Multimodal Machine Learning: Principles, Challenges, and Open Questions

Paul Pu Liang
Machine Learning Department
Carnegie Mellon University

https://www.cs.cmu.edu/~pliang/
pliang@cs.cmu.edu
https://github.com/pliang279
@pliang279

Real-world Artificial Intelligence

Digital intelligence

Multimedia

Image/video description

[Rui et al., 1999; Huang 2004]

Physical intelligence

Embodied AI, autonomous driving

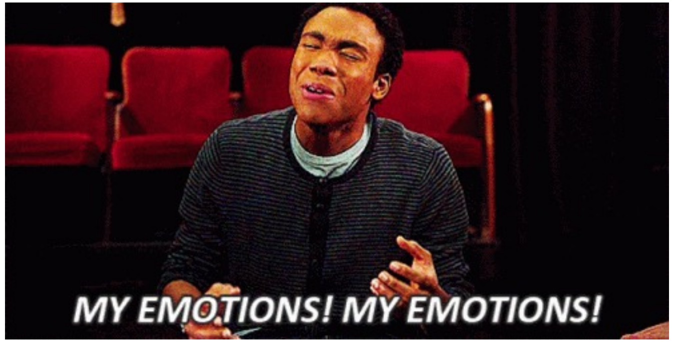
[Xu et al., 2017; Szot et al., 2021]

Social intelligence

Affective computing

Human-AI interaction

[Picard 1997; Jaimes & Sebe 2007]



Multimodal Artificial Intelligence

Digital intelligence

Multimedia

Image/video description

[Rui et al., 1999; Huang 2004]

Physical intelligence

Embodied AI, autonomous driving

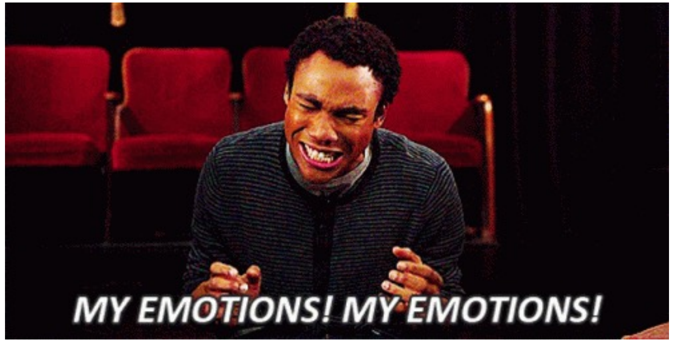
[Xu et al., 2017; Szot et al., 2021]

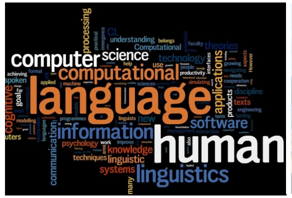
Social intelligence

Affective computing

Human-AI interaction

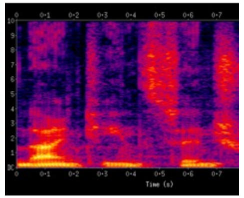
[Picard 1997; Jaimes & Sebe 2007]



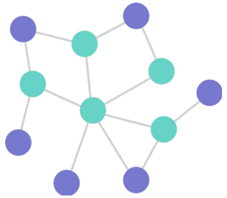


Language (written)

Image

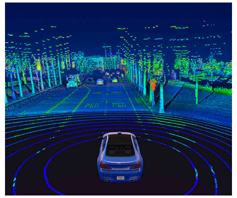


Audio

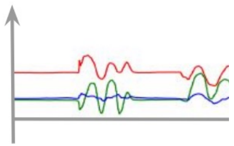


Graphs

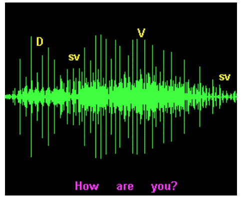
Video



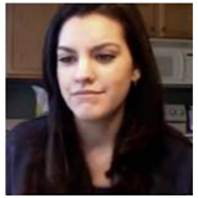
LIDAR



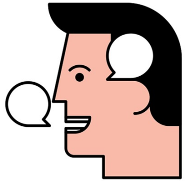
Sensors



Speech



Video (faces)



Language (spoken)

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**

[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Prior Research in Multimodal

Four eras of multimodal research

- The "behavioral" era (1970s until late 1980s)
- The "computational" era (late 1980s until 2000)
- The "interaction" era (2000 - 2010)
- The "deep learning" era (2010s until …)
  - Focus of this talk: last 5 years

[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Behavioral Study of Multimodal

Language
and gestures

David McNeill

“For McNeill, gestures are in effect the speaker’s thought in action, and integral components of speech, not merely accompaniments or additions.”

McGurk effect

[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Behavioral Study of Multimodal

Language
and gestures

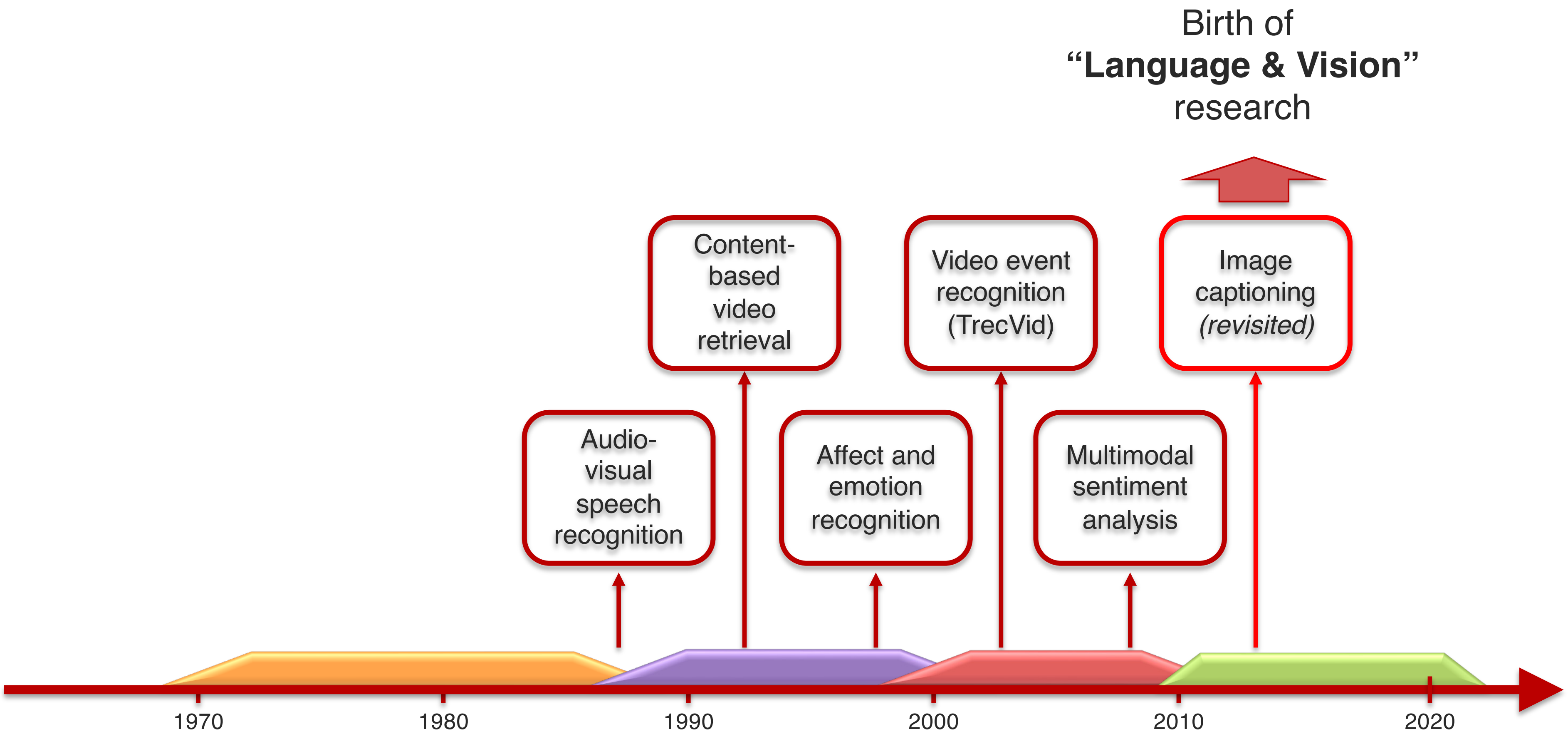
David McNeill

"For McNeill, gestures are in effect the speaker's thought in action, and integral components of speech, not merely accompaniments or additions."

McGurk effect

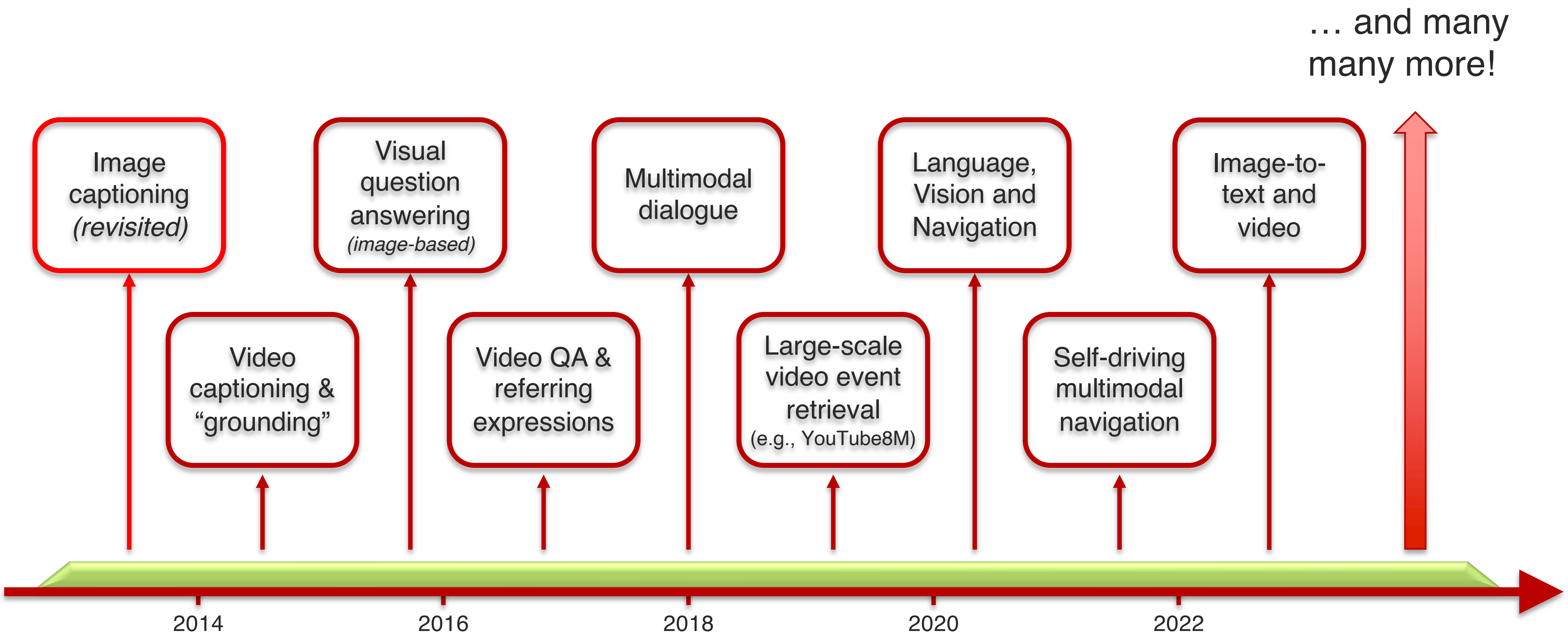
[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Multimodal Research Tasks



[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Multimodal Research Tasks



[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Multimodal ML – Surveys, Tutorials and Courses

2016

Multimodal Machine Learning: A Survey and Taxonomy

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency

(Arxiv 2017, IEEE TPAMI journal, February 2019)

https://arxiv.org/abs/1705.09406

Tutorials: CVPR 2016, ACL 2016, ICMI 2016, …

Graduate-level courses:

Multimodal Machine Learning **(11th edition)**
https://cmu-multicomp-lab.github.io/mmml-course/fall2020/

Advanced Topics in Multimodal Machine Learning
https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2022/

2022

Foundations and Trends in Multimodal ML

Paul Liang, Amir Zadeh, and Louis-Philippe Morency

- 6 core challenges
- 50+ taxonomic classes
- 600+ referenced papers

Tutorials: CVPR 2022, NAACL 2022, …

Updated graduate-level course:

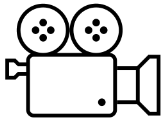
Multimodal Machine Learning **(12th edition)**
https://cmu-multicomp-lab.github.io/mmml-course/fall2022/

[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. arXiv 2022]

What is a Modality?

Modality

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



from a sensor

Raw Modalities
(closest from sensor)

Abstract Modalities
(farthest from sensor)

Examples:

Speech signal	Language	Sentiment intensity
Image	Detected objects	Object categories

[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. CVPR & NAACL 2022 Tutorials]

What is Multimodal?

A dictionary definition…

Multimodal: with multiple modalities

A research-oriented definition…

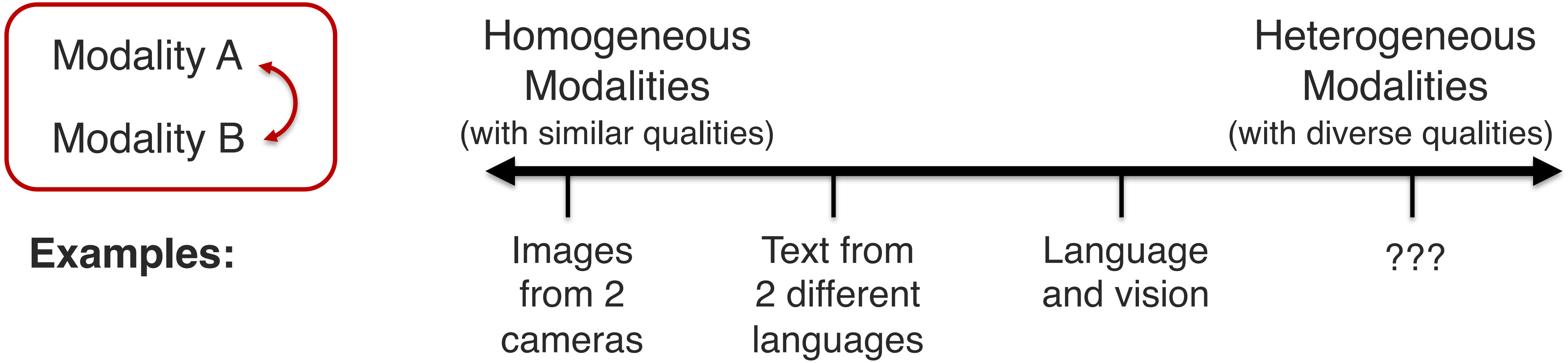
***Multimodal* is the science of**

heterogeneous and interconnected data

[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. CVPR & NAACL 2022 Tutorials]

Heterogeneous Modalities

Information present in different modalities will often show diverse qualities, structures, and representations.



Abstract modalities are more likely to be homogeneous

[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. CVPR & NAACL 2022 Tutorials]

Dimensions of Heterogeneity

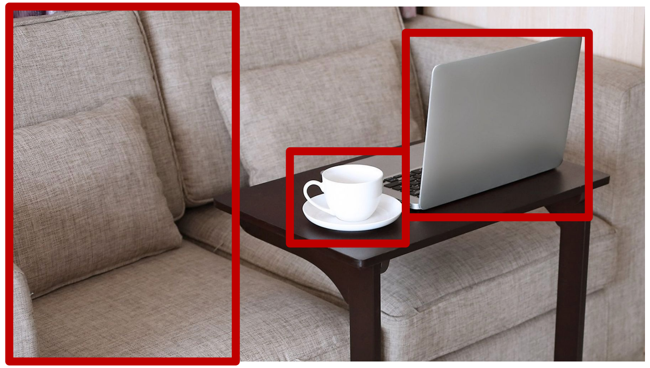
Information present in different modalities will often show diverse qualities, structures, and representations.

A teacup on the right of a laptop in a clean room.

[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. CVPR & NAACL 2022 Tutorials]

Dimensions of Heterogeneity

Information present in different modalities will often show diverse qualities, structures, and representations.



*A **teacup** on the **right** of a **laptop** in a **clean room**.*

1 **Distribution:** discrete or continuous, support

{  ,  ,  }

{teacup, right, laptop, clean, room}

Dimensions of Heterogeneity

Information present in different modalities will often show diverse qualities, structures, and representations.

A teacup on the right of a laptop in a clean room.

2 **Granularity:** sampling rate and frequency

objects per image

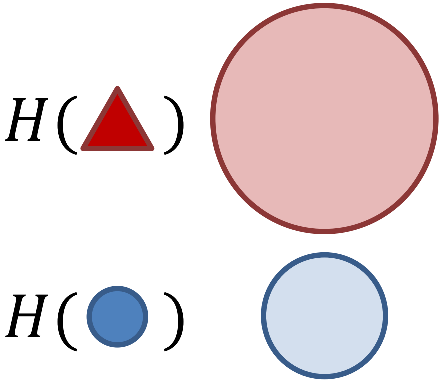
words per minute

Dimensions of Heterogeneity

Information present in different modalities will often show diverse qualities, structures, and representations.

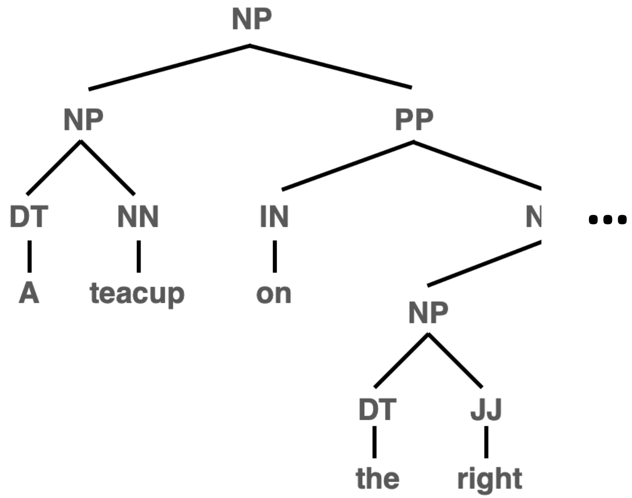
A teacup on the right of a laptop in a clean room.

3 **Information:** entropy and density

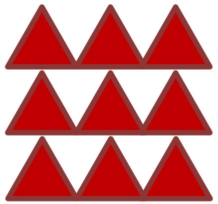


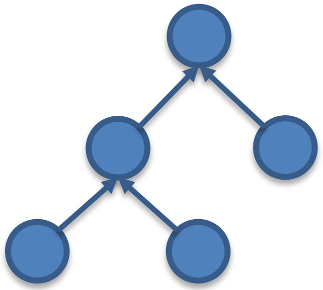
Dimensions of Heterogeneity

Information present in different modalities will often show diverse qualities, structures, and representations.



4 **Structure:** static, temporal, spatial, hierarchical





Dimensions of Heterogeneity

Information present in different modalities will often show diverse qualities, structures, and representations.

A teacup on the right of a laptop in a clean room.

5 **Noise:** uncertainty, signal-to-noise ratio, missing data

teacup → teacip

right → rihjt

Dimensions of Heterogeneity

Information present in different modalities will often show diverse qualities, structures, and representations.

A teacup on the right of a laptop in a clean room.

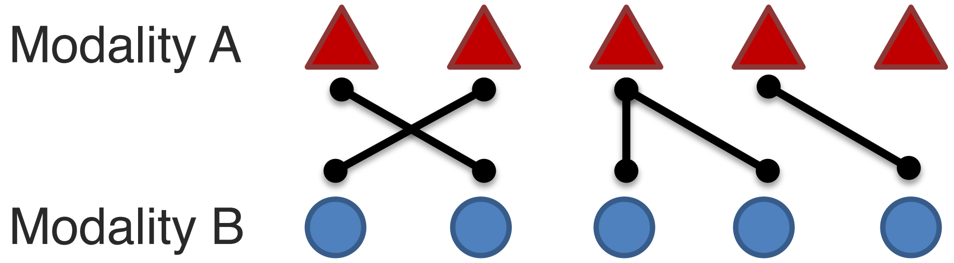
6 **Relevance:** task relevance, context dependence

→ recreational
→ living room
→ right-handed

A teacup on the right of a laptop in a clean room.

→ workspace
→ study room

Interconnected Modalities



Modalities are often related and share complementary information that interact

① **Connections**

Which elements are connected and why?

② **Cross-modal interactions**

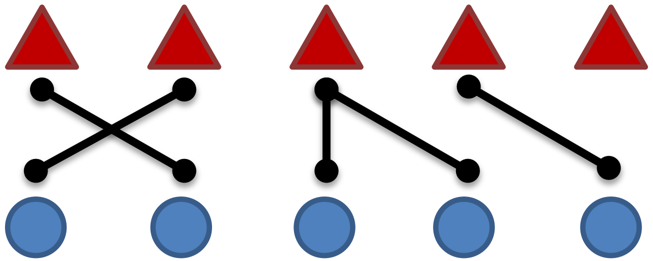
How are connected elements interacting during inference?

[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. CVPR & NAACL 2022 Tutorials]

Interconnected Modalities

Modality A

Modality B



① **Connections**

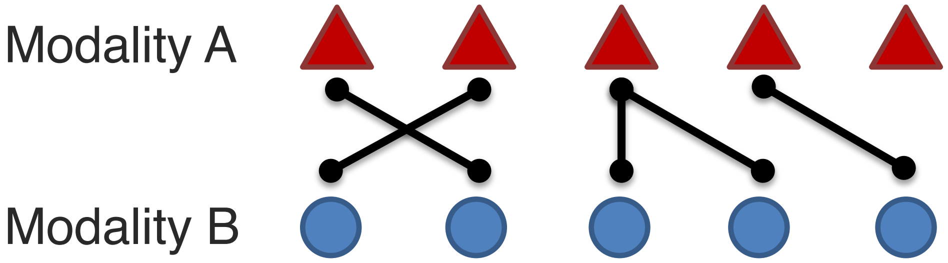
Which elements are connected and why?

A teacup on the right of a laptop in a clean room.

teacup

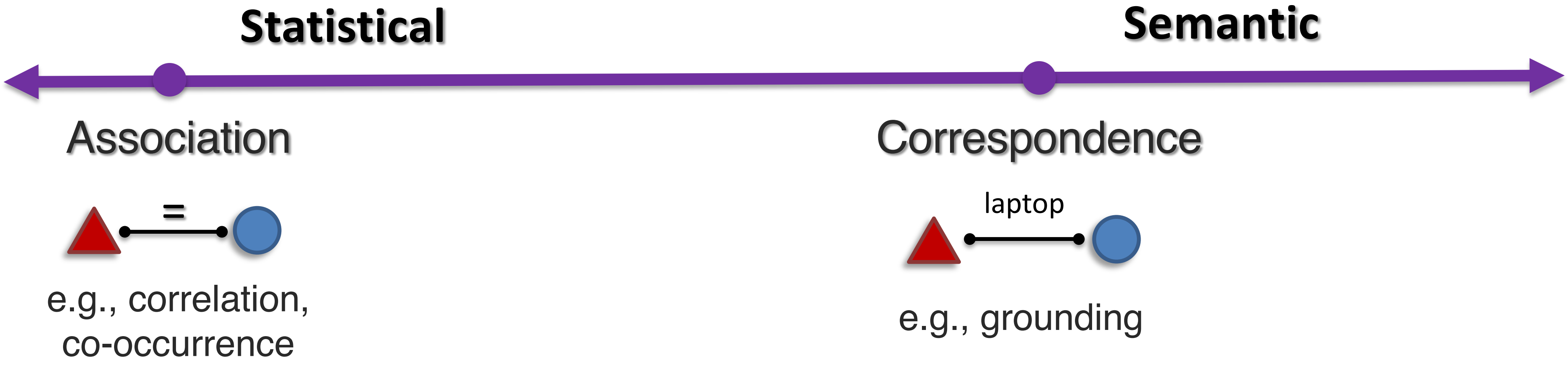
laptop

Interconnected Modalities



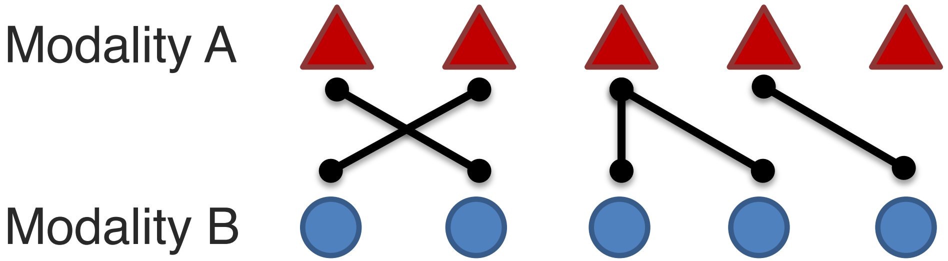
① **Connections**

Which elements are connected and why?



[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. CVPR & NAACL 2022 Tutorials]

Interconnected Modalities



1 **Connections**

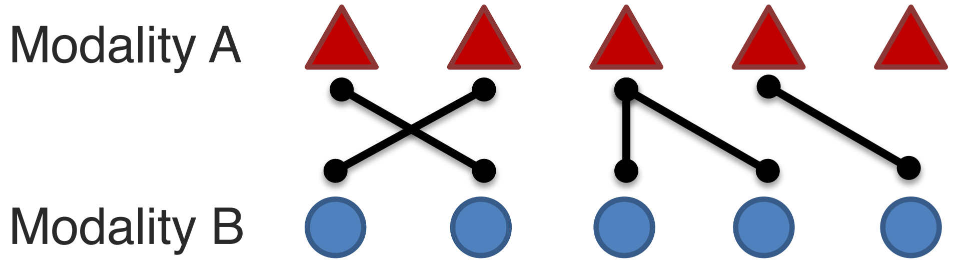
Which elements are connected and why?

A teacup on the right of a laptop in a clean room.

clean

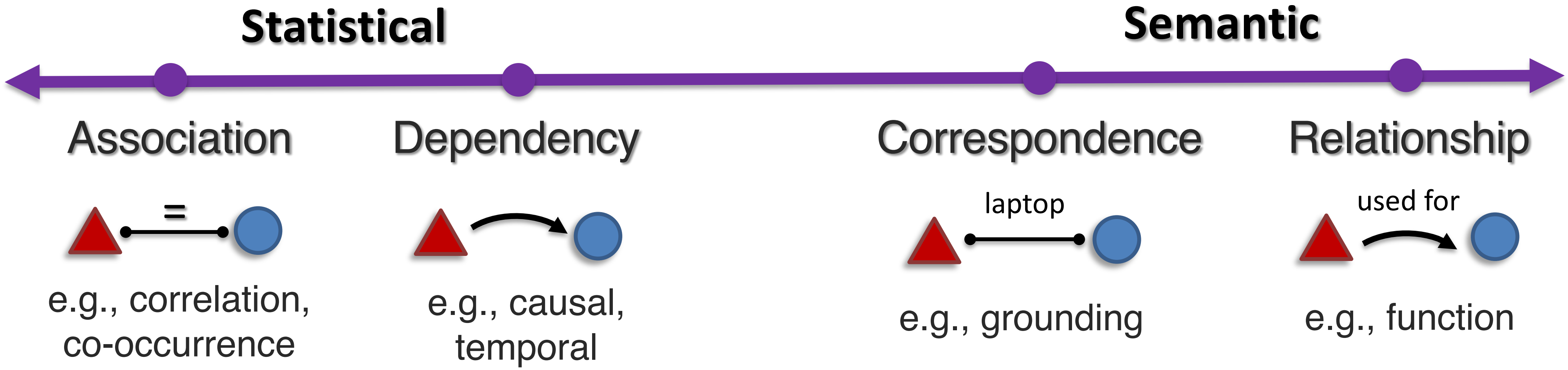
room

Interconnected Modalities



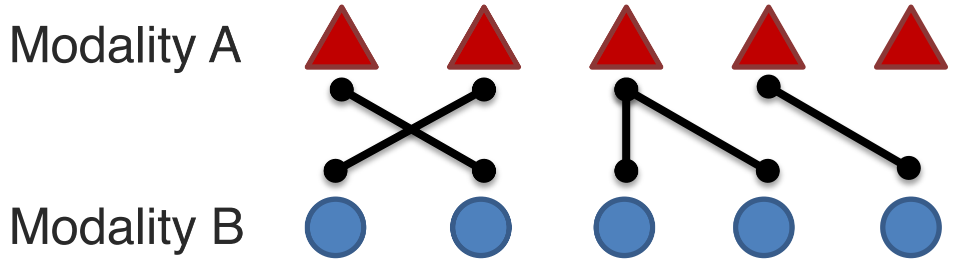
① **Connections**

Which elements are connected and why?



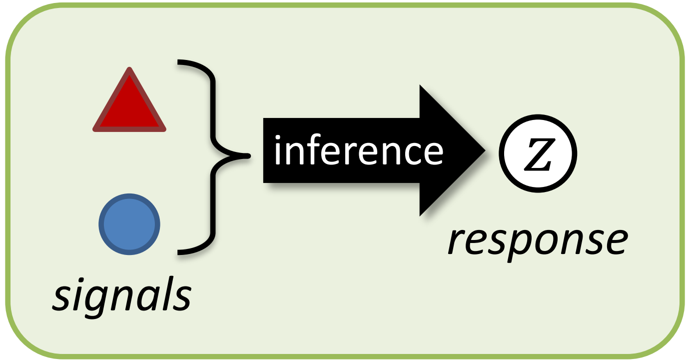
[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. CVPR & NAACL 2022 Tutorials]

Interconnected Modalities



① **Connections**

Which elements are connected and why?



Interactions happen during inference!

② **Cross-modal interactions**

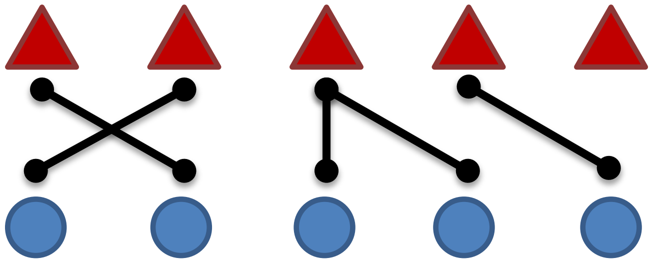
How are connected elements interacting during inference?

