MTAG: Modal-Temporal Attention Graph for Unaligned Human Multimodal Language Sequences

Jianing Yang, Yongxin Wang, Ruitao Yi, Yuying Zhu, Azaan Rehman,
Amir Zadeh, Soujanya Poria, Louis-Philippe Morency

Presentation Outline

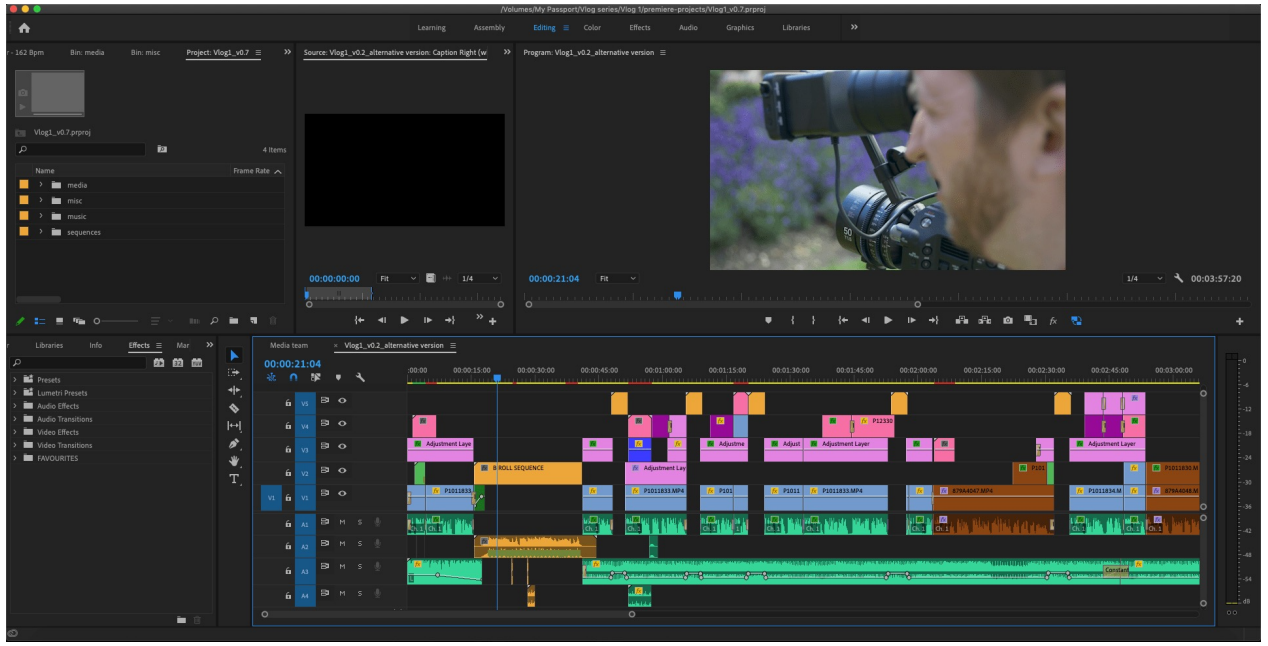
- Motivation: Challenges in multimodal language learning
- Method
- Datasets
- Results
- Analysis
- Applications

What is multimodal sequence?

- Data that is ***sequential*** and recorded in ***multiple*** channels

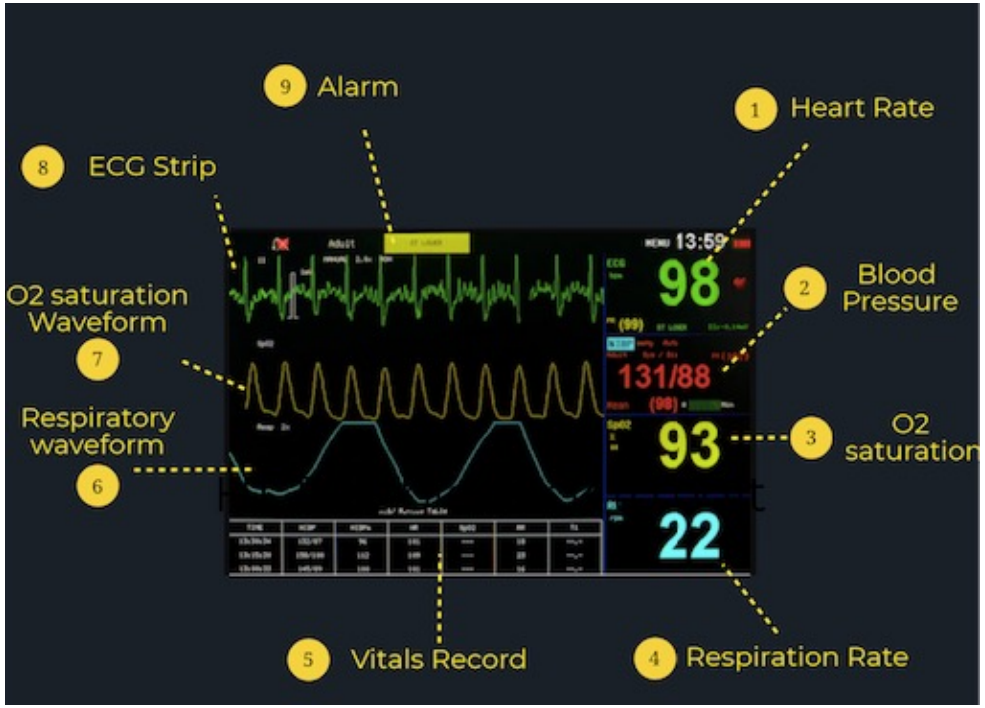
What is multimodal sequence?

- Data that is ***sequential*** and recorded in ***multiple*** channels
- Examples:
  - Video



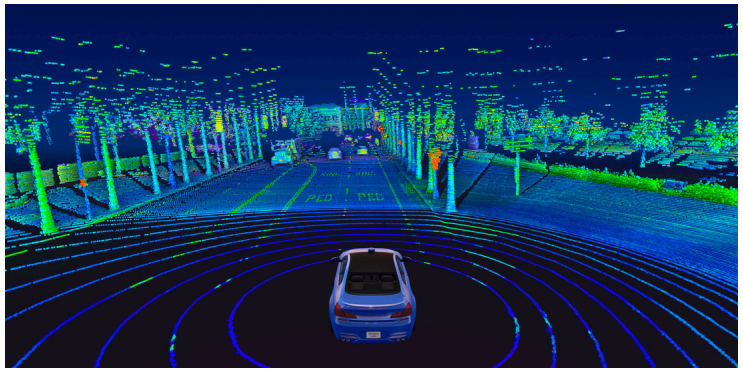
What is multimodal sequence?

