Since 2004
UET
ĐẠI HỌC CÔNG NGHỆ, ĐHQGHN
VNU-University of Engineering and Technology

Since 1906
VNU
ĐẠI HỌC QUỐC GIA HÀ NỘI
Vietnam National University, Hanoi

THE 4TH SoICT's INTERNATIONAL SUMMER SCHOOL ON AGENTIC AI

Multimodal Learning: Foundations, Emerging Trends and Its Applications in Emotion Recognition

Duc-Trong Le

VNU University of Engineering and Technology, Hanoi, Vietnam

Hanoi, 09/2025

About Presenter

- **Education**: Ph.D in Computer Science, Singapore Management University, Singapore (2019)
- **Present Positions**:
  - Head of Computer Science Department, VNU-UET-FIT
  - Head of MORAI Research Group@VNU-UET
  - Co-Head of L2R Research Group@VNU-UET
  - Leader@ AI for Health Research Group, VNU-UET
  - Mentor @ FPT Software AI Residency Program
- **Research Interests**: *Reliable AI, Multimodal Learning, Recommendation Systems, Medical Image Analysis*
- Website: https://www.trongld.com
- Teams: 3 PhD students (1 more in 10/2025), 10+ MSc students

Duc-Trong Le

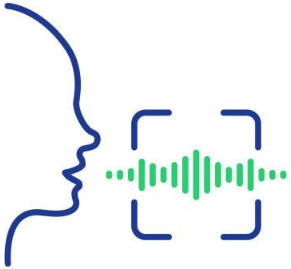
Google Scholar

Outline

- Foundations of Multimodal Learning
  - What is Multimodal?
  - Multimodal Machine Learning
  - Core Research Challenges
- Emerging Trends in Multimodal Learning
- Applications in Multimodal Emotion Recognition

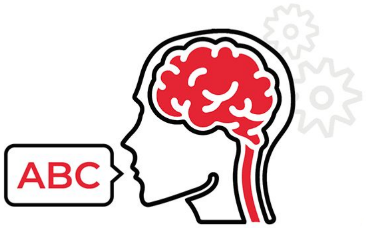
1. Foundations of Multimodal Learning

From Human Capabilities to AI Capabilities



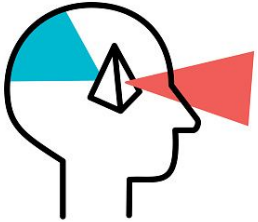
Speech Recognition & Processing

(Sound/Acoustic)



Natural Language Comprehension

(Text/Language)



Visual Perception & Analytics

(Visual)

Thinking, Reasoning & Decision-making

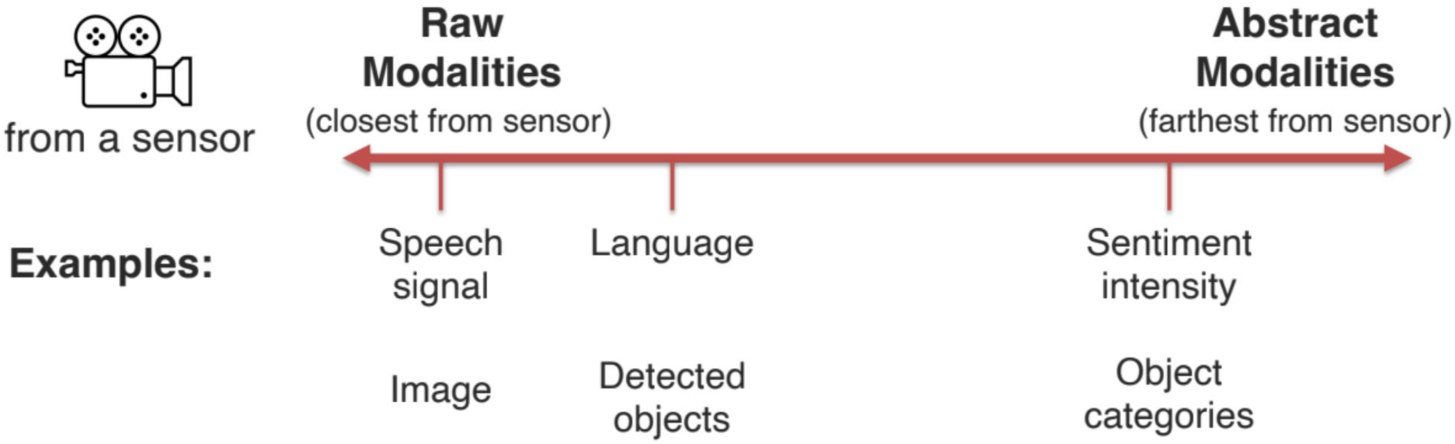
(Brain Signal)

Multimodal Behaviors and Signals

- **Language**
  - Lexicon:
    - Words
  - Syntax
    - Part-of-speech
    - Dependencies
  - Pragmatics
    - Discourse acts
- **Acoustic**
  - Prosody:
    - Intonation
    - Voice quality
  - Vocal expressions
    - Laughter, moans
- **Visual**
  - Gestures:
    - Head gestures
    - Eye gestures
    - Arm gestures
  - Body language
    - Body posture
    - Proxemics
  - Eye contact:
    - Head gaze
    - Eye gaze
  - Facial expressions
    - FACS action units
    - Smile , frowning
- **Touch**
  - Haptics:
  - Motion
- **Physiological**
  - Skin conductance
  - Electrocardiogram
- **Mobile**:
  - GPS location
  - Accelerometer
  - Light Sensors

What is a Modality?

Modality refers to **the way** in which something **expressed or perceived**



What is a Multimodal?

Multimodal refers to situations where **multiple modalities are involved**.

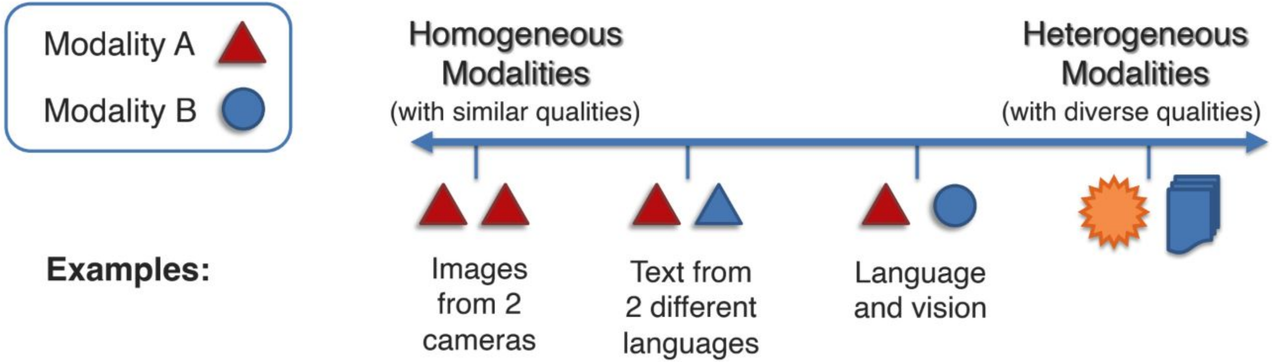
Research-oriented

Multimodal is the scientific study of **heterogeneous** and **interconnected** data

Connected + Interacting

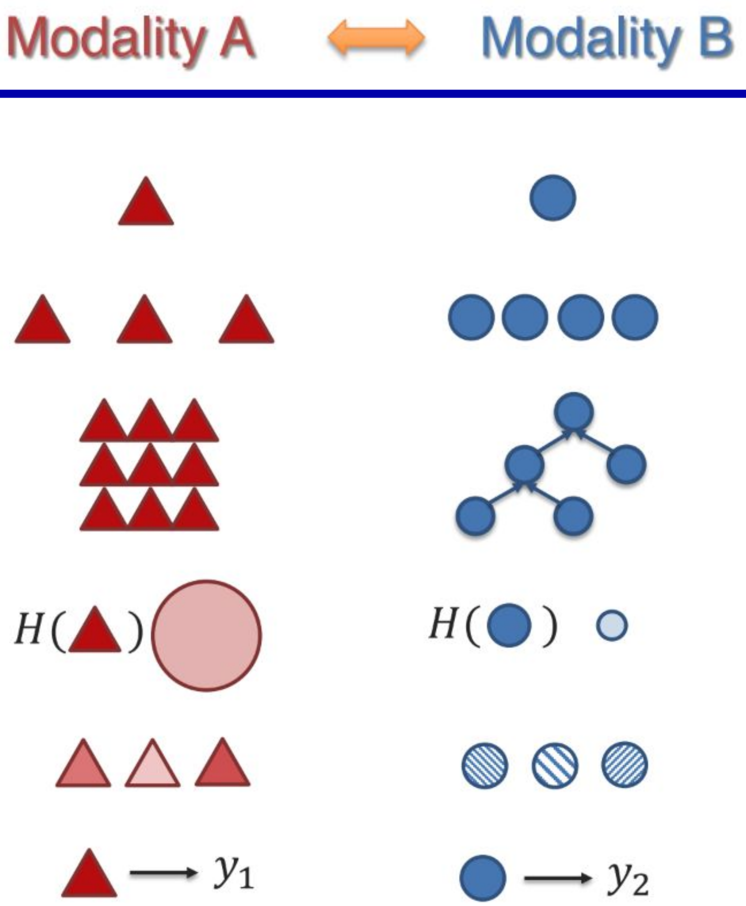
Heterogeneous Modalities

Heterogeneous: Diverse qualities, structures and representations.



Abstract modalities are more likely to be homogeneous

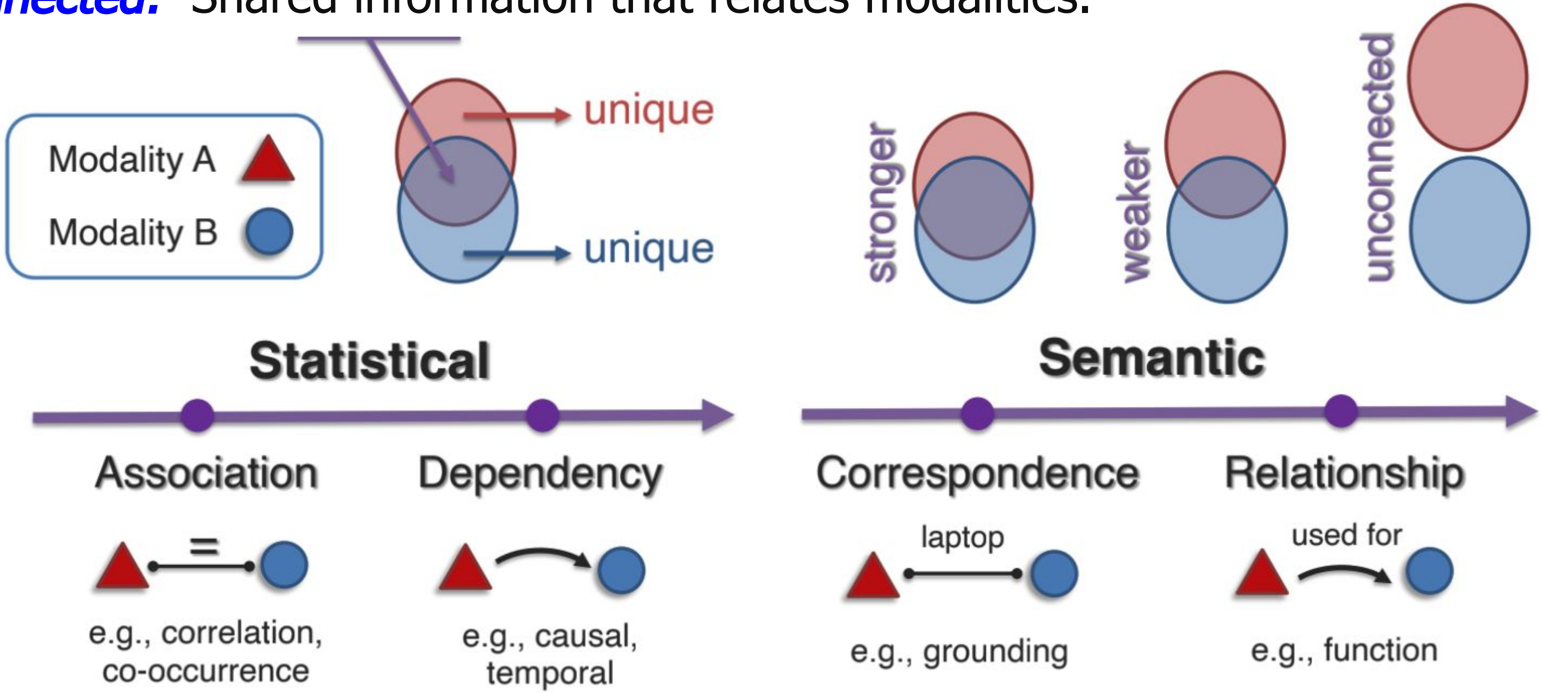
Dimensions of Heterogeneity



1. **Element representations:**
   Discrete, continuous, granularity
2. **Element distributions:**
   Density, frequency
3. **Structure:**
   Temporal, spatial, latent, explicit
4. **Information:**
   Abstraction, entropy
5. **Noise:**
   Uncertainty, noise, missing data
6. **Relevance:**
   Task, context dependence

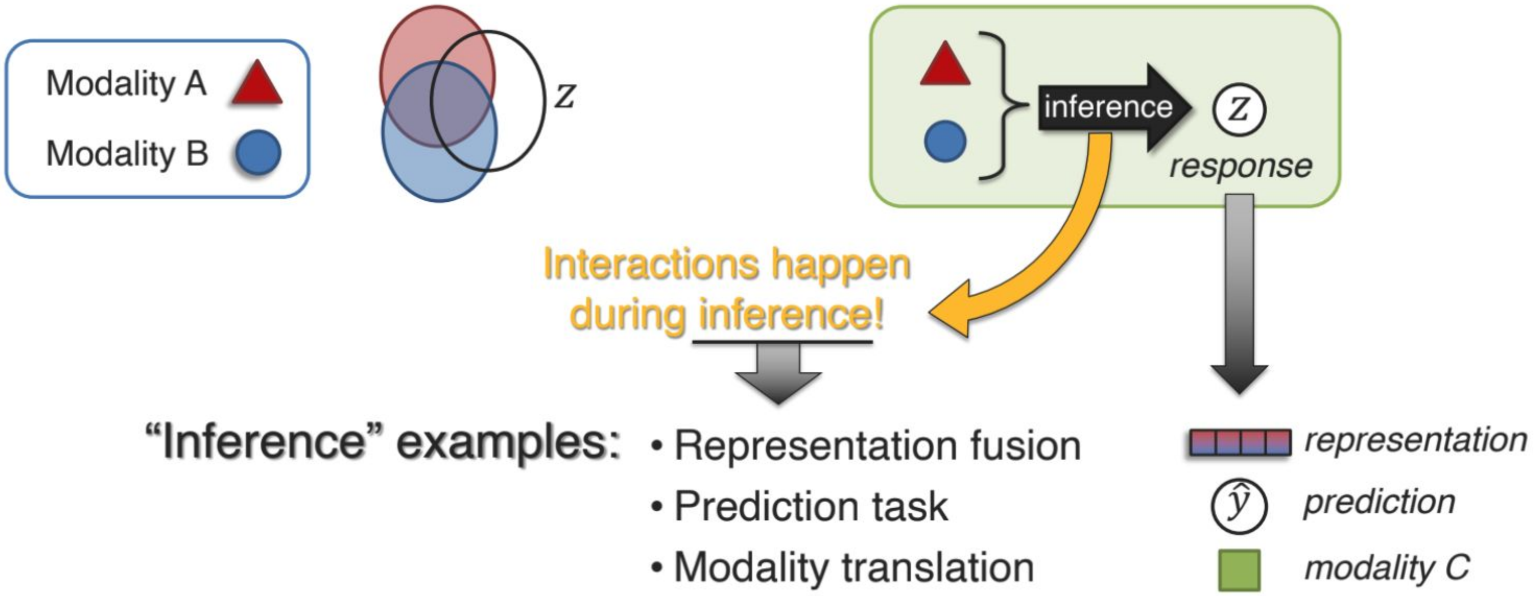
Connected Modalities

Connected: Shared information that relates modalities.