[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. CVPR & NAACL 2022 Tutorials]

Interconnected Modalities

Modality A

Modality B



② **Cross-modal interactions**

How are connected elements interacting during inference?

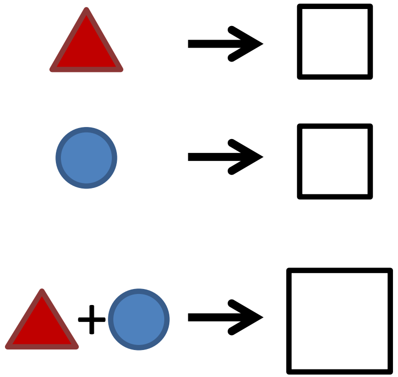
Is this indoors?

inference → *Yes!*

A teacup on the right of a laptop in a clean room.

inference → *Yes!*

Redundant



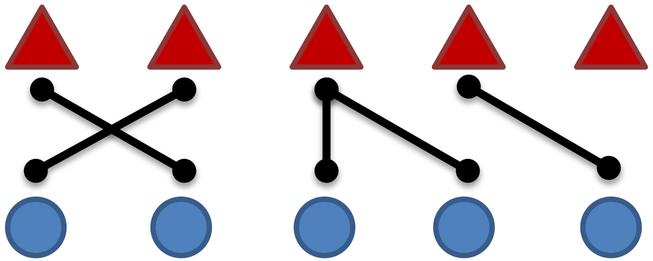
Enhancement

[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. CVPR & NAACL 2022 Tutorials]

Interconnected Modalities

Modality A

Modality B



Is this indoors?

A teacup on the right of a laptop in a clean room.

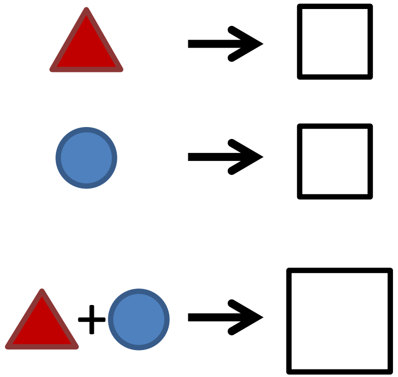
② **Cross-modal interactions**

How are connected elements interacting during inference?

inference → ***Yes!***

inference → ***Yes!***

Redundant



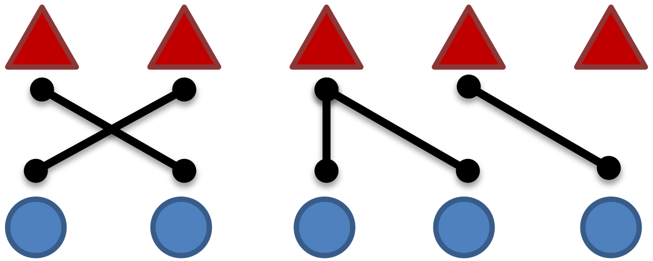
Enhancement

[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. CVPR & NAACL 2022 Tutorials]

Interconnected Modalities

Modality A

Modality B



② **Cross-modal interactions**

How are connected elements interacting during inference?

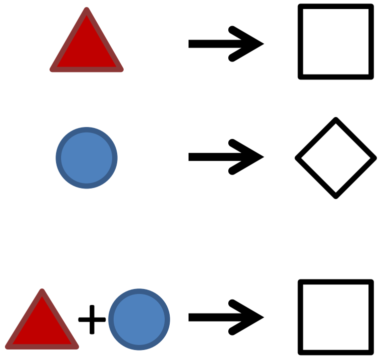
Is this a living room?

A teacup on the right of a laptop in a clean room.

inference → *Yes!*

inference → *No, probably study room.*

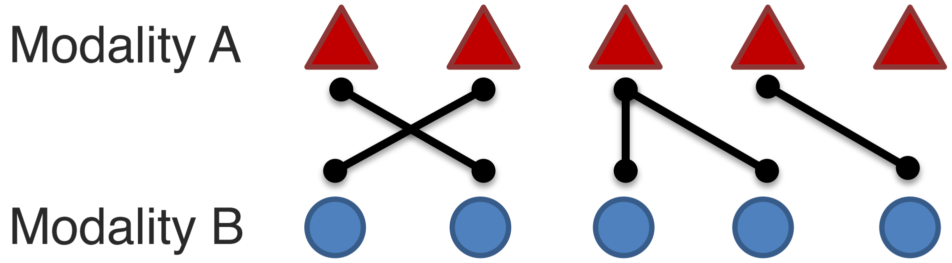
Non-redundant



Dominance

[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. CVPR & NAACL 2022 Tutorials]

Interconnected Modalities



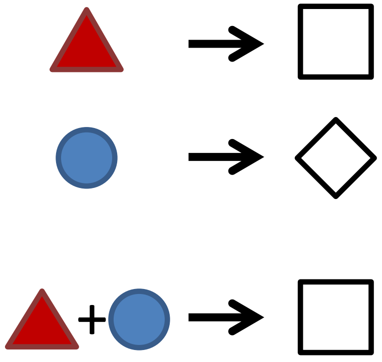
② Cross-modal interactions

How are connected elements interacting during inference?

Is this a living room?

inference → ***Yes!***

Non-redundant



A teacup on the right of a laptop in a clean room.

inference → *No, probably study room.*

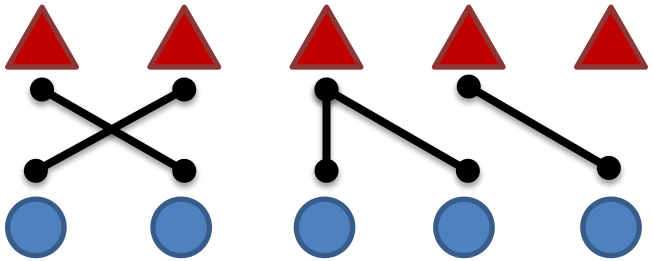
Dominance

[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. CVPR & NAACL 2022 Tutorials]

Interconnected Modalities

Modality A

Modality B



Should I work here?

A teacup on the right of a laptop in a clean room.

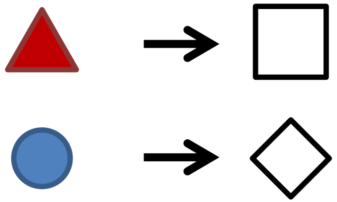
② Cross-modal interactions

How are connected elements interacting during inference?

inference

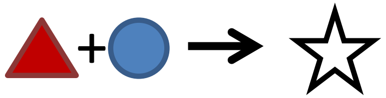
Maybe? Comfy sofa but table's too small.

Non-redundant



inference

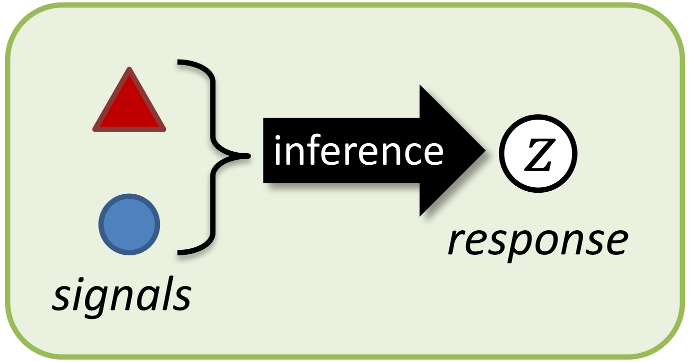
Maybe? Clean and there's tea.



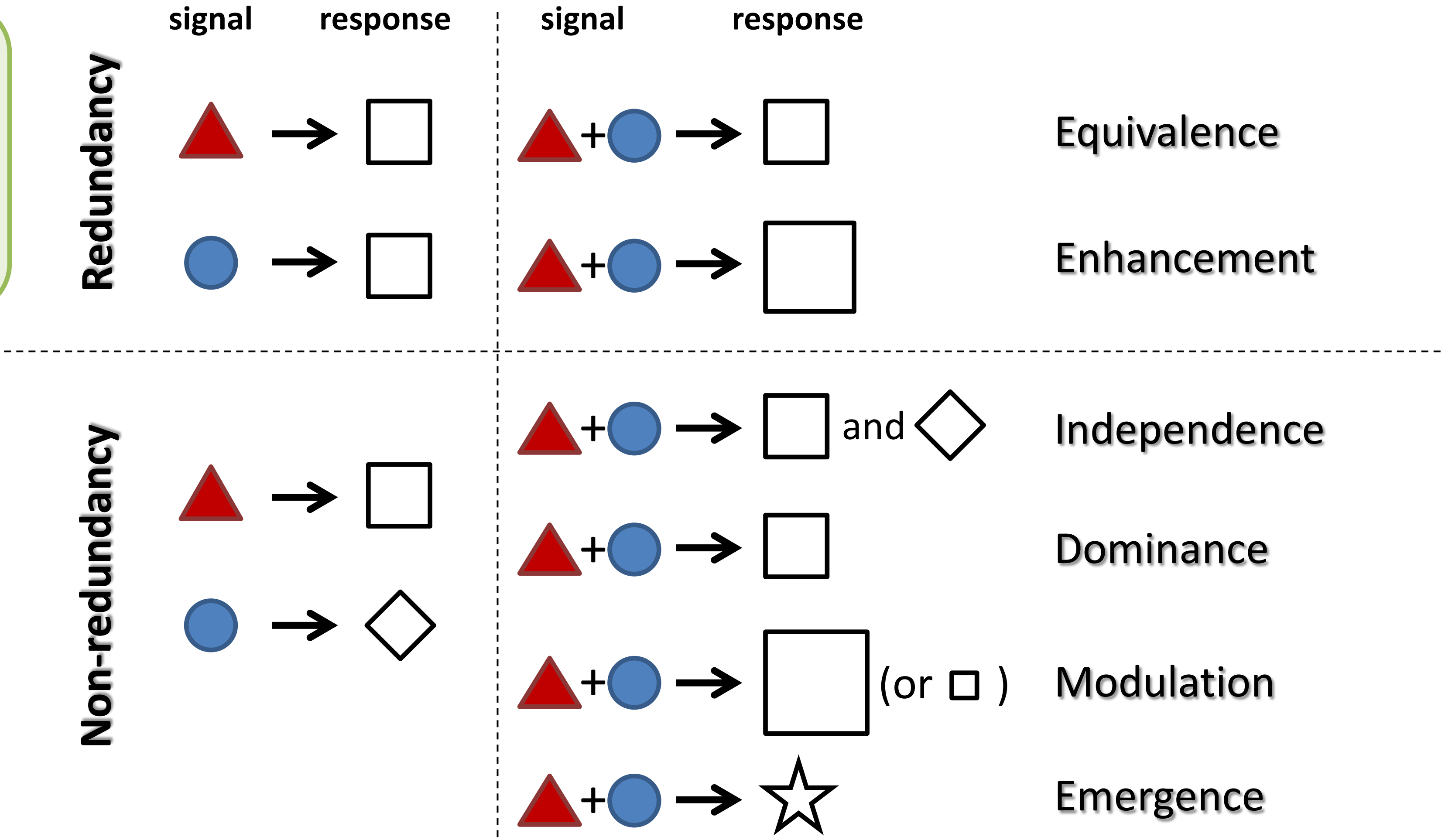
Emergence

[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. CVPR & NAACL 2022 Tutorials]

Cross-modal Interactions – A Behavioral Science View

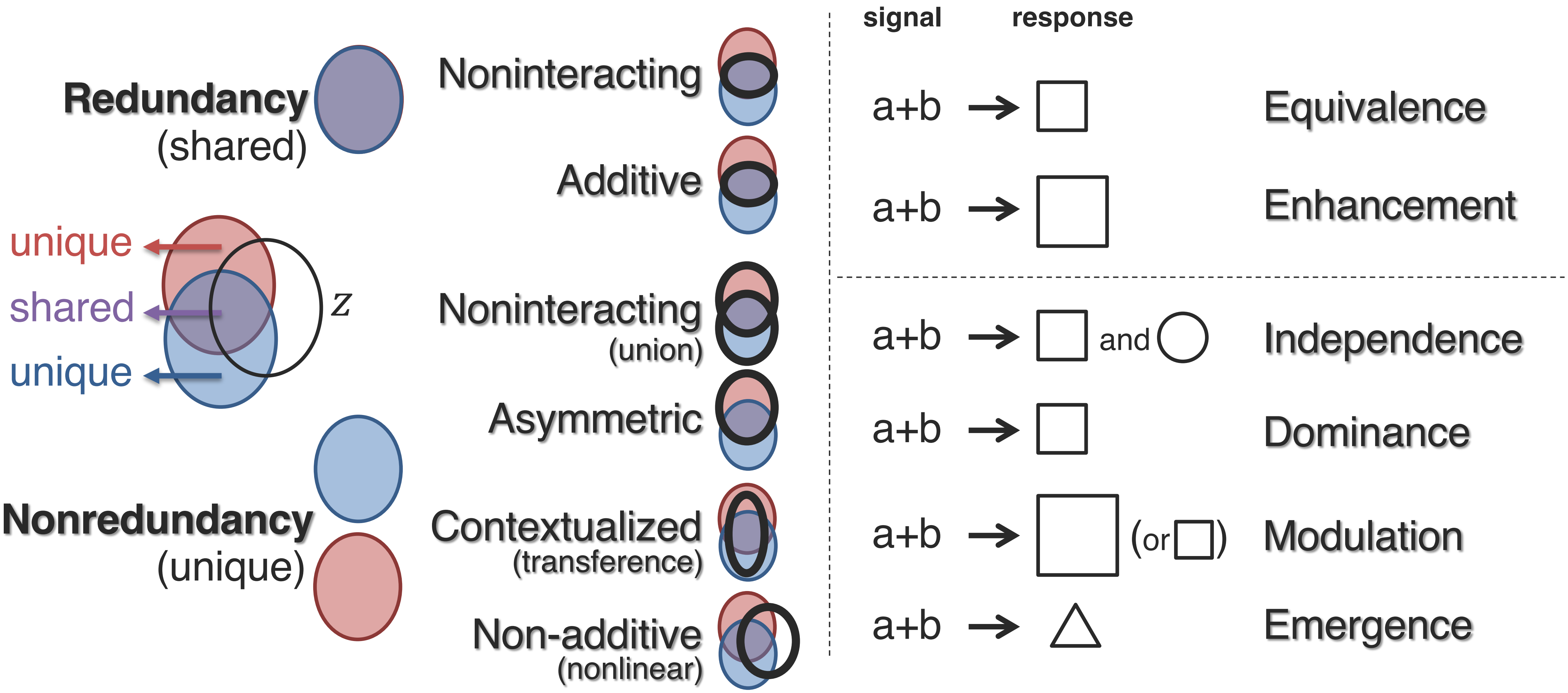


Multimodal Communication



[Partan and Marler. Communication Goes Multimodal. Science 1999]

Cross-modal Interaction Mechanics



[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. CVPR & NAACL 2022 Tutorials]

What is Multimodal? → *Why is it hard?* → *What is next?*

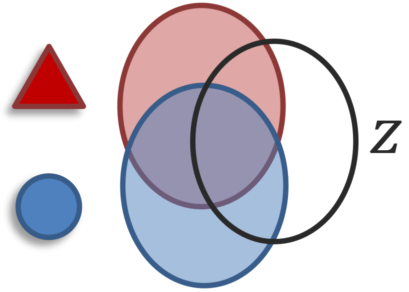
Heterogeneous

+

Connected

+

Interacting



Multimodal is the scientific study of heterogeneous and interconnected data ☺

Multimodal Machine Learning

Language

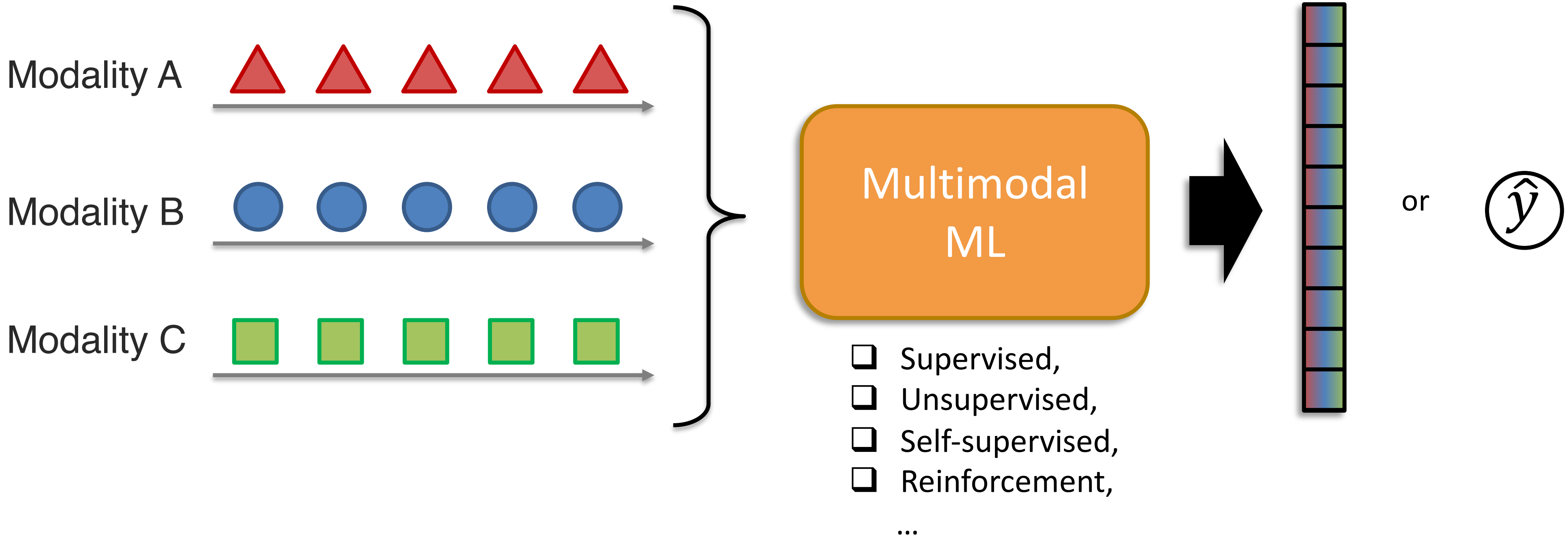
I really like this talk

Vision

Acoustic

[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. CVPR & NAACL 2022 Tutorials]

Multimodal Machine Learning



[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. CVPR & NAACL 2022 Tutorials]

Multimodal Machine Learning

What are the core multimodal technical challenges, understudied in conventional machine learning?

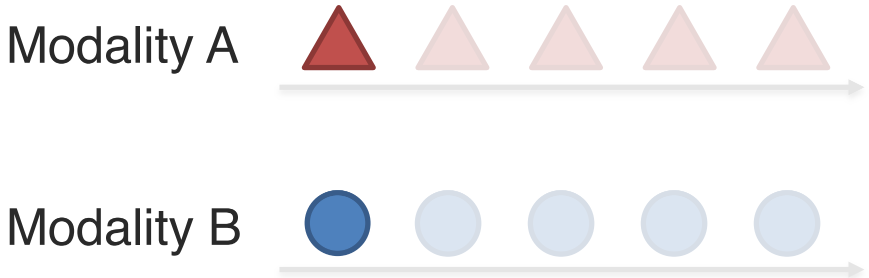
[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. CVPR & NAACL 2022 Tutorials]

Challenge 1: Representation

Definition: Learning representations that reflect cross-modal interactions between individual elements, across different modalities

➡ This is a core building block for most multimodal modeling problems!

Individual elements:

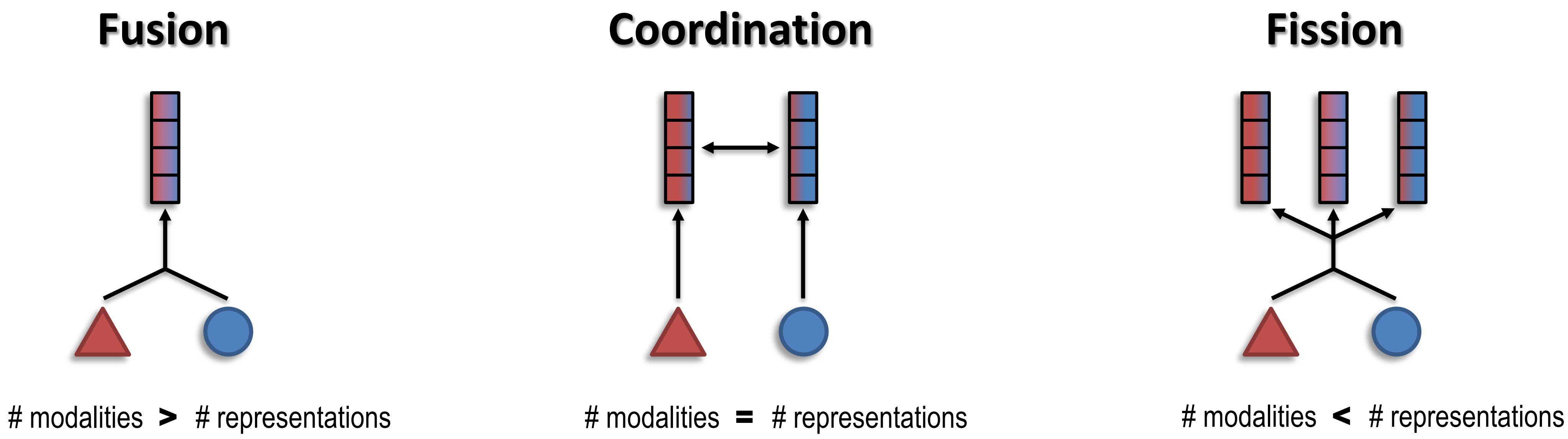


[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. CVPR & NAACL 2022 Tutorials]

Challenge 1: Representation

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

Sub-challenges:



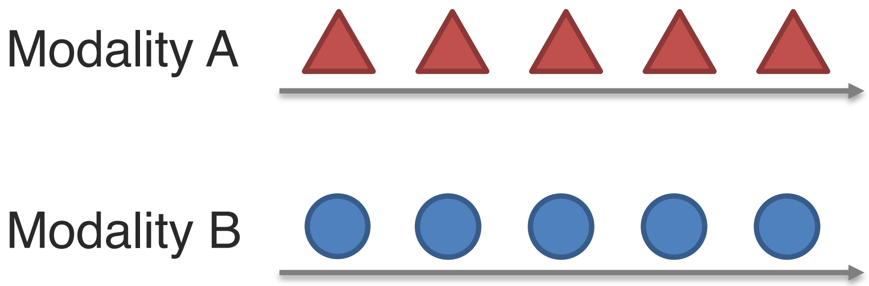
[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. CVPR & NAACL 2022 Tutorials]

Challenge 2: Alignment

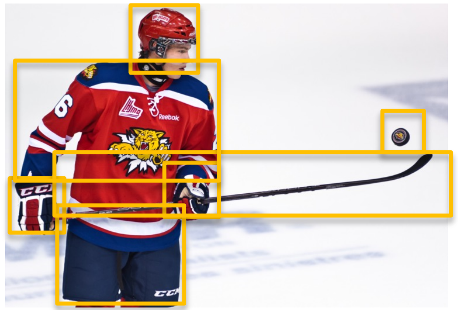
Definition: Identifying and modeling cross-modal connections between all elements of multiple modalities, building from the data structure.

➡ Most modalities have internal structure with multiple elements

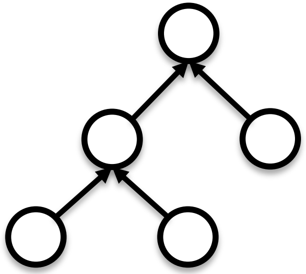
Elements with temporal structure:



Other structured examples:



Spatial



Hierarchical

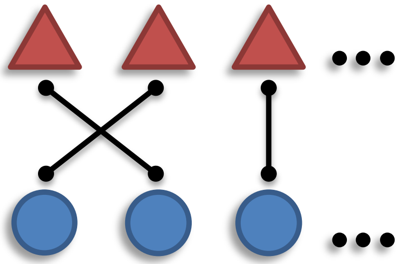
[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. CVPR & NAACL 2022 Tutorials]

Challenge 2: Alignment

Definition: Identifying and modeling cross-modal connections between all elements of multiple modalities, building from the data structure.

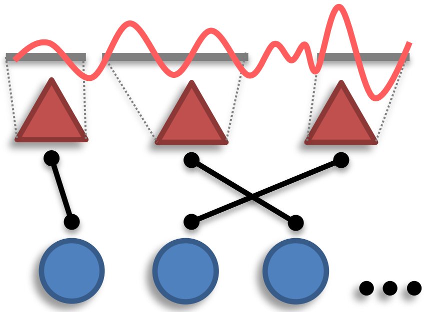
Sub-challenges:

Discrete connections



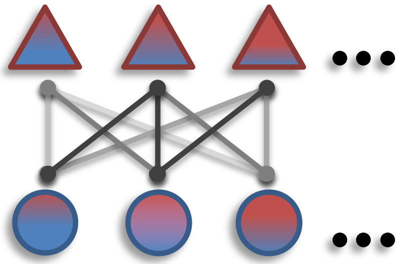
Explicit alignment
(e.g., grounding)

Continuous alignment



Granularity of
individual elements

Contextualized representation

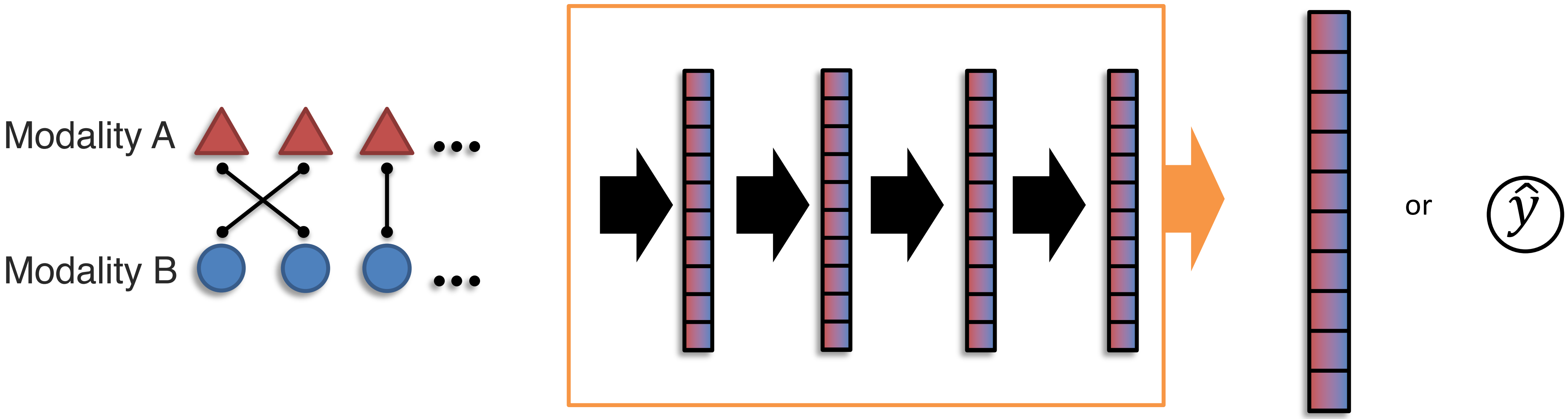


Implicit alignment
+ representation

[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. CVPR & NAACL 2022 Tutorials]

Challenge 3: Reasoning

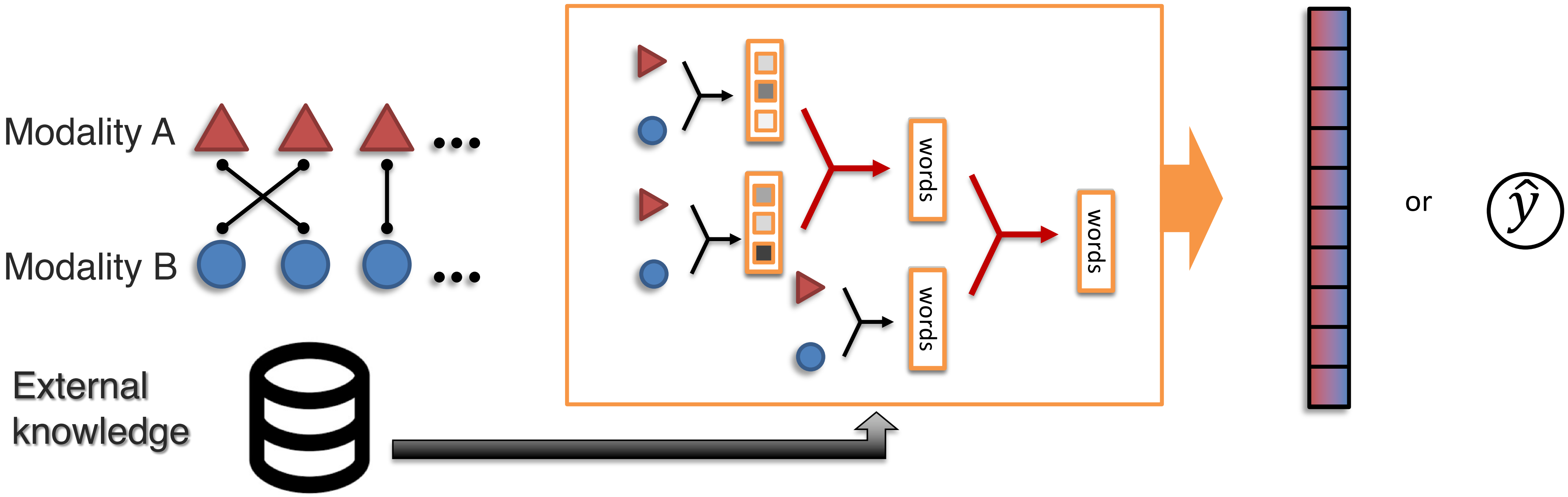
Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure.



[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. CVPR & NAACL 2022 Tutorials]

Challenge 3: Reasoning

Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure.



