
General	Window length	200	400	3
	Learning rate	0.0001	0.01	4
	Hop length	50	300	3
	Model complexity	2	128	7
	Number of layers	2	16	4
Personal	Window length	2	60	10
	Learning rate	0.0001	0.05	14
	Hop length	2	25	7

3. RESULTS

3. RESULTS

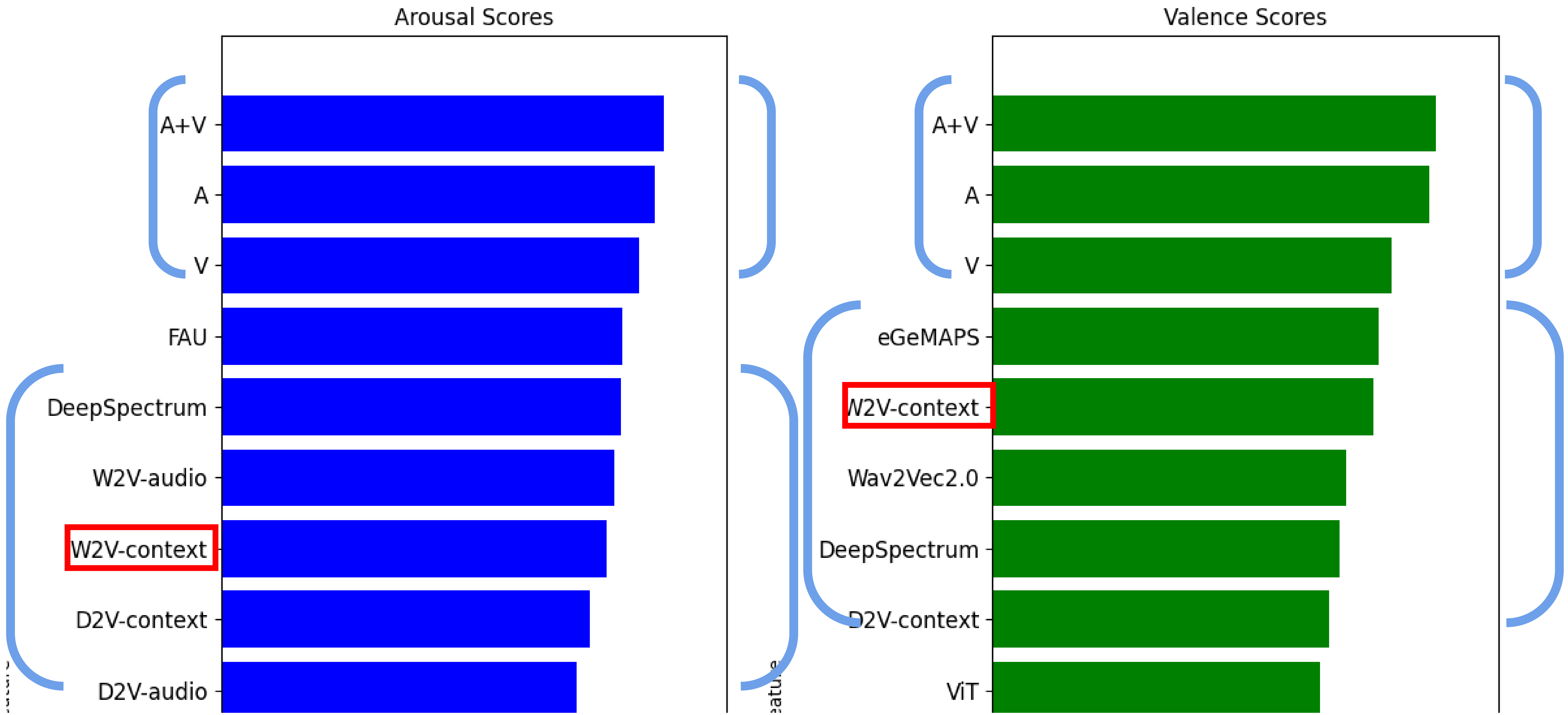




03
Results

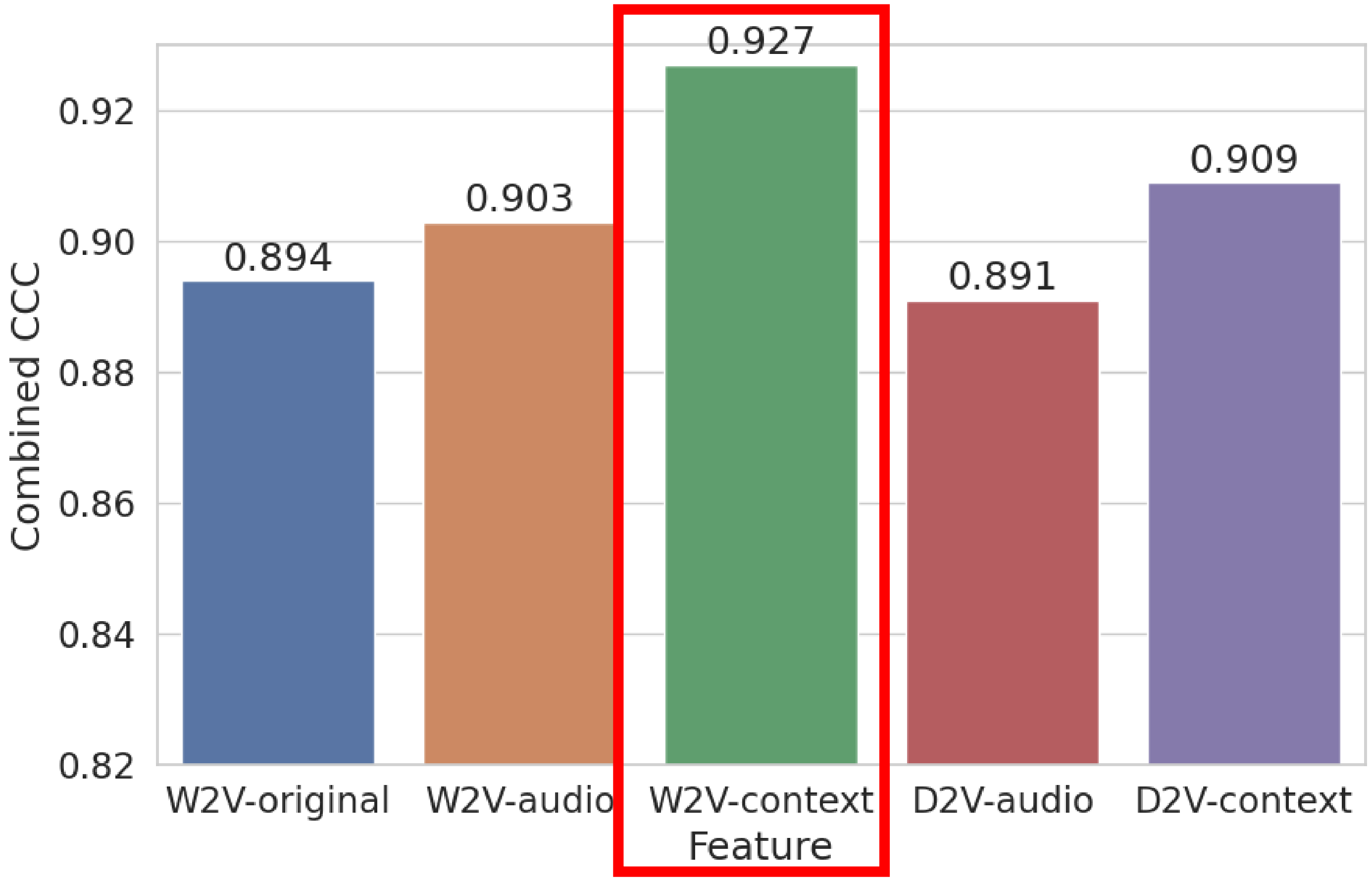
3. RESULTS

