



Language
Technologies
Institute

Carnegie
Mellon
University

Multimodal Language Analysis with Recurrent Multistage Fusion

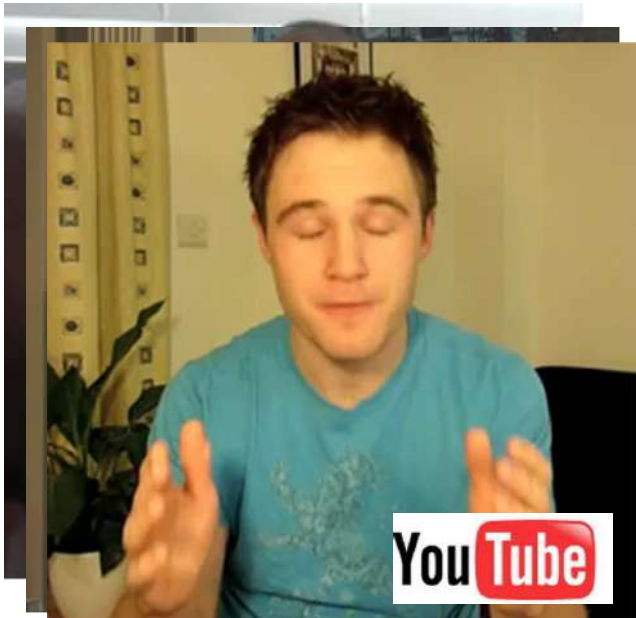
Presenter: Paul Pu Liang

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

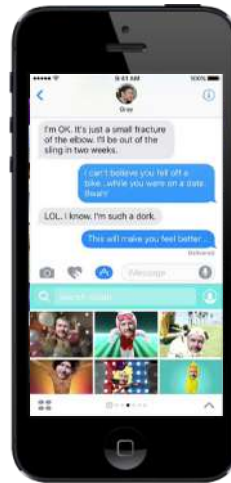


Progress of Artificial Intelligence

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

Acoustic

- Prosody
- Vocal expressions

Visual

- Gestures
- Body language
- Eye contact
- Facial expressions



Sentiment

- Positive
- Negative

Emotion

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

Personality

- Confidence
- Persuasion
- Passion

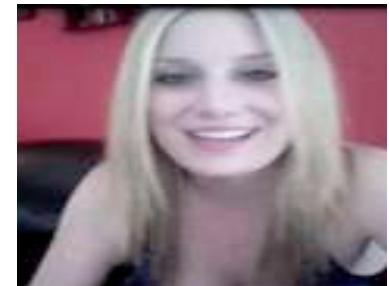
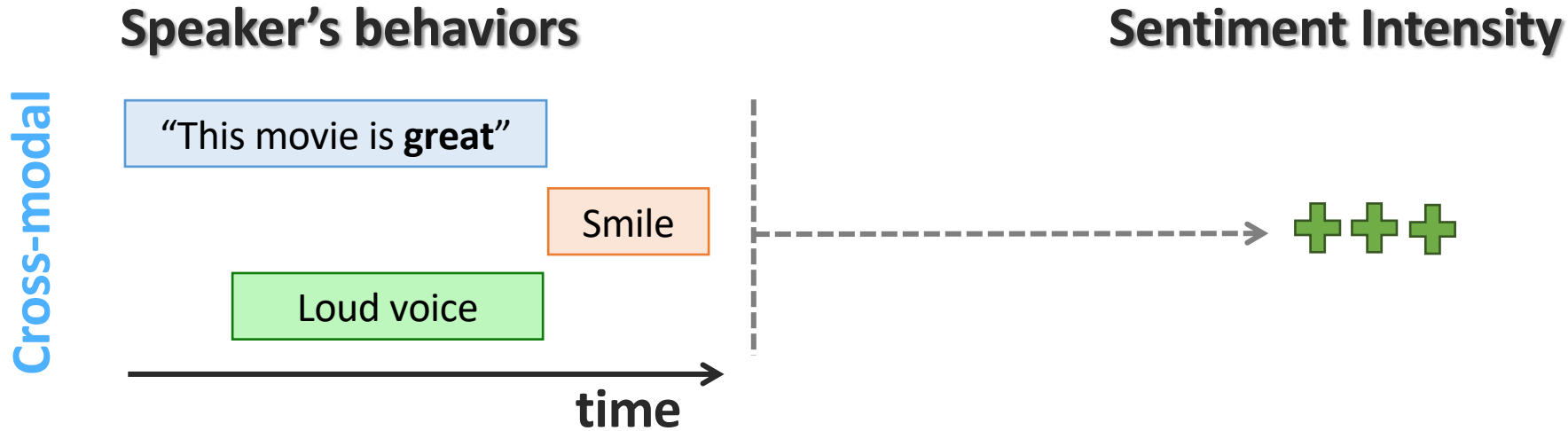
Challenge 1: Intra-modal Interactions

a) Temporal sequences



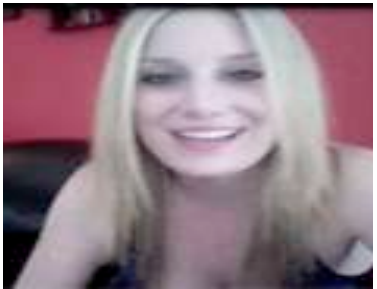
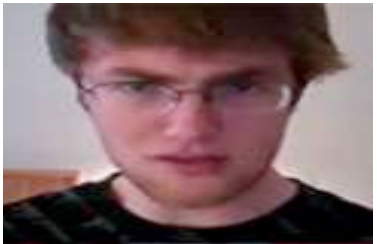
Challenge 2: Cross-modal Interactions

- 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

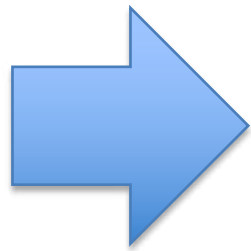
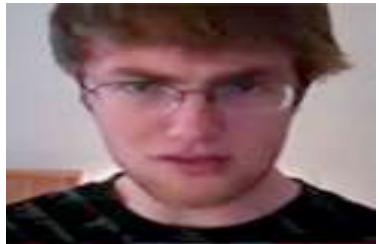


positive reaction
positive words



Multistage Aggregation in Humans

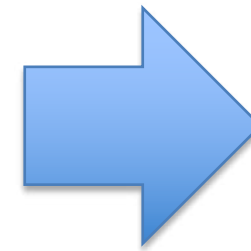
(Parsini et al. 2015,
Taylor et al. 2017)



wide smile
loud voice

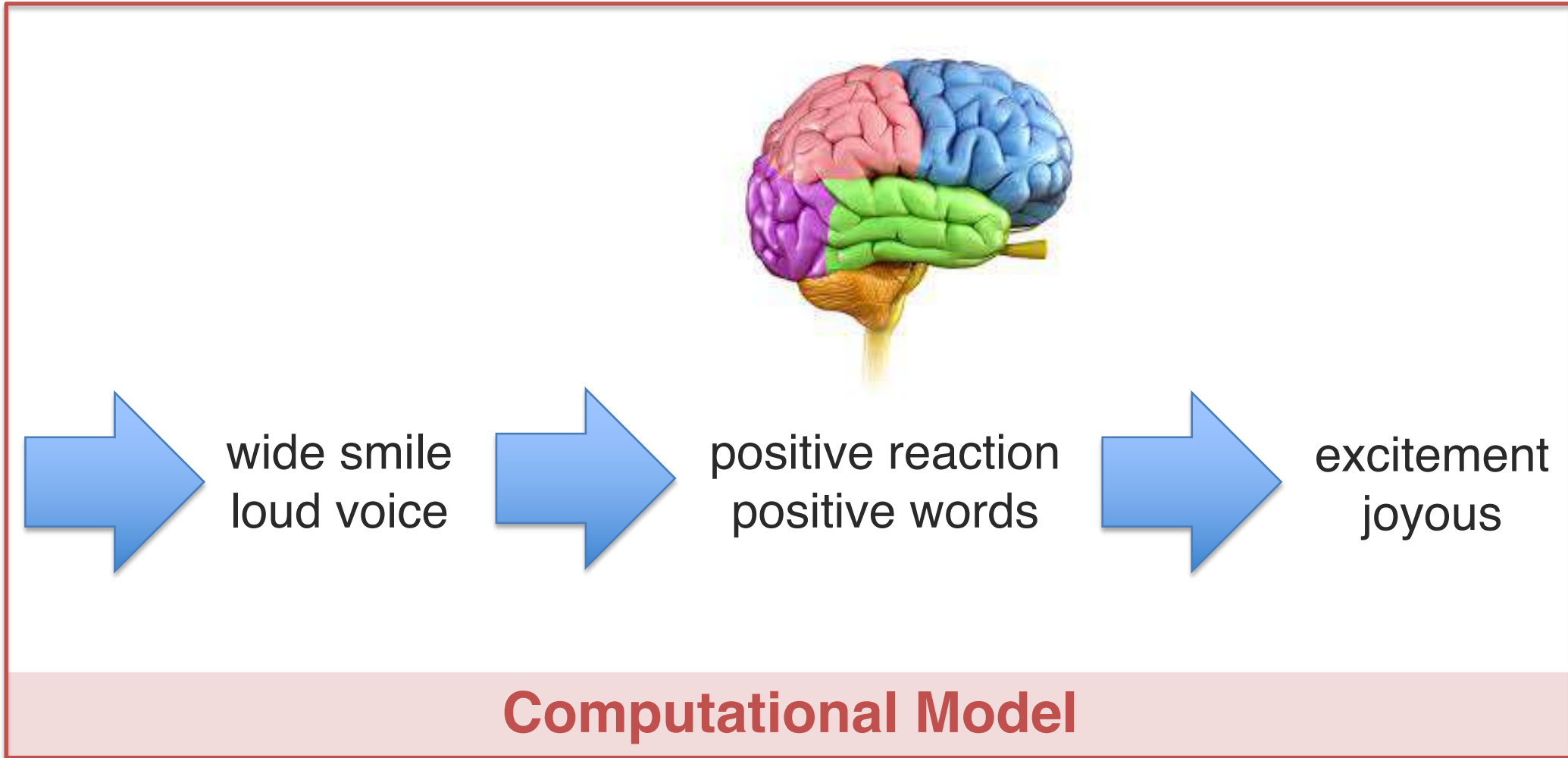
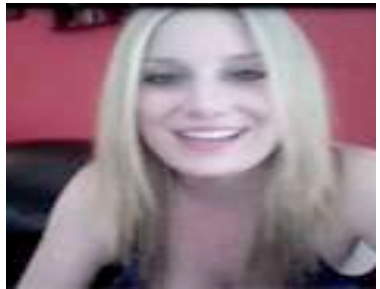
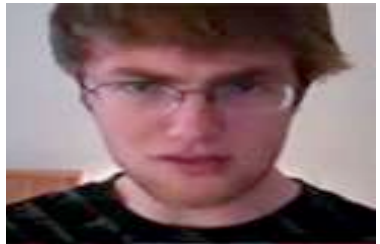


positive reaction
positive words

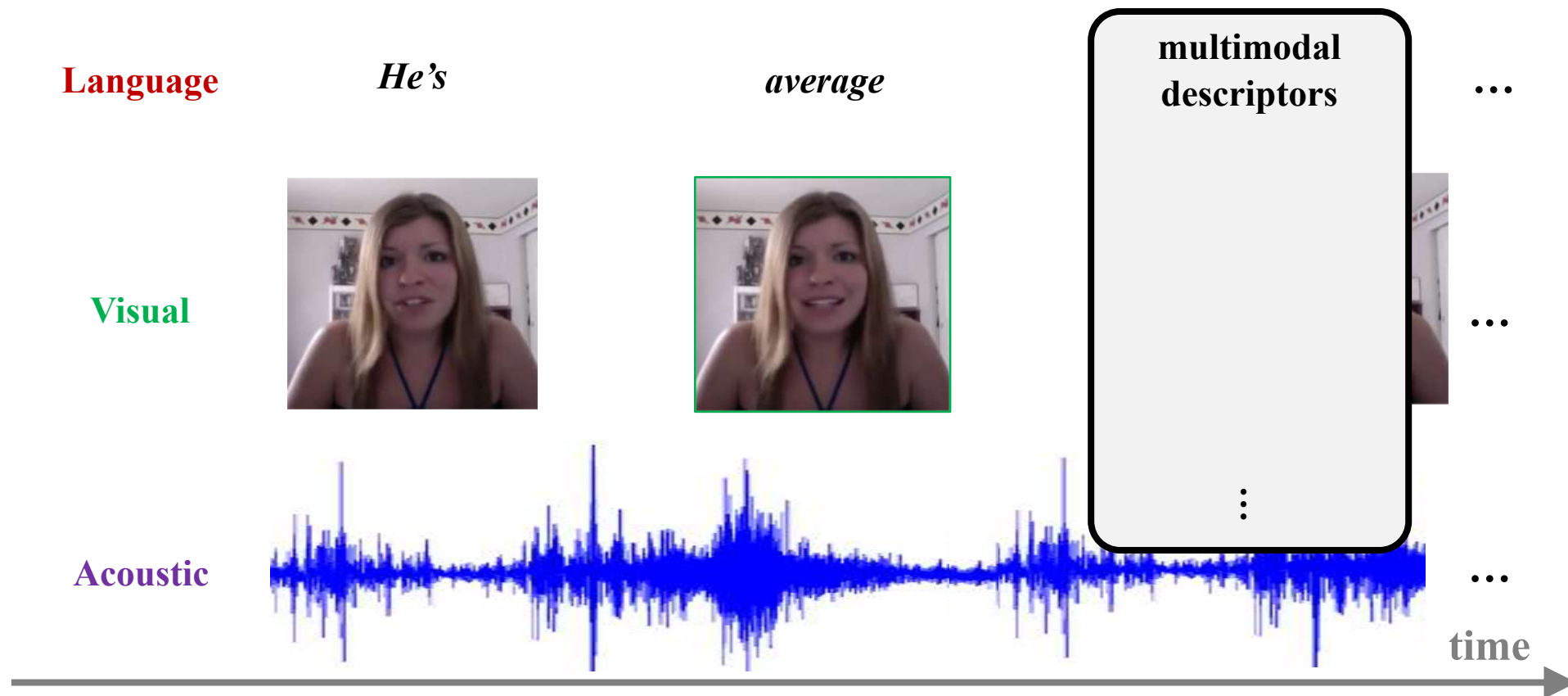


excitement
joyous

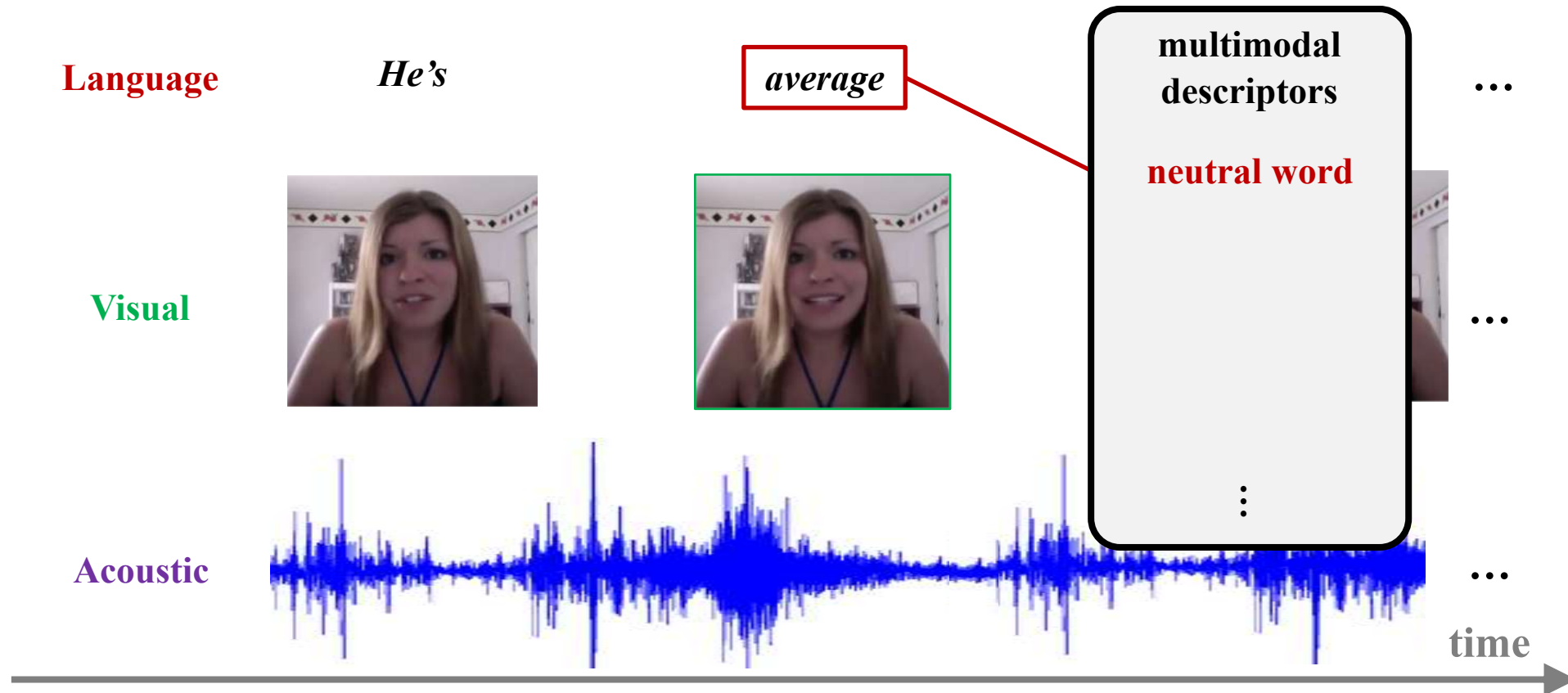
Computational Model for Multistage Fusion