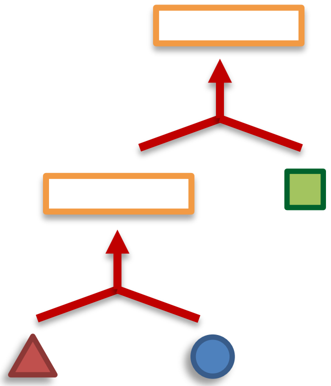
[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. CVPR & NAACL 2022 Tutorials]

Challenge 3: Reasoning

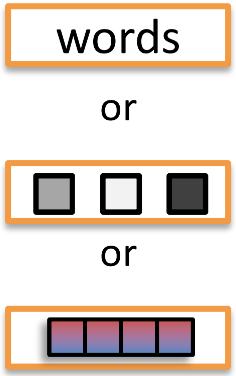
Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure.

Sub-challenges:

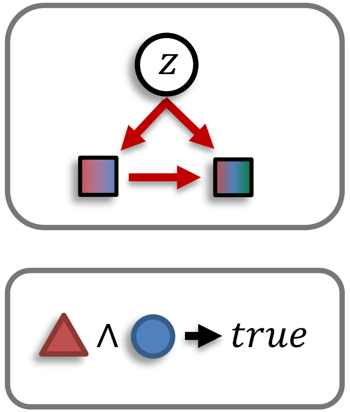
Structure modeling



Intermediate concepts



Inference paradigm



External knowledge

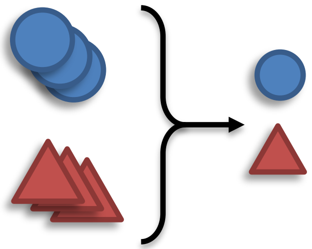
[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. CVPR & NAACL 2022 Tutorials]

Challenge 4: Generation

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

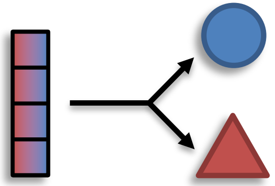
Sub-challenges:

Summarization



Translation

Creation



Information: (content)

Reduction: □ > □

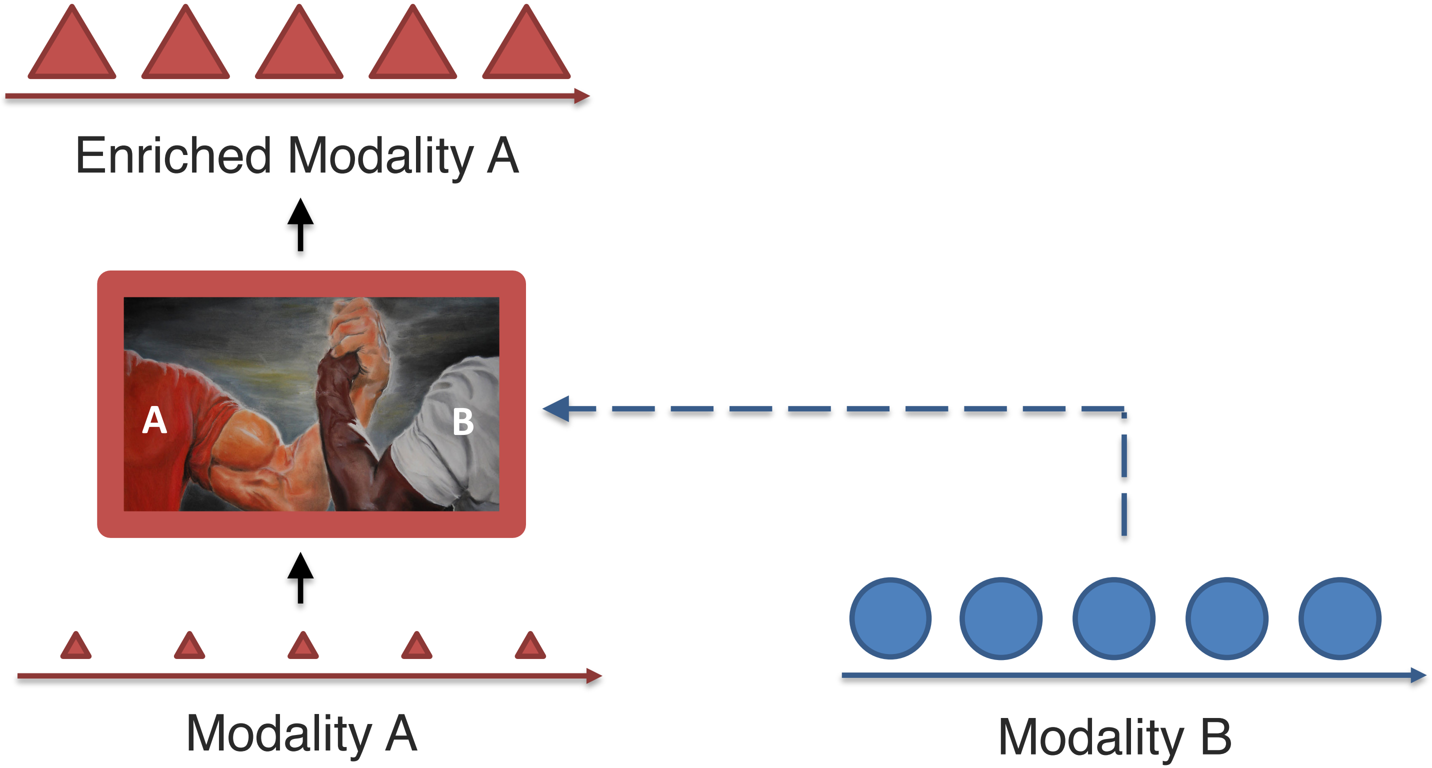
Maintenance: □ = □

Expansion: □ < □

[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. CVPR & NAACL 2022 Tutorials]

Challenge 5: Transference

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



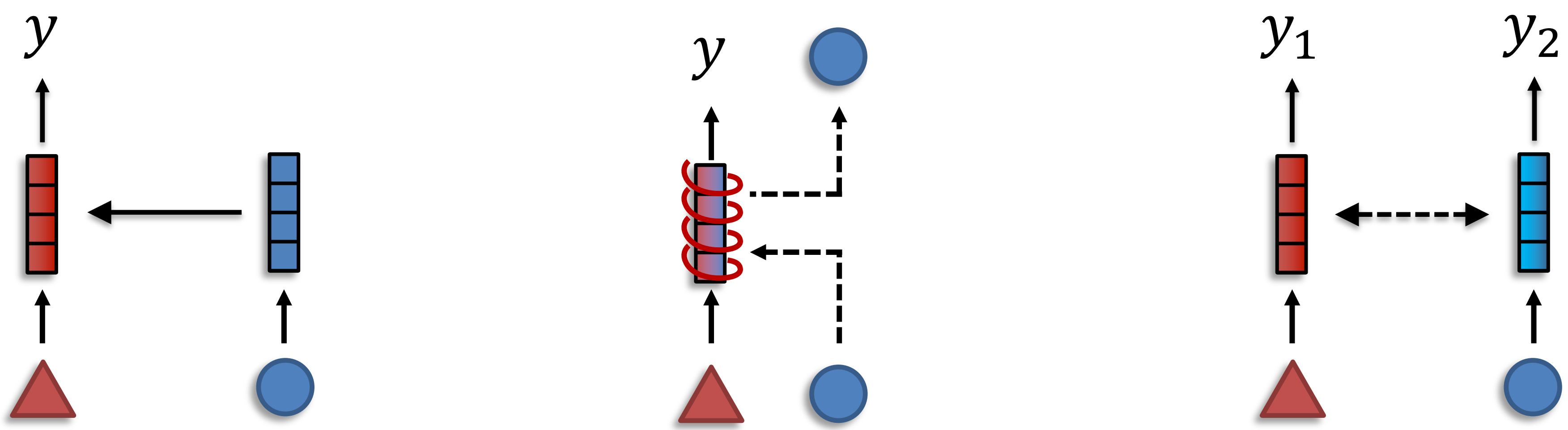
[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. CVPR & NAACL 2022 Tutorials]

Challenge 5: Transference

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

Sub-challenges:

Transfer **Co-learning** **Model Induction**



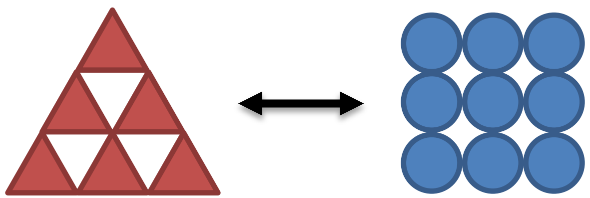
[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. CVPR & NAACL 2022 Tutorials]

Challenge 6: Quantification

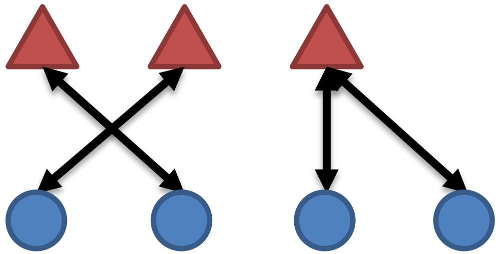
Definition: Empirical and theoretical study to better understand heterogeneity, cross-modal interactions, and the multimodal learning process.

Sub-challenges:

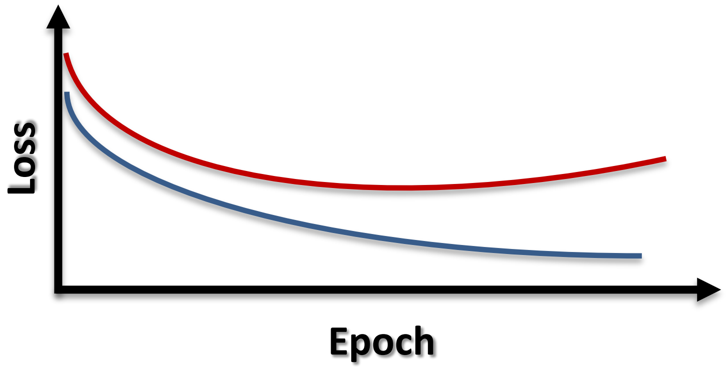
Heterogeneity



Interactions

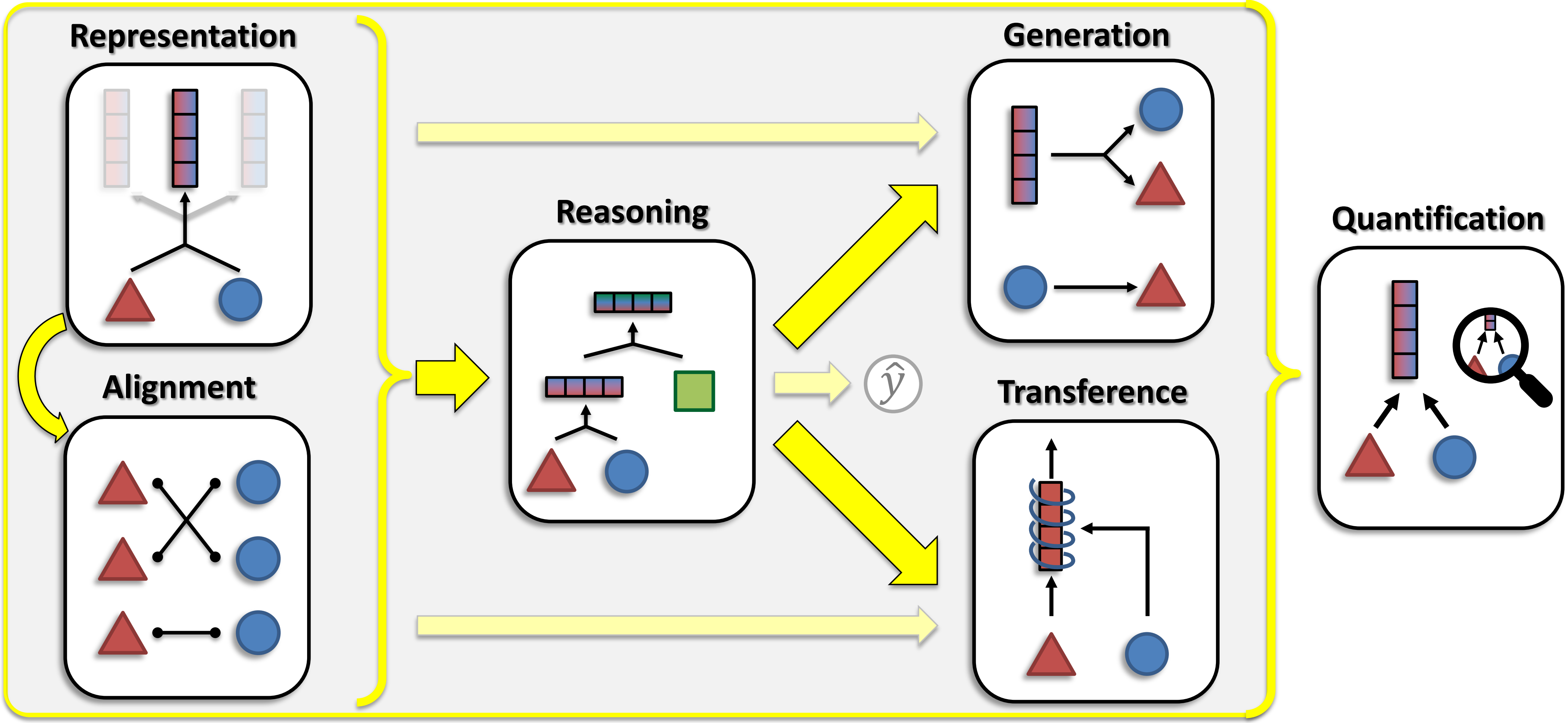


Learning



[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. CVPR & NAACL 2022 Tutorials]

Core Multimodal Challenges



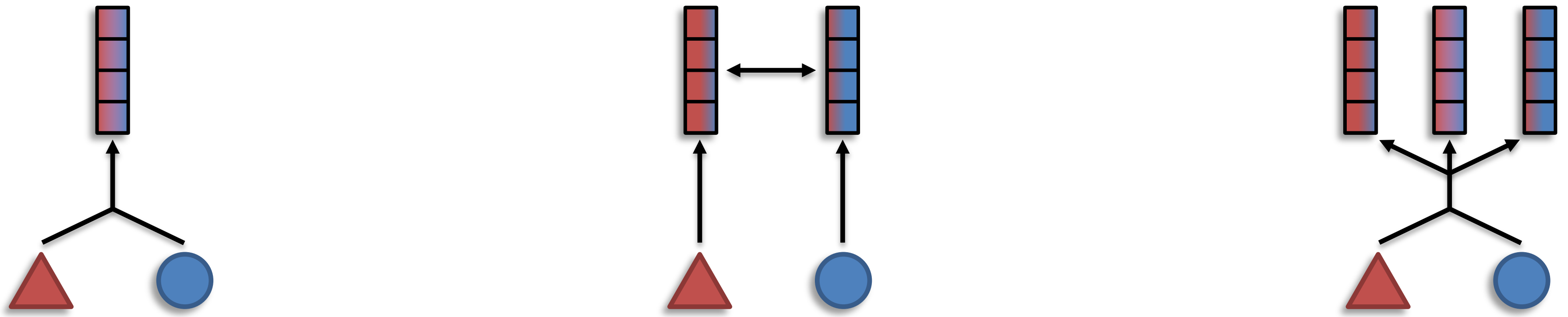
Challenge 1: Representation

Challenge 1: Representation

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

Sub-challenges:

Fusion **Coordination** **Fission**



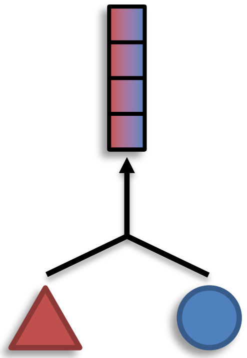
modalities > # representations

modalities = # representations

modalities < # representations

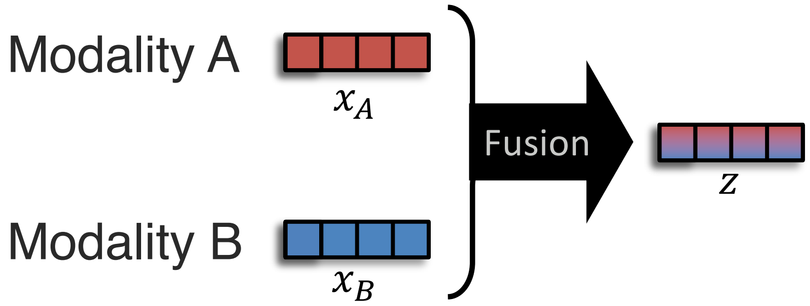
[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. CVPR & NAACL 2022 Tutorials]

Sub-Challenge 1a: Representation Fusion



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

Concepts for Representation Fusion



Goal: Model *cross-modal interactions* between the multimodal elements

Let's study the univariate case first

(only 1-dimensional features)

Linear regression:

$$z = \underbrace{w_0}_{\text{constant}} + \underbrace{w_1 x_A + w_2 x_B}_{\text{Additive terms}} + \underbrace{w_3 (x_A \times x_b)}_{\text{Multiplicative term}} + \underbrace{\epsilon}_{\text{error}}$$

1. Additive interaction:

$$z = w_1 x_A + w_2 x_B + \epsilon$$

2. Multiplicative interaction:

$$z = w_3 (x_A \times x_b) + \epsilon$$

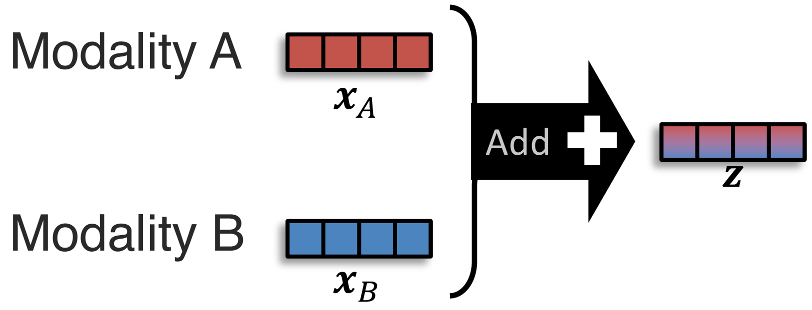
3. Additive and multiplicative interactions:

$$z = w_1 x_A + w_2 x_B + w_3 (x_A \times x_b) + \epsilon$$

Additive Fusion

Back to multivariate case!

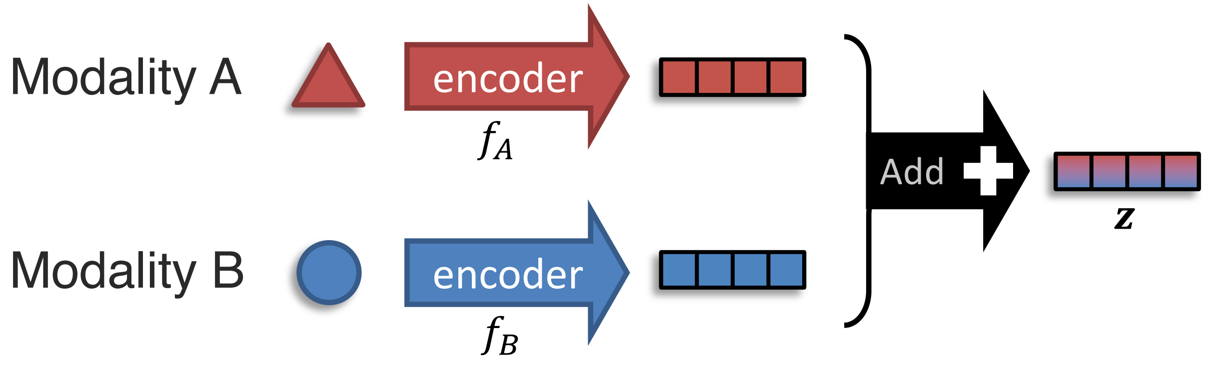
(multi-dimensional features)



Additive fusion:

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

With unimodal encoders:

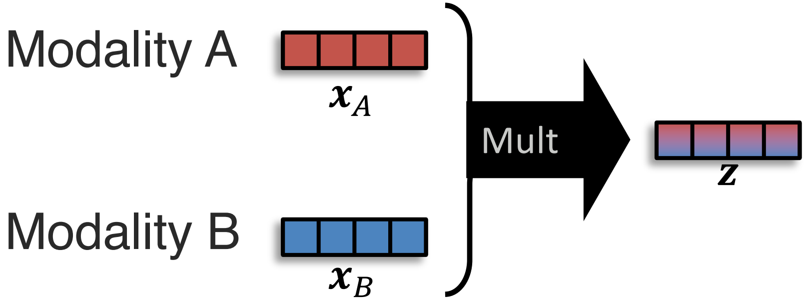


Additive fusion:

$$\boldsymbol{z} = f_A(\triangle) + f_B(\bigcirc)$$

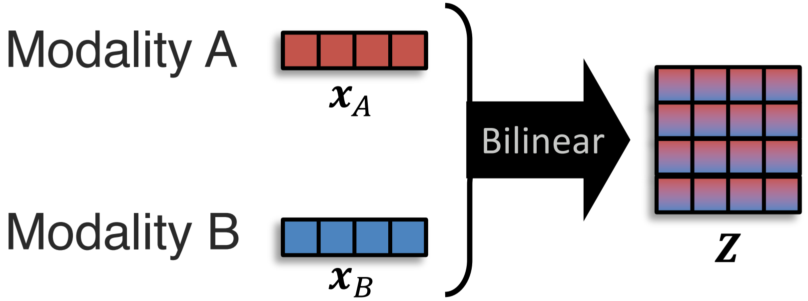
It could be seen as an ensemble approach (late fusion)

Multiplicative Fusion



Multiplicative fusion:

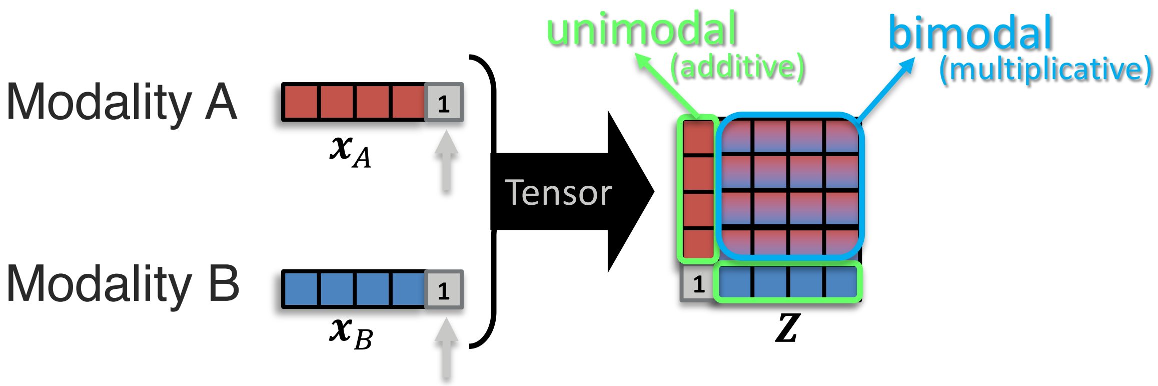
$$z = w(x_A \times x_B)$$



Bilinear Fusion:

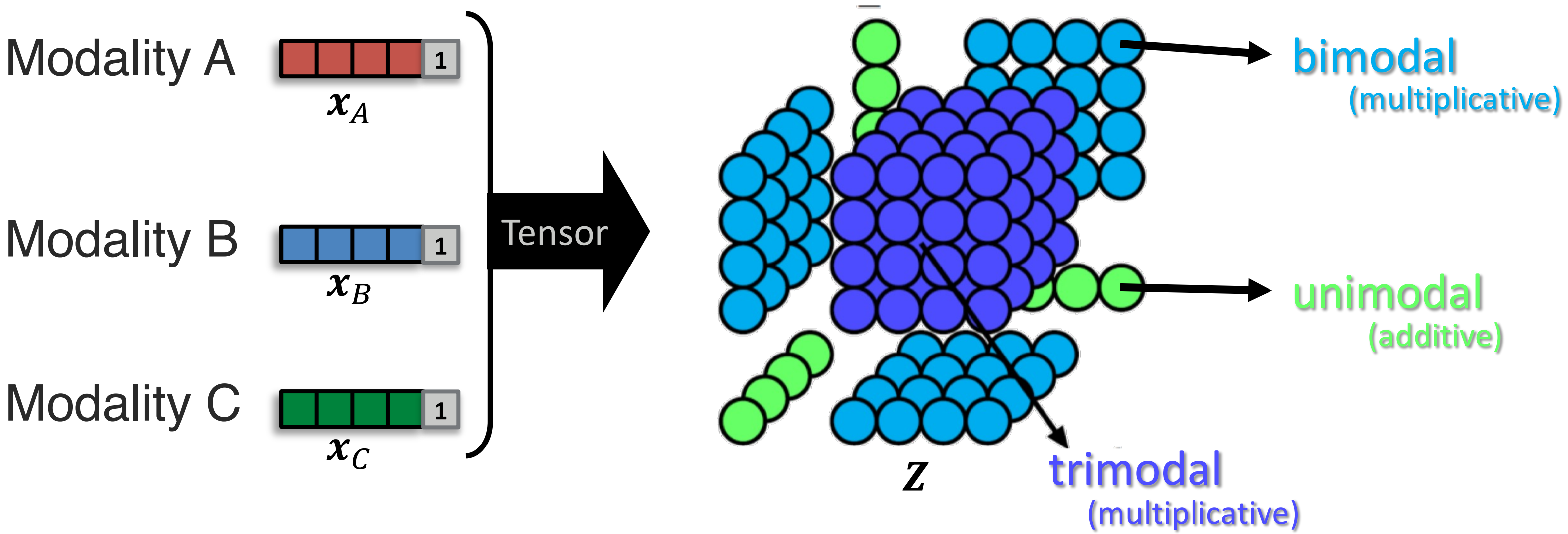
$$Z = w(x_A^T \cdot x_B)$$

Tensor Fusion



Tensor Fusion (bimodal):

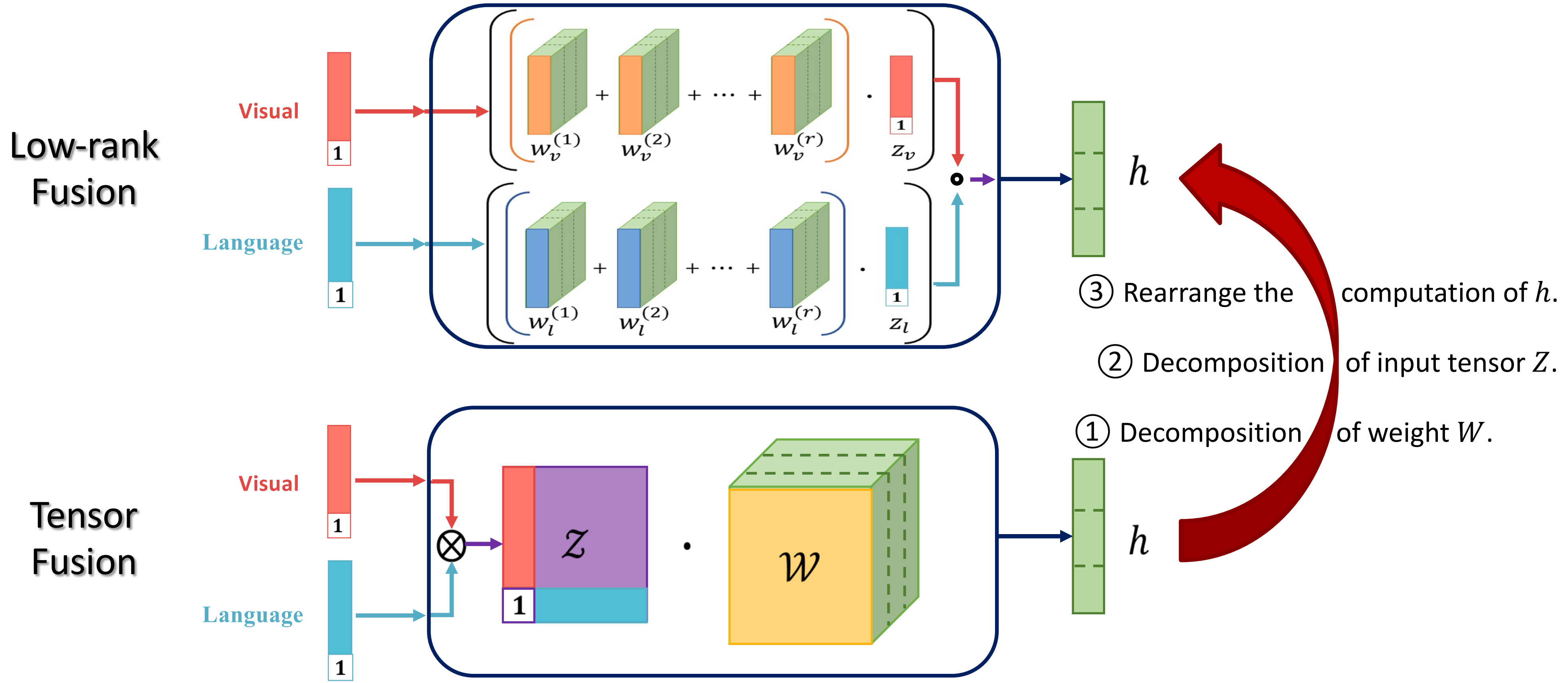
$$\boldsymbol{Z} = \boldsymbol{w}([\boldsymbol{x}_A \quad 1]^T \cdot [\boldsymbol{x}_B \quad 1])$$



... but the weight matrix may end up quite large!