Audio features excel in unimodal level
Fusion shows consistent highest scores

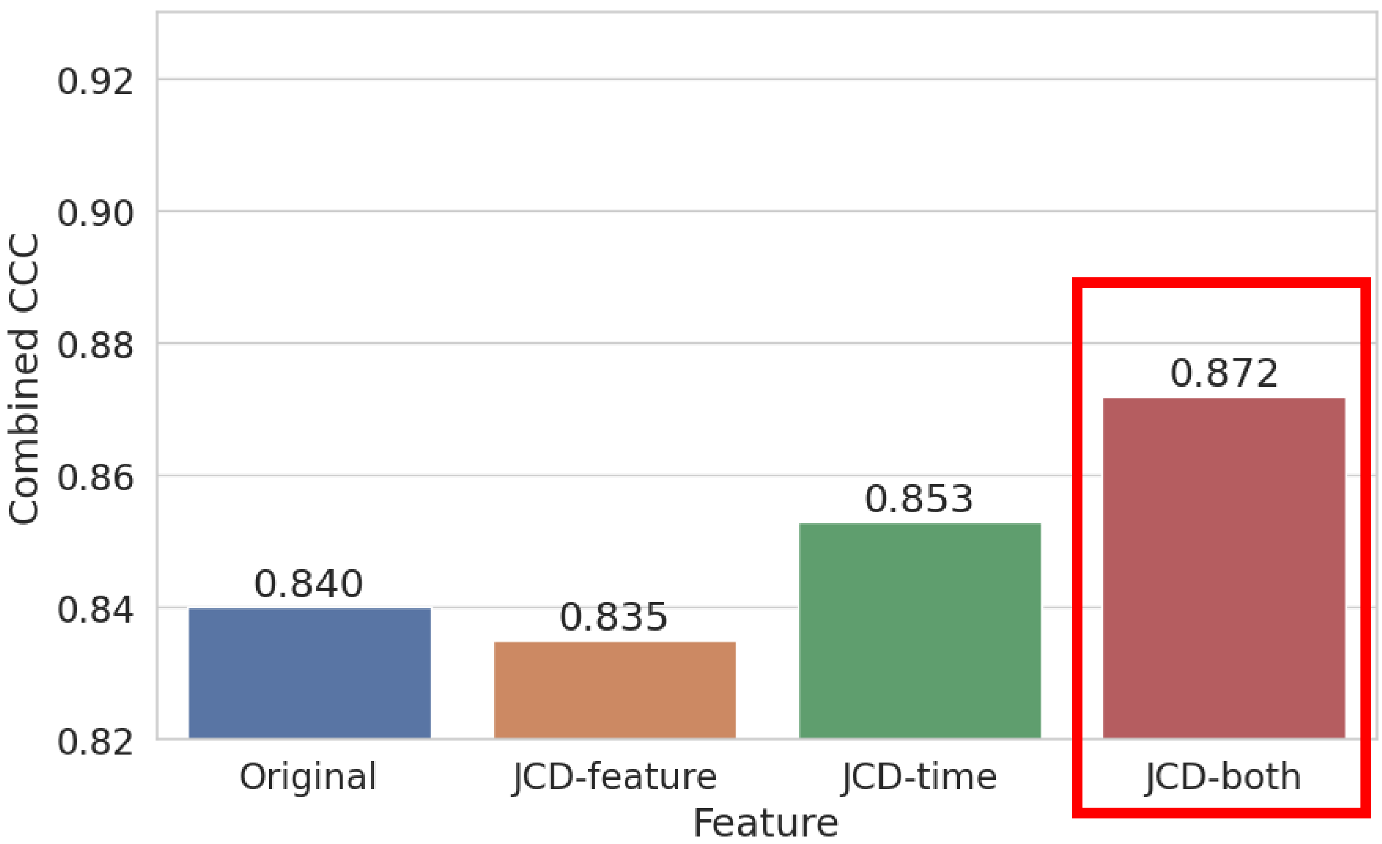


03
Results

3. RESULTS



SSL-based Features

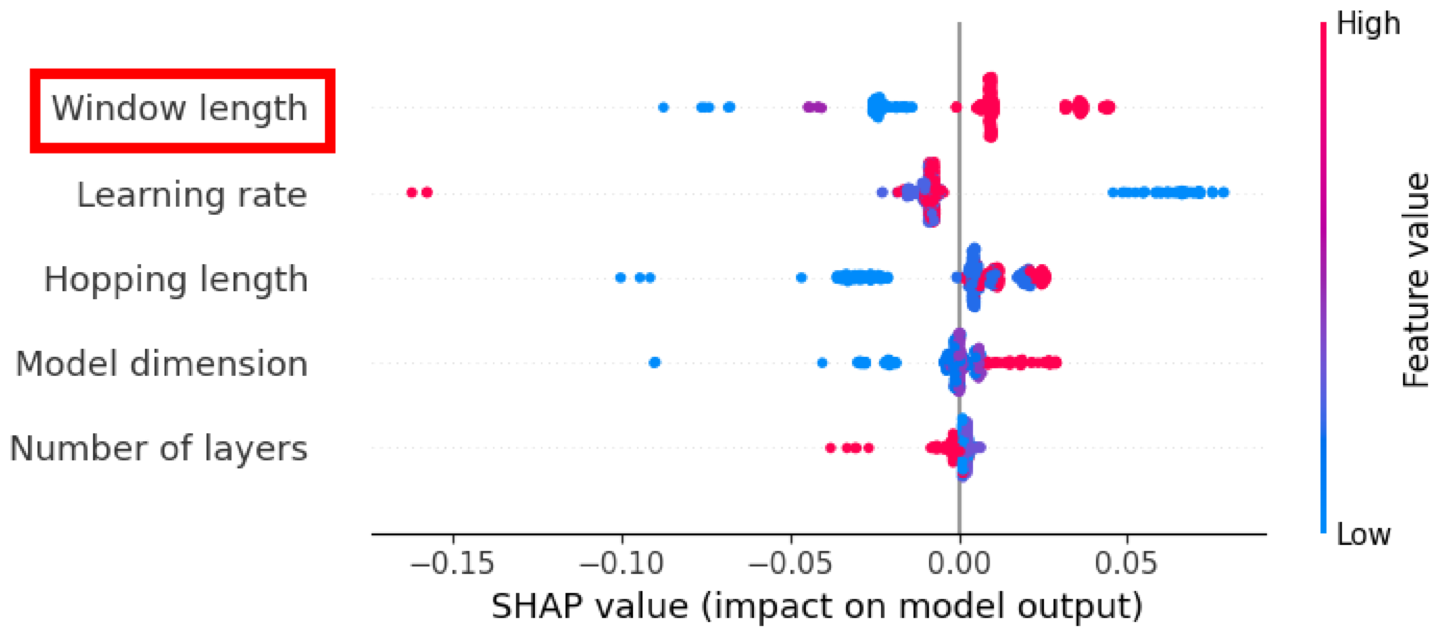


