MACHINE LEARNING DEPARTMENT

Language Technologies Institute

Carnegie Mellon University

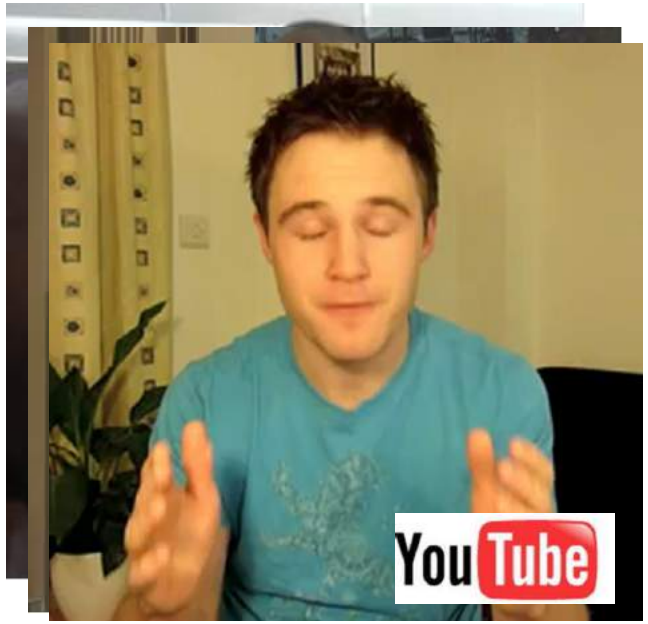
Modeling Spatiotemporal Multimodal Language with Recurrent Multistage Fusion

Presenter: Paul Pu Liang

Paul Pu Liang, Ziyin Liu, Amir Zadeh, Louis-Philippe Morency

Progress of Artificial Intelligence

Multimedia Content



Intelligent
Personal Assistants

Robots and
Virtual Agents

Multimodal Language Modalities

Language

- Lexicon
- Syntax
- Pragmatics

Visual

- Gestures
- Body language
- Eye contact
- Facial expressions

Acoustic

- Prosody
- Vocal expressions

Multimodal Language Modalities

Language

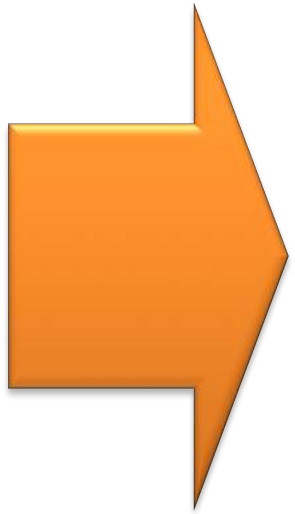
- Lexicon
- Syntax
- Pragmatics

Visual

- Gestures
- Body language
- Eye contact
- Facial expressions

Acoustic

- Prosody
- Vocal expressions



Sentiment

- Positive
- Negative

Emotion

- Anger
- Disgust
- Fear
- Happiness
- Sadness
- Surprise

Personality

- Confidence
- Persuasion
- Passion

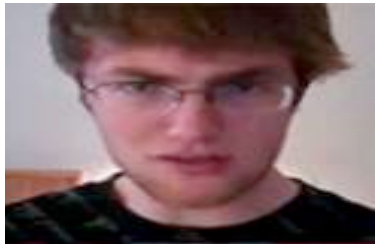
Challenge 1: Temporal Intra-modal Interactions

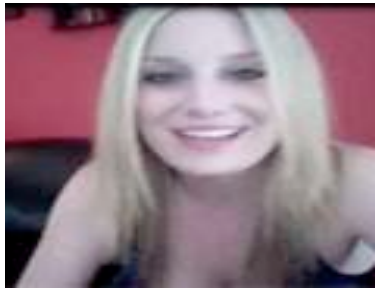
a) Temporal sequences

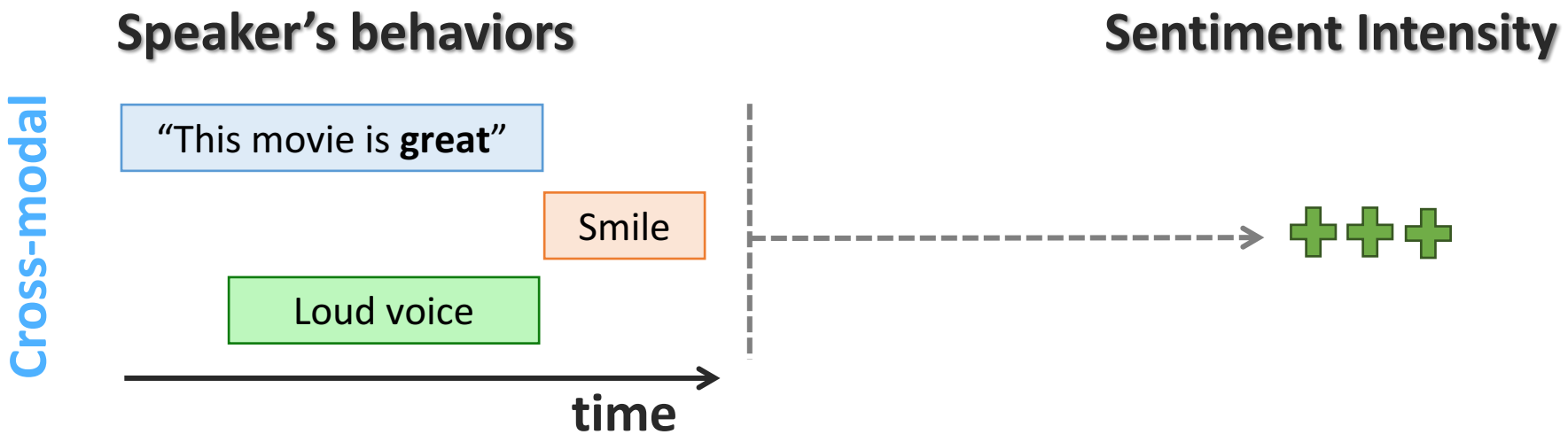
Challenge 2: Spatial Cross-modal Interactions

a) Multiple co-occurring interactions

b) Different weighted combinations

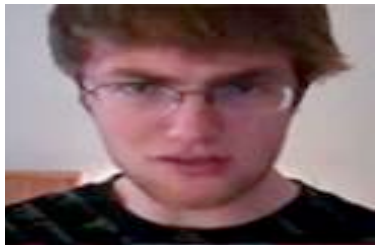


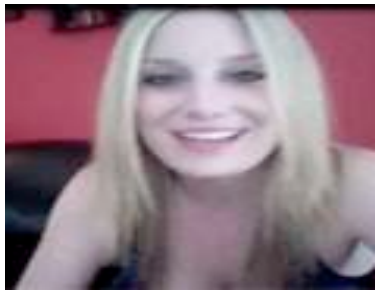




Multistage Aggregation in Humans

(Parsini et al. 2015,
Taylor et al. 2017)

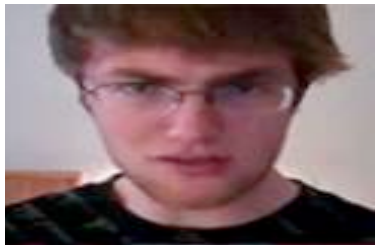


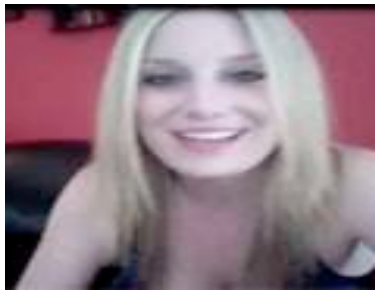


wide smile
loud voice

Multistage Aggregation in Humans

(Parsini et al. 2015,
Taylor et al. 2017)



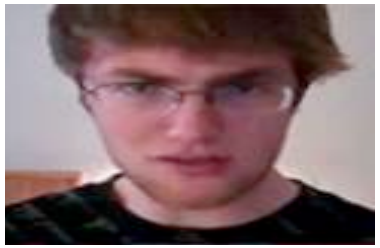


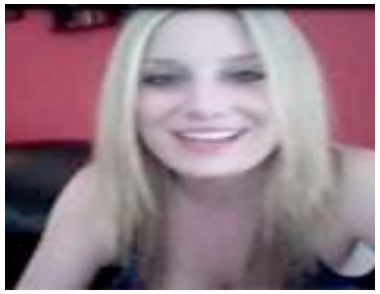
wide smile
loud voice

positive reaction
positive words

Multistage Aggregation in Humans

(Parsini et al. 2015,
Taylor et al. 2017)