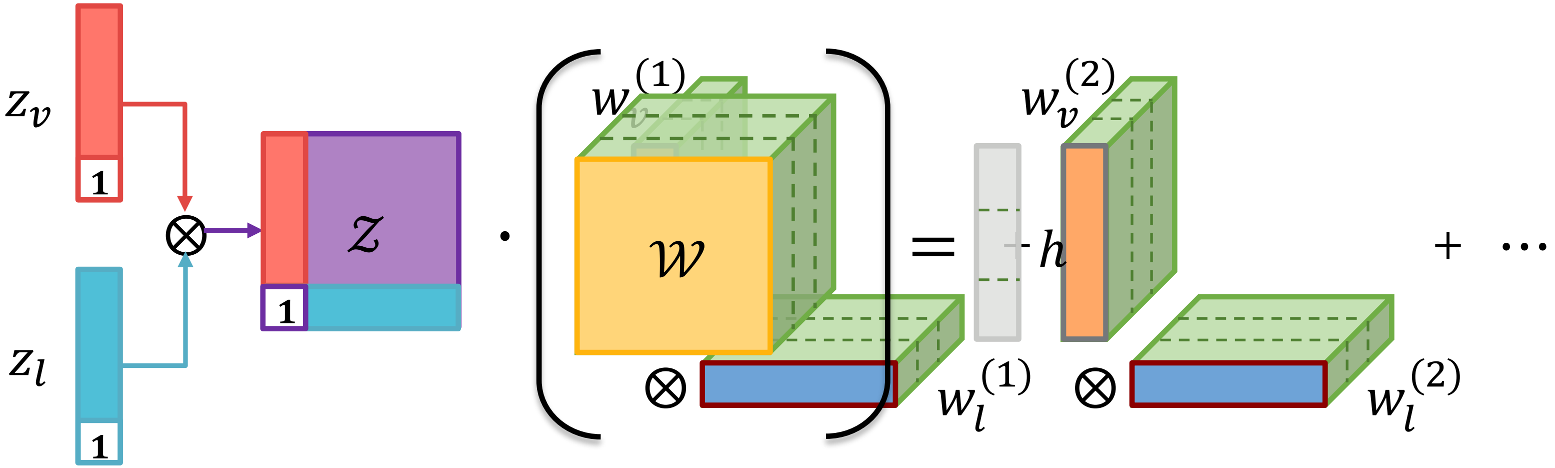
[Zadeh et al., Tensor Fusion Network for Multimodal Sentiment Analysis. EMNLP 2017]

Low-rank Fusion



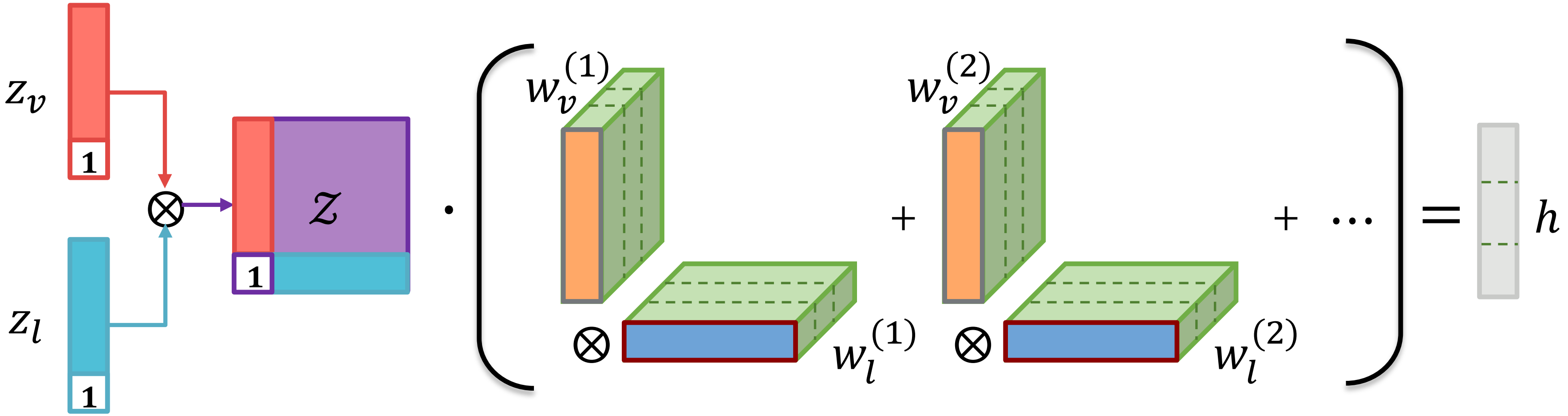
[Liu et al., Efficient Low-rank Multimodal Fusion with Modality-Specific Factors. ACL 2018]

Low-rank Fusion



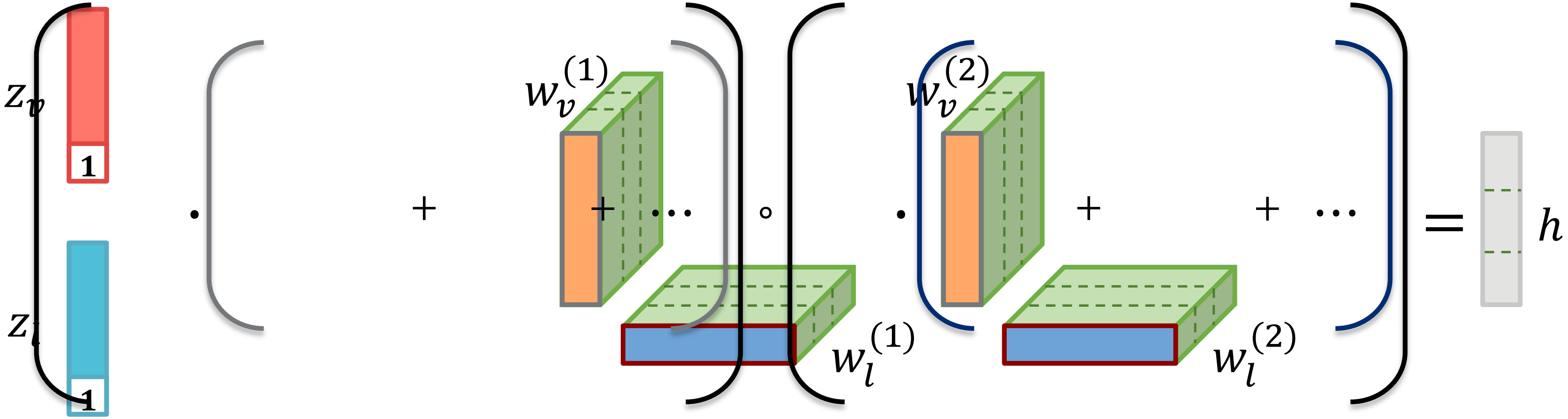
[Liu et al., Efficient Low-rank Multimodal Fusion with Modality-Specific Factors. ACL 2018]

Low-rank Fusion



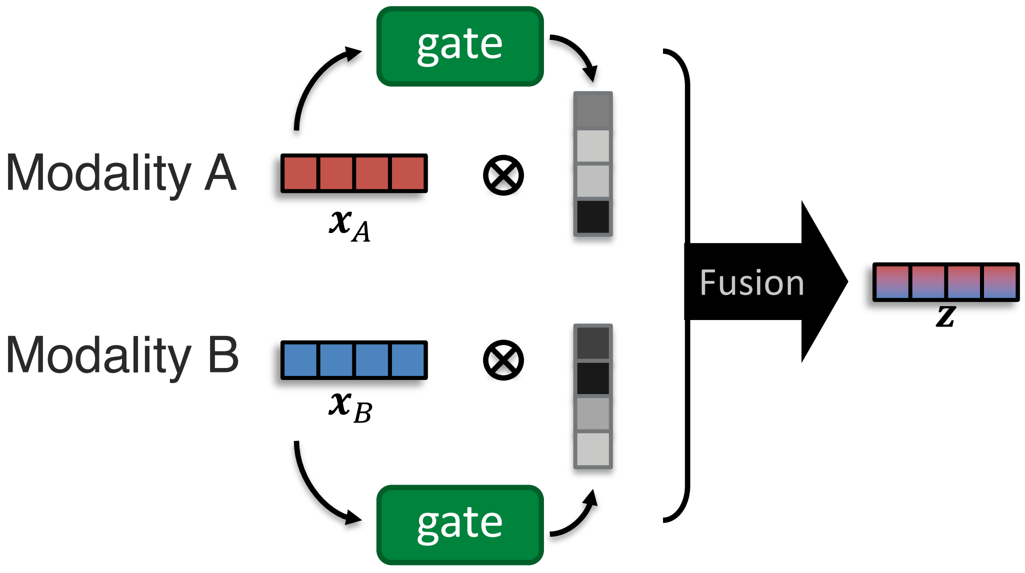
[Liu et al., Efficient Low-rank Multimodal Fusion with Modality-Specific Factors. ACL 2018]

Low-rank Fusion



[Liu et al., Efficient Low-rank Multimodal Fusion with Modality-Specific Factors. ACL 2018]

Gated Fusion



Example with additive fusion:

$$z = g_A(\boldsymbol{x}_A, \boldsymbol{x}_B) \cdot \boldsymbol{x}_A + g_B(\boldsymbol{x}_A, \boldsymbol{x}_B) \cdot \boldsymbol{x}_B$$

→ g_A and g_B can be seen as attention functions

Linear: $\boldsymbol{x}_A \boldsymbol{w}_1 \cdot (\boldsymbol{x}_B \boldsymbol{W}_2)^T$

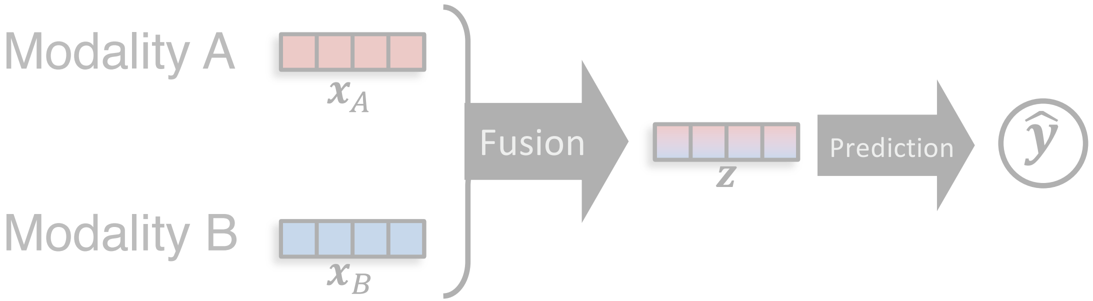
Nonlinear: $f_A(\boldsymbol{x}_A) \cdot (f_B(\boldsymbol{x}_B))^T$

Kernel: $k(\boldsymbol{x}_A, \boldsymbol{x}_B)$

- Linear
- Polynomial
- Exponential
- RBF

[Arevalo et al., Gated Multimodal Units for information fusion. ICLR-workshop 2017]

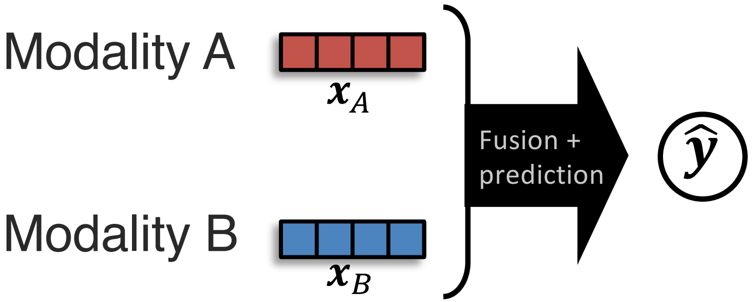
Nonlinear Fusion



Nonlinear fusion:

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

where f could be a multi-layer perceptron or any nonlinear model

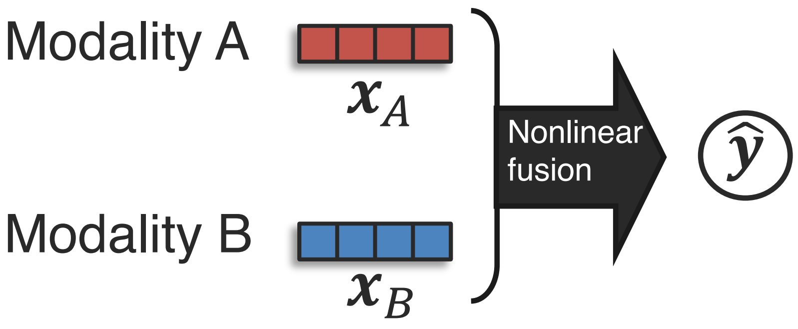


This could be seen as *early fusion*:

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

… but will our neural network learn the nonlinear interactions?

Measuring Non-Additive Interactions



Nonlinear fusion:

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

Projection?

Additive fusion:

$$\hat{y}' = f_A(\boldsymbol{x}_A) + f_B(\boldsymbol{x}_B)$$

Projection from nonlinear to additive (using EMAP):

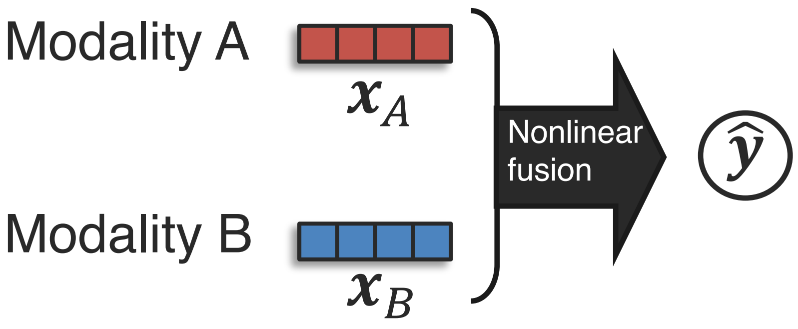
$$\tilde{f}(\boldsymbol{x}_A, \boldsymbol{x}_B) = \underbrace{\mathbb{E}_{\boldsymbol{x}_B}[f(\boldsymbol{x}_A, \boldsymbol{x}_B)]}_{f_A(\boldsymbol{x}_A)} + \underbrace{\mathbb{E}_{\boldsymbol{x}_A}[f(\boldsymbol{x}_A, \boldsymbol{x}_B)]}_{f_B(\boldsymbol{x}_B)}$$

Modality A + Modality B

Additive fusion (approximation)

[Hessel and Lee, Does my multimodal model learn cross-modal interactions? It’s harder to tell than you might think! EMNLP 2020]

Measuring Non-Additive Interactions



Nonlinear fusion:

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

EMAP projection

Additive fusion:

$$\hat{\boldsymbol{y}}' = \hat{f}_A(\boldsymbol{x}_A) + \hat{f}_B(\boldsymbol{x}_B) + \hat{\boldsymbol{\mu}}$$

		I-INT	I-SEM	I-CTX	T-VIS	R-POP	T-ST1	T-ST2
Nonlinear	Neural Network	90.4	69.2	78.5	51.1	63.5	71.1	79.9
Polynomial	Polykernel SVM	**91.3**	**74.4**	**81.5**	50.8	–	72.1	**80.9**
Nonlinear	FT LXMERT	83.0	68.5	76.3	**53.0**	63.0	66.4	78.6
Nonlinear	↳ + Linear Logits	89.9	73.0	80.7	**53.4**	**64.1**	**75.5**	80.3
Additive	Linear Model	90.4	72.8	80.9	51.3	63.7	**75.6**	76.1
	Best Model	**91.3**	**74.4**	**81.5**	**53.4**	*64.2*	75.5	**80.9**
Additive	↳ + EMAP	**91.1**	**74.2**	**81.3**	51.0	***64.1***	**75.9**	**80.7**

Always a good baseline!

Differences are small!!!

[Hessel and Lee, Does my multimodal model learn cross-modal interactions? It's harder to tell than you might think! EMNLP 2020]

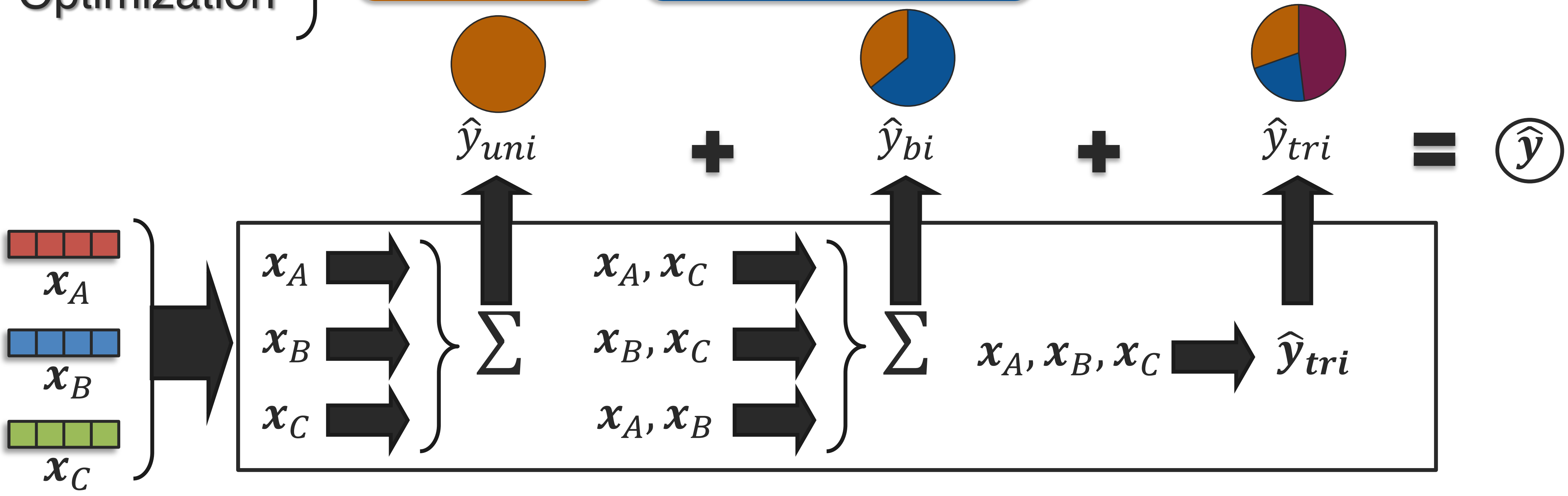
Non-Additive Interactions

Idea: prioritize simpler interactions

Unimodal (additive) residual **Bimodal** (non-additive) residual **Trimodal** (non-additive)

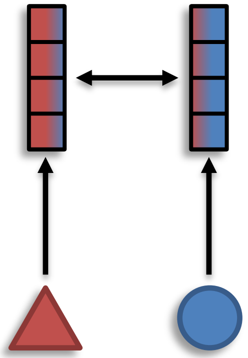
Multimodal Residual Optimization

$$\boxed{\mathcal{L}(y, \hat{y}_{uni})} + \boxed{\mathcal{L}(\boxed{y - \hat{y}_{uni}}, \hat{y}_{bi})} + \mathcal{L}(\boxed{y - \hat{y}_{uni} - \hat{y}_{bi}}, \hat{y}_{tri})$$

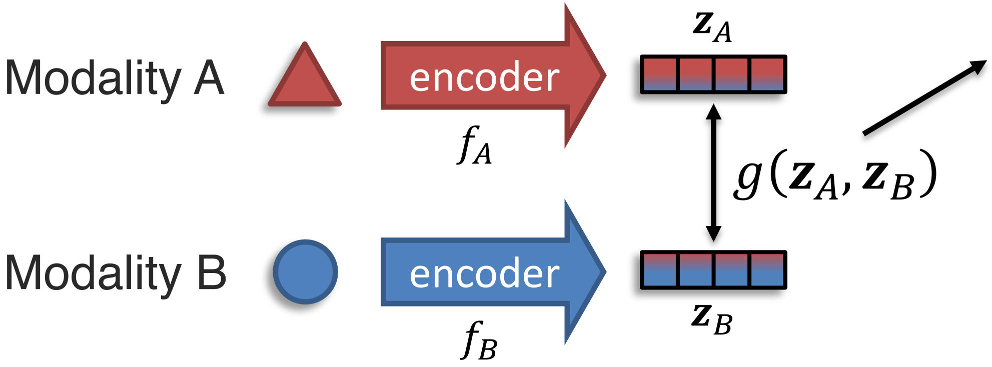


[Wortwein et al., Beyond Additive Fusion: Learning Non-Additive Multimodal Interactions, Findings-EMNLP 2022]

Sub-Challenge 1b: Representation Coordination



Definition: Learn multimodal contextualized representations coordinated through their interconnections.



2. Coordination function captures **interconnections**

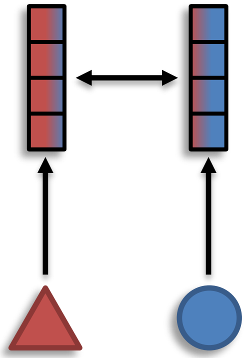
1. Specialized encoders capture **heterogeneity**

Learning with coordination function:

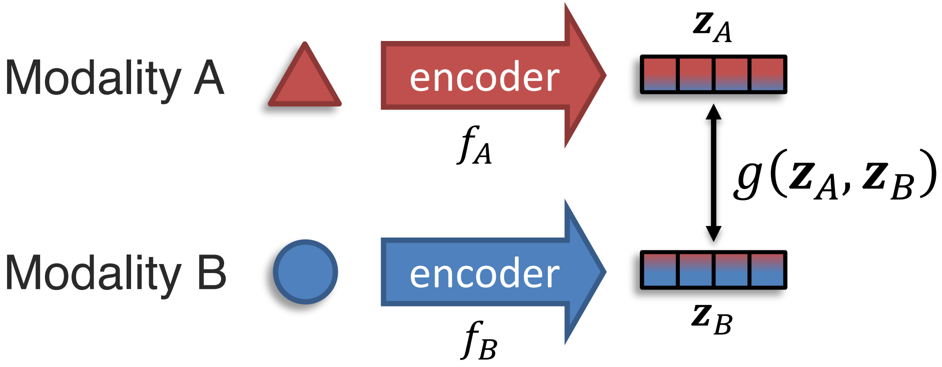
$$\mathcal{L} = g(f_A(\triangle), f_B(\bigcirc))$$

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

Coordinated Representations



Definition: Learn multimodal contextualized representations coordinated through their interconnections.



Learning with coordination function:

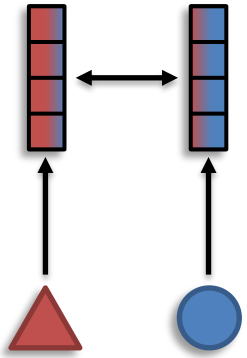
$$\mathcal{L} = g\big(f_A(\triangle), f_B(\bigcirc)\big)$$

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

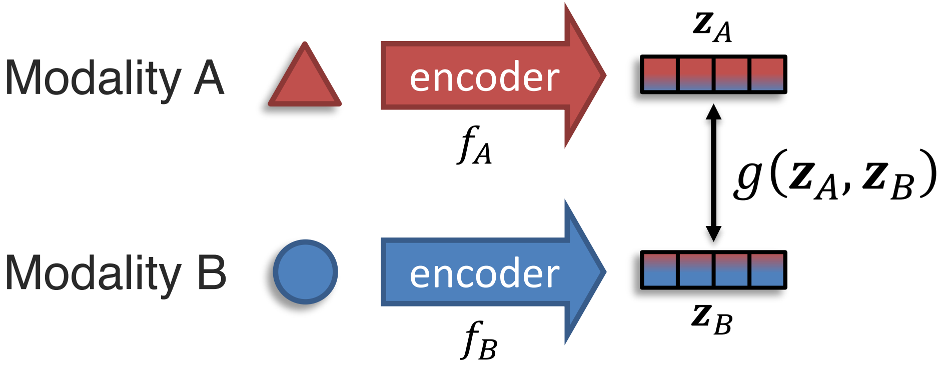
1 Cosine similarity:

$$g(\boldsymbol{z}_A, \boldsymbol{z}_B) = \frac{\boldsymbol{z}_A \cdot \boldsymbol{z}_B}{\|\boldsymbol{z}_A\| \|\boldsymbol{z}_B\|}$$

Coordinated Representations



Definition: Learn multimodal contextualized representations coordinated through their interconnections.



Learning with coordination function:

$$\mathcal{L} = g\left(f_A(\triangle), f_B(\bigcirc)\right)$$

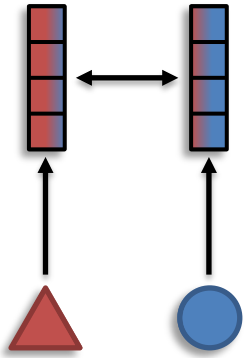
with model parameters θ_g, θ_{f_A} and θ_{f_B}

② Kernel similarity functions:

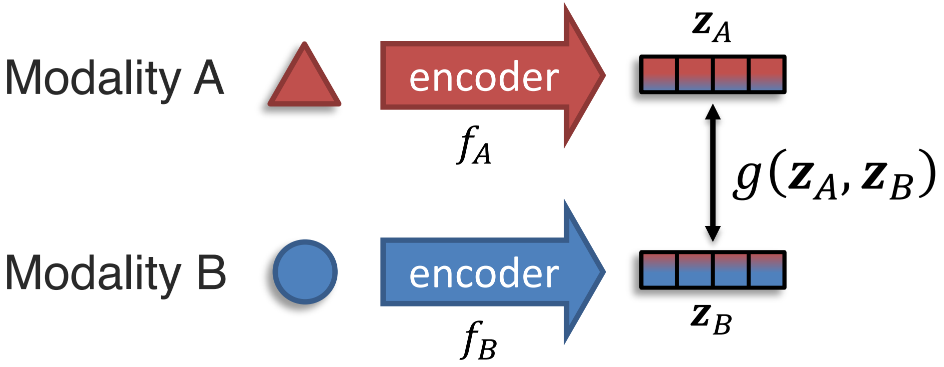
$$g(\boldsymbol{z}_A, \boldsymbol{z}_B) = k(\boldsymbol{z}_A, \boldsymbol{z}_B)$$

- Linear
- Polynomial
- Exponential
- RBF

Coordinated Representations



Definition: Learn multimodally-contextualized representations coordinated through their cross-modal connections.



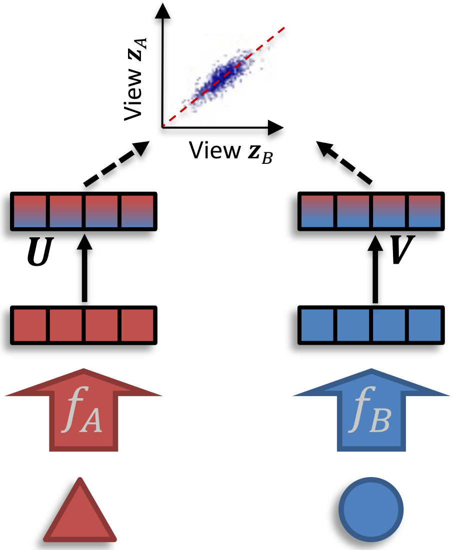
Learning with coordination function:

$$\mathcal{L} = g\big(f_A(\triangle), f_B(\bigcirc)\big)$$

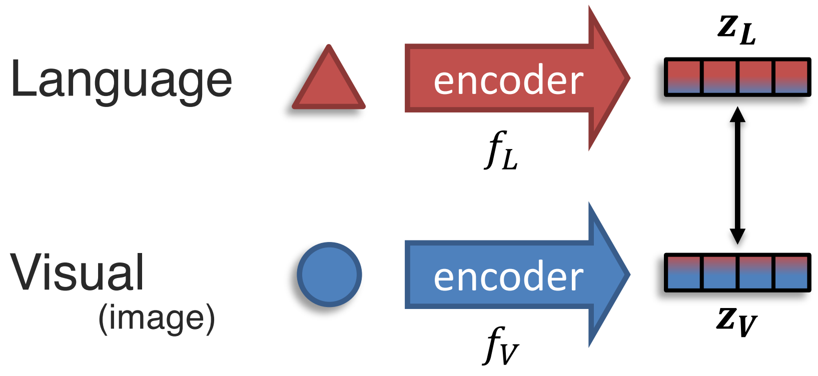
with model parameters θ_g, θ_{f_A} and θ_{f_B}

③ Canonical Correlation Analysis (CCA):

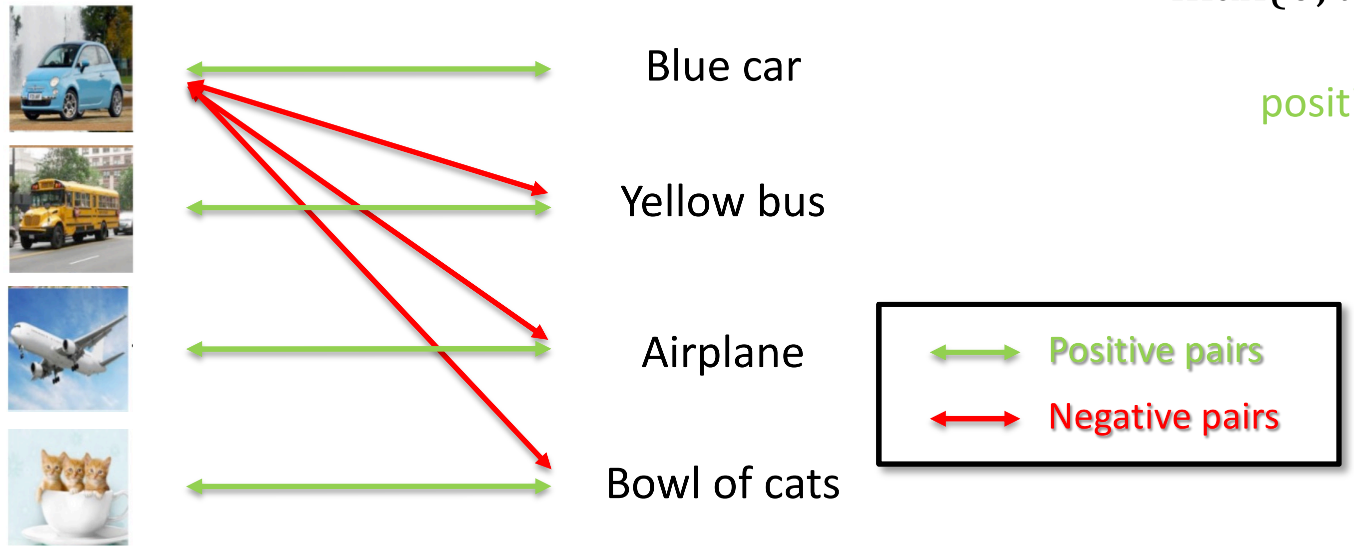
$$\underset{\boldsymbol{V},\boldsymbol{U},f_A,f_B}{\operatorname{argmax}}\ corr(\boldsymbol{z}_A, \boldsymbol{z}_B)$$



Coordination with Contrastive Learning



Paired data: { , } (e.g., images and text)



Contrastive loss:

brings positive pairs closer and pushes negative pairs apart

Simple contrastive loss:

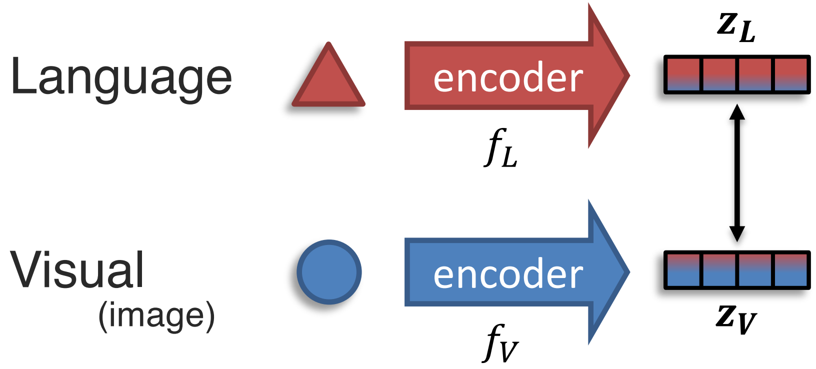
$$\max\{0, \alpha + \boxed{g(\boldsymbol{z}_A, \boldsymbol{z}_B^+)} - \boxed{g(\boldsymbol{z}_A, \boldsymbol{z}_B^-)}\}$$

positive pairs

negative pair

Coordination function (e.g., cosine similarity)

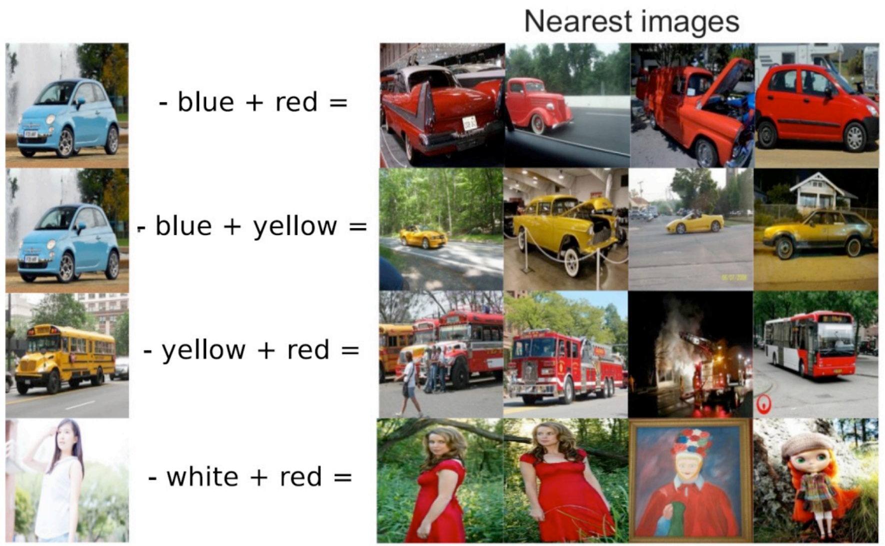
Visual-Semantic Representations

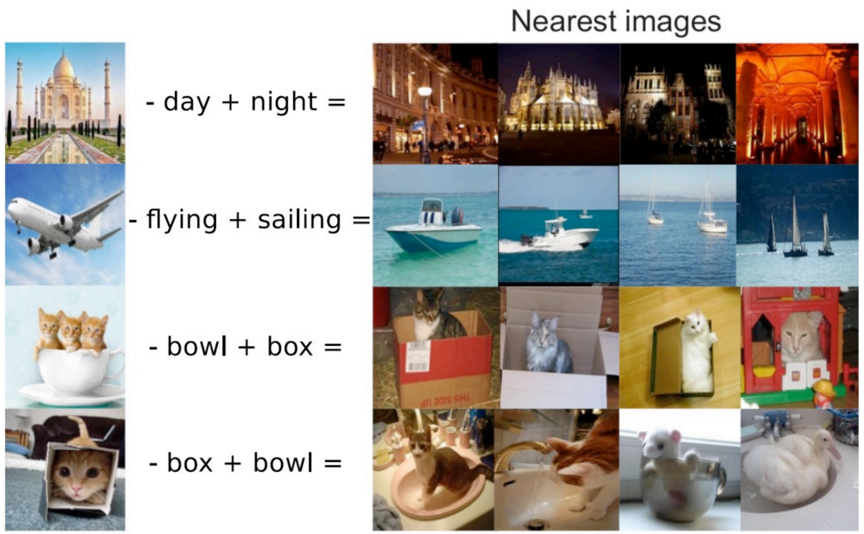


Two contrastive loss terms:

$$\max\{0, \alpha + sim(\mathbf{z}_L, \mathbf{z}_V^+) - sim(\mathbf{z}_L, \mathbf{z}_V^-)\}$$

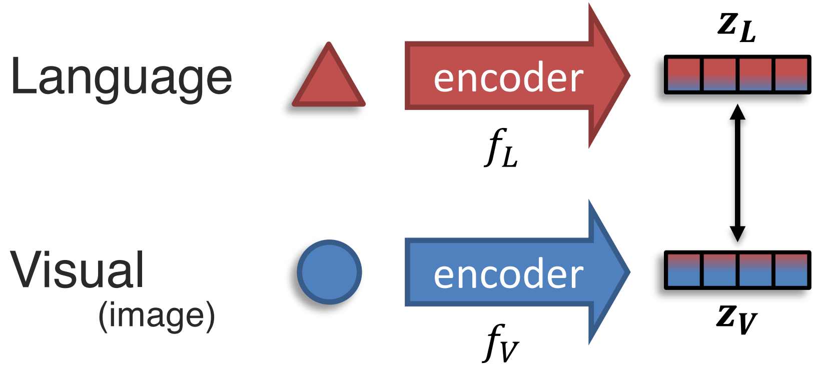
$$+ \max\{0, \alpha + sim(\mathbf{z}_V, \mathbf{z}_L^+) - sim(\mathbf{z}_V, \mathbf{z}_L^-)\}$$





[Kiros et al., Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models, NeurIPS 2014]

Contrastive Language Image Pretraining



Popular contrastive loss: InfoNCE

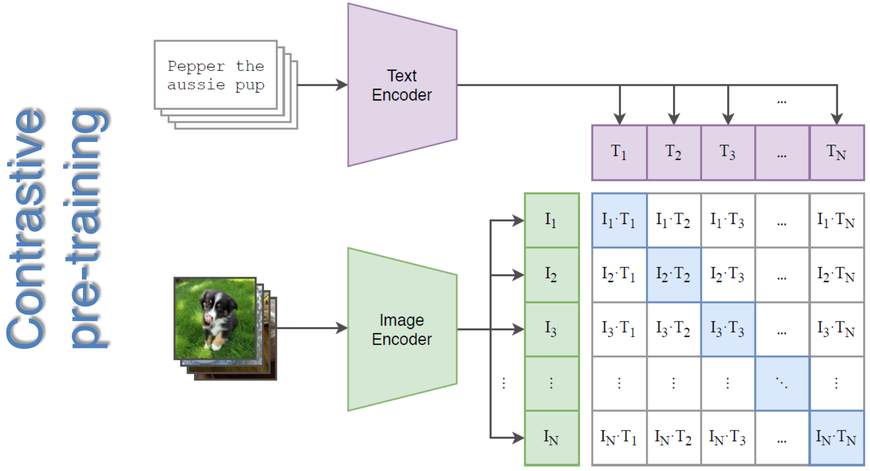
$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{\boxed{sim(\mathbf{z}_A^i, \mathbf{z}_B^i)}}{\boxed{\sum_{j=1}^{N} sim(\mathbf{z}_A^i, \mathbf{z}_B^j)}}$$

positive pairs

negative pairs and positive pairs

Similarity function can be cosine similarity

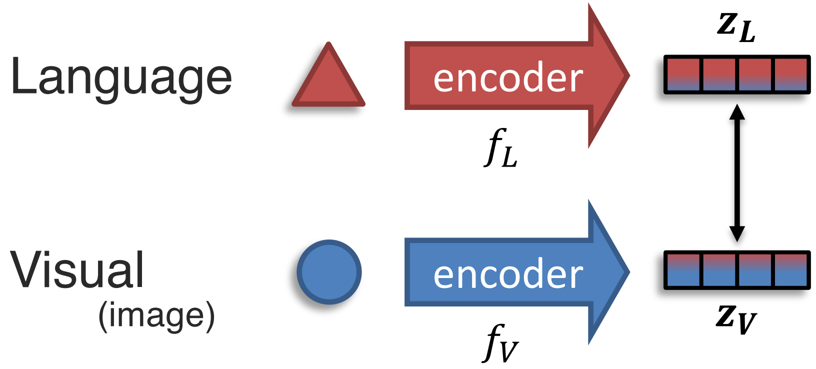
Positive and negative pairs:



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

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

Contrastive Learning and Connected Modalities

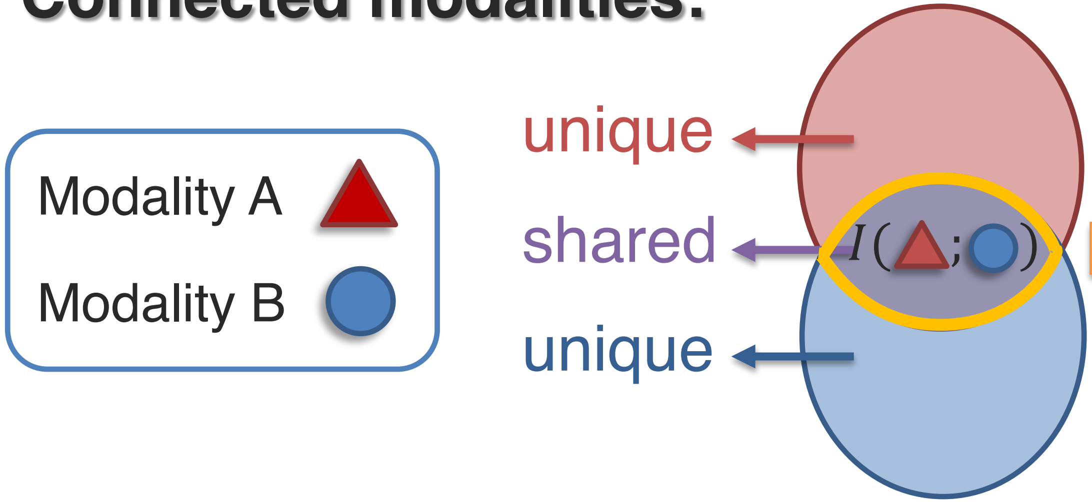


Popular contrastive loss: InfoNCE

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

CLIP focuses on shared connections

Connected modalities:



Mutual information $I(X;Y)$

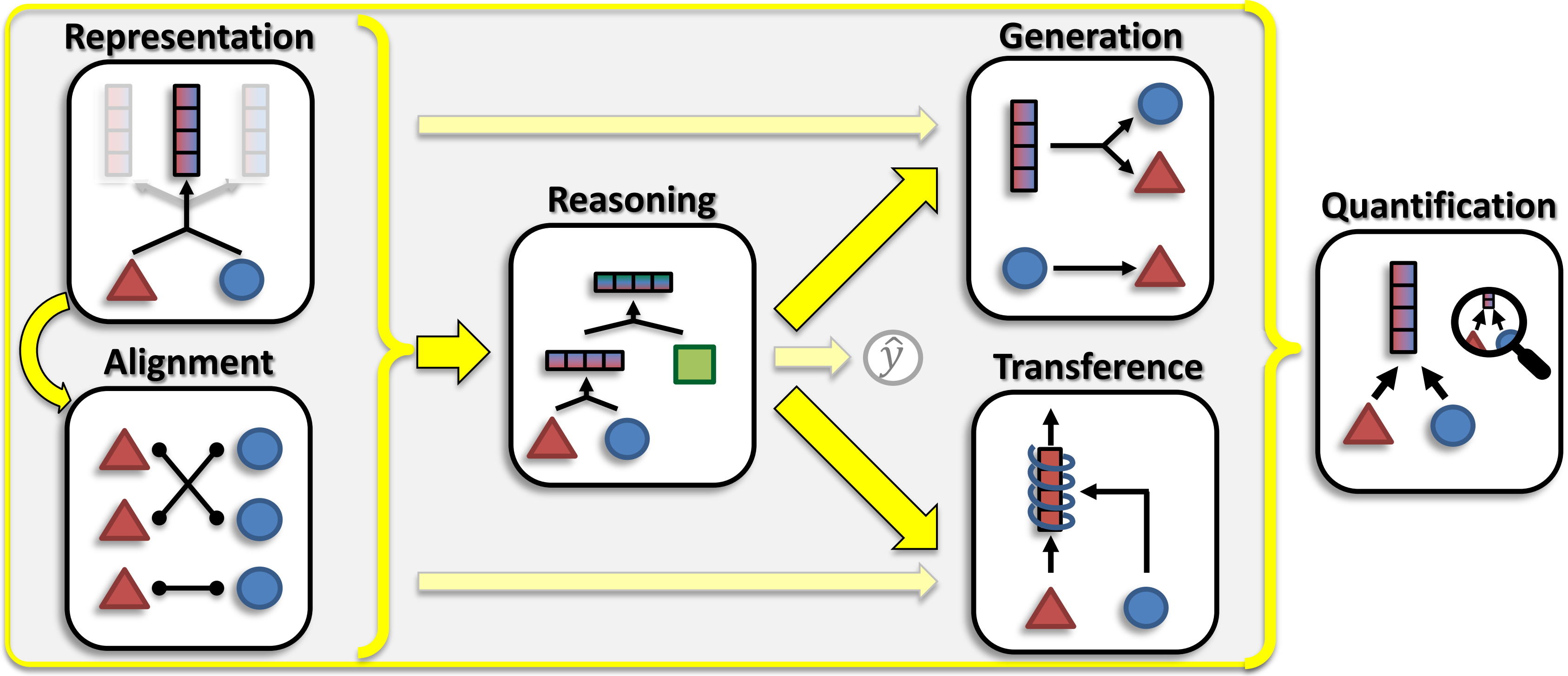
$$\mathbb{E}_{X,Y}\left[\log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}\right]$$

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

Summary

Resources: https://cs.cmu.edu/~pliang/

***Multimodal* is the science of heterogeneous and interconnected data.**



What is Multimodal? → Why is it hard? → What is next?

What is Multimodal?

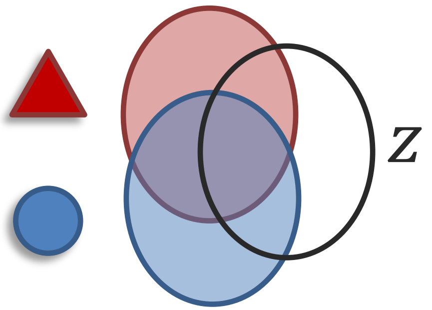
Heterogeneous

+

Connected

+

Interacting



Why is it hard?

Representation

Alignment

Reasoning

Generation

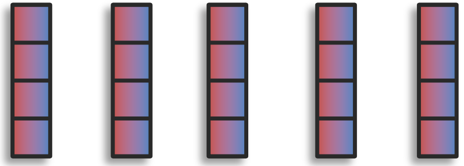
Transference

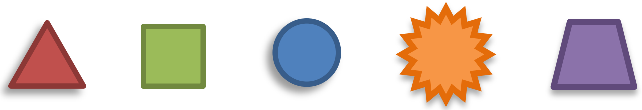
Quantification

What is next?

Future Direction: Heterogeneity & Interactions

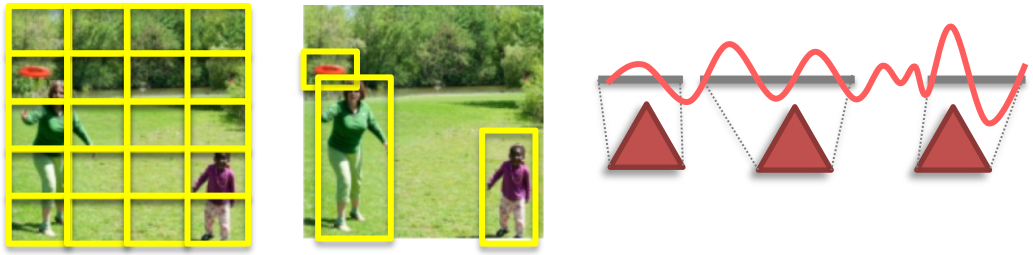
Homogeneity vs Heterogeneity





Challenges:

Arbitrary tokenization

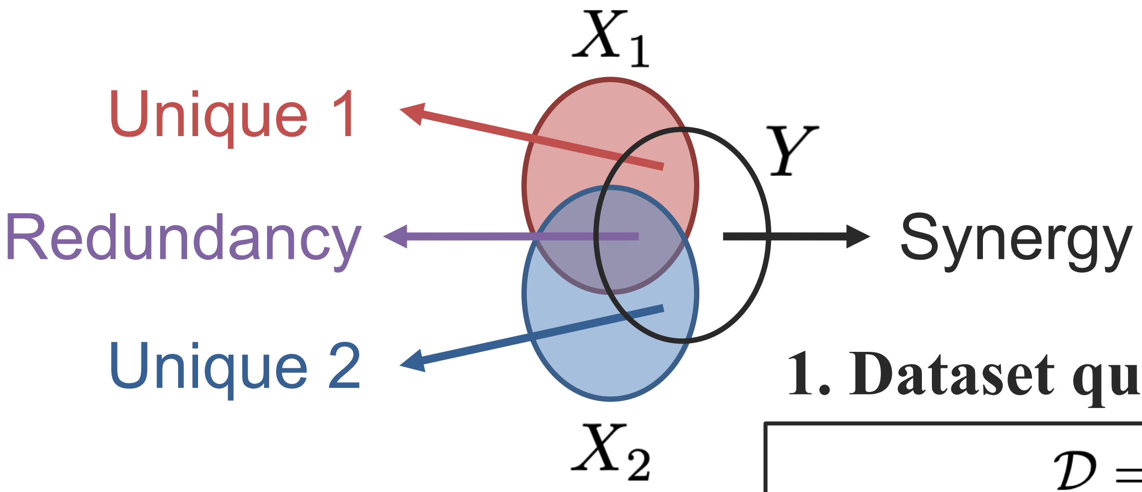


Beyond differentiable interactions

Causal, logical, brain-inspired

Theoretical foundations of interactions

Quantifying Interactions



1. Dataset quantification:

$$\mathcal{D} = \{(x_1, x_2, y)\} \longrightarrow \{R, U_1, U_2, S\}_{\mathcal{D}}$$

2. Model quantification:

$$f(\mathcal{D}) = \{(x_1, x_2, \hat{y} = f(x_1, x_2))\} \longrightarrow \{R, U_1, U_2, S\}_{f(\mathcal{D})}$$

$$\{R, U_1, U_2, S\}_{f(\mathcal{D}_1)}, \ldots, \{R, U_1, U_2, S\}_{f(\mathcal{D}_k)} \longrightarrow \{R, U_1, U_2, S\}_f$$

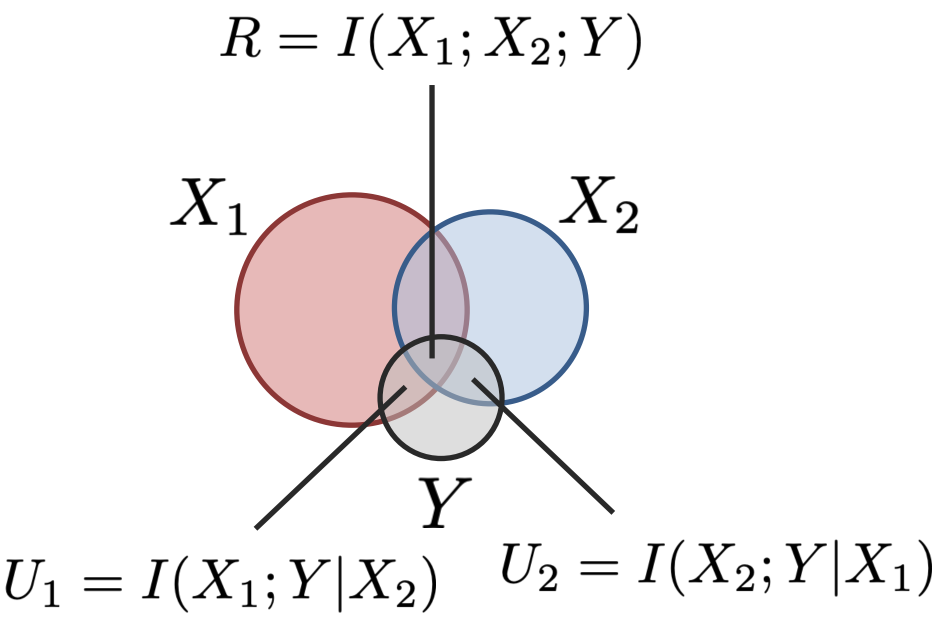
3. Model selection:

$$\{R, U_1, U_2, S\}_{\mathcal{D}} \longleftrightarrow \{R, U_1, U_2, S\}_f$$

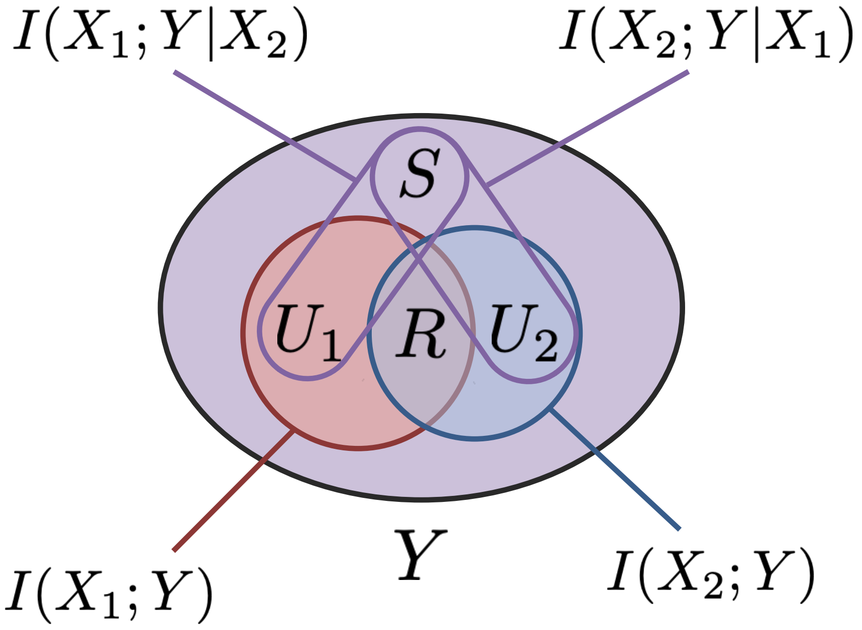
[Liang et al., Quantifying & Modeling Feature Interactions: An Information Decomposition Framework. arXiv 2023]

Quantifying Interactions

Classical Information Theory



Partial Information Decomposition

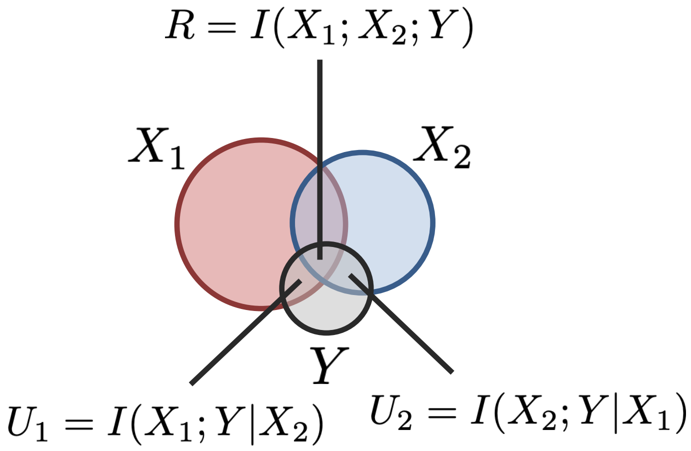


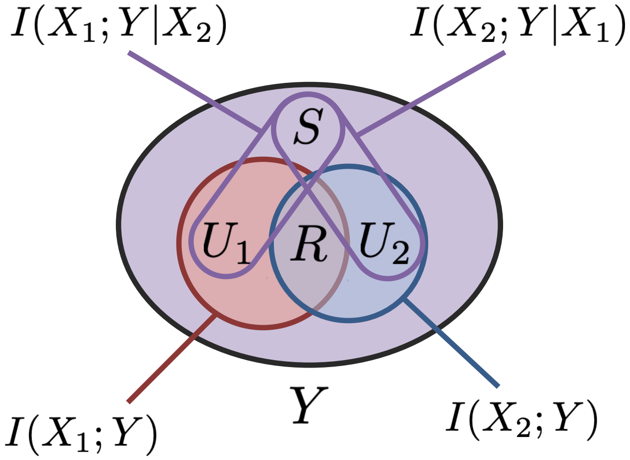
$$R - S = I(X_1; X_2; Y)$$

$$R + U_1 + U_2 + S = I(X_1, X_2; Y)$$

[Liang et al., Quantifying & Modeling Feature Interactions: An Information Decomposition Framework. arXiv 2023]

Quantifying Interactions





$$\Delta_p = \{q \in \Delta : q(x_i, y) = p(x_i, y)\ \forall y, x_i, i \in \{1, 2\}\}$$

Marginal-matching distributions

$$R = \max_{q \in \Delta_p} I_q(X_1; X_2; Y) \qquad U_2 = \max_{q \in \Delta_p} I_q(X_2; Y|X_1)$$

$$U_1 = \max_{q \in \Delta_p} I_q(X_1; Y|X_2) \qquad S = I_p(X_1, X_2; Y) - \min_{q \in \Delta_p} I_q(X_1, X_2; Y)$$

[Liang et al., Quantifying & Modeling Feature Interactions: An Information Decomposition Framework. arXiv 2023]

Quantifying Interactions

$$q^* = \arg\max_{q \in \Delta_p} H_q(Y | X_1, X_2)$$

If X1, X1, Y have small and discrete support,
exact solution via convex programming with linear constraints.

Else, neural network estimator.

$$\Delta_p = \{q \in \Delta : q(x_i, y) = p(x_i, y)\ \forall y, x_i, i \in \{1, 2\}\}$$

Marginal-matching distributions

$$R = \max_{q \in \Delta_p} I_q(X_1; X_2; Y) \qquad U_2 = \max_{q \in \Delta_p} I_q(X_2; Y | X_1)$$

$$U_1 = \max_{q \in \Delta_p} I_q(X_1; Y | X_2) \qquad S = I_p(X_1, X_2; Y) - \min_{q \in \Delta_p} I_q(X_1, X_2; Y)$$

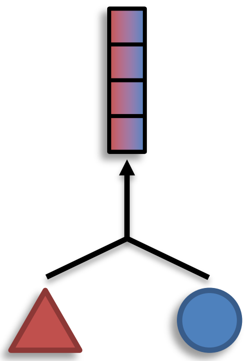
[Liang et al., Quantifying & Modeling Feature Interactions: An Information Decomposition Framework. arXiv 2023]

Representation Models

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

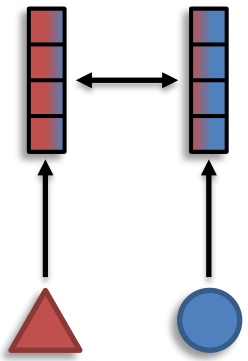
Sub-challenges:

Fusion



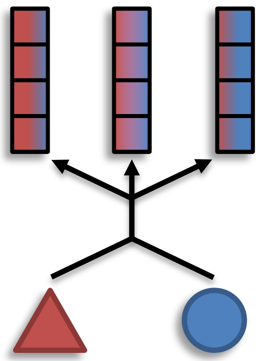
modalities **>** # representations

Coordination



modalities **=** # representations

Fission



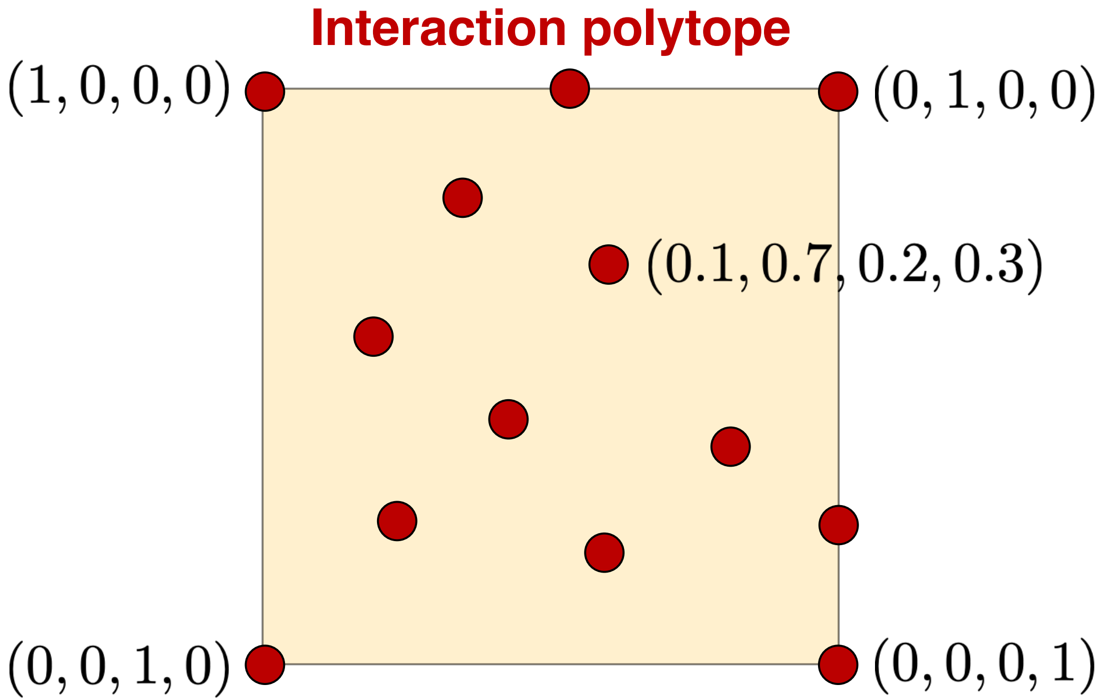
modalities **<** # representations

[Liang et al., Quantifying & Modeling Feature Interactions: An Information Decomposition Framework. arXiv 2023]

Model Selection

1. Dataset quantification:

$$\mathcal{D} = \{(x_1, x_2, y)\} \longrightarrow \{R, U_1, U_2, S\}_{\mathcal{D}}$$



[Liang et al., Quantifying & Modeling Feature Interactions: An Information Decomposition Framework. arXiv 2023]

