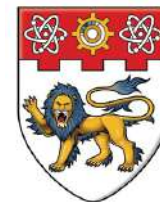




Language  
Technologies  
Institute



Carnegie  
Mellon  
University



NANYANG  
TECHNOLOGICAL  
UNIVERSITY

# Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph

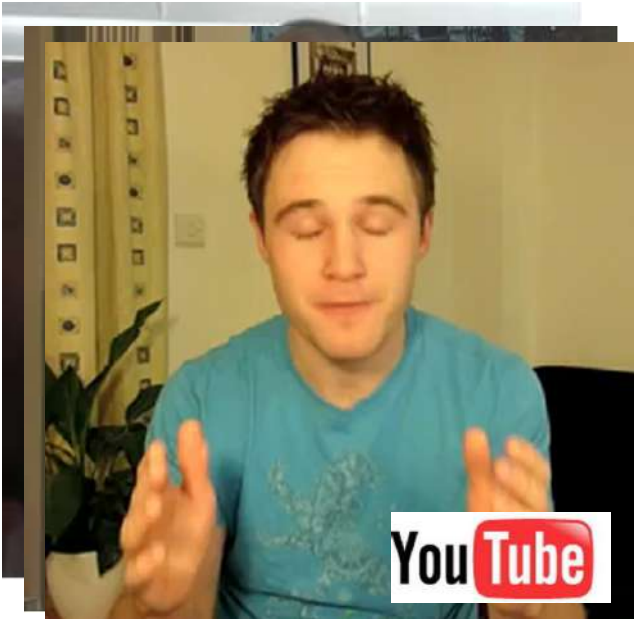
**Presenter: Paul Pu Liang**

Amir Zadeh, Paul Pu Liang, Jonathan Vanbriessen, Soujanya Poria,  
Edmund Tong, Erik Cambria, Minghai Chen, Louis-Philippe Morency

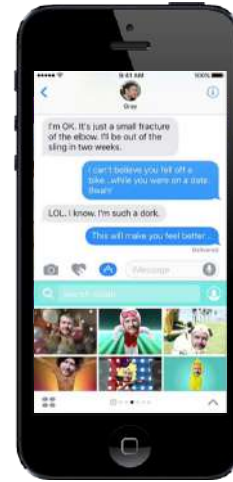
# Progress of Artificial Intelligence

---

## Multimedia Content



## Intelligent Personal Assistants



## Robots and Virtual Agents



# Continuous Theories of (Multimodal) Language

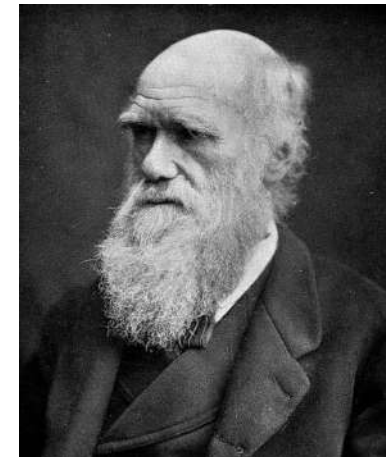
Throughout evolution language and nonverbal behaviors developed together.



Cries and Imitations



Modern Language



# 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

# Multimodal Language Modalities

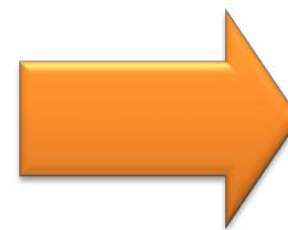
---



# Multimodal Language Modalities

---

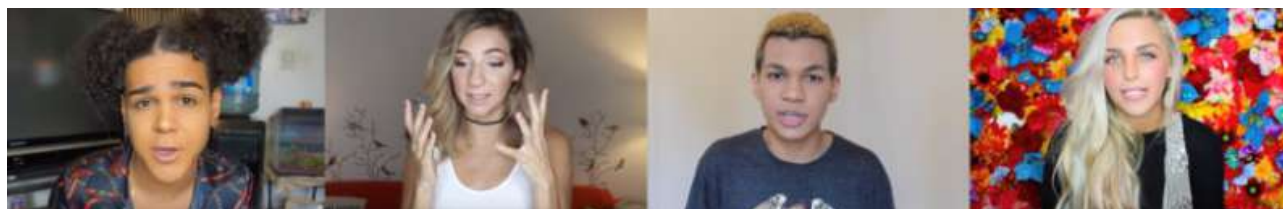
Language Visual Acoustic



Sentiment  
Emotion  
Personality

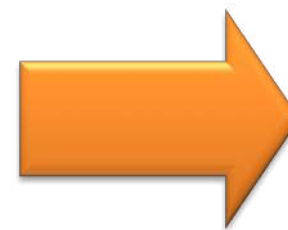
**Datasets**

**Models**



# Multimodal Language Modalities

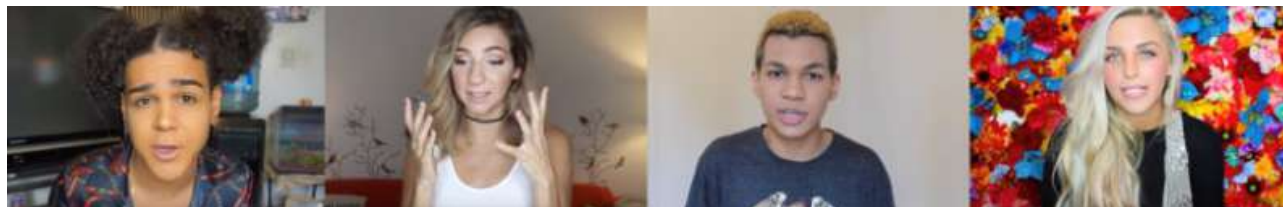
Language Visual Acoustic



Sentiment  
Emotion  
Personality

**Datasets**

**Models**

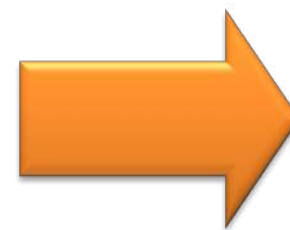


- ✓ Large-scale
- ✓ Diverse



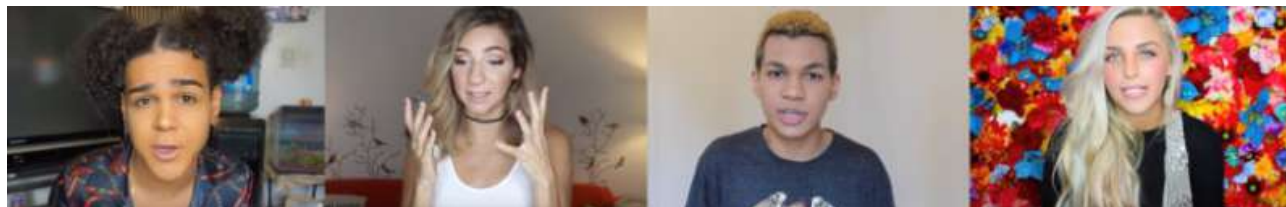
# Multimodal Language Modalities

Language Visual Acoustic



Sentiment  
Emotion  
Personality

## Datasets



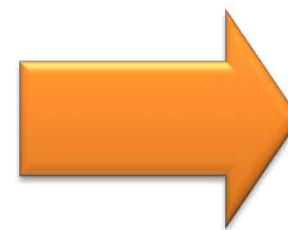
- ✓ Large-scale
- ✓ Diverse

## Models

- Word-level alignment
- Attention models
- Memory-based models

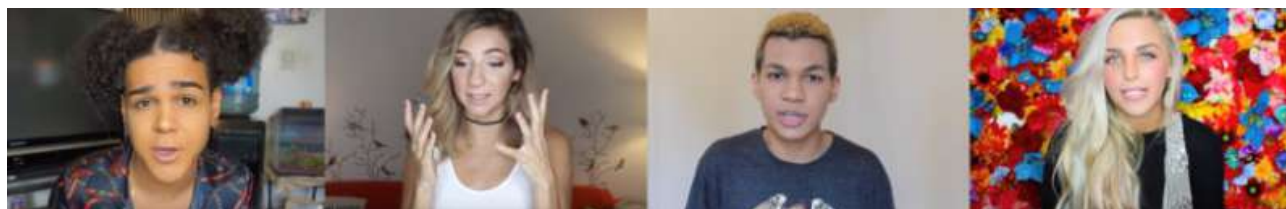
# Multimodal Language Modalities

Language Visual Acoustic



Sentiment  
Emotion  
Personality

## Datasets



- ✓ Large-scale
- ✓ Diverse

## Models

- Word-level alignment
  - Attention models
  - Memory-based models
- ✓ Good Performance
  - ✓ Interpretable

# Datasets for Multimodal Language

---

- Require large and diverse amounts of data: **(Novelty)**

- Diversity in **samples**



# Datasets for Multimodal Language

---

- Require large and diverse amounts of data: **(Novelty)**

- Diversity in **samples**



- Diversity in **topics**

reviews debates movies drama speech politics science



# Datasets for Multimodal Language

- Require large and diverse amounts of data: **(Novelty)**

- Diversity in **samples**
- Diversity in **topics**
- Diversity in **speakers**



reviews debates movies drama speech politics science





# Datasets for Multimodal Language

- Require large and diverse amounts of data: **(Novelty)**

- Diversity in **samples**



- Diversity in **topics**

reviews debates movies drama speech politics science

- Diversity in **speakers**



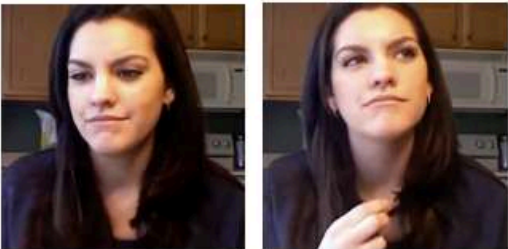
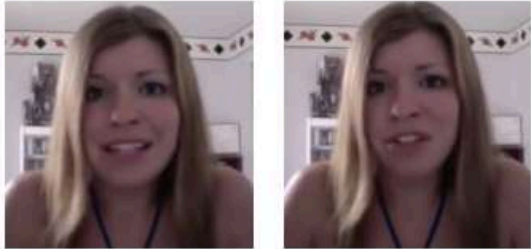
- Diversity in **annotations**



# New Dataset: CMU-MOSEI

---

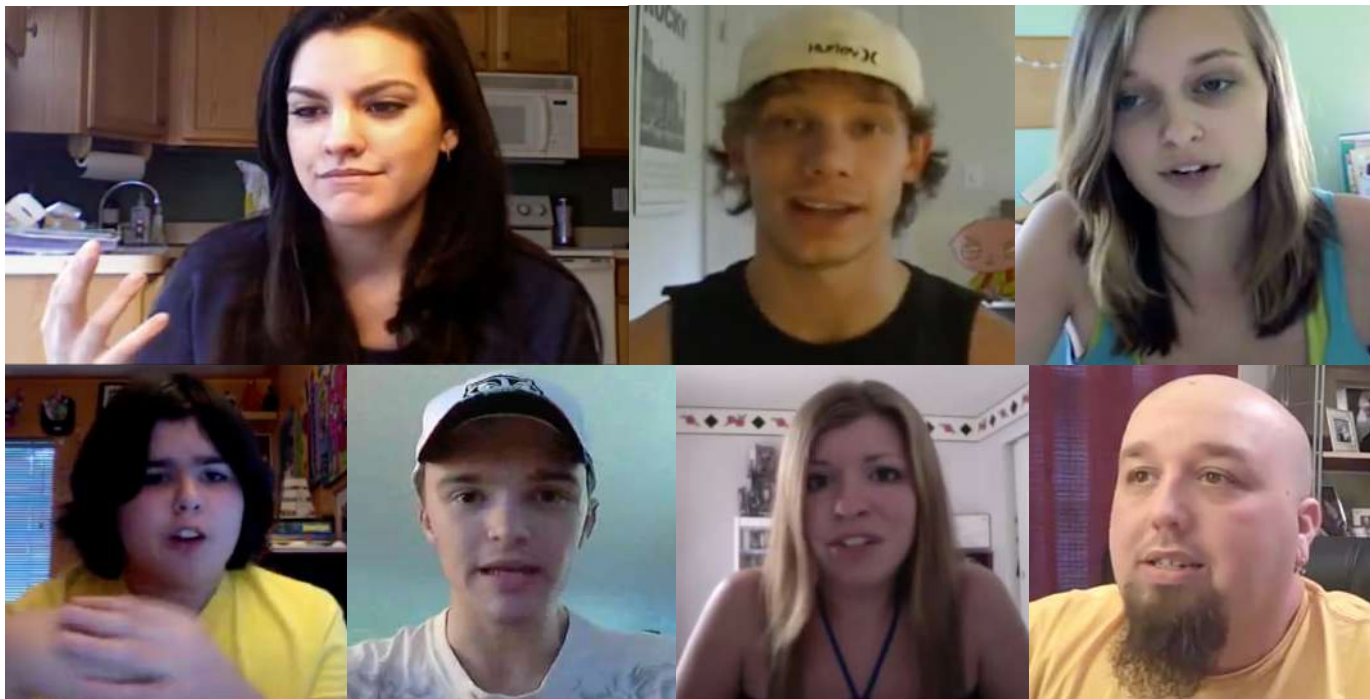
23,000 video segments  
3 modalities

<b>Language:</b>	<i>And he I don't think he got mad when hah I don't know maybe.</i>		<i>Too much too fast, I mean we basically just get introduced to this character...</i>		<i>All I can say is he's a pretty average guy.</i>	
<b>Vision:</b>						
<b>Acoustic:</b>	(frustrated voice)		(angry voice)		(disappointed voice)	