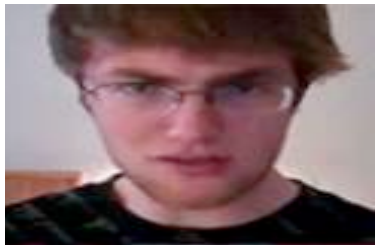


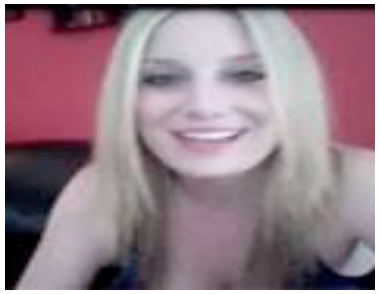
wide smile
loud voice

positive reaction
positive words

excitement
joyous

Computational Model for Multistage Fusion





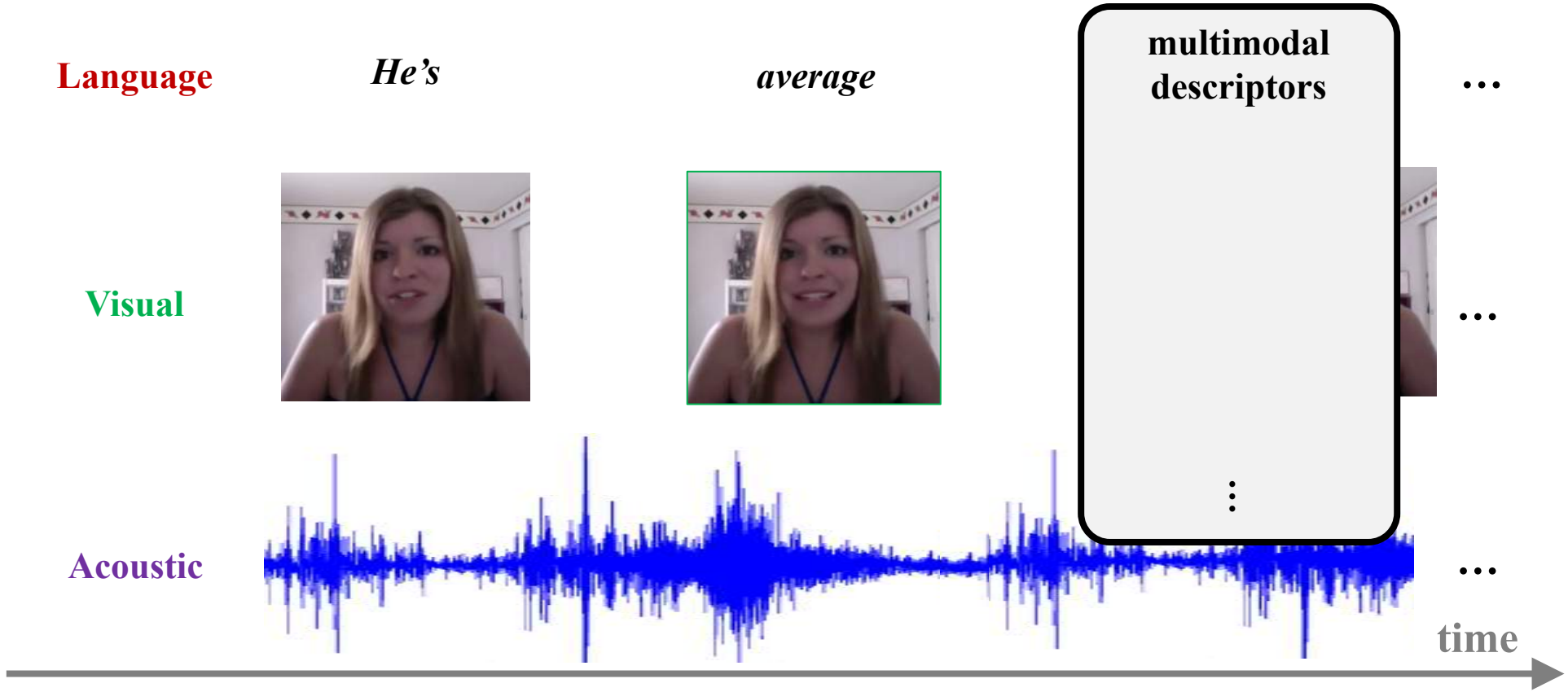
wide smile
loud voice

positive reaction
positive words

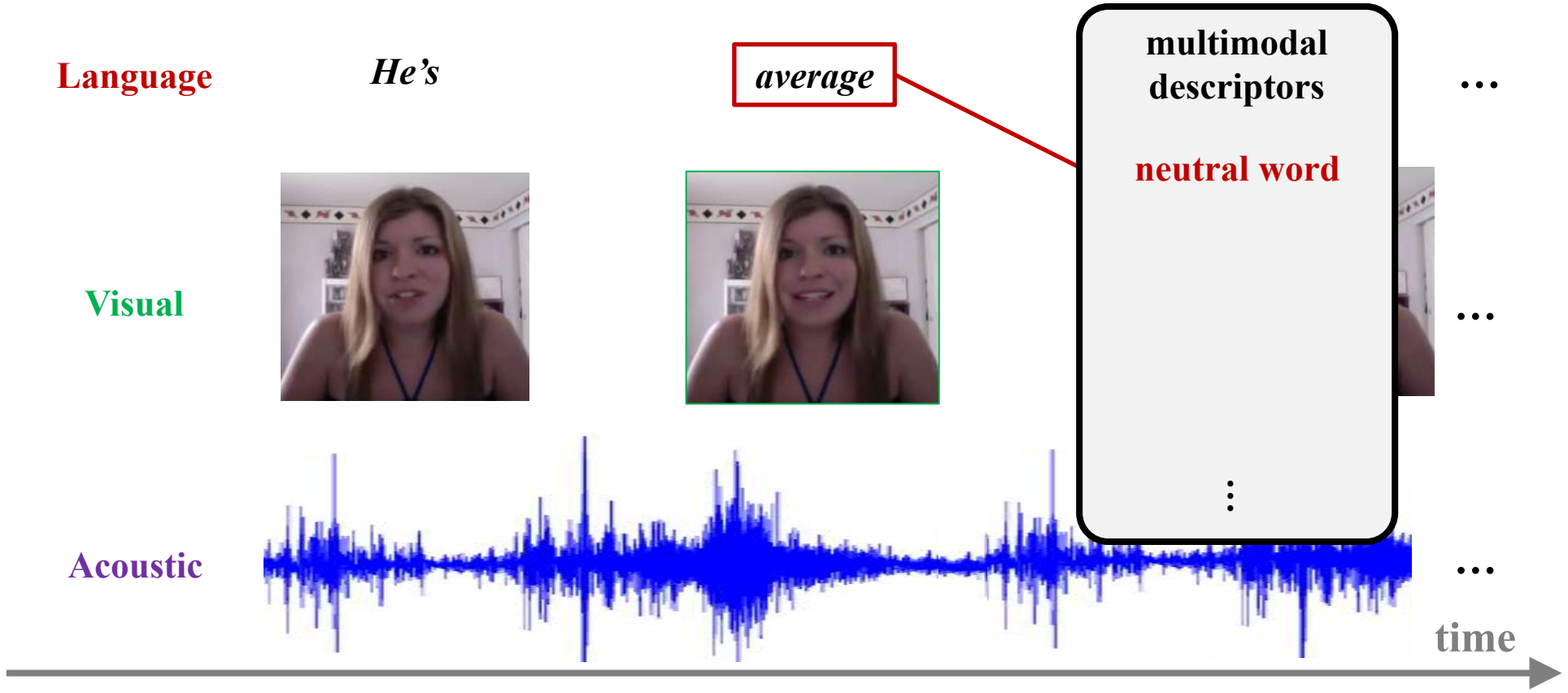
excitement
joyous

Computational Model

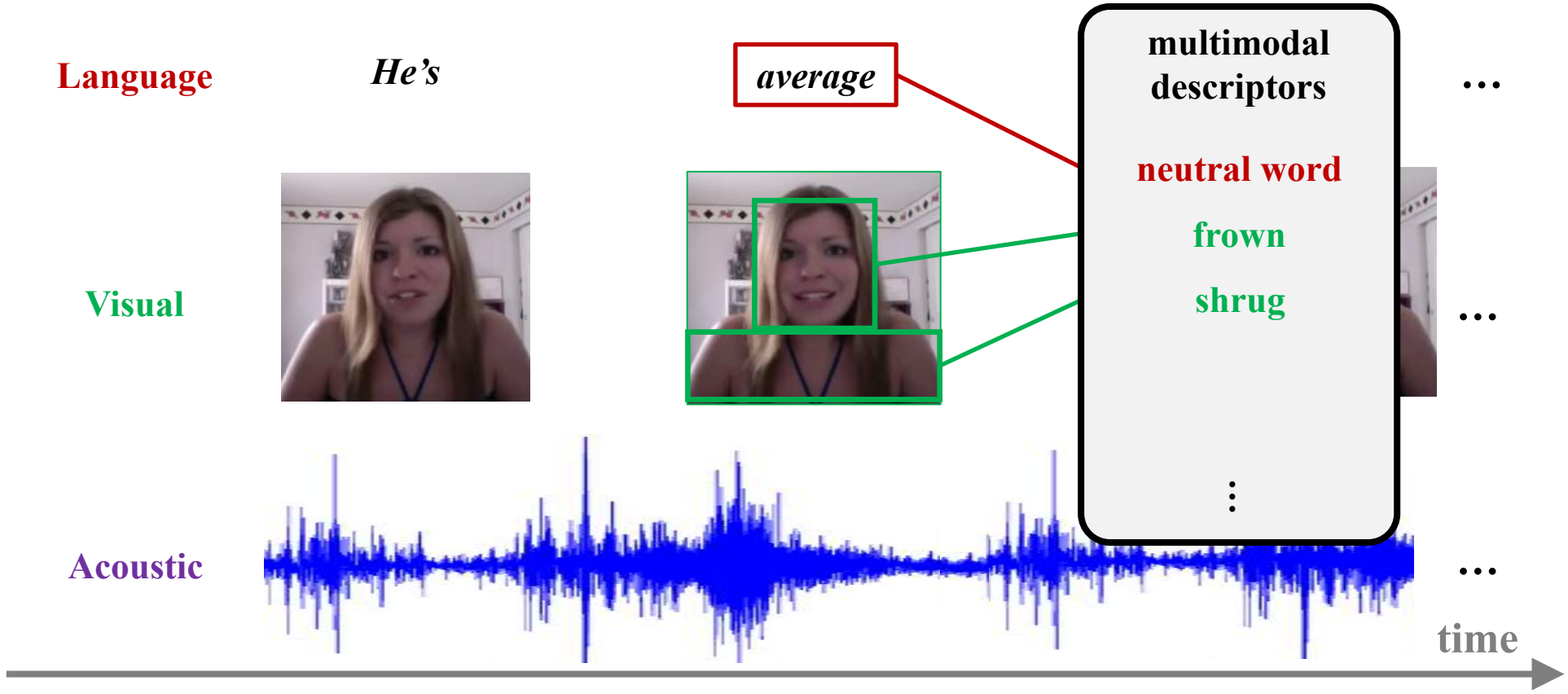
Multimodal Descriptors



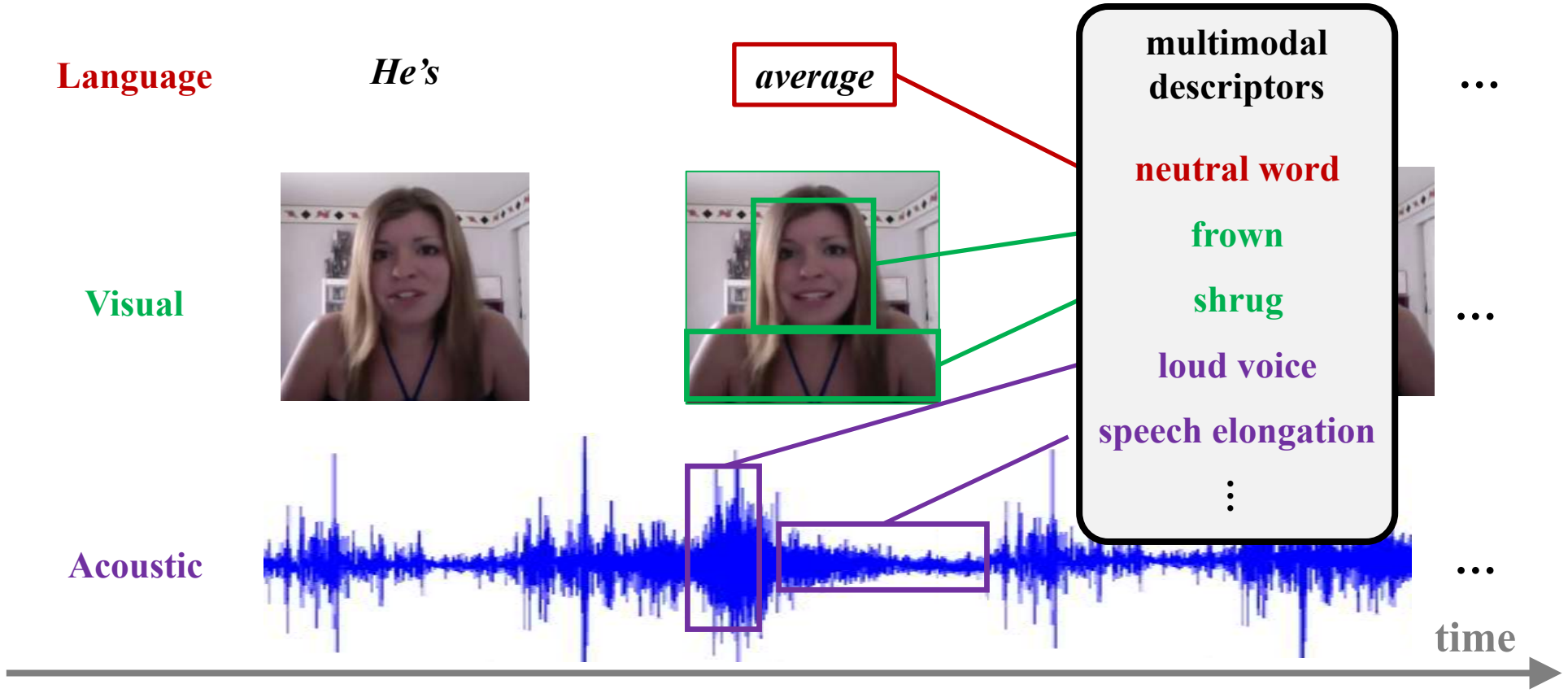
Language Descriptors



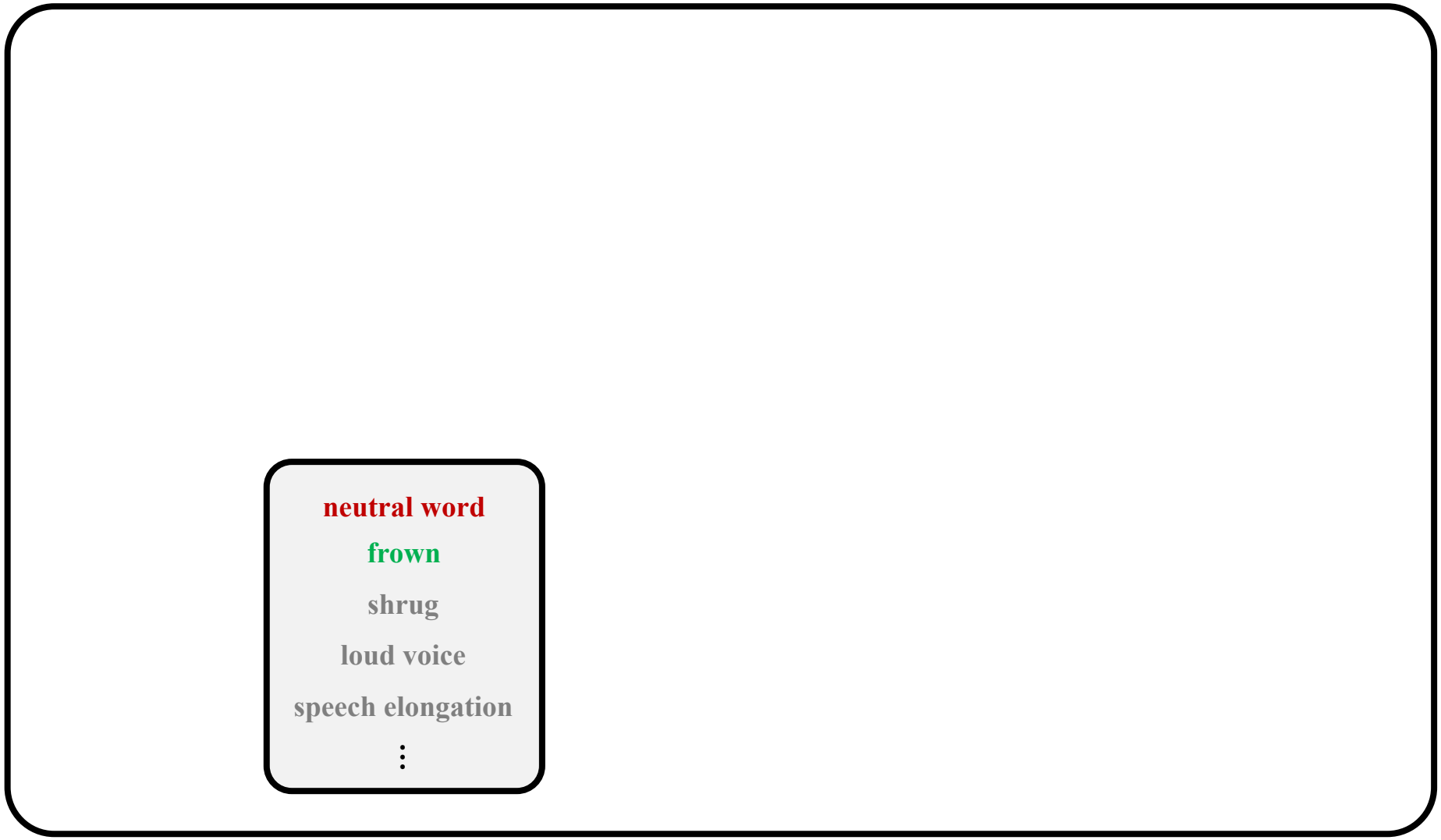
Visual Descriptors



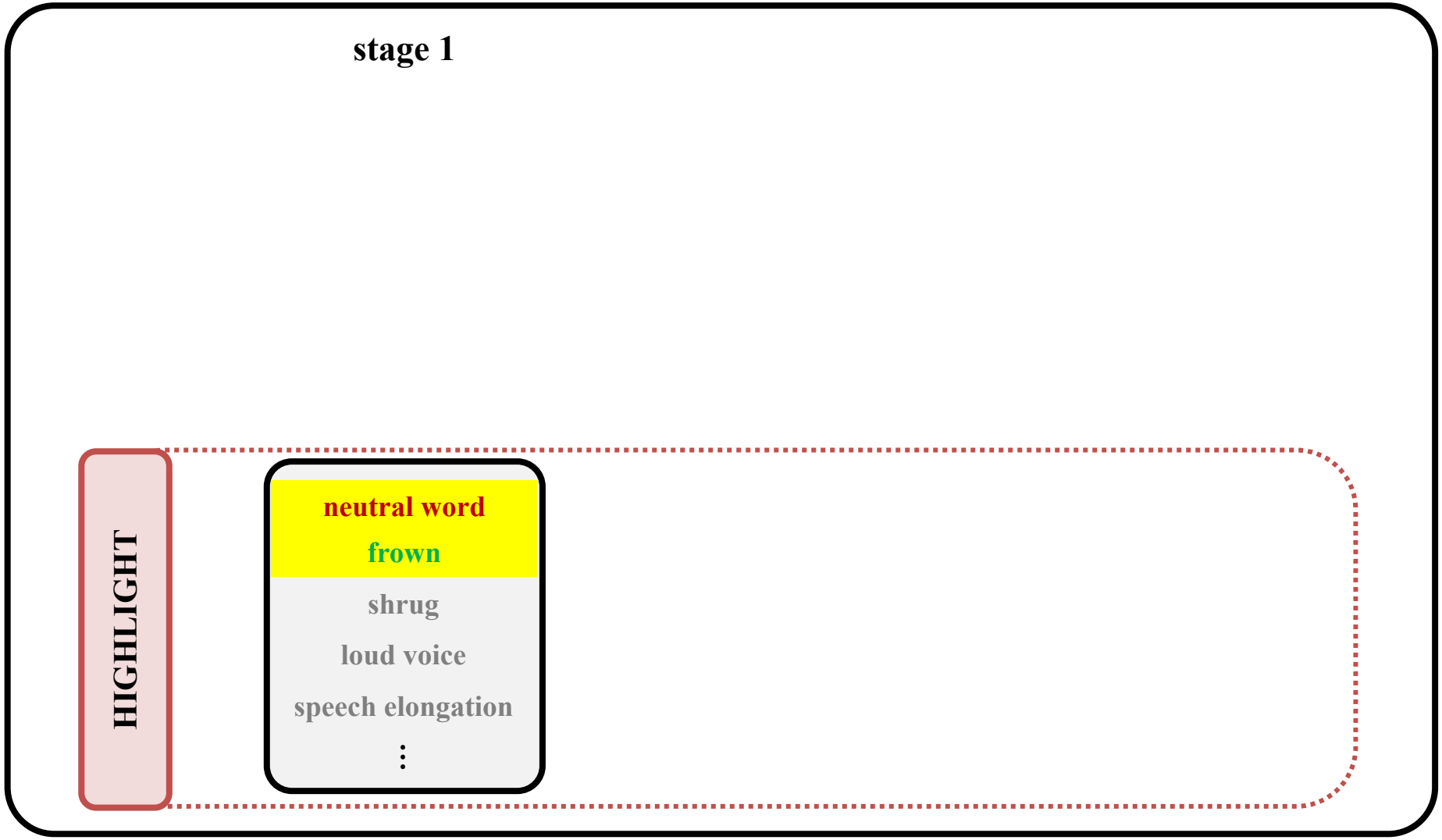
