Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 12.1: Quantification

Paul Liang

** Co-lecturer: Louis-Philippe Morency.*
Original course co-developed with Tadas Baltrusaitis.
Spring 2021 edition taught by Yonatan Bisk

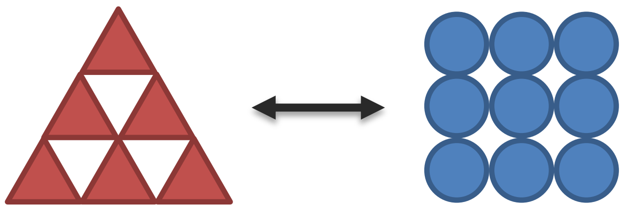
Reading Assignments are Back!

- Four main steps for the reading assignments
  - Monday 8pm: Official start of the assignment
  - Wednesday 8pm: Select your paper
  - **Friday 8pm:** Post your summary
  - **Monday 8pm:** Post your extra comments (5 posts)
- **4 papers:** Latent diffusion, explanations on VQA models, interaction quantification, benchmark quantification

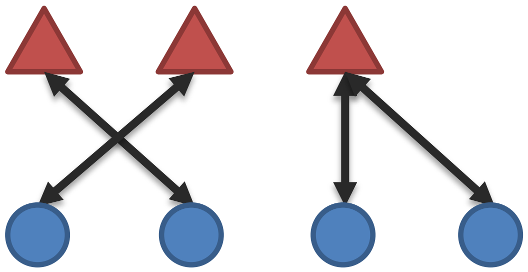
Quantification

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

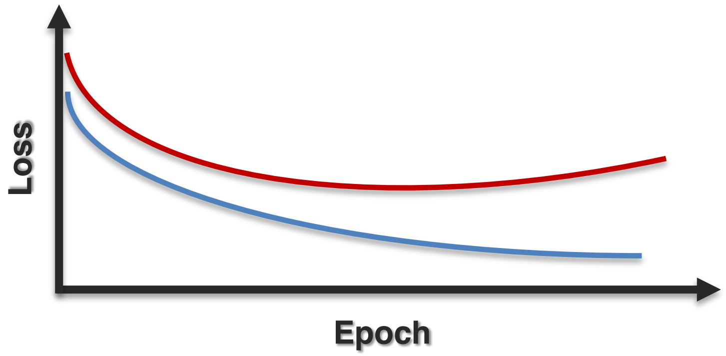
A Heterogeneity



B Interactions

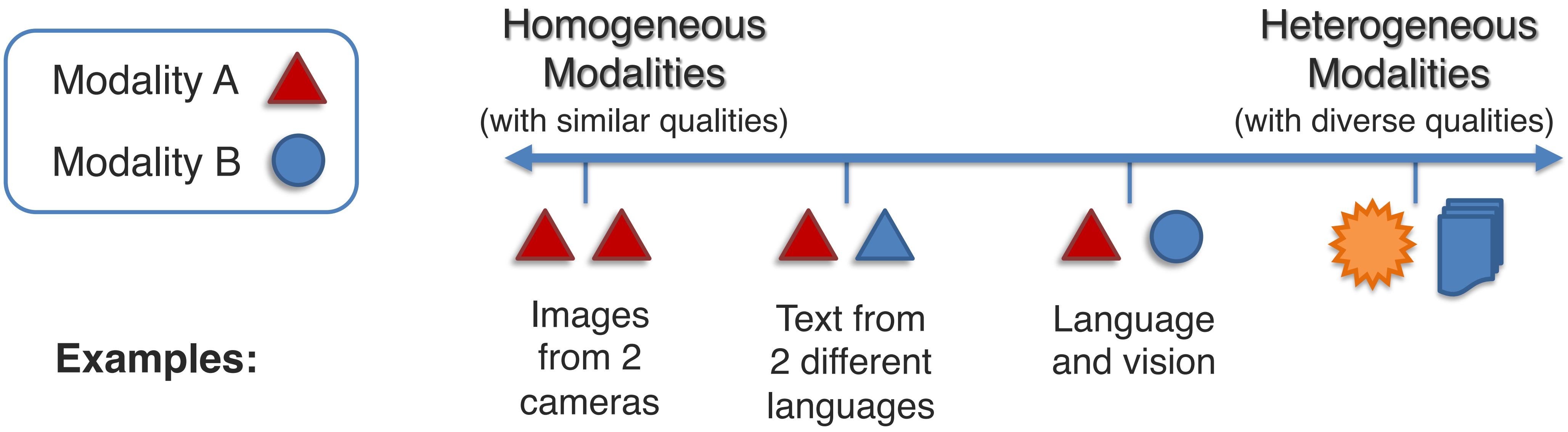


C Learning



Sub-Challenge 6a: Heterogeneity

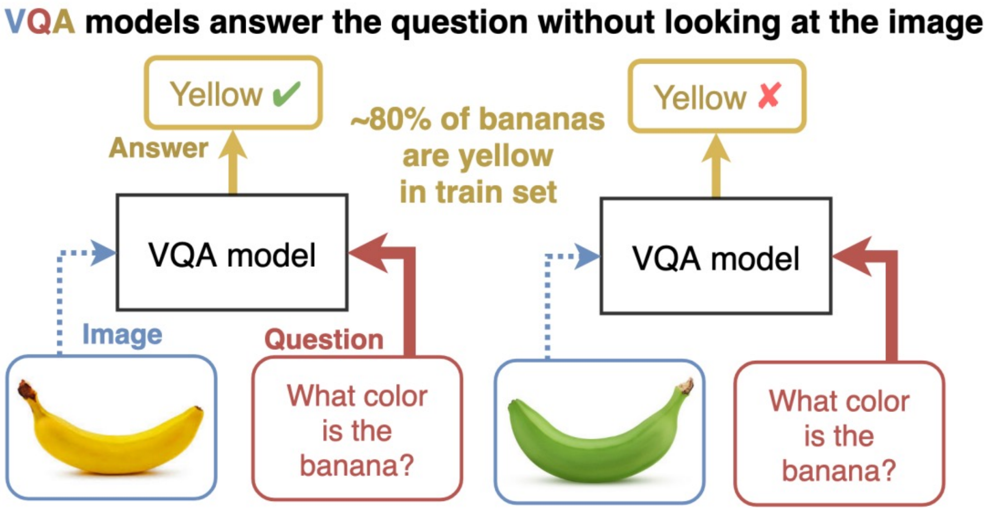
Definition: Quantifying the dimensions of heterogeneity in multimodal datasets and how they subsequently influence modeling and learning.



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

Modality Biases

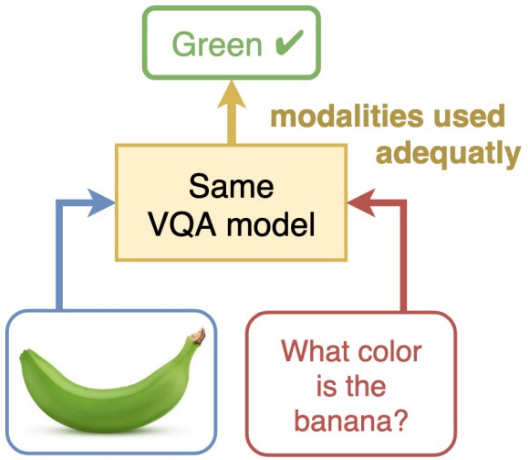
Heterogeneity in information and relevance
Unimodal biases and modality collapse



Balancing modalities

Balancing training

Not the case when trained with RUBi



[Wu et al., Characterizing and Overcoming the Greedy Nature of Learning in Multi-modal Deep Neural Networks. ICML 2022]
[Javaloy et al., Mitigating Modality Collapse in Multimodal VAEs via Impartial Optimization. ICML 2022]
[Goyal et al., Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. CVPR 2017]

Modality Biases

Heterogeneity in information and relevance

Fairness and social biases – unimodal social biases

Finding: Image captioning models capture spurious correlations between gender and generated actions

Wrong

Baseline:
*A **man** sitting at a desk with a laptop computer.*

[Hendricks et al., Women also Snowboard: Overcoming Bias in Captioning Models. ECCV 2018]

Modality Biases

Heterogeneity in information and relevance

Fairness and social biases – unimodal social biases

Finding: Image captioning models capture spurious correlations between gender and generated actions

Baseline:
*A **man** sitting at a desk with a laptop computer.*

Our Model:
*A **woman** sitting in front of a laptop computer.*

[Hendricks et al., Women also Snowboard: Overcoming Bias in Captioning Models. ECCV 2018]

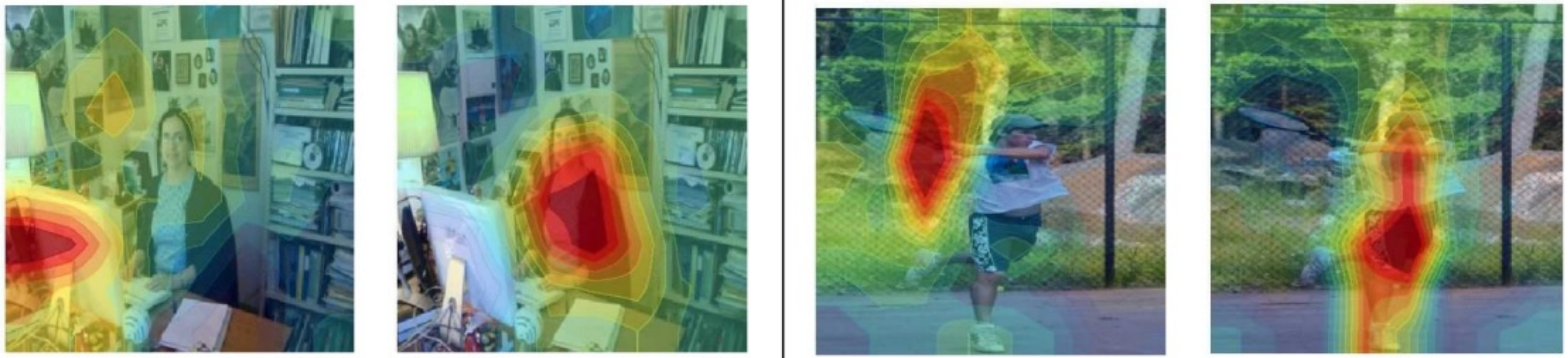
Modality Biases

Heterogeneity in information and relevance

Fairness and social biases – unimodal social biases

Finding: Image captioning models capture spurious correlations between gender and generated actions

Wrong	Right for the Right Reasons	Right for the Wrong Reasons	Right for the Right Reasons



Baseline:	Our Model:	Baseline:	Our Model:
*A **man** sitting at a desk with a laptop computer.*	*A **woman** sitting in front of a laptop computer.*	*A **man** holding a tennis racquet on a tennis court.*	*A **man** holding a tennis racquet on a tennis court.*

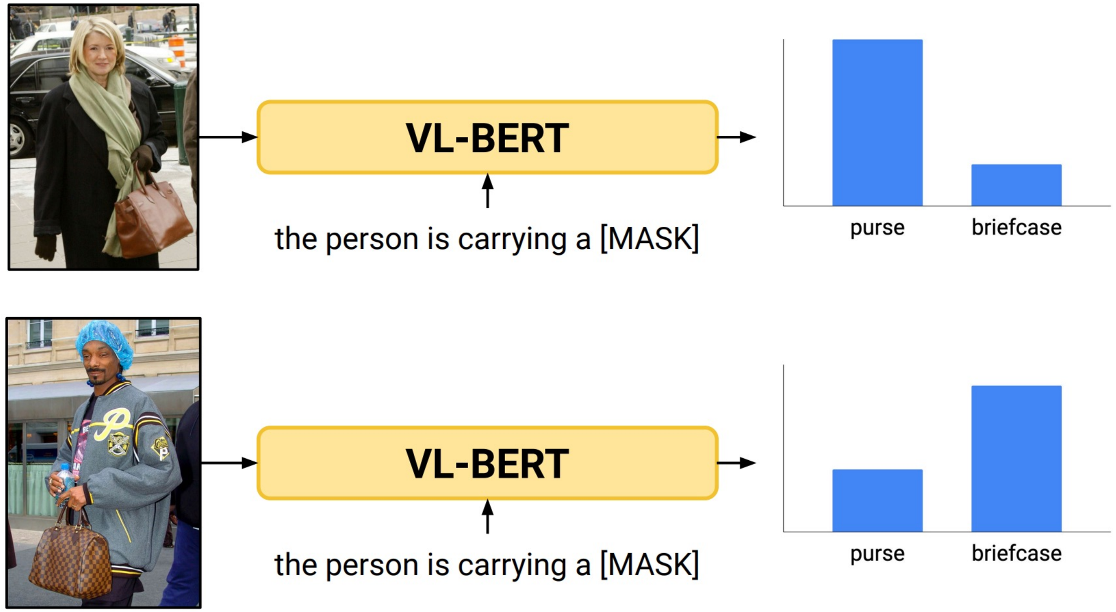
[Hendricks et al., Women also Snowboard: Overcoming Bias in Captioning Models. ECCV 2018]

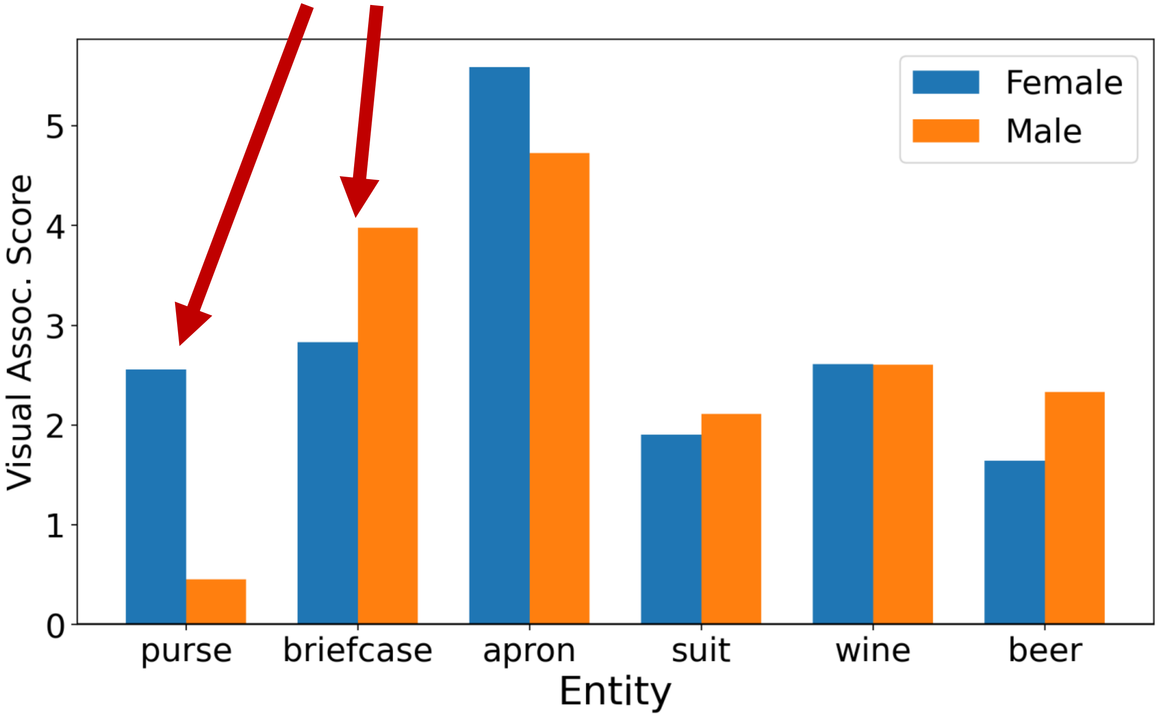
Modality Biases

Heterogeneity in information and relevance

Fairness and social biases – cross-modal interactions worsen social biases

Visual information makes model more confident
in reinforcing gender stereotypes





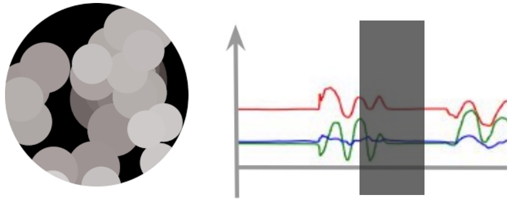
[Srinivasan and Bisk, Worst of Both Worlds: Biases Compound in Pre-trained Vision-and-Language Models. NAACL 2022]

Noise Topologies and Robustness

Heterogeneity in noise

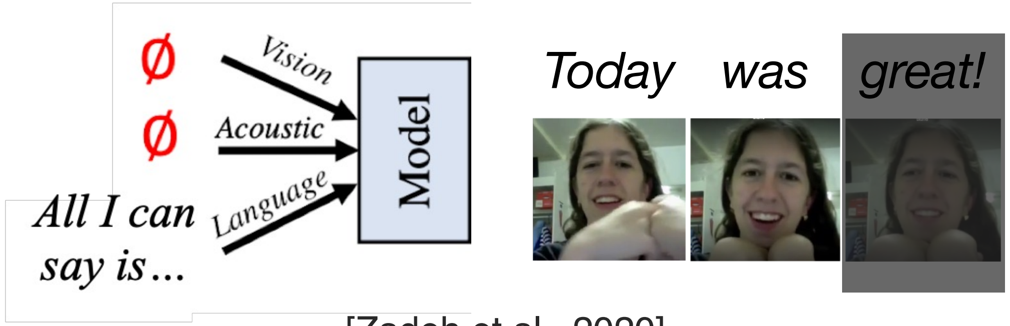
Modality-specific robustness

noise ⟶ nosie



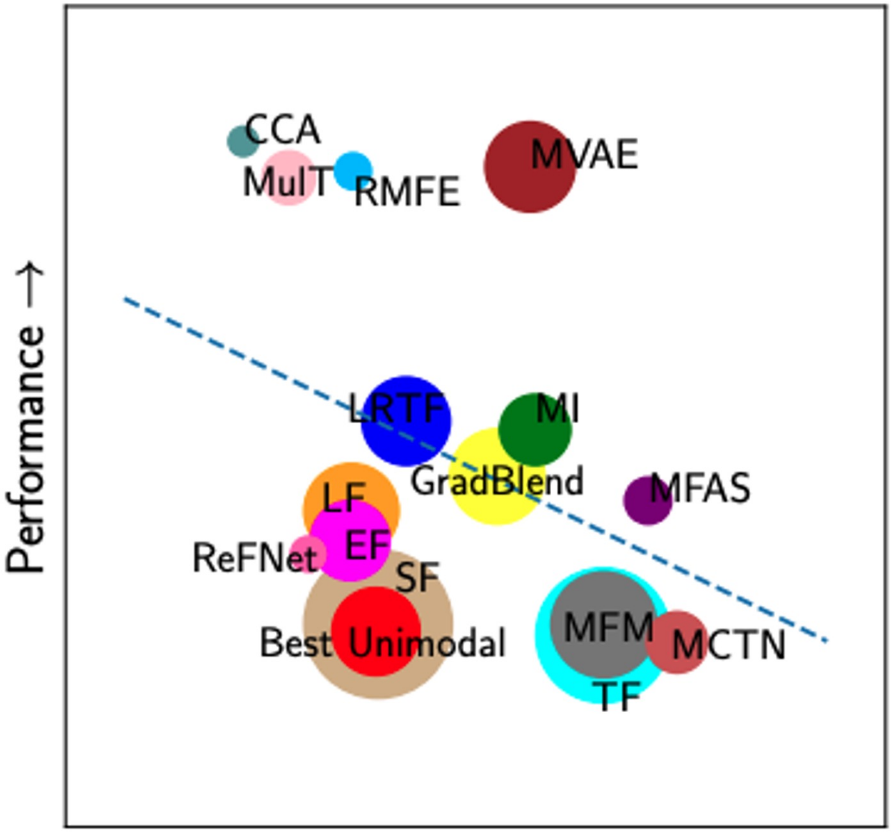
[Belinkov & Bisk, 2018; Subramaniam et al., 2009; Boyat & Joshi, 2015]

Multimodal robustness



[Zadeh et al., 2020]

Strong tradeoffs between performance and robustness



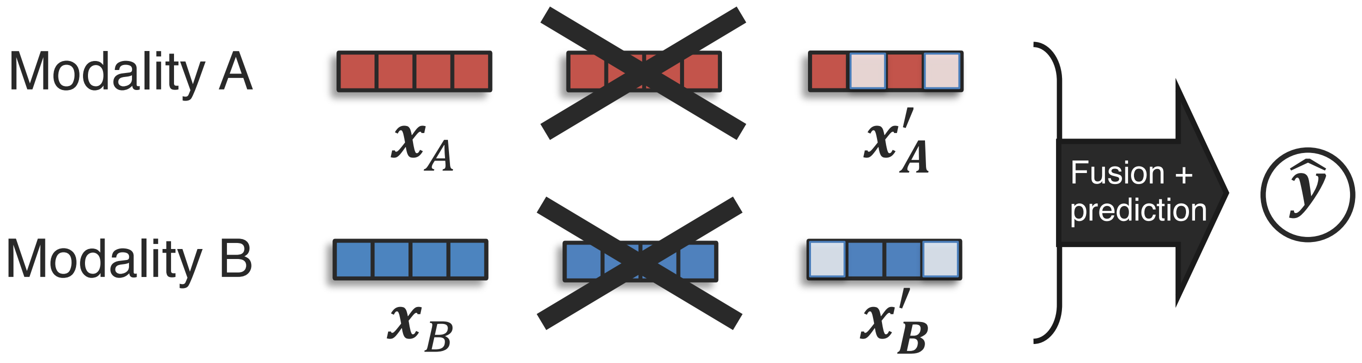
rate of accuracy drops

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

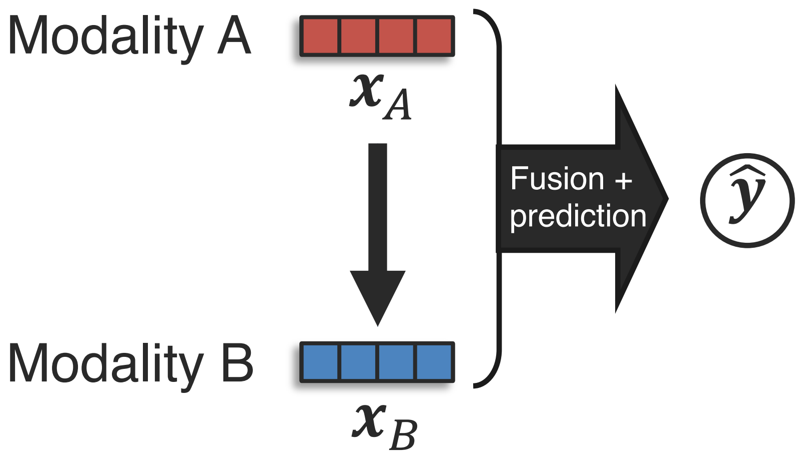
Noise Topologies and Robustness

Several approaches towards more robust models

Robust data + training



Infer missing modalities



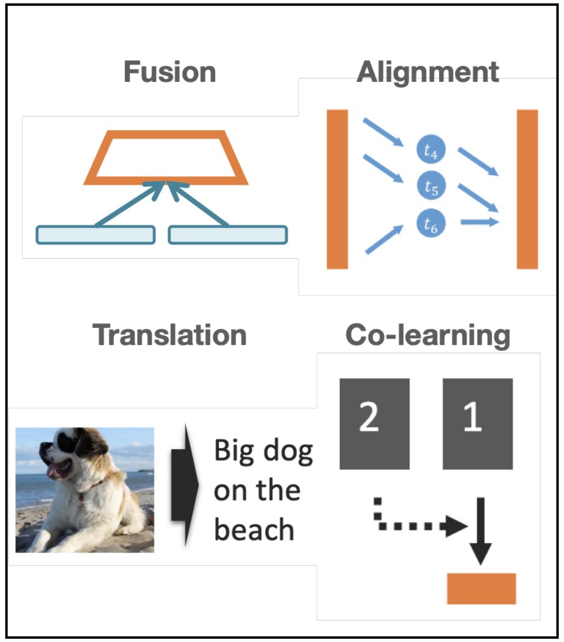
Translation model
Joint probabilistic model

