A Multimodal Self-supervised AI Framework for Monitoring Challenge Behavior Risks in Children with ASD

Zhenhao Zhao[1], Eunsun Chung[2], Kyoug-Mee Chung[2], Michelle Crawford[3], Chung Hyuk Park[1]

[1]Department of Biomedical Engineer, The George Washington University Washington, DC, USA
[2]Department of Psychology, Yonsei University, Seoul, Korea
[3]Neurobehavioral Unit, Kennedy Krieger Institute, Baltimore, MD, USA

Background

Autism Spectrum Disorder (ASD)
- A life-long neuro-developmental condition
- Experience difficulties in communication and social interaction
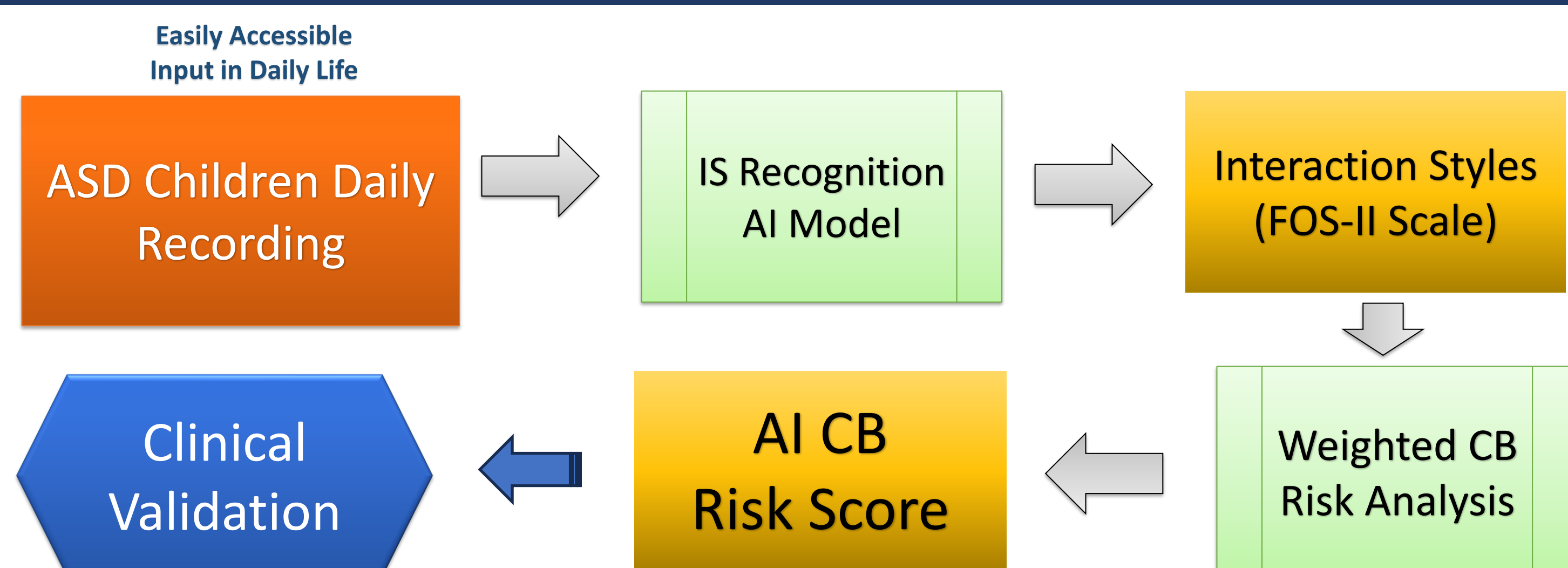- 1 in 36 children is diagnosed with ASD.

