MACHINE LEARNING DEPARTMENT

Language Technologies Institute

Carnegie Mellon University

Computational Modeling of Human Multimodal Language

Paul Pu Liang

Joint work with: Amir Zadeh, Yao-Hung Hubert Tsai, Zhun Liu, Ying Shen, Hai Pham, Varun Lakshminarasimhan, Edmund Tong, Jon Vanbriessen, Ruslan Salakhutdinov, Louis-Philippe Morency

Contents

- Human Multimodal Language

Contents

- Human Multimodal Language
- 5 directions:
  - Intra-modal and Cross-modal
  - Unimodal, Bimodal and Trimodal
  - Direct and Relative
  - Multimodal Representation Learning
  - Robust Multimodal Representation Learning

Contents

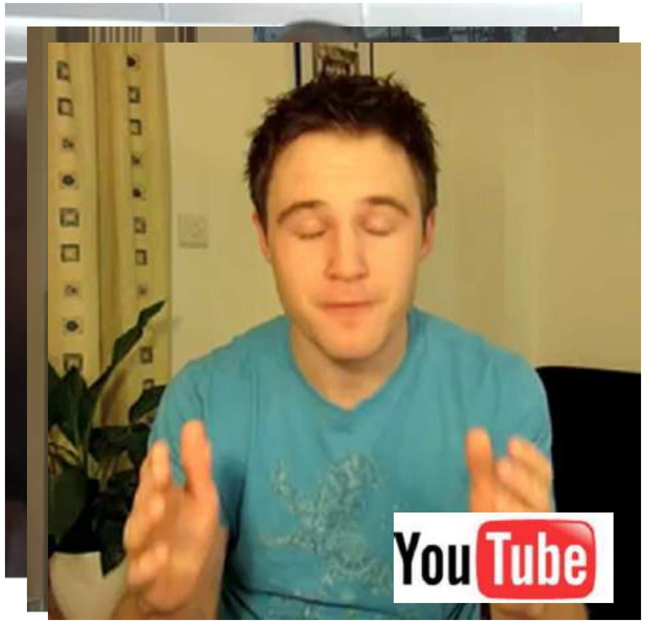
- Human Multimodal Language
- 5 directions:
  - Intra-modal and Cross-modal
  - Unimodal, Bimodal and Trimodal
  - Direct and Relative
  - Multimodal Representation Learning
  - Robust Multimodal Representation Learning
- New Multimodal Dataset: MOSEI

Contents

- Human Multimodal Language
- 5 directions:
  - Intra-modal and Cross-modal
  - Unimodal, Bimodal and Trimodal
  - Direct and Relative
  - Multimodal Representation Learning
  - Robust Multimodal Representation Learning
- New Multimodal Dataset: MOSEI
- Future directions

Progress of Artificial Intelligence

Multimedia Content



Intelligent
Personal Assistants

Robots and
Virtual Agents

Multimodal Communicative Behaviors

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

Direction 1: Intra-modal and Cross-modal

Challenge 1: Intra-modal dynamics

Intra-modal

Speaker's behaviors

"This movie is great" → ++

time

Sentiment Intensity

Challenge 1: Intra-modal dynamics

Intra-modal

Speaker's behaviors

Sentiment Intensity

"This movie is great" → ++

time

Smile | Head nod → +

time

Challenge 1: Intra-modal dynamics

Intra-modal

Speaker's behaviors

"This movie is great"

time

Smile

Head nod

time

Sentiment Intensity

Challenge 2: Cross-modal Dynamics

a) Multiple co-occurring interactions

Speaker's behaviors

Sentiment Intensity

Cross-modal

"This movie is **great**"

Smile

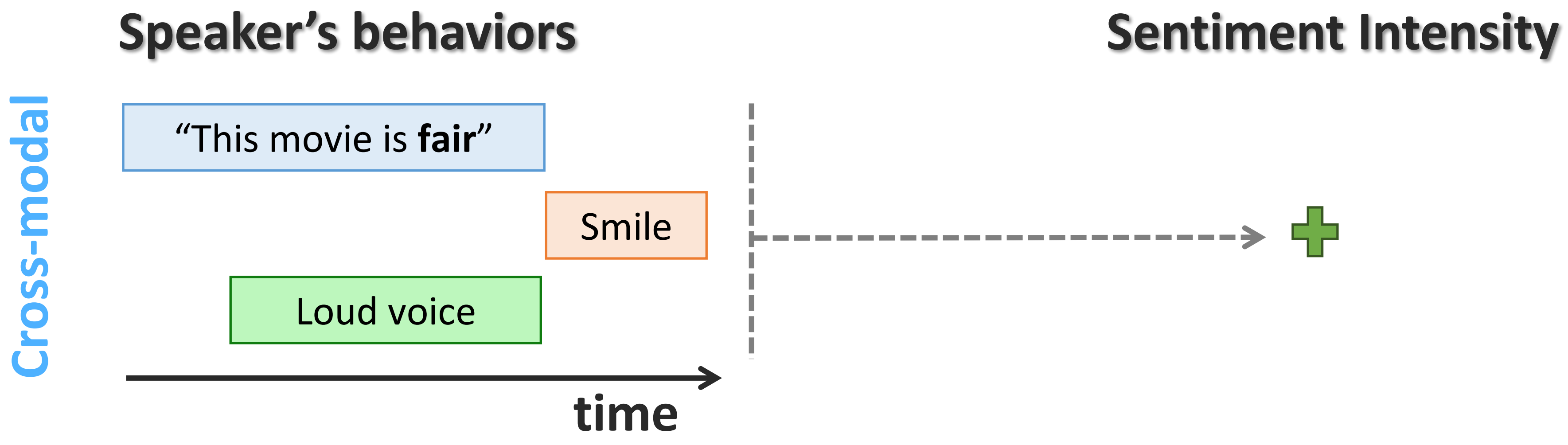
Loud voice

time

Challenge 2: Cross-modal Dynamics

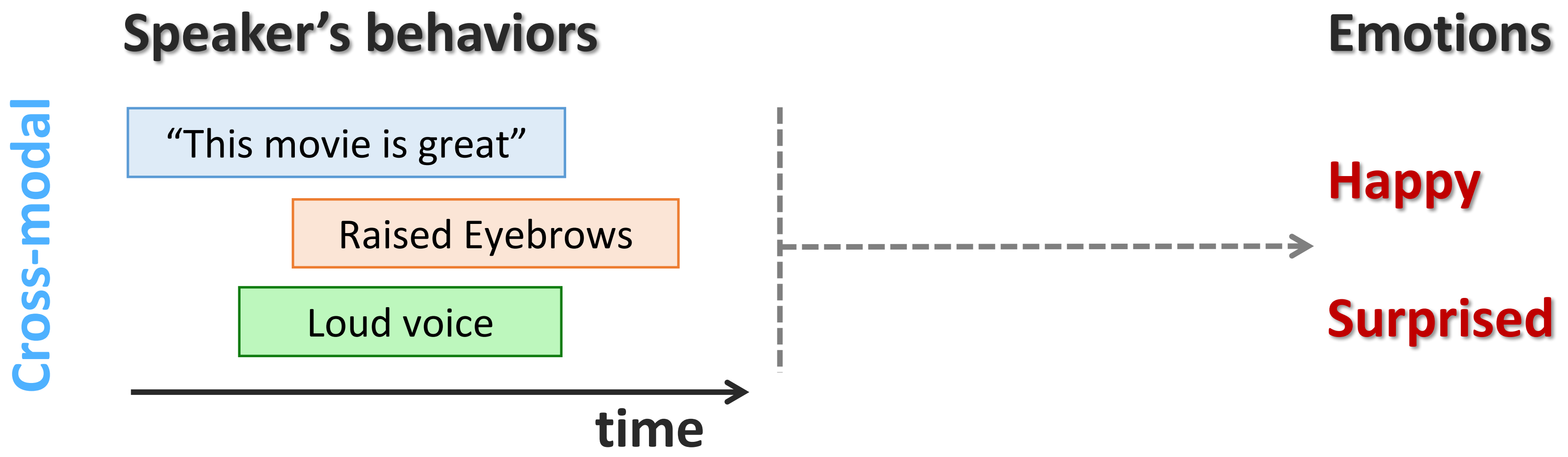
a) Multiple co-occurring interactions

b) Different weighted combinations



Challenge 2: Cross-modal Dynamics

a) Multiple co-occurring interactions
b) Different weighted combinations
c) Multiple prediction targets



Multi-attention Recurrent Network (MARN)

1 **Modeling intra-modal dynamics**

Set of Long-short Term Memories

Multi-attention Recurrent Network (MARN)

1 **Modeling intra-modal dynamics**

→ Set of Long-short Term Memories

2 **Modeling cross-modal dynamics**

→ Set of Long-short Term **Hybrid** Memories + Single-attention Block

Multi-attention Recurrent Network (MARN)

1 **Modeling intra-modal dynamics**

→ Set of Long-short Term Memories

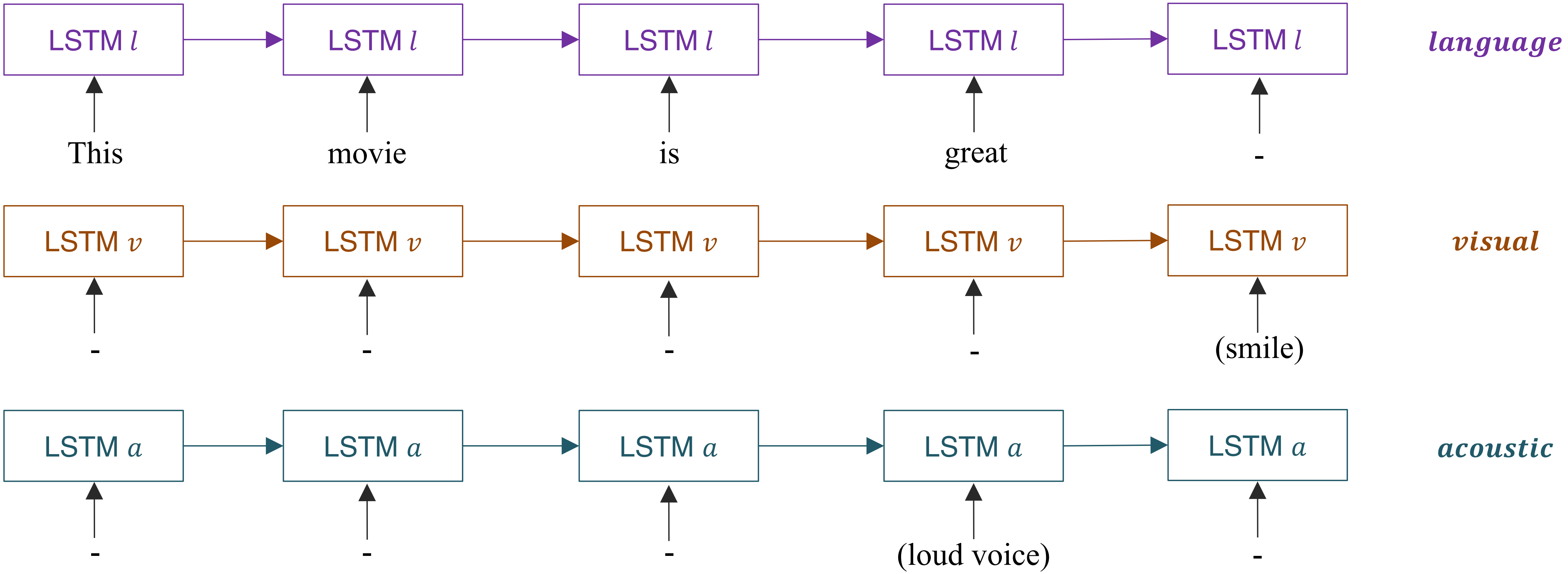
2 **Modeling cross-modal dynamics**

→ Set of Long-short Term **Hybrid** Memories + Single-attention Block

Modeling multiple cross-modal dynamics

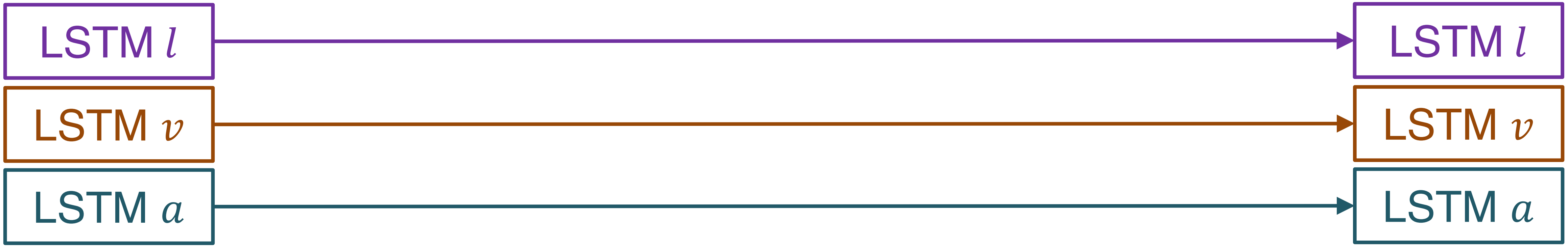
→ Set of Long-short Term **Hybrid** Memories + **Multi-attention** Block

Challenge 1: Intra-modal Dynamics

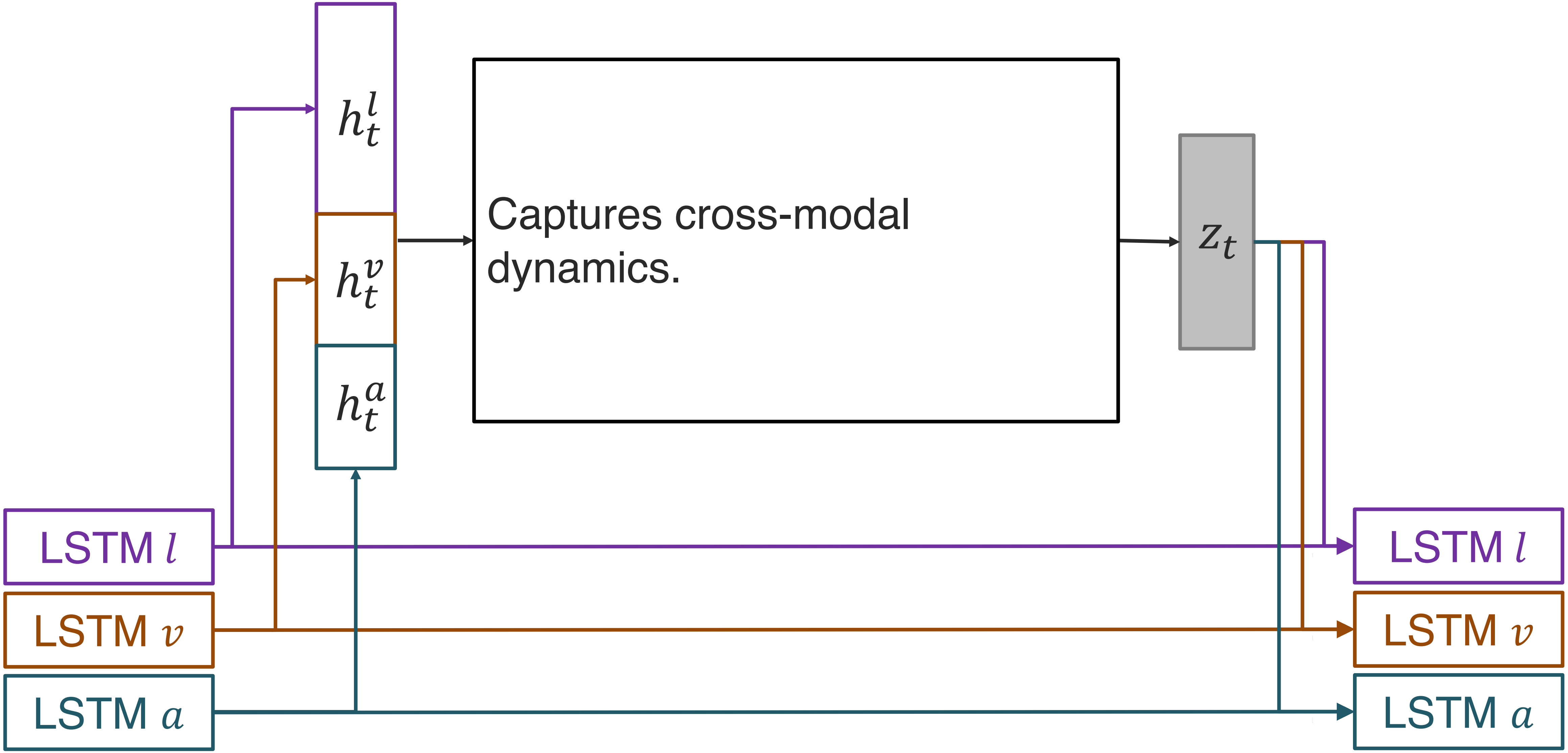


Challenge 2: Cross-modal Dynamics

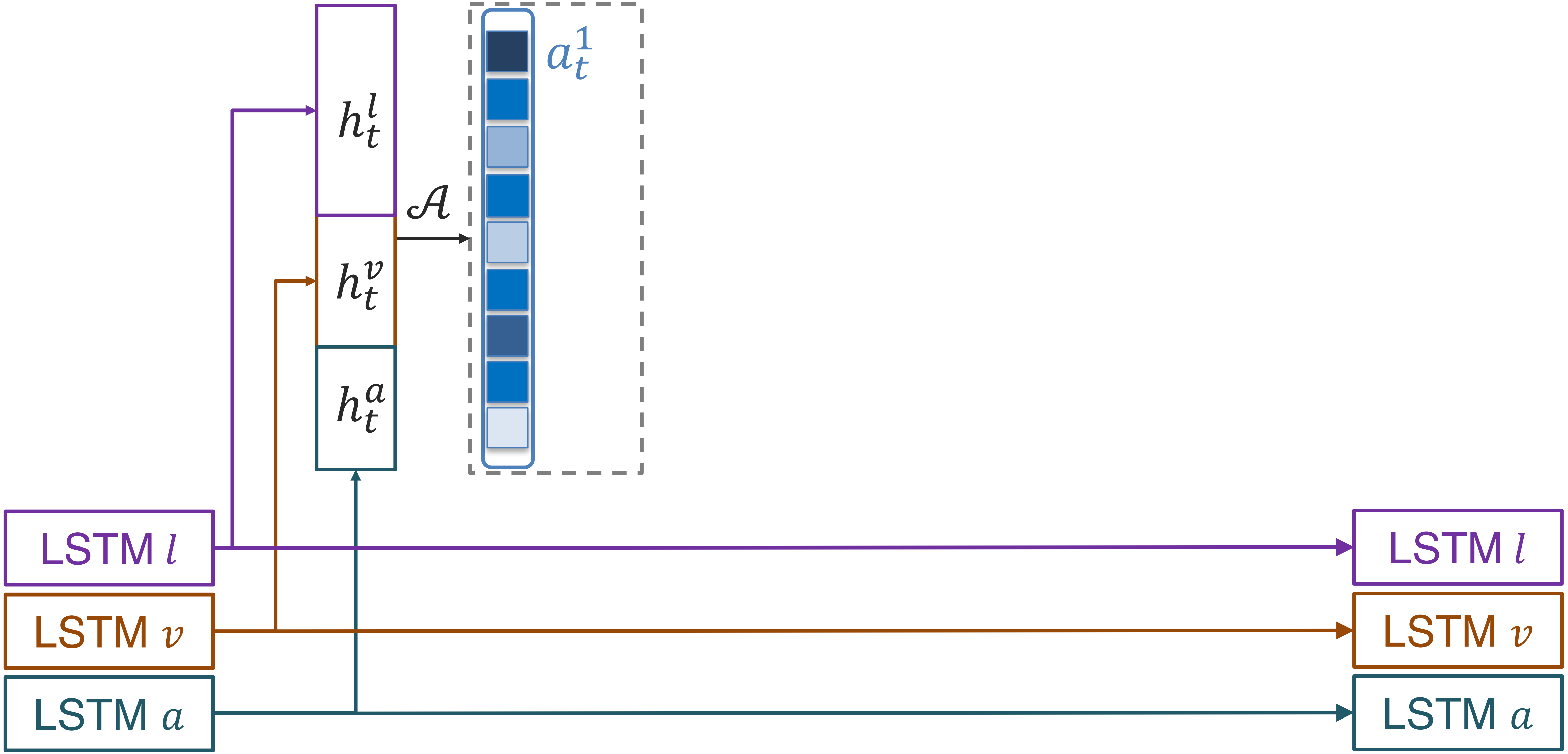
- How do we capture cross-modal dynamics continuously across time?



