Modeling Spatiotemporal Multimodal Language with Recurrent Multistage Fusion

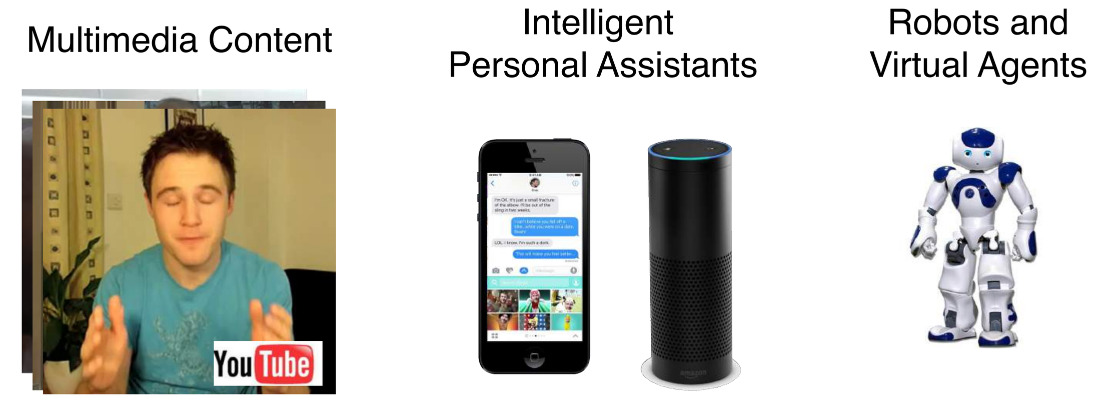
Paul Pu Liang, Ziyin Liu, Amir Zadeh, Louis-Philippe Morency

School of Computer Science, Carnegie Mellon University

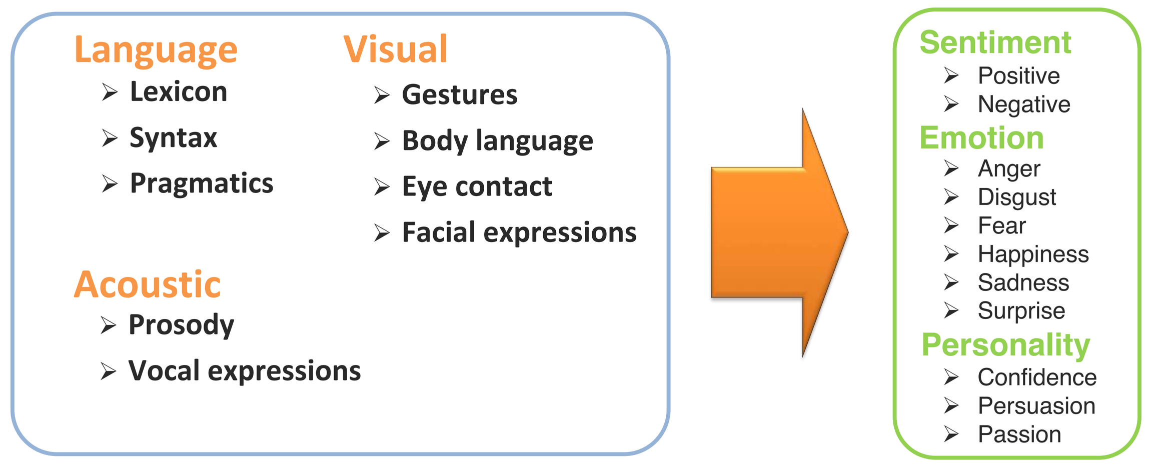
{pliang,ziyinl,abagherz,morency}@cs.cmu.edu

Artificial Intelligence and Multimodal Language

Giving Artificial Intelligence the power to model human language is a core research challenge.



Multimodal Language Modalities



Challenge 1: Intra-modal Interactions

Intra-modal interactions exist within each modality independent of other modalities (**temporal interactions**).

