Language Technologies Institute

Carnegie Mellon University

# Multimodal Machine Learning

## Lecture 12.2: Quantification

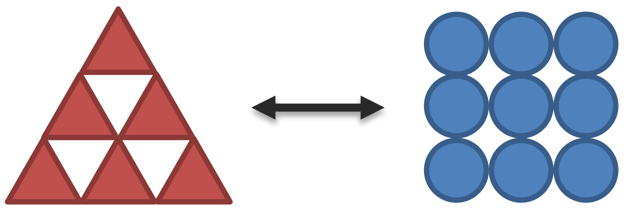
Paul Liang

** Co-lecturer: Louis-Philippe Morency. Original course co-developed with Tadas Baltrusaitis. Spring 2021 and 2022 editions taught by Yonatan Bisk. Spring 2023 edition taught by Yonatan and Daniel Fried*

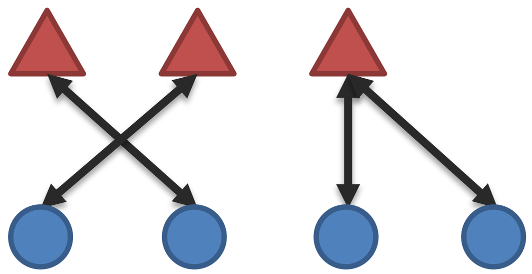
# Quantification

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

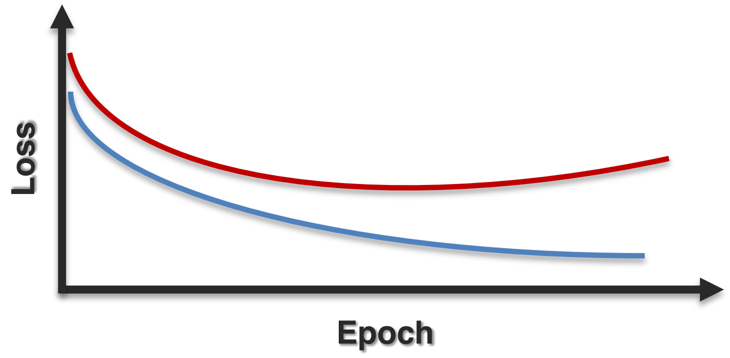
(A) Heterogeneity



(B) Interactions

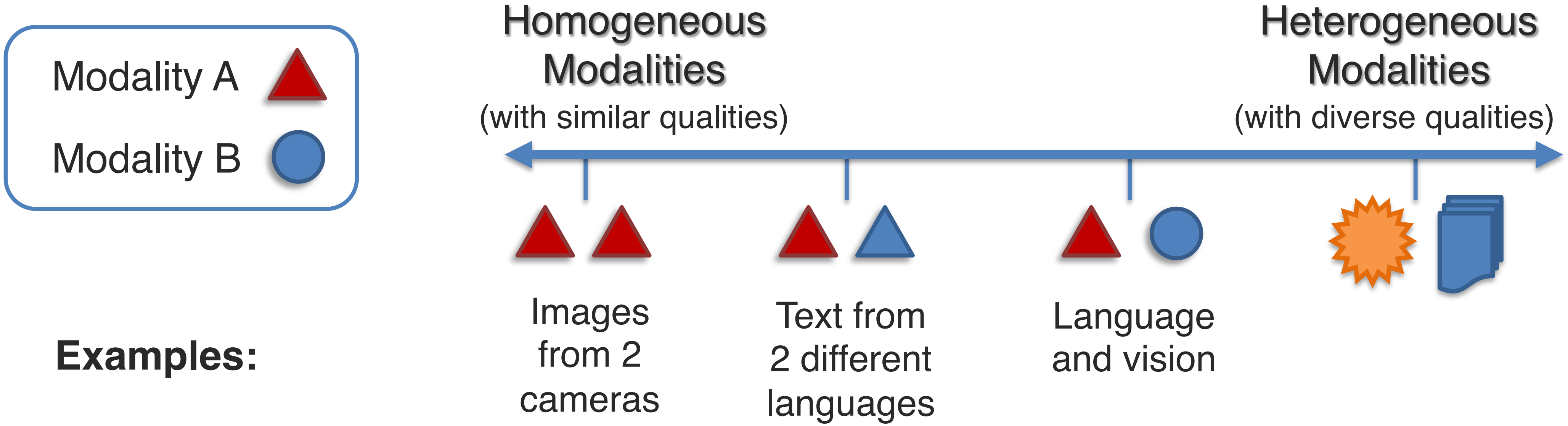


(C) Learning



# Sub-Challenge 6a: Heterogeneity

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

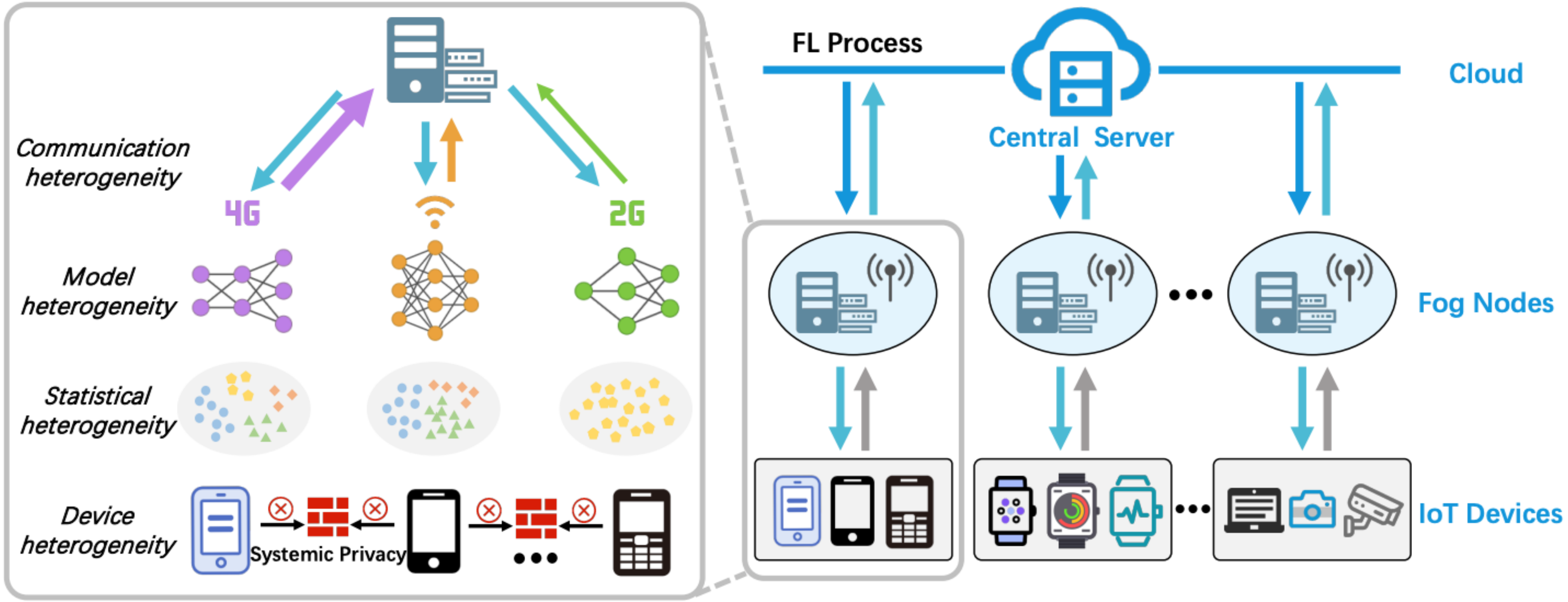



**Examples:**

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

# Distribution heterogeneity

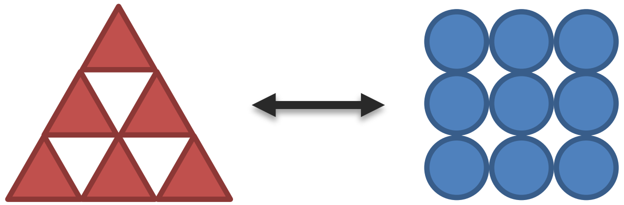
## Inspired by distributed learning




[Ye et al., Heterogeneous Federated Learning: State-of-the-art and Research Challenges, 2023]

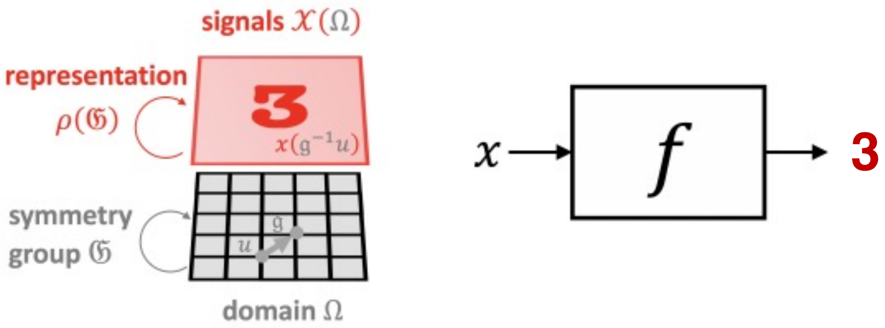
# Structure heterogeneity

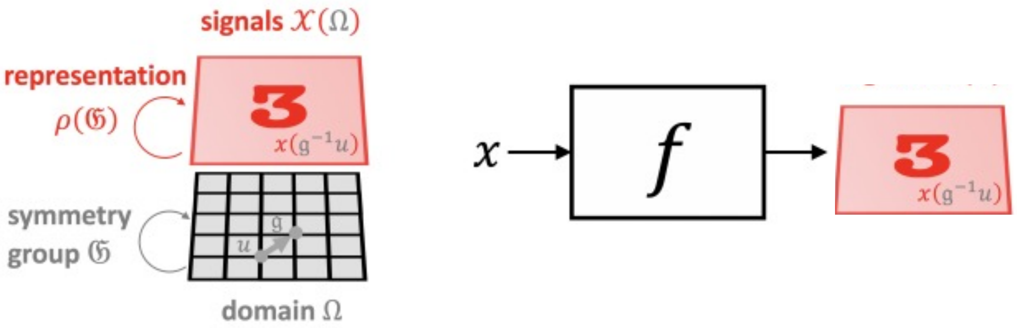
## Inspired by structure learning



A function $f : \mathcal{X}(\Omega) \to \mathcal{Y}$ is $\mathfrak{G}$-*invariant* if $f(\rho(\mathfrak{g})x) = f(x)$ for all $\mathfrak{g} \in \mathfrak{G}$ and $x \in \mathcal{X}(\Omega)$, i.e., its output is unaffected by the group action on the input.

A function $f : \mathcal{X}(\Omega) \to \mathcal{X}(\Omega)$ is $\mathfrak{G}$-*equivariant* if $f(\rho(\mathfrak{g})x) = \rho(\mathfrak{g})f(x)$ for all $\mathfrak{g} \in \mathfrak{G}$, i.e., group action on the input affects the output in the same way.




Derives deep learning architectures on grids, graphs, sets, etc.

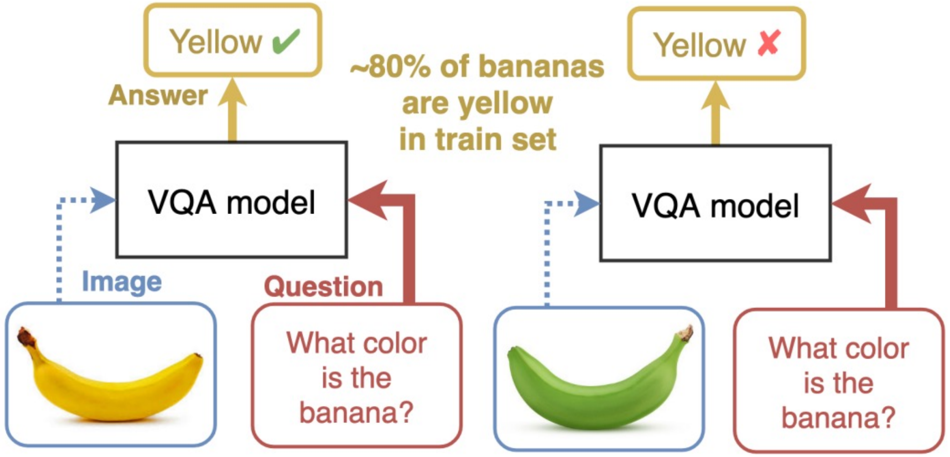
[Bronstein et al., Geometric Deep Learning Grids, Groups, Graphs, Geodesics, and Gauges, 2021]

# Modality Biases

**Heterogeneity in information and relevance**

Unimodal biases and modality collapse

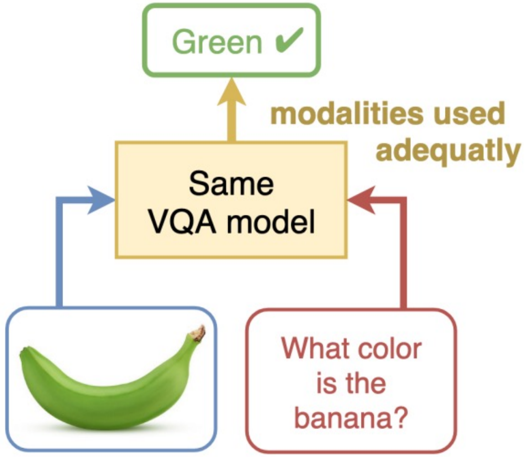
**VQA models answer the question without looking at the image**




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]

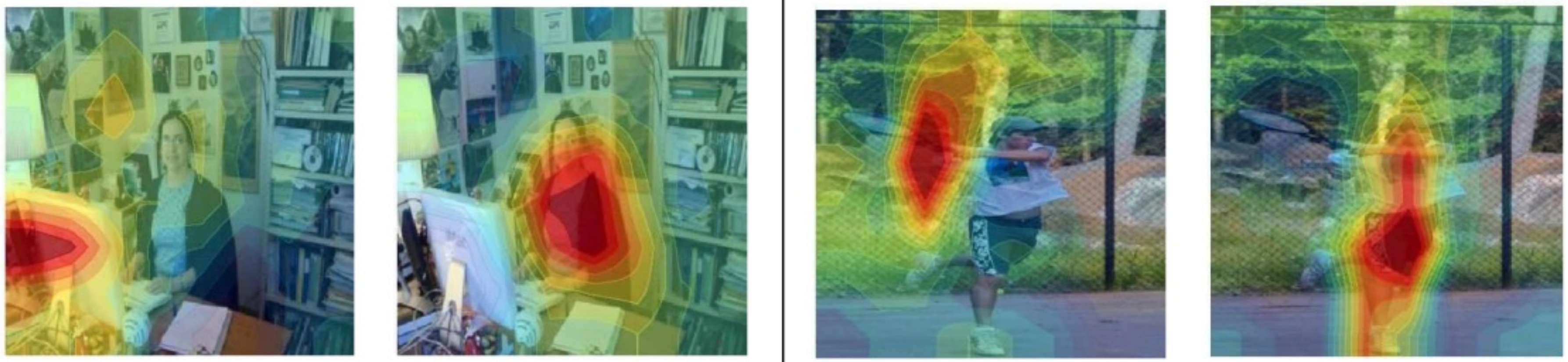
# Modality Biases

**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 | Right for the Right Reasons | Right for the Wrong Reasons | Right for the Right Reasons



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

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

Baseline:
*A **man** holding a tennis racquet on a tennis court.*

Our Model:
*A **man** holding a tennis racquet on a tennis court.*

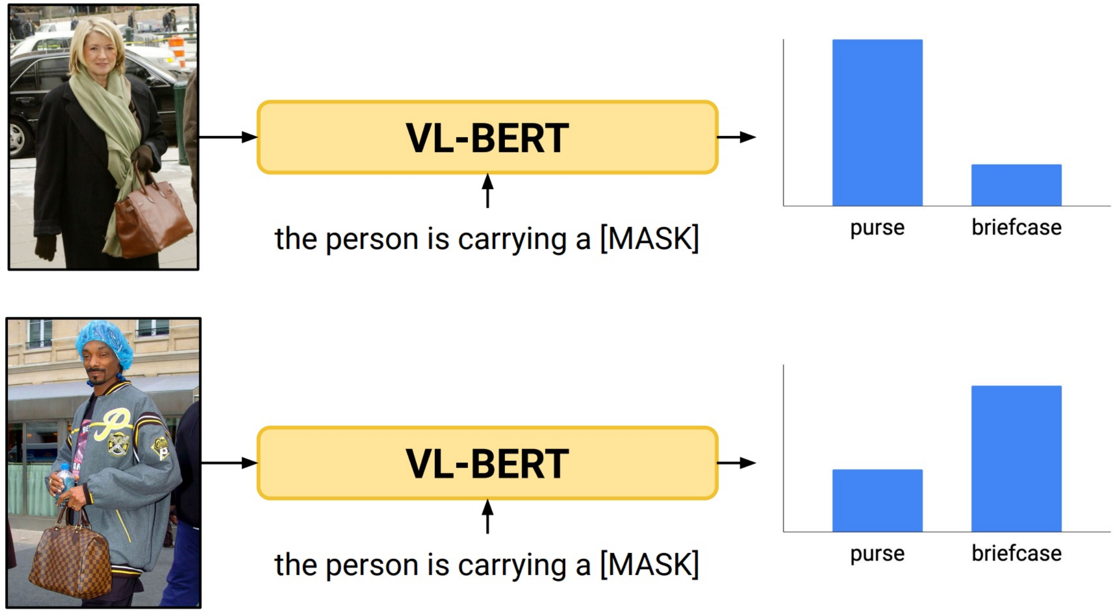
[Hendricks et al., Women also Snowboard: Overcoming Bias in Captioning Models. ECCV 2018]

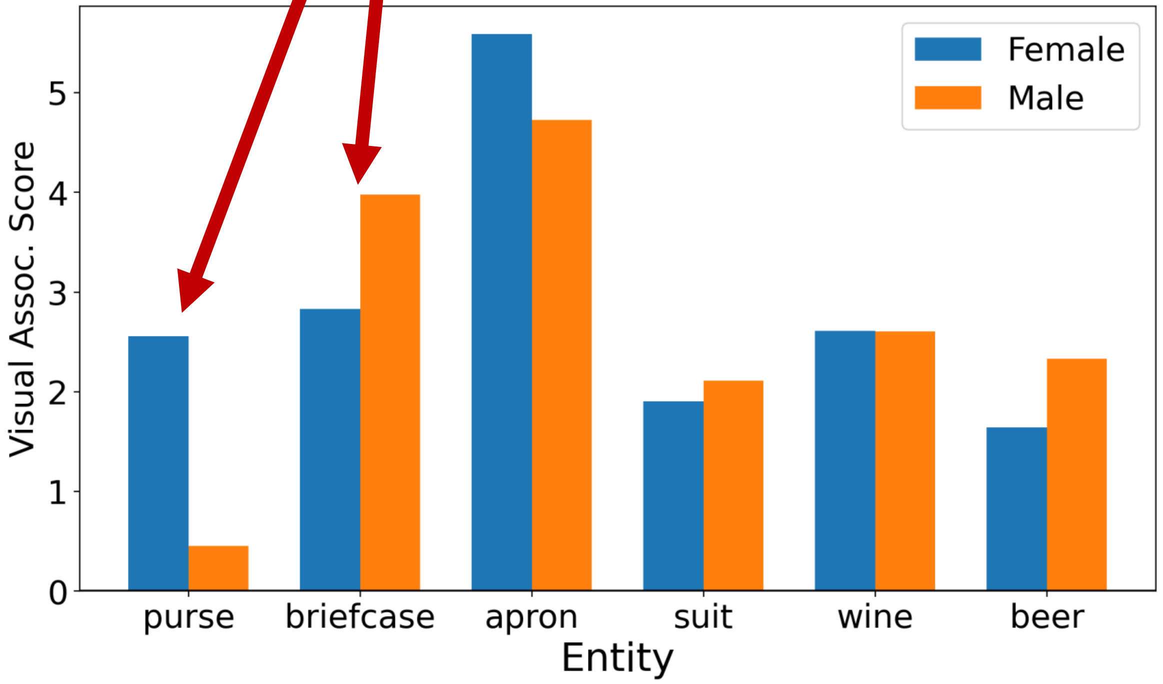
# Modality Biases

## 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



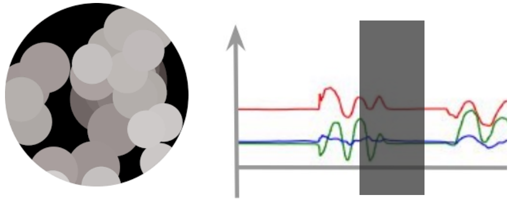

[Srinivasan and Bisk, Worst of Both Worlds: Biases Compound in Pre-trained Vision-and-Language Models. NAACL 2022]