# CMU-MOSEI Dataset

1,000 speakers

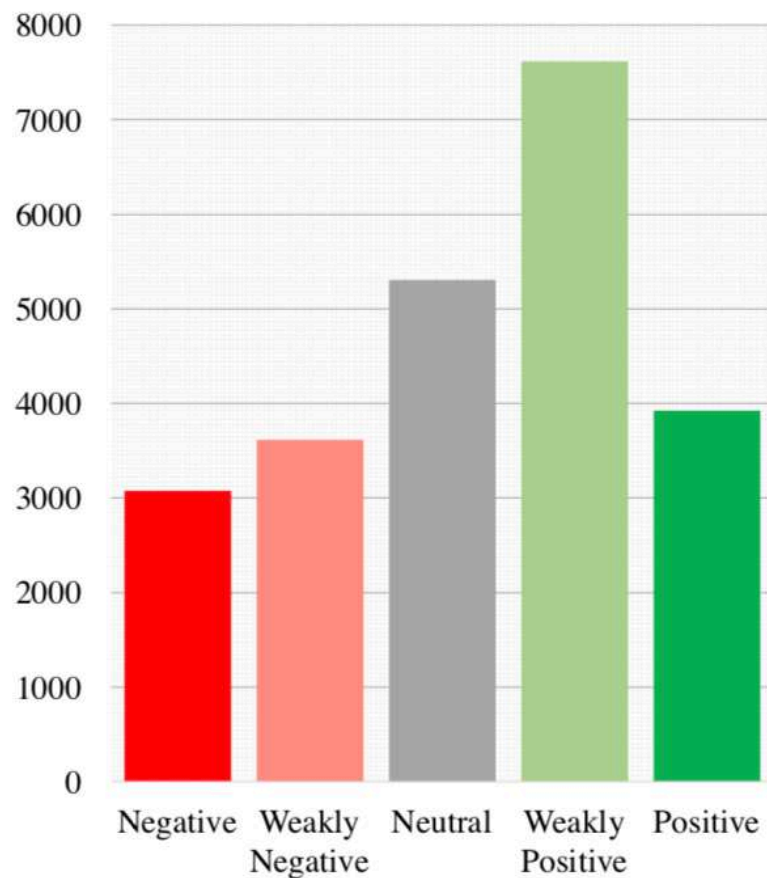


# 250 topics

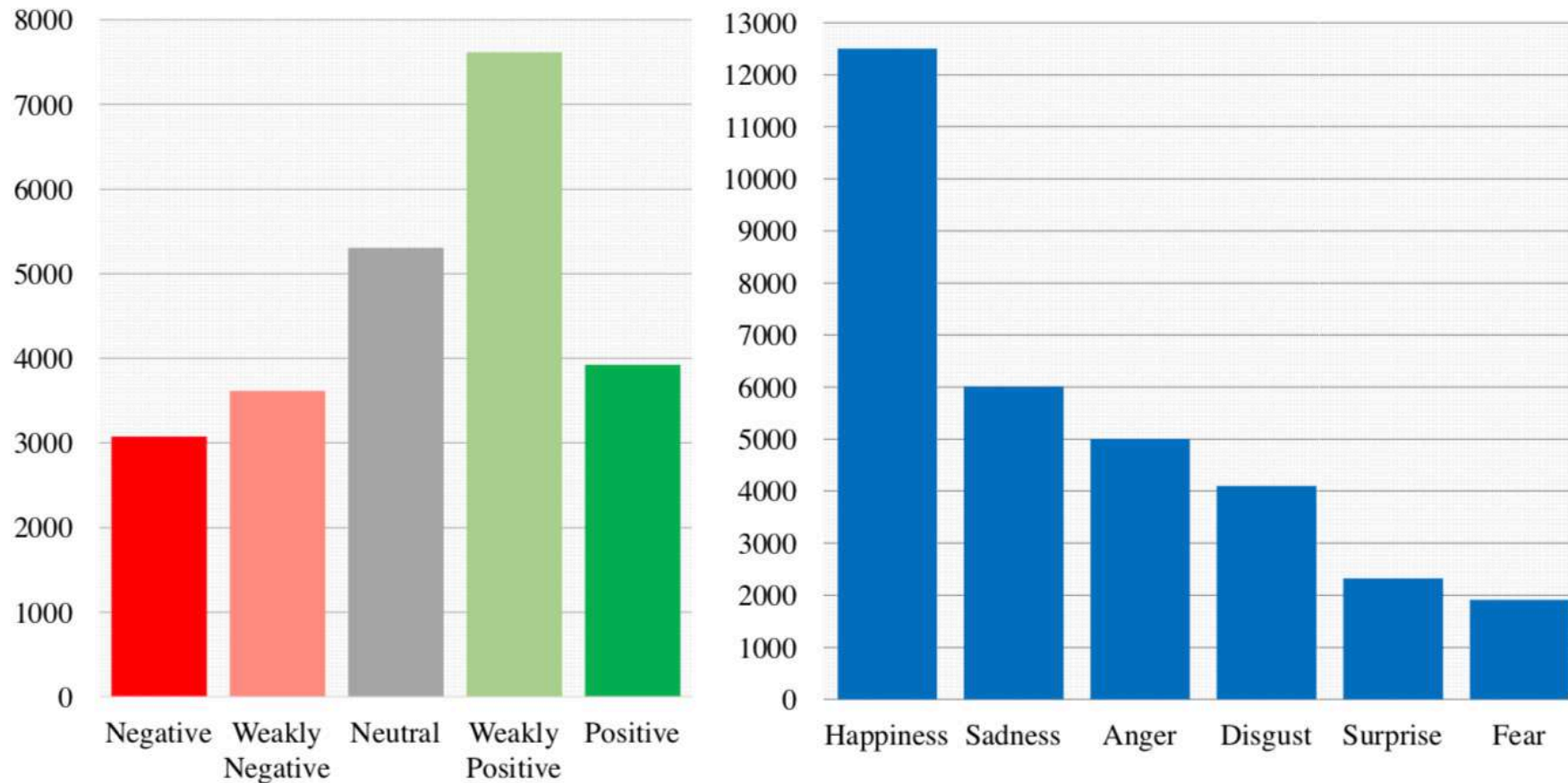


# Annotation Distributions

---

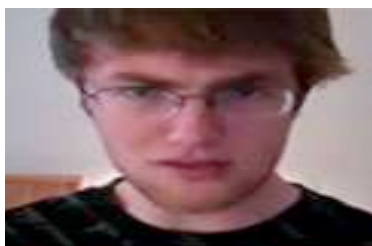


# Annotation Distributions





# Feature Extraction



## Language

- Glove word embeddings

## Visual

- Facet features
- MultiComp OpenFace
- Face embeddings

## Acoustic

- COVAREP features
  - MFCCs
  - Pitch tracking

## Alignment

- Word level
- P2FA



## Sentiment

- Positive
- Negative

## Emotion

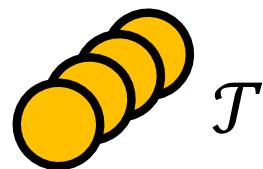
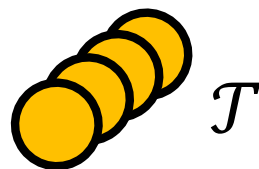
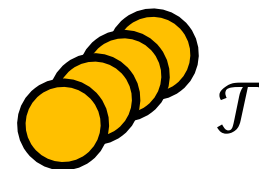
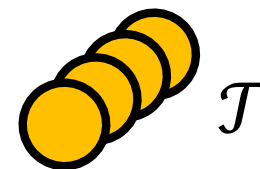
- Anger
- Disgust
- Fear
- Happiness
- Sadness
- Surprise

# CMU-MOSEI Dataset

	Dataset	# S	# Sp	Mod	Sent	Emo	TL (hh:mm:ss)
Multimodal	<b>CMU-MOSEI</b>	<b>23,453</b>	<b>1,000</b>	$\{l, v, a\}$	✓	✓	<b>65:53:36</b>
	CMU-MOSI	2,199	98	$\{l, v, a\}$	✓	✗	02:36:17
	ICT-MMMO	340	200	$\{l, v, a\}$	✓	✗	13:58:29
	YouTube	300	50	$\{l, v, a\}$	✓	✗	00:29:41
	MOUD	400	101	$\{l, v, a\}$	✓	✗	00:59:00
Language	IEMOCAP	10,000	10	$\{l, v, a\}$	✗	✓	11:28:12
	SST	11,855	—	$\{l\}$	✓	✗	—
	Cornell	2,000	—	$\{l\}$	✓	✗	—
Audio-visual	HUMAINE	50	4	$\{v, a\}$	✗	✓	04:11:00
	RECOLA	46	46	$\{v, a\}$	✗	✓	03:50:00
	SEWA	538	408	$\{v, a\}$	✗	✓	04:39:00
	SEMAINE	80	20	$\{v, a\}$	✗	✓	06:30:00
	AFEW	1,645	330	$\{v, a\}$	✗	✓	02:28:03

# Models for Multimodal Language

multimodal

 $\mathcal{T}$  $\mathcal{T}$  $\mathcal{T}$  $\mathcal{T}$ 

Multimodal Fusion

Transcript

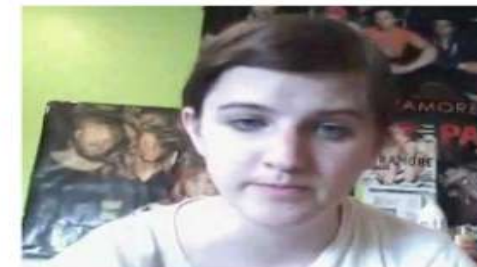
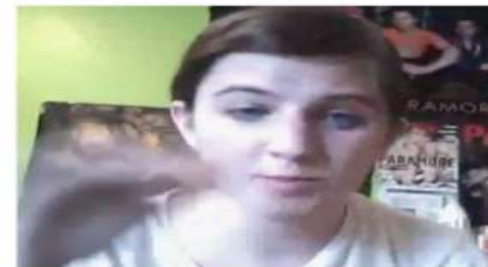
*Um...*

*...mm*

*this movie*

*is dumb.*

Video clips



Visual gestures

Gaze Aversion

Frown

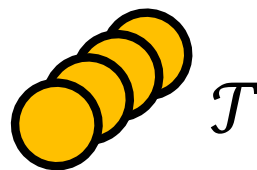
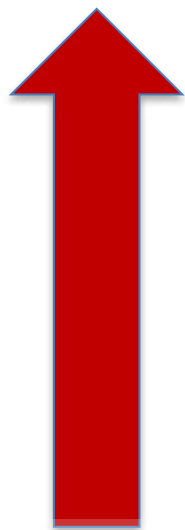
-

Frustration



# Models for Multimodal Language

multimodal

 $\mathcal{T}$ 

## Multimodal Fusion

## Interpretation

- Importance of each modality
- Interactions between modalities

Transcript

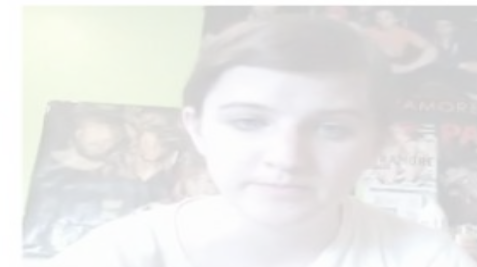
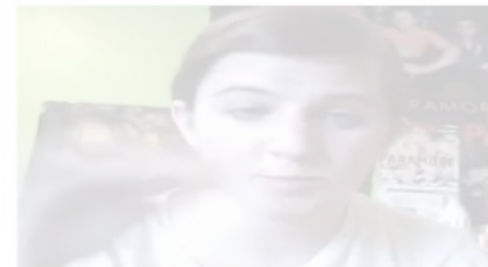
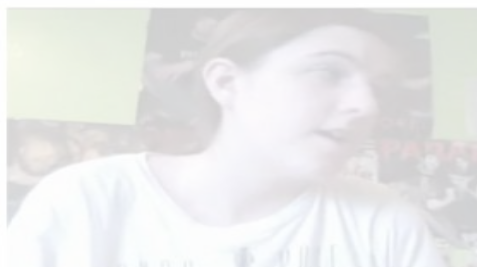
*Um...*

*...mm*

*this movie*

*is dumb.*

Video clips



Visual gestures

Gaze Aversion

Frown

-

Frustration

# Dynamic Fusion Graph (DFG)

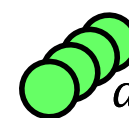
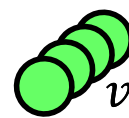
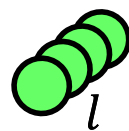
multimodal



## Interpretation

- Importance of each modality
- Interactions between modalities

unimodal



Transcript

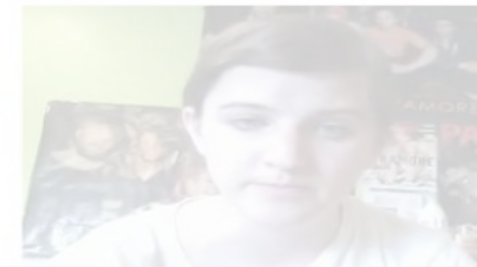
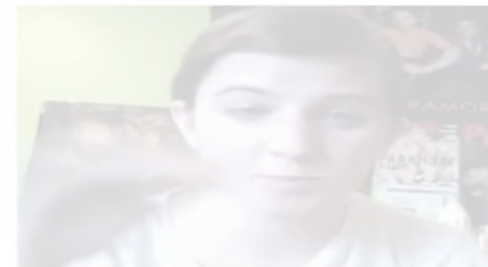
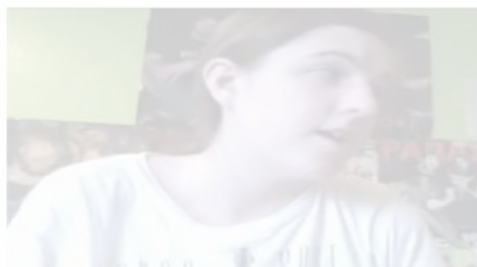
*Um...*

*...mm*

*this movie*

*is dumb.*

Video clips



Visual gestures

Gaze Aversion

Frown

-

Frustration



# Dynamic Fusion Graph (DFG)

multimodal



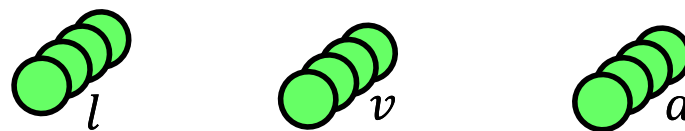
## Interpretation

- Importance of each modality
- Interactions between modalities

bimodal



unimodal



Transcript

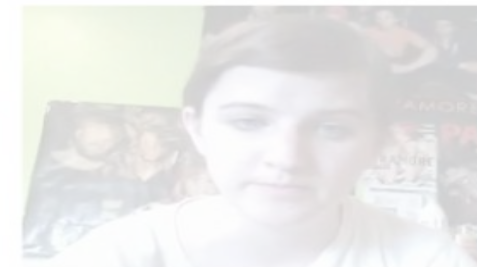
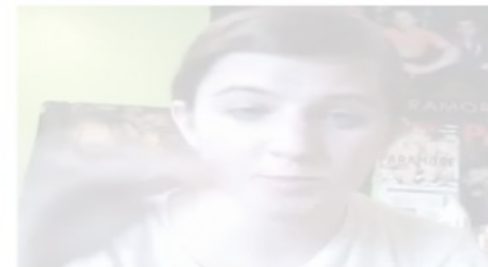
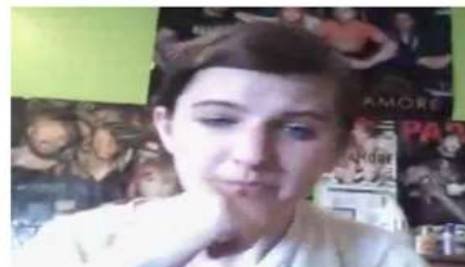
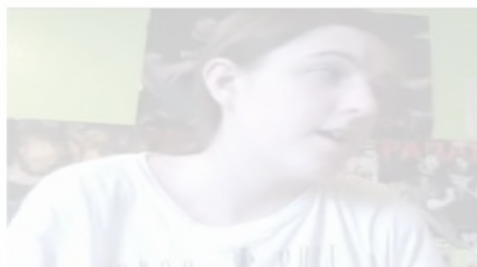
*Um...*