Pose Features

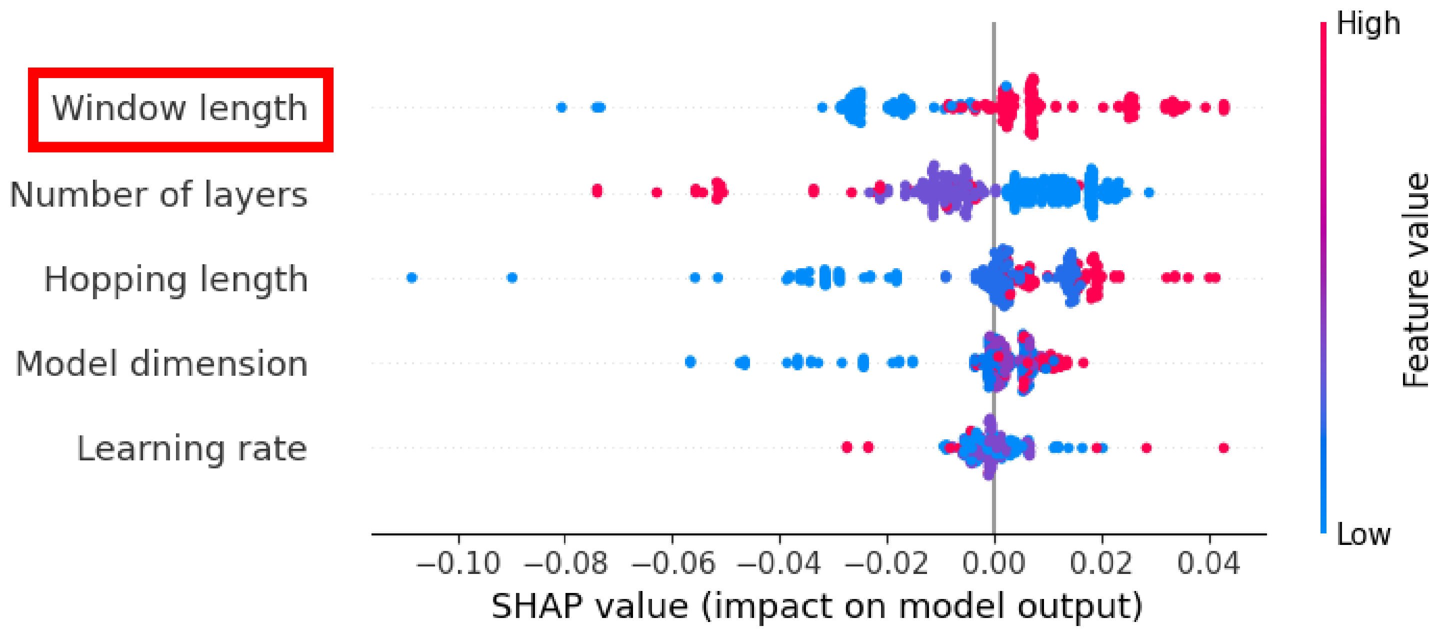
03
Results

3. RESULTS

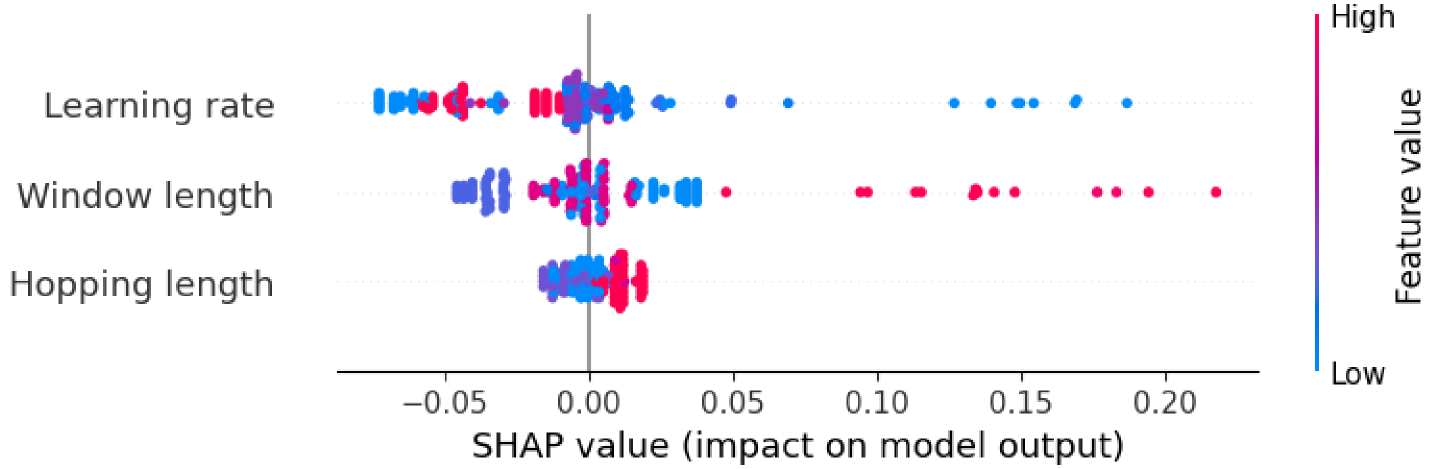
Positive correlation with development CCC