# Noise Topologies and Robustness

## Heterogeneity in noise

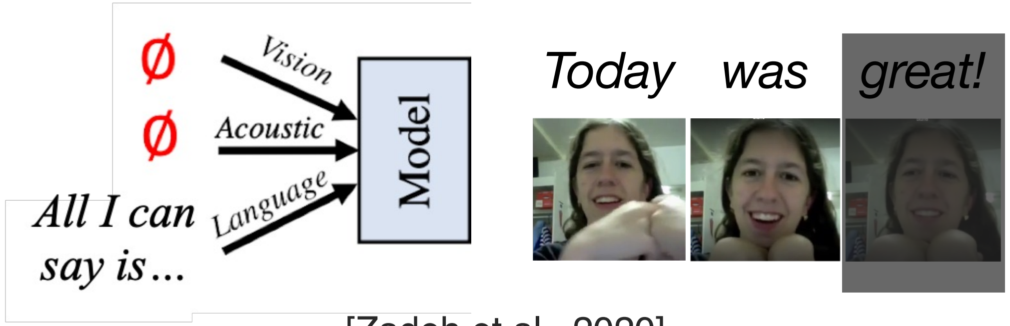
**Modality-specific robustness**

noise ⟶ nosie



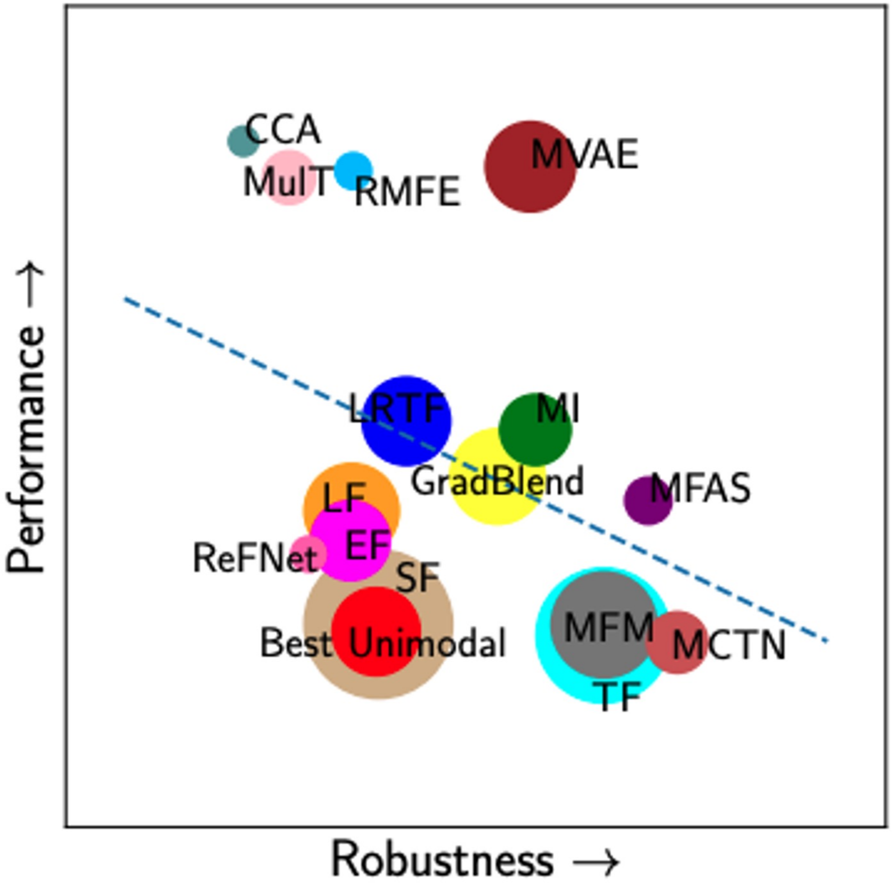
[Belinkov & Bisk, 2018; Subramaniam et al., 2009; Boyat & Joshi, 2015]

**Multimodal robustness**




[Zadeh et al., 2020]

## Strong tradeoffs between performance and robustness



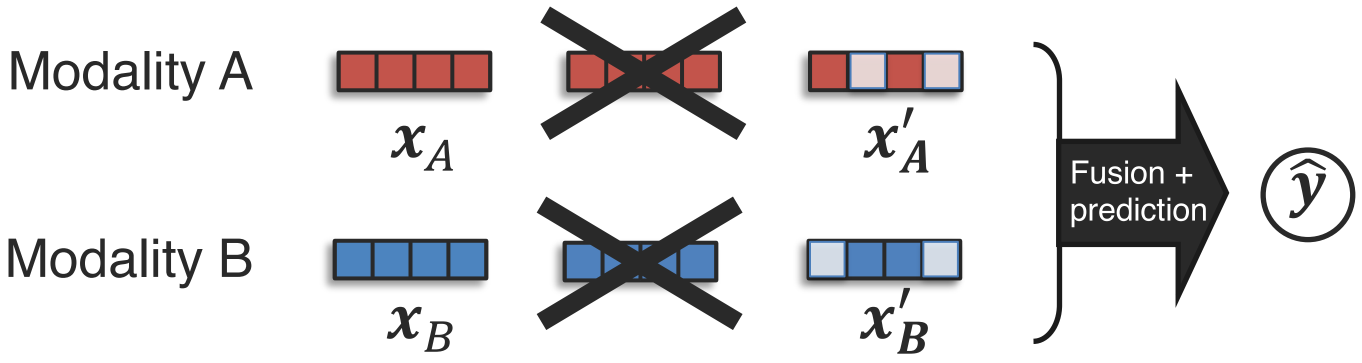

rate of accuracy drops

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

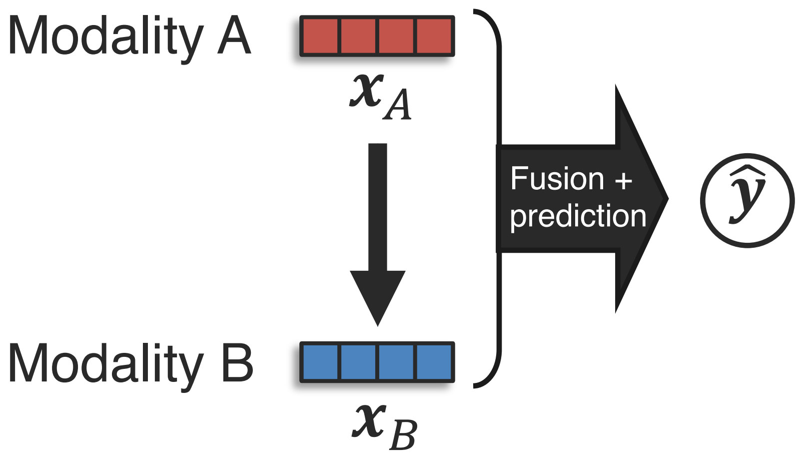
# Noise Topologies and Robustness

## Several approaches towards more robust models

Robust data + training




Infer missing modalities

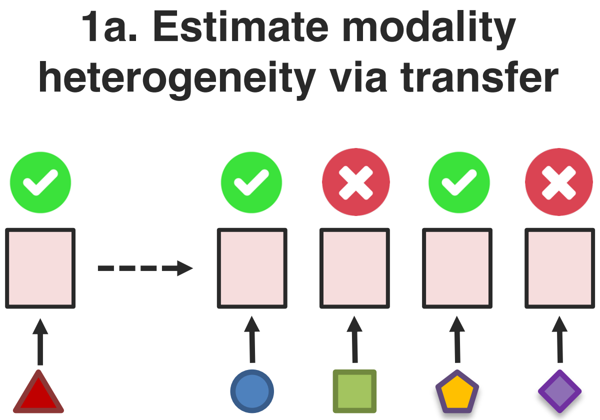



Translation model
Joint probabilistic model

[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]

# Quantifying Heterogeneity via Transfer

## Information transfer, transfer learning perspective




In practice, efficient by pre-trained models and few-shot transfer

**Implicitly captures these:**

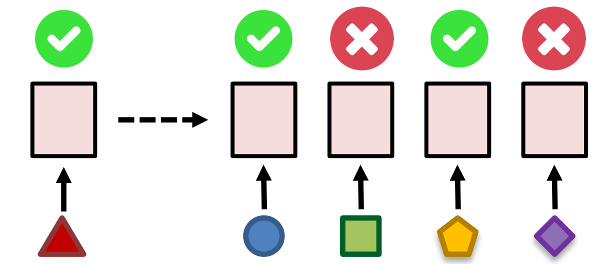
1. Element representation
2. Element distribution
3. Structure
4. Information
5. Noise
6. Relevance

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

# Heterogeneity-aware Fusion

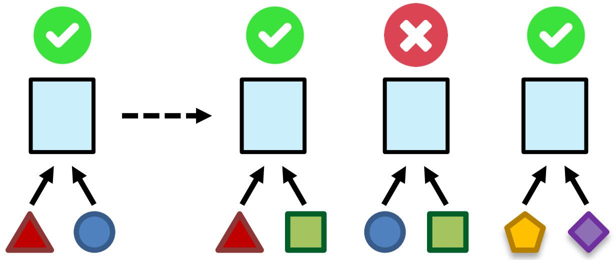
## Information transfer, transfer learning perspective

**1a. Estimate modality heterogeneity via transfer**

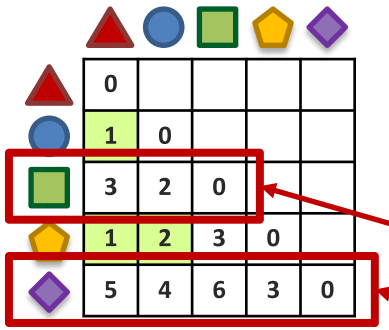


**(Implicitly captures heterogeneity)**

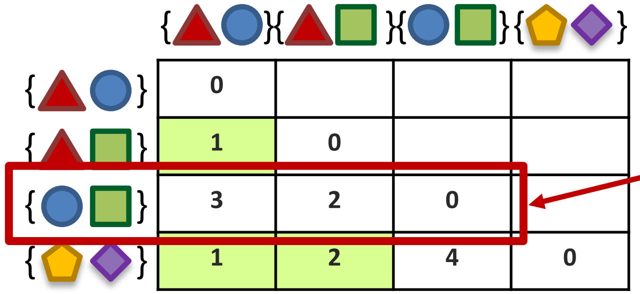
**1b. Estimate interaction heterogeneity via transfer**



**2a. Compute modality heterogeneity matrix**




**2b. Compute interaction heterogeneity matrix**




**3. Determine parameter clustering**

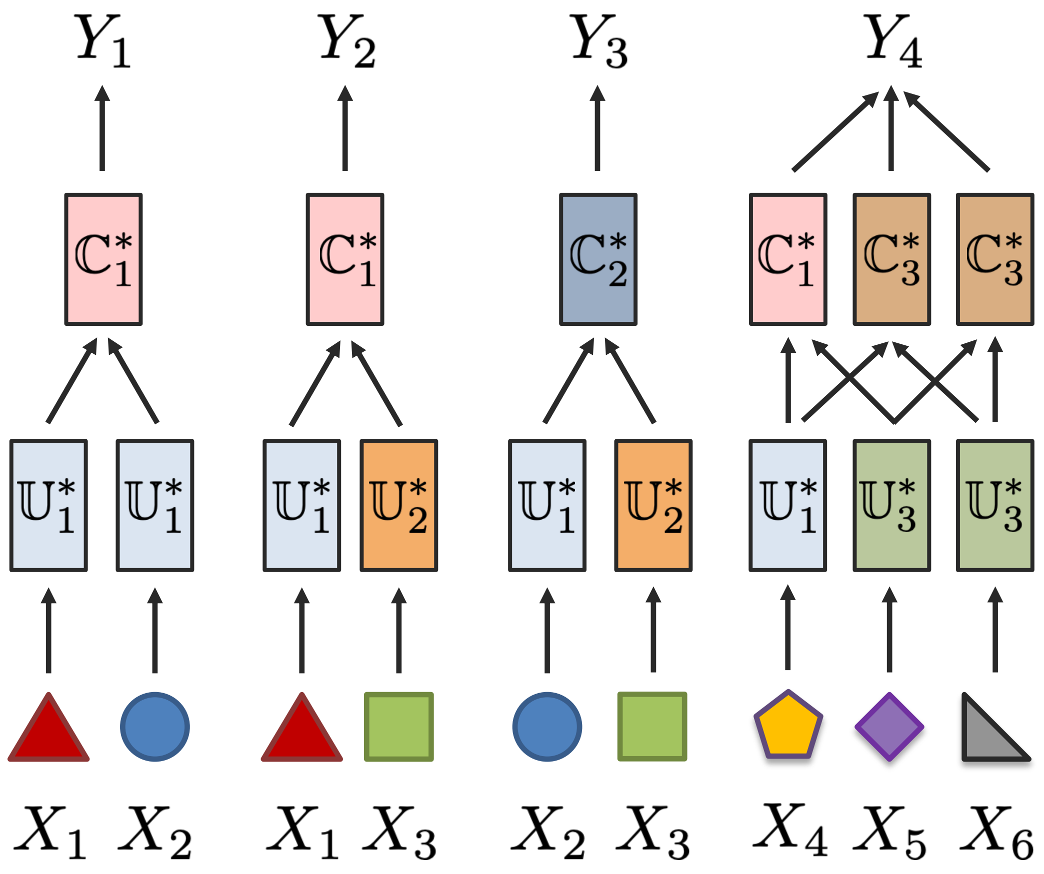
$$\mathbb{U}_1 = \{U_1, U_2, U_4\}$$
$$\mathbb{U}_2 = \{U_3\}$$
$$\mathbb{U}_3 = \{U_5\}$$

$$\mathbb{C}_1 = \{C_{12}, C_{13}, C_{45}\}$$
$$\mathbb{C}_2 = \{C_{23}\}$$

[Zamir et al., Taskonomy: Disentangling Task Transfer Learning. CVPR 2018]
[Liang et al., HighMMT: Quantifying Modality & Interaction Heterogeneity for High-Modality Learning. TMLR 2022]

# Heterogeneity-aware Fusion

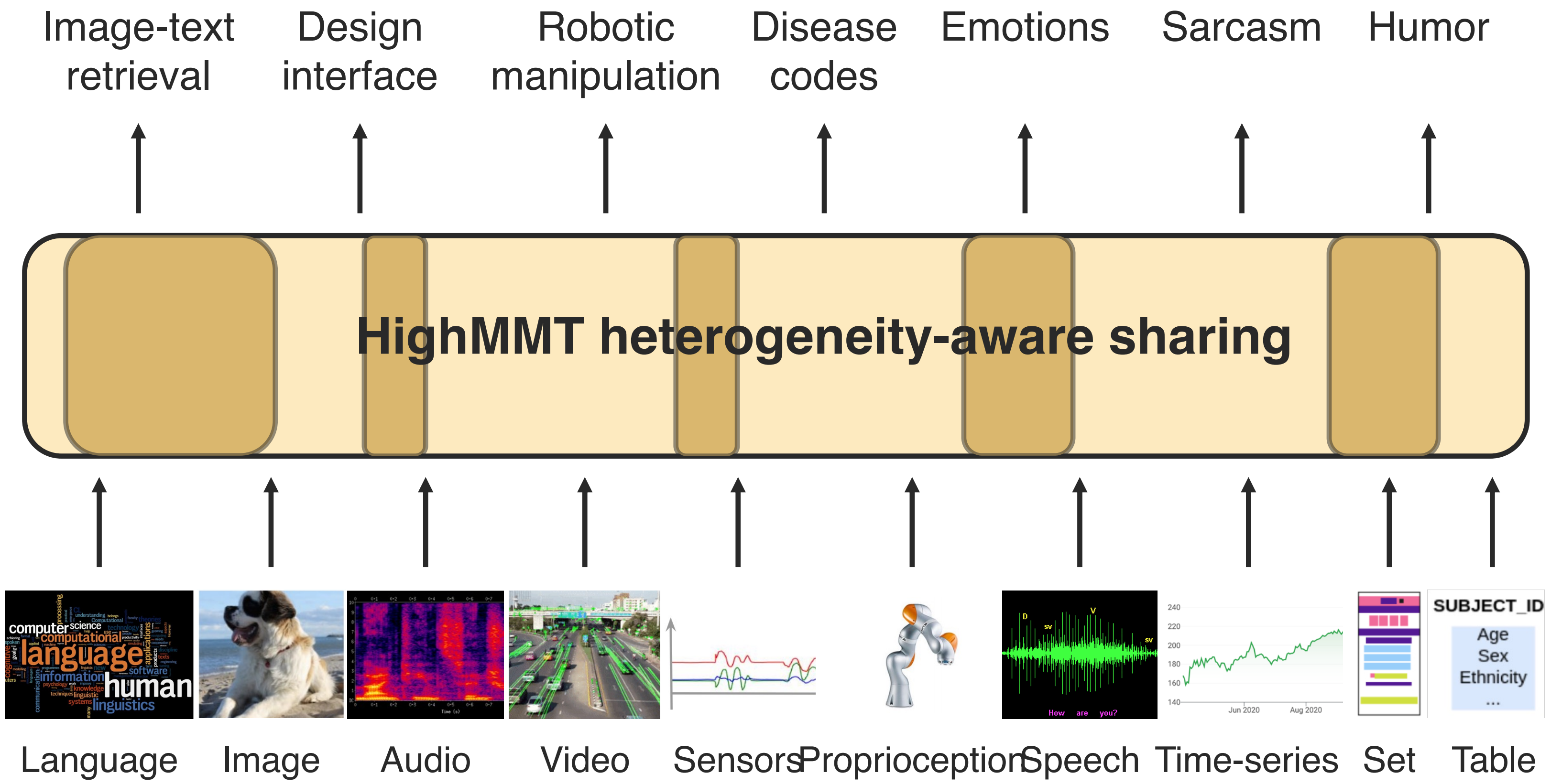
## Information transfer, transfer learning perspective

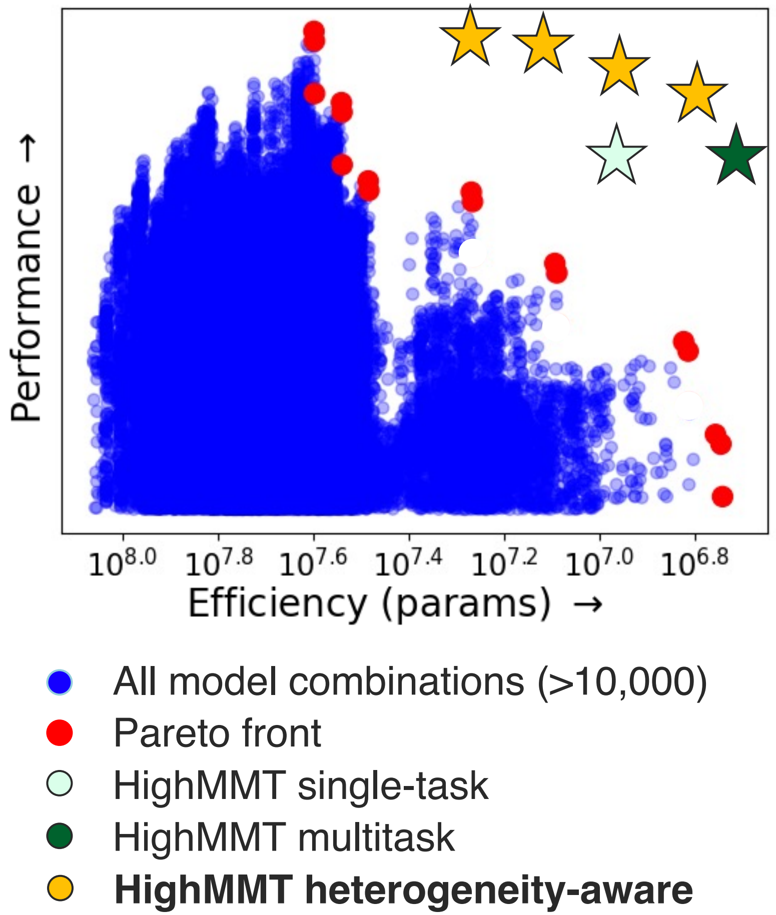



[Liang et al., HighMMT: Quantifying Modality & Interaction Heterogeneity for High-Modality Learning. TMLR 2022]

# Quantifying Modality Heterogeneity

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



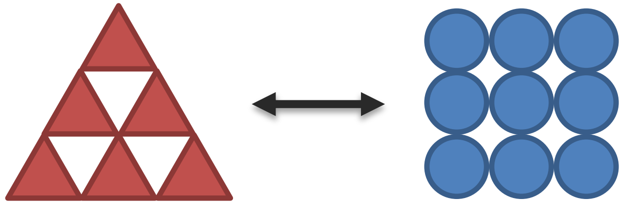

[Liang et al., HighMMT: Quantifying Modality & Interaction Heterogeneity for High-Modality Learning. TMLR 2022]

# Challenges: Quantifying Heterogeneity

Open challenges

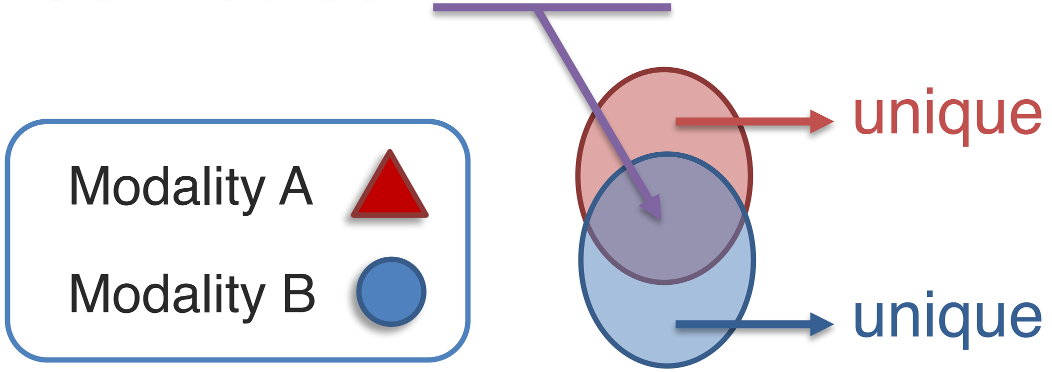
**Open challenges:**
- Noisy and missing modalities.
- New and understudied modalities.
- Large number of modalities.
- Cases where its unclear which modalities are useful – active selection

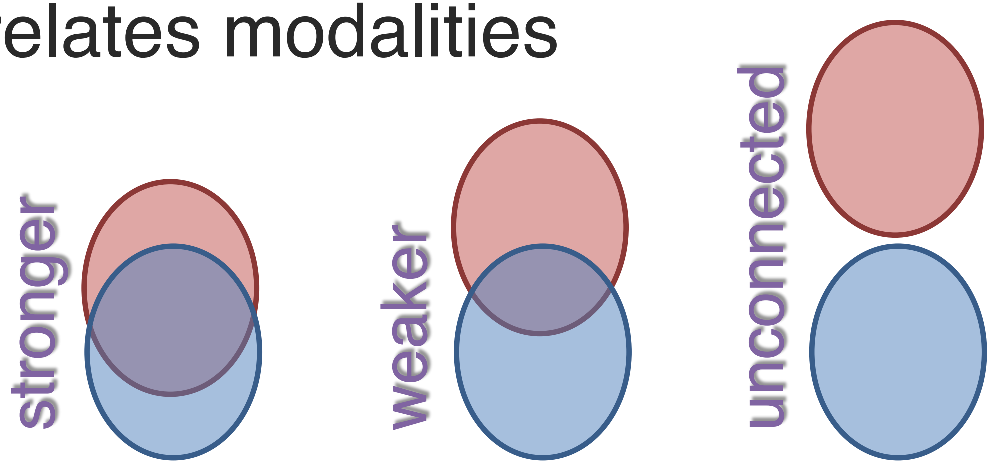
- Related fields: federated learning, active learning, distributed systems, structure & invariances