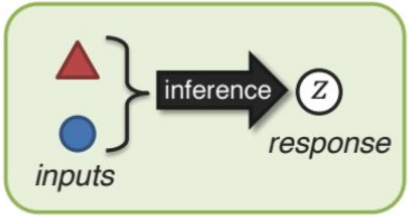


Interacting Modalities

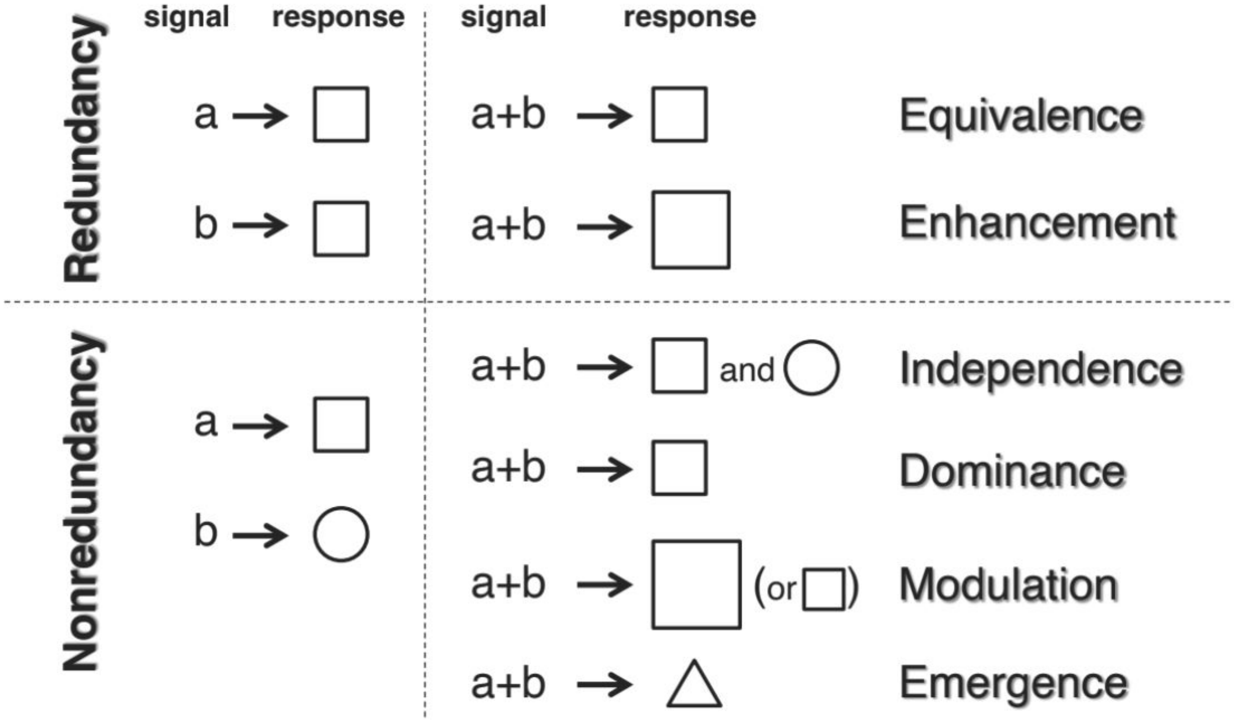
Interacting: process affecting each modality, creating new response



Taxonomy of Interaction Responses - A Behavioral Science View

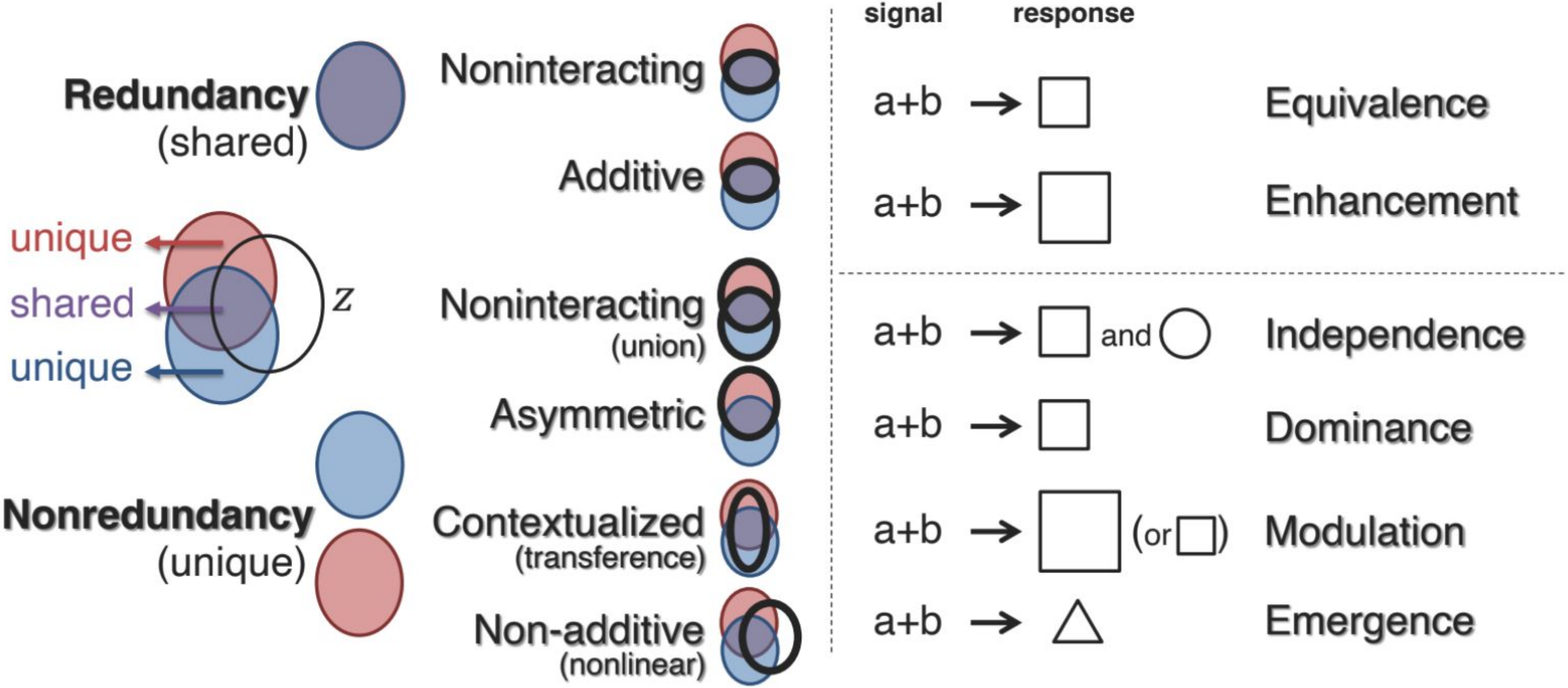


Multimodal Communication

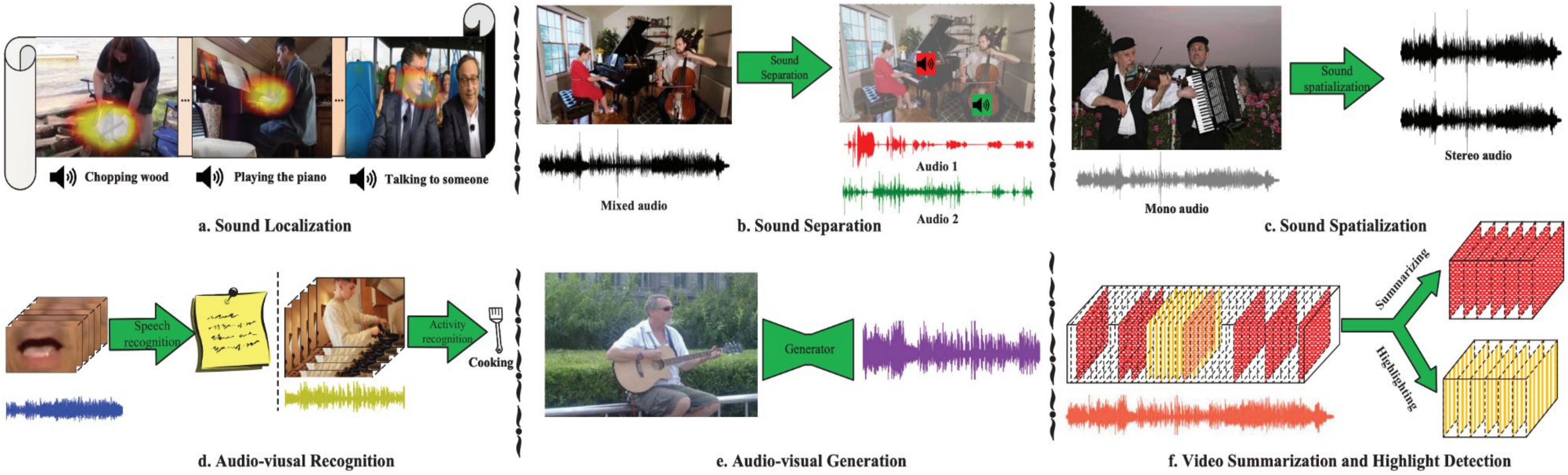


Partan and Marler (2005). Issues in the classification of multimodal communication signals. American Naturalist, 166(2)

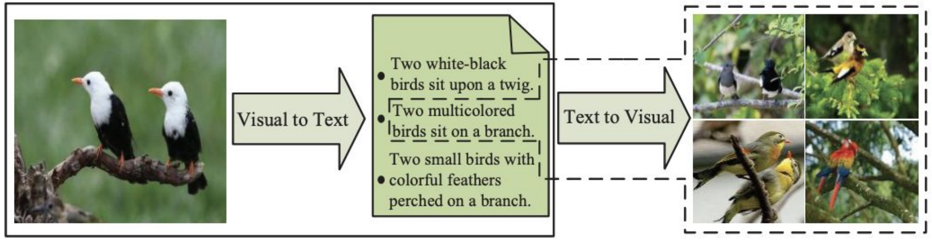
Cross-modal Interaction Mechanics



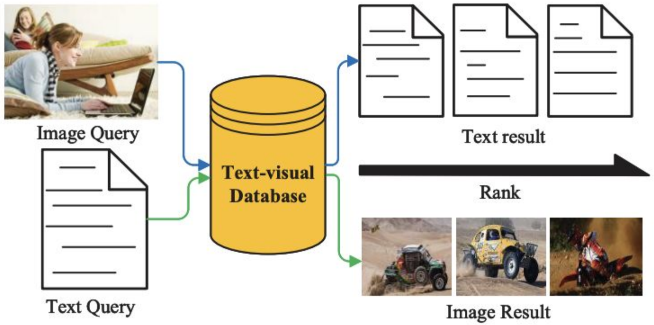
Tasks: Audio-Visual Modalities



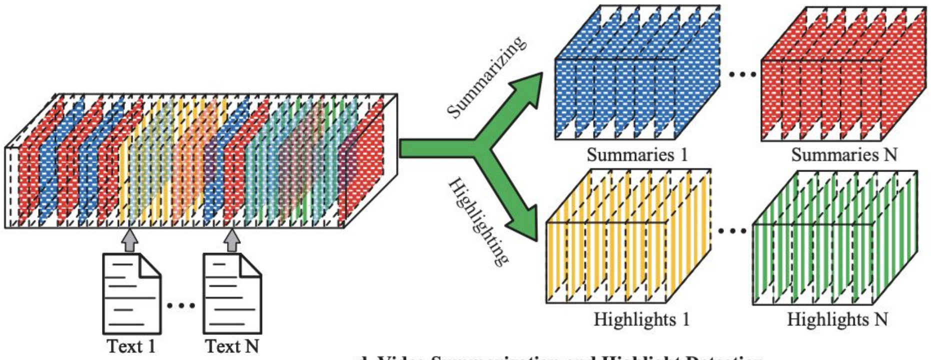
Tasks: Visual-Text Modalities



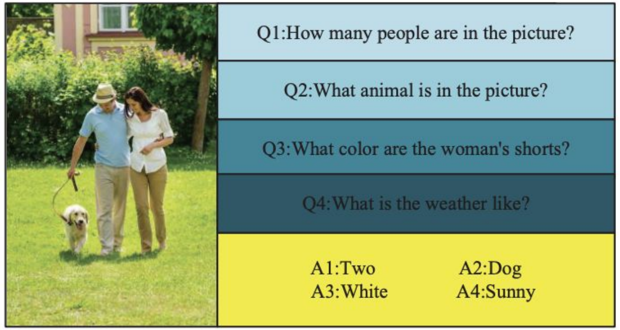
a. Visual to Text/Text to Visual Generation