*...mm*

*this movie*

*is dumb.*

Video clips



Visual gestures

Gaze Aversion

Frown

-

Frustration

# Dynamic Fusion Graph (DFG)

multimodal



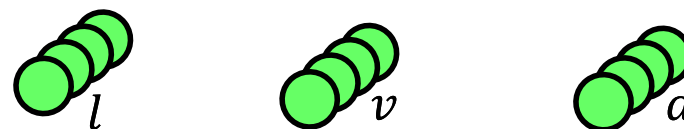
trimodal



bimodal



unimodal



## Interpretation

- Importance of each modality
- Interactions between modalities

Transcript

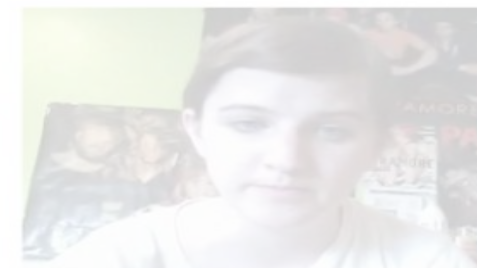
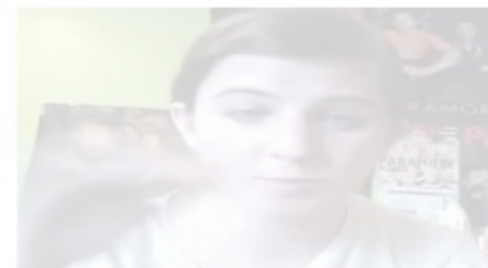
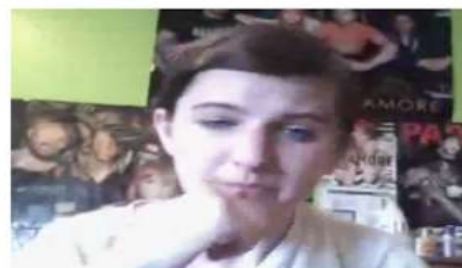
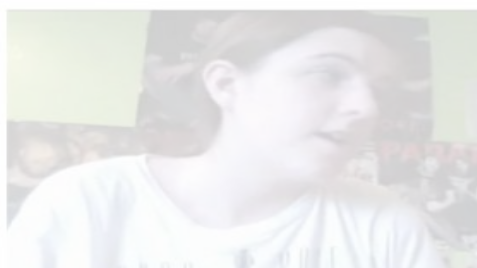
*Um...*

*...mm*

*this movie*

*is dumb.*

Video clips



Visual gestures

Gaze Aversion

Frown

-

Frustration

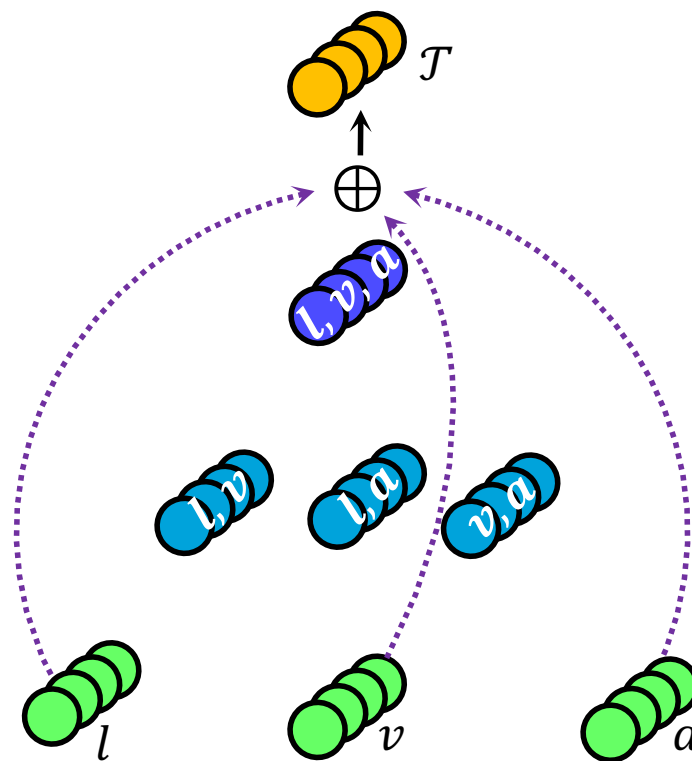
# Dynamic Fusion Graph (DFG)

multimodal

trimodal

bimodal

unimodal



## Interpretation

- Importance of each modality
- Interactions between modalities

fusion weights

Transcript

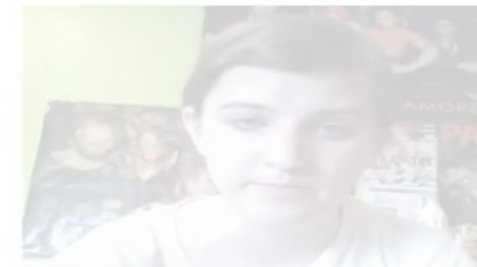
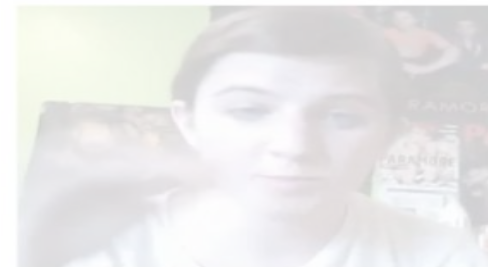
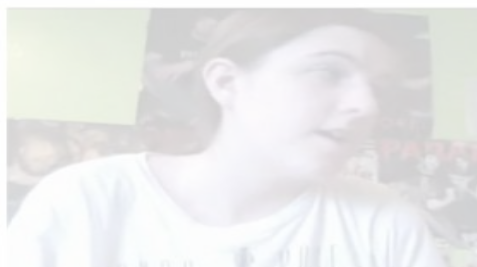
*Um...*

*...mm*

*this movie*

*is dumb.*

Video clips



Visual gestures

Gaze Aversion

Frown

-

Frustration

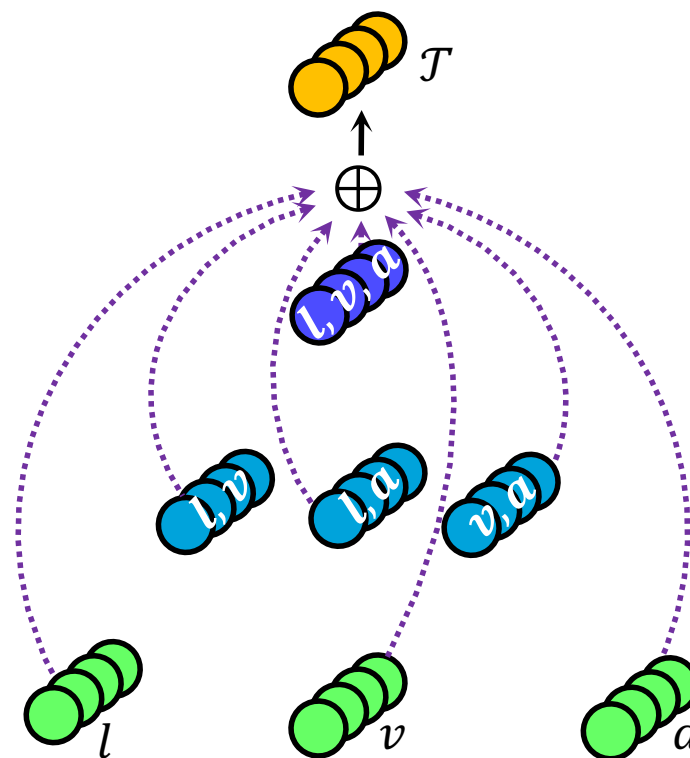
# Dynamic Fusion Graph (DFG)

multimodal

trimodal

bimodal

unimodal



## Interpretation

- Importance of each modality
- **Interactions between modalities**

**fusion weights**

Transcript

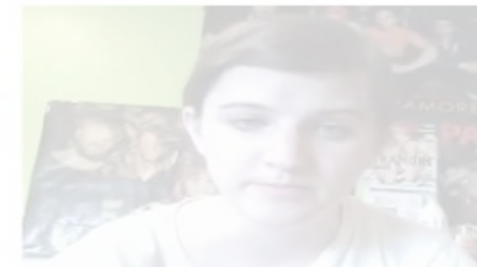
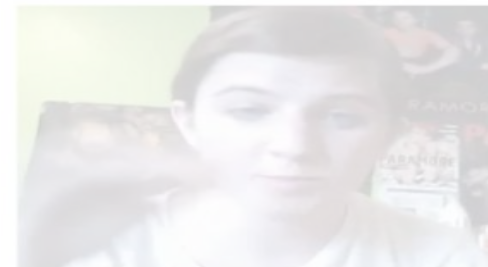
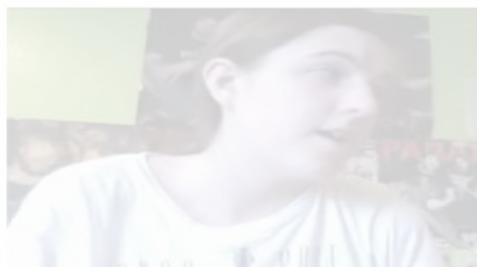
*Um...*

*...mm*

*this movie*

*is dumb.*

Video clips



Visual gestures

Gaze Aversion

Frown

-

Frustration



# Dynamic Fusion Graph (DFG)

multimodal

trimodal

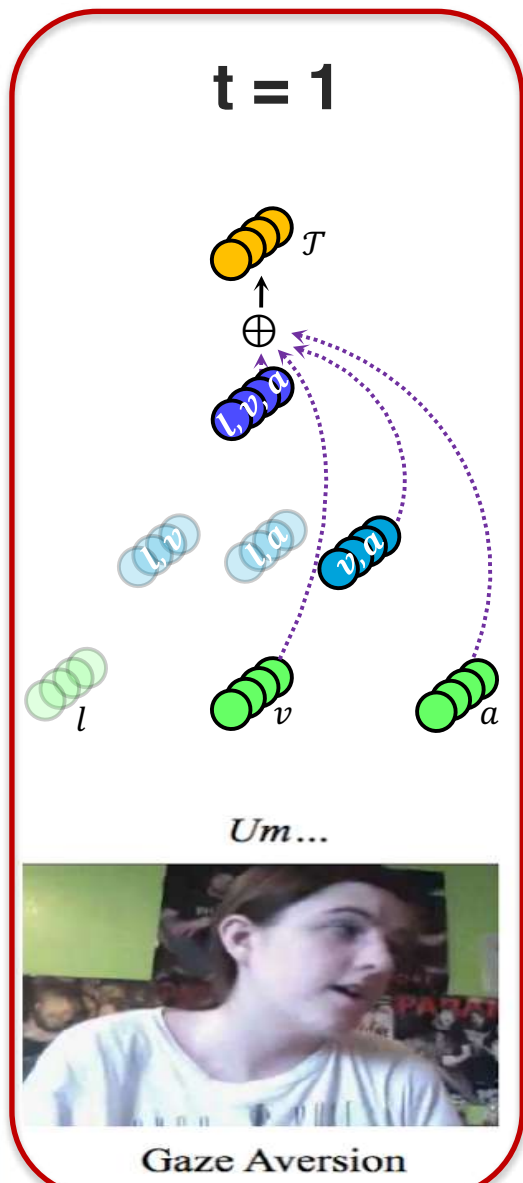
bimodal

unimodal

Transcript

Video clips

Visual gestures



Gaze Aversion

...mm

Frown

this movie

-

is dumb.

Frustration



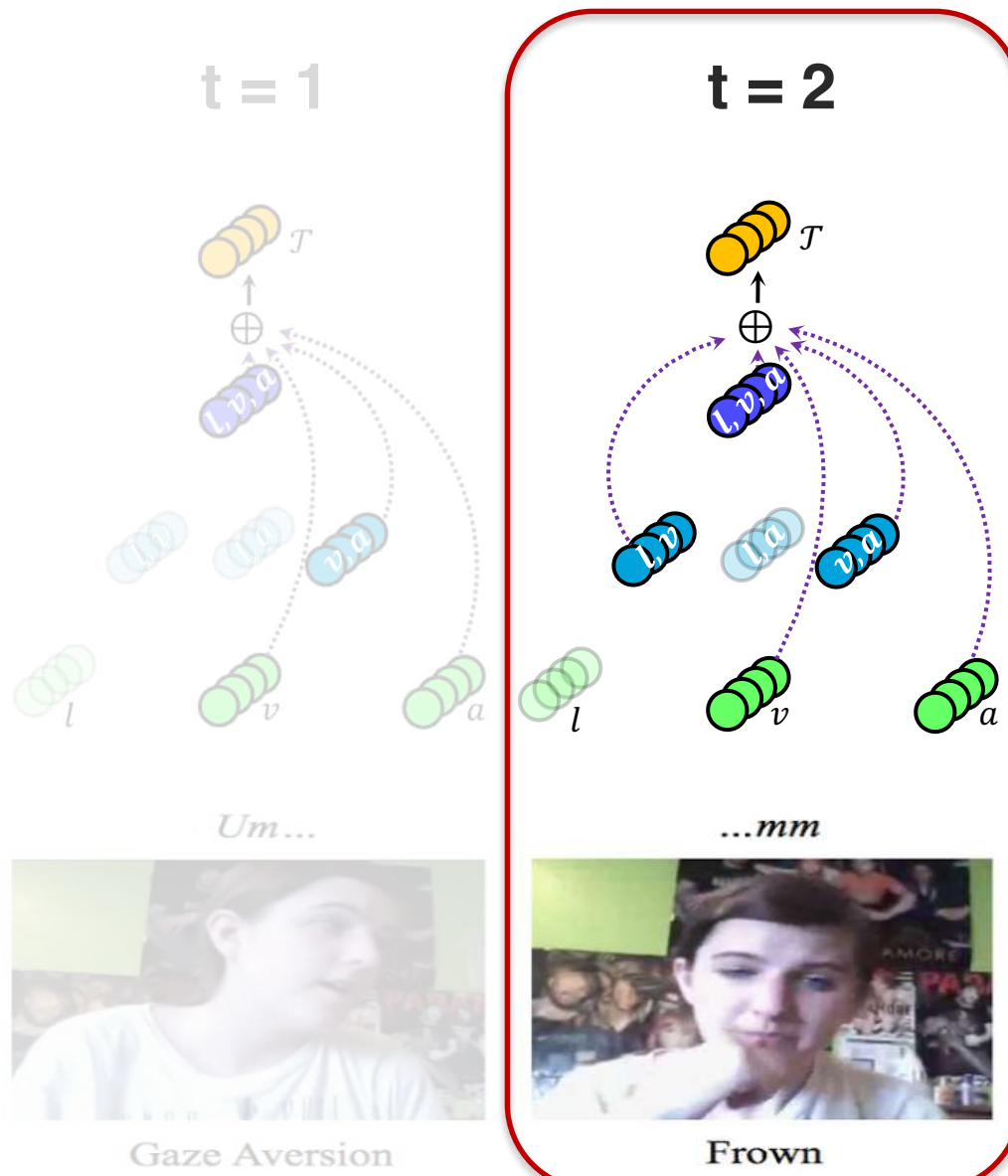
# Dynamic Fusion Graph (DFG)