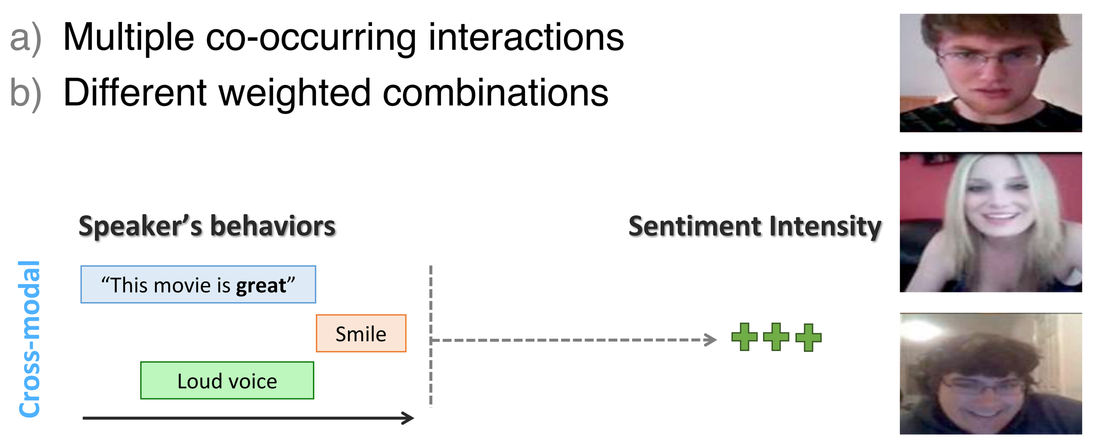
a) Temporal sequences

Challenge 2: Cross-modal Interactions

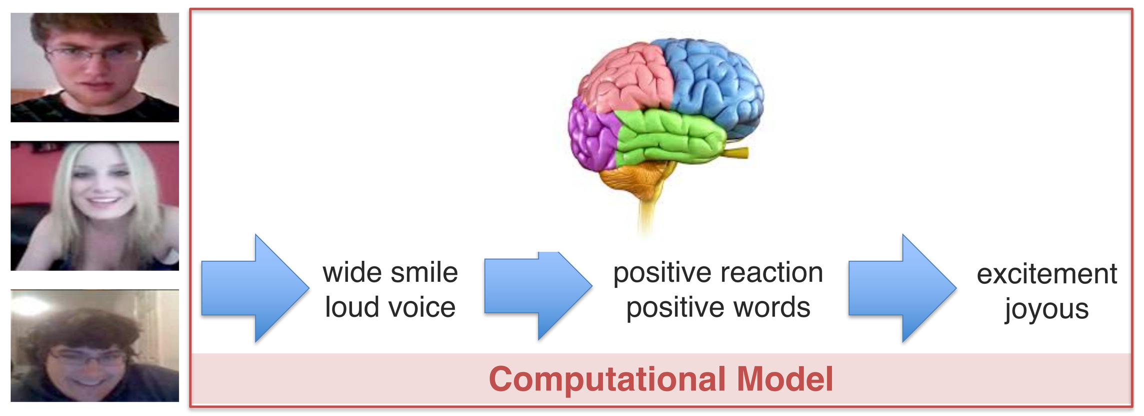
Cross-modal interactions refer to interactions between modalities (**spatial interactions**).