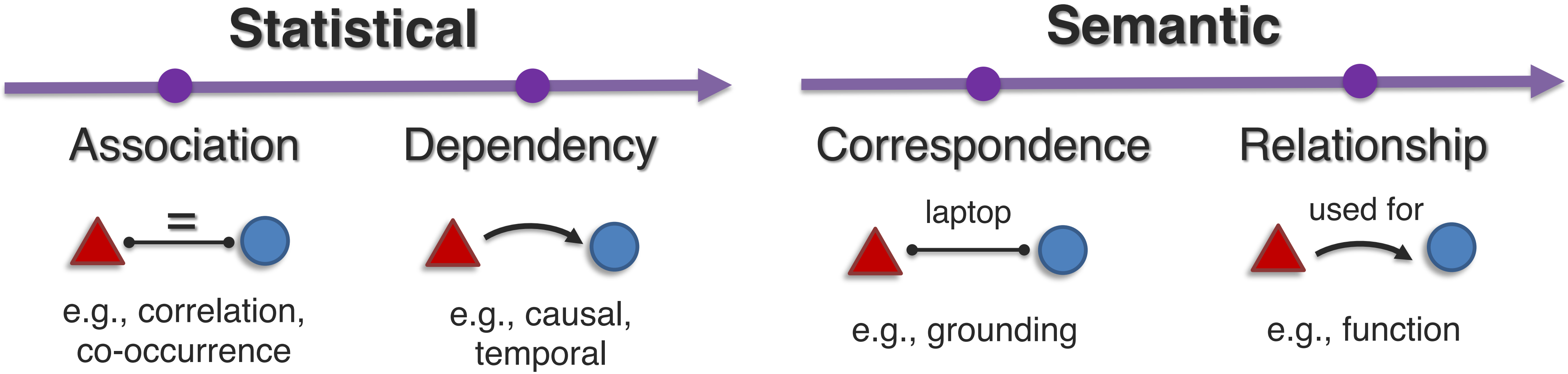


# Sub-Challenge 6b: Cross-modal Connections

Connected: Shared information that relates modalities

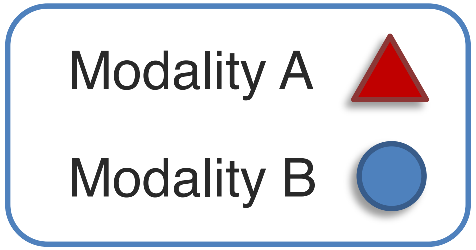


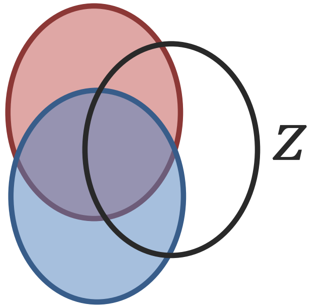



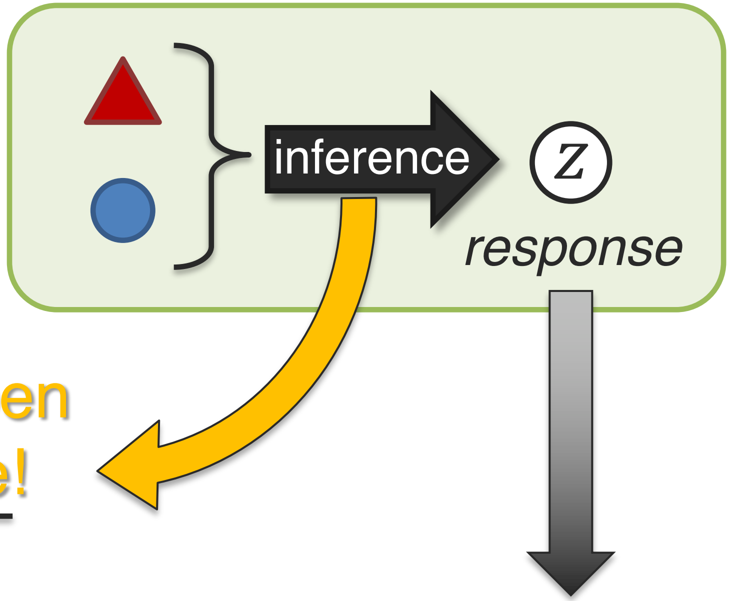



# Sub-Challenge 6b: Cross-modal Interactions

**Interacting:** process affecting each modality, creating new response



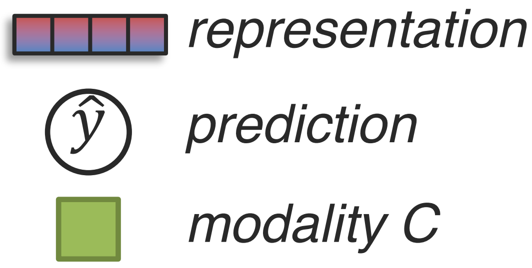




Interactions happen during inference!

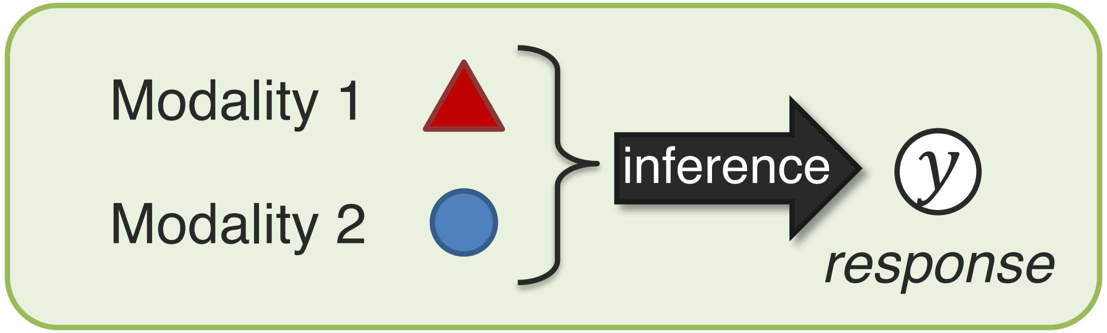
"Inference" examples:

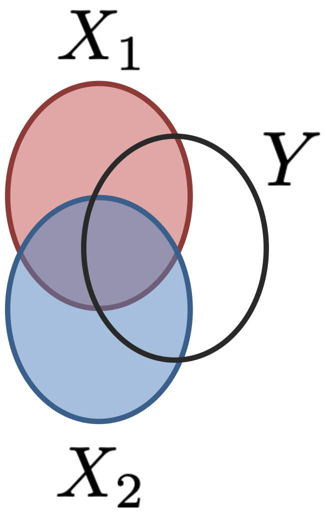
- Representation fusion
- Prediction task
- Modality translation



# Part 1: Multimodal Interactions

**Interactions:** Understanding *commonalities* between modalities and how they *combine* to provide information for a task.

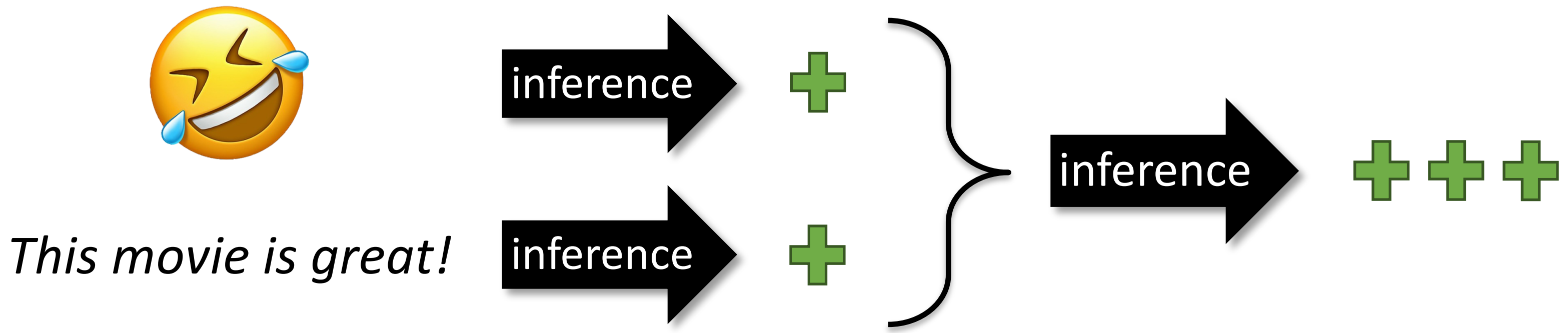


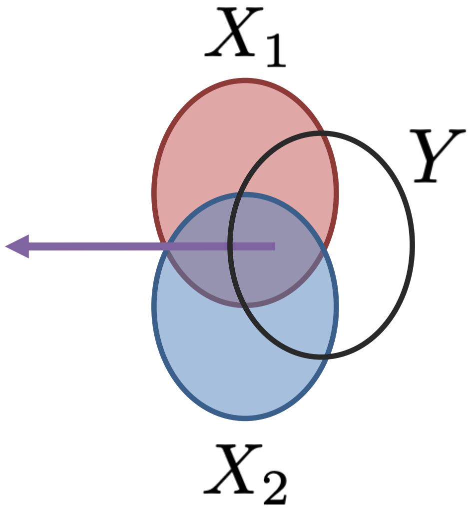

[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. ICML 2023, CVPR 2022, NAACL 2022 Tutorials]

# Multimodal Interactions

**Interactions:** Understanding *commonalities* between modalities and how they *combine* to provide information for a task.



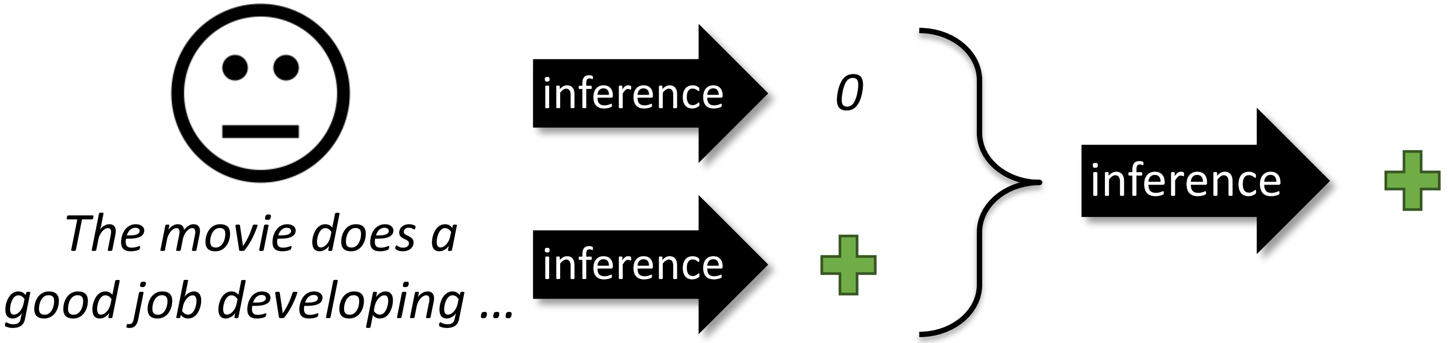

Redundancy: Shared by both modalities and task

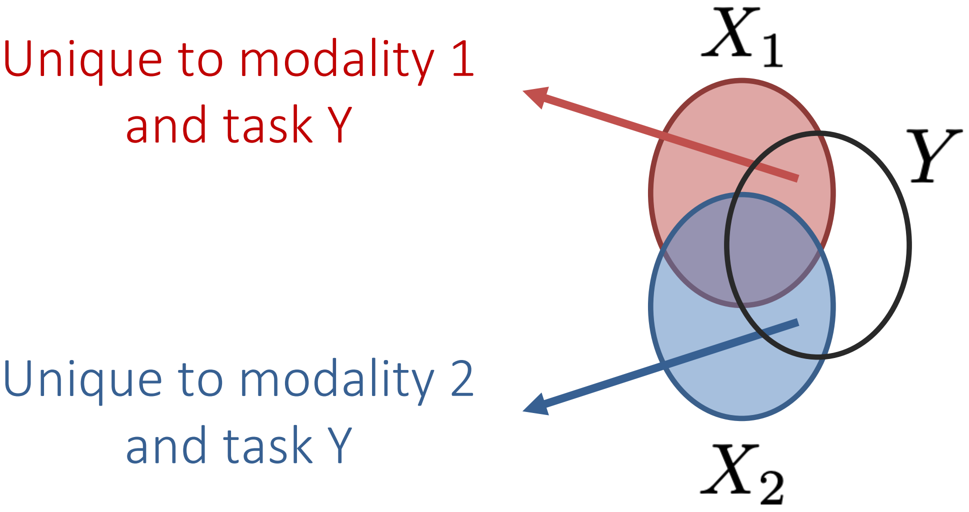



[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. ICML 2023, CVPR 2022, NAACL 2022 Tutorials]

# Multimodal Interactions

**Interactions:** Understanding *commonalities* between modalities and how they *combine* to provide information for a task.

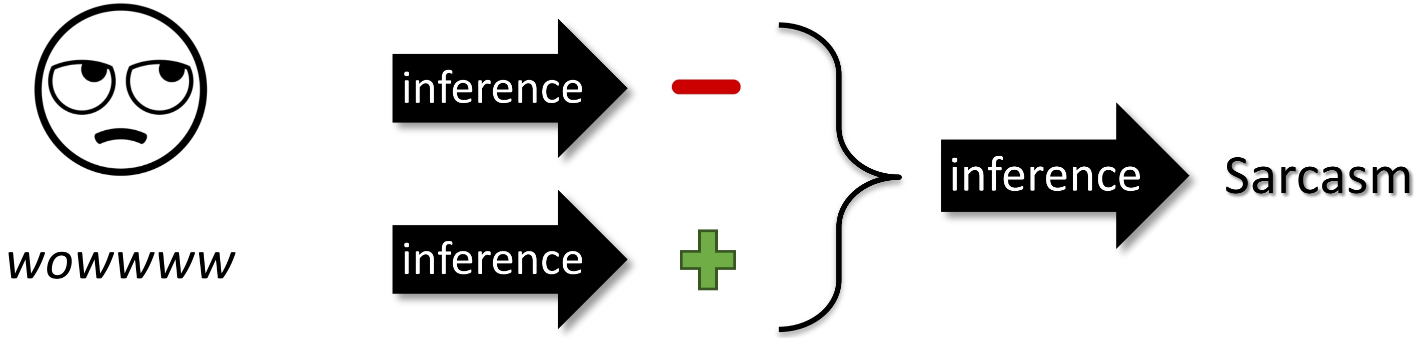


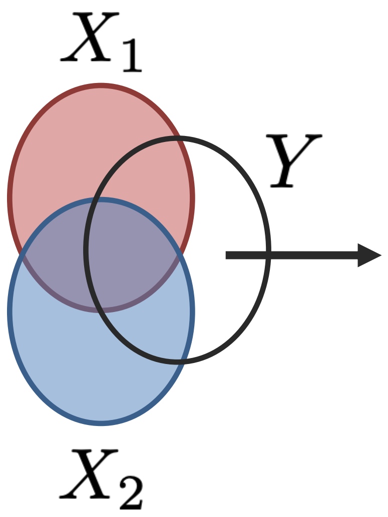




[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. ICML 2023, CVPR 2022, NAACL 2022 Tutorials]

# Multimodal Interactions

**Interactions:** Understanding *commonalities* between modalities and how they *combine* to provide information for a task.




Synergy: Emerging information from multimodal interaction

[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. ICML 2023, CVPR 2022, NAACL 2022 Tutorials]

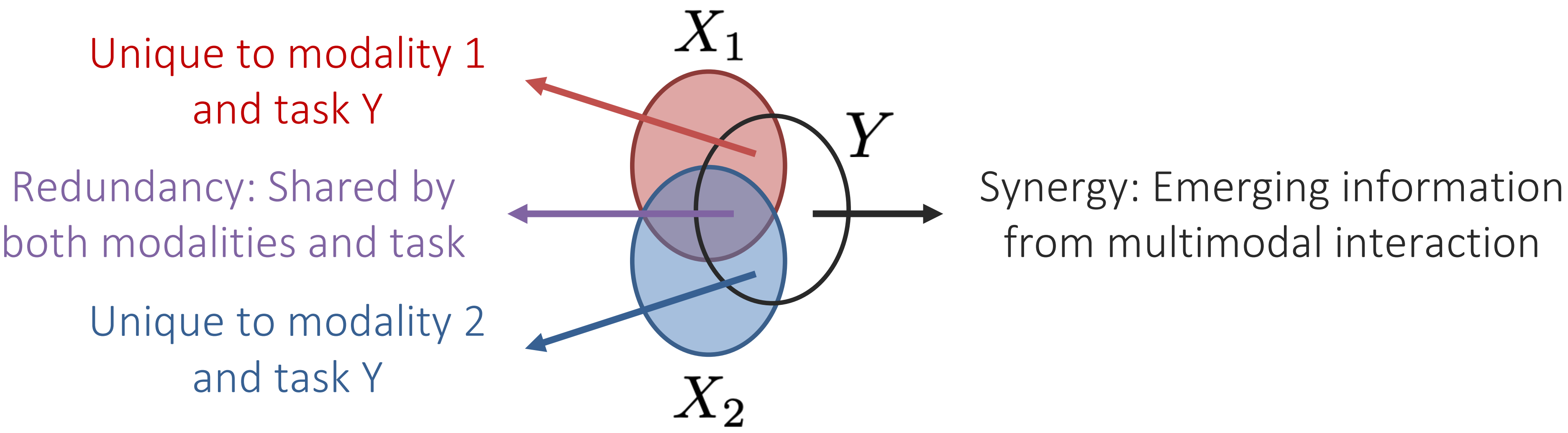
# Quantifying Multimodal Interactions

**Fundamental questions in multimodal learning**

*What interactions are in my data?*

*What interactions do models learn?*

*What models are suitable for my data?*




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

# Mathematical Framework for Multimodal Interactions

**Estimating synergy**

Unimodal marginal 1

| Verbal | Sarcasm |
|---|---|
| − | Yes/No |
| + | Yes/No |

Original multimodal distribution *p*

| Verbal | Nonverbal | Sarcasm |
|---|---|---|
| − | − | No |
| − | + | Yes |
| + | − | Yes |
| + | + | No |

Synergy = Original multimodal information about the task
− multimodal information given by the ***worst*** distribution combining the same modalities

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

# Mathematical Framework for Multimodal Interactions

**Estimating synergy**

Original multimodal distribution *p*

| Verbal | Nonverbal | Sarcasm |
|---|---|---|
| − | − | No |
| − | + | Yes |
| + | − | Yes |
| + | + | No |

Unimodal marginal 1

| Verbal | Sarcasm |
|---|---|
| − | Yes/No |
| + | Yes/No |

Unimodal marginal 2

| Nonverbal | Sarcasm |
|---|---|
| − | Yes/No |
| + | Yes/No |

Synergy = Original multimodal information about the task
– multimodal information given by the ***worst*** distribution combining the same modalities

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

# Mathematical Framework for Multimodal Interactions

**Many ways of combining these 2 unimodal marginals into a multimodal distribution!**

Unimodal marginal 1

| Verbal | Sarcasm |
|---|---|
| − | Yes/No |
| + | Yes/No |

Original multimodal distribution *p*

| Verbal | Nonverbal | Sarcasm |
|---|---|---|
| − | − | No |
| − | + | Yes |
| + | − | Yes |
| + | + | No |

Unimodal marginal 2

| Nonverbal | Sarcasm |
|---|---|
| − | Yes/No |
| + | Yes/No |

Multimodal is **very** informative about task $I_p(\{X_2, X_2\}; Y) = 1$

Synergy = Original multimodal information about the task
− multimodal information given by the ***worst*** distribution combining the same modalities

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

# Mathematical Framework for Multimodal Interactions

**Many ways of combining these 2 unimodal marginals into a multimodal distribution!**

Unimodal marginal 1

| Verbal | Sarcasm |
|---|---|
| − | Yes/No |
| + | Yes/No |

Another distribution *q* combining modalities

| Verbal | Nonverbal | Sarcasm |
|---|---|---|
| − | − | Yes/No |
| − | + | Yes/No |
| + | − | Yes/No |
| + | + | Yes/No |

Unimodal marginal 2

| Nonverbal | Sarcasm |
|---|---|
| − | Yes/No |
| + | Yes/No |

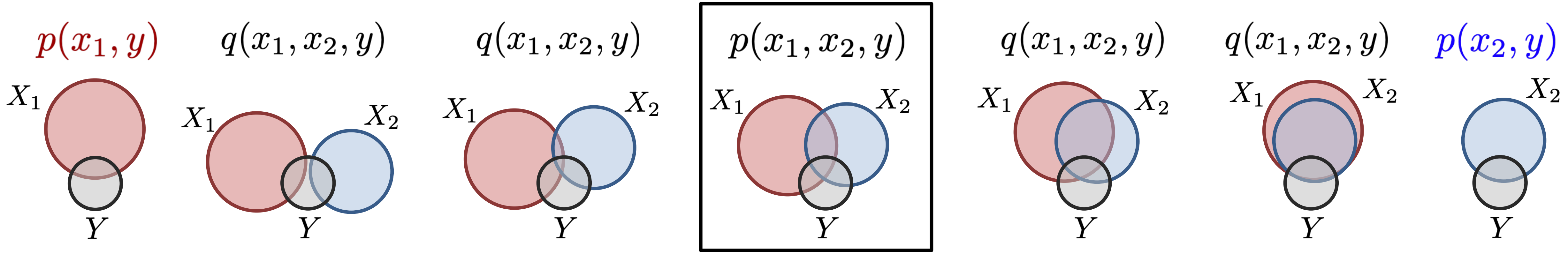
Multimodal is **less** informative about task $I_q(\{X_2, X_2\}; Y) = 0$

Synergy = Original multimodal information about the task
− multimodal information given by the ***worst*** distribution combining modalities = 1 – 0 = **1**

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

# Mathematical Framework for Interactions

**More formally as partial information decomposition:** [Bertschinger et al., 2014]




*q* must be a ***coupling*** of the unimodal marginals:

$$\Delta_p = \{q(x_1, x_2, y) : q(x_1, y) = p(x_1, y), q(x_2, y) = p(x_2, y)\}$$

$$\boxed{S = \underbrace{I_p(X_1, X_2; Y)}_{\text{Task-relevant multimodal info}} - \underbrace{\min_{q \in \Delta_p} I_q(X_1, X_2; Y)}_{\text{Task-relevant multimodal info without synergy:}}}$$

$$S_{q^*} = I_{q^*}(X_1, X_2; Y) - \min_{q \in \Delta_p} I_q(X_1, X_2; Y) = 0$$

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

# Estimating Partial Information Decomposition

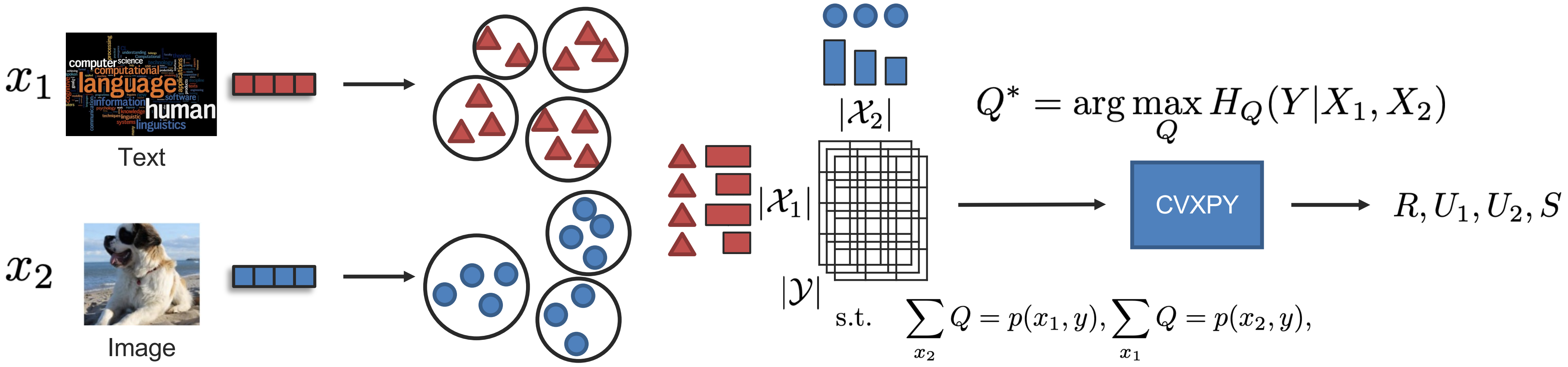
Equivalent formulation as max-entropy optimization:

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

$$\Delta_p = \{q(x_1, x_2, y) : q(x_1, y) = p(x_1, y), q(x_2, y) = p(x_2, y)\}$$

If $X_1, X_2, Y$ have small discrete support: exact solution via convex programming.