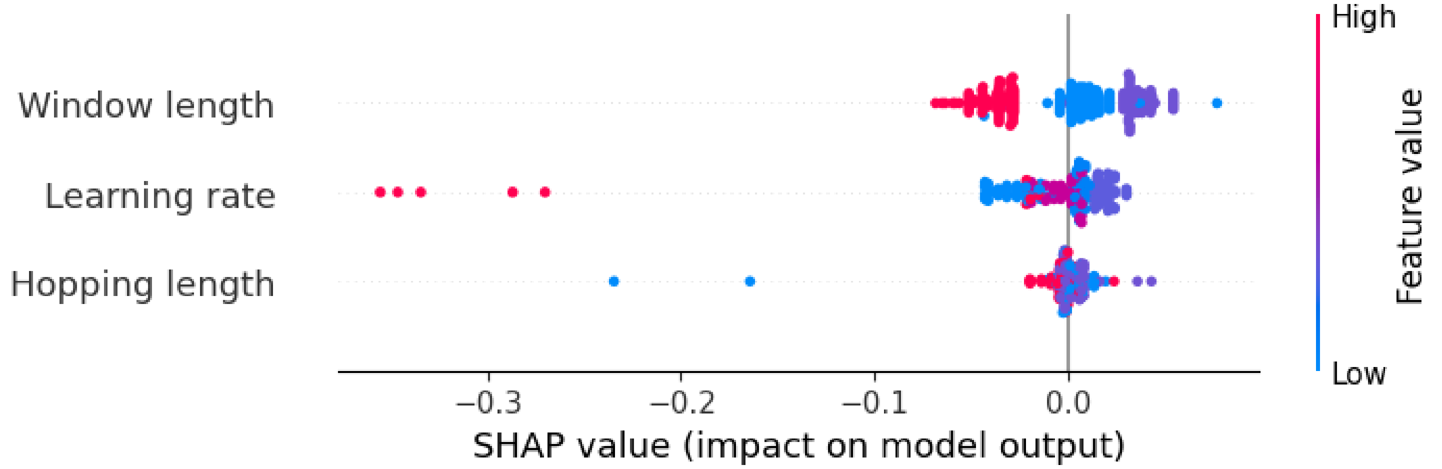
General GRU



General Transformer-encoder



Personal GRU



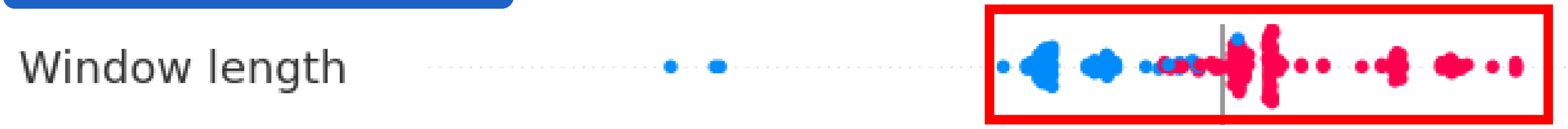
Personal Transformer-encoder

03
Results

3. RESULTS

Reverse trend observed between General model and Personal model

General Transformer-encoder



Personal Transformer-encoder

03

Results

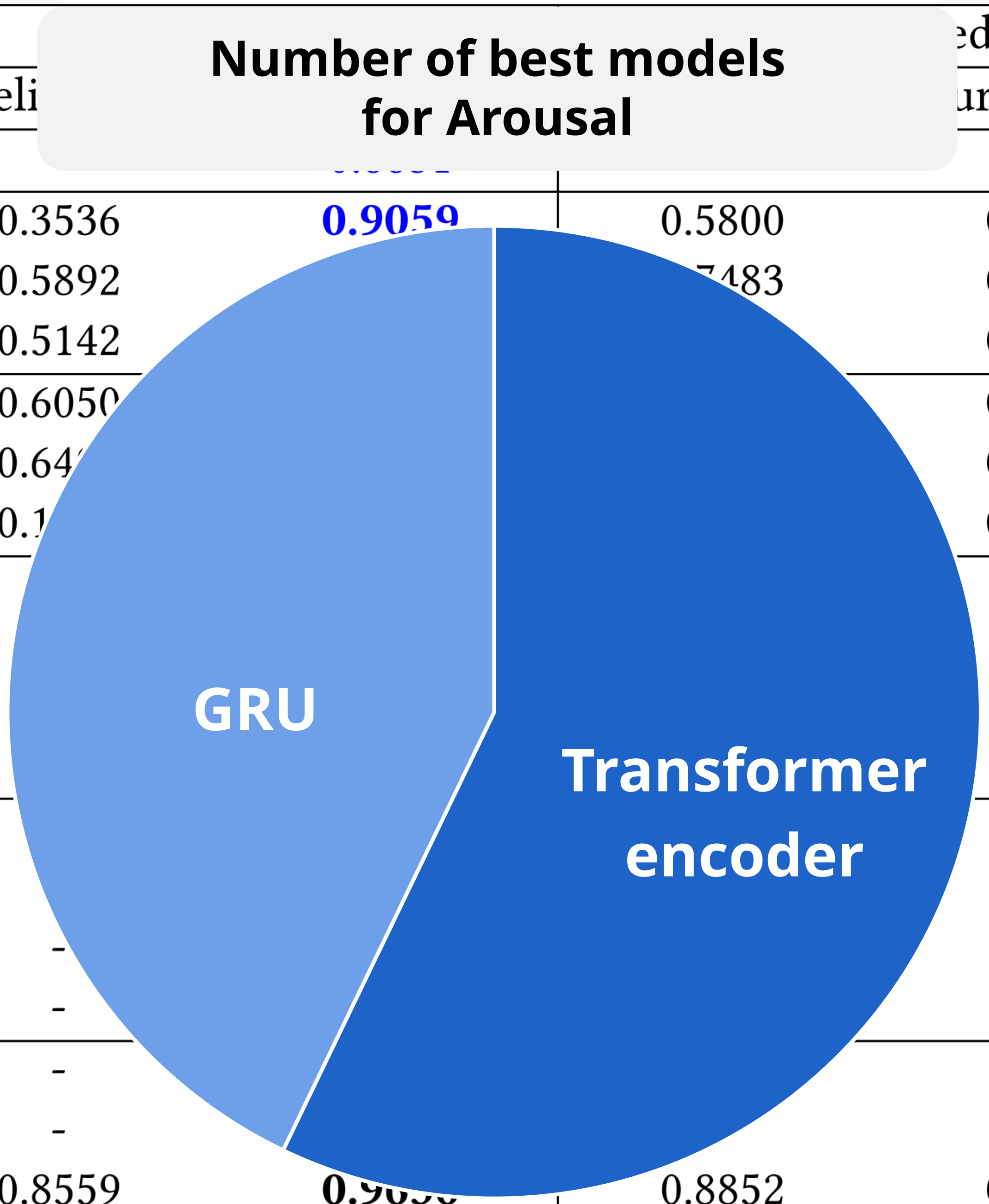
3. RESULTS

Feature info			Arousal		Valence		Combined	
Feature name	Type	Dim	Baseline CCC	Our best CCC	Baseline CCC	Our best CCC	Baseline CCC	Our best CCC
Biosignal	S	3	-	0.8716	-	0.6651	-	0.7684
DeepSpectrum	A	1024	0.8064	**0.9376**	0.3536	**0.9059**	0.5800	**0.9218**
eGeMAPS	A	78	0.9073	0.8783	0.5892	**0.9296**	0.7483	**0.9040**