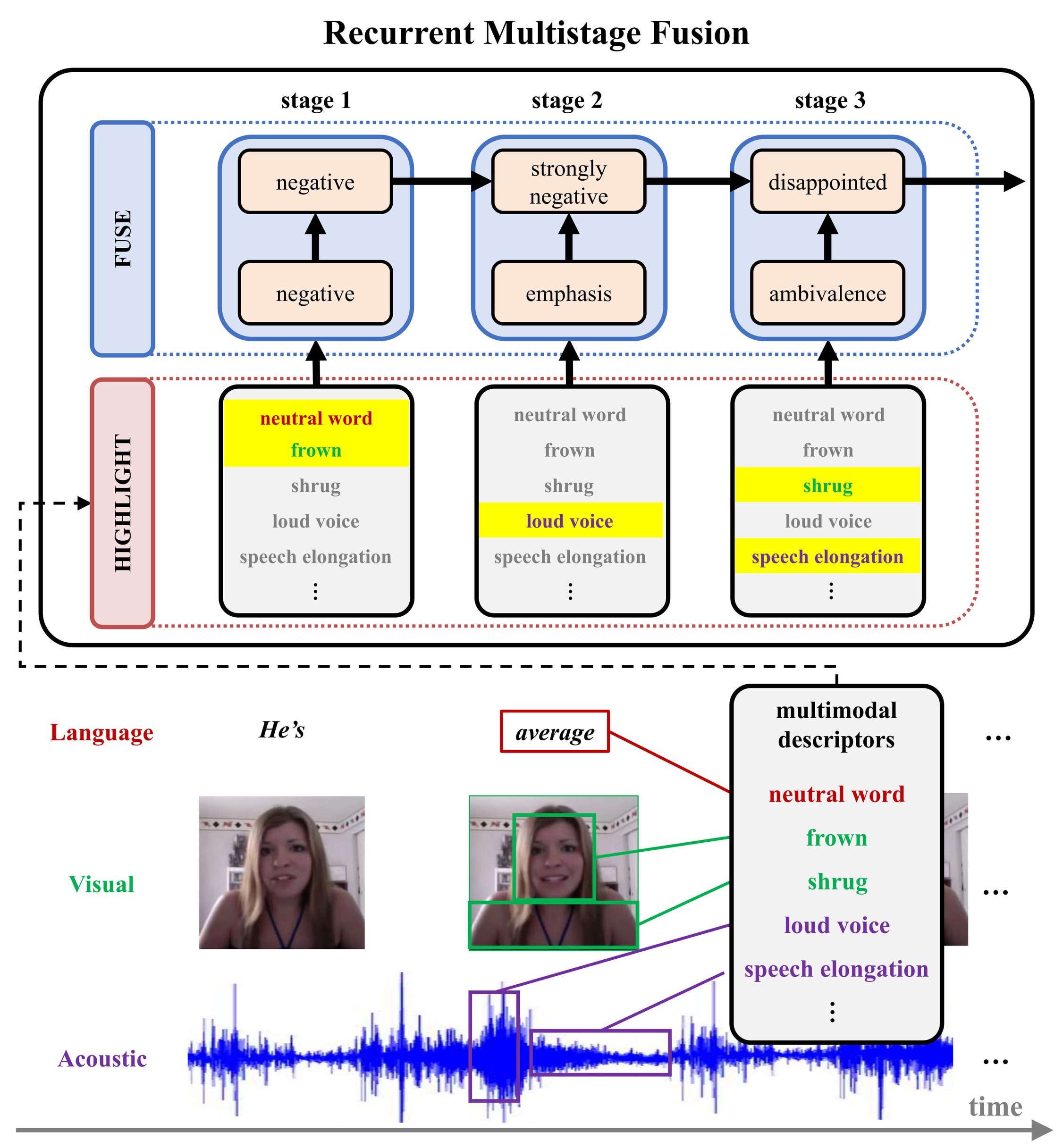
a) Multiple co-occurring interactions

b) Different weighted combinations



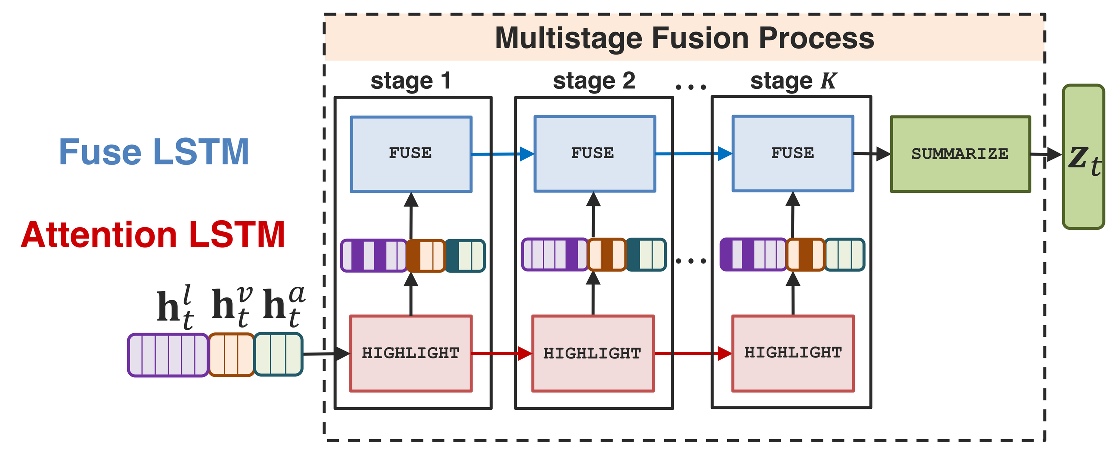
Recurrent Multistage Fusion





Recurrent Multistage Fusion Network

Multistage Fusion Process



HIGHLIGHT: At each stage k, a subset of the multimodal signals represented in $\mathbf{h}_t$ will be automatically highlighted for fusion.