**Clustering** **Convex programming with linear constraints**




$$Q^* = \arg\max_{Q} H_Q(Y|X_1, X_2)$$

$$\text{s.t.} \quad \sum_{x_2} Q = p(x_1, y), \sum_{x_1} Q = p(x_2, y),$$

$$Q \geq 0, \sum_{x_1, x_2, y} Q = 1.$$

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

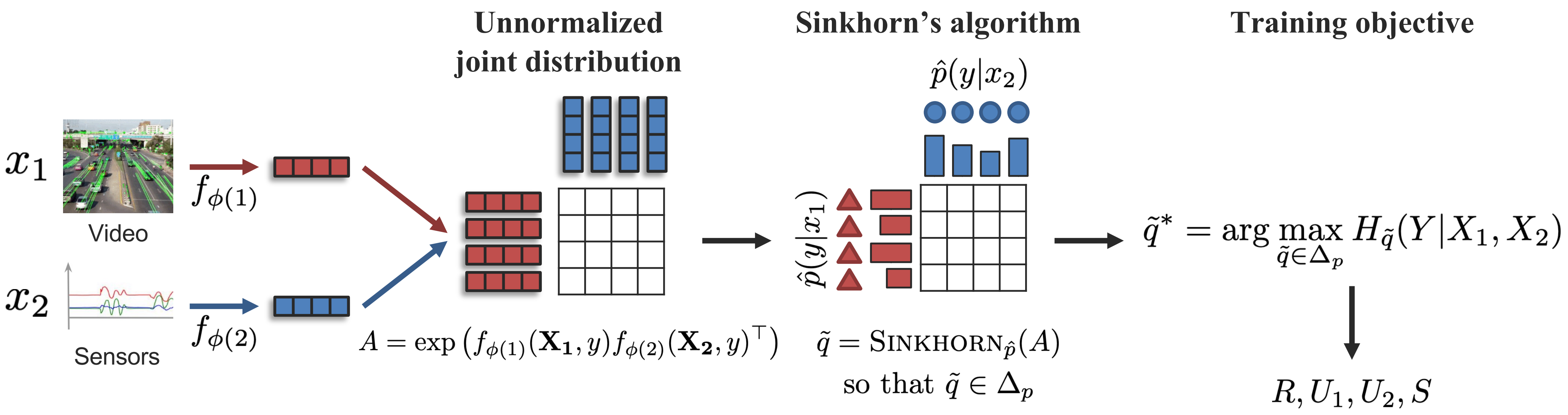
# Estimating Partial Information Decomposition

Equivalent formulation as max-entropy optimization:

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

$$\Delta_p = \{q(x_1, x_2, y) : q(x_1, y) = p(x_1, y), q(x_2, y) = p(x_2, y)\}$$

If $X_1, X_2, Y$ high-dimensional & continuous: an approximate neural network estimator.




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

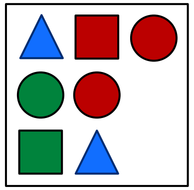
# Quantifying Multimodal Datasets

## 1. Dataset quantification:

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


**Sheldon** :
Its just a *privilege* to watch your mind at work.

- **Text** : suggests a compliment.
- **Audio** : neutral tone.
- **Video** : straight face.



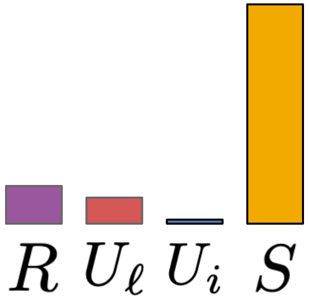
*Is there a red shape above a circle?*

Sentiment

Sarcasm

VQA





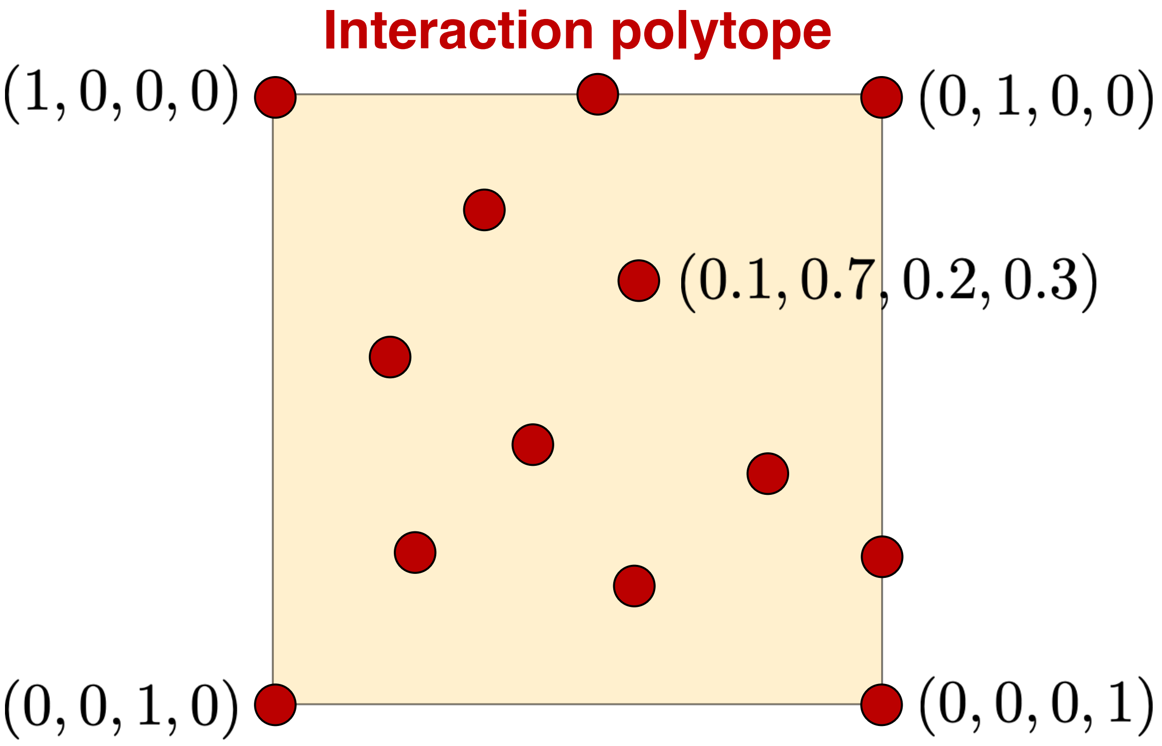

Also matches human judgment of interactions, and other sanity checks on synthetic datasets

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

# Model Selection

**1. Dataset quantification:**

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




Can be done with synthetic data

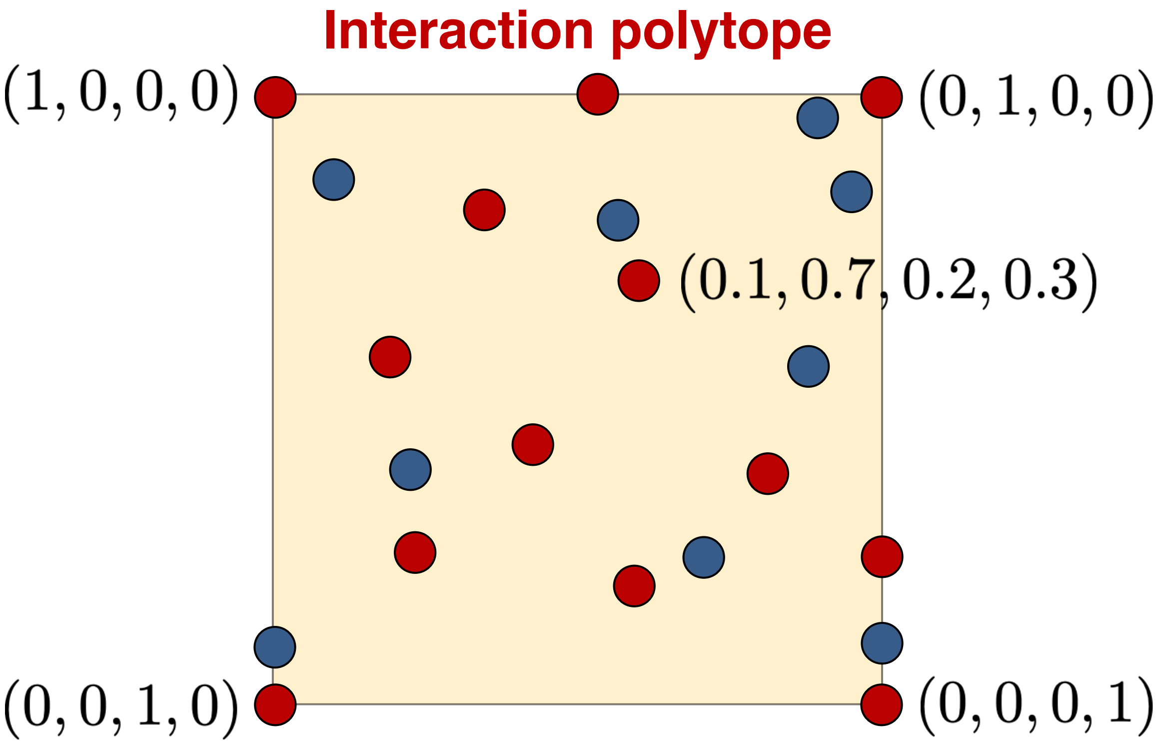
[Liang et al., Quantifying & Modeling Multimodal Interactions: An Information Decomposition Framework. NeurIPS 2023]

# Model Selection

## 2. Model quantification:

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

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




Model families trained on synthetic data

- Unimodal models
- Ensemble
- Multiplicative interactions
- and many more…

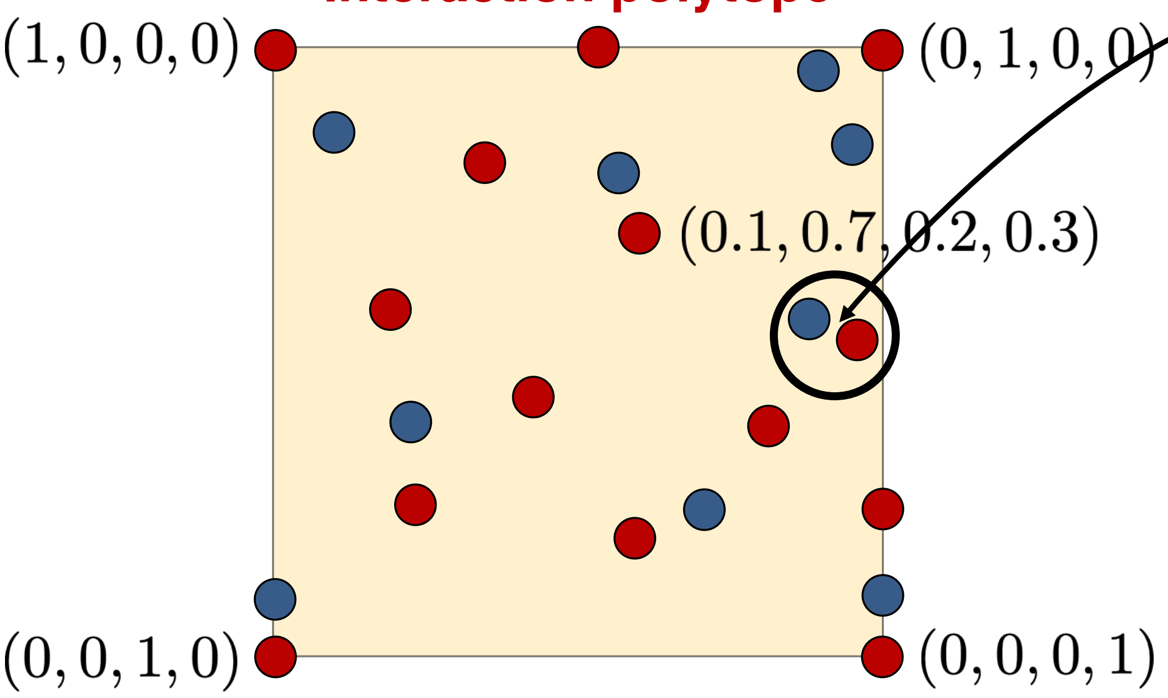
[Liang et al., Quantifying & Modeling Multimodal Interactions: An Information Decomposition Framework. NeurIPS 2023]

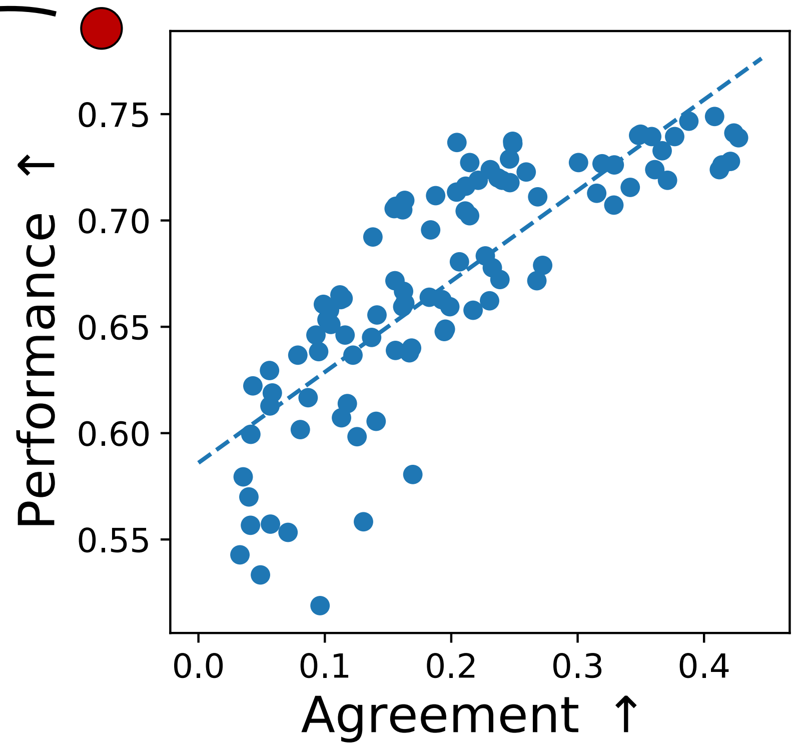
# Model Selection

## 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 Multimodal Interactions: An Information Decomposition Framework. NeurIPS 2023]

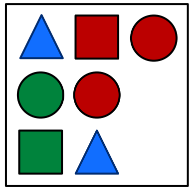
# Model Selection

## 3. Model selection:

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


**Sheldon** :
Its just a *privilege* to watch your mind at work.

- **Text** : suggests a compliment.
- **Audio** : neutral tone.
- **Video** : straight face.



*Is there a red shape above a circle?*

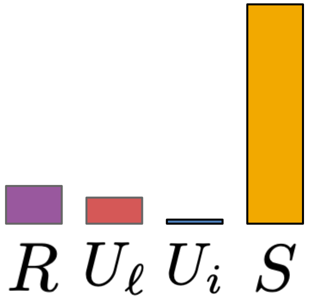
Sentiment


Language/Agreement

Sarcasm


Multimodal Transformer

VQA

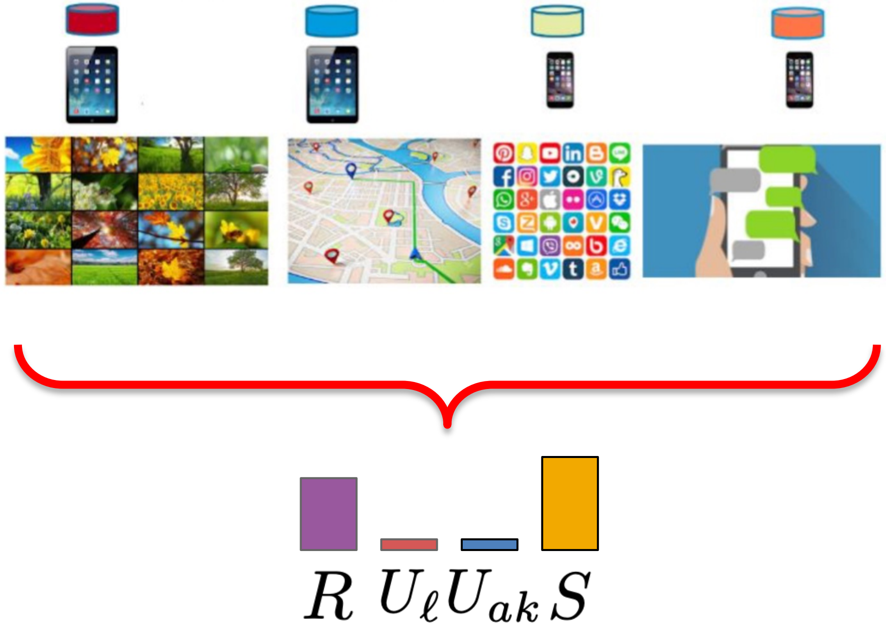



Multiplicative/Transformer

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

# Application 1: Mental Health

**Daily mood prediction as a stepping-stone towards real-time assessment of suicide ideation.**




Text + app + keystroke interactions

| Slower implies positive | Faster implies positive |
|---|---|
| just | why, thank, haha |
| next, was, into, people | making, work, idk |
| stuff, cute, phone, want, talk, see | they, send, dont, man, going |
| don't, talk | think, you, all, love |

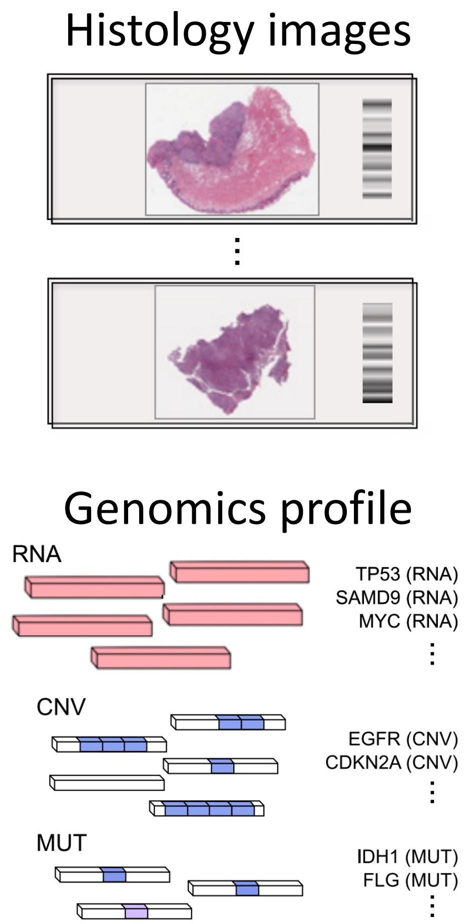
+ words like 'love', 'thanks', 'haha' become more positive when typed faster

- words like 'don't, 'just' become more negative when typed faster

[Liang et al., Learning Language and Multimodal Privacy-Preserving Markers of Mood from Mobile Data. ACL 2021]

