Carnegie Mellon University

Language Technologies Institute

MultiComp Lab

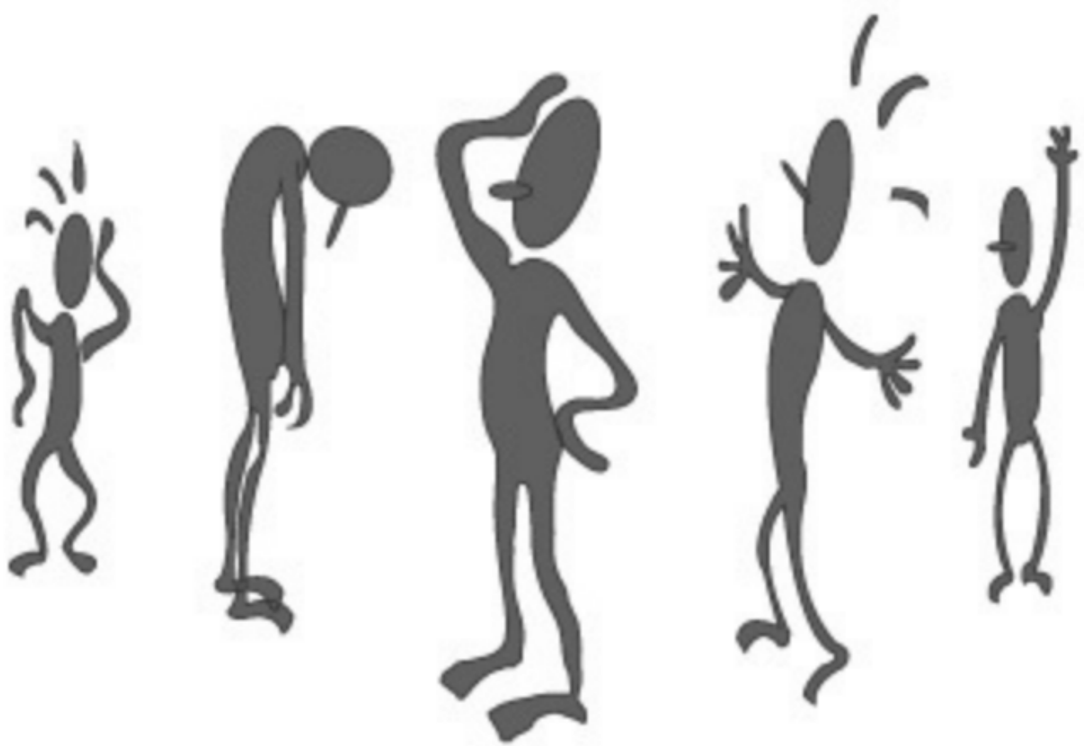
Efficient Low-rank Multimodal Fusion With Modality-specific Factors