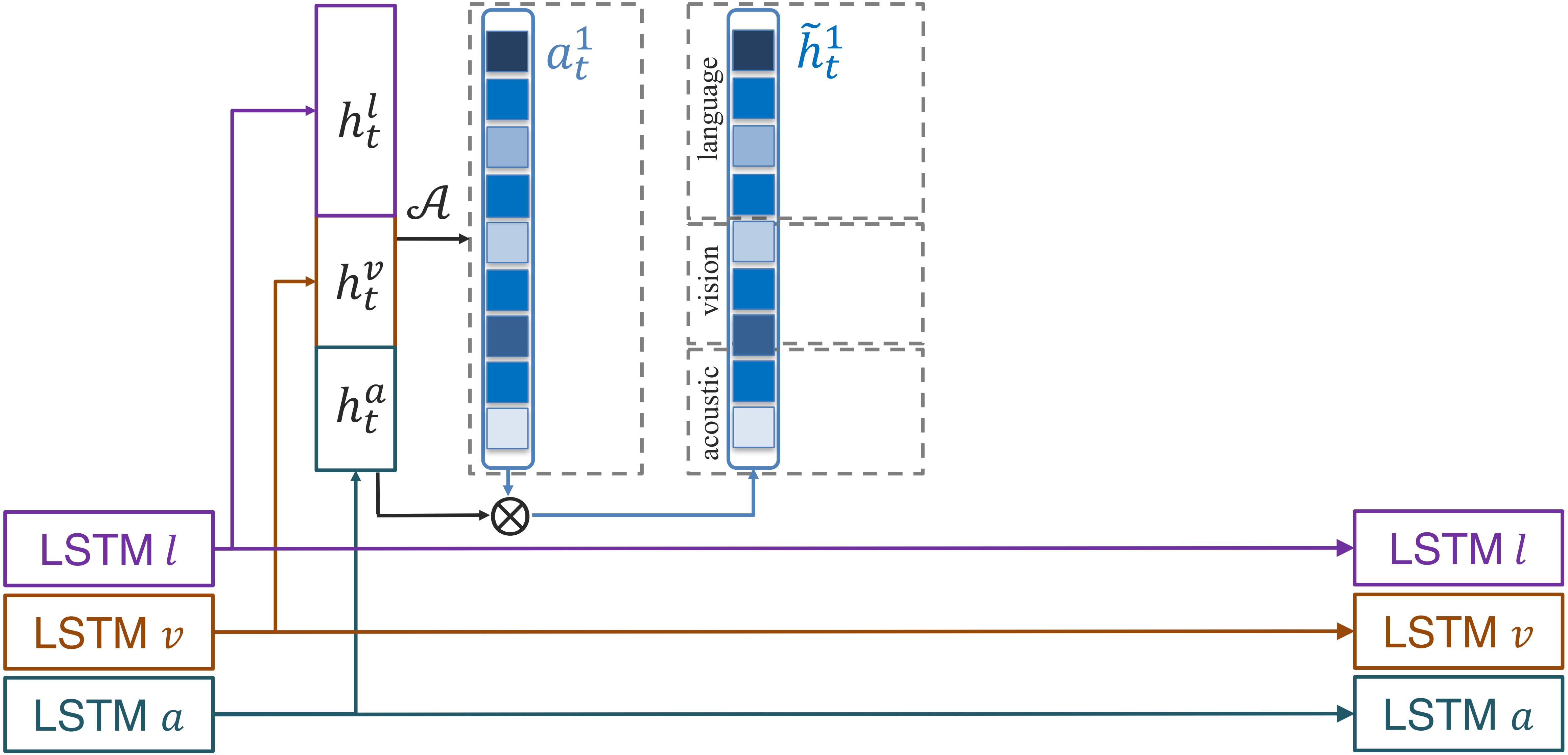
Challenge 2: Single-attention Block



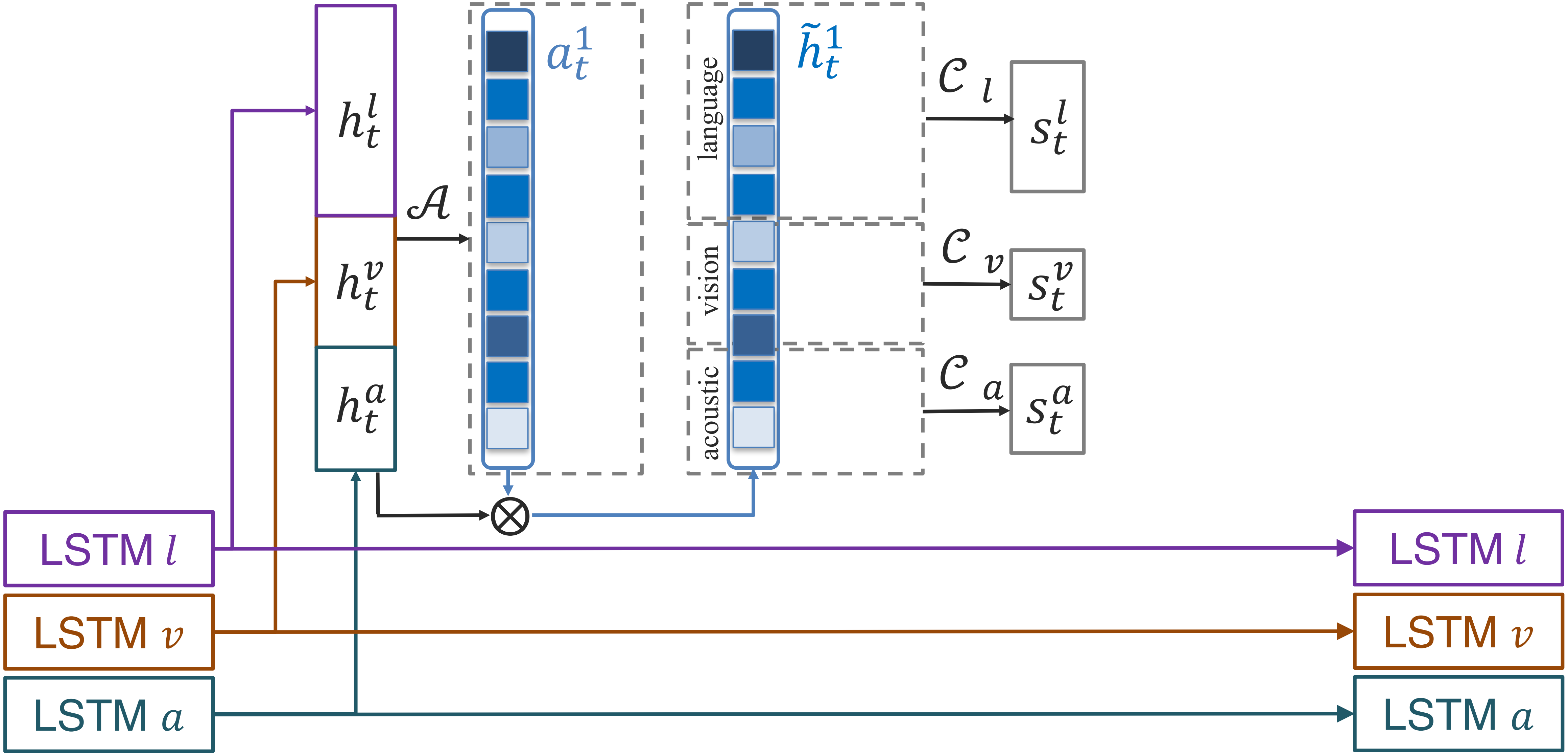
Challenge 2: Single-attention Block



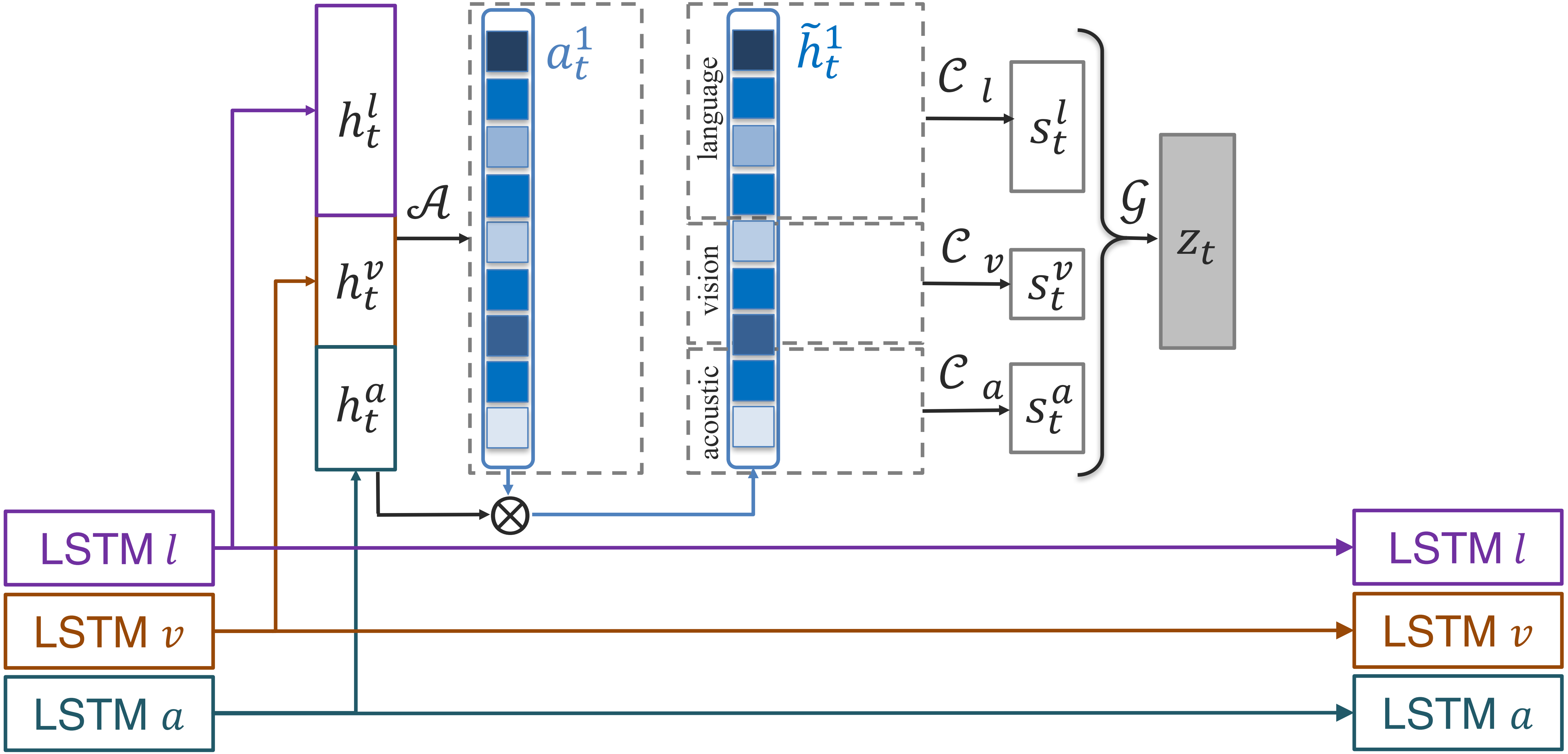
Challenge 2: Single-attention Block



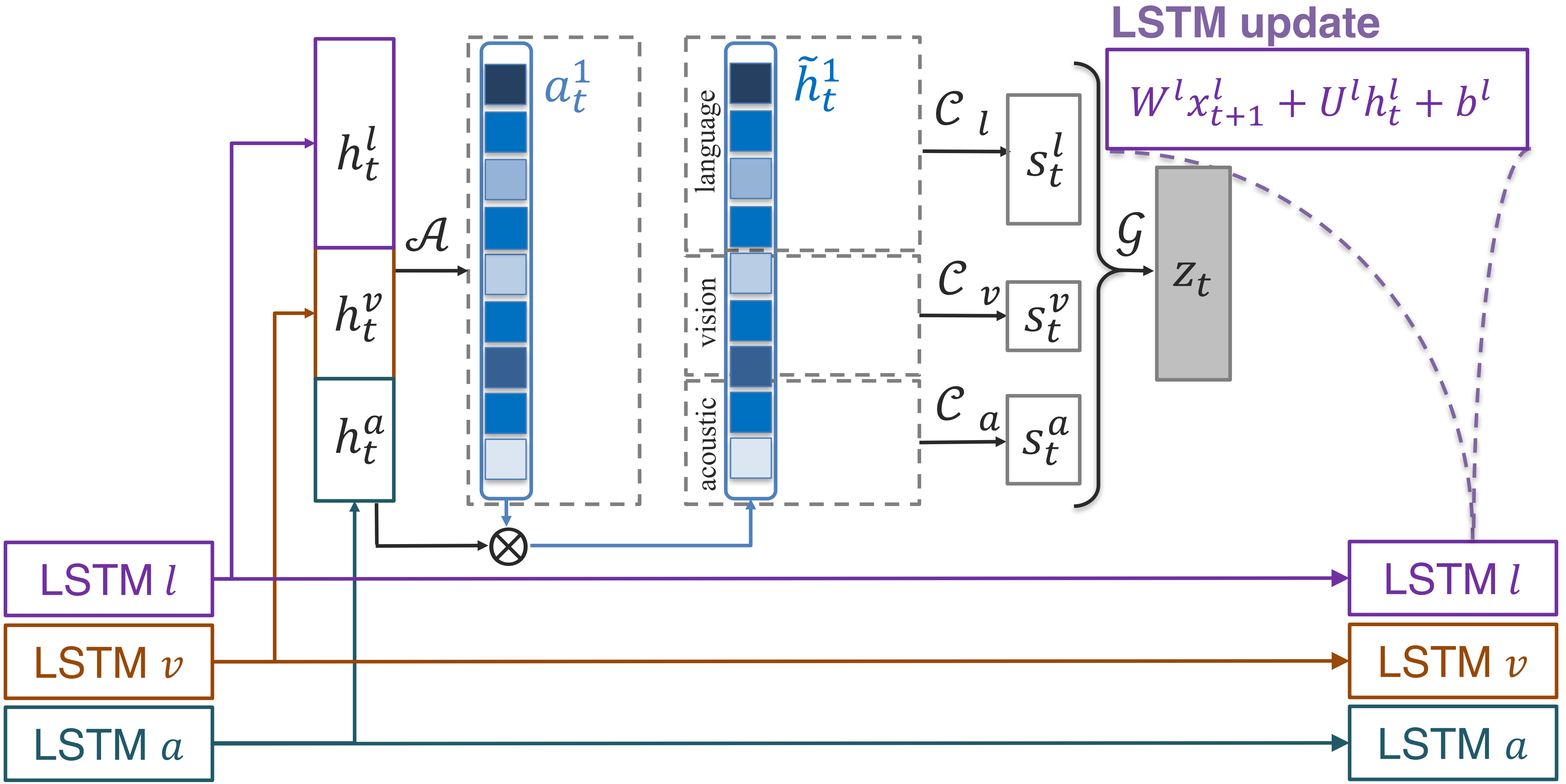
Challenge 2: Single-attention Block



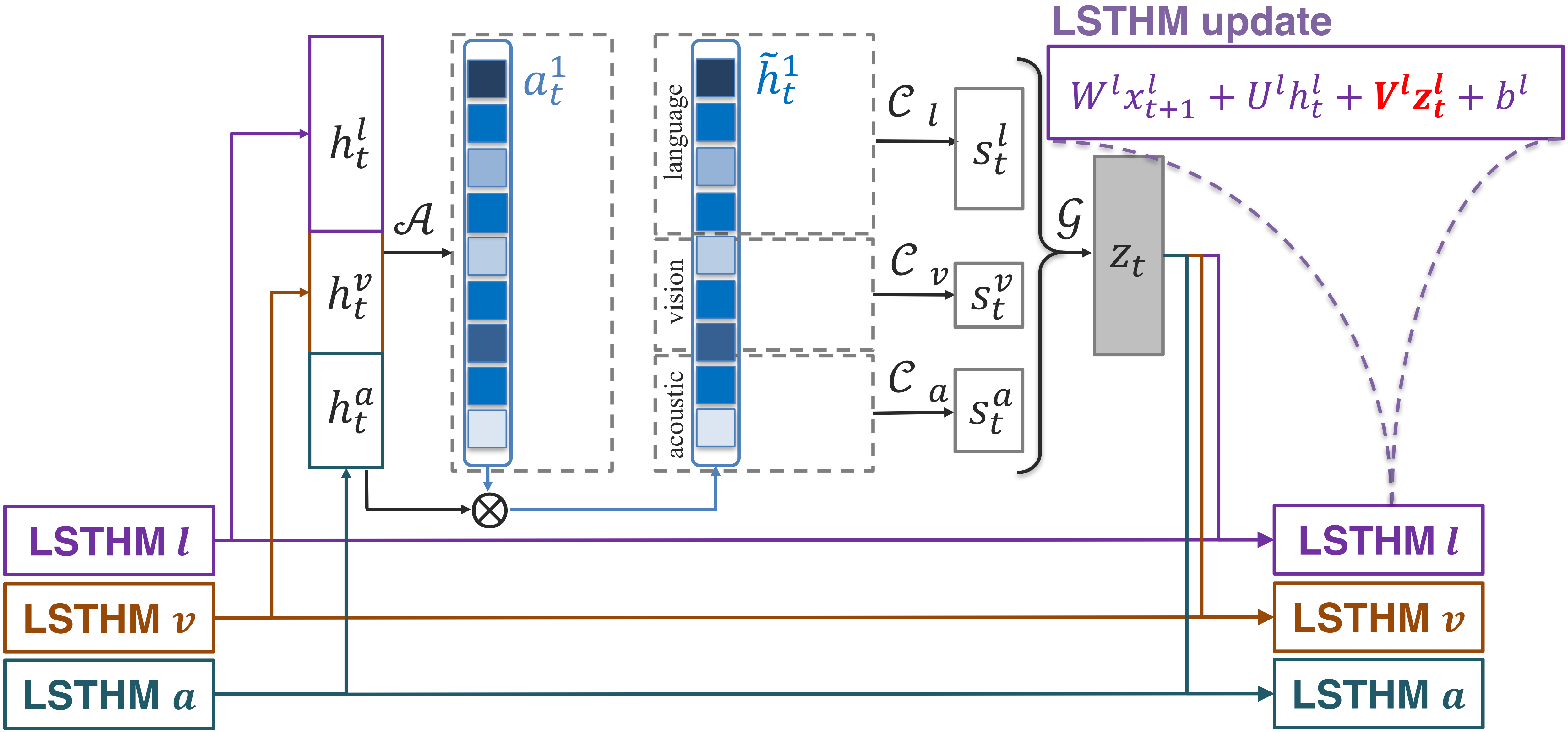
Challenge 2: Single-attention Block



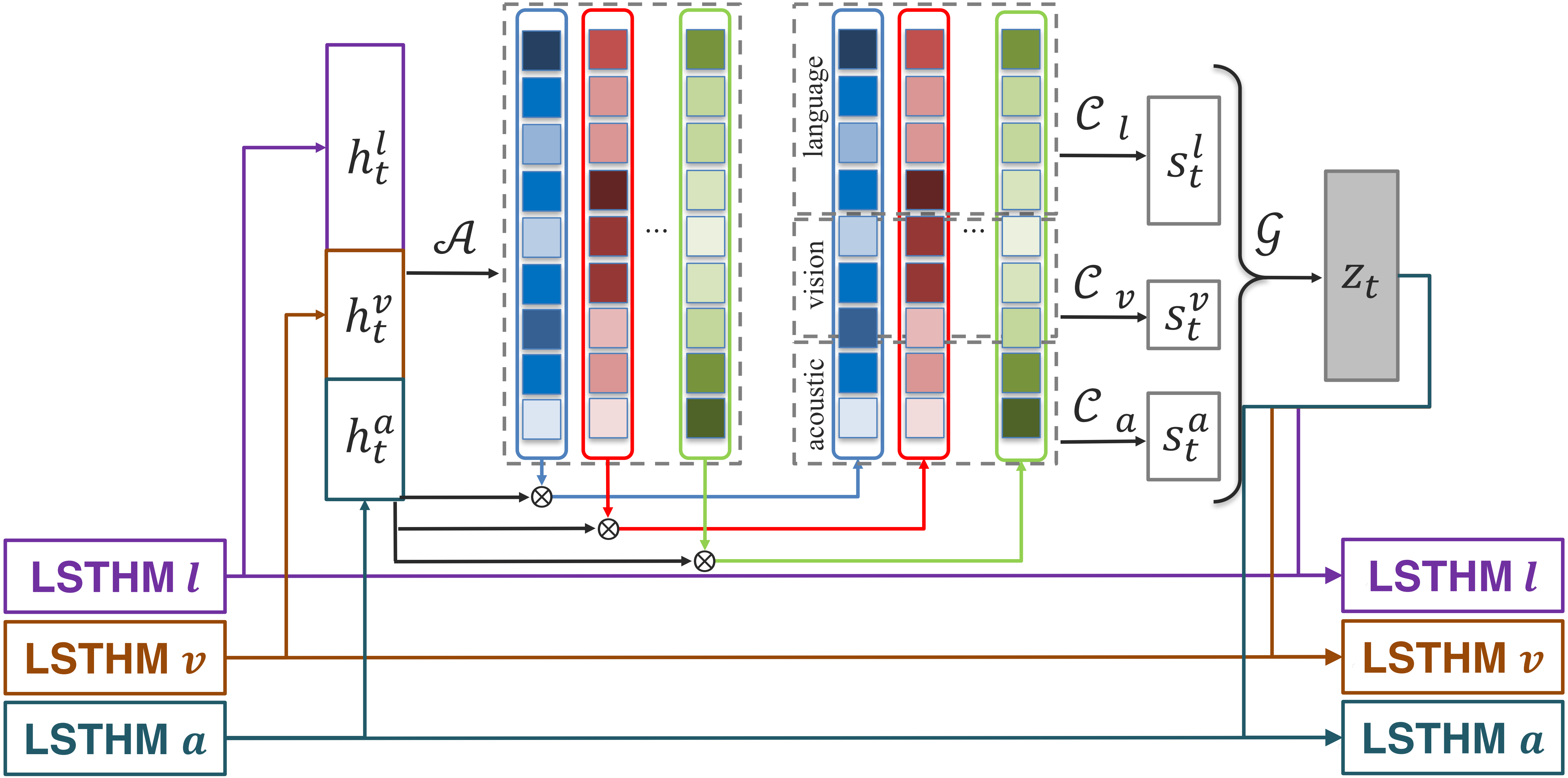
Challenge 2: Single-attention Block



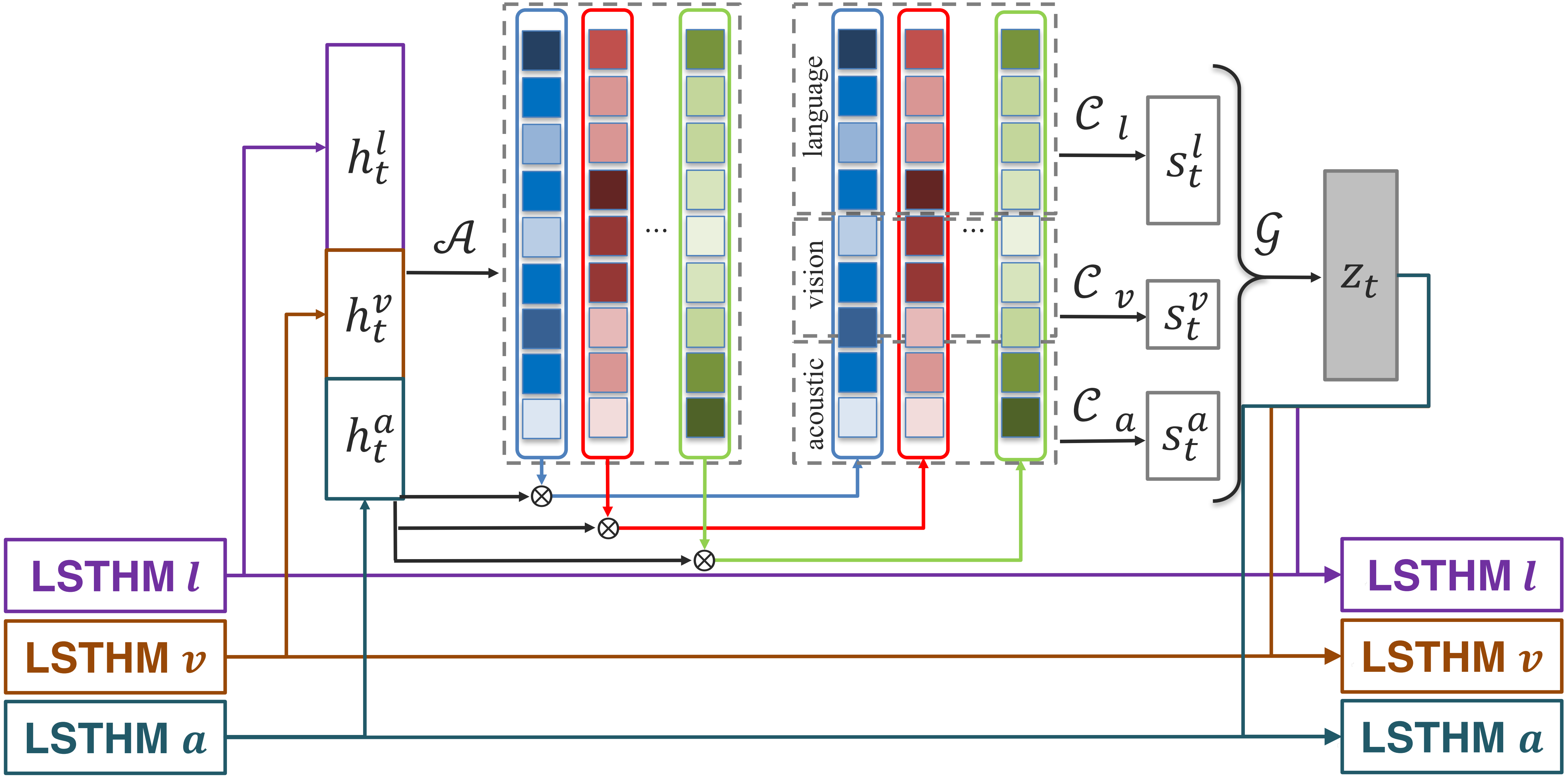
Challenge 2: Long-short Term Hybrid Memory



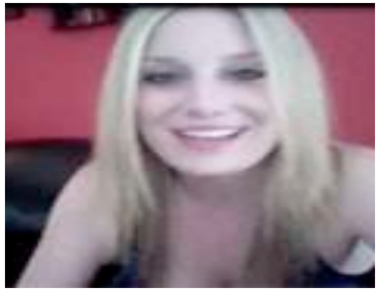
Challenge 2: Multi-attention Block



Multi-attention Recurrent Network (MARN)



Experiments



Language

- Glove word embeddings

Visual

- Facet features
  - FACS action units
  - Emotions

Acoustic

- COVAREP features
  - MFCCs
  - Pitch tracking

Alignment

- Word level
- P2FA

Sentiment

- Positive
- Negative

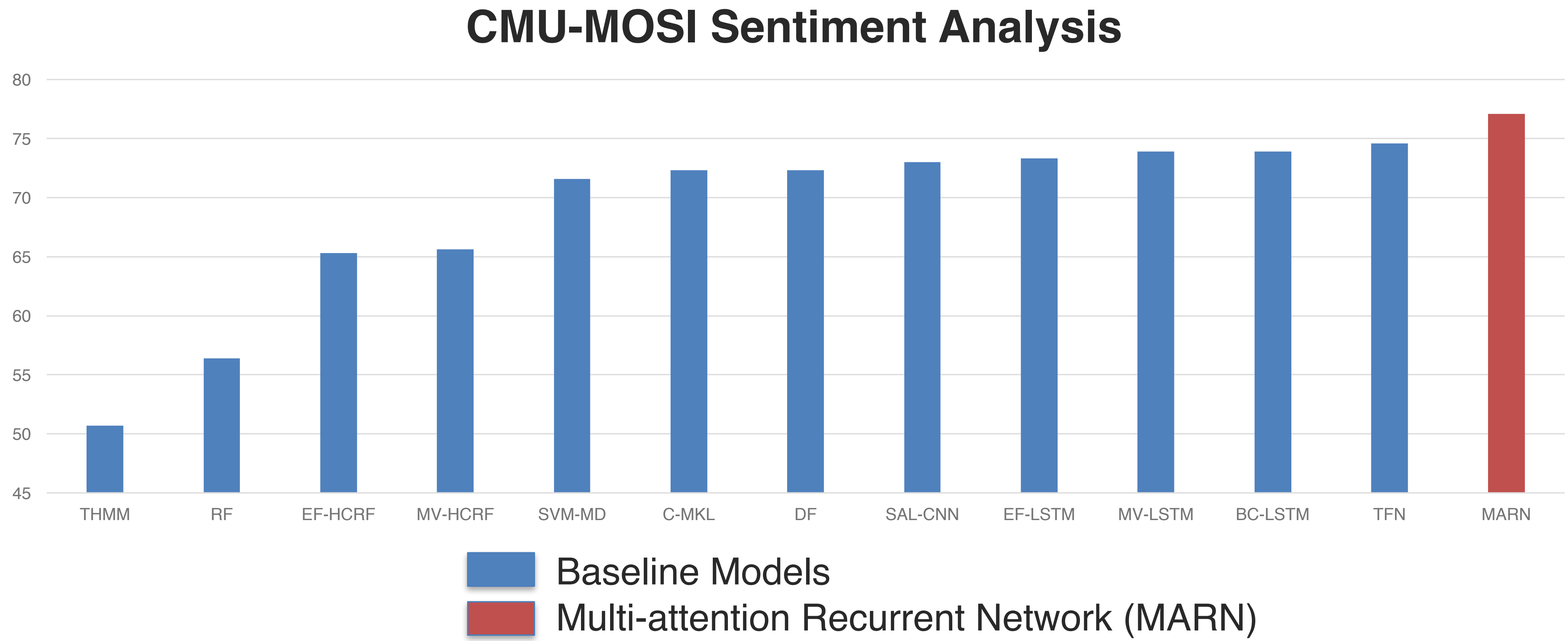
Emotion

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

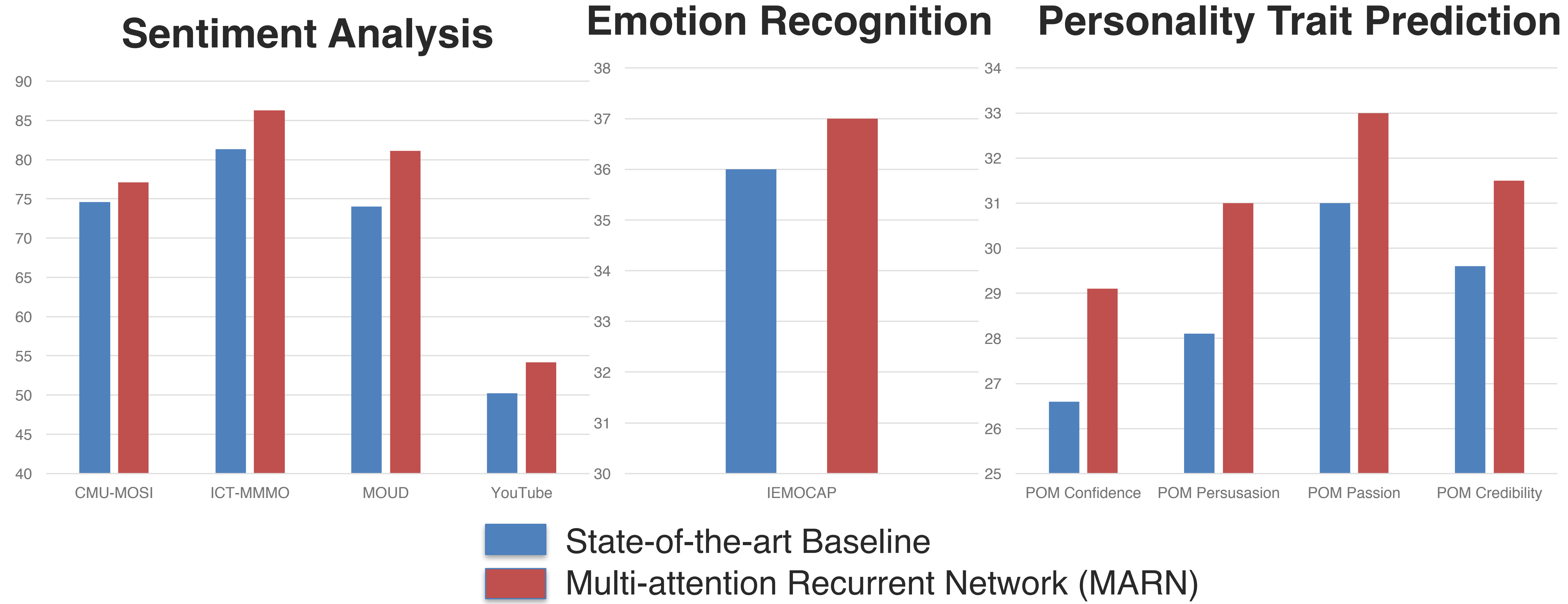
Personality

- Confidence
- Persuasion
- Passion

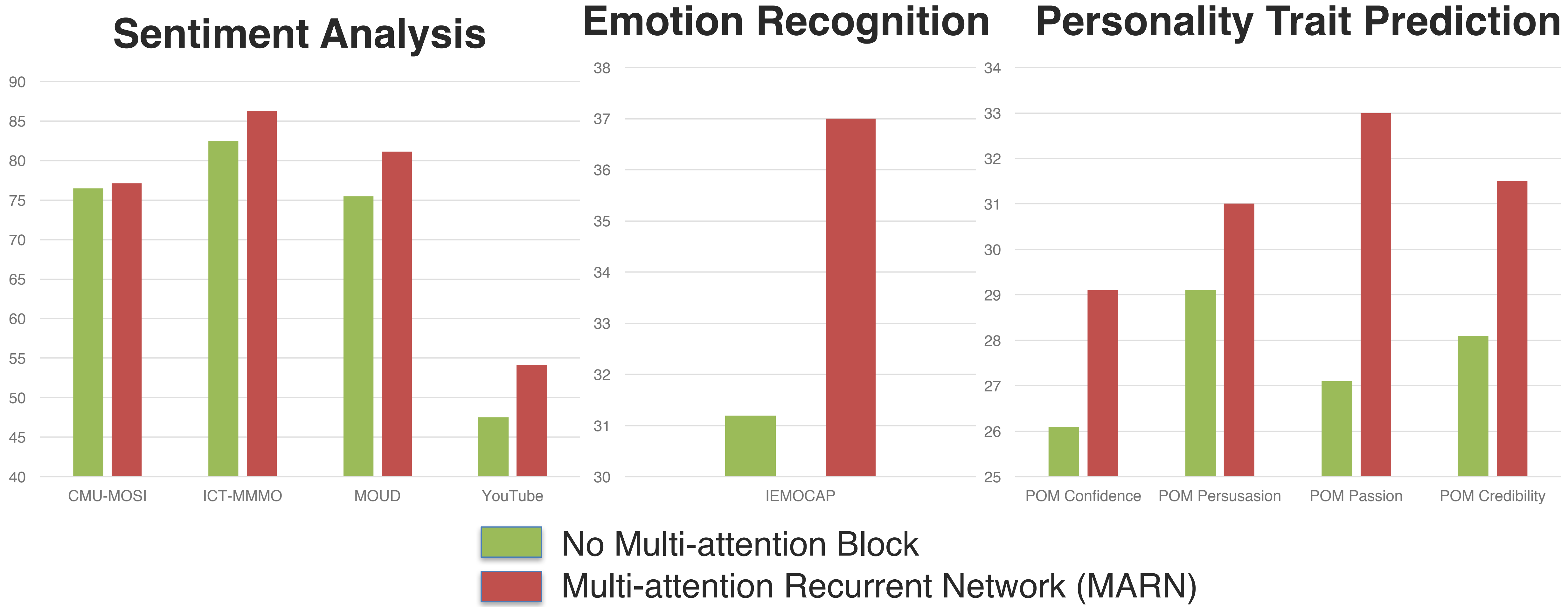
State-of-the-art Results



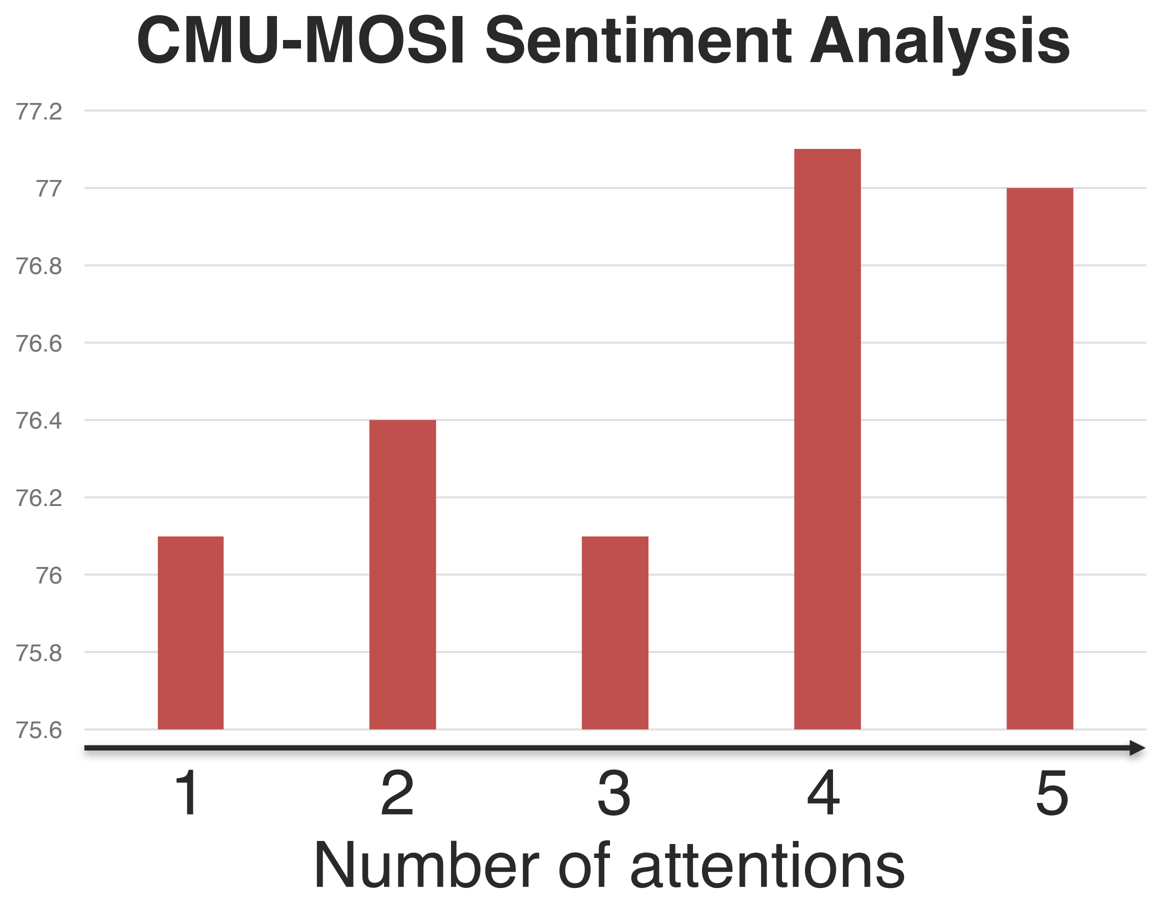
State-of-the-art Results

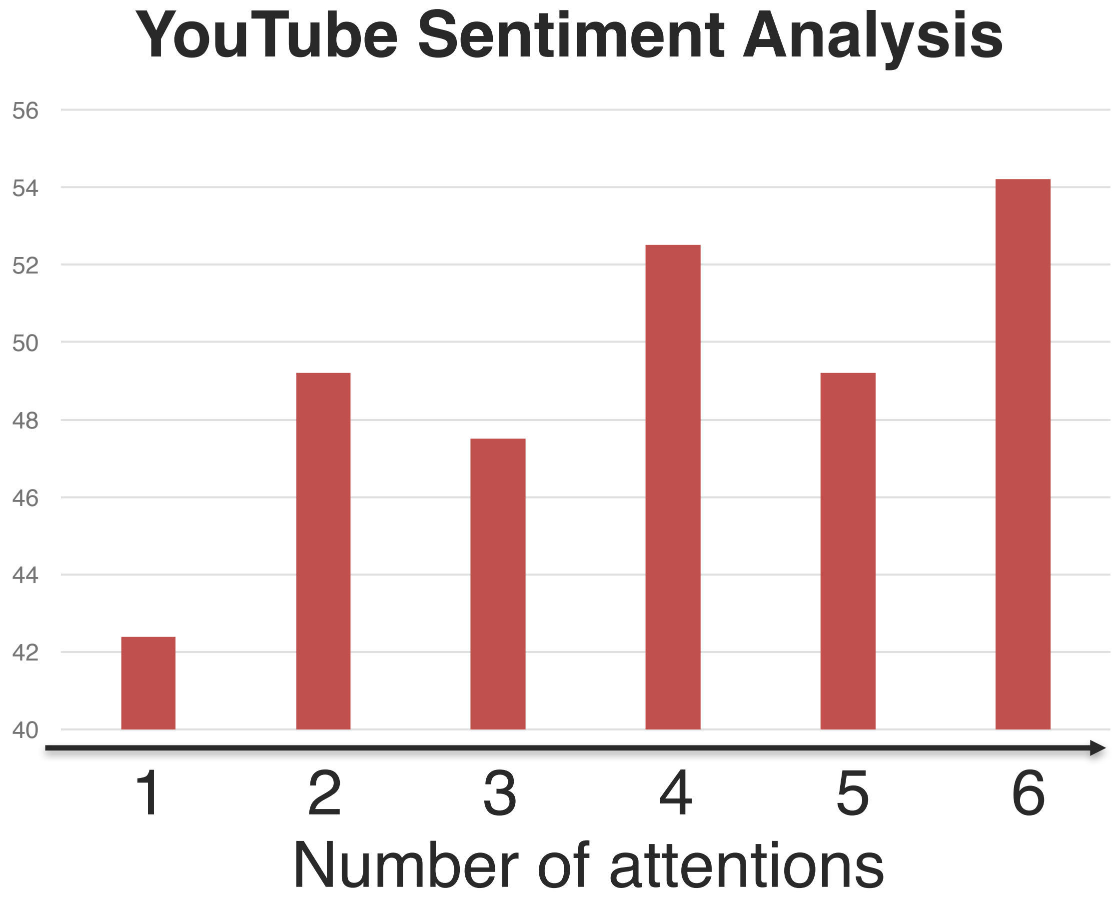


Multi-attention Block is Important

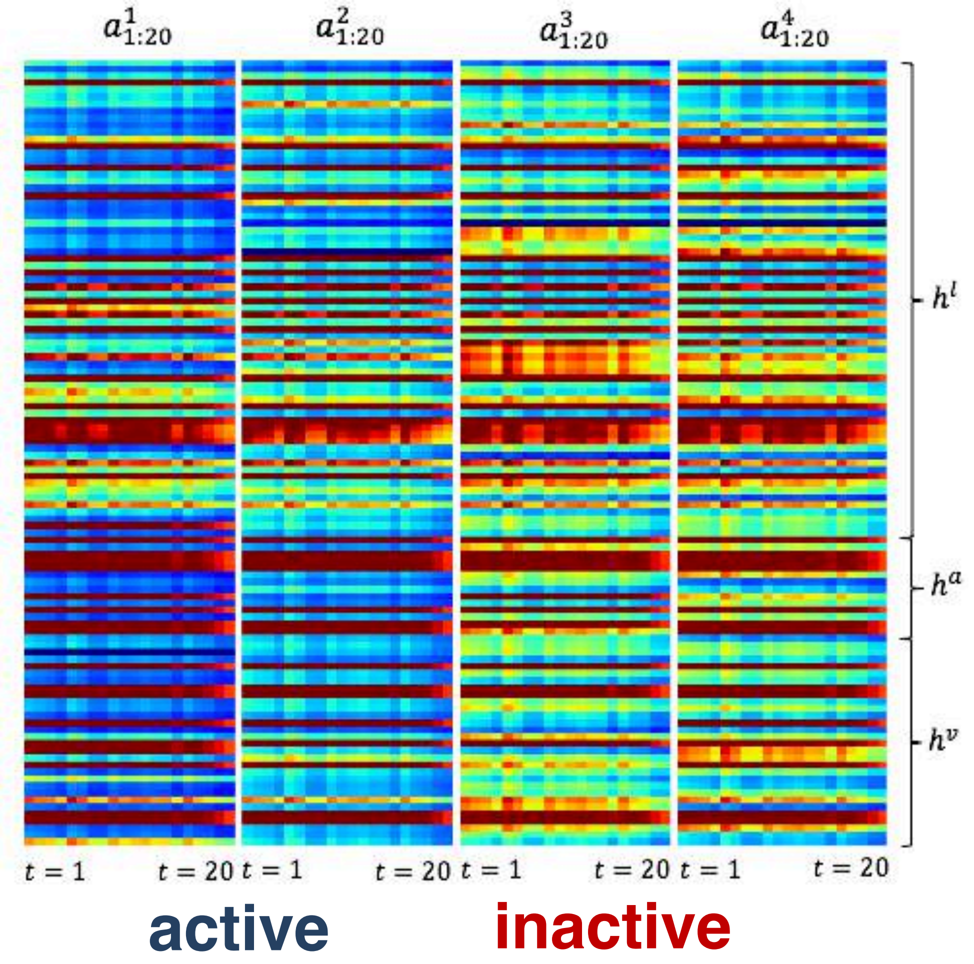


Multiple Attentions are Important



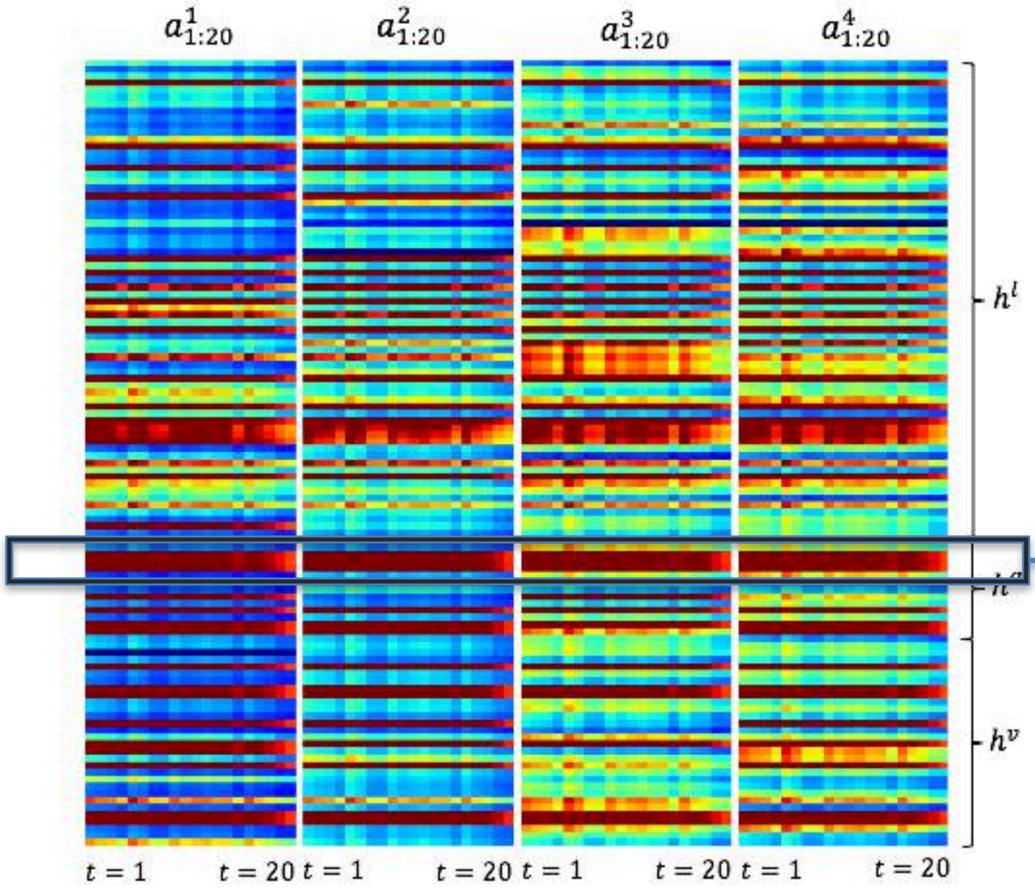


Visualization



Attentions *show diversity* and are sensitive to different cross-modal dynamics

Visualization

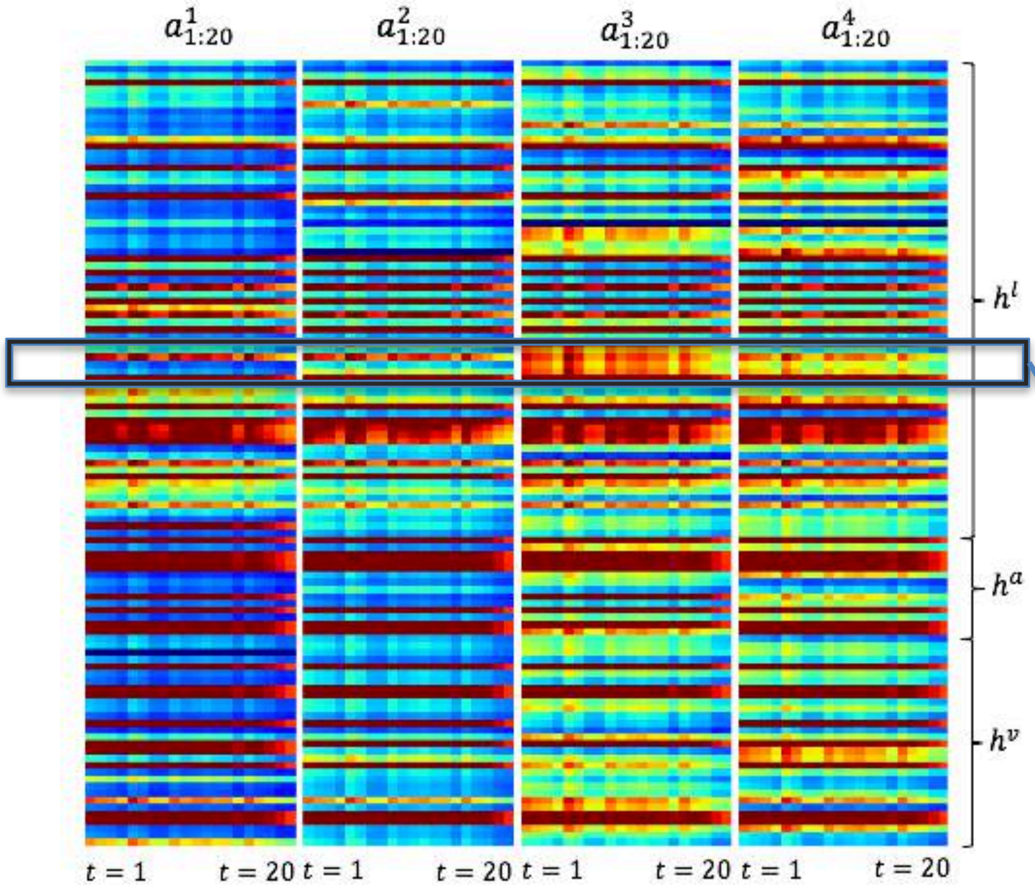


active **inactive**

Some attentions *always* *inactive*

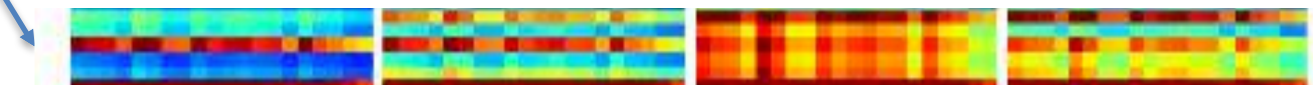
- Carry only intra-modal dynamics
- No cross-modal dynamics

Visualization



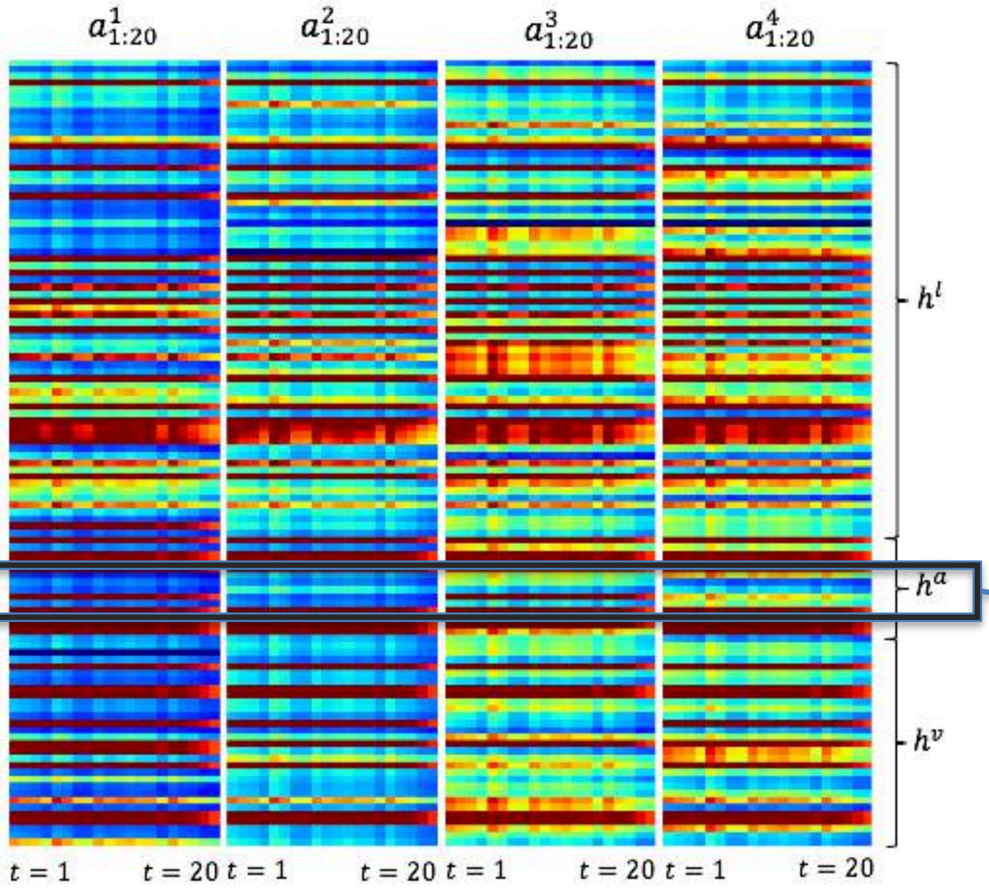
active **inactive**

Attentions *change behaviors across time*, some changes are more drastic than others.



time ⟶

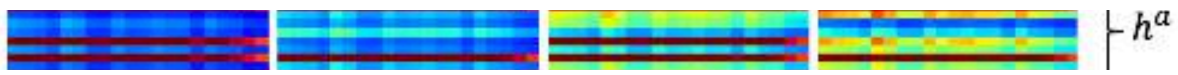
Visualization



active inactive

Different attentions *focus on different modalities.*

active inactive



Multi-attention Recurrent Network (MARN)

1 **Modeling intra-modal dynamics**

→ Set of Long-short Term Memories

2 **Modeling cross-modal dynamics**

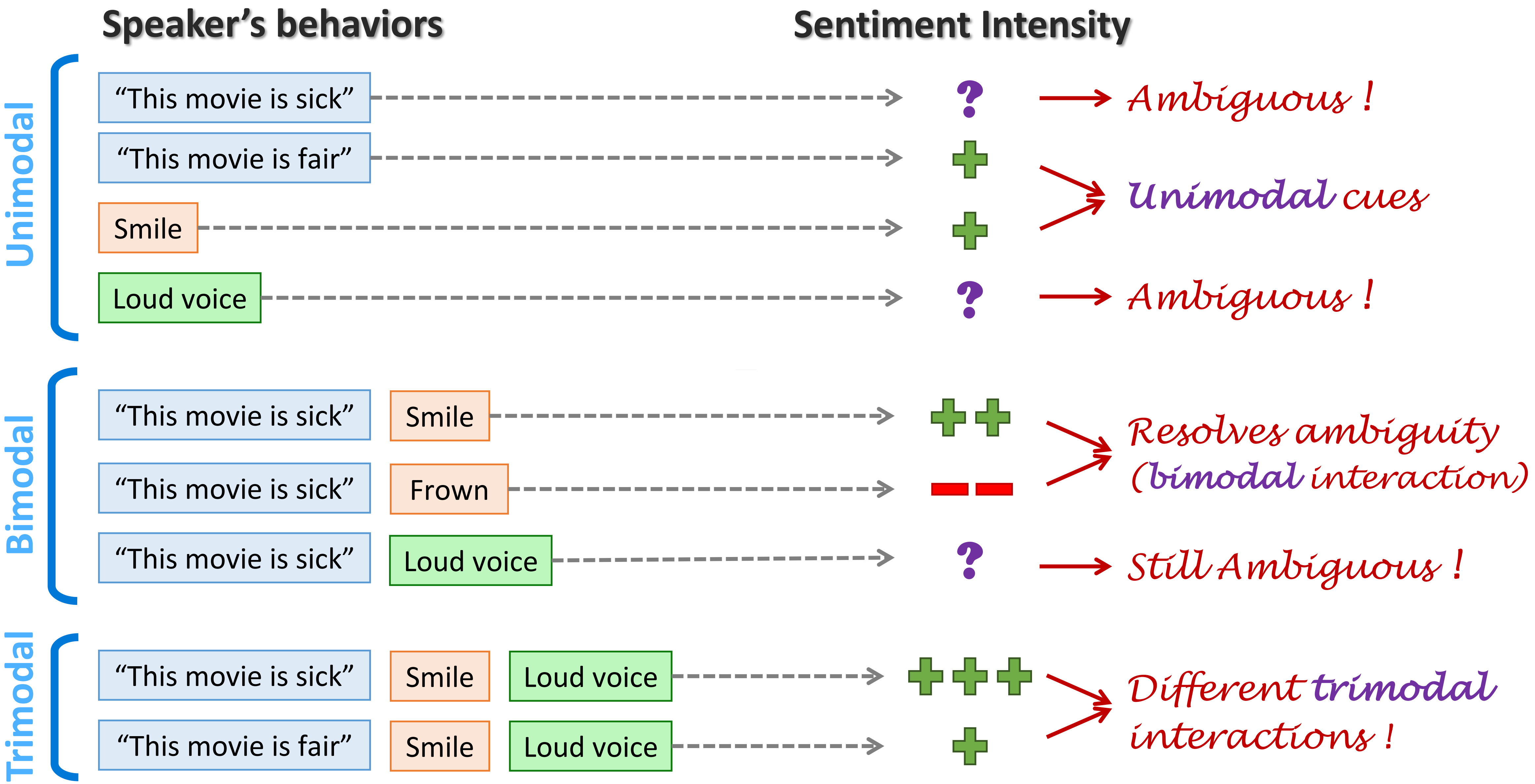
→ Set of Long-short Term **Hybrid** Memories + Single-attention Block

Modeling multiple cross-modal dynamics

→ Set of Long-short Term **Hybrid** Memories + **Multi-attention** Block

Direction 2: Unimodal, Bimodal and Trimodal

Speaker's behaviors | Sentiment Intensity

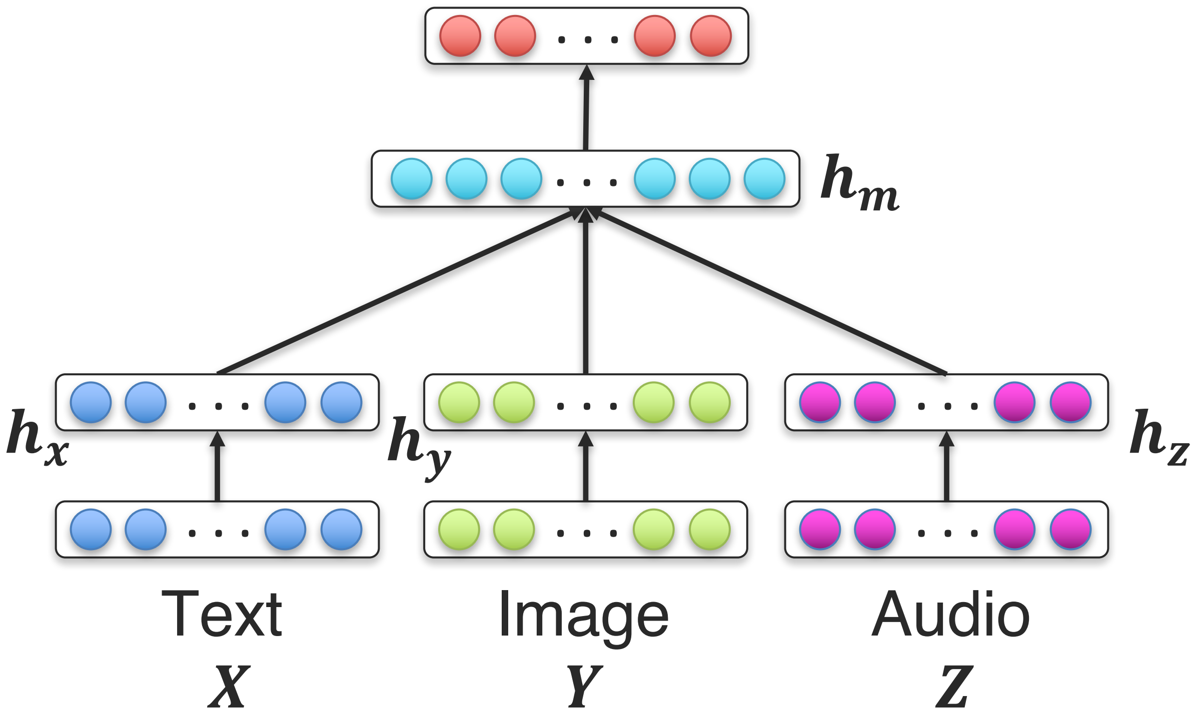

Speaker's behaviors
Sentiment Intensity
Unimodal
"This movie is sick"
Ambiguous !
"This movie is fair"
Smile
Unimodal cues
Loud voice
Ambiguous !
Bimodal
"This movie is sick"
Smile
"This movie is sick"
Frown
Resolves ambiguity
(bimodal interaction)
"This movie is sick"
Loud voice
Still Ambiguous !
Trimodal
"This movie is sick"
Smile
Loud voice
"This movie is fair"
Smile
Loud voice
Different trimodal
interactions !