# Application 2: Computational Pathology

HARVARD MEDICAL SCHOOL · BRIGHAM AND WOMEN'S HOSPITAL · BROAD INSTITUTE





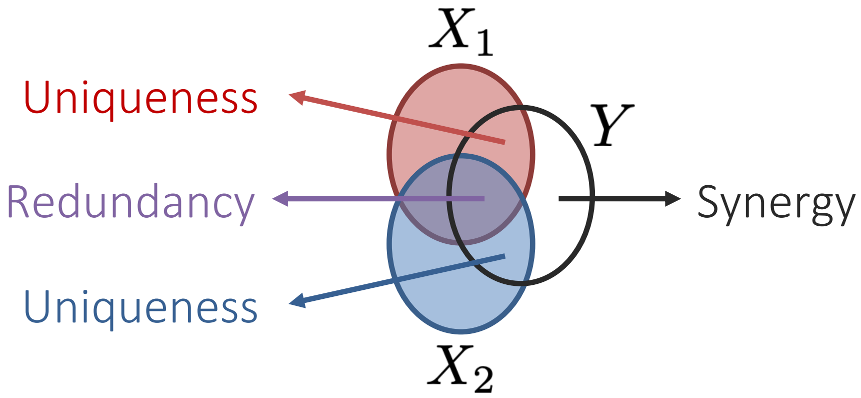
Glioma: Genomics unimodal


Pancreas: Histology + genomics interaction

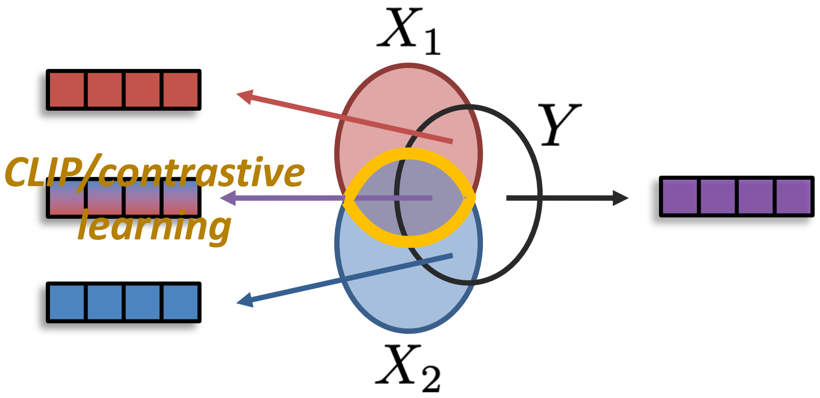
Understanding the models and adoption in practice by doctors

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

# Implications of Studying Multimodal Interactions

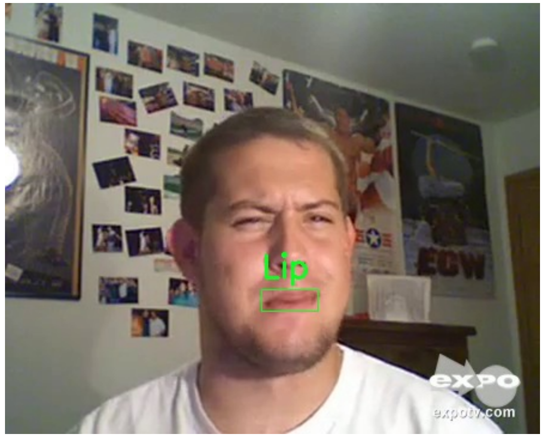



Optimizing these interactions as training objectives:




[Liang et al., FactorCL. NeurIPS 2023]

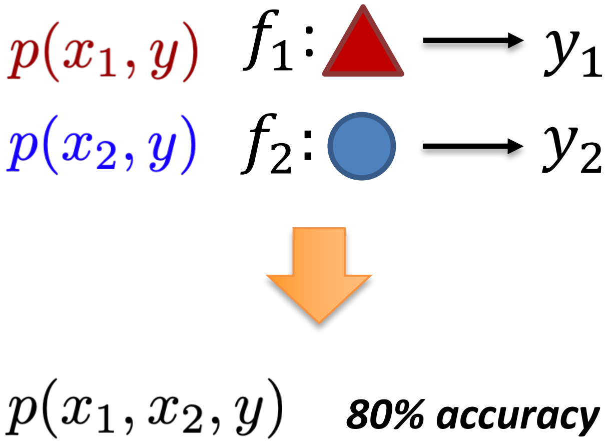
Visualizing the interactions learned in individual neurons:




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

[Liang et al., MultiViz. ICLR 2023]

Predicting multimodal performance to decide modality utility:



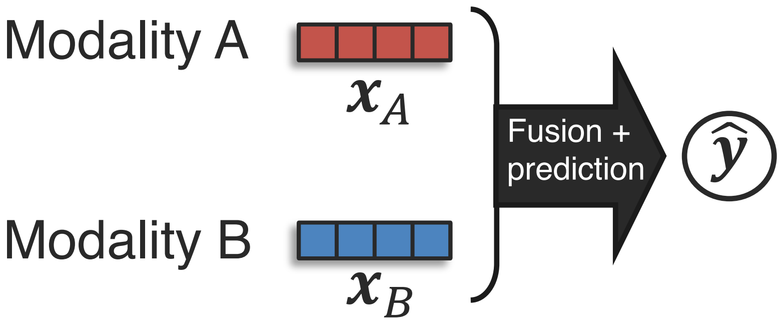

[Liang et al., Semi-supervised. arXiv 2023]

# 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.

$\mu$ 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 $\frac{\partial f^2}{\partial \boldsymbol{x}_A \partial \boldsymbol{x}_B} > 0.$

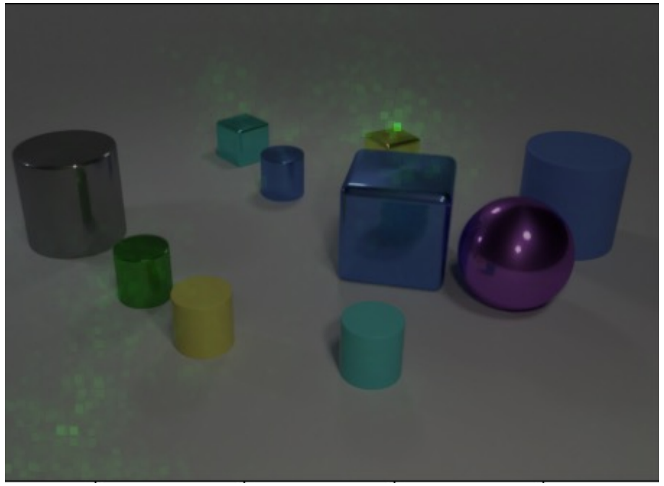
Natural second-order extension of gradient-based approaches!

[Liang et al., MultiViz: Visualizing and Understanding Multimodal Models. ICLR 2023]

# Quantifying Cross-modal Interactions

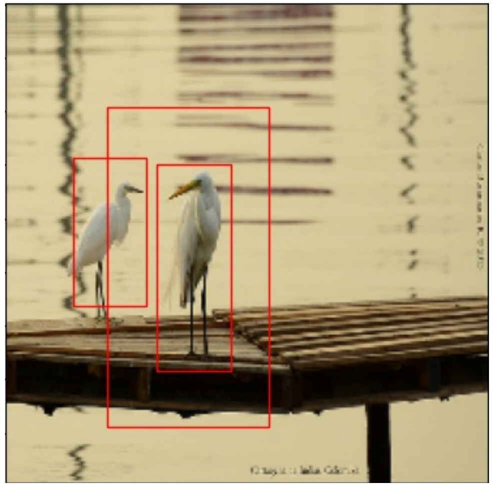
## Identifying individual cross-modal interactions

**CLEVR**



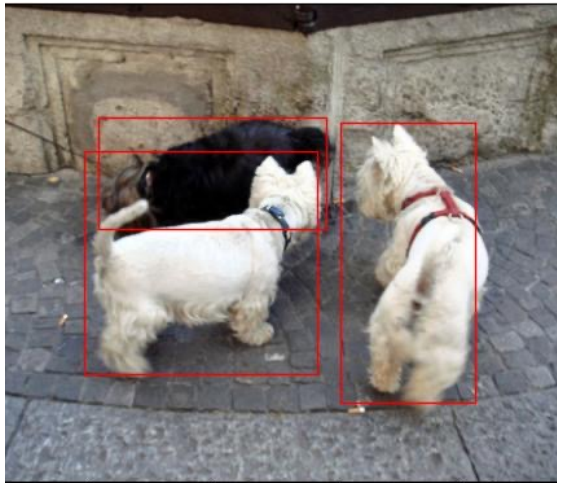
*The other small shiny thing that is the same shape as the **tiny yellow shiny object** is what color?*

**VQA 2.0**



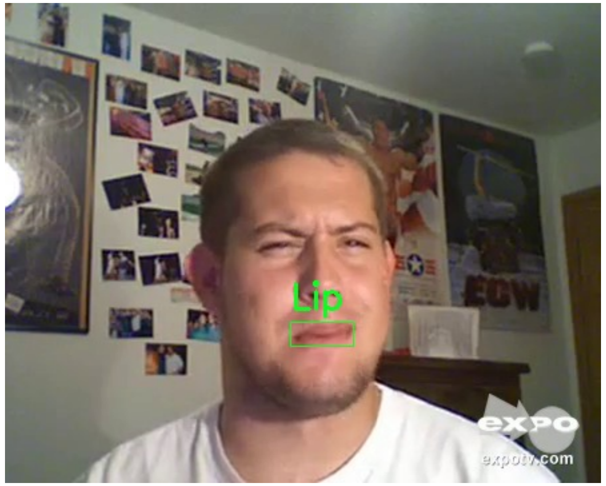
*How many **birds**?*

**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**…*

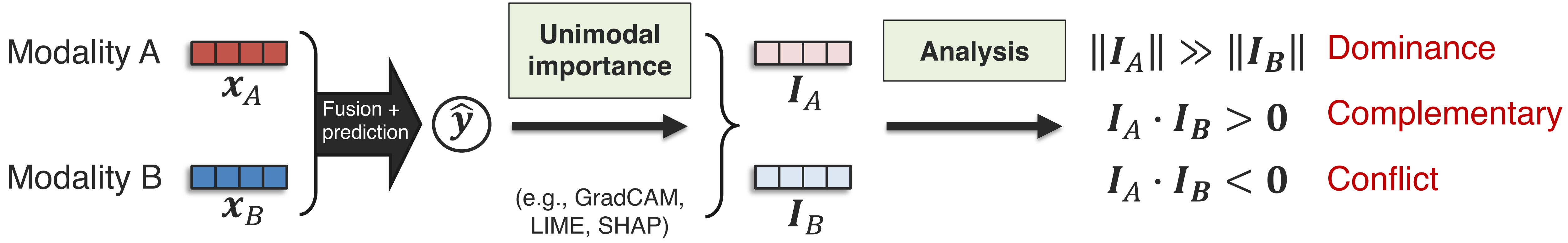
Correspondence

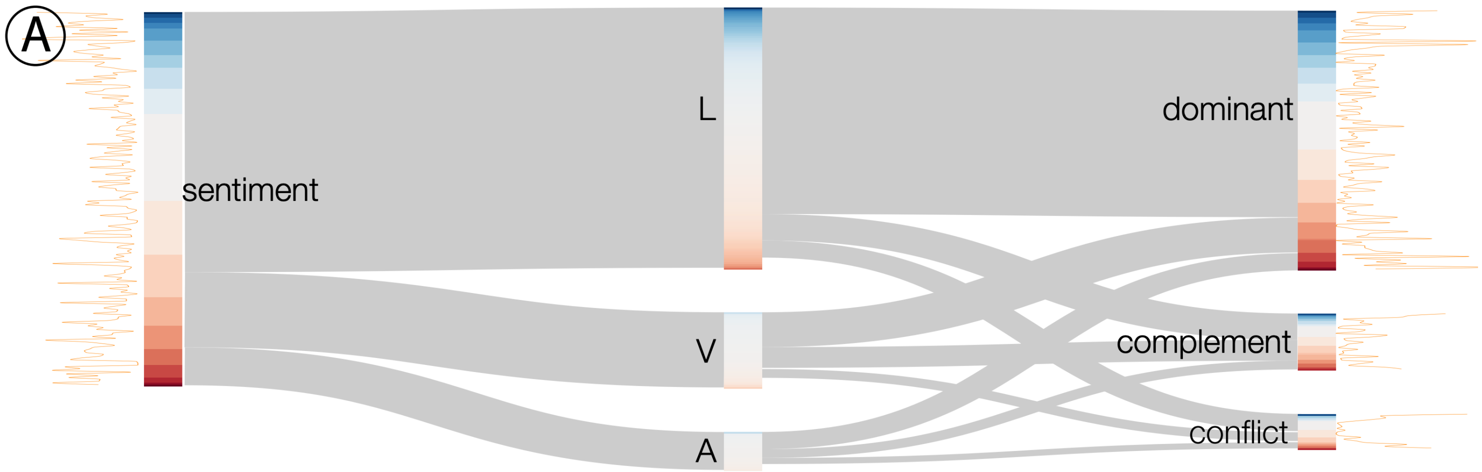
Relationships

[Liang et al., MultiViz: Visualizing and Understanding Multimodal Models. ICLR 2023]

# Quantifying Cross-modal Interactions

## Classification of cross-modal interactions



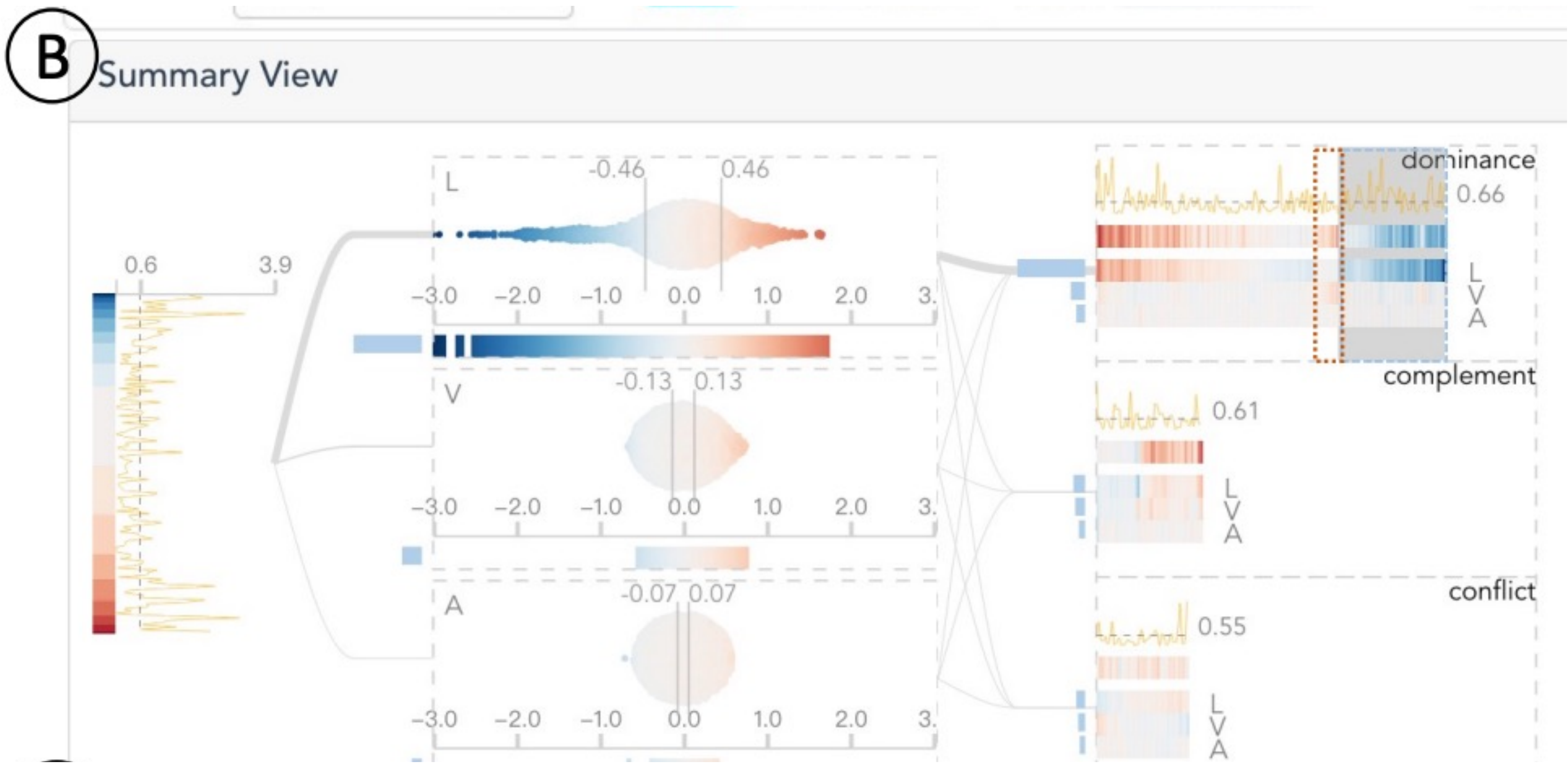

Language is often **dominant** in multimodal sentiment analysis

[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/](https://andy-xingbowang.com/m2lens/)



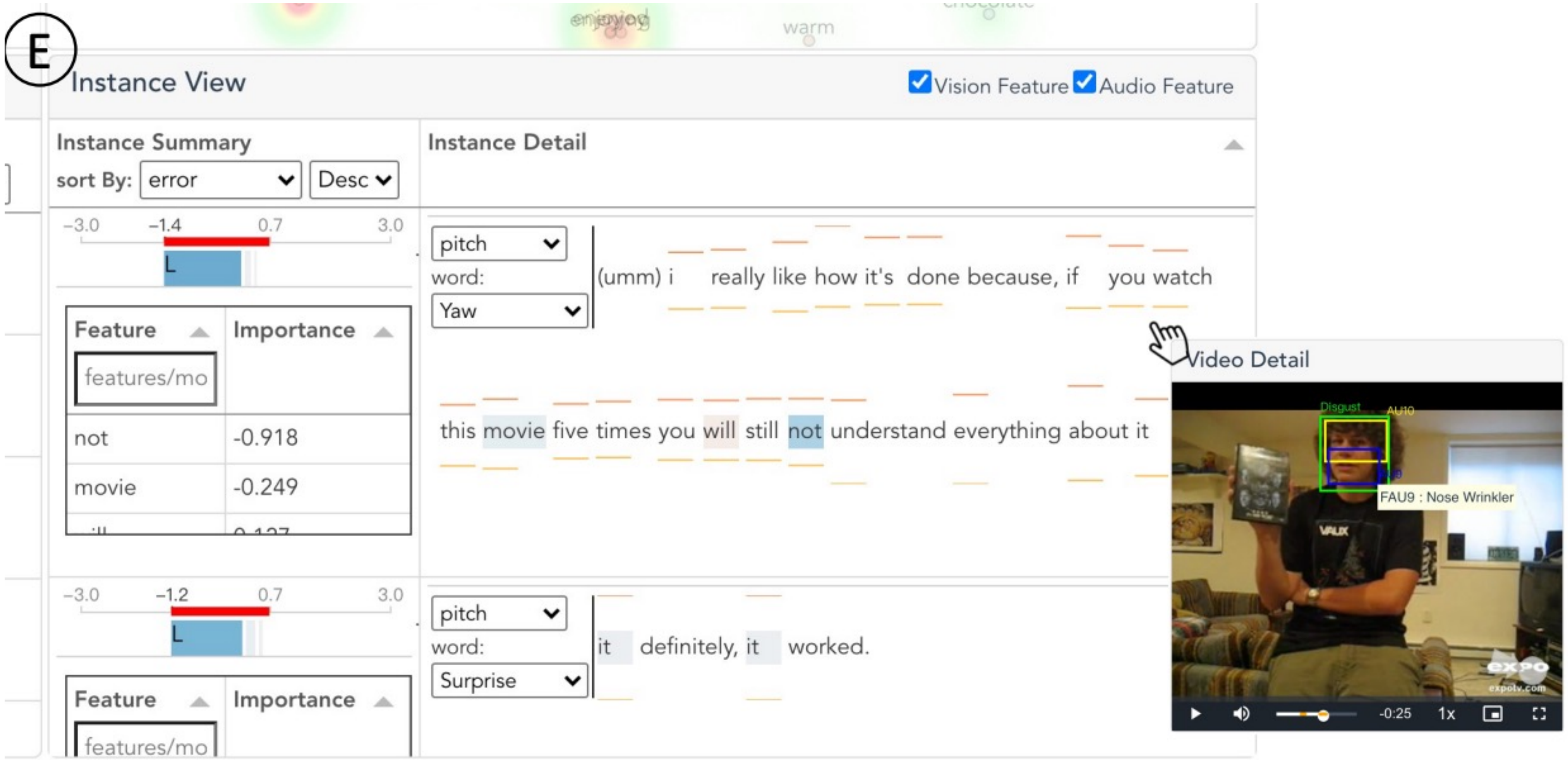

Summary of cross-modal interactions across entire dataset.

[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/](https://andy-xingbowang.com/m2lens/)

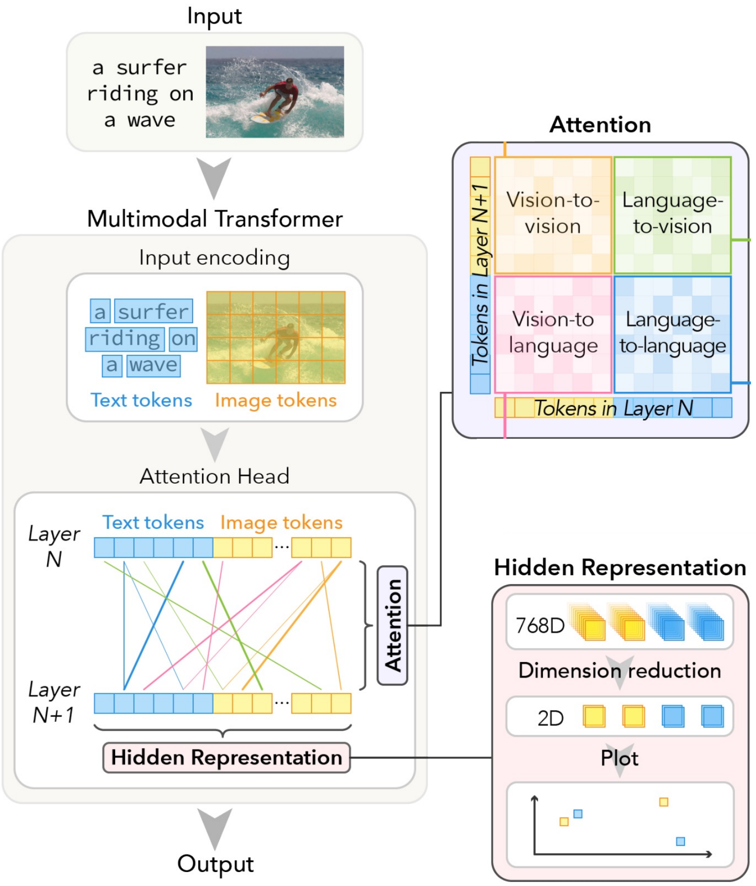



Summary of
cross-modal interactions
in a single instance.