Family Observation Schedule 3rd Edition (FOS-R-III): A structured interaction styles (IS) scale in clinical settings.
- Assesses the severity of autism.
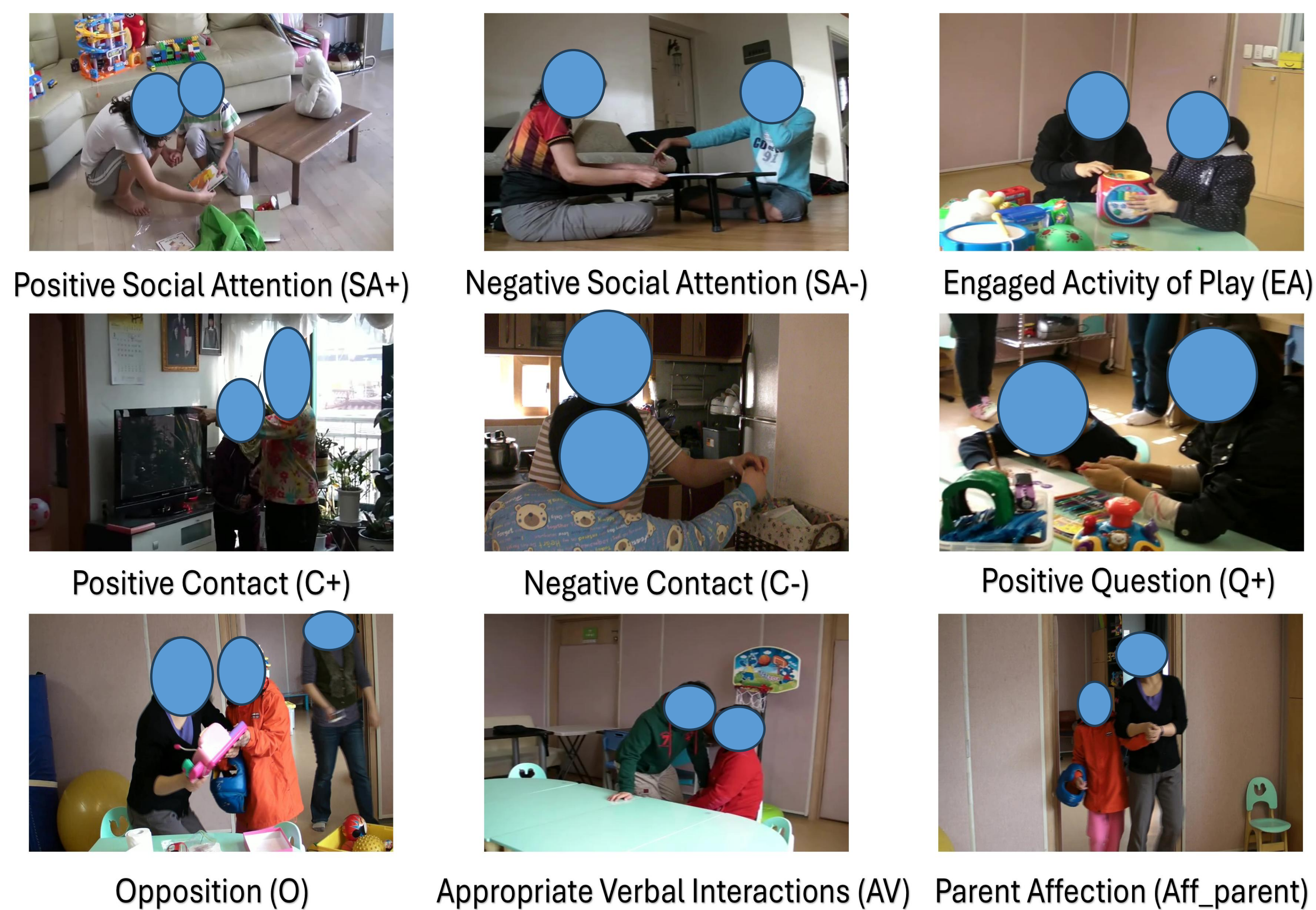- Identifies challenging behaviors.