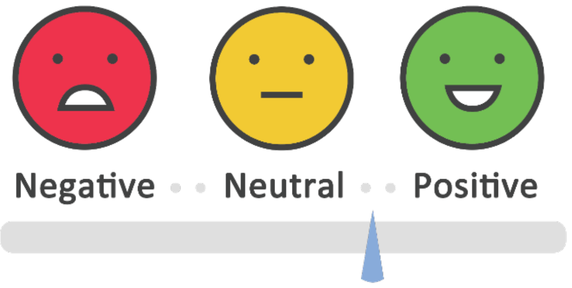
Zhun Liu, Ying Shen,

Varun Bharadwaj, Paul Pu Liang,

Amir Zadeh, Louis-Philippe Morency

Artificial Intelligence

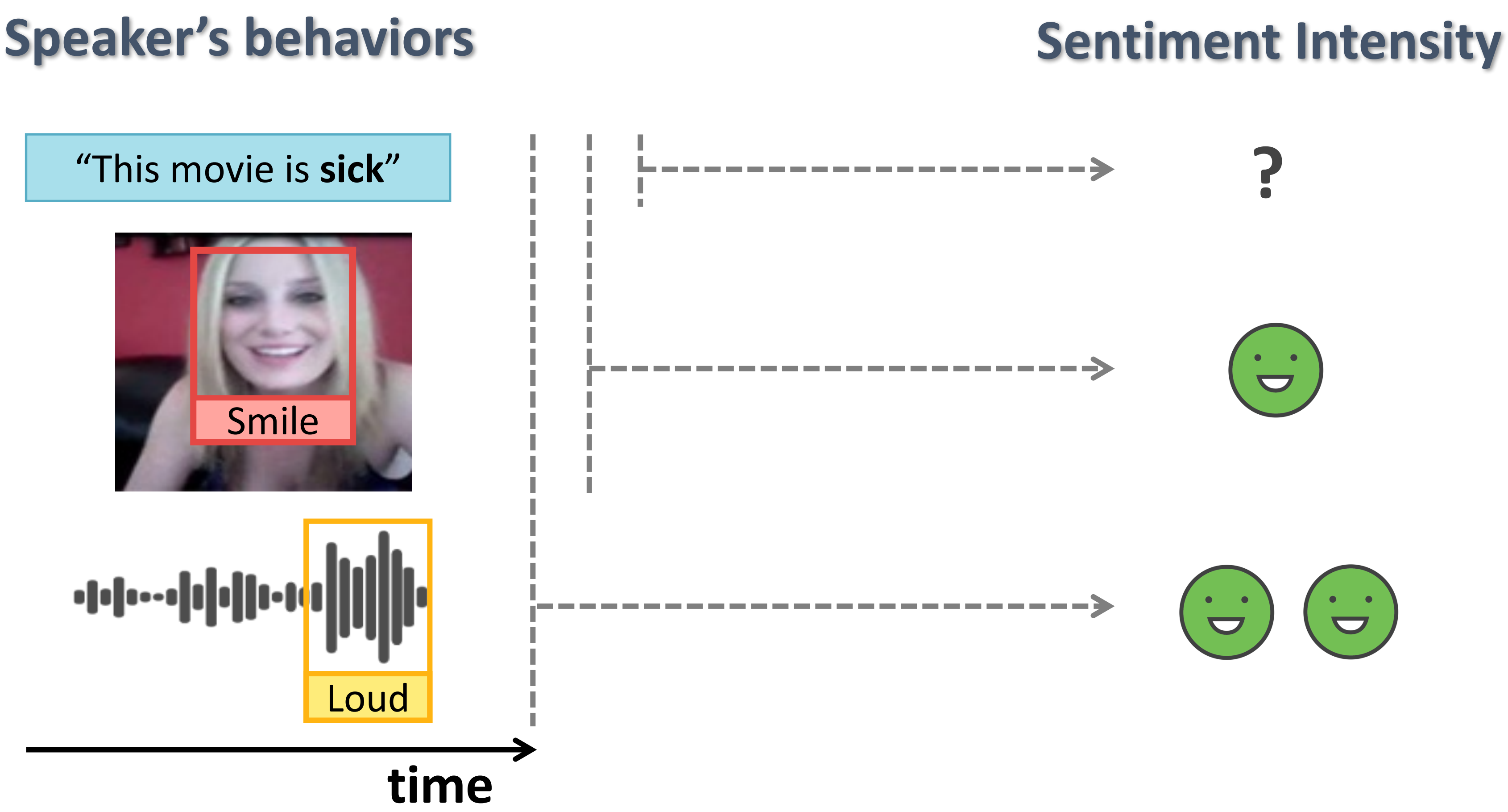




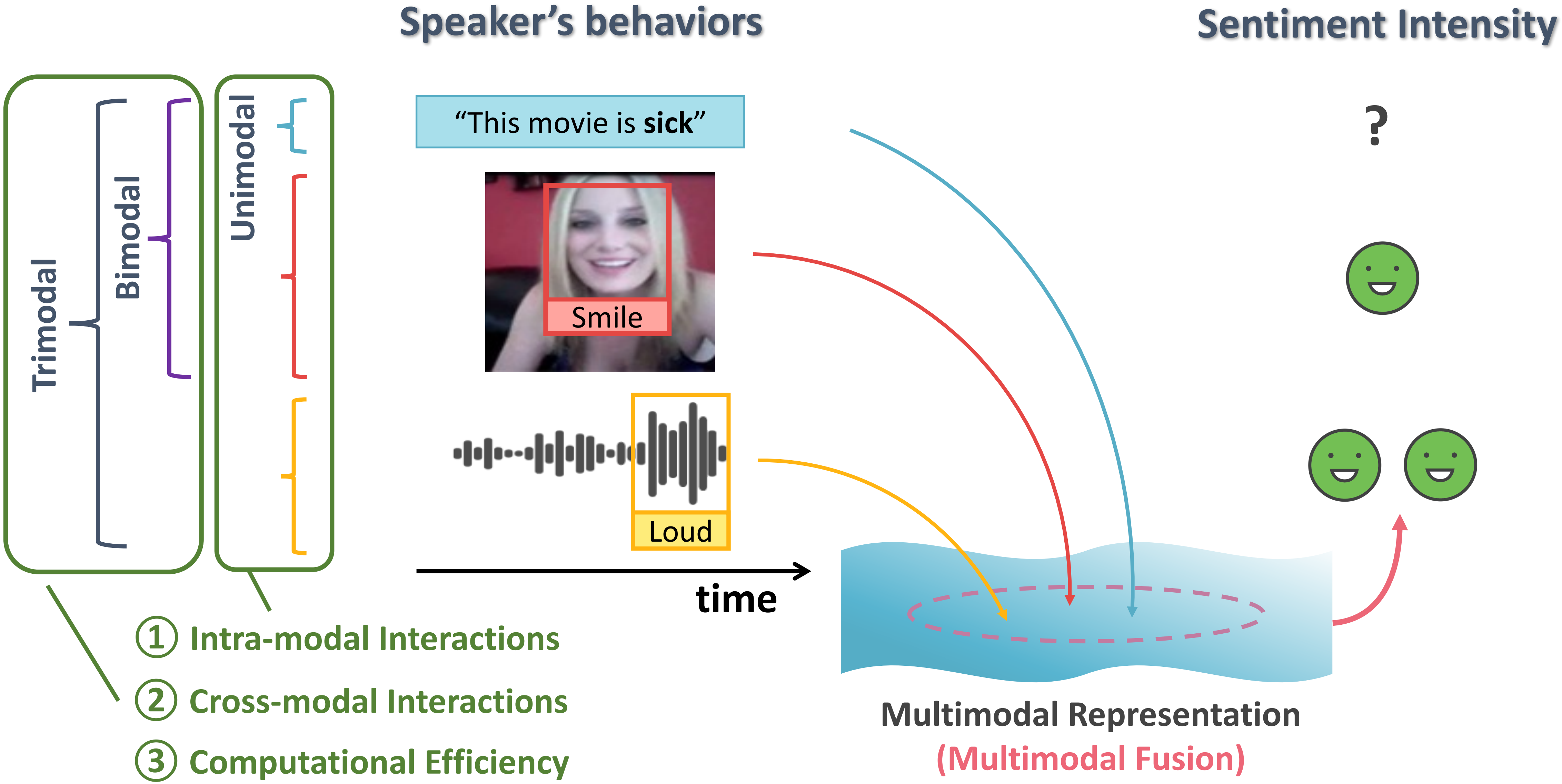
Hey Siri

Cortana

Sentiment and Emotion Analysis



Multimodal Sentiment and Emotion Analysis

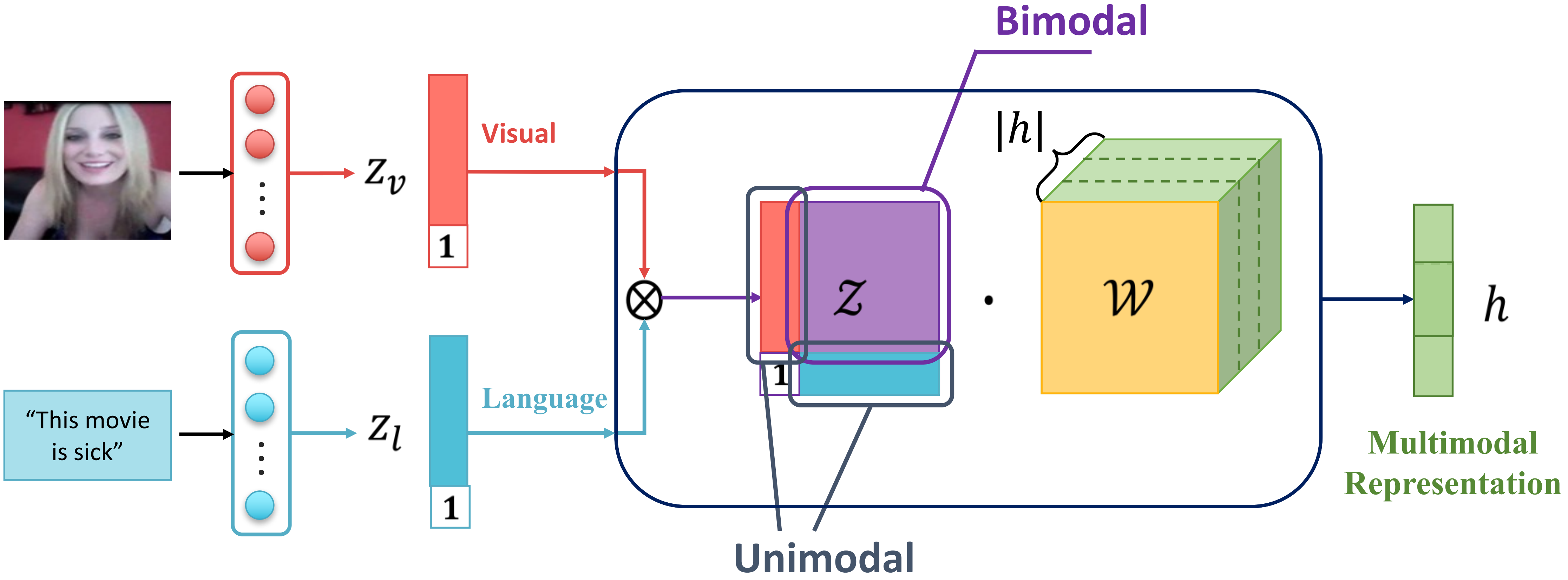


① Intra-modal Interactions

② Cross-modal Interactions

③ Computational Efficiency

Multimodal Fusion using Tensor Representation

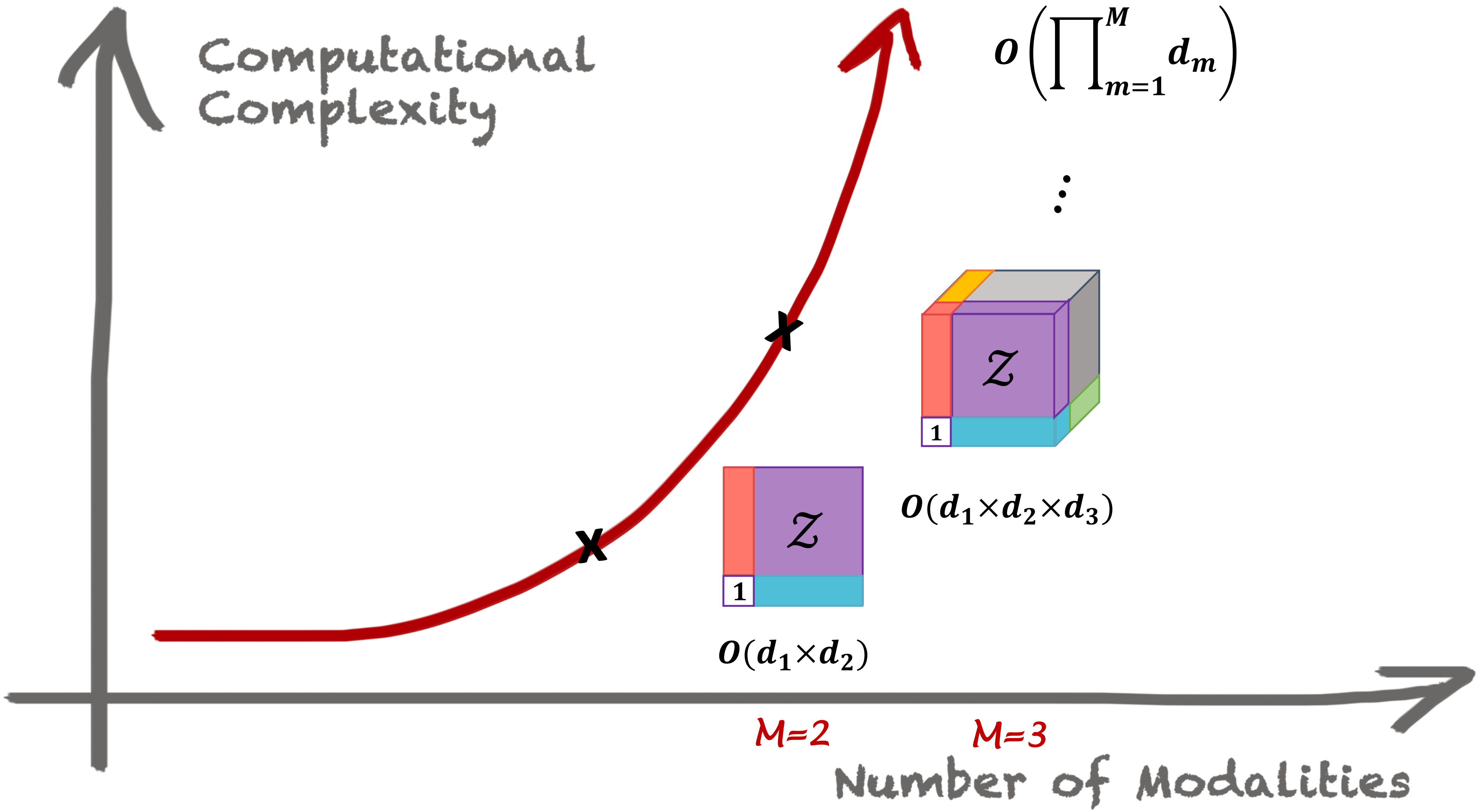


- ☑ Intra-modal interactions
- ☑ Cross-modal interactions
- ☒ Computational efficiency

$$\mathcal{Z} = \begin{bmatrix} z_v \\ 1 \end{bmatrix} \otimes \begin{bmatrix} z_l \\ 1 \end{bmatrix} = \begin{bmatrix} \boxed{z_v} & \boxed{z_v \otimes z_l} \\ 1 & \boxed{z_l} \end{bmatrix}$$