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

# Quantifying Cross-modal Interactions

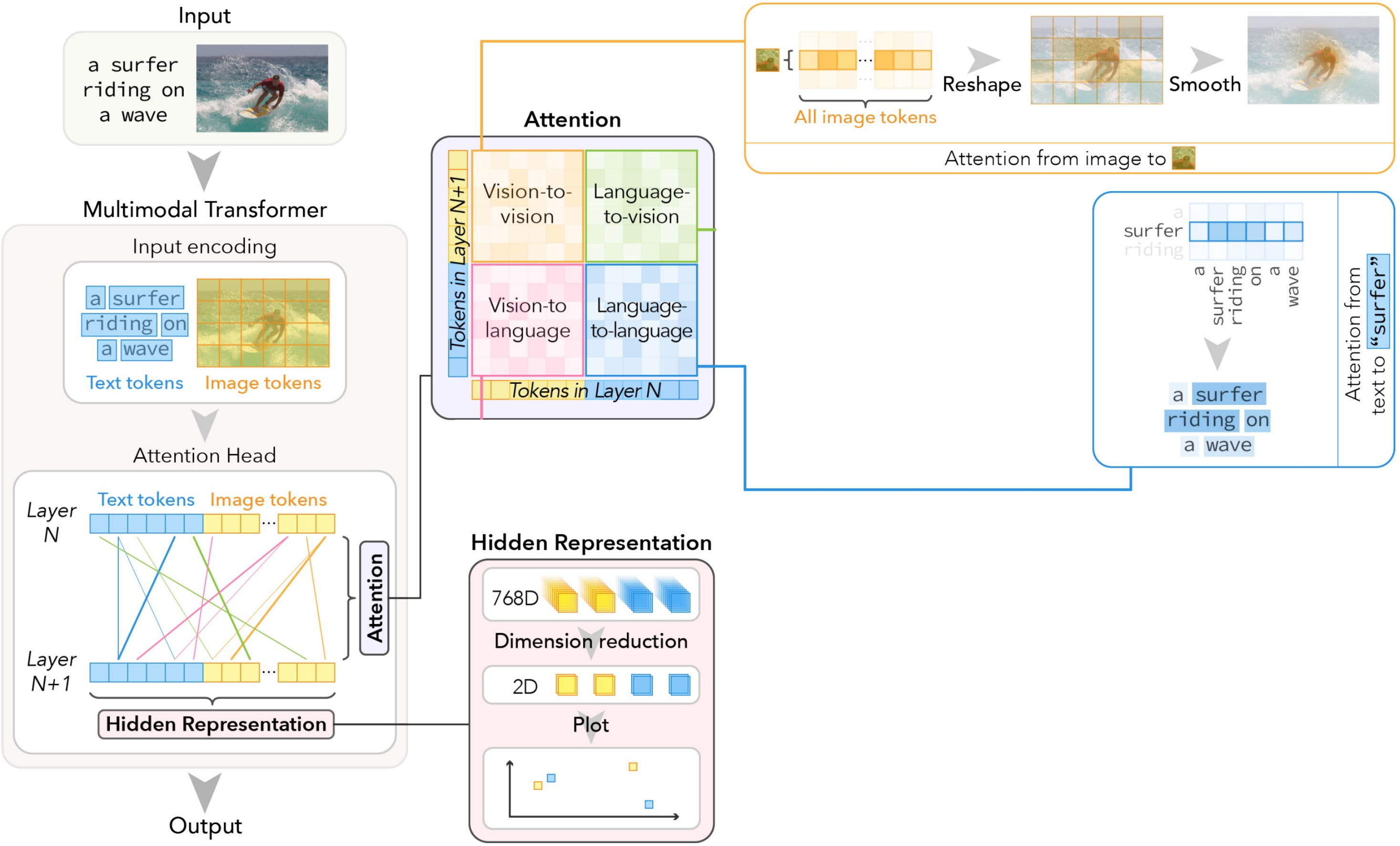
**Visualizing multimodal transformers** See interactive website: https://github.com/IntelLabs/VL-InterpreT




[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



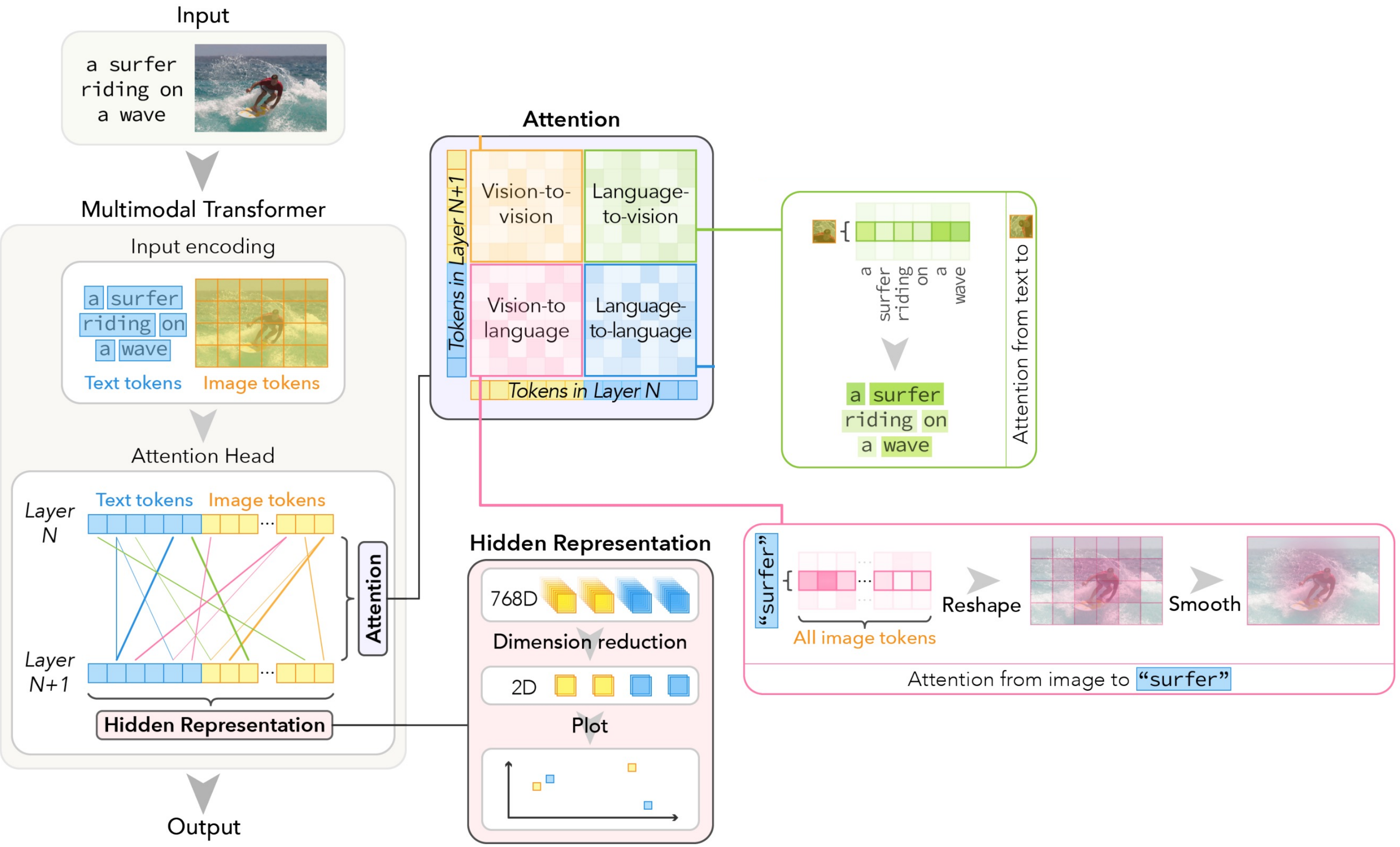

Unimodal image importance

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



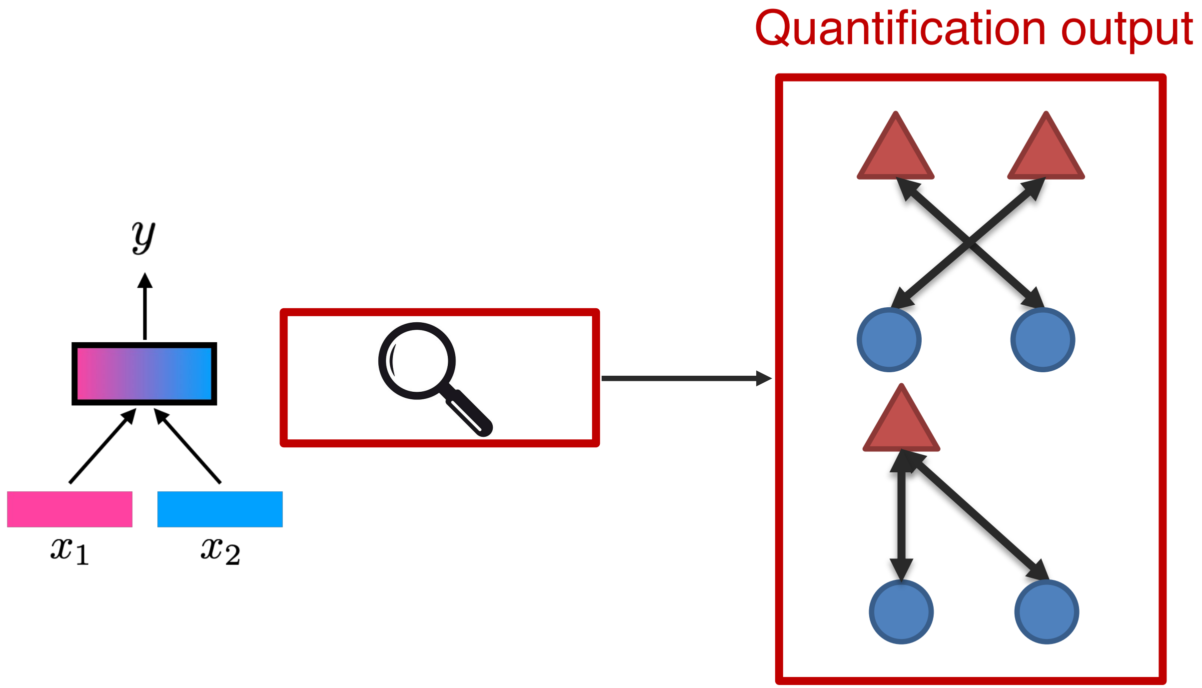

Correspondence and complementary interactions

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

# Evaluating Quantification

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

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

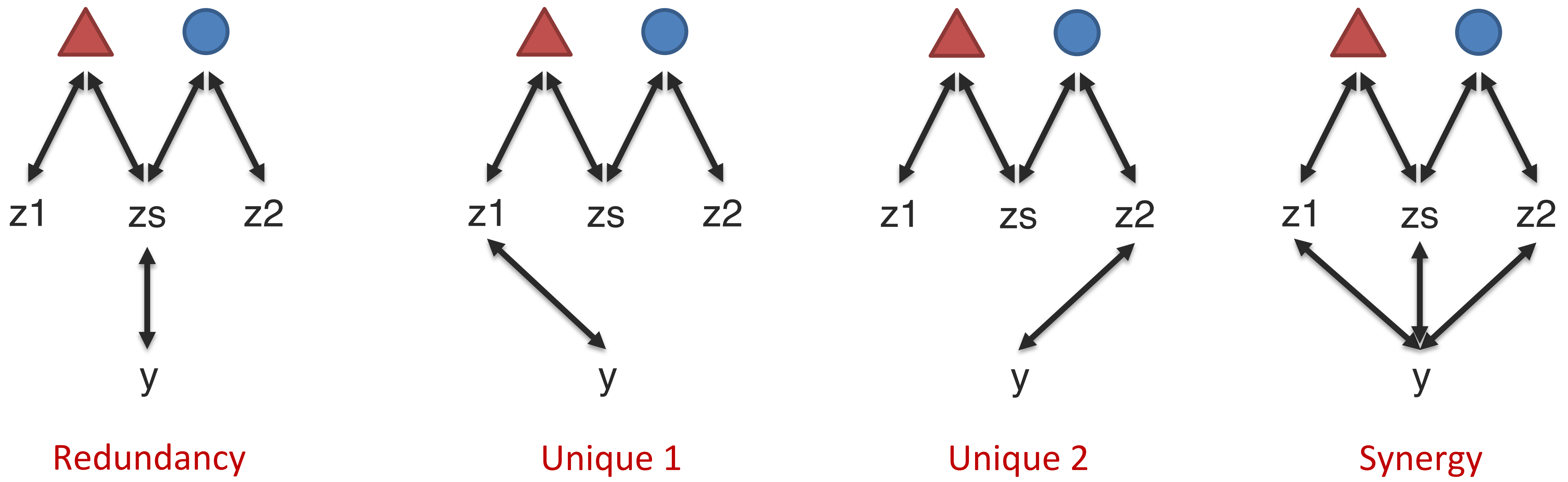



[Liang et al., MultiViz: Towards Visualizing and Understanding Multimodal Models. ICLR 2023]

# Directly Evaluating Quantification

**Direct evaluation: Create datasets for each tested quality, but limited to synthetic data**

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



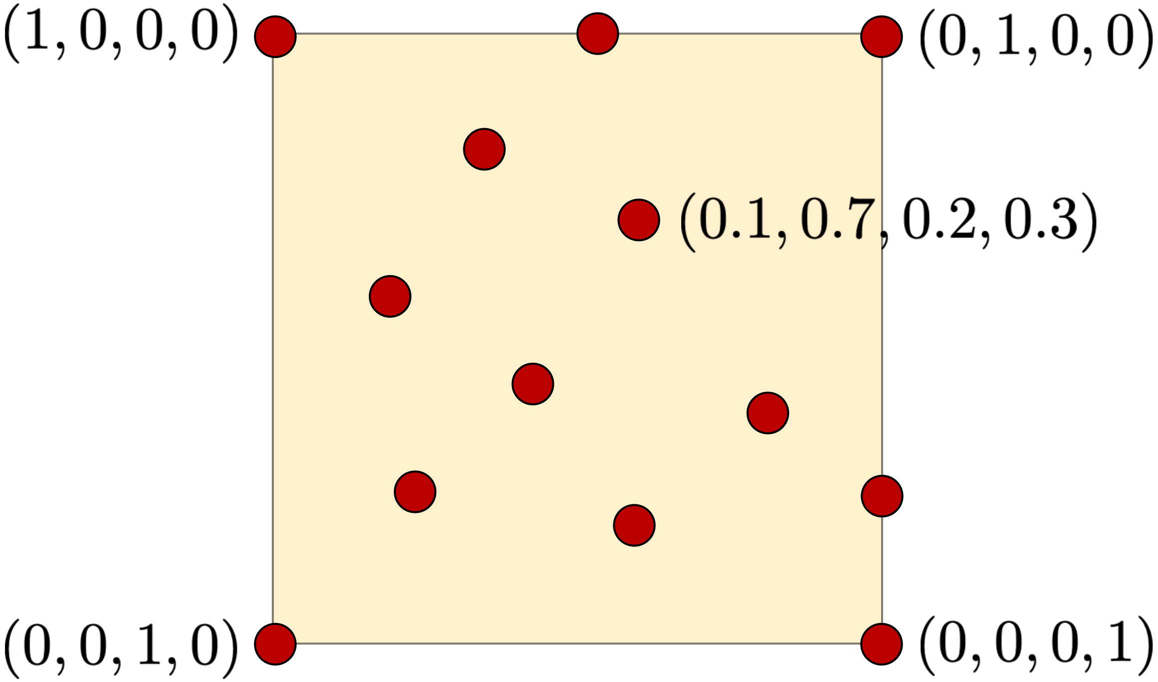
# Directly Evaluating Quantification

**Direct evaluation: Create datasets for each tested quality, but limited to synthetic data**

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

Can be done with synthetic data

**Interaction polytope**

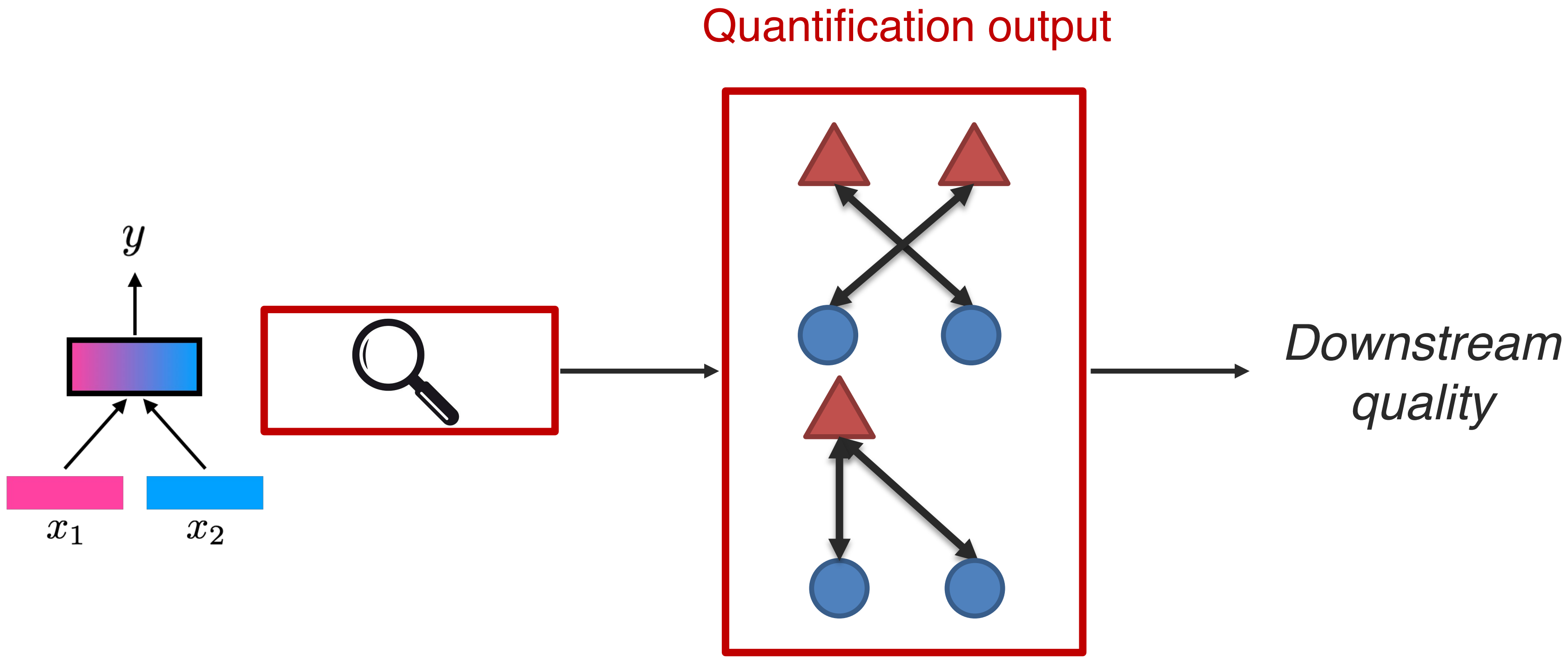



[Liang et al., MultiViz: Towards Visualizing and Understanding Multimodal Models. ICLR 2023]

# Indirectly Evaluating Quantification

## Indirect evaluation

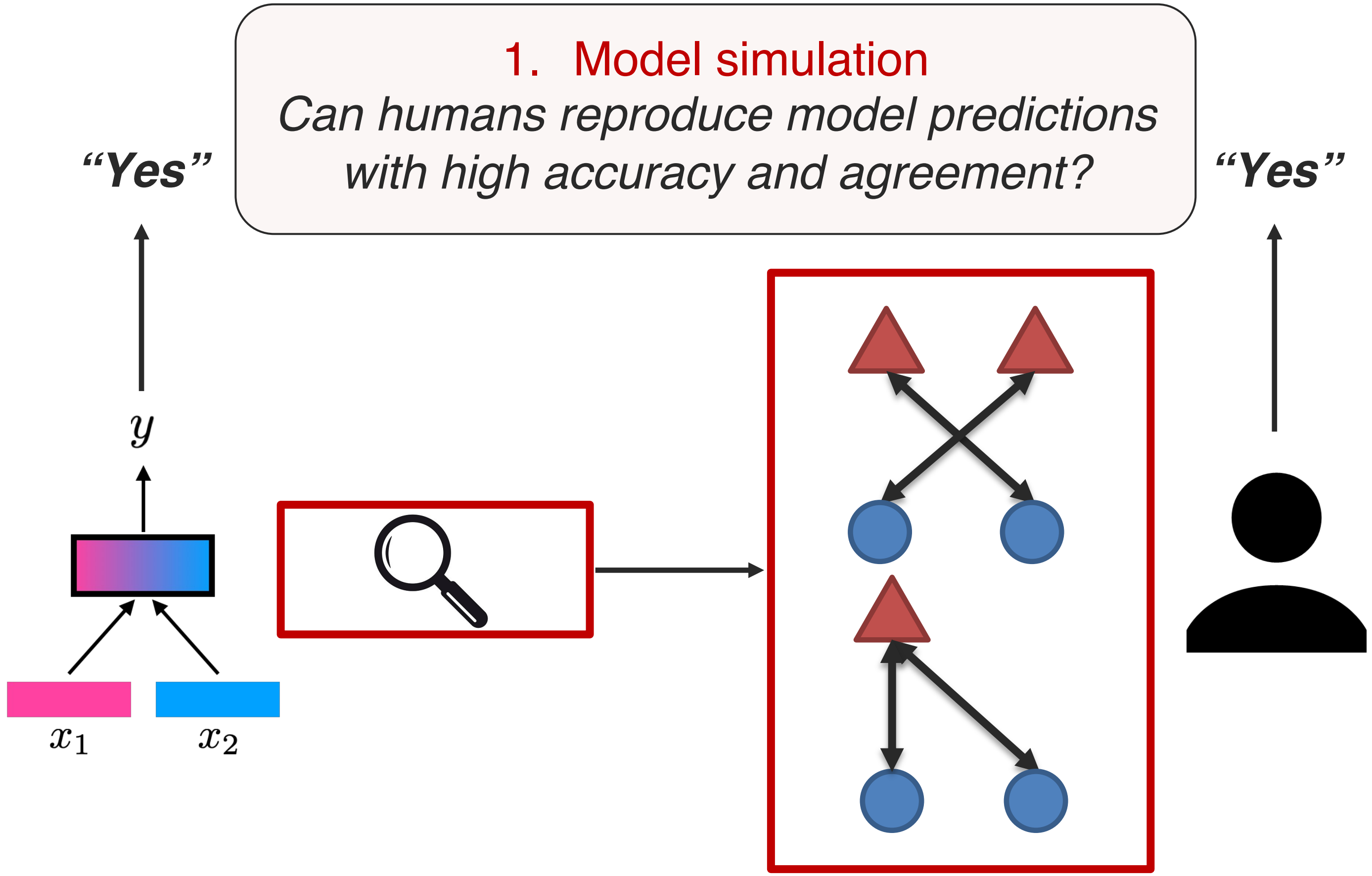
*Find some downstream quality that practitioners find useful and can be easily evaluated.*