$$\mathbf{a}_t^{[k]} = f_H(\mathbf{h}_t ; \mathbf{a}_t^{[1:k-1]}, \Theta) \tag{1}$$

$$\tilde{\mathbf{h}}_t^{[k]} = \mathbf{h}_t \odot \mathbf{a}_t^{[k]} \tag{2}$$

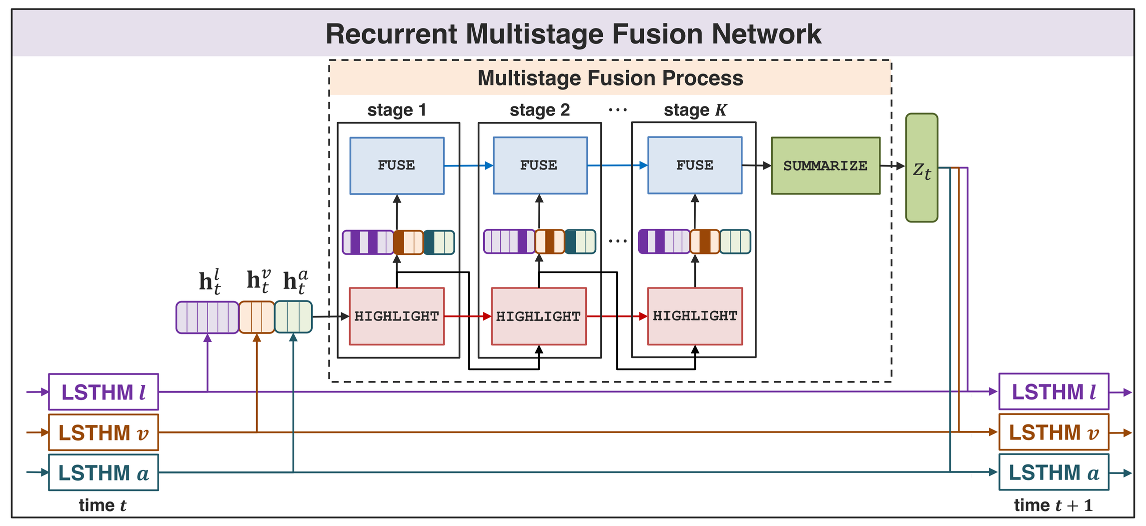
FUSE: The highlighted multimodal signals are simultaneously fused in a local fusion and then integrated with fusion representations from previous stages.

$$\mathbf{s}_t^{[k]} = f_F(\tilde{\mathbf{h}}_t^{[k]} ; \mathbf{s}_t^{[1:k-1]}, \Theta) \tag{3}$$

SUMMARIZE: After completing K stages of HIGHLIGHT and FUSE, the SUMMARIZE operation generates a cross-modal representation using all final fusion representations $\mathbf{s}_t^{[1:K]}$.