"Tensor Fusion Network for Multimodal Sentiment Analysis" by Zadeh, A., et, al. (2017)

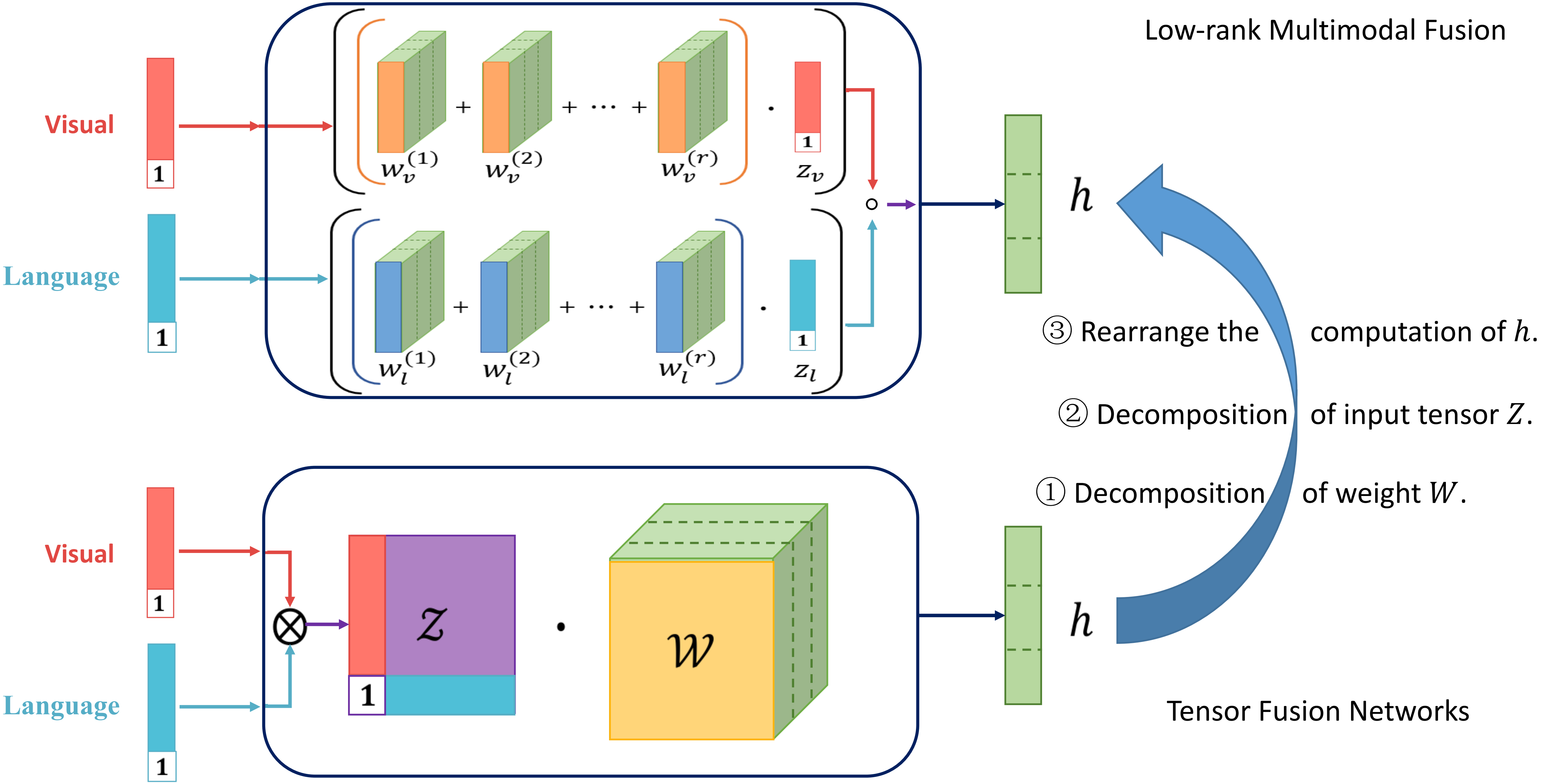
Computational Complexity – Tensor Product



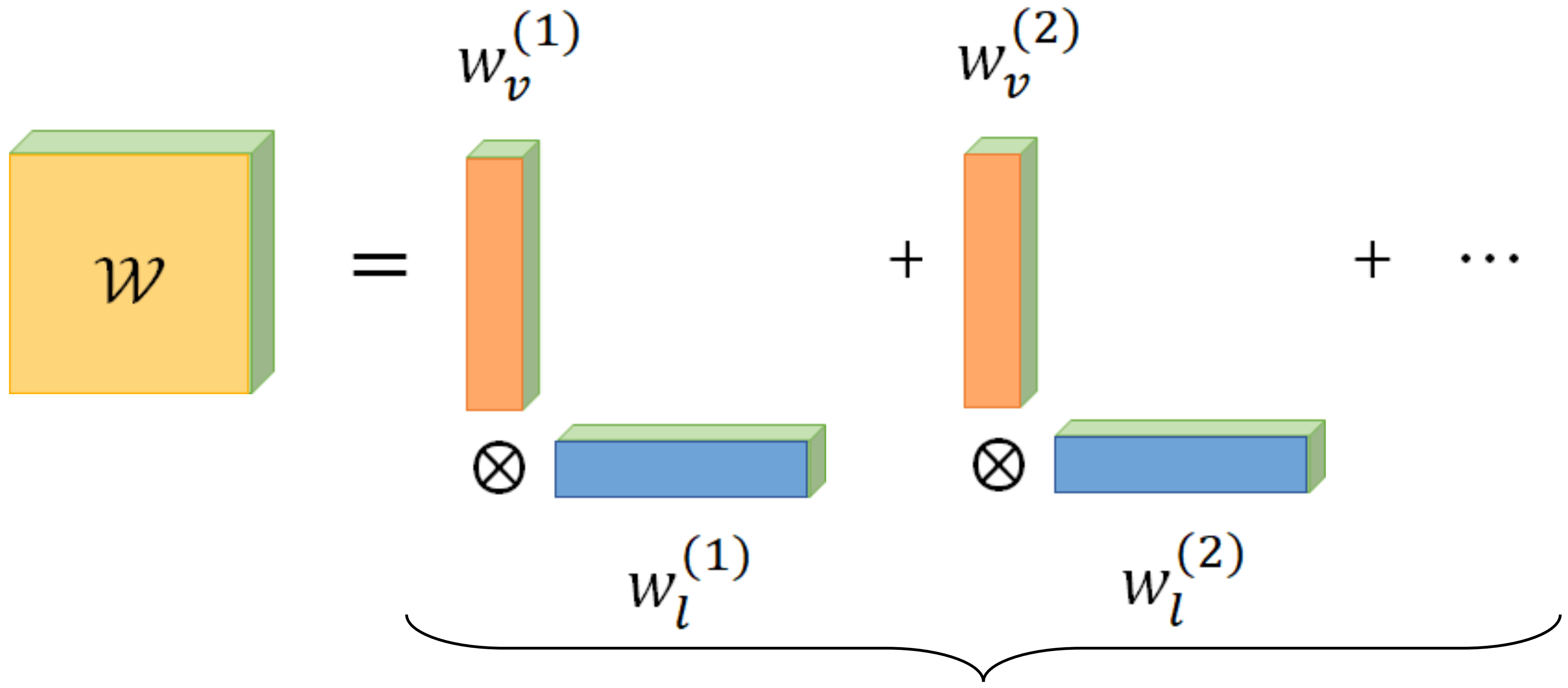
CORE CONTRIBUTIONS

Low-rank Multimodal Fusion (LMF)