Simple Neural Network

Joint Multimodal Representation

Simply concatenates all three individual representations:

$$\boldsymbol{h_m} = \boldsymbol{f}\left(\boldsymbol{W} \cdot \left[\boldsymbol{h_x}, \boldsymbol{h_y}, \boldsymbol{h_z}\right]\right)$$



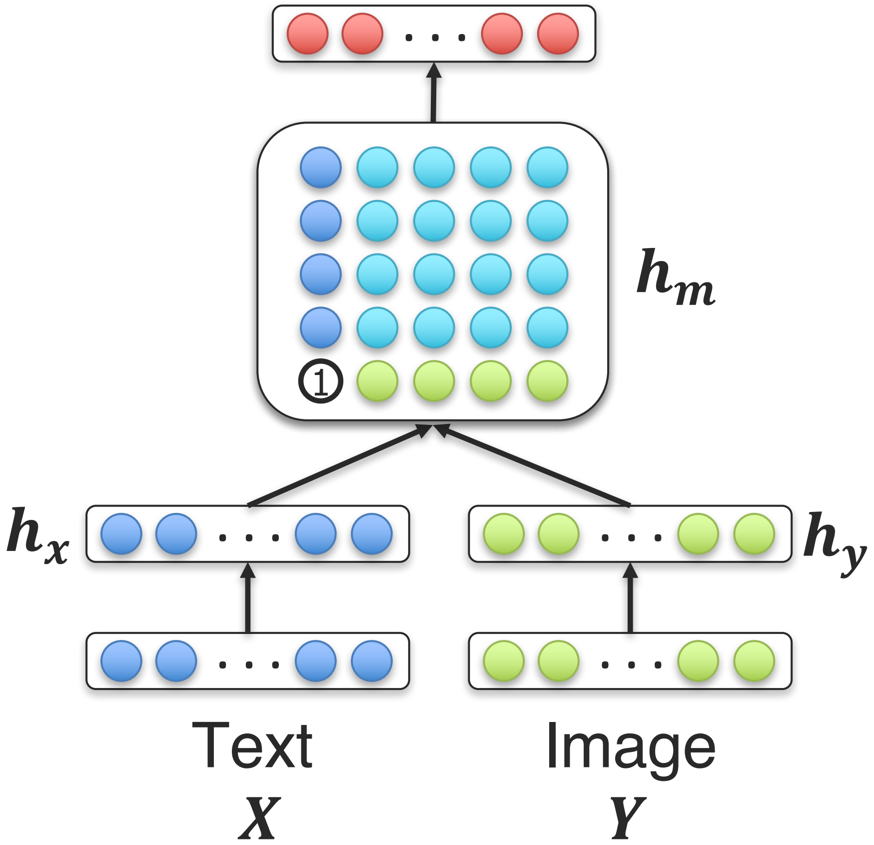
- Similar to early fusion

Multimodal Tensor Fusion Network (TFN)

Models both unimodal and bimodal interactions:

$$\boldsymbol{h}_m = \begin{bmatrix} \boldsymbol{h}_x \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \boldsymbol{h}_y \\ 1 \end{bmatrix} = \begin{bmatrix} \boxed{\boldsymbol{h}_x} & \boxed{\boldsymbol{h}_x \otimes \boldsymbol{h}_y} \\ 1 & \boxed{\boldsymbol{h}_y} \end{bmatrix}$$

Important !

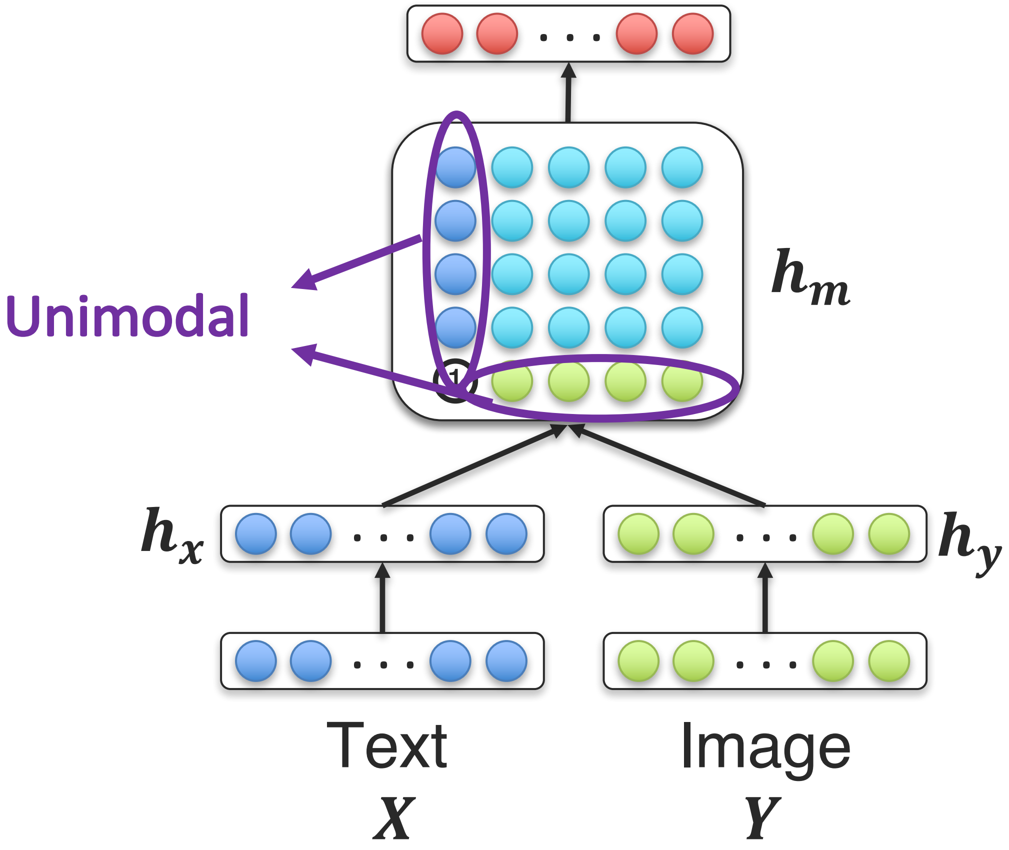


Multimodal Tensor Fusion Network (TFN)

Models both unimodal and bimodal interactions:

$$h_m = \begin{bmatrix} h_x \\ 1 \end{bmatrix} \otimes \begin{bmatrix} h_y \\ 1 \end{bmatrix} = \begin{bmatrix} \boxed{h_x} & \boxed{h_x \otimes h_y} \\ 1 & \boxed{h_y} \end{bmatrix}$$

Important !

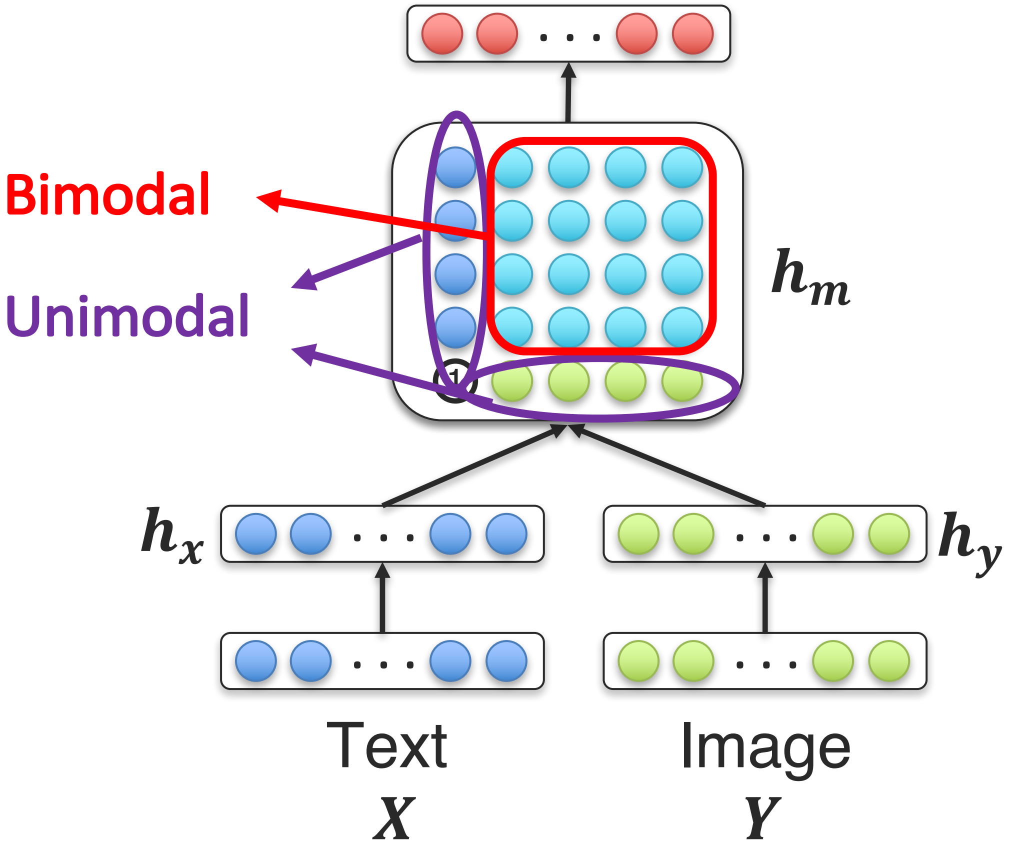


Multimodal Tensor Fusion Network (TFN)

Models both unimodal and bimodal interactions:

$$\boldsymbol{h}_m = \begin{bmatrix} \boldsymbol{h}_x \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \boldsymbol{h}_y \\ 1 \end{bmatrix} = \begin{bmatrix} \boxed{\boldsymbol{h}_x} & \boxed{\boldsymbol{h}_x \otimes \boldsymbol{h}_y} \\ 1 & \boxed{\boldsymbol{h}_y} \end{bmatrix}$$

Important !

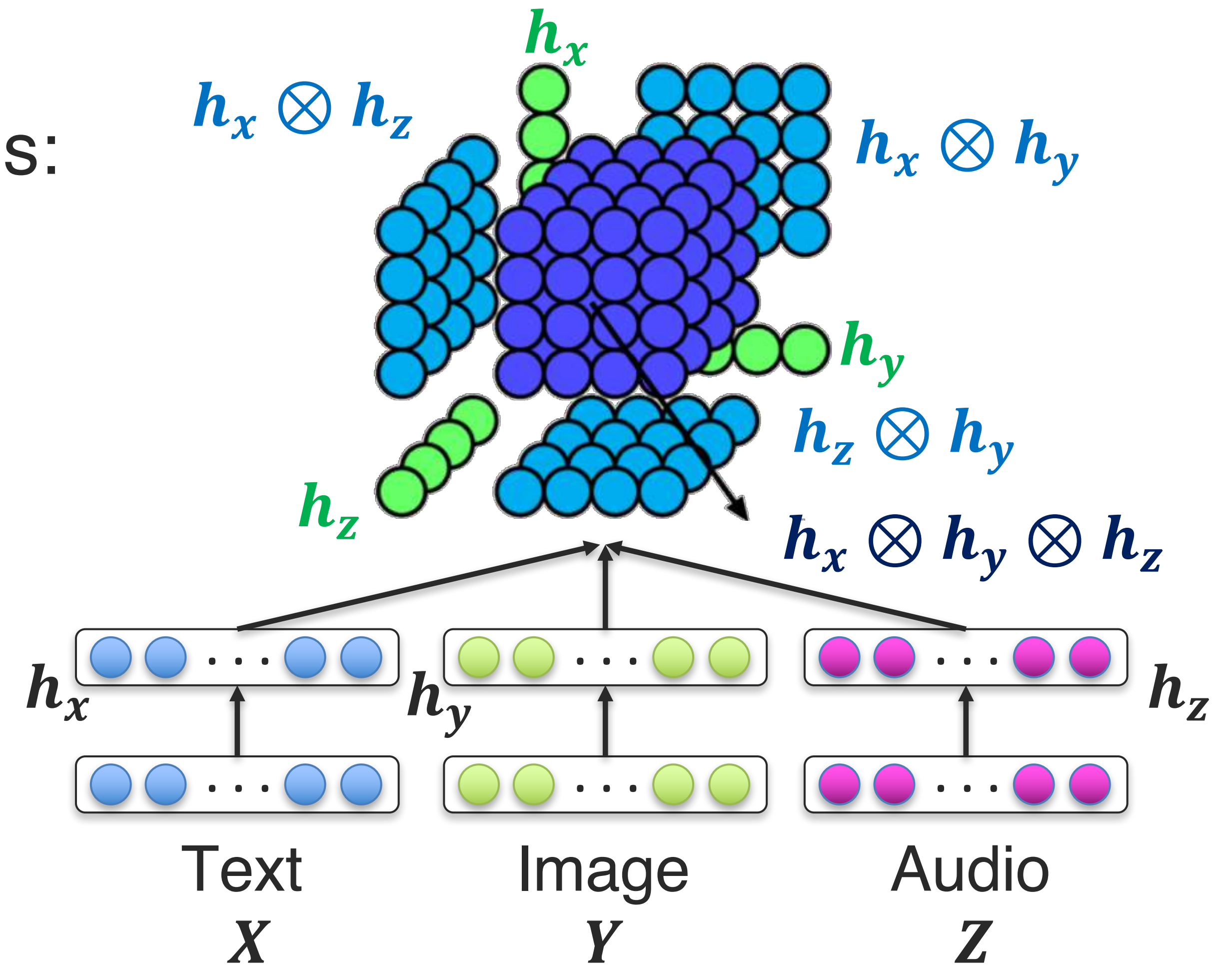


Multimodal Tensor Fusion Network (TFN)

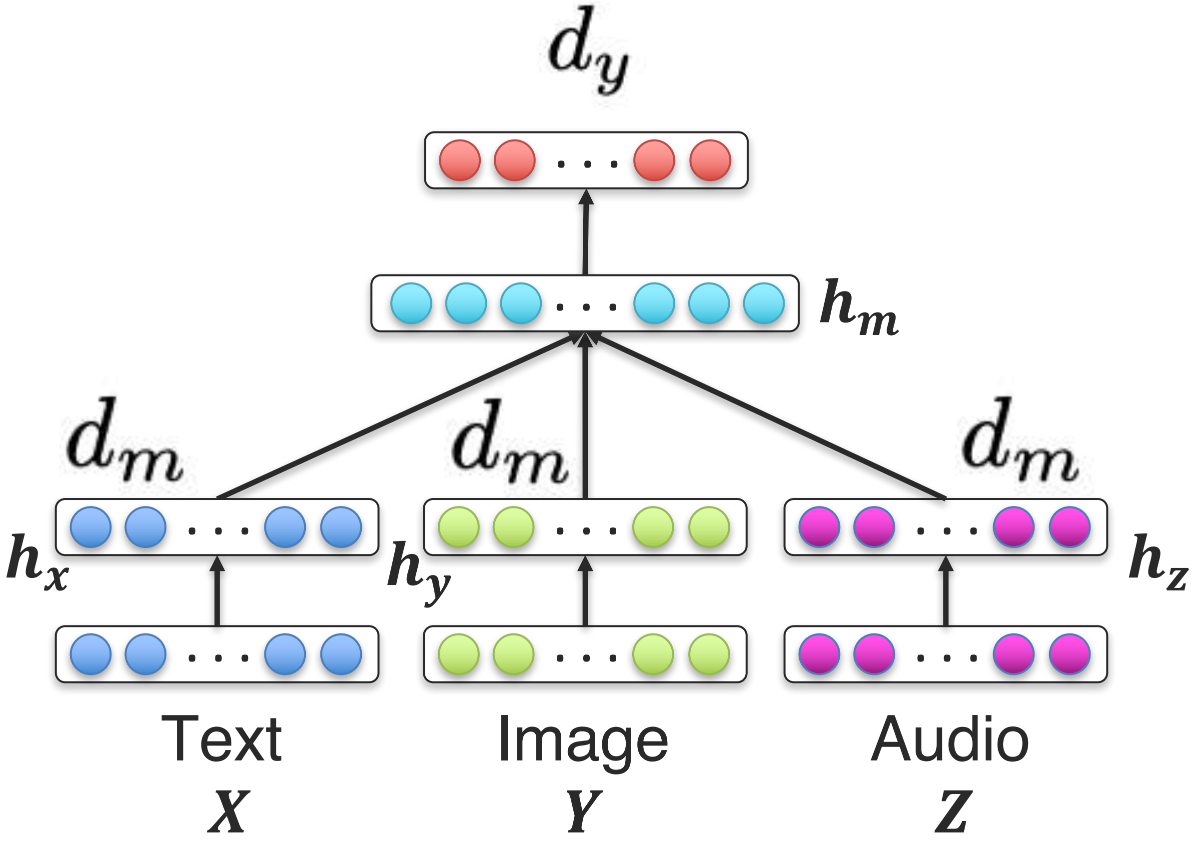
Can be extended to three modalities:

$$h_m = \begin{bmatrix} h_x \\ 1 \end{bmatrix} \otimes \begin{bmatrix} h_y \\ 1 \end{bmatrix} \otimes \begin{bmatrix} h_z \\ 1 \end{bmatrix}$$

Explicitly models unimodal, bimodal and trimodal interactions!

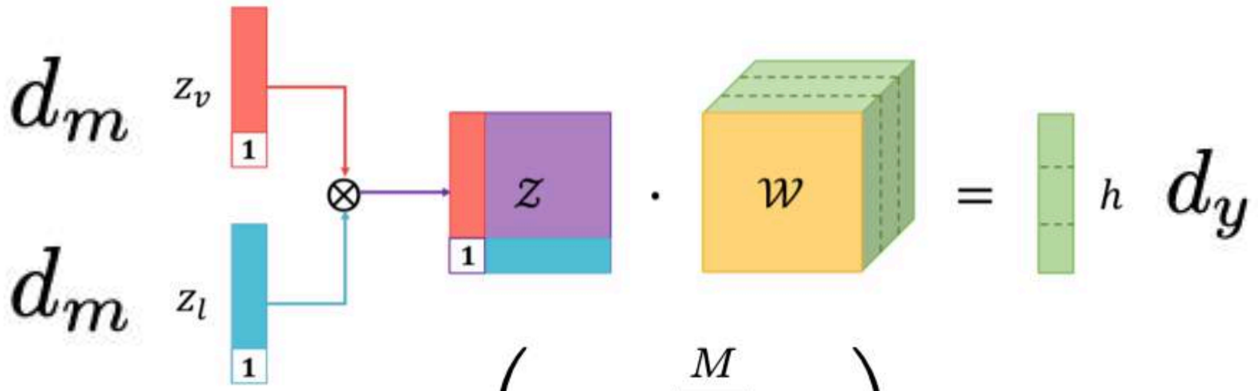


Number of Parameters



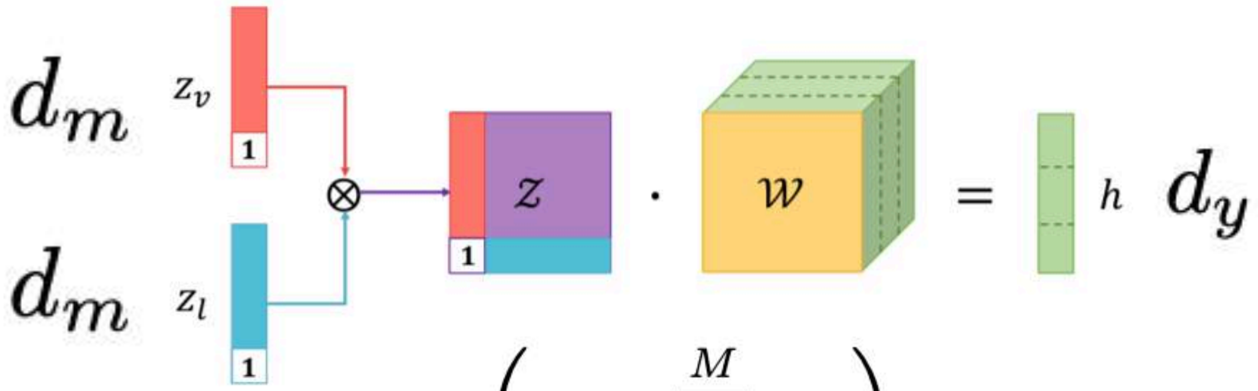
$$O\left(d_y \times \sum_{m=1}^{M} d_m\right)$$

Number of Parameters



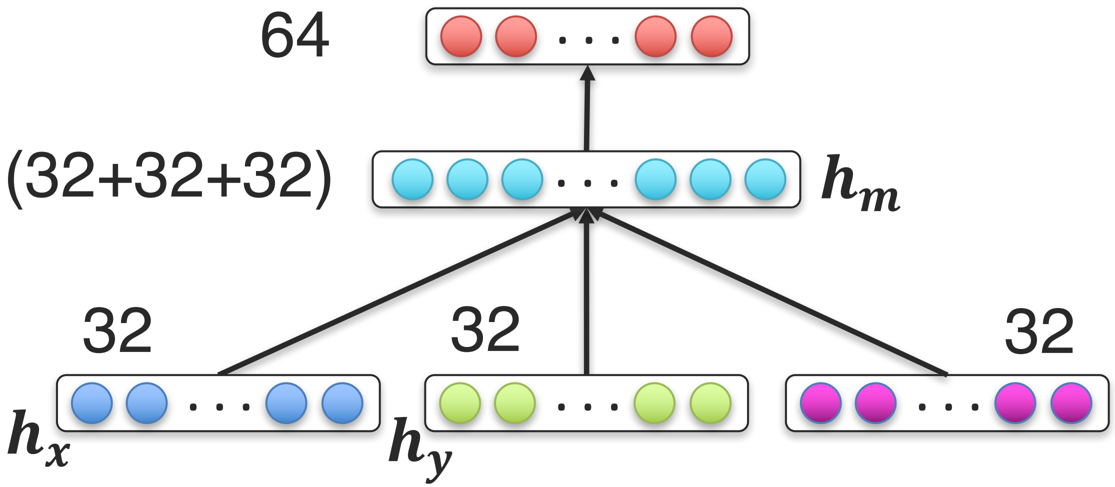
$$O\left(d_y \times \prod_{m=1}^{M} d_m\right)$$

Number of Parameters

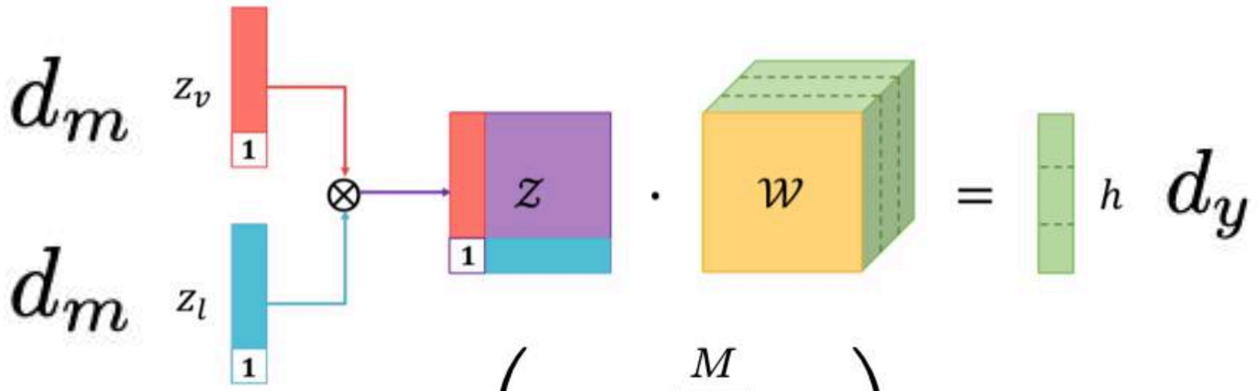


$$O\left(d_y \times \prod_{m=1}^{M} d_m\right)$$

64 x (32+32+32) = 6000



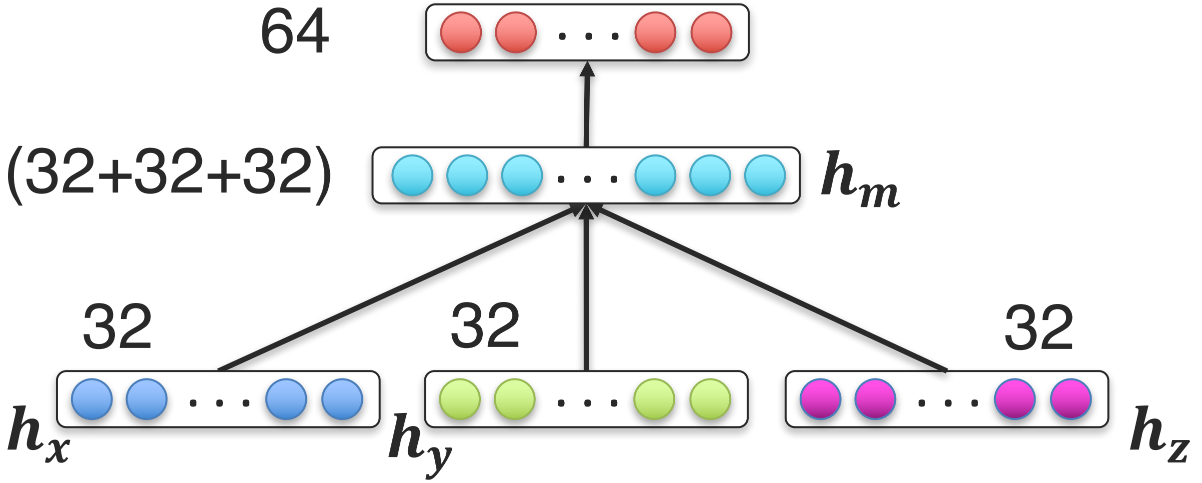
Number of Parameters

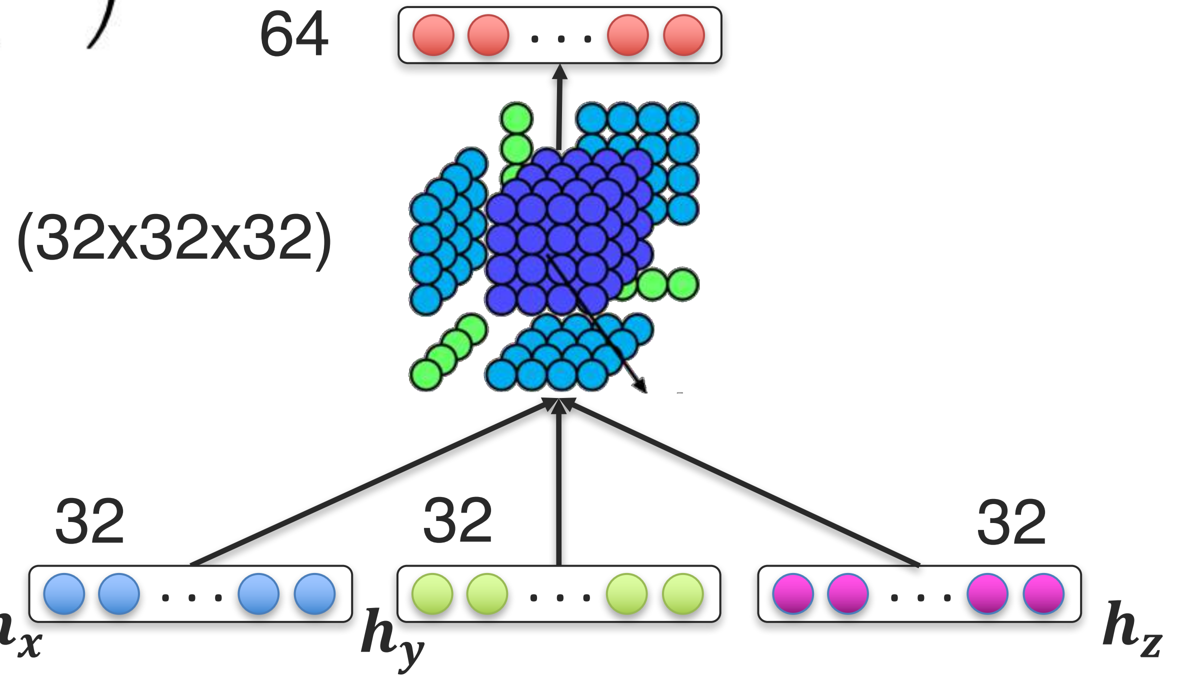


$$O\left(d_y \times \prod_{m=1}^{M} d_m\right)$$

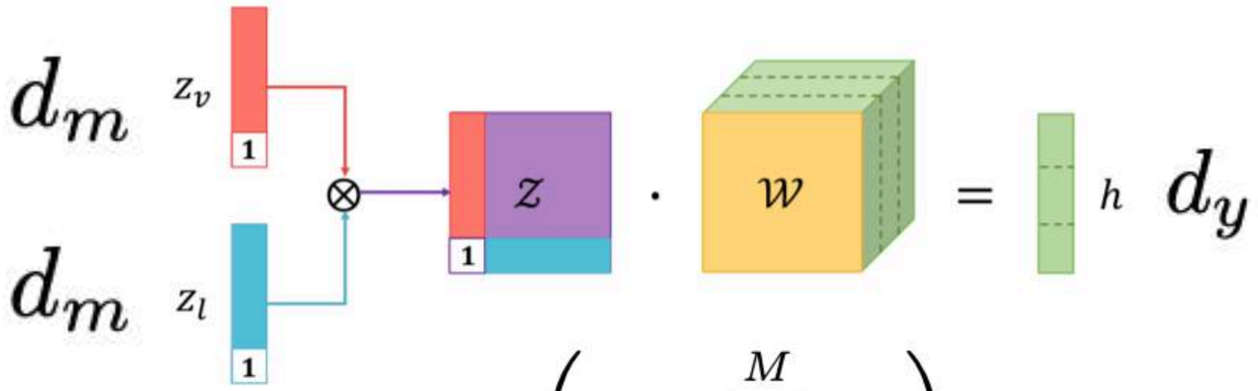
64 x (32x32x32)
= 2000000

64 x (32+32+32) = 6000

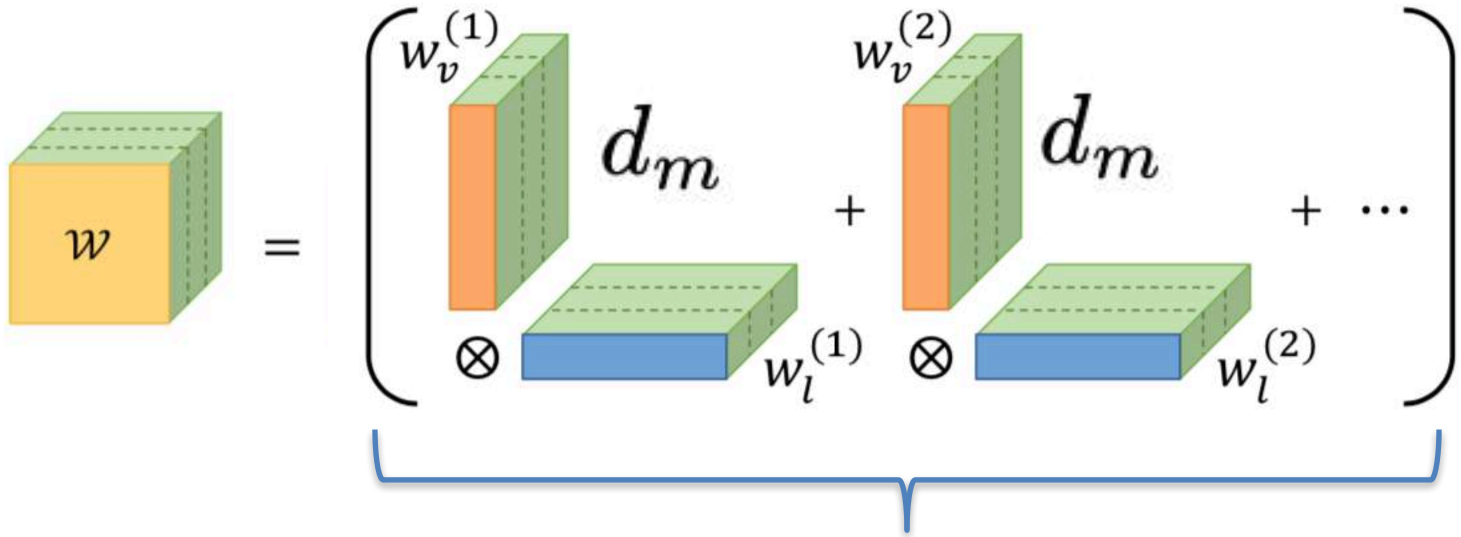




Low-rank Tensor Approximation

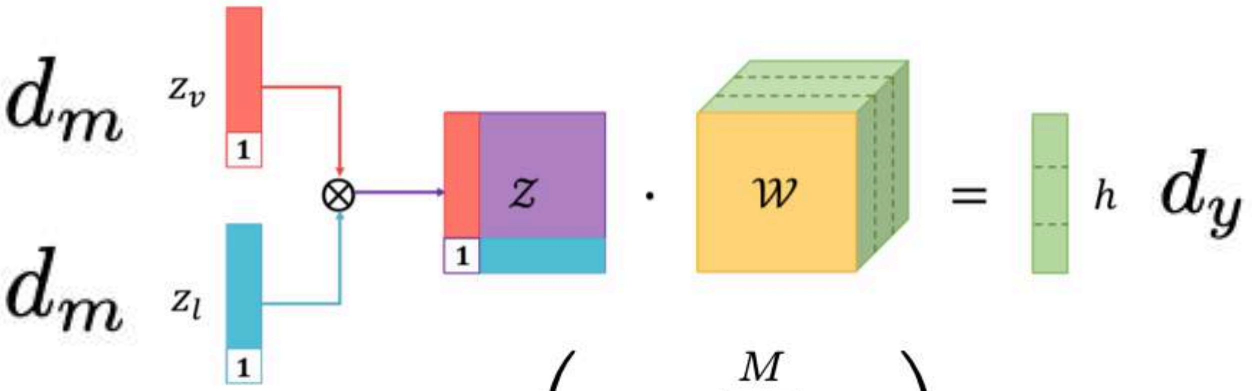


- Rank-r approximation $O\left(d_y \times \prod_{m=1}^{M} d_m\right)$

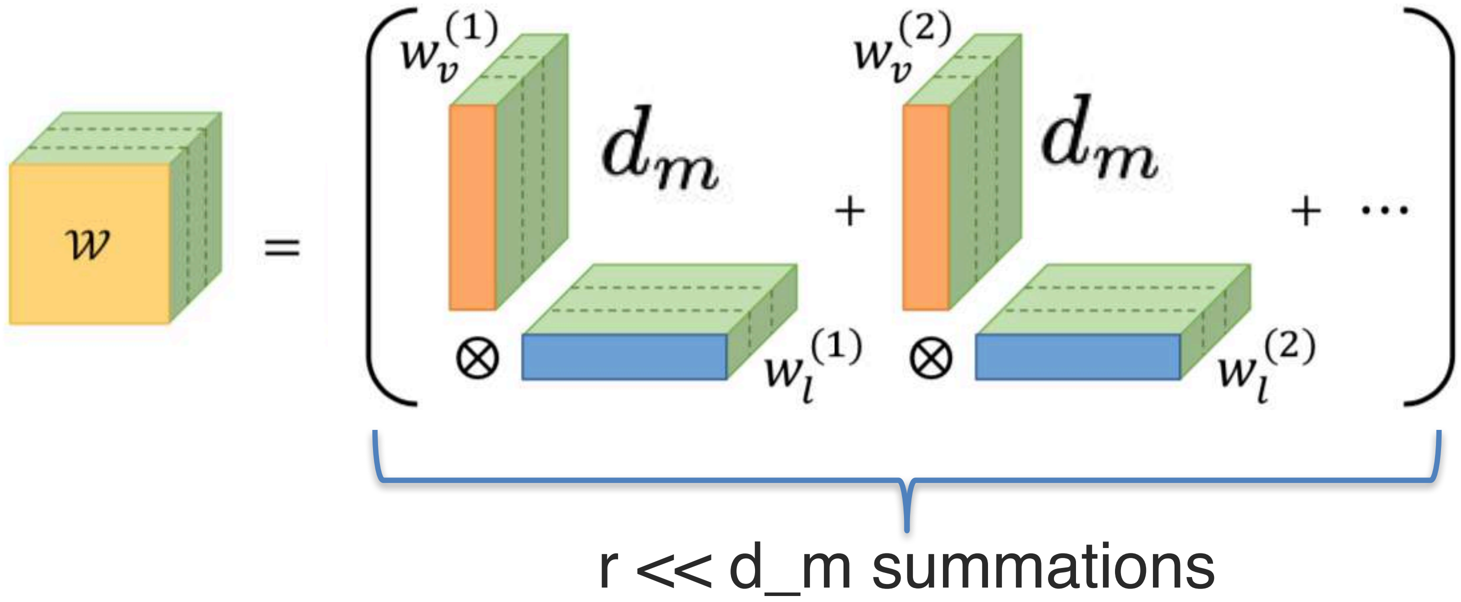


r summations

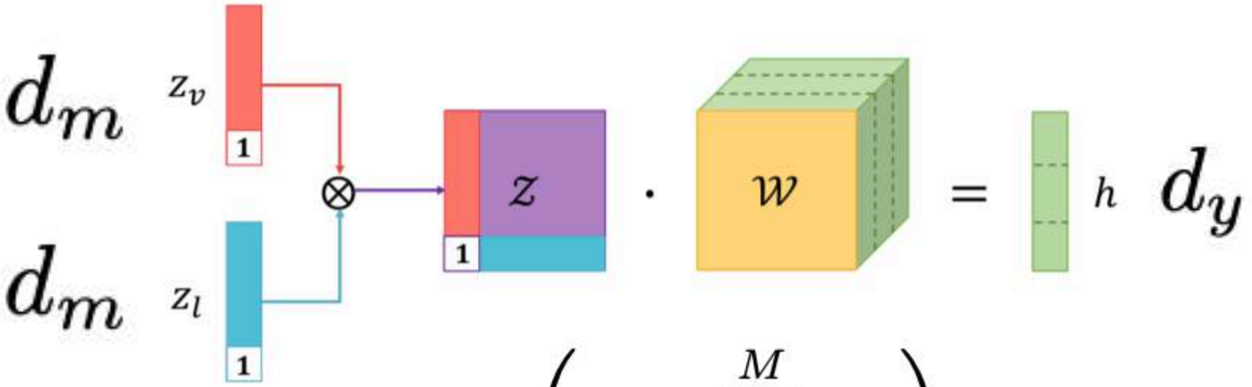
Low-rank Tensor Approximation



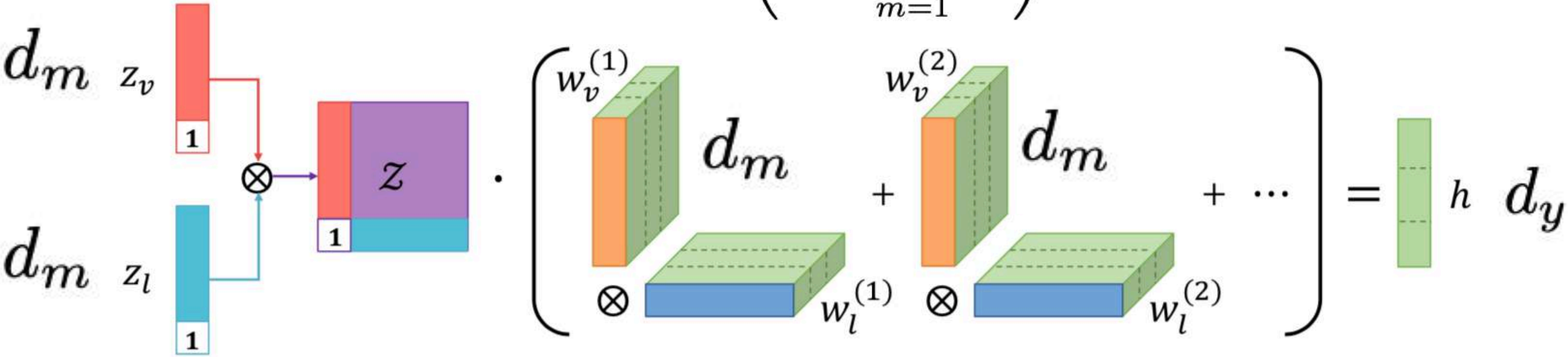
- Rank-r approximation $O\left(d_y \times \prod_{m=1}^{M} d_m\right)$