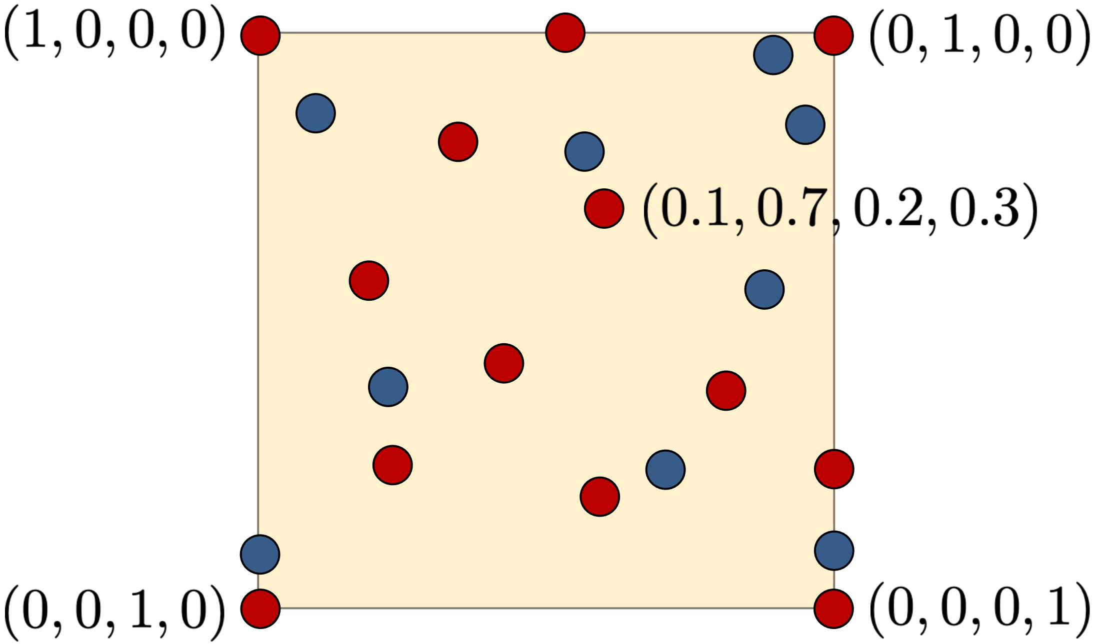
Model Selection

2. Model quantification:

$$f(\mathcal{D}) = \{(x_1, x_2, \hat{y} = f(x_1, x_2))\} \longrightarrow \{R, U_1, U_2, S\}_{f(\mathcal{D})}$$

$$\{R, U_1, U_2, S\}_{f(\mathcal{D}_1)}, \ldots, \{R, U_1, U_2, S\}_{f(\mathcal{D}_k)} \longrightarrow \{R, U_1, U_2, S\}_f$$

Interaction polytope



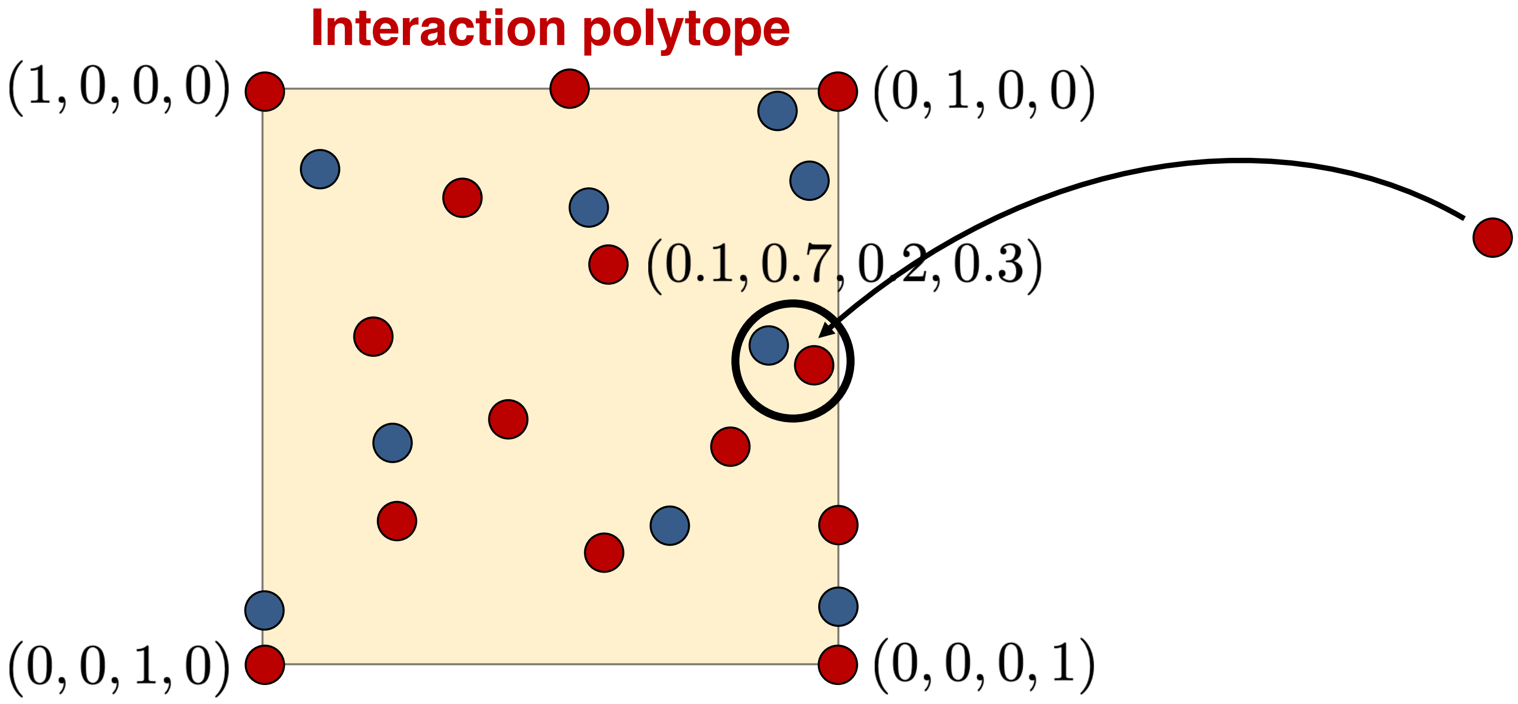
[Liang et al., Quantifying & Modeling Feature Interactions: An Information Decomposition Framework. arXiv 2023]

Model Selection

3. Model selection:

$$\{R, U_1, U_2, S\}_{\mathcal{D}} \longleftrightarrow \{R, U_1, U_2, S\}_f$$

Selects models with >96% performance



[Liang et al., Quantifying & Modeling Feature Interactions: An Information Decomposition Framework. arXiv 2023]

Future Direction: High-modality

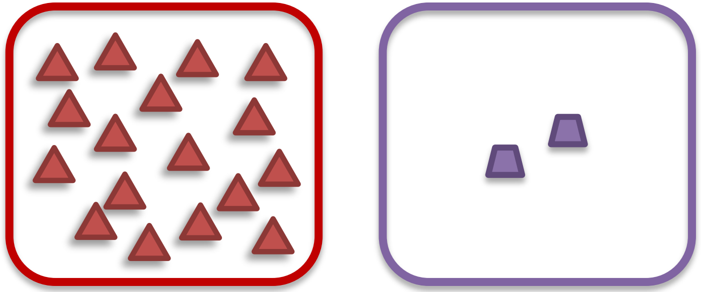
MultiBench

https://github.com/pliang279/MultiBench

Few modalities → High-modality

Language Vision Audio Graphs Control LIDAR Sensors Set Table Financial Medical

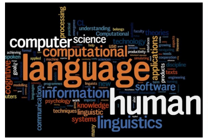
Challenges: **Non-parallel learning** **Limited resources**

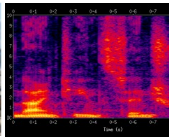


High-Modality Learning

How can we transfer knowledge across multiple tasks, each over a different subset of modalities?

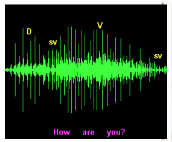
Video classification





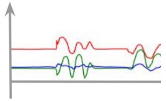
Language Video Audio

Sentiment, emotions



Audio Video

Robot dynamics



Video Time-series

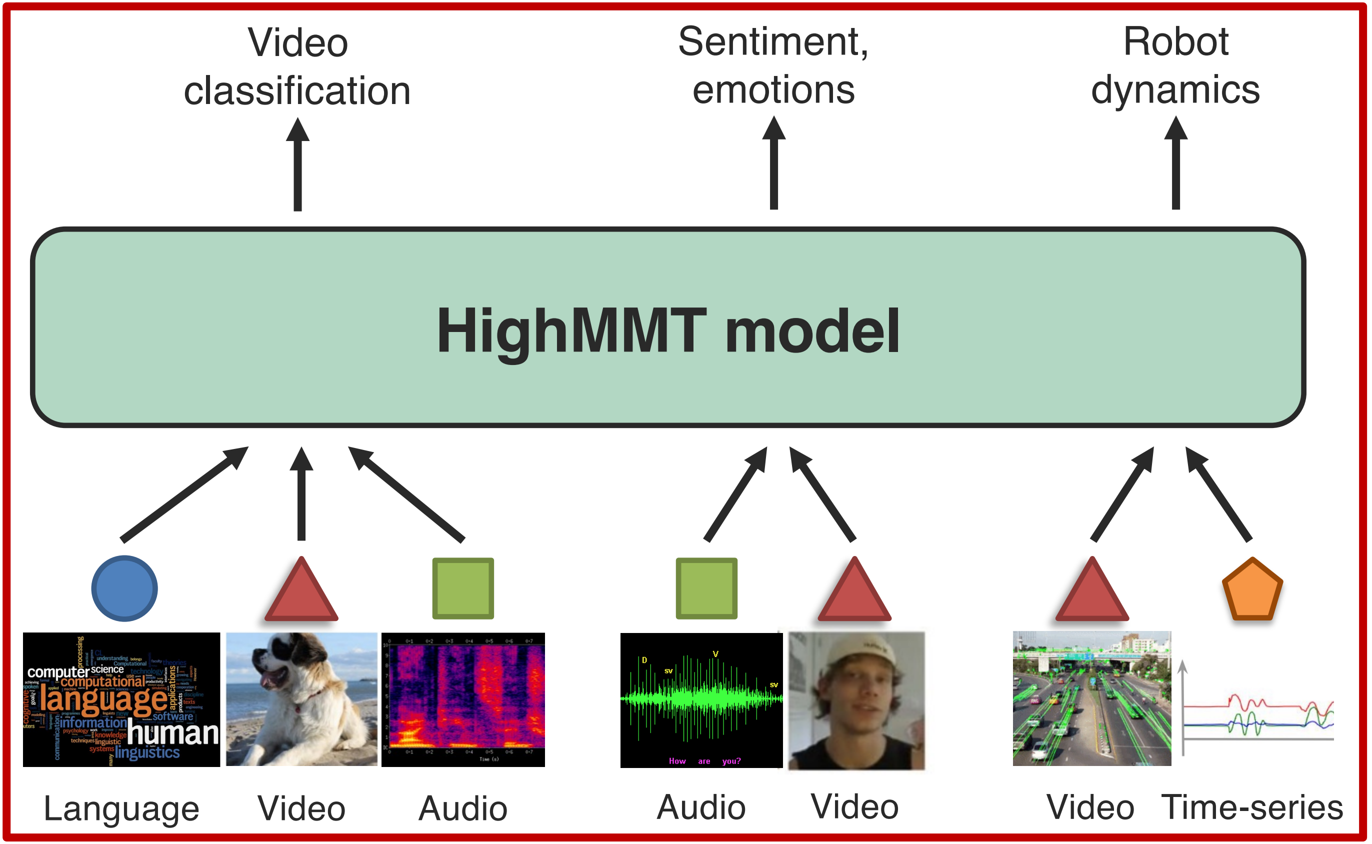
Generalization across modalities and tasks
Important if some tasks are low-resource

[Liang et al., MultiBench: Multiscale Benchmarks for Multimodal Representation Learning. NeurIPS 2021]

HighMMT

Transfer across partially observable modalities

HighMMT: unified model + parameter sharing + multitask and transfer learning



Non-parallel multitask learning

Task-specific classifiers

Shared multimodal model

Modality-specific embeddings

Standardized input sequence

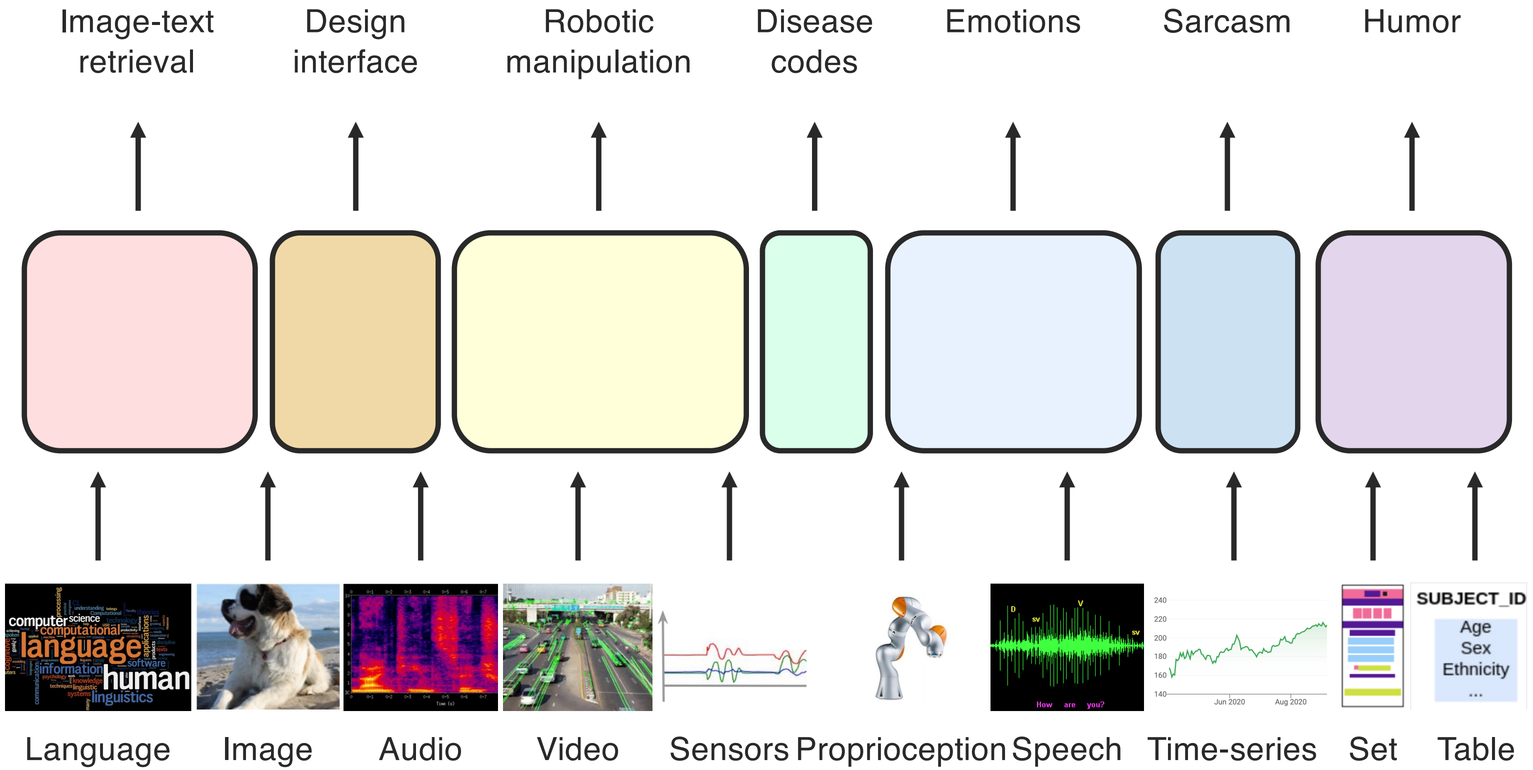
Same model architecture!

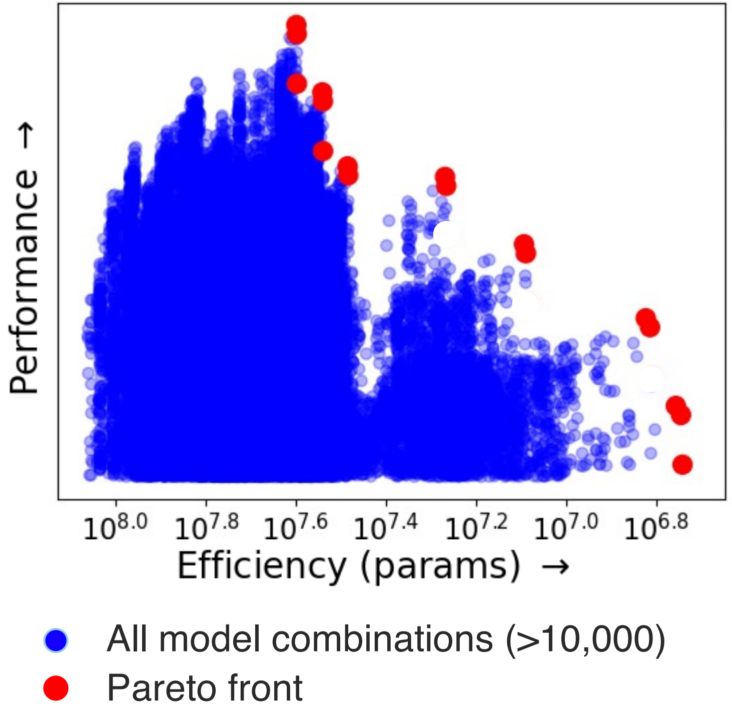
Same parameters!

[Liang et al., HighMMT: Quantifying Modality & Interaction Heterogeneity for High-Modality Learning. TMLR 2022 revisions]

HighMMT

Traditional approaches: different model + different parameters

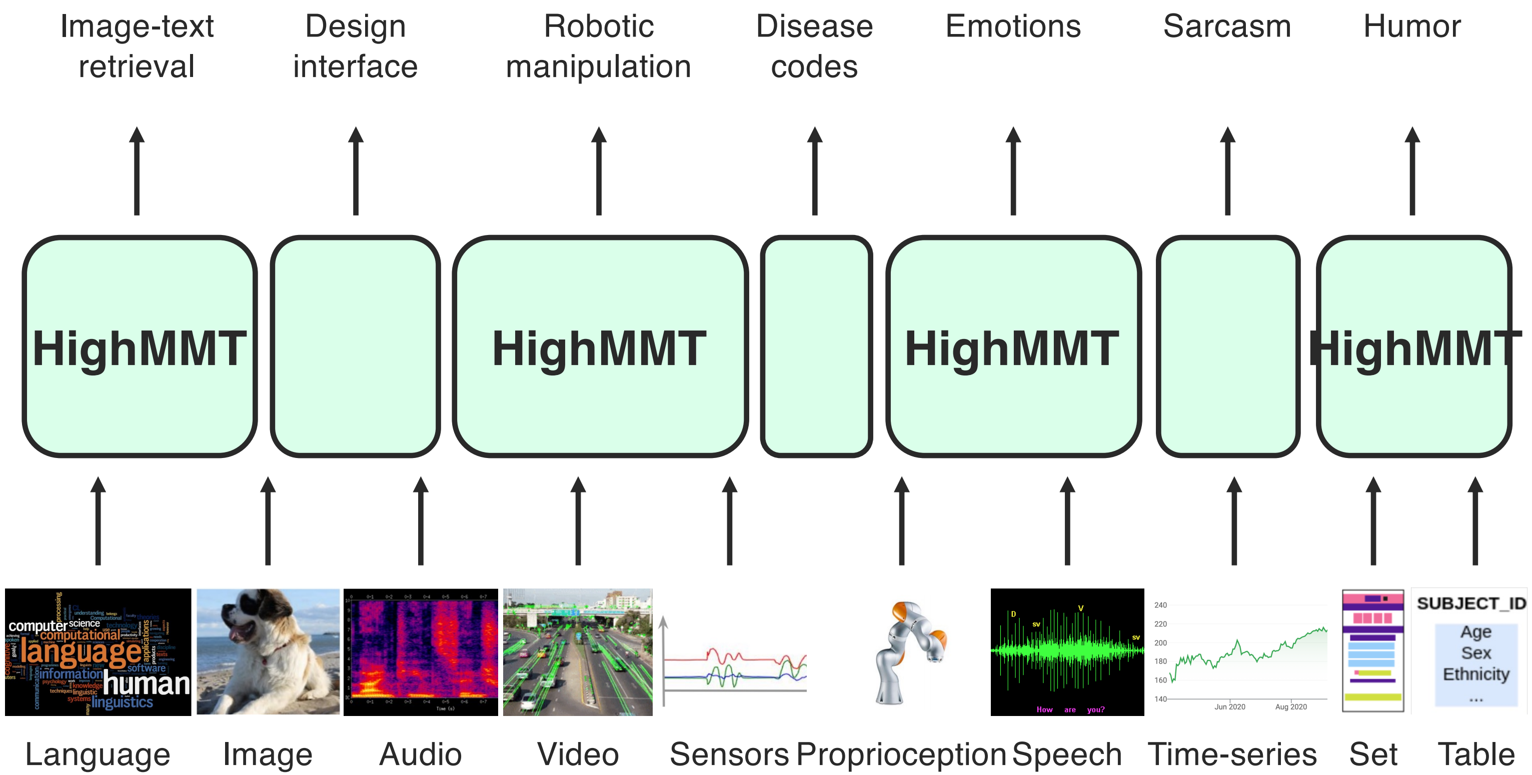


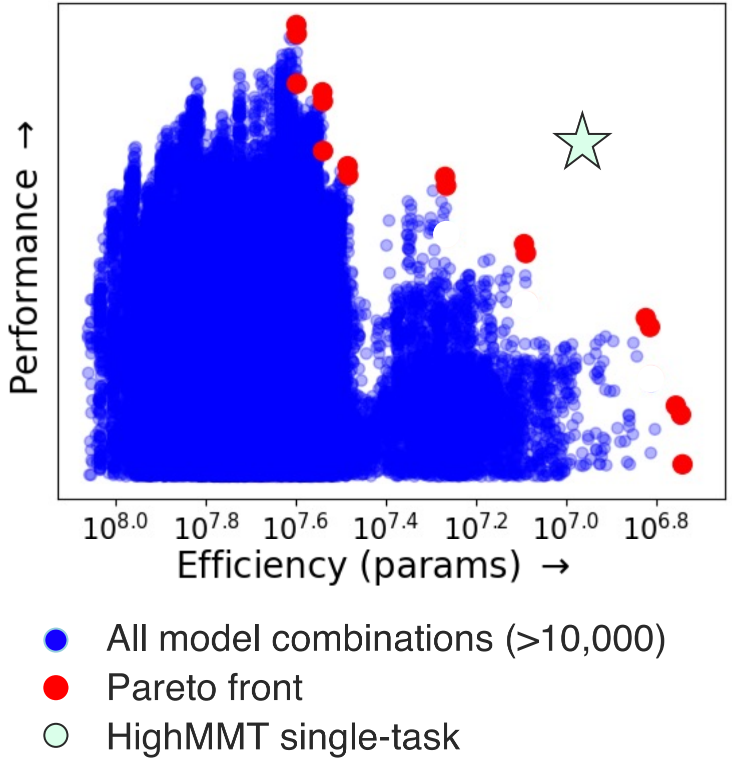


[Liang et al., HighMMT: Quantifying Modality & Interaction Heterogeneity for High-Modality Learning. TMLR 2022 revisions]

HighMMT

Traditional approaches: different model + different parameters

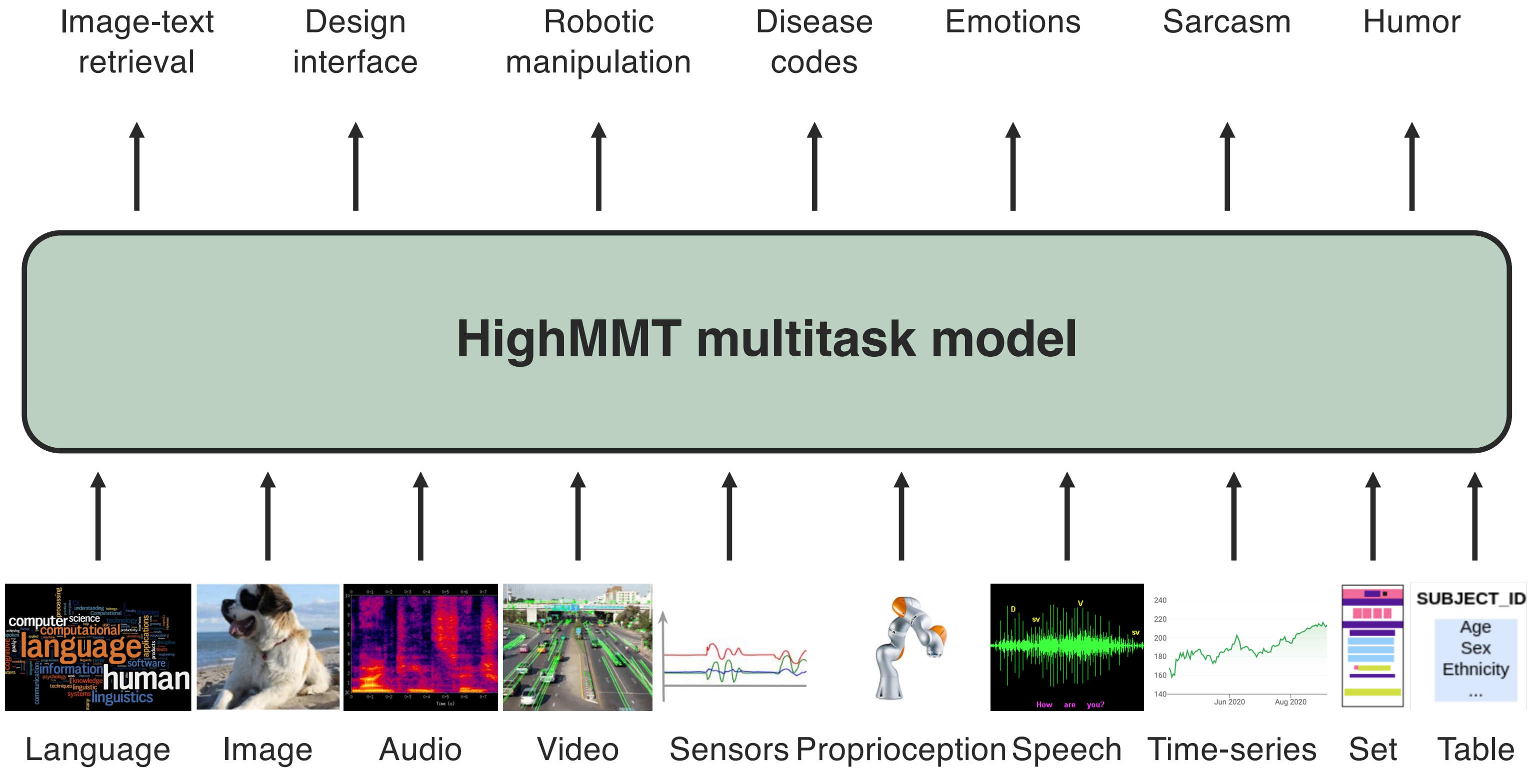


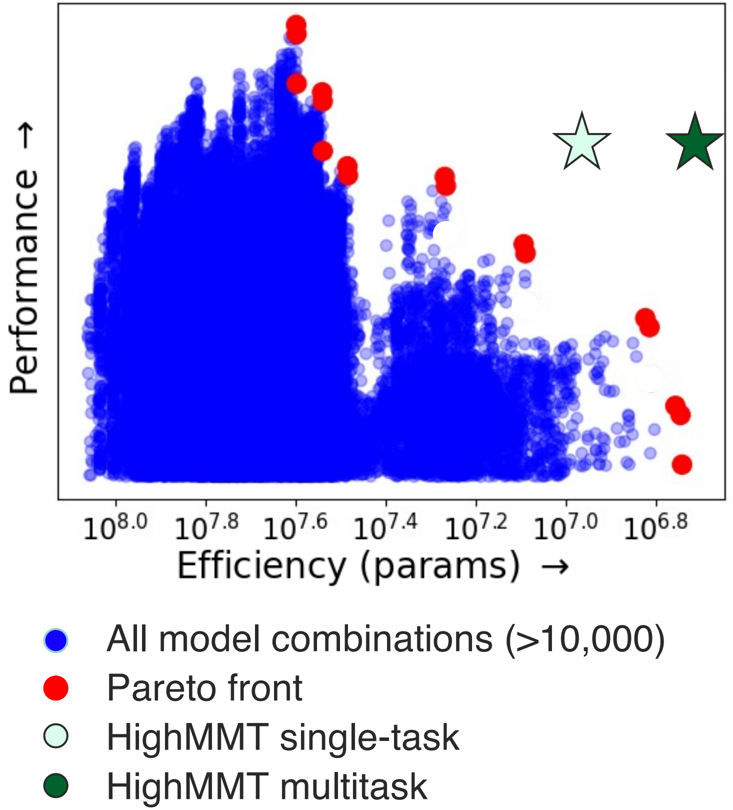


[Liang et al., HighMMT: Quantifying Modality & Interaction Heterogeneity for High-Modality Learning. TMLR 2022 revisions]

HighMMT

Traditional approaches: different model + different parameters



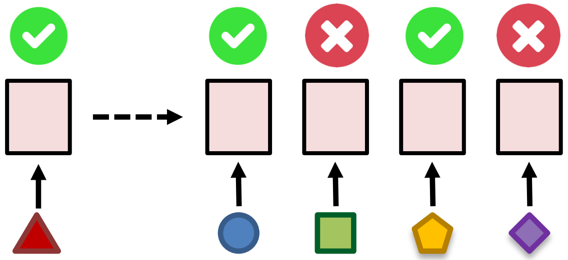


[Liang et al., HighMMT: Quantifying Modality & Interaction Heterogeneity for High-Modality Learning. TMLR 2022 revisions]

Quantifying Modality Heterogeneity

Information transfer, transfer learning perspective

1a. Estimate modality heterogeneity via transfer



Implicitly captures these:

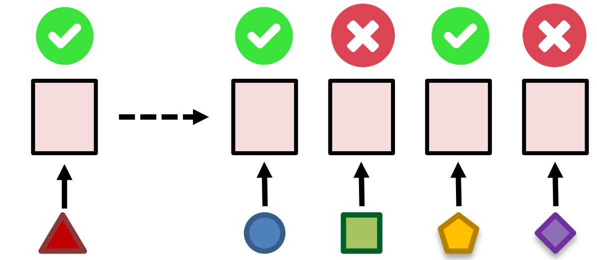
1. **Element representation**
2. **Element distribution**
3. **Structure**
4. **Information**
5. **Noise**
6. **Relevance**