From Tensor Representation to Low-rank Fusion

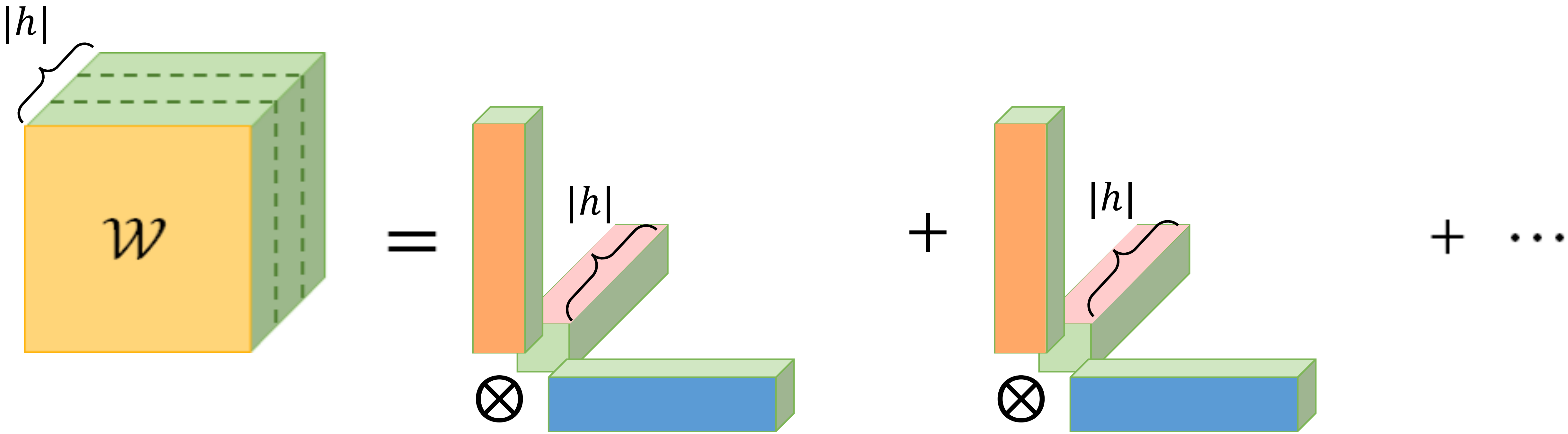


Canonical Polyadic (CP) Decomposition of tensors

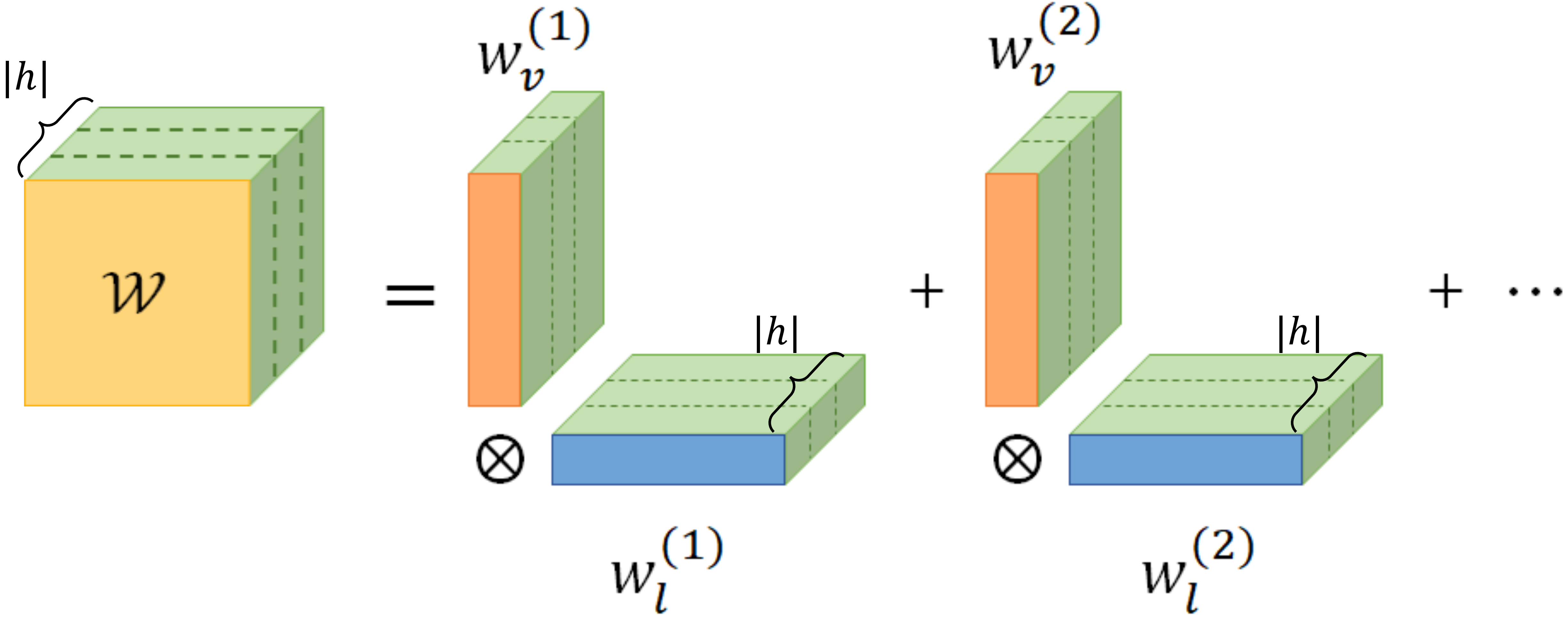


Rank of tensor W: minimum number of vector tuples needed for exact reconstruction