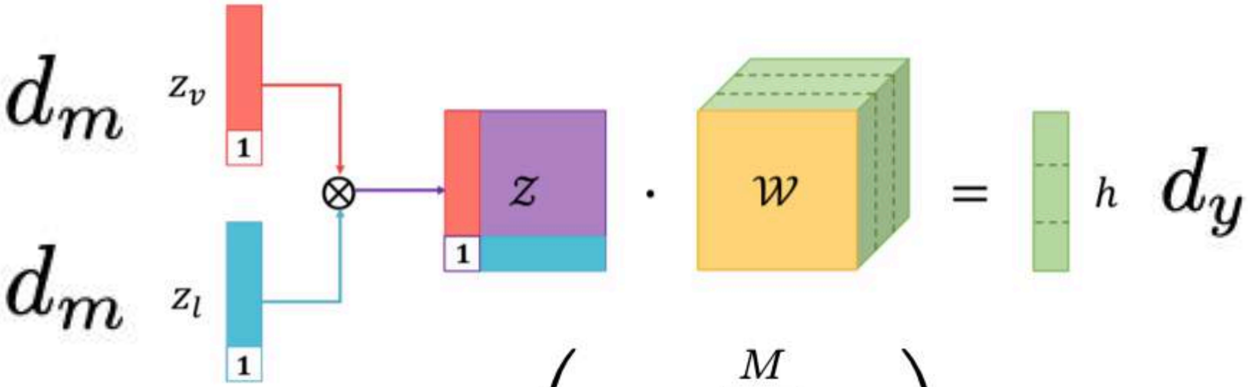
Low-rank Tensor Approximation



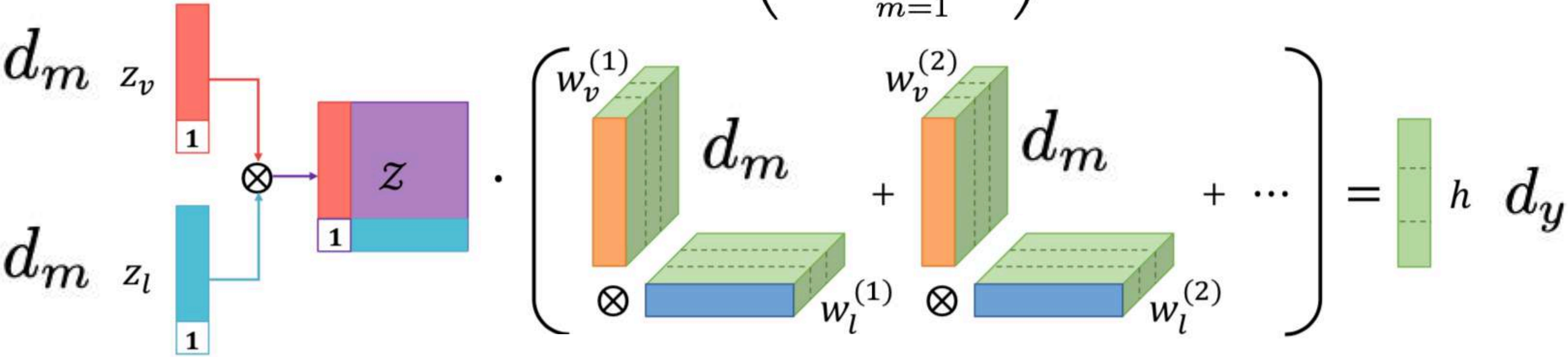
- Rank-r approximation $O\left(d_y \times \prod_{m=1}^{M} d_m\right)$



Low-rank Tensor Approximation

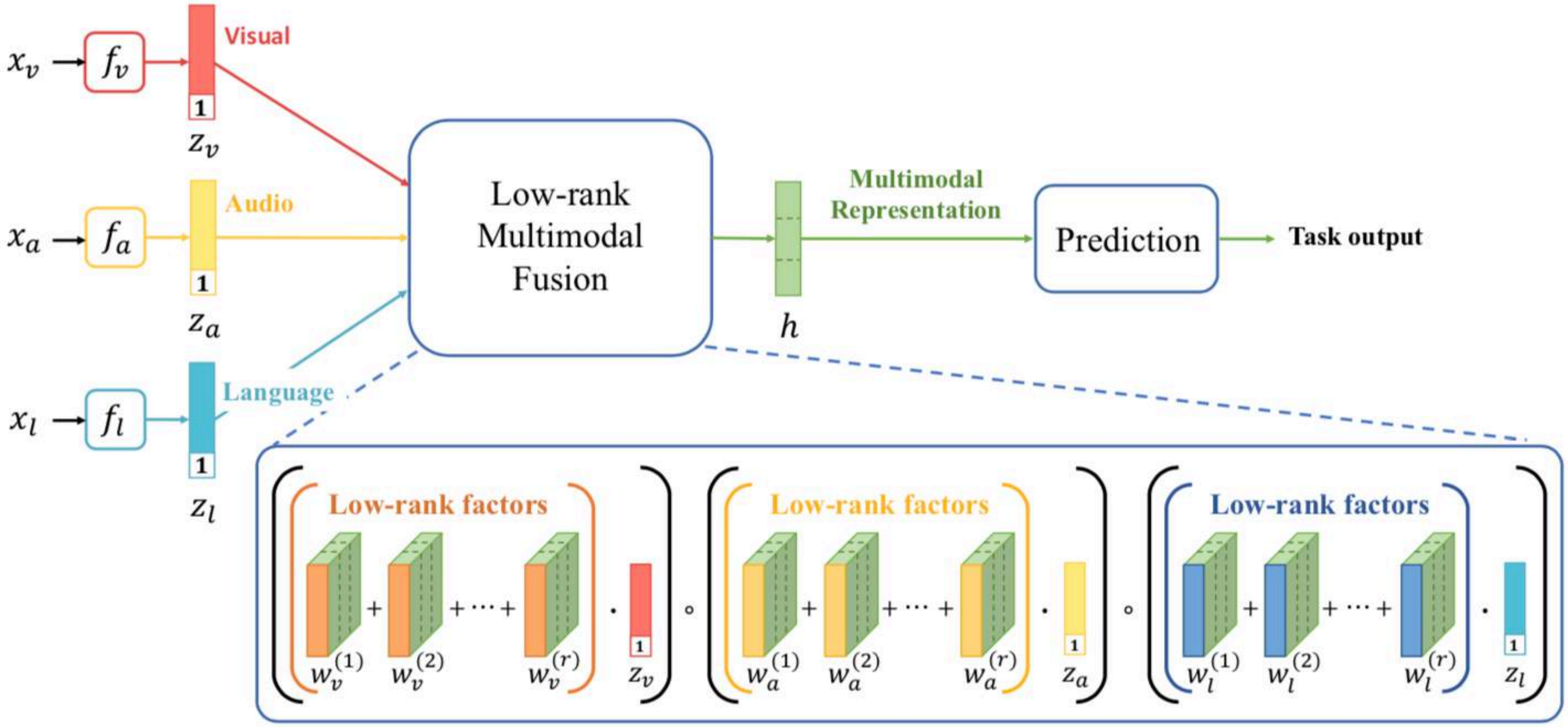


- Rank-r approximation $O\left(d_y \times \prod_{m=1}^{M} d_m\right)$



$$O\left(d_y \times r \times \sum_{m=1}^{M} d_m\right)$$

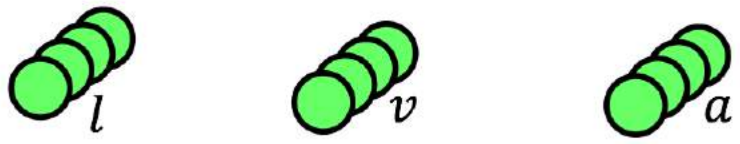
Low-rank Multimodal Fusion



Results

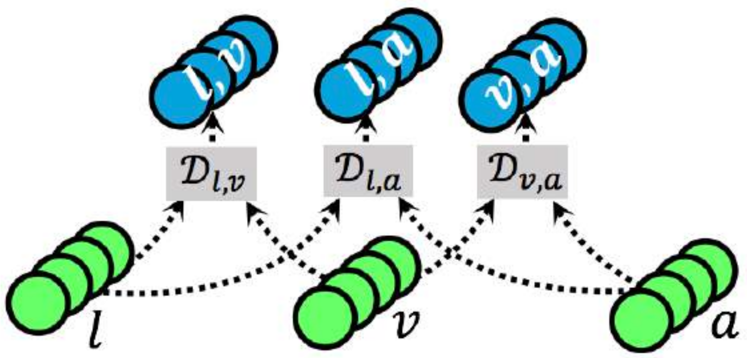
Dataset	CMU-MOSI					POM			IEMOCAP			
Metric	MAE	Corr	Acc-2	F1	Acc-7	MAE	Corr	Acc	F1-Happy	F1-Sad	F1-Angry	F1-Neutral
SVM	1.864	0.057	50.2	50.1	17.5	0.887	0.104	33.9	81.5	78.8	82.4	64.9
DF	1.143	0.518	72.3	72.1	26.8	0.869	0.144	34.1	81.0	81.2	65.4	44.0
BC-LSTM	1.079	0.581	73.9	73.9	28.7	0.840	0.278	34.8	81.7	81.7	84.2	64.1
MV-LSTM	1.019	0.601	73.9	74.0	33.2	0.891	0.270	34.6	81.3	74.0	84.3	66.7
MARN	0.968	0.625	77.1	77.0	34.7	-	-	39.4	83.6	81.2	84.2	65.9
MFN	0.965	0.632	**77.4**	**77.3**	**34.1**	0.805	0.349	41.7	84.0	82.1	83.7	69.2
TFN	0.970	0.633	73.9	73.4	32.1	0.886	0.093	31.6	83.6	82.8	84.2	65.4
LMF	**0.912**	**0.668**	76.4	75.7	32.8	**0.796**	**0.396**	**42.8**	**85.8**	**85.9**	**89.0**	**71.7**

Dynamic Fusion Graph



unimodal

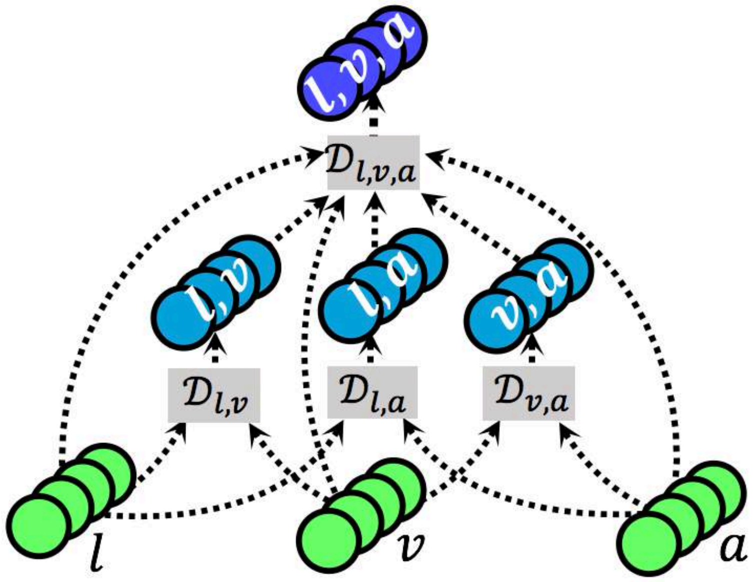
Dynamic Fusion Graph



bimodal

unimodal

Dynamic Fusion Graph

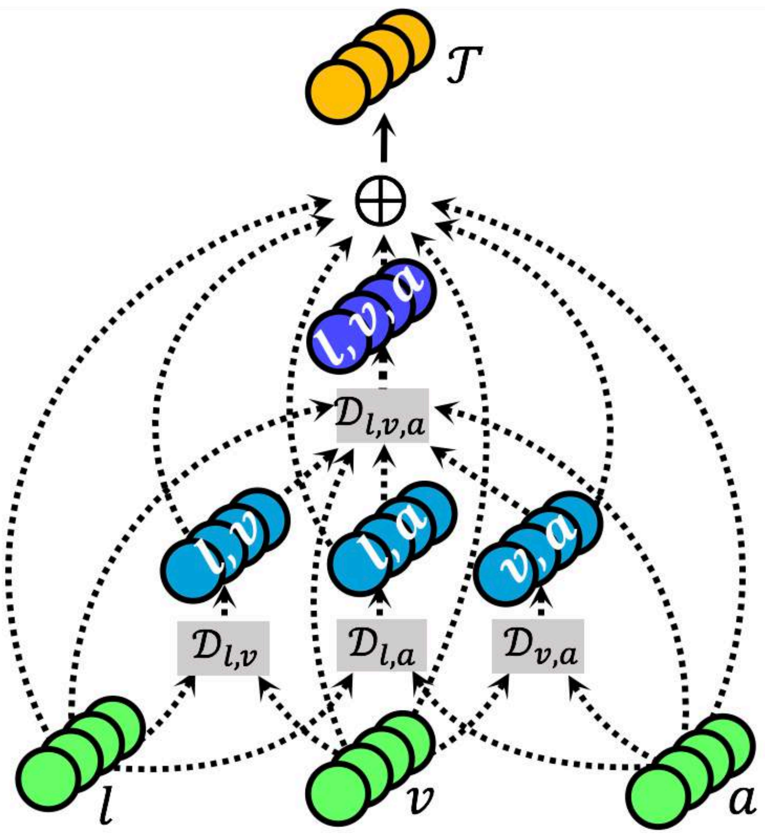


trimodal

bimodal

unimodal

Dynamic Fusion Graph



multimodal representation

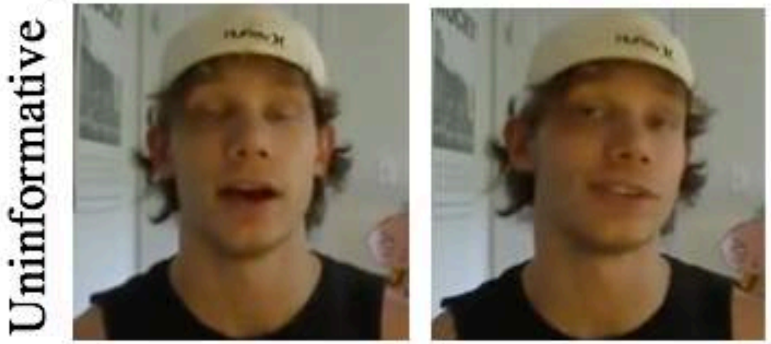
trimodal

bimodal

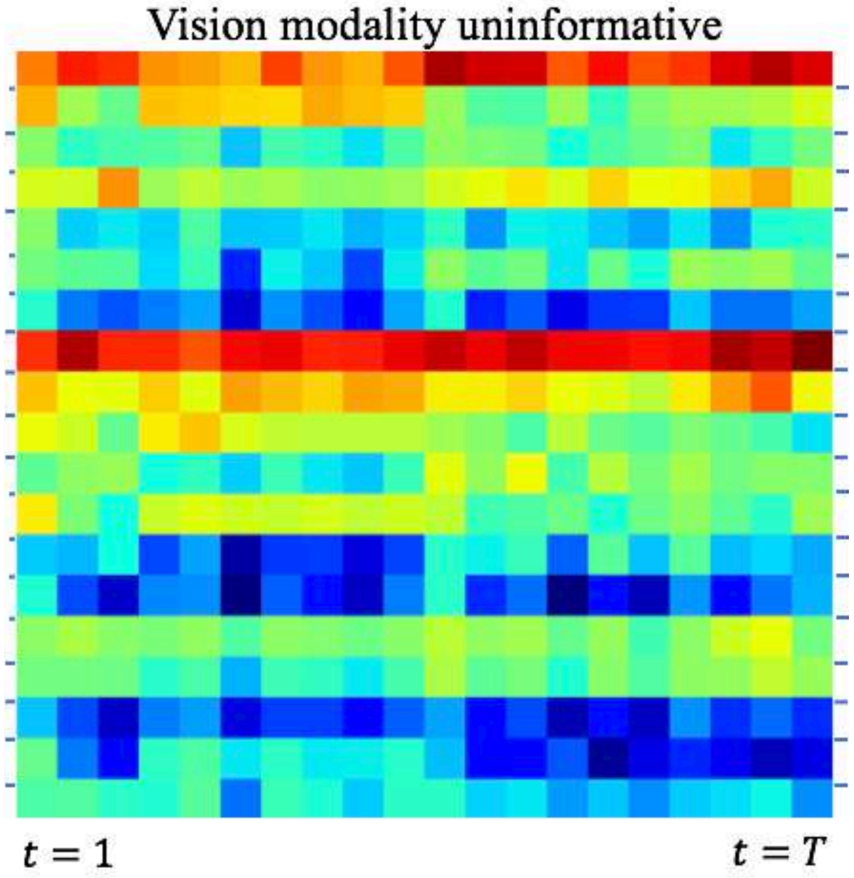
unimodal

Interpretable Fusion

Too much too fast, I mean we basically just get introduced to this character…



(angry voice)



unimodal visual

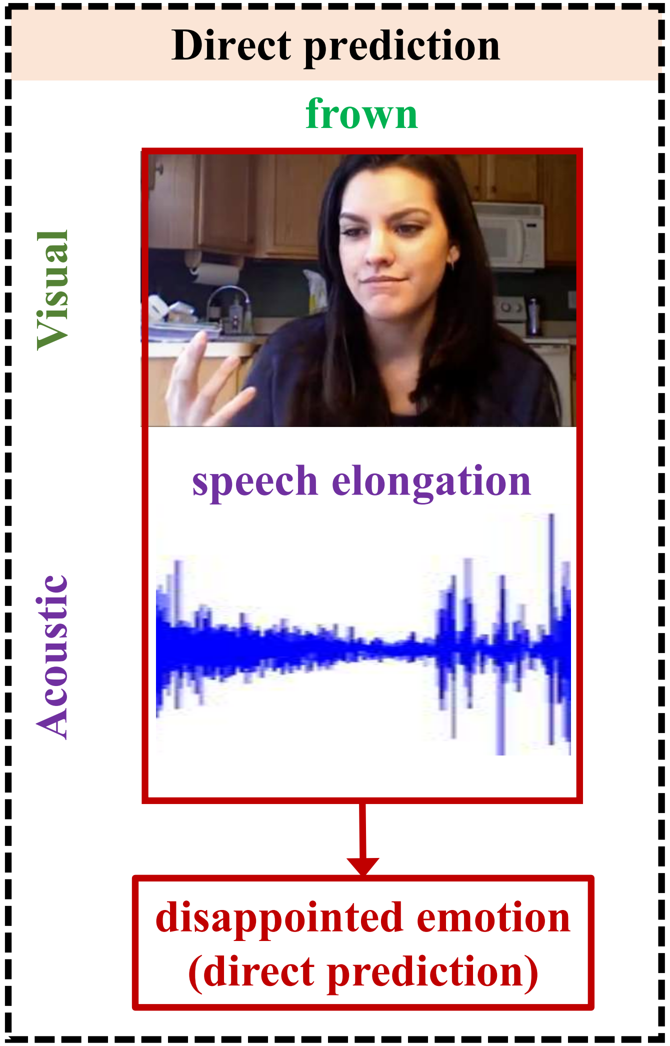
bimodal visual

trimodal visual

Direction 3: Direct and Relative

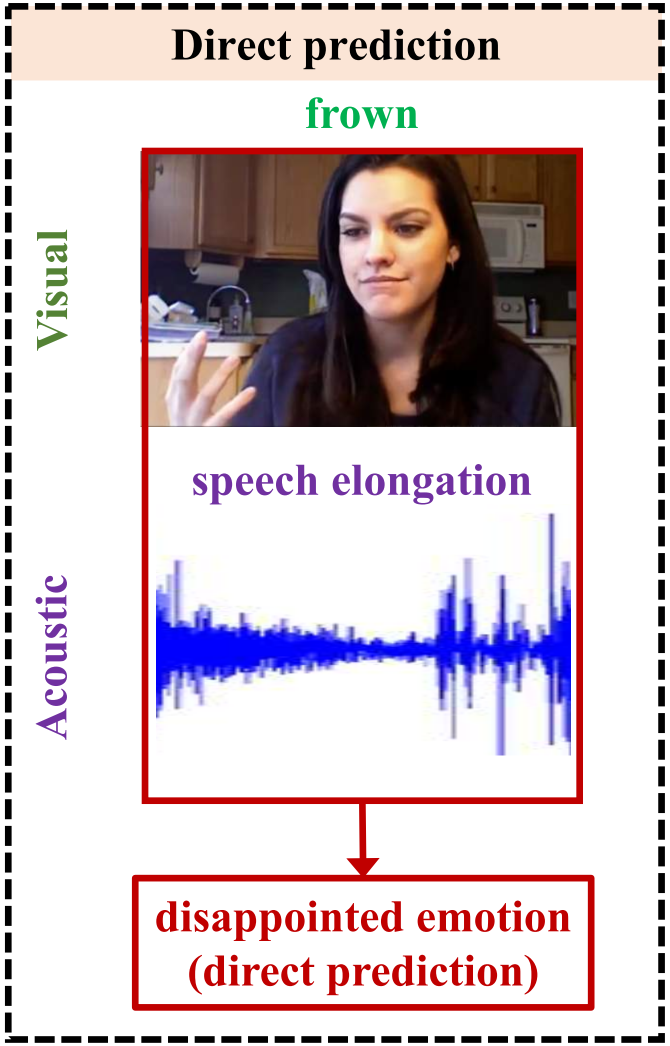
Person-independent Features

- Universal emotion expressions



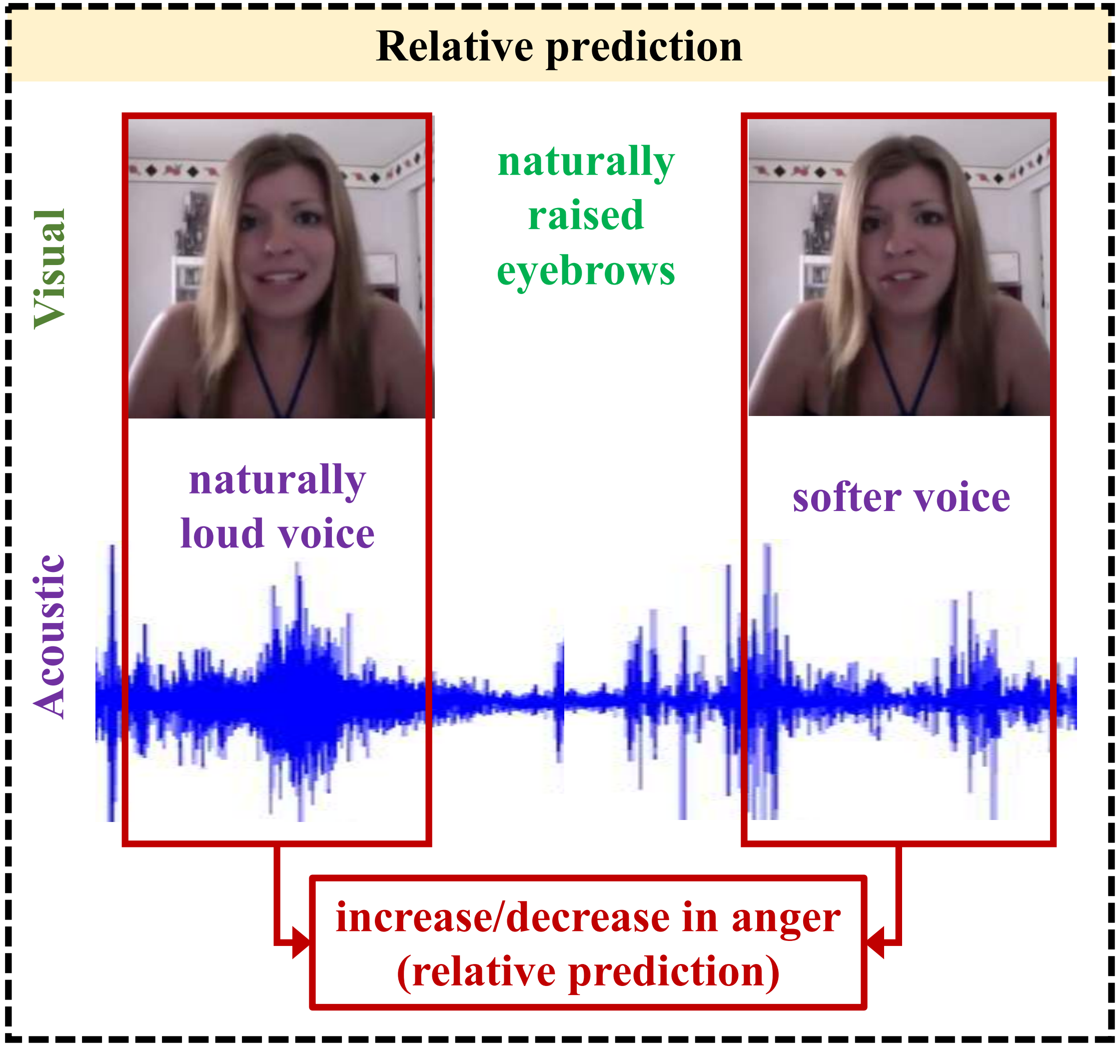
Person-independent Features

- Universal emotion expressions
- Absolute emotions can be **directly** inferred from these observed behaviors



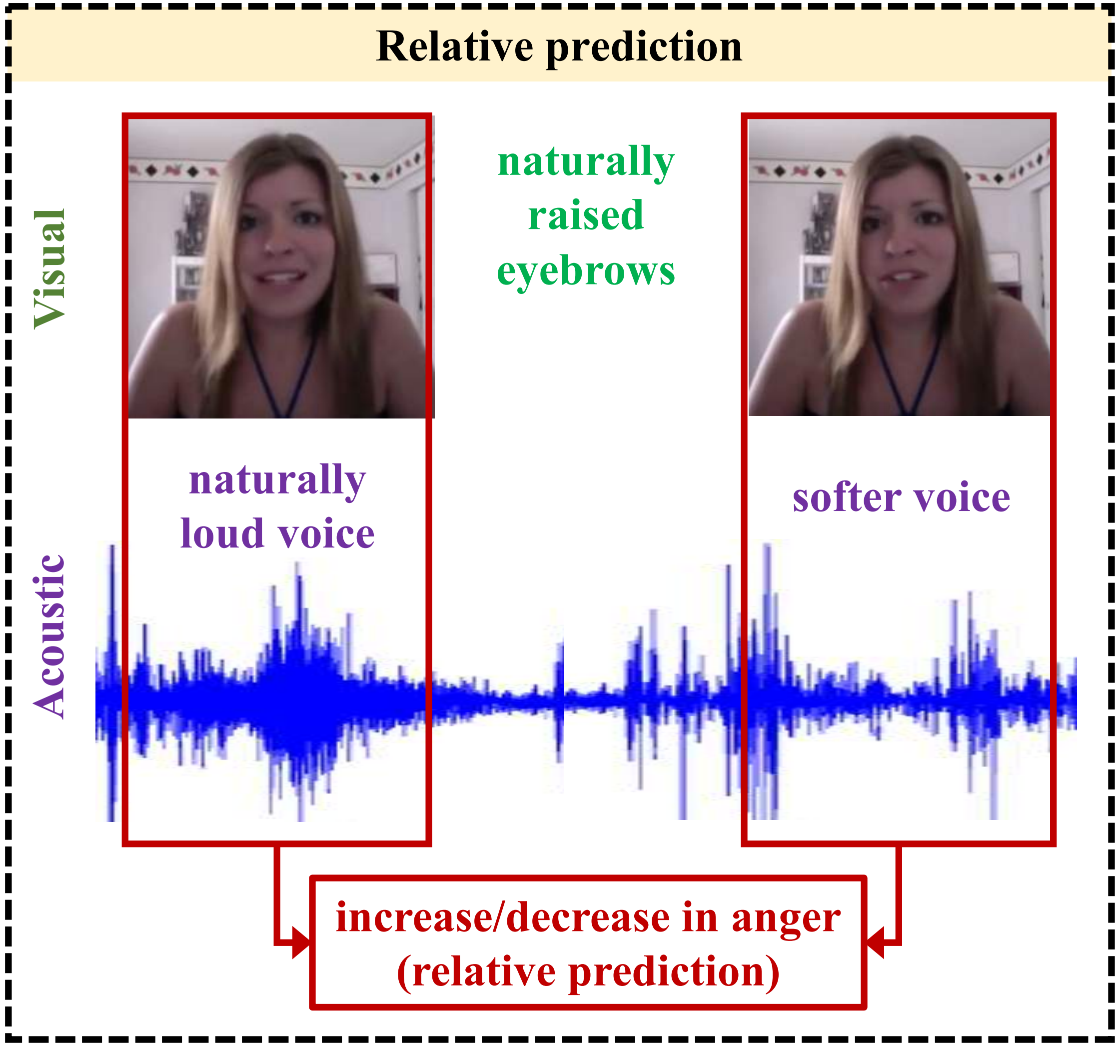
Person-dependent Features

- Emotions are also expressed with idiosyncratic behaviors



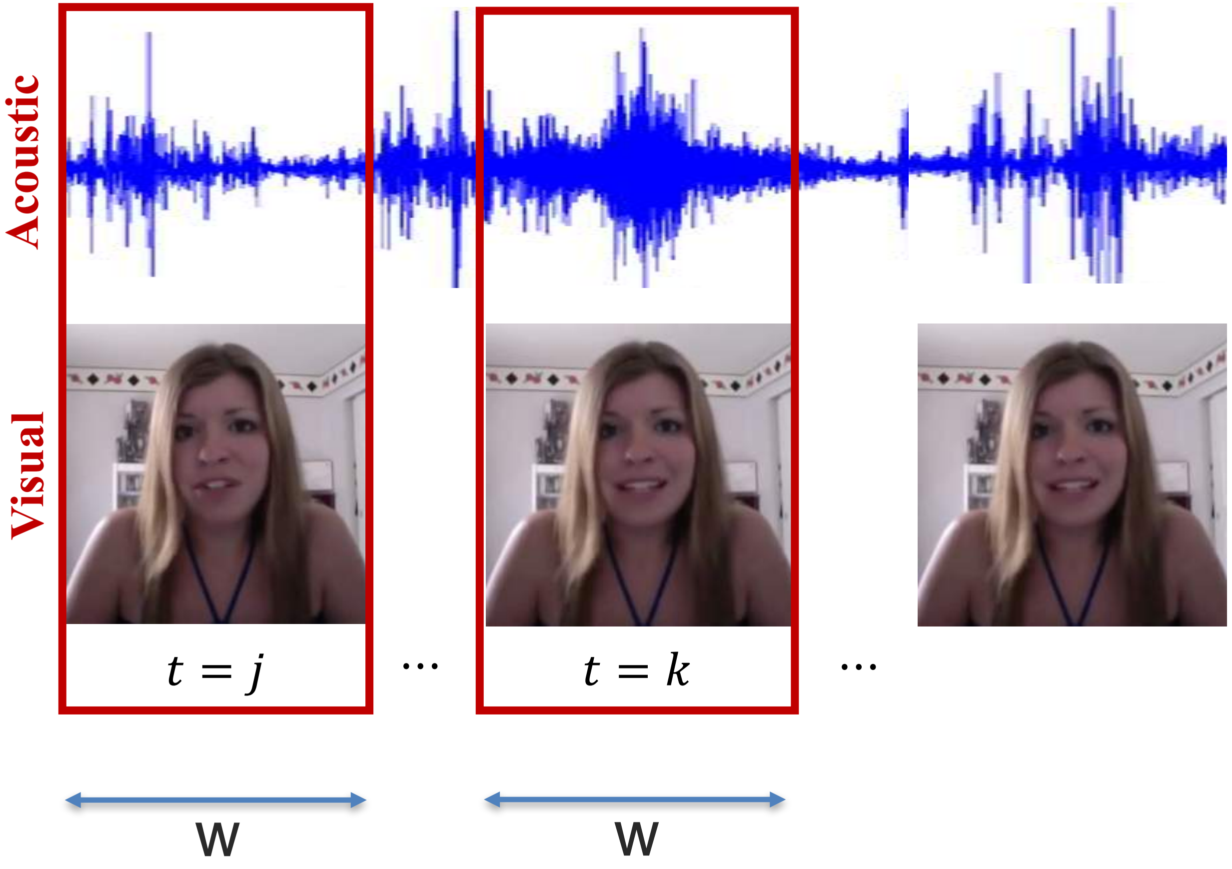
Person-dependent Features

- Emotions are also expressed with idiosyncratic behaviors
- Estimate **relative** changes by comparing behaviors