Acoustic Descriptors



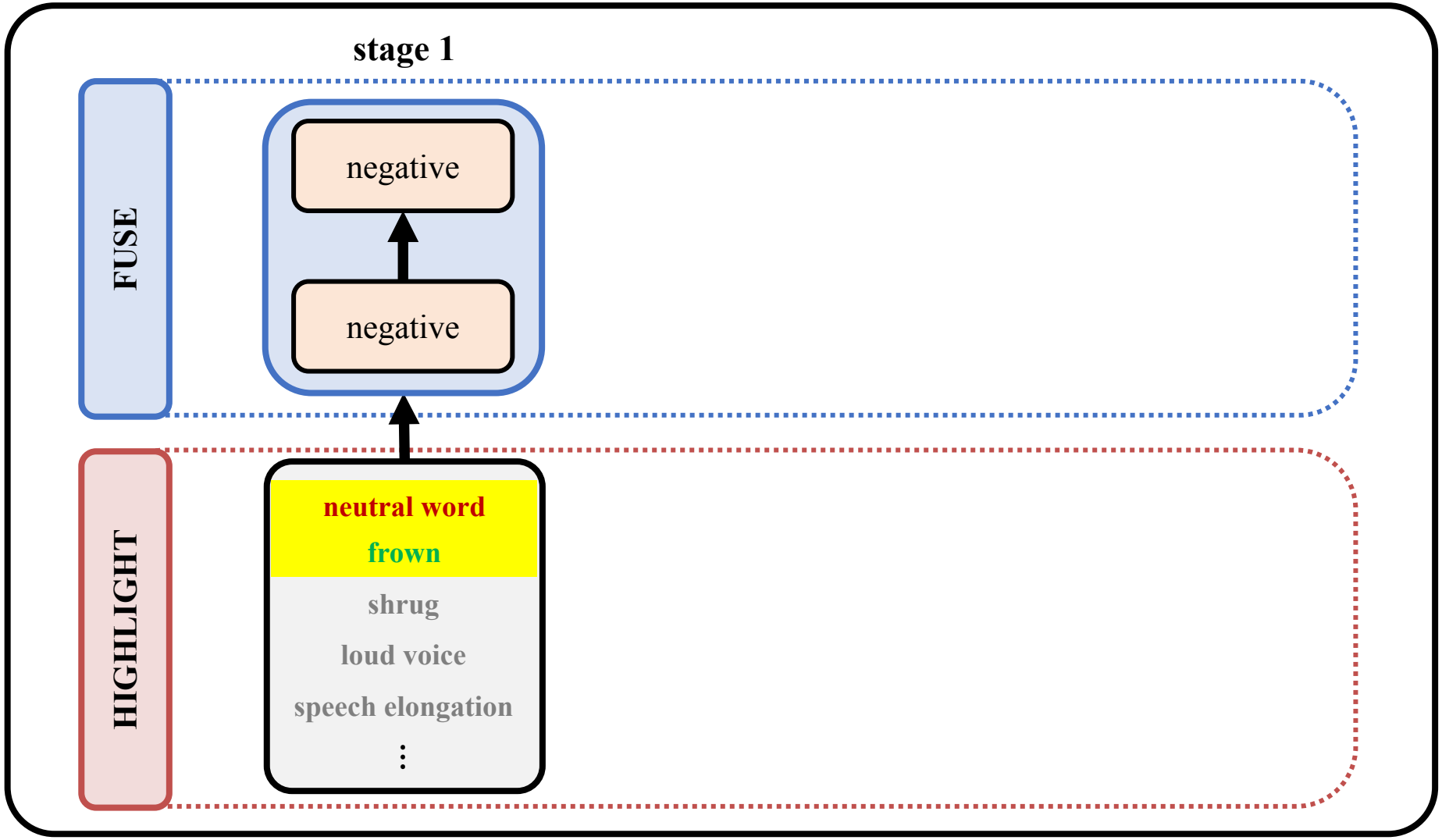
Multistage Fusion



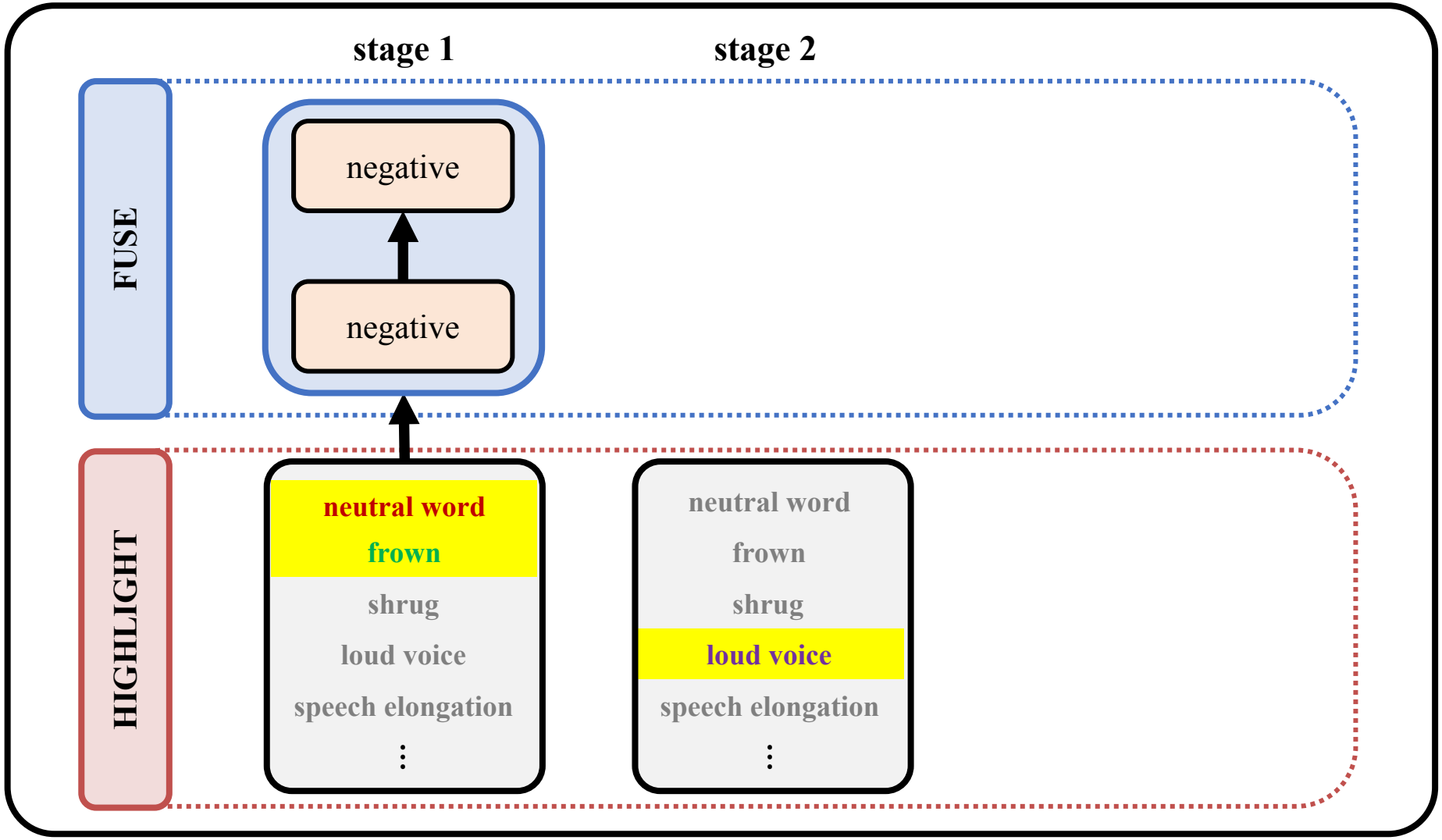
Multistage Fusion



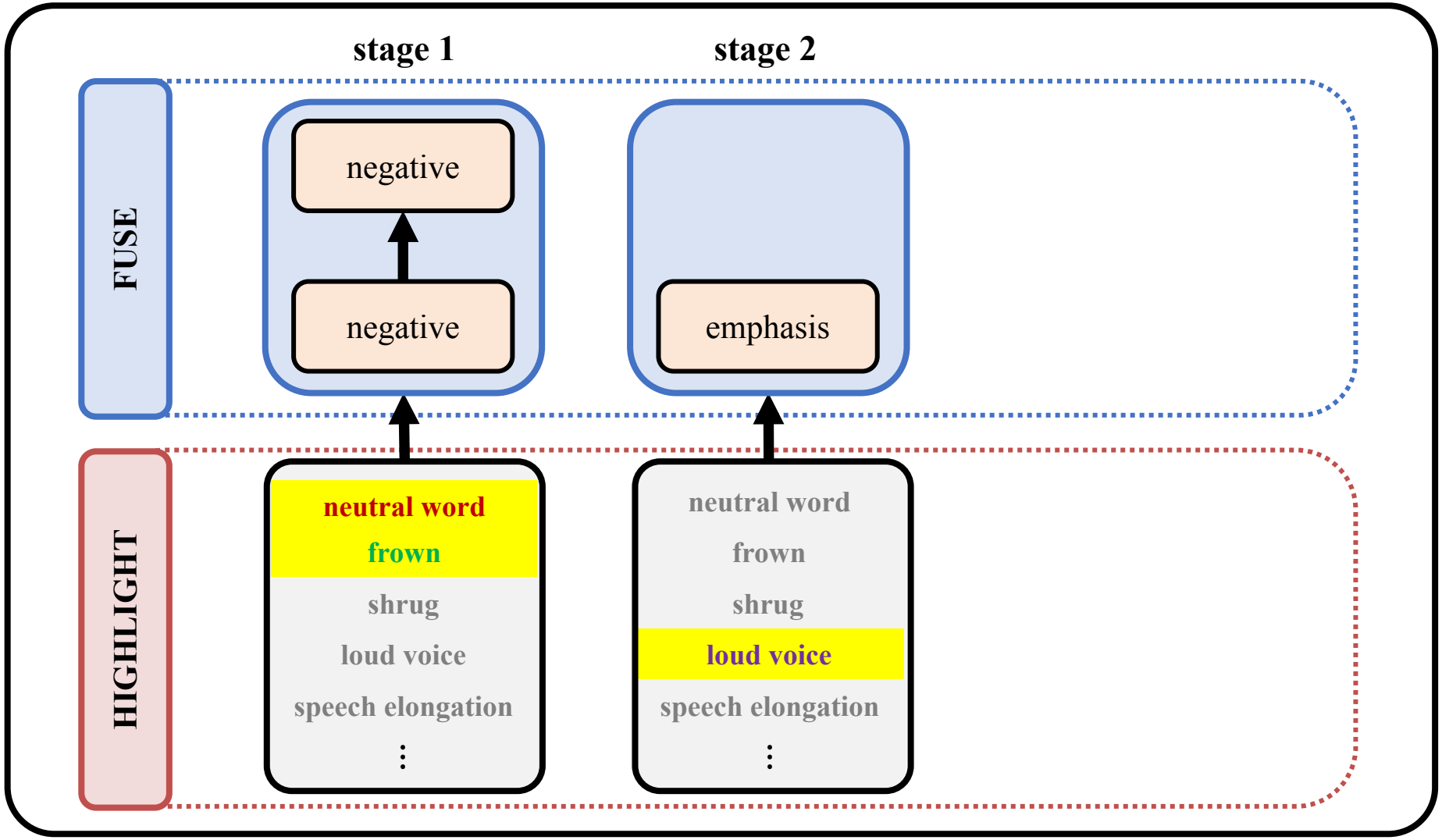
Multistage Fusion



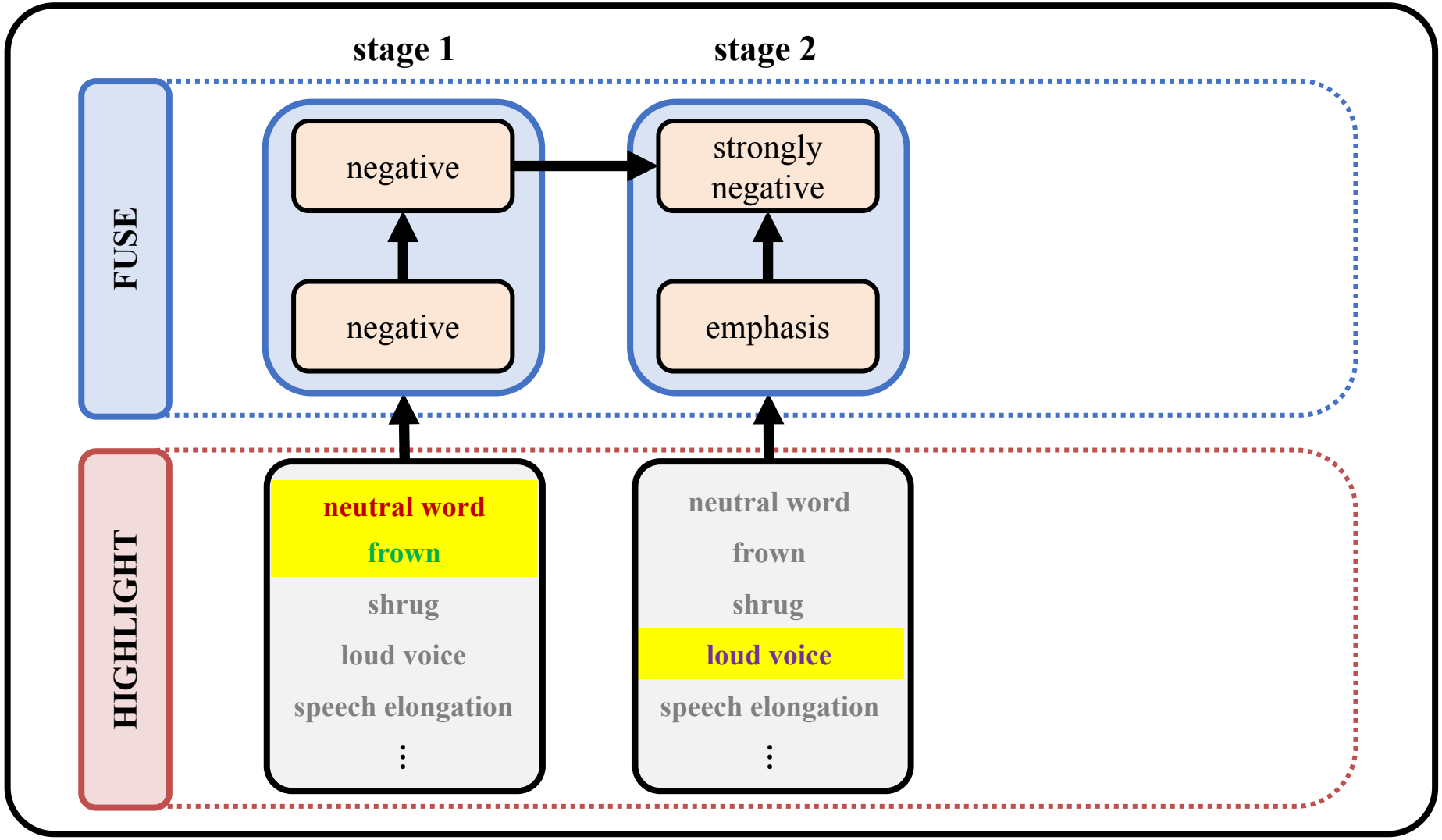
Multistage Fusion



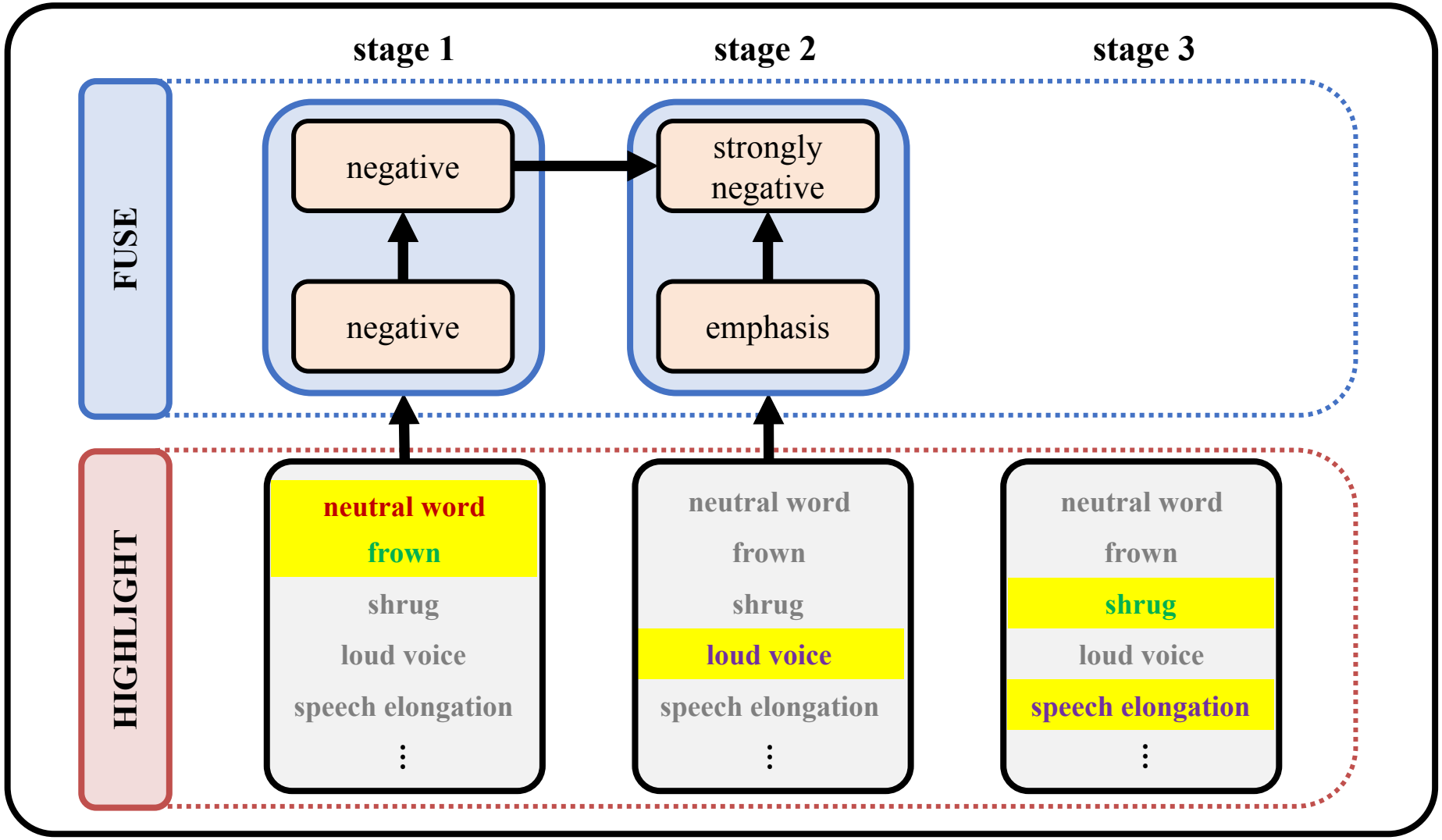
Multistage Fusion



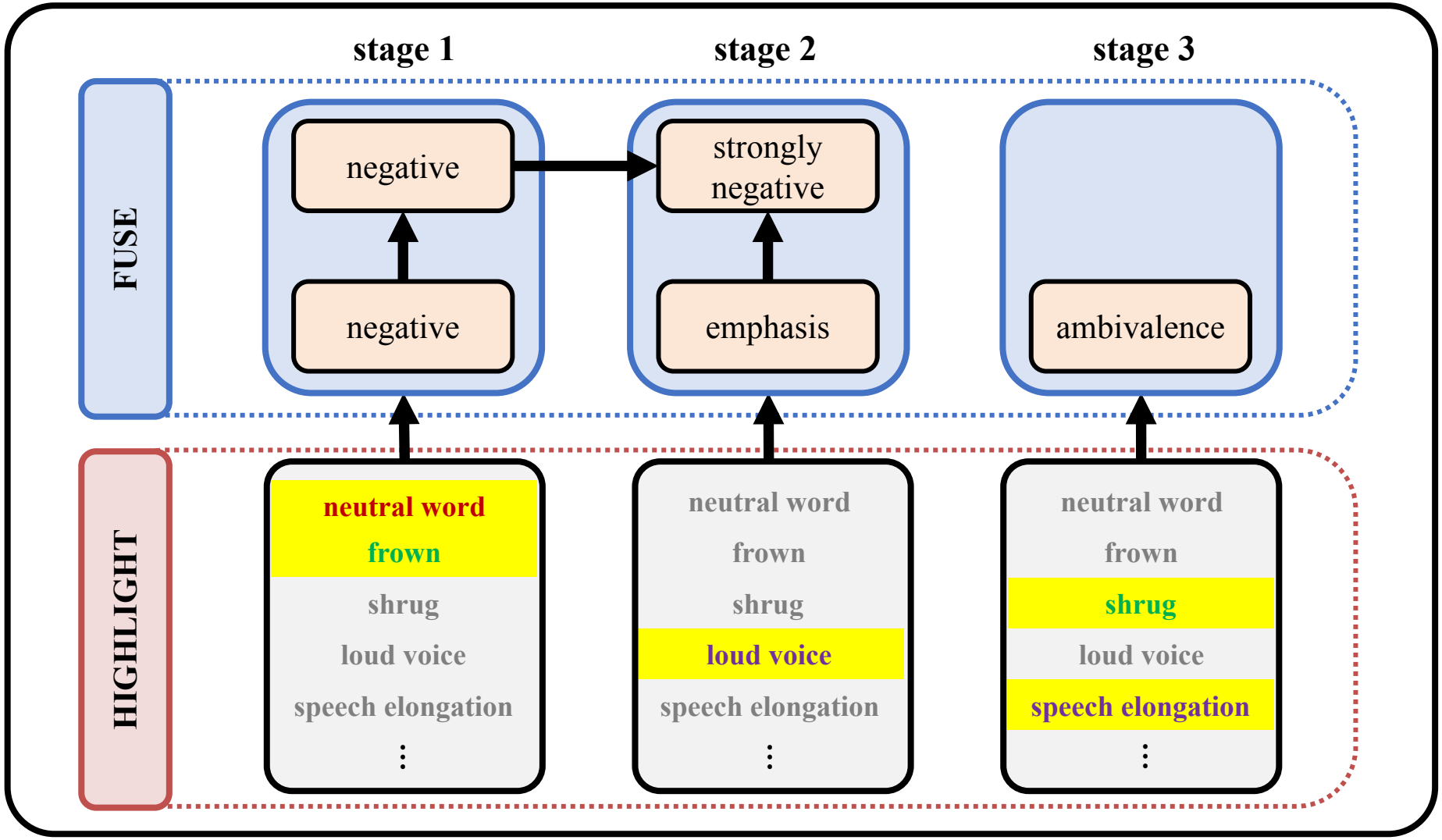
Multistage Fusion



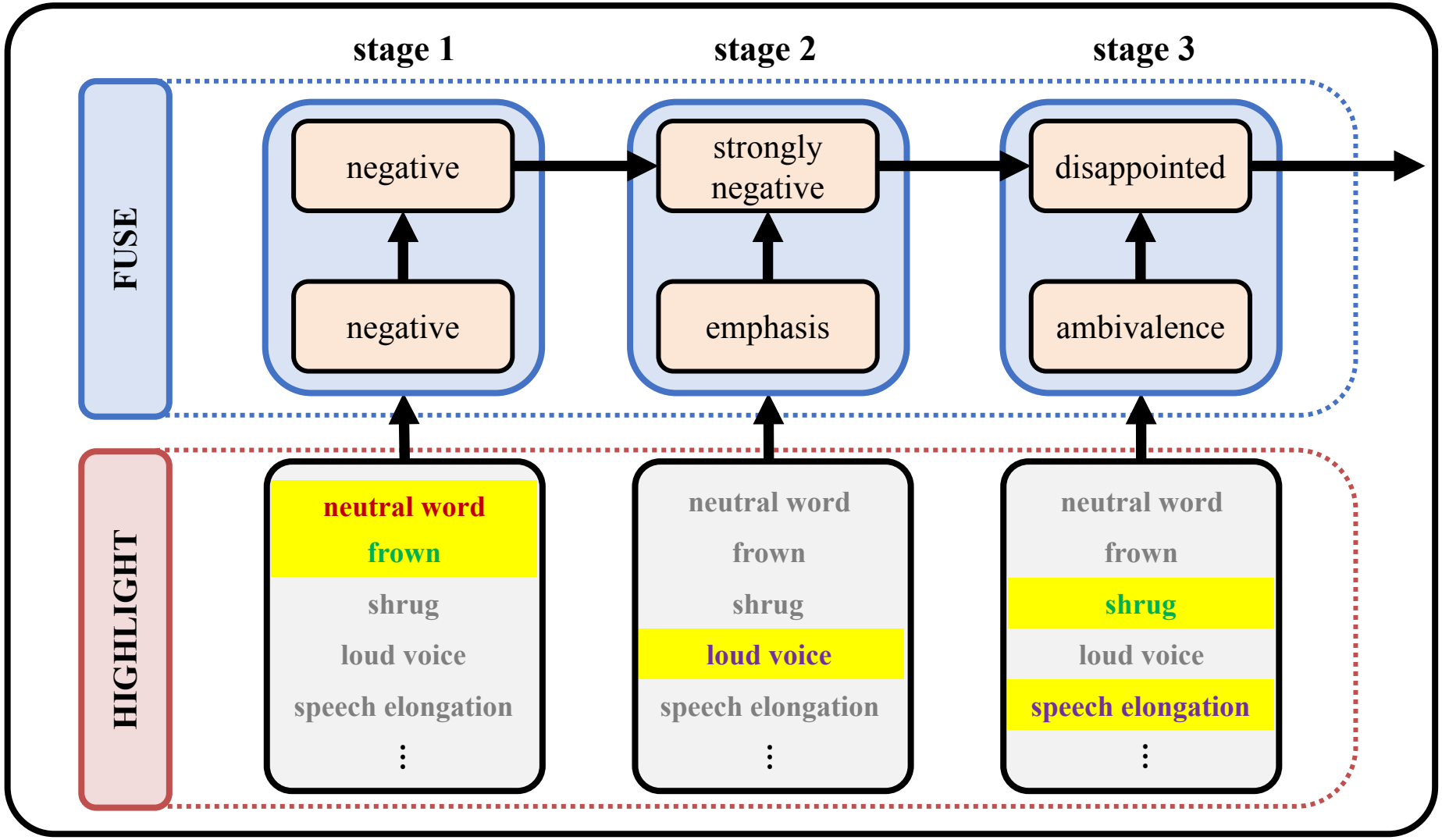
Multistage Fusion



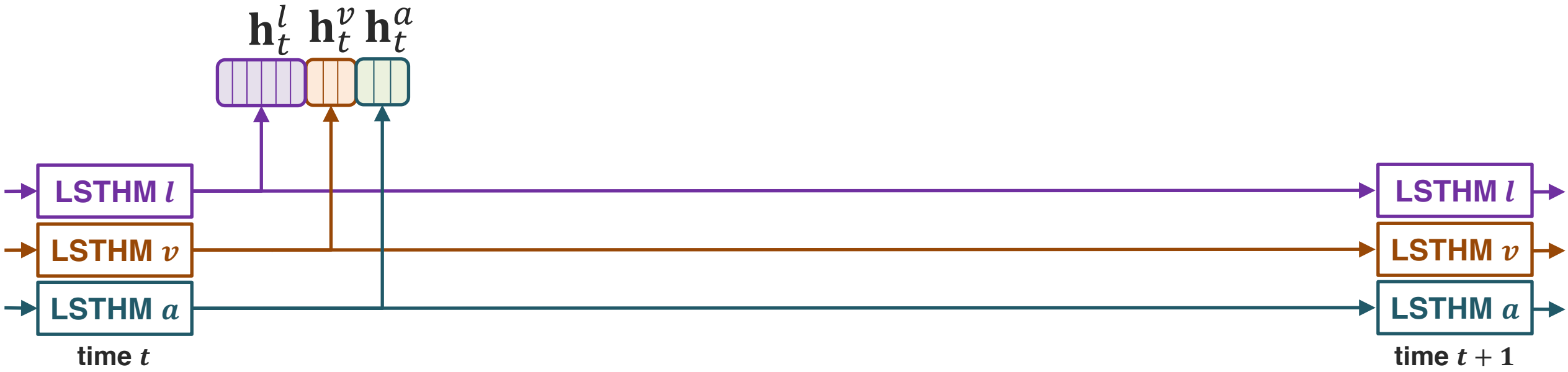
Multistage Fusion



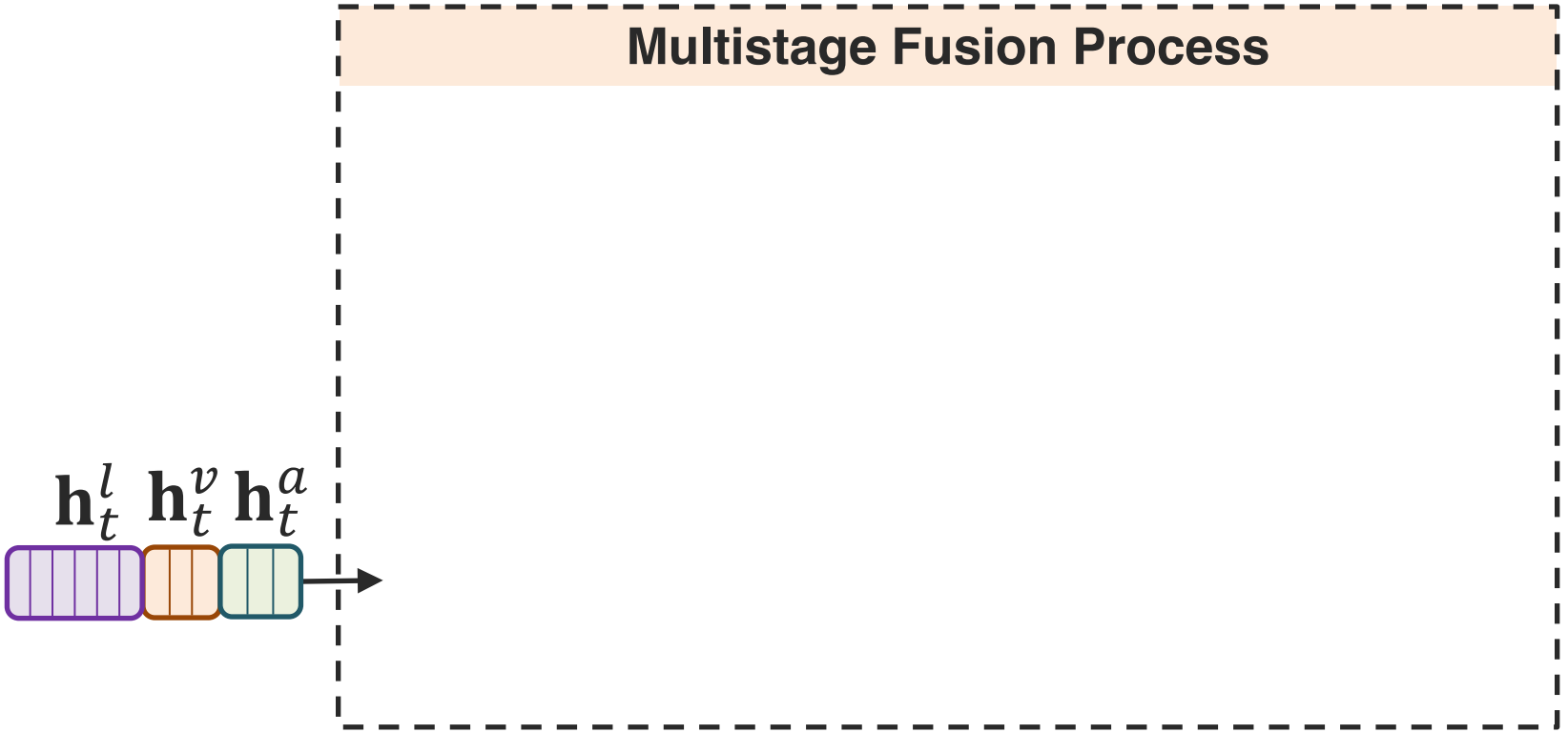
Multistage Fusion



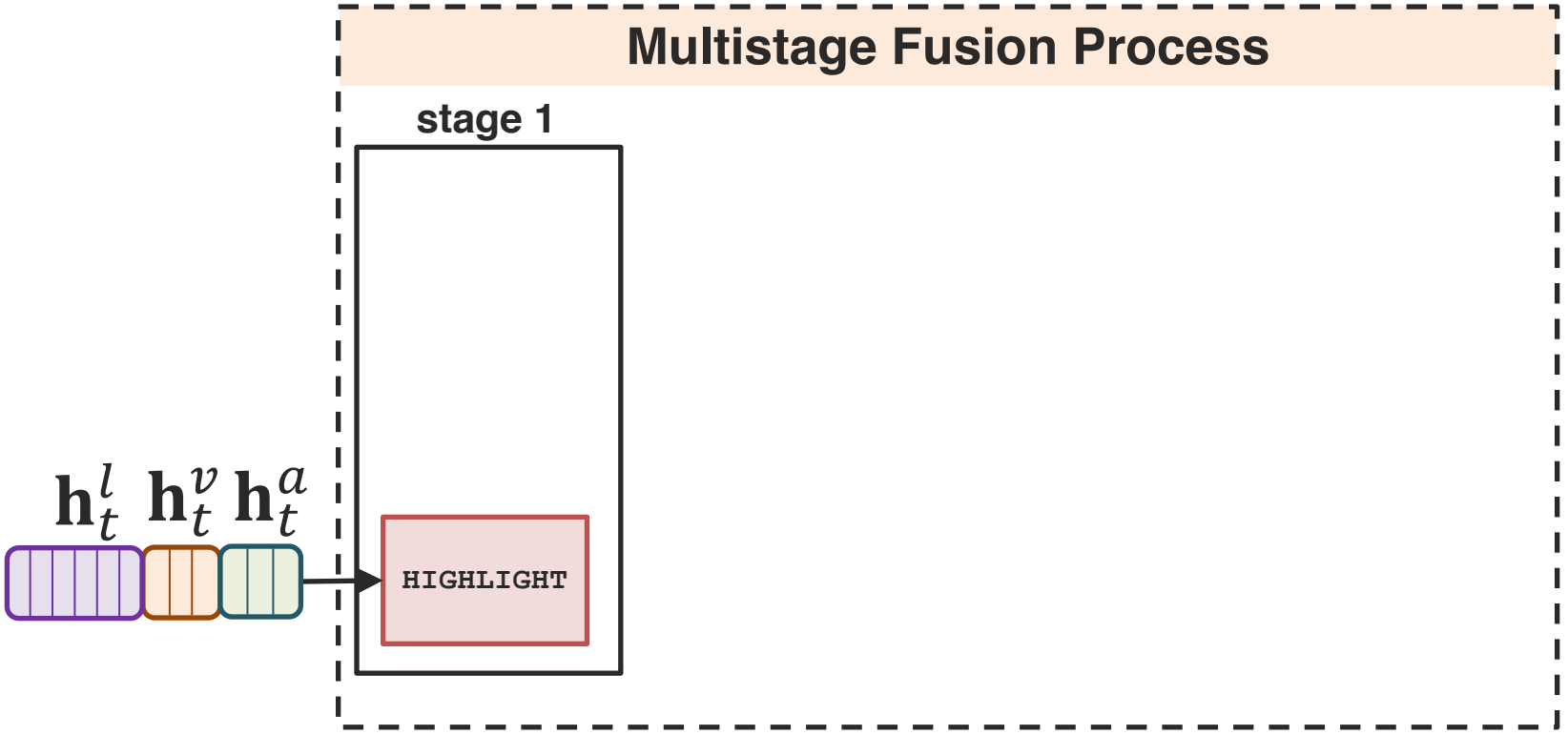