- Data that is ***sequential*** and recorded in ***multiple*** channels
- Examples:
  - Video
  - Healthcare: Vital signals



What is multimodal sequence?

- Data that is ***sequential*** and recorded in ***multiple*** channels
- Examples:
  - Video
  - Healthcare: Vital signals
  - Autonomous vehicles



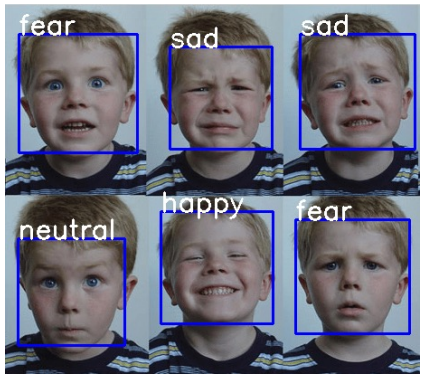
Why do we care about multimodal sequence?

Why do we care about multimodal sequence?

Abundant data

Why do we care about multimodal sequence?

Disease prediction



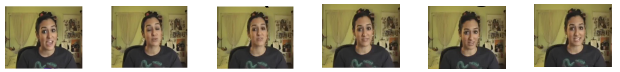
Emotion detection

Abundant data

Diverse applications

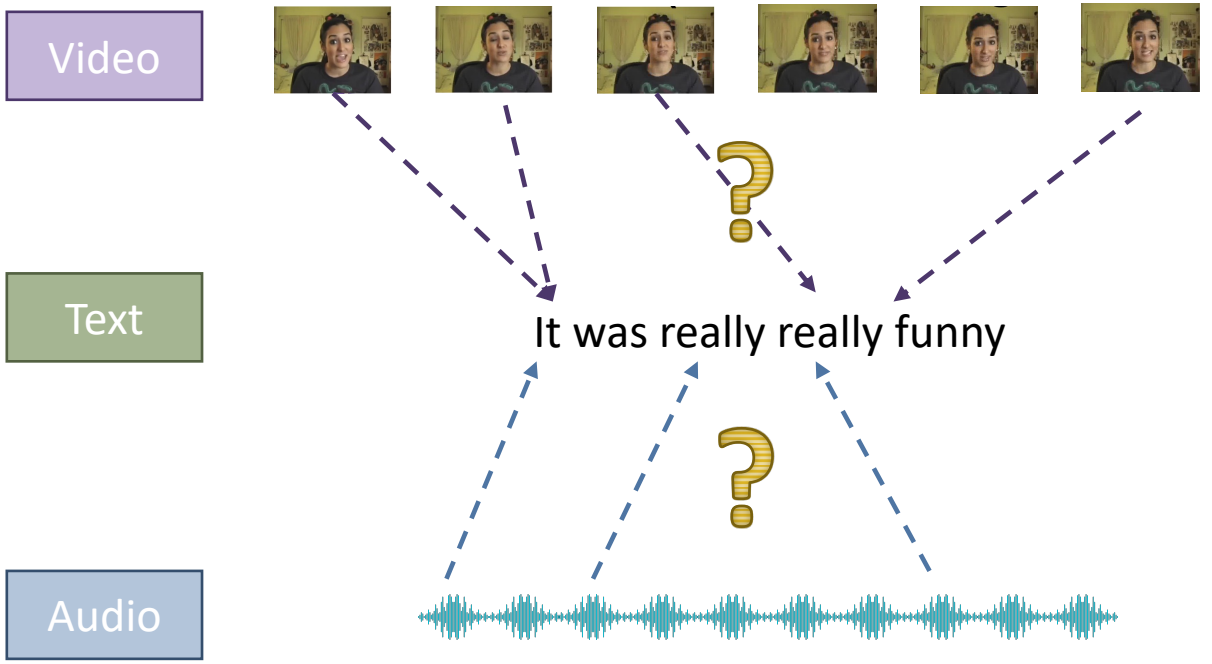
Multimodal Sequence Challenges

- Irregular lengths
- Misalignment
- Fusion of more than 2 modalities
- Long-term temporal dependency

Modality	Feature Length
Video 	6
Text: It was really really funny	5
Audio 	10

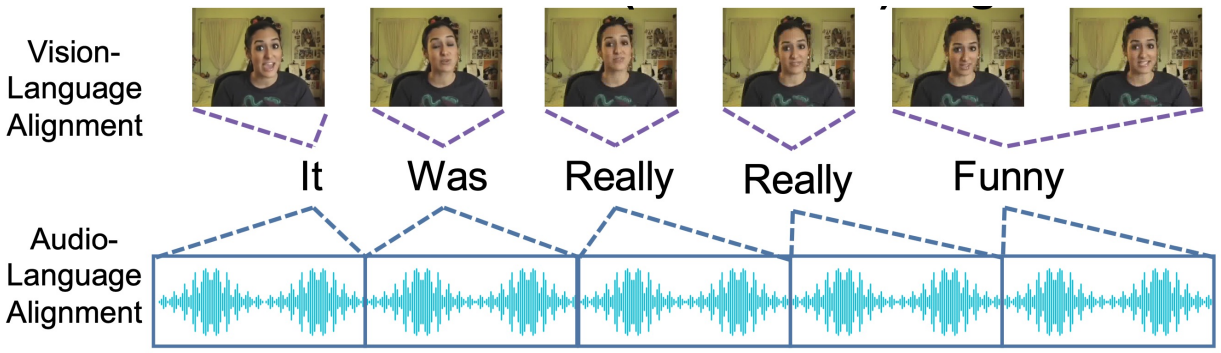
Multimodal Sequence Challenges

- Irregular lengths
- Misalignment
- Fusion of more than 2 modalities
- Long-term temporal dependency



Multimodal Sequence Challenges

- Irregular lengths
- Misalignment
- Fusion of more than 2 modalities
- Long-term temporal dependency