multimodal

trimodal

bimodal

unimodal



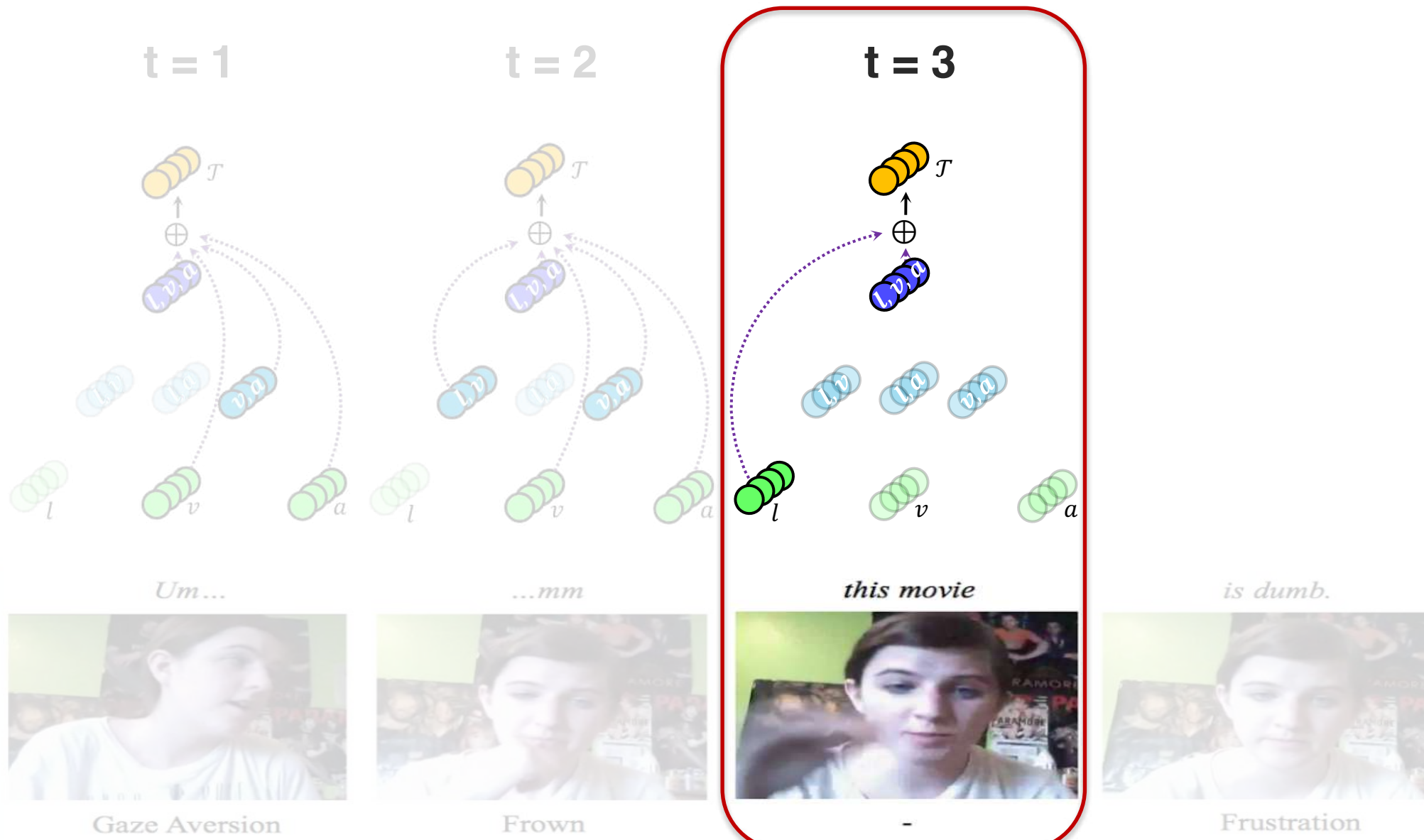
# Dynamic Fusion Graph (DFG)

multimodal

trimodal

bimodal

unimodal



# Dynamic Fusion Graph (DFG)

multimodal

trimodal

bimodal

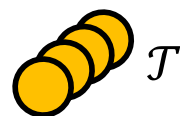
unimodal





# Dynamic Fusion Graph (DFG)

multimodal



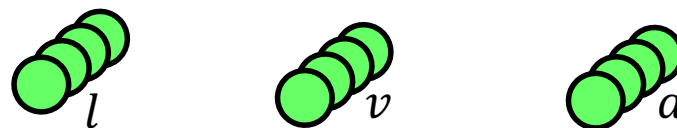
trimodal



bimodal



unimodal



## Interpretation

- Importance of each modality
- Interactions between modalities

fusion weights

Transcript

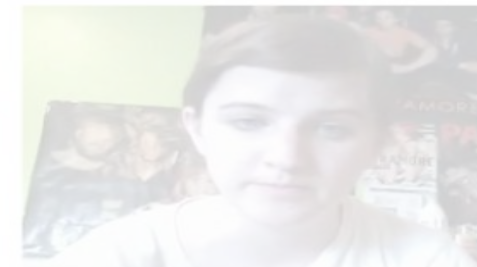
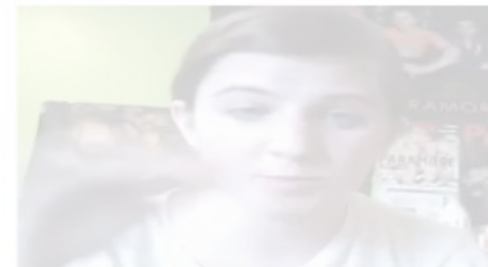
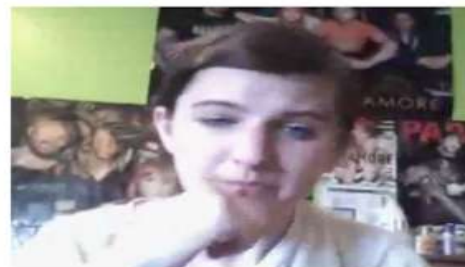
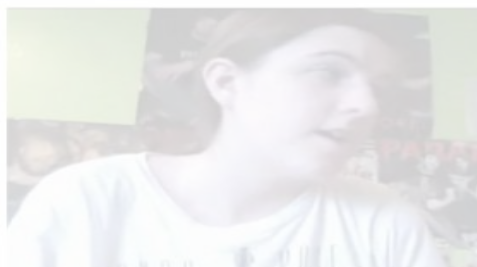
*Um...*

*...mm*

*this movie*

*is dumb.*

Video clips



Visual gestures

Gaze Aversion

Frown

-

Frustration



# Dynamic Fusion Graph (DFG)

multimodal



trimodal

bimodal

unimodal

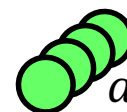
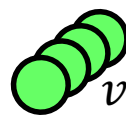
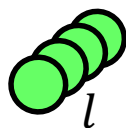
## Interpretation

- Importance of each modality
- Interactions between modalities

fusion weights

- Construction of bimodal and trimodal representations

construction weights



Transcript

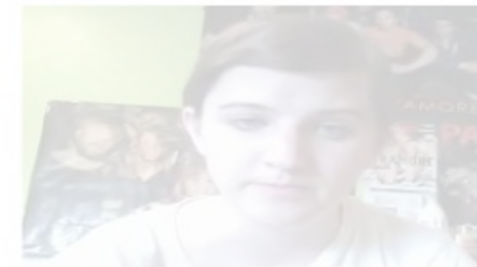
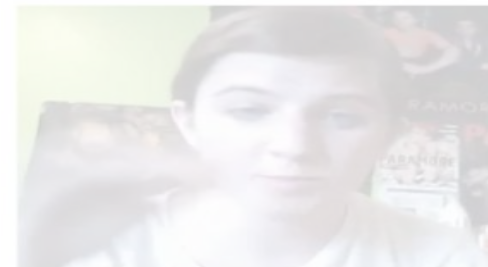
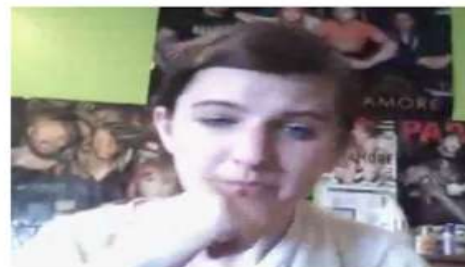
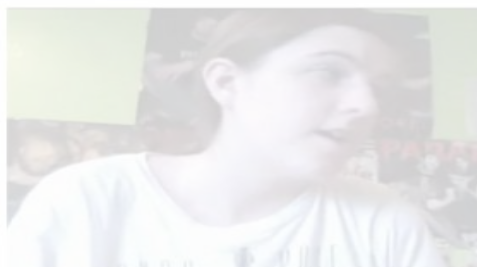
*Um...*

*...mm*

*this movie*

*is dumb.*

Video clips



Visual gestures

Gaze Aversion

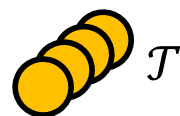
Frown

-

Frustration

# Dynamic Fusion Graph (DFG)

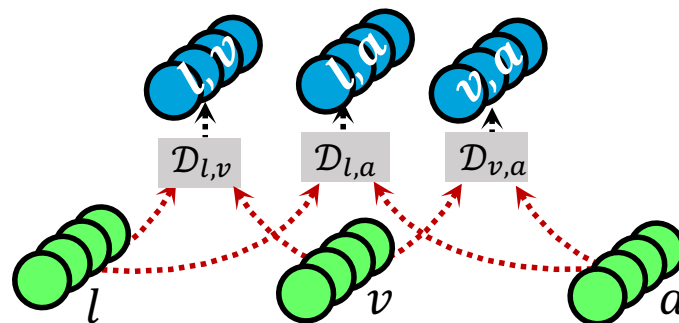
multimodal



trimodal

bimodal

unimodal



## Interpretation

- Importance of each modality
- Interactions between modalities

fusion weights

- Construction of bimodal and trimodal representations**

construction weights

Transcript

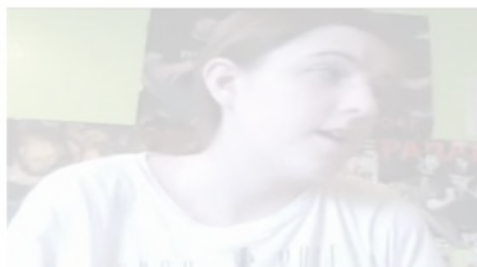
*Um...*

*...mm*

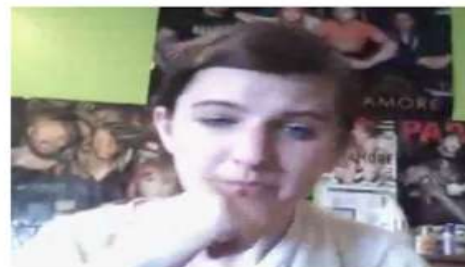
*this movie*

*is dumb.*

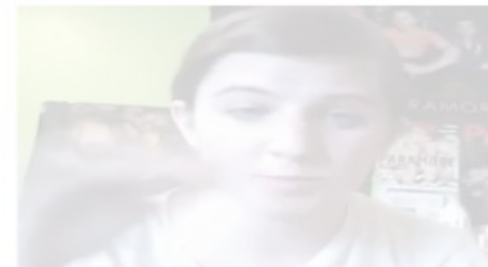
Video clips



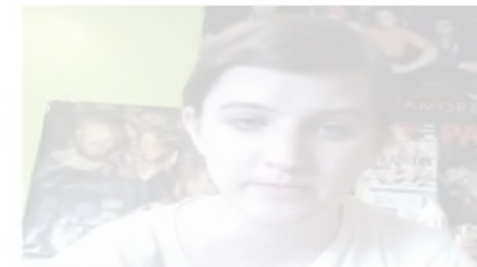
Gaze Aversion



Frown



-



Frustration

Visual gestures

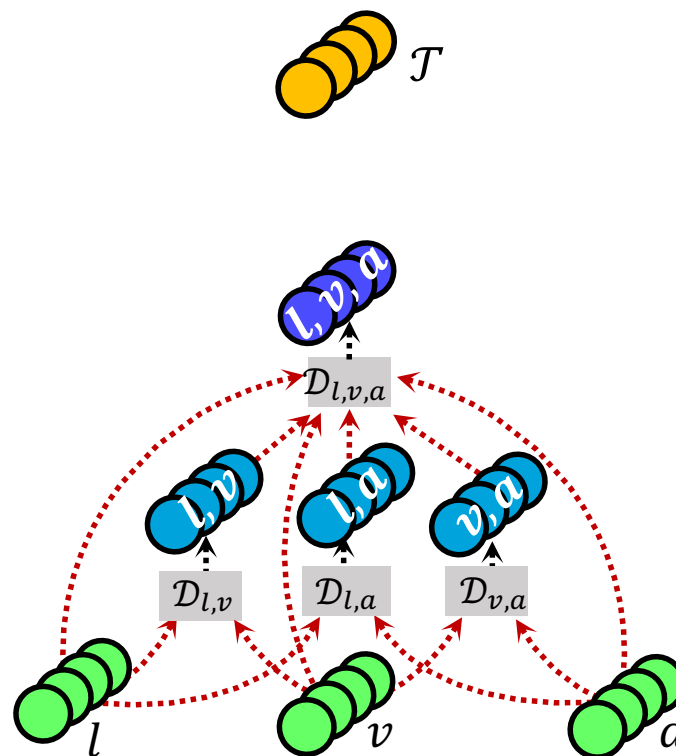
# Dynamic Fusion Graph (DFG)

multimodal

trimodal

bimodal

unimodal



## Interpretation

- Importance of each modality
- Interactions between modalities

fusion weights

- Construction of bimodal and trimodal representations

construction weights

Transcript

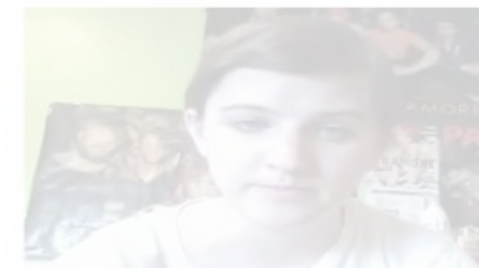
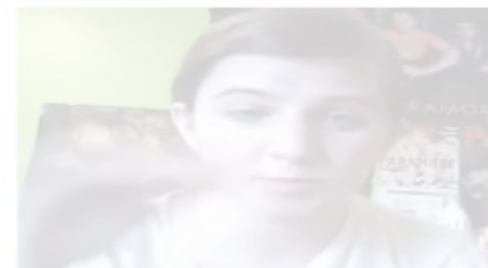
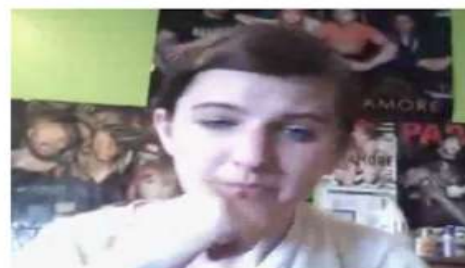
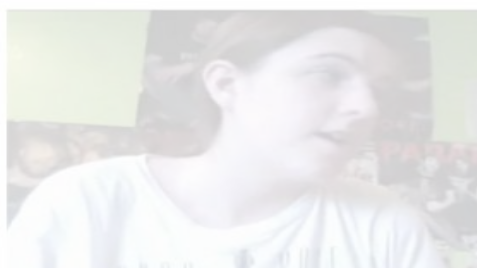
*Um...*