Intra-modal Recurrent Networks



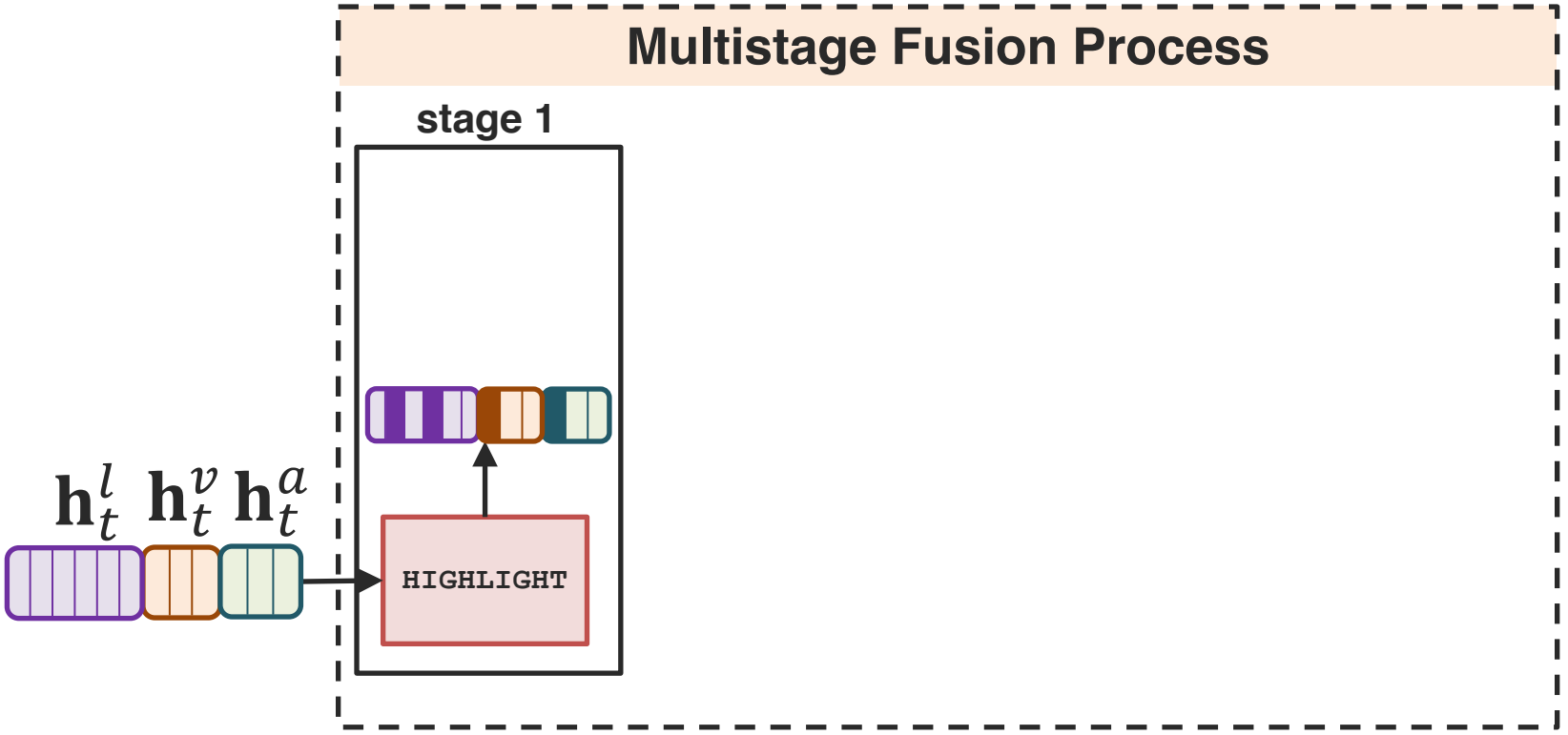
Multistage Fusion Process



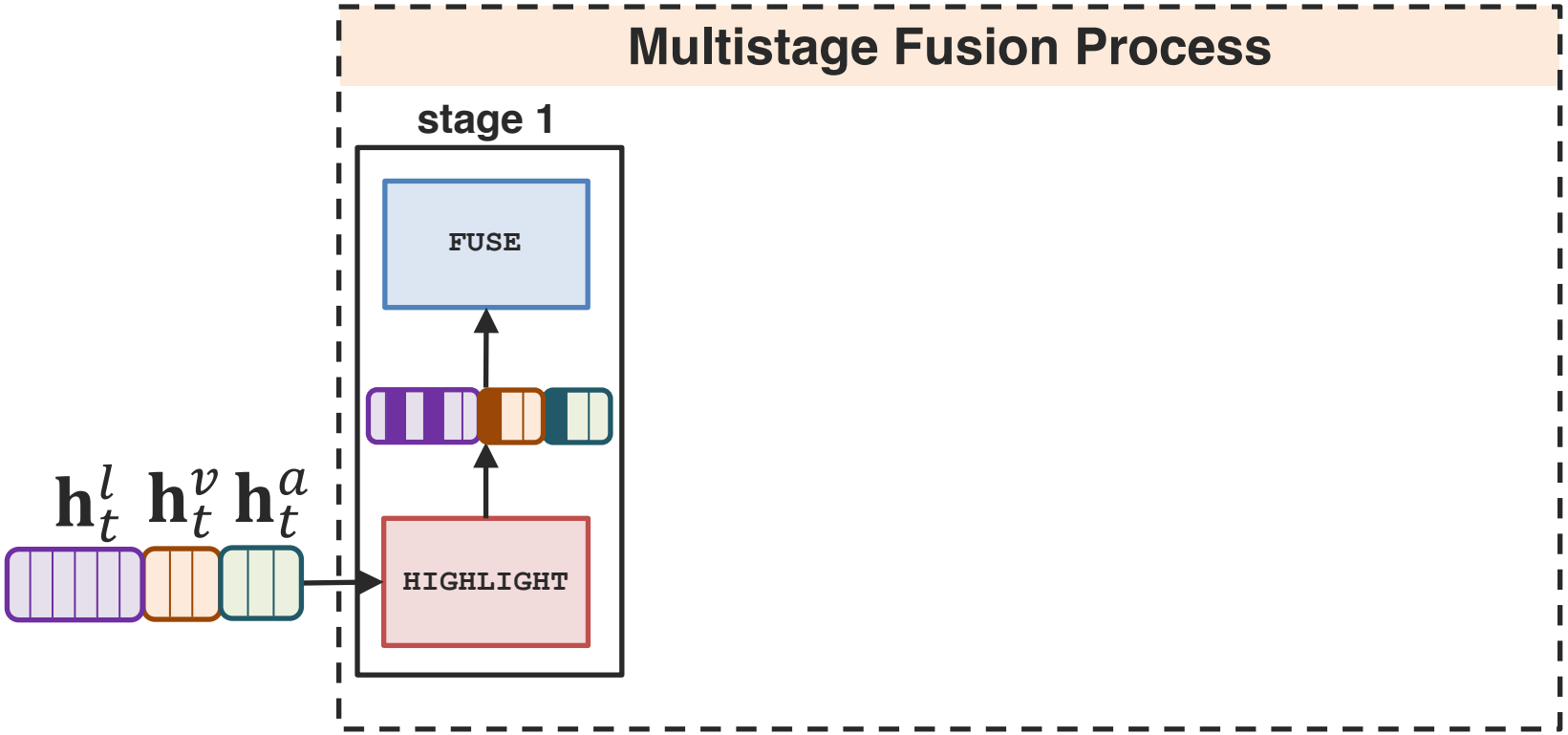
Multistage Fusion Process



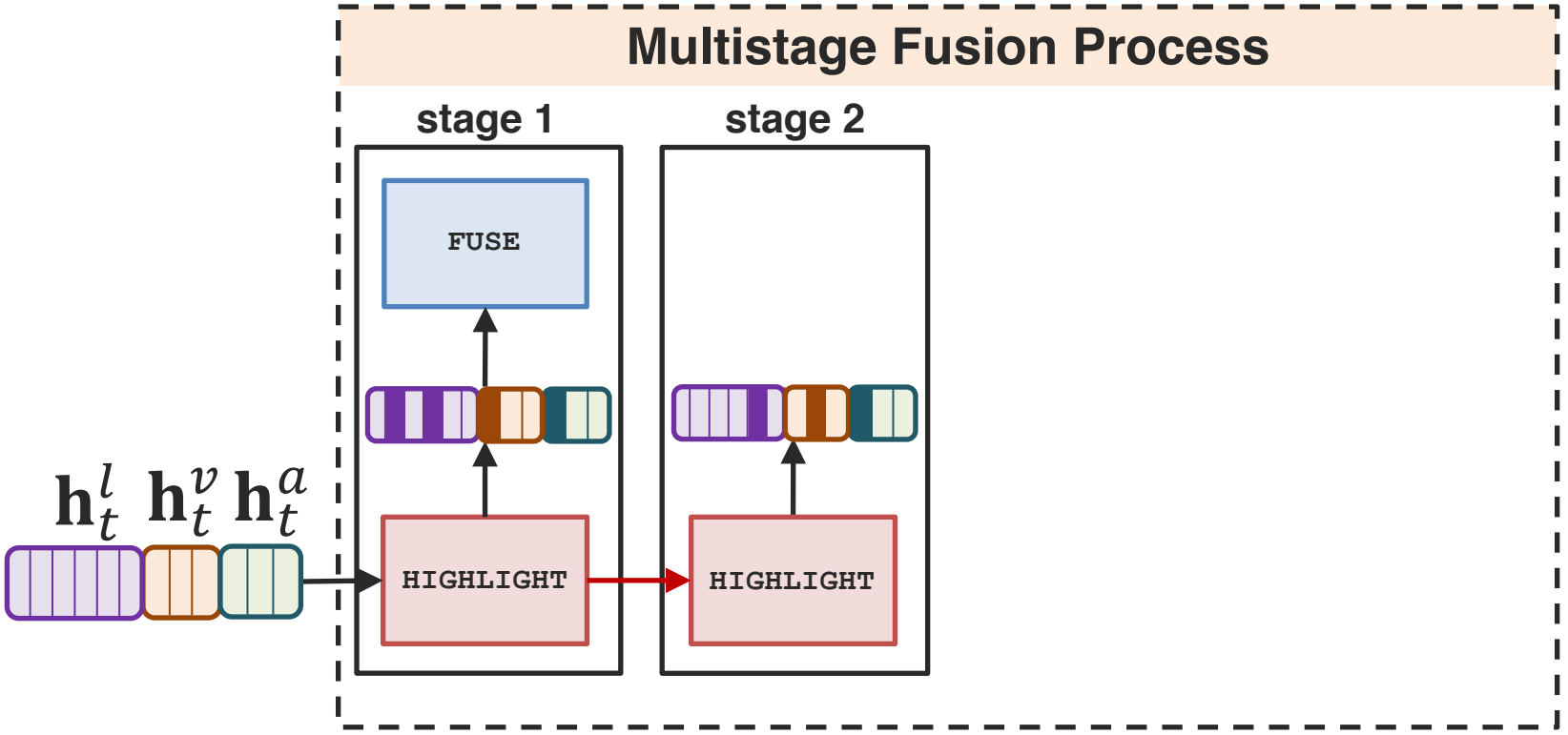
Multistage Fusion Process



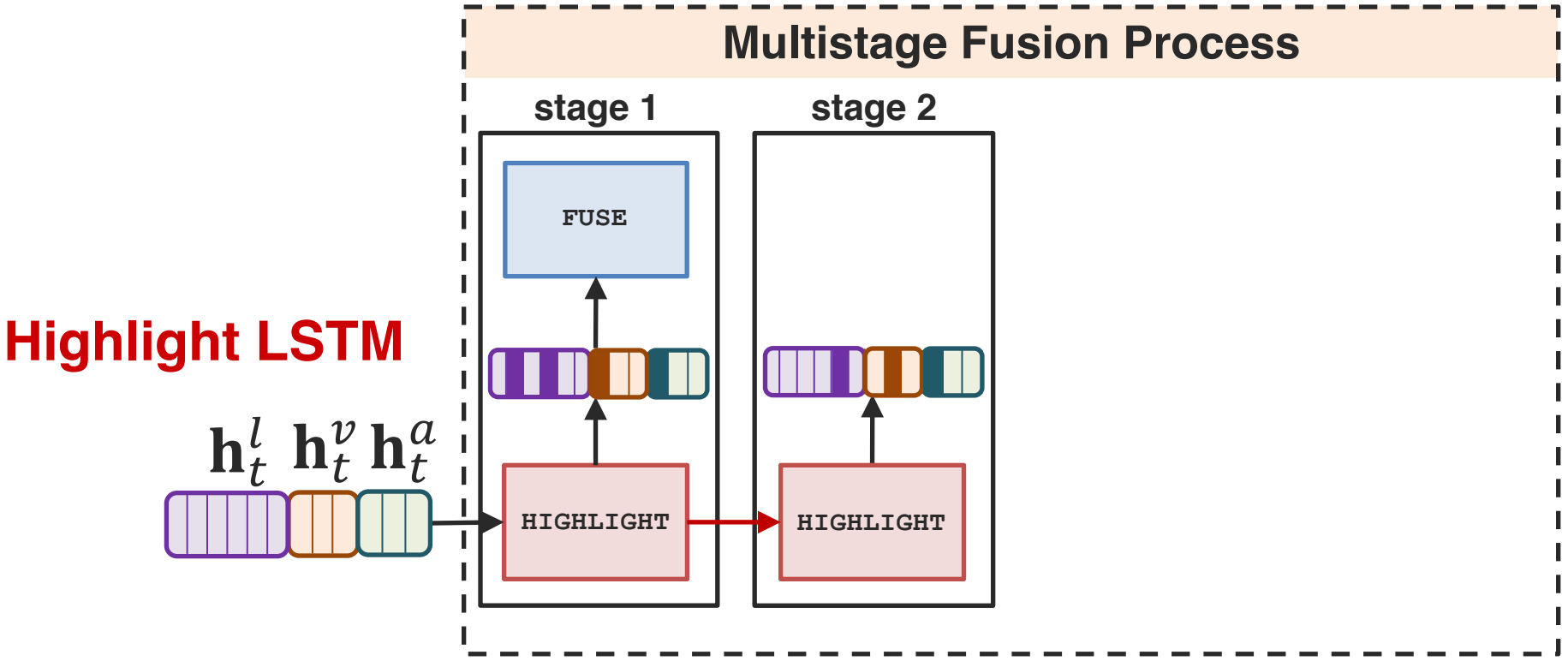
Multistage Fusion Process



Multistage Fusion Process



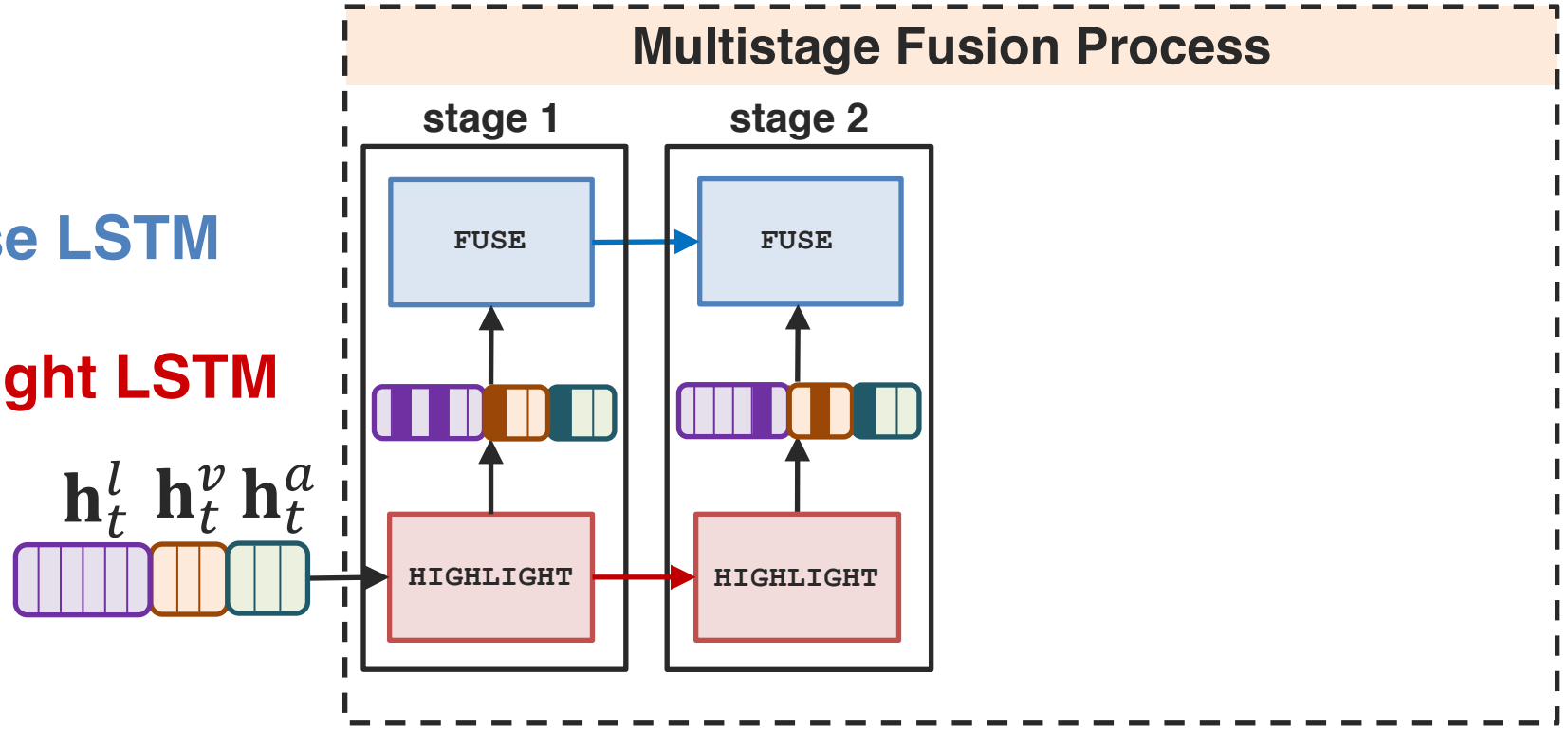
Multistage Fusion Process



Multistage Fusion Process

Fuse LSTM

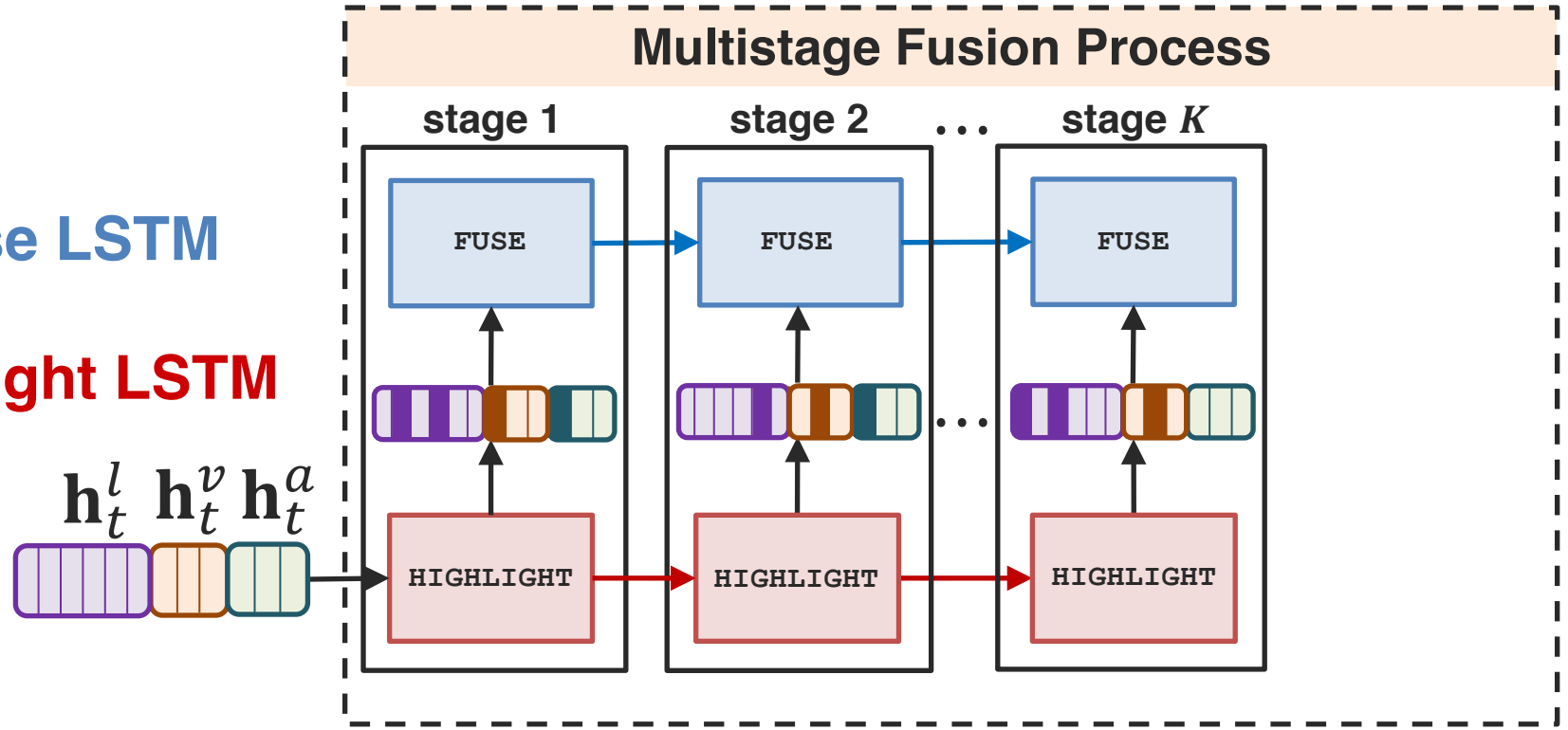
Highlight LSTM



Multistage Fusion Process

Fuse LSTM

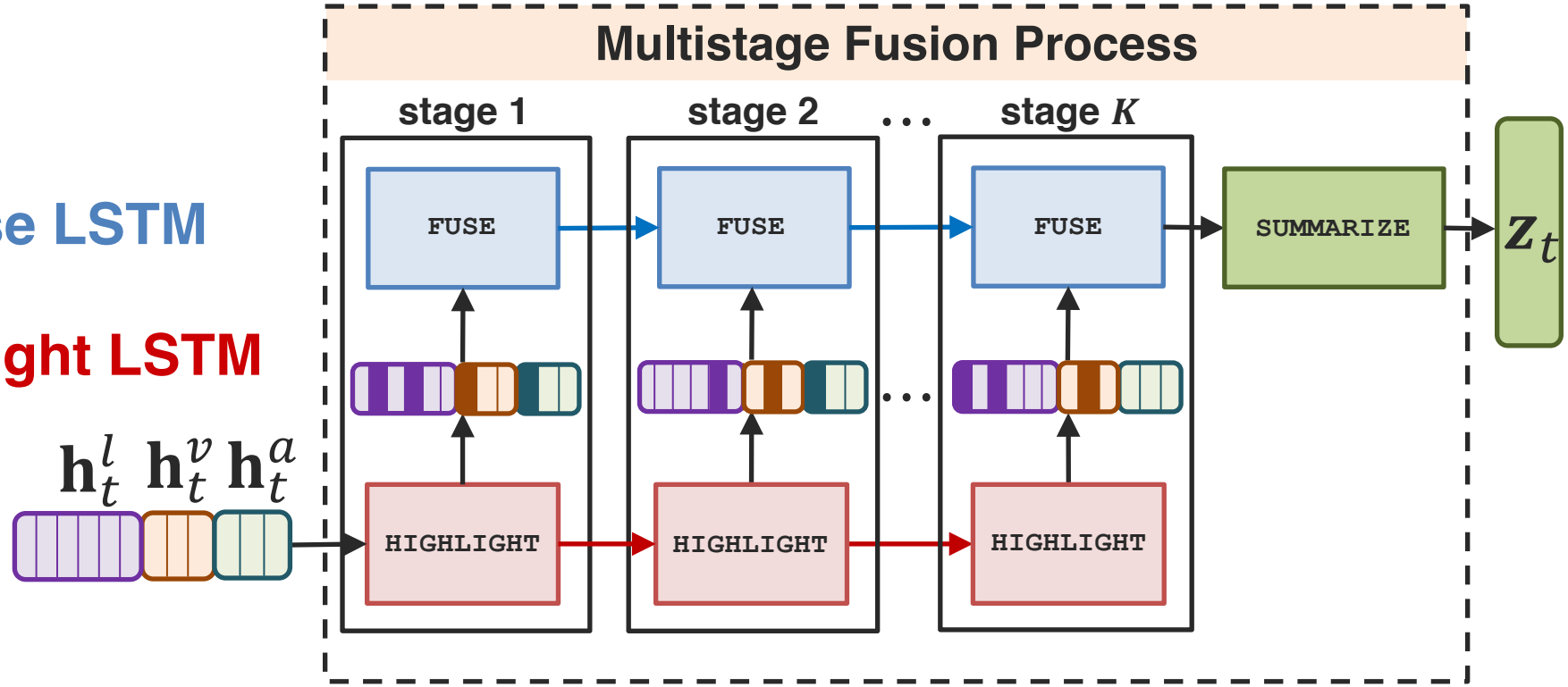
Highlight LSTM