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

# Indirectly Evaluating Quantification

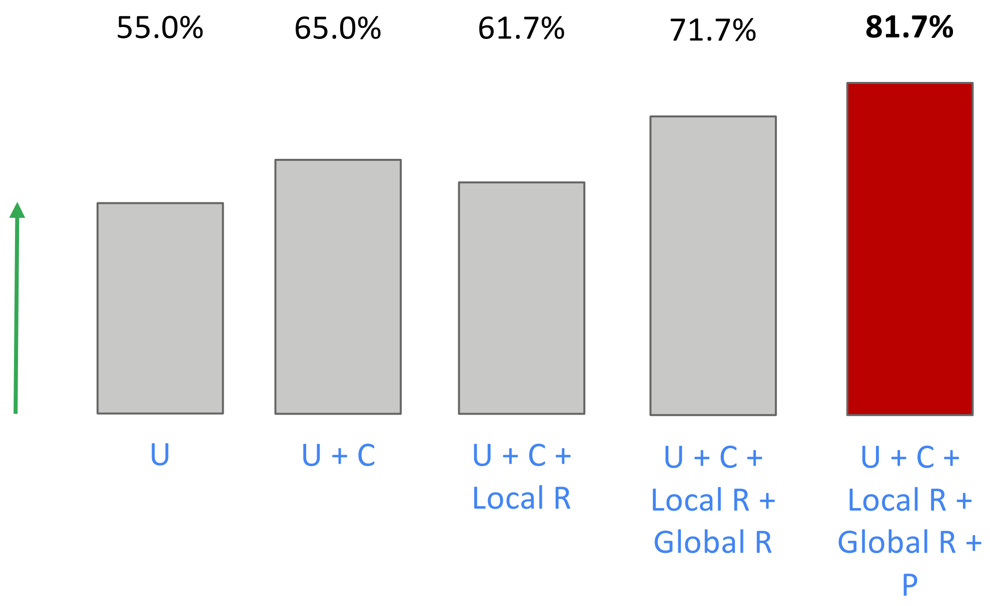
## Indirect evaluation: Model simulation




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

# Indirectly Evaluating Quantification

## Indirect evaluation: Model simulation



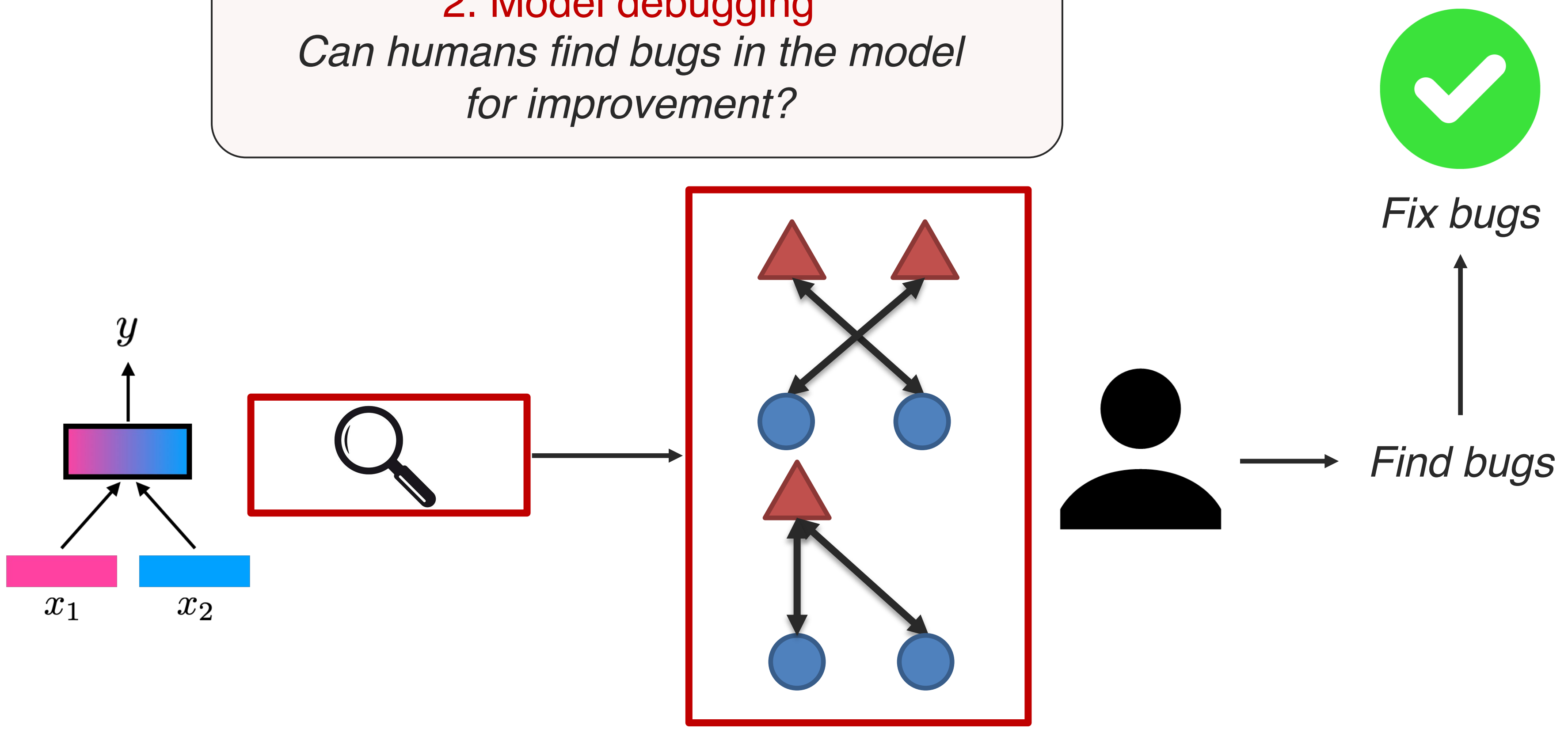

MultiViz stages leads to higher accuracy and agreement
Blind test + reasonable baselines + measurable outcome

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

# Indirectly Evaluating Quantification

## Indirect evaluation: Model error analysis and debugging

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




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

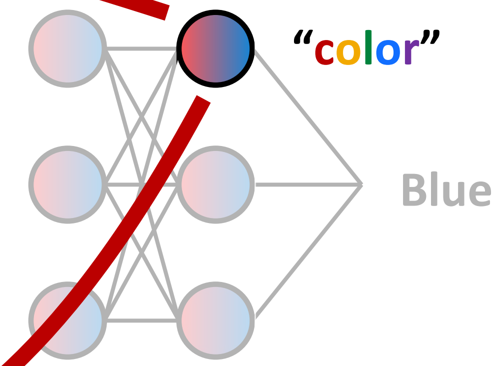
# Indirectly Evaluating Quantification

## Indirect evaluation: Model error analysis and debugging

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

Local analysis

3. Multimodal representations





*What color is the Salisbury Rd sign?*


*What color is the building?*

*What color are the checkers on the wall?*

Global analysis

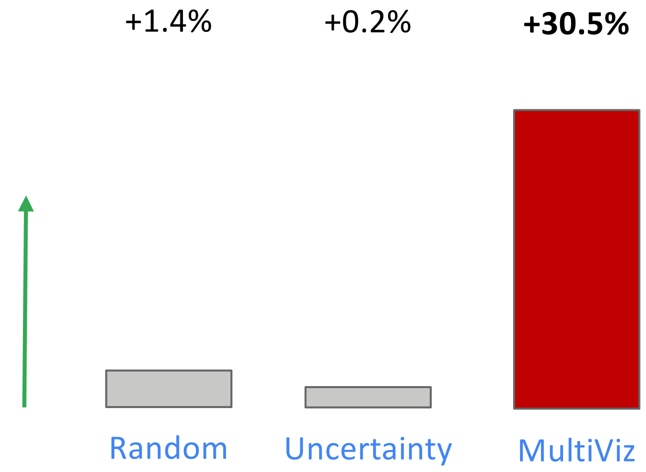
*"Models pick up cross-modal interactions but fail in identifying color!"*

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

# Indirectly Evaluating Quantification

## Indirect evaluation: 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: Towards Visualizing and Understanding Multimodal Models. ICLR 2023]

# Indirectly Evaluating Quantification

**More ways for indirect evaluation:**

- Model selection: given fixed budget, try randomly or try models in order based on what quantification tells me.
- Data/modality selection: given fixed budged, collect random data or collect based on what quantification tells me.
- If quantification gives theoretical result, check how well the theory matches experiments.

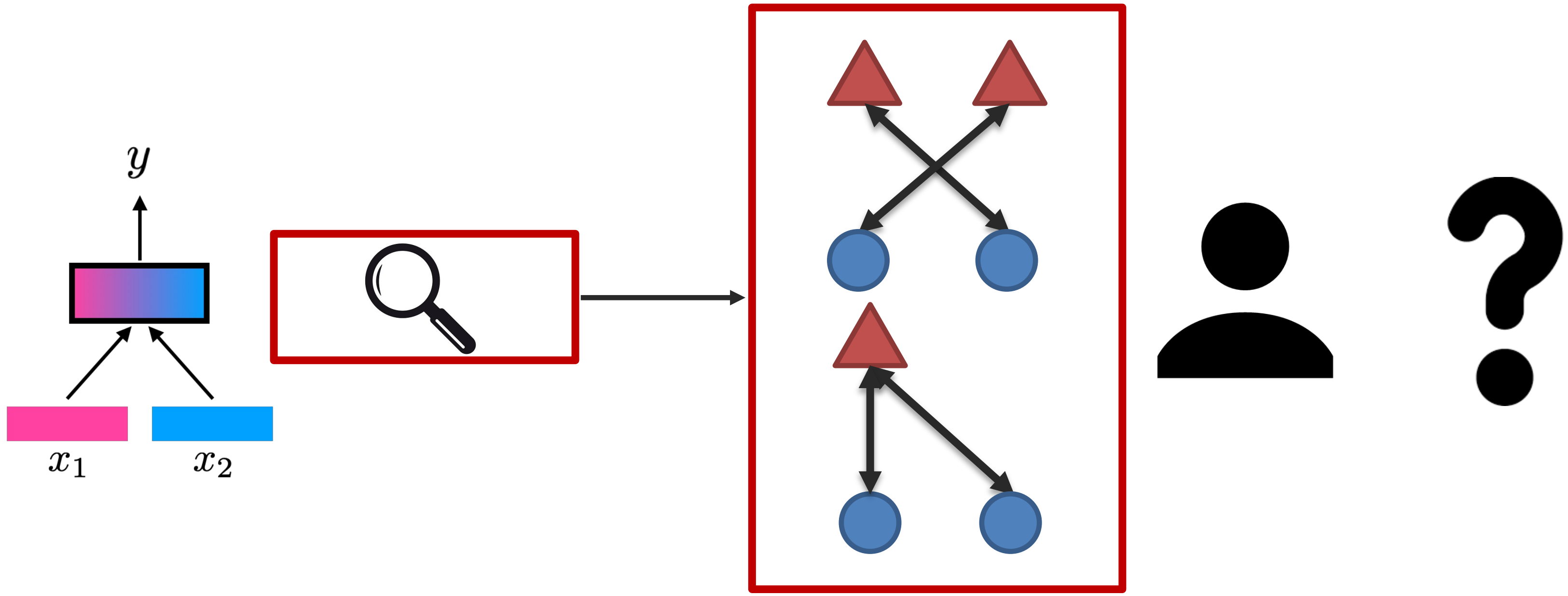
[Liang et al., MultiViz: Towards Visualizing and Understanding Multimodal Models. ICLR 2023]

# 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



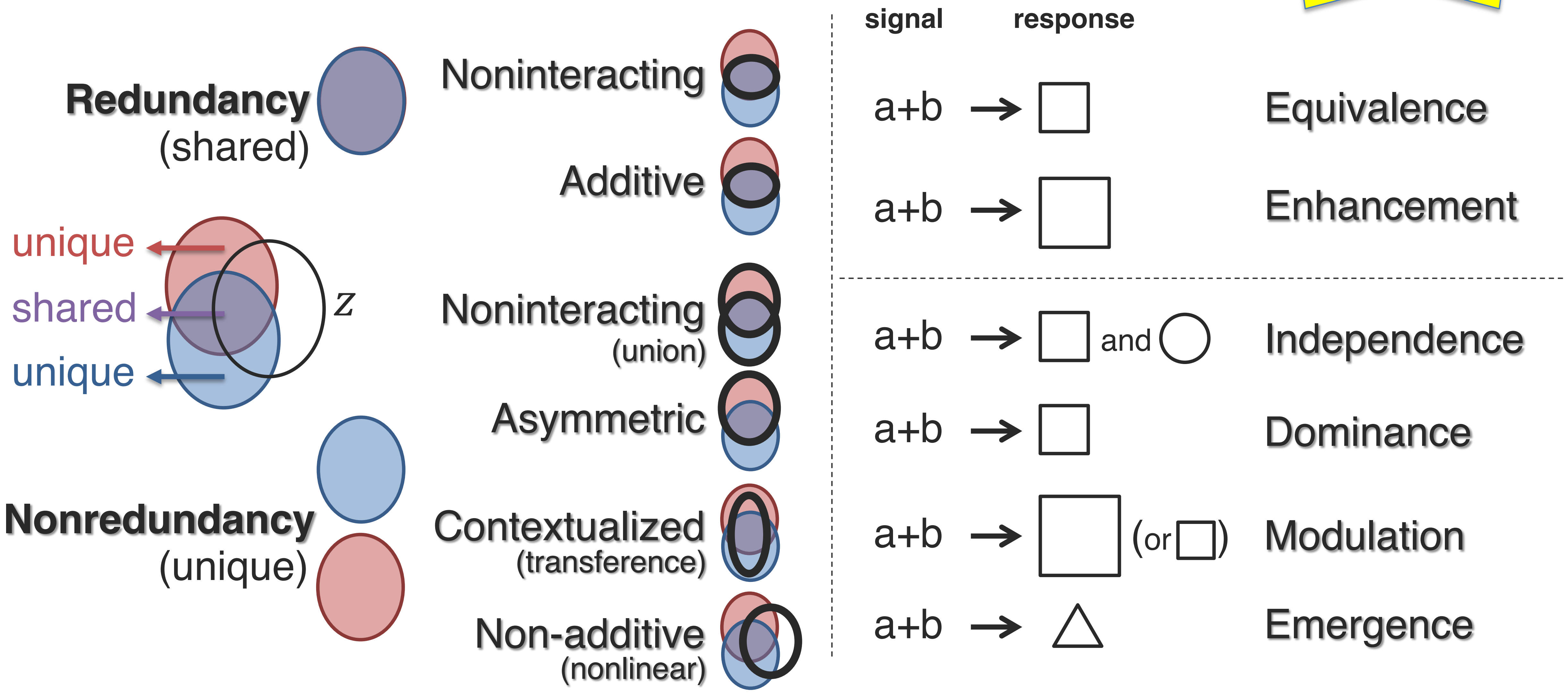


[Chandrasekaran et al., Do explanations make VQA models more predictable to a human? EMNLP 2018]

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

# Challenges: Quantifying Multimodal Interactions

**Open challenges**



# Challenges: Quantifying Multimodal Interactions

**Recall error analysis!**

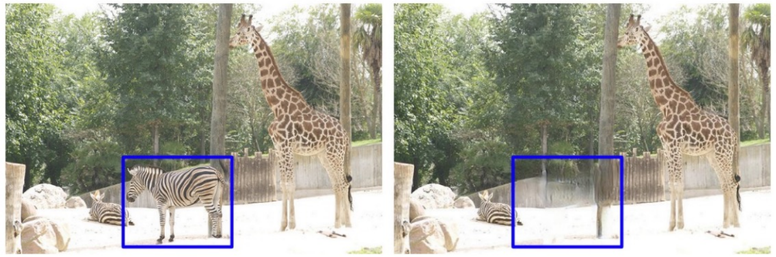
## Causal, logical interactions beyond additive/multiplicative

**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:** $p(y|do(zebras = 1))$

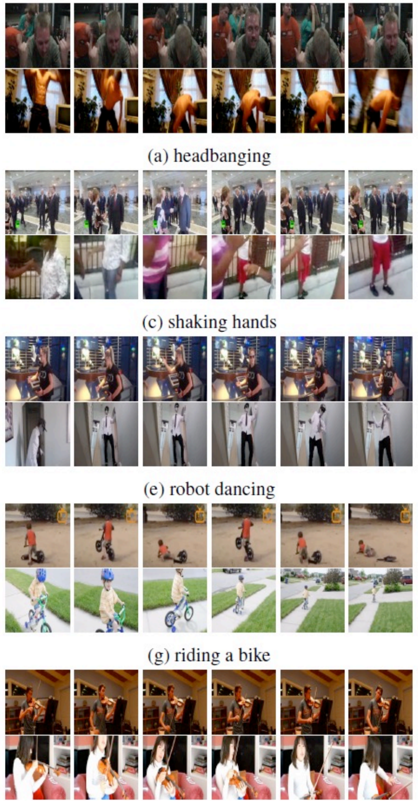
Existing models struggle to adapt to targeted causal interventions.
How can we make them more robust to spurious correlations?

[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?**

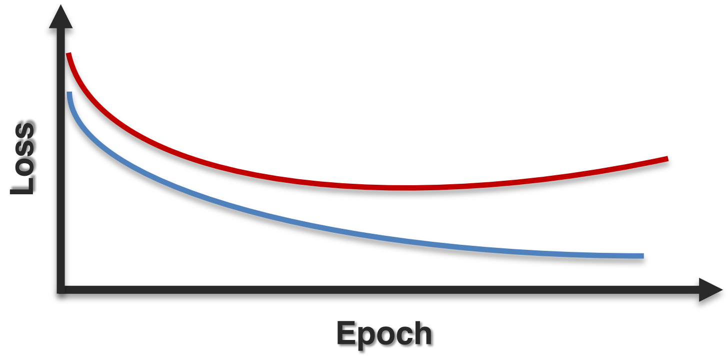
[Wang et al., What Makes Training Multi-modal Classification Networks Hard? CVPR 2020]

# 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

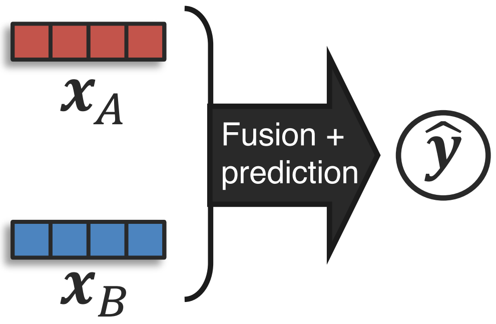
OGR wrt each modality tells us how much to train that modality

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

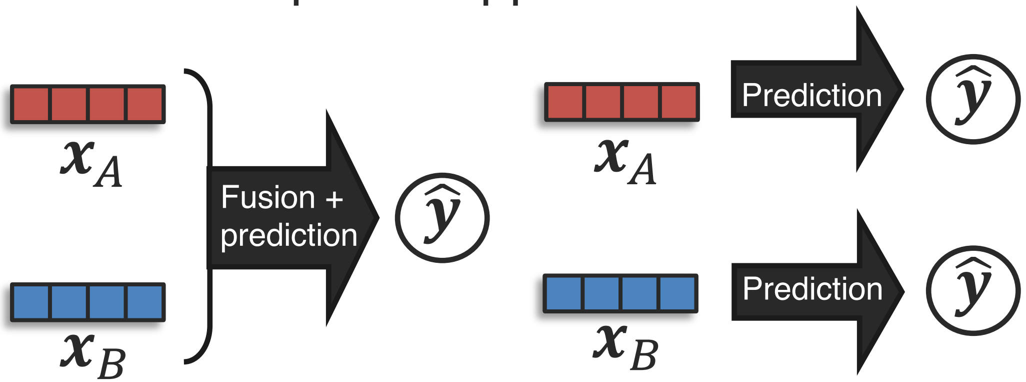
# Optimization challenges

## 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

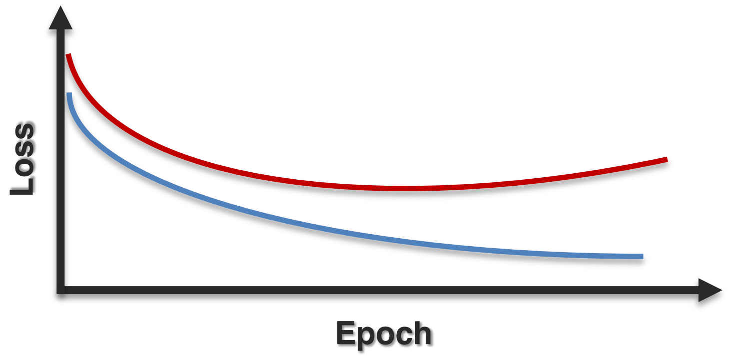
[Wang et al., What Makes Training Multi-modal Classification Networks Hard? CVPR 2020]

# 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.

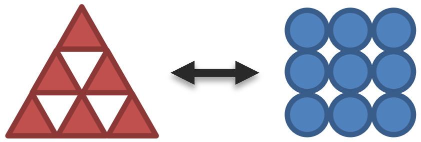


# More Quantification

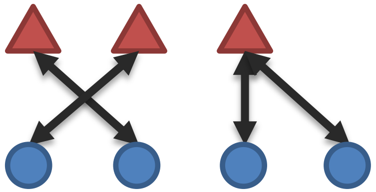
## Dimensions of quantification

Representation Alignment Reasoning Transference Generation

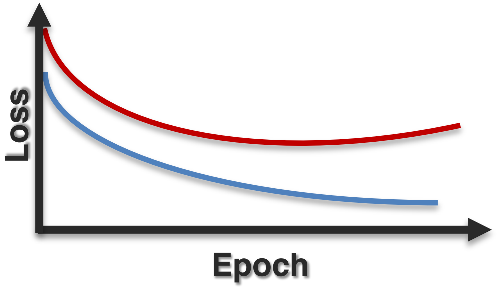
Heterogeneity



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 Multimodal?