[Liang et al., HighMMT: Quantifying Modality & Interaction Heterogeneity for High-Modality Learning. TMLR 2022 revisions]

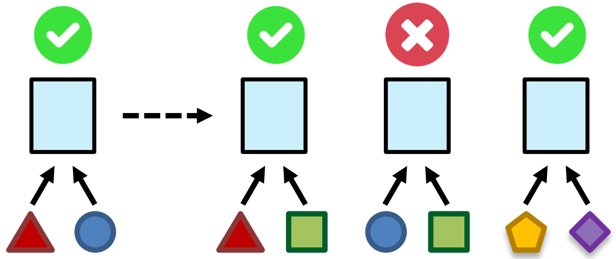
Quantifying Modality Heterogeneity

Information transfer, transfer learning perspective

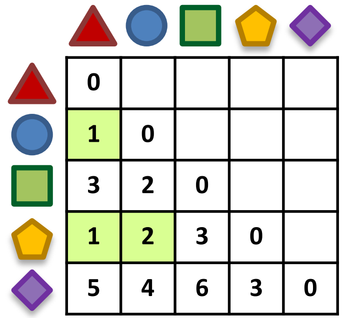
1a. Estimate modality heterogeneity via transfer



1b. Estimate interaction heterogeneity via transfer

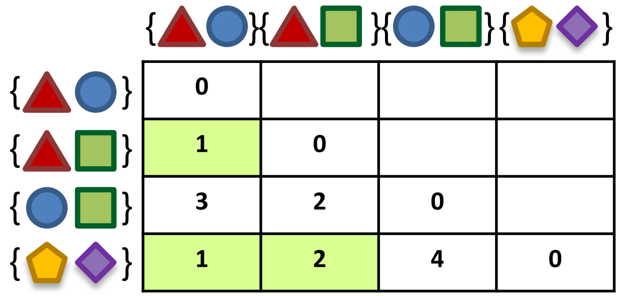


2a. Compute modality heterogeneity matrix



0				
1	0			
3	2	0		
1	2	3	0	
5	4	6	3	0

2b. Compute interaction heterogeneity matrix



0			
1	0		
3	2	0	
1	2	4	0

[Liang et al., HighMMT: Quantifying Modality & Interaction Heterogeneity for High-Modality Learning. TMLR 2022 revisions]

Quantifying Modality Heterogeneity

Information transfer, transfer learning perspective

2a. Compute modality heterogeneity matrix

	0				
	1	0			
	3	2	0		
	1	2	3	0	
	5	4	6	3	0

2b. Compute interaction heterogeneity matrix

{}	0			
{}	1	0		
{}	3	2	0	
{}	1	2	4	0

3. Determine parameter clustering

$$\mathbb{U}_1 = \{U_1, U_2, U_4\} \quad \mathbb{C}_1 = \{C_{12}, C_{13}, C_{45}\}$$

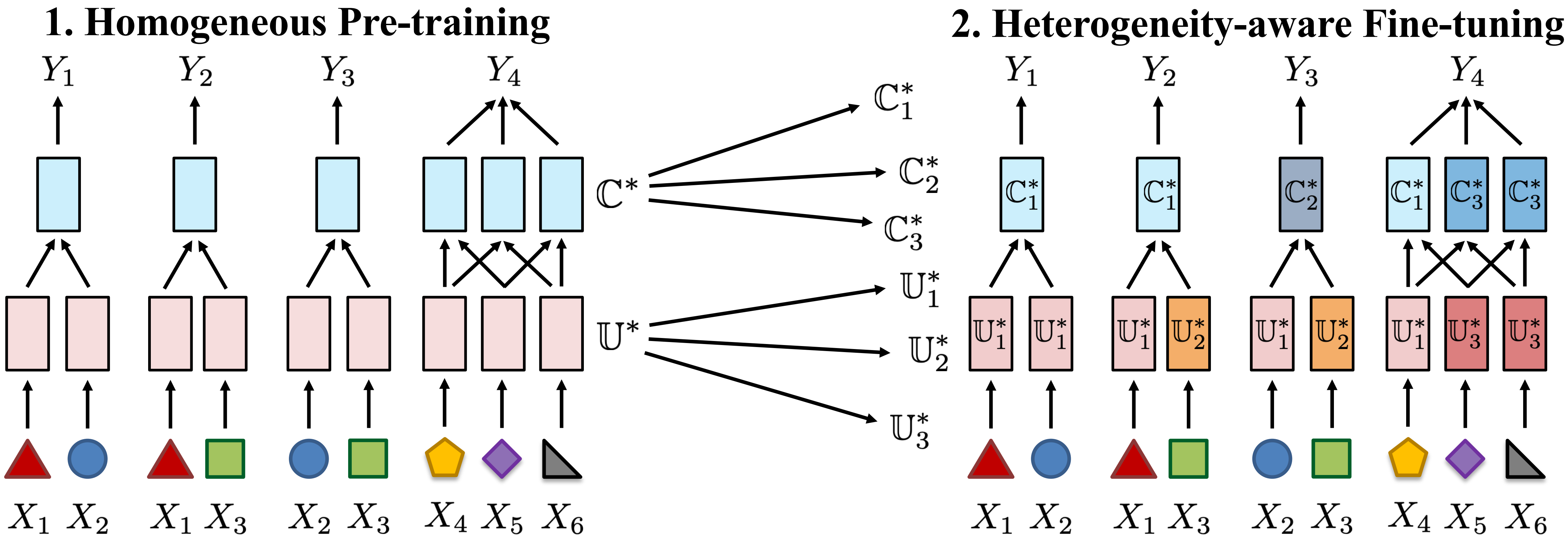
$$\mathbb{U}_2 = \{U_3\} \quad \mathbb{C}_2 = \{C_{23}\}$$

$$\mathbb{U}_3 = \{U_5\}$$

[Liang et al., HighMMT: Quantifying Modality & Interaction Heterogeneity for High-Modality Learning. TMLR 2022 revisions]

Quantifying Modality Heterogeneity

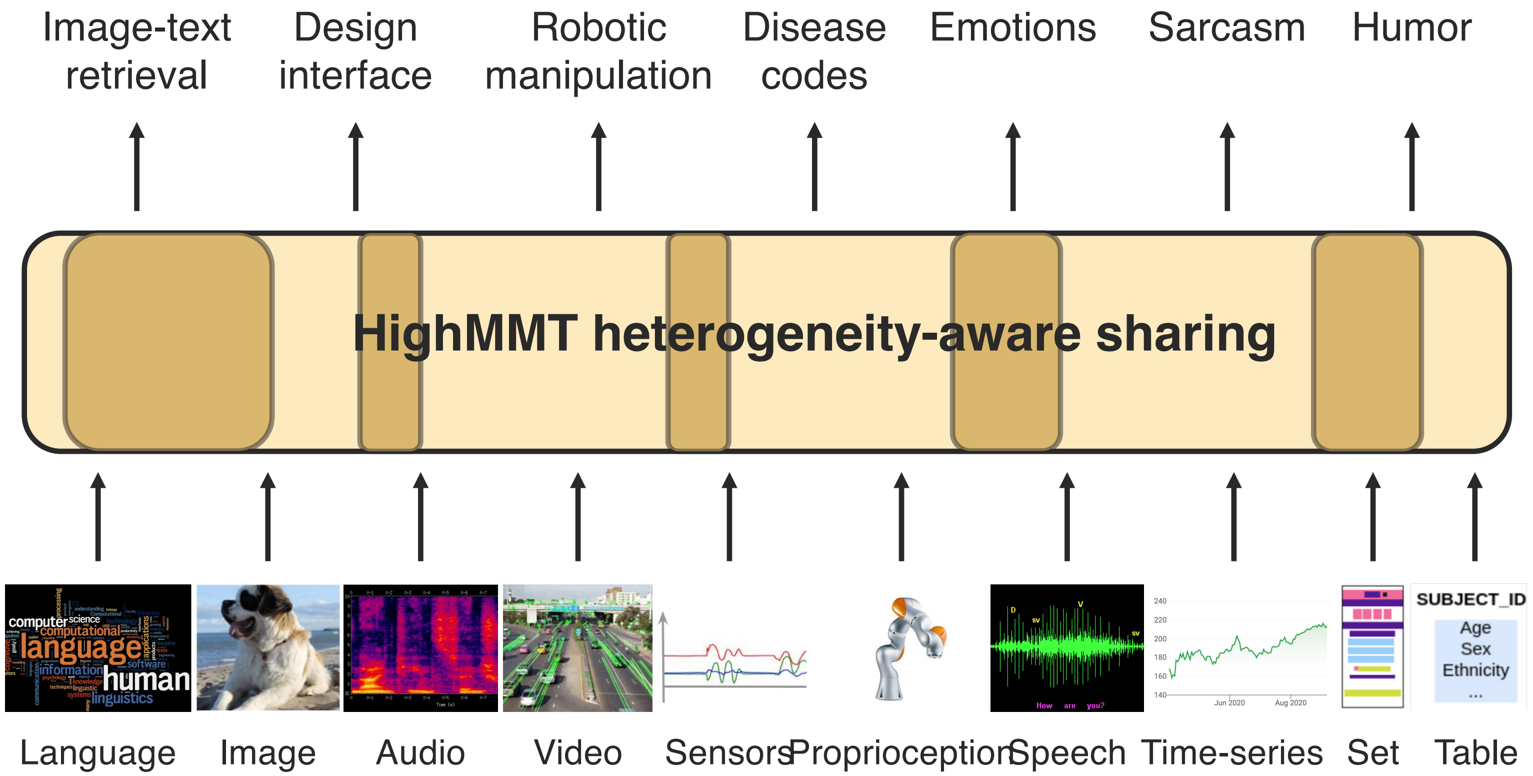
Information transfer, transfer learning perspective

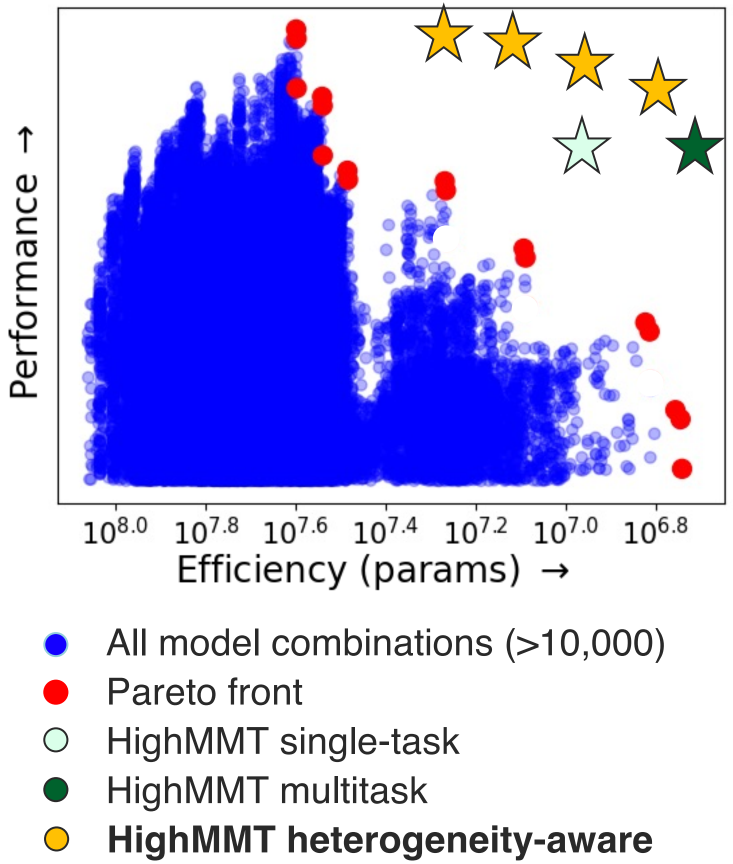


[Liang et al., HighMMT: Quantifying Modality & Interaction Heterogeneity for High-Modality Learning. TMLR 2022 revisions]

HighMMT + Quantifying Modality Heterogeneity

HighMMT heterogeneity-aware sharing: estimate heterogeneity to determine parameter sharing



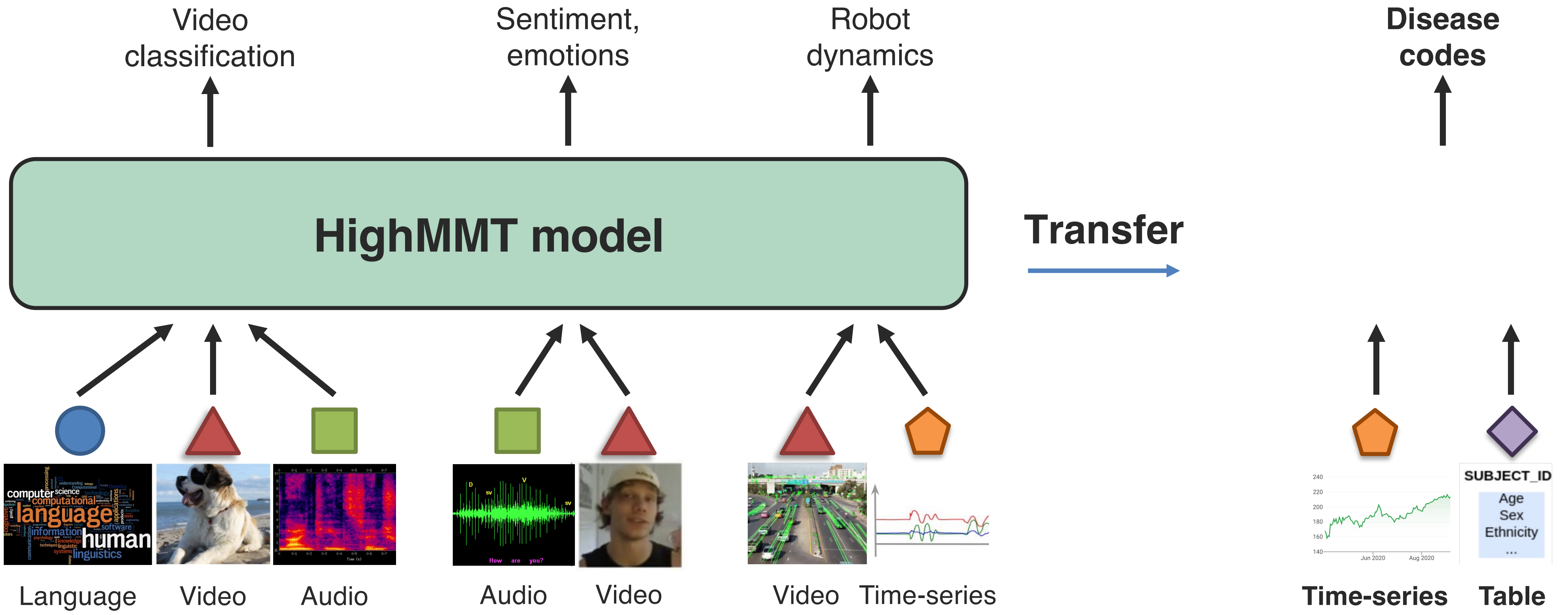


[Liang et al., HighMMT: Quantifying Modality & Interaction Heterogeneity for High-Modality Learning. TMLR 2022 revisions]

HighMMT

Transfer across partially observable modalities

HighMMT: unified model + parameter sharing + multitask and transfer learning



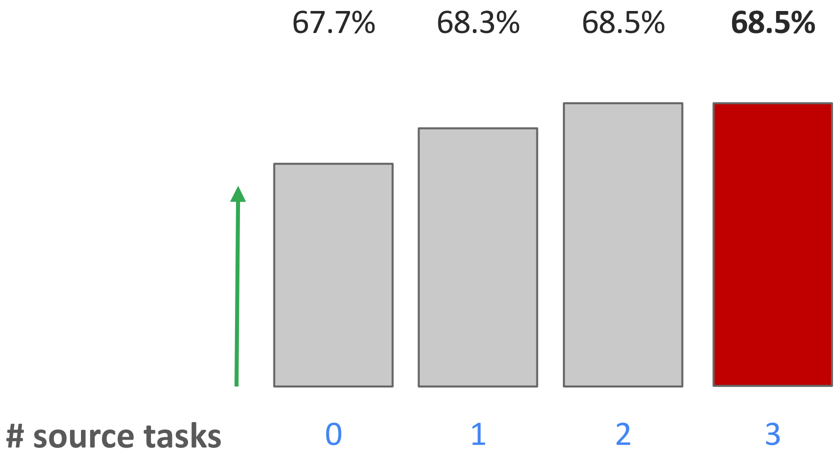
[Liang et al., HighMMT: Quantifying Modality & Interaction Heterogeneity for High-Modality Learning. TMLR 2022 revisions]

HighMMT

Transfer across partially observable modalities

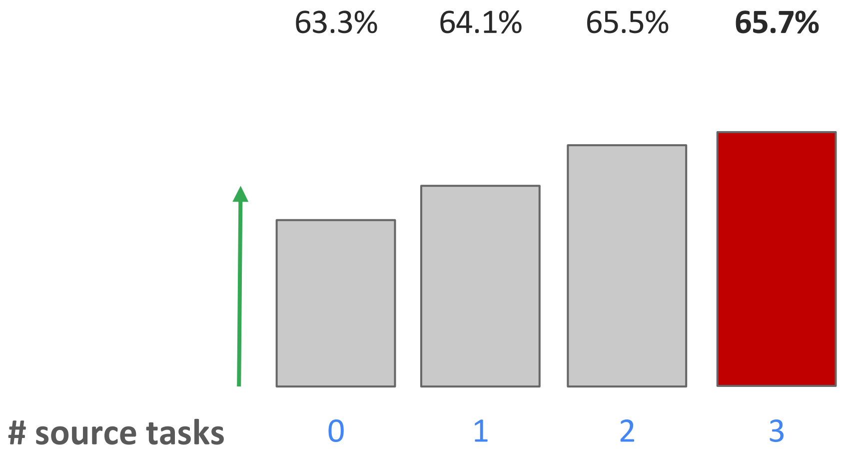
HighMMT: unified model + parameter sharing + multitask and transfer learning

Target task: MIMIC



(from different modalities, research areas, and tasks)

Target task: UR-FUNNY



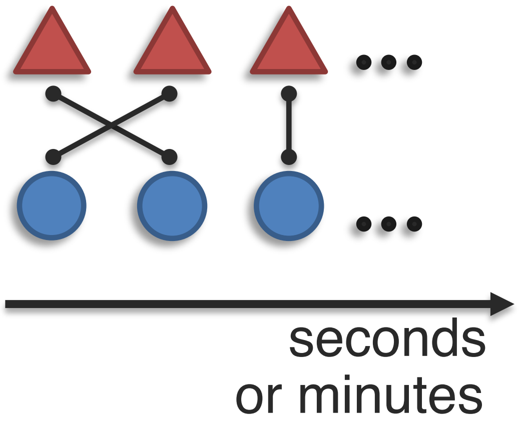
(from different modalities, research areas, and tasks)

Achieves both multitask and transfer capabilities across modalities and tasks

[Liang et al., HighMMT: Quantifying Modality & Interaction Heterogeneity for High-Modality Learning. TMLR 2022 revisions]

Future Direction: Long-term

Short-term

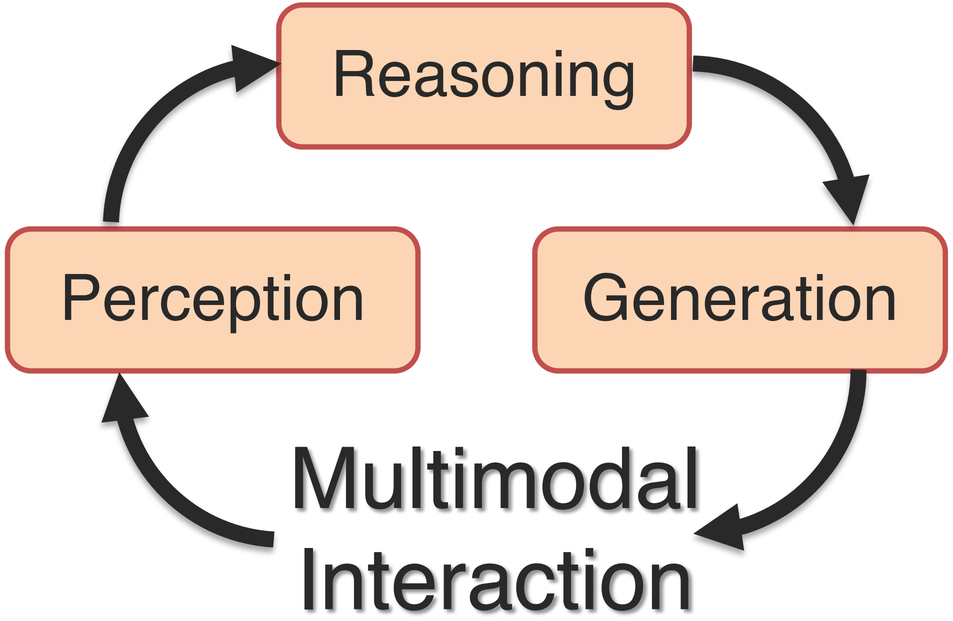


Long-term

Challenges:

Compositionality **Memory** **Personalization**

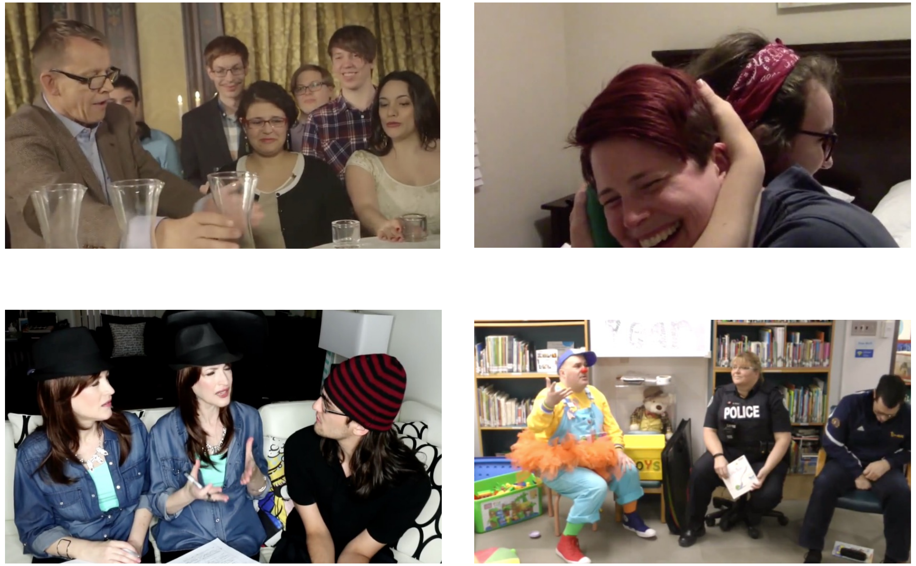
Future Direction: Interaction



Social-IQ

https://www.thesocialiq.com/

Social Intelligence



Challenges:

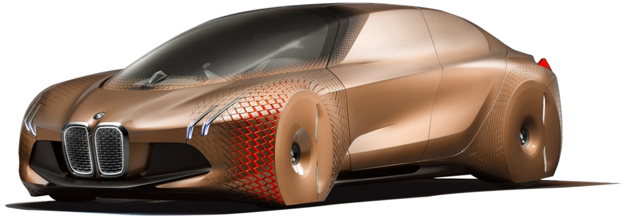
Multi-Party **Generation** **Ethics**

Future Direction: Real-world

MultiViz

https://github.com/pliang279/MultiViz

Healthcare Decision Support



Intelligent Interfaces and Vehicles

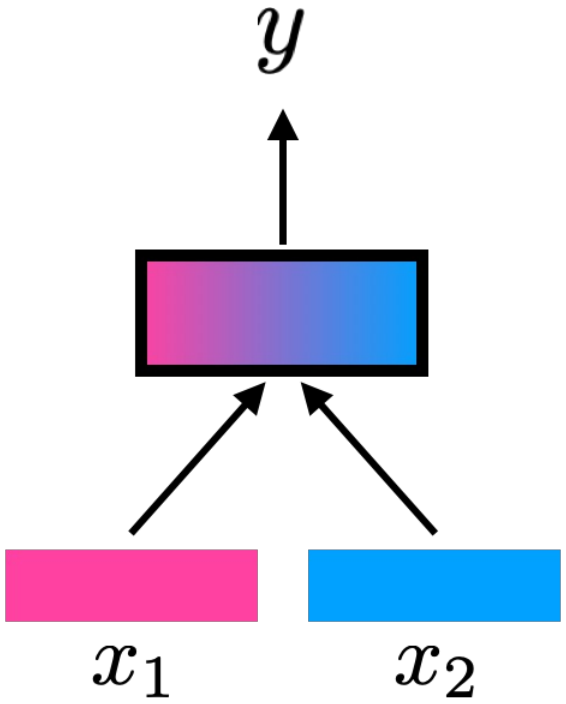
Online Learning and Education

Challenges:

Robustness **Fairness** **Generalization** **Interpretation**

Real-World Quantification

*How can we understand the modeling of **heterogeneity** and **interconnections** and gain insights for safe real-world deployment?*

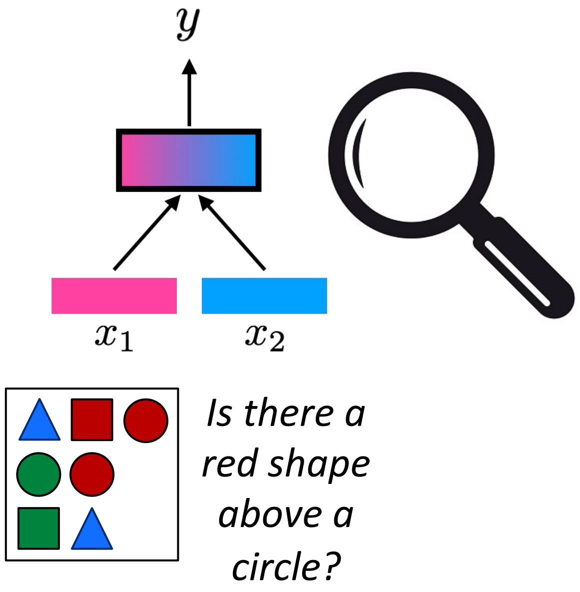


Internal mechanics

MultiViz: Interpreting Internal Mechanics

*How can we understand the modeling of **heterogeneity** and **interconnections** and gain insights for safe real-world deployment?*

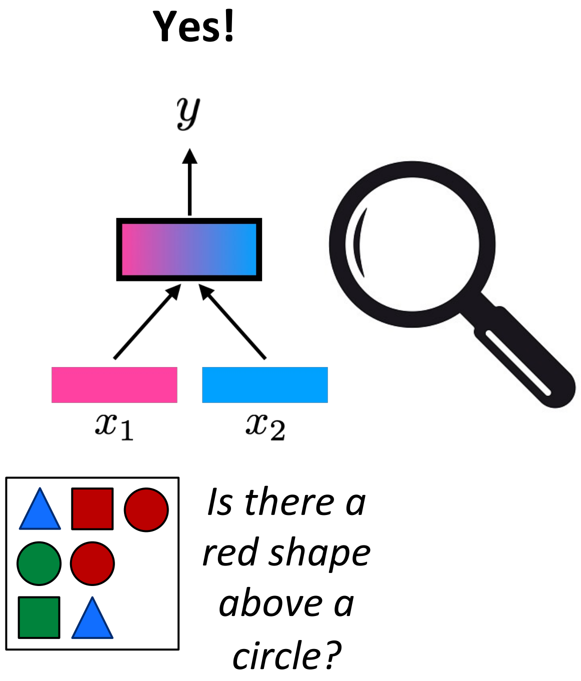
Yes!

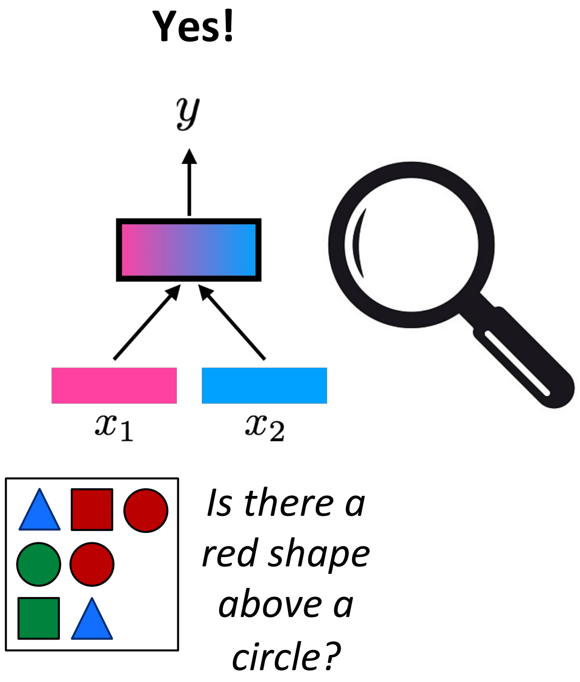


[Liang et al., MultiViz: Towards Visualizing and Understanding Multimodal Models. ICLR 2023, CHI 2023 Late Breaking Work]

MultiViz: Interpreting Internal Mechanics

*How can we understand the modeling of **heterogeneity** and **interconnections** and gain insights for safe real-world deployment?*



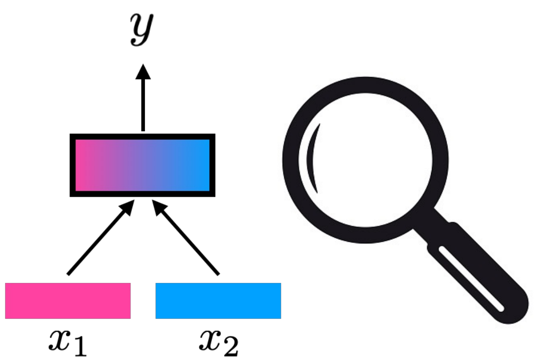


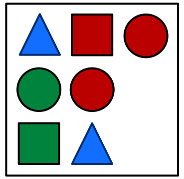
[Liang et al., MultiViz: Towards Visualizing and Understanding Multimodal Models. ICLR 2023, CHI 2023 Late Breaking Work]

MultiViz: Interpreting Internal Mechanics

Unimodal importance: Does the model correctly identify keywords in the question?