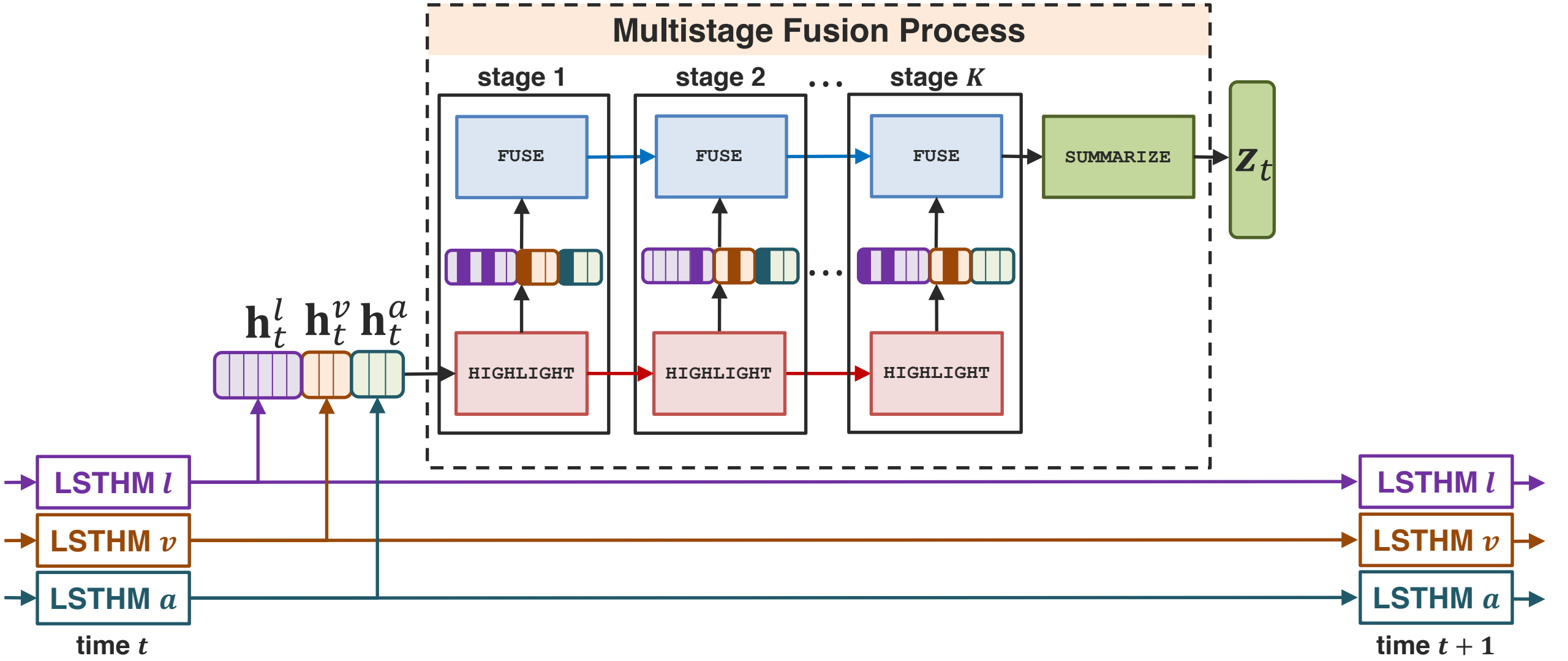
Multistage Fusion Process

Fuse LSTM

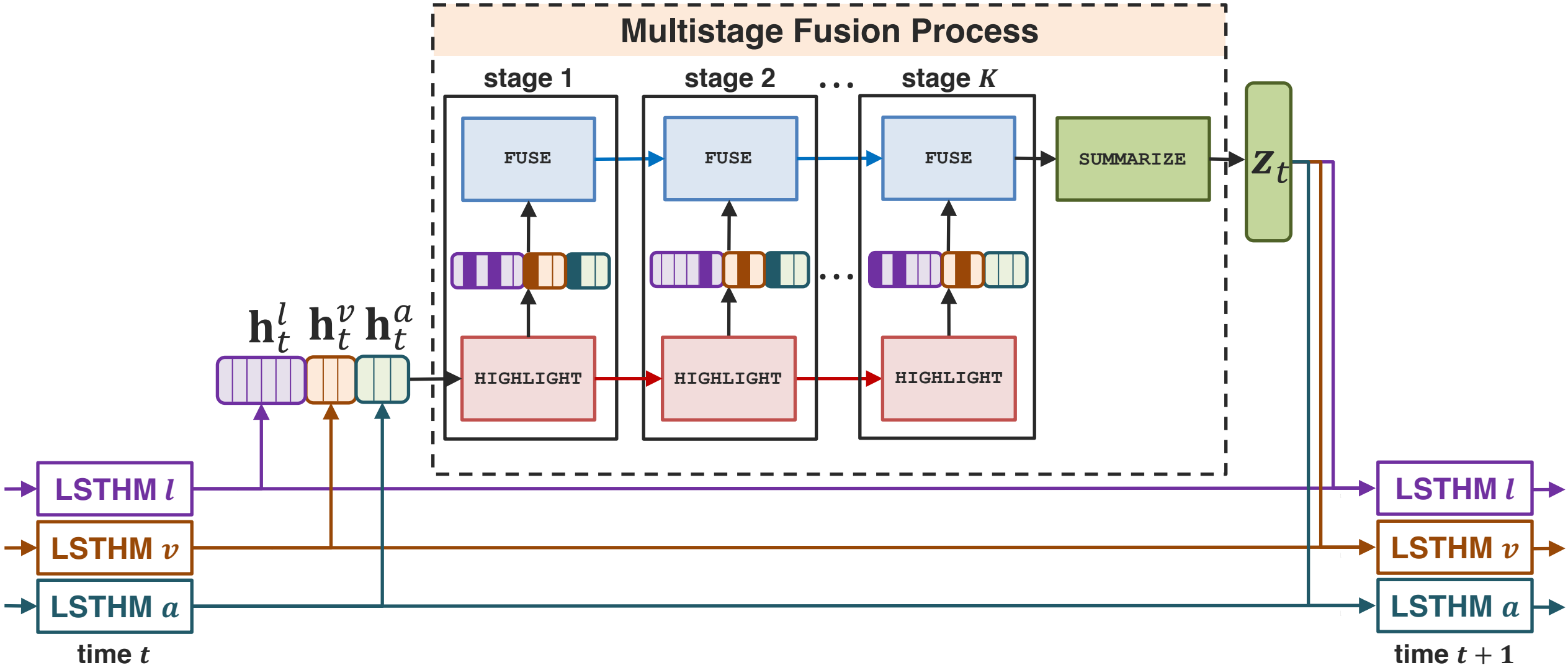
Highlight LSTM



Recurrent Multistage Fusion Network



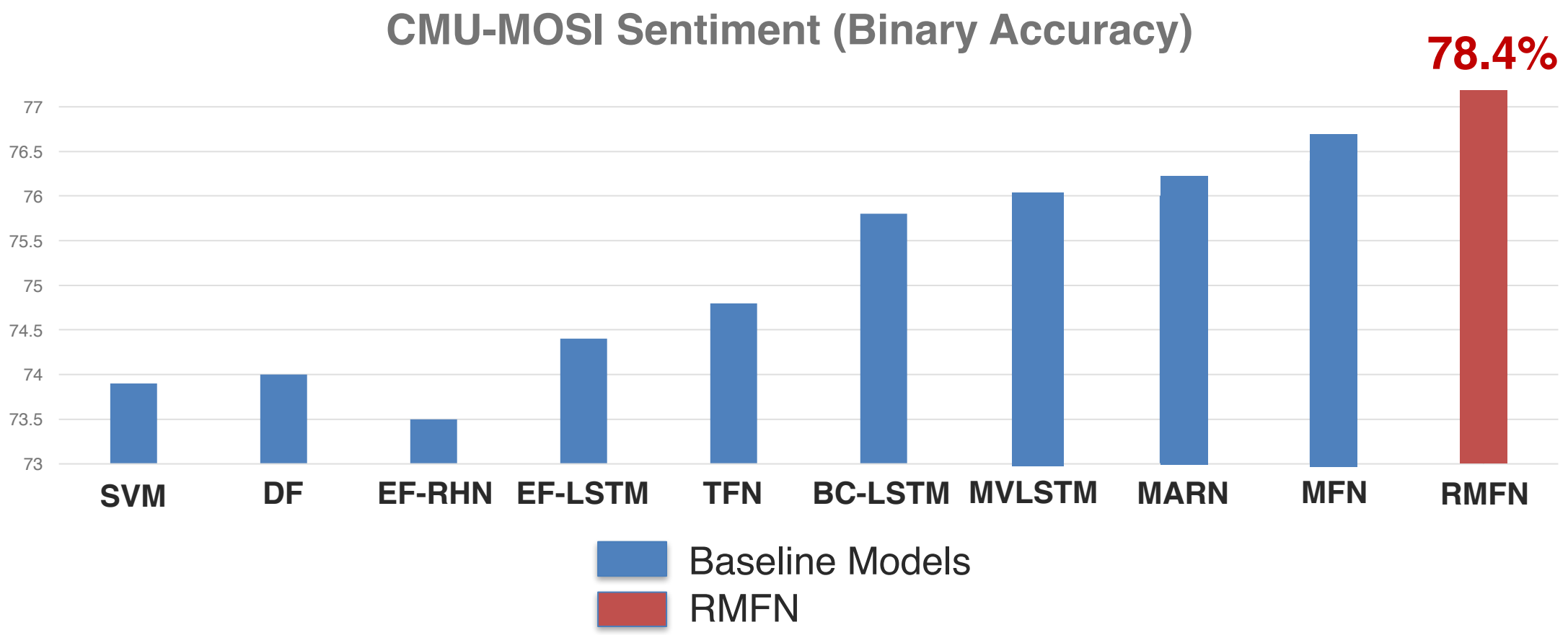
Recurrent Multistage Fusion Network



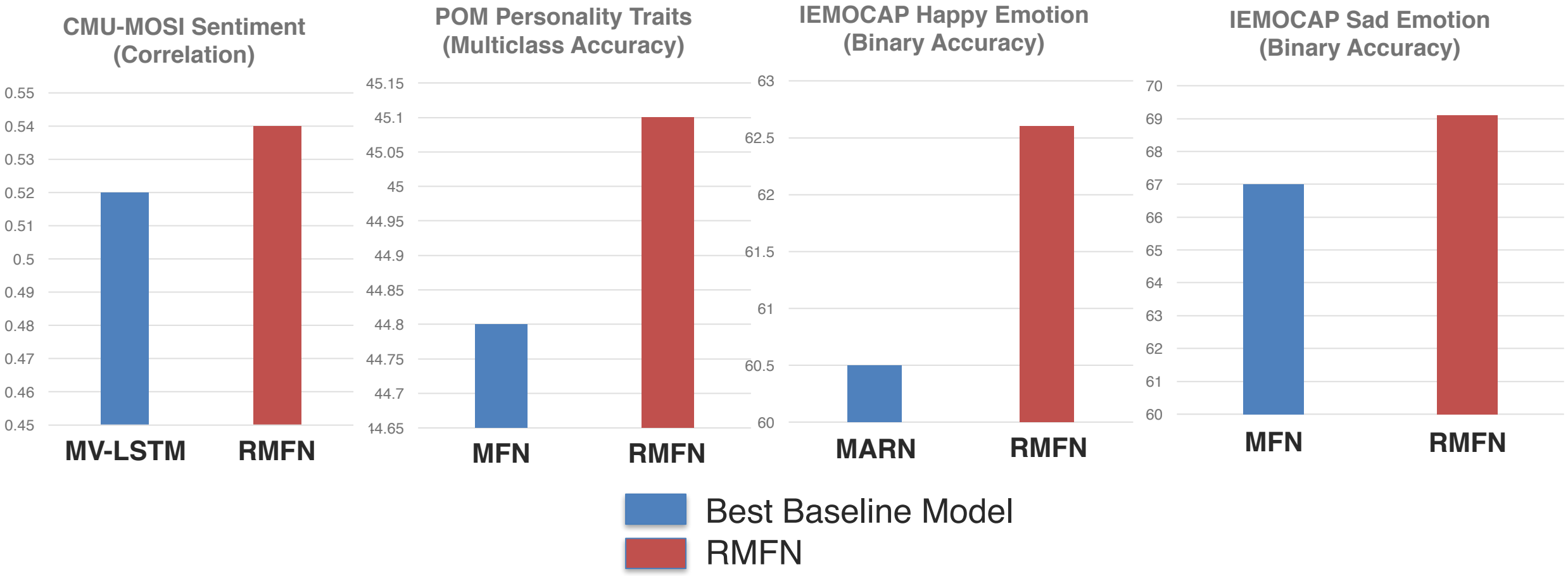
Baseline Models

1. Non-temporal Models
   - SVM (Cortes and Vapnik, 1995), DF (Nojavanasghari et al., 2016)
2. Early Fusion
   - EF-LSTM (Hochreiter and Schmidhuber, 1997), EF-RHN (Zilly et al., 2016)
3. Late Fusion
   - LMF (Liu et al., 2018), TFN (Zadeh et al., 2017), BC-LSTM (Poria et al., 2017)