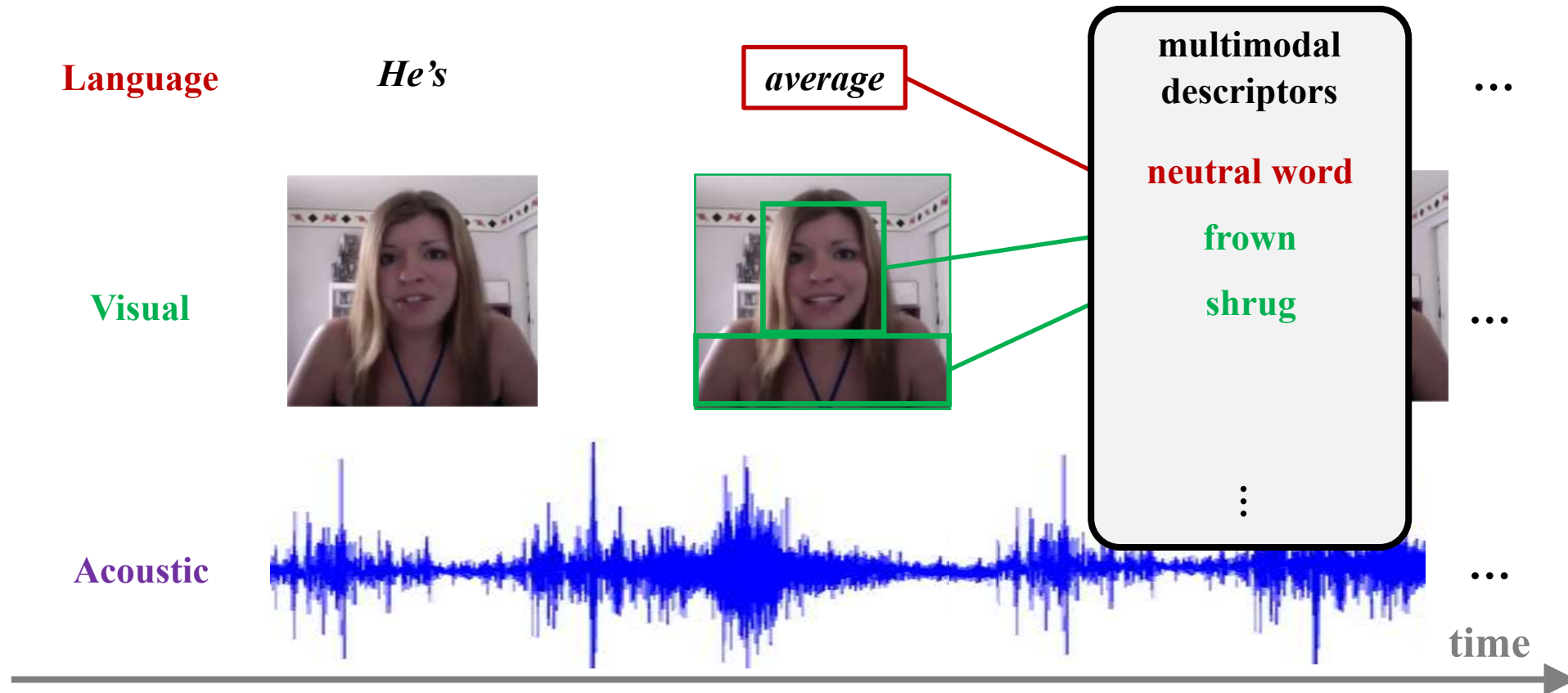
Multimodal Descriptors



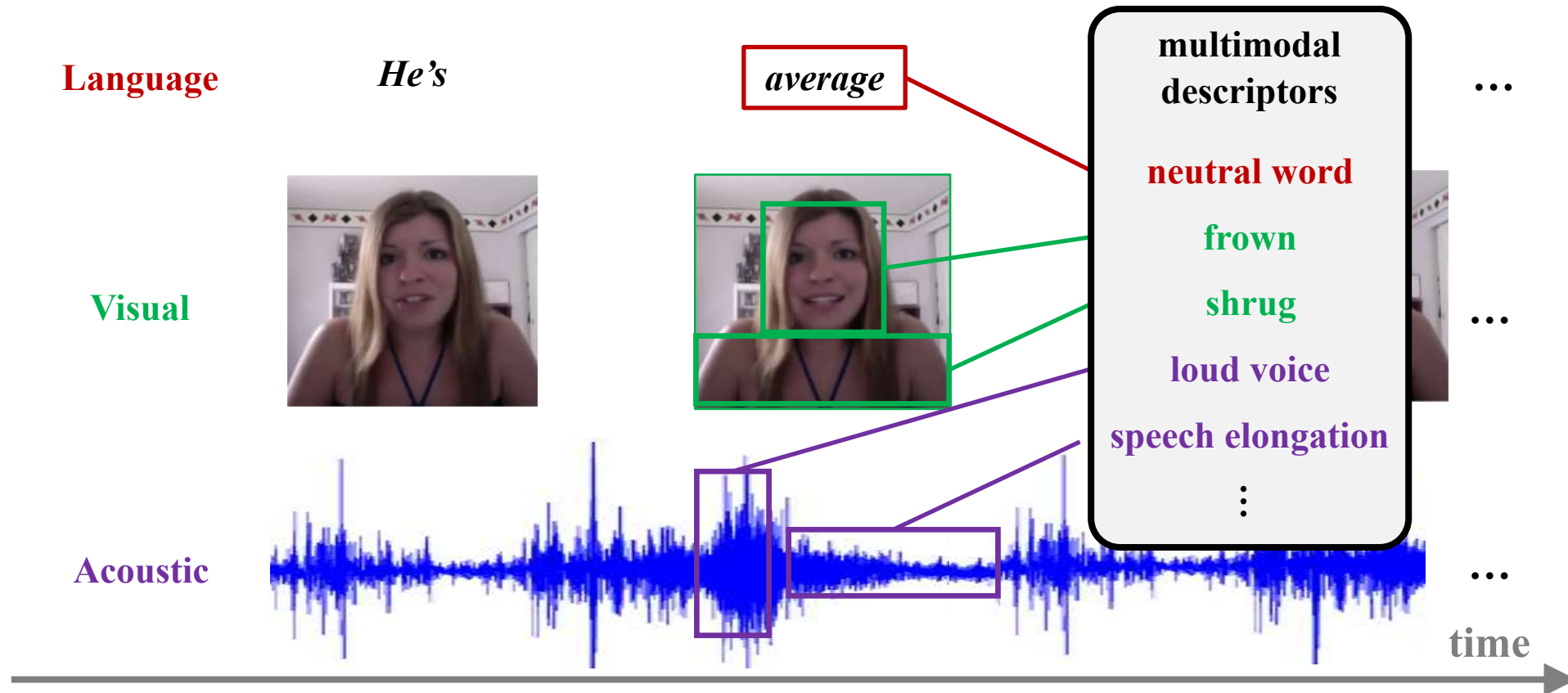
Language Descriptors



Visual Descriptors



Acoustic Descriptors



Multistage Fusion

neutral word

frown

shrug

loud voice

speech elongation

⋮

Multistage Fusion

stage 1

HIGHLIGHT

neutral word

frown

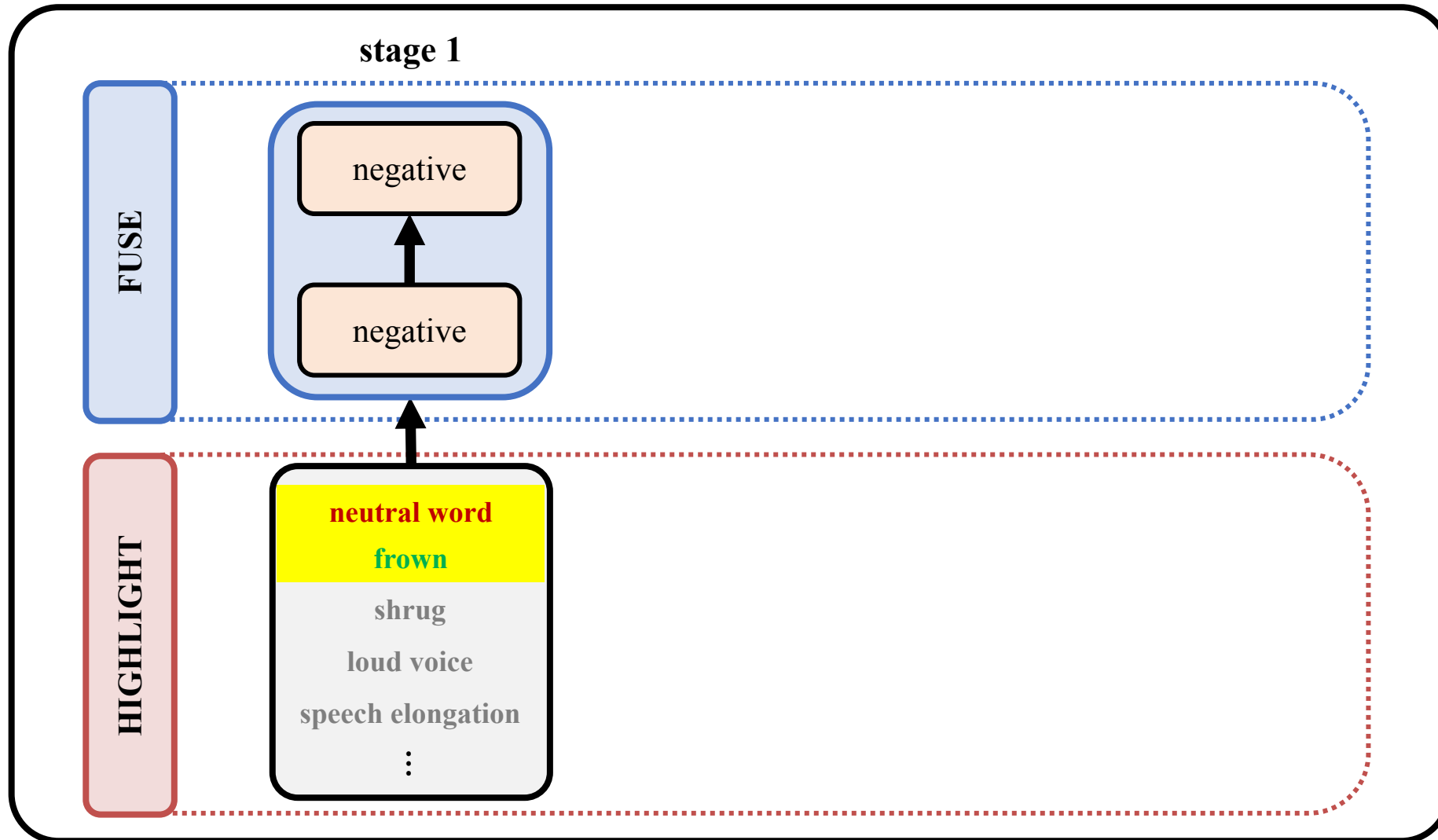
shrug

loud voice

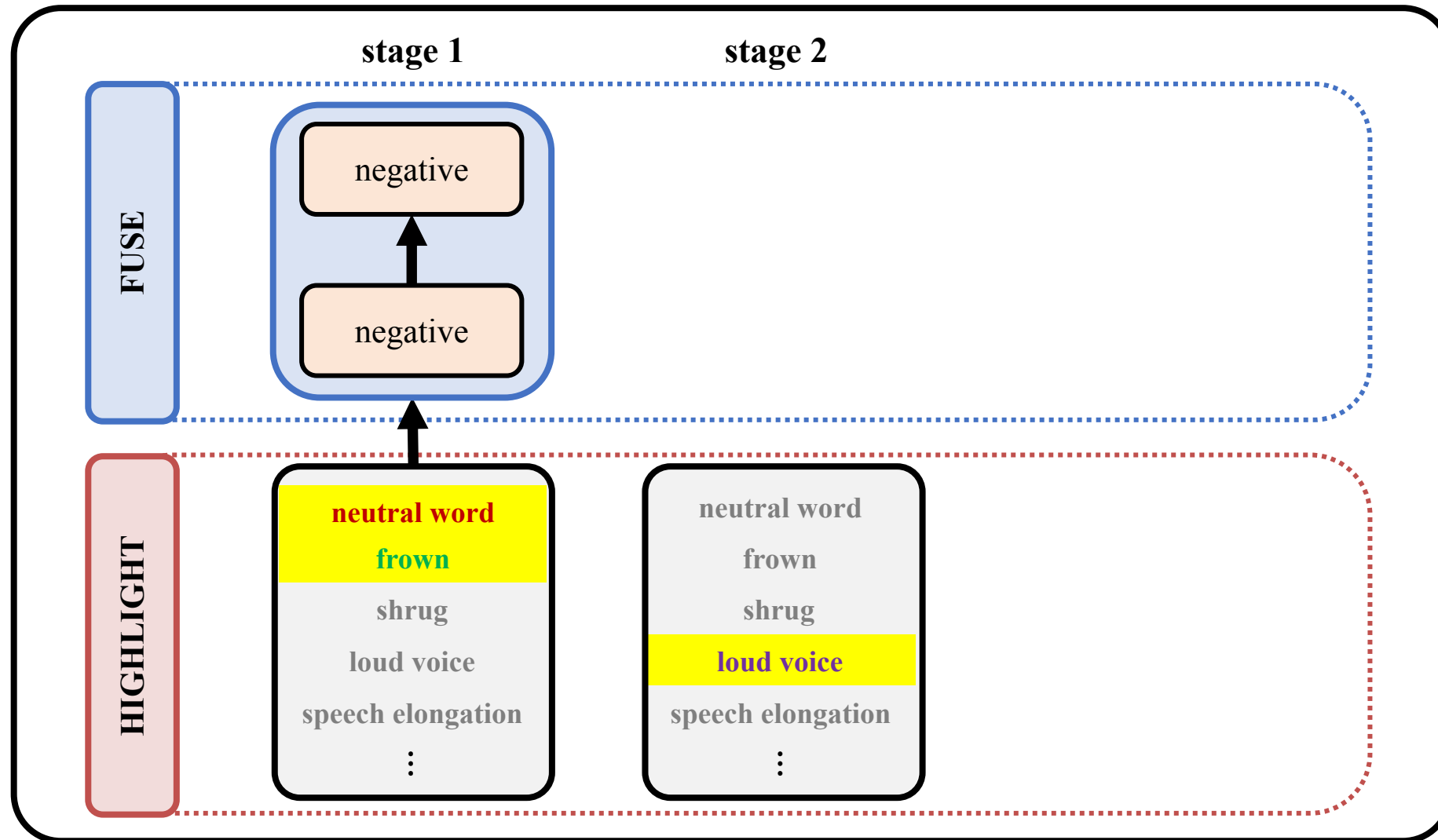
speech elongation

⋮

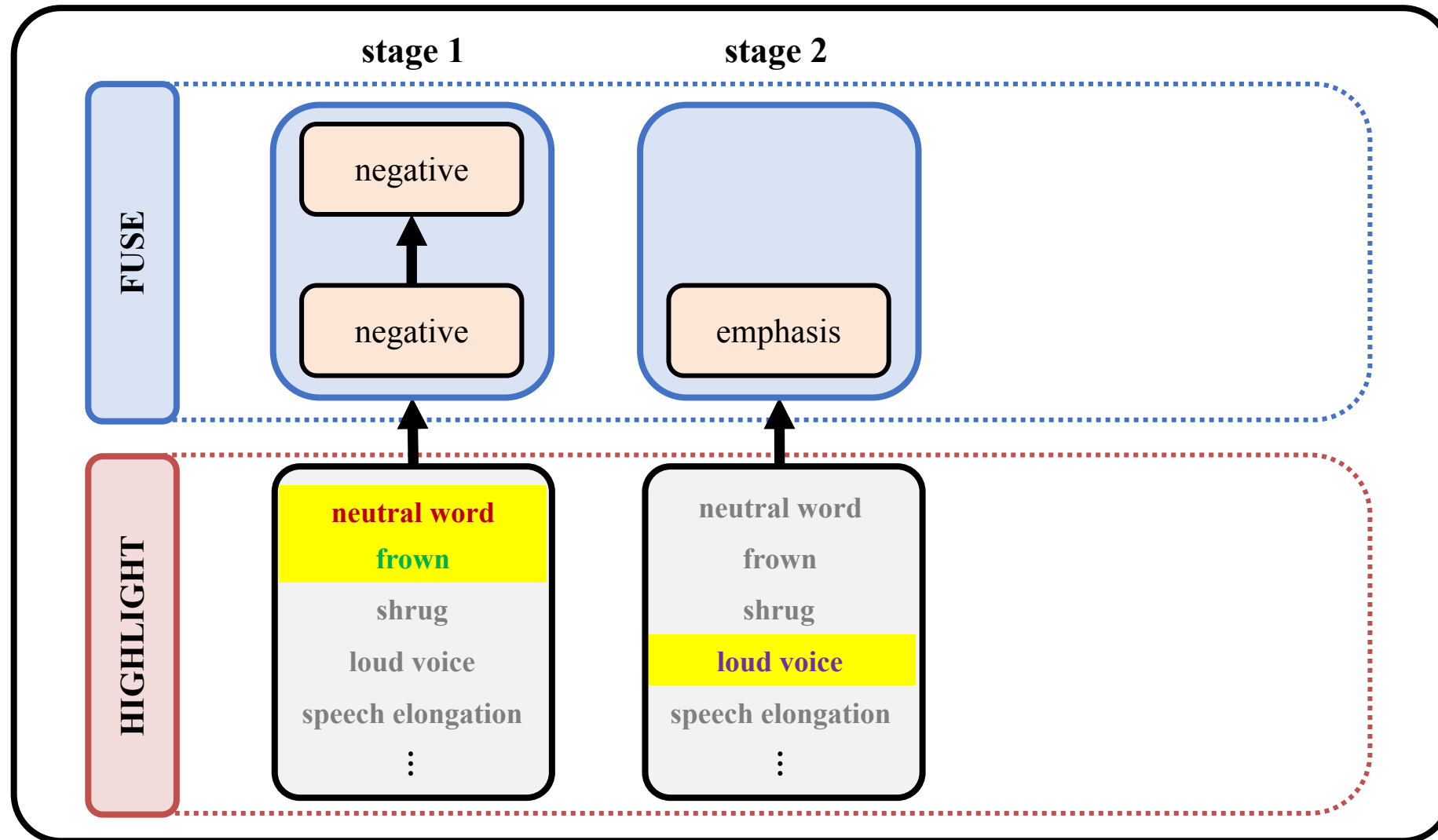
Multistage Fusion



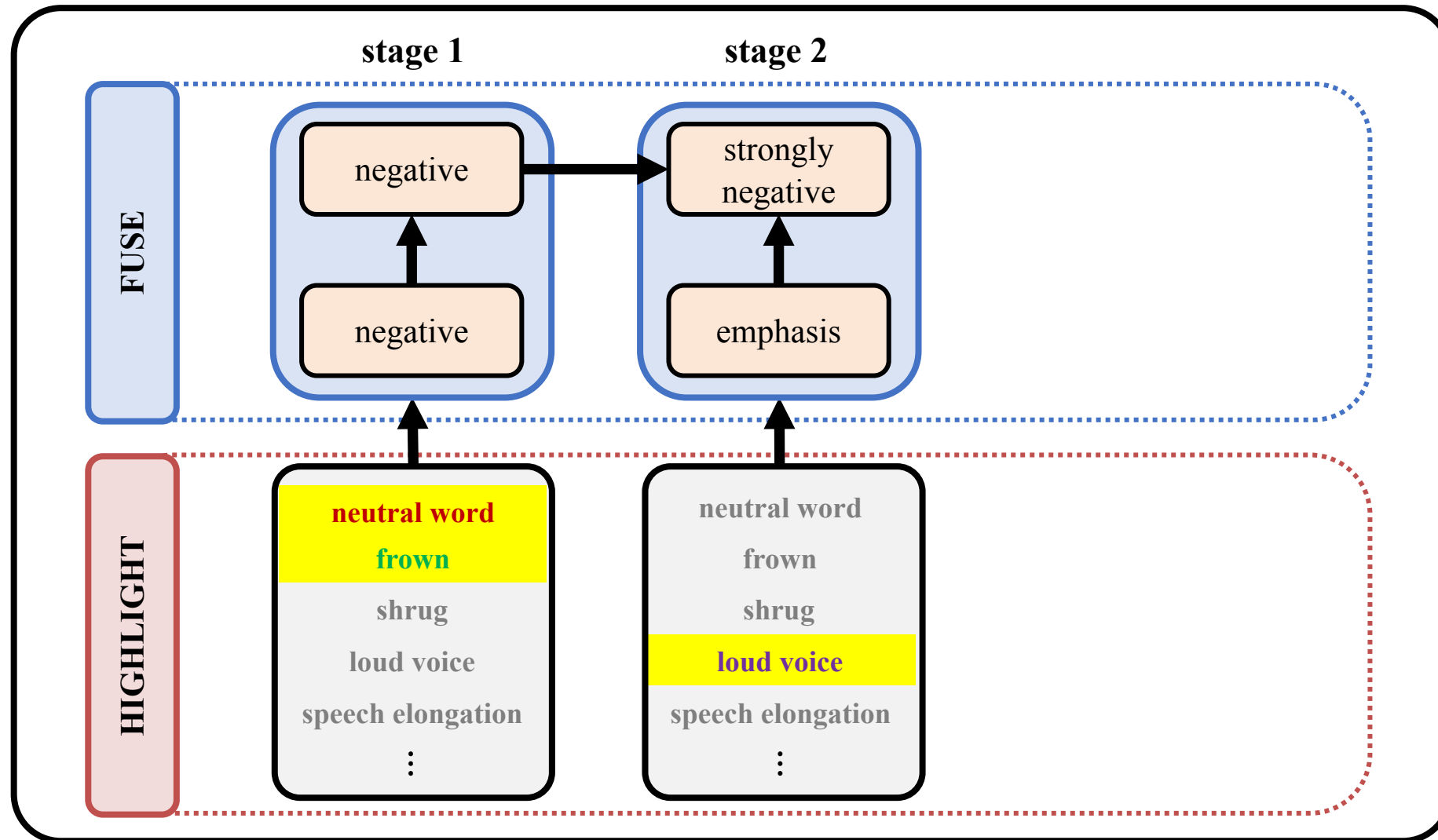