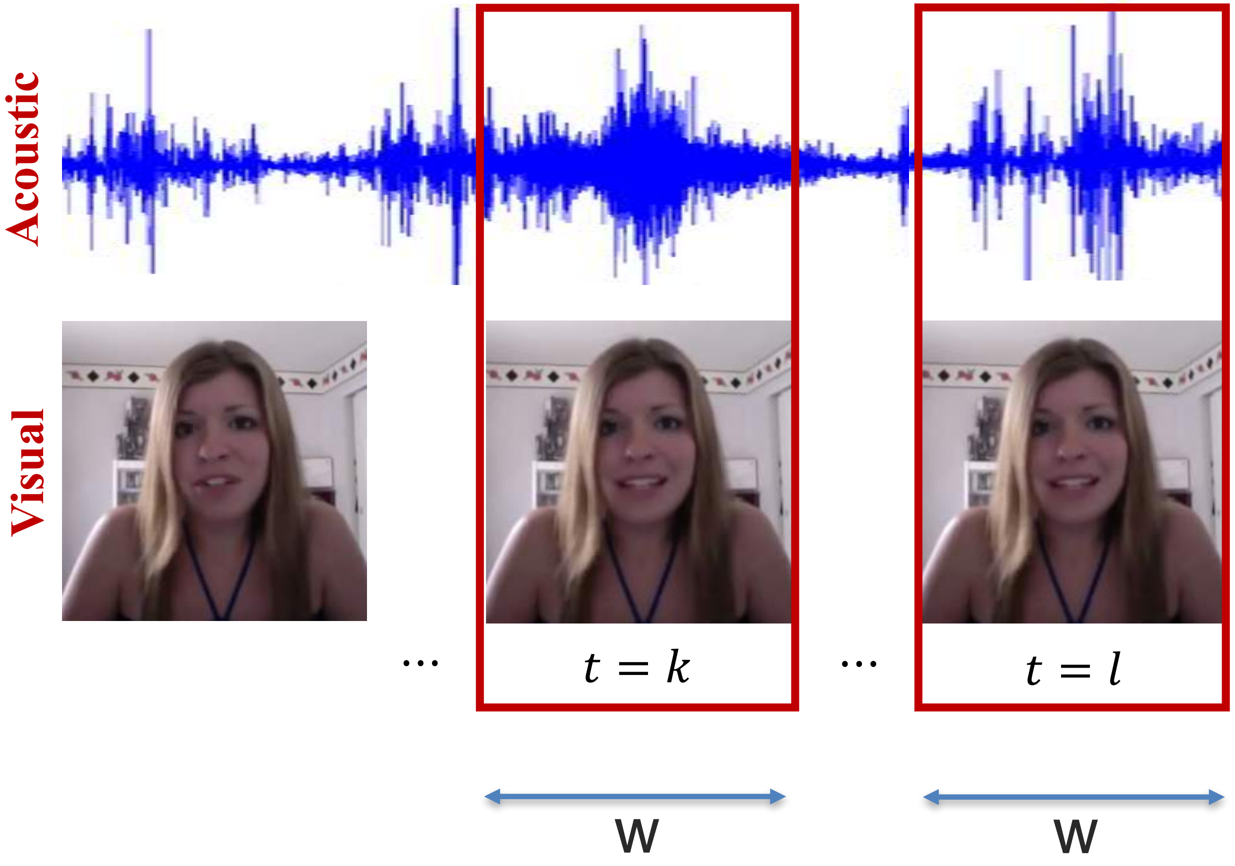
Multimodal Local Ranking

- Video centered at 2 random indices j,k
- **w** window size



Multimodal Local Ranking

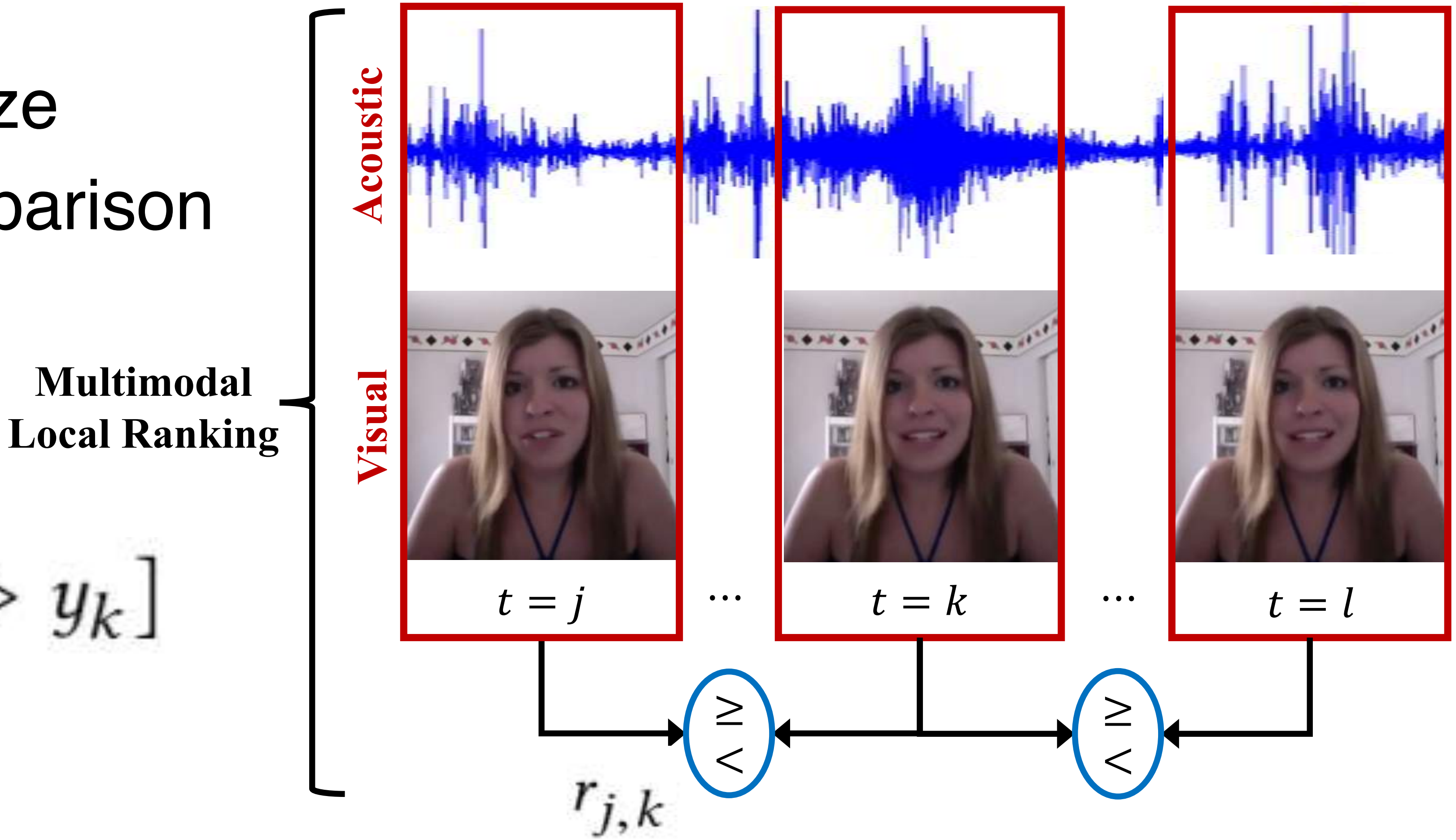
- Video centered at 2 random indices k,l
- **w** window size



Multimodal Local Ranking

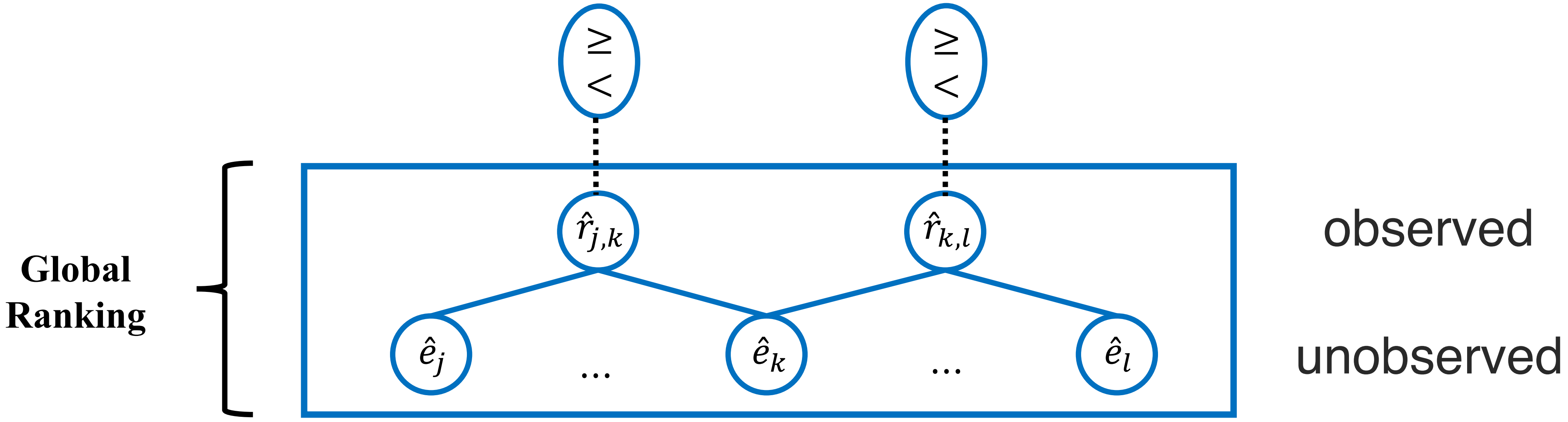
- **w** window size
- **m** local comparison pairs

$$r_{j,k} = \mathbb{I}[y_j > y_k]$$



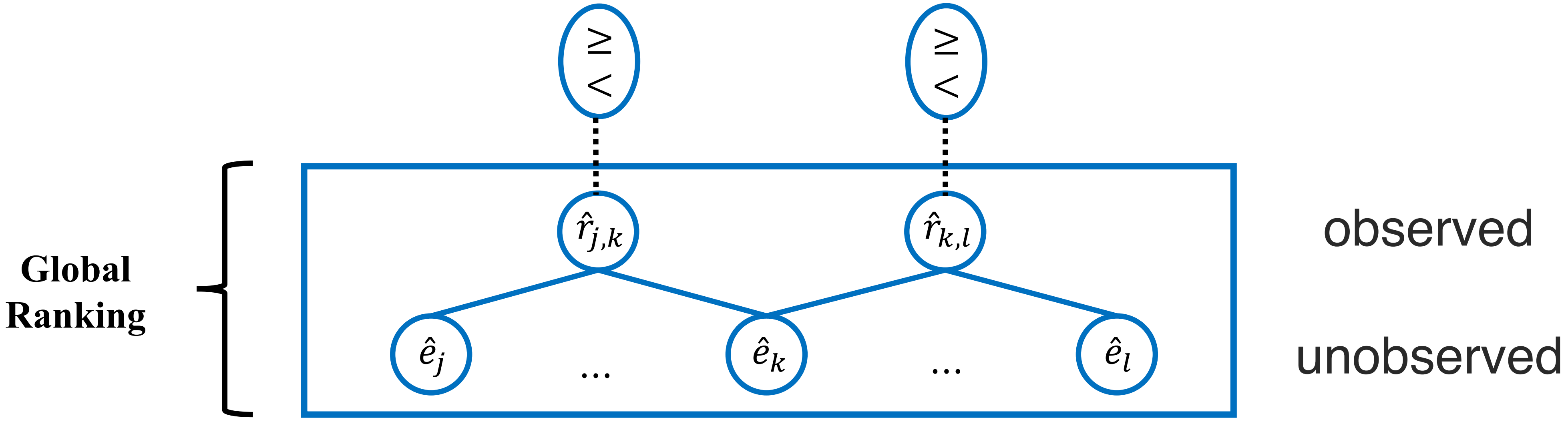
Global Ranking

- Bayesian ranking algorithm



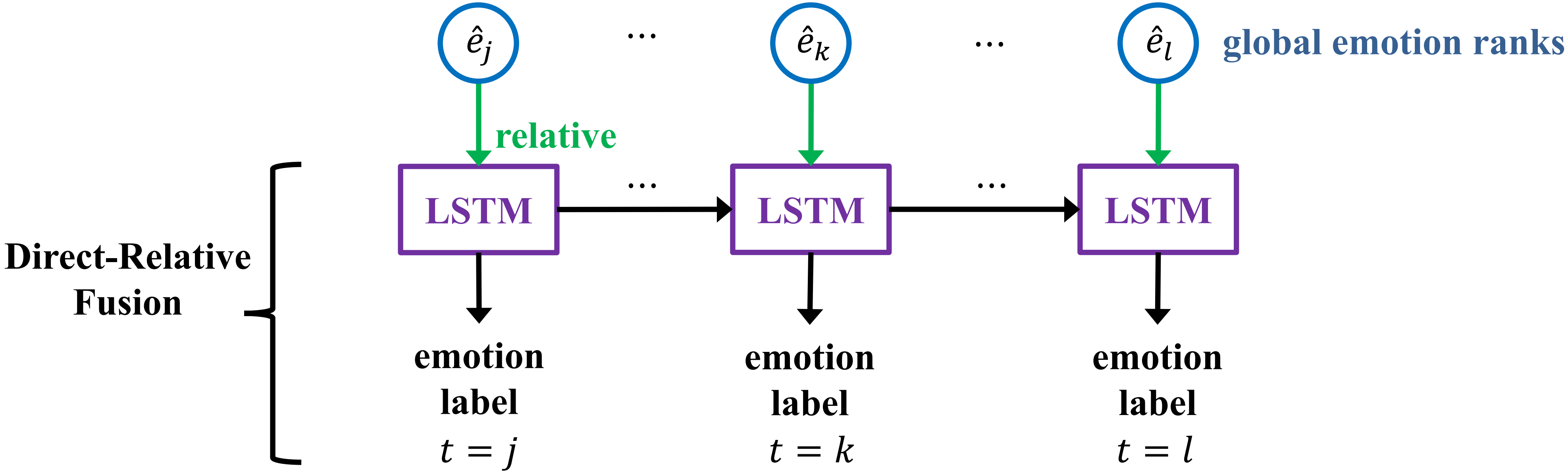
Global Ranking

- Bayesian ranking algorithm

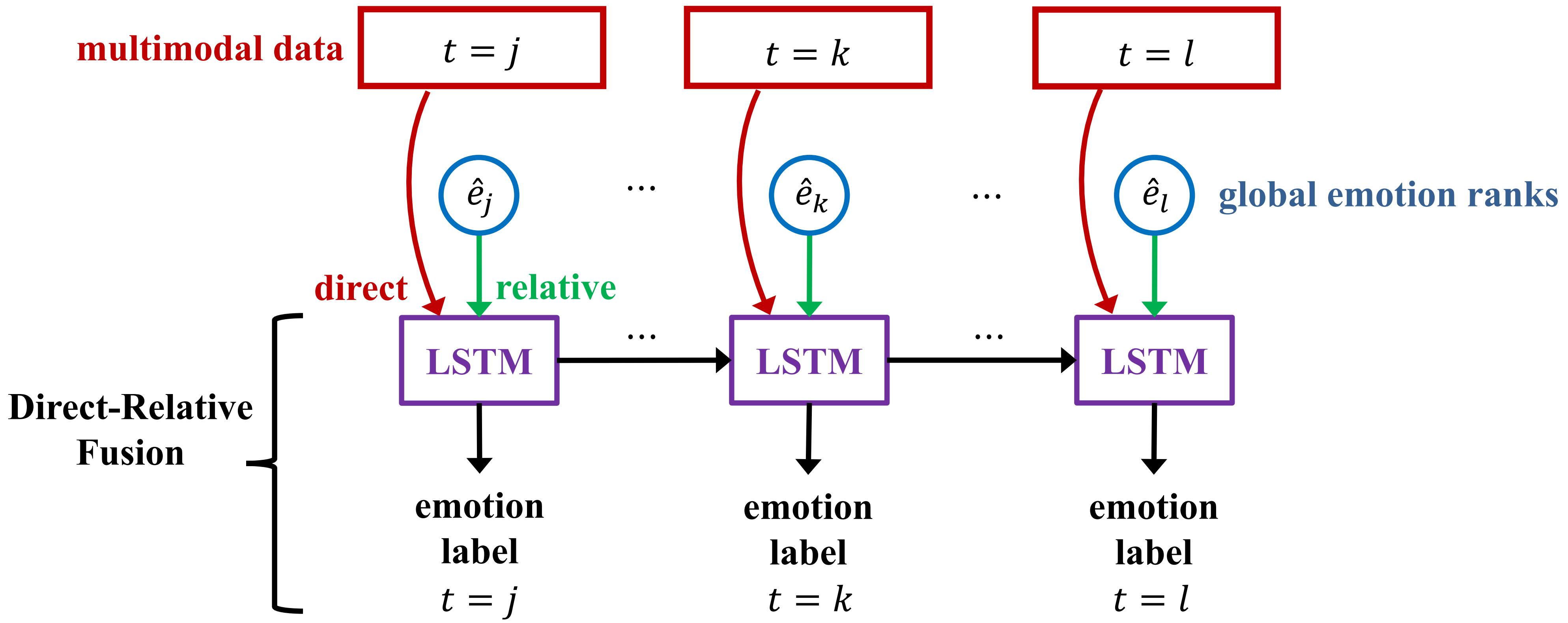


$$r_{j,k} = \mathbb{I}[y_j > y_k] \qquad p(r_{j,k} = 1 | e_j, e_k) = p(e_j > e_k)$$

Direct-Relative Fusion



Direct-Relative Fusion



Results

Dataset	AVEC16	
Task	Arousal	Valence
Metric	CCC	CCC
EF-(-/S/B/SB)LSTM [9, 11, 29]	0.4327	0.4667
Gated-LSTM [38]	0.3210	0.4667
MV-LSTM, view-specific [27]	0.4530	0.4431
MV-LSTM, coupled [27]	0.4300	0.4477
MV-LSTM, hybrid [27]	0.4729	0.4924
MV-LSTM, fully connected [27]	0.4293	0.4896
MLRF-500	0.4732	0.5063
MLRF-1000	**0.5049**	**0.5432**
Improvement over baselines	↑ 0.032	↑ 0.0508

Effect of Window Size

Dataset	AVEC16	
Task	Arousal	Valence
Metric	CCC	CCC
MLRF-500 $w = 10$	0.4165	0.2377
MLRF-500 $w = 50$	0.4168	0.4175
MLRF-500 $w = 100$	0.4196	0.4340
MLRF-500 $w = 200$	**0.4732**	**0.5063**

Effect of Direct and Relative Approaches

Dataset	AVEC16	
Task	Arousal	Valence
Metric	CCC	CCC
MLRF-500 direct predictions only	0.4327	0.4667
MLRF-500 relative predictions only	0.3646	0.0402
MLRF-500	**0.4732**	**0.5063**
MLRF-1000 direct predictions only	0.4327	0.4667
MLRF-1000 relative predictions only	0.4297	0.0846
MLRF-1000	**0.5049**	**0.5432**

Direction 4: Multimodal Representation Learning

Representation Learning

- Discriminative: $P(\mathbf{Y}|\mathbf{X}_1, \cdots, \mathbf{X}_M)$

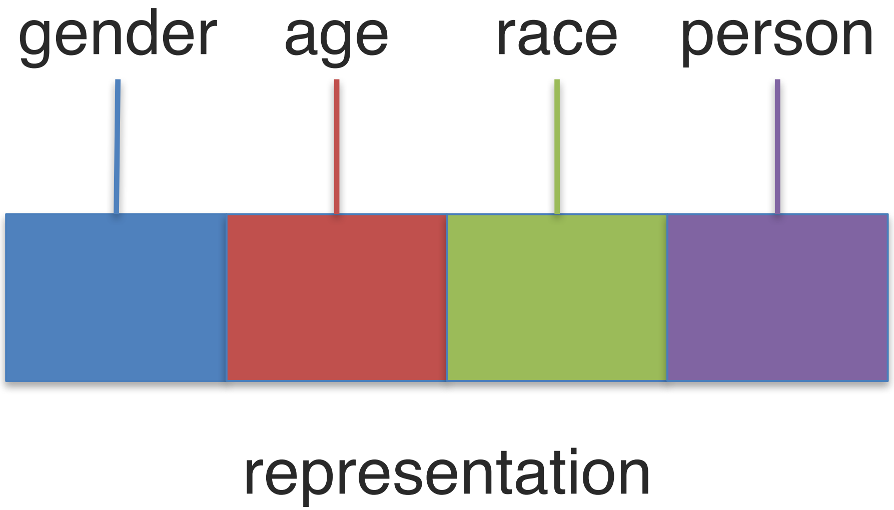
Representation Learning

- Discriminative: $P(\mathbf{Y}|\mathbf{X}_1, \cdots, \mathbf{X}_M)$
- Generative: $P(\mathbf{X}_1, \cdots, \mathbf{X}_M)$

Representation Learning

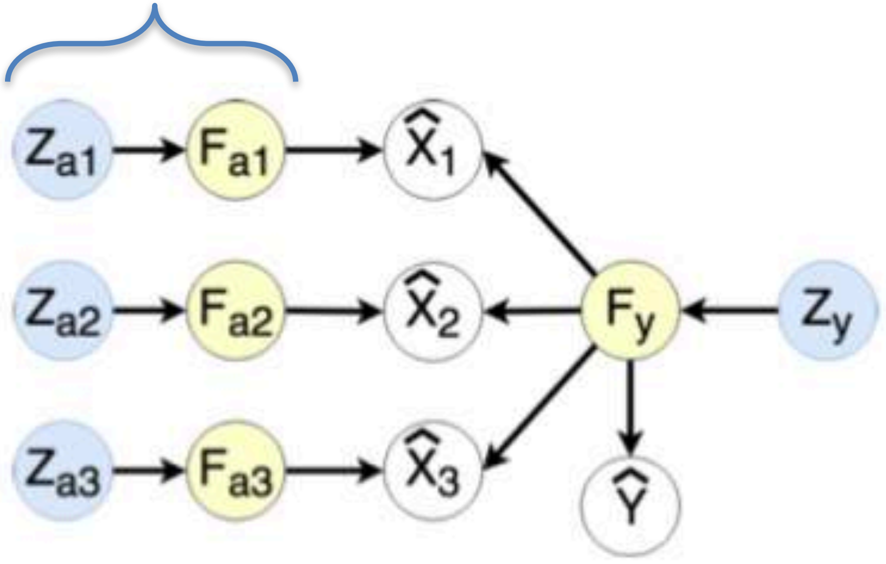
- Discriminative: $P(\mathbf{Y}|\mathbf{X}_1, \cdots, \mathbf{X}_M)$
- Generative: $P(\mathbf{X}_1, \cdots, \mathbf{X}_M)$
- Specificity: modality-specific and multimodal

Factorized Representations



Multimodal Factorization Model

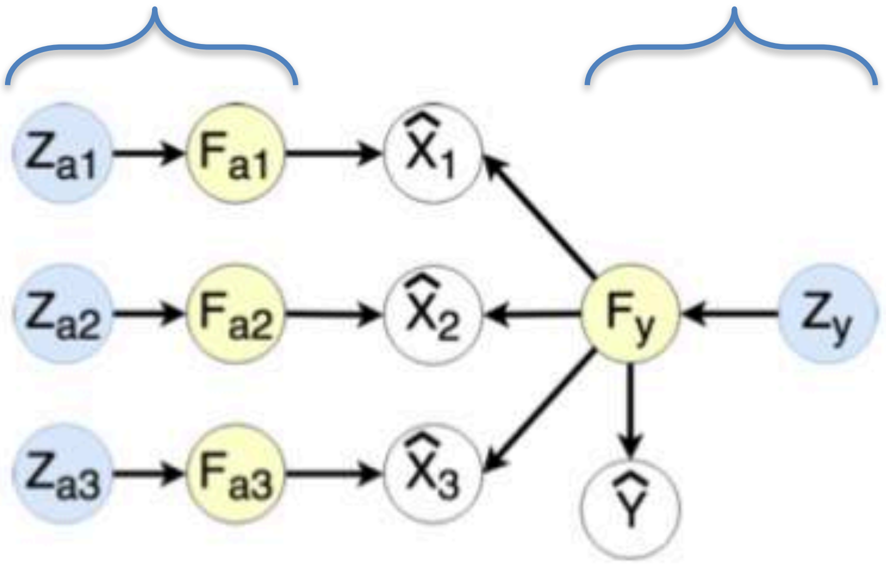
Modality-specific generative factors



Multimodal Factorization Model

Modality-specific generative factors

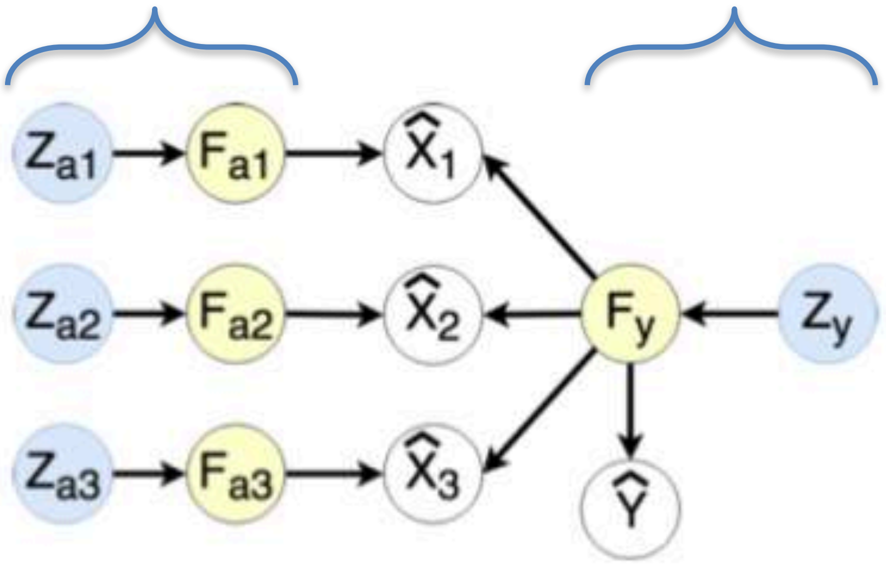
Multimodal discriminative factor



Generative-Discriminative Objective

Modality-specific generative factors

Multimodal discriminative factor

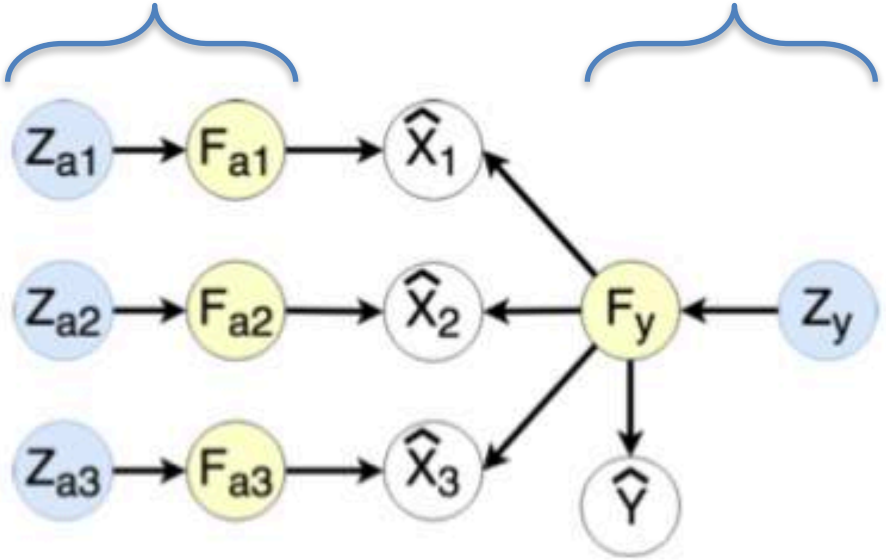


$$\underbrace{\left[\sum_{i=1}^{M} c_{X_i}\Big(\mathbf{X}_i, F\big(G_{ai}(\mathbf{Z_{ai}}), G_y(\mathbf{Z_y})\big)\Big)\right.}_{\text{Generative}}$$

Generative-Discriminative Objective

Modality-specific generative factors

Multimodal discriminative factor

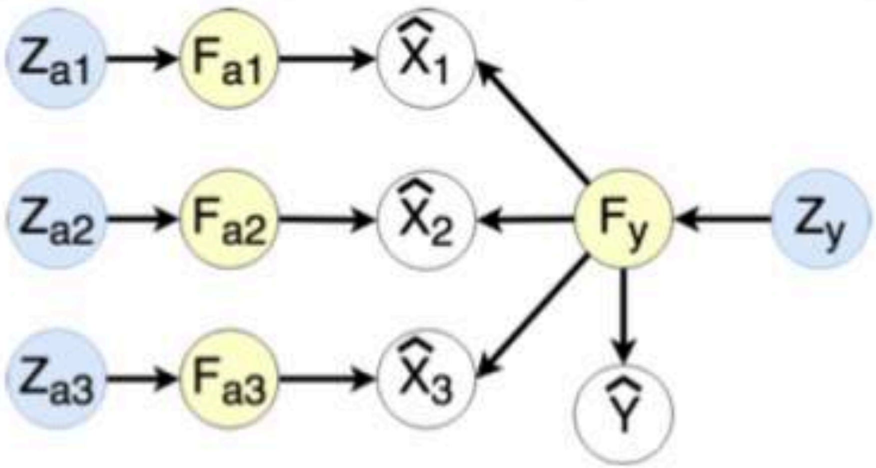


$$\left[\underbrace{\sum_{i=1}^{M} c_{X_i}\Big(\mathbf{X}_i, F(G_{ai}(\mathbf{Z_{ai}}), G_y(\mathbf{Z_y}))\Big)}_{\text{Generative}} + \underbrace{c_Y\Big(\mathbf{Y}, D(G_y(\mathbf{Z_y}))\Big)}_{\text{Discriminative}}\right]$$

Generative-Discriminative Objective

Modality-specific generative factors

Multimodal discriminative factor

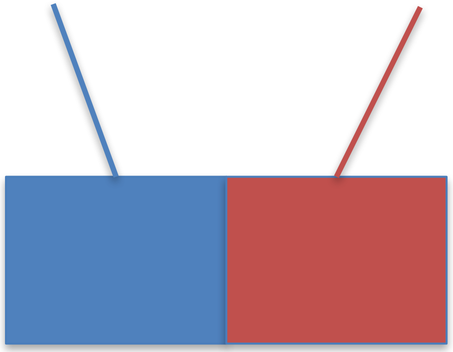


$$\underbrace{\left[\sum_{i=1}^{M} c_{X_i}\Big(\mathbf{X}_i, F(G_{ai}(\mathbf{Z_{ai}}), G_y(\mathbf{Z_y}))\Big)}_{\text{Generative}} + \underbrace{c_Y\Big(\mathbf{Y}, D(G_y(\mathbf{Z_y}))\Big)\right]}_{\text{Discriminative}} + \underbrace{\lambda \mathcal{MMD}(Q_{\mathbf{Z}}, P_{\mathbf{Z}}),}_{\text{Regularizer}}$$