b. Text-visual Mutual Retrieval

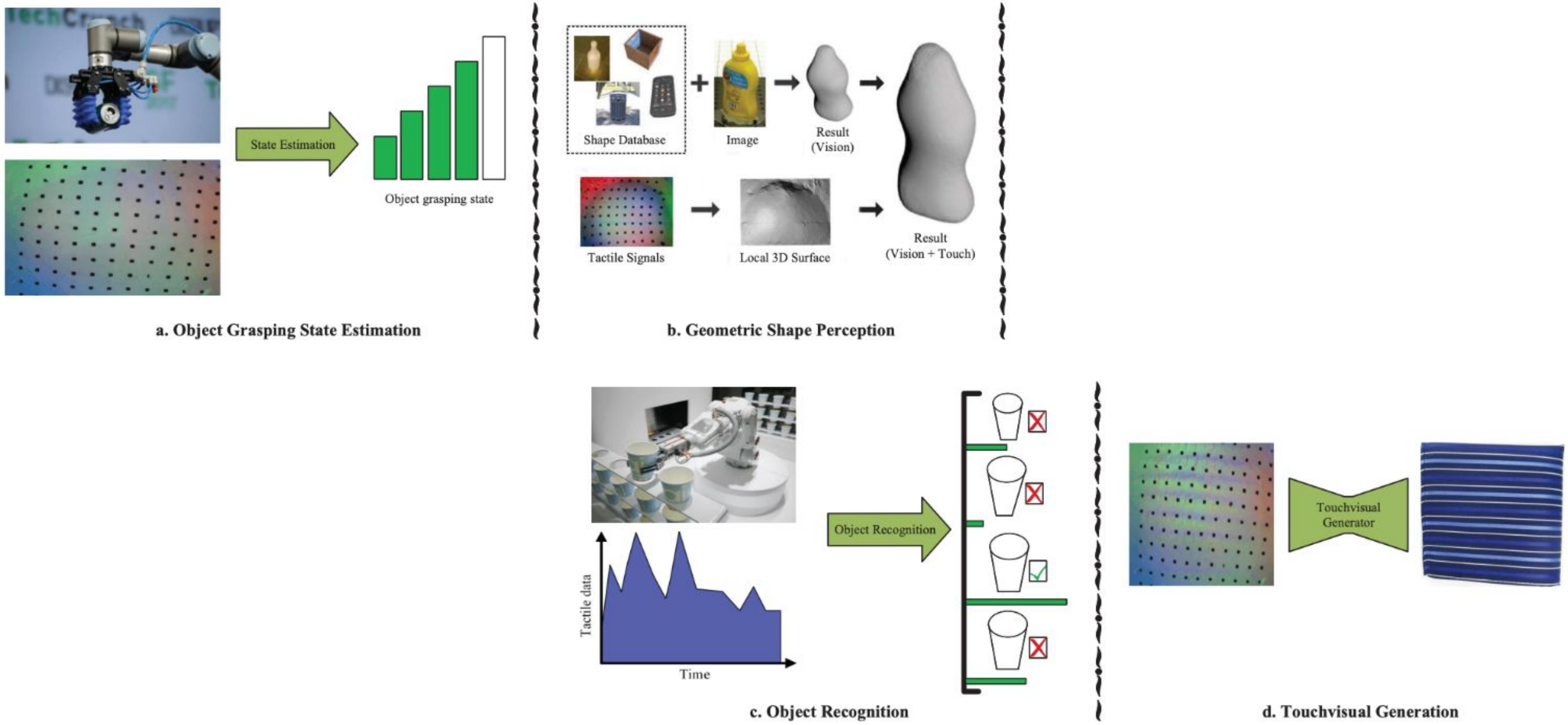


d. Video Summarization and Highlight Detection

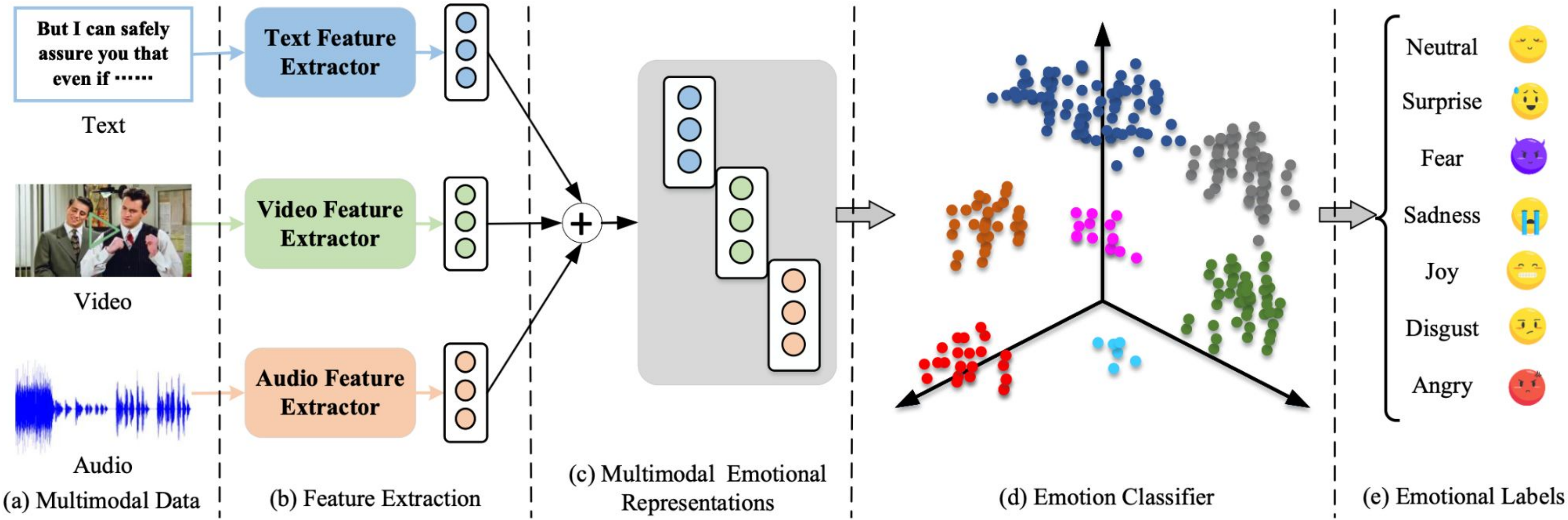


c. Visual Question Answering System

Tasks: Touch-Visual Modalities

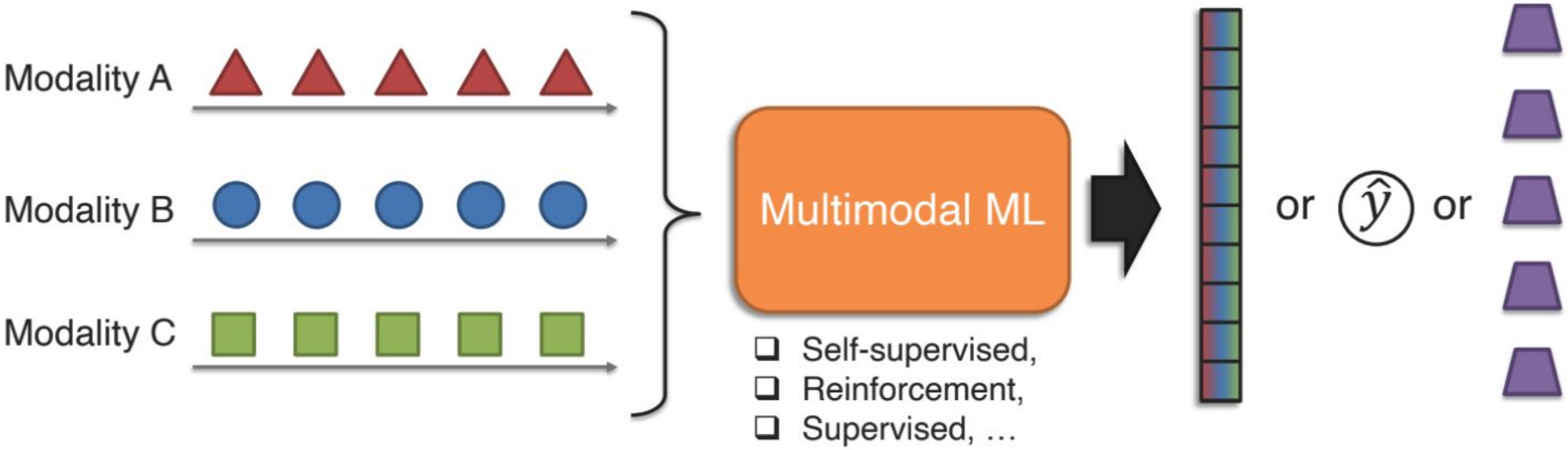


Multimodal Emotion Recognition Task

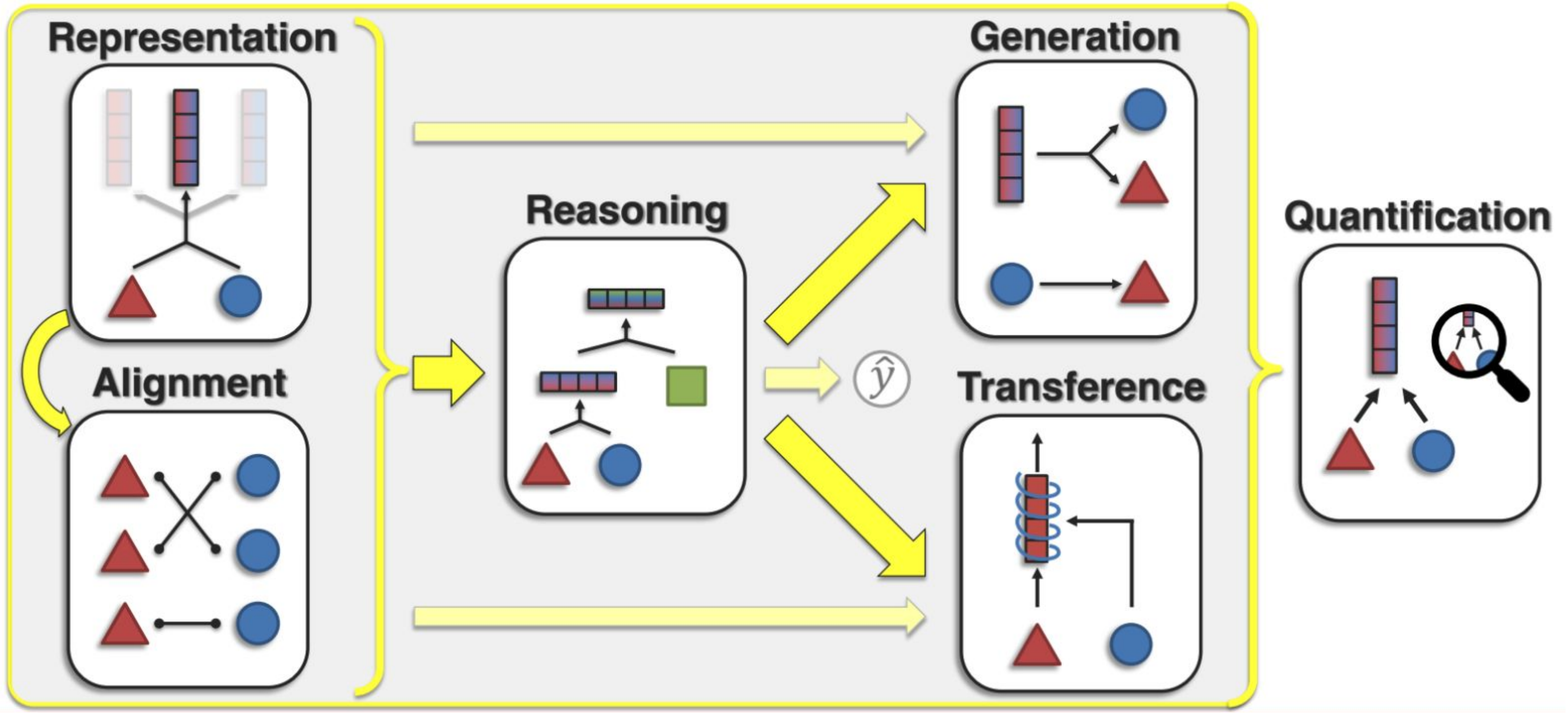


Multimodal Machine Learning

Multimodal Machine Learning aims to **build models** that can **process** and **relate information** from **multiple modalities.**



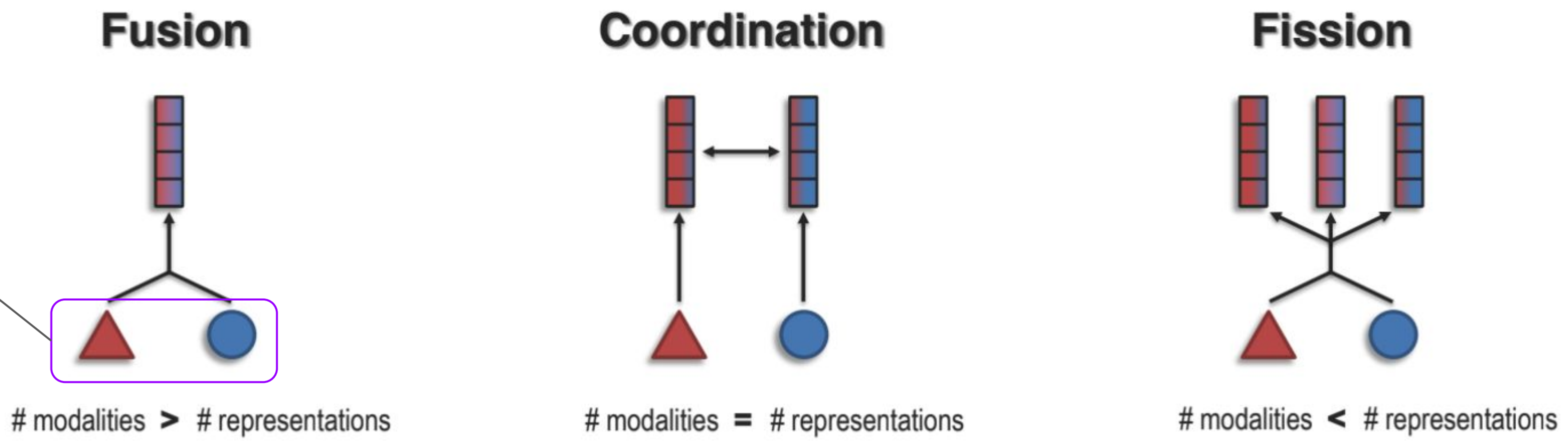
Core Multimodal Learning Challenges



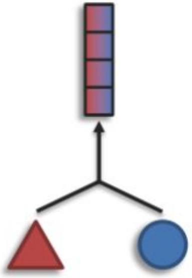
Challenge 1: Representation

Learning representations that reflect **cross-modal interactions** between **individual elements**, across **different modalities.**

From same element

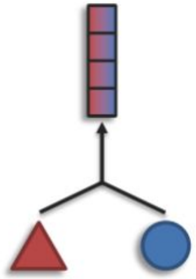


Sub-Challenge 1a: Representation Fusion



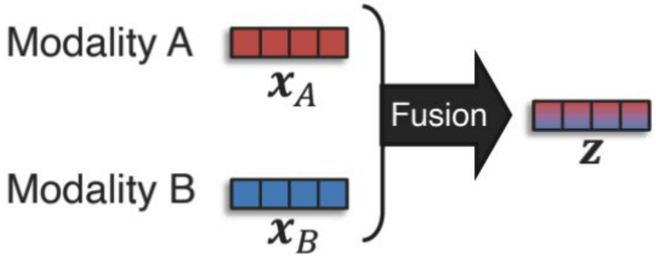
Learn a **joint representation** that models **cross-modal interactions** between **individual elements** of **different modalities**.

Sub-Challenge 1a: Representation Fusion



Learn a **joint representation** that models **cross-modal interactions** between **individual elements** of **different modalities**.