$$\mathbf{z}_t = \mathcal{S}(\mathbf{s}_t^{[1:K]} ; \Theta) \tag{4}$$

Recurrent Multistage Fusion Network



Experiments

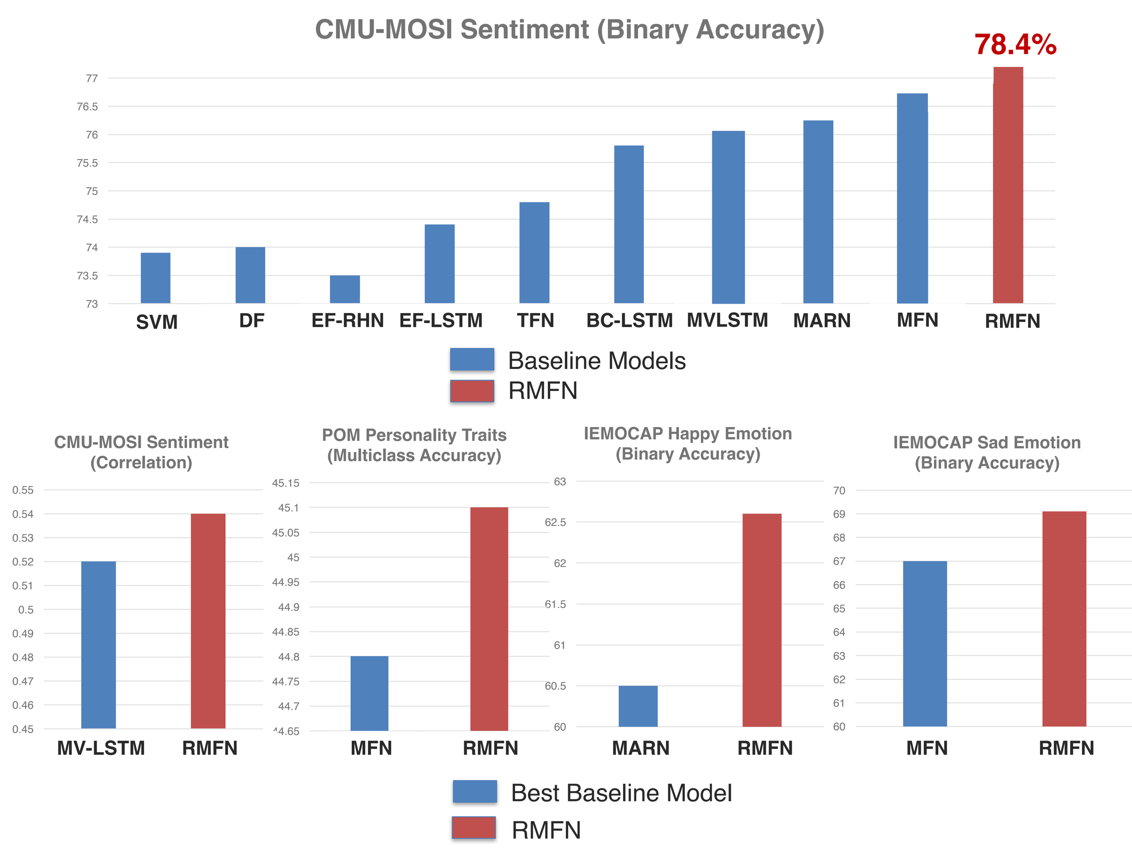
Datasets

- Multimodal Sentiment Analysis: CMU-MOSI
- Multimodal Emotion Recognition: IEMOCAP
- Multimodal Personality Traits Prediction: POM
- Language, visual and acoustic features extracted and aligned by P2FA.

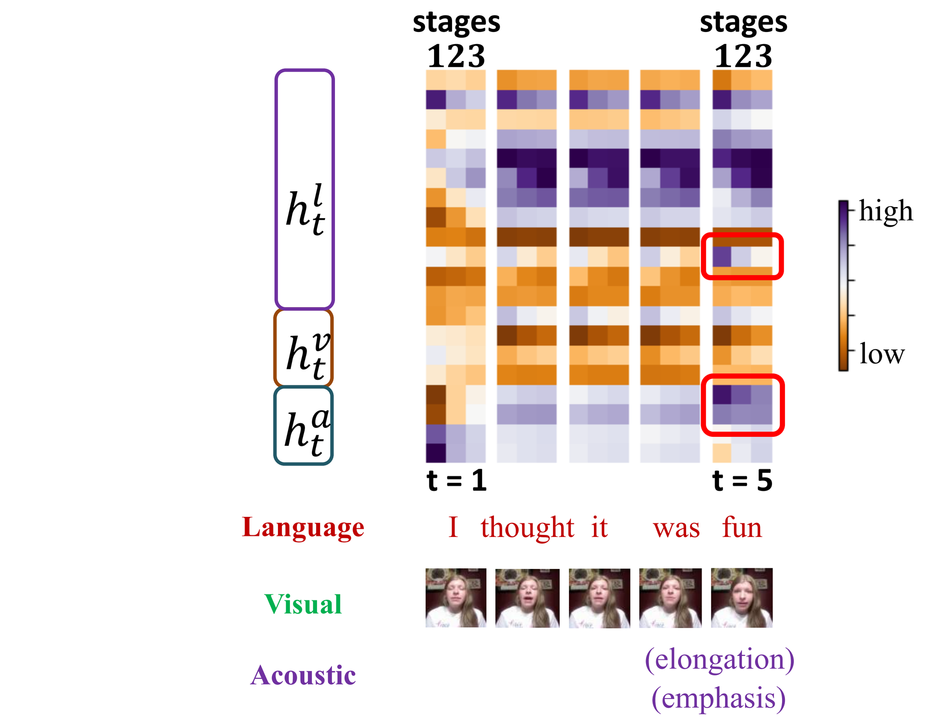
Baseline Models

- Non-temporal Models
- Multimodal Temporal Graphical Models
- Multimodal Temporal Neural Networks