A dictionary definition…

**Multimodal:** with multiple modalities

A research-oriented definition…

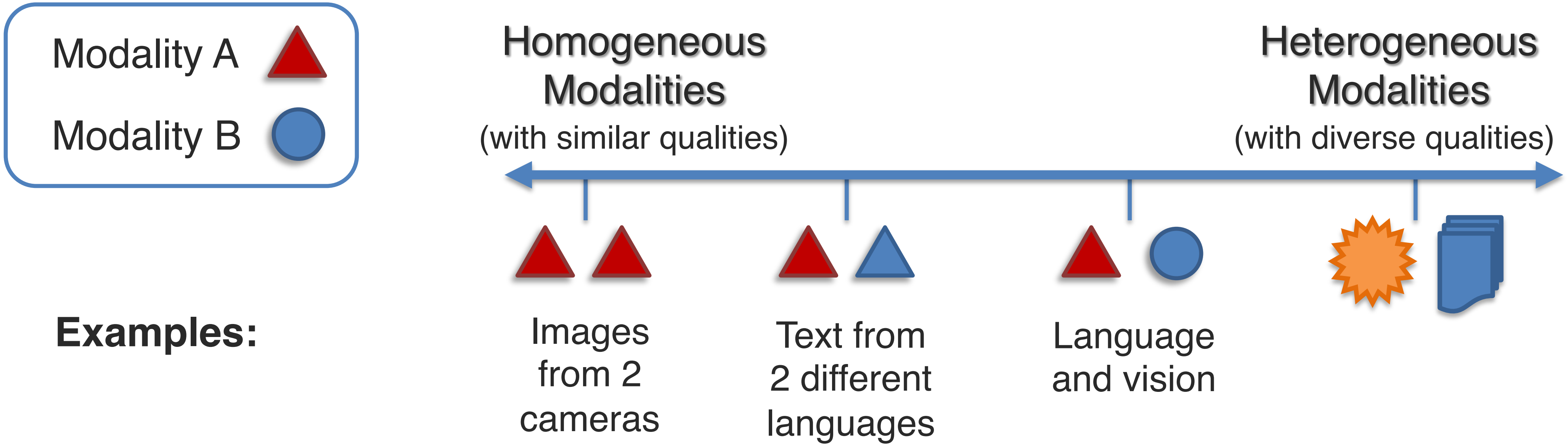
***Multimodal* is the scientific study of**

**heterogeneous and interconnected data**

**Connected + Interacting**

# Heterogeneous Modalities

**Heterogeneous:** Diverse qualities, structures and representations.



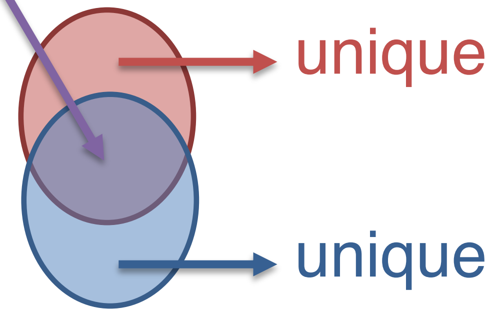

Abstract modalities are more likely to be homogeneous

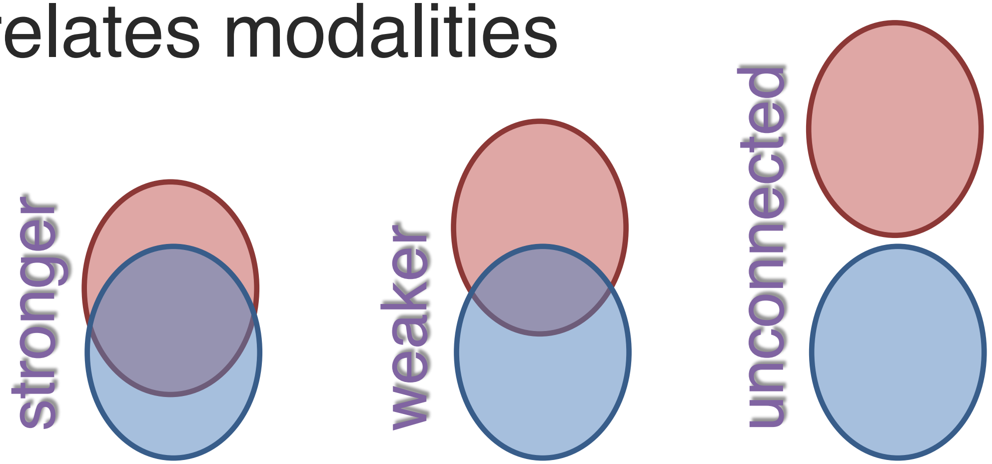
# Connected Modalities

Connected: Shared information that relates modalities

Modality A

Modality B




## Statistical

### Association

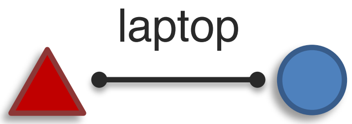

e.g., correlation, co-occurrence

### Dependency

e.g., causal, temporal

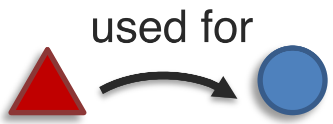
## Semantic

### Correspondence




e.g., grounding

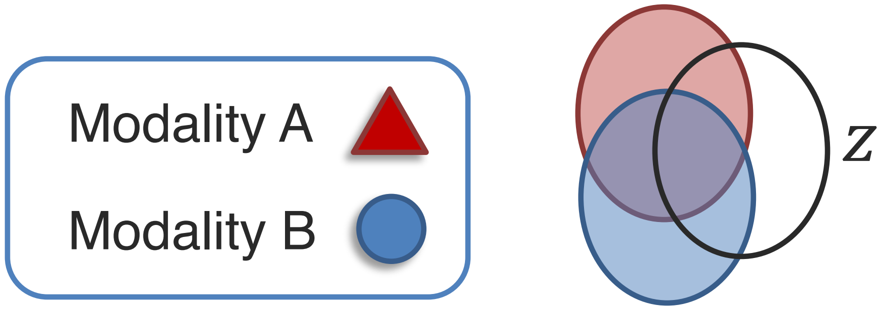
### Relationship

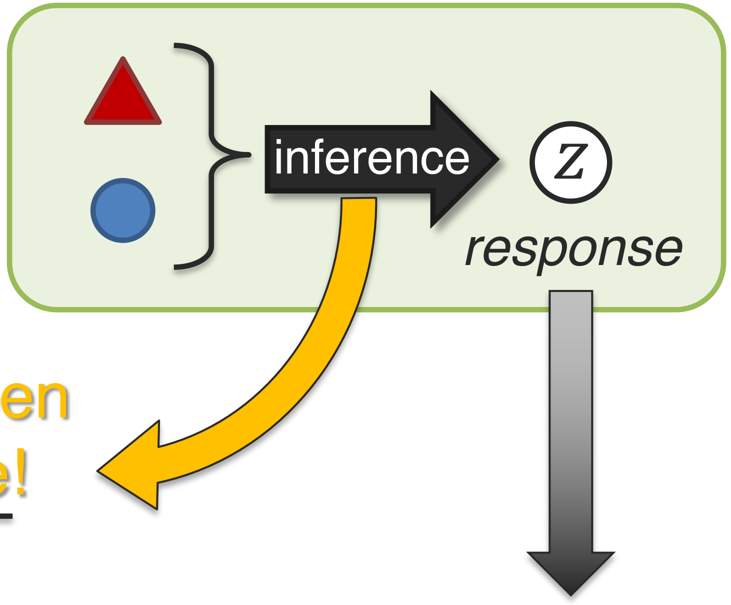



e.g., function

# Interacting Modalities

**Interacting**: process affecting each modality, creating new response




Interactions happen during inference!

"Inference" examples:

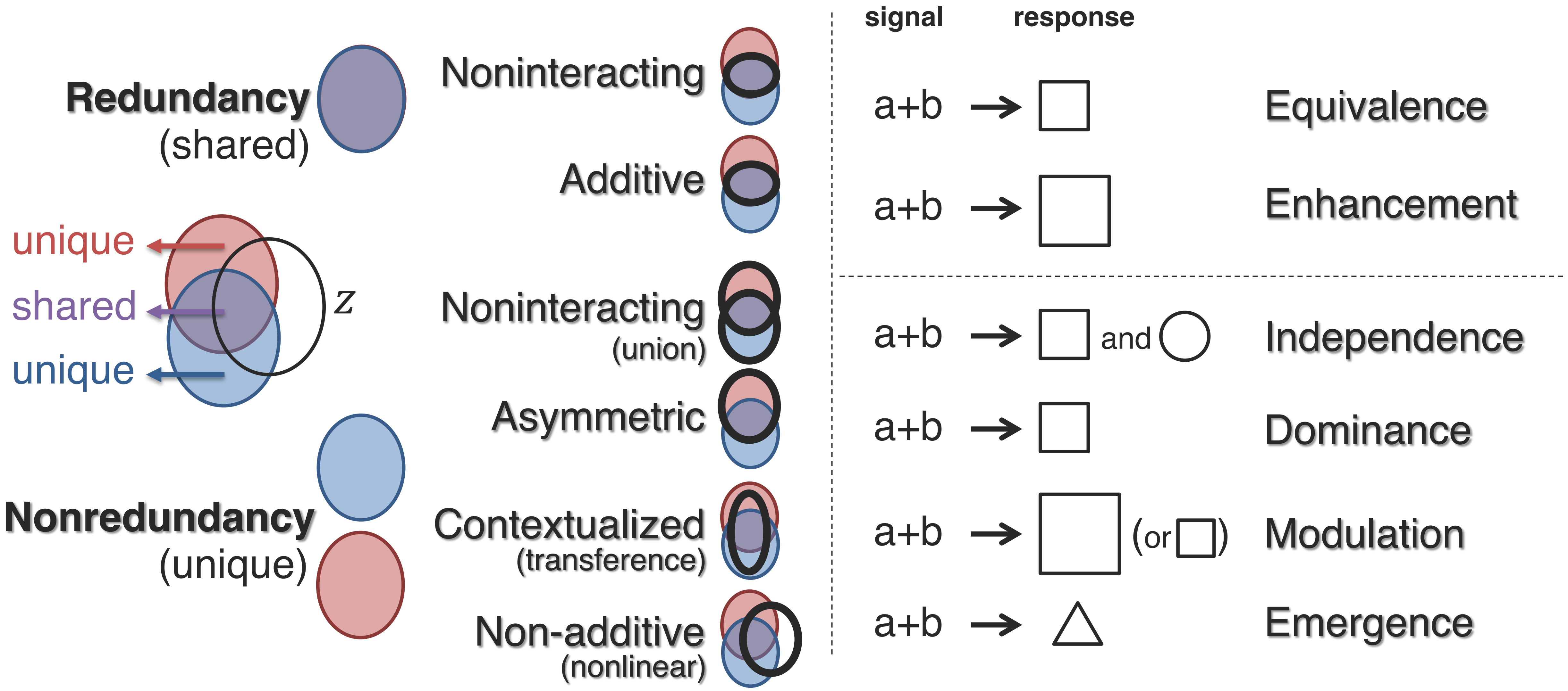
- Representation fusion
- Prediction task
- Modality translation

*representation*

$\hat{y}$ *prediction*

*modality C*

# Interacting Modalities



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

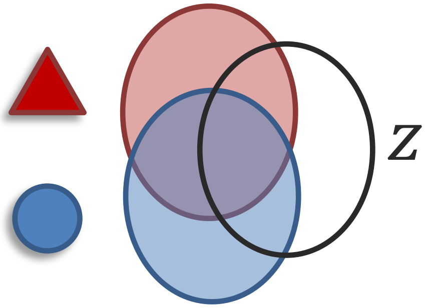
**Heterogeneous**

➕

**Connected**

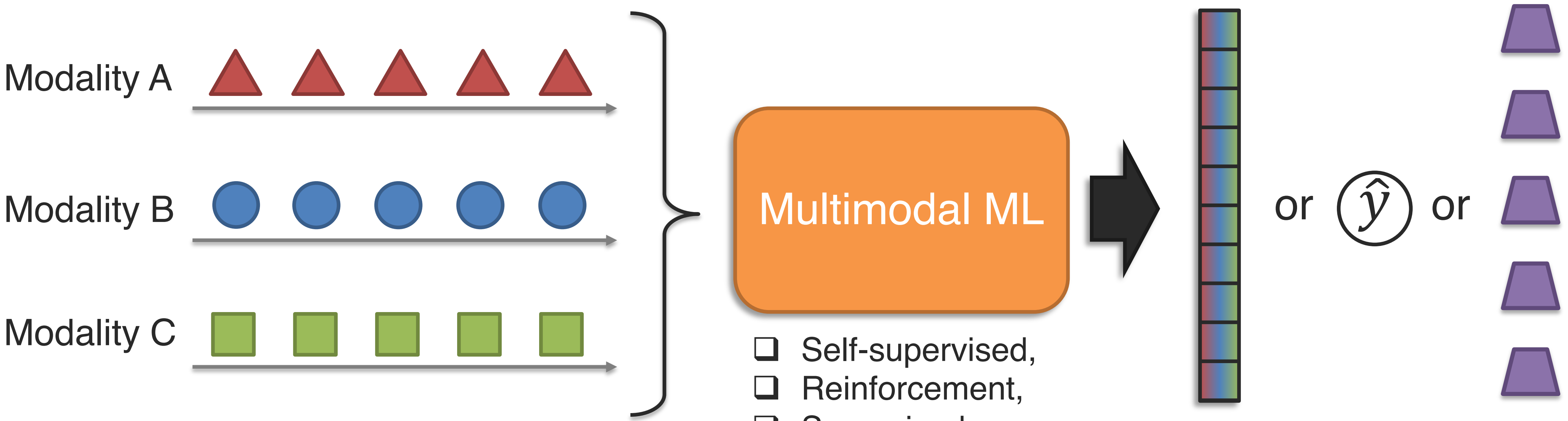
➕

**Interacting**




# Multimodal is the scientific study of heterogeneous and interconnected data ☺

# Multimodal Machine Learning




- ❑ Self-supervised,
- ❑ Reinforcement,
- ❑ Supervised, …

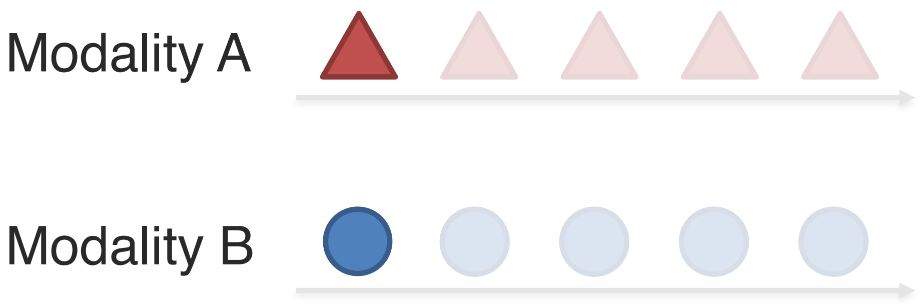
***What are the core multimodal technical challenges,***
***understudied in conventional machine learning?***

# 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

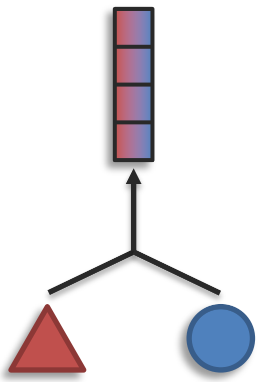
*representation using holistic features*

# Challenge 1: Representation

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

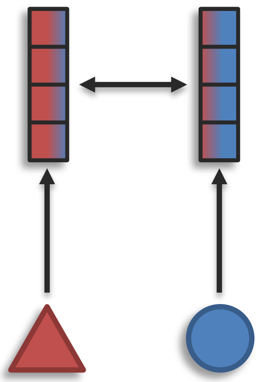
## Sub-challenges:

### Fusion



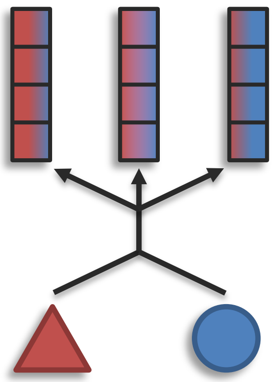
# modalities **>** # representations

### Coordination



# modalities **=** # representations

### Fission



# modalities **<** # representations

# Challenge 2: Alignment

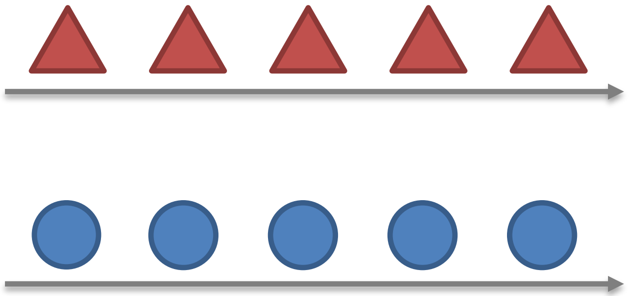
**Definition:** Identifying and modeling cross-modal connections between all elements of multiple modalities, building from the data structure

Most modalities have internal structure with multiple elements

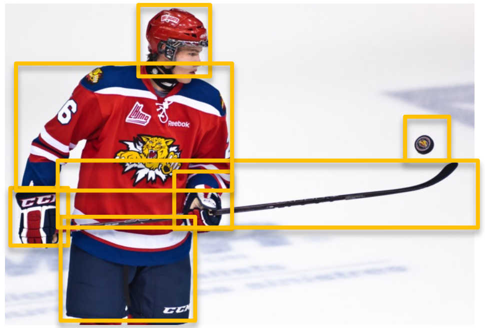
**Elements with temporal structure:**

Modality A

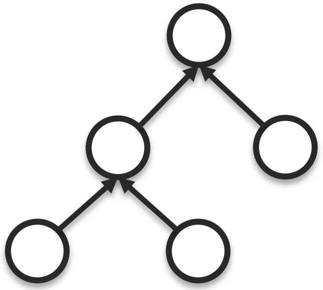
Modality B



**Other structured examples:**



Spatial



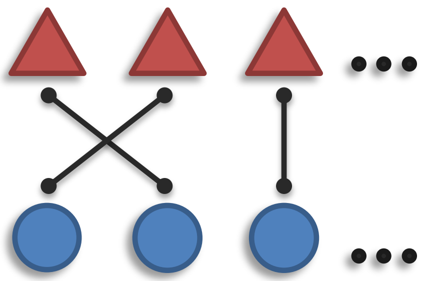
Hierarchical

# Challenge 2: Alignment

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

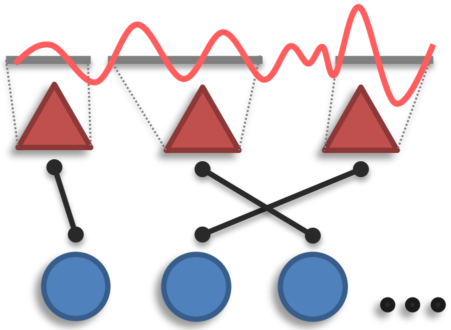
## Sub-challenges:

### Discrete Alignment



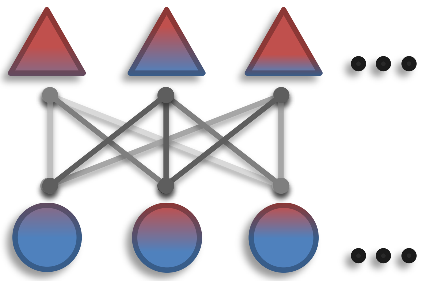
Discrete elements and connections

### Continuous Alignment



Segmentation and continuous warping

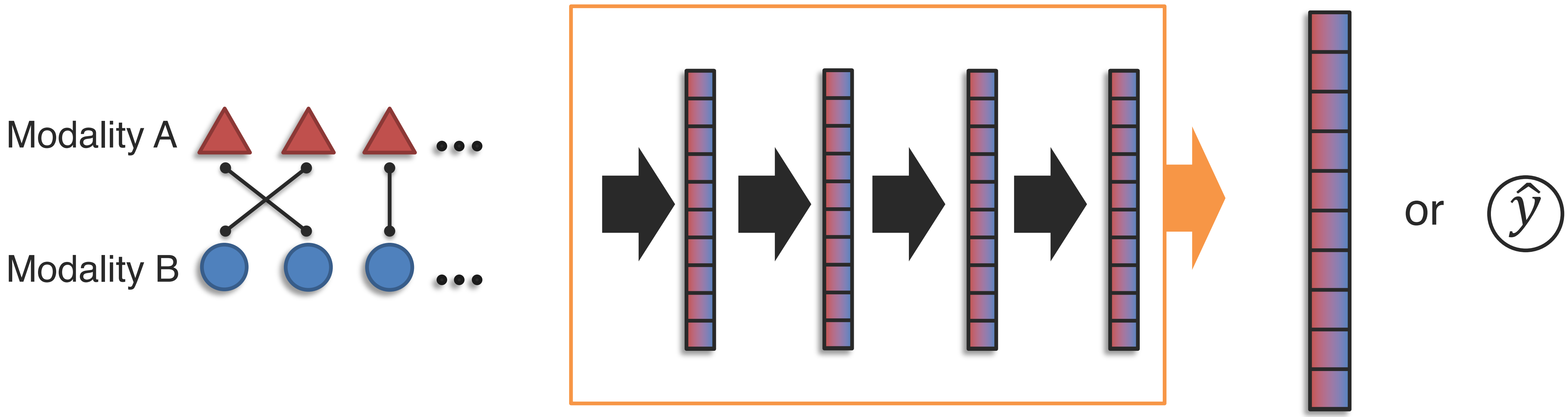
### Contextualized Representation



Alignment + representation