Canonical Polyadic (CP) Decomposition of 3D tensors

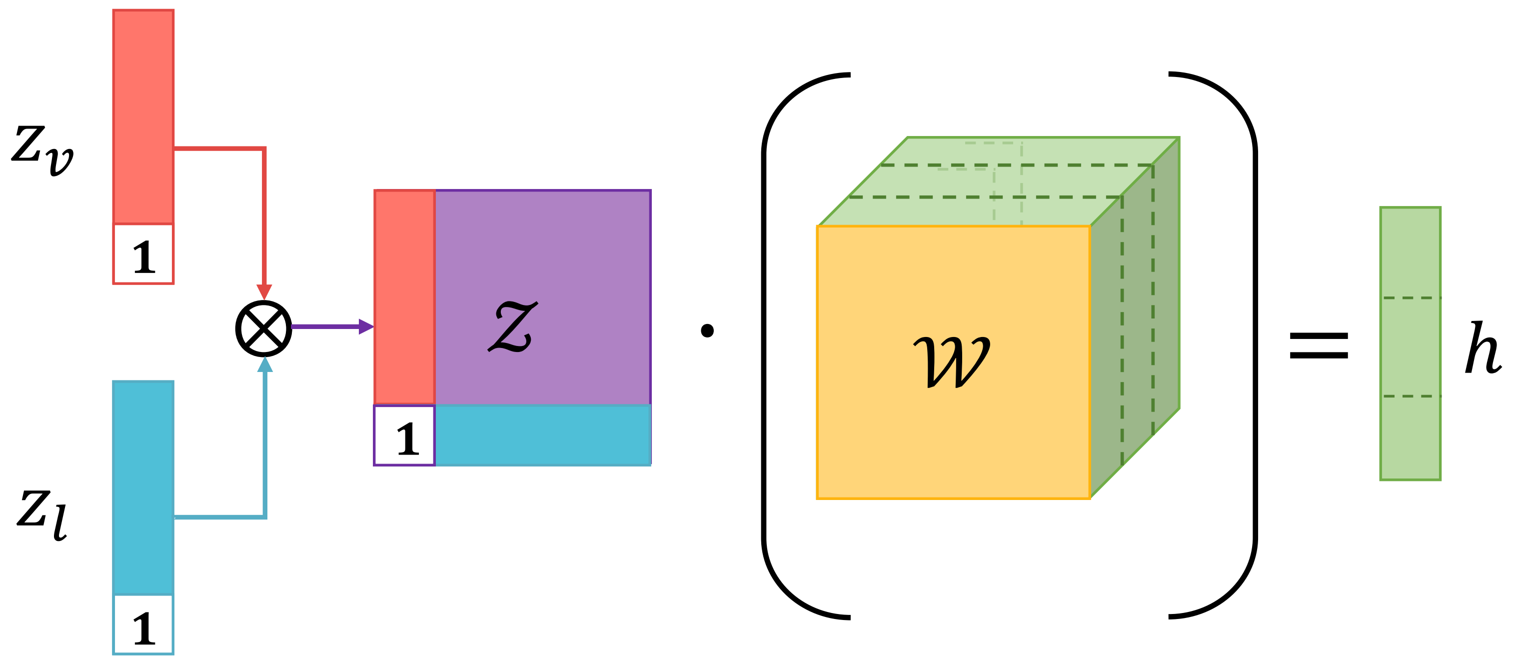


Modality-specific Decomposition

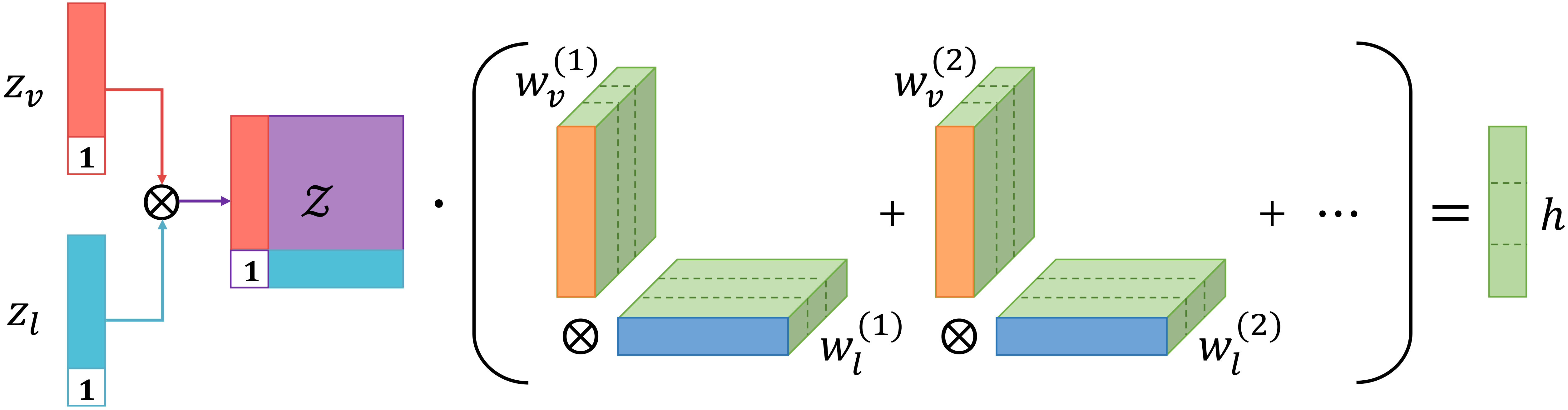


➡ Retain the dimension for the multimodal representation h during decomposition

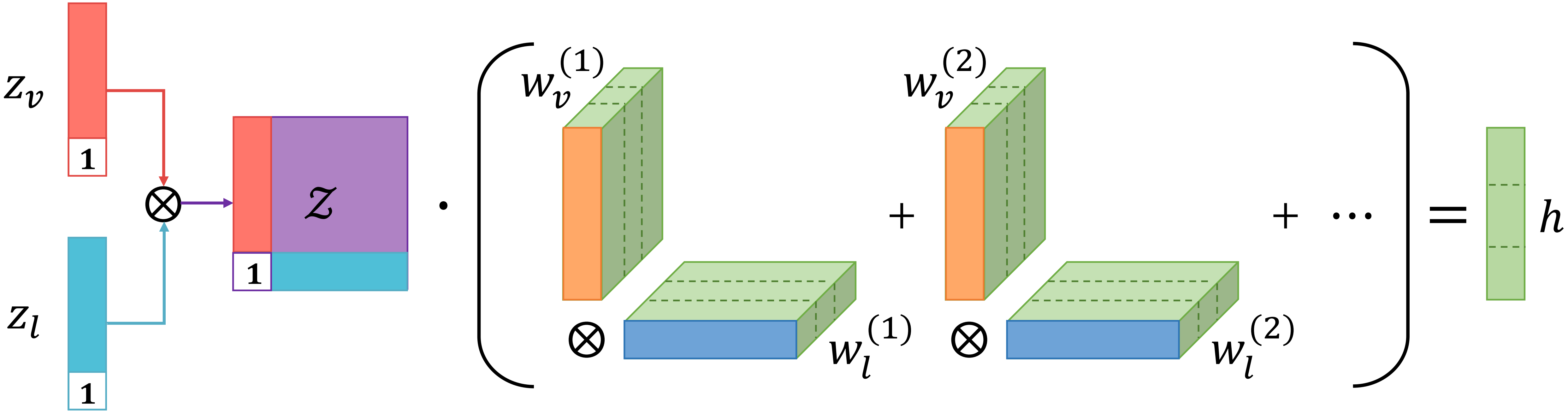
① Decomposition of weight tensor W



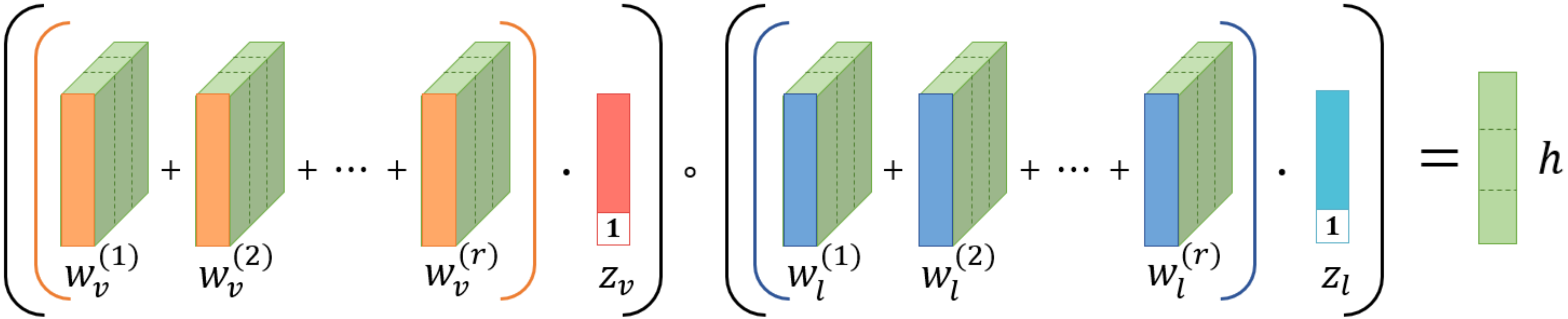
① Decomposition of weight tensor W



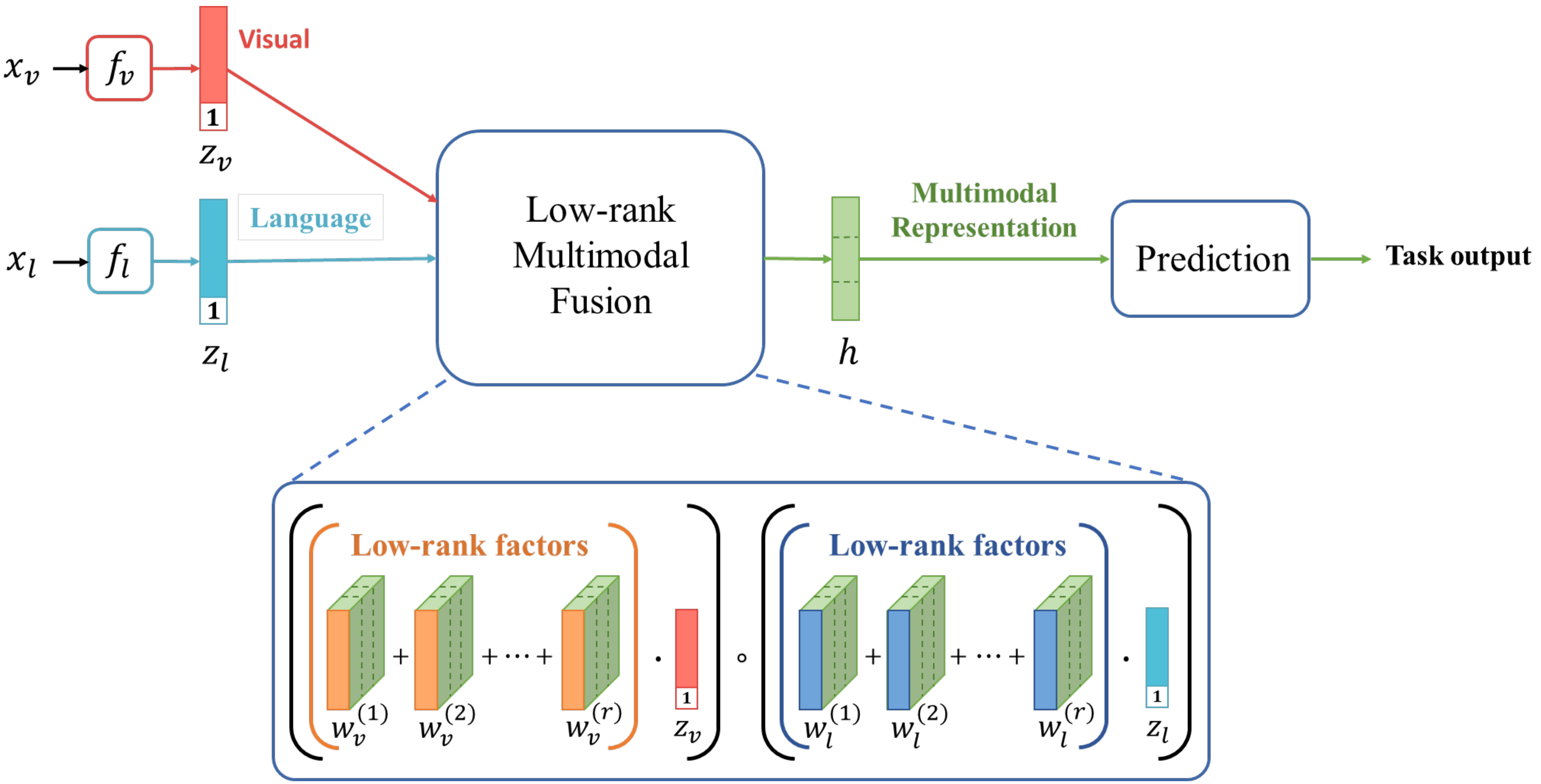
② Decomposition of Z



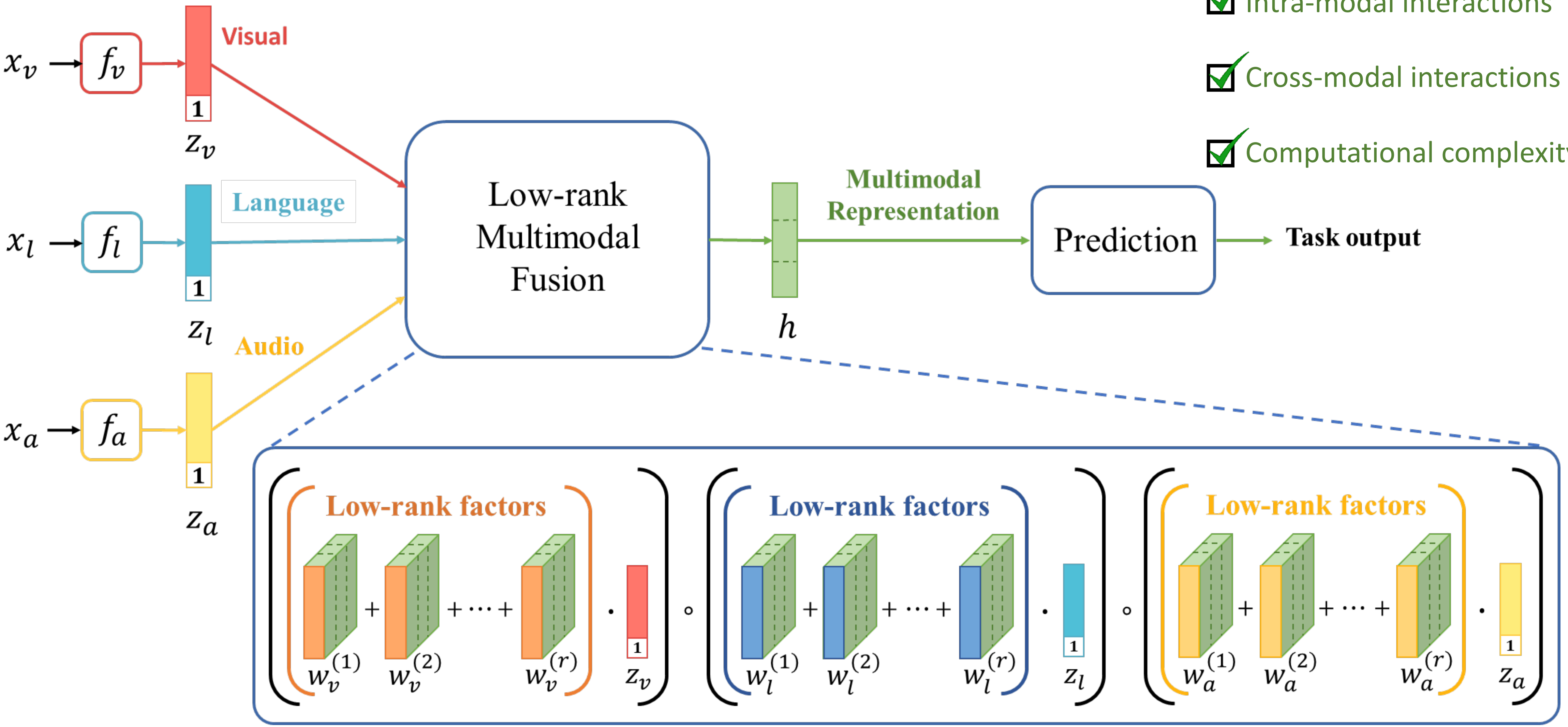
③ Rearranging computation



Low-rank Multimodal Fusion



Easily scales to more modalities

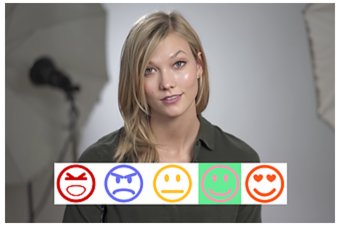


- ☑ Intra-modal interactions
- ☑ Cross-modal interactions
- ☑ Computational complexity