Multistage Fusion



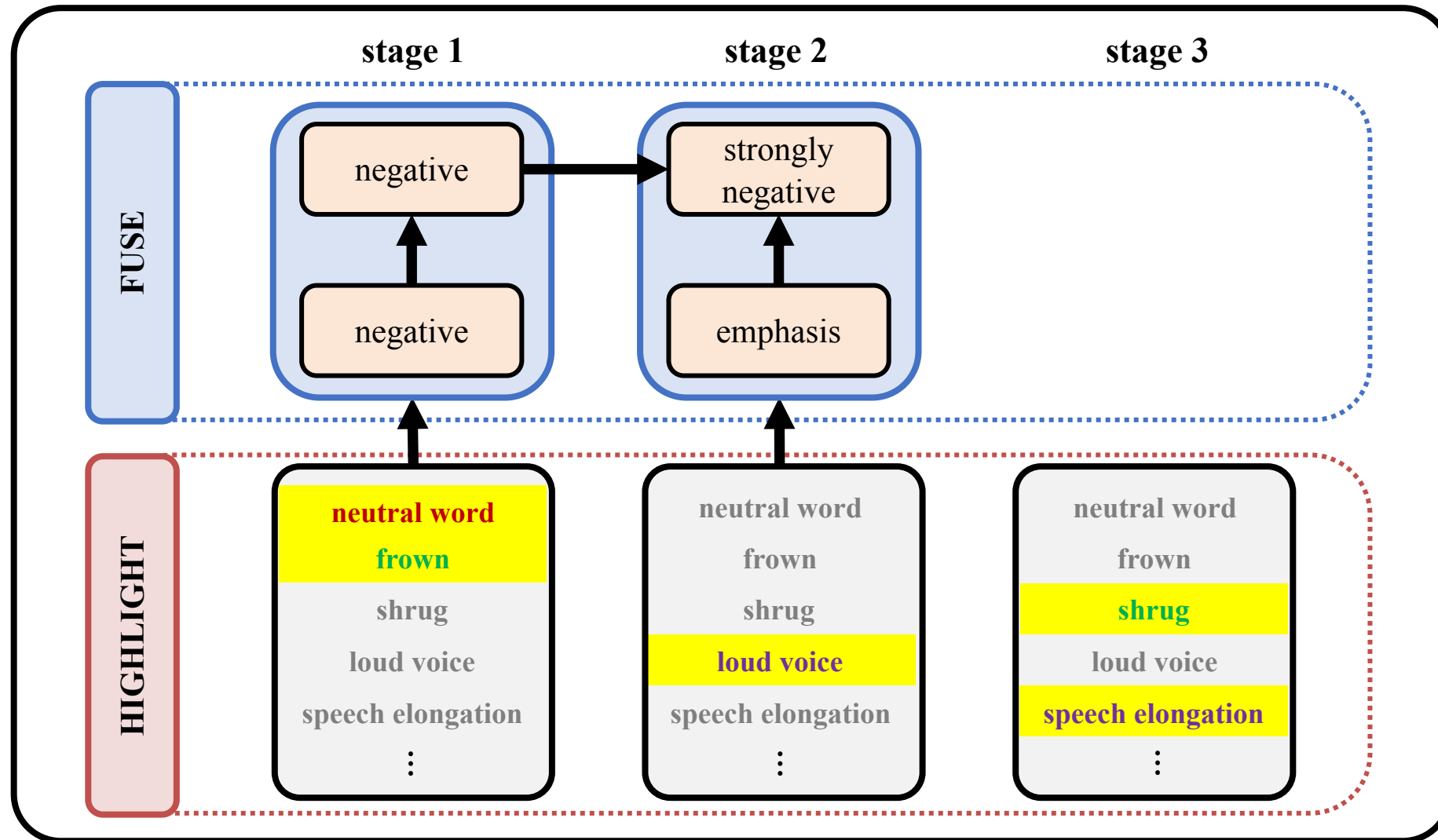
Multistage Fusion



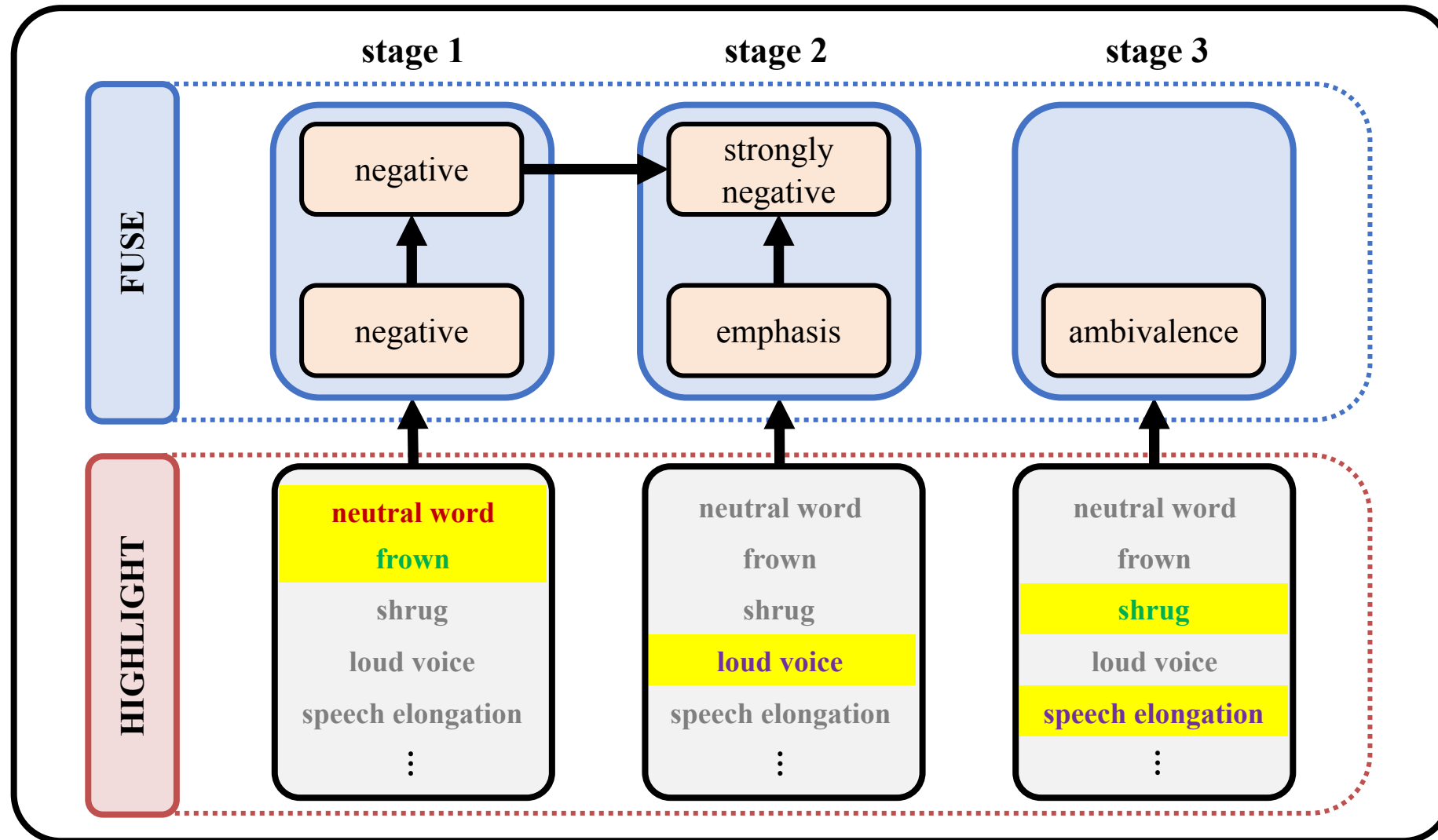
Multistage Fusion



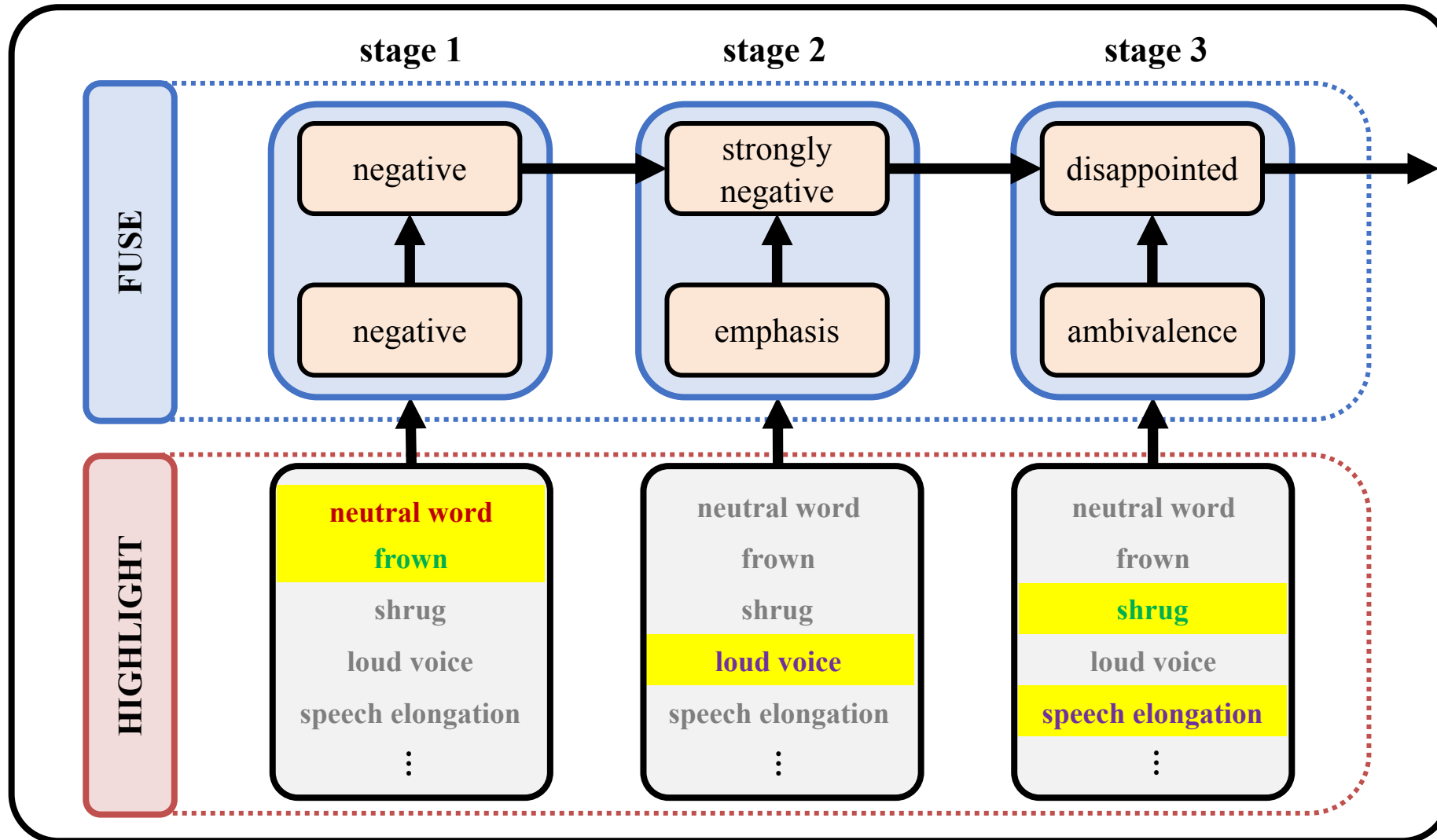
Multistage Fusion



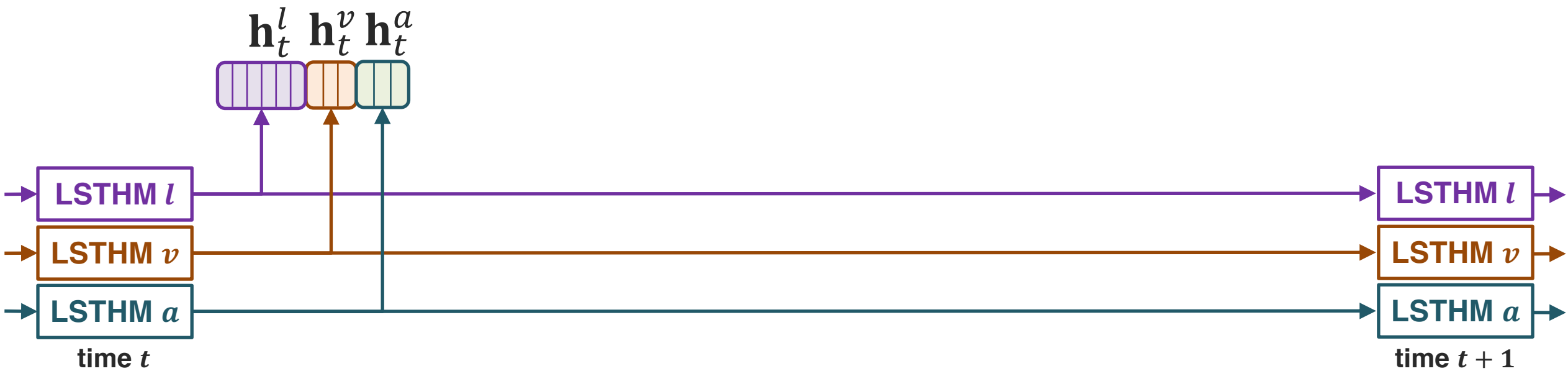
Multistage Fusion



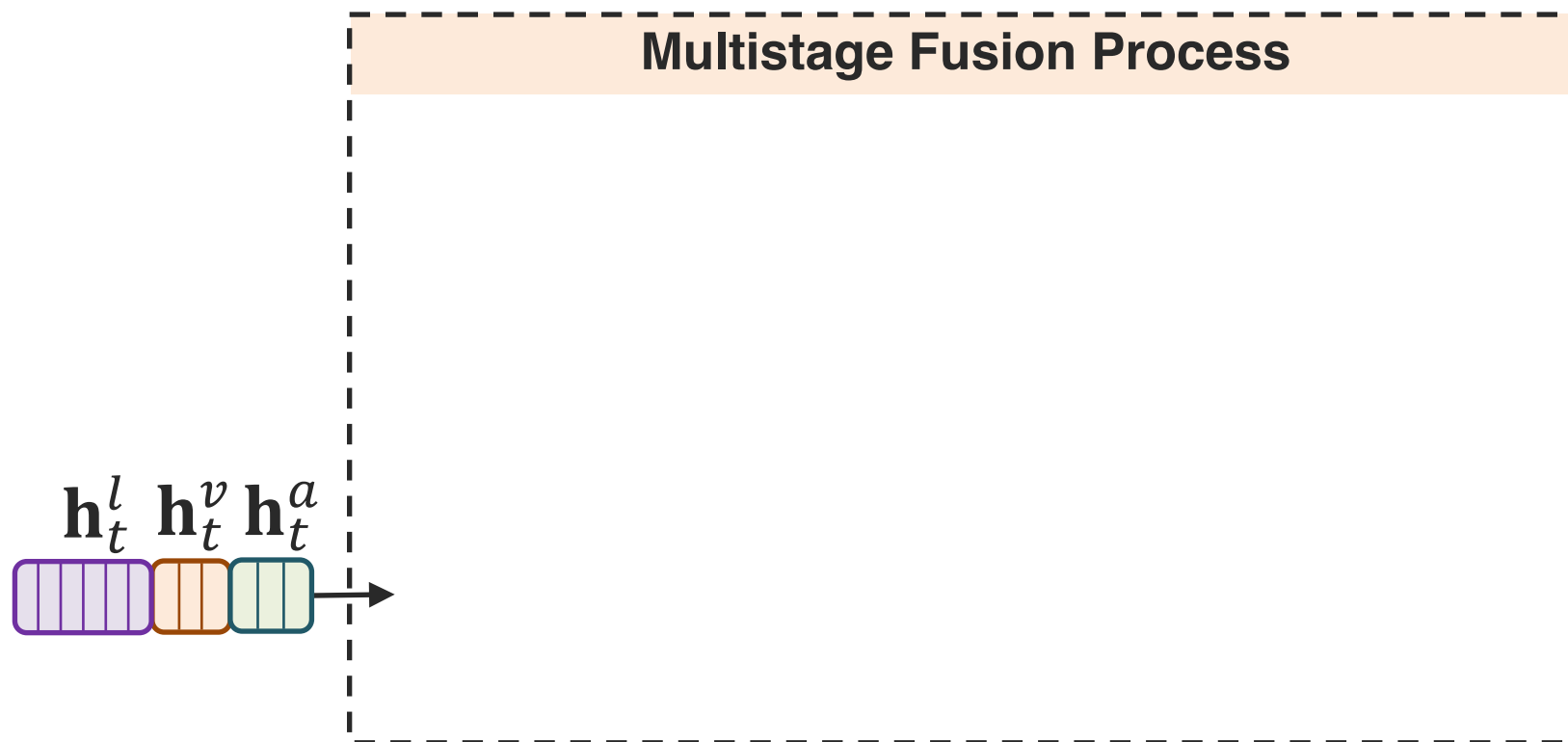
Multistage Fusion



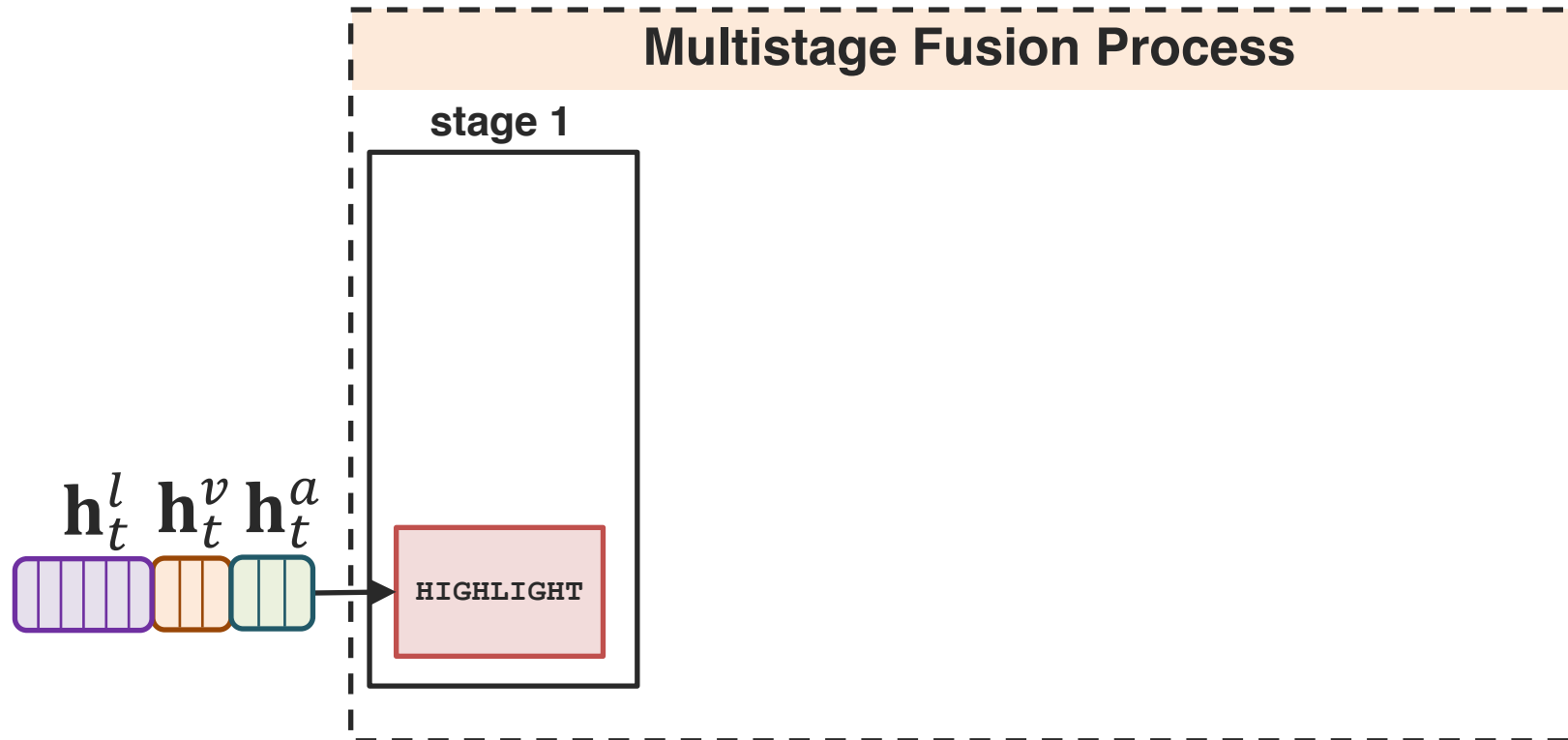
Intra-modal Recurrent Networks



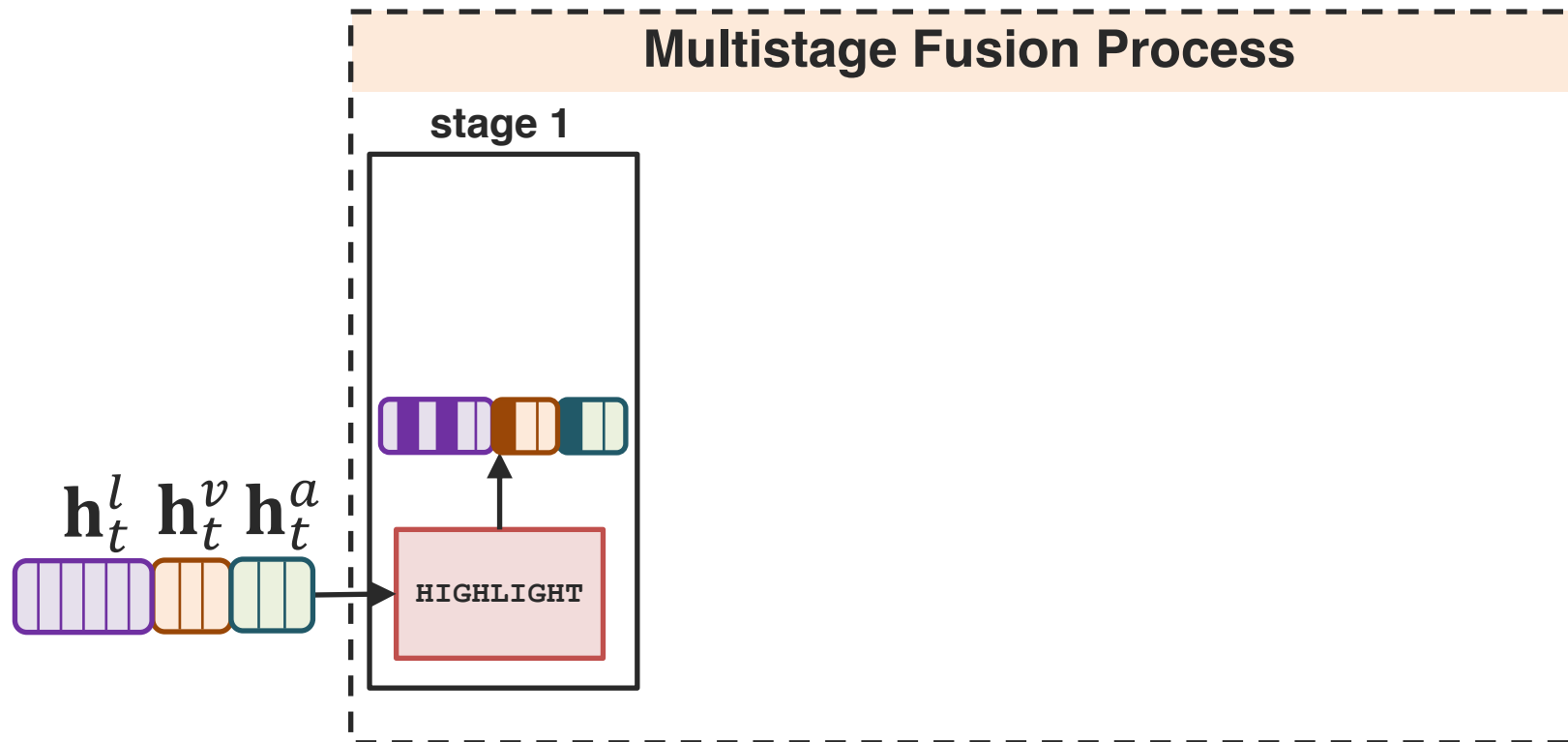
Multistage Fusion Process



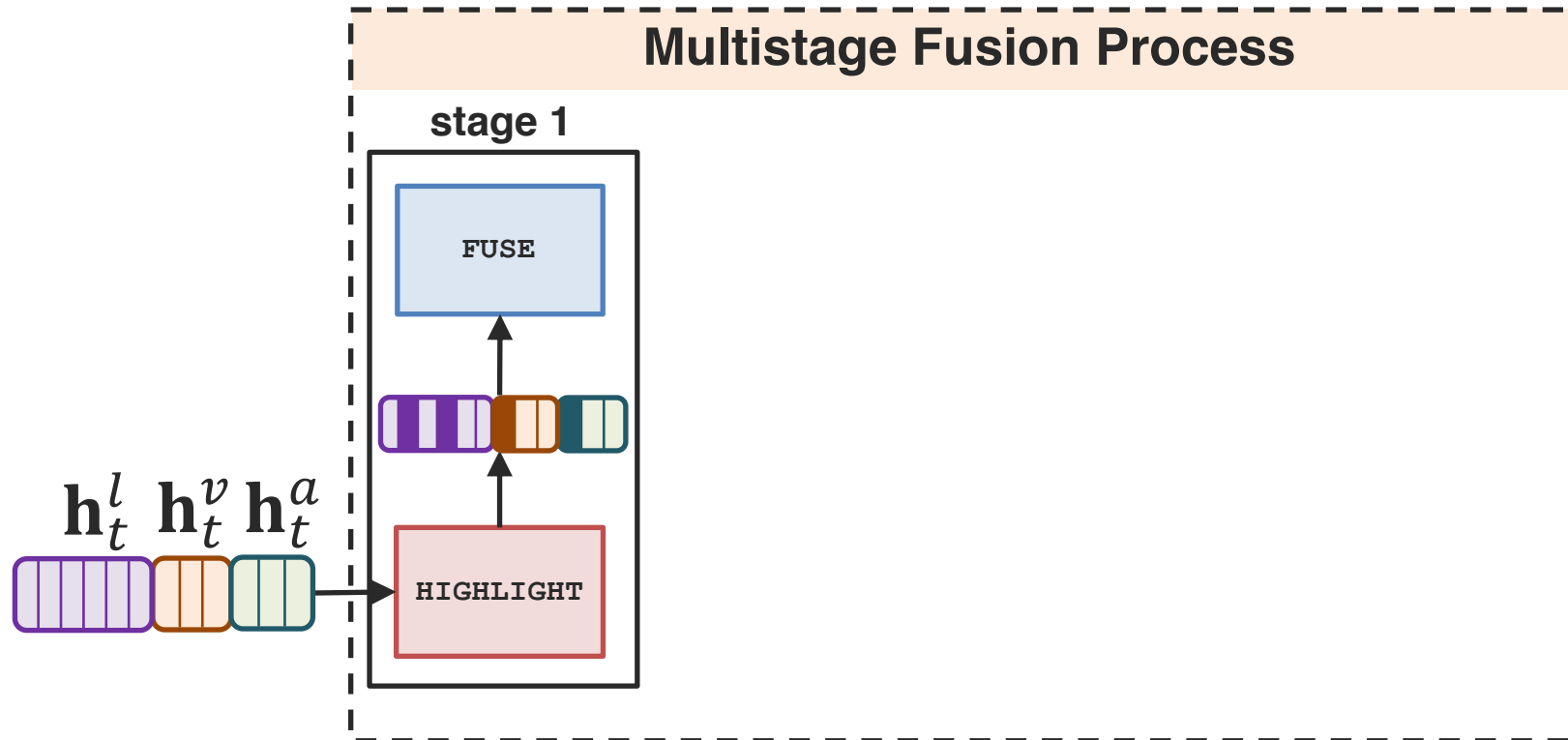
Multistage Fusion Process