Challenging behaviors (CBs) include **self-injury, aggression, and disruptive conduct**
- Severely affecting children's health and social interactions.
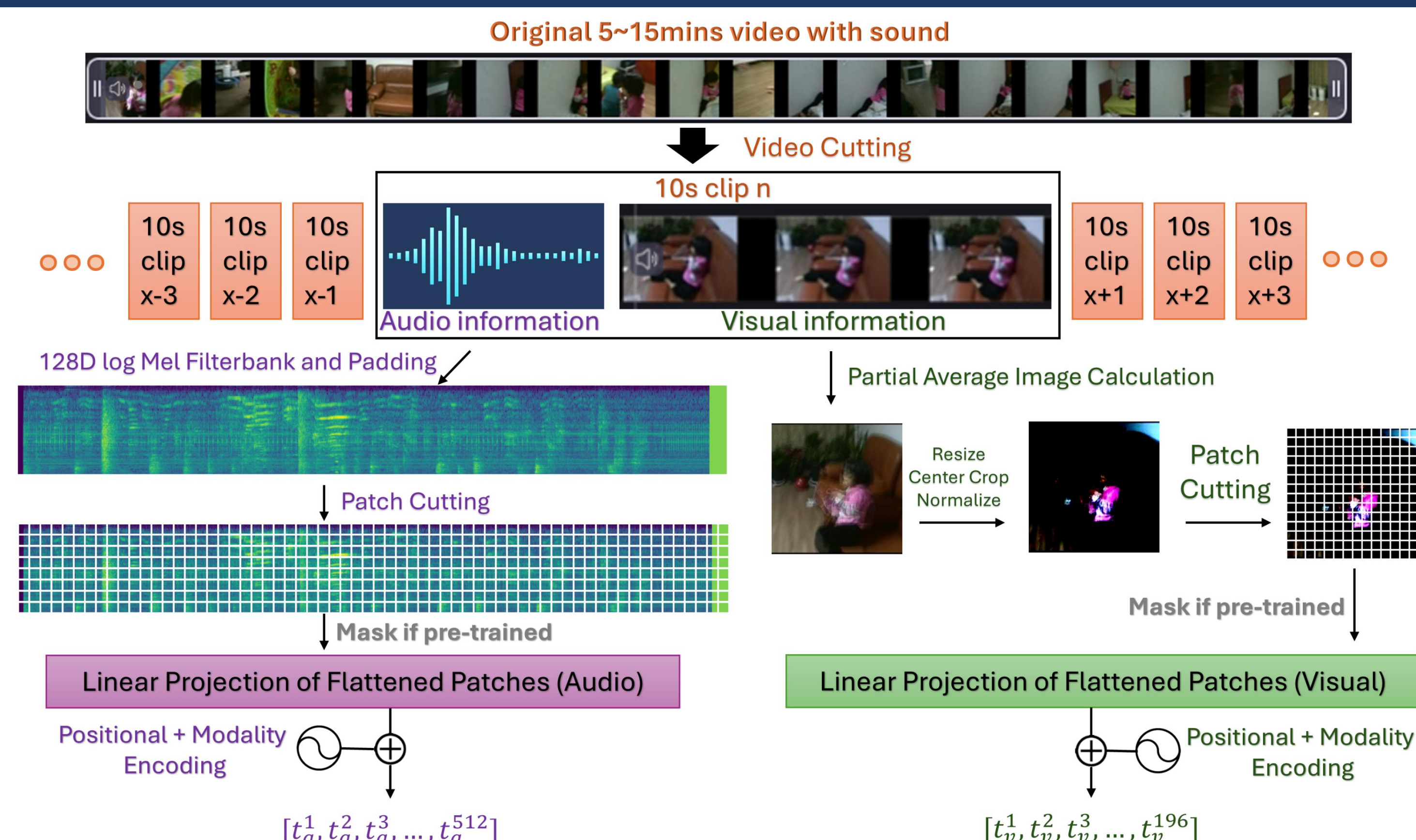- Traditional CBs risk evaluation **relies on clinicians**, making it **time-consuming and labor-intensive**.