EXPERIMENTS AND RESULTS

Datasets

CMU-MOSI



POM

IEMOCAP

Sentiment Analysis

2199 video segments
- Single-speaker
- From 93 Movie reviews

Segment level annotations
- Sentiment
- Real-valued

Speaker Trait Recognition

1000 full video clips
- Single-speaker
- Movie reviews

Video level annotations
- 16 types of speaker traits
- Categorical annotations

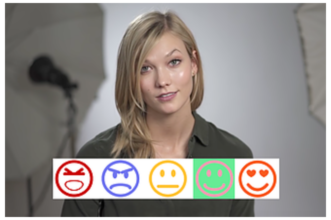
Emotion Recognition

10039 video segments
- Dyadic interaction
- From 302 videos

Segment level annotations
- 10 classes of emotions
- Categorical annotations

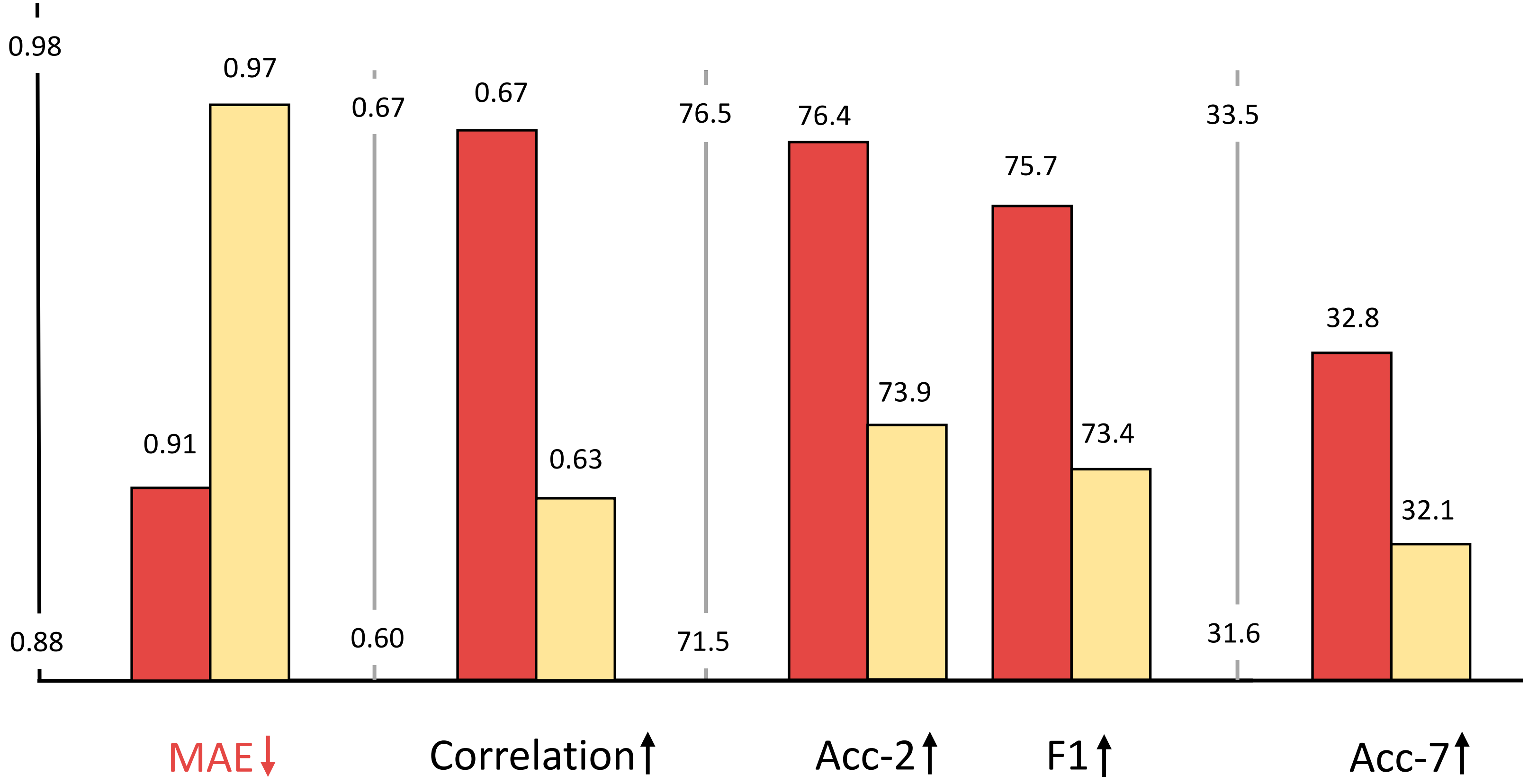
Compare to full rank tensor fusion

CMU-MOSI



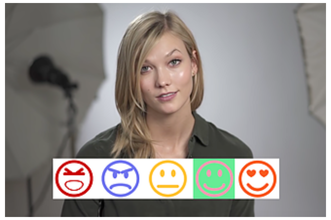
Low-rank Multimodal Fusion (Our Model) LMF