Wav2Vec2.0	A	1024	0.7421	**0.8775**	0.5142	**0.9096**	0.6282	**0.8936**
ViT	V	384	0.2691	**0.8999**	0.6050	**0.8947**	0.4371	**0.8973**
FaceNet	V	512	0.8260	**0.8766**	0.6491	**0.8936**	0.7376	**0.8851**
FAU	V	20	0.6382	**0.9378**	0.1468	**0.8124**	0.3925	**0.8751**
W2V-audio	A	512	-	0.9336	-	0.8718	-	0.9027
W2V-context	A	768	-	0.9287	-	0.9258	-	0.9273
D2V-audio	A	512	-	0.9110	-	0.8713	-	0.8912
D2V-context	A	768	-	0.9186	-	0.9001	-	0.9093
Pose, original	V	26	-	0.8022	-	0.8775	-	0.8399
Pose, JCD-feature	V	105	-	0.8684	-	0.8017	-	0.8351
Pose, JCD-time	V	105	-	0.8884	-	0.8180	-	0.8532
Pose, JCD-both	V	105	-	0.8649	-	0.8649	-	0.8718
A	Fusion	3	-	0.9577	-	0.9590	-	0.9584
V	Fusion	3	-	0.9478	-	0.9373	-	0.9426
A+V	Fusion	6	0.9145	**0.9625**	0.8559	**0.9636**	0.8852	**0.9631**