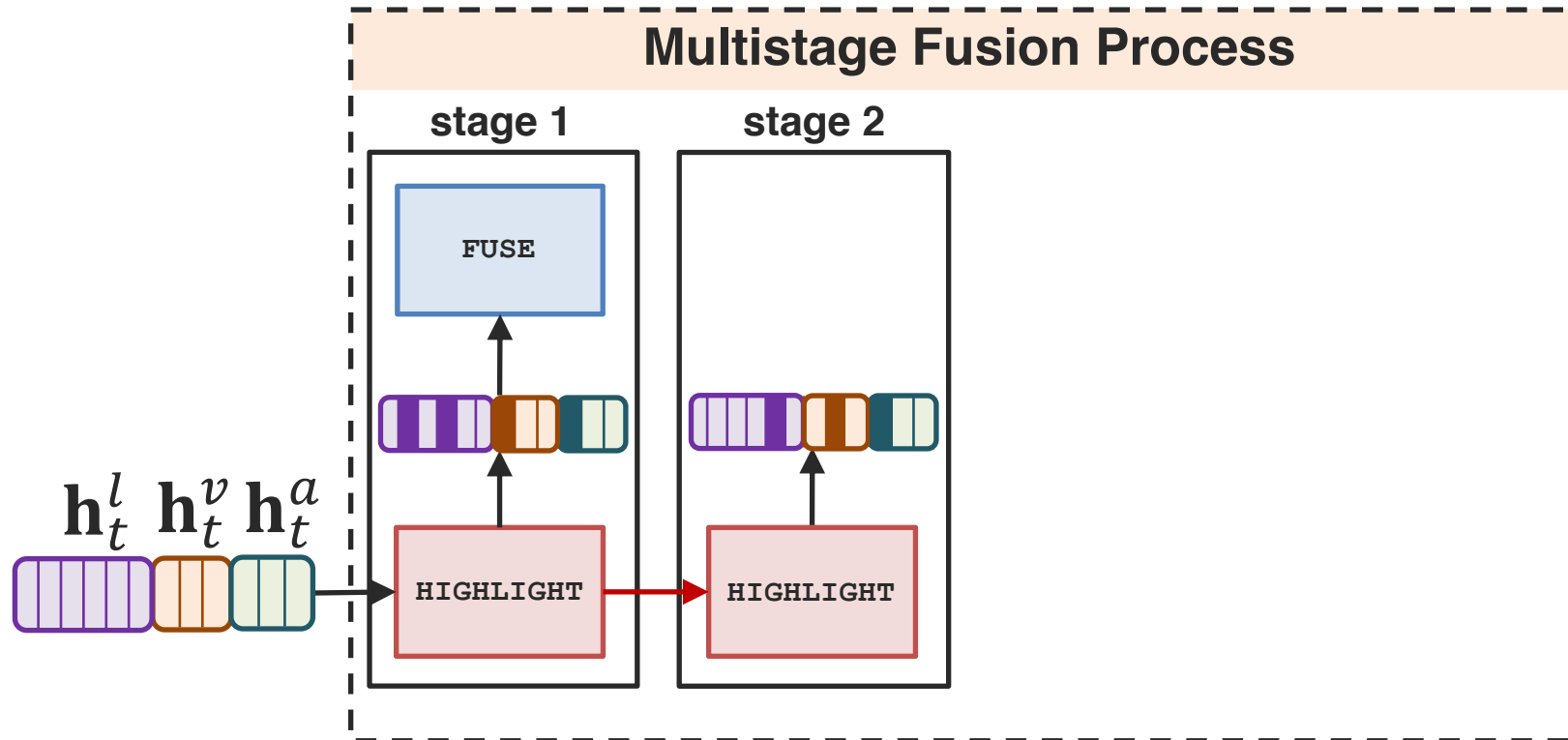
Multistage Fusion Process



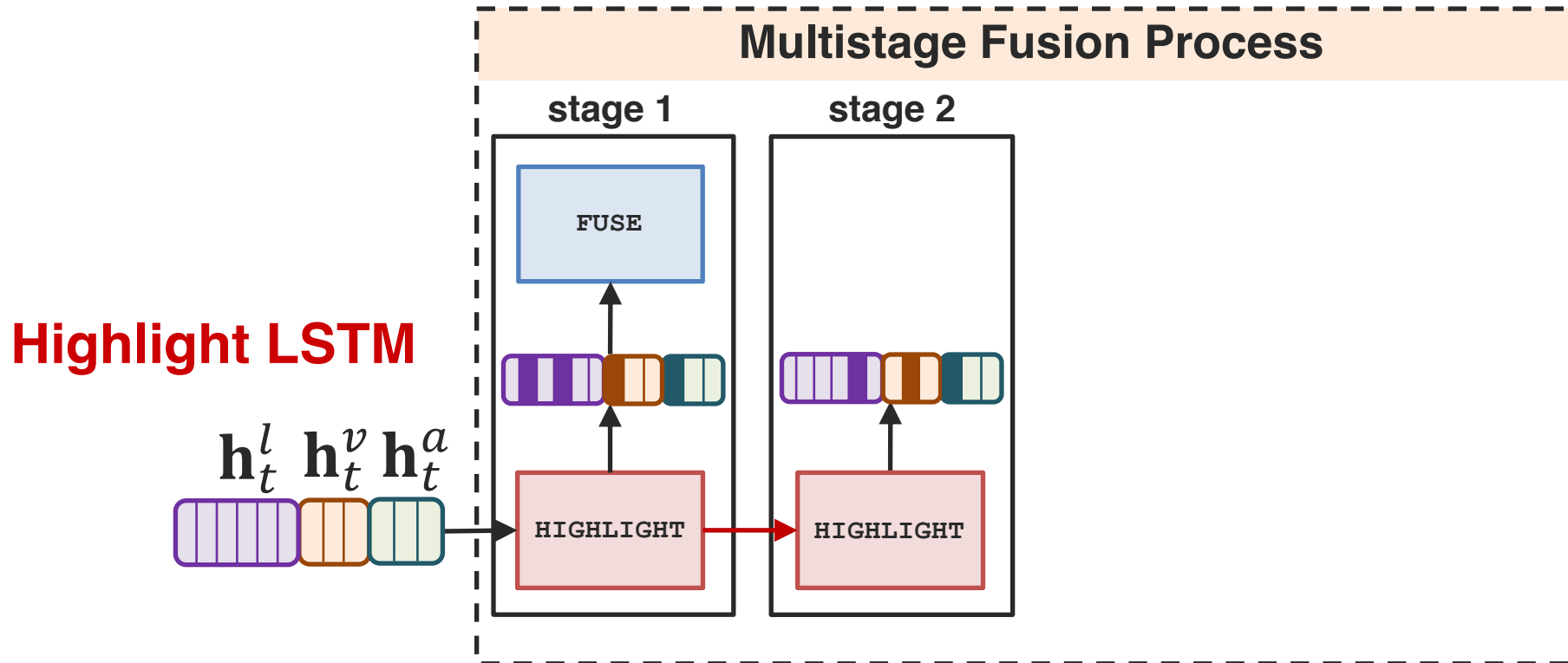
Multistage Fusion Process



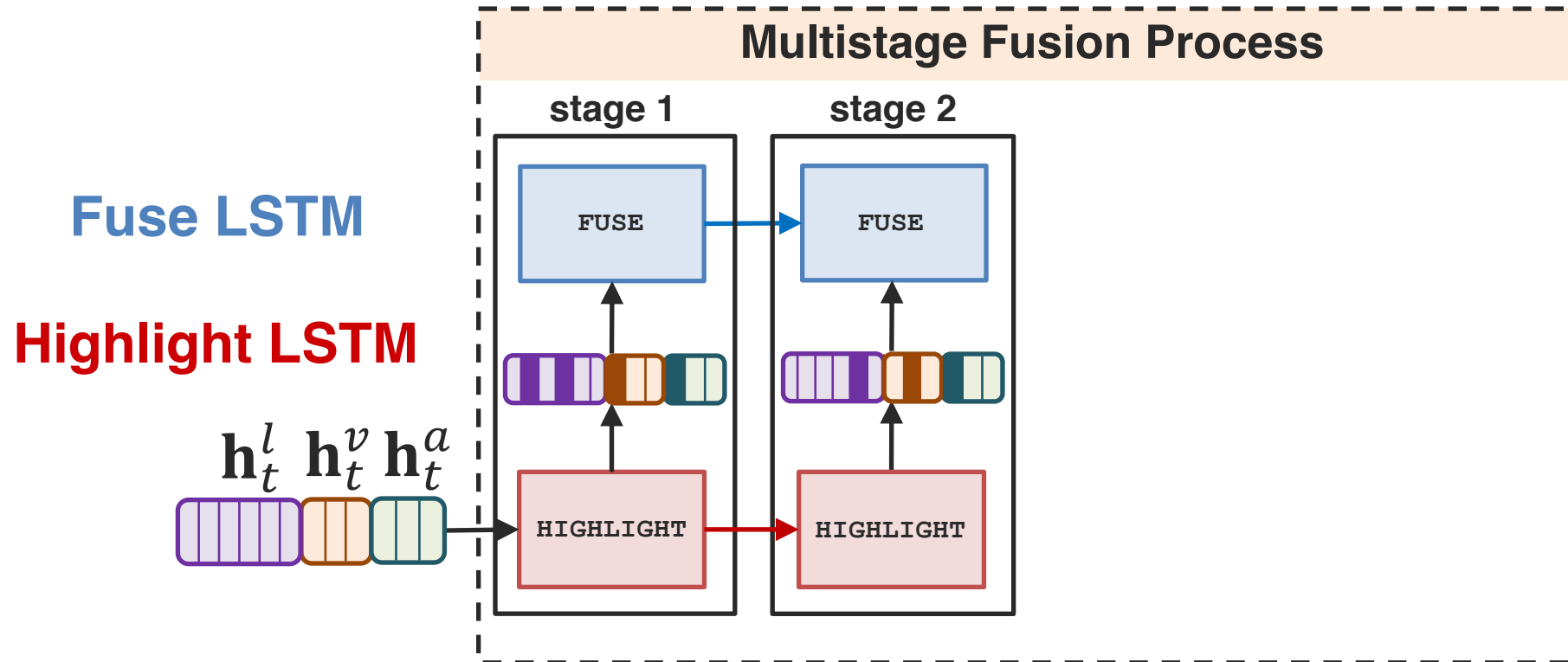
Multistage Fusion Process



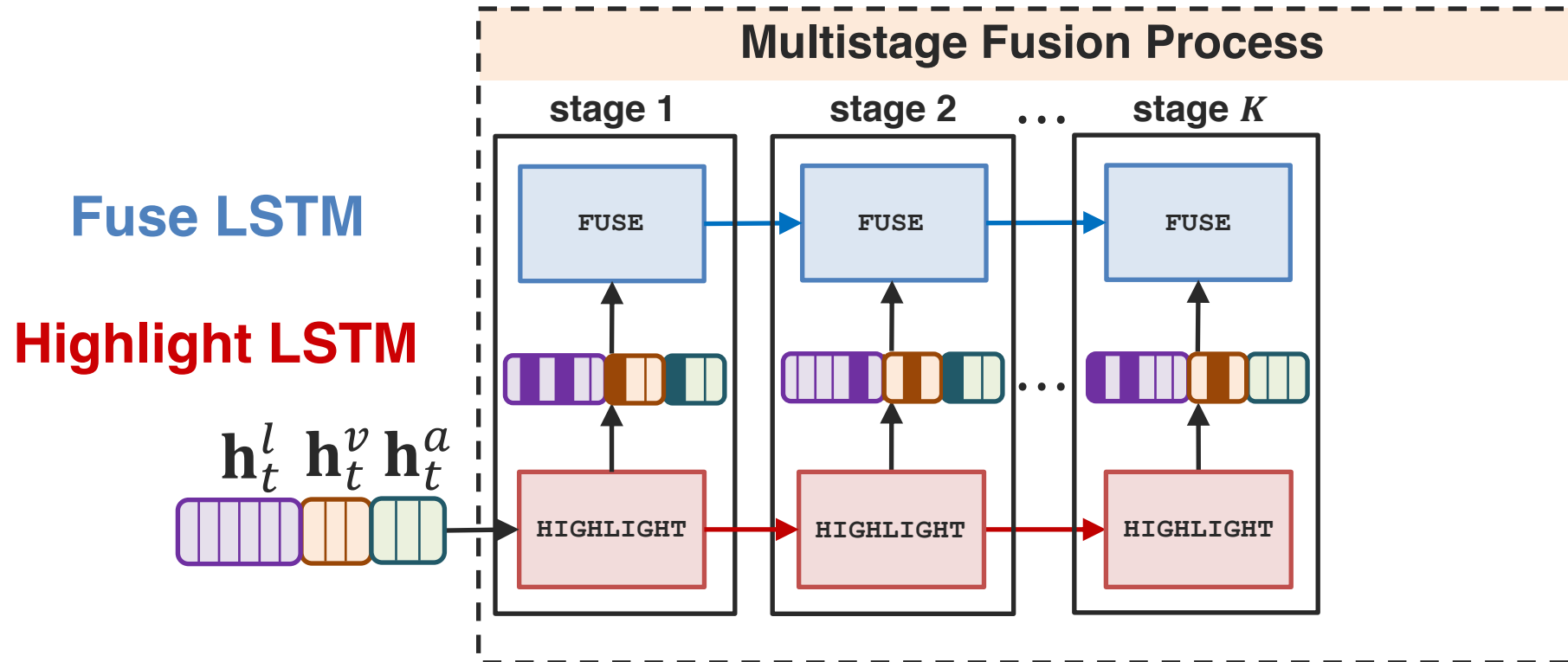
Multistage Fusion Process



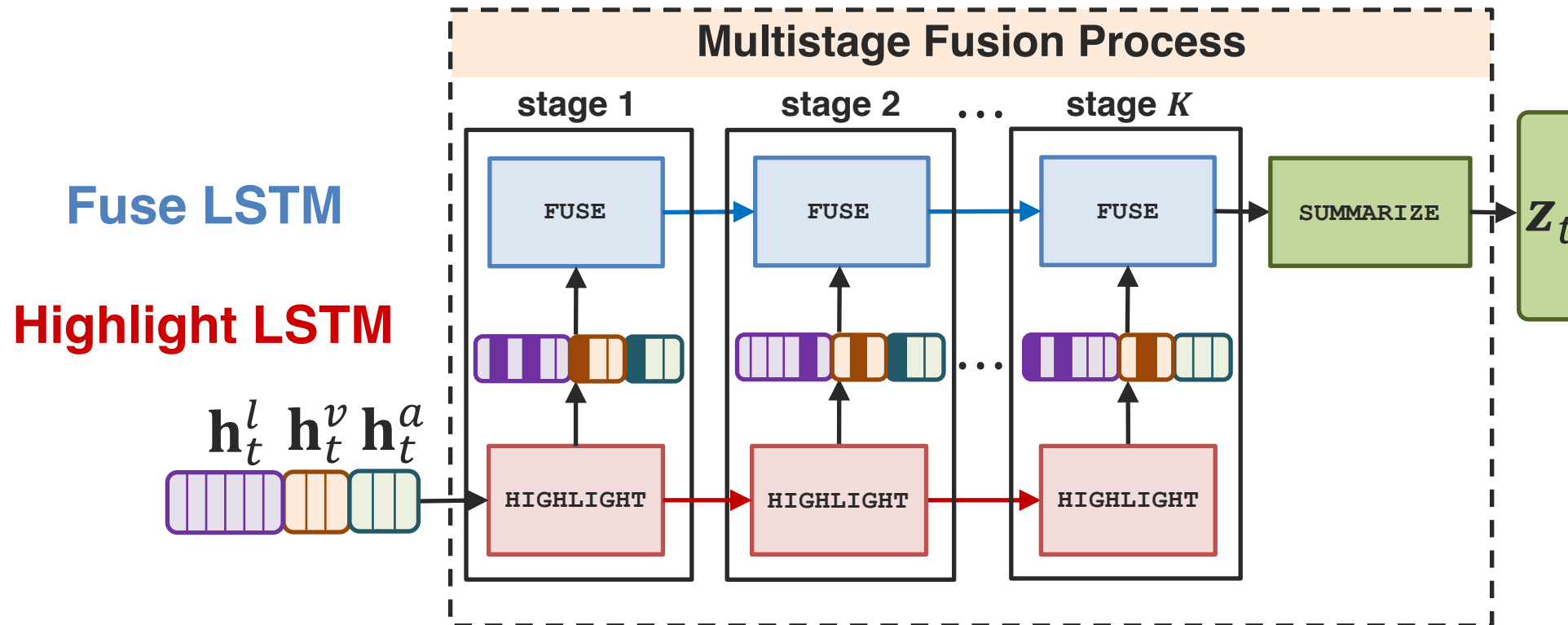
Multistage Fusion Process



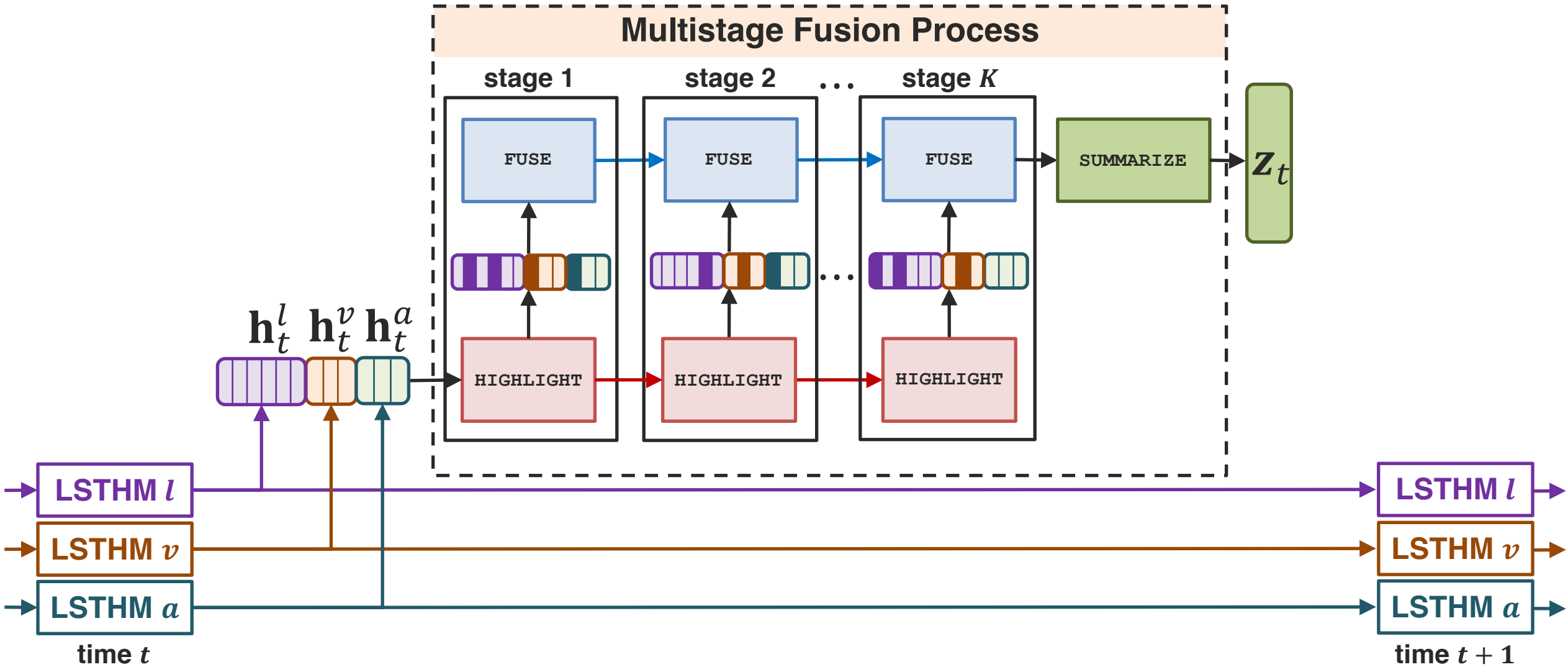
Multistage Fusion Process



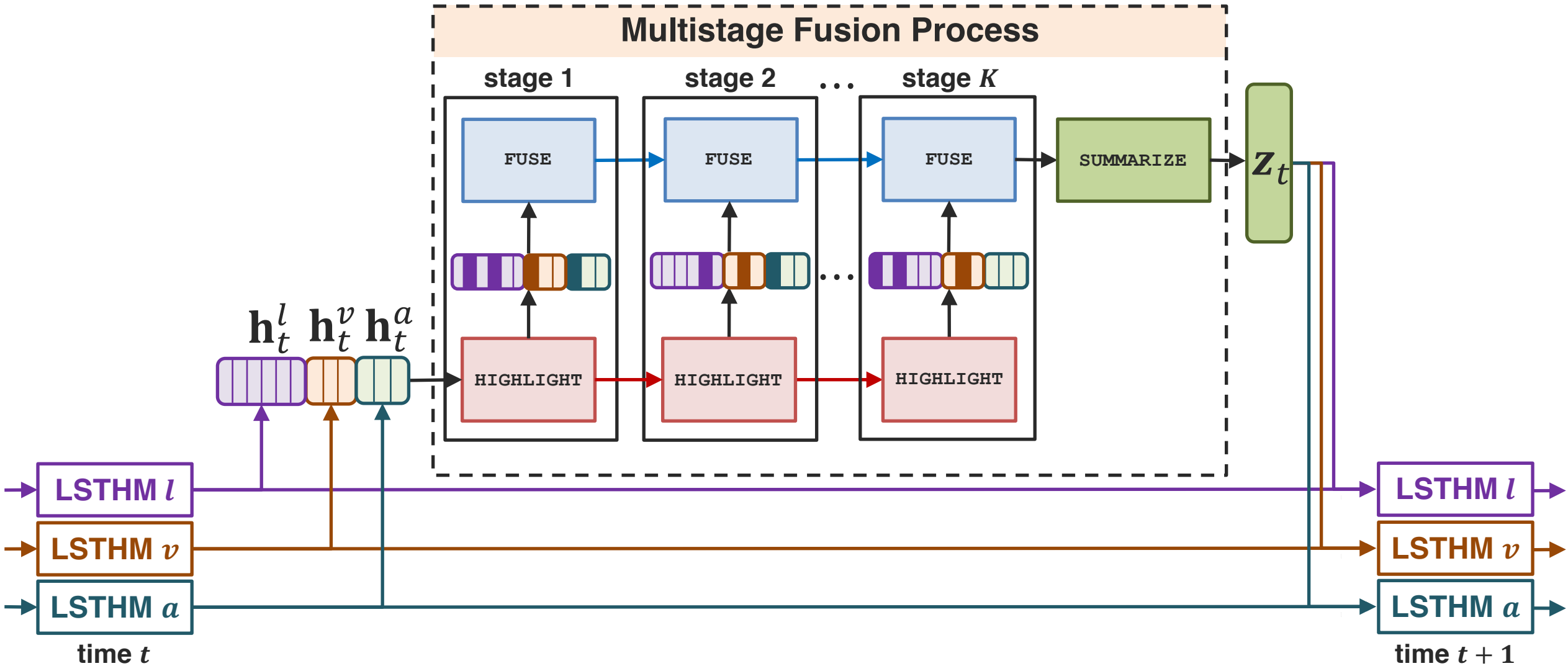
Multistage Fusion Process



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