Tensor Fusion Networks (Zadeh, et al., 2017) TFN



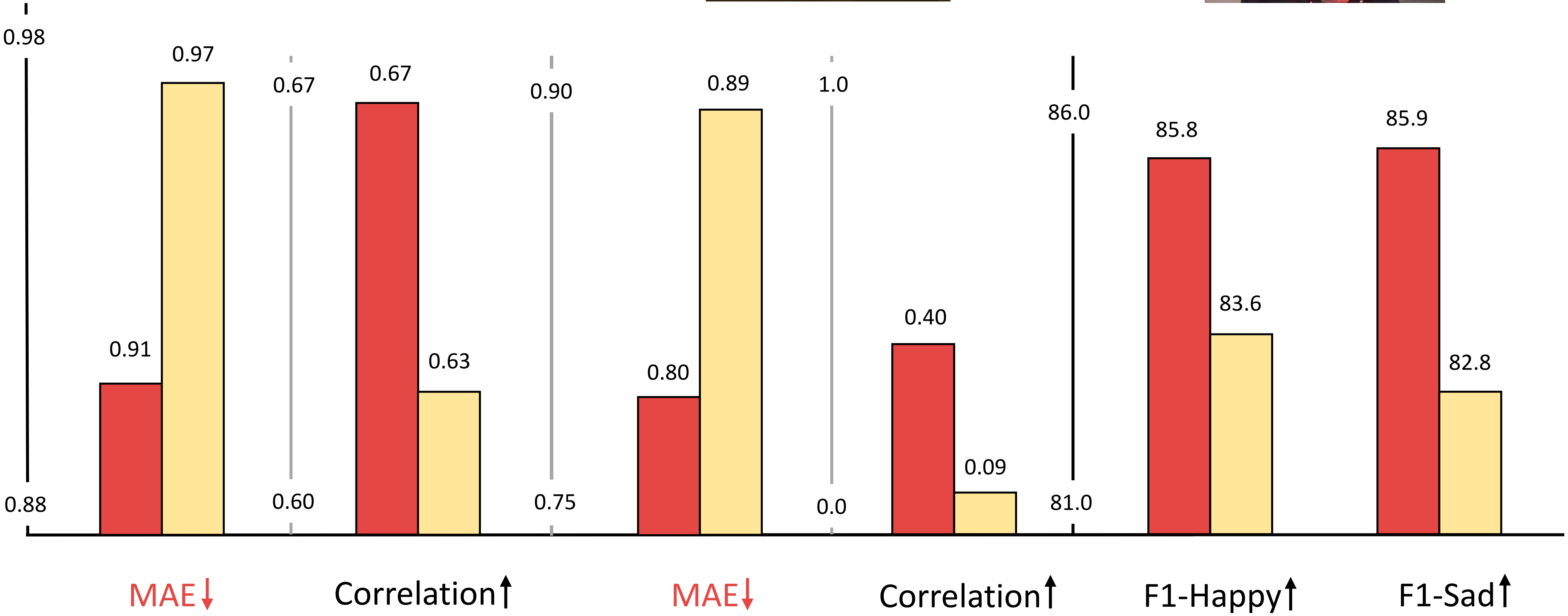
Compare to full rank tensor fusion

CMU-MOSI



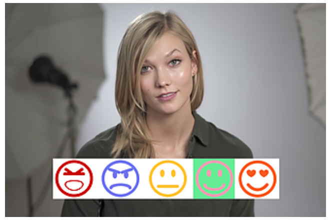
POM

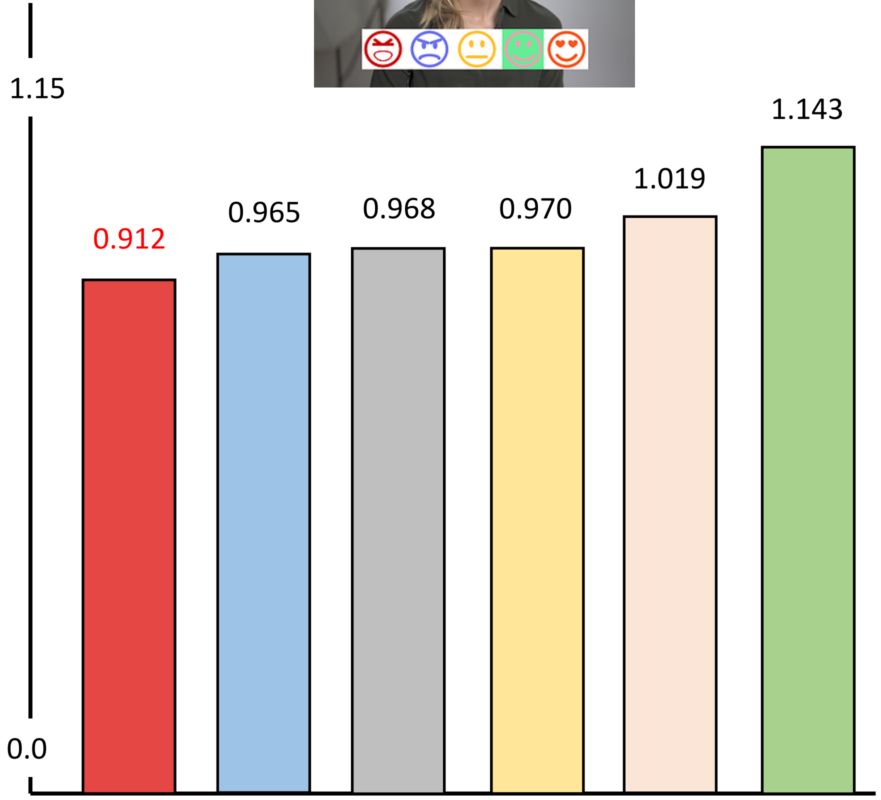
IEMOCAP



Compare with State-of-the-Art Approaches

CMU-MOSI





Mean Average Error (MAE)

Low-rank Multimodal Fusion (our model) — LMF

Memory Fusion Networks (Zadeh, et al., 2018) — MFN

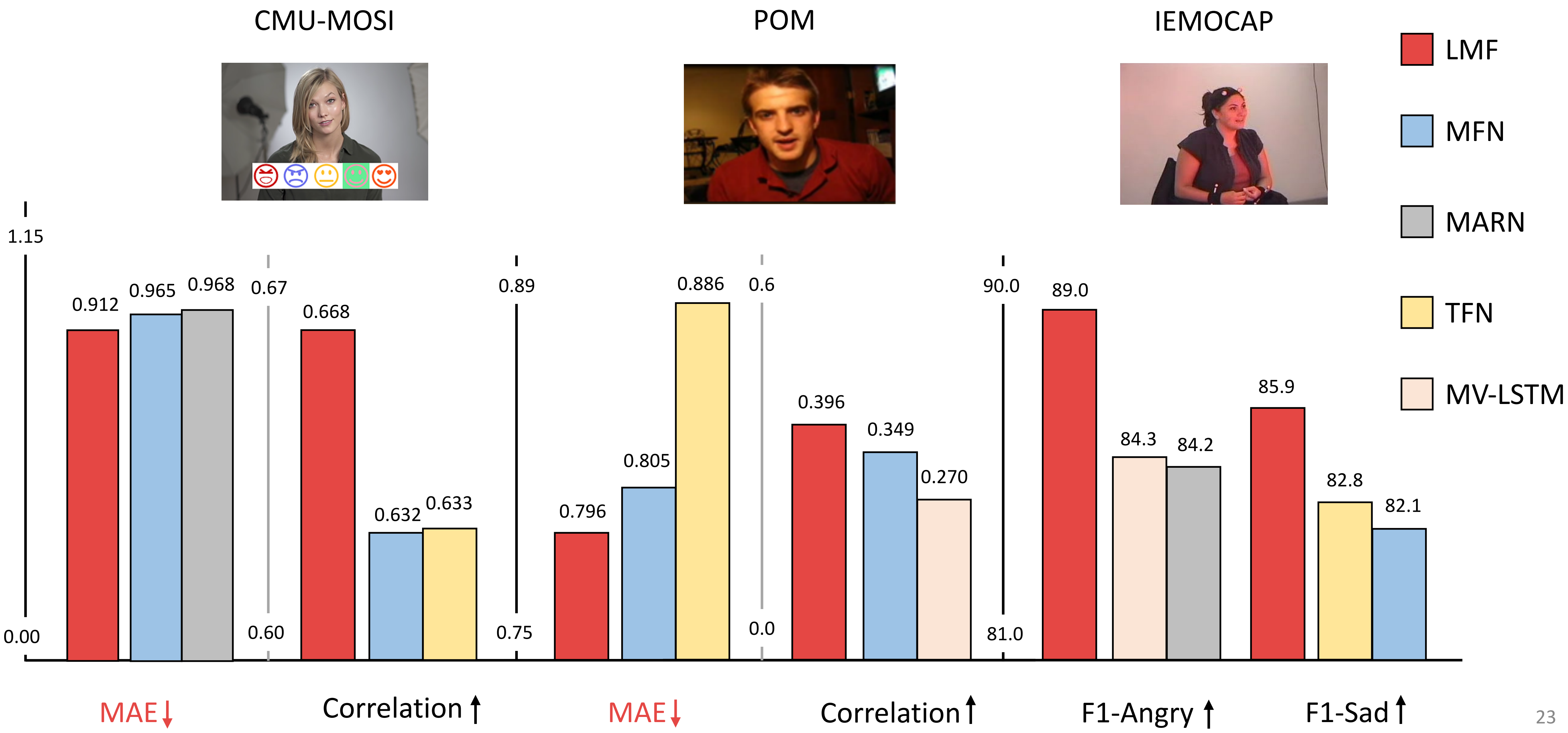
Multi-attention Recurrent Networks (Zadeh, et al., 2018) — MARN