03
Results

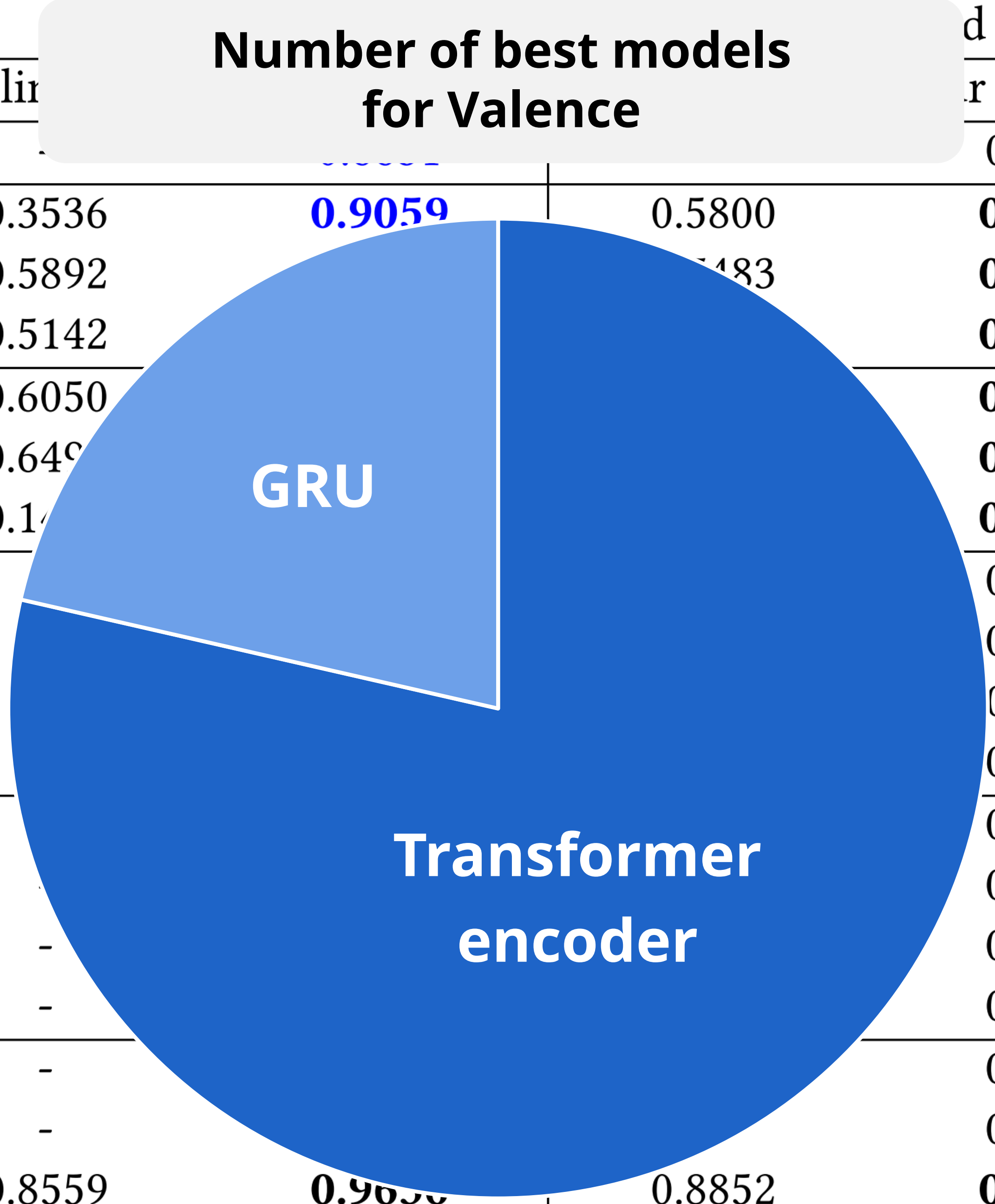
3. RESULTS

Feature info			Arousal		Baseli[illegible]		[illegible]ed	
Feature name	Type	Dim	Baseline CCC	Our best CCC	Baseli[illegible]			[illegible]ur best CCC
Biosignal	S	3	-	0.8716		[illegible]		0.7684
DeepSpectrum	A	1024	0.8064	**0.9376**	0.3536	**0.9059**	0.5800	**0.9218**
eGeMAPS		78	0.9073	0.8783	0.5892		[illegible]	**0.9040**
Wav2Vec2.0		1024	0.7421	**0.8775**	0.5142			**0.8936**
ViT	V	384	0.2691	**0.8999**	[illegible]			**0.8973**
FaceNet		512	0.8260	**0.8766**	[illegible]			**0.8851**
FAU		20	0.6382	**0.9378**	[illegible]			**0.8751**
W2V-audio	A	512	-	0.9336				0.9027
W2V-context		768	-	0.9287				0.9273
D2V-audio		512	-	0.9110				0.8912
D2V-context		768	-	0.9186				0.9093
Pose, original	V	26	-	0.8022				0.8399
Pose, JCD-feature		105	-	0.8684				0.8351
Pose, JCD-time		105	-	0.8884	-			0.8532
Pose, JCD-both		105	-	0.8649	-			0.8718
A	Fusion	3	-	0.9577	-			0.9584
V		3	-	0.9478	-			0.9426
A+V		6	0.9145	**0.9625**	0.8559	[illegible]	0.8852	**0.9631**