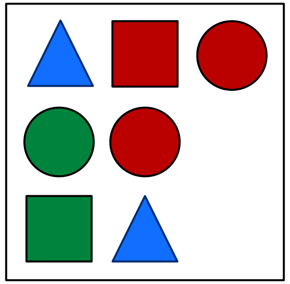
Yes!





Is there a red shape above a circle?

1. Unimodal importance



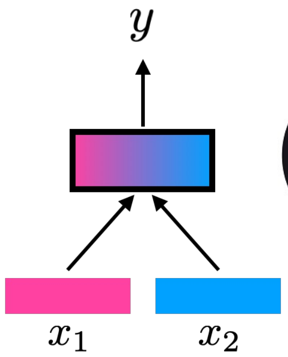
Is there a ***red shape above*** *a* ***circle****?*

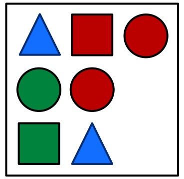
[Liang et al., MultiViz: Towards Visualizing and Understanding Multimodal Models. ICLR 2023, CHI 2023 Late Breaking Work]

MultiViz: Interpreting Internal Mechanics

Cross-modal interactions: Does the model correctly relate the question with the image?

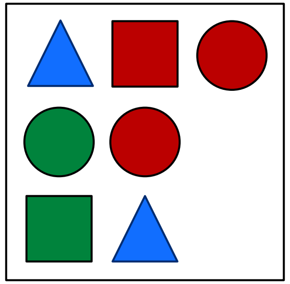
Yes!





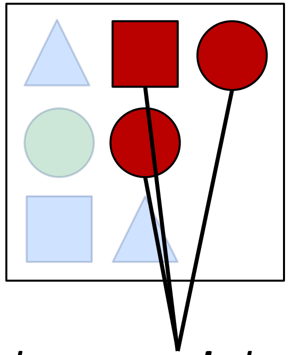
Is there a red shape above a circle?

1. Unimodal importance

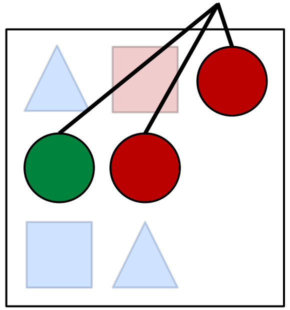


*Is there a **red shape above** a **circle**?*

2. Cross-modal interactions



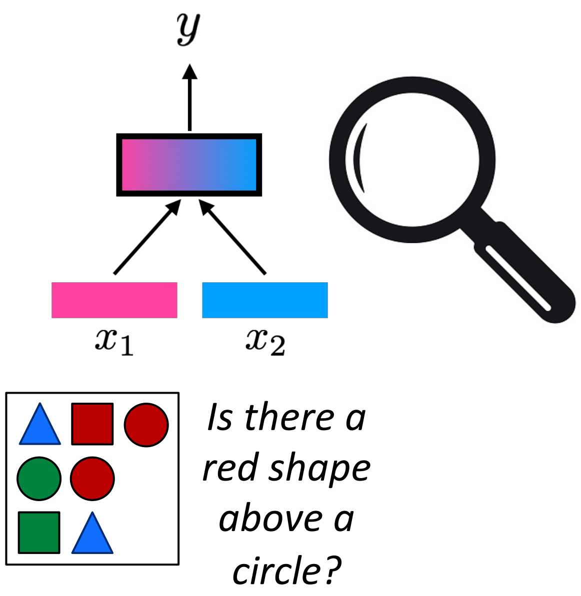
*Is there a **red** shape above a **circle**?*



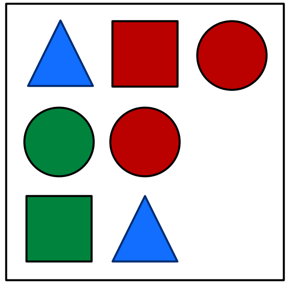
MultiViz: Interpreting Internal Mechanics

Multimodal representations: Does the model consistently assign concepts to features?

Yes!

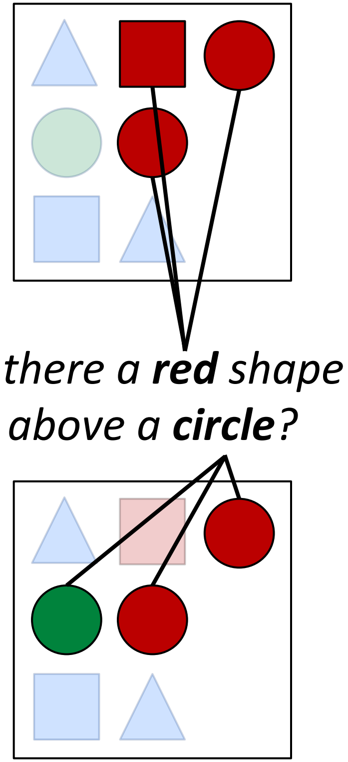


1. Unimodal importance

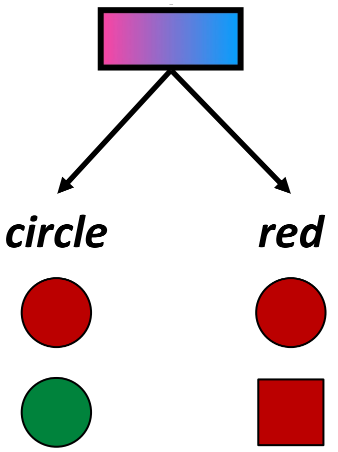


*Is there a **red shape above** a **circle**?*

2. Cross-modal interactions

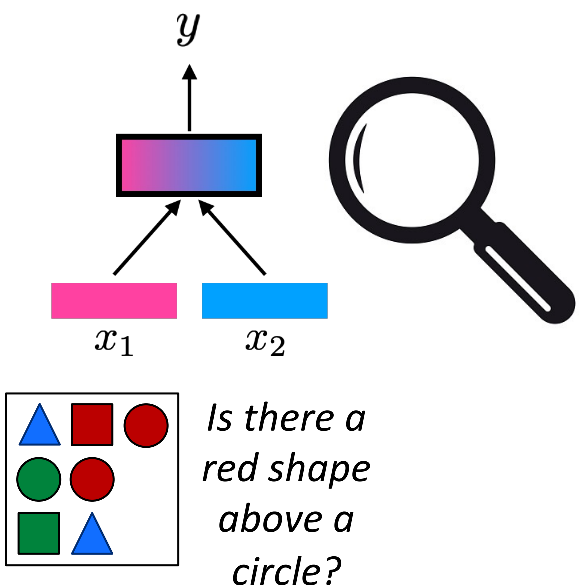


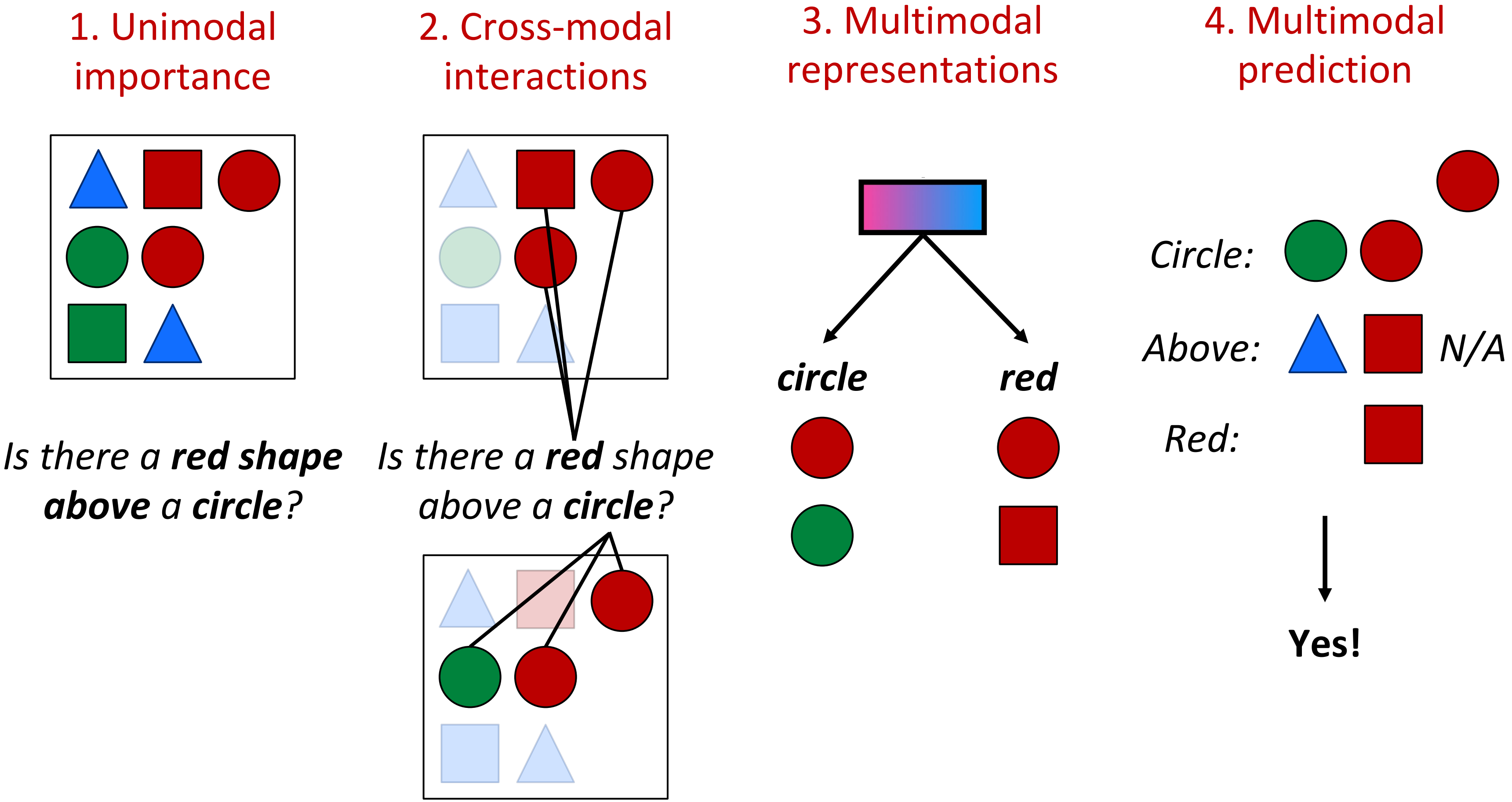
3. Multimodal representations



MultiViz: Interpreting Internal Mechanics

Multimodal prediction: Does the model correctly compose question and image information?





MultiViz: Interpreting Internal Mechanics

How can we interpret cross-modal interactions in multimodal models?

2. Cross-modal interactions

Statistical non-additive interactions [Friedman & Popescu, 2008, Sorokina et al., 2008]

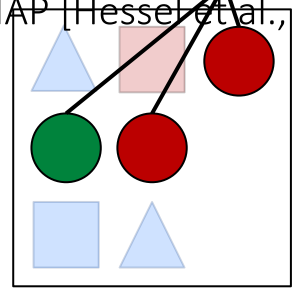
f exhibits interactions between 2 features x_1 and x_2 iff
f cannot be decomposed into a sum of unimodal subfunctions g_1, g_2
such that $f(x_1, x_2) = g_1(x_1) + g_2(x_2)$.

$$f \text{ exhibits interactions between 2 features } x_1 \text{ and } x_2 \text{ iff } \frac{\partial f^2}{\partial x_1 \partial x_2} > 0.$$

Natural second-order extension of gradient-based approaches!

*Is there a **red** shape above a **circle**?*

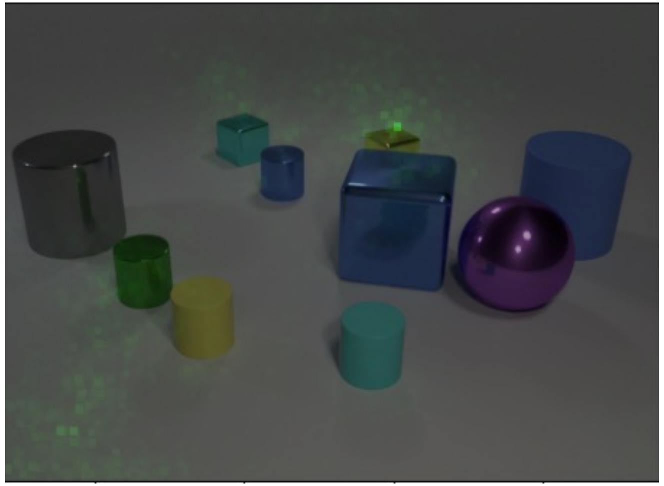
Also related: EMAP [Hessel et al., 2020], DIME [Lyu et al., 2022]



MultiViz: Interpreting Internal Mechanics

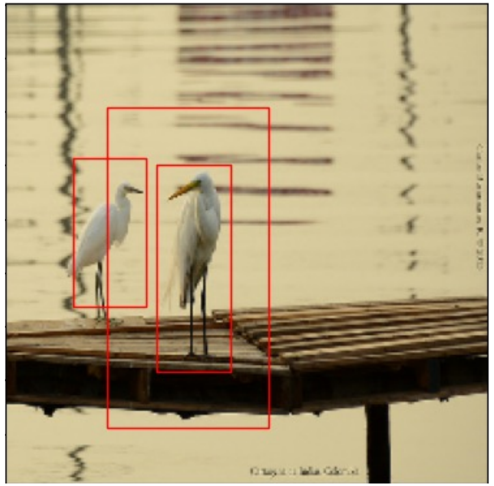
How can we interpret cross-modal interactions in multimodal models?

CLEVR



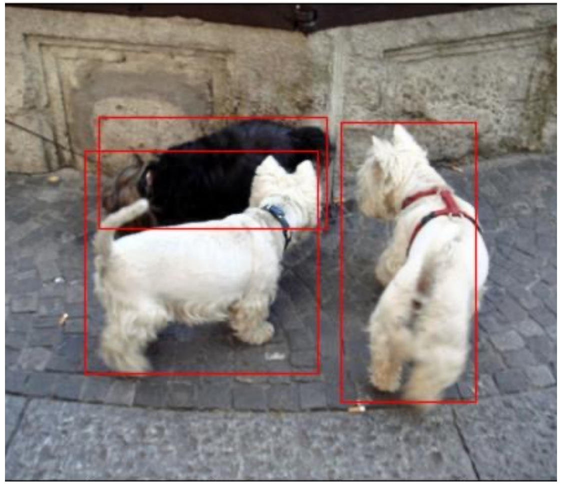
*The other small shiny thing that is the same shape as the **tiny yellow shiny object** is what color?*

VQA 2.0



*How many **birds**?*

Flickr-30k



***Three small dogs**, two white and one black and white, on a sidewalk.*

CMU-MOSEI

*Why am I spending my money watching this? **(sigh)** I think I was more **sad**…*

Correspondences

Relationships

[Liang et al., MultiViz: Towards Visualizing and Understanding Multimodal Models. ICLR 2023, CHI 2023 Late Breaking Work]

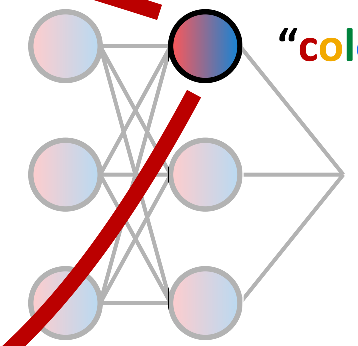
MultiViz: Interpreting Internal Mechanics

How can we understand multimodal representations?

What color is the tie of the second man to the left?

Local analysis

3. Multimodal representations



Global analysis

What color is the Salisbury Rd sign?

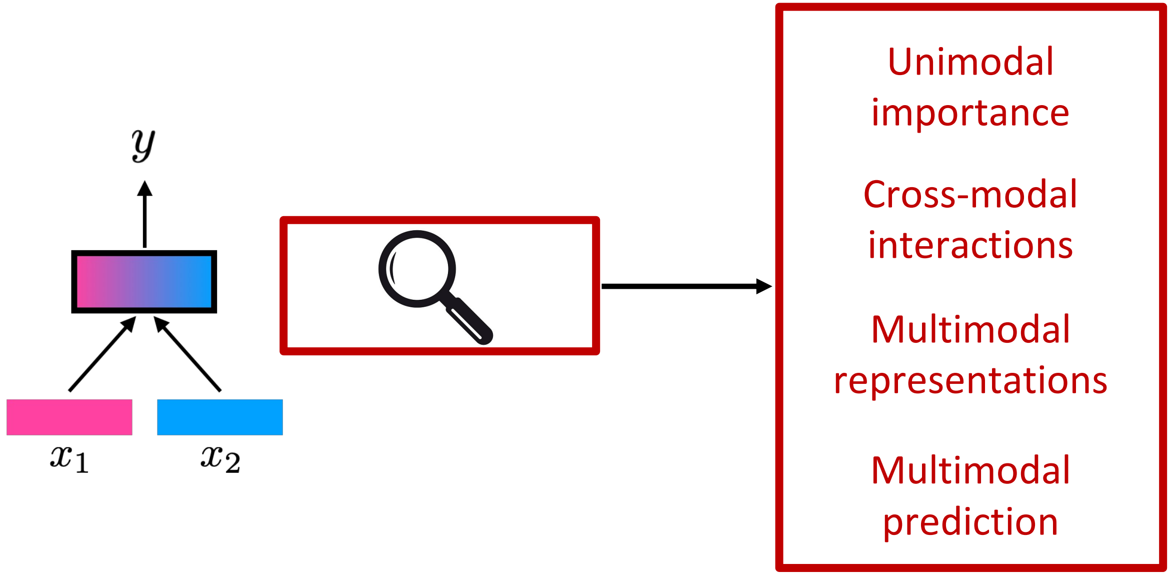
What color is the building?

What color are the checkers on the wall?

MultiViz: Interpreting Internal Mechanics

How can we evaluate the success of interpreting internal mechanics?

Problem: real-world datasets and models do not have
unimodal importance, cross-modal interactions, representations annotated!



[Liang et al., MultiViz: Towards Visualizing and Understanding Multimodal Models. ICLR 2023, CHI 2023 Late Breaking Work]

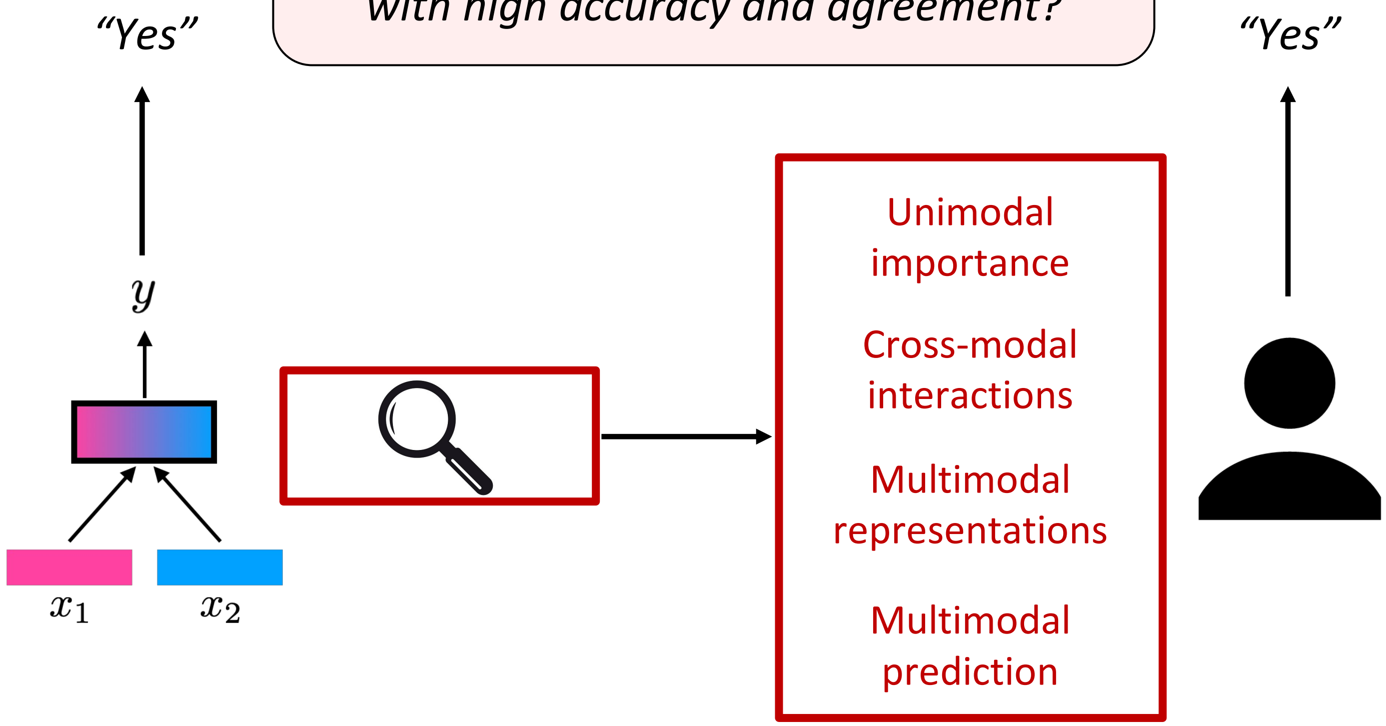
MultiViz: Interpreting Internal Mechanics

Open challenges

How can we evaluate the success of interpreting internal mechanics?

1. Model simulation

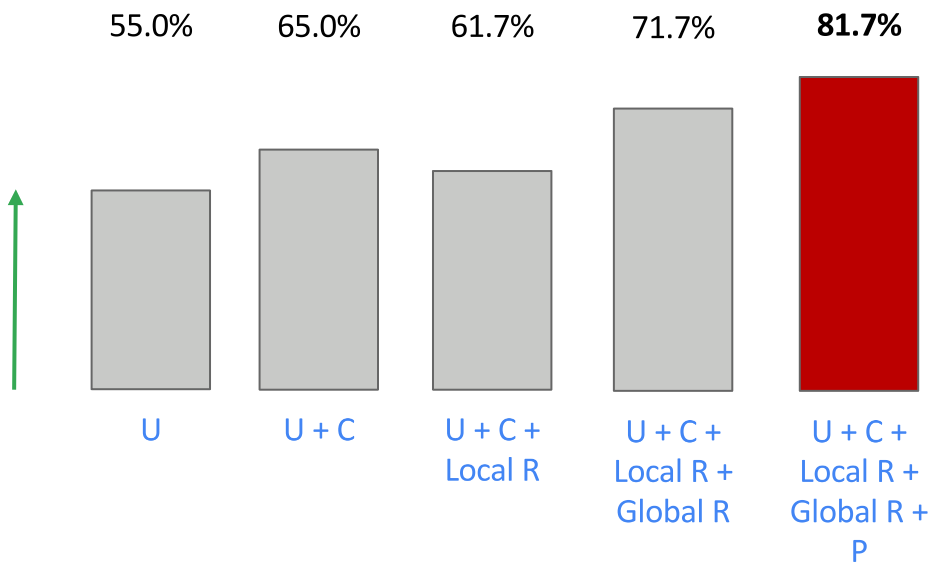
Can humans reproduce model predictions with high accuracy and agreement?



[Liang et al., MultiViz: Towards Visualizing and Understanding Multimodal Models. ICLR 2023, CHI 2023 Late Breaking Work]

MultiViz: Interpreting Internal Mechanics

How can we evaluate the success of interpreting internal mechanics?



MultiViz stages leads to higher accuracy and agreement

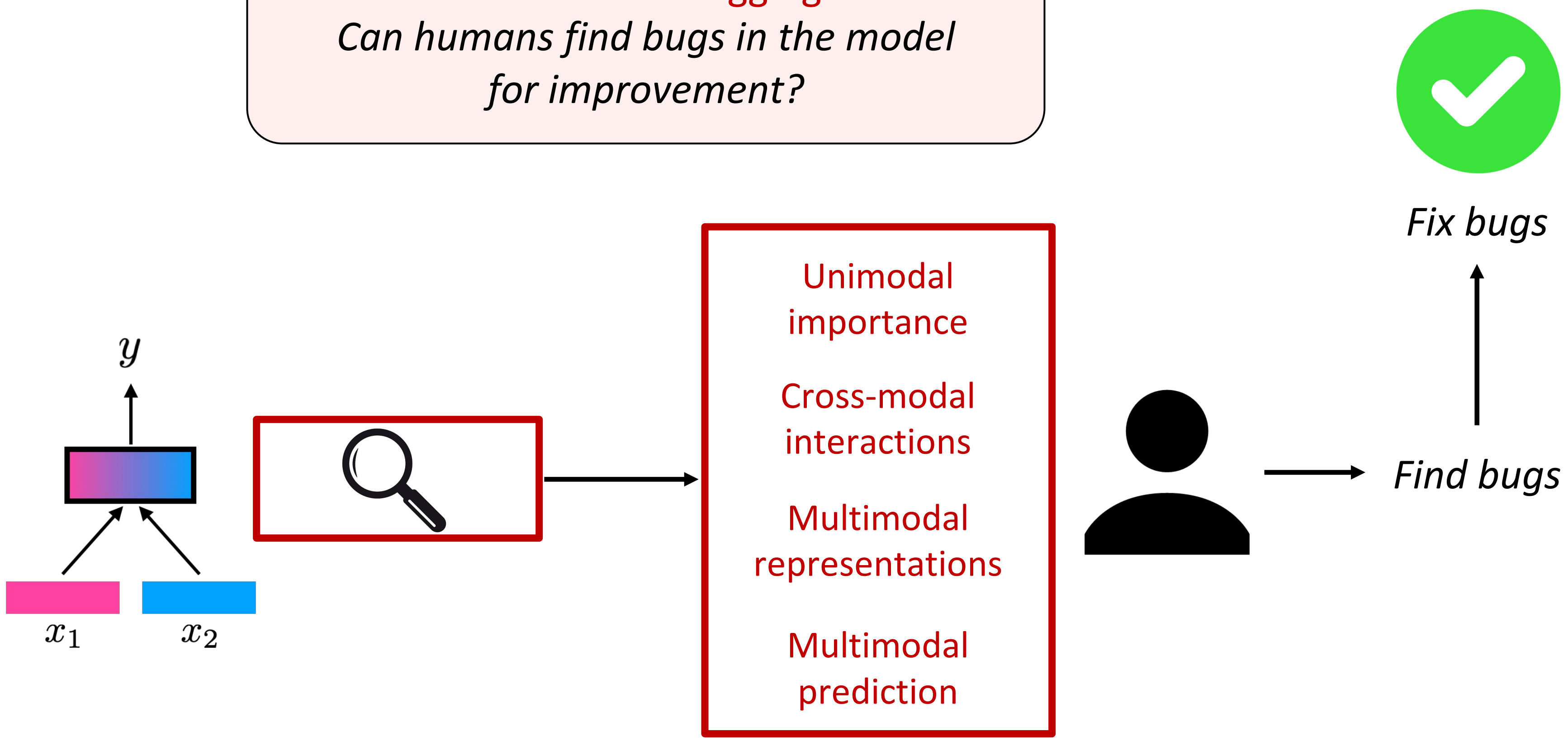
[Liang et al., MultiViz: Towards Visualizing and Understanding Multimodal Models. ICLR 2023, CHI 2023 Late Breaking Work]

MultiViz: Interpreting Internal Mechanics

Open challenges

How can we evaluate the success of interpreting internal mechanics?

2. Model debugging
Can humans find bugs in the model for improvement?



[Liang et al., MultiViz: Towards Visualizing and Understanding Multimodal Models. ICLR 2023, CHI 2023 Late Breaking Work]

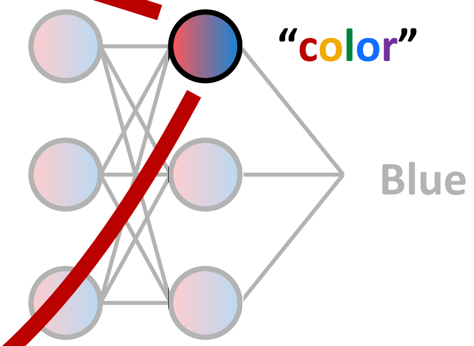
MultiViz: Interpreting Internal Mechanics

How can we understand multimodal representations?

What color is the tie of the second man to the left?

Local analysis

3. Multimodal representations



Global analysis

What color is the Salisbury Rd sign?

What color is the building?

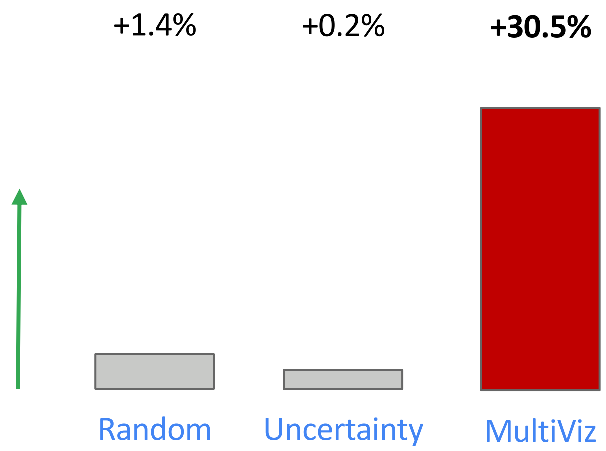
What color are the checkers on the wall?

“Models pick up cross-modal interactions but fail in identifying color!”

MultiViz: Interpreting Internal Mechanics

How can we evaluate the success of interpreting internal mechanics?

"Models pick up cross-modal interactions but fail in identifying color!" → *Add targeted examples involving color.*



Side note: we used this to discover a bug in a popular deep learning code repository.

Transformers

MultiViz enables error analysis and debugging of multimodal models

[Liang et al., MultiViz: Towards Visualizing and Understanding Multimodal Models. ICLR 2023, CHI 2023 Late Breaking Work]

What is Multimodal? → Why is it hard? → What is next?

What is Multimodal?

Heterogeneous
+
Connected
+
Interacting

Why is it hard?

Representation
Alignment
Reasoning
Generation
Transference
Quantification

What is next?

Heterogeneity
High-modality
Long-term
Interaction
Real-world

https://cmu-multicomp-lab.github.io/mmml-course/fall2022/

[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. arXiv 2022]

https://www.cs.cmu.edu/~pliang/
pliang@cs.cmu.edu
@pliang279