*...mm*

*this movie*

*is dumb.*

Video clips



Visual gestures

Gaze Aversion

Frown

-

Frustration

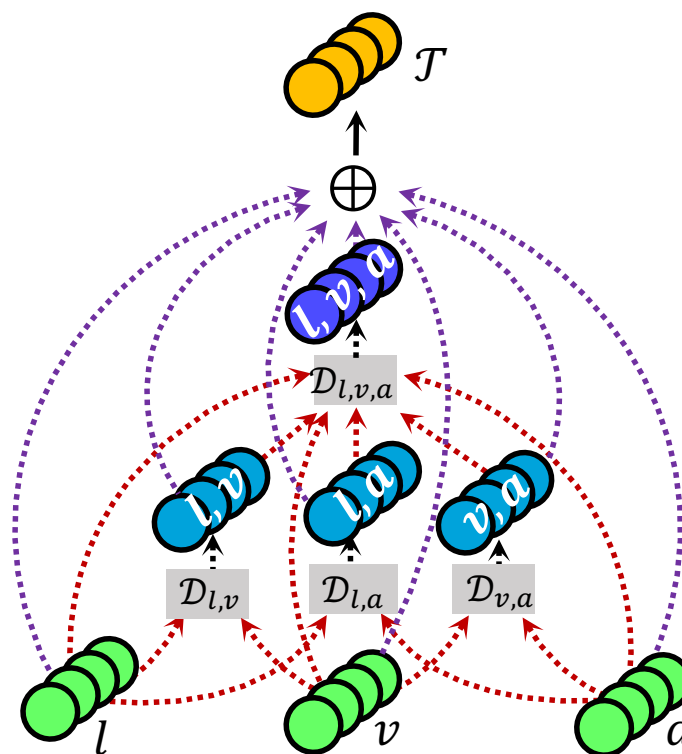
# Dynamic Fusion Graph (DFG)

multimodal

trimodal

bimodal

unimodal



## Interpretation

- Importance of each modality
- Interactions between modalities

fusion weights

- Construction of bimodal and trimodal representations

construction weights

Transcript

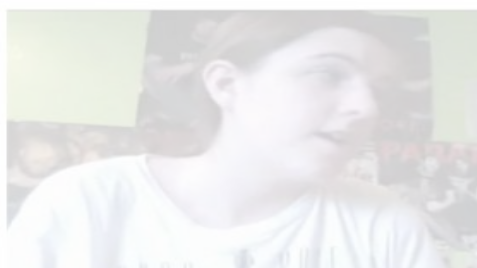
*Um...*

*...mm*

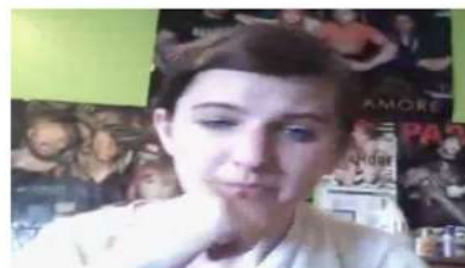
*this movie*

*is dumb.*

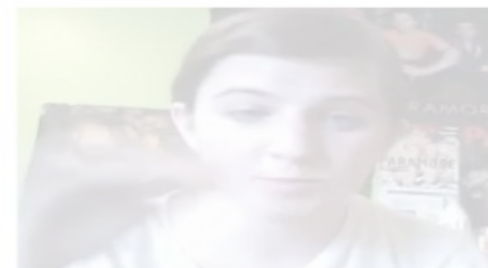
Video clips



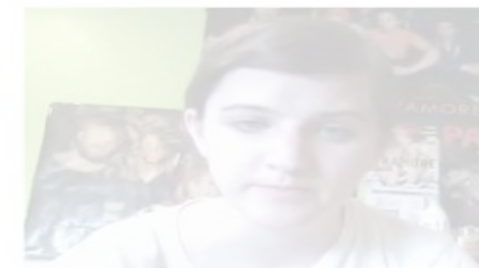
Gaze Aversion



Frown



-



Frustration

Visual gestures



# Dynamic Fusion Graph (DFG)

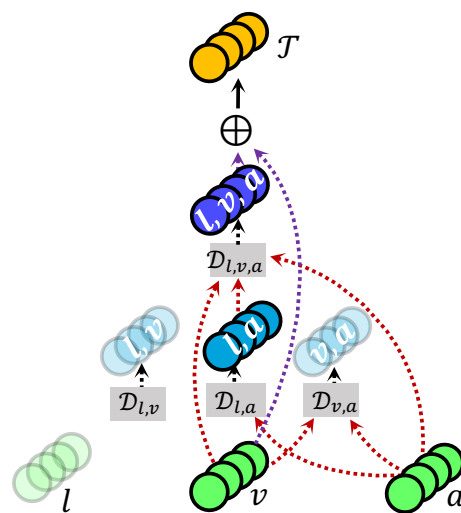
multimodal

trimodal

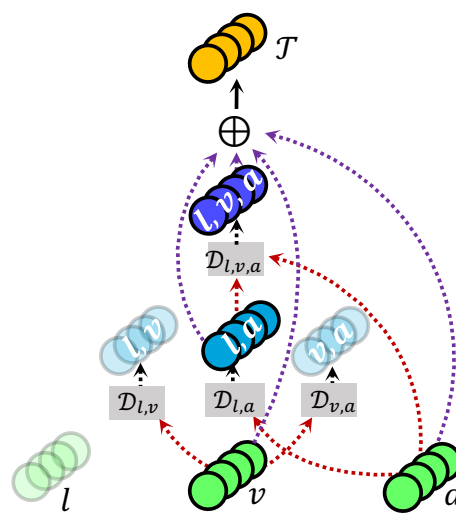
bimodal

unimodal

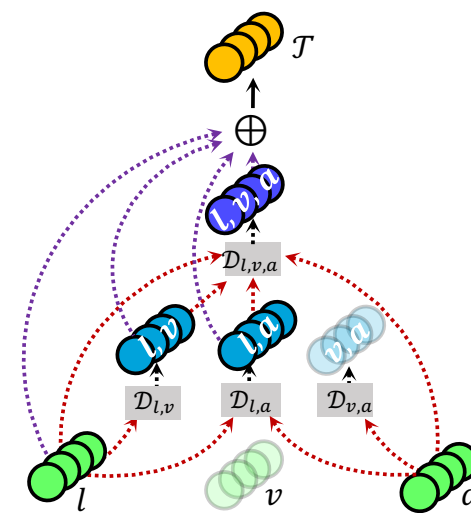
$t = 1$



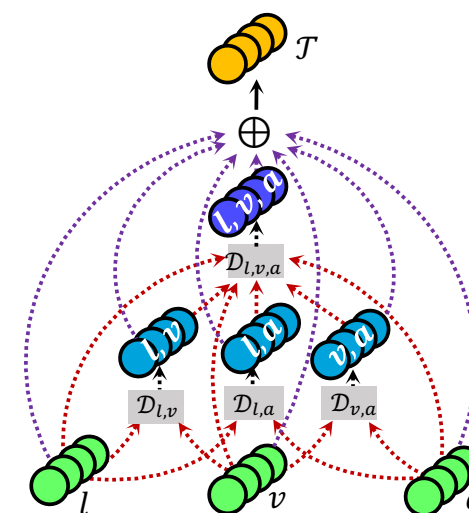
$t = 2$



$t = 3$



$t = 4$



Transcript

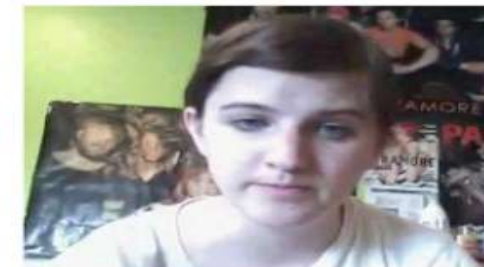
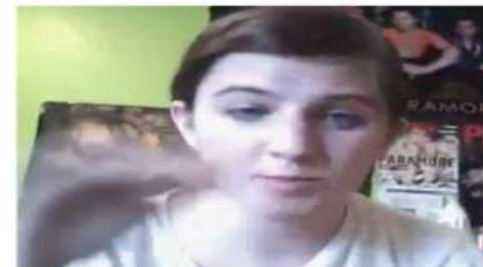
*Um...*

*...mm*

*this movie*

*is dumb.*

Video clips



Visual gestures

Gaze Aversion

Frown

-

Frustration

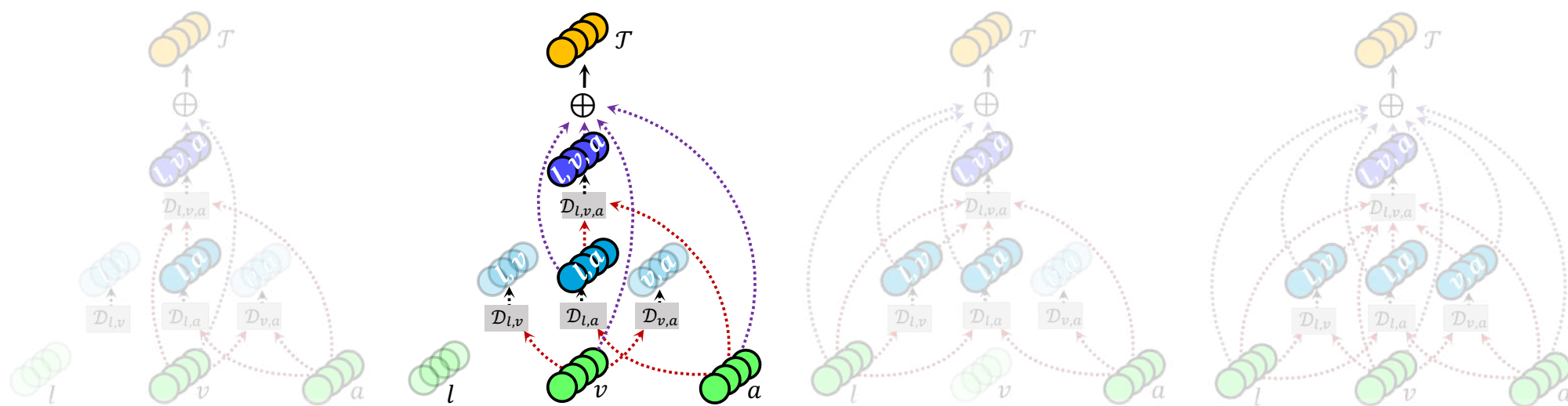
# Graph-Memory Fusion Network (Graph-MFN)

multimodal

trimodal

bimodal

unimodal



Transcript

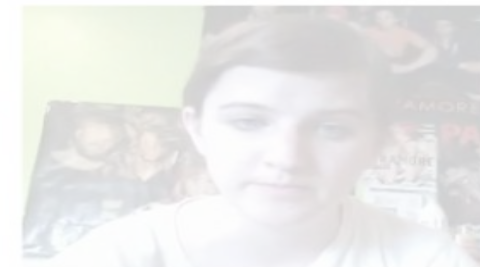
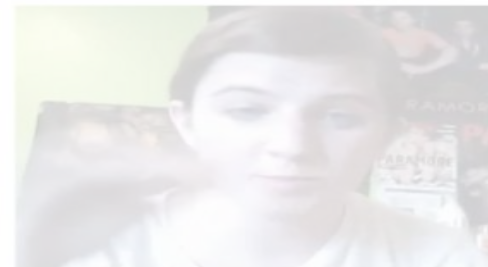
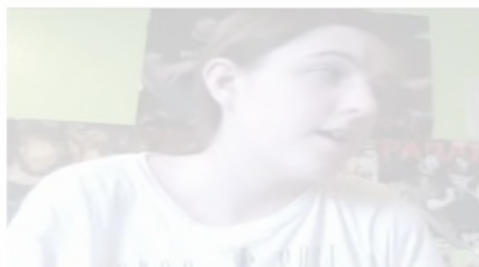
*Um...*