Basic Fusion - Additive Interaction

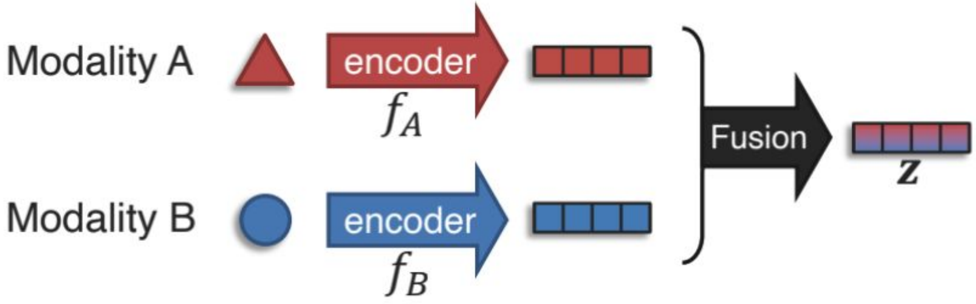


Additive fusion:

$$\boldsymbol{z} = \boldsymbol{w}_1 \boldsymbol{x}_A + \boldsymbol{w}_2 \boldsymbol{x}_B$$

➡ 1-layer neural network can be seen as additive

With unimodal encoders:

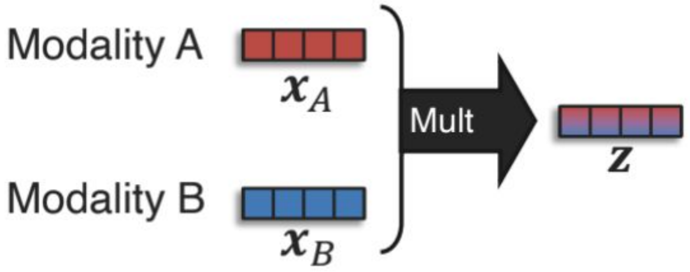


Additive fusion:

$$\boldsymbol{z} = f_A(\blacktriangle) + f_B(\bullet)$$

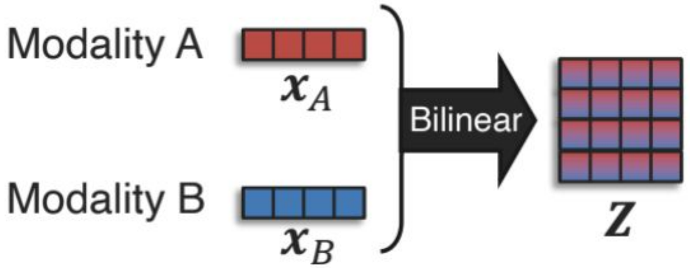
➡ It could be seen as an ensemble approach (late fusion)

Basic Fusion - Multiplicative Interactions



Simple multiplicative fusion:

$$z = w(x_A \times x_B)$$

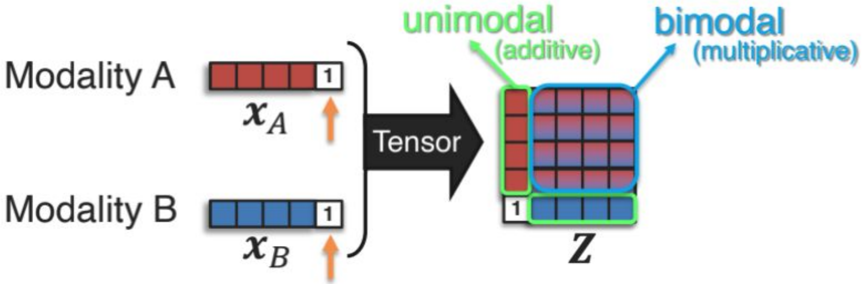


Bilinear Fusion:

$$Z = W(x_A^T \cdot x_B)$$

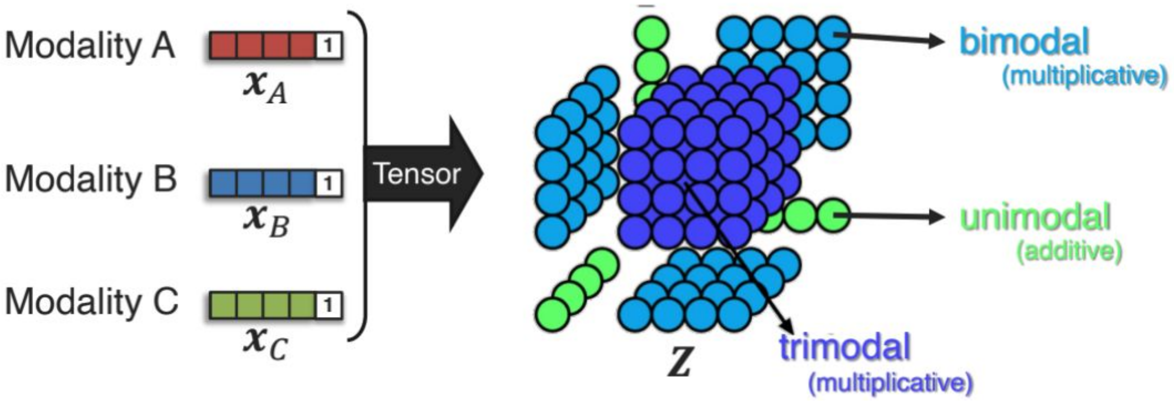
[Jayakumar et al., Multiplicative Interactions and Where to Find Them. ICLR 2020]

Tensor Fusion



Tensor Fusion (bimodal):

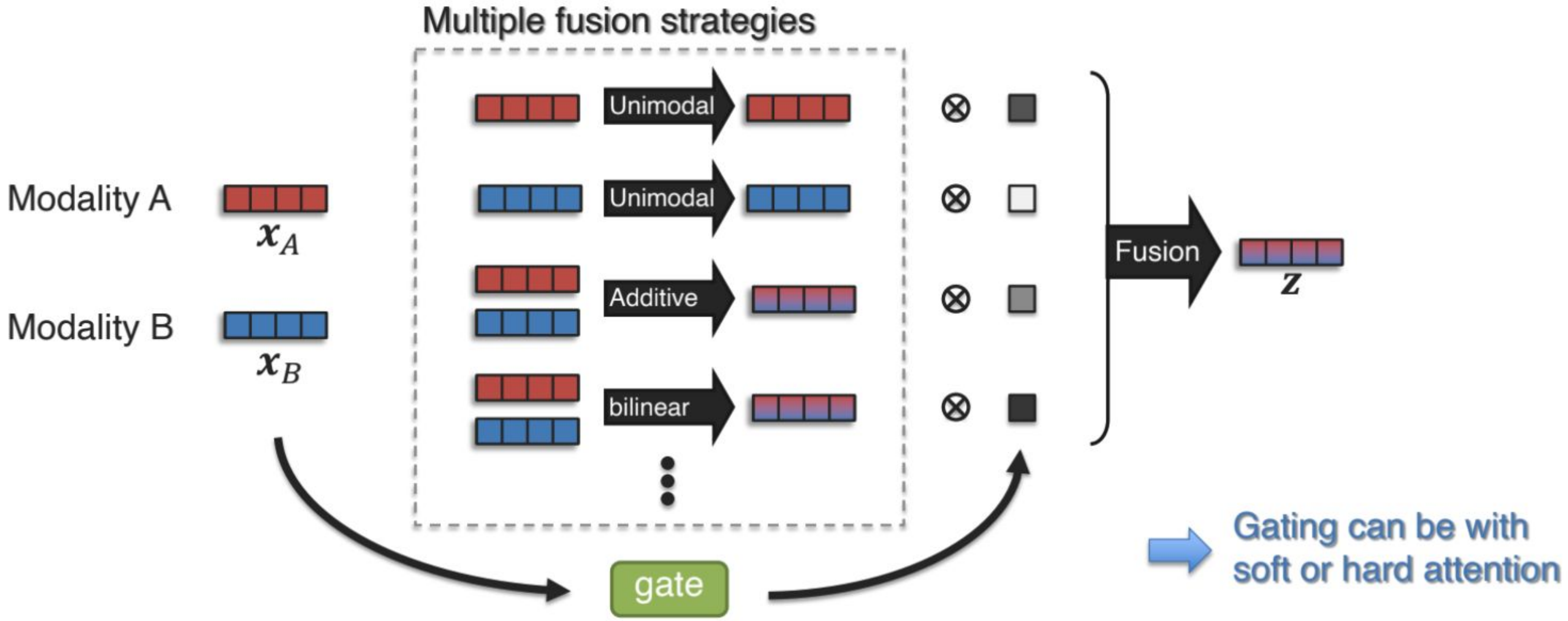
$$Z = w([x_A \quad 1]^T \cdot [x_B \quad 1])$$



… but the weight matrix may end up quite large!

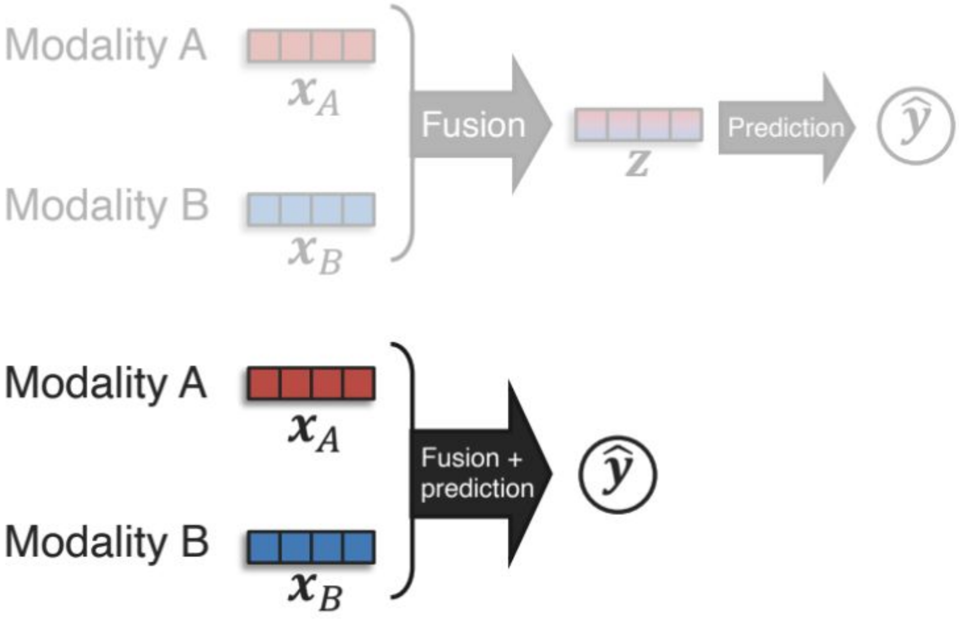
[Hou et al., Deep Multimodal Multilinear Fusion with High-order Polynomial Pooling. NeurIPS 2019]

Mixture of Fusions



[Xu et al., MUFASA: Multimodal Fusion Architecture Search for Electronic Health Records, AAAI 2021]

Nonlinear Fusion



Nonlinear fusion:

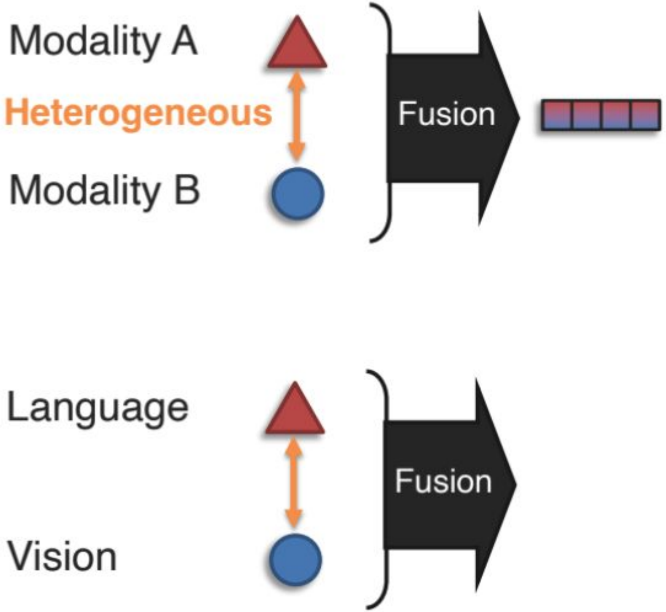
$$\hat{\boldsymbol{y}} = f(\boldsymbol{x}_A, \boldsymbol{x}_B) \in \mathbb{R}^d$$

For any nonlinear model

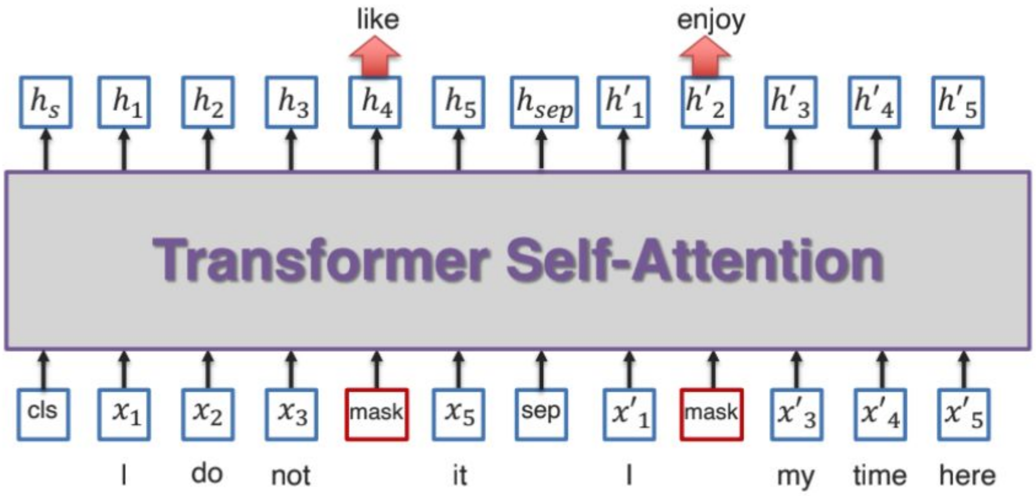
This could be seen as *early fusion*:

$$\hat{\boldsymbol{y}} = f([\boldsymbol{x}_A, \boldsymbol{x}_B])$$

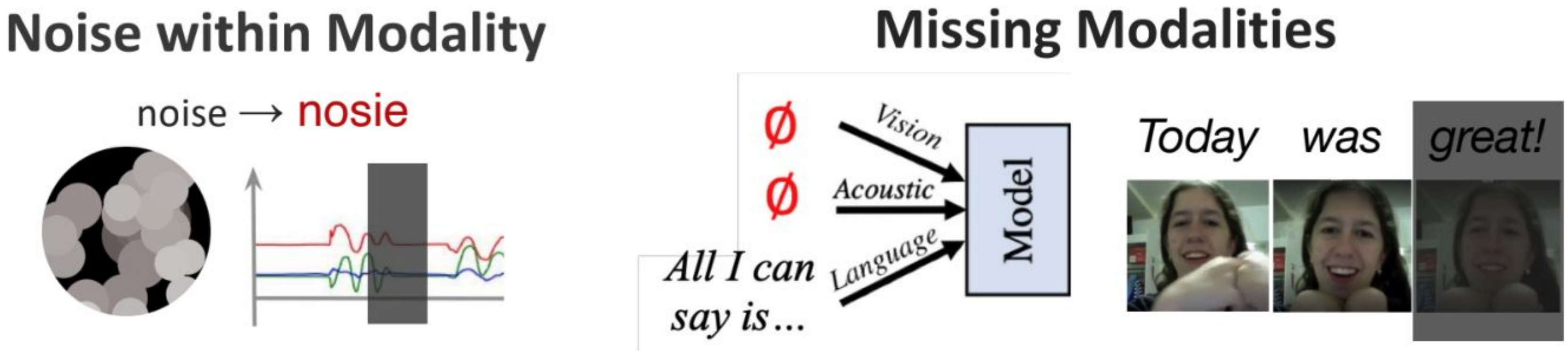
Fusion with Heterogenous Modalities



Can the same fusion algorithm handle raw heterogeonous modalities?



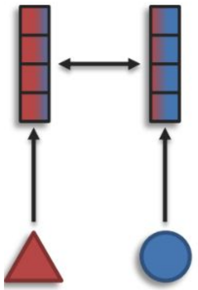
Heterogeneity in Noise



How to remove noise or inferring missing modalities from noised input.