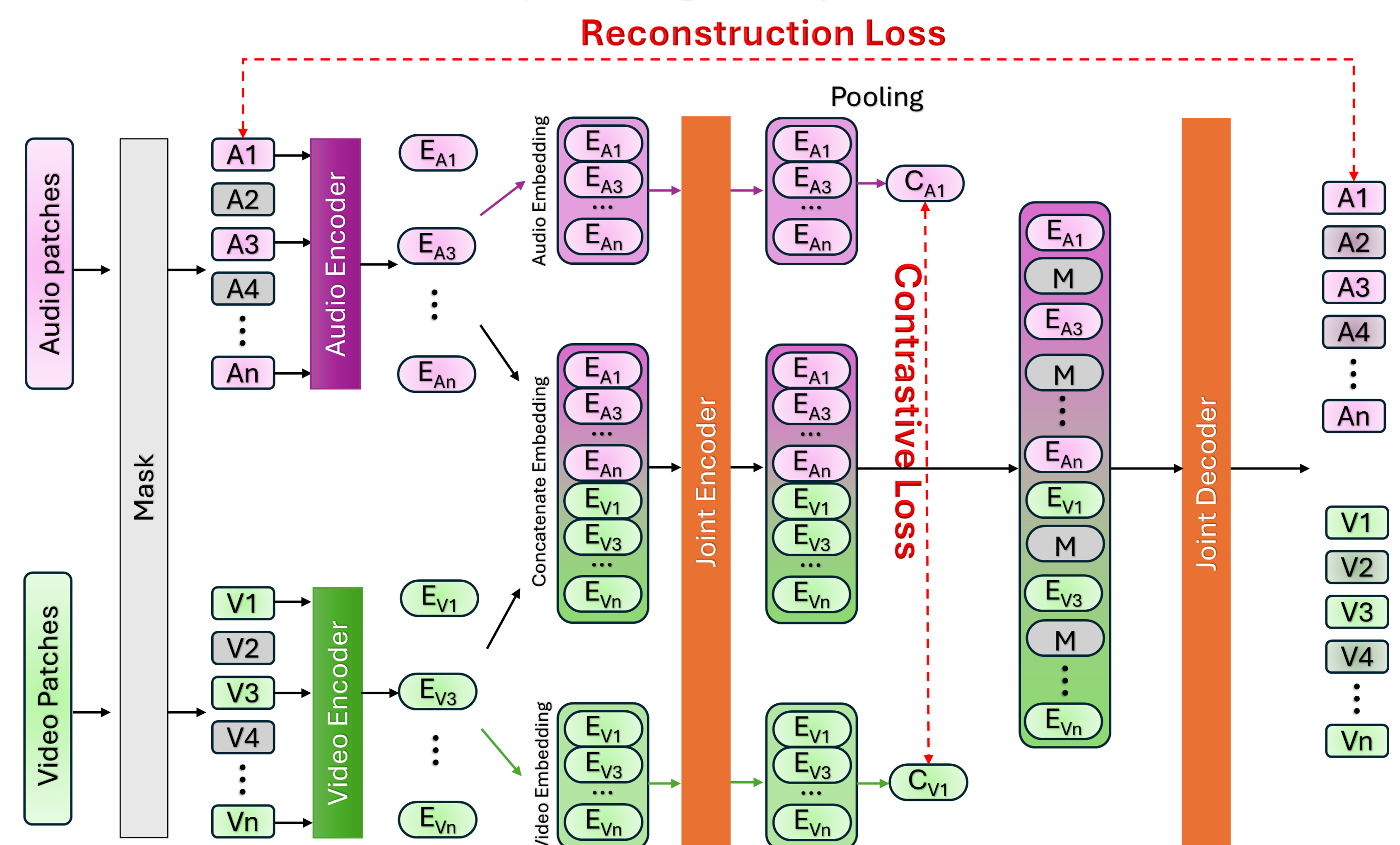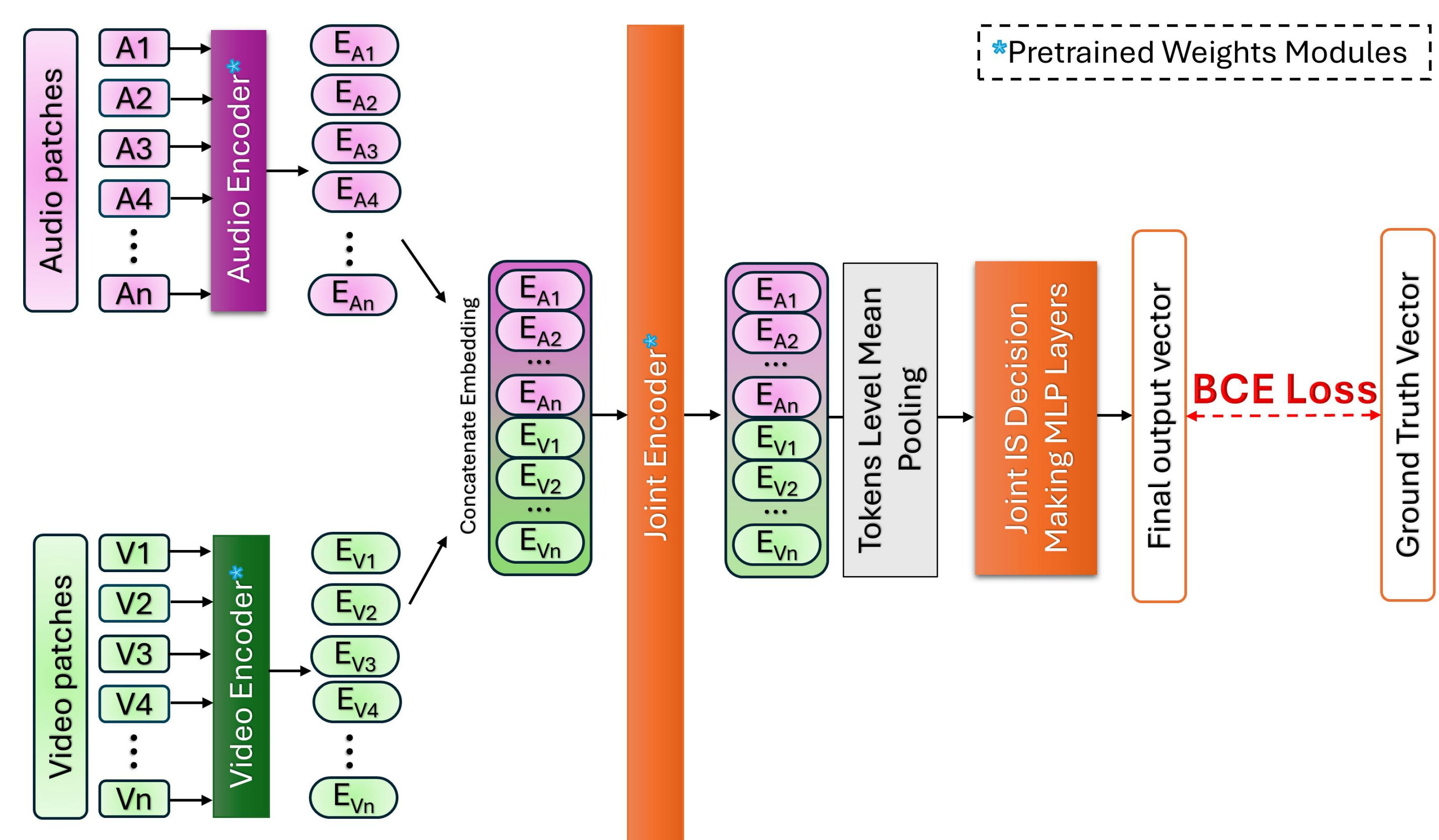## Study Overview