*...mm*

*this movie*

*is dumb.*

Video clips



Visual gestures

Gaze Aversion

Frown

-

Frustration

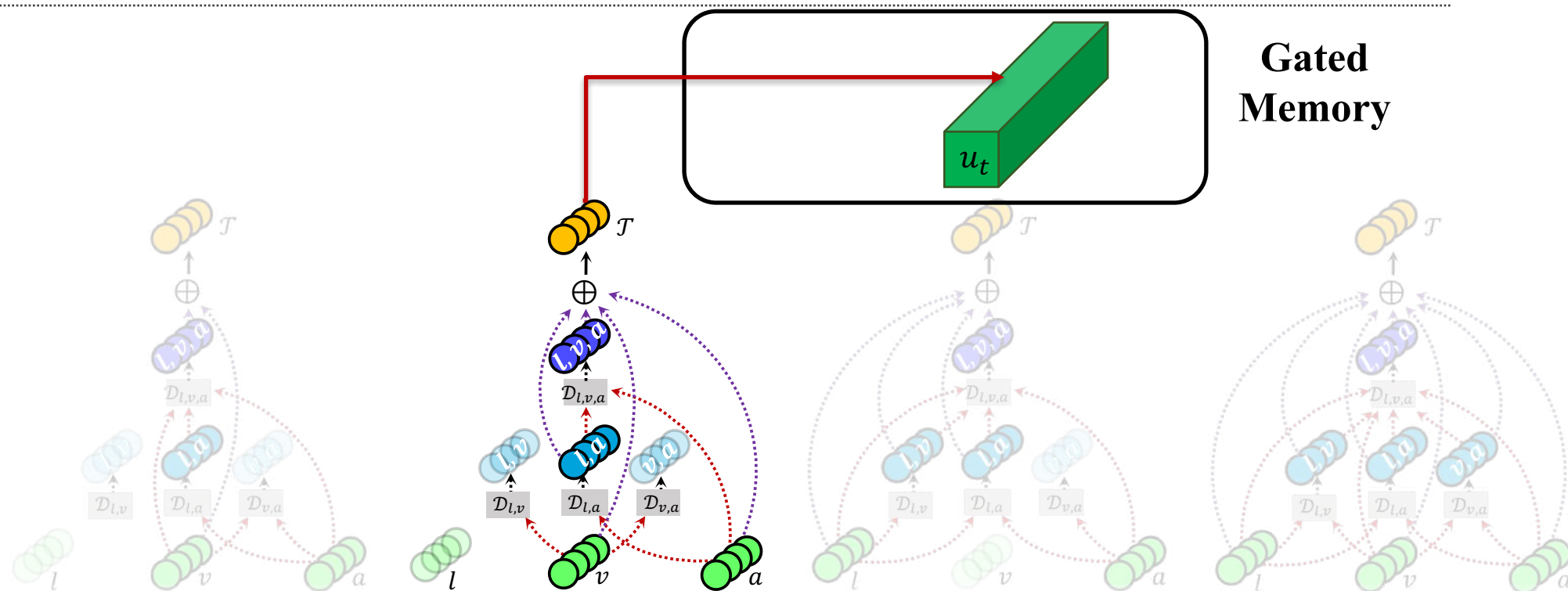
# Graph-Memory Fusion Network (Graph-MFN)

multimodal

trimodal

bimodal

unimodal



Transcript

Video clips

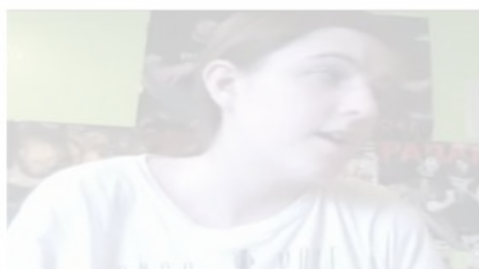
Visual gestures

*Um...*

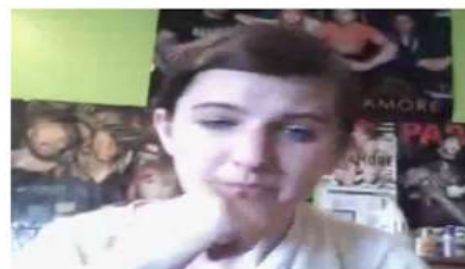
*...mm*

*this movie*

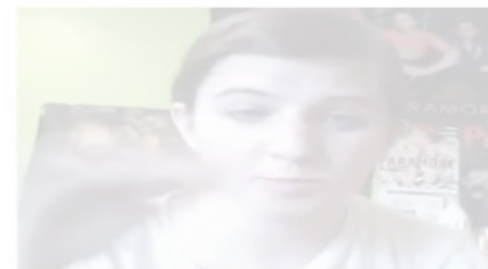
*is dumb.*



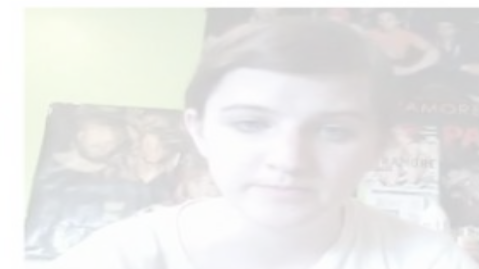
Gaze Aversion



Frown



-



Frustration



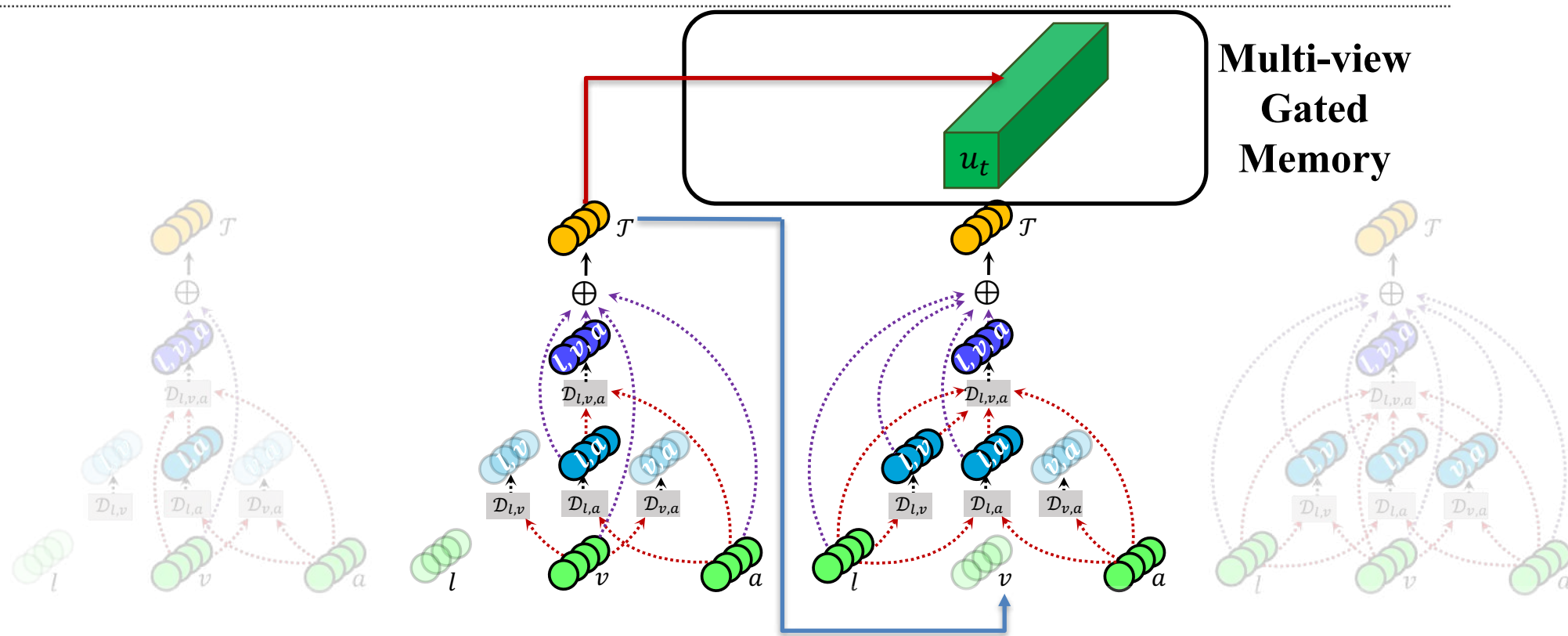
# Graph-Memory Fusion Network (Graph-MFN)

multimodal

trimodal

bimodal

unimodal



Transcript

Video clips

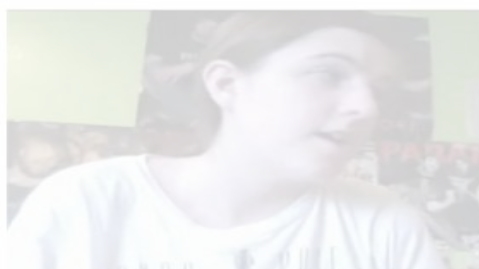
Visual gestures

*Um...*

*...mm*

*this movie*

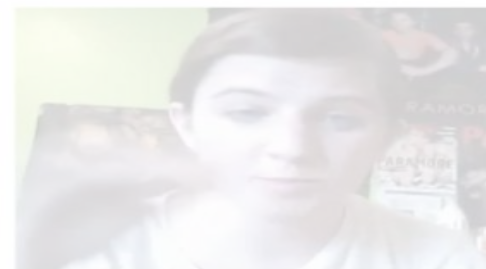
*is dumb.*



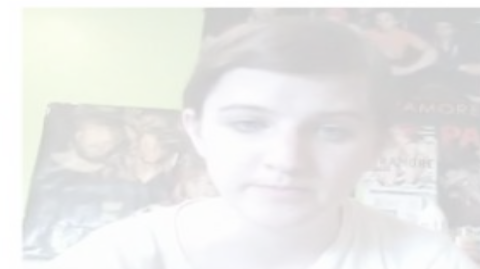
Gaze Aversion



Frown



-



Frustration



# 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

- TFN (Zadeh et al., 2017), BC-LSTM (Poria et al., 2017)

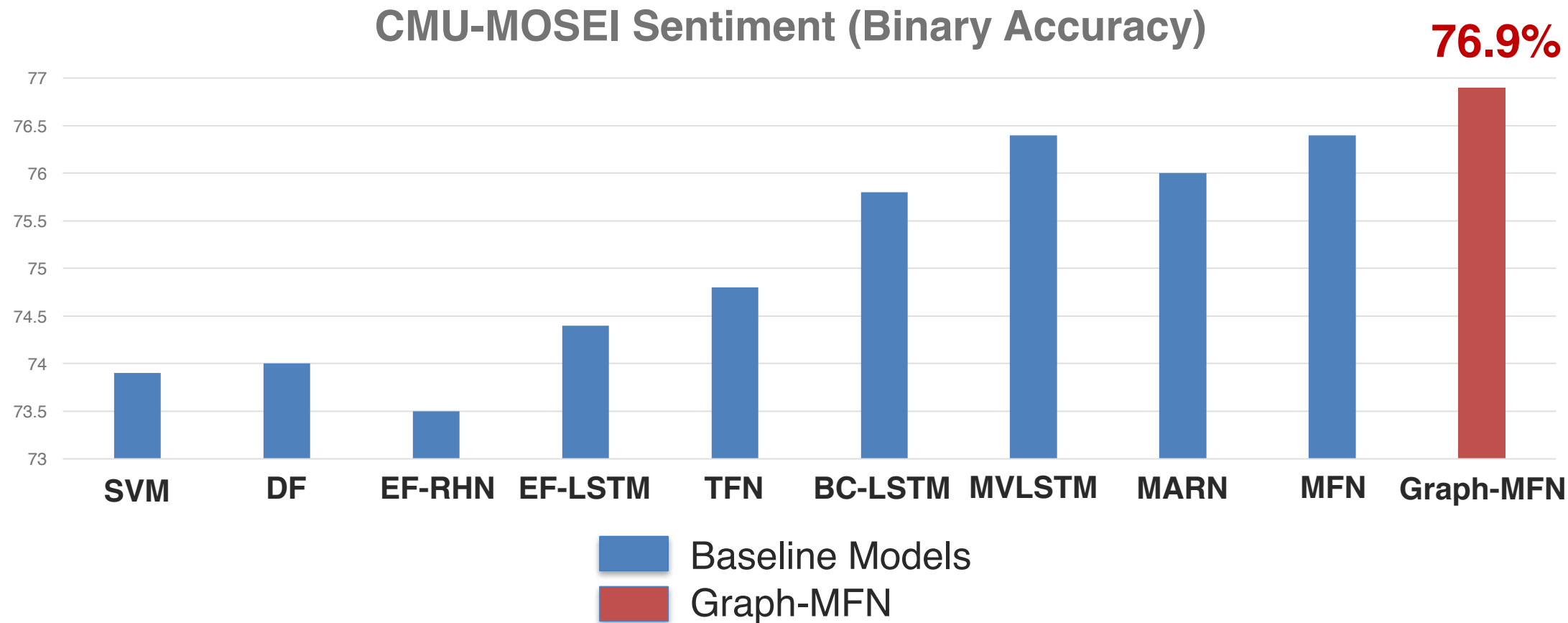
## 4. Multi-view Learning

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

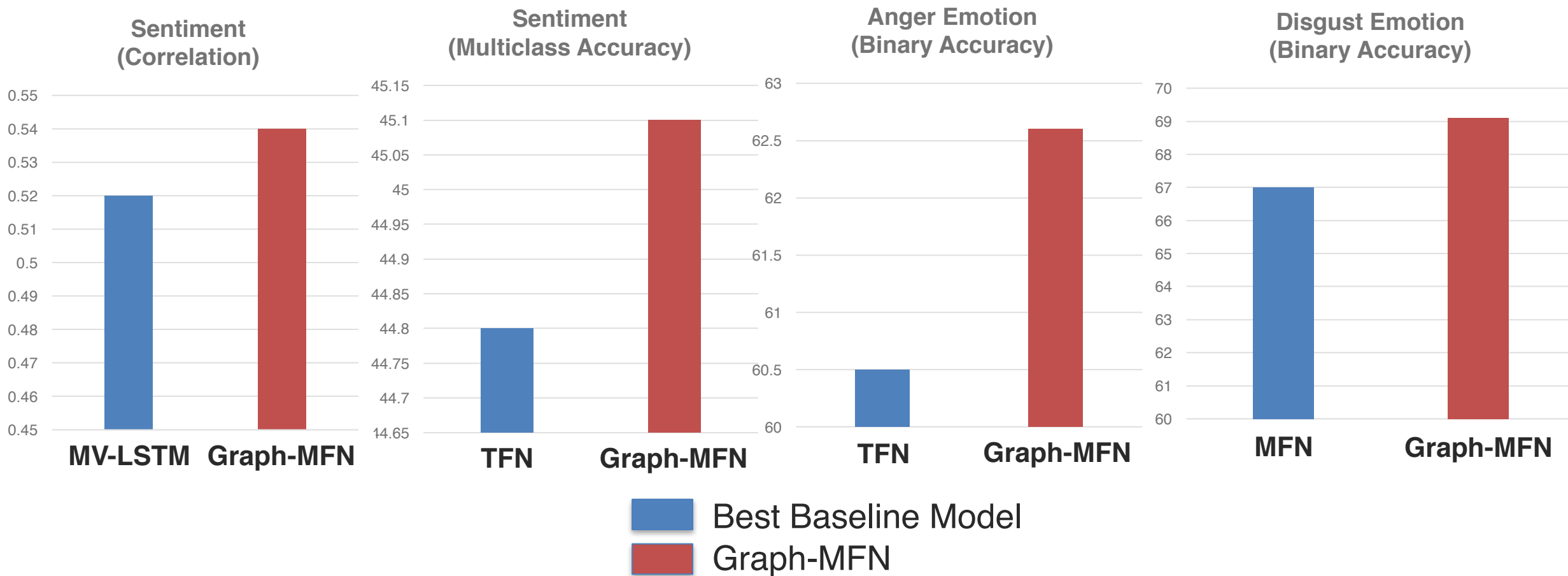
## 5. Memory-based models

- MFN (Zadeh et al., 2018)