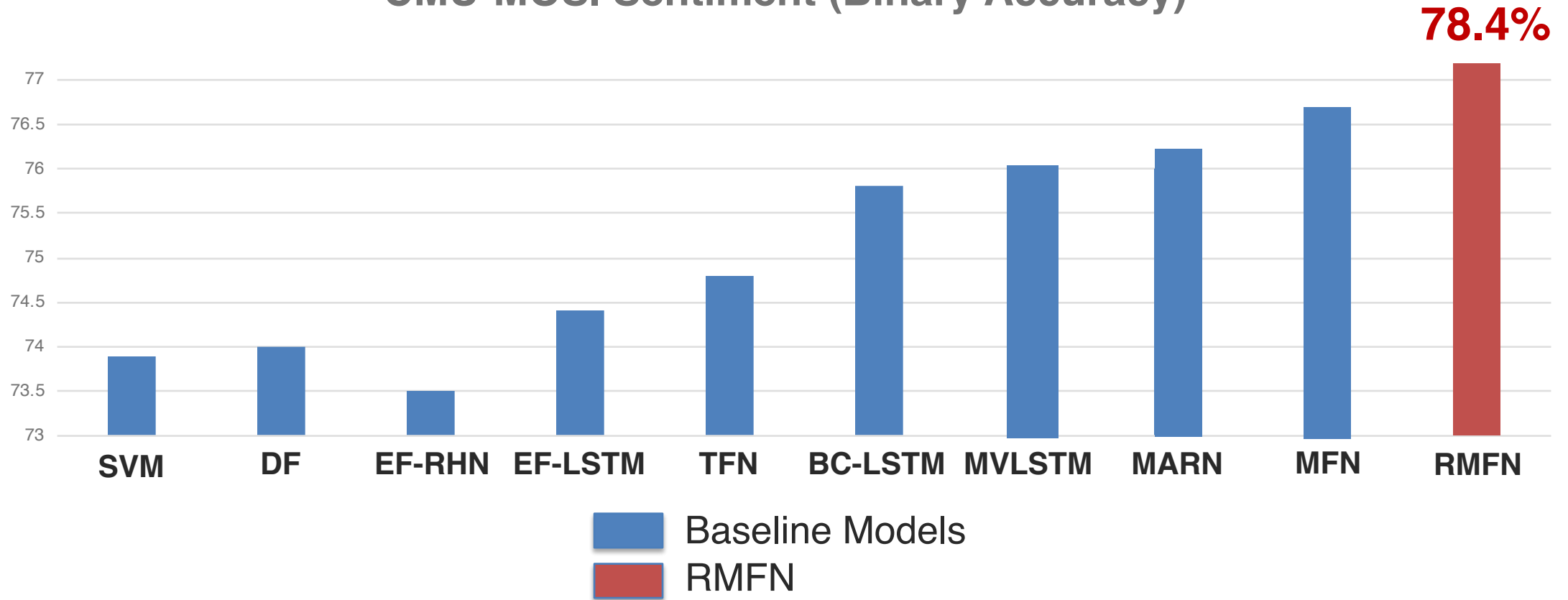
- MV-LSTM (Rajagopalan et al., 2016)

5. Memory-based models

- MARN, MFN (Zadeh et al., 2018)

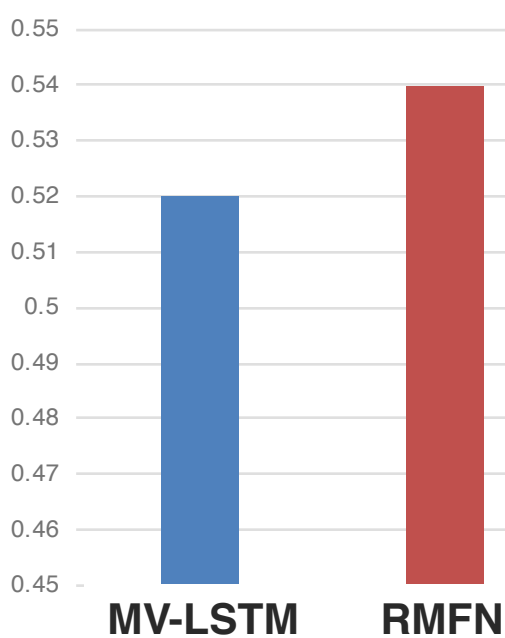
State-of-the-art Results

CMU-MOSI Sentiment (Binary Accuracy)

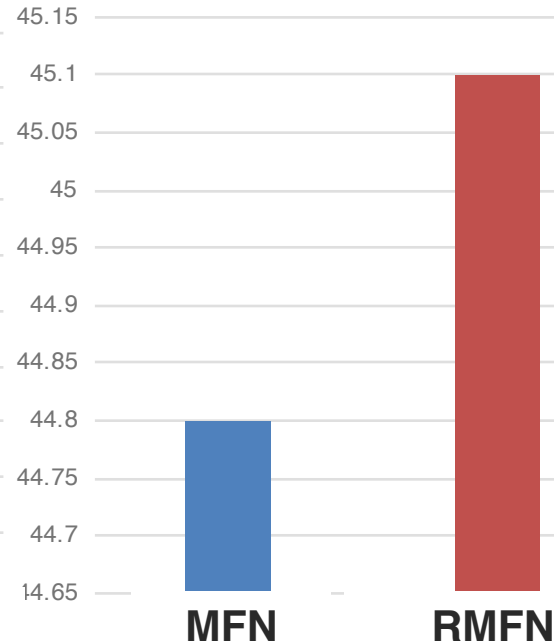


State-of-the-art Results

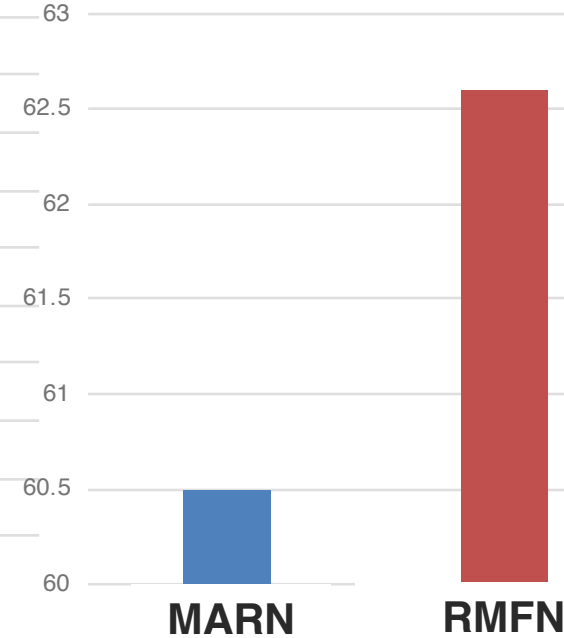
CMU-MOSI Sentiment
(Correlation)



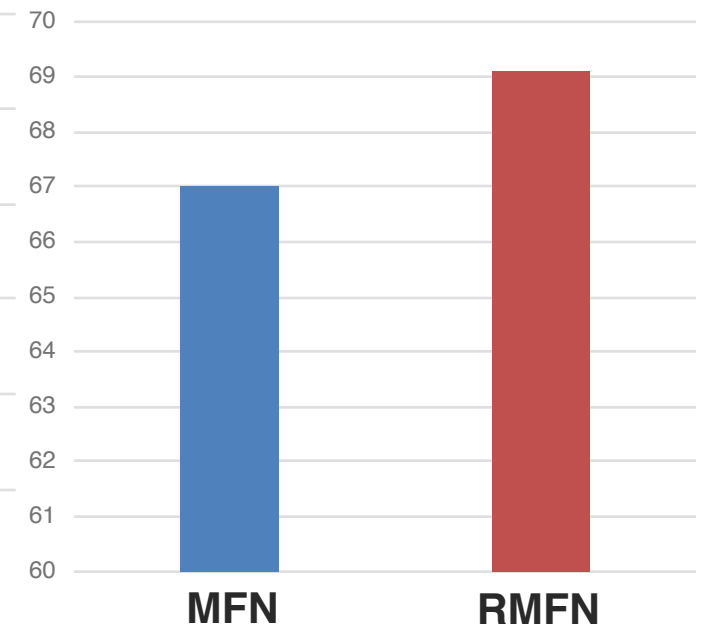
POM Personality Traits
(Multiclass Accuracy)



IEMOCAP Happy Emotion
(Binary Accuracy)