[Ngiam et al., Multimodal Deep Learning. ICML 2011]
[Srivastava and Salakhutdinov, Multimodal Learning with Deep Boltzmann Machines. JMLR 2014]
[Tran et al., Missing Modalities Imputation via Cascaded Residual Autoencoder. CVPR 2017]
[Pham et al., Found in Translation: Learning Robust Joint Representations via Cyclic Translations Between Modalities. AAAI 2019]

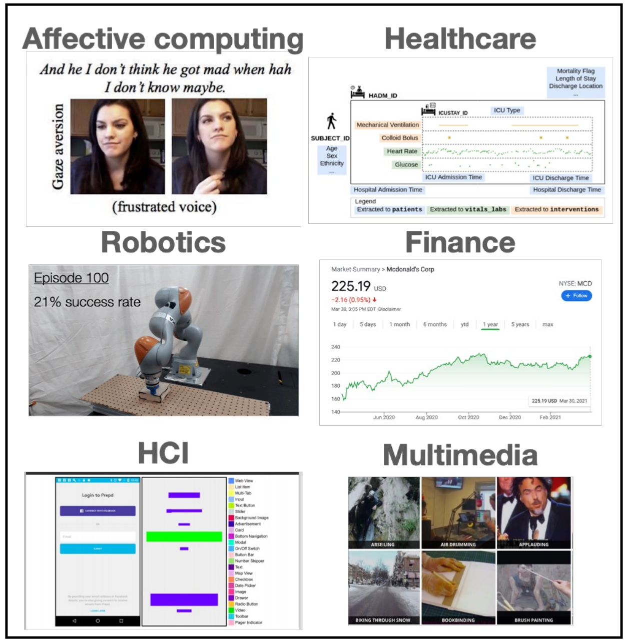
Going Beyond Language, Vision, and Audio

MultiBench datasets → MultiBench data loader → MultiZoo model → MultiBench evaluator → MultiBench leaderboard

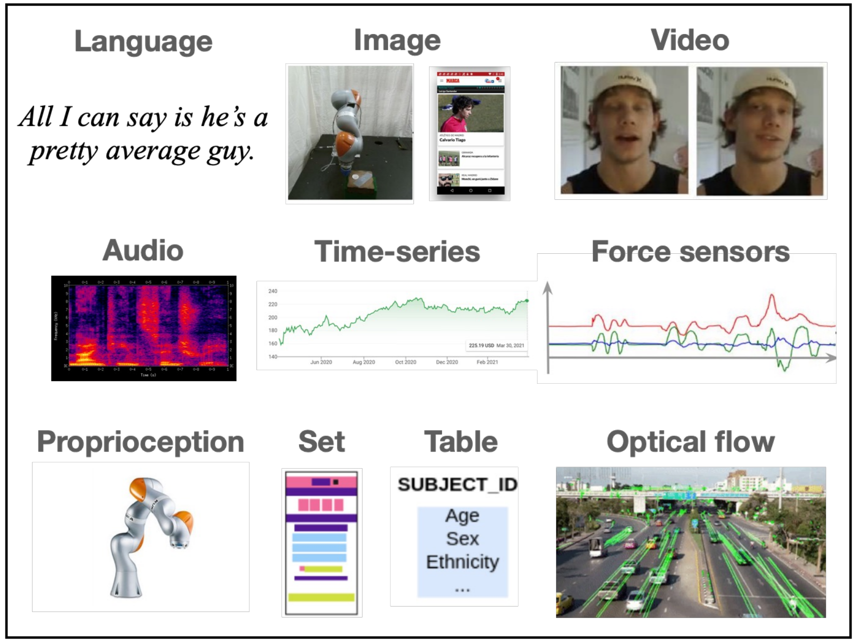
Challenges



Domains



Modalities

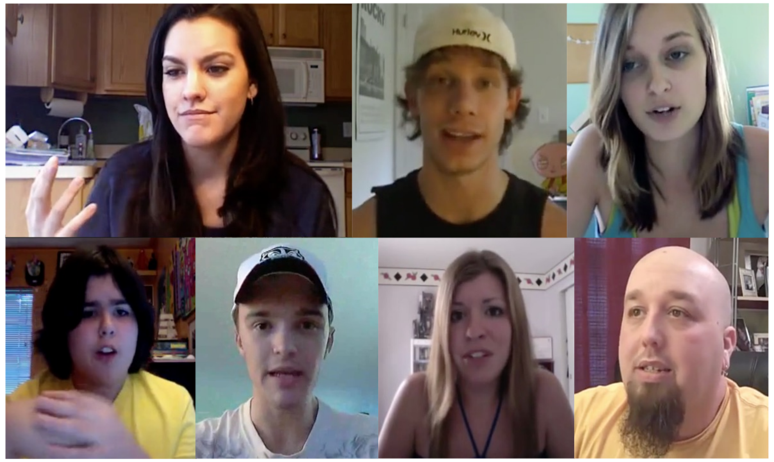


[Liang et al., MultiBench: Multiscale Benchmarks for Multimodal Representation Learning. NeurIPS 2021]

MultiBench: Generalization to Diverse Modalities

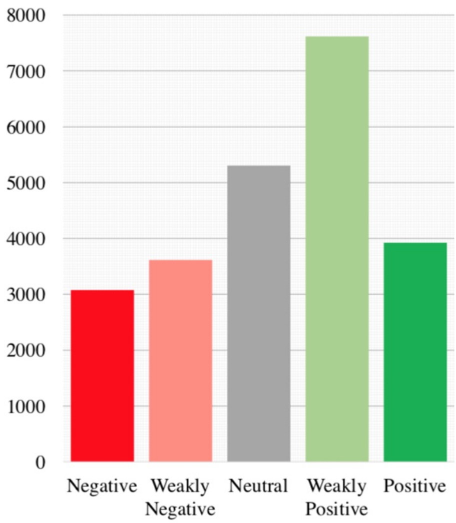
Multimodal affect recognition

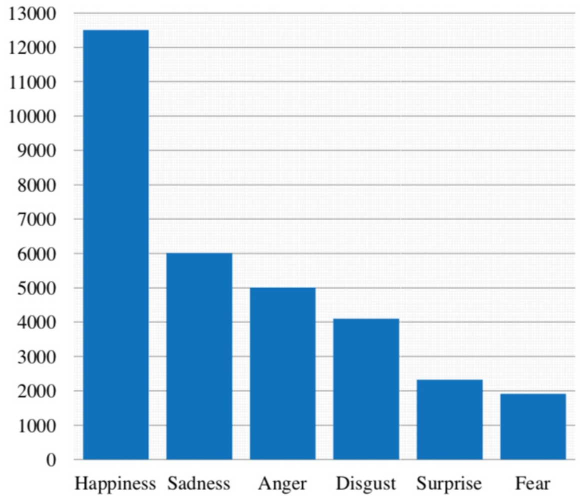
1,000 speakers



250 topics

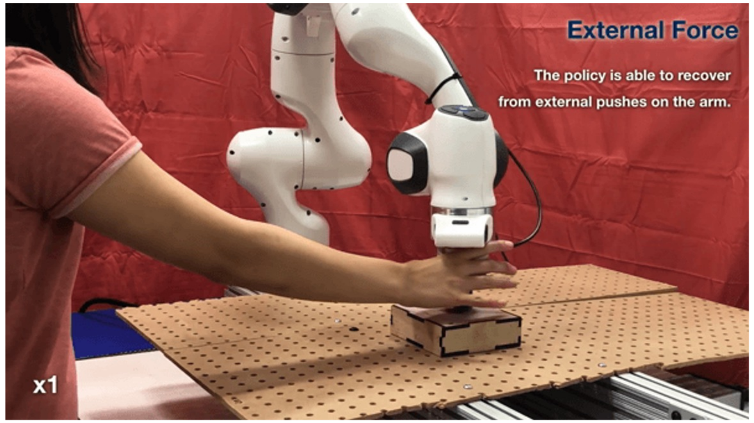
Diverse annotations





MultiBench: Generalization to Diverse Modalities

Multisensor fusion in robotics

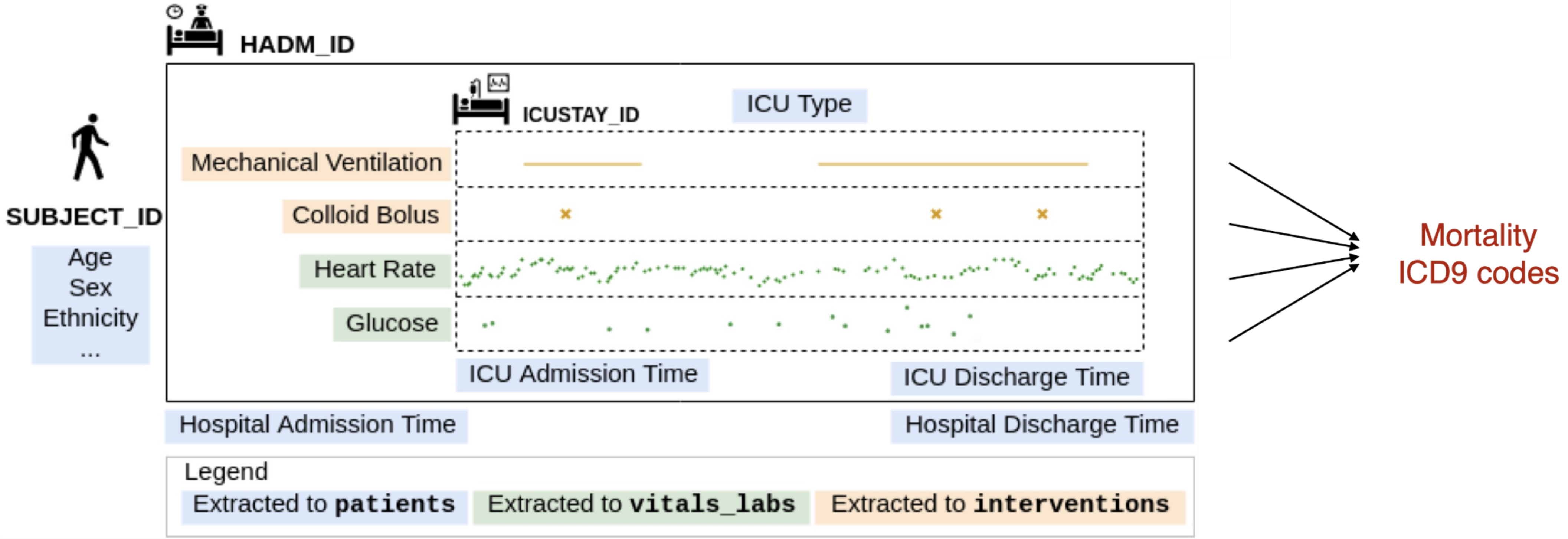


[Lee et al., ICRA 2019]

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

MultiBench: Generalization to Diverse Modalities

Multimodal learning in healthcare



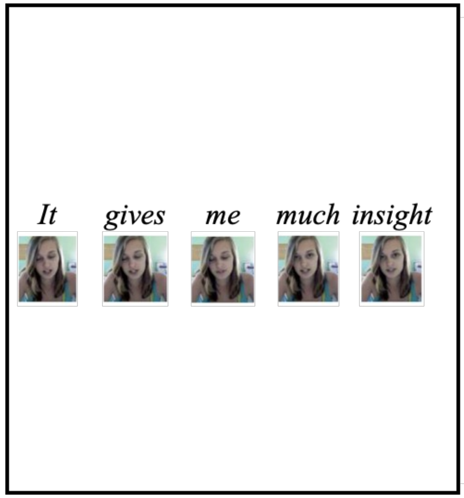
[Johnson et al., Nature 2016]

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

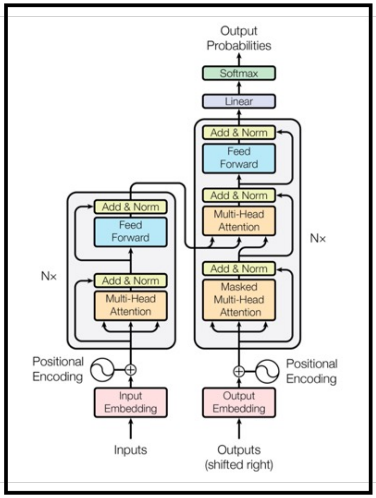
MultiBench: Generalization to Diverse Modalities

MultiBench datasets → MultiBench data loader → **MultiZoo model** → MultiBench evaluator → MultiBench leaderboard

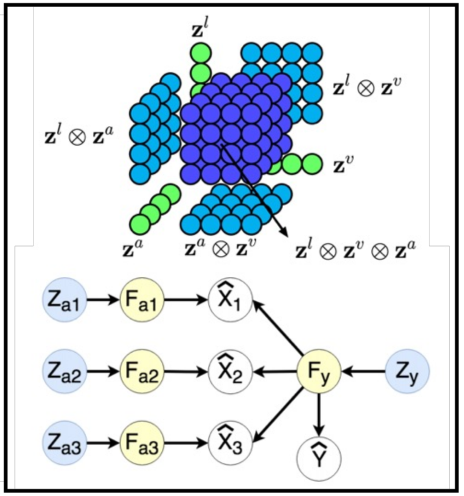
Data preprocessing	Unimodal models	Fusion paradigms	Optimization objectives	Training procedures



[Chen et al., 2017]



[Vaswani et al., 2017]



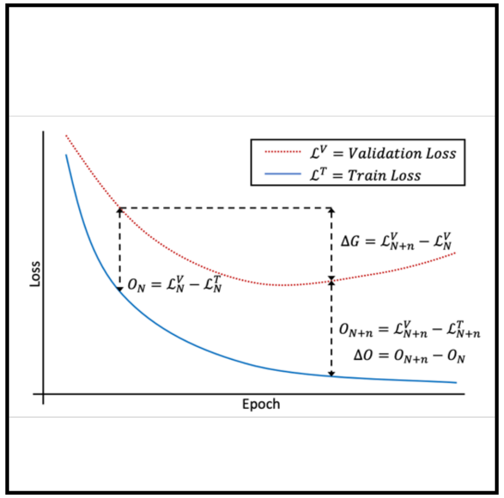
[Zadeh et al., 2017]

$$\mathcal{L}_{\text{sup}} = -\log p(y|\mathbf{x}_1, \mathbf{x}_2)$$

$$\mathcal{L}_{\text{CCA}} = -\text{corr}\left(g_1(\mathbf{z}_1), g_2(\mathbf{z}_2)\right)$$

$$\mathcal{L}_{\text{rec}} = \|g_1(\mathbf{z}_{\text{mm}}) - \mathbf{x}_1\|_2 + \|g_2(\mathbf{z}_{\text{mm}}) - \mathbf{x}_2\|_2$$

[Sankaran et al., 2021]



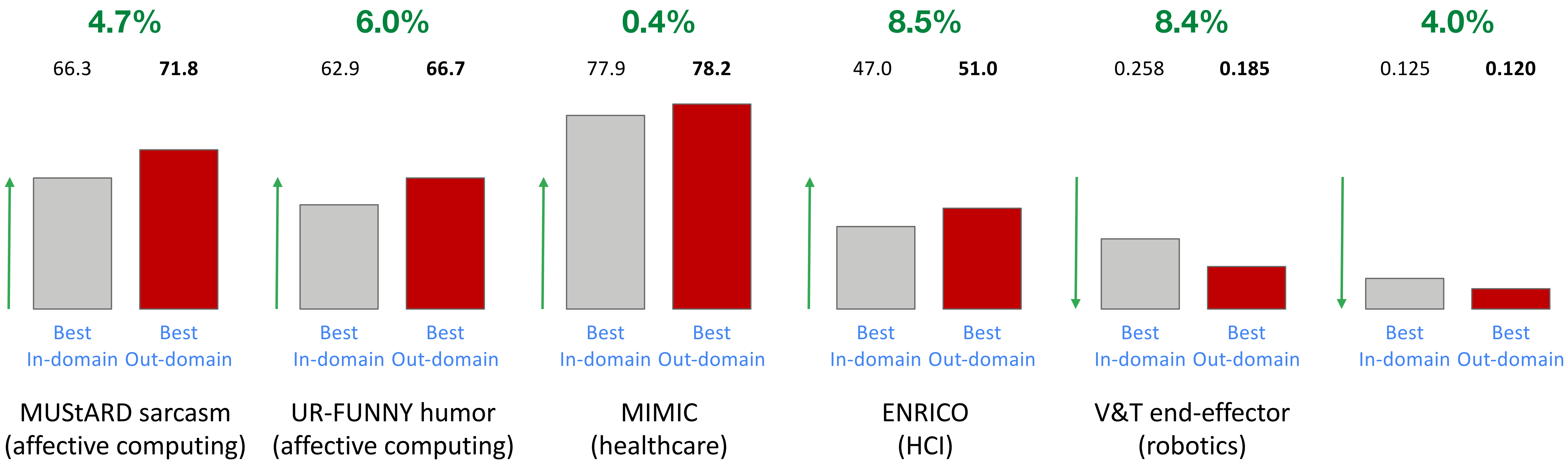
[Wang et al., 2020]

20 recent models, but composable to up to >200 combinations!

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

MultiBench: Generalization to Diverse Modalities

Benefits of standardization

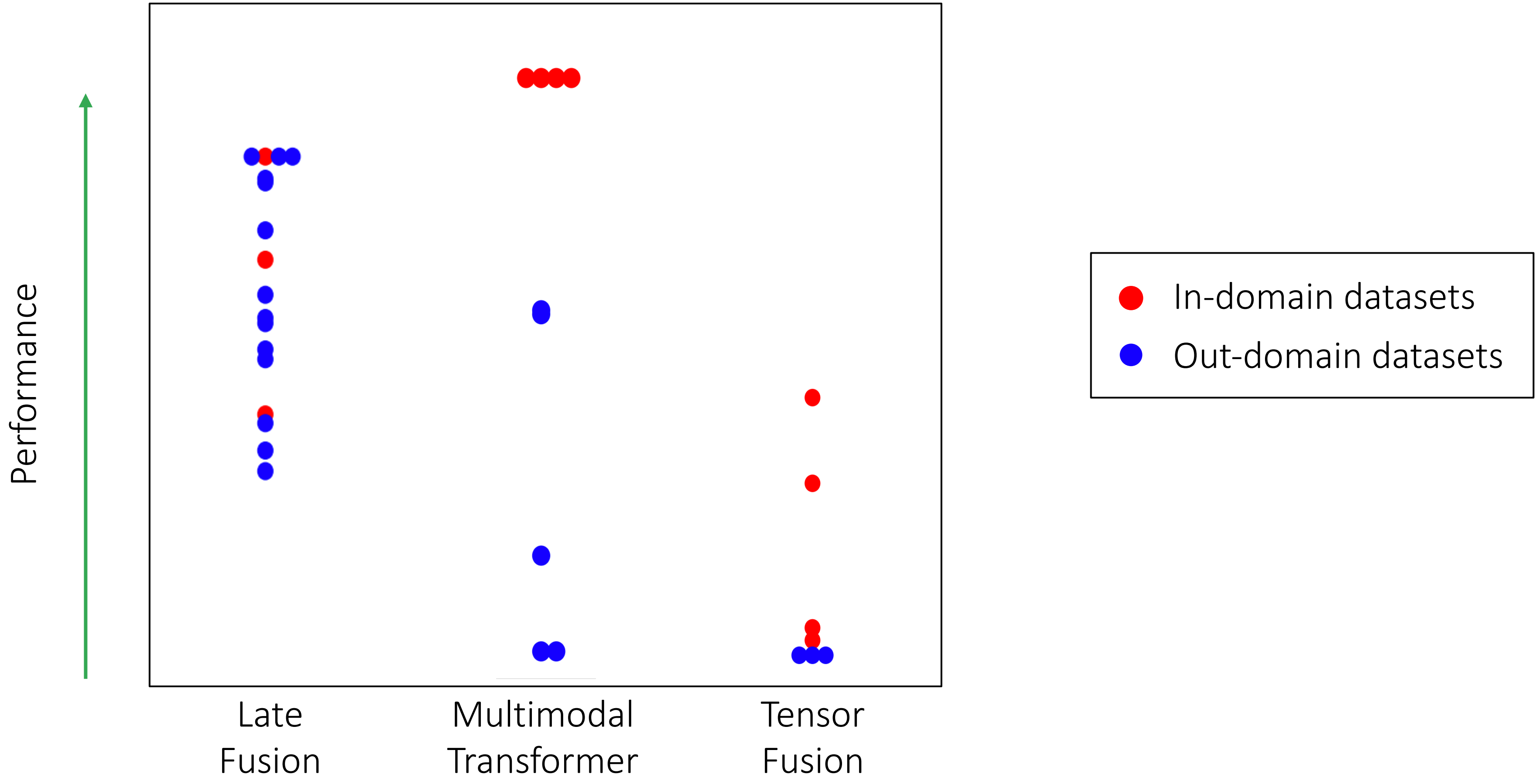


Simply applying methods in other areas improves performance on 9/15 datasets

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

MultiBench: Generalization to Diverse Modalities

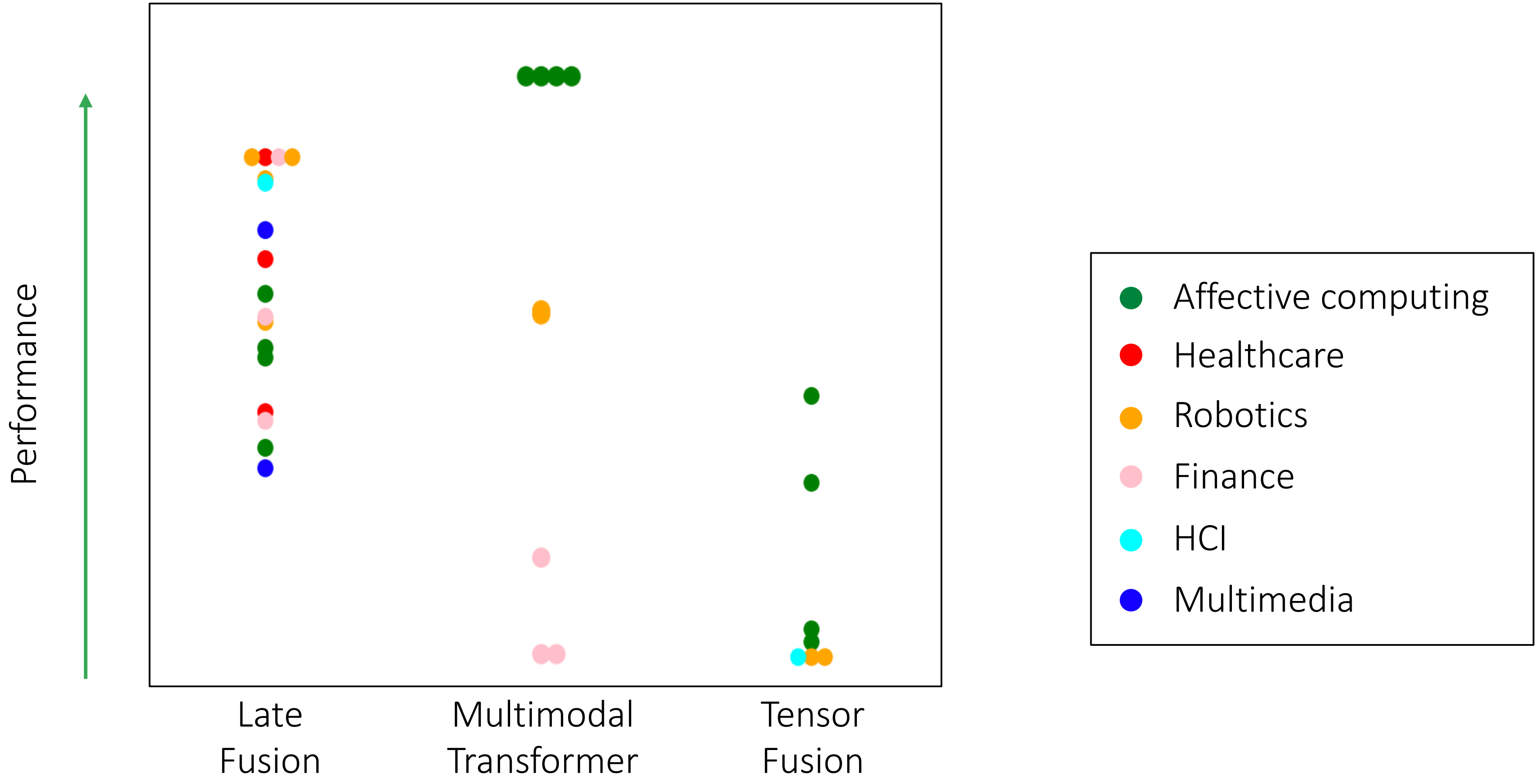
Methods struggle to perform outside of their own domain



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

MultiBench: Generalization to Diverse Modalities

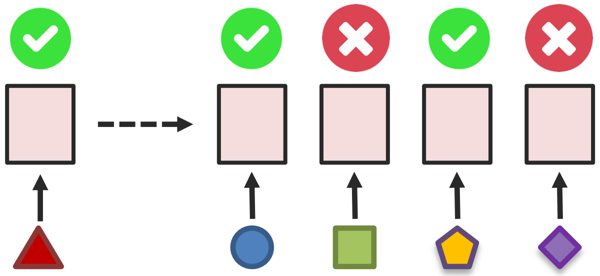
Generalization across modalities and tasks is difficult!