3. RESULTS

Feature info			Arousal		[illegible]		[illegible]	
Feature name	Type	Dim	Baseline CCC	Our best CCC	[illegible]	[illegible]	[illegible]	[illegible]
Biosignal	S	3	-	0.8716	[illegible]	[illegible]	[illegible]	0.7684
DeepSpectrum	A	1024	0.8064	**0.9376**	0.3536	**0.9059**	0.5800	**0.9218**
eGeMAPS	A	78	0.9073	0.8783	0.5892	[illegible]	[illegible]	**0.9040**
Wav2Vec2.0	A	1024	0.7421	**0.8775**	0.5142	[illegible]	[illegible]	**0.8936**
ViT	V	384	0.2691	**0.8999**	0.6050	[illegible]	[illegible]	**0.8973**
FaceNet	V	512	0.8260	**0.8766**	[illegible]	[illegible]	[illegible]	**0.8851**
FAU	V	20	0.6382	**0.9378**	[illegible]	[illegible]	[illegible]	**0.8751**
W2V-audio	A	512	-	0.9336	[illegible]	[illegible]	[illegible]	0.9027
W2V-context	A	768	-	0.9287	[illegible]	[illegible]	[illegible]	0.9273
D2V-audio	A	512	-	0.9110	[illegible]	[illegible]	[illegible]	0.8912
D2V-context	A	768	-	0.9186	[illegible]	[illegible]	[illegible]	0.9093
Pose, original	V	26	-	0.8022	[illegible]	[illegible]	[illegible]	0.8399
Pose, JCD-feature	V	105	-	0.8684	[illegible]	[illegible]	[illegible]	0.8351
Pose, JCD-time	V	105	-	0.8884	-	[illegible]	[illegible]	0.8532
Pose, JCD-both	V	105	-	0.8649	-	[illegible]	[illegible]	0.8718
A	Fusion	3	-	0.9577	-	[illegible]	[illegible]	0.9584
V	Fusion	3	-	0.9478	-	[illegible]	[illegible]	0.9426
A+V	Fusion	6	0.9145	**0.9625**	0.8559	[illegible]	0.8852	**0.9631**



3. RESULTS

Features	Arousal [CCC]		Valence [CCC]		Combined [CCC]	
	Base	Ours	Base	Ours	Base	Ours
A+V, A+V	0.7450	**0.8262**	0.7827	0.8844	0.7639	0.8553
A+V normalize, A+V-FAU	-	0.7875	-	**0.8892**	-	0.8384
A+V-FaceNet-eGeMAPS, A	-	0.8046	-	0.8434	-	0.8240
A+V+FAU+DeepSpectrum+W2V-context, A+V-FAU+eGeMAP+W2V-context	-	0.8258	-	0.8847	-	0.8553
A+V, A+V-FAU	-	**0.8262**	-	**0.8892**	-	**0.8577**

2nd place in the competition

03
Results

4. DISCUSSION & CONCLUSIONS

4. DISCUSSION

Transformer-encoder

- **Transformer-encoder** architecture **excels in personalization** tasks, particularly in **Valence** predictions

- Except for the FAU feature, the model achieved the highest development CCC scores across all features

- Ability to capture **long range dependencies** using attention mechanism led to success

  ✓ **Window length** was important

04
Conclusions

4. DISCUSSION

Hyperparameter Tuning

- managed to surpass the baseline development CCC in all unimodal predictions, except for the Arousal-eGeMAPS

  ✓ Meta-learning discovered that **learning rate and window length are crucial factors**

  ✓ Increase in window length negatively impacted development CCC of personalized Transformer-encoder model

04
Conclusions

4. DISCUSSION

Newly Crafted Features

- Pose features extracted through JCD and the different SSL-based features (Wav2Vec2.0 and Data2Vec), showed considerable promise in improving emotional dimension predictions

  ✓ **JCD based features** demonstrated a notable **enhancement over** the **original Pose feature**

  ✓ SSL-based features, particularly context-based ones, consistently scored higher CCC scores compared to their audio counterparts

04
Conclusions

4. CONCLUSIONS

In summary

- Given the MuSe-Stress 2023 baseline, we investigated three different approaches (New Features, Transformer-encoder, Hyperparameter Tuning), reaching the **2nd place in the competition**

  - ➢ still difficult to answer **why** and **how** the different approaches affect the CCC values

Future work

- Investigate the generalizability of our newly engineered pose features by testing them across different use cases that involve stress detection (e.g., driver behavior monitoring)

THANK YOU. QUESTIONS?