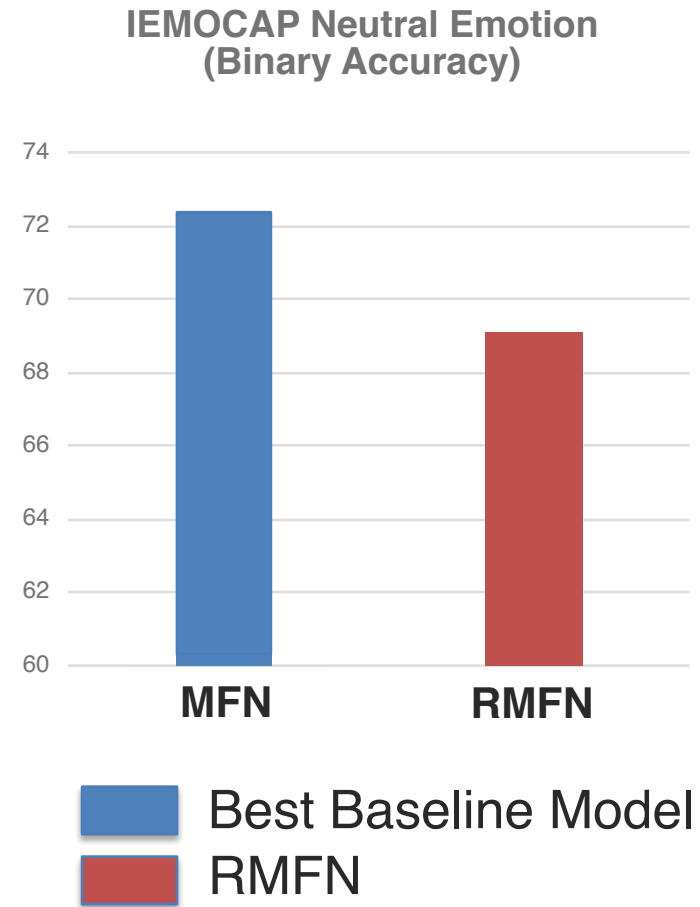


IEMOCAP Sad Emotion
(Binary Accuracy)



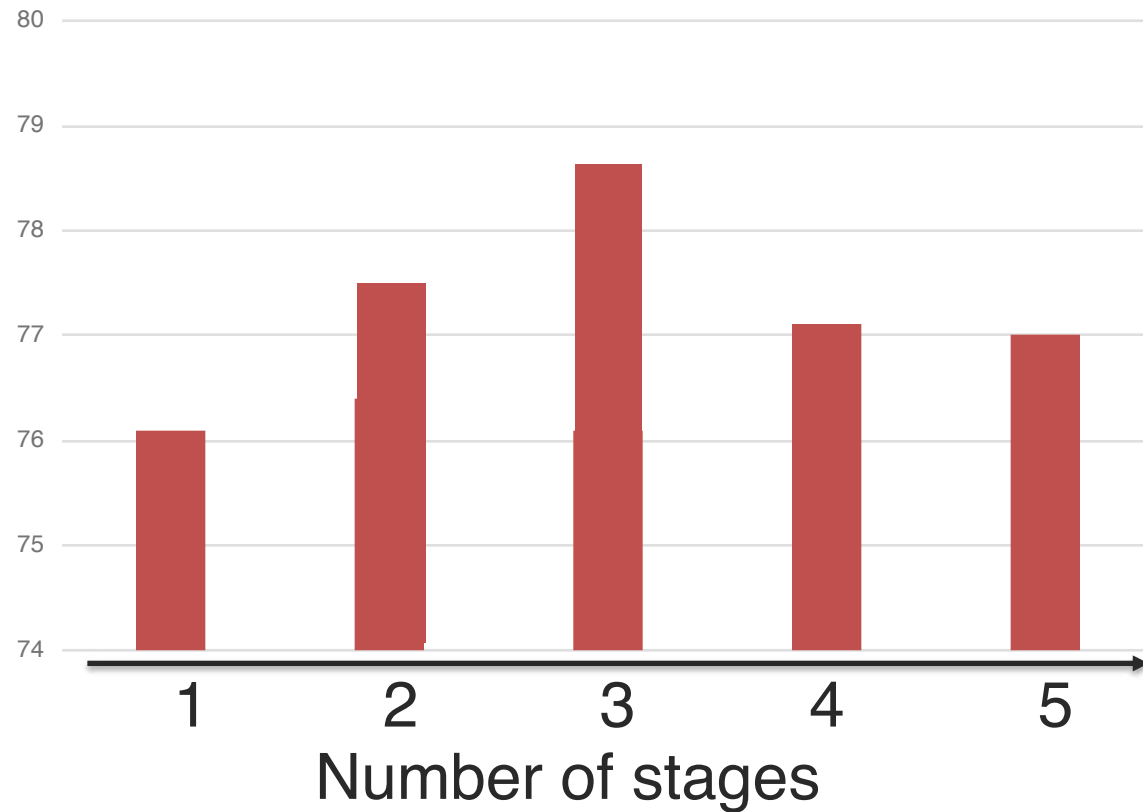
Best Baseline Model
 RMFN

Results

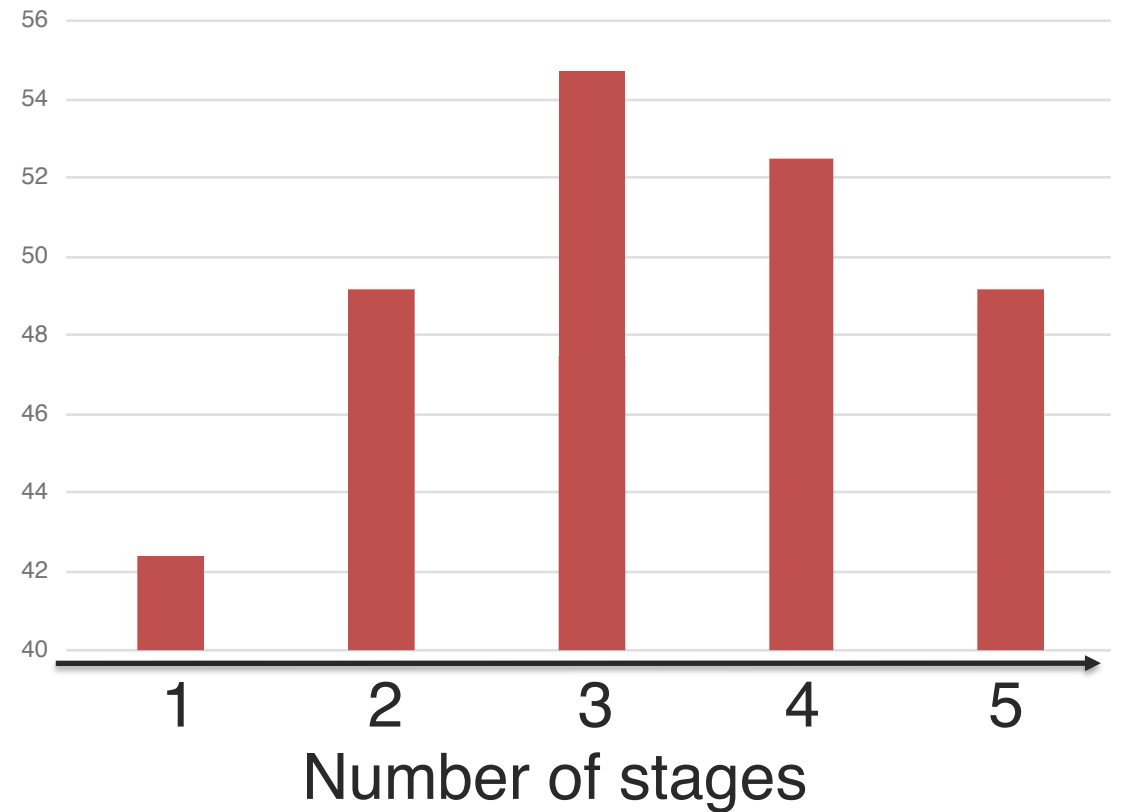


Multiple Stages are Important

CMU-MOSI Sentiment Analysis
(Binary Accuracy)



CMU-MOSI Sentiment Analysis
(Multiclass Accuracy)



Ablation Studies

Dataset	CMU-MOSI				
	Sentiment				
Task					
Metric	A2 ↑	F1 ↑	A7 ↑	MAE ↓	Corr ↑
MARN	77.1	77.0	34.7	0.968	0.625
RMFN (no MFP)	76.5	76.5	30.8	0.998	0.582
RMFN (no HIGHLIGHT)	77.9	77.9	35.9	0.952	0.666
RMFN	78.4	78.0	38.3	0.922	0.681

Interpretable Fusion

Language

I thought it was fun

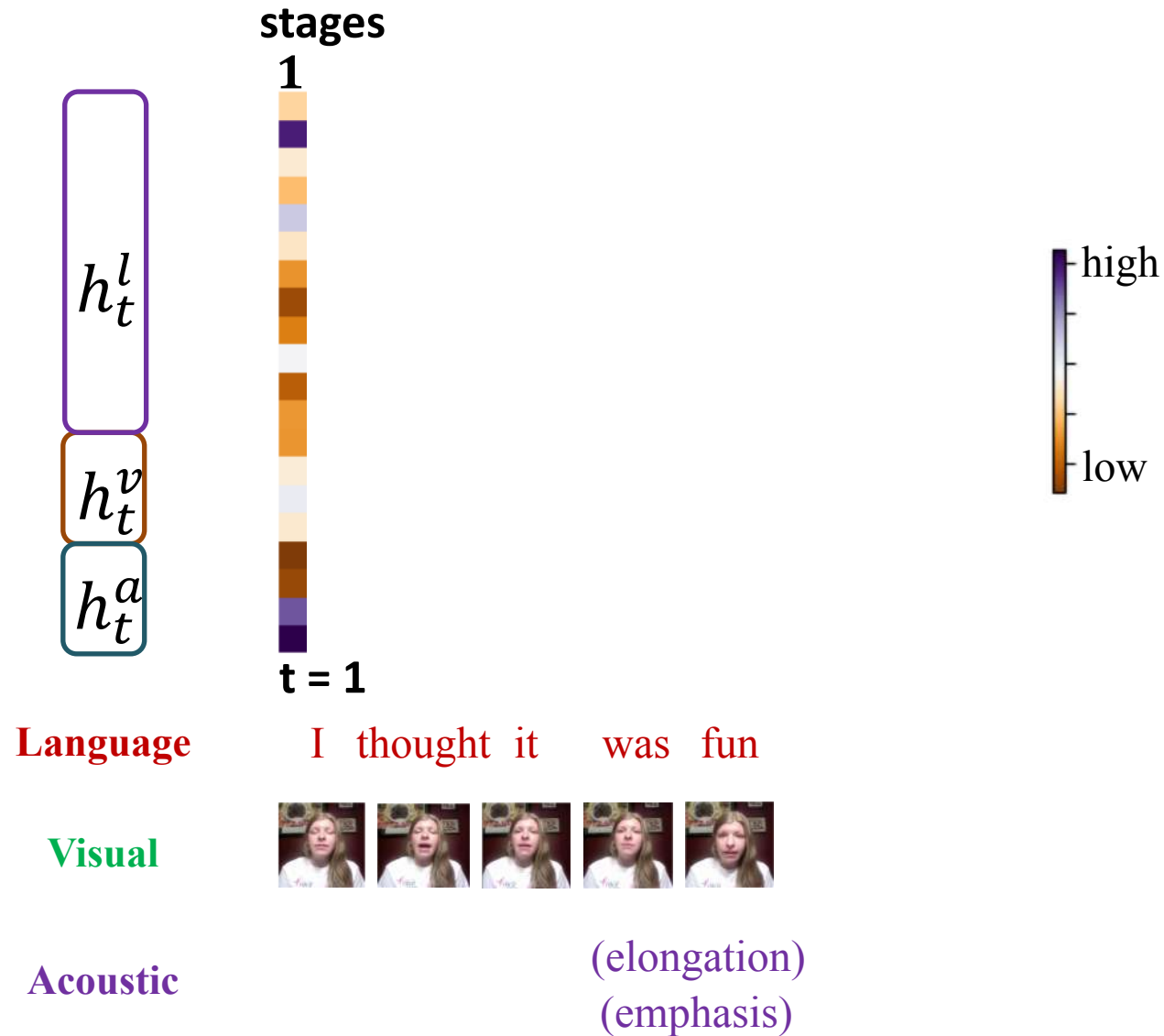
Visual



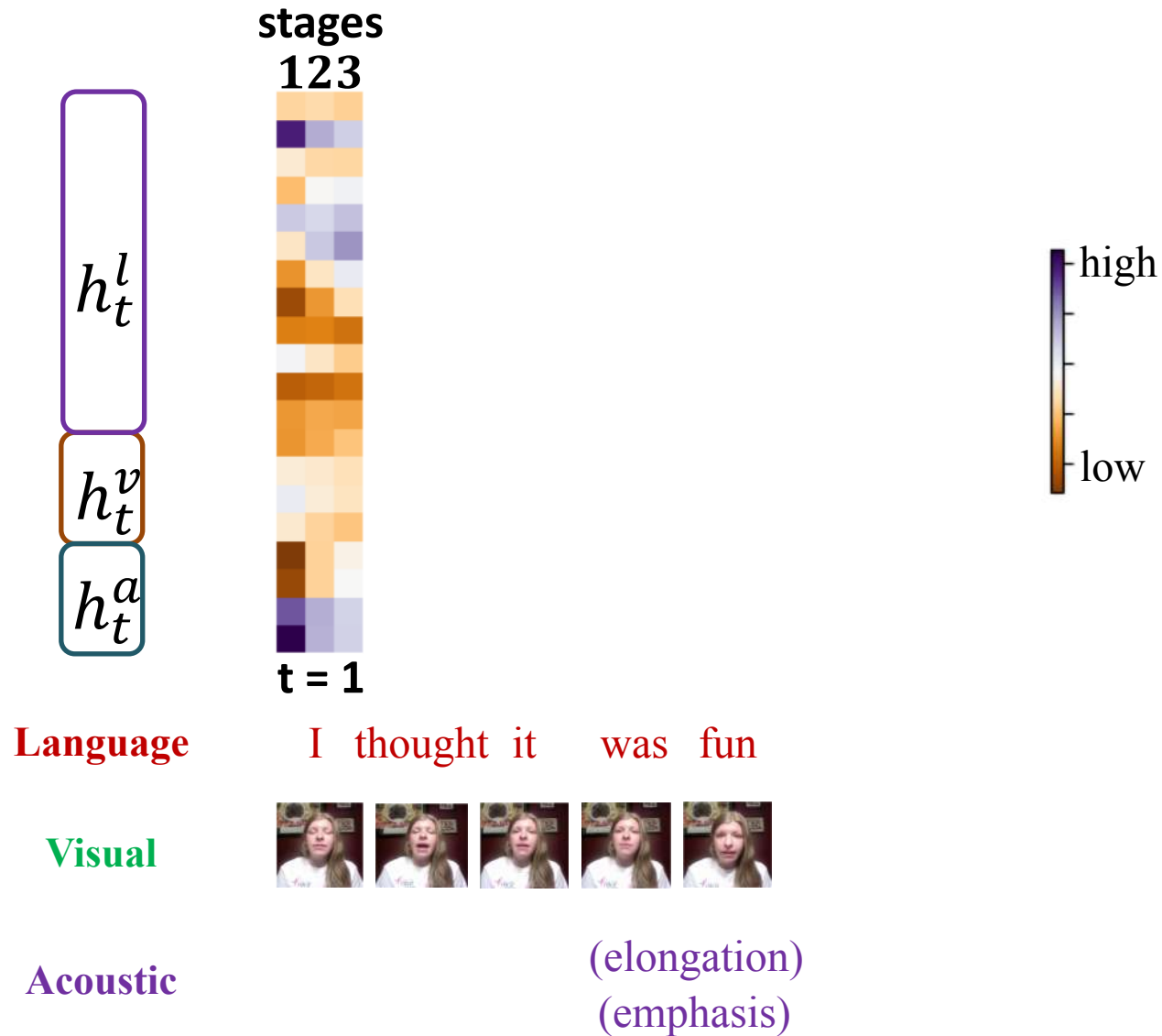
Acoustic

(elongation)
(emphasis)