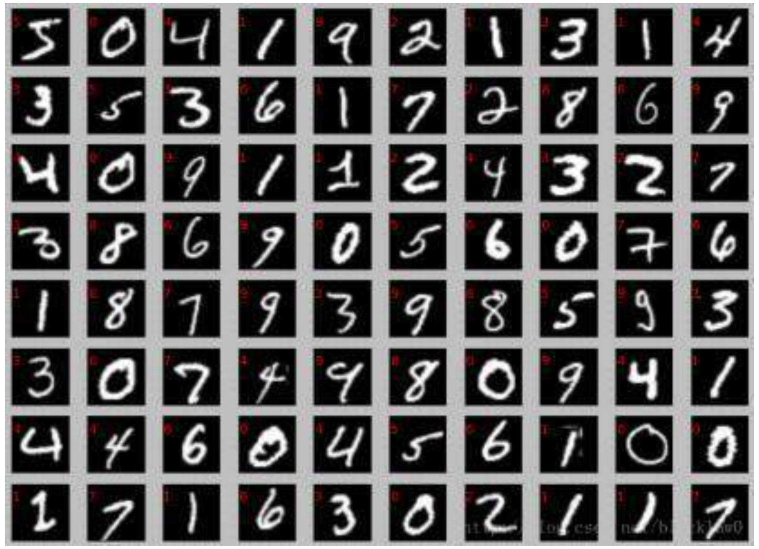
Unimodal Generation Results

za: style

zy: label 0-9

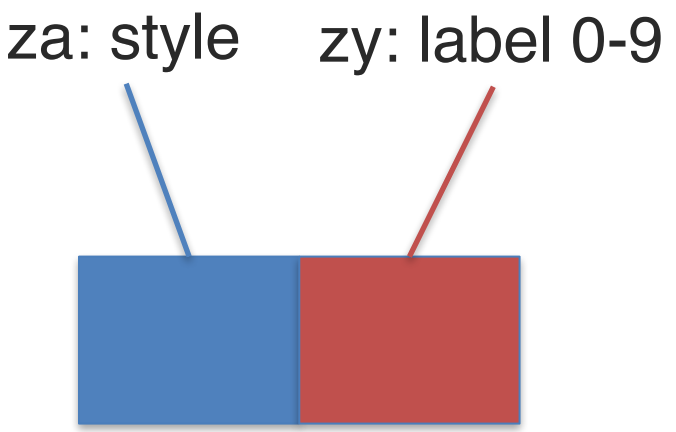


MNIST

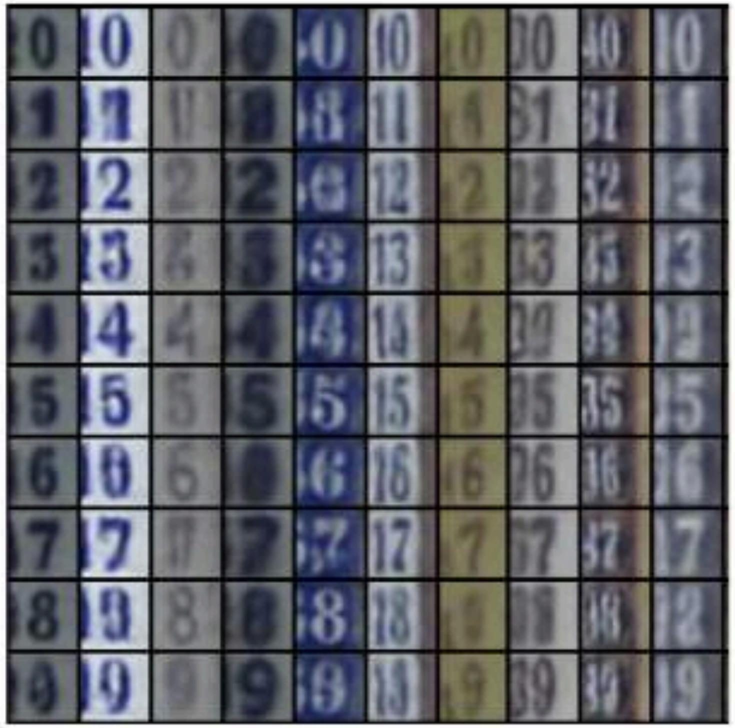


SVHN

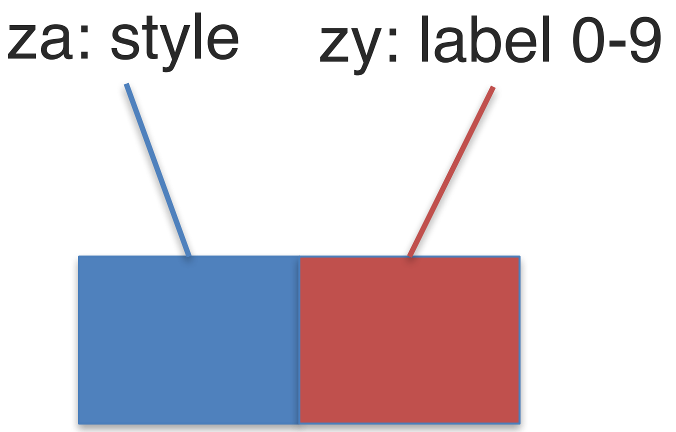
Unimodal Generation Results

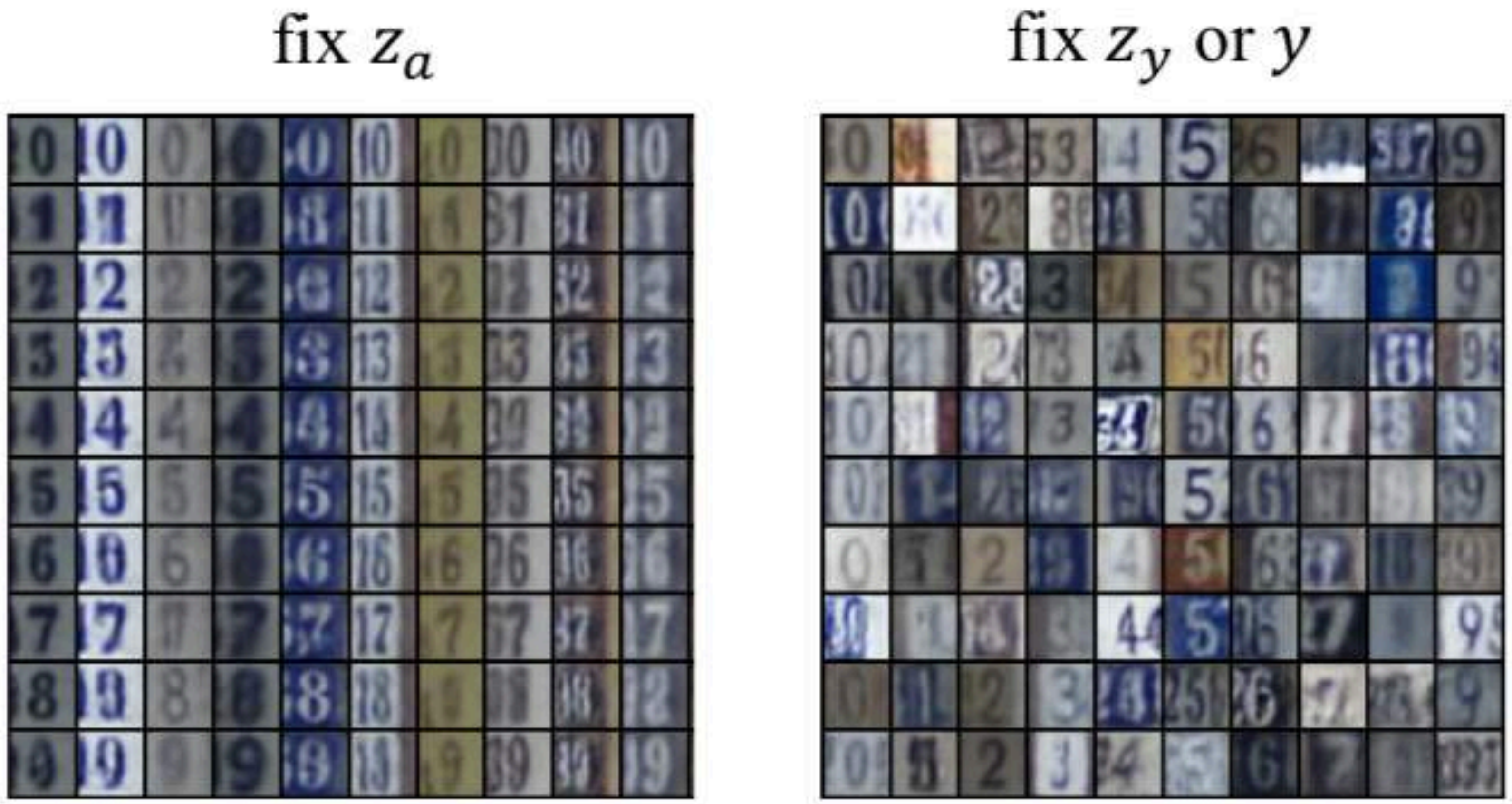


fix z_a

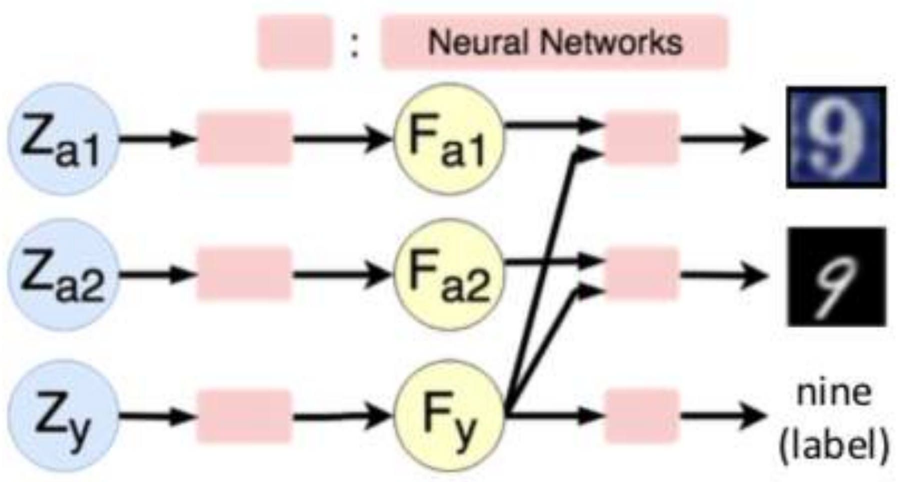


Unimodal Generation Results

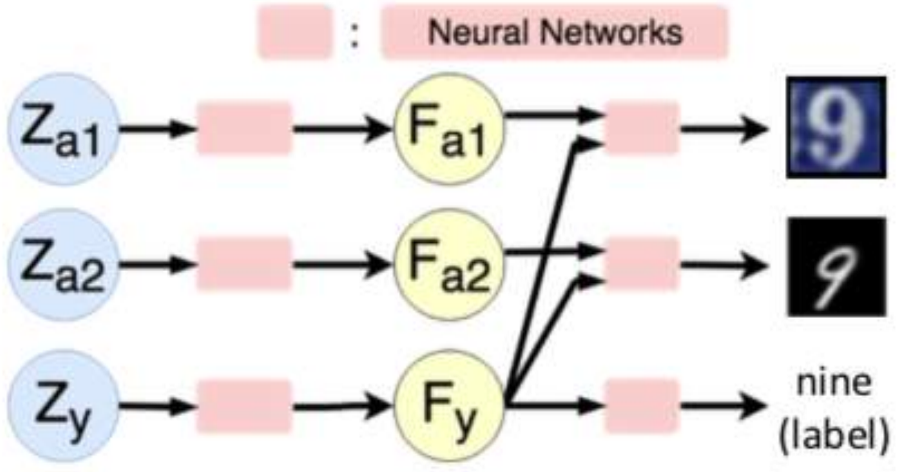




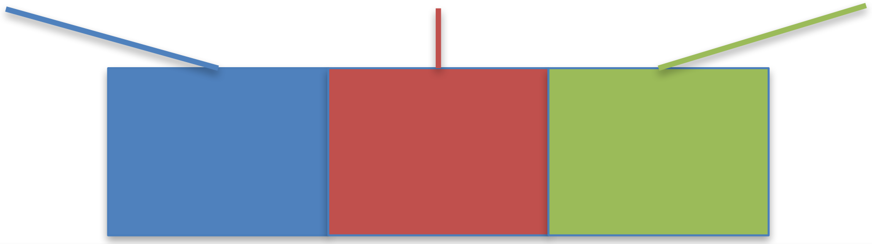
Multimodal Generative Results



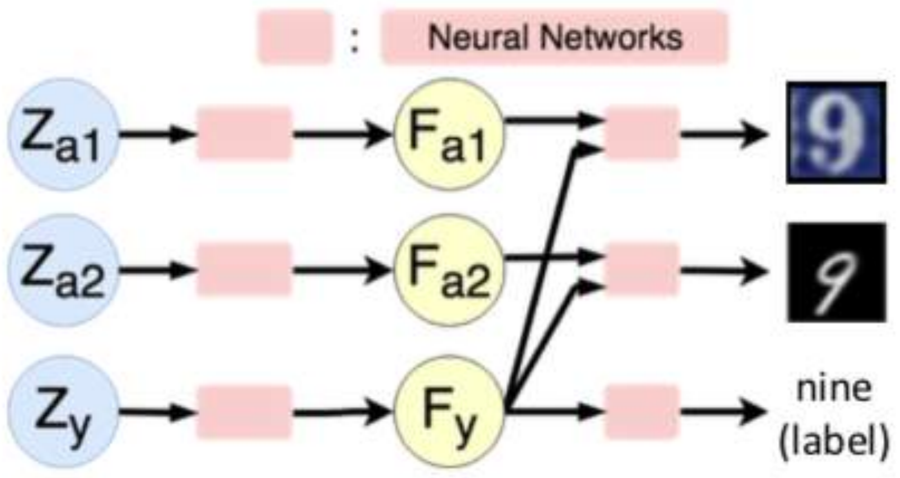
Multimodal Generative Results

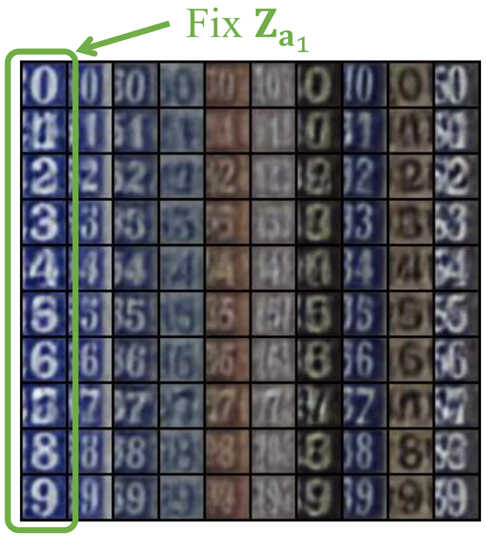


za1: SVHN style zy: label za2: MNIST style

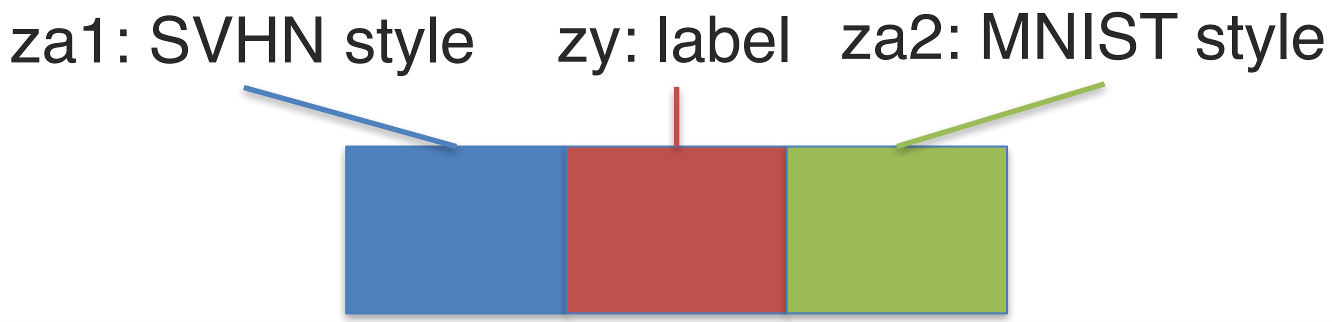


Multimodal Generative Results

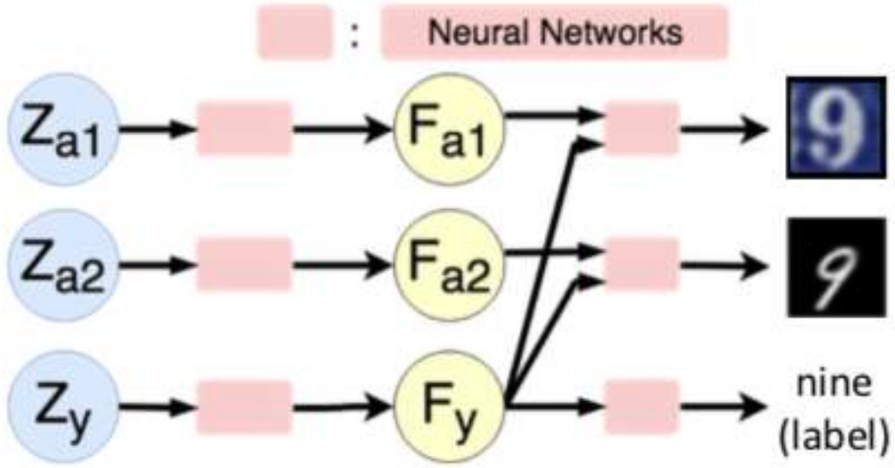


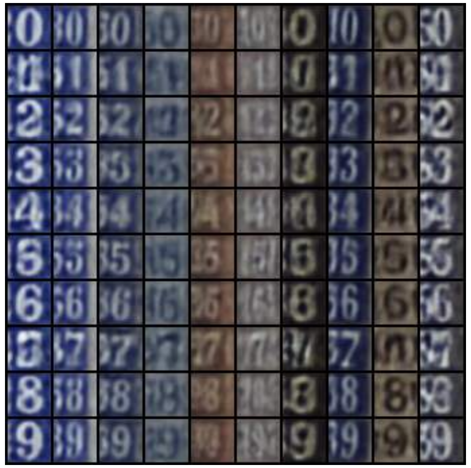


Modality 1 (**SVHN**)



Multimodal Generative Results

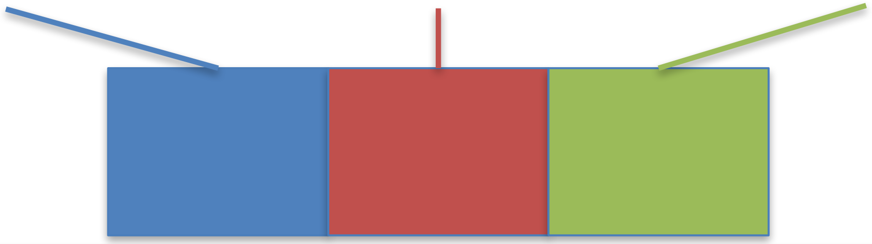




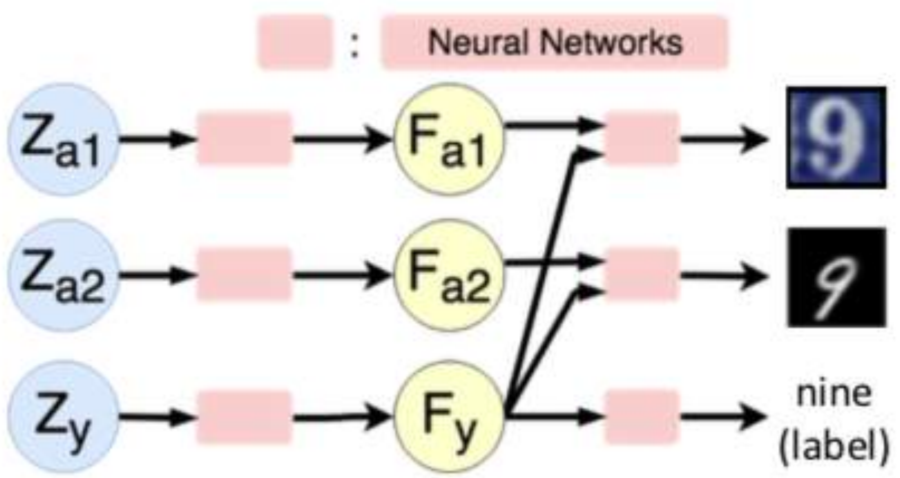
Modality 1 (**SVHN**)

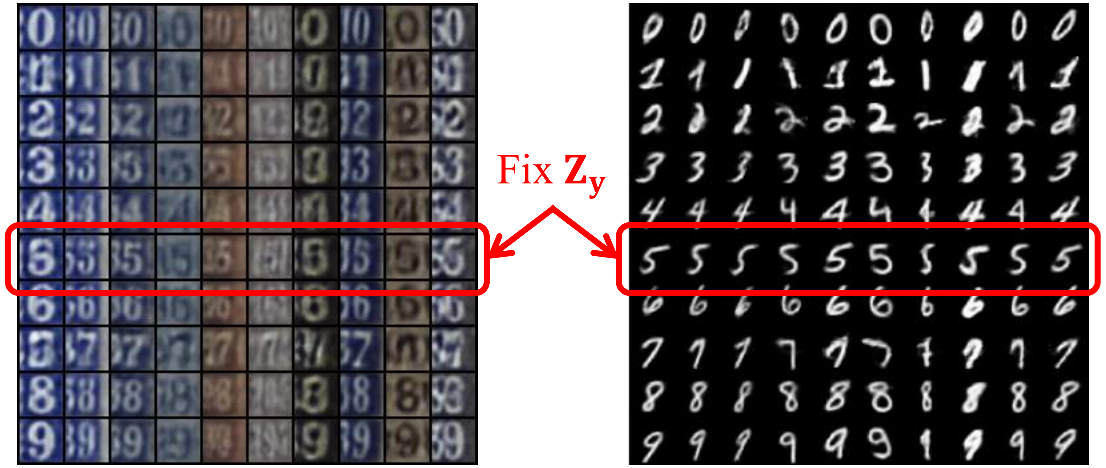
Modality 2 (**MNIST**)

za1: SVHN style zy: label za2: MNIST style

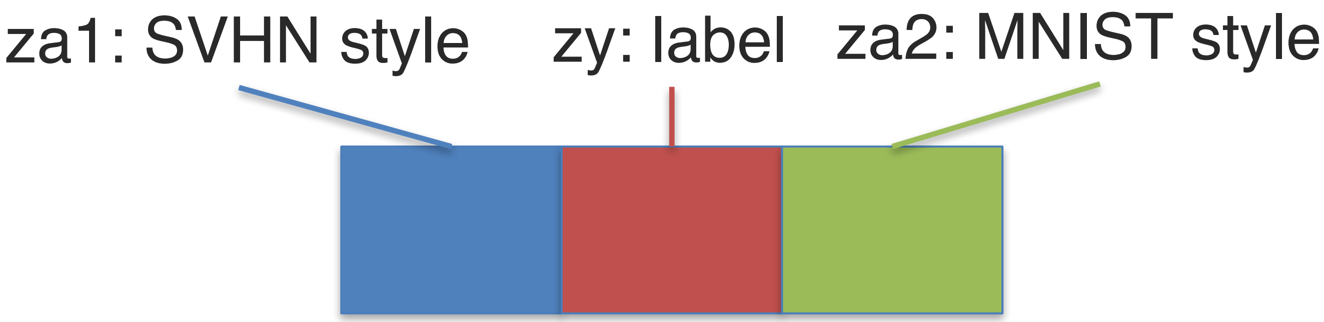


Multimodal Generative Results





Modality 1 (**SVHN**) Modality 2 (**MNIST**)



Multimodal Discriminative Results

Table 2: Results for ablation studies on CMU-MOSI. Best results in bold. All the components in MFM are necessary for best performance.

Variants	Multimodal Discriminative Factor	Hybrid Gen.-Discr. Objective	Factorized Gen.-Discr. Factors	Modal.-Speci. Generative Factors	Dataset : CMU-MOSI Task : Sentiment A^2	F1	A^7	MAE	r
M_E	yes	no	–	–	76.1	76.0	28.7	1.043	0.634
M_D	no	no	–	–	74.6	74.7	28.7	1.024	0.626
M_C	yes	yes	no	–	76.5	76.5	31.9	1.071	0.647
M_B	no	yes	no	–	74.9	75.0	33.1	1.023	0.627
M_A	yes	yes	yes	no	75.1	75.1	32.4	1.039	0.645
MFM	yes	yes	yes	yes	**77.3**	**77.2**	**35.4**	**0.961**	**0.661**

Direction 5: Robust Multimodal Representation Learning

Learning Joint Representations: 2 modalities

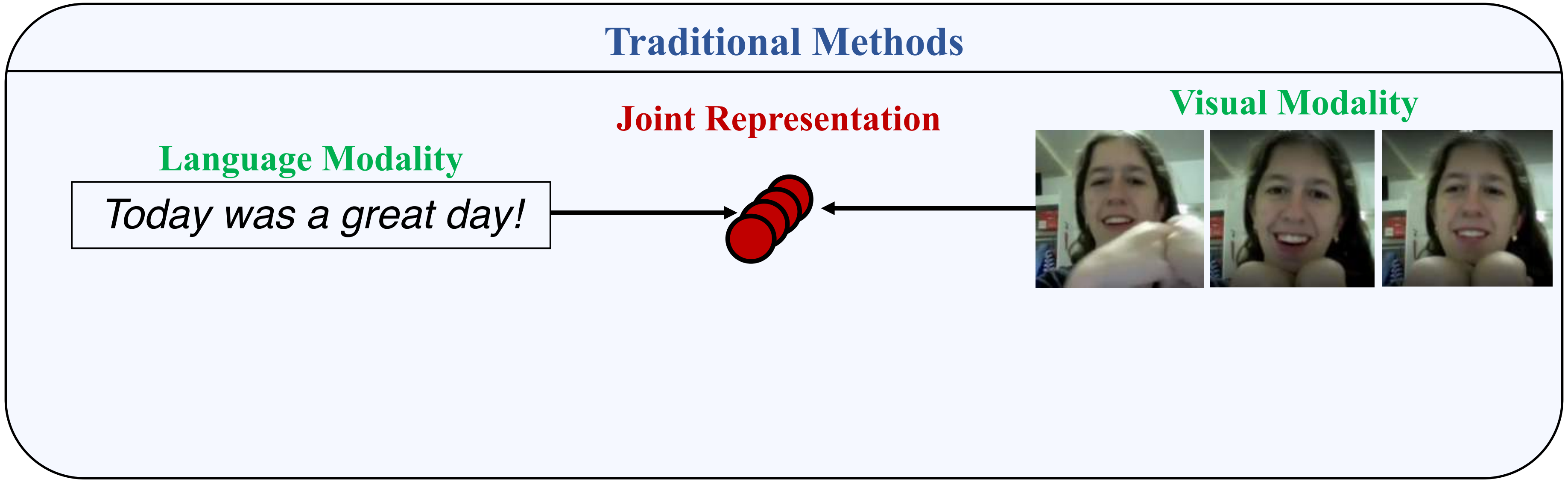
Traditional Methods

Language Modality

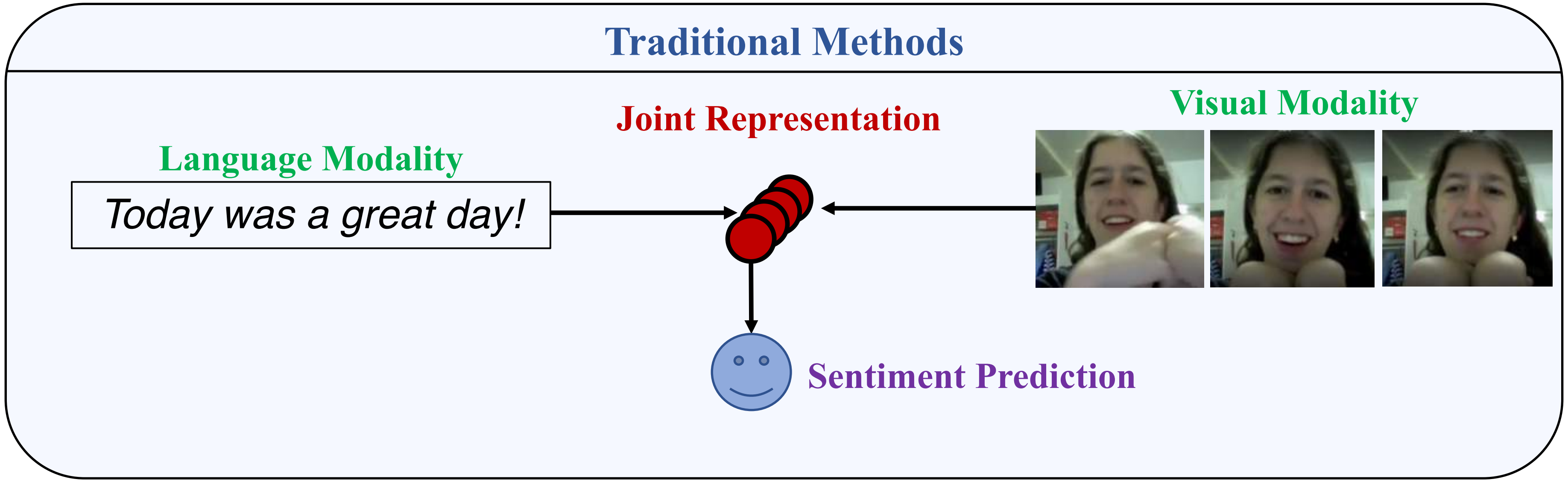
Today was a great day!

Visual Modality

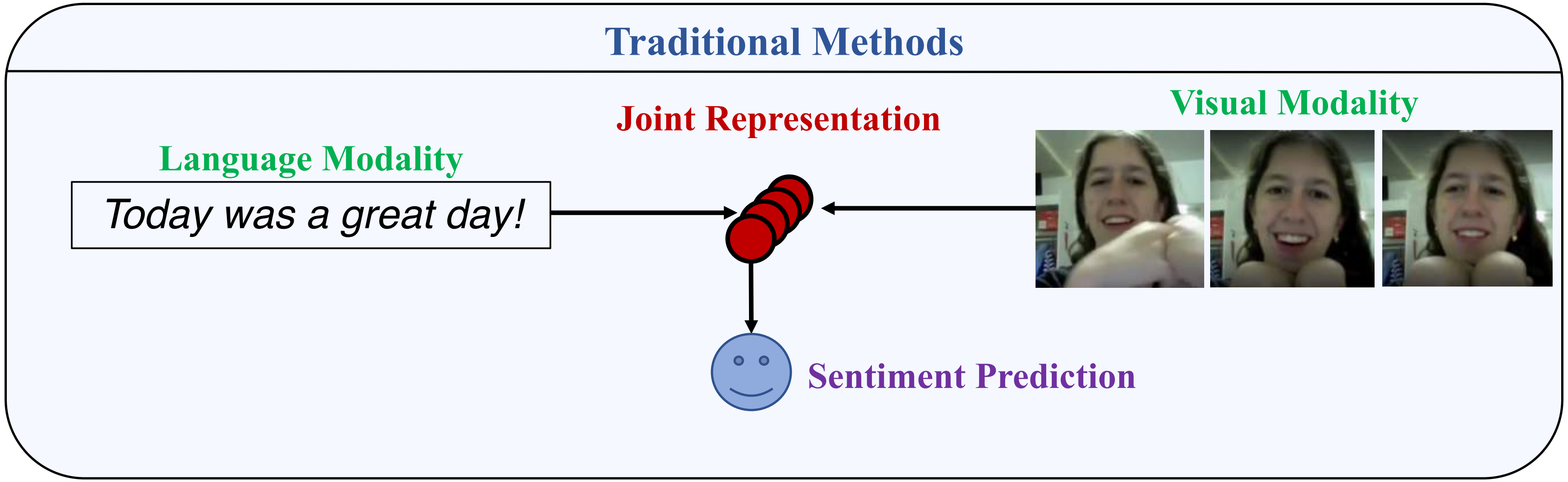
Learning Joint Representations: 2 modalities



Learning Joint Representations: 2 modalities

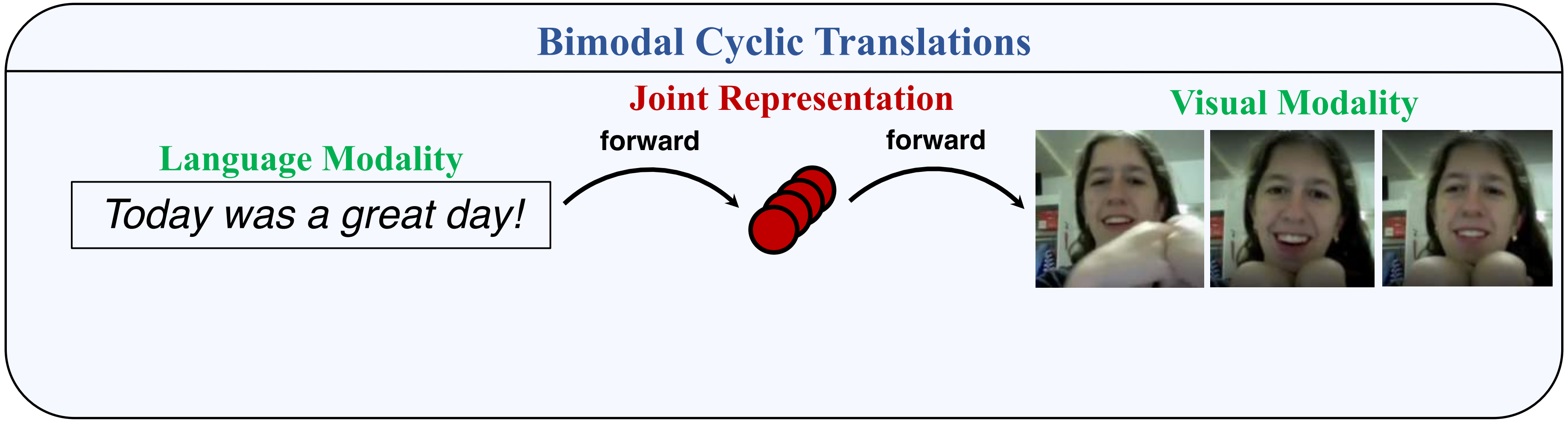


Learning Joint Representations: 2 modalities

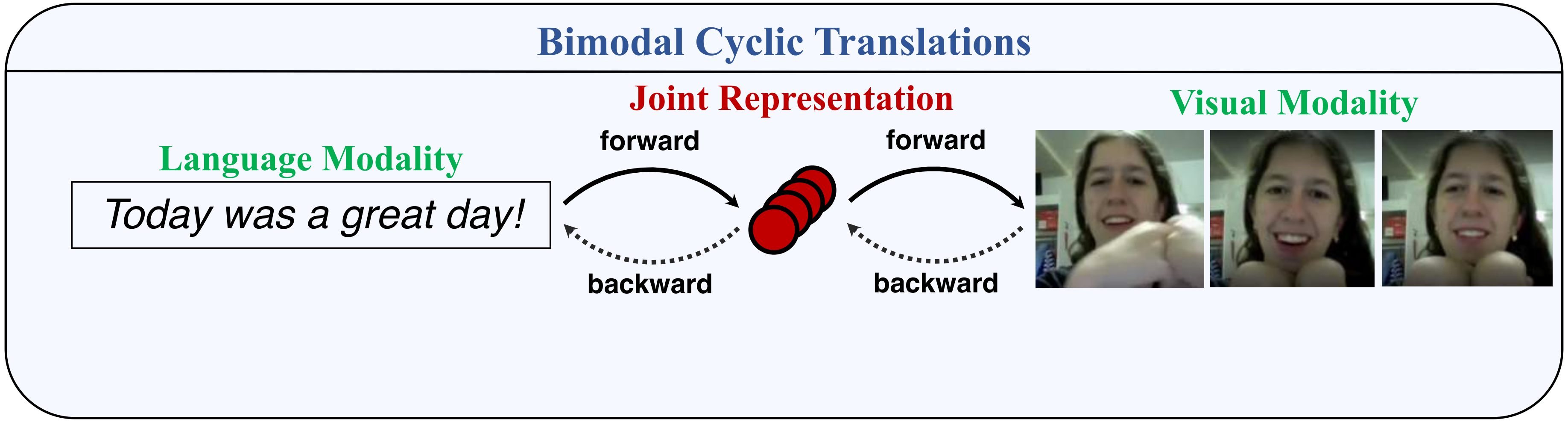


Both modalities required at test time!
Sensitive to missing/noisy visual modality.

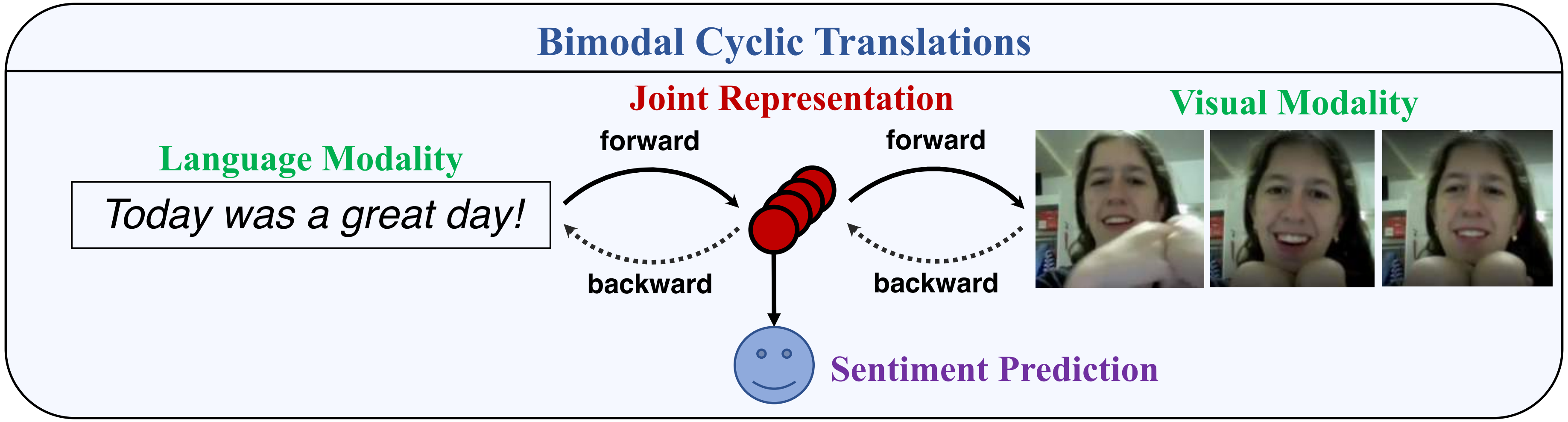
Learning Robust Joint Representations: 2 modalities



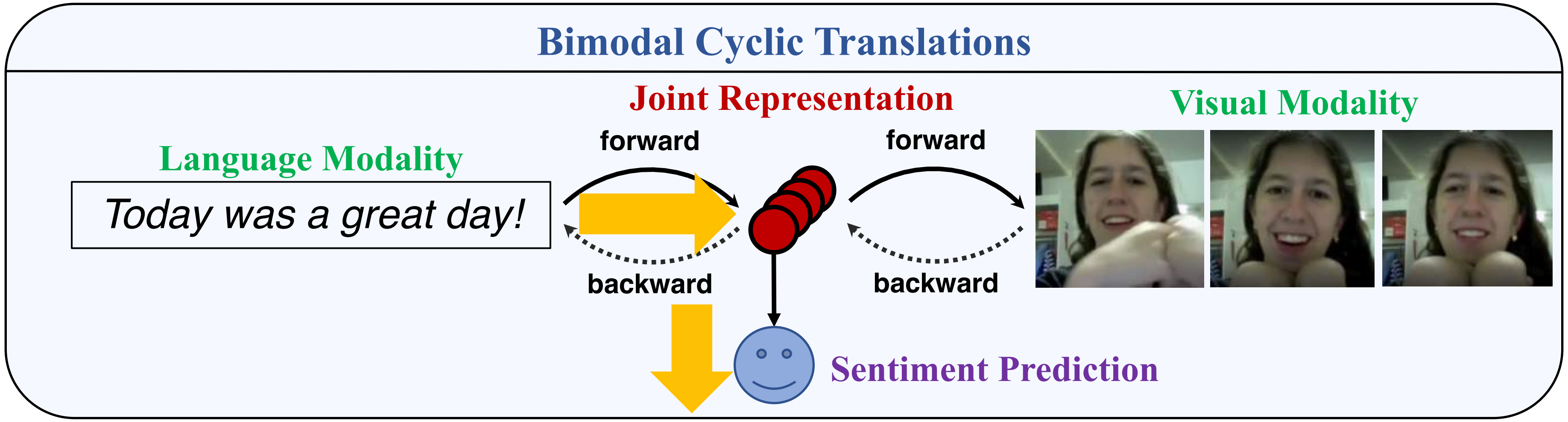
Learning Robust Joint Representations: 2 modalities



Learning Robust Joint Representations: 2 modalities



Learning Robust Joint Representations: 2 modalities



Only language modality required at test time!

Learning Robust Joint Representations: 3 modalities

Trimodal Cyclic Translations

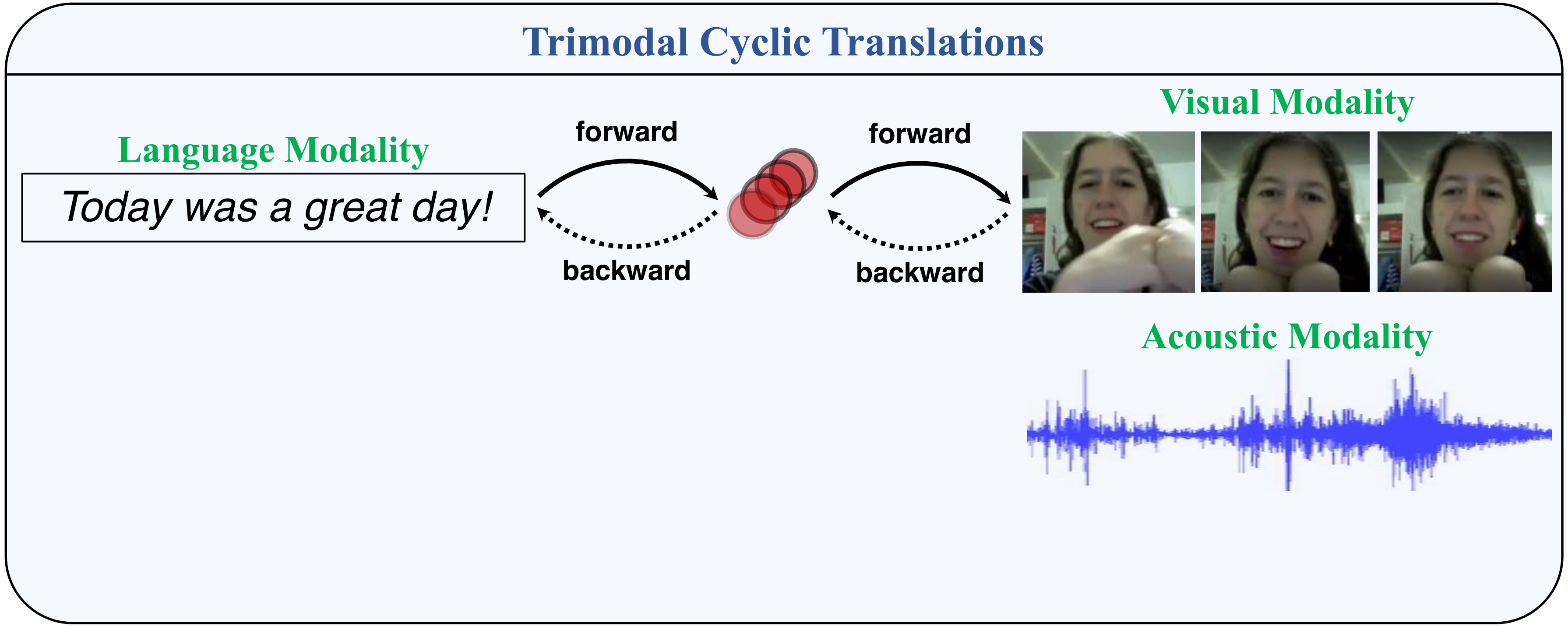
Language Modality

Today was a great day!

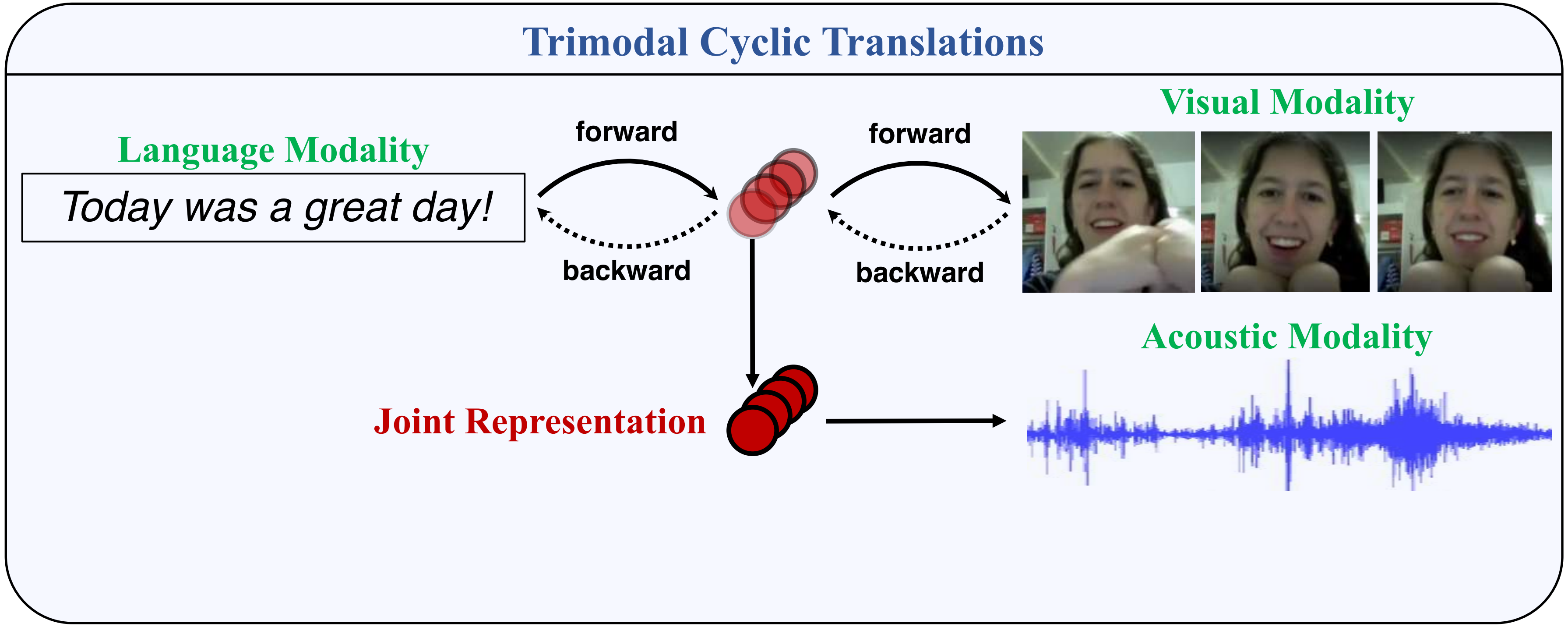
Visual Modality

Acoustic Modality

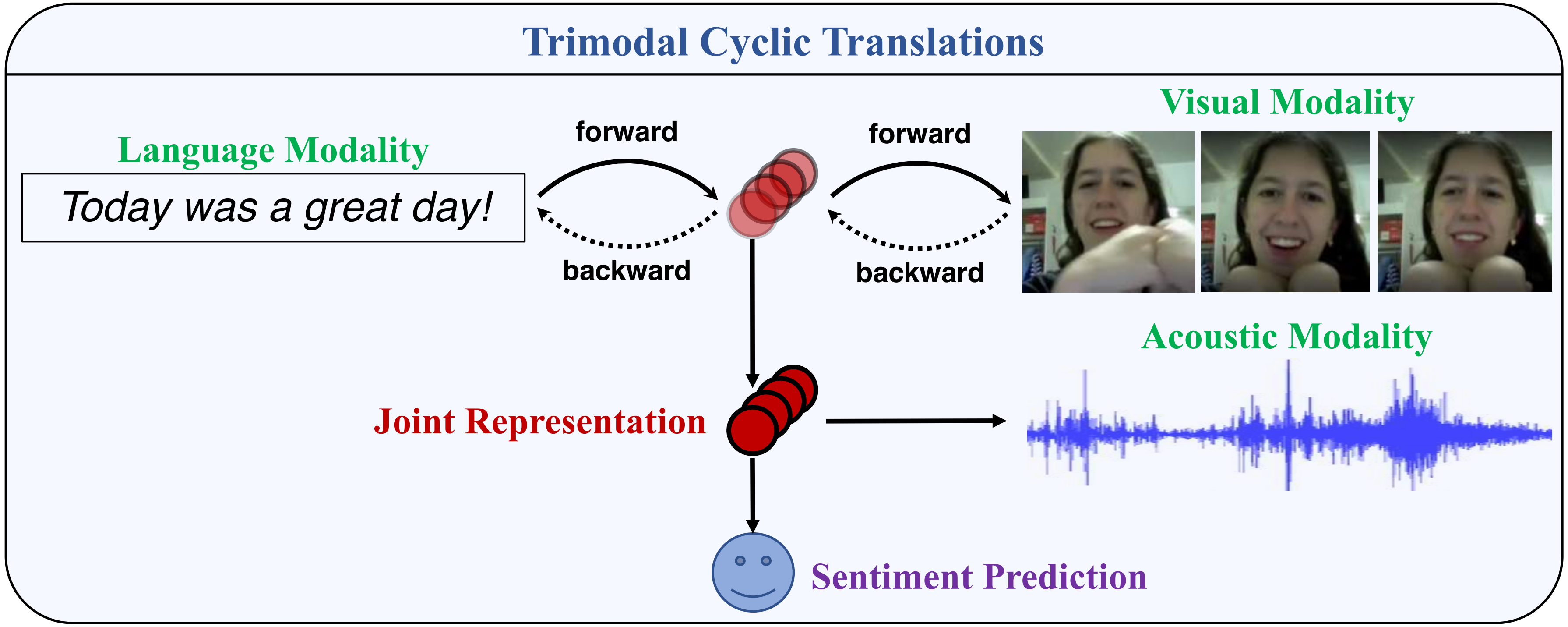
Learning Robust Joint Representations: 3 modalities



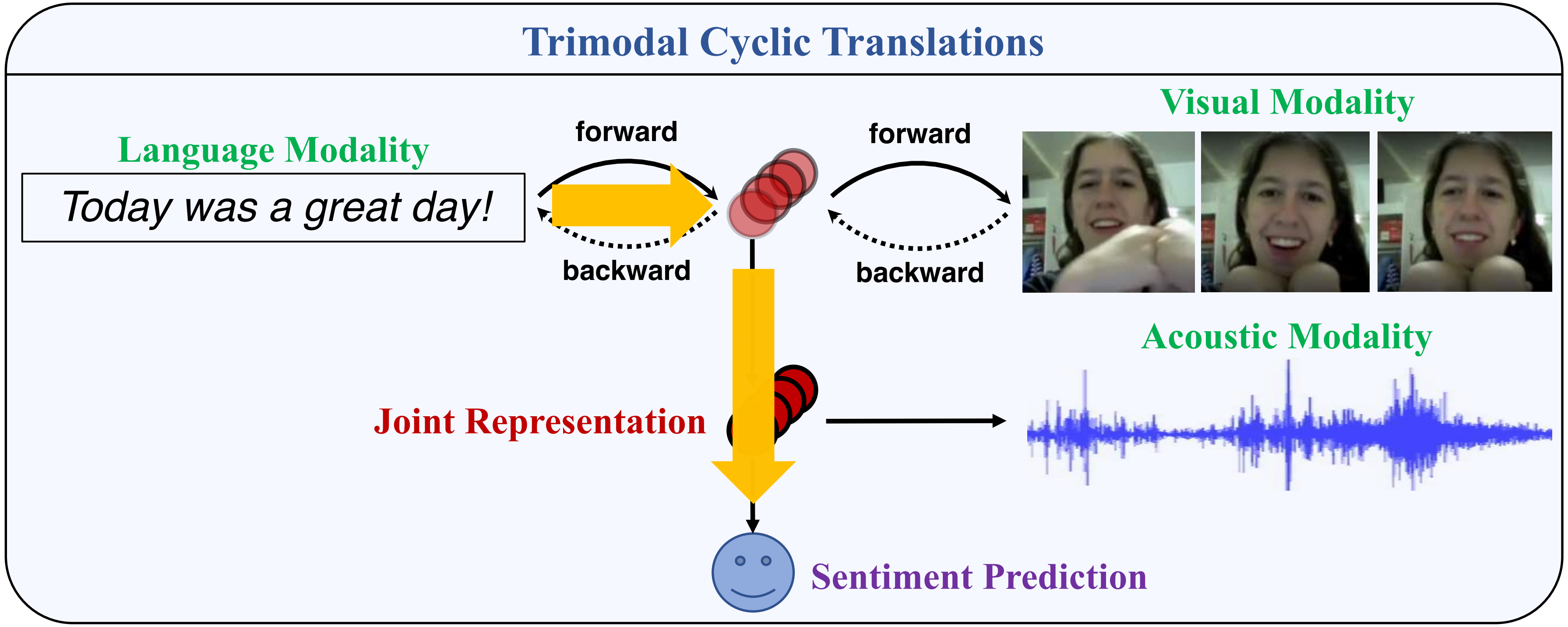
Learning Robust Joint Representations: 3 modalities



Learning Robust Joint Representations: 3 modalities



Learning Robust Joint Representations: 3 modalities



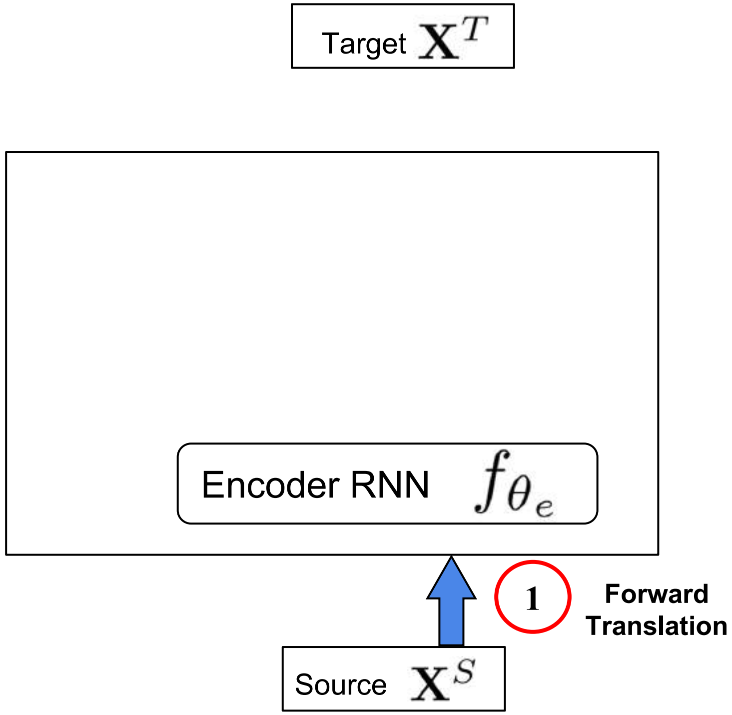
Only language modality required at test time!