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

Quantifying Modality Heterogeneity

Information transfer, transfer learning perspective



Implicitly captures these:

1 Element representation

2 Element distribution

3 Structure

4 Information

5 Noise

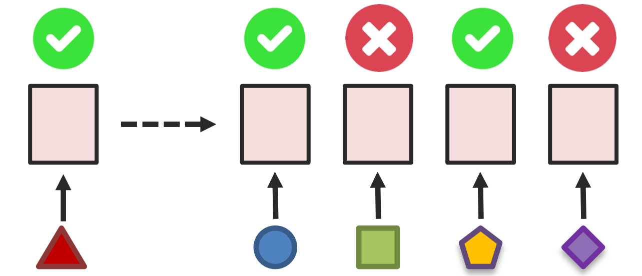
6 Relevance

[Liang et al., HighMMT: Quantifying Modality and Interaction Heterogeneity for High-Modality Representation Learning. arXiv 2022]

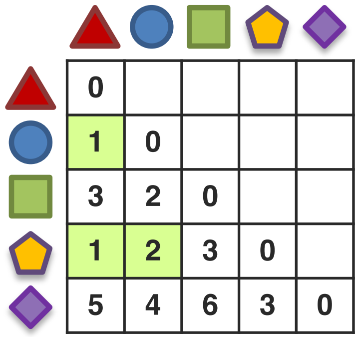
Quantifying Modality Heterogeneity

Information transfer, transfer learning perspective

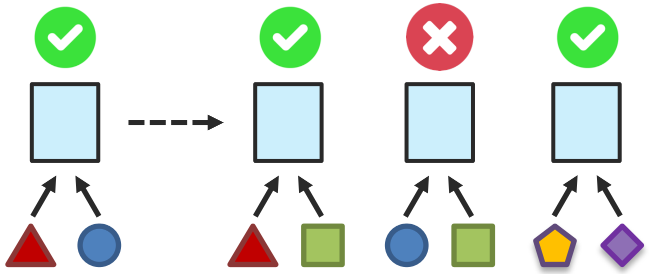
1a. Estimate modality heterogeneity via transfer



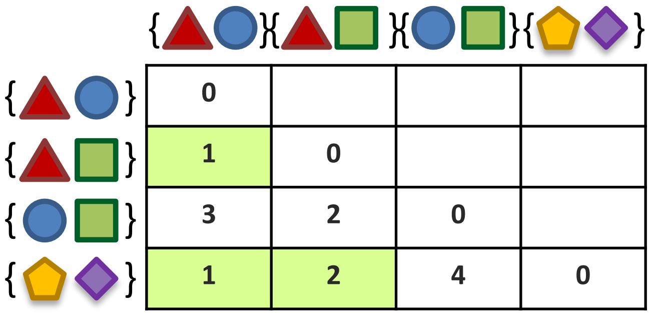
2a. Compute modality heterogeneity matrix



1b. Estimate interaction heterogeneity via transfer



2b. Compute interaction heterogeneity matrix



[Liang et al., HighMMT: Quantifying Modality and Interaction Heterogeneity for High-Modality Representation Learning. arXiv 2022]

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

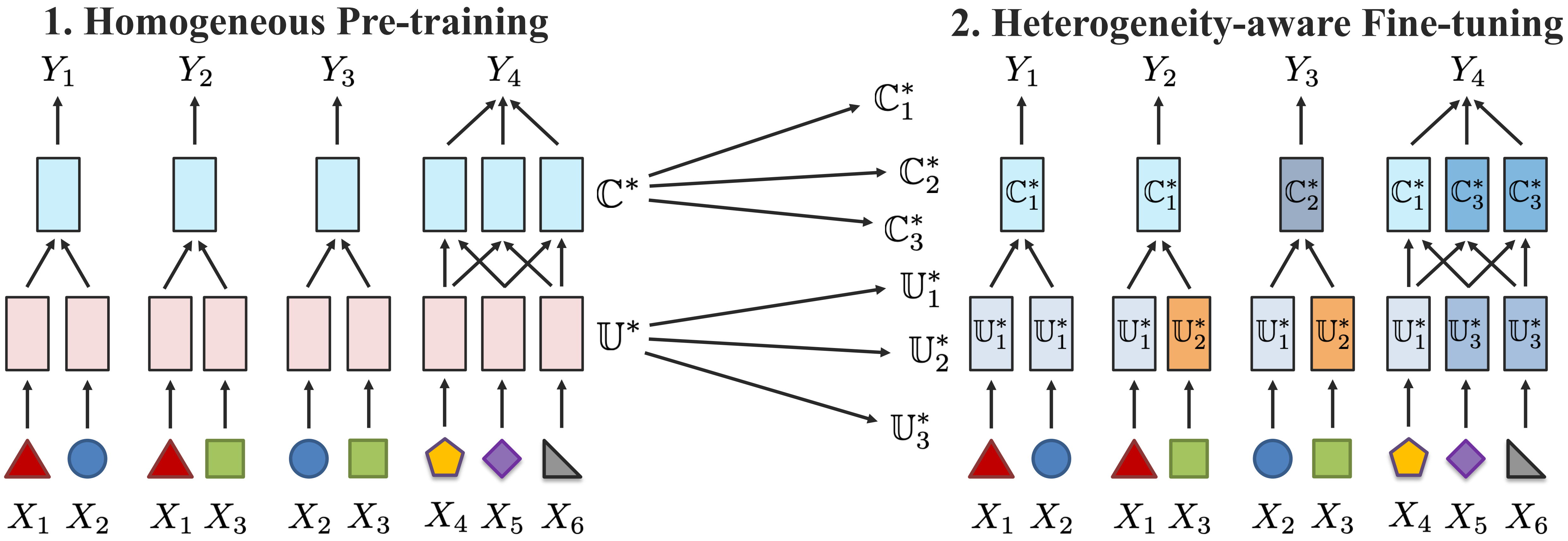
3. Determine parameter clustering

$$\mathbb{U}_1 = \{U_1, U_2, U_4\} \quad \mathbb{C}_1 = \{C_{12}, C_{13}, C_{45}\}$$
$$\mathbb{U}_2 = \{U_3\} \quad \mathbb{C}_2 = \{C_{23}\}$$
$$\mathbb{U}_3 = \{U_5\}$$

[Liang et al., HighMMT: Quantifying Modality and Interaction Heterogeneity for High-Modality Representation Learning. arXiv 2022]

Quantifying Modality Heterogeneity

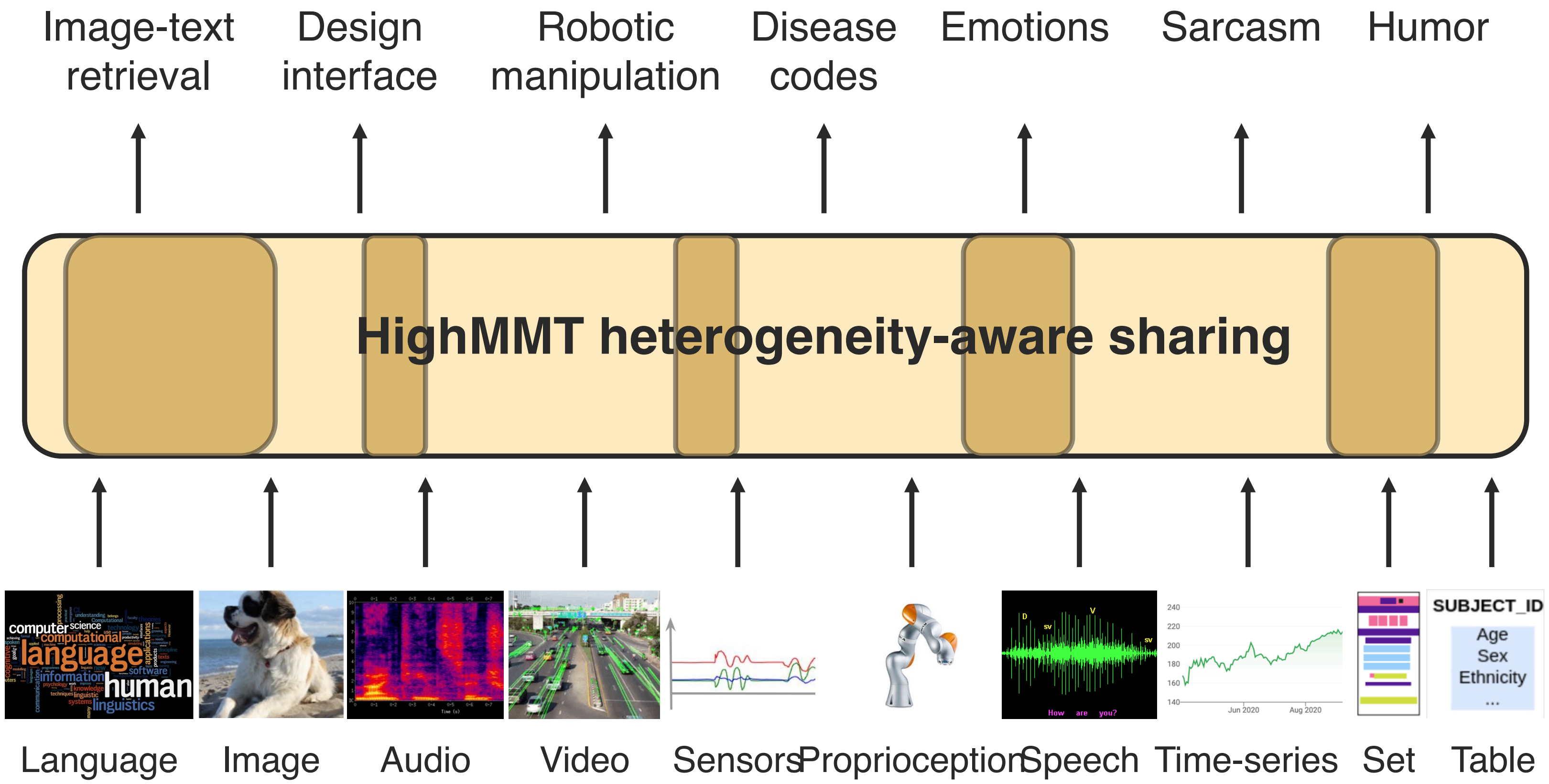
Information transfer, transfer learning perspective

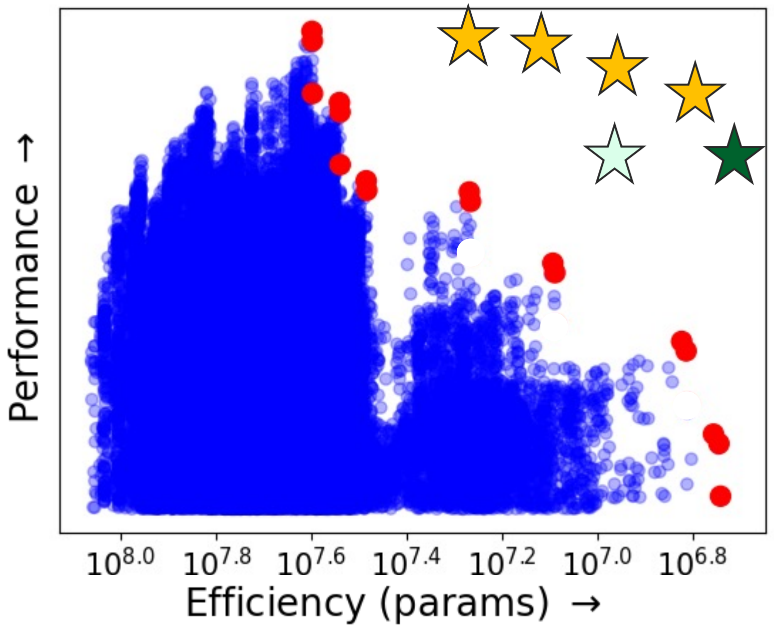


[Liang et al., HighMMT: Quantifying Modality and Interaction Heterogeneity for High-Modality Representation Learning. arXiv 2022]

Quantifying Modality Heterogeneity

HighMMT heterogeneity-aware: estimate heterogeneity to determine parameter sharing





- All model combinations (>10,000)
- Pareto front
- HighMMT single-task
- HighMMT multitask
- **HighMMT heterogeneity-aware**

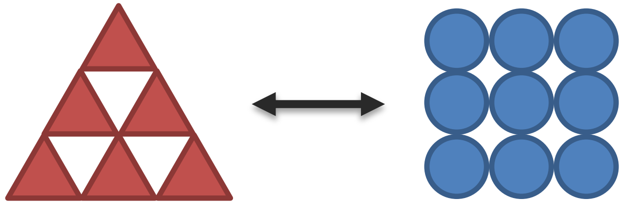
[Liang et al., HighMMT: Towards Modality and Task Generalization for High-Modality Representation Learning. arXiv 2022]

Challenges: Quantifying Heterogeneity

Open challenges

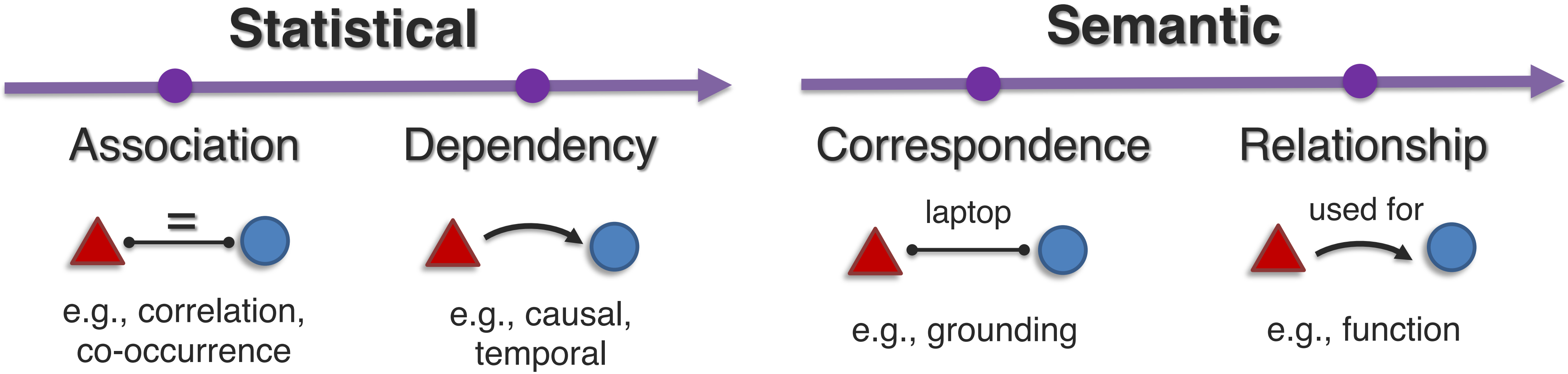
Open challenges:

- Quantifying and modeling: chicken and egg problem.
- Bottom-up vs top-down, data-driven vs hypothesis-driven.
- Noisy and missing modalities.
- New and understudied modalities.
- Large number of modalities.
- Cases where its unclear which modalities are useful.



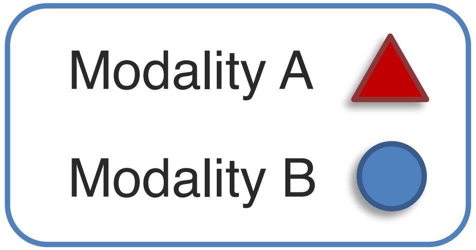
Sub-Challenge 6b: Cross-modal Connections

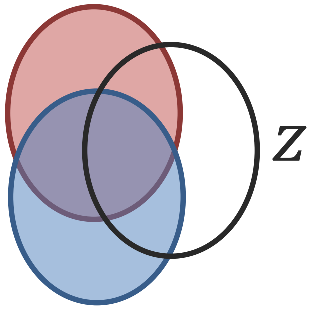
Connected: Shared information that relates modalities

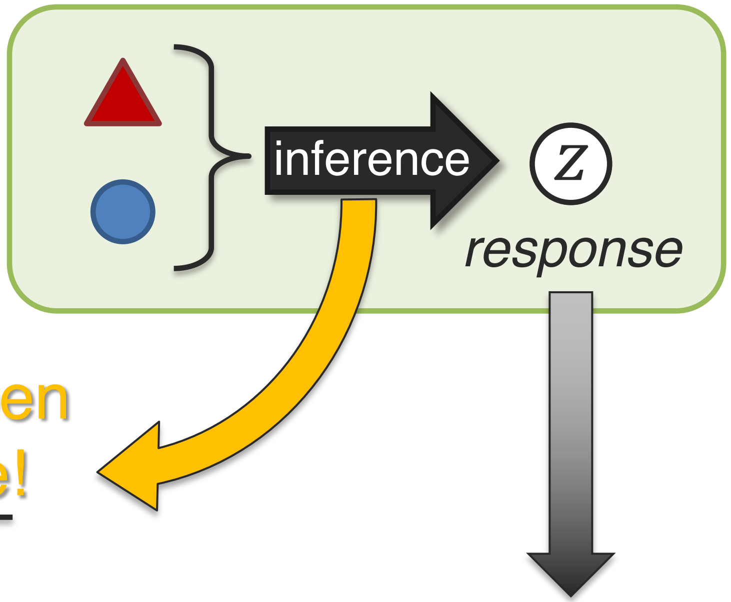


Sub-Challenge 6b: Cross-modal Interactions

Interacting: process affecting each modality, creating new response



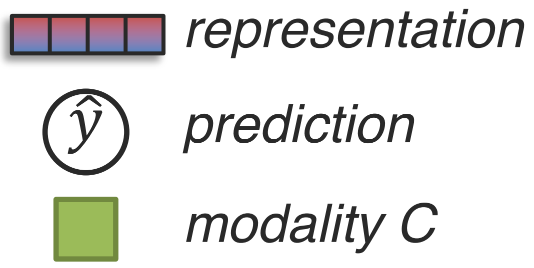




Interactions happen during inference!

"Inference" examples:

- Representation fusion
- Prediction task
- Modality translation

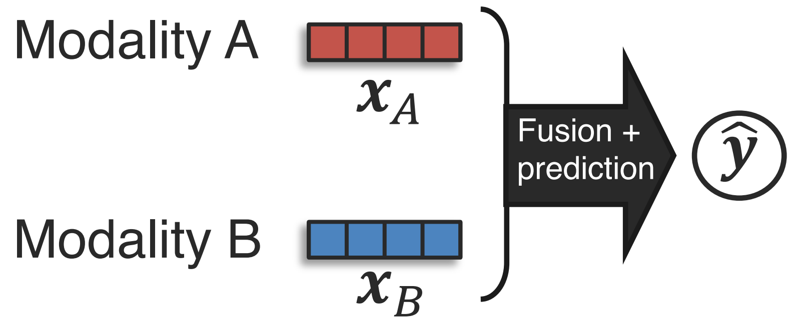


Quantifying Cross-modal Interactions

Identifying overall presence of cross-modal interactions

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

f exhibits interactions between 2 features $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$ iff f cannot be decomposed into a sum of unimodal subfunctions f_A, f_B such that $f(\boldsymbol{x}_A, \boldsymbol{x}_B) = f_A(\boldsymbol{x}_A) + f_B(\boldsymbol{x}_B)$.



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

If the additive projection $\tilde{f}(\boldsymbol{x}_A, \boldsymbol{x}_B)$ is equal to nonlinear fusion $f(\boldsymbol{x}_A, \boldsymbol{x}_B)$ then the non-additive interactions are not modeled.

μ measures **overall quantity** of cross-modal interactions on a trained model + dataset.

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

Quantifying Cross-modal Interactions

Identifying individual cross-modal interactions

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

f exhibits interactions between 2 features $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$ iff f cannot be decomposed into a sum of unimodal subfunctions f_A, f_B such that $f(\boldsymbol{x}_A, \boldsymbol{x}_B) = f_A(\boldsymbol{x}_A) + f_B(\boldsymbol{x}_B)$.

f exhibits interactions between 2 features $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$ iff $\frac{\partial f^2}{\partial \boldsymbol{x}_A \partial \boldsymbol{x}_B} > 0$.

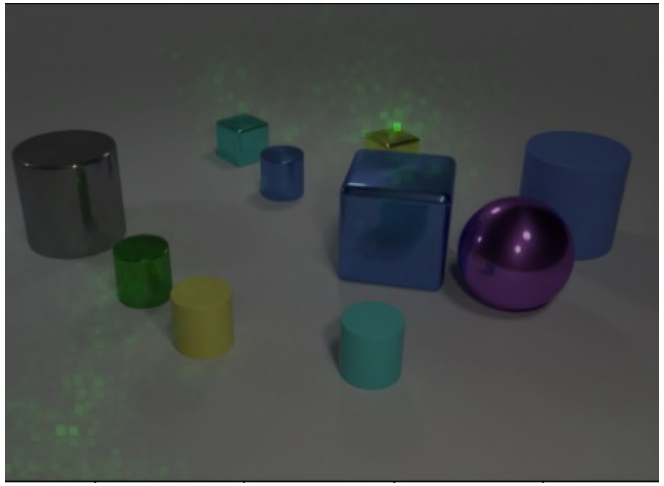
Natural second-order extension of gradient-based approaches!