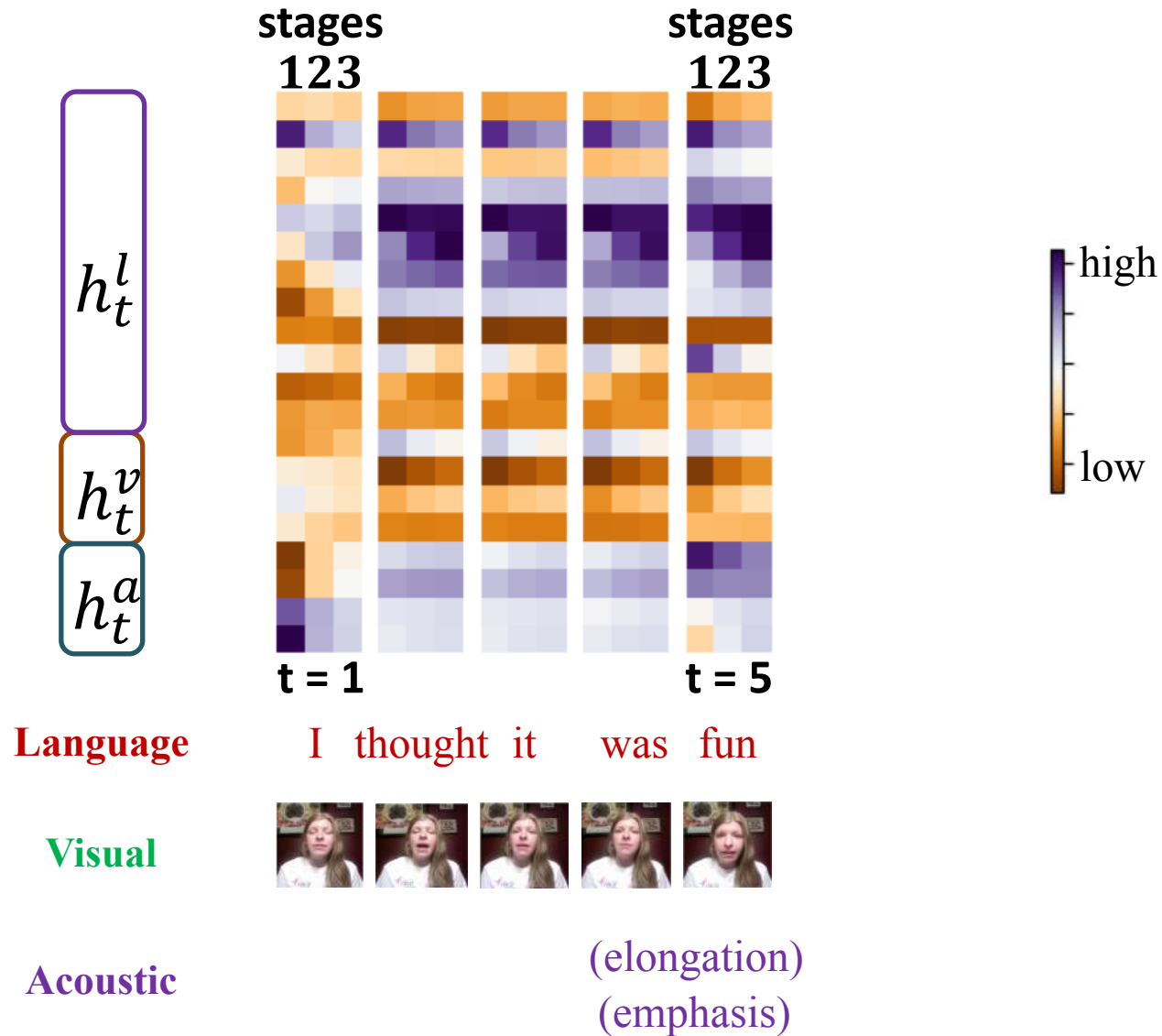
Interpretable Fusion



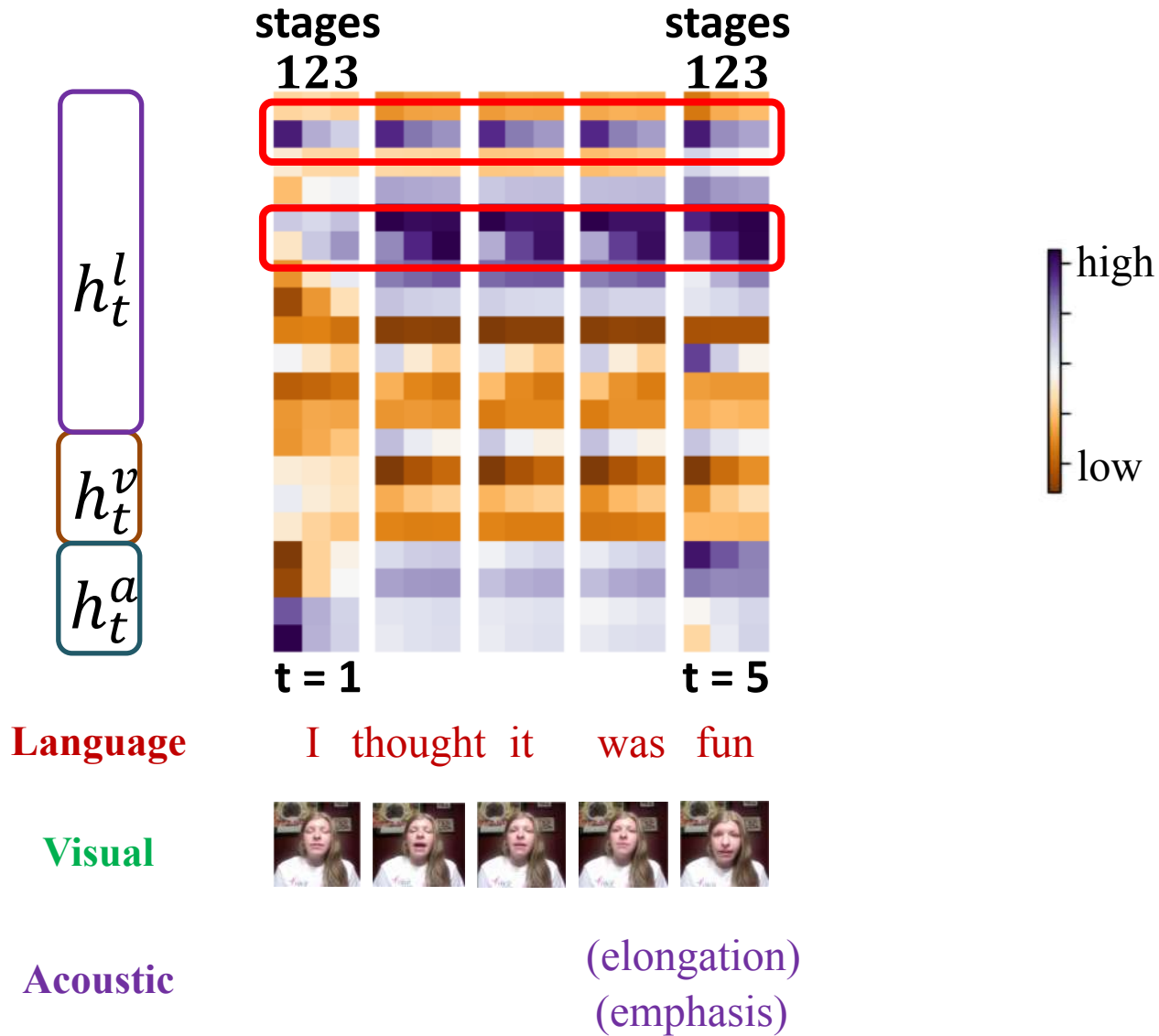
Interpretable Fusion



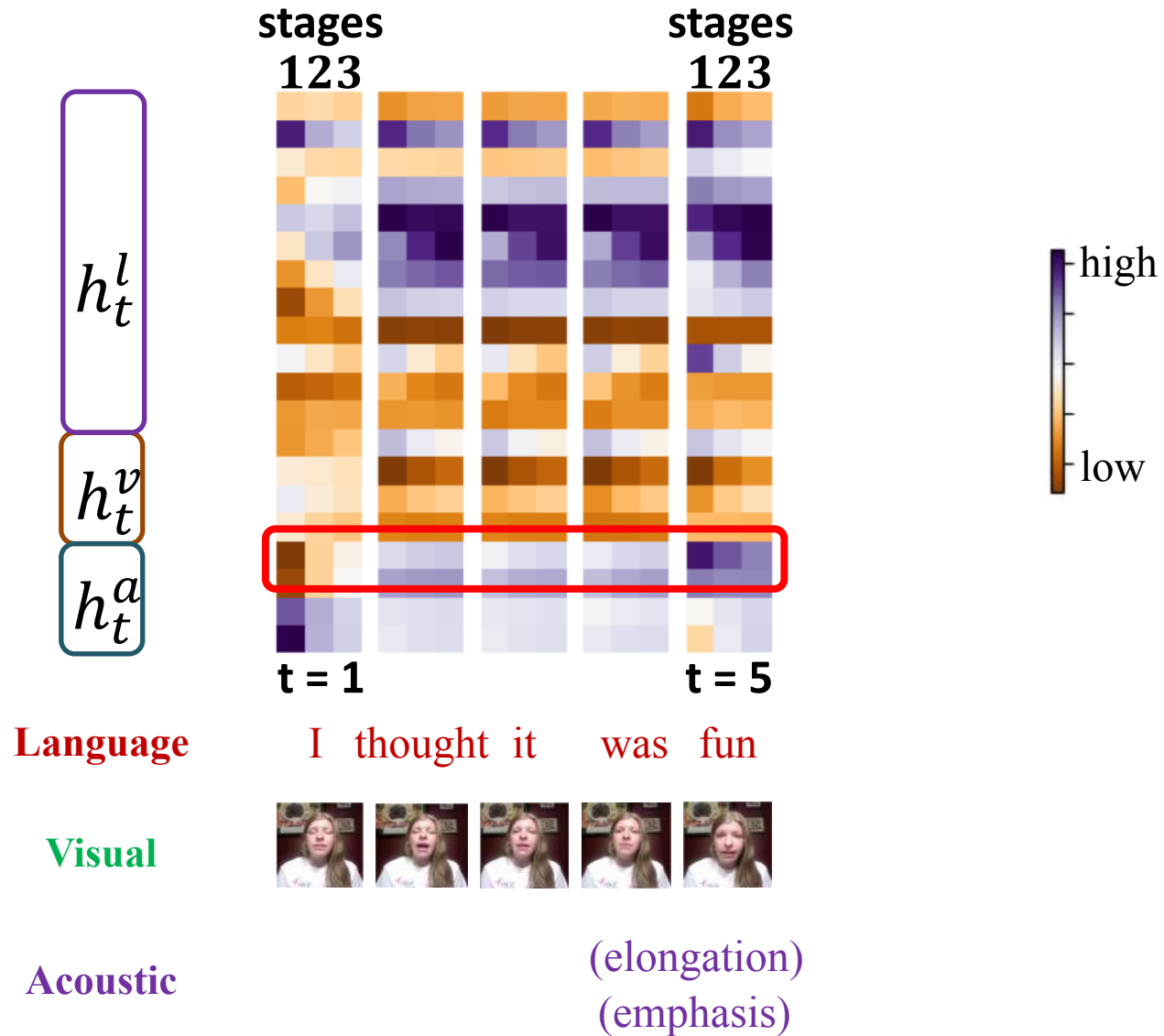
Interpretable Fusion



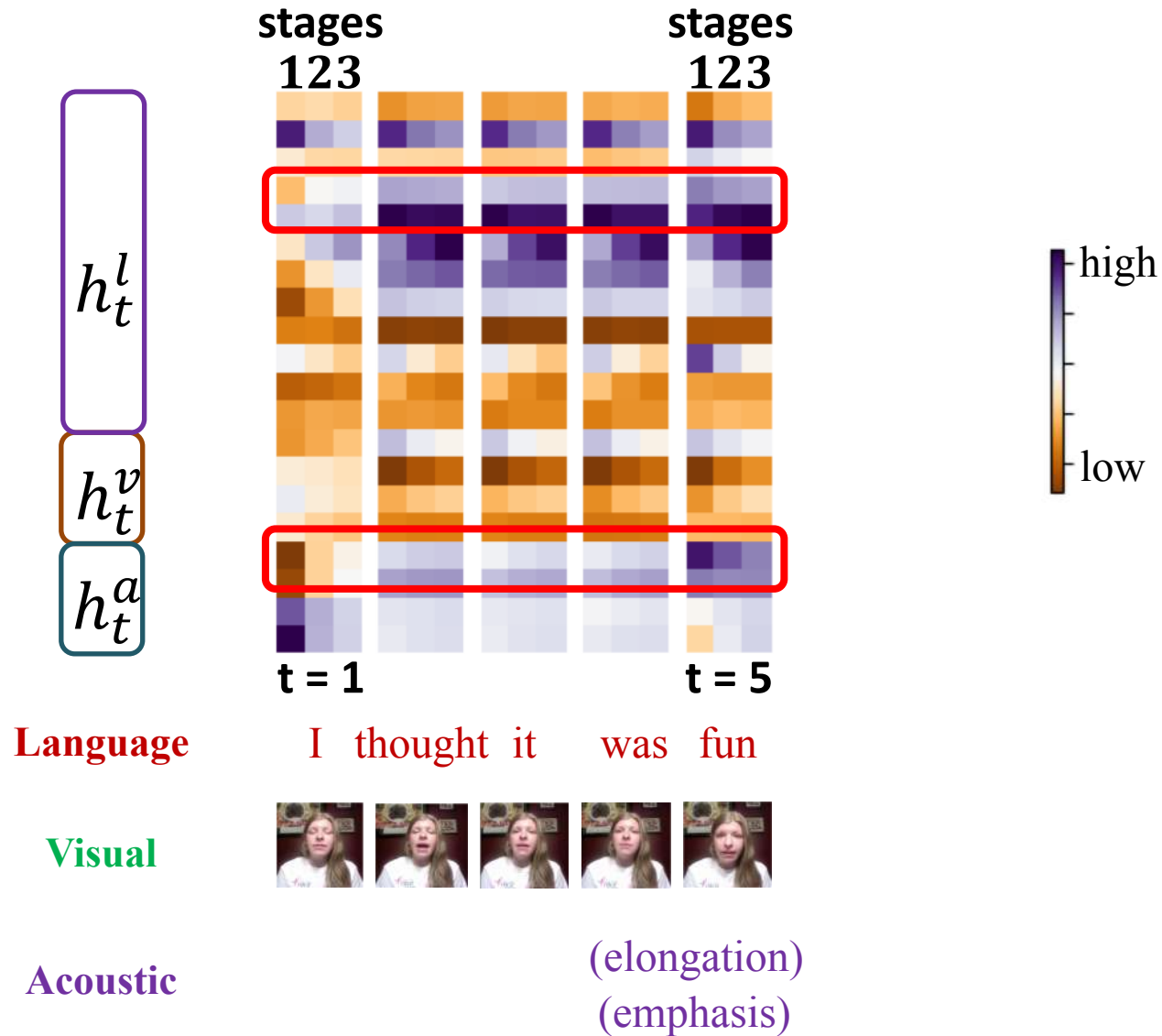
Across Stages



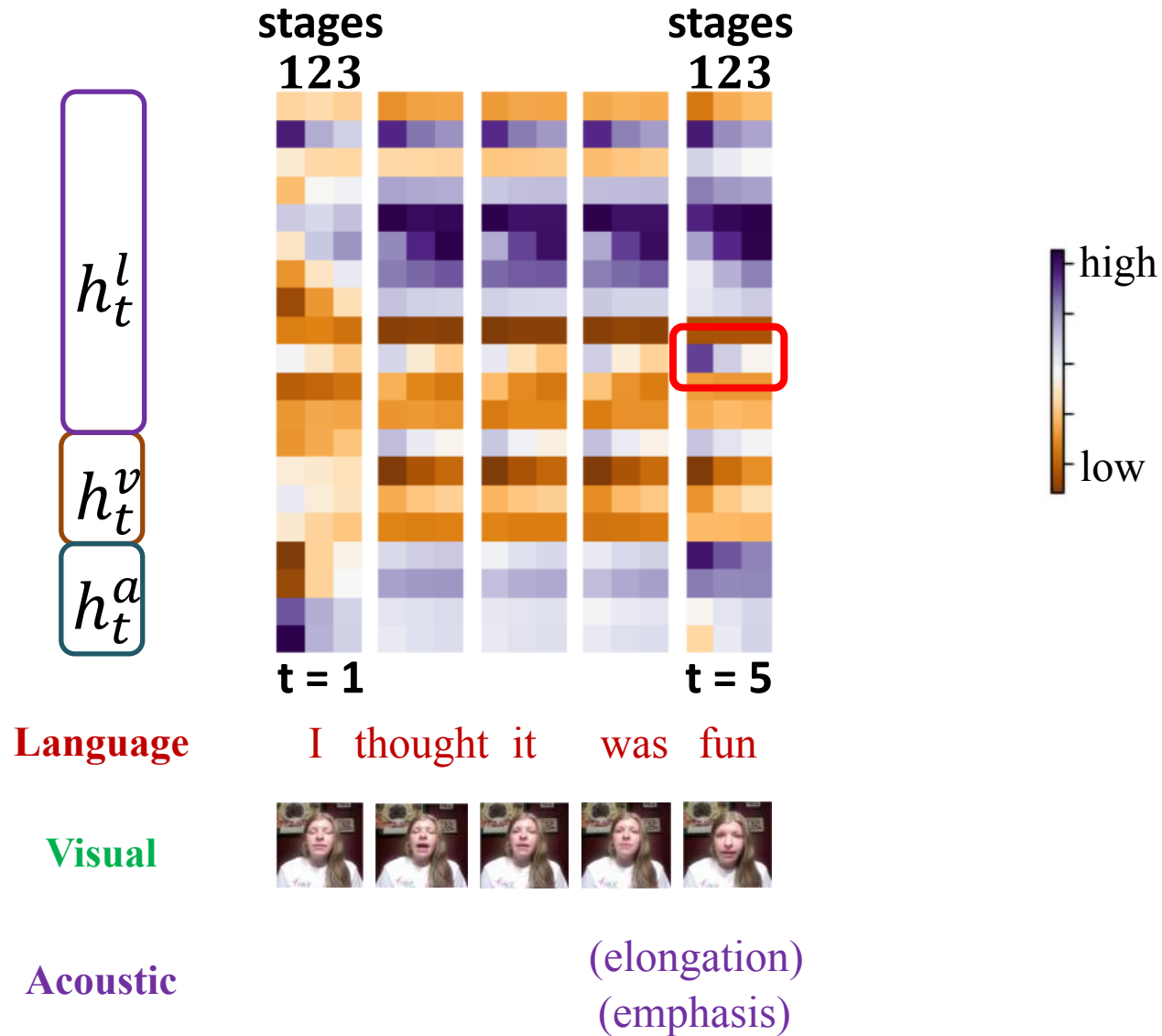
Across Time



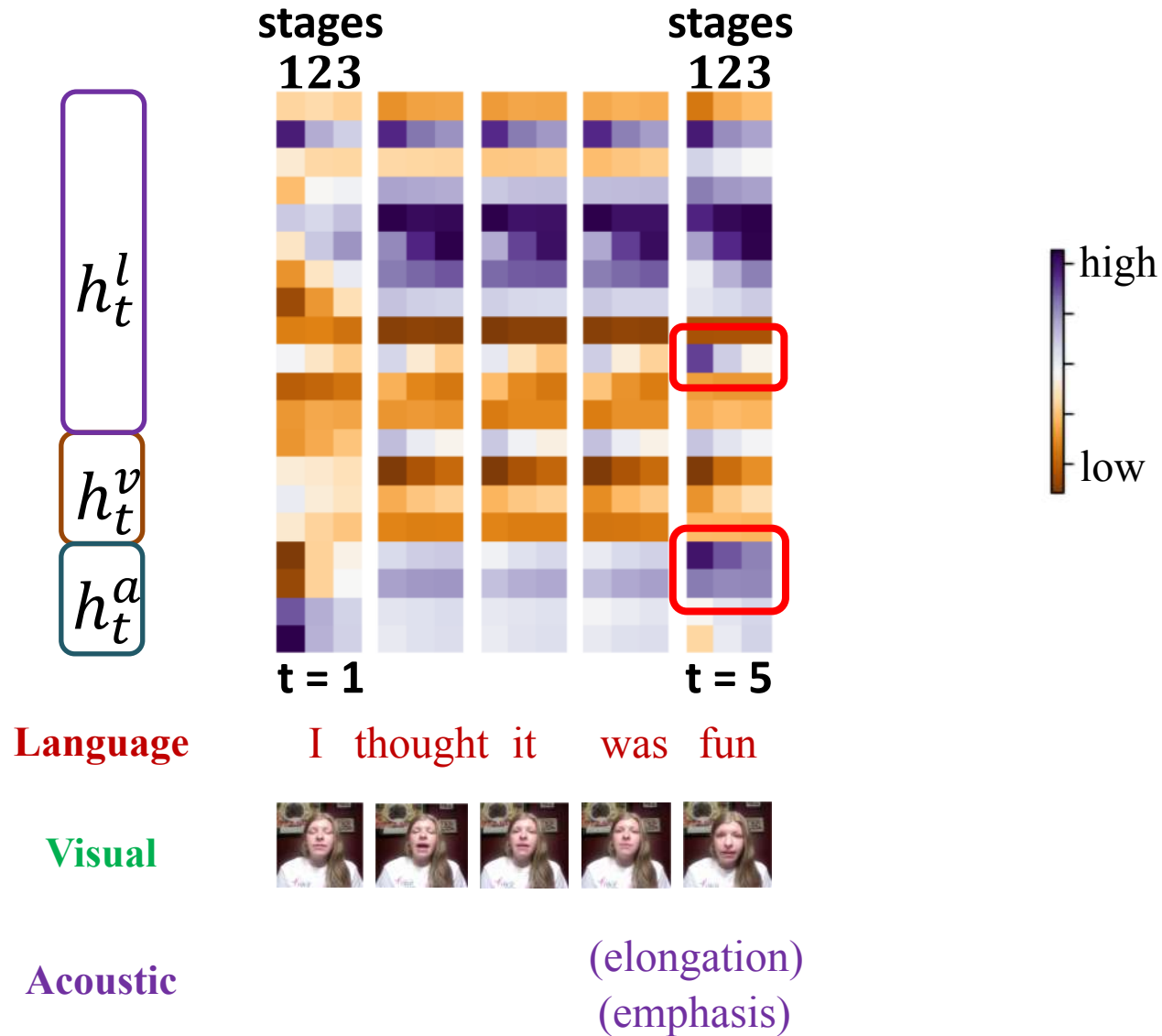
Multimodal Priors



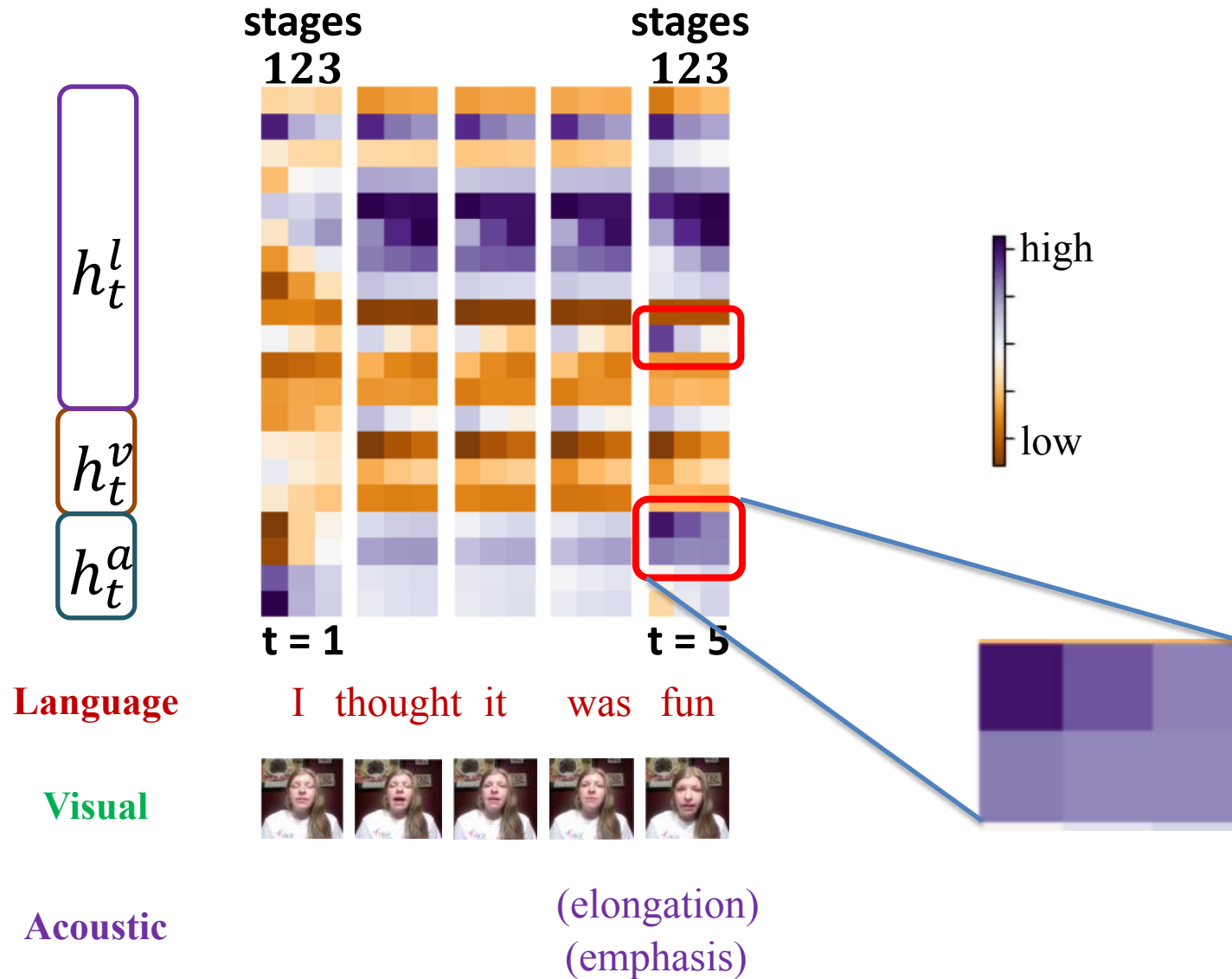
Synchronized Interactions



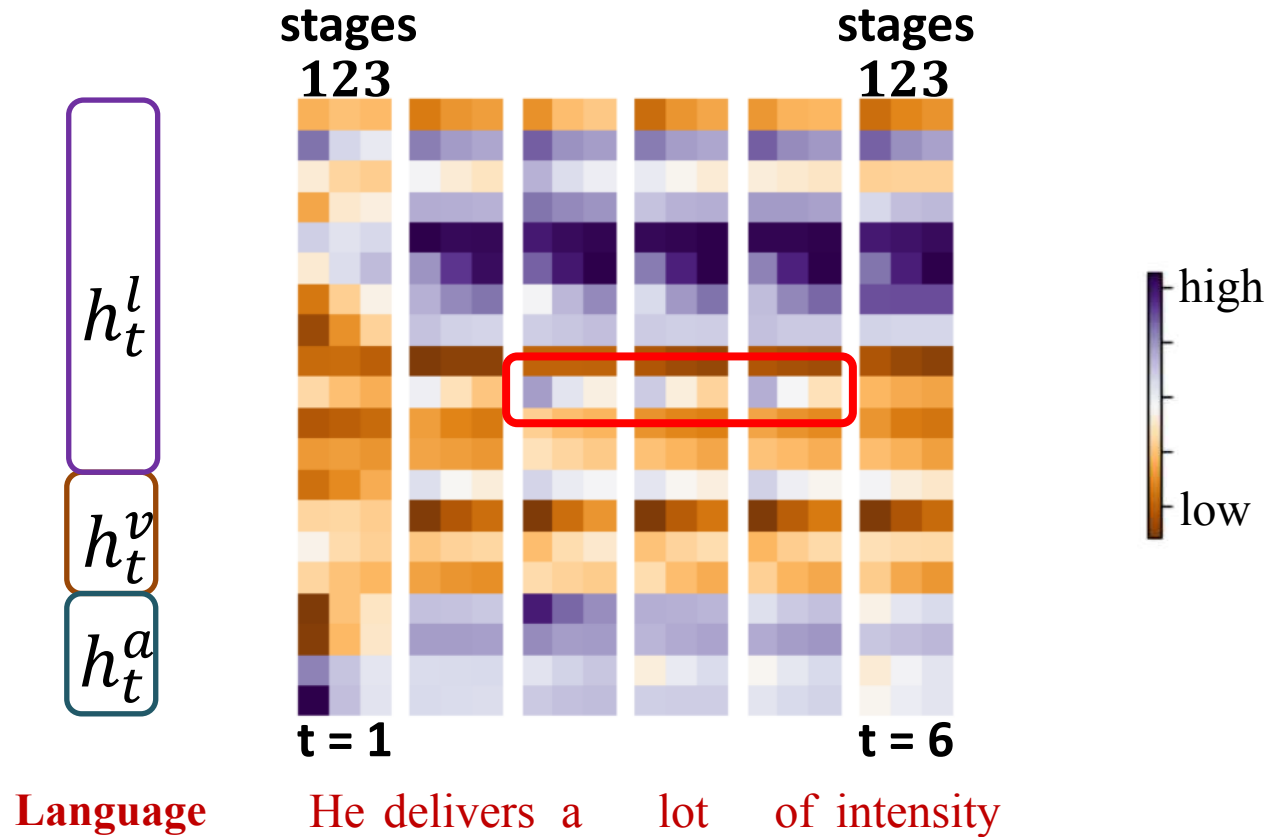
Synchronized Interactions