4. Multi-view Learning
   - MV-LSTM (Rajagopalan et al., 2016)
5. Memory-based models
   - MARN, MFN (Zadeh et al., 2018)

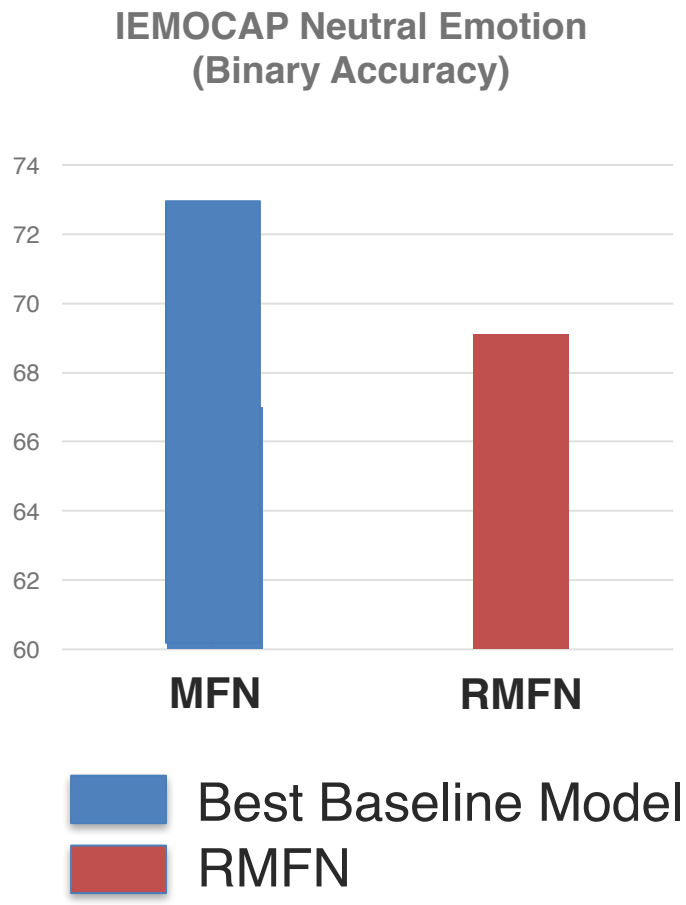
State-of-the-art Results



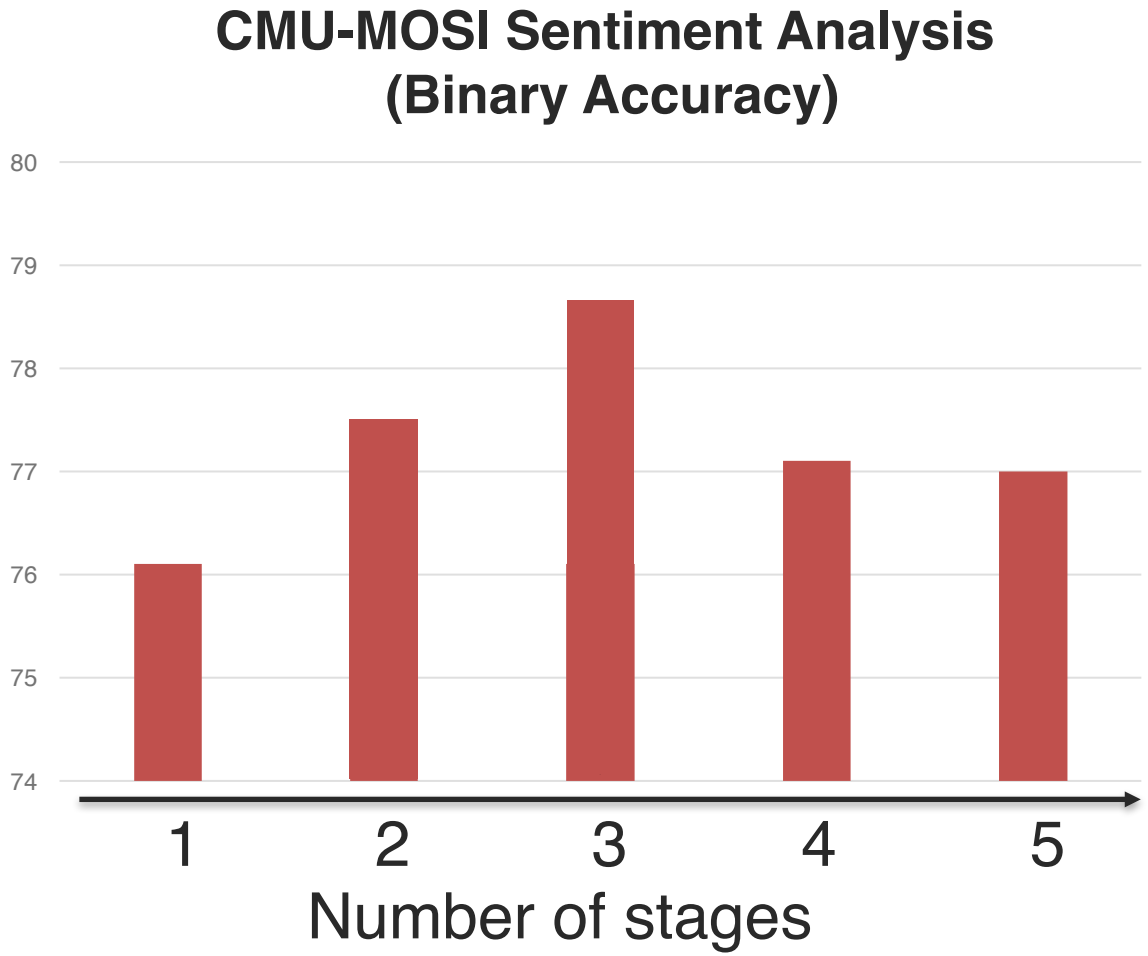
State-of-the-art Results

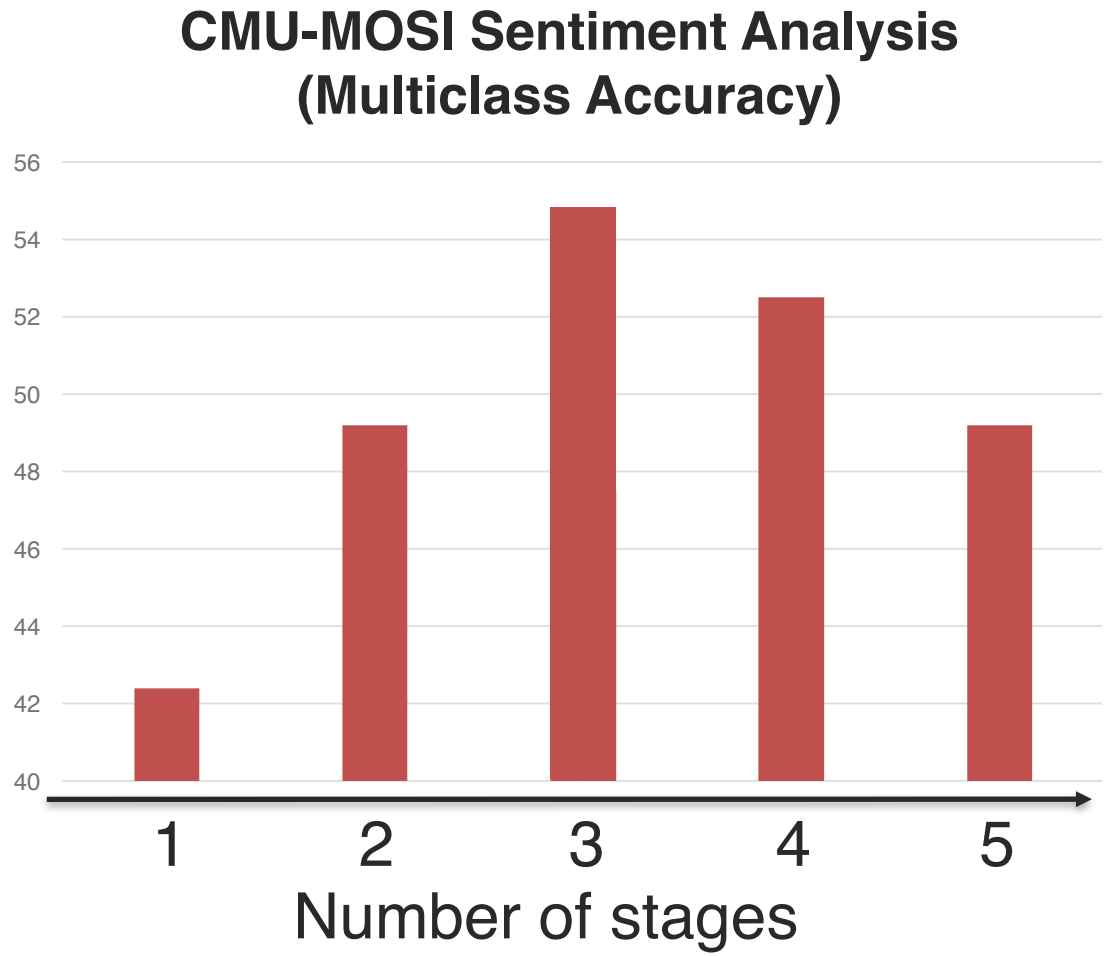


Results


IEMOCAP Neutral Emotion
(Binary Accuracy)
74
72
70
68
66
64
62
60
MFN
RMFN
Best Baseline Model
RMFN

Multiple Stages are Important





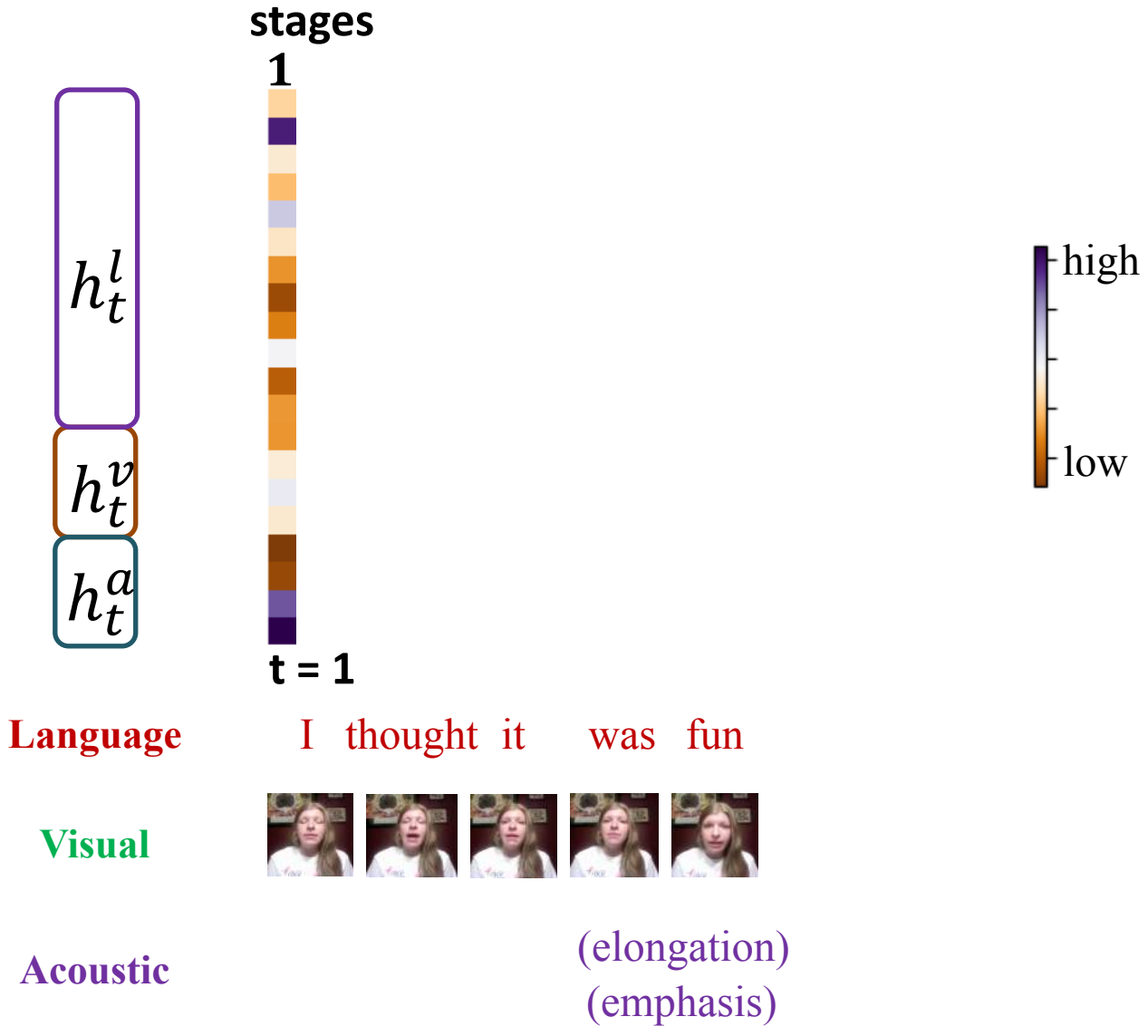
Interpretable Fusion

Language I thought it was fun

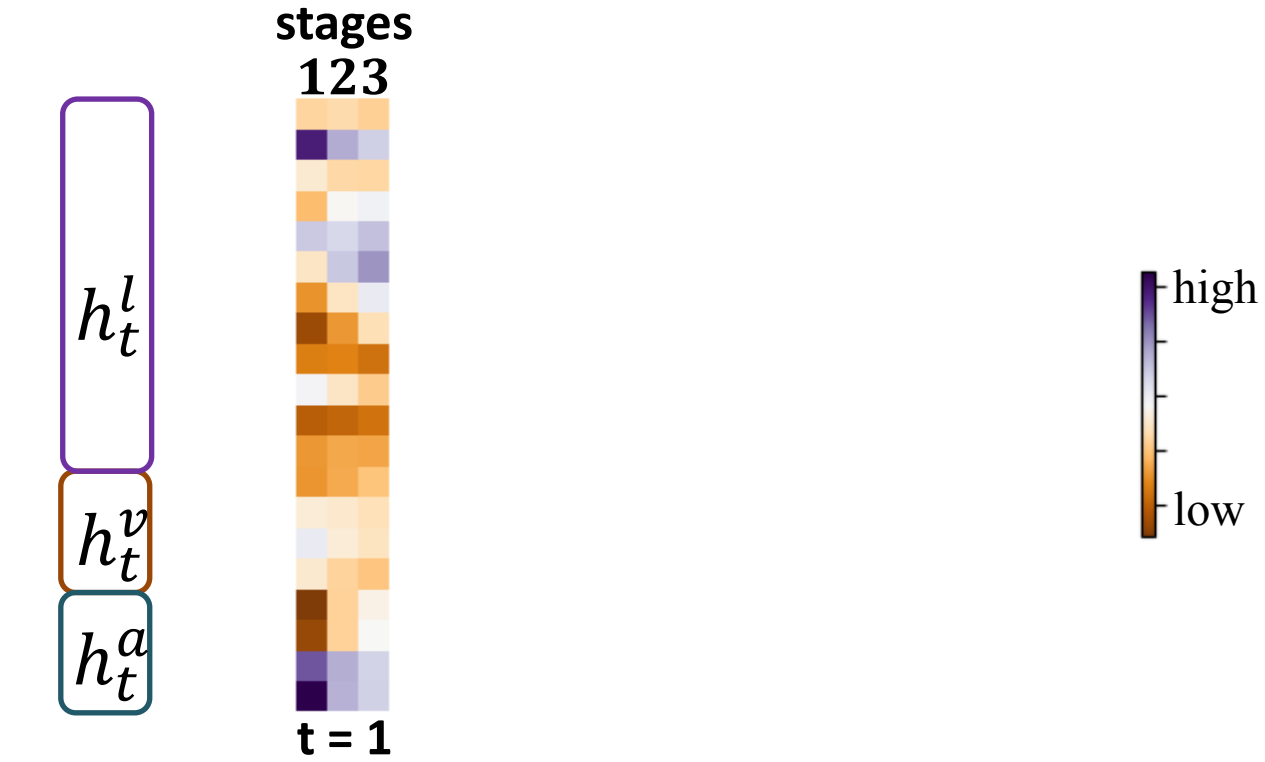
Visual 

Acoustic (elongation) (emphasis)