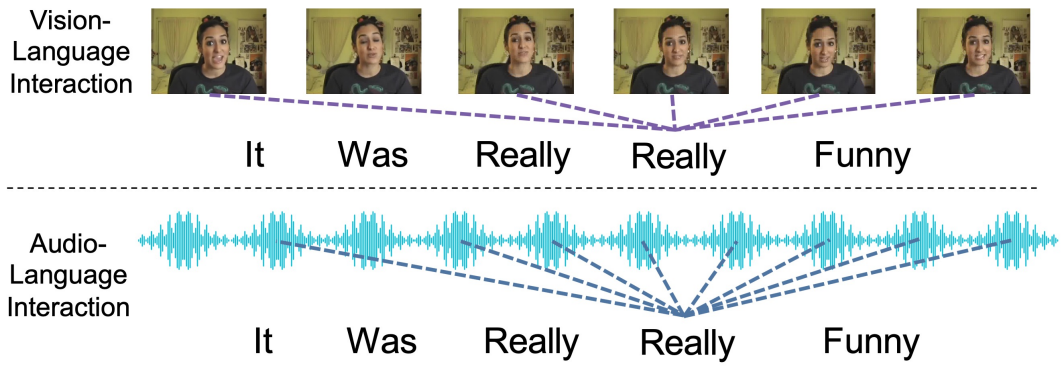
Hard (word-level) Alignment (prior approach)

Cons:

- More supervision
  - (e.g. time intervals)
- More engineering effort

Multimodal Sequence Challenges

- Irregular lengths
- Misalignment
- Fusion of more than 2 modalities
- Long-term temporal dependency



Pairwise Cross-modal Attention (Prior approach)

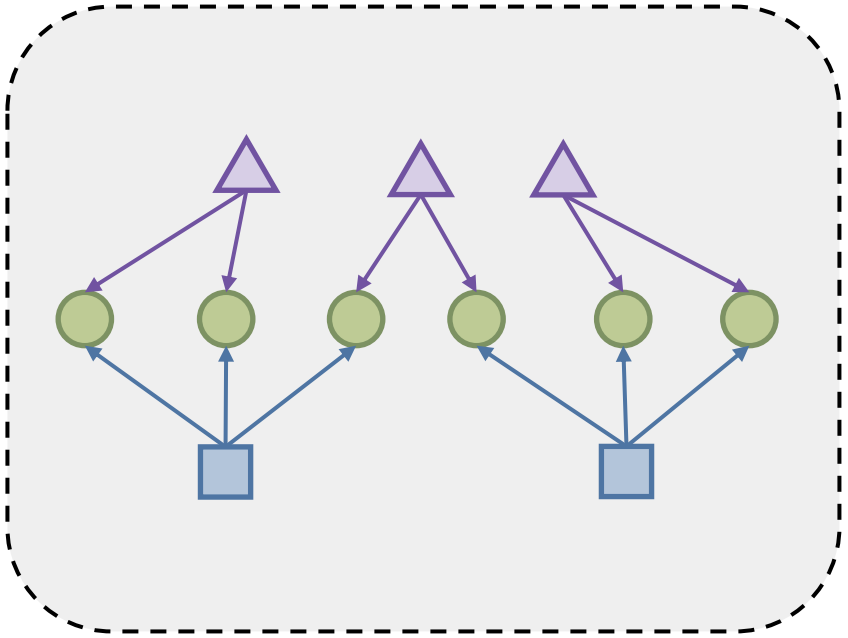
Cons:

- Only two modalities at a time
- Repeated many times for each modality pair
  --> Lots of model parameters

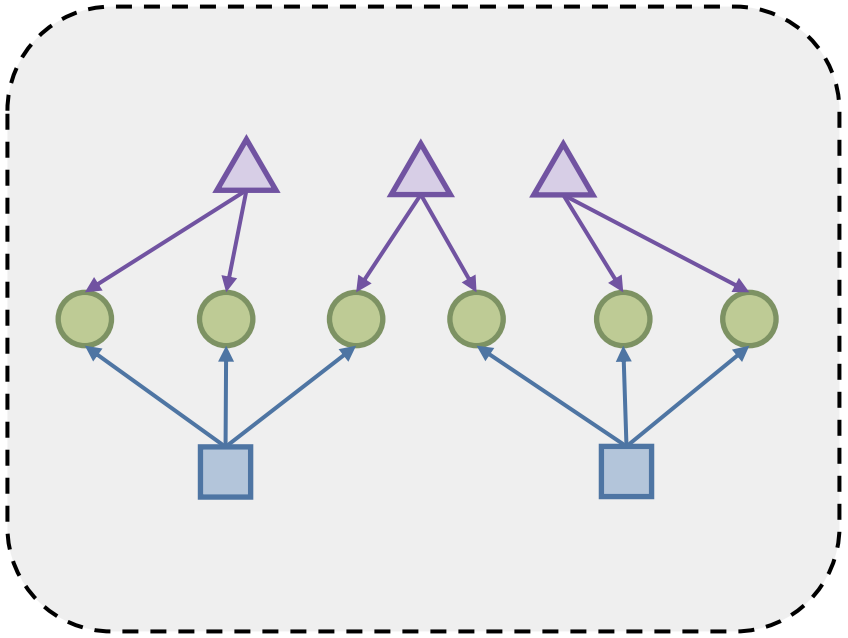
Multimodal Sequence Challenges

- Irregular lengths
- Misalignment
- Fusion of more than 2 modalities
- **Long-term** temporal dependency

Why Use Graph for Multimodal Sequence?



Why Use Graph for Multimodal Sequence?



Graph:

- Add nodes freely
- Build edges freely

Why Use Graph for Multimodal Sequence?

Challenges:

- Irregular lengths ✅
- Misalignment
- Fusion of more than 2 modalities
- Long-term temporal dependency

Graph:

- Add nodes freely
- Build edges freely

Why Use Graph for Multimodal Sequence?

Challenges:

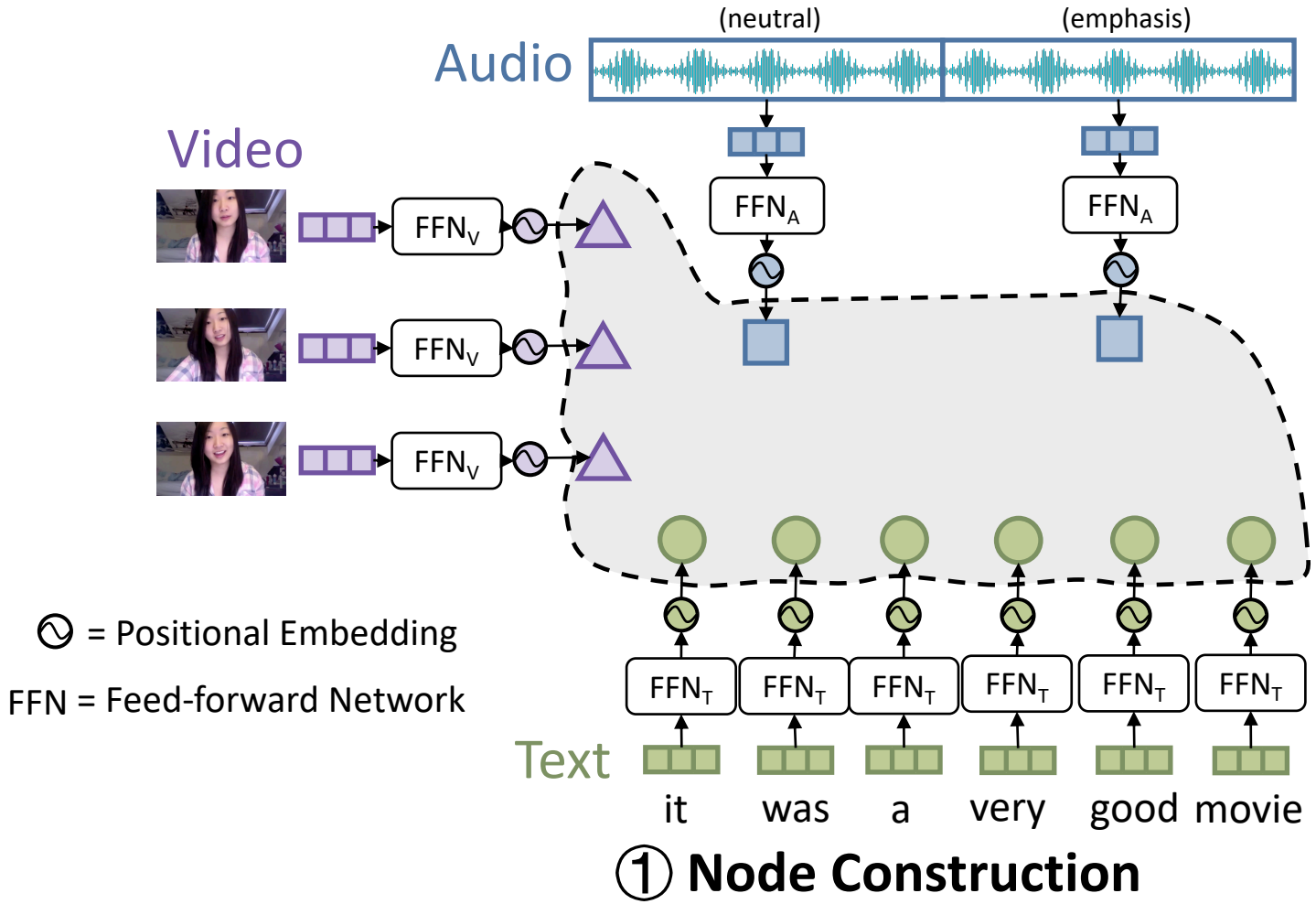
- Irregular lengths
- Misalignment ✅
- Fusion of more than 2 modalities ✅
- Long-term temporal dependency ✅

Graph:

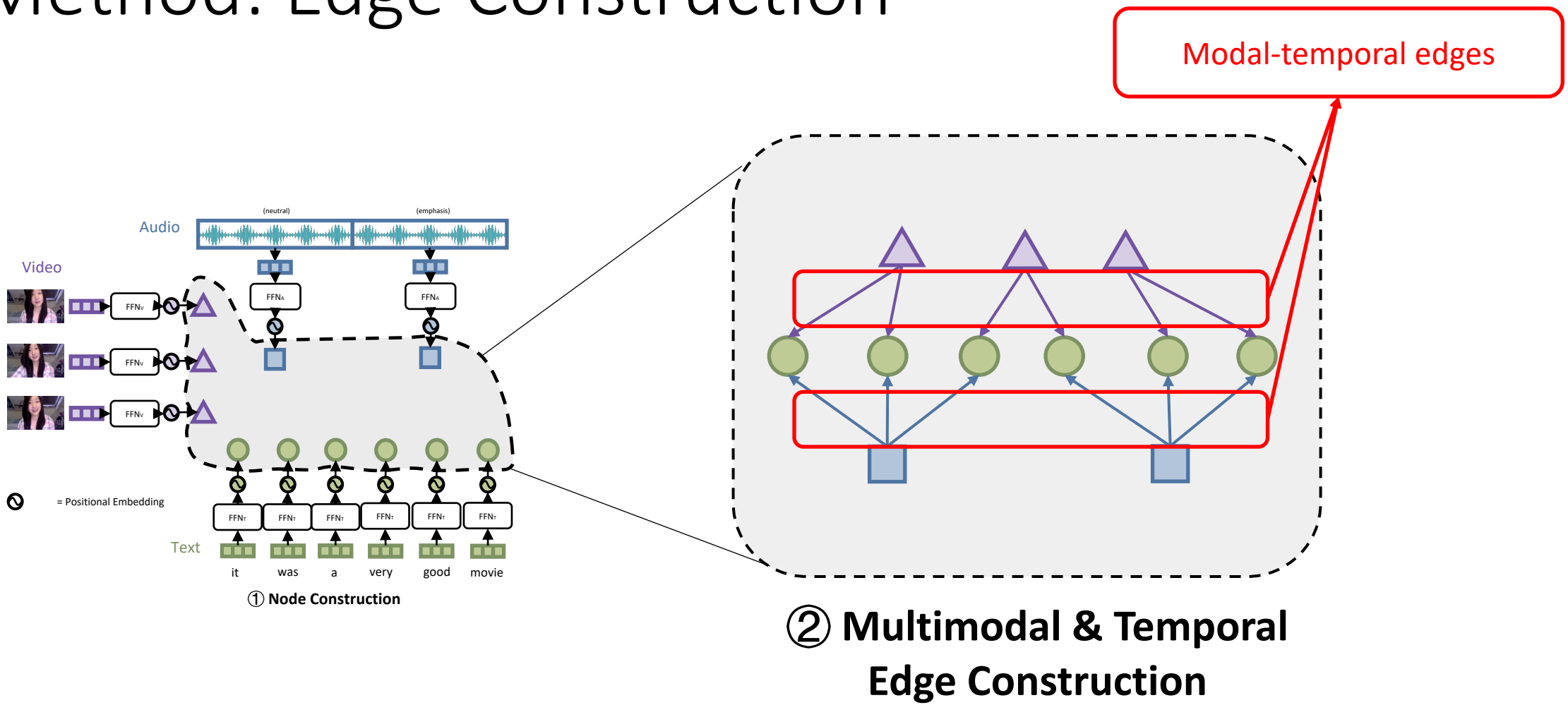
- Add nodes freely
- Build edges freely

Method

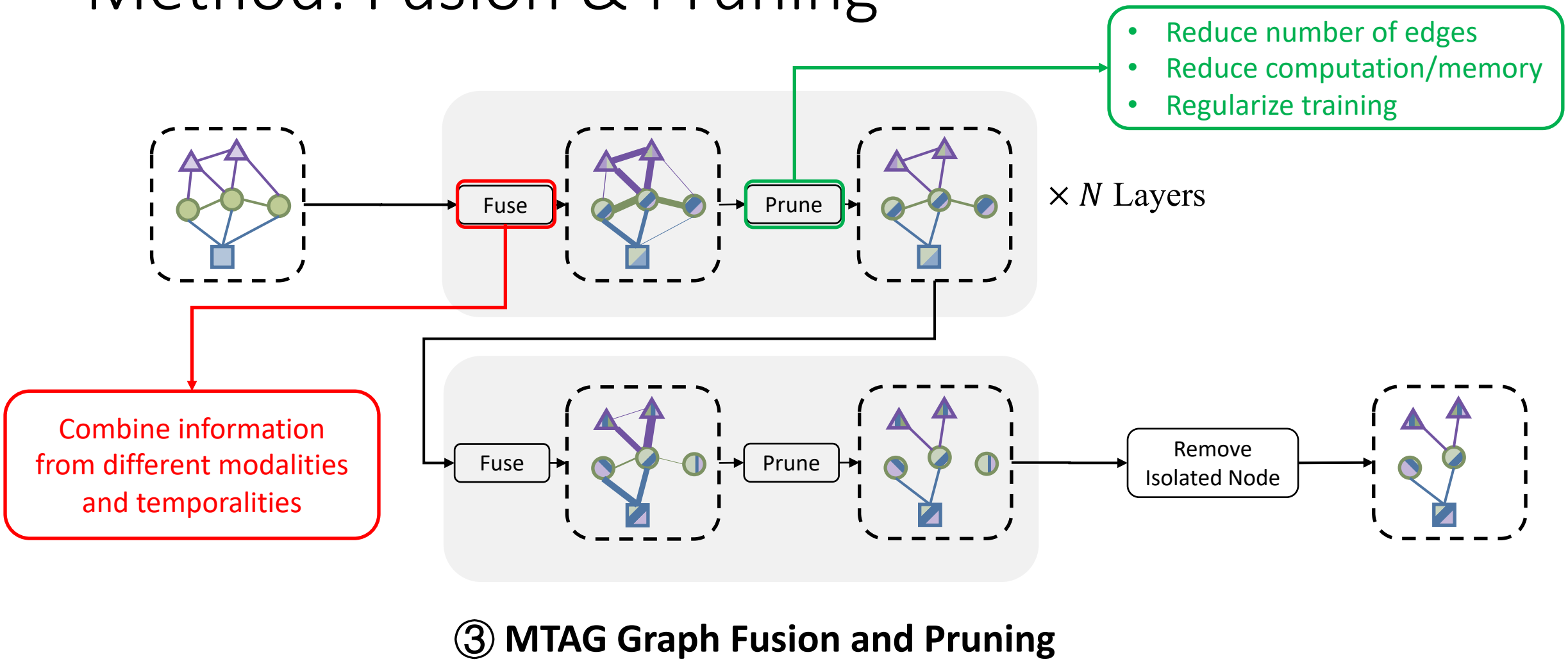
Method: Node Construction



Method: Edge Construction

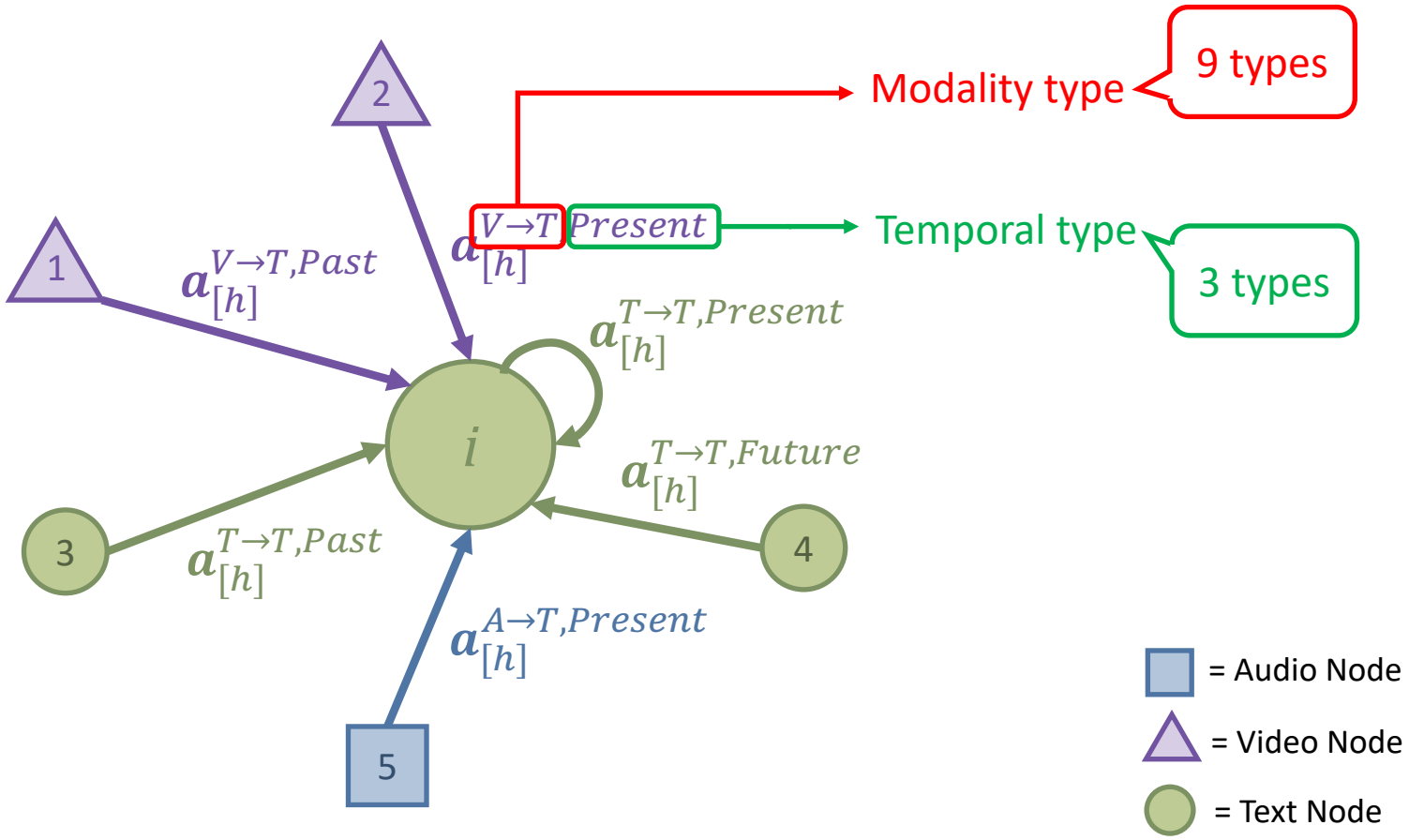


Method: Fusion & Pruning



③ MTAG Graph Fusion and Pruning

Method: Close-up of Modal-Temporal Attention

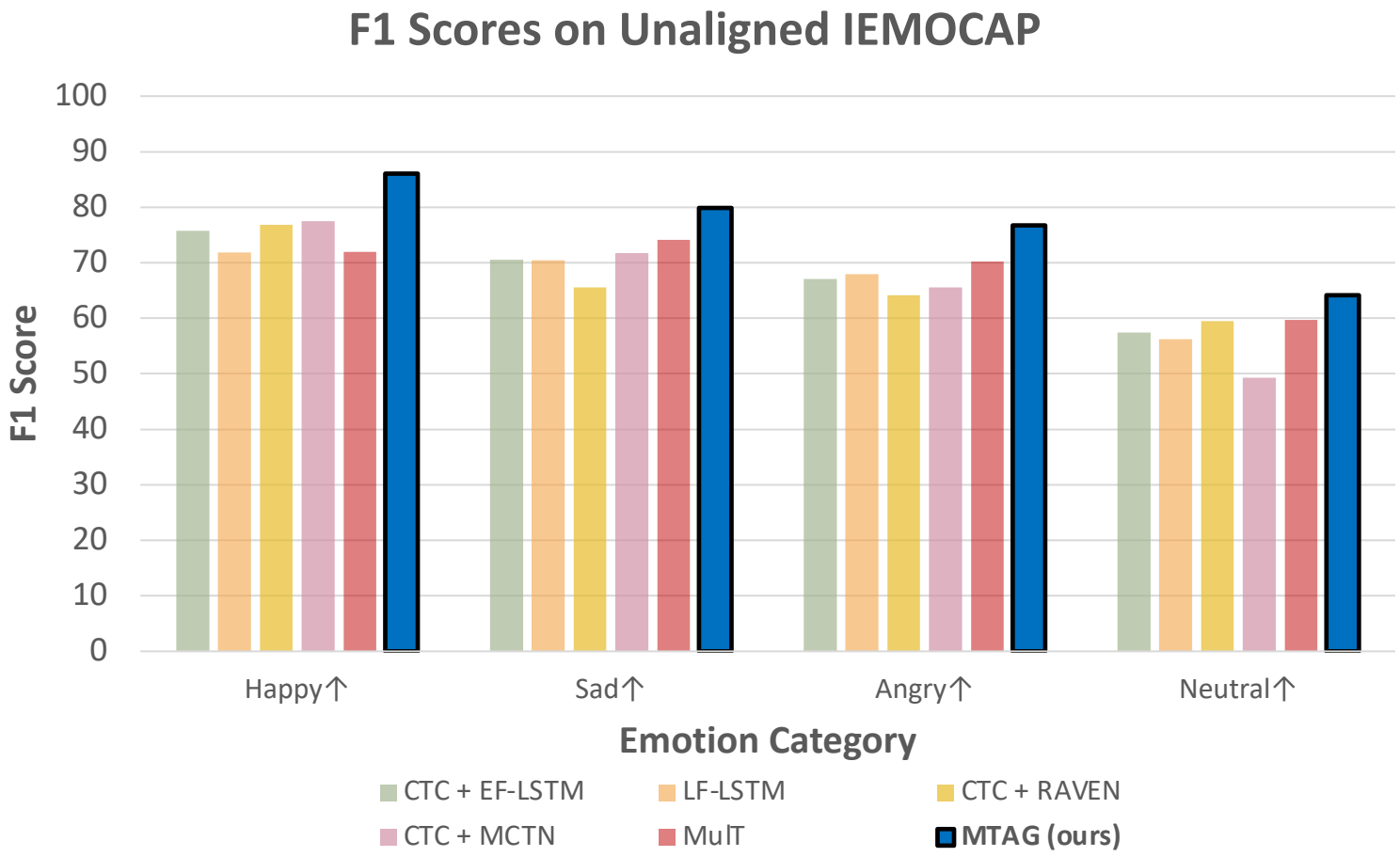


Datasets

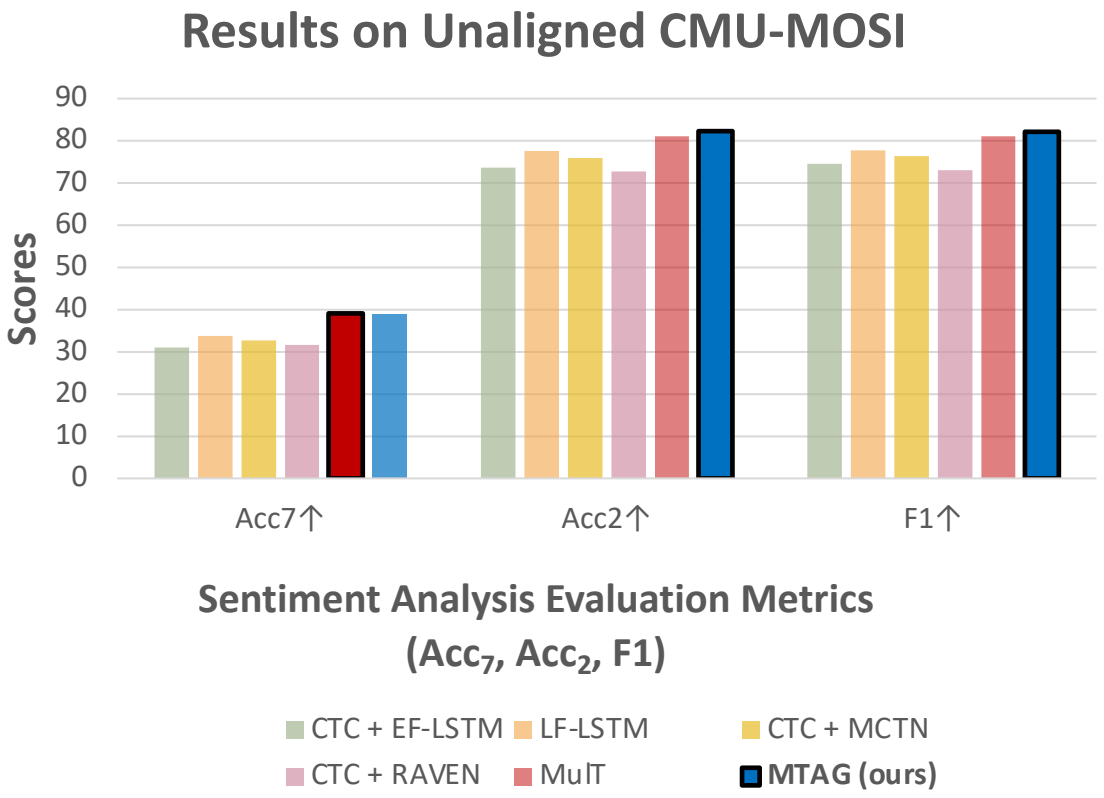
- IEMOCAP: Video Emotion Classification
  - Happy, sad, angry, neutral, etc.

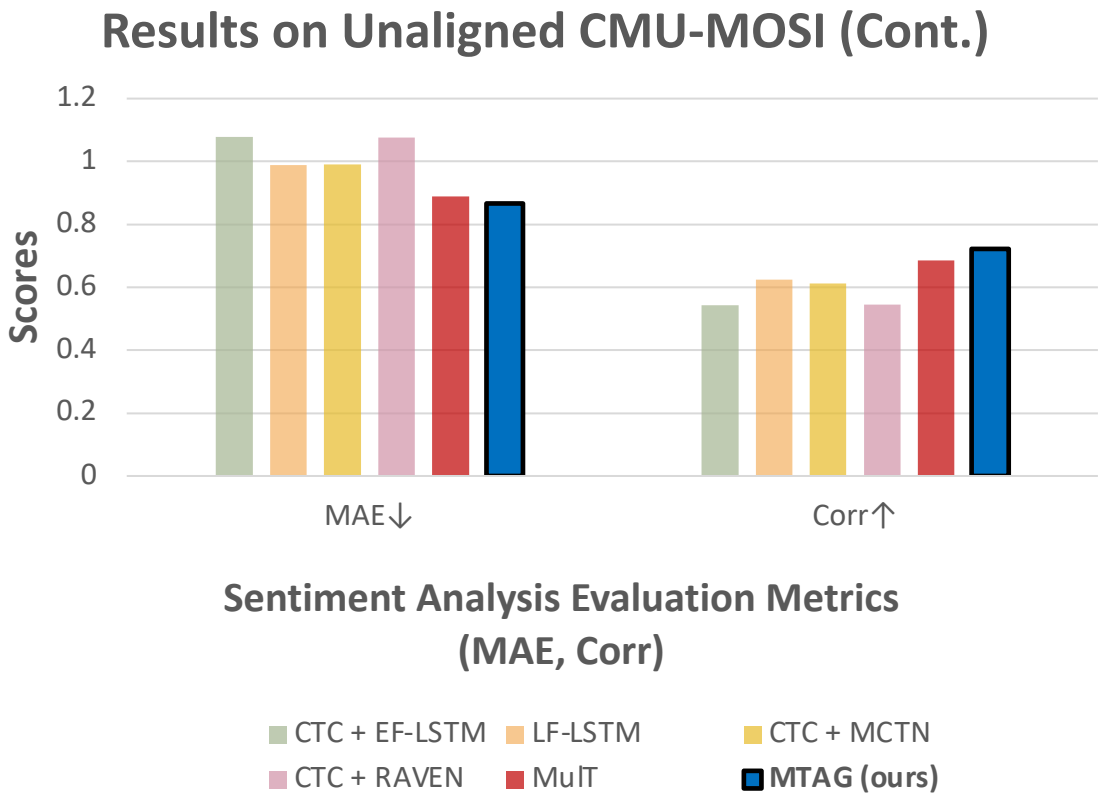
- CMU-MOSI: Video Sentiment Analysis
  - Sentiment score $\in$ [-3, +3]

Results: IEMOCAP



Results: CMU-MOSI





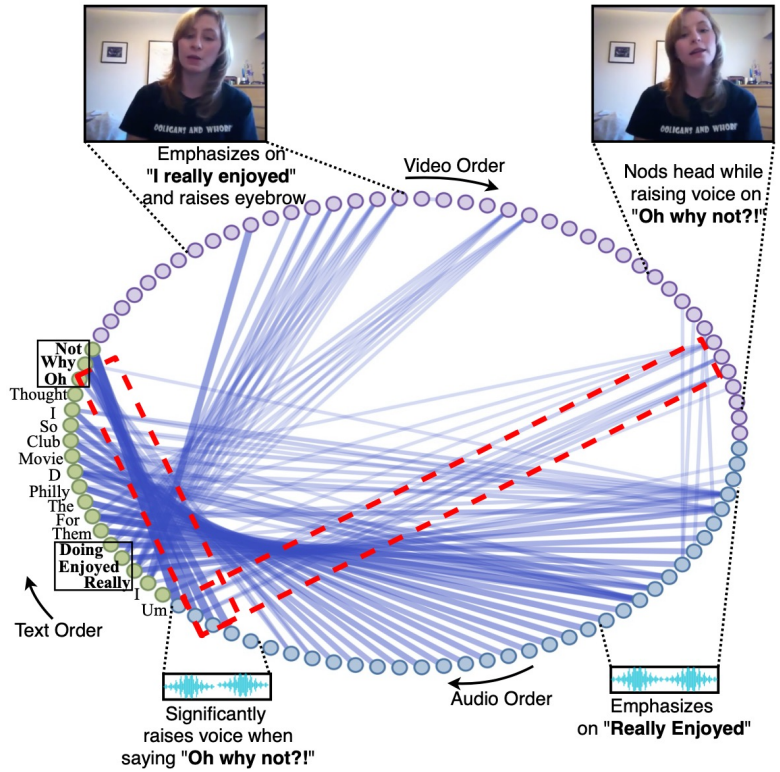
Acc_7 = 7-way classification accuracy
Acc_2 = 2-way (pos/neg) classification accuracy

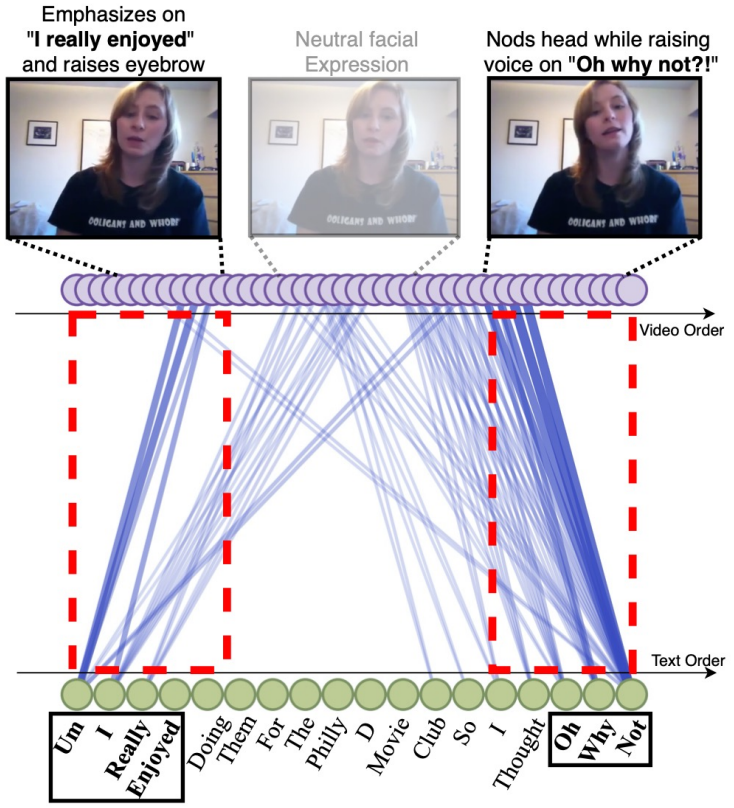
Results: Parameter Efficiency

Model	# Parameters
MulT (previous SOTA)	2.24 M
MTAG (ours)	**0.14 M**

Table 4: Number of model parameters (M = Million).

Qualitative Analysis





(a) Text-to-vision edge attention weights.

Ablation Study

Ablation	$Acc_2 \uparrow$	F1 ↑	MAE ↓
Edge Types			
No Edge Types	82.4	82.5	0.937
Multimodal Edges Only	85.6	85.7	**0.859**
Temporal Edges Only	85.2	85.2	0.887
Pruning			
Random Pruning Keep 80%	75.5	74.5	1.080
No Pruning	84.7	84.7	0.908
Modalities			
Language Only	81.5	81.4	0.911
Vision Only	57.0	57.1	1.41
Audio Only	58.1	58.1	1.37
Vision, Audio	62.0	59.2	1.360
Language, Audio	85.9	85.7	0.915
Language, Vision	86.6	86.6	0.896
Full Model, All Modalities	**87.0**	**87.0**	**0.859**

Finding 1:

Adding modality and temporal specific edges improves performance

Ablation Study

Ablation	$Acc_2 \uparrow$	F1 ↑	MAE ↓
Edge Types			
No Edge Types	82.4	82.5	0.937
Multimodal Edges Only	85.6	85.7	**0.859**
Temporal Edges Only	85.2	85.2	0.887
Pruning			
Random Pruning Keep 80%	75.5	74.5	1.080
No Pruning	84.7	84.7	0.908
Modalities			
Language Only	81.5	81.4	0.911
Vision Only	57.0	57.1	1.41
Audio Only	58.1	58.1	1.37
Vision, Audio	62.0	59.2	1.360
Language, Audio	85.9	85.7	0.915
Language, Vision	86.6	86.6	0.896
Full Model, All Modalities	**87.0**	**87.0**	**0.859**

Finding 2:

Top-K% pruning improves performance;
Random pruning decreases performance

Ablation Study

Ablation	$Acc_2 \uparrow$	F1 ↑	MAE ↓
Edge Types			
No Edge Types	82.4	82.5	0.937
Multimodal Edges Only	85.6	85.7	**0.859**
Temporal Edges Only	85.2	85.2	0.887
Pruning			
Random Pruning Keep 80%	75.5	74.5	1.080
No Pruning	84.7	84.7	0.908
Modalities			
Language Only	81.5	81.4	0.911
Vision Only	57.0	57.1	1.41
Audio Only	58.1	58.1	1.37
Vision, Audio	62.0	59.2	1.360
Language, Audio	85.9	85.7	0.915
Language, Vision	86.6	86.6	0.896
Full Model, All Modalities	**87.0**	**87.0**	**0.859**

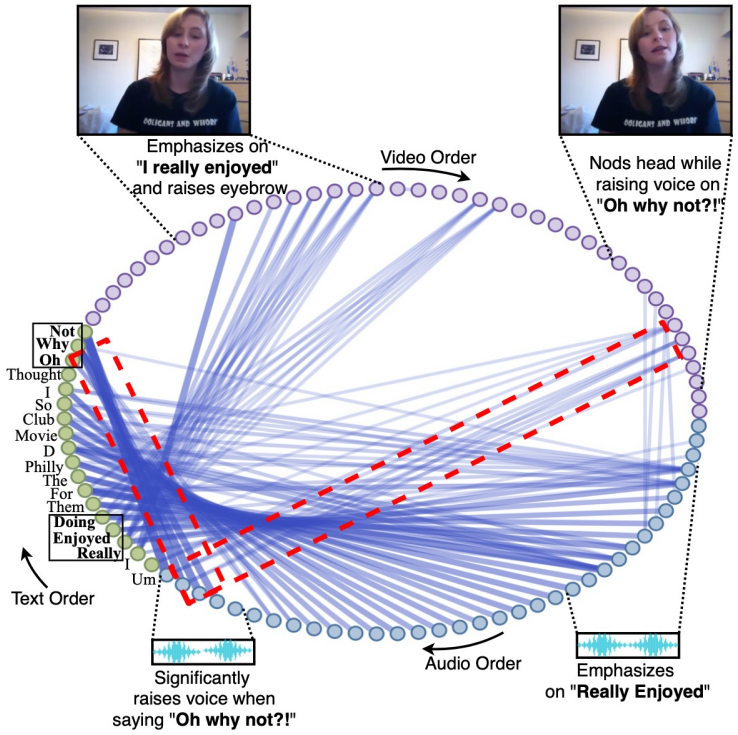
Finding 3:
Language is the most helpful modality

MTAG increases its performance as more modalities are provided

Summary of Contributions

- A new pipeline to model unaligned multimodal sequence data
- A new graph convolution operation called MTAG fusion
- State-of-the-art results on two datasets
- Much fewer model parameters than previous SOTA

MTAG: Modal-Temporal Attention Graph for Unaligned Human Multimodal Language Sequences



Paper

Code