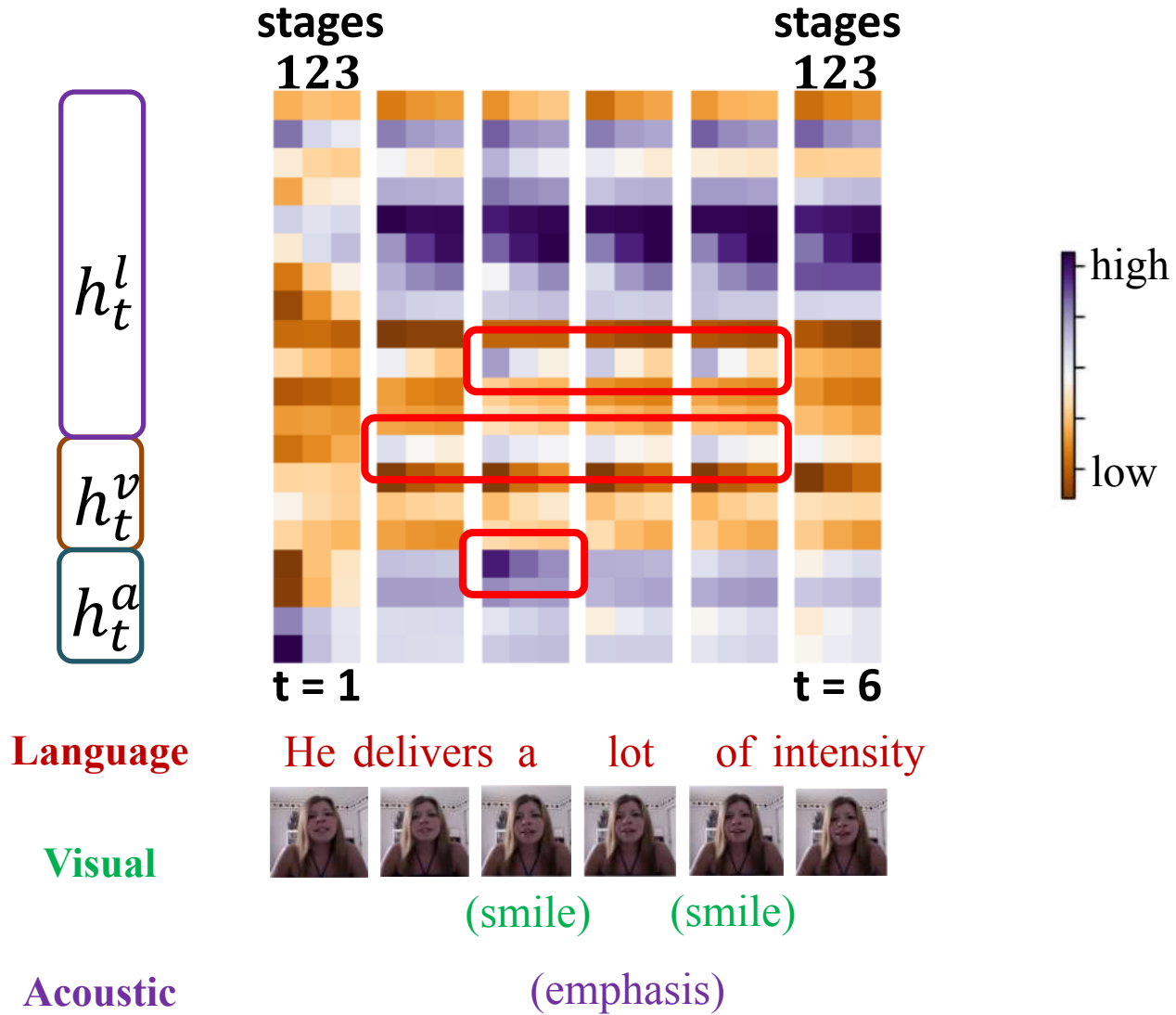
Synchronized Interactions



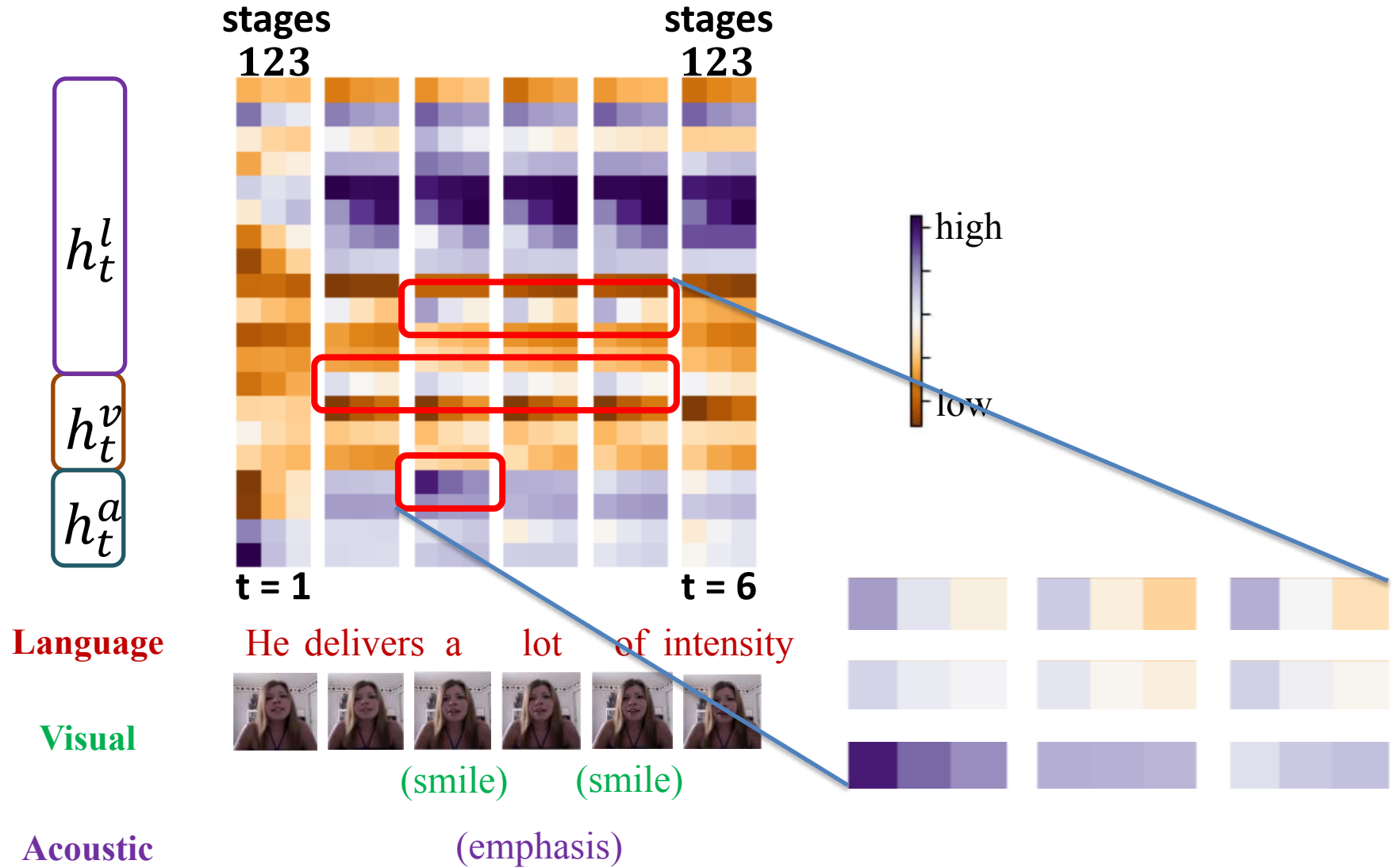
Asynchronous Trimodal Interactions



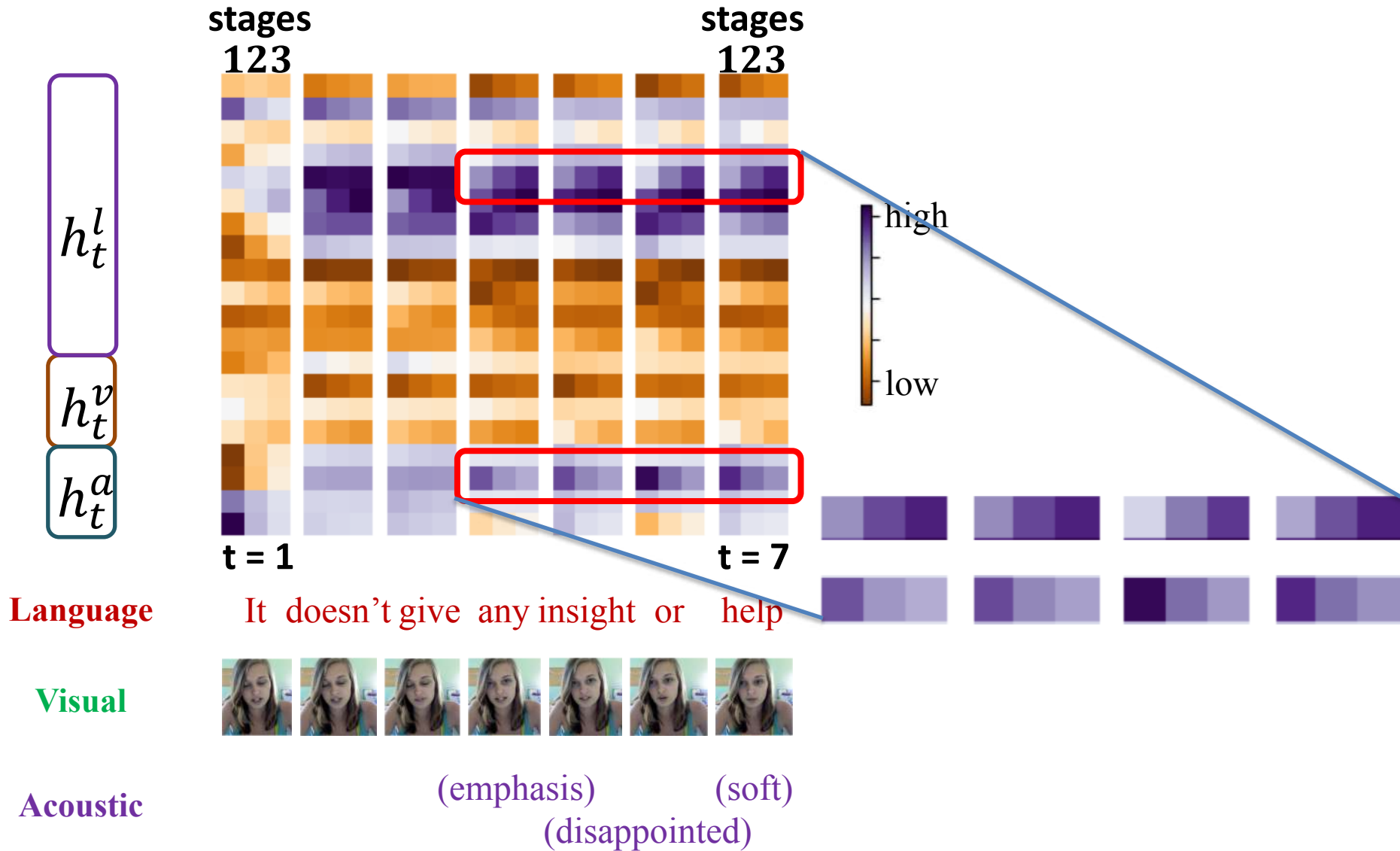
Asynchronous Trimodal Interactions



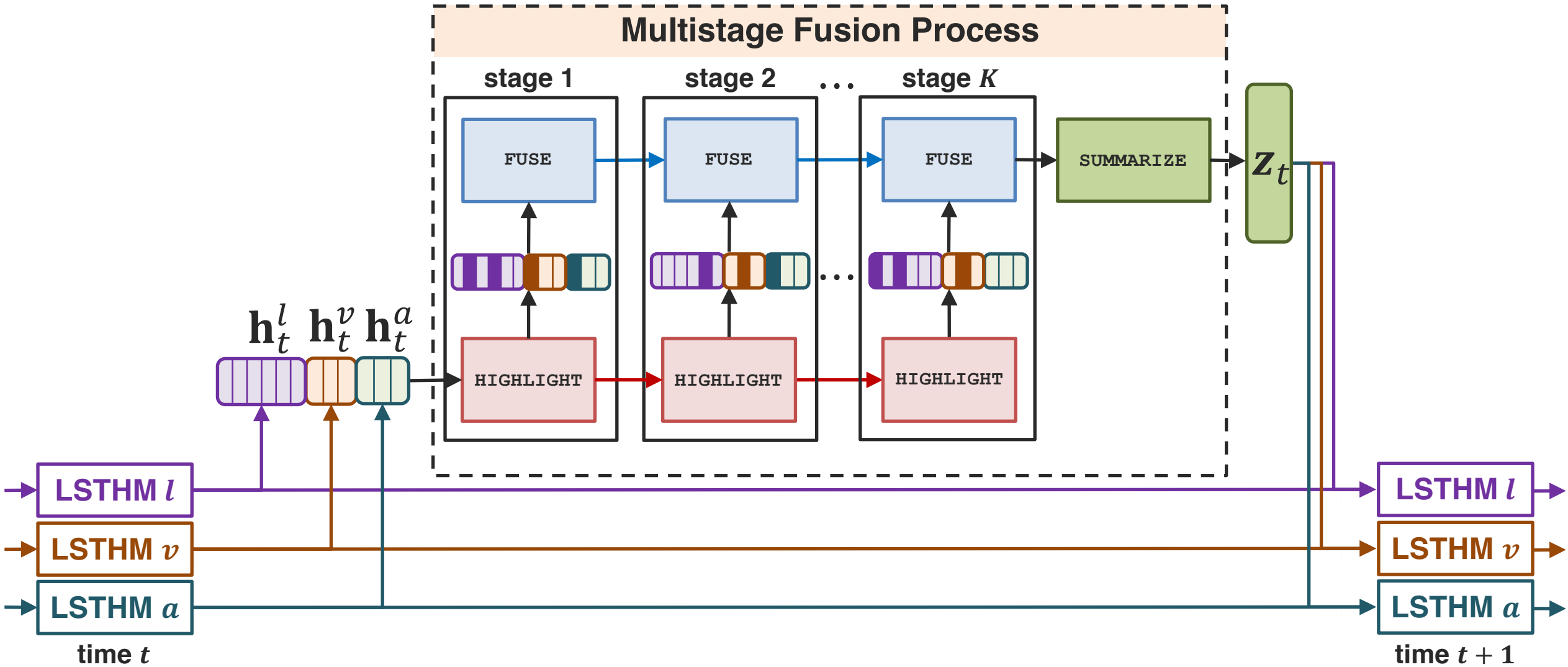
Asynchronous Trimodal Interactions



Bimodal Interactions



Recurrent Multistage Fusion Network



The End!

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

Email: pliang@cs.cmu.edu

Twitter: [@pliang279](https://twitter.com/pliang279)