Tensor Fusion Networks (Zadeh, et al., 2017) — TFN

Multi-view LSTM (Rajagopalan, et al., 2016) — MV-LSTM

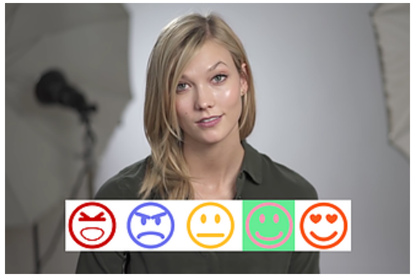
Deep Fusion (Nojavanasghari, et al., 2016) — Deep Fusion

Compare with Top 2 State-of-the-Art Approaches



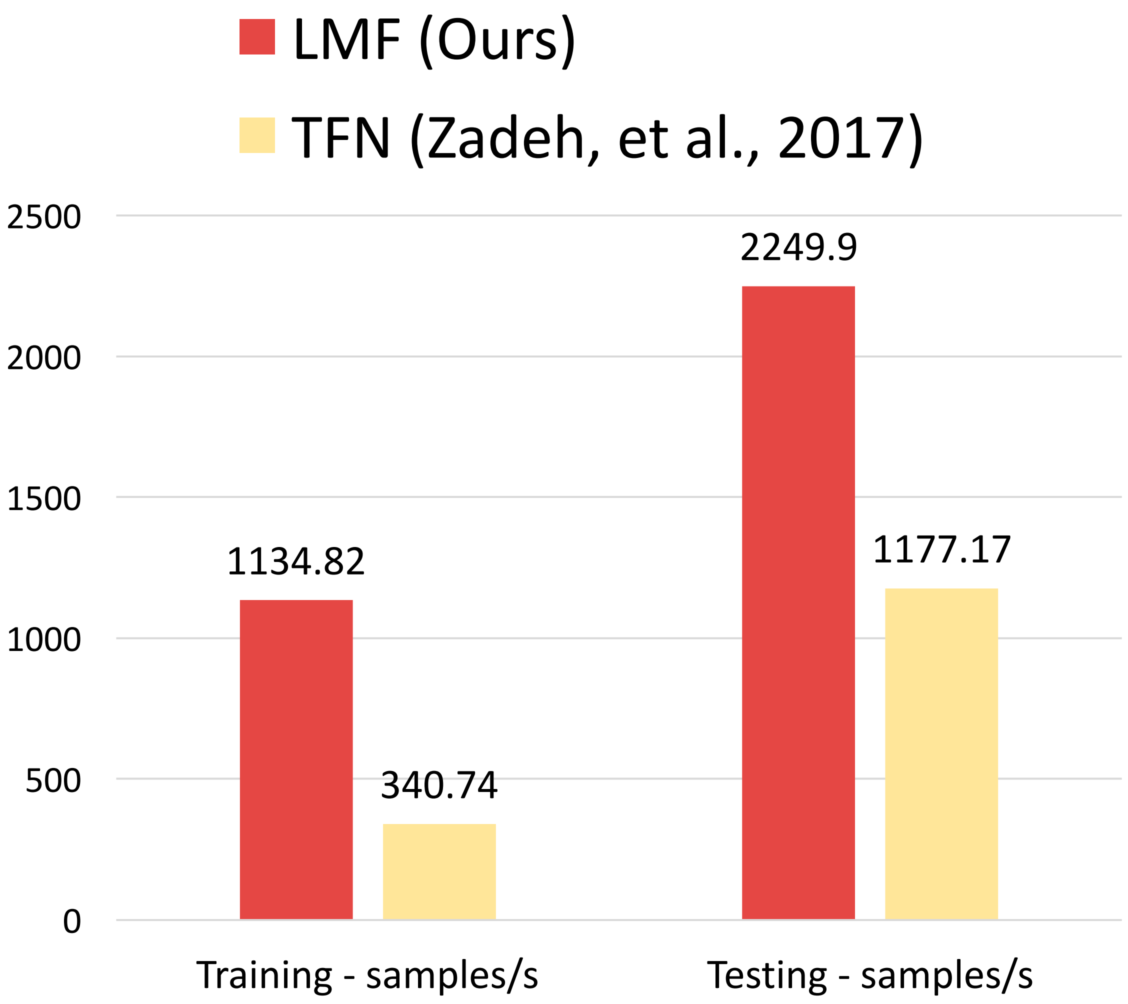
Efficiency Improvement

CMU-MOSI

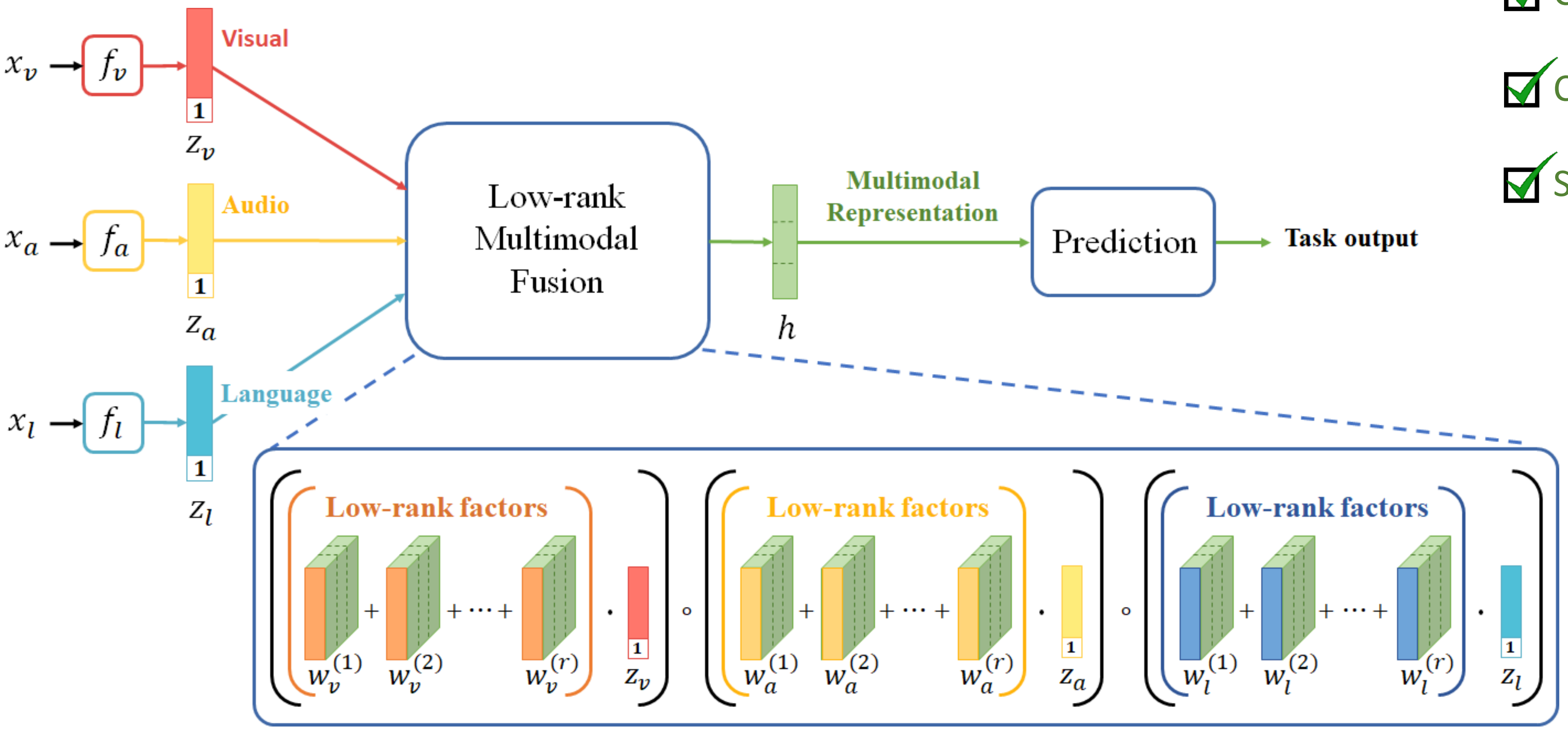


Efficiency Metric: Number of data samples processed per second

- Training Efficiency
- Testing Efficiency



Conclusions



- ☑ Intra-modal interactions
- ☑ Cross-modal interactions
- ☑ Computational complexity
- ☑ State-of-the-art results

Thank you!

Code: https://github.com/Justin1904/Low-rank-Multimodal-Fusion

MultiComp Lab

http://multicomp.cs.cmu.edu/