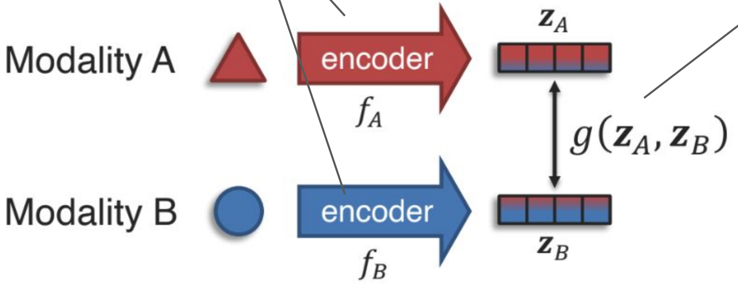
Sub-Challenge 1b: Representation Coordination

Learn **multimodally-contextualized representations** that are **coordinated** through their **cross-modal interactions.**



Capture Heterogeneity

Capture interconnections



Learning with coordination function:

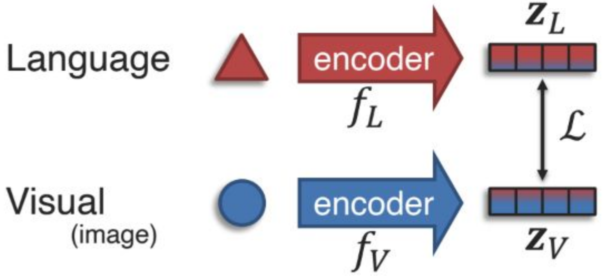
$$\mathcal{L} = g\big(f_A(\blacktriangle), f_B(\bullet)\big)$$

with model parameters θ_g, θ_{f_A} and θ_{f_B}

g ~ (cosine, kernel similarity functions, …

or contrastive loss)

Sub-Challenge 1b: Representation Coordination



Popular contrastive loss: InfoNCE

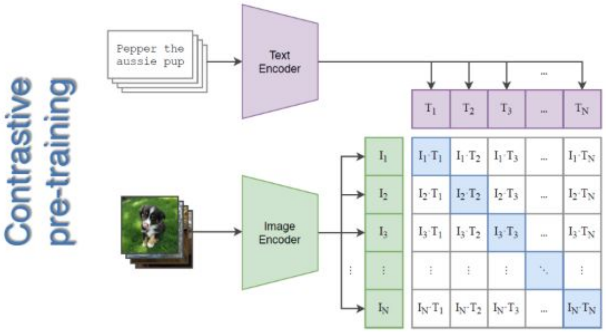
$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\boxed{\mathrm{sim}(\boldsymbol{z}_A^i, \boldsymbol{z}_B^i)}}{\boxed{\sum_{j=1}^{N}\mathrm{sim}(\boldsymbol{z}_A^i, \boldsymbol{z}_B^j)}}$$

positive pairs

negative pairs
and positive pairs

Similarity function can
be cosine similarity

Positive and negative pairs:

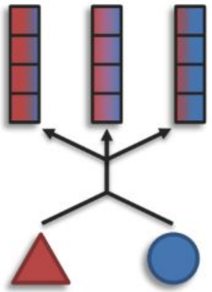


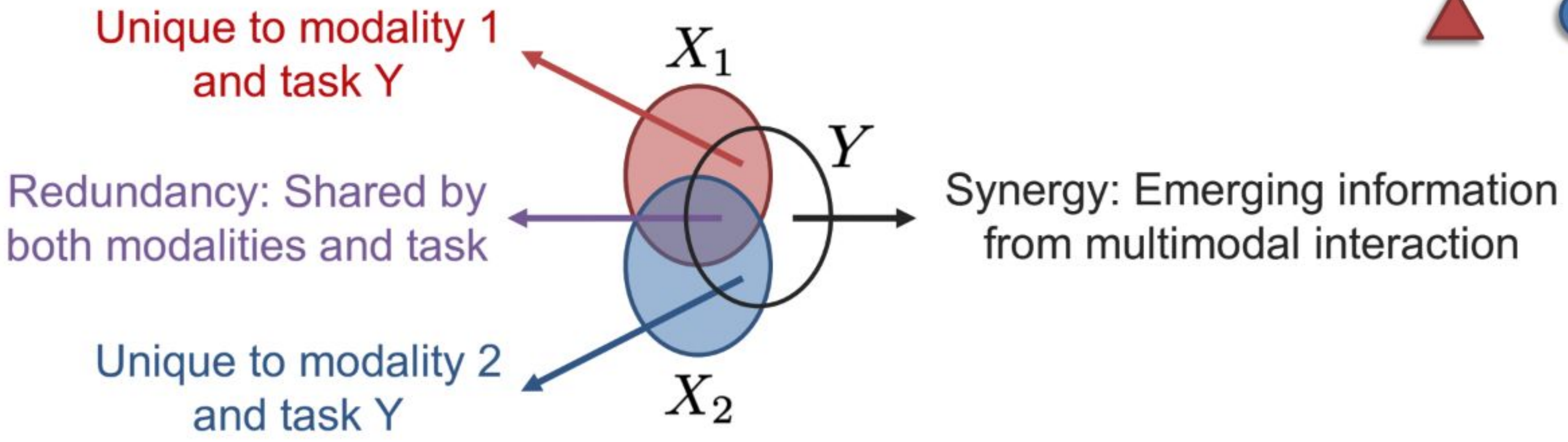
- CLIP encoders (f_L and f_V) are great for language-vision tasks
- $\boldsymbol{z}_L$ and $\boldsymbol{z}_V$ are coordinated but not identical representation spaces

[Radford et al., Learning Transferable Visual Models From Natural Language Supervision. ICML 2021]

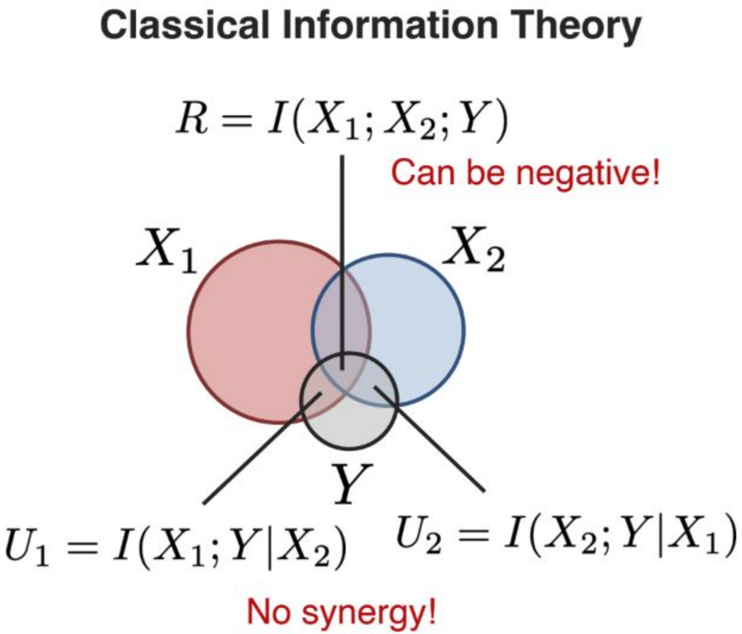
Sub-Challenge 1c: Representation Fission

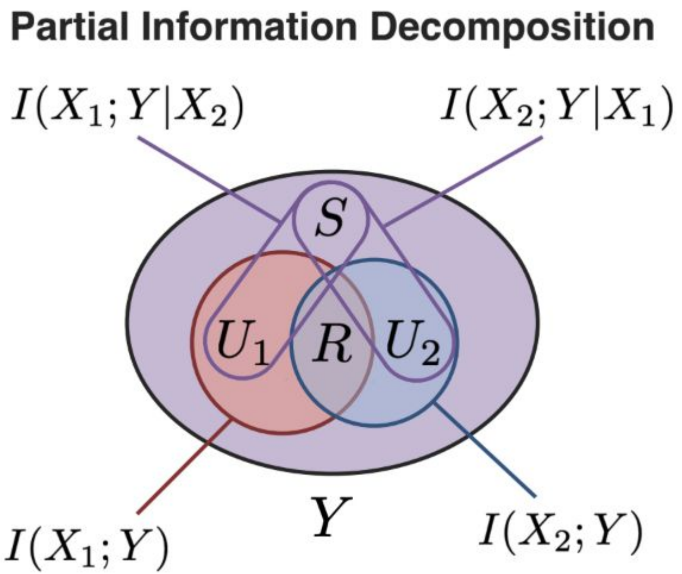
Learning a **new set of representations** that reflects **individual multimodal interactions** and **data clustering**.





Partial Information Decomposition



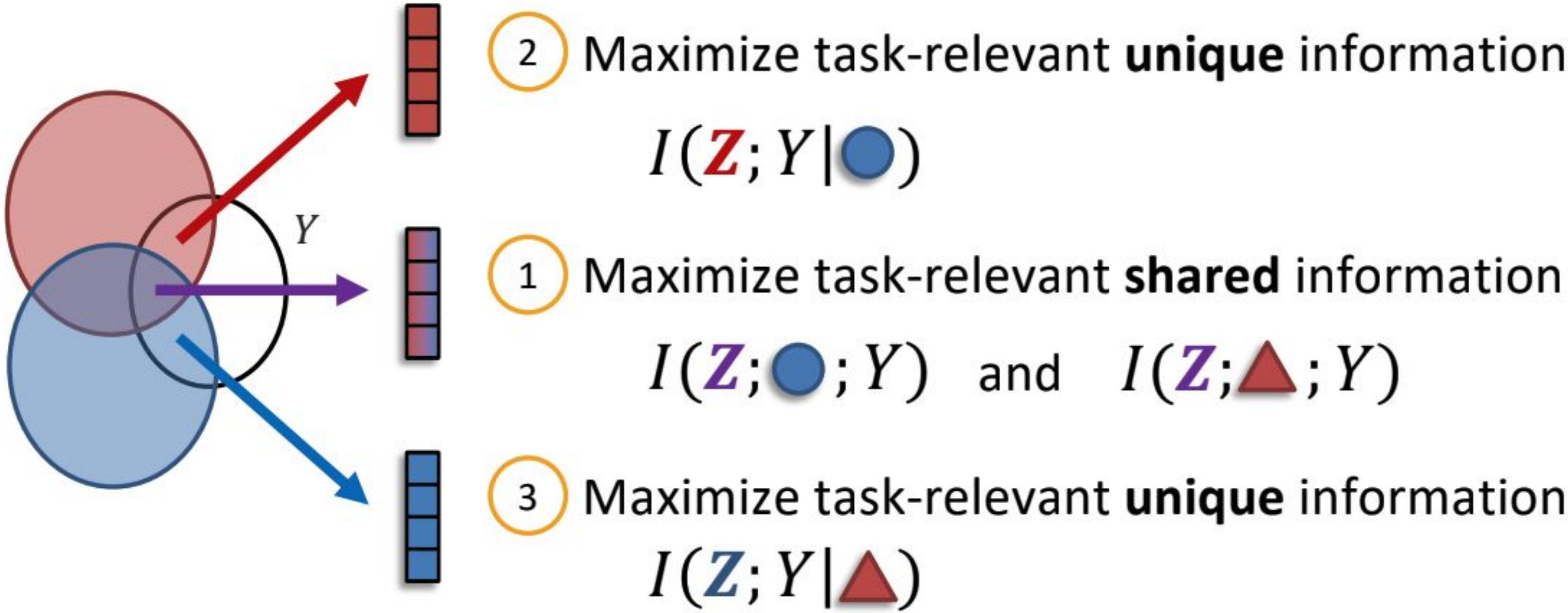


$R - S = I(X_1; X_2; Y)$ Explains negative!

$R + U_1 + U_2 + S = I(X_1, X_2; Y)$ Task-relevant multimodal info

[Williams and Beer. Non-negative Decomposition of Mutual Information. 2010]

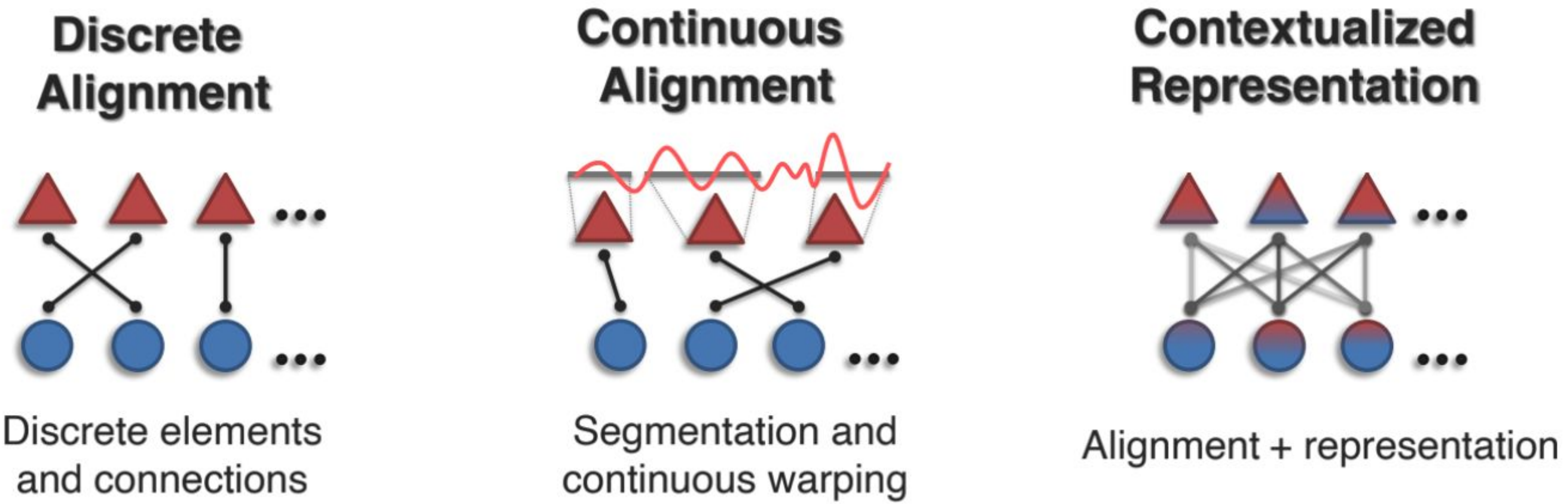
Learning Task-relevant Unique Information



[[Liang et al., Factorized Contrastive Learning: Going Beyond Multi-view Redundancy,, NeurIPS'23]

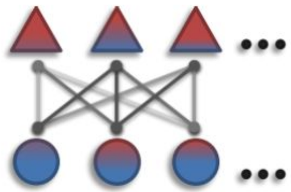
Challenge 2: Alignment

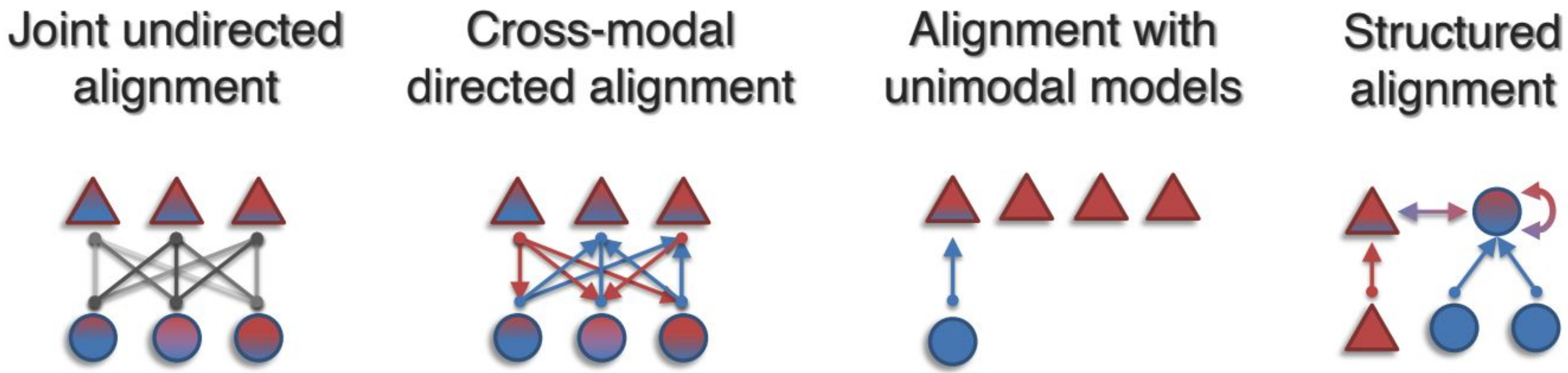
Identifying and modeling **cross-modal connections** between **all elements** of **multiple modalities**, building from the **data structure,** e.g., temporal, spatial, hierarchical.