[Liang et al., MultiViz: An Analysis Benchmark for Visualizing and Understanding Multimodal Models. arXiv 2022]

Quantifying Cross-modal Interactions

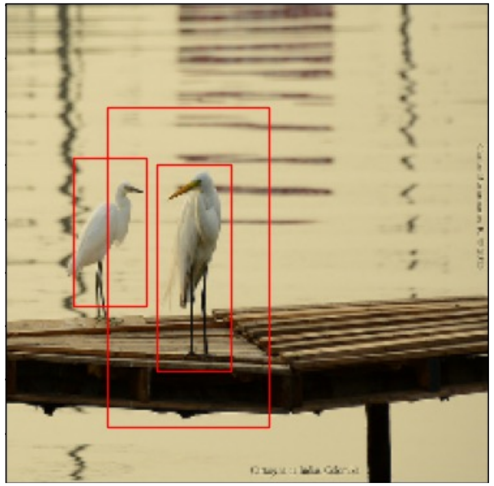
Identifying individual cross-modal interactions

CLEVR



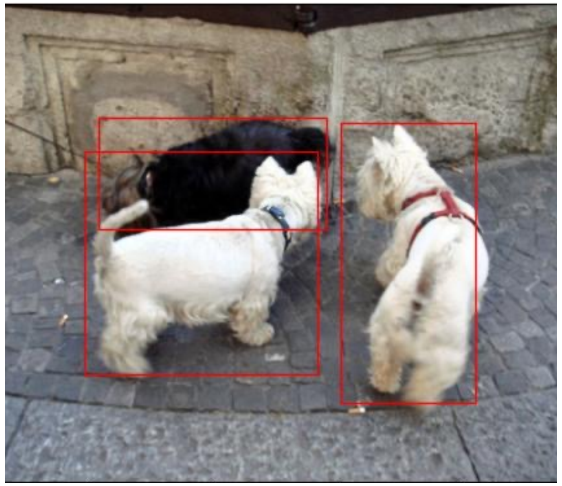
*The other small shiny thing that is the same shape as the **tiny yellow shiny object** is what color?*

VQA 2.0



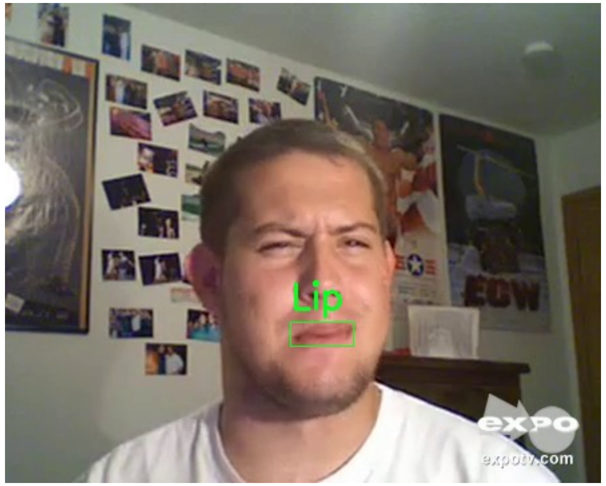
*How many **birds**?*

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**…*

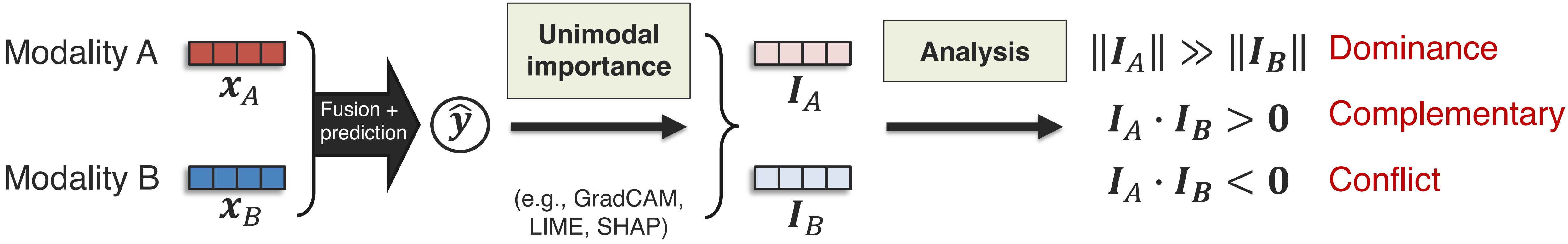
Correspondence

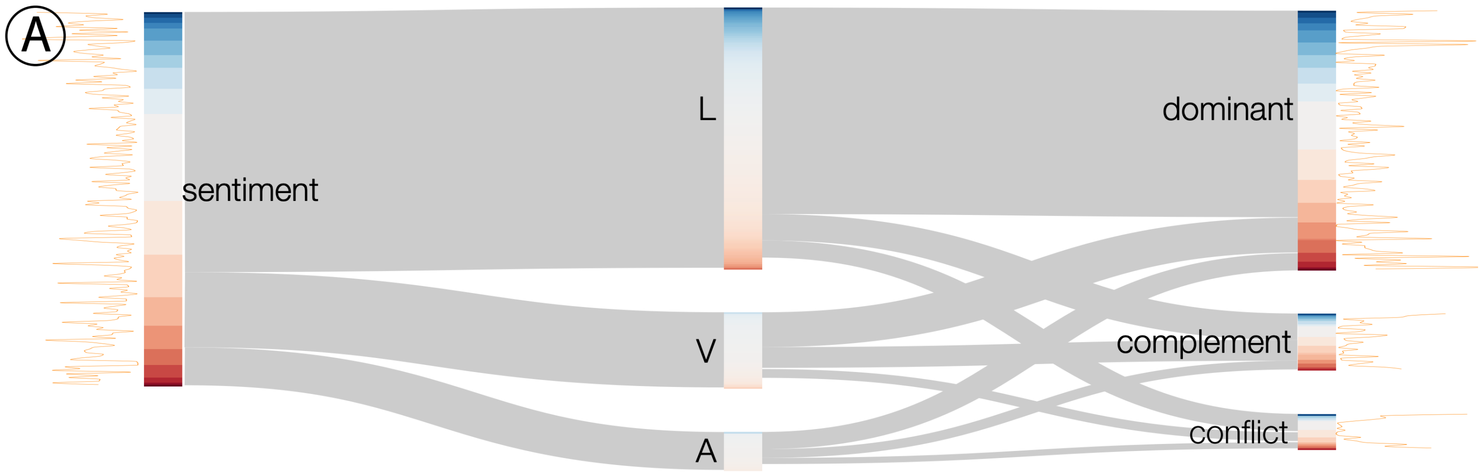
Relationships

[Liang et al., MultiViz: An Analysis Benchmark for Visualizing and Understanding Multimodal Models. arXiv 2022]

Quantifying Cross-modal Interactions

Classification of cross-modal interactions





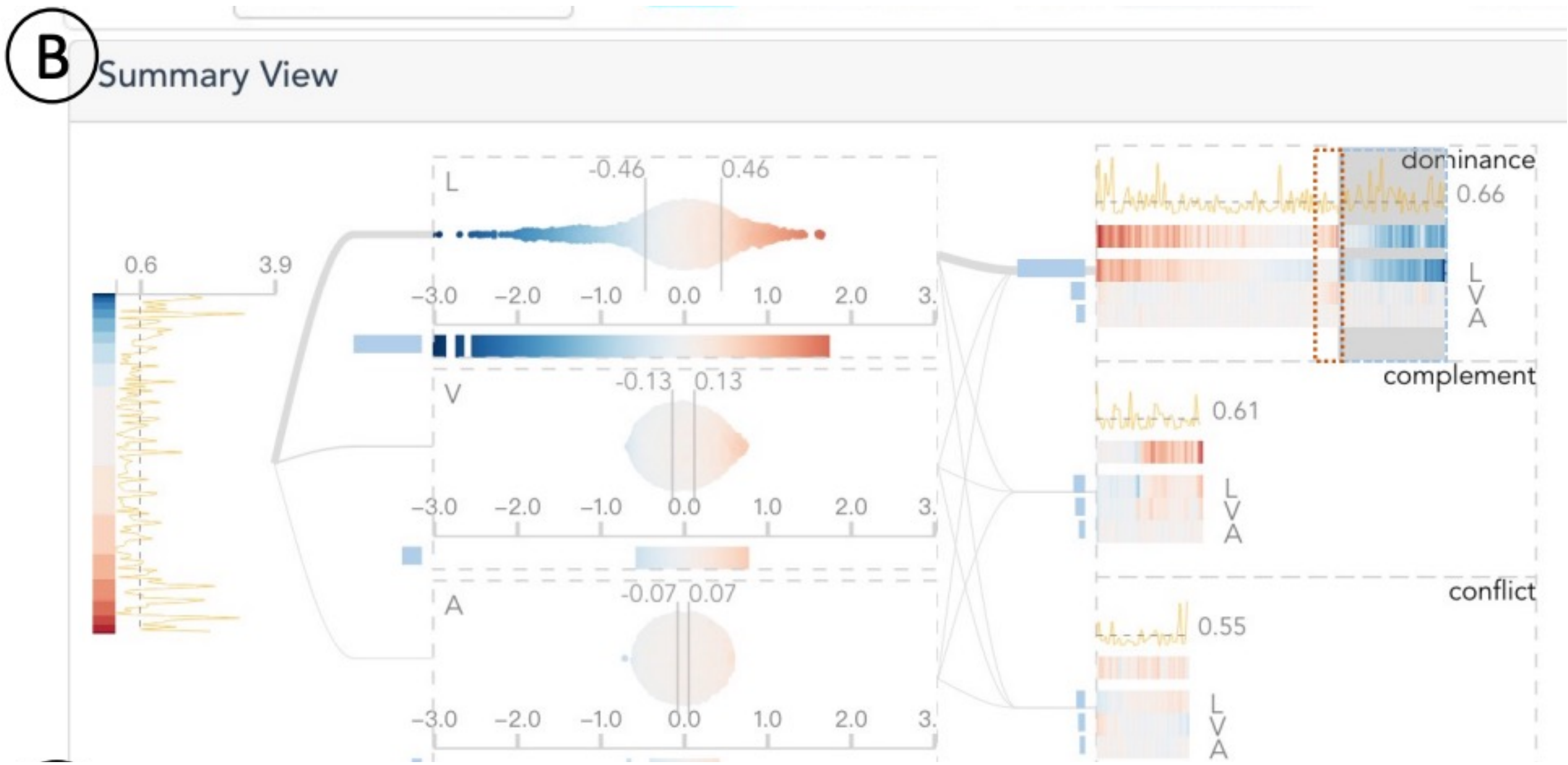
Language is often **dominant** in multimodal sentiment analysis

[Wang et al., M2Lens: Visualizing and Explaining Multimodal Models for Sentiment Analysis. IEEE Trans Visualization and Computer Graphics 2021]

Quantifying Cross-modal Interactions

Visualization website

See interactive website: https://andy-xingbowang.com/m2lens/



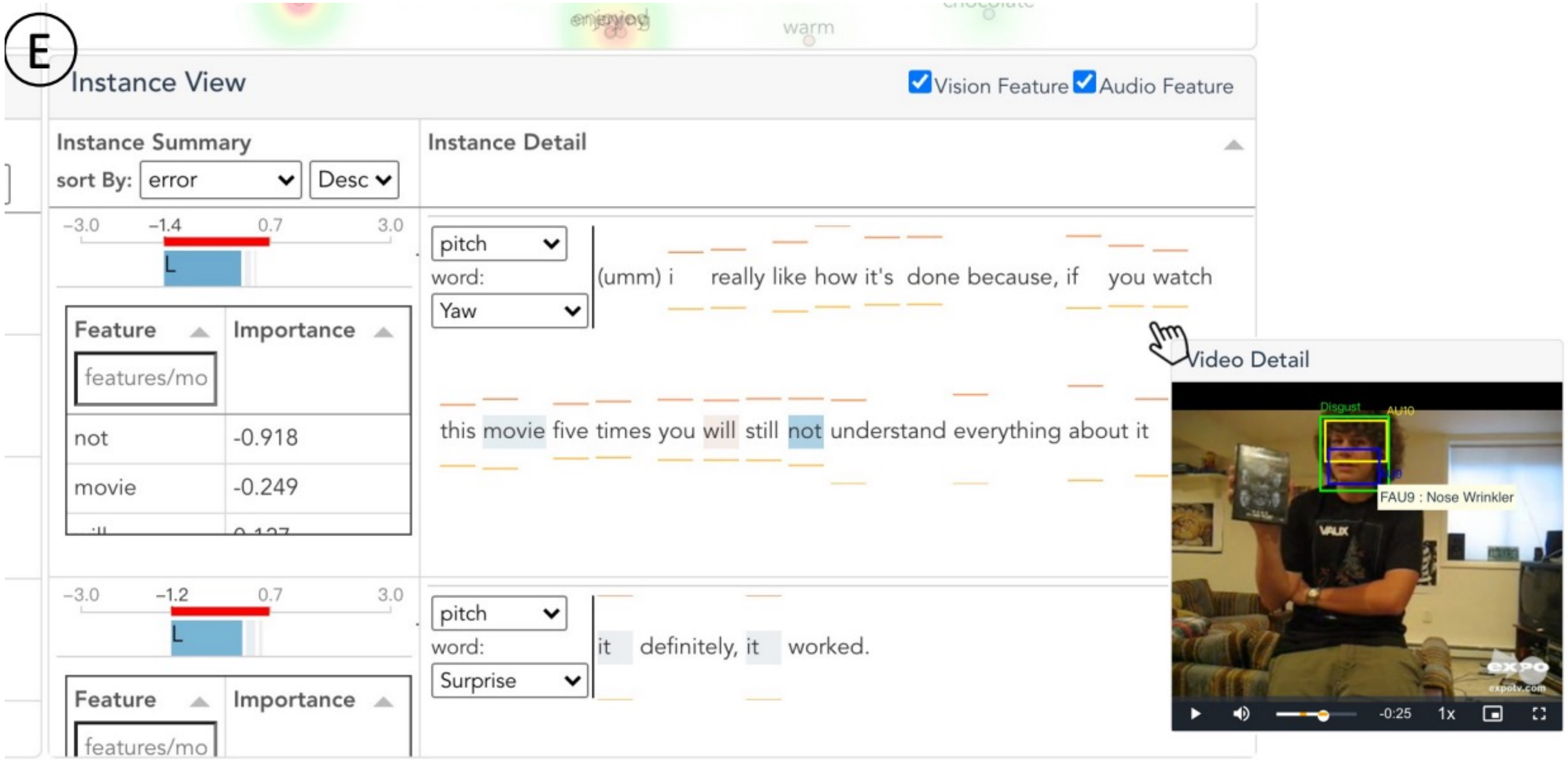
Summary of cross-modal interactions across entire dataset.

[Wang et al., M2Lens: Visualizing and Explaining Multimodal Models for Sentiment Analysis. IEEE Trans Visualization and Computer Graphics 2021]

Quantifying Cross-modal Interactions

Visualization website

See interactive website: https://andy-xingbowang.com/m2lens/

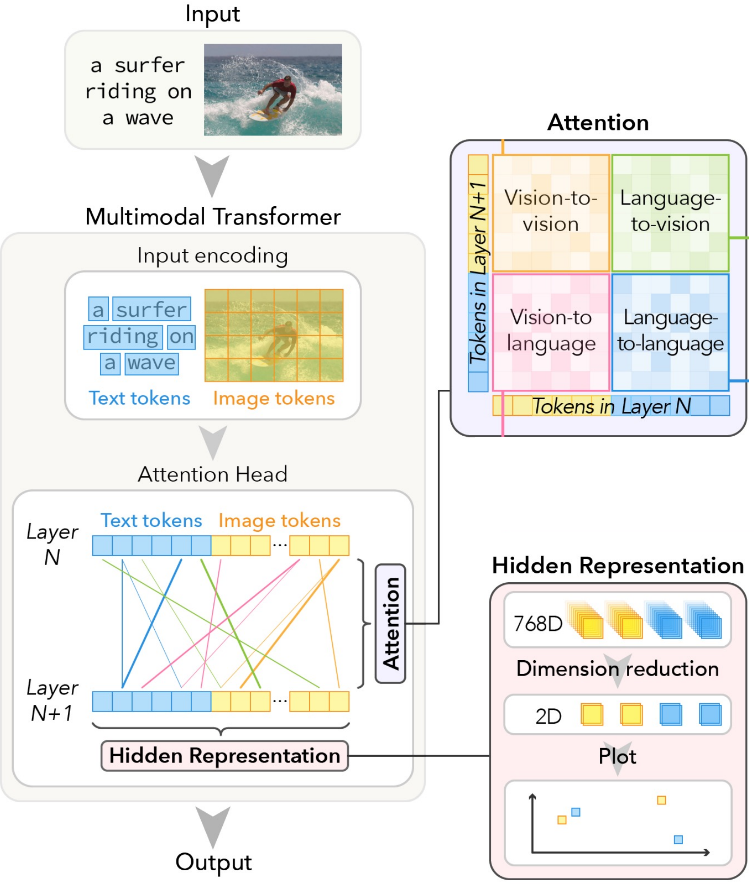


Summary of cross-modal interactions in a single instance.

[Wang et al., M2Lens: Visualizing and Explaining Multimodal Models for Sentiment Analysis. IEEE Trans Visualization and Computer Graphics 2021]

Quantifying Cross-modal Interactions

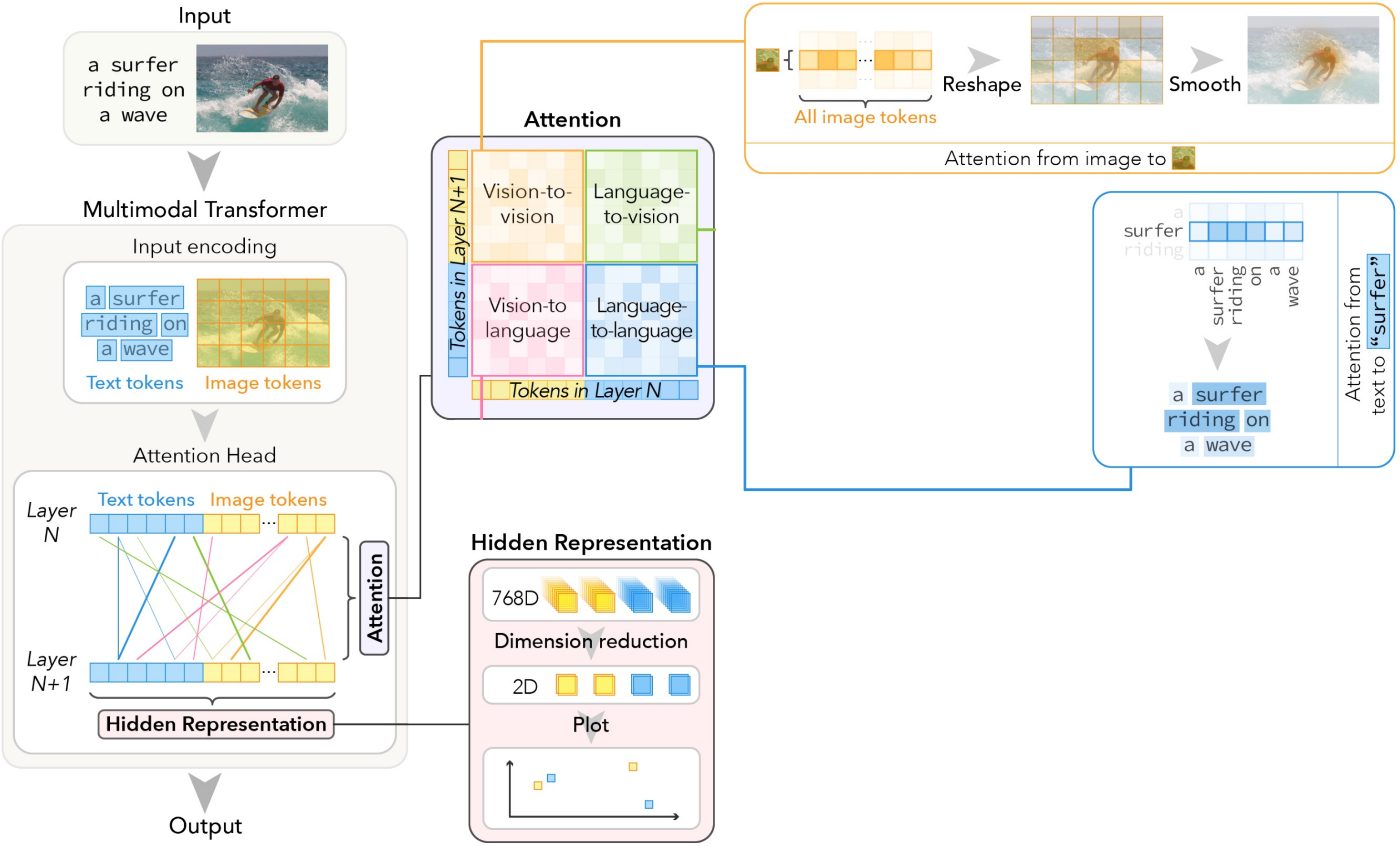
Visualizing multimodal transformers See interactive website: https://github.com/IntelLabs/VL-InterpreT



[Aflalo et al., VL-InterpreT: An Interactive Visualization Tool for Interpreting Vision-Language Transformers. CVPR 2022]

Quantifying Cross-modal Interactions

Visualizing multimodal transformers See interactive website: https://github.com/IntelLabs/VL-InterpreT



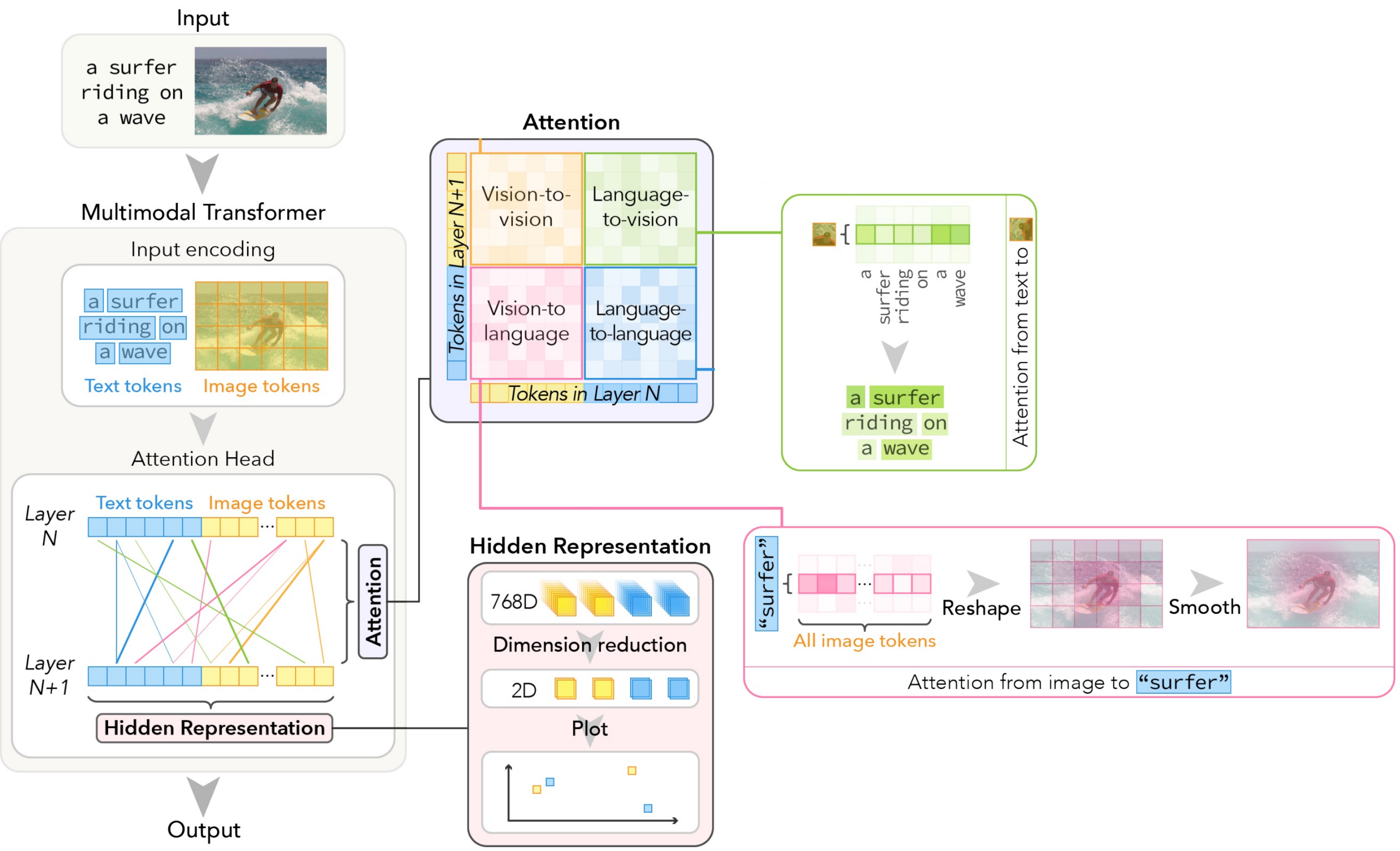
Unimodal image importance

Unimodal text importance

[Aflalo et al., VL-InterpreT: An Interactive Visualization Tool for Interpreting Vision-Language Transformers. CVPR 2022]

Quantifying Cross-modal Interactions

Visualizing multimodal transformers See interactive website: https://github.com/IntelLabs/VL-InterpreT



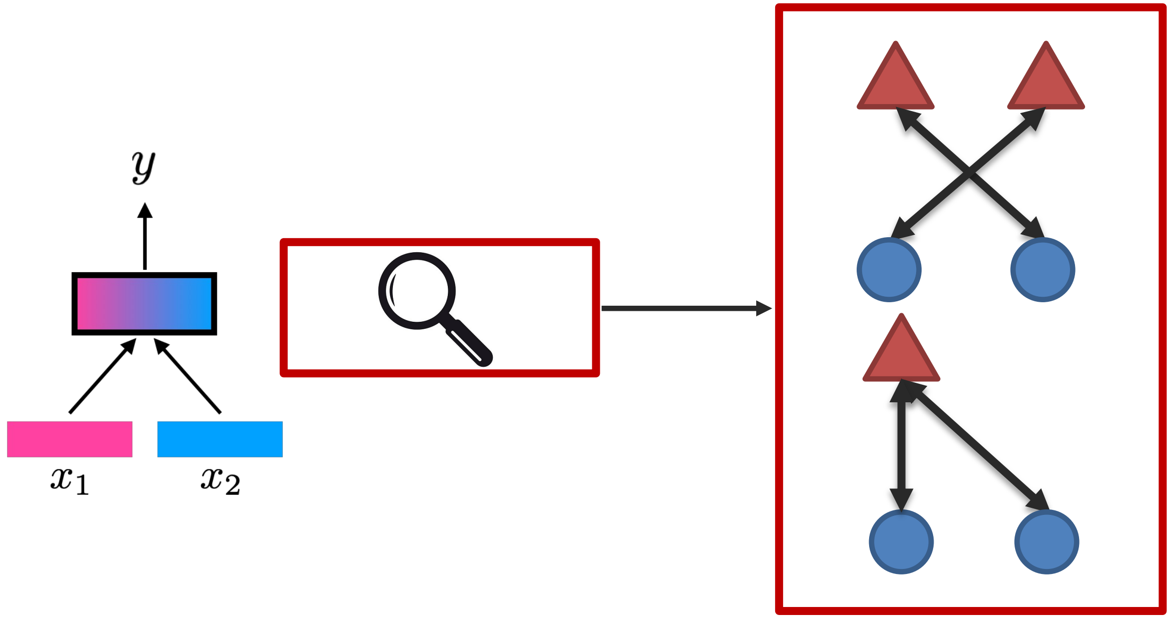
Correspondence and complementary interactions

[Aflalo et al., VL-InterpreT: An Interactive Visualization Tool for Interpreting Vision-Language Transformers. CVPR 2022]

Evaluating Interpretability

How can we evaluate the success of interpreting cross-modal interactions?