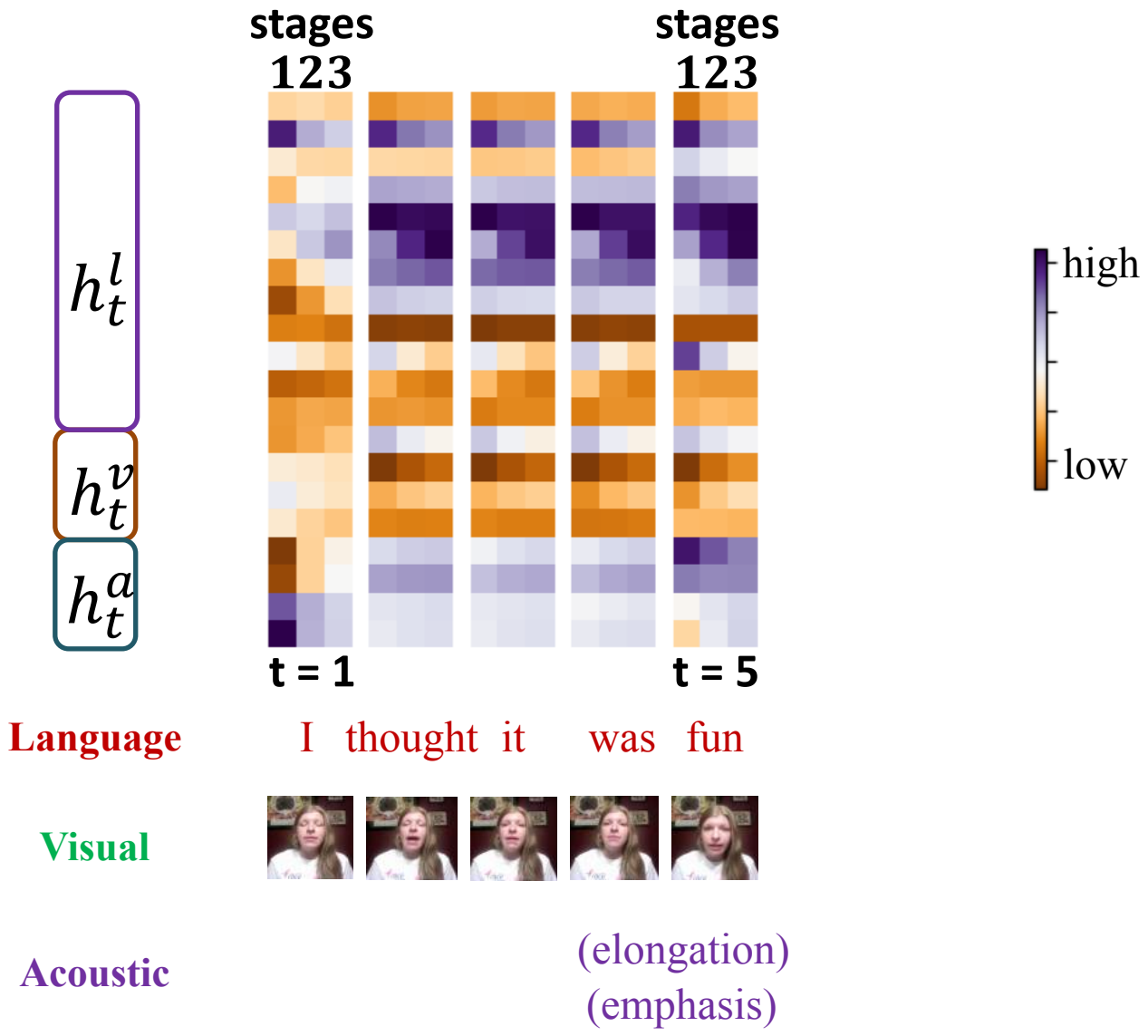
Interpretable Fusion



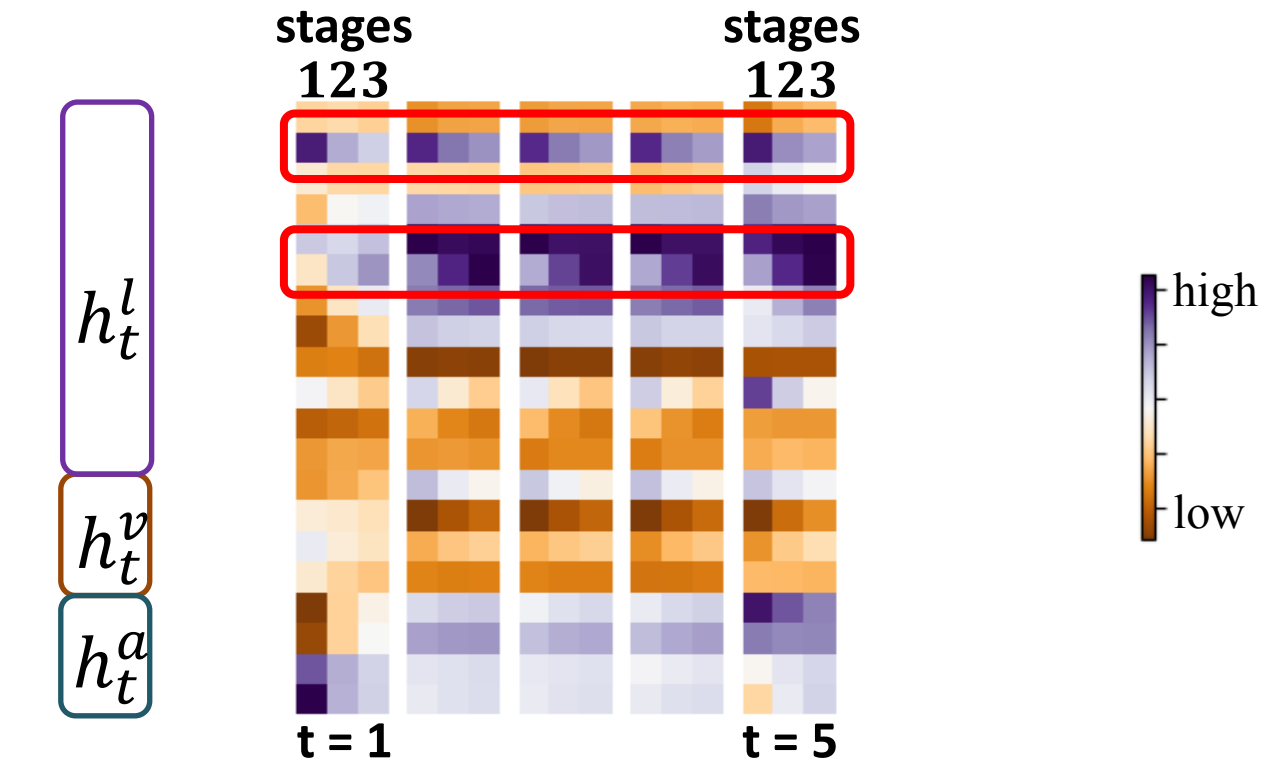
Interpretable Fusion



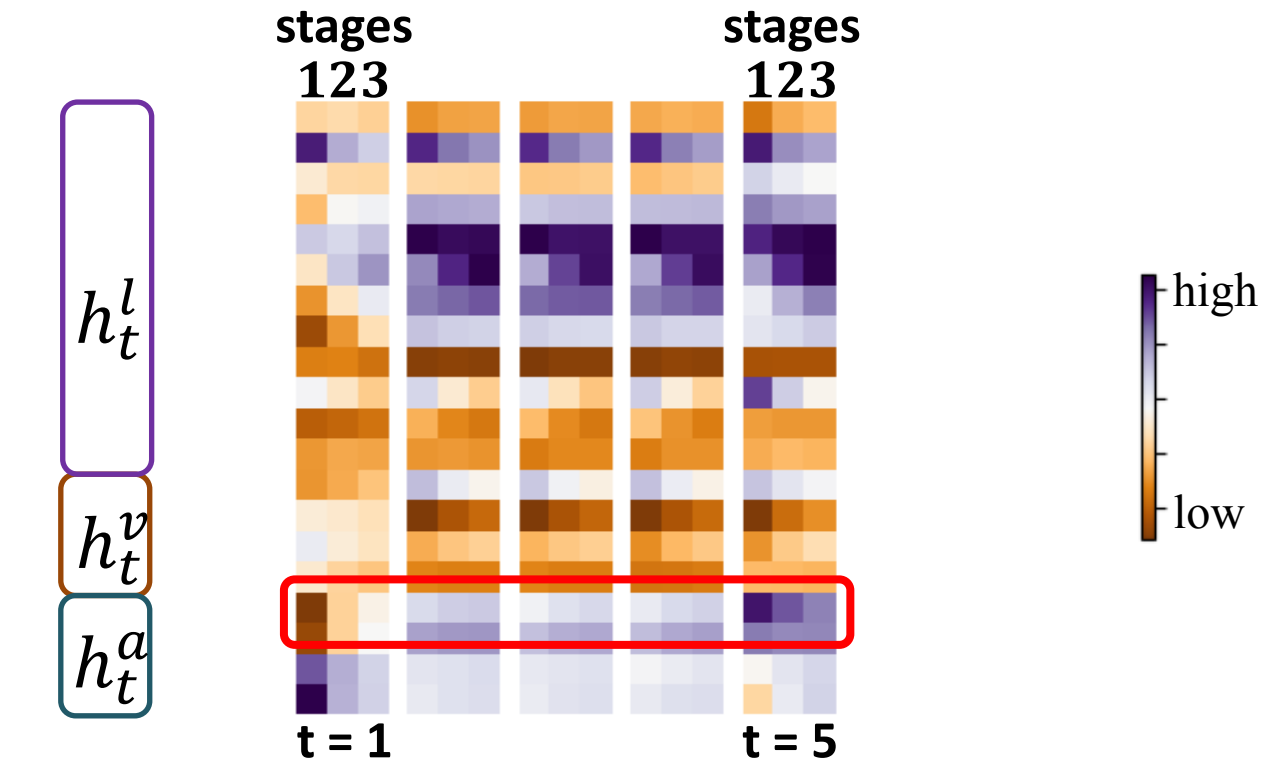
Interpretable Fusion



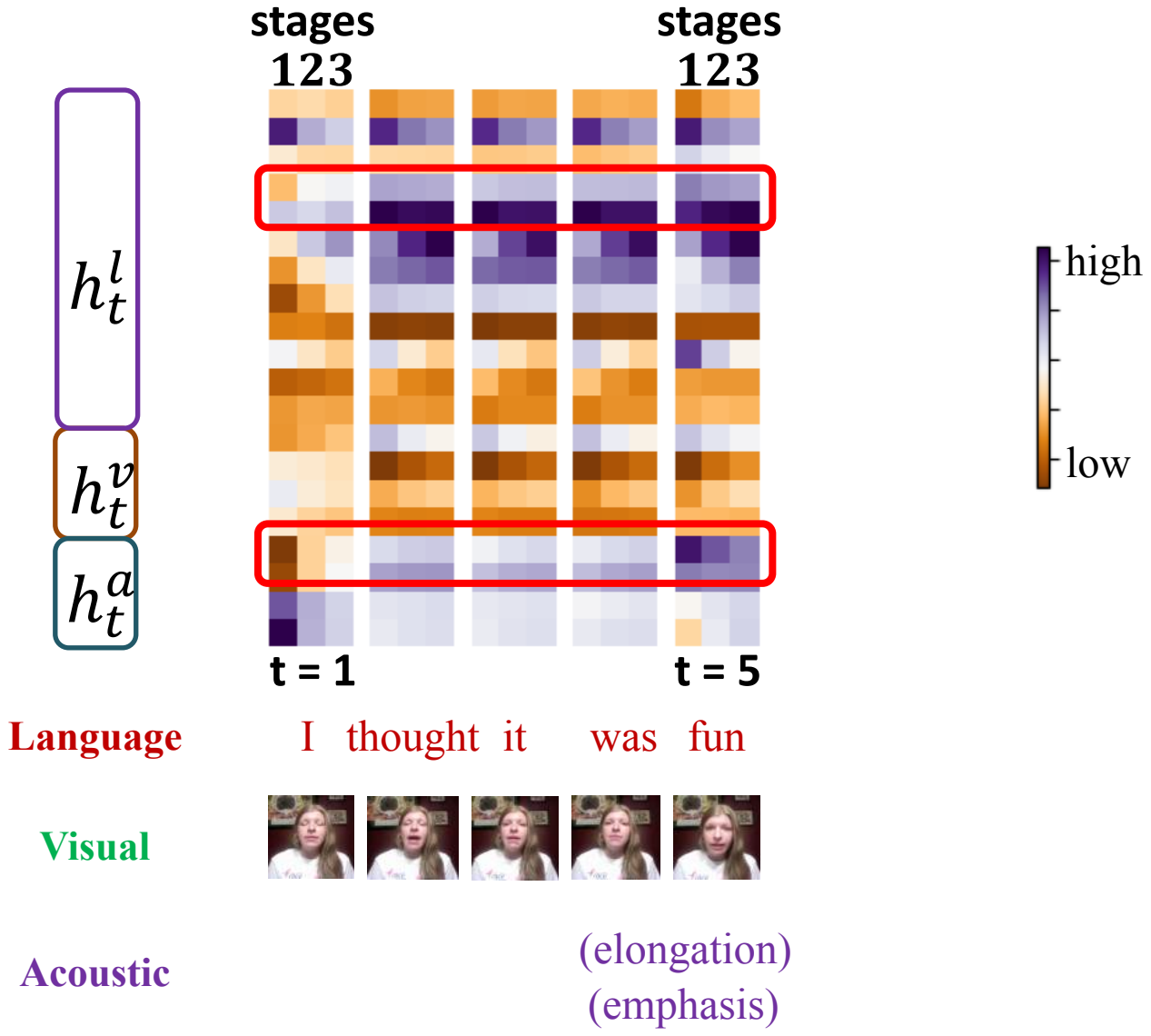
Across Stages



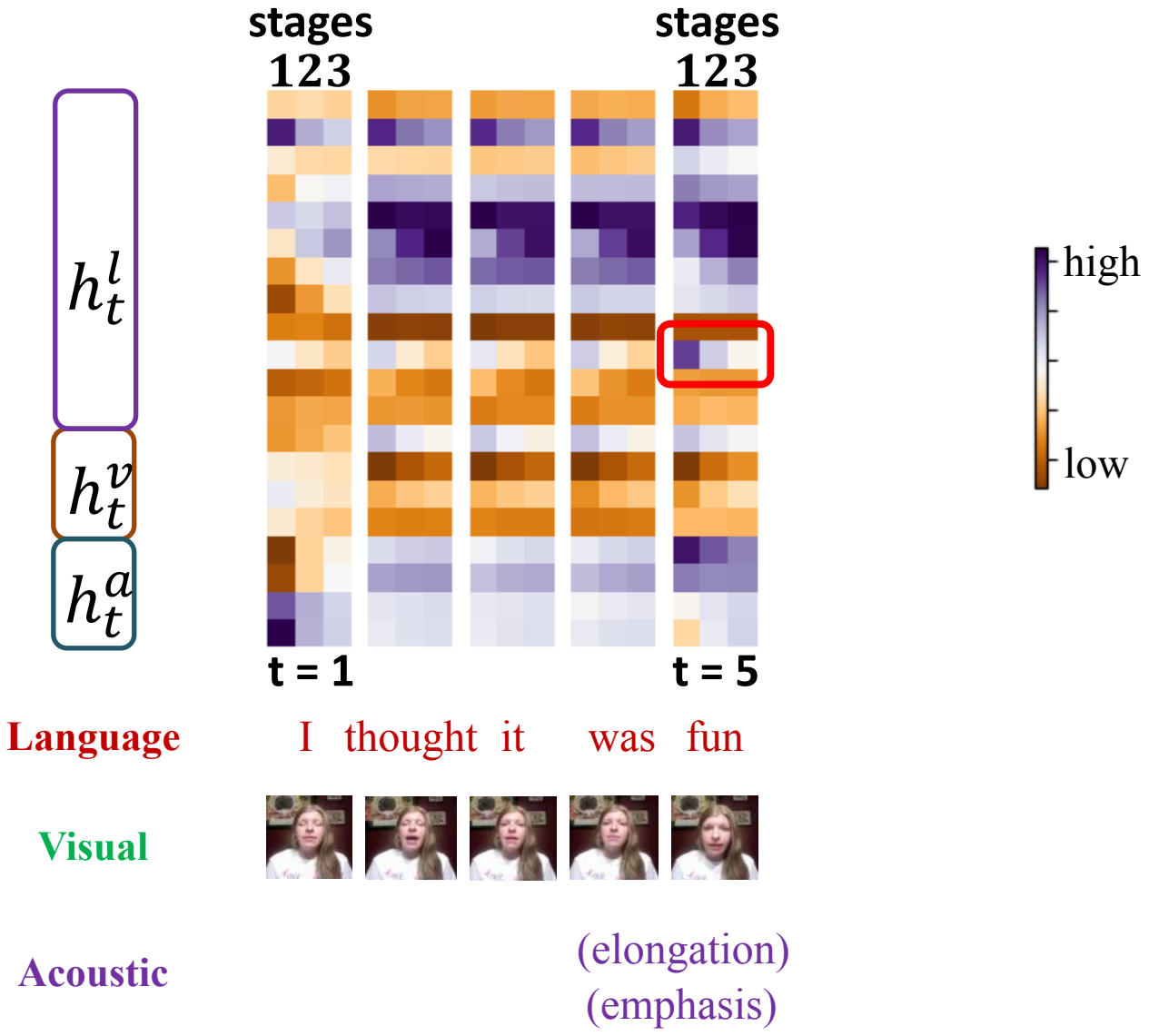
Across Time


stages
123
stages
123
h_t^l
h_t^v
h_t^a
t = 1
t = 5
high
low
Language
I thought it was fun
Visual
Acoustic
(elongation)
(emphasis)