# State-of-the-art Results



# State-of-the-art Results



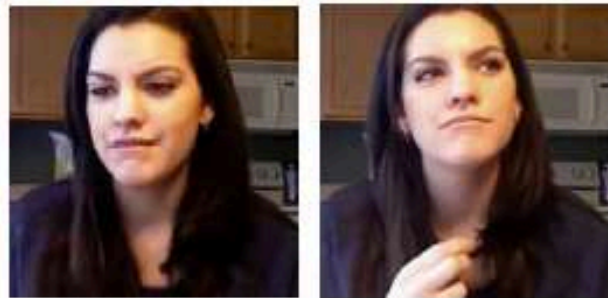
# Interpretable Fusion

---

$t = 1$   $t = T$

*And he I don't think he got mad when hah  
I don't know maybe.*

Gaze aversion



(frustrated voice)

$t = 1$   $t = T$

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

Uninformative



(angry voice)



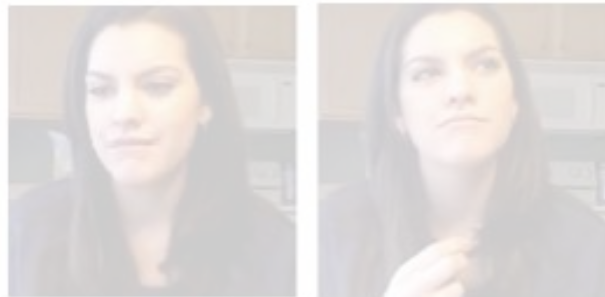
# Interpretable Fusion

---

 $t = 1$  $t = T$ 

*And he I don't think he got mad when hah  
I don't know maybe.*

Gaze aversion



(frustrated voice)

 $t = 1$  $t = T$ 

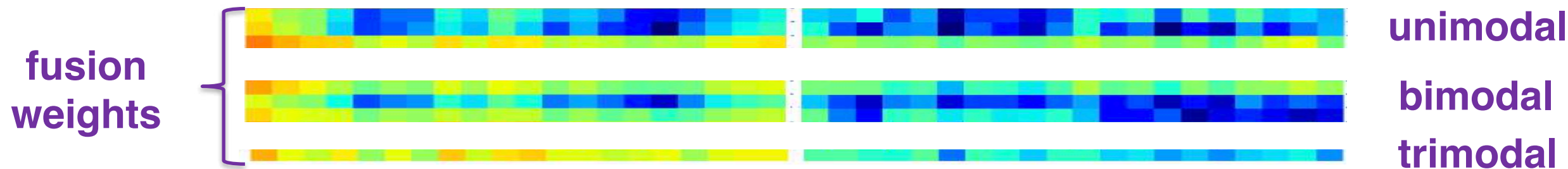
*Too much too fast, I mean we basically just  
get introduced to this character...*

Uninformative



(angry voice)

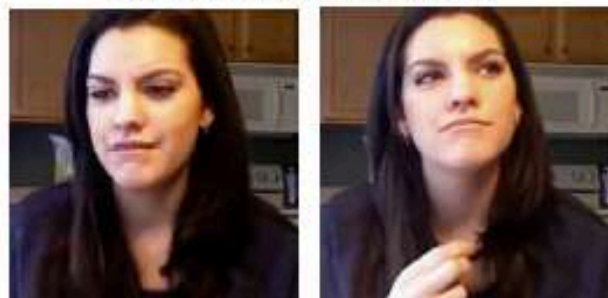
# Interpretable Fusion


 $t = 1$ 
 $t = T$ 
 $t = 1$ 
 $t = T$ 

*And he I don't think he got mad when hah  
I don't know maybe.*

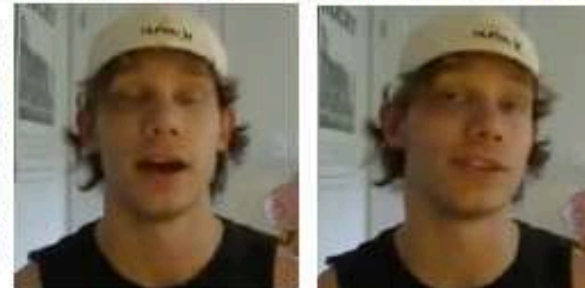
*Too much too fast, I mean we basically just  
get introduced to this character...*

Gaze aversion



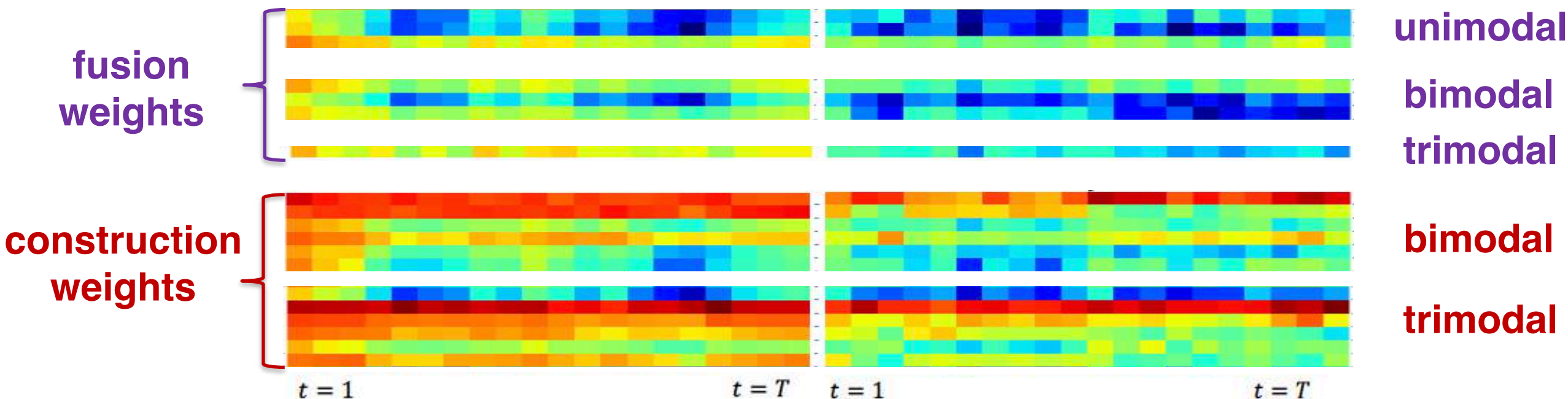
(frustrated voice)

Uninformative



(angry voice)

# Interpretable Fusion

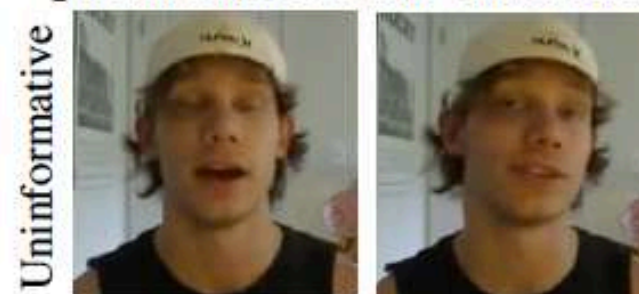


*And he I don't think he got mad when hah  
I don't know maybe.*



(frustrated voice)

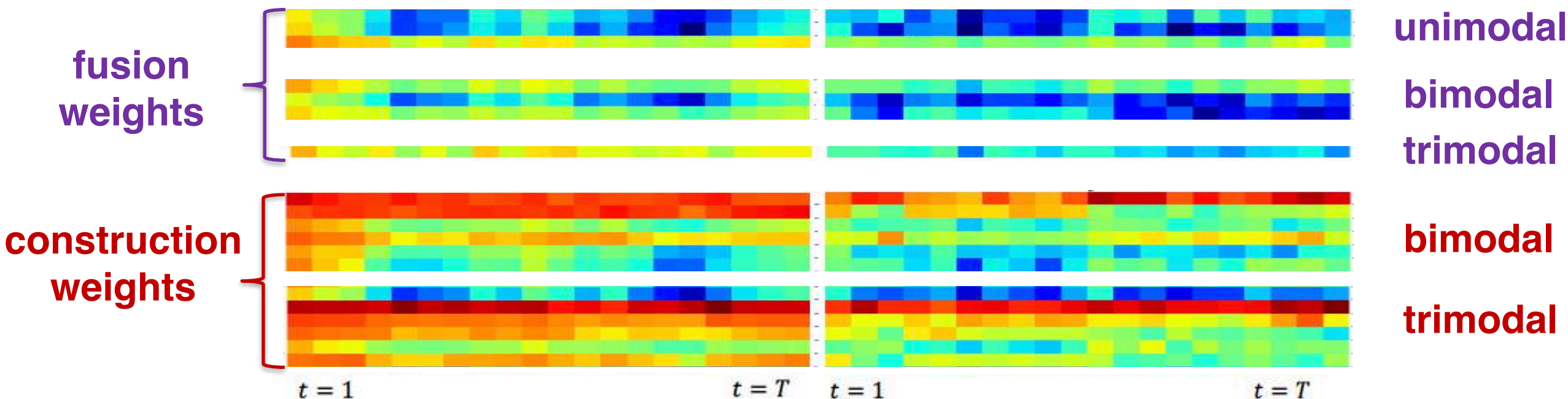
*Too much too fast, I mean we basically just  
get introduced to this character...*



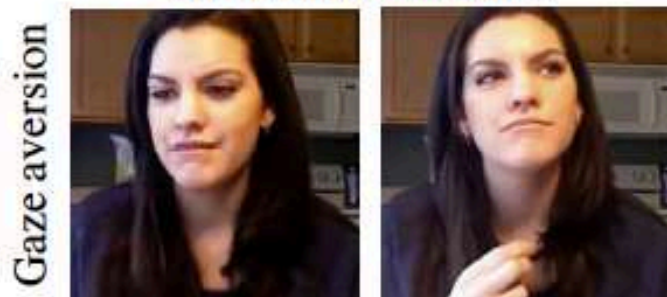
(angry voice)



# Multimodal Fusion has a Dynamic Nature



*And he I don't think he got mad when hah  
I don't know maybe.*



(frustrated voice)

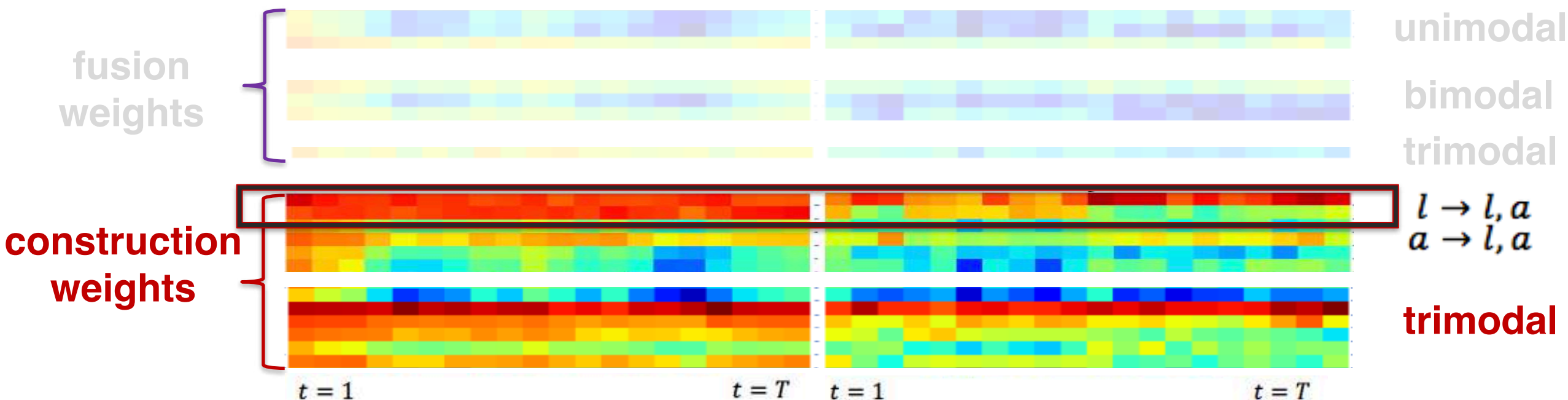
*Too much too fast, I mean we basically just  
get introduced to this character...*



(angry voice)

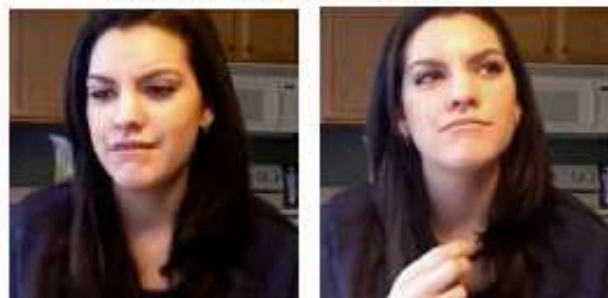


# Priors in Human Multimodal Language



*And he I don't think he got mad when hah  
I don't know maybe.*

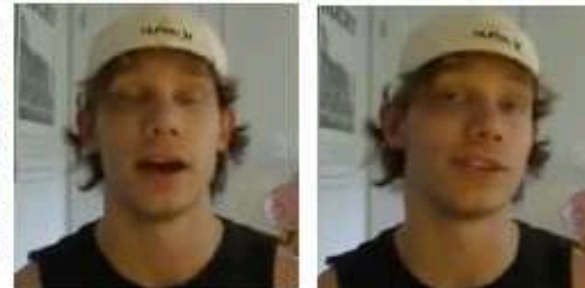
Gaze aversion



(frustrated voice)

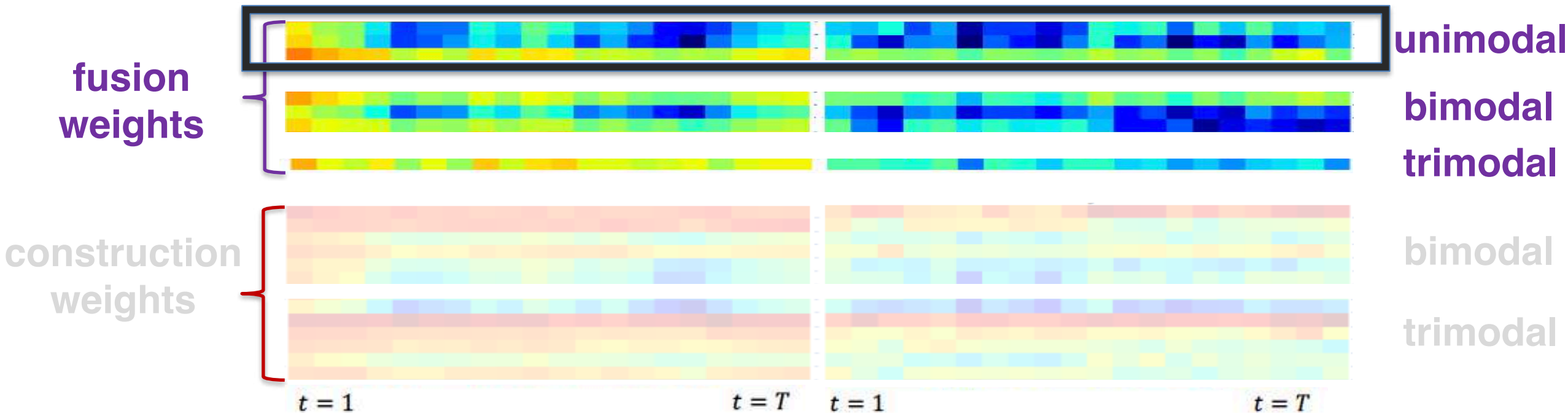
*Too much too fast, I mean we basically just  
get introduced to this character...*