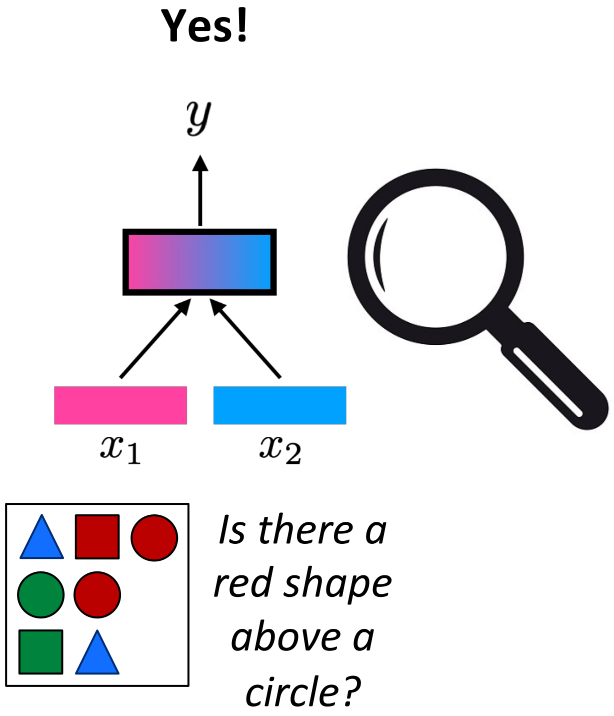
Problem: real-world datasets and models do not have cross-modal interactions annotated!



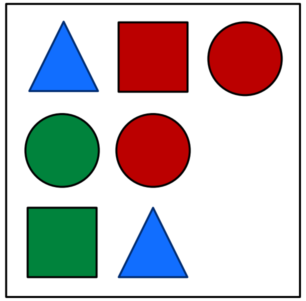
[Liang et al., MultiViz: A Framework for Visualizing and Understanding Multimodal Models. arXiv 2022]

Evaluating Interpretability: A Multi-stage Approach

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



1. Unimodal importance



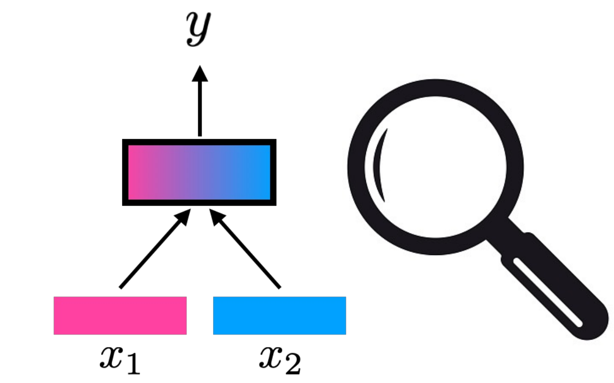
Is there a ***red shape above*** *a* ***circle****?*

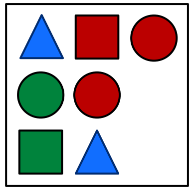
[Liang et al., MultiViz: An Analysis Benchmark for Visualizing and Understanding Multimodal Models. arXiv 2022]

Evaluating Interpretability: A Multi-stage Approach

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

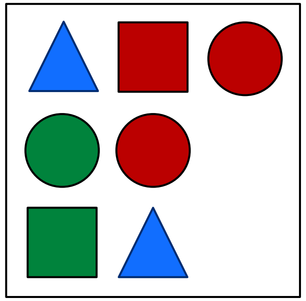
Yes!





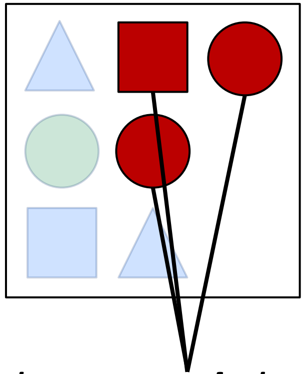
Is there a red shape above a circle?

1. Unimodal importance

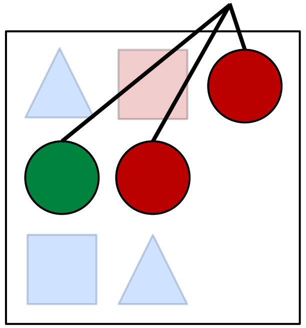


*Is there a **red shape** **above** a **circle**?*

2. Cross-modal interactions



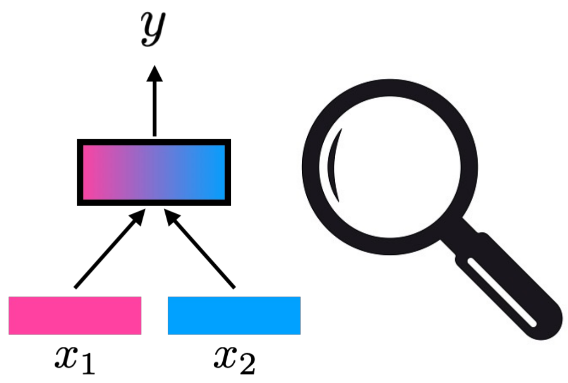
*Is there a **red** shape above a **circle**?*

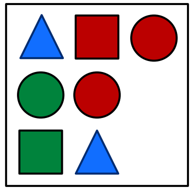


Evaluating Interpretability: A Multi-stage Approach

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

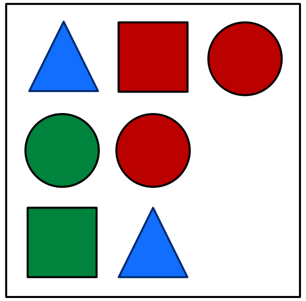
Yes!





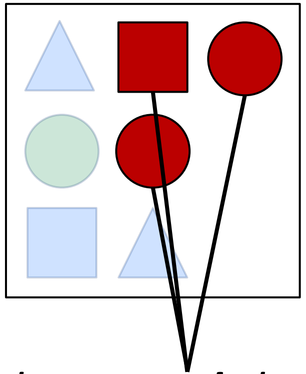
Is there a red shape above a circle?

1. Unimodal importance

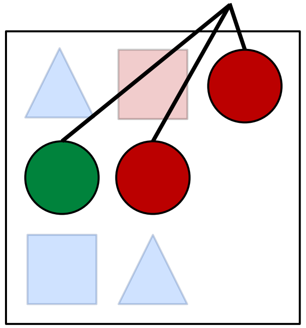


*Is there a **red shape above** a **circle**?*

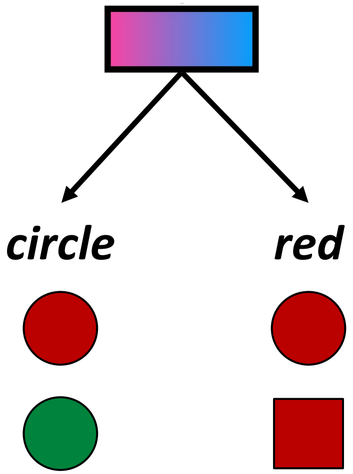
2. Cross-modal interactions



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

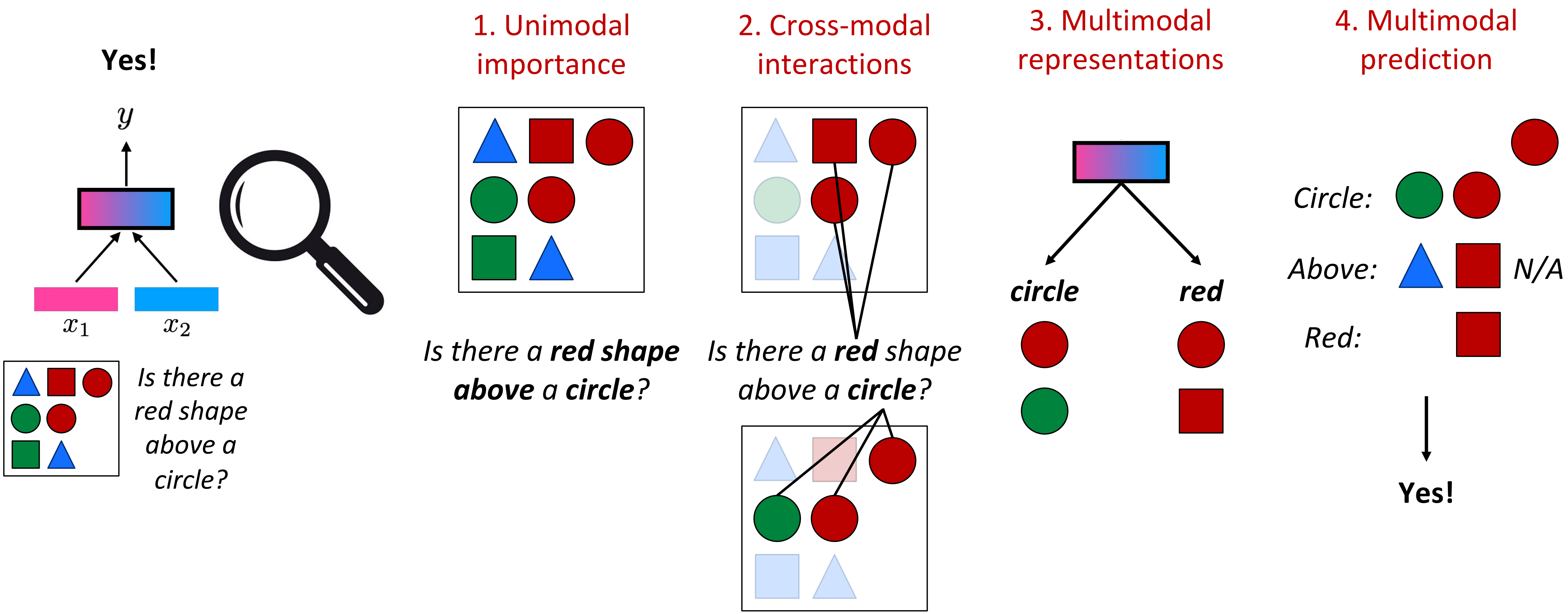


3. Multimodal representations



Evaluating Interpretability: A Multi-stage Approach

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



Evaluating Interpretability: A Multi-stage Approach

Identifying individual cross-modal connections

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

f exhibits interactions between 2 features $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$ iff f cannot be decomposed into a sum of unimodal subfunctions f_A, f_B such that $f(\boldsymbol{x}_A, \boldsymbol{x}_B) = f_A(\boldsymbol{x}_A) + f_B(\boldsymbol{x}_B)$.

f exhibits interactions between 2 features $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$ iff $\frac{\partial f^2}{\partial \boldsymbol{x}_A \partial \boldsymbol{x}_B} > 0$.

Natural second-order extension of gradient-based approaches!

[Liang et al., MultiViz: An Analysis Benchmark for Visualizing and Understanding Multimodal Models. arXiv 2022]

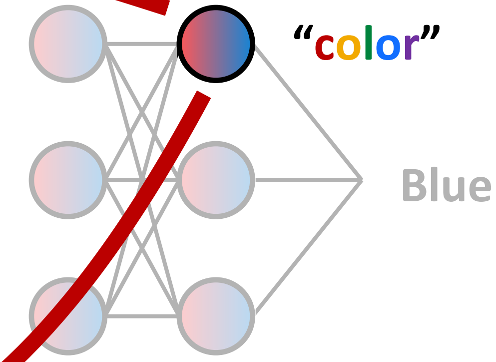
Evaluating Interpretability: A Multi-stage Approach

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?

[Liang et al., MultiViz: An Analysis Benchmark for Visualizing and Understanding Multimodal Models. arXiv 2022]

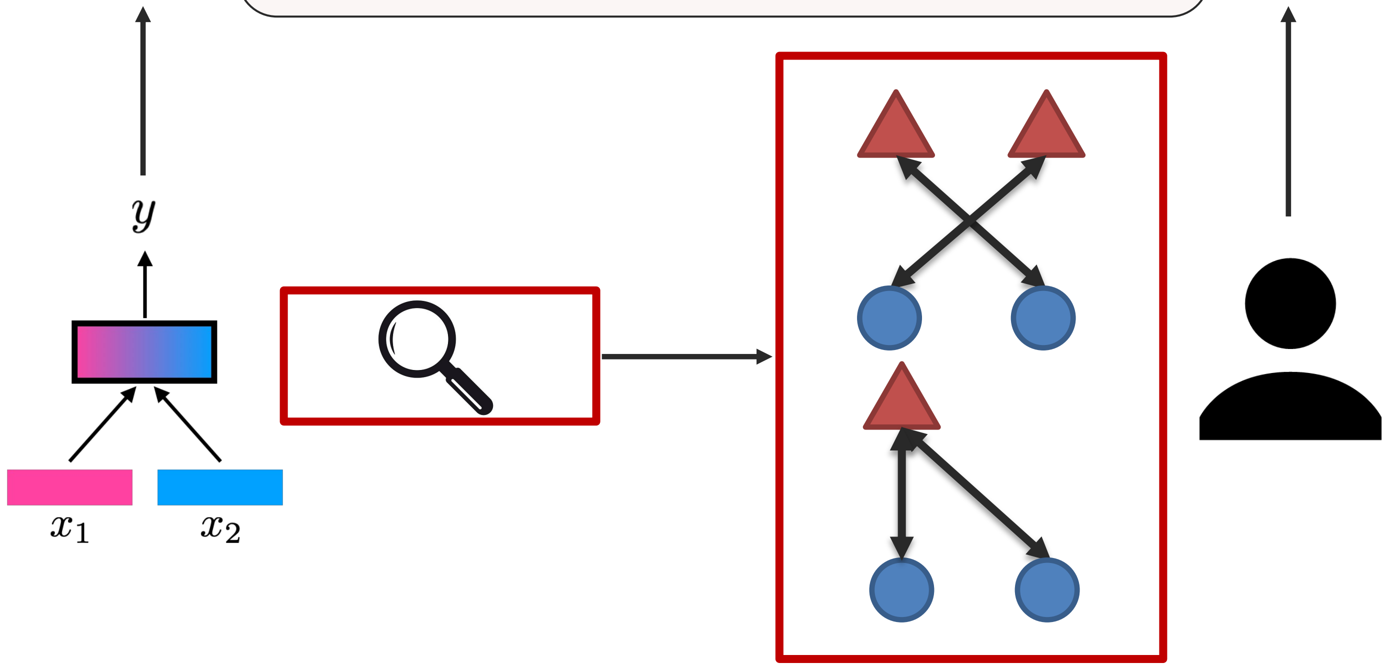
Evaluating Interpretability: MultiViz

Model simulation

1. Model simulation

Can humans reproduce model predictions with high accuracy and agreement?

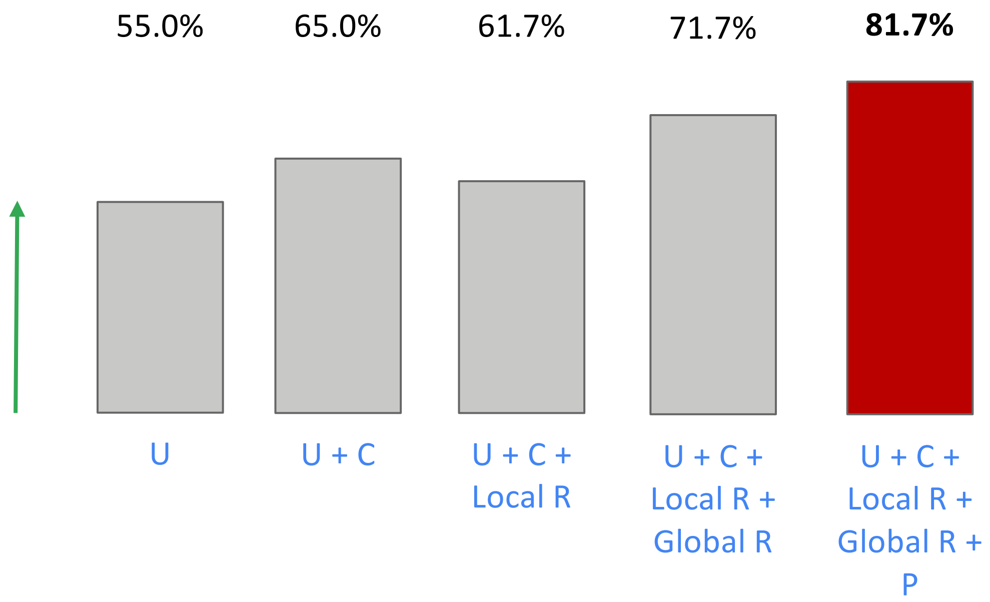
“Yes” “Yes”



[Liang et al., MultiViz: A Framework for Visualizing and Understanding Multimodal Models. arXiv 2022]

Evaluating Interpretability: MultiViz

Model simulation



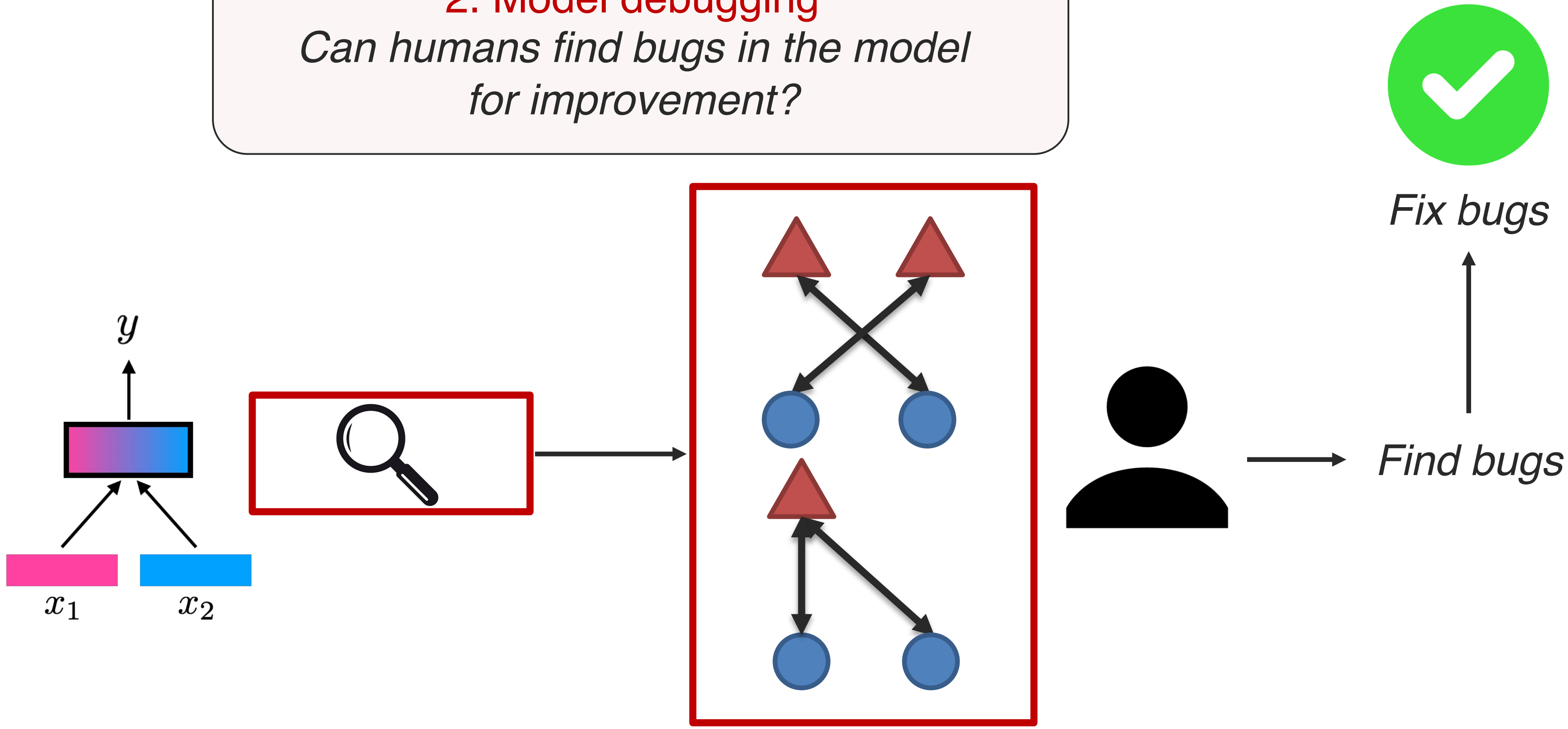
MultiViz stages leads to higher accuracy and agreement
Blind test + reasonable baselines + quantifiable outcome

[Liang et al., MultiViz: A Framework for Visualizing and Understanding Multimodal Models. arXiv 2022]

Evaluating Interpretability: MultiViz

Model error analysis and debugging

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



[Liang et al., MultiViz: A Framework for Visualizing and Understanding Multimodal Models. arXiv 2022]

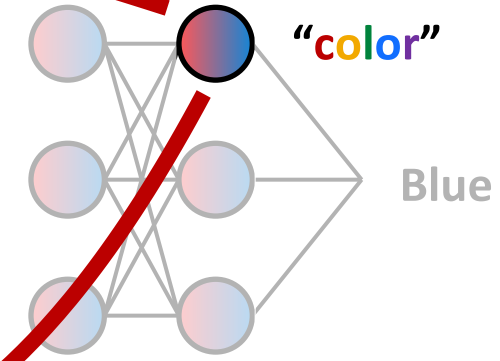
Evaluating Interpretability: MultiViz

Model error analysis and debugging

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?