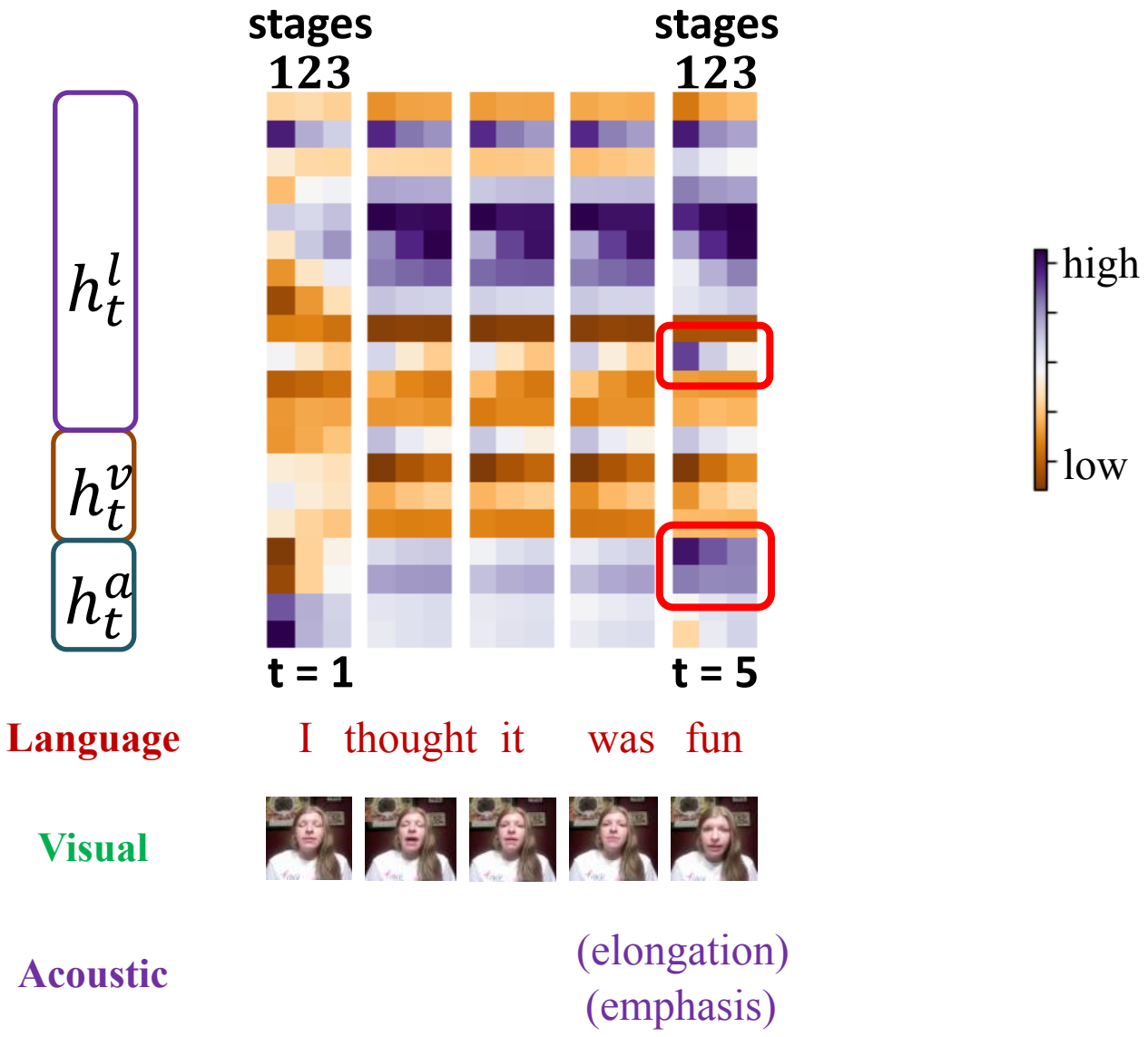
Multimodal Priors



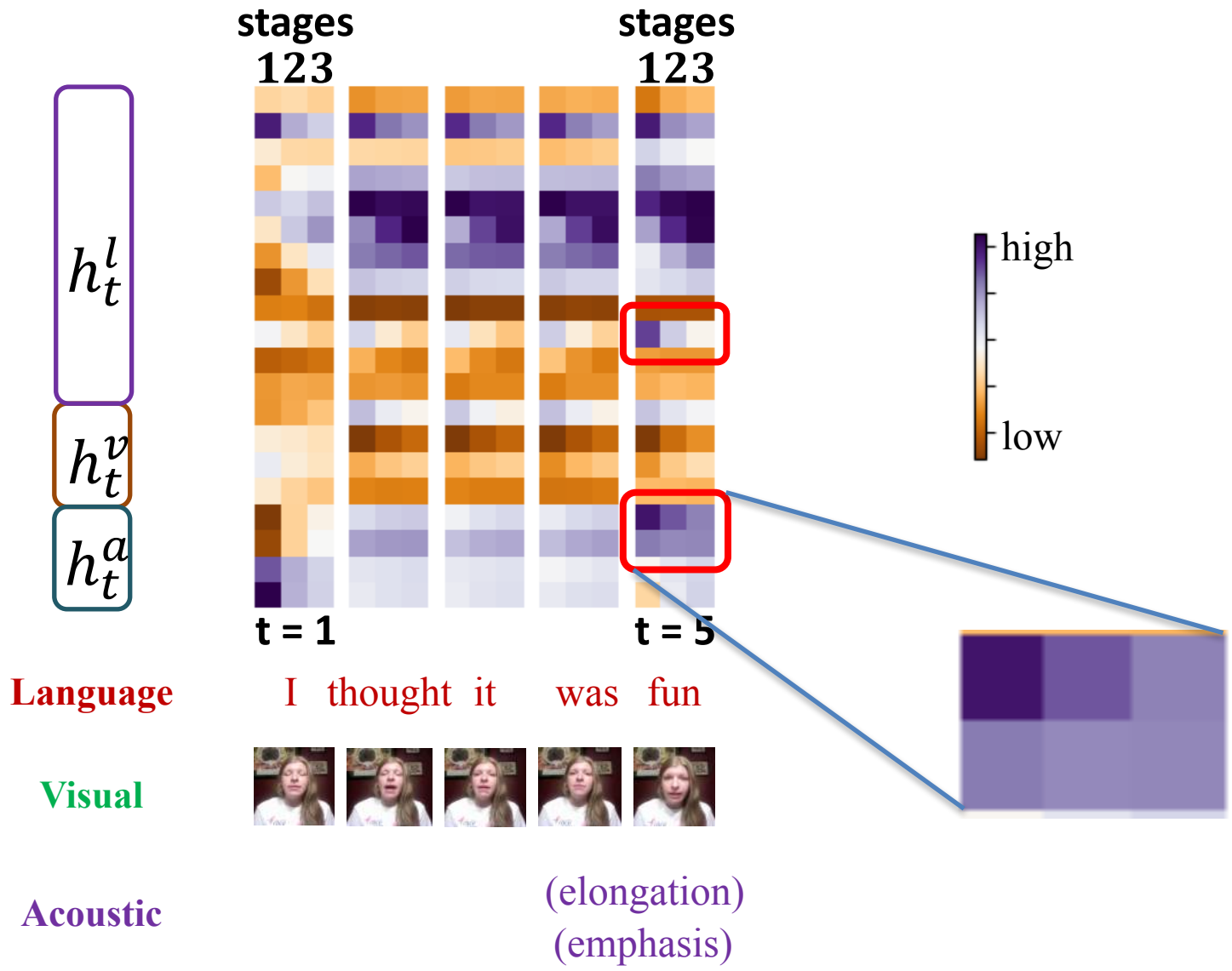
Synchronized Interactions



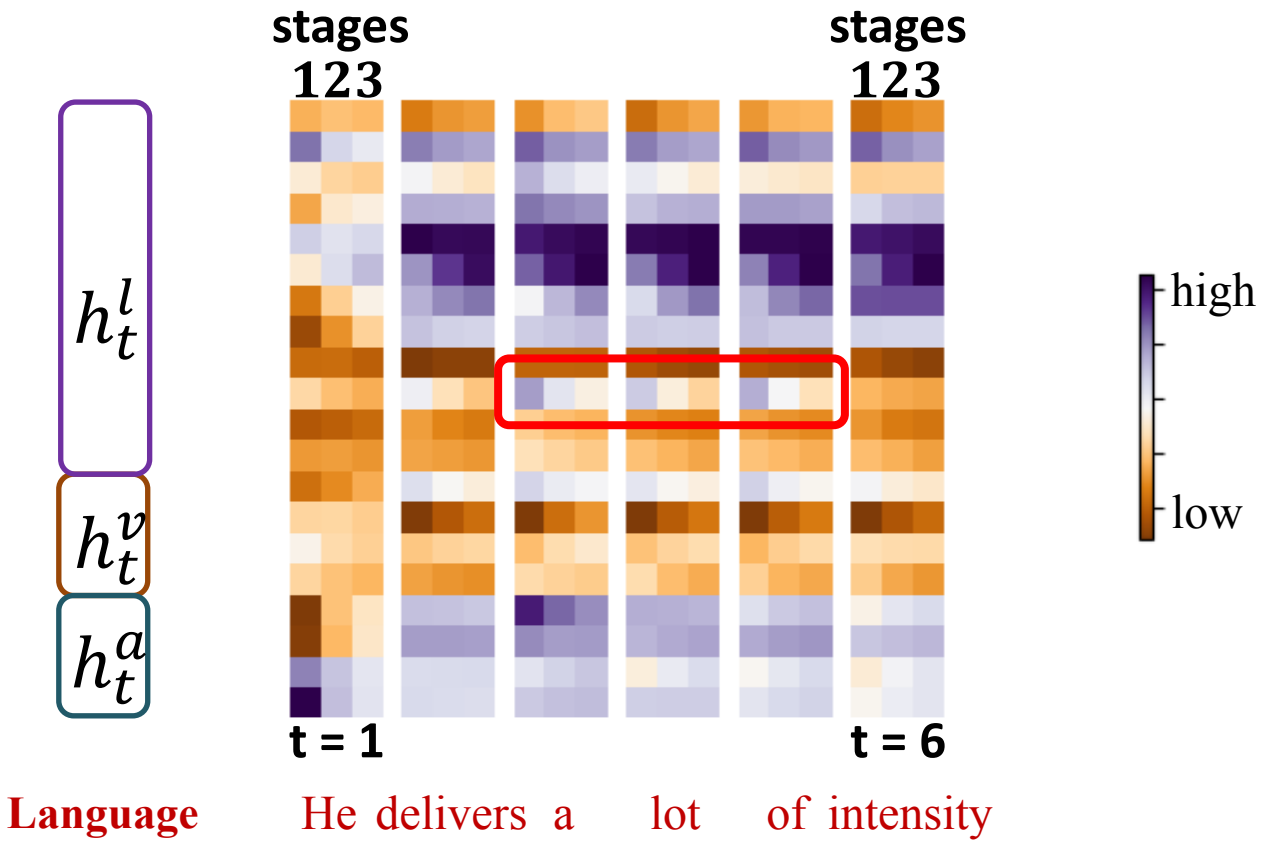
Synchronized Interactions



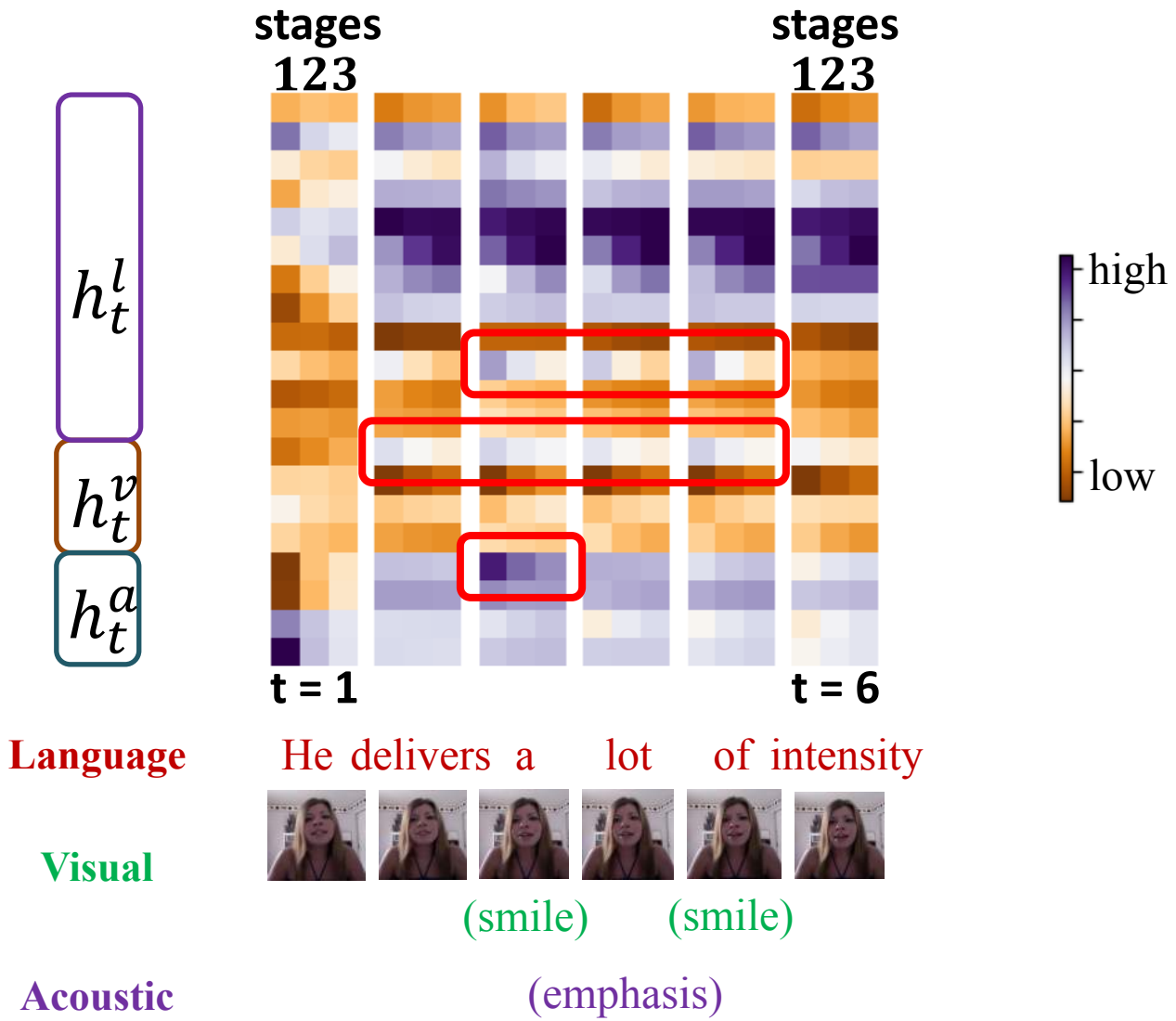
Synchronized Interactions



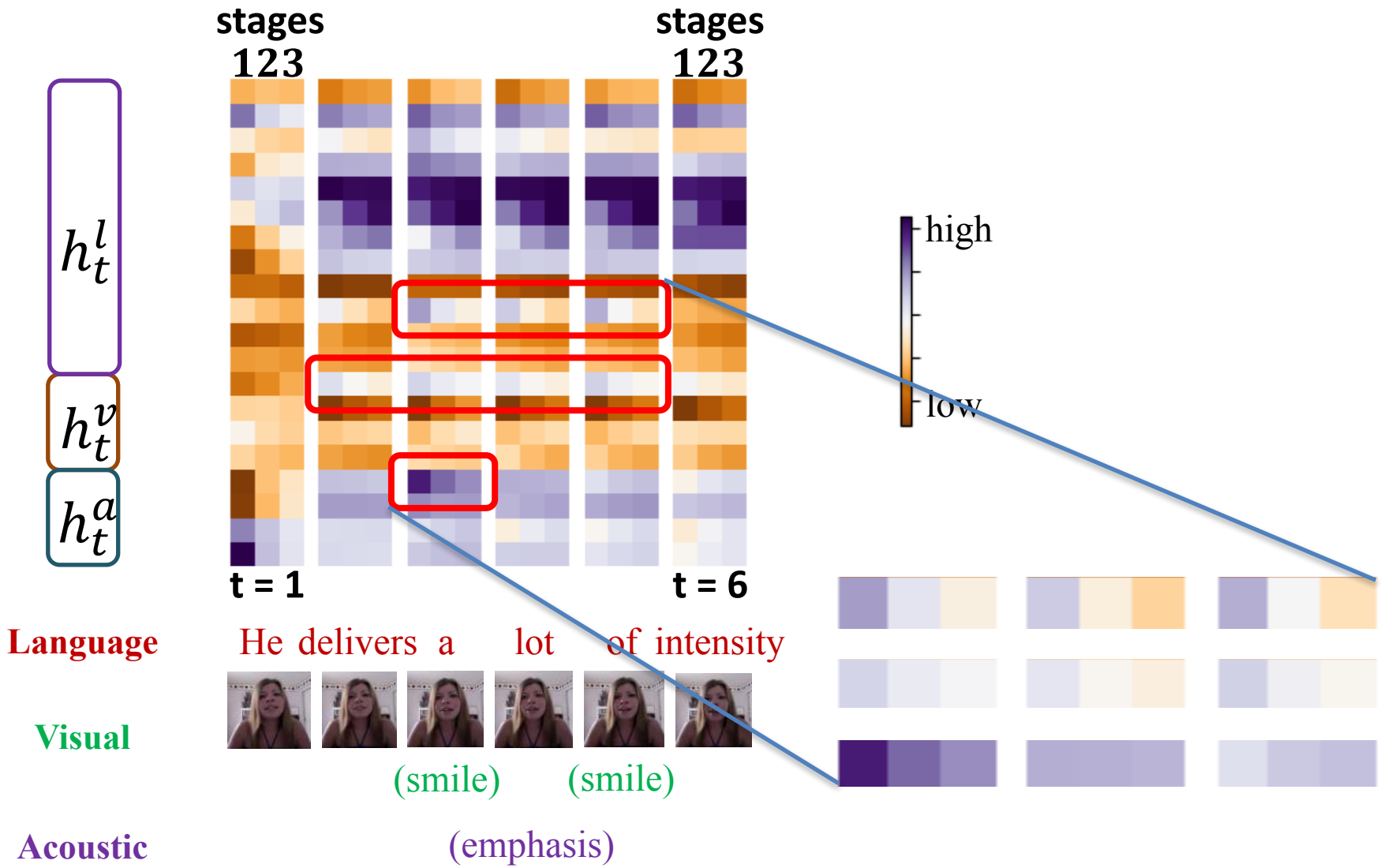
Asynchronous Trimodal Interactions



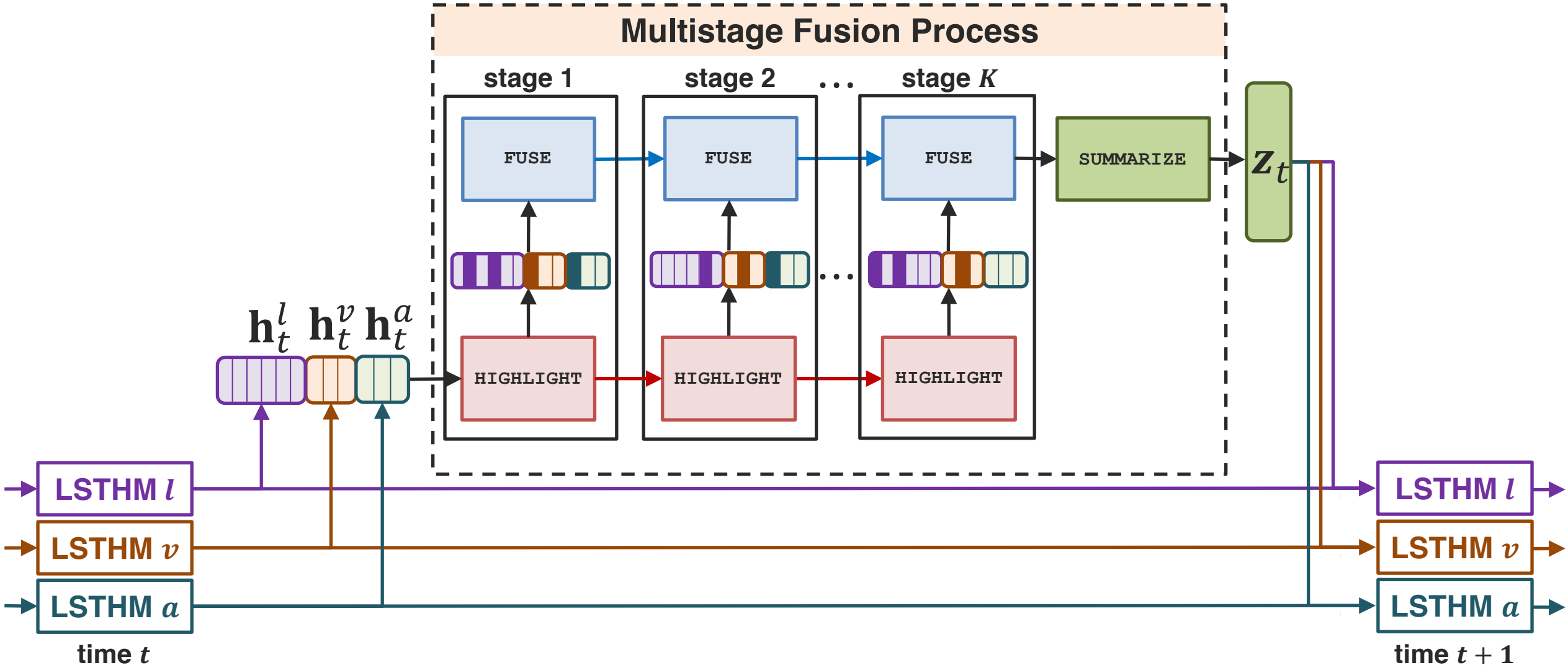
Asynchronous Trimodal Interactions



Asynchronous Trimodal Interactions



Recurrent Multistage Fusion Network



The End!

Website: www.cs.cmu.edu/~pliang

Email: pliang@cs.cmu.edu

Twitter: @pliang279