Sub-Challenge 2c: Contextualized Representation

Learn representations that **reflect the cross-modal interactions** of the **structured multimodal data.**

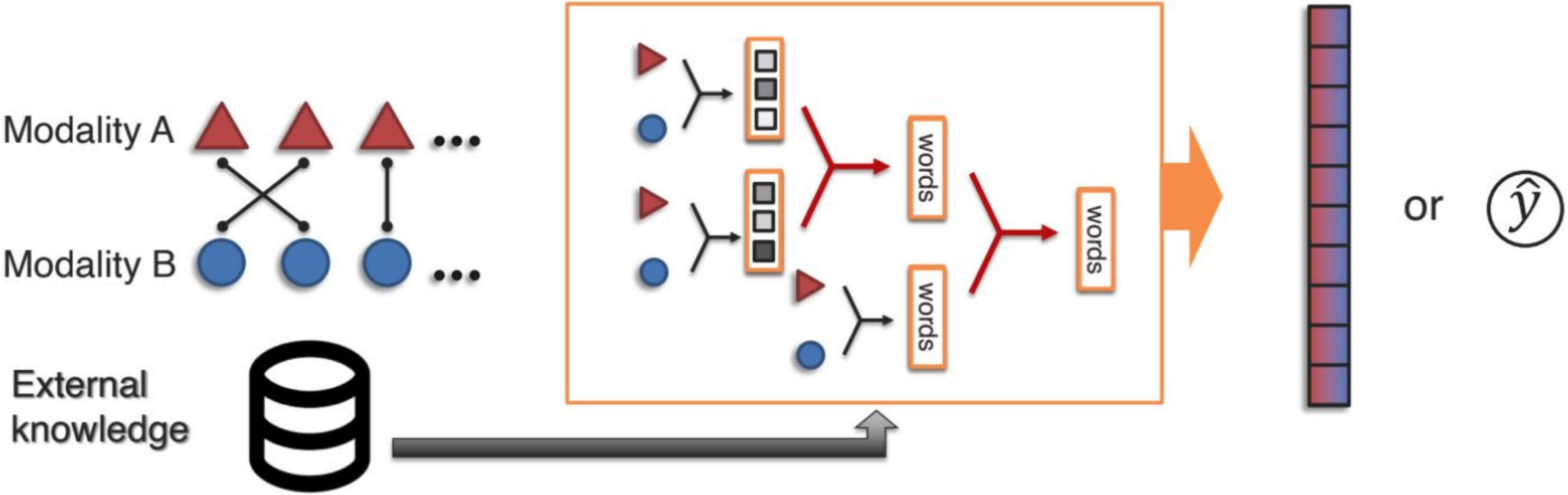




(+ knowledge graphs)

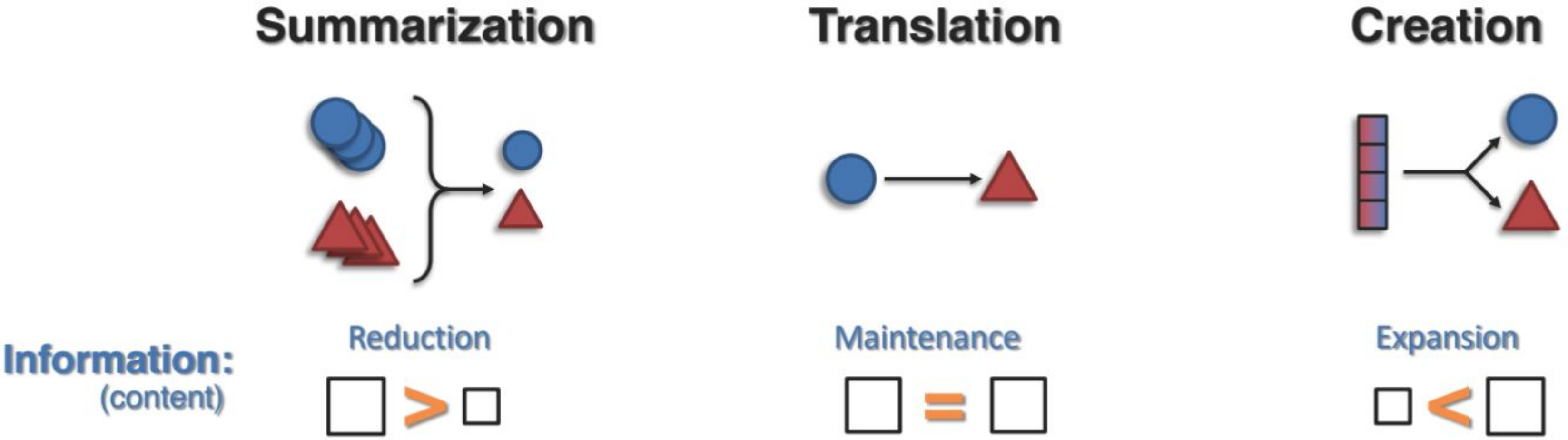
Challenge 3: Reasoning

Combining knowledge, usually through multiple inferential steps, exploiting **multimodal alignment** and **problem structure**



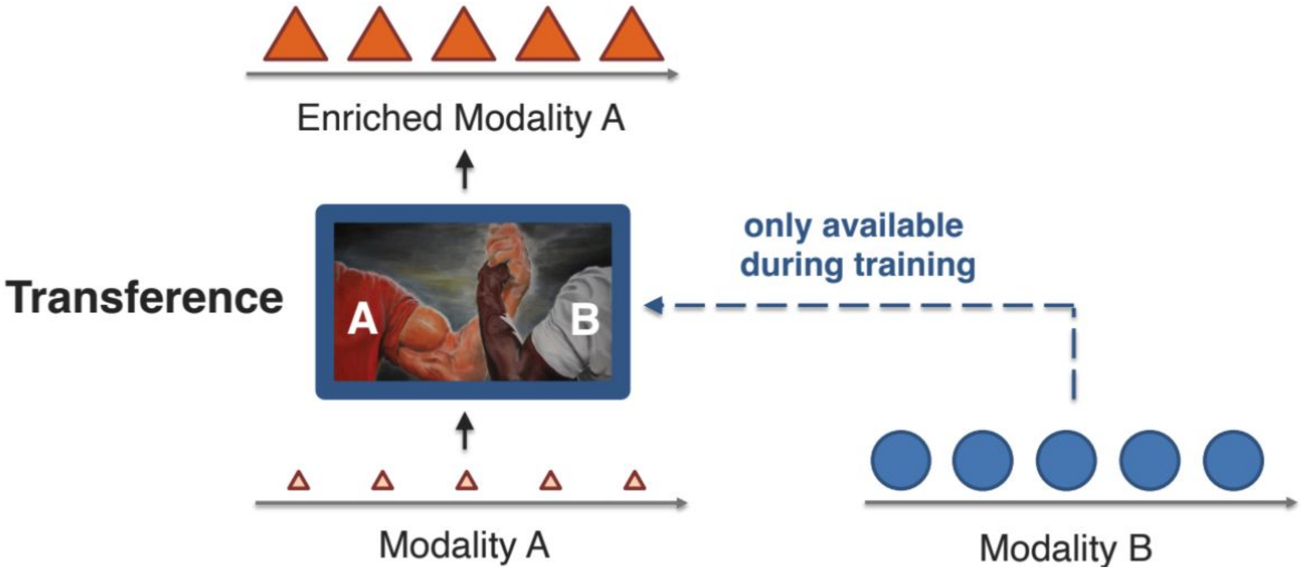
Challenge 4: Generation

Learning a **generative process** to produce **raw modalities** that reflects **cross-modal interactions**, **structure** and **coherence**



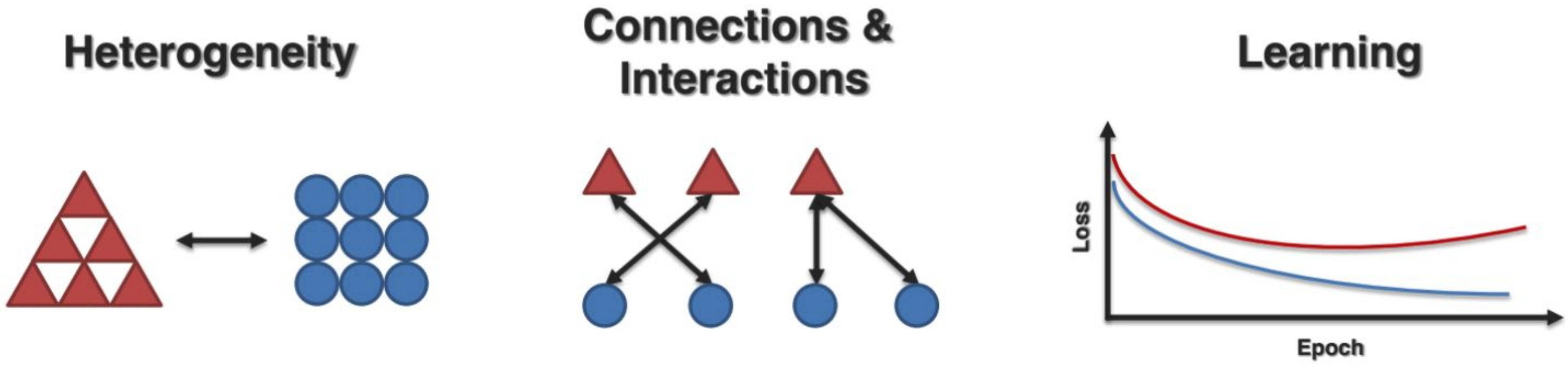
Challenge 5: Transference

Transfer knowledge between modalities, usually to help the **target modality** which may be **noisy** or with **limited resources.**



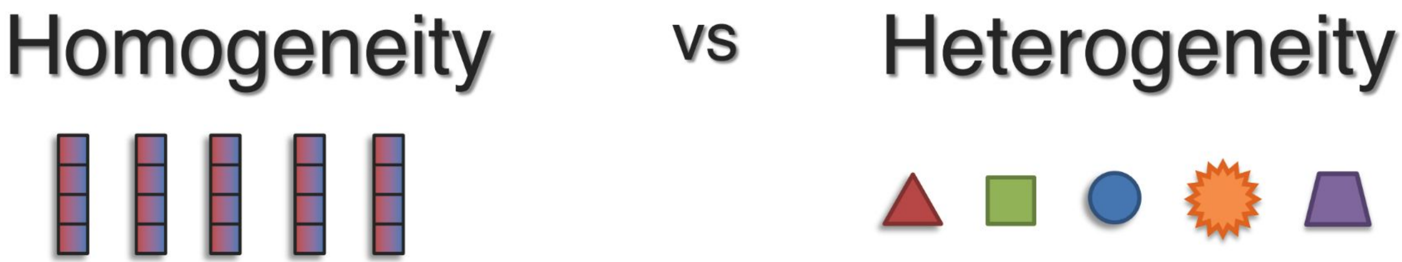
Challenge 6: Quantification

Empirical and theoretical study to **better understand heterogeneity**, **cross-modal interactions** and the **multimodal learning process**.



2. Emerging Trends in Multimodal Learning

Heterogeneity



Challenges:

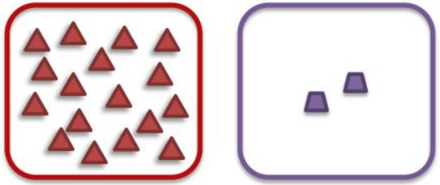
- **Arbitrary tokenization** between modalities
- **Beyond differentiable interactions**: Causal, logical, brain-inspired interactions, theoretical study of interactions

Multi-modality to High-modality

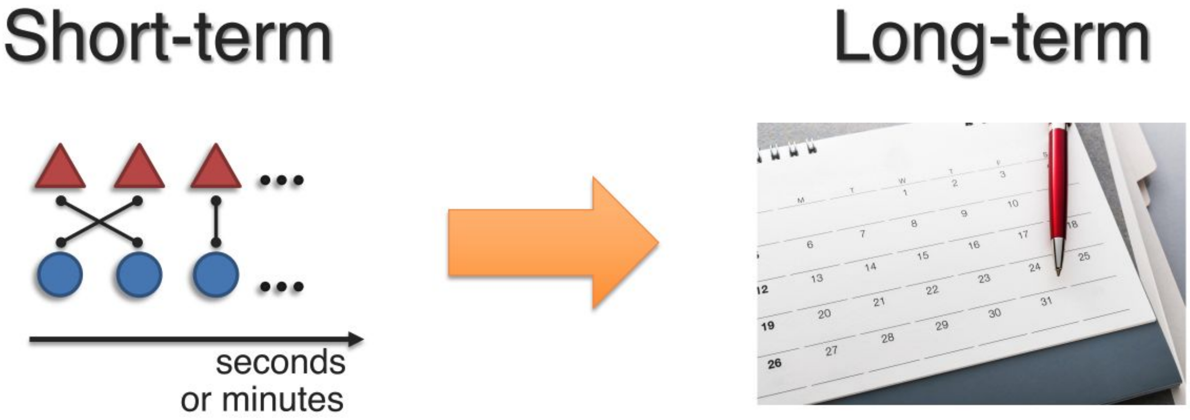
Language Vision Audio Graphs Control LIDAR Sensors Set Table Financial Medical

Challenges:

- Non-parallel learning
- Limited Resources

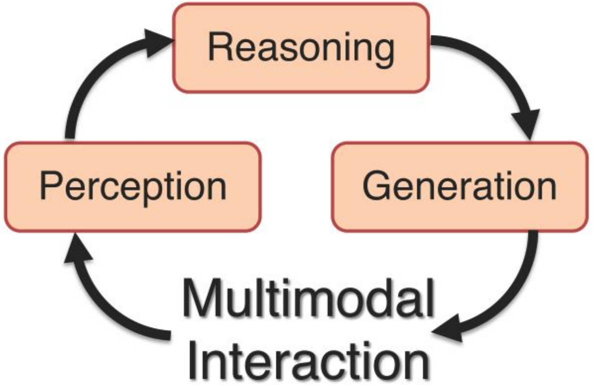


Short-term to Long-term



Challenges: Compositionality, Memory (Continual learning), Personalization

Complex Interaction



Challenges: Multi-party, Causality, Bias/Fairness

Reliability

Healthcare Decision Support

Intelligent Interfaces and Vehicles

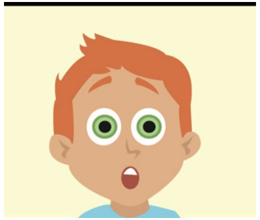
Online Learning and Education

Challenges:

- Robustness
- Fairness
- Interpretation

3. Applications in Multimodal Emotion Recognition

Multimodal Emotion Recognition



Visual

Audio

Text

. . .