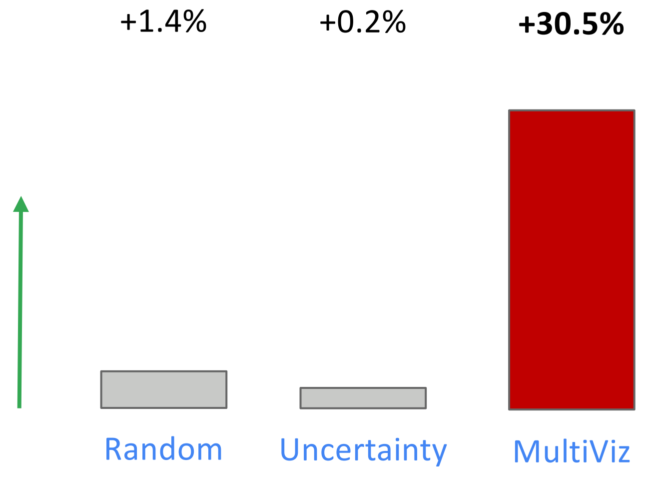
“Models pick up cross-modal interactions but fail in identifying color!”

[Liang et al., MultiViz: A Framework for Visualizing and Understanding Multimodal Models. arXiv 2022]

Evaluating Interpretability: MultiViz

Model error analysis and debugging

"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

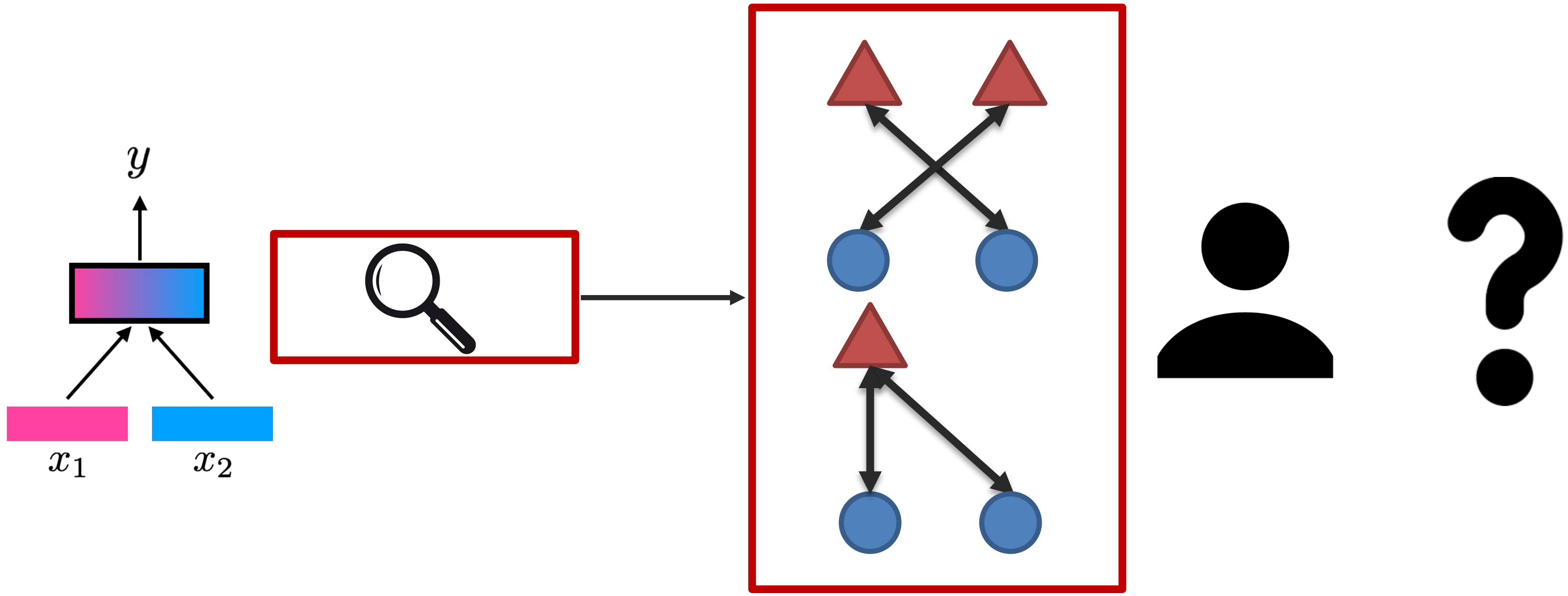
[Liang et al., MultiViz: A Framework for Visualizing and Understanding Multimodal Models. arXiv 2022]

Challenges: Quantifying Multimodal Interactions

Open challenges

Open challenges:

- Faithfulness: do explanations accurately reflect model's internal mechanics?
- Usefulness: unclear if explanations help humans
- Disagreement: different interpretation methods may generate different explanations
- Evaluate: how to best evaluate interpretation methods

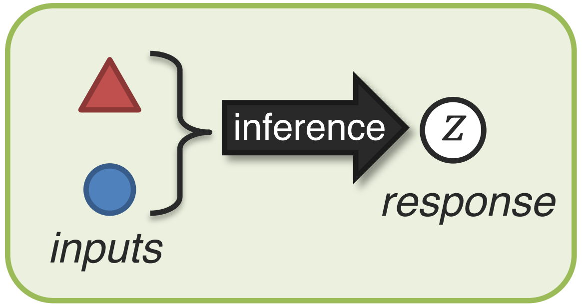


[Chandrasekaran et al., Do explanations make VQA models more predictable to a human? EMNLP 2018]

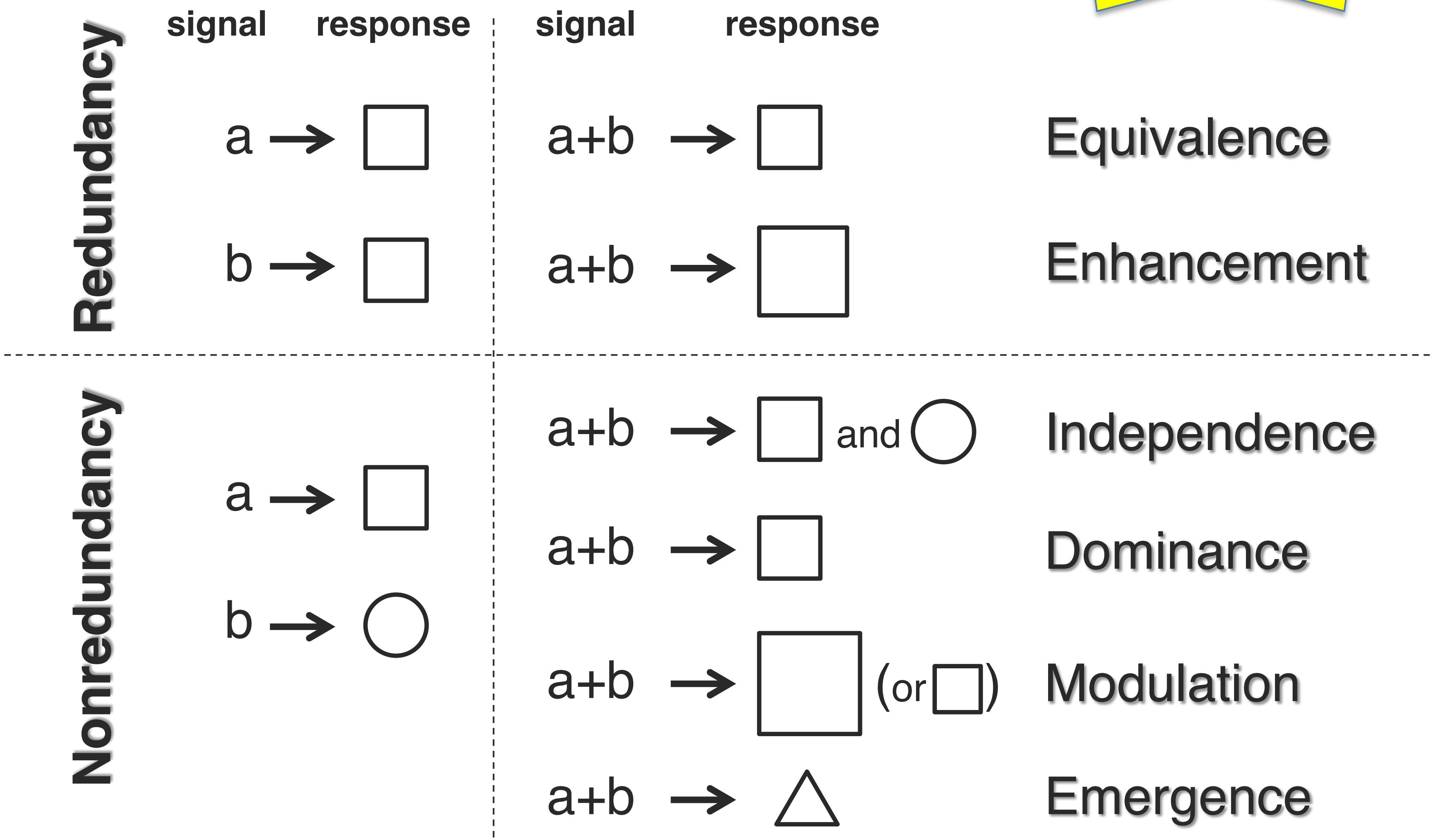
[Krishna et al., The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective. arXiv 2022]

Challenges: Quantifying Multimodal Interactions

Open challenges



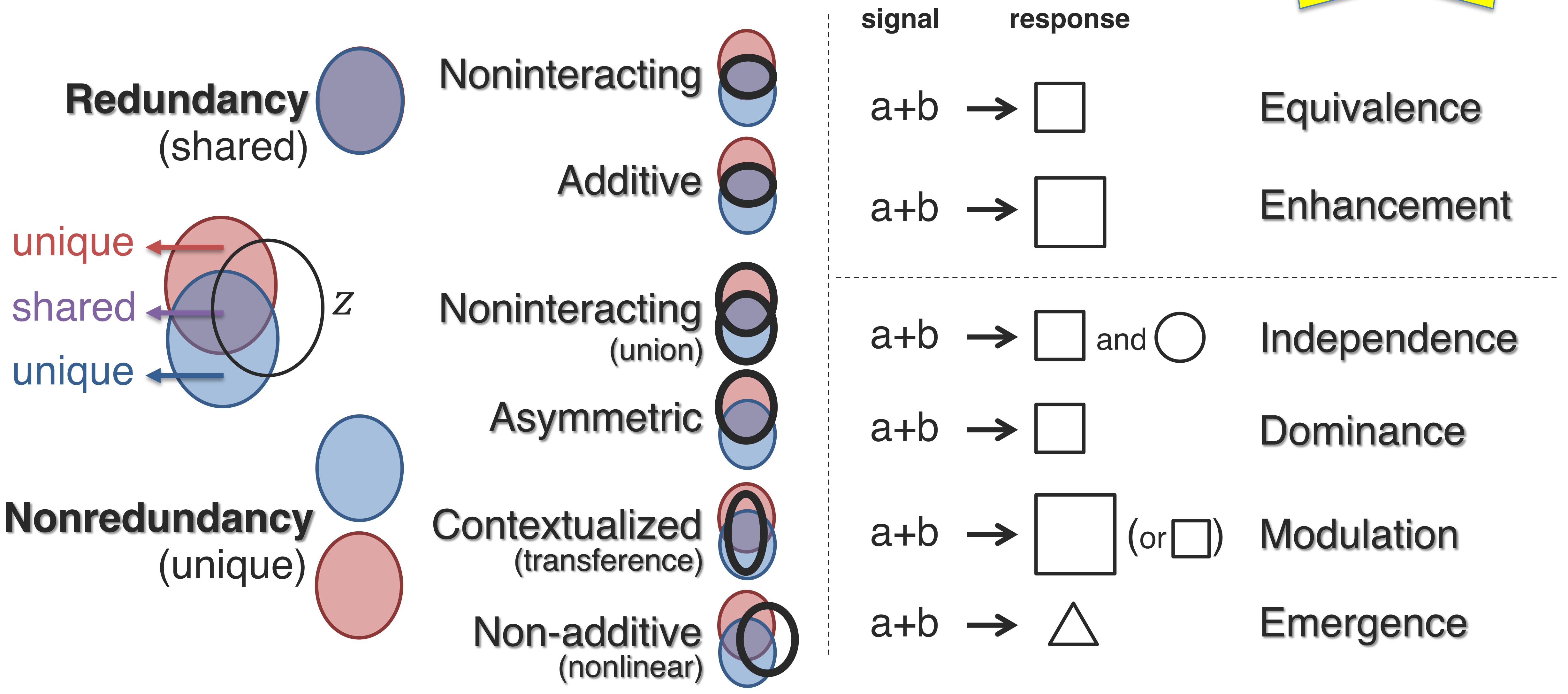
Multimodal Communication



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

Challenges: Quantifying Multimodal Interactions

Open challenges



Challenges: Quantifying Multimodal Interactions

Recall error analysis!

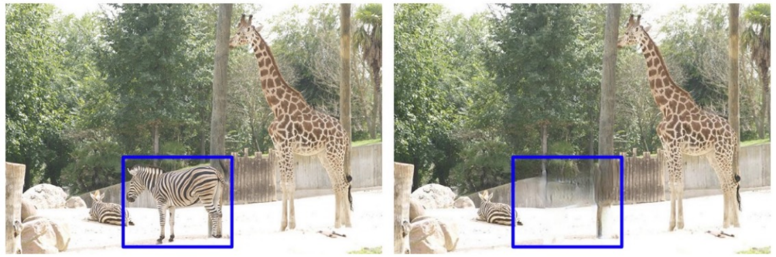
Causal, logical interactions beyond additive/multiplicative

Covariant VQA

Target object in question

Q: How many zebras are there in the picture?

A: 2 *zebra removed* A: 1



Baselines: 2 2

i.e., treatment variable

zebras → **prediction**

Interventional conditional: $p(y|do(zebras = 1))$

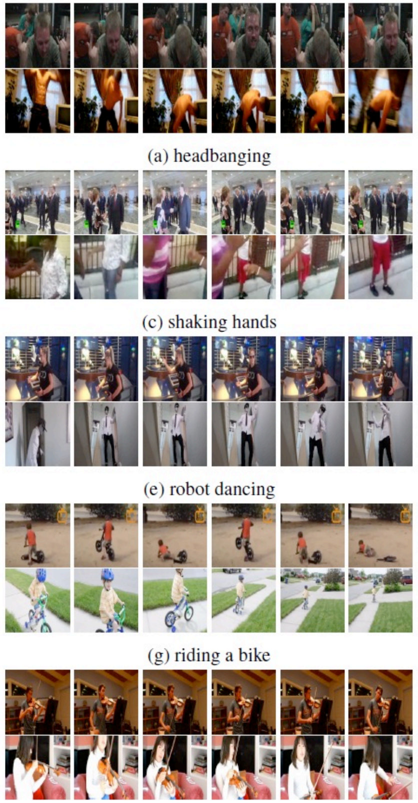
Existing models struggle to adapt to targeted causal interventions.
How can we make them more robust to spurious correlations?

[Agarwal et al., Towards Causal VQA: Revealing & Reducing Spurious Correlations by Invariant & Covariant Semantic Editing. CVPR 2020]

Sub-Challenge 6c: Multimodal Learning Process

Definition: Characterizing the learning and optimization challenges involved when learning from heterogeneous data.

Kinetics dataset



Adding more modalities should always help?

Modalities: **RGB** (video clips)
A (Audio features)
OF (optical flow - motion)

Dataset	Multi-modal	V@1	Best Uni	V@1	Drop
Kinetics	A + RGB	71.4	RGB	**72.6**	-1.2
	RGB + OF	71.3	RGB	**72.6**	-1.3
	A + OF	58.3	OF	**62.1**	-3.8
	A + RGB + OF	70.0	RGB	**72.6**	-2.6

But sometimes multimodal doesn't help! **Why?**

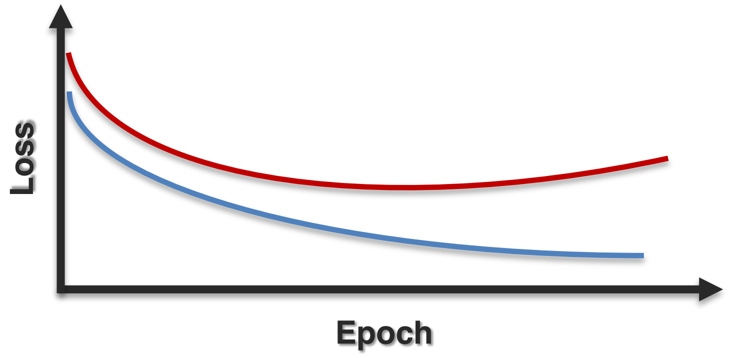
[Wang et al., What Makes Training Multi-modal Classification Networks Hard? CVPR 2020]

Optimization challenges

Learning and optimization challenges

2 explanations for drop in performance:

1. Multimodal networks are more prone to overfitting due to increased complexity
2. Different modalities overfit and generalize at different rates



Key idea 1: compute overfitting-to-generalization ratio (OGR)

Gap between training and valid loss

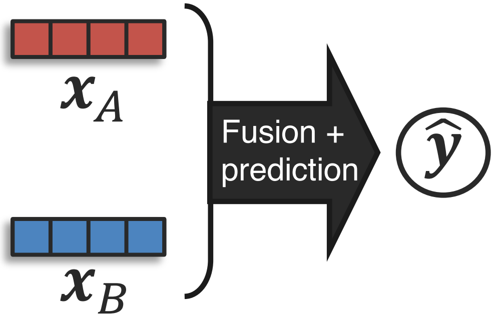
OGR wrt each modality tells us how much to train that modality

[Wang et al., What Makes Training Multi-modal Classification Networks Hard? CVPR 2020]

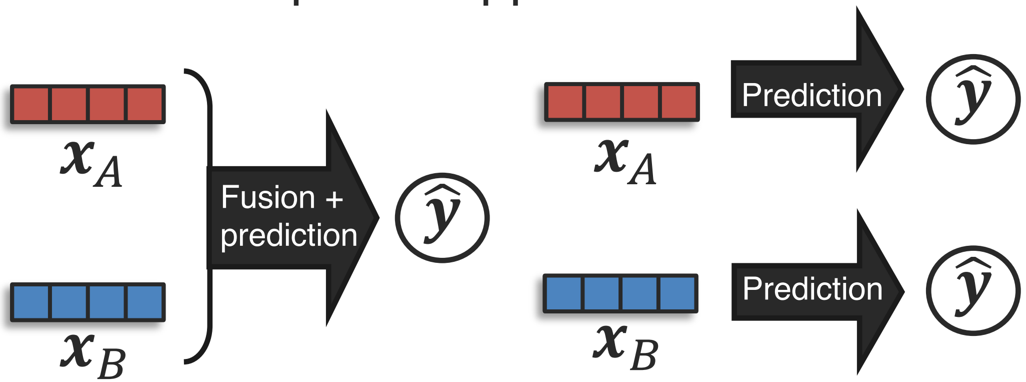
Optimization challenges

Learning and optimization challenges

Conventional approach



Proposed approach



Key idea 2: Simultaneously train unimodal networks to estimate OGR wrt each modality

Reweight multimodal loss using unimodal OGR values

Allows to better balance generalization & overfitting rate of different modalities

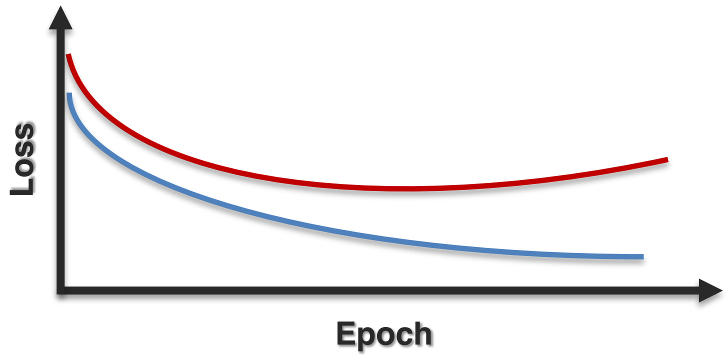
[Wang et al., What Makes Training Multi-modal Classification Networks Hard? CVPR 2020]

Challenges

Open challenges

Open challenges:

- Learning, generalization, and optimization in high-dimensional settings (p >> n).
- Modality shortcuts and biases.
- Dimensionality reduction, modality selection, approximate inference.
- Reducing time and space complexity, model compression and efficiency.

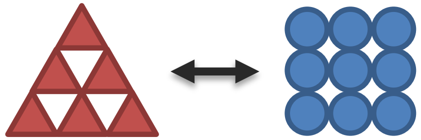


More Quantification

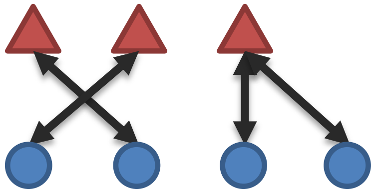
Dimensions of quantification

Representation | Alignment | Reasoning | Transference | Generation

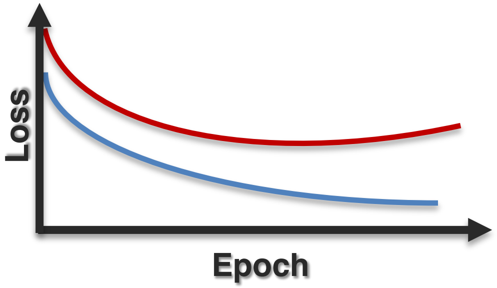
Heterogeneity



Interactions



Learning



Open challenges

Conclusion

What is a Modality?

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**

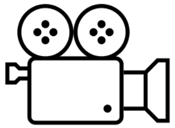
Physiological
- **Skin conductance**
- **Electrocardiogram**

Mobile
- **GPS location**
- **Accelerometer**
- **Light sensors**

What is a Modality?

Definition

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

What is Multimodal?