Multimodal Cyclic Translation Network

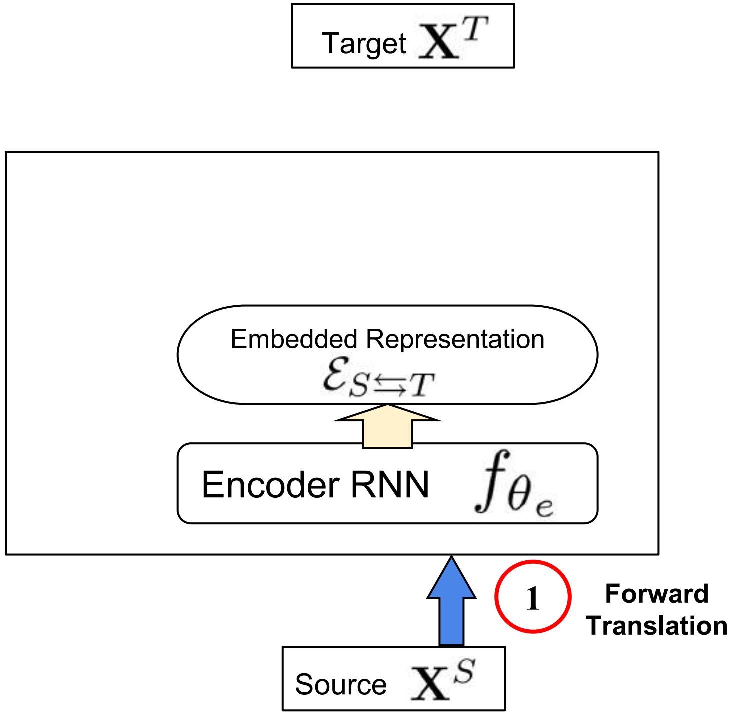
Target $\mathbf{X}^T$

Source $\mathbf{X}^S$

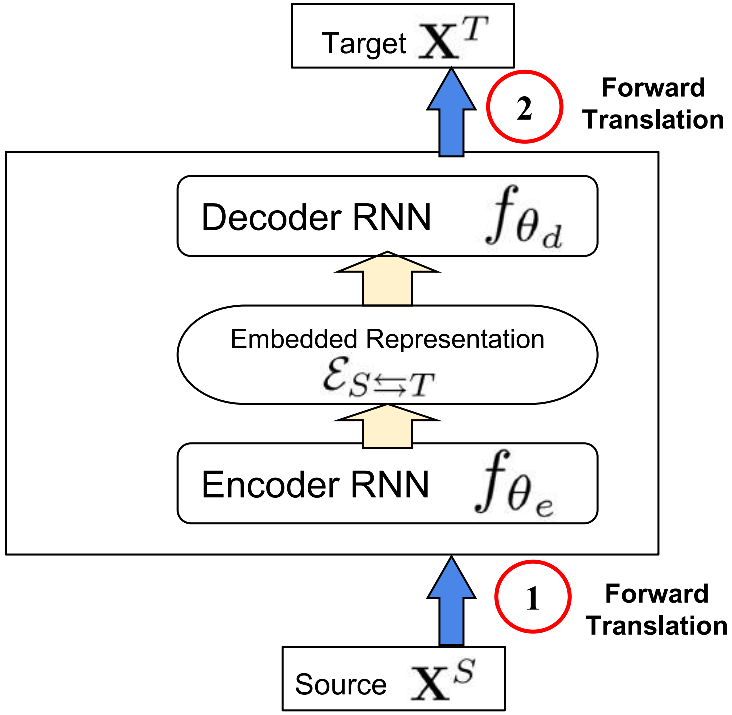
Multimodal Cyclic Translation Network



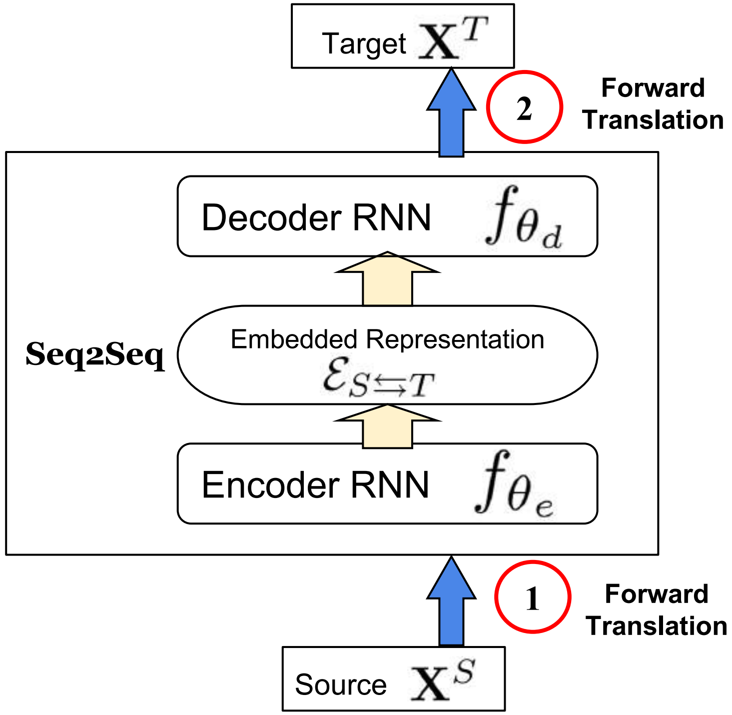
Multimodal Cyclic Translation Network



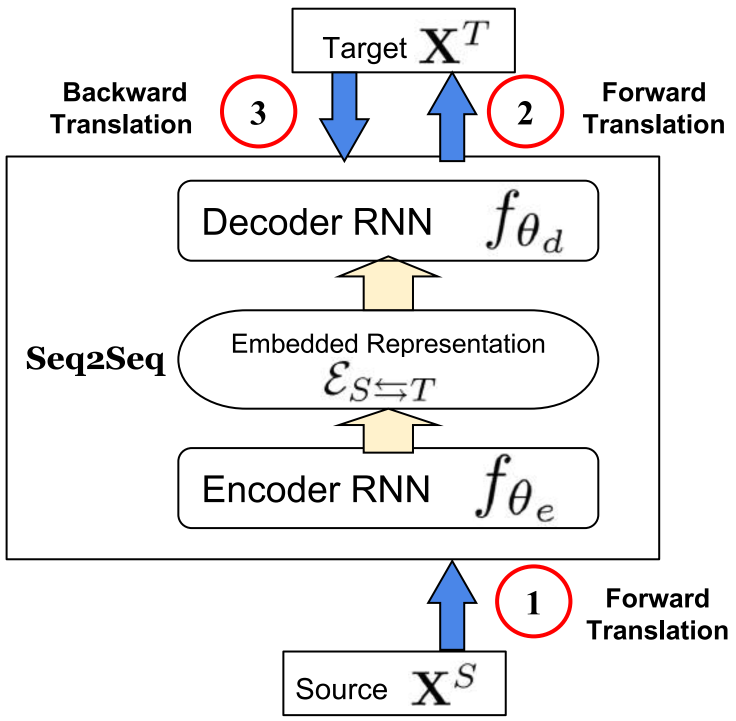
Multimodal Cyclic Translation Network



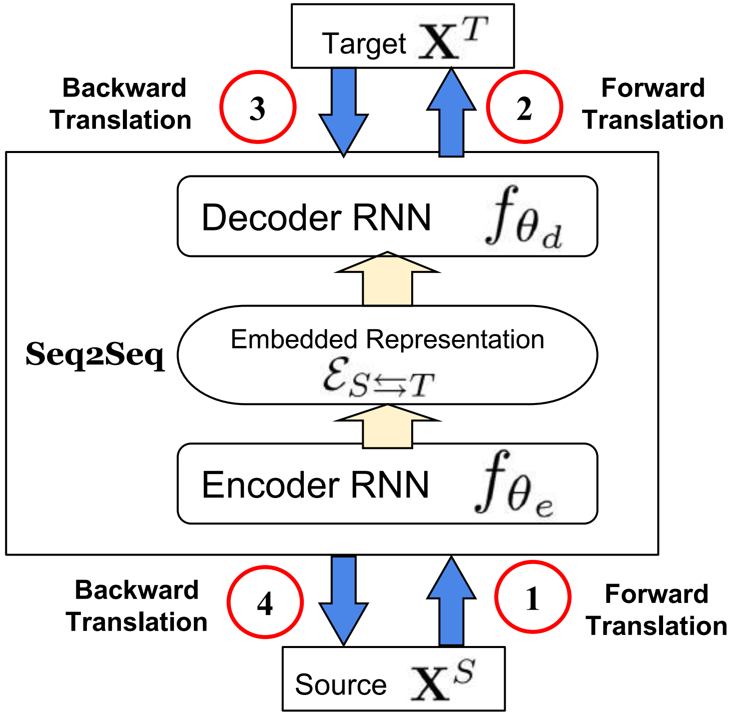
Multimodal Cyclic Translation Network



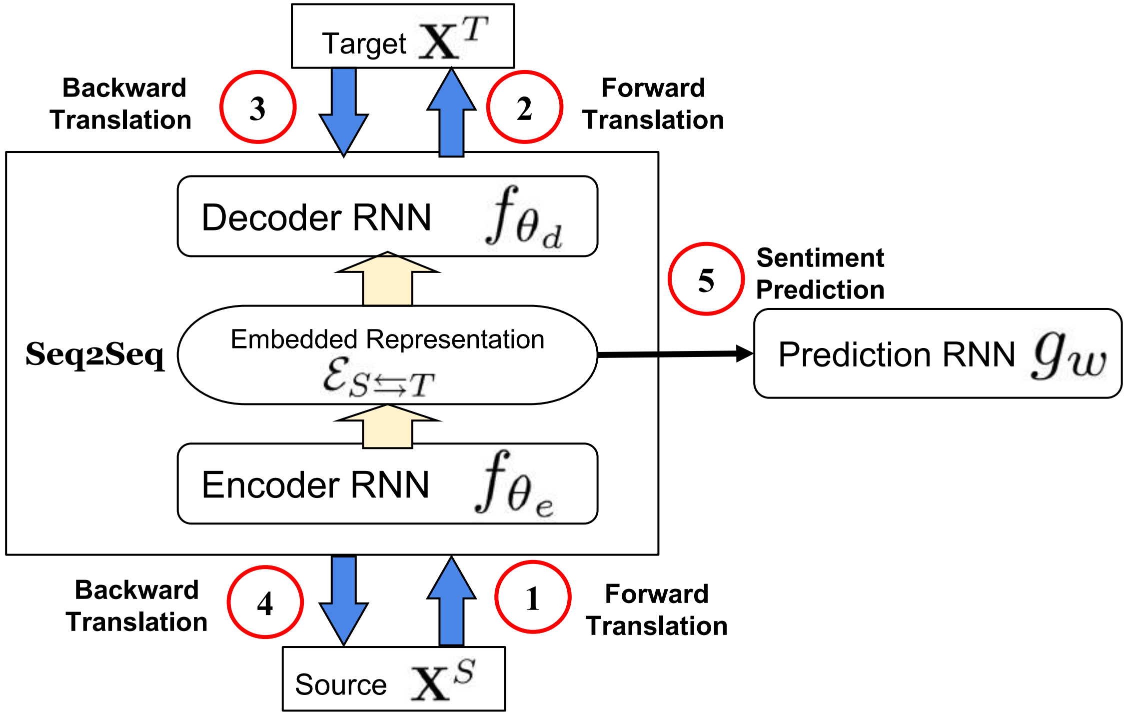
Multimodal Cyclic Translation Network



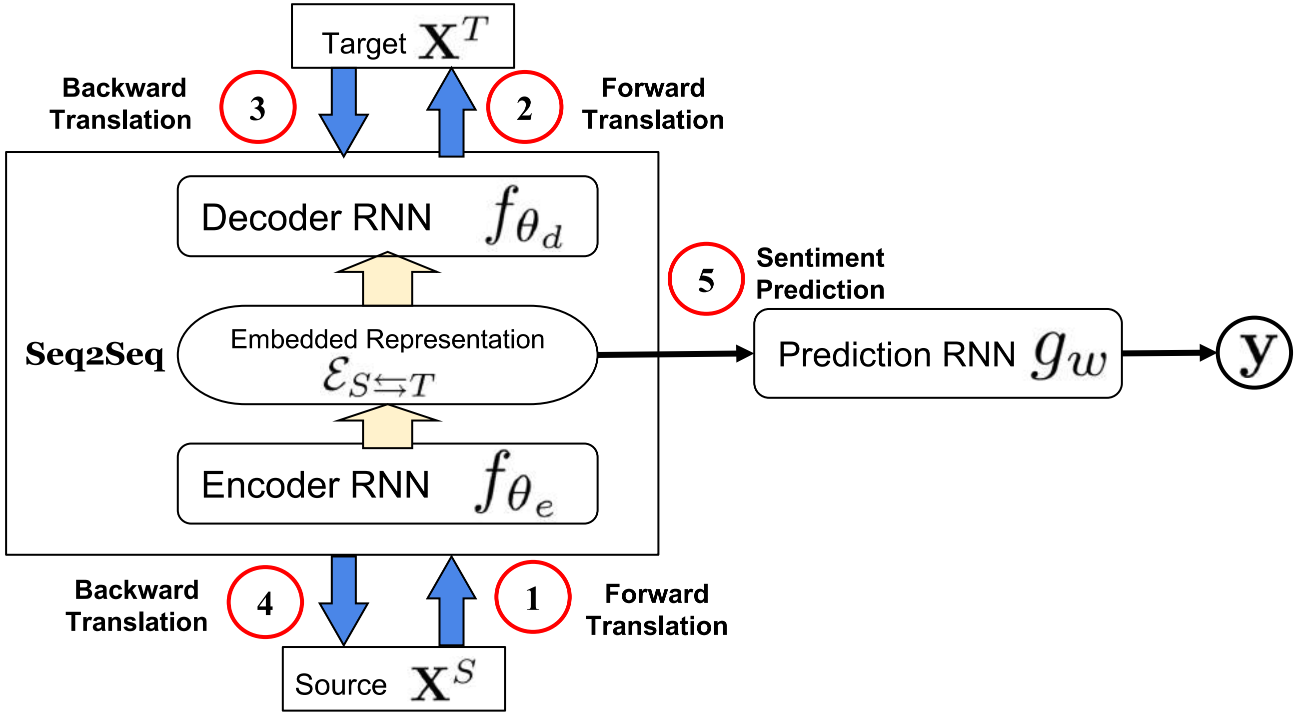
Multimodal Cyclic Translation Network



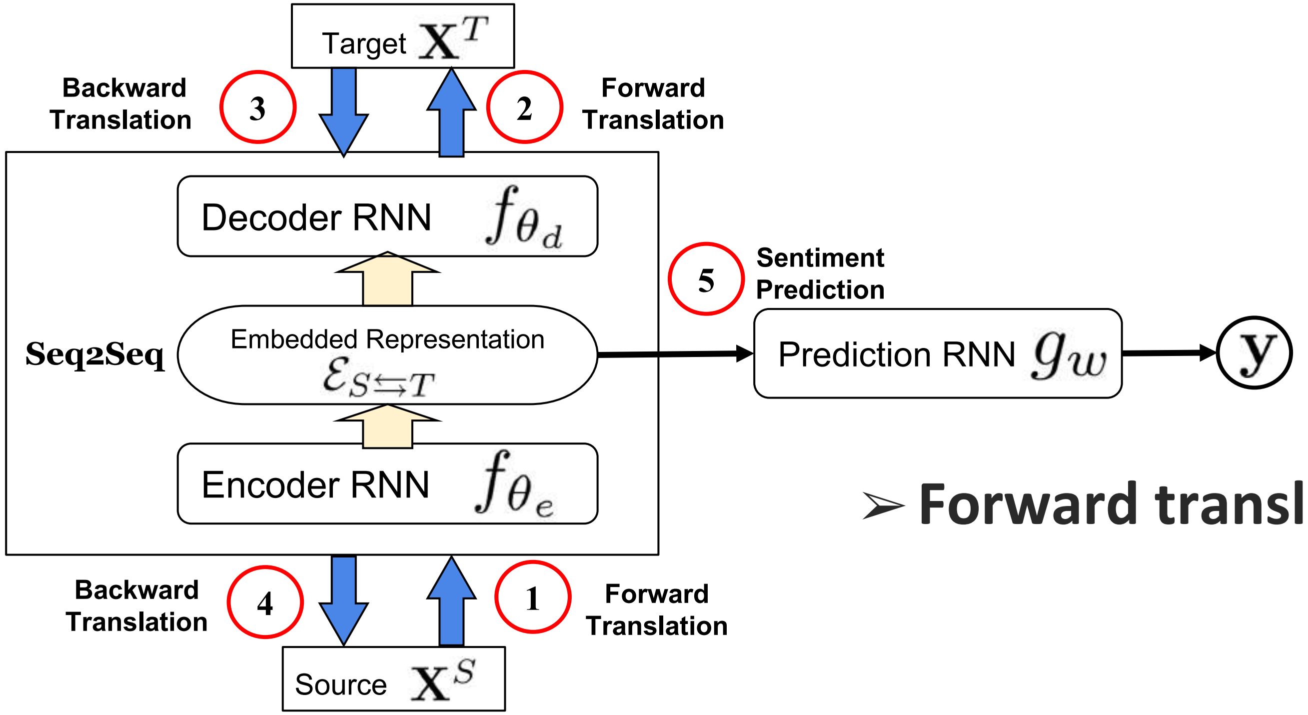
Multimodal Cyclic Translation Network



Multimodal Cyclic Translation Network

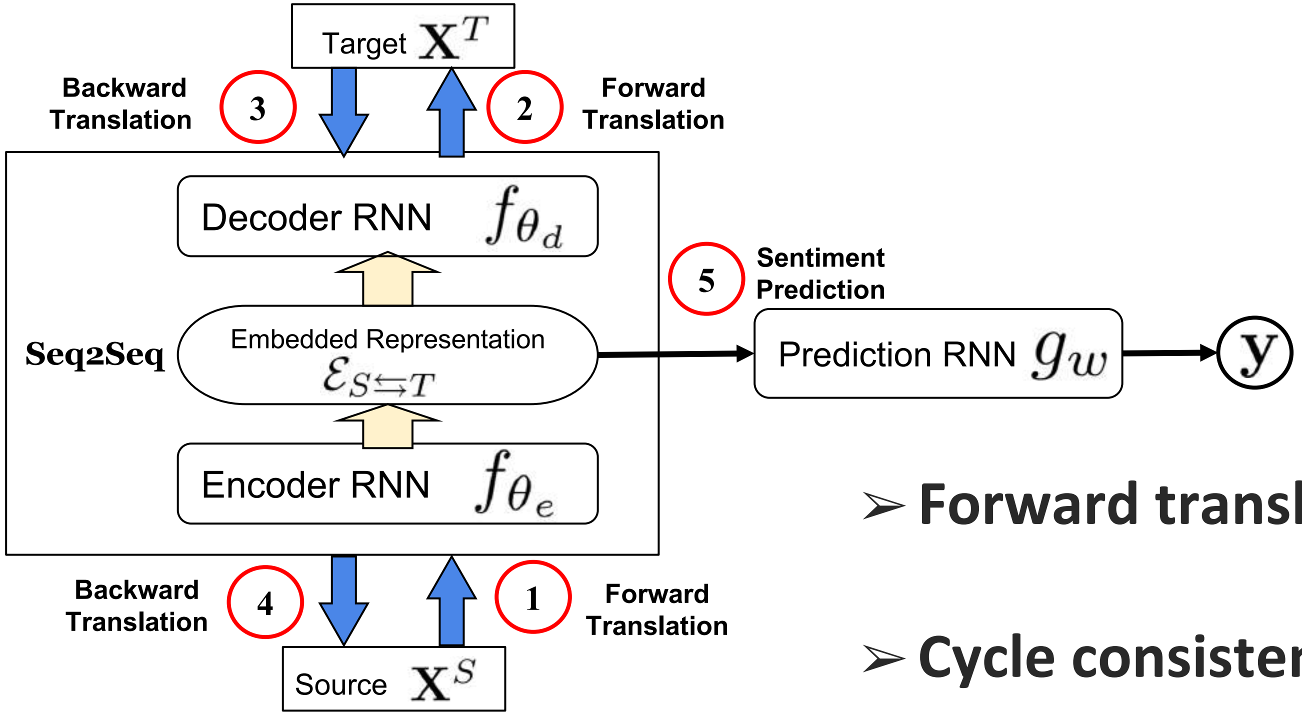


Coupled Translation-Prediction Objective



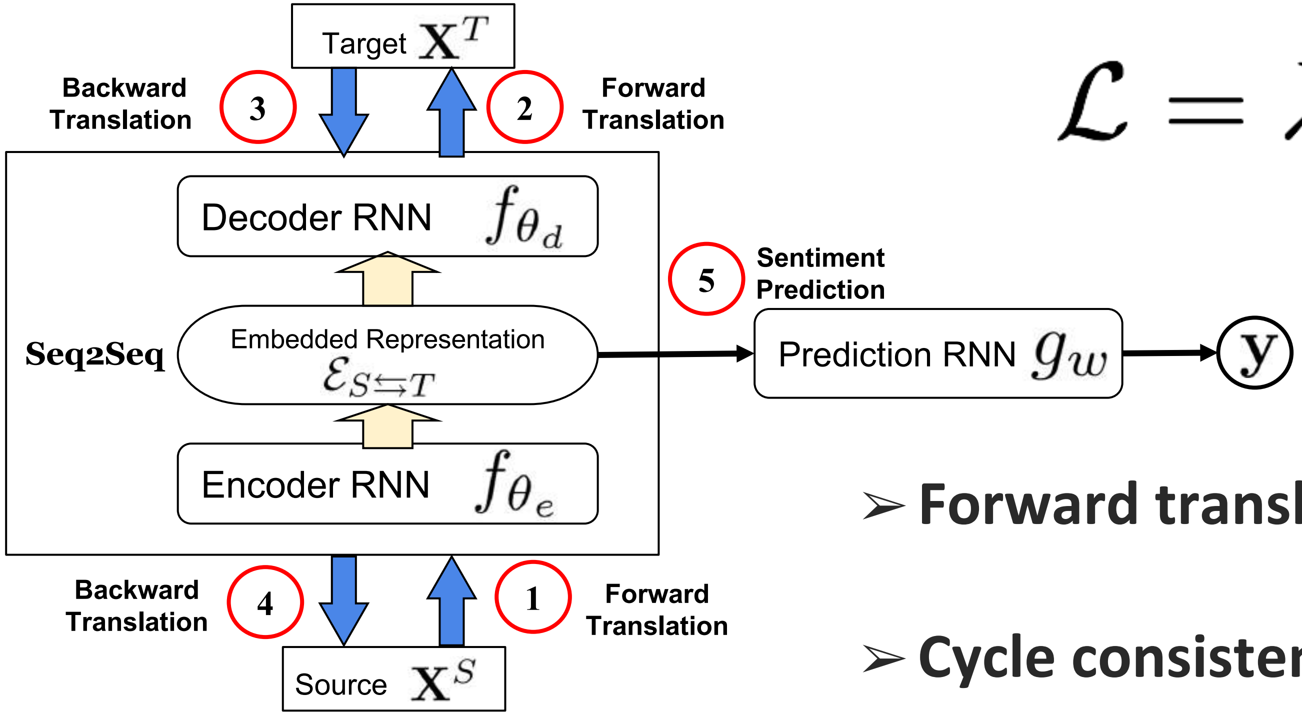
➢ **Forward translation loss** $\mathcal{L}_t = \mathbb{E}[\ell_{\mathbf{X}^T}(\hat{\mathbf{X}}^T, \mathbf{X}^T)]$

Coupled Translation-Prediction Objective



- **Forward translation loss** $\mathcal{L}_t = \mathbb{E}[\ell_{\mathbf{X}^T}(\hat{\mathbf{X}}^T, \mathbf{X}^T)]$
- **Cycle consistent loss** $\mathcal{L}_c = \mathbb{E}[\ell_{\mathbf{X}^S}(\hat{\mathbf{X}}^S, \mathbf{X}^S)]$

Coupled Translation-Prediction Objective



$$\mathcal{L} = \lambda_t \mathcal{L}_t + \lambda_c \mathcal{L}_c + \mathcal{L}_p$$

- **Forward translation loss** $\mathcal{L}_t = \mathbb{E}[\ell_{\mathbf{X}^T}(\hat{\mathbf{X}}^T, \mathbf{X}^T)]$
- **Cycle consistent loss** $\mathcal{L}_c = \mathbb{E}[\ell_{\mathbf{X}^S}(\hat{\mathbf{X}}^S, \mathbf{X}^S)]$
- **Prediction loss** $\mathcal{L}_p = \mathbb{E}[\ell_{\mathbf{y}}(\hat{\mathbf{y}}, \mathbf{y})]$

Hierarchical Multimodal Cyclic Translation Network

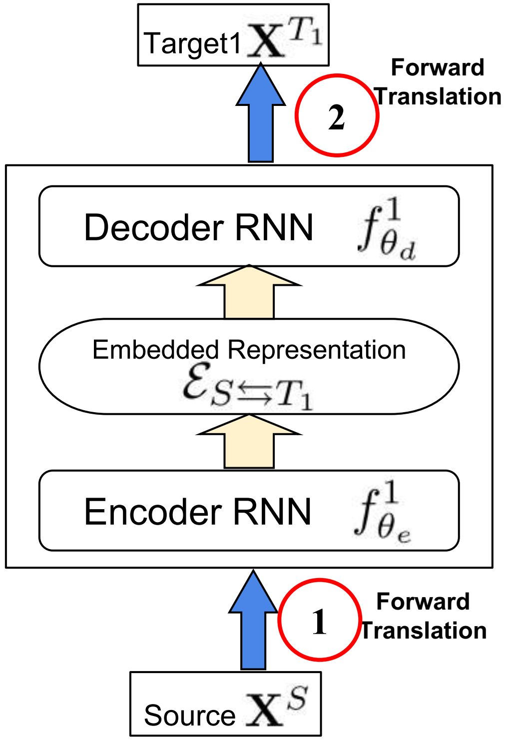
Target2 $\mathbf{X}^{T_2}$

Target1

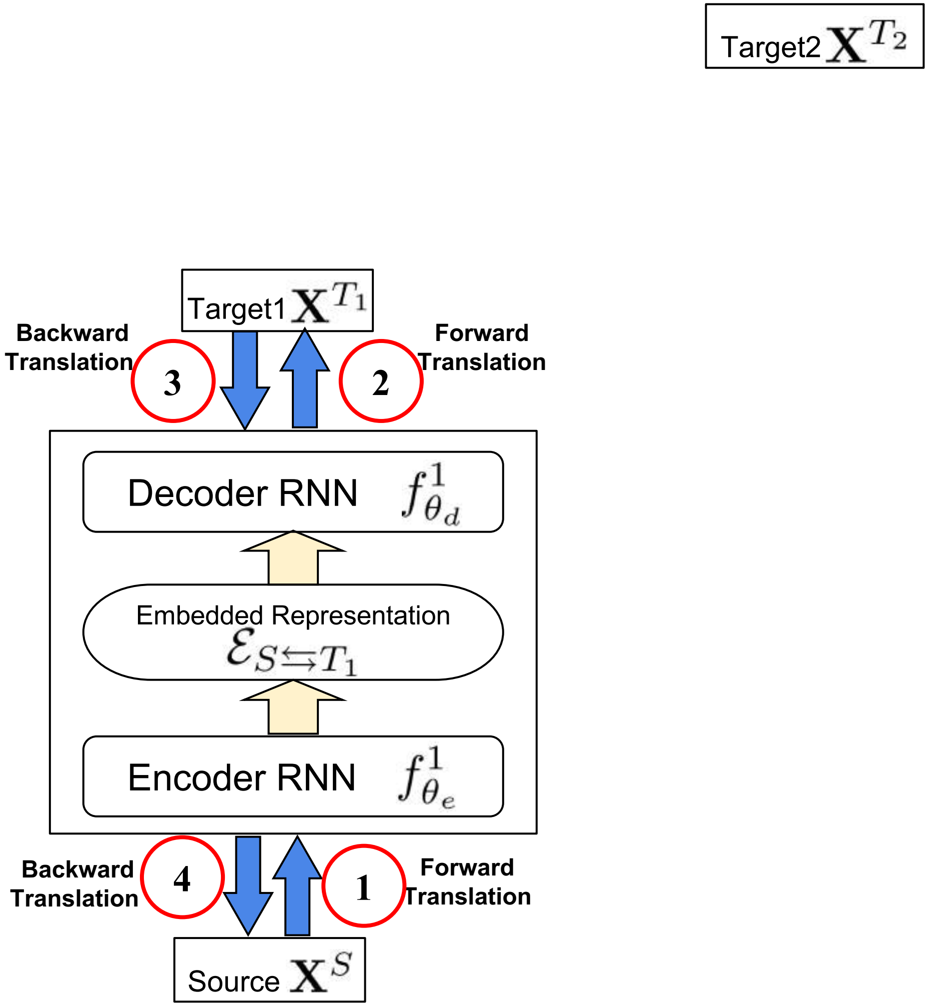
Source $\mathbf{X}^{S}$

Hierarchical Multimodal Cyclic Translation Network

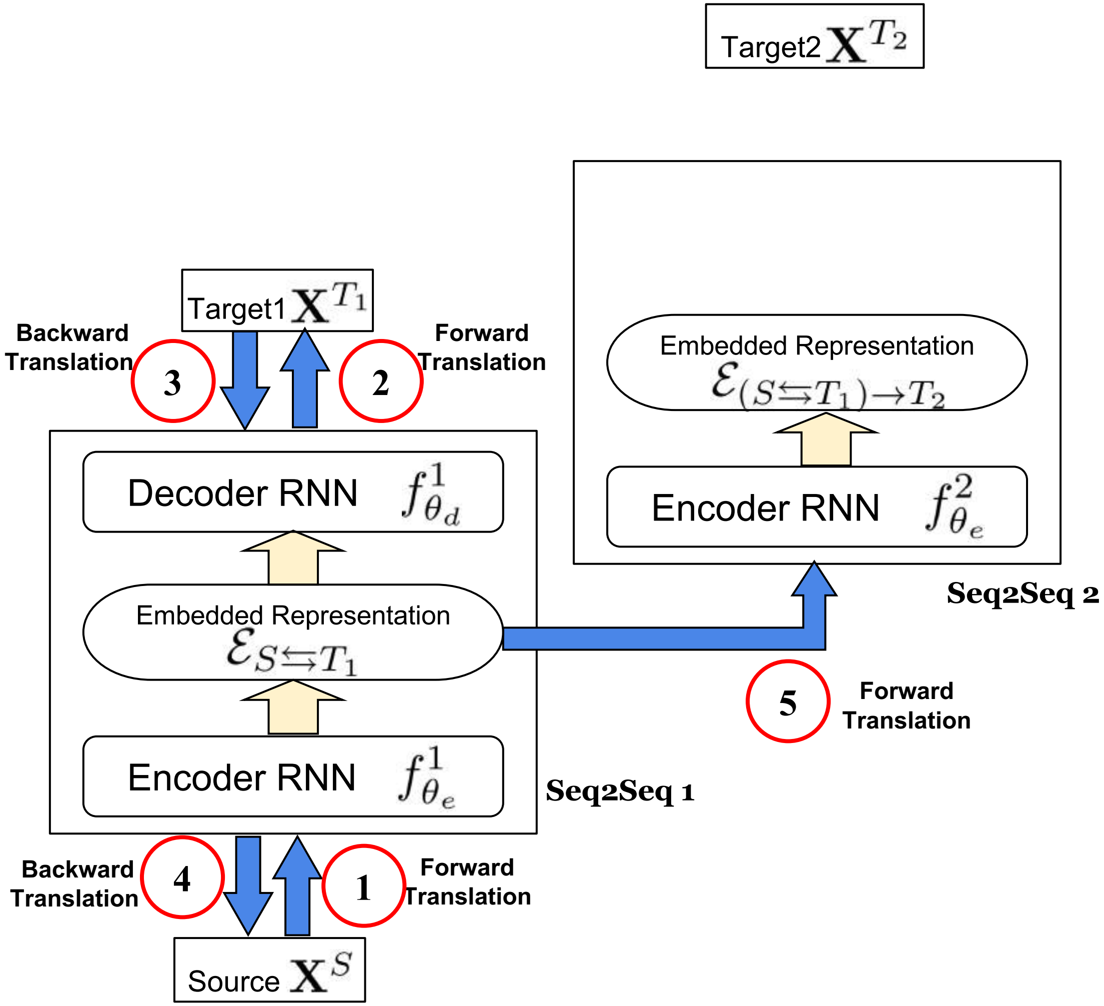
Target2 $\mathbf{X}^{T_2}$



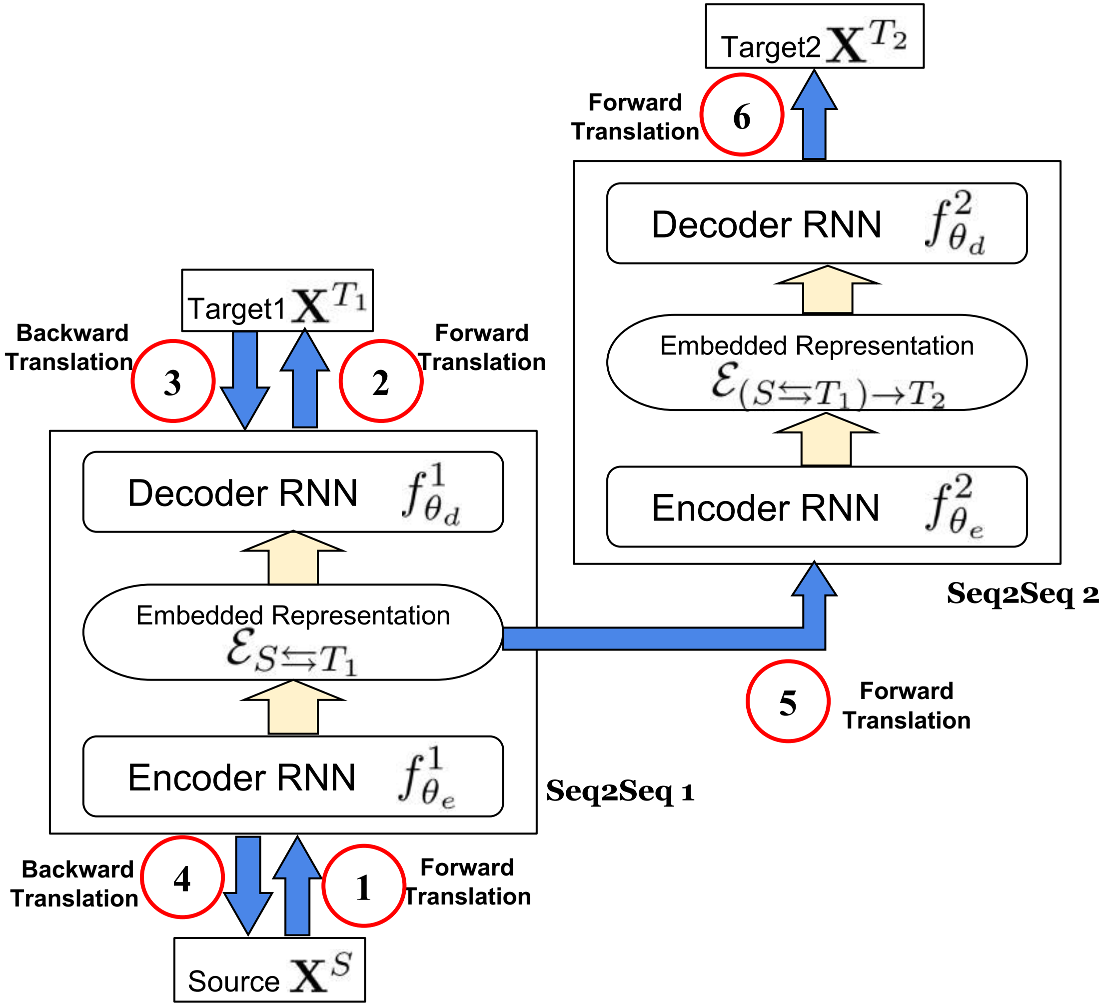
Hierarchical Multimodal Cyclic Translation Network