Uninformative



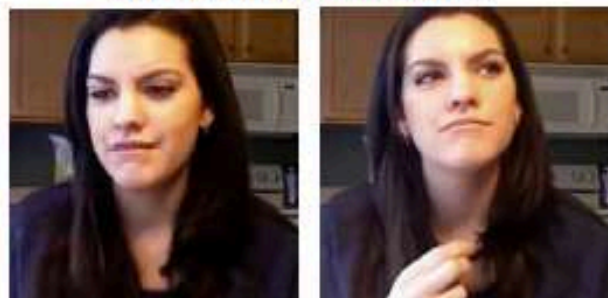
(angry voice)

# Priors in Human Multimodal Language



*And he I don't think he got mad when hah  
I don't know maybe.*

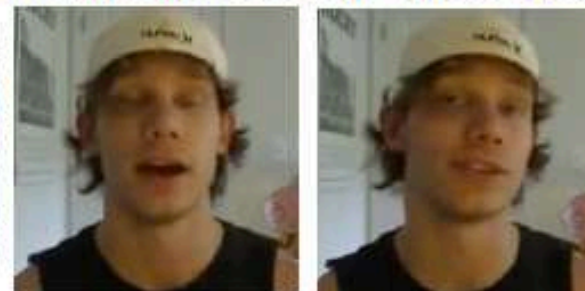
Gaze aversion



(frustrated voice)

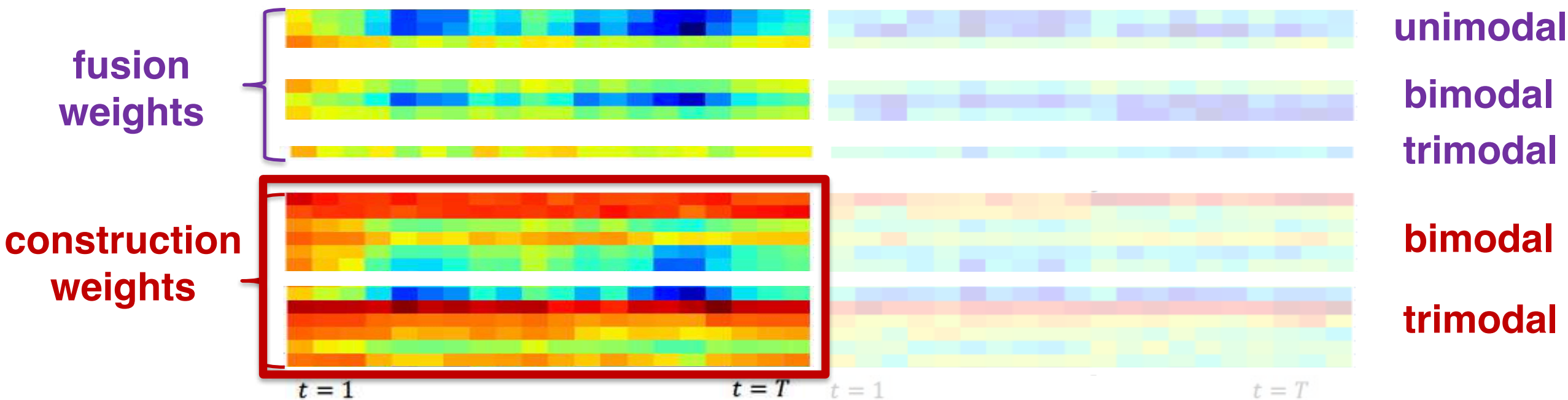
*Too much too fast, I mean we basically just  
get introduced to this character...*

Uninformative

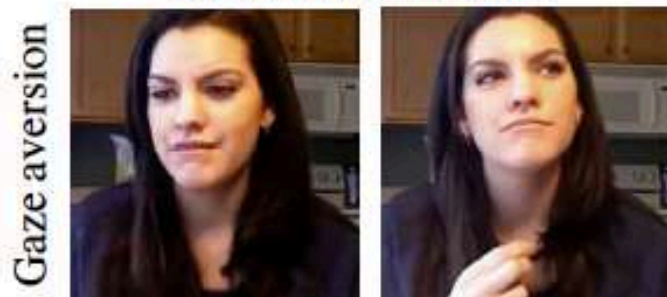


(angry voice)

# Dynamic Selection of Modalities

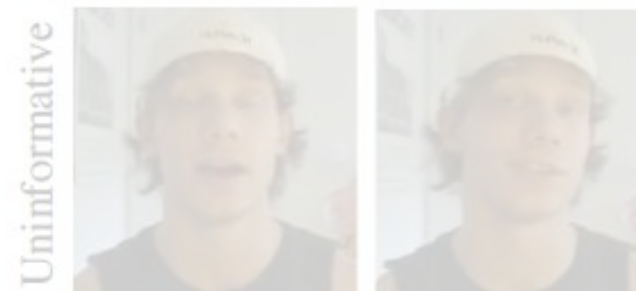


*And he I don't think he got mad when hah  
I don't know maybe.*



(frustrated voice)

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

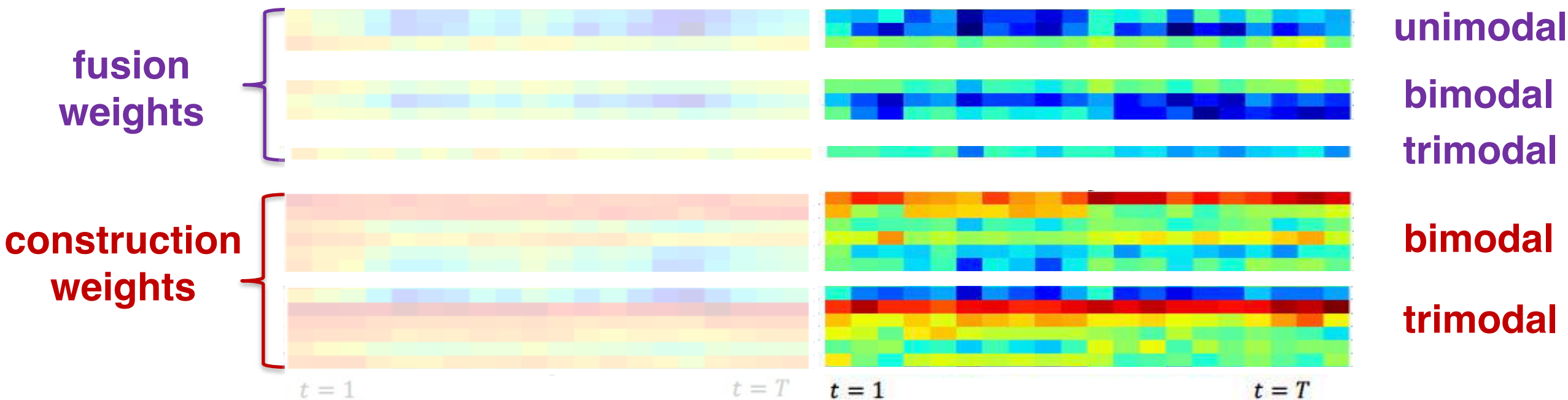


(angry voice)

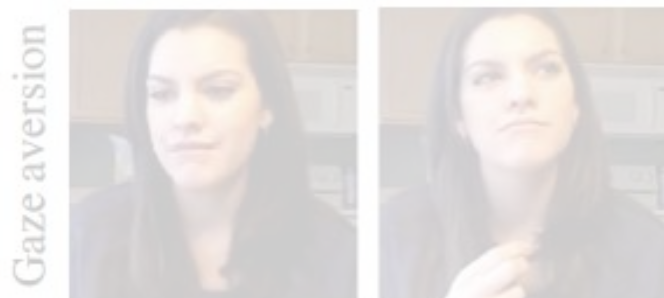
**all modalities  
are informative**



# Dynamic Selection of Modalities



*And he I don't think he got mad when hah  
I don't know maybe.*



(frustrated voice)

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



(angry voice)

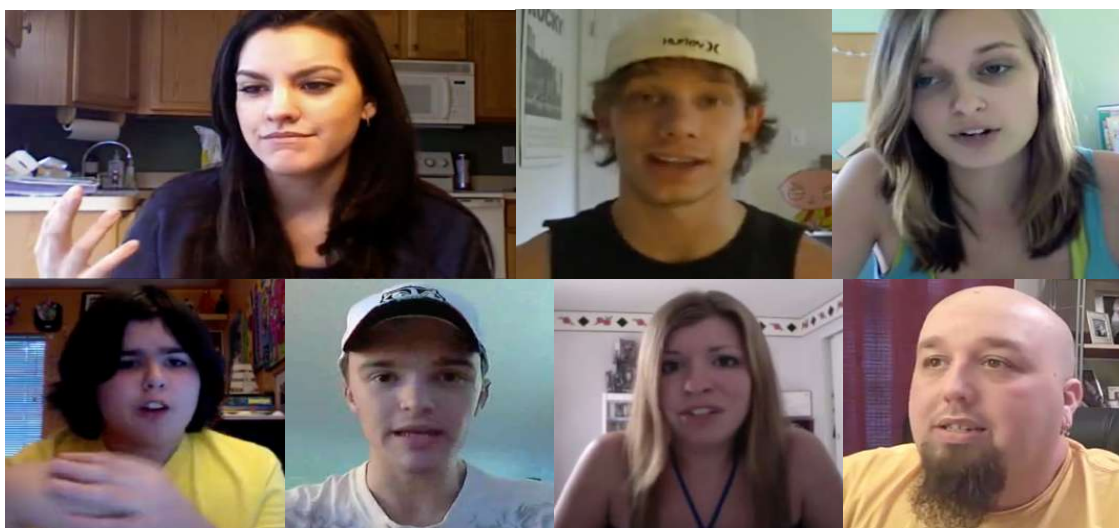
**visual modality  
uninformative**



# Computational Modeling of Multimodal Language

1

## CMU-MOSEI Dataset

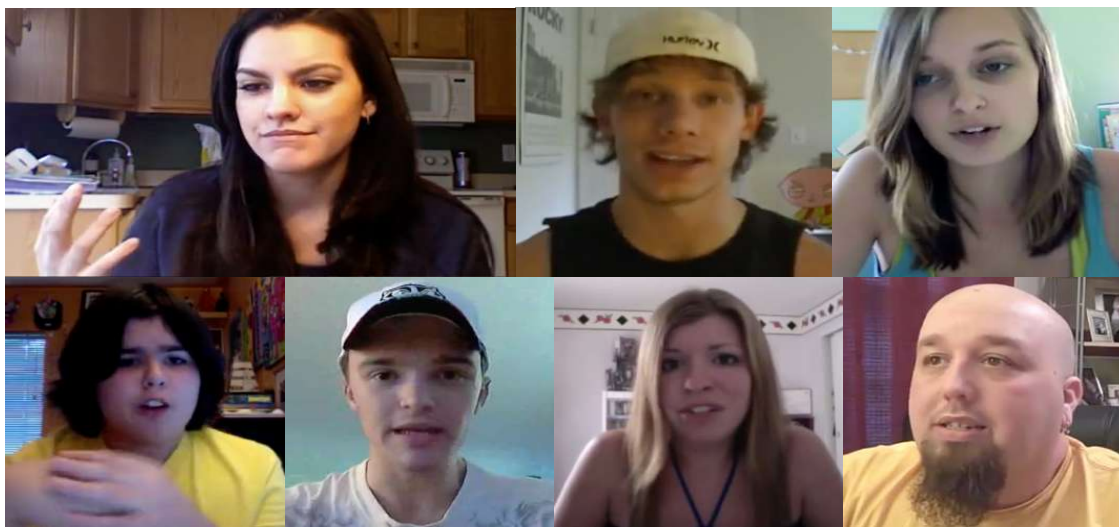


- ✓ Large-scale
- ✓ Diverse

# Computational Modeling of Multimodal Language

1

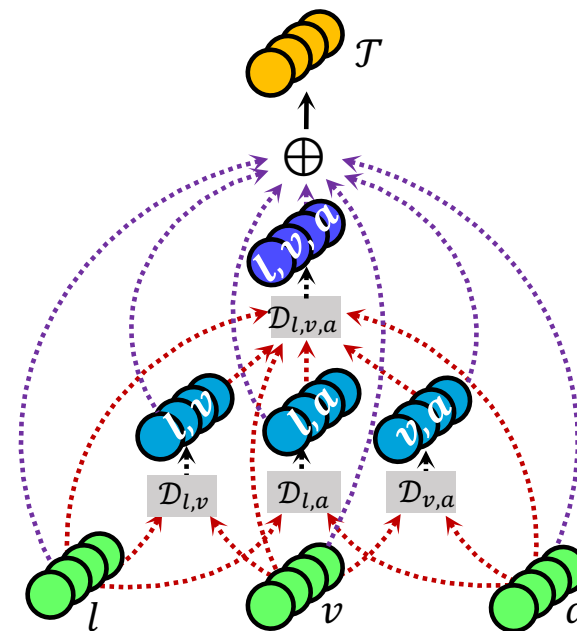
## CMU-MOSEI Dataset



- ✓ Large-scale
- ✓ Diverse

2

## Dynamic Fusion Graph



- ✓ Good Performance
- ✓ Interpretable

# The End!

**Data:** <https://github.com/A2Zadeh/CMU-MultimodalSDK>

**Website:** [www.cs.cmu.edu/~pliang](http://www.cs.cmu.edu/~pliang)

**Email:** [pliang@cs.cmu.edu](mailto:pliang@cs.cmu.edu)

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

# The End!

**Data:** <https://github.com/A2Zadeh/CMU-MultimodalSDK>

**Website:** [www.cs.cmu.edu/~pliang](http://www.cs.cmu.edu/~pliang)

**Email:** [pliang@cs.cmu.edu](mailto:pliang@cs.cmu.edu)

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

**Workshop @ 20 July 9am – 3pm, Room 217**

**First Grand Challenge and Workshop on  
Human Multimodal Language**