A dictionary definition…

Multimodal: with multiple modalities

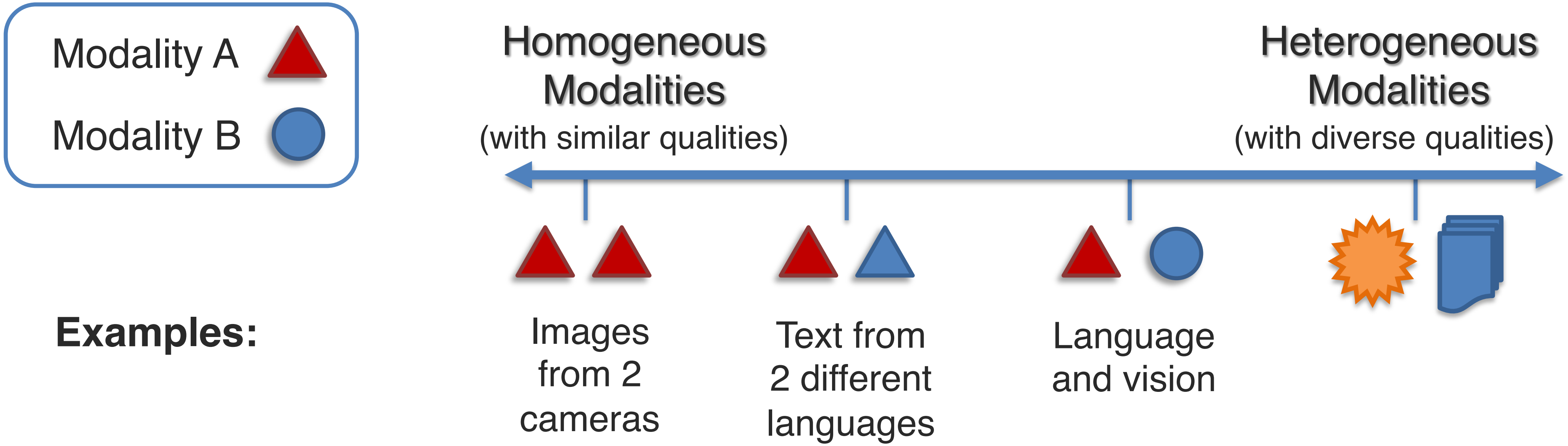
A research-oriented definition…

Multimodal **is the scientific study of**

$\textbf{heterogeneous and } \underbrace{\textbf{interconnected}}_{\textbf{Connected + Interacting}} \textbf{ data}$

Heterogeneous Modalities

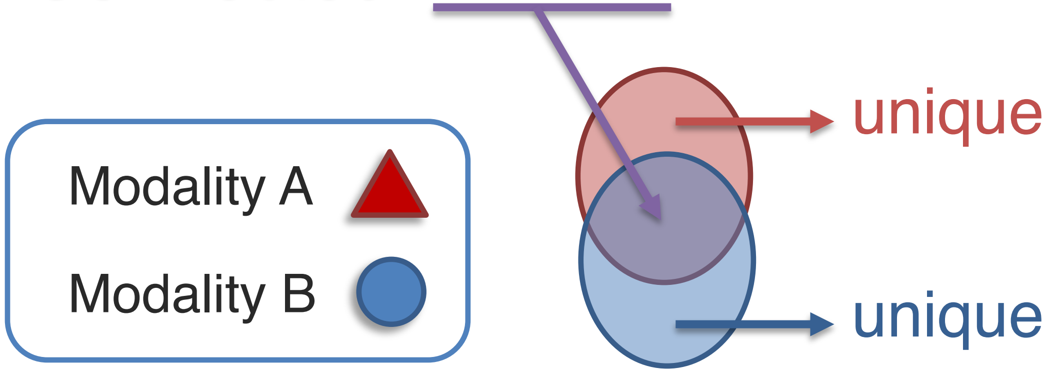
Heterogeneous: Diverse qualities, structures and representations.

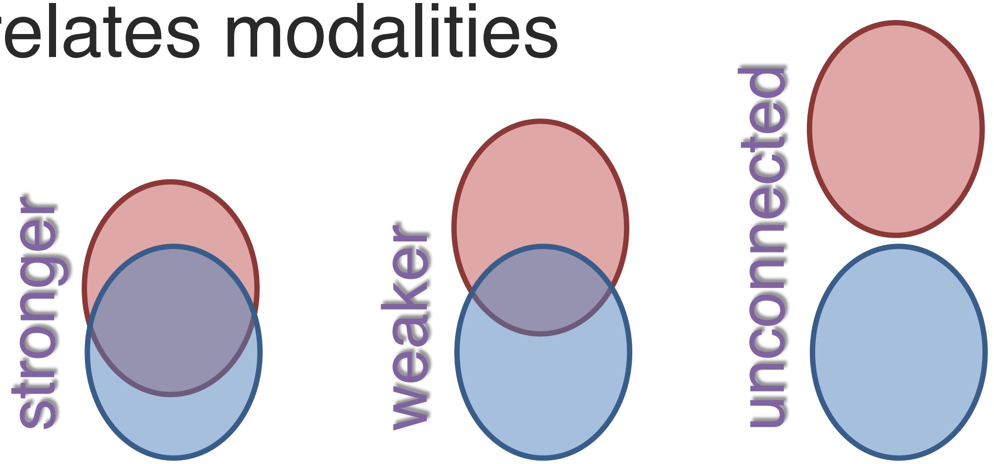


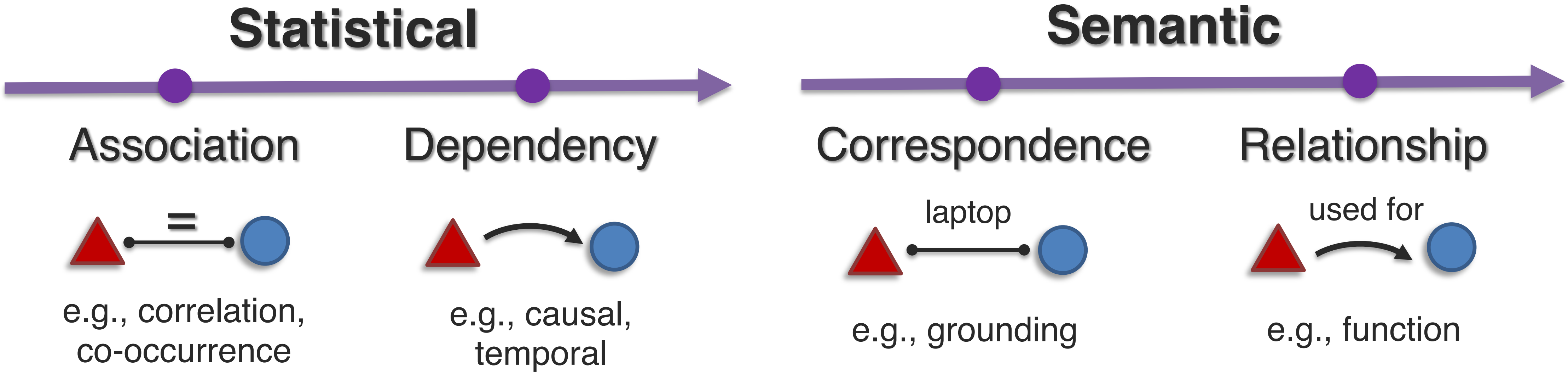
Abstract modalities are more likely to be homogeneous

Connected Modalities

Connected: Shared information that relates modalities





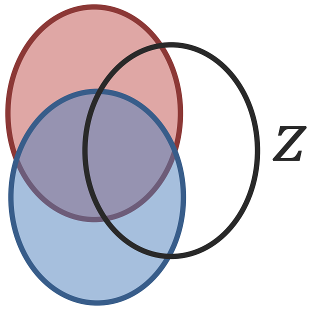


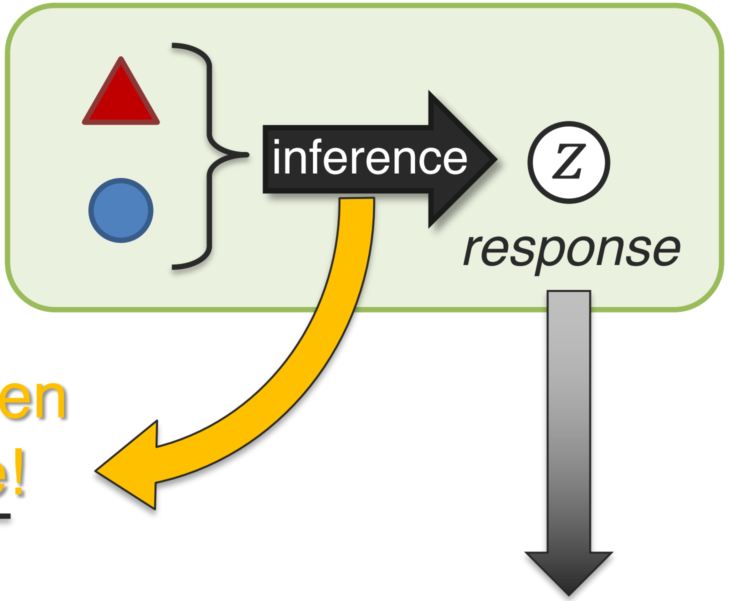
Interacting Modalities

Interacting: process affecting each modality, creating new response

Modality A

Modality B

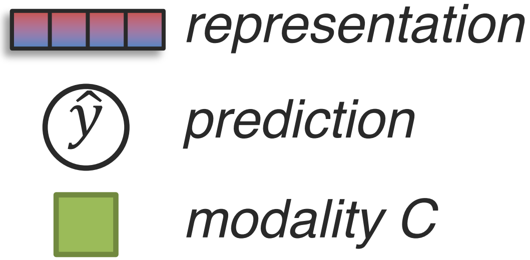




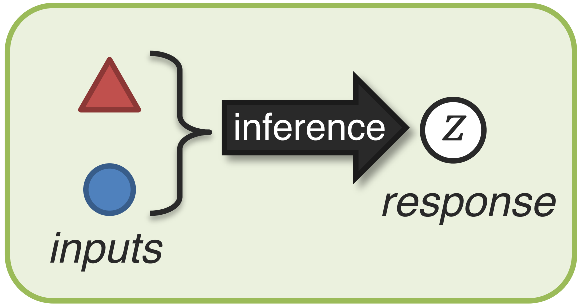
Interactions happen during inference!

"Inference" examples:

- Representation fusion
- Prediction task
- Modality translation



Taxonomy of Interaction Responses – A Behavioral Science View

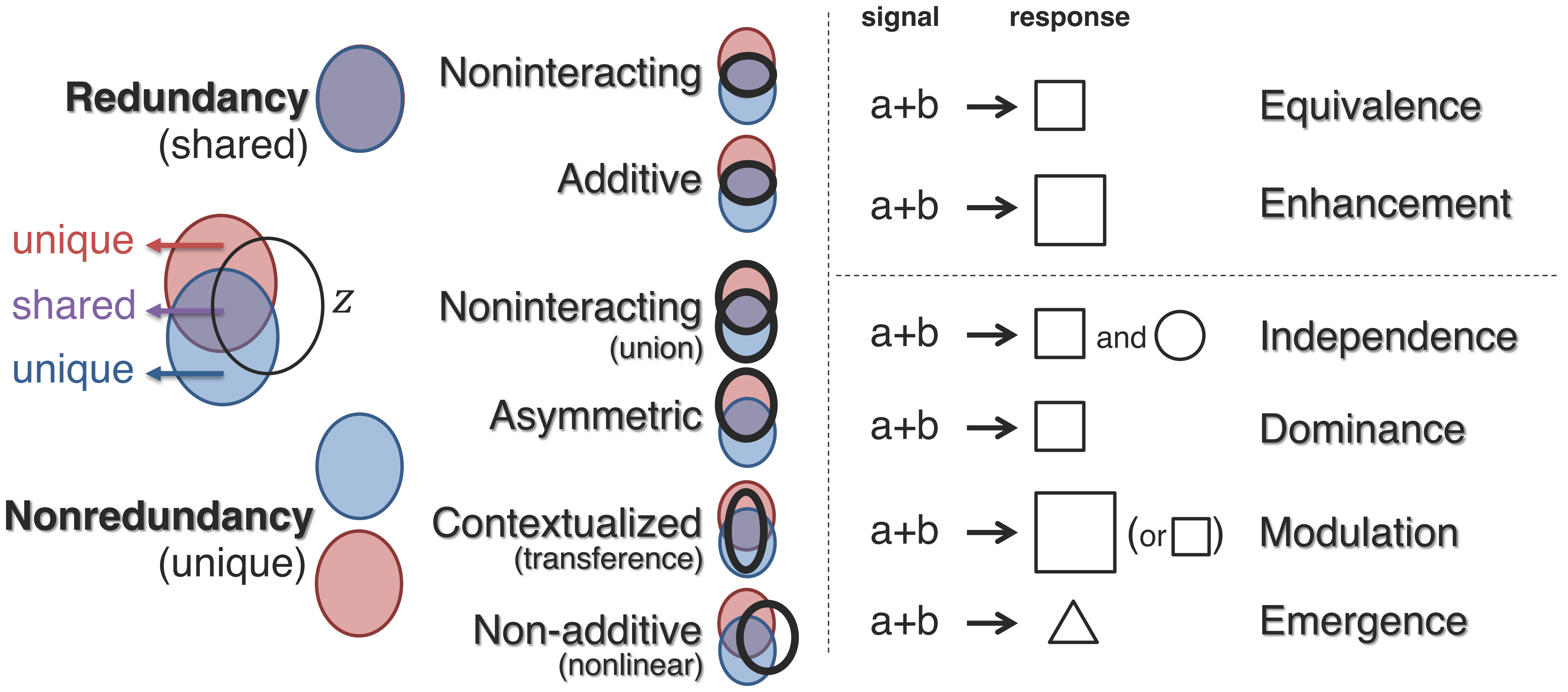


Multimodal Communication

	signal	response	signal	response	
Redundancy	a	□	a+b	□	Equivalence
	b	□	a+b	□	Enhancement
Nonredundancy	a	□	a+b	□ and ○	Independence
	b	○	a+b	□	Dominance
			a+b	□ (or □)	Modulation
			a+b	△	Emergence

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

Interacting Modalities



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

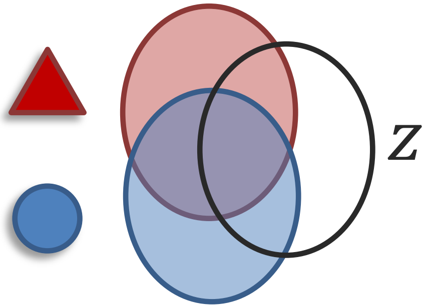
Heterogeneous

+

Connected

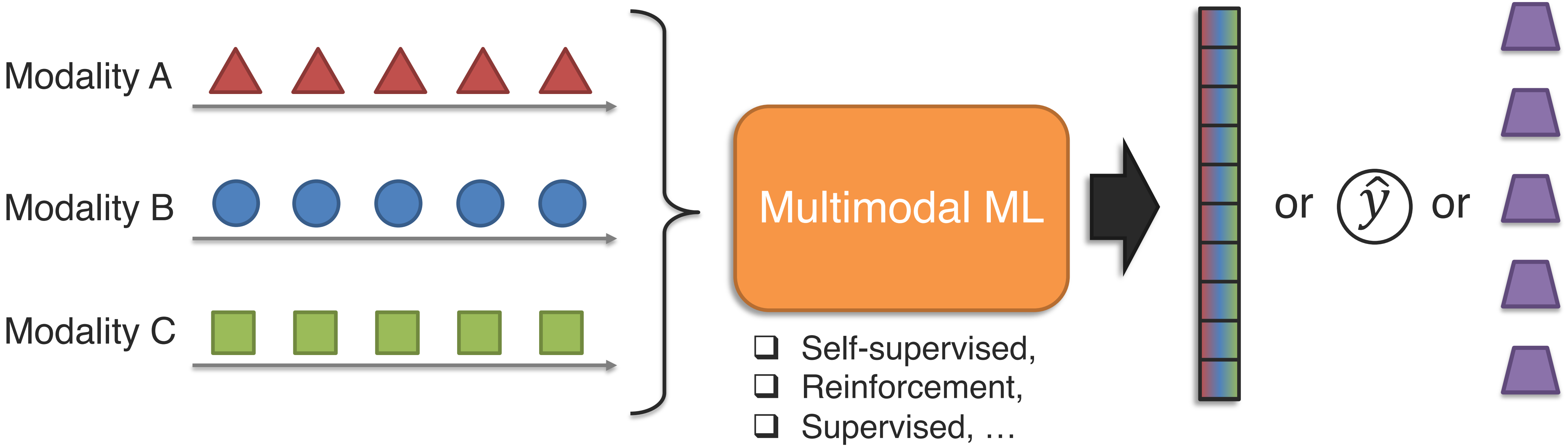
+

Interacting



Multimodal is the scientific study of heterogeneous and interconnected data ☺

Multimodal Machine Learning



What are the core multimodal technical challenges,

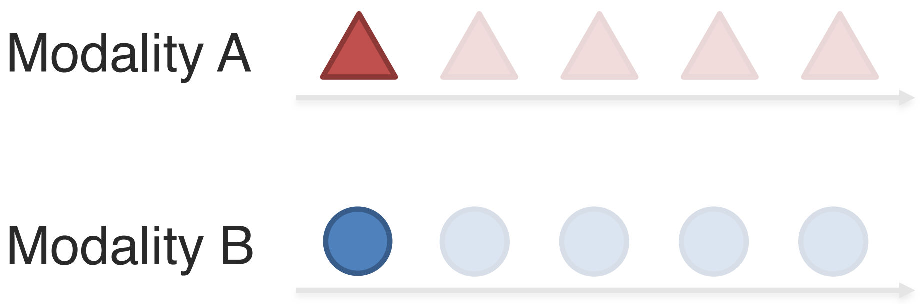
understudied in conventional machine learning?

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:



It can be seen as a "local" representation

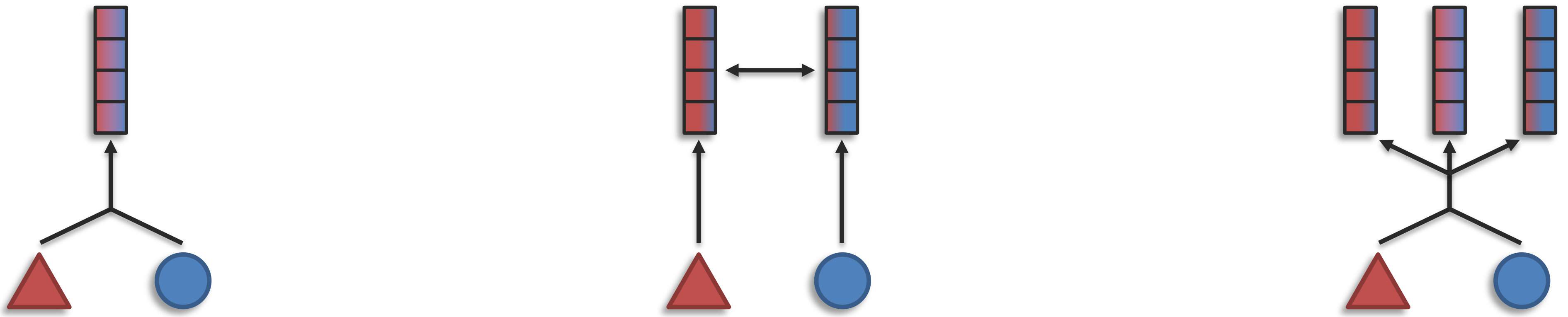
or

representation using holistic features

Challenge 1: Representation

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

Challenge 2: Alignment

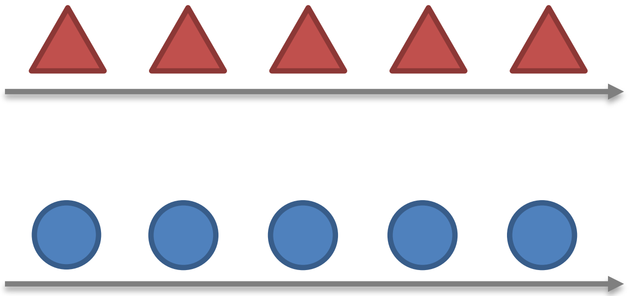
Definition: Identifying and modeling cross-modal connections between all elements of multiple modalities, building from the data structure

➡ Most modalities have internal structure with multiple elements

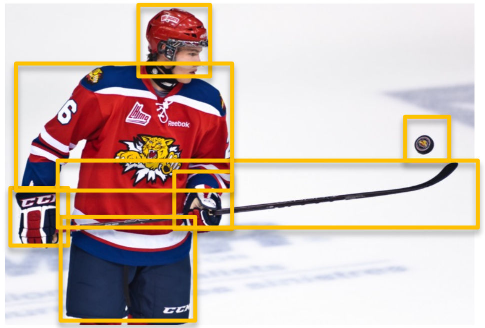
Elements with temporal structure:

Modality A

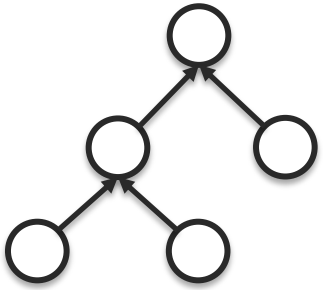
Modality B



Other structured examples:



Spatial



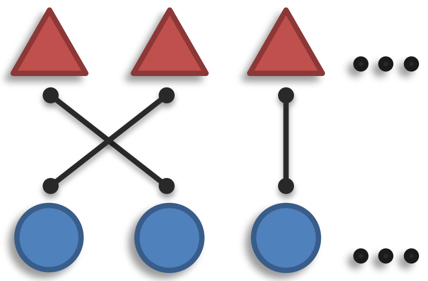
Hierarchical

Challenge 2: Alignment

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

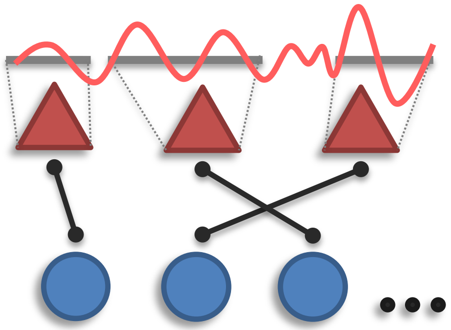
Sub-challenges:

Discrete Alignment



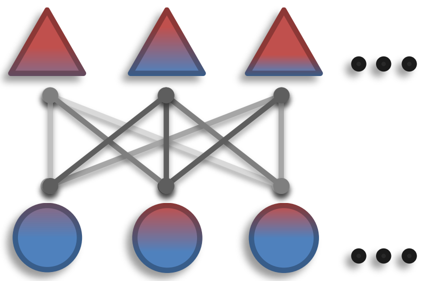
Discrete elements and connections

Continuous Alignment



Segmentation and continuous warping

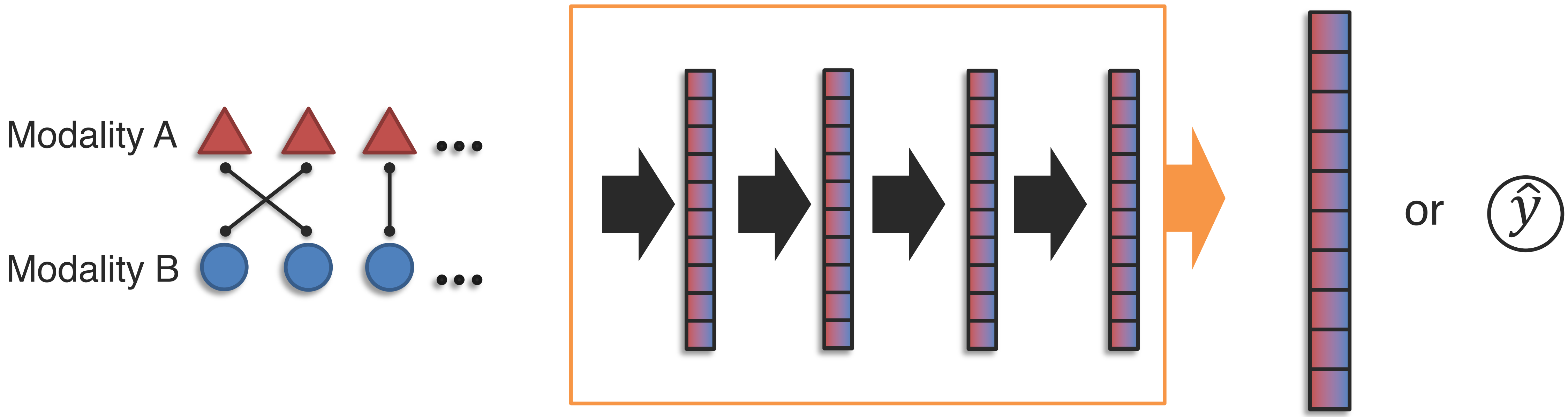
Contextualized Representation



Alignment + representation

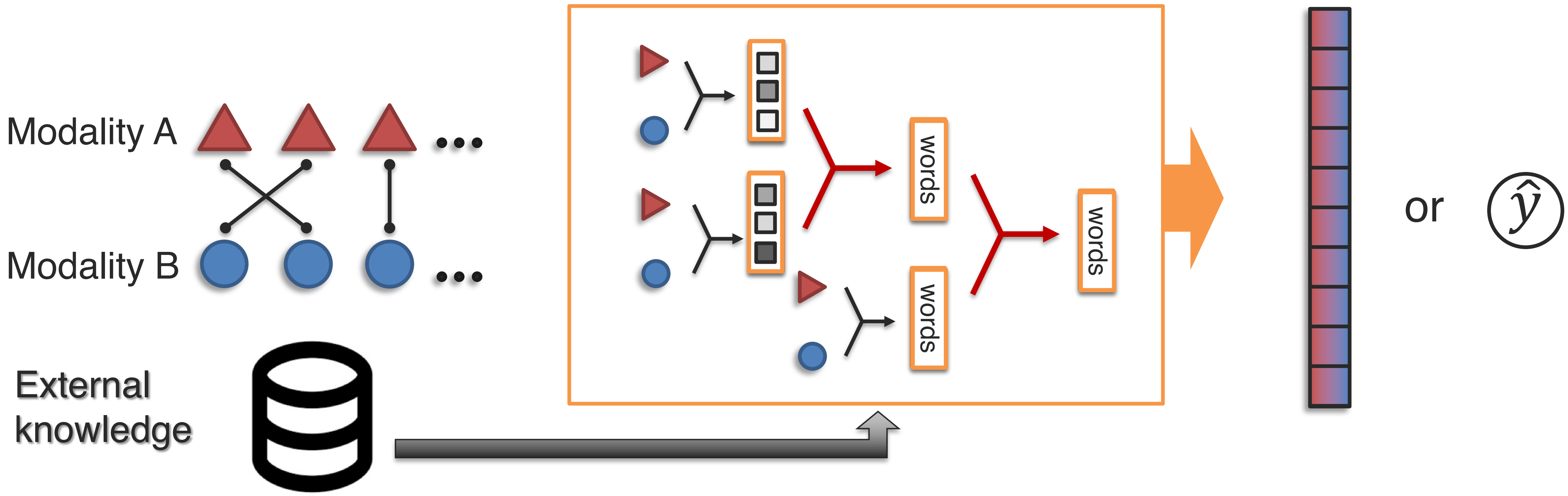
Challenge 3: Reasoning

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



Challenge 3: Reasoning

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

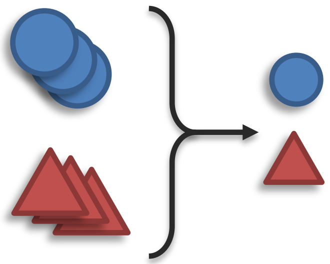


Challenge 4: Generation

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

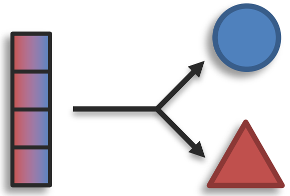
Sub-challenges:

Summarization



Translation

Creation



Information: (content)

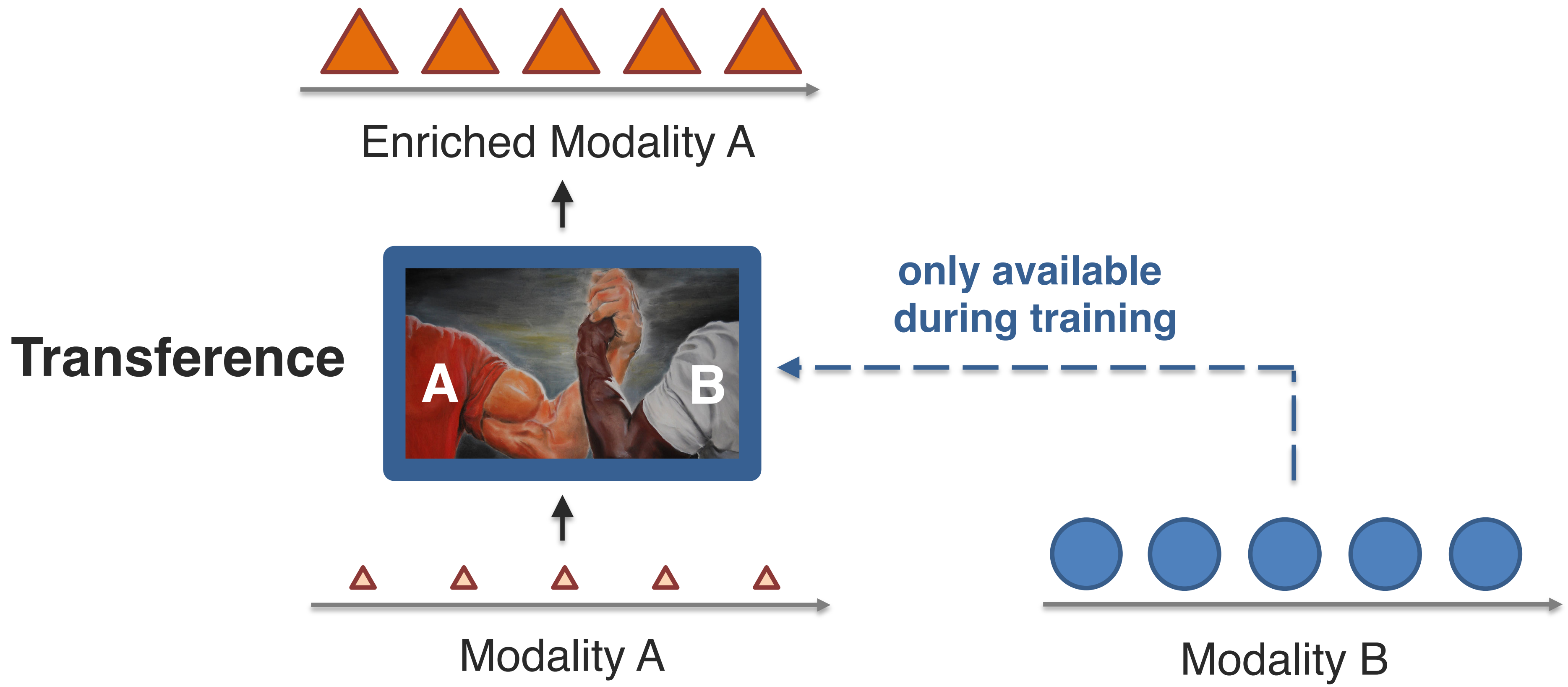
Reduction

Maintenance

Expansion

Challenge 5: Transference

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

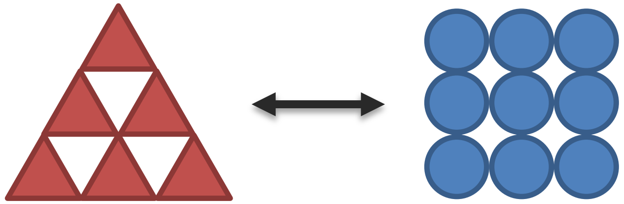


Challenge 6: Quantification

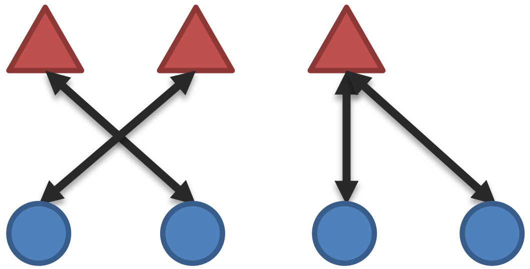
Definition: Empirical and theoretical study to better understand heterogeneity, cross-modal interactions and the multimodal learning process

Sub-challenges:

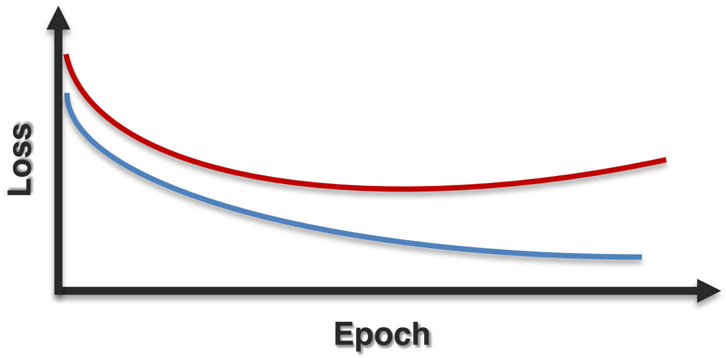
Heterogeneity



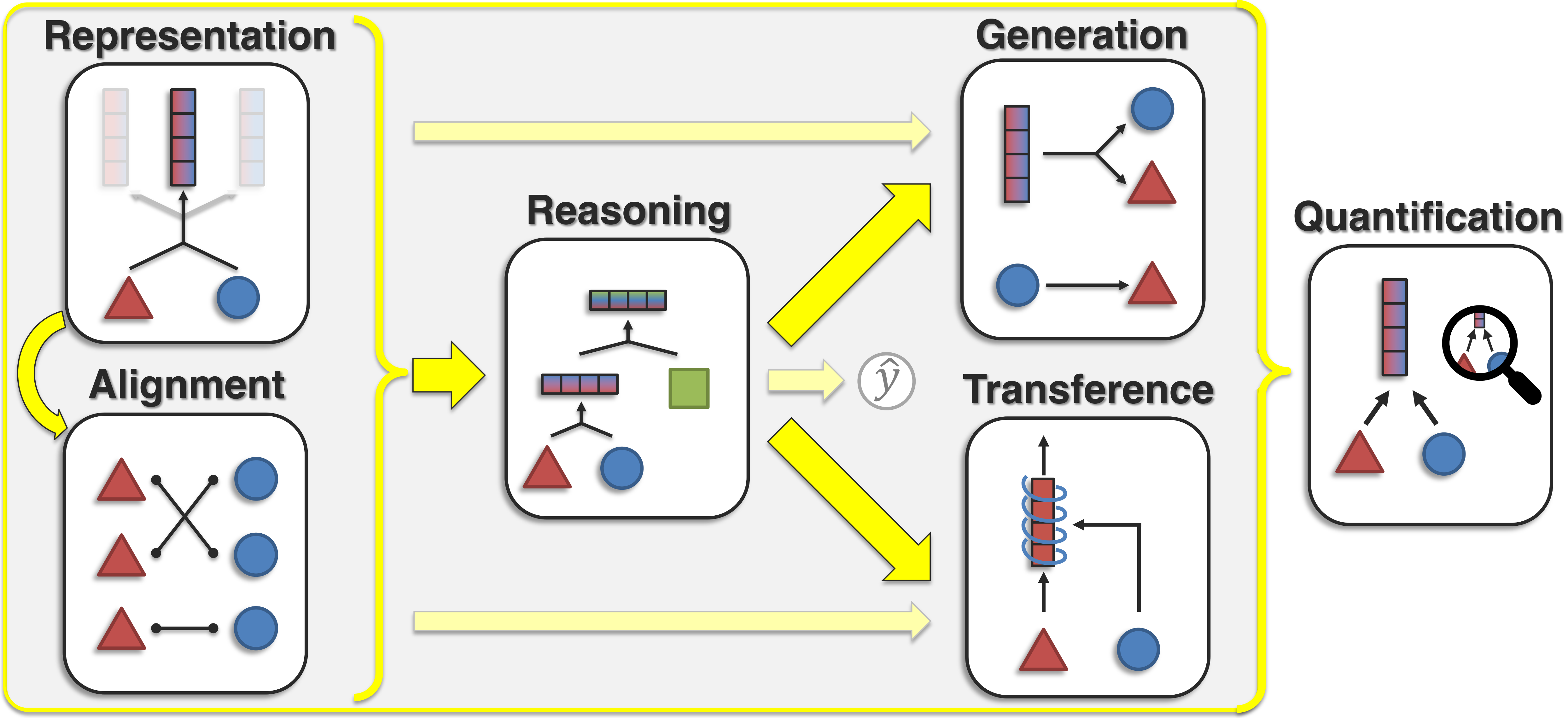
Connections & Interactions



Learning

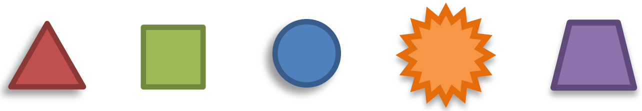


Core Multimodal Challenges



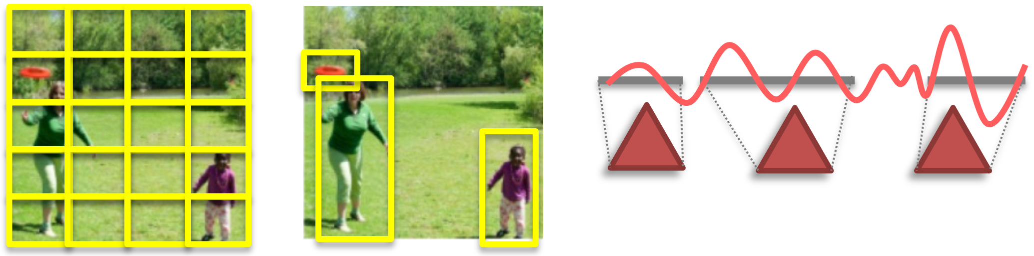
Future Direction: Heterogeneity

Homogeneity vs Heterogeneity



Examples:

Arbitrary Tokenization



Beyond Additive Interactions

Causal, logical interactions

Brain-inspired representations

Future Direction: High-modality

MultiBench

https://github.com/pliang279/MultiBench

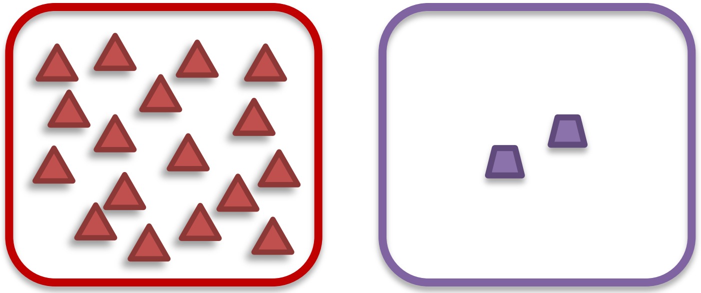
Few modalities → High-modality

Language Vision Audio Graphs Control LIDAR Sensors Set Table Financial Medical

Examples:

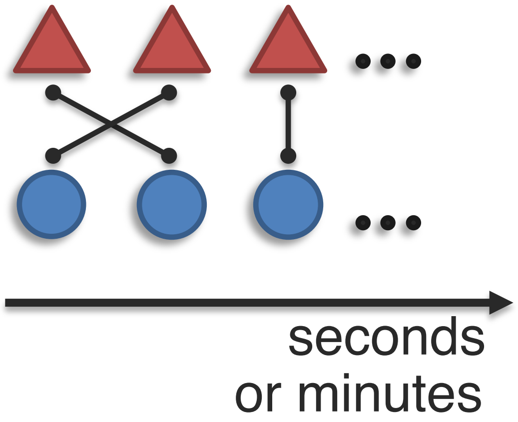
Non-parallel learning

Limited resources



Future Direction: Long-term

Short-term



Long-term

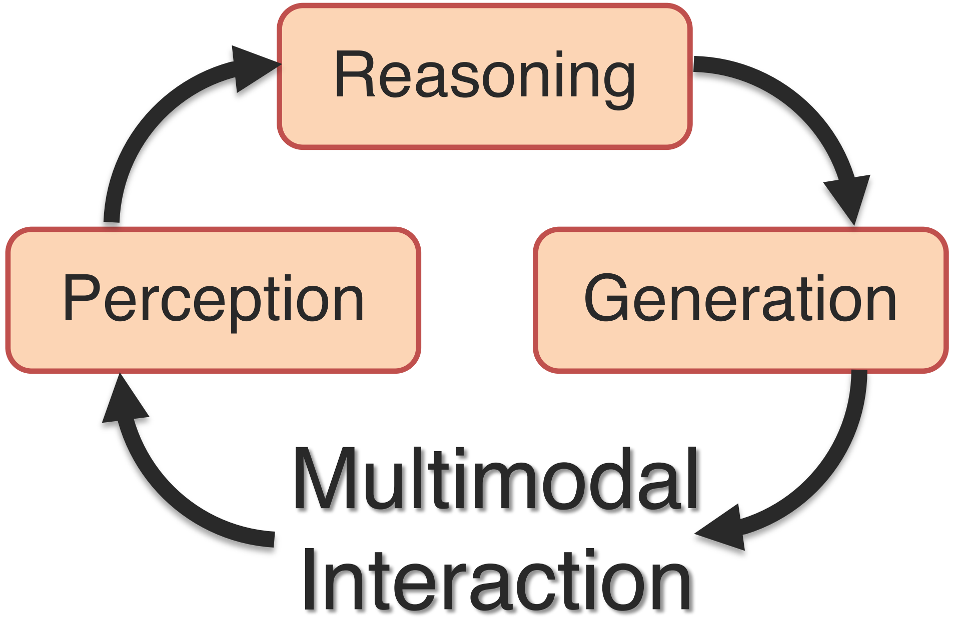
Examples:

Compositionality **Memory** **Personalization**

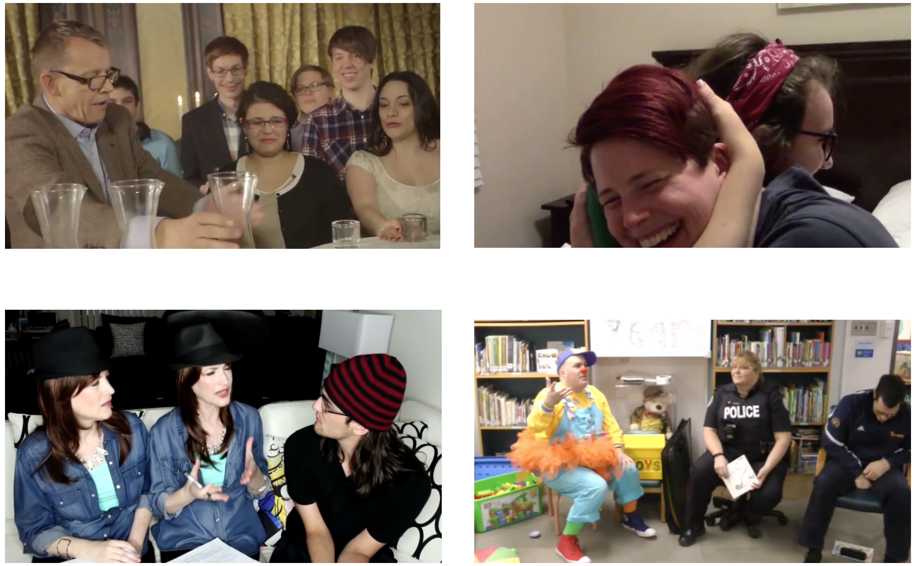
Future Direction: Interaction

Social-IQ

https://www.thesocialiq.com/



Social Intelligence



Examples:

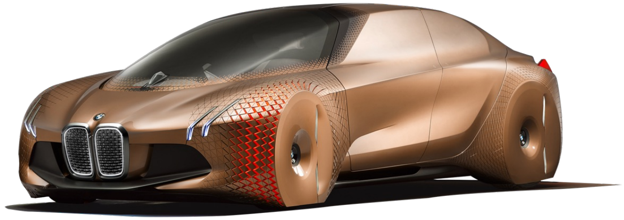
Multi-Party **Causality** **Ethical**

Future Direction: Real-world

MultiViz

https://github.com/pliang279/MultiViz

Healthcare Decision Support



Intelligent Interfaces and Vehicles

Online Learning and Education

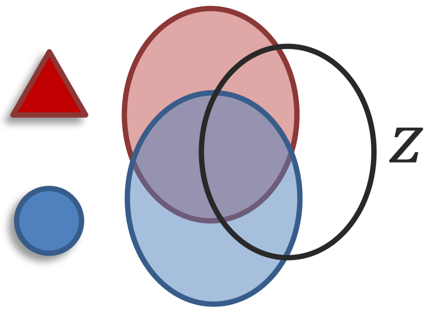
Examples:

Robustness **Fairness** **Generalization**

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

Advanced Topics in Multimodal ML @ CMU

Advanced Topics in MultiModal Machine Learning

11-877 • Spring 2022 • Carnegie Mellon University

Multimodal machine learning (MMML) is a vibrant multi-disciplinary research field which addresses some of the original goals of artificial intelligence by integrating and modeling multiple communicative modalities, including language, vision, and acoustic. This research field brings some unique challenges for multimodal researchers given the heterogeneity of the data and the contingency often found between modalities. This course is designed to be a graduate-level course covering recent research papers in multimodal machine learning, including technical challenges with representation, alignment, reasoning, generation, co-learning and quantifications. The main goal of the course is to increase critical thinking skills, knowledge of recent technical achievements, and understanding of future research directions.

- **Time:** Friday 10:10-11:30 am
- **Location:** Virtual for the first 2 weeks (find zoom link in piazza), GHC 5222 thereafter
- **Discussion and Q&A:** Piazza
- **Assignment submissions:** Canvas (for registered students only)
- **Contact:** Students should ask all course-related questions on Piazza, where you will also find announcements.

Instructor Louis-Philippe Morency
Email: morency@cs.cmu.edu

Instructor Amir Zadeh
Email: abagherz@cs.cmu.edu

Instructor Paul Liang
Email: pliang@cs.cmu.edu

1/28 Week 2: **Cross-modal interactions** [synopsis]

- What are the different ways in which modalities can interact with each other in multimodal tasks? Can we formalize a taxonomy of such cross-modal interactions, which will enable us to compare and contrast them more precisely?
- What are the design decisions (aka inductive biases) that can be used when modeling these cross-modal interactions in machine learning models?
- What are the advantages and drawbacks of designing models to capture each type of cross-modal interaction? Consider not just prediction performance, but tradeoffs in time/space complexity, interpretability, etc.
- Given an arbitrary dataset and prediction task, how can we systematically decide what type of cross-modal interactions exist, and how can that inform our modeling decisions?
- Given trained multimodal models, how can we understand or visualize the nature of cross-modal interactions?

Readings:

- Does my multimodal model learn cross-modal interactions? It's harder to tell than you might think!
- What Does BERT with Vision Look At?
- Multiplicative Interactions and Where to Find Them
- Cooperative Learning for Multi-view Analysis
- Vision-and-Language or Vision-for-Language? On Cross-Modal Influence in Multimodal Transformers
- Seeing past words: Testing the cross-modal capabilities of pretrained V&L models on counting tasks

2/4 Week 3: **Multimodal co-learning** [synopsis]

- What are the types of cross-modal interactions involved to enable such co-learning scenarios where multimodal training ends up generalizing to unimodal testing?
- What are some design decisions (inductive bias) that could be made to promote transfer of information from one modality to another?
- How do we ensure that during co-learning, only useful information is transferred, and not some undesirable bias? This may become a bigger issue in low-resource settings.
- How can we know if co-learning has succeeded? Or failed? What approaches could we develop to visualize and probe the success of co-learning?
- How can we formally, empirically, or intuitively measure the additional information provided by auxiliary modality? How can we design controlled experiments to test these hypotheses?
- What are the advantages and drawbacks of information transfer during co-learning? Consider not just prediction performance, but also tradeoffs with complexity, interpretability, fairness, etc.

Readings:

- Multimodal Prototypical Networks for Few-shot Learning
- SMIL: Multimodal Learning with Severely Missing Modality
- Multimodal Co-learning: Challenges, Applications with Datasets, Recent Advances and Future Directions
- Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision
- What Makes Multi-modal Learning Better than Single (Provably)
- Found in Translation: Learning Robust Joint Representations by Cyclic Translations Between Modalities
- Zero-Shot Learning Through Cross-Modal Transfer
- 12-in-1: Multi-Task Vision and Language Representation Learning
- A Survey of Reinforcement Learning Informed by Natural Language

https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2022/