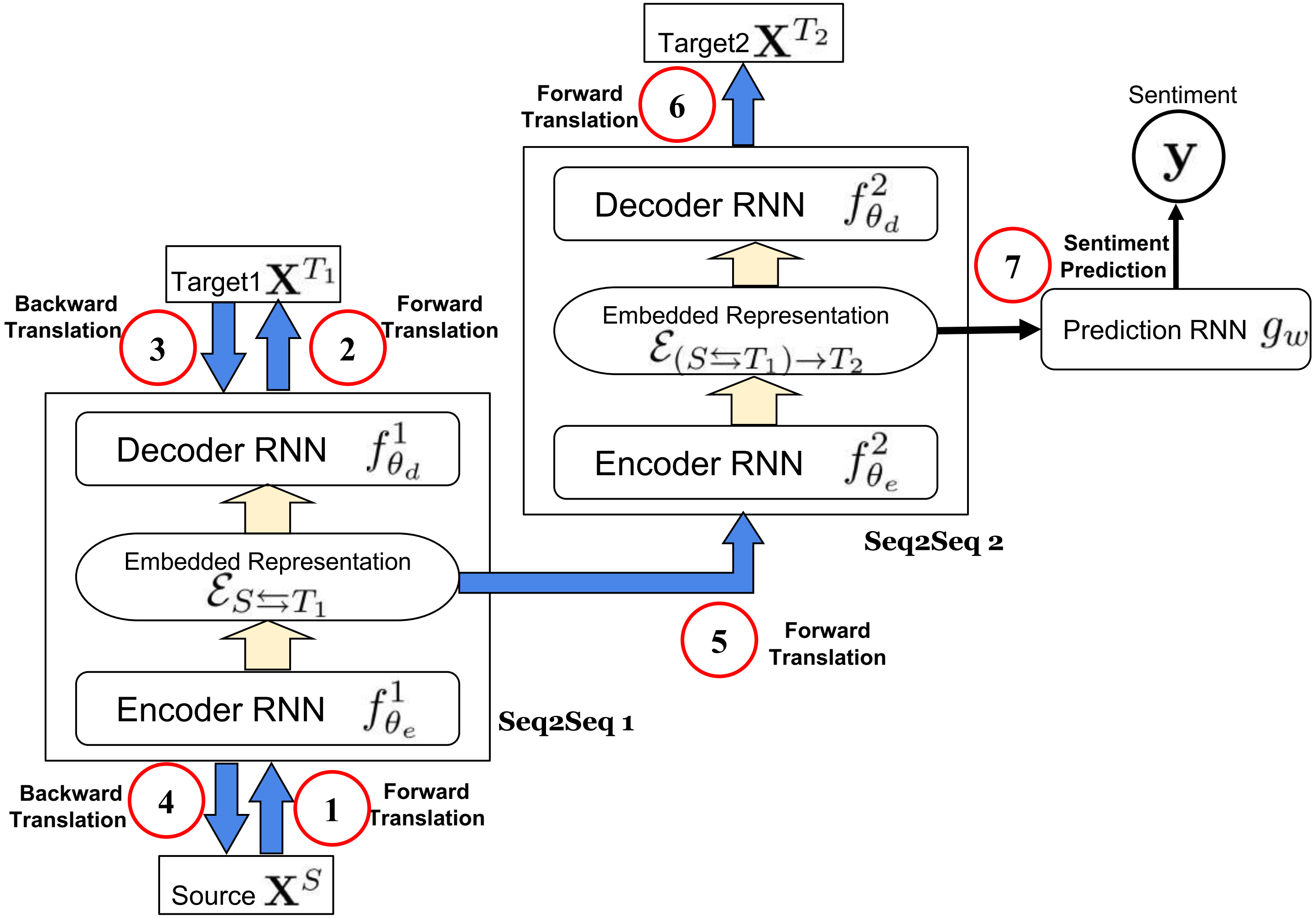
Hierarchical Multimodal Cyclic Translation Network



Hierarchical Multimodal Cyclic Translation Network



Hierarchical Multimodal Cyclic Translation Network



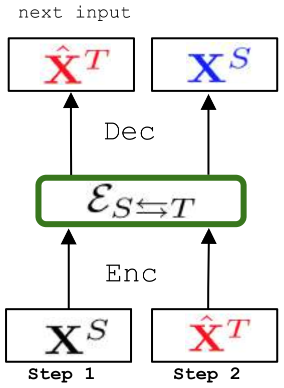
State-of-the-art Results: CMU-MOSI

Dataset		CMU-MOSI			
Model	Test Inputs	Acc(↑)	F1(↑)	MAE(↓)	Corr(↑)
RF	$\{\ell, v, a\}$	56.4	56.3	-	-
SVM	$\{\ell, v, a\}$	71.6	72.3	1.100	0.559
THMM	$\{\ell, v, a\}$	50.7	45.4	-	-
EF-HCRF	$\{\ell, v, a\}$	65.3	65.4	-	-
MV-HCRF	$\{\ell, v, a\}$	65.6	65.7	-	-
DF	$\{\ell, v, a\}$	74.2	74.2	1.143	0.518
EF-LSTM	$\{\ell, v, a\}$	74.3	74.3	1.023	0.622
MV-LSTM	$\{\ell, v, a\}$	73.9	74.0	1.019	0.601
BC-LSTM	$\{\ell, v, a\}$	75.2	75.3	1.079	0.614
TFN	$\{\ell, v, a\}$	74.6	74.5	1.040	0.587
GME-LSTM(A)	$\{\ell, v, a\}$	76.5	73.4	0.955	-
MARN	$\{\ell, v, a\}$	77.1	77.0	0.968	0.625
MFN	$\{\ell, v, a\}$	77.4	77.3	0.965	0.632
LMF	$\{\ell, v, a\}$	76.4	75.7	0.912	0.668
RMFN	$\{\ell, v, a\}$	78.4	78.0	0.922	**0.681**
MCTN	$\{\ell\}$	**79.3**	**79.1**	**0.909**	0.676

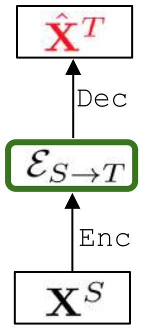
State-of-the-art Results: ICT-MMMO and YouTube

Dataset		ICT-MMMO		YouTube	
Model	Test Inputs	Acc(↑)	F1(↑)	Acc(↑)	F1(↑)
RF	$\{\ell, v, a\}$	70.0	69.8	33.3	32.3
SVM	$\{\ell, v, a\}$	68.8	68.7	42.4	37.9
THMM	$\{\ell, v, a\}$	53.8	53.0	42.4	27.9
EF-HCRF	$\{\ell, v, a\}$	73.8	73.1	45.8	45.0
MV-HCRF	$\{\ell, v, a\}$	68.8	67.1	44.1	44.0
DF	$\{\ell, v, a\}$	65.0	58.7	45.8	32.0
EF-LSTM	$\{\ell, v, a\}$	72.5	70.9	44.1	43.6
MV-LSTM	$\{\ell, v, a\}$	72.5	72.3	45.8	43.3
BC-LSTM	$\{\ell, v, a\}$	70.0	70.1	45.0	45.1
TFN	$\{\ell, v, a\}$	72.5	72.6	45.0	41.0
MARN	$\{\ell, v, a\}$	71.3	70.2	48.3	44.9
MFN	$\{\ell, v, a\}$	73.8	73.1	**51.7**	51.6
MCTN	$\{\ell\}$	**81.3**	**80.8**	**51.7**	**52.4**

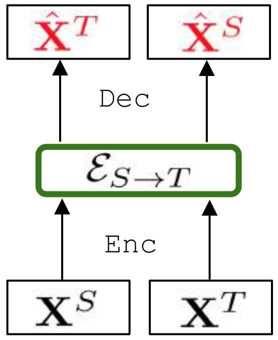
Bimodal Variations



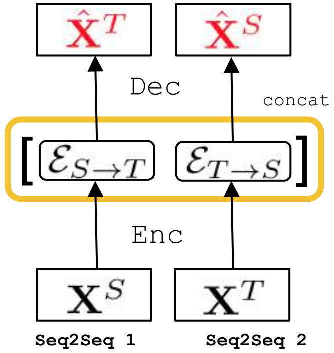
MCTN Bi



Simple Bi

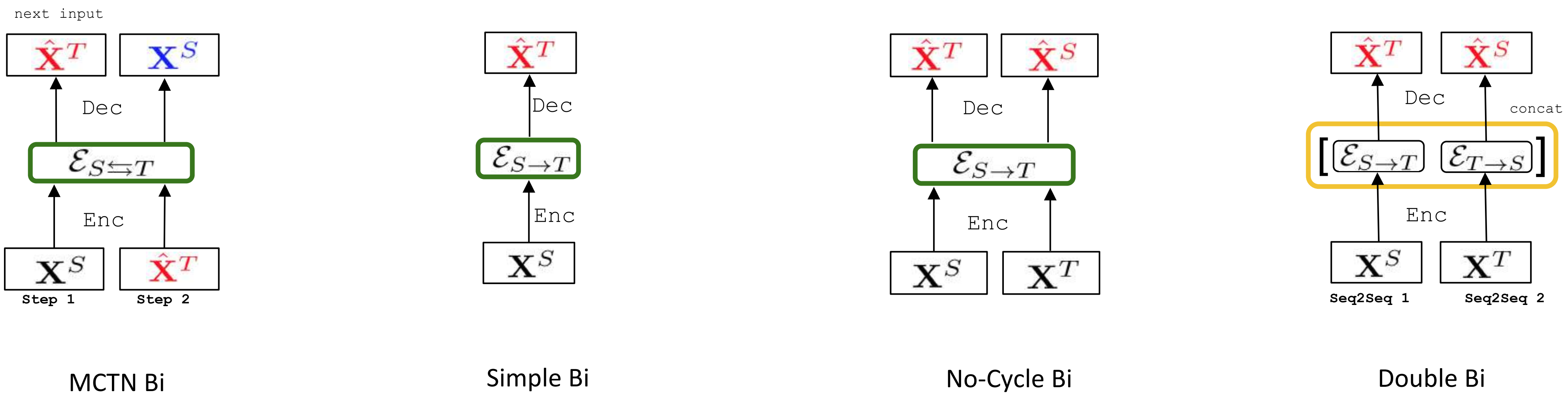


No-Cycle Bi



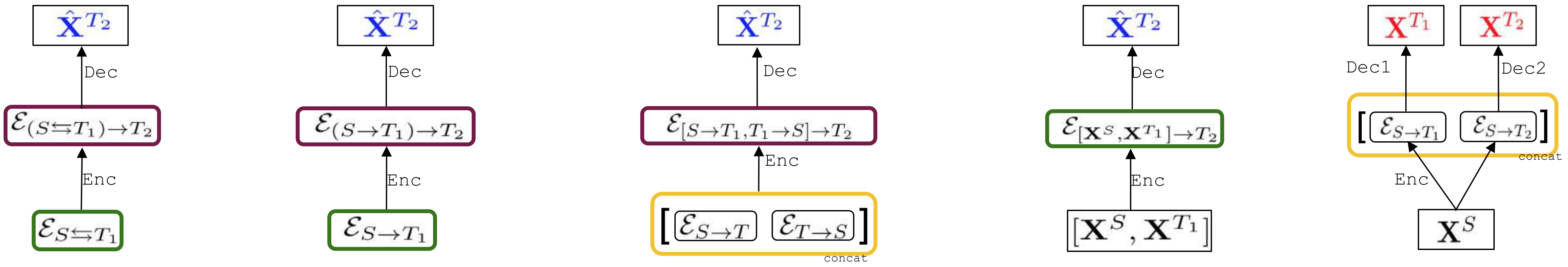
Double Bi

Bimodal Variations



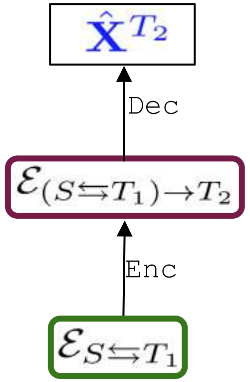
+ Cyclic translations
+ Parameter sharing

Trimodal Variations

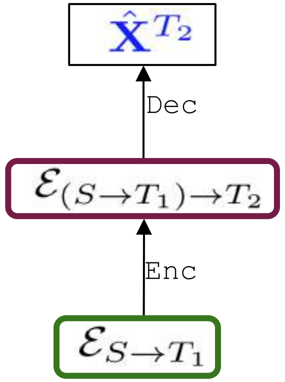


MCTN Tri | Simple Tri | Double Tri | Concat Tri | Paired Tri

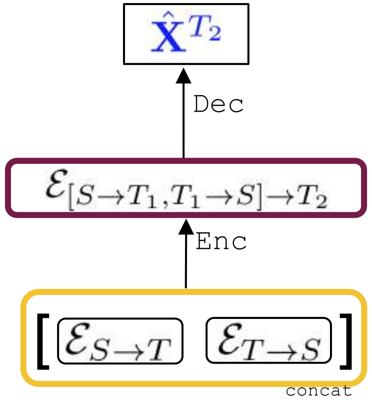
Trimodal Variations



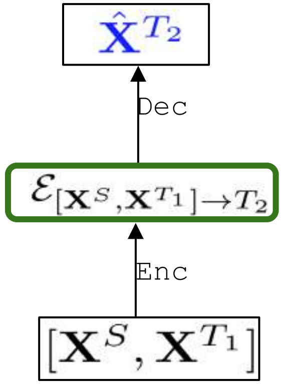
MCTN Tri



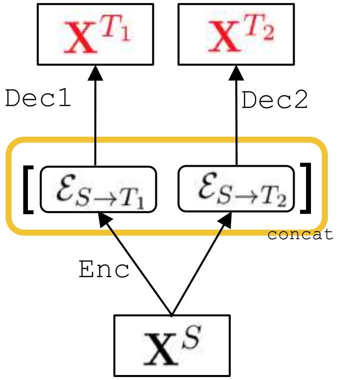
Simple Tri



Double Tri



Concat Tri



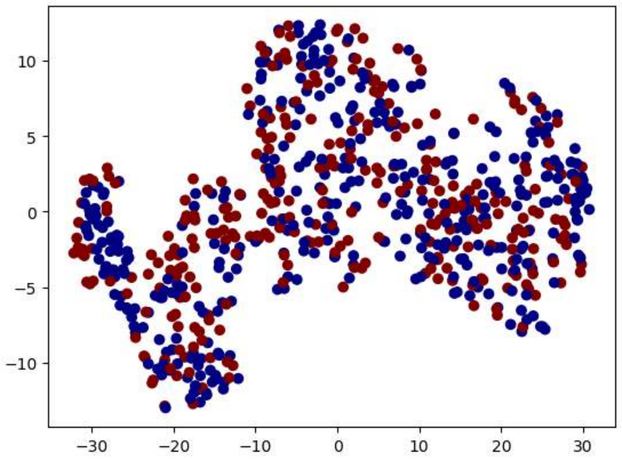
Paired Tri

+ Cyclic translations
+ Parameter sharing
+ Hierarchical structure

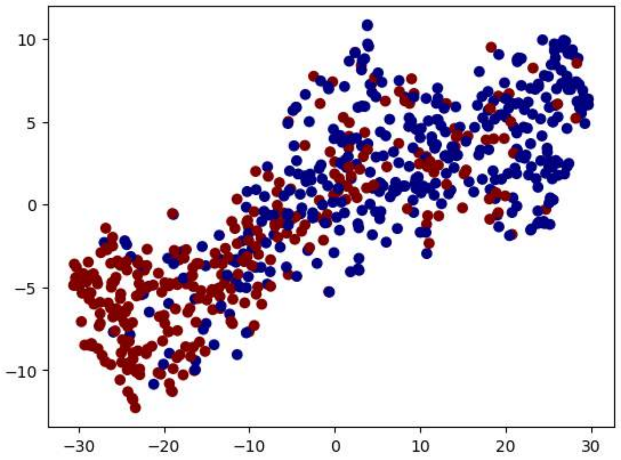
Adding More Modalities

Dataset	CMU-MOSI				
Model	Translation	Acc	F1	MAE	Corr
MCTN Bi (Fig. 4a)	$V \leftrightarrows A$	53.1	53.2	1.420	0.034
	$T \leftrightarrows A$	76.4	76.4	0.977	0.636
	$T \leftrightarrows V$	76.8	76.8	1.034	0.592
MCTN Tri (Fig. 4e)	$(V \leftrightarrows A) \rightarrow T$	56.4	56.3	1.455	0.151
	$(T \leftrightarrows A) \rightarrow V$	78.7	78.8	0.960	0.650
	$(T \leftrightarrows V) \rightarrow A$	**79.3**	**79.1**	**0.909**	**0.676**

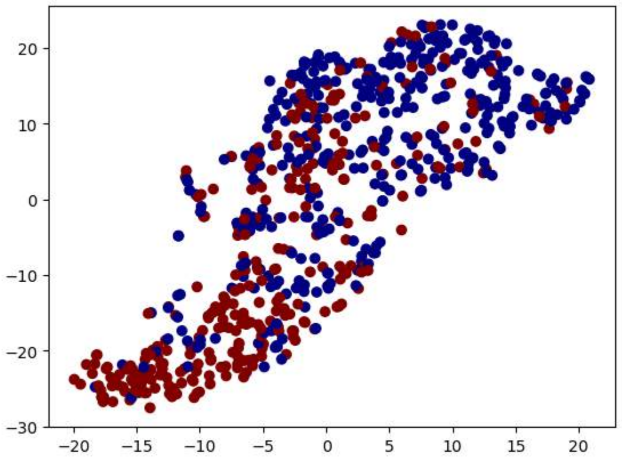
Adding More Modalities



Bimodal MCTN
without
cyclic translation



Bimodal MCTN
with
cyclic translation



Trimodal MCTN
with
cyclic translation

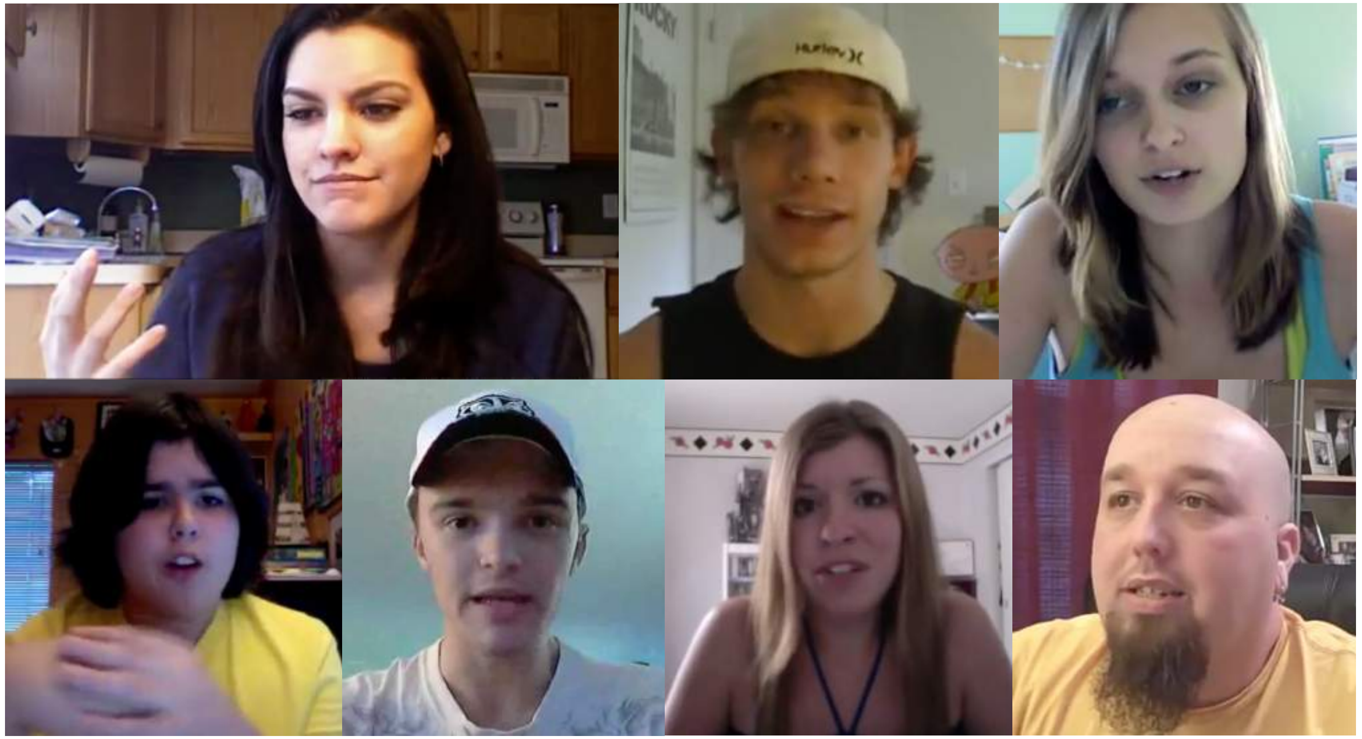
New Multimodal Dataset: MOSEI

New Dataset: MOSEI

23,000 video segments
3 modalities

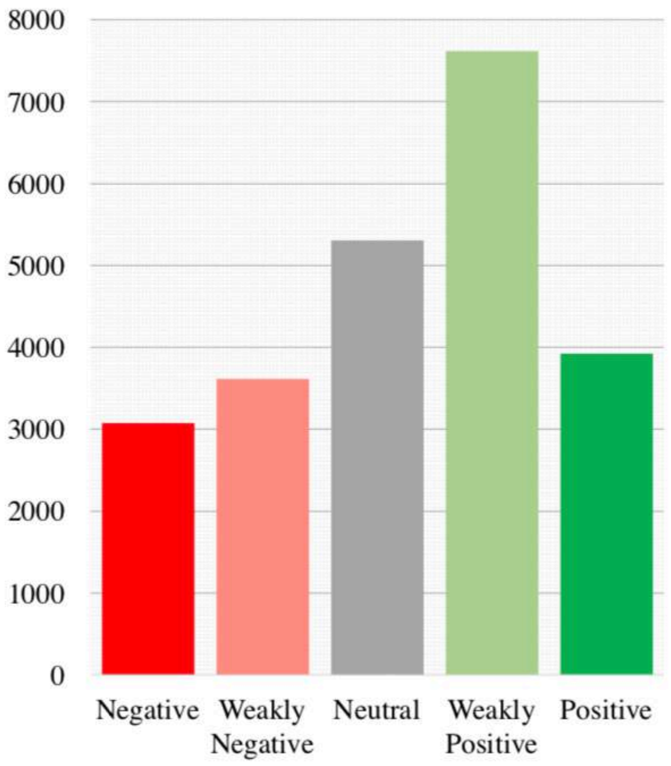
MOSEI Dataset

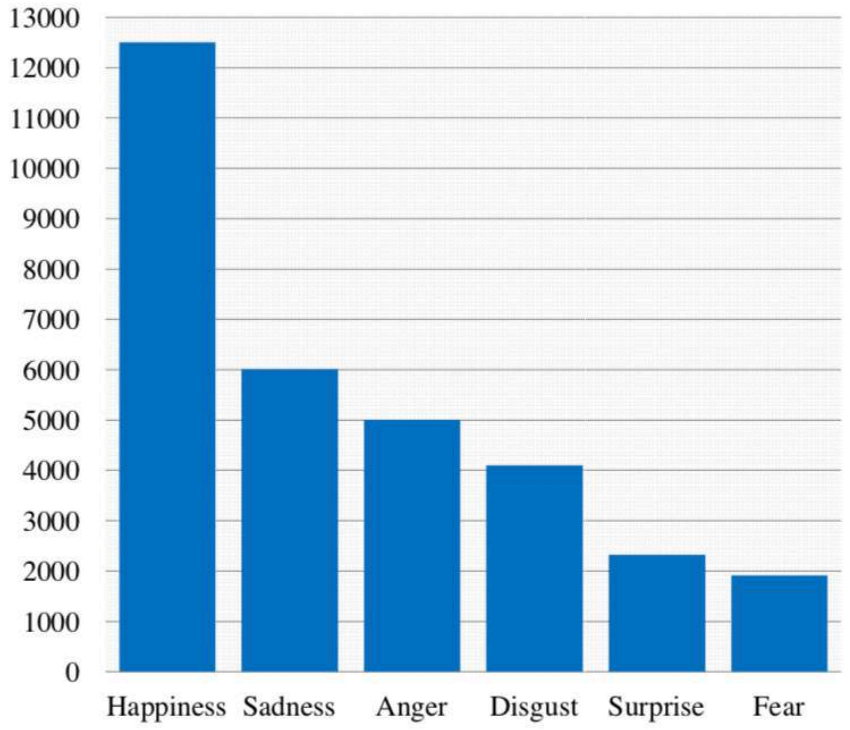
1,000 speakers



250 topics

Annotation Distributions



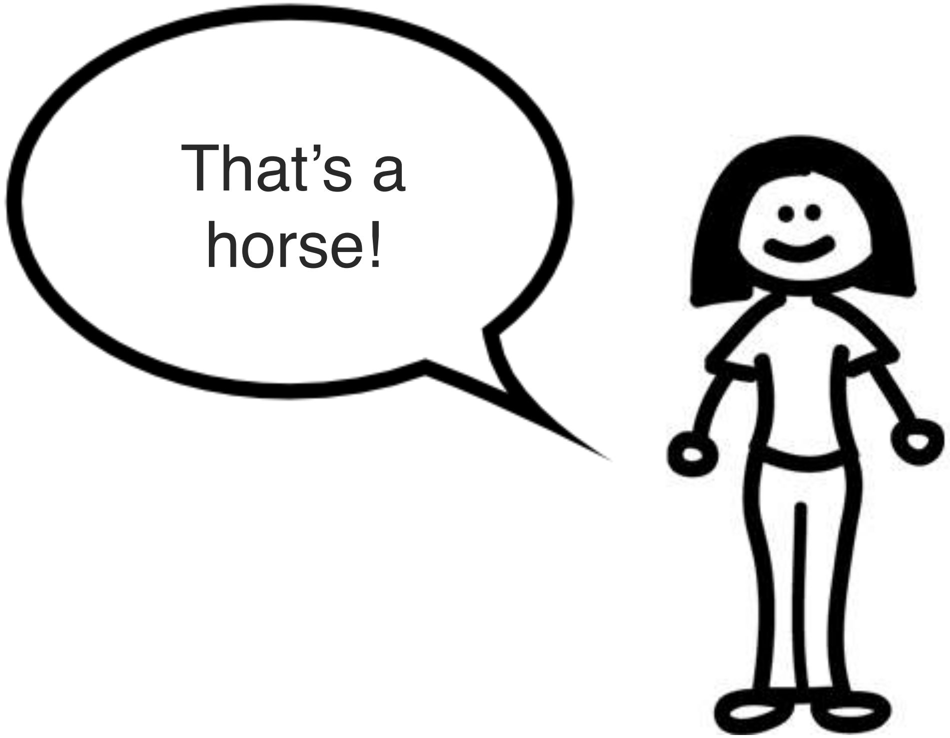


Future Directions

- Learning from limited/missing multimodal data

Future Directions

- Learning from limited/missing multimodal data

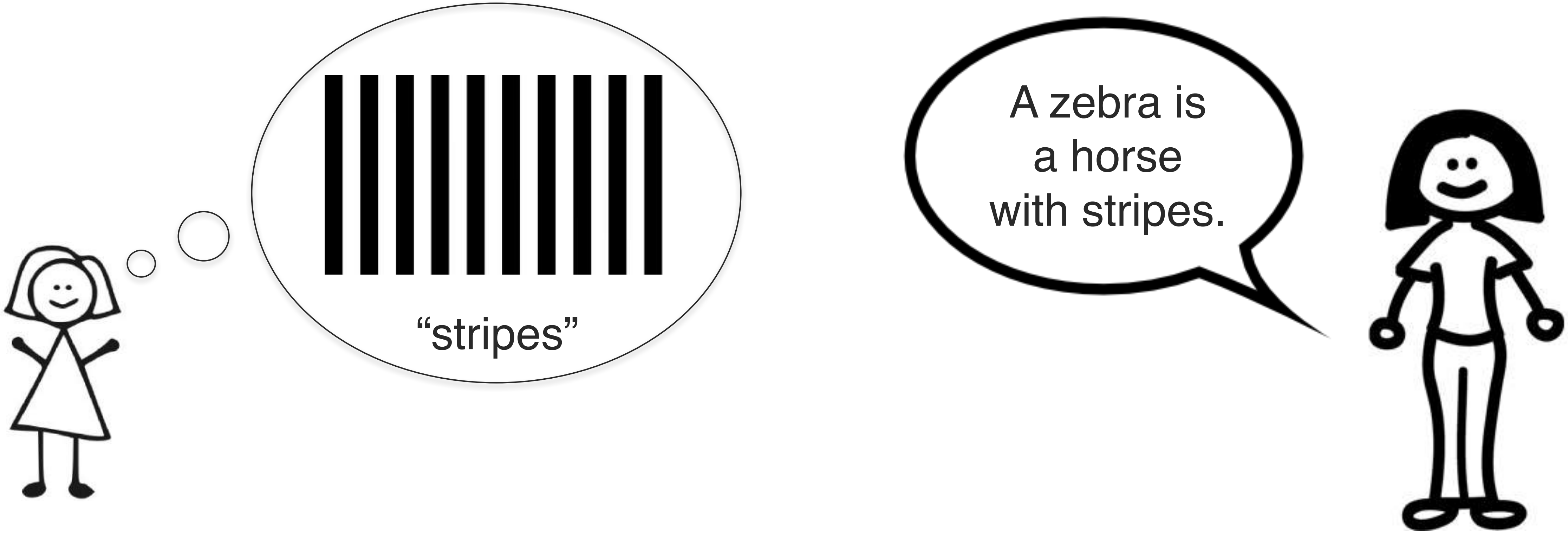


Future Directions

- Learning from limited/missing multimodal data

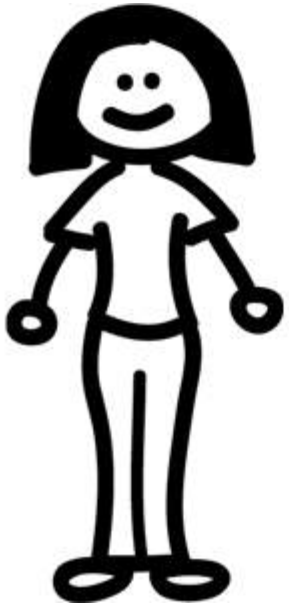
Future Directions

- Learning from limited/missing multimodal data



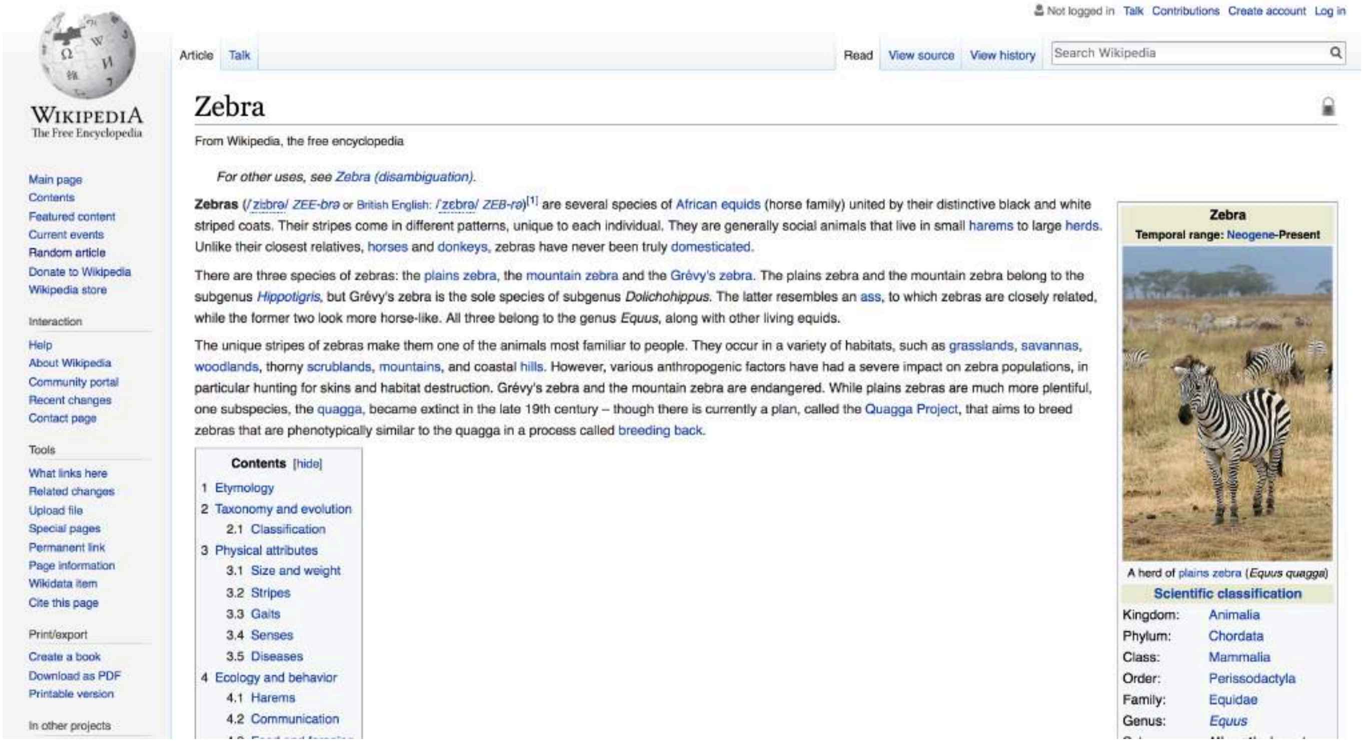
Future Directions

- Learning from limited/missing multimodal data



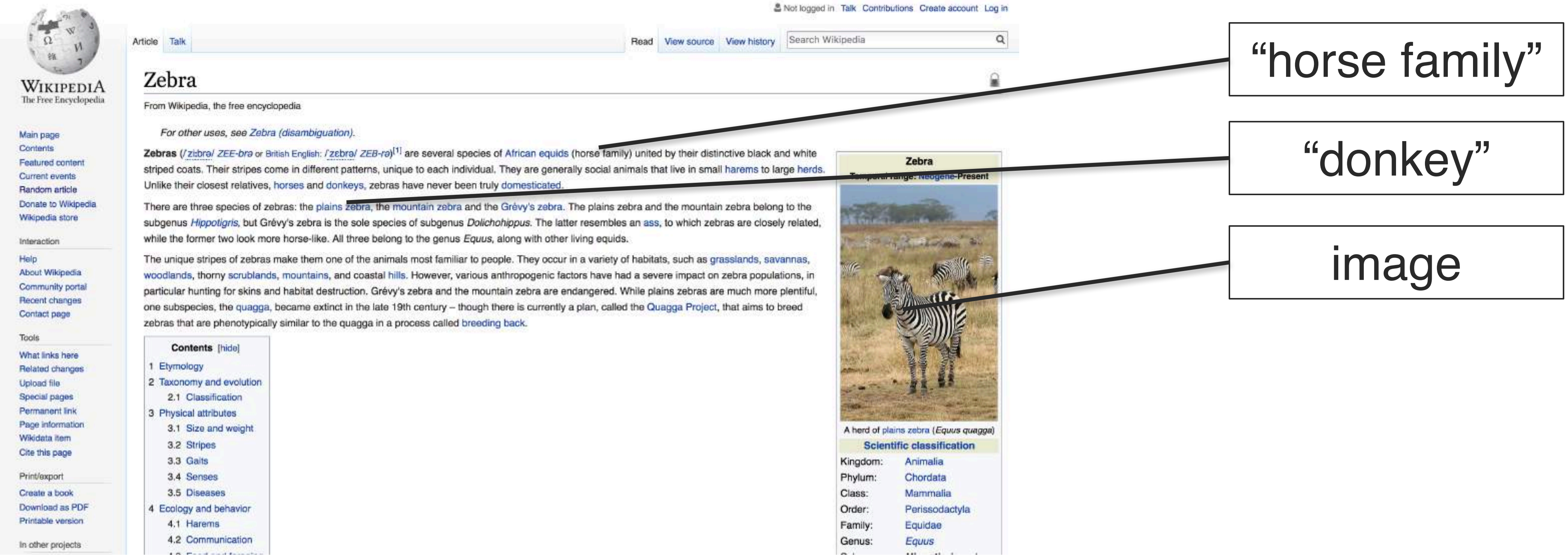
Future Directions

- Learning from unstructured, semi-supervised multimodal data



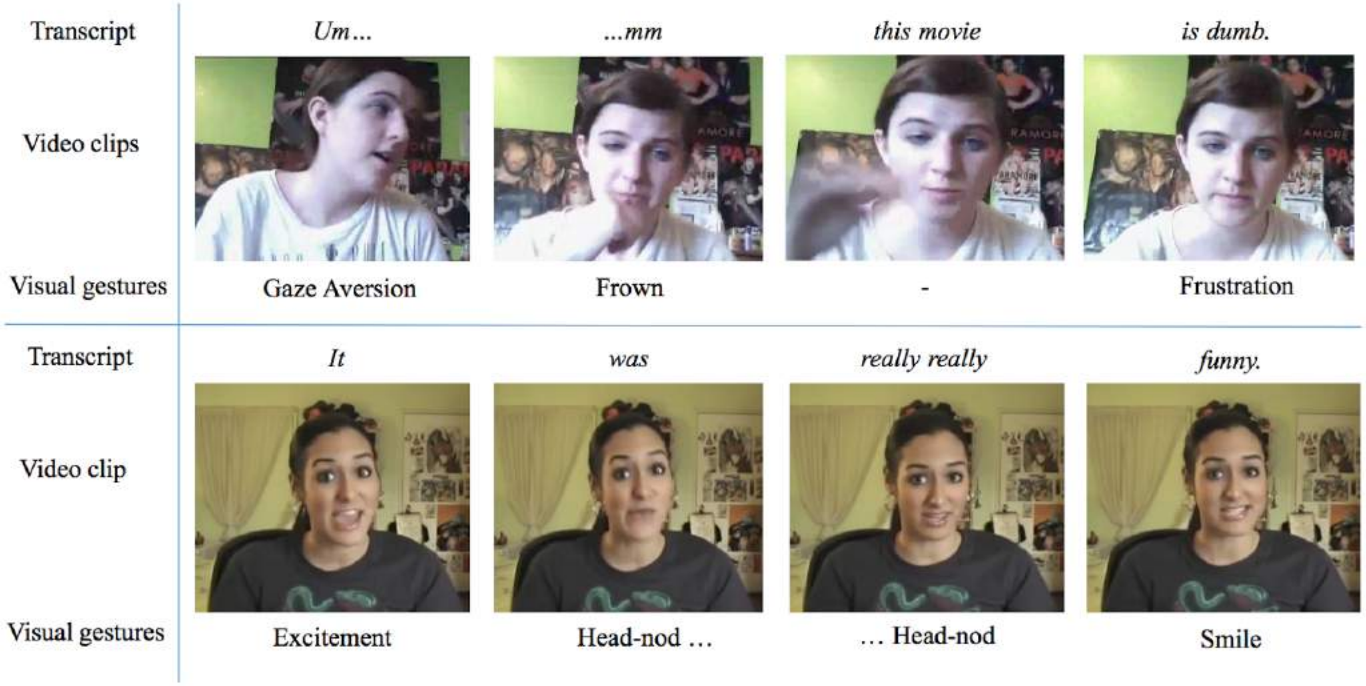
Future Directions

- Learning from unstructured, semi-supervised multimodal data



Future Directions

- Multimodal generation, style transfer, video prediction



Computational Modeling of Multimodal Language

1 **5 Directions**

- Intra-modal and Cross-modal
- Unimodal, Bimodal and Trimodal
- Direct and Relative
- Multimodal Representation Learning
- Robust Multimodal Representation Learning

2 **MOSEI Dataset**

- Diversity in samples, topics, speakers and annotations

The End!

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

Email: pliang@cs.cmu.edu