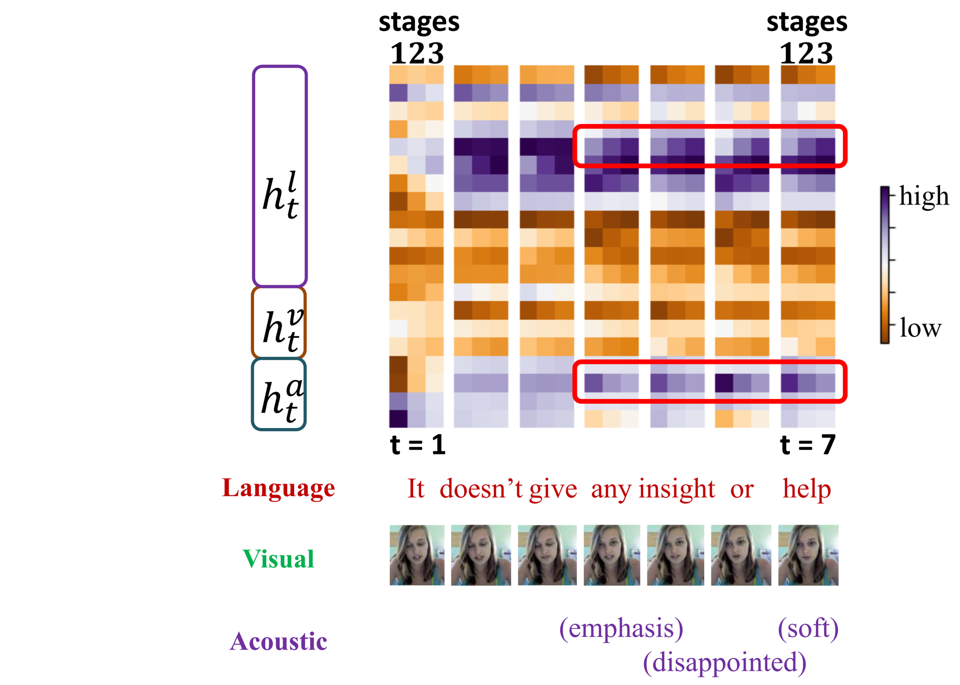
Results



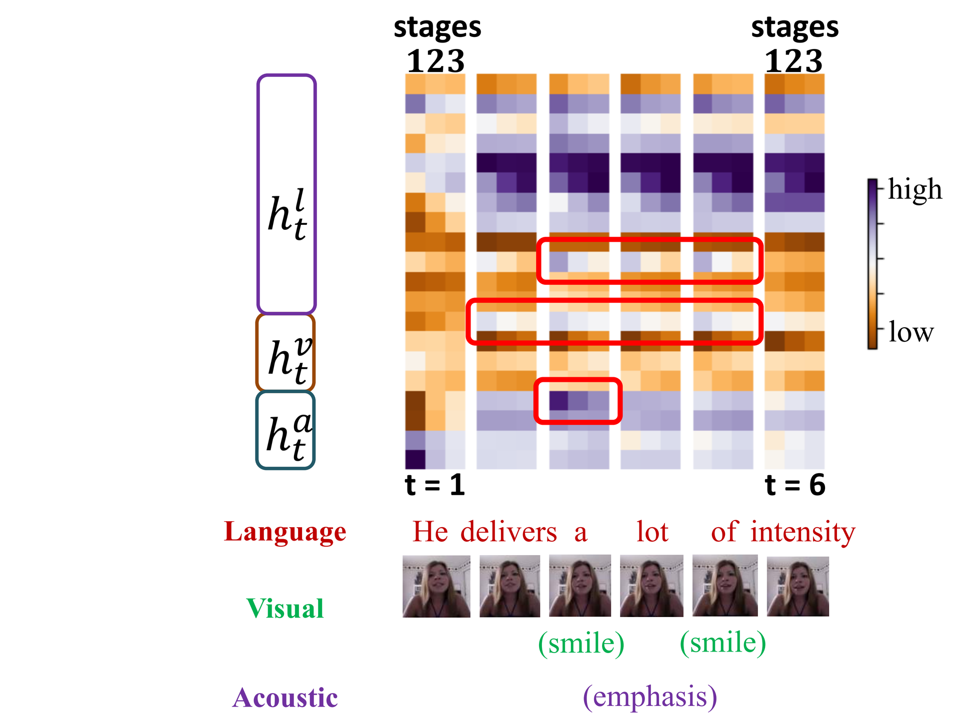
Synchronized Interactions



Bimodal Interactions



Asynchronous Trimodal Interactions



Visualizations

- Attention weights change across **multiple stages of fusion**.
- Attention weights **vary over time** and adapt to the multimodal inputs.
- **Language and acoustic** modalities most commonly highlighted.

Conclusion

RMFN decomposes the multimodal fusion problem into multiple stages, each focused on a subset of multimodal signals. Multiple stages coordinate to capture both **synchronous and asynchronous** multimodal interactions.