Interaction Style (IS)

Positive Social Attention (SA+)   Negative Social Attention (SA-)   Engaged Activity of Play (EA)

Positive Contact (C+)   Negative Contact (C-)   Positive Question (Q+)

Opposition (O)   Appropriate Verbal Interactions (AV)   Parent Affection (Aff_parent)

Tokenization

Original 5~15mins video with sound

Video Cutting

10s clip n

Audio information   Visual information

128D log Mel Filterbank and Padding   Partial Average Image Calculation

Patch Cutting   Resize Center Crop Normalize   Patch Cutting

Mask if pre-trained   Mask if pre-trained

Linear Projection of Flattened Patches (Audio)   Linear Projection of Flattened Patches (Visual)

Positional + Modality Encoding   Positional + Modality Encoding

$[t_a^1, t_a^2, t_a^3, …, t_a^{512}]$ $[t_v^1, t_v^2, t_v^3, …, t_v^{196}]$

Self-supervised IS Recognition Model (AV-FOS)

Pretrained Stage: Using nearly endless internet data.

Reconstruction Loss

Pooling

Contrastive Loss

Tuning Stage: Learning from clinical annotations

*Pretrained Weights Modules

BCE Loss

Results: Model Performance

TABLE I
AV-FOS Performance on the FOS Dataset.

Model	mAP ↑	Accuracy ↑	Strict Accuracy ↑	AUC ↑	F1 Score ↑	Time Cost →
GPT4V + Prompt V1	0.3181	0.7965	0.1355	0.6624	0.4581	4.3349
GPT4V + Prompt V2	0.2481	0.7668	0.1468	0.5896	0.3330	3.9792
SlowFast	0.6138	0.8287	0.1125	0.8445	0.5437	0.0031
ViT	0.6167	0.8172	0.0889	0.8486	0.5448	0.0011
AV-FOS Model	**0.6879**	**0.8590**	**0.2003**	**0.8868**	**0.5936**	**0.0027**

Performance Comparison by Metrics   Time Cost

AV-FOS   GPT4V+ Prompt V1   GPT4V+ Prompt V2   SlowFast   ViT