ML/DL Classifier

Anger

Happiness

Excitement

Sadness

Frustration

Surprise (99%)

Fear

Multimodal Emotion Recognition refers to the identification and understanding of human emotional states by combining various modalities, e.g., visual, audio, text

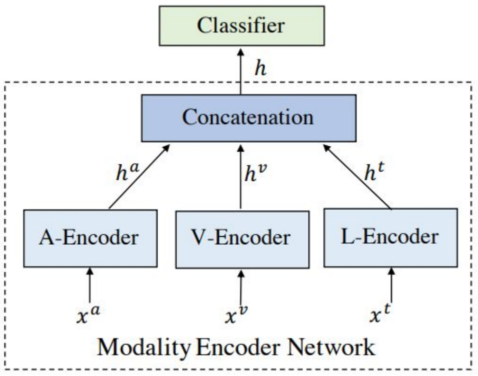
Multimodal Emotion Recognition in Conversations (ERC)

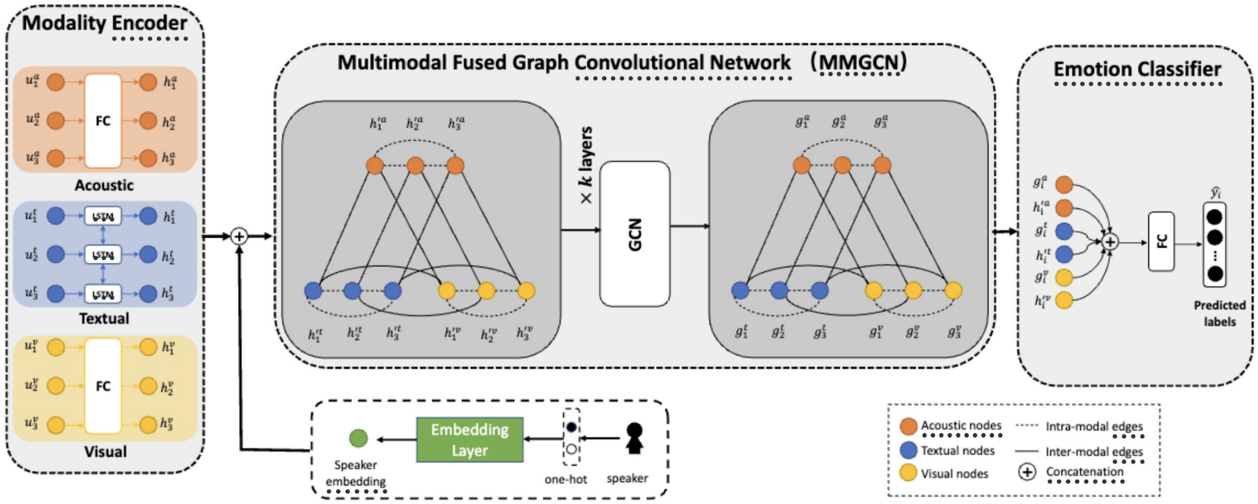
Multimodal Emotion Recognition in Conversation refers to the process of understanding and interpreting emotions expressed in the context of conversations using multiple modes of communication

Source: Refashioning Emotion Recognition Modelling: The Advent of Generalised Large Models

Multimodal Learning for ERC

Main idea: Seek to exploit modalities separately and/or jointly to enhance the multimodal representation for the ERC task.

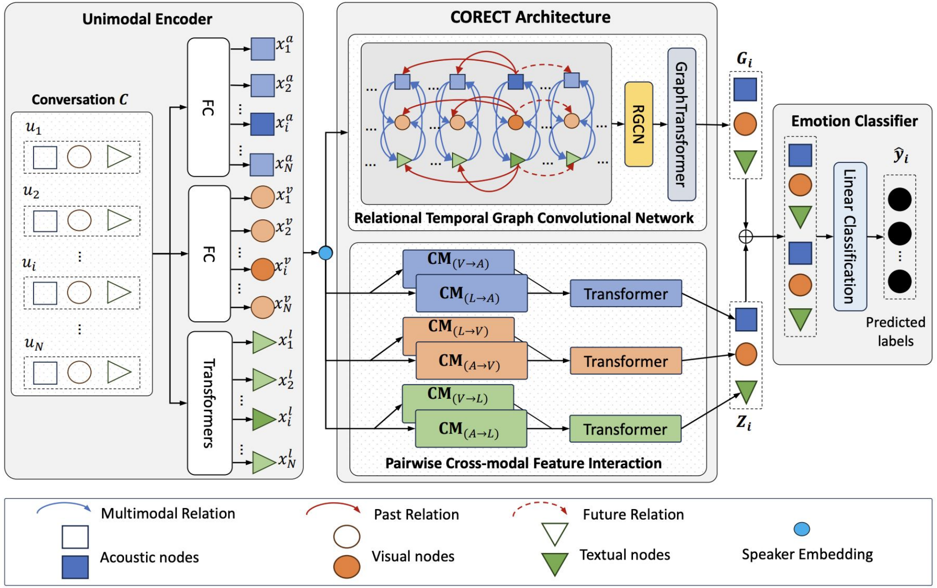




Source: MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation

CORRECT

Contextualized Representation



Main idea: Exploit relational temporal information & pairwise cross-modality feature interactions.

[Conversation Understanding using Relational Temporal Graph Neural Networks with Auxiliary Cross-Modality Interaction (EMNLP 2023, A*)]

	IEMOCAP (6-way)							
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Acc. (%)	w-F1 (%)
	32.63	70.34	51.14	63.44	67.91	61.06	59.58	59.10
	30.38	62.41	52.39	59.83	60.25	60.69	56.56	56.13
ICON (Hazarika et al., 2018a)	29.91	64.57	57.38	63.04	63.42	60.81	59.09	58.54
DialogueRNN (Majumder et al., 2019)	33.18	78.80	59.21	65.28	71.86	58.91	63.40	62.75
DialogueGCN (Ghosal et al., 2019)	47.10	**80.88**	58.71	66.08	70.97	61.21	65.54	65.04
MMGCN (Wei et al., 2019)	45.45	77.53	61.99	66.70	72.04	64.12	65.56	65.71
DialogueCRN (Hu et al., 2021)	51.59	74.54	62.38	67.25	73.96	59.97	65.31	65.34
COGMEN (Joshi et al., 2022)	55.76	80.17	63.21	61.69	**74.91**	63.90	67.04	67.27
CORECT (Ours)	**59.30**	80.53	**66.94**	**69.59**	72.69	**68.50**	**69.93** (↑ 2.89)	**70.02** (↑ 2.75)

MI-TPA

Contextualized Representation

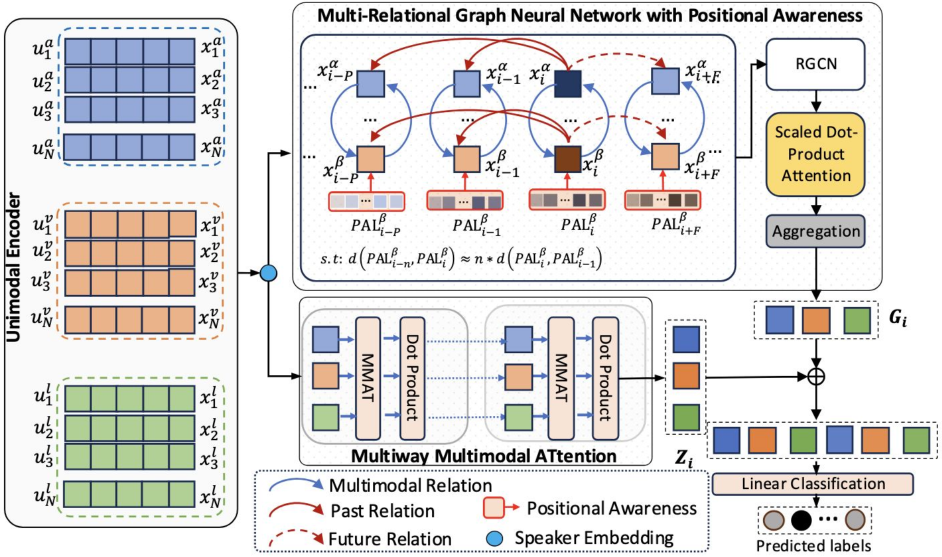


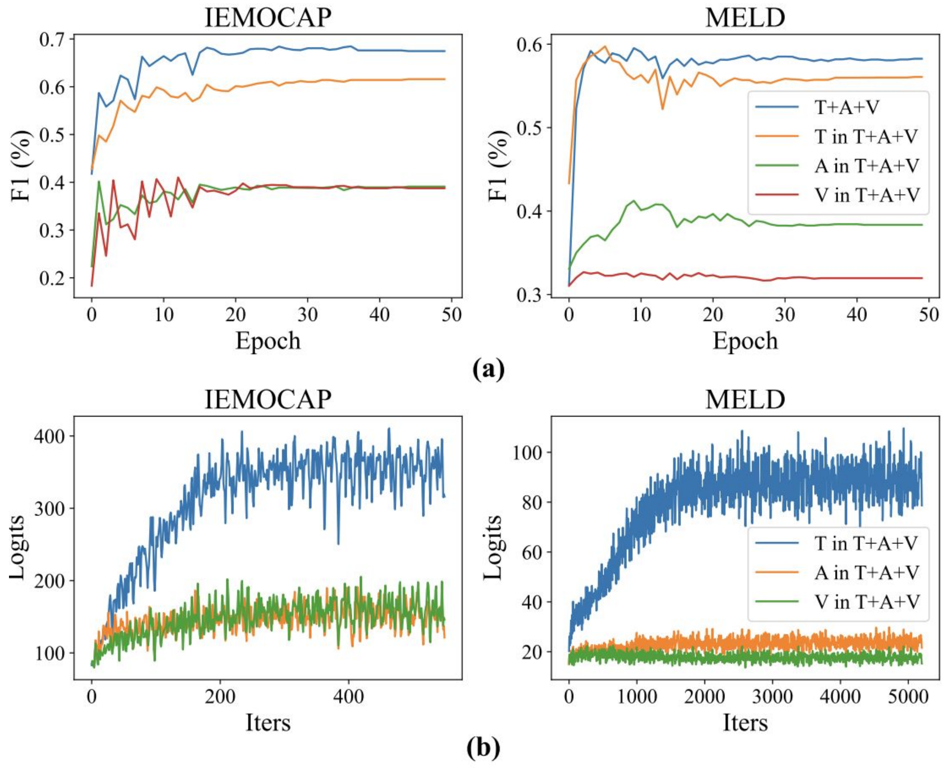
Fig. 2: Overall framework of MI-TPA

Main idea: Exploit relational temporal information with positional awareness & multiway-multimodal feature interactions.

	Labels						Overall Performance		
	Happy	Sad	Neutral	Angry	Excited	Frustrated	w-F1 (%)	Acc. (%)	Std dev
	32.63	70.34	51.14	63.44	67.91	61.06	59.10	59.58	11.85
	30.38	62.41	52.39	59.83	60.25	60.69	56.13	56.56	10.38
	29.91	64.57	57.38	63.04	63.42	60.81	58.54	59.09	11.28
	-	-	-	-	-	-	62.41	-	-
	52.10	73.30	58.40	61.90	69.70	62.30	63.50	63.50	6.98
]	-	-	-	-	-	-	66.18	-	-
69]	-	-	-	-	-	-	69.70	69.50	-
	-	-	-	-	-	-	67.61	-	-
	-	-	-	-	-	-	-	-	-
72]	-	-	-	-	-	-	-	-	-
SDT []	48.14	74.84	57.31	64.68	64.78	61.15	62.19	61.68	...
MM-DFN [73]	42.22	78.98	66.42	<u>69.77</u>	<u>75.56</u>	66.33	66.18	68.21	-
CFN-ESA [74]	53.29	<u>80.72</u>	**69.69**	68.73	74.60	<u>67.29</u>	<u>70.14</u>	70.06	8.36
GraphSmile []	49.66	70.67	62.15	65.32	69.91	65.06	64.77	64.57	...
MMGCN [7]	-	-	-	-	-	-	65.71	65.71	9.28
I-GCN [75]	50.00	**83.80**	59.30	64.60	74.30	59.00	65.40	65.50	11.06
COGMEN [8]	55.76	80.17	63.21	61.69	74.91	63.90	67.27	67.04	7.69
CORECT [9]	<u>59.30</u>	80.53	<u>66.94</u>	69.59	72.69	**68.50**	70.02	<u>69.93</u>	<u>5.90</u>
MI-TPA (Ours)	**62.69**	77.59	<u>67.17</u>	**71.04**	**77.52**	66.04	**70.39**	**70.36**	**5.23**

[MI-TPA: Integrating Multimodal Interaction and Temporal Positional Awareness for Enhancing Conversational Emotion Recognition, TAFFC (Q1, IF 9.8, Round 2)]

The Presence of Imbalance Modality Learning

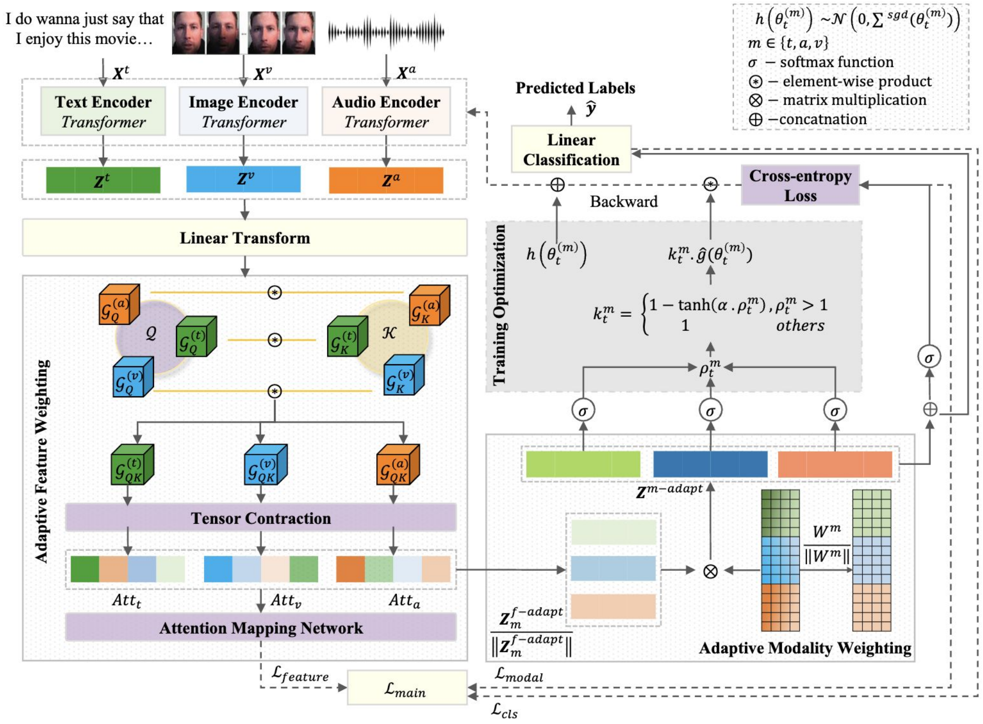


The text modality quickly addresses the overall model performance, whereas the visual and audio modalities remain under-optimized throughout the training process

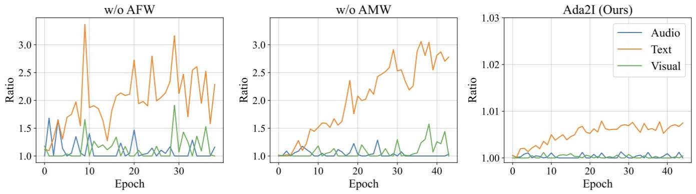
Ada2I

Quantification

Main idea: Adaptive Feature-level Balancing (intra-modal) & Adaptive Modality-Level Balancing (inter-modal)



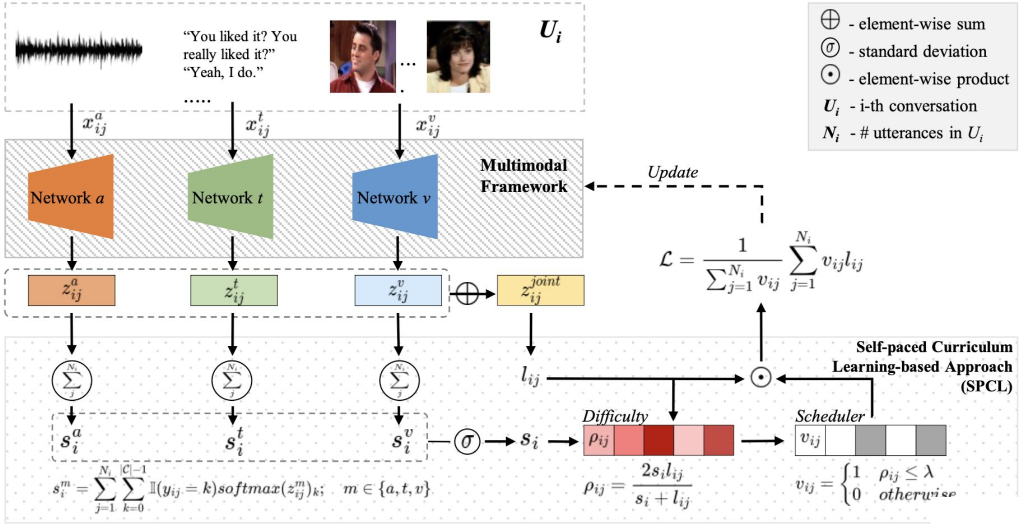
Methods	IEMOCAP							
	TAV		TA		TV		AV	
	W-F1	Acc	W-F1	Acc	W-F1	Acc	W-F1	Acc
DialogueRNN†	61.31	61.61	61.90	61.98	60.19	59.95	48.31	50.71
DialogueGCN†	62.76	63.22	64.36	64.39	61.25	62.23	49.20	49.85
BiDDIN†	58.81	58.84	58.88	58.16	59.04	58.96	46.36	46.77
MM-DFN†	64.92	64.57	63.91	64.20	61.02	60.60	54.48	55.03
MMGCN†	64.53	64.51	63.25	63.40	61.02	61.06	54.14	54.90
Ada2I	**68.97**	**68.76**	**66.91**	**67.28**	**65.48**	**65.43**	**55.16**	**55.64**
Δ	↑4.05	↑4.19	↑2.55	↑2.89	↑4.23	↑3.20	↑0.68	↑0.61



[Ada2I: Enhancing Modality Balance for Multimodal Conversational Emotion Recognition (ACM MM 2024, A*)]

SPCL

Quantification



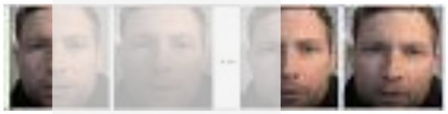
Main idea: Model-agnostic, Difficulty Measure quantifies sample complexity (utterance and conversation-level) & Learning Scheduler adaptively regulates the training curriculum (easier -> complex)

MMGCN [6]

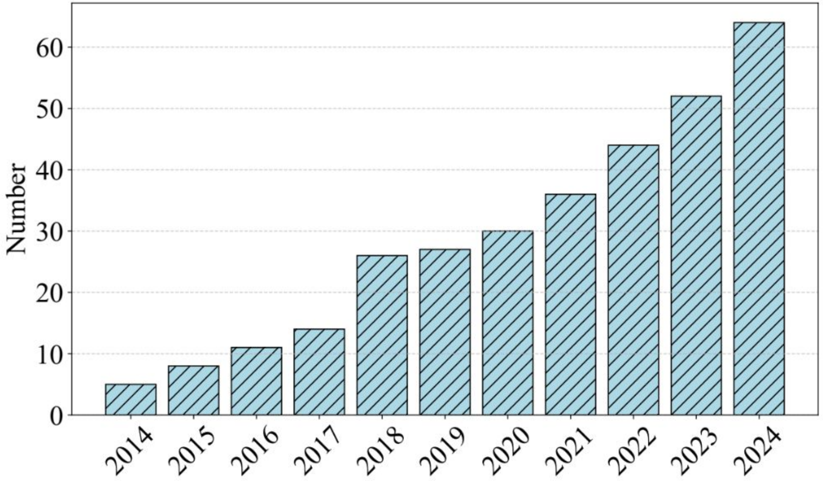
Baseline	62.67	62.67	62.66	62.72	58.99	59.14	47.22	49.23
+ RNA loss	63.13	63.28	59.25	59.27	56.30	56.50	50.35	51.20
+ OGM-GE	62.42	62.69	62.33	62.42	58.83	59.03	51.90	53.54
+ FAGM	64.53	64.51	63.25	63.40	61.02	61.06	54.14	54.90
+SPCL	**67.84**	**68.02**	**66.07**	**66.05**	**66.24**	**66.24**	**54.91**	**55.14**
Δ	*3.31*	*3.51*	*2.82*	*2.65*	*5.22*	*5.18*	*0.77*	*0.24*
Δ_{Base}	*5.17*	*5.35*	*3.41*	*3.33*	*7.25*	*7.10*	*7.69*	*5.91*

[Leveraging Self-Paced Curriculum Learning for Enhanced Modality Balance in Multimodal Conversational Emotion Recognition (NCAA, Q1, IF5.6, Round 2)]

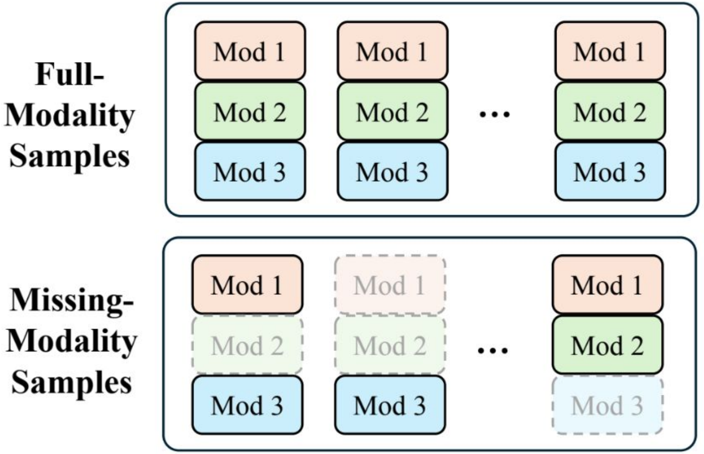
The Presence of Incomplete Modalities

Modality	Demonstration	Possible Reasons
Text	…They act like they are too cool to talk to me…	• Unfamiliar terms • Automated speech recognition fault
Audio		• Background noise • Sensor failure
Video		• Face undetected • Fast motions

Recent Work dealing with Incomplete Modalities

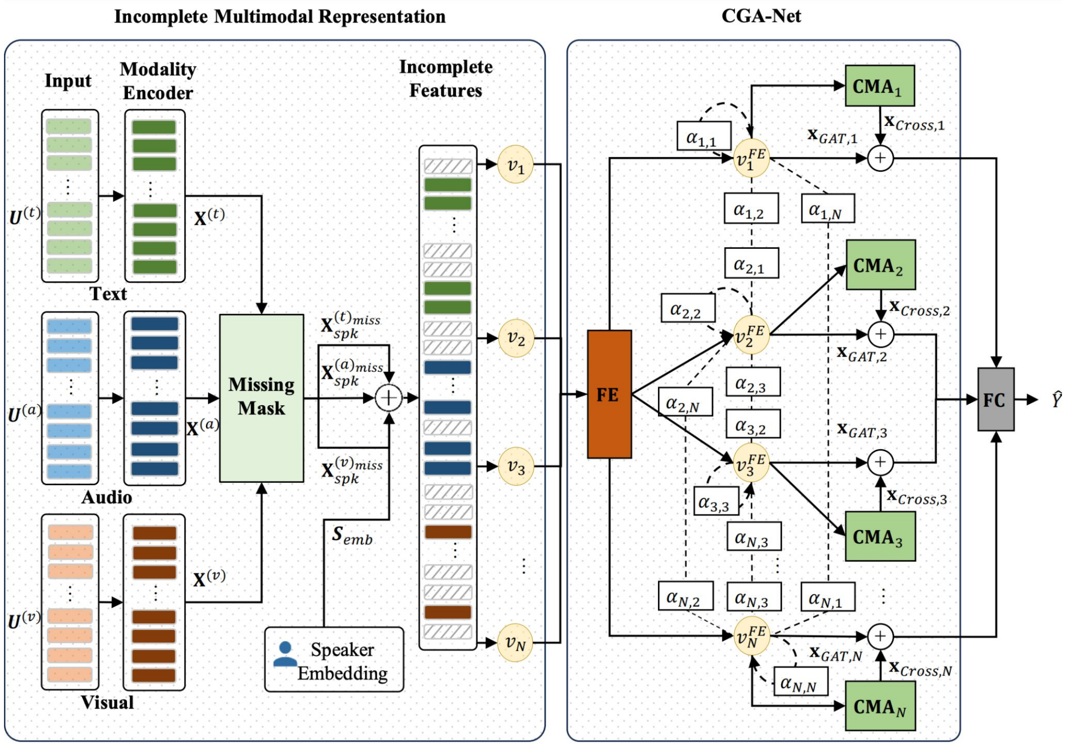


(a) Trends in Missing Modality Publications (2014-2024 October)



(b) Full-Modality Samples vs. Missing-Modality Samples

Mi-CGA



Main idea: Reconstruct incomplete multimodality & pairwise cross-modality feature interactions.

Dataset	Models	Missing Rates 0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	Average
IEMOCAP (4-way)	CPM-Net	58.00	55.29	53.65	52.52	51.01	49.09	47.38	44.76	51.46
	AE	74.82	71.36	67.40	62.02	57.24	50.56	43.04	39.86	58.29
	CRA	76.26	71.28	67.34	62.24	57.04	49.86	43.22	38.56	58.23
	MMIN	74.94	71.84	69.36	66.34	63.30	60.54	57.52	55.44	64.91
	GCNet	78.36	77.48	77.34	76.22	75.14	73.80	71.88	71.38	75.20
	Mi-CGA	**83.42**	**82.83**	**82.27**	**81.50**	**83.17**	**80.08**	**79.96**	**79.35**	**81.50**
	Δ	5.06	5.35	4.93	5.28	8.03	6.28	8.08	7.97	6.30
IEMOCAP (6-way)	CPM-Net	41.05	37.33	36.22	35.73	35.11	33.64	32.26	31.25	35.32
	AE	56.76	52.82	48.66	42.26	35.18	29.12	25.08	23.18	39.13
	CRA	58.68	53.50	49.76	45.88	39.94	32.88	28.08	26.16	41.86
	MMIN	56.96	53.94	51.46	48.42	45.60	42.82	40.18	37.84	47.15
	GCNet	58.64	58.50	57.64	57.08	56.12	54.40	53.60	53.46	56.18
	Mi-CGA	**66.04**	**65.83**	**64.07**	**63.08**	**61.72**	**59.96**	**59.52**	**59.18**	**62.65**
	Δ	7.36	7.33	6.43	6.00	5.60	5.56	5.92	5.72	6.47
CMU-MOSI	CPM-Net	71.90	68.91	71.12	70.59	64.95	65.88	64.02	61.79	67.77
	AE	56.76	52.82	48.66	42.26	35.18	29.12	25.08	23.18	39.13
	CRA	58.68	53.50	49.76	45.88	39.94	32.88	28.08	26.16	41.86
	MMIN	85.20	81.91	78.22	74.60	70.14	67.72	64.04	61.53	72.92
	GCNet	85.01	82.54	80.17	78.54	76.48	73.45	69.46	68.35	76.75
	DiCMoR	85.60	83.90	82.00	80.20	77.70	76.40	73.00	70.08	78.70
	IMDer	85.60	84.80	**83.40**	81.00	78.50	75.90	74.00	71.20	79.30
	Mi-CGA	**87.21**	**85.02**	83.28	**81.83**	**79.56**	**78.62**	**75.63**	**73.05**	**80.05**
	Δ	1.61	0.22	-0.12	0.83	1.06	2.22	1.63	1.85	0.75

Source: Mi-CGA: Cross-Modal Graph Attention Network for Robust Emotion Recognition in the Presence of Incomplete Modalities (Neurocomputing Jan 2025, Q1, IF5.5)

Future Research

- **Missing or noisy modalities**: Real-world scenarios often have incomplete or corrupted modality data (e.g., video with poor lighting)
- **Emotion dynamics over time**: Recognizing transitions, persistence, and context-dependent changes in emotions is still hard.
- **Explainability**: Understanding how each modality contribute to the classification.

Summary

- Foundations of Multimodal Learning
  - What is Multimodal?
  - Multimodal Machine Learning
  - Core Research Challenges
- Emerging Trends in Multimodal Learning
- Applications in Multimodal Emotion Recognition

Thank you

Duc-Trong Le
trongld@vnu.edu.vn

References

- Louis-Philippe Morency, Paul Pu Liang, Amir Zadeh. Tutorials on Multimodal Machine Learning. ICML 2023
- Yuan, Yuan, Zhaojian Li, and Bin Zhao. "A survey of multimodal learning: Methods, applications, and future." ACM Computing Surveys 57.7 (2025): 1-34.
- Zong, Yongshuo, Oisin Mac Aodha, and Timothy M. Hospedales. "Self-supervised multimodal learning: A survey." IEEE Transactions on Pattern Analysis and Machine Intelligence 47.7 (2024): 5299-5318.
- Zhu, Ye, et al. "Vision+ x: A survey on multimodal learning in the light of data." IEEE Transactions on Pattern Analysis and Machine Intelligence 46.12 (2024): 9102-9122.
- Xu, Peng, Xiatian Zhu, and David A. Clifton. "Multimodal learning with transformers: A survey." IEEE Transactions on Pattern Analysis and Machine Intelligence 45.10 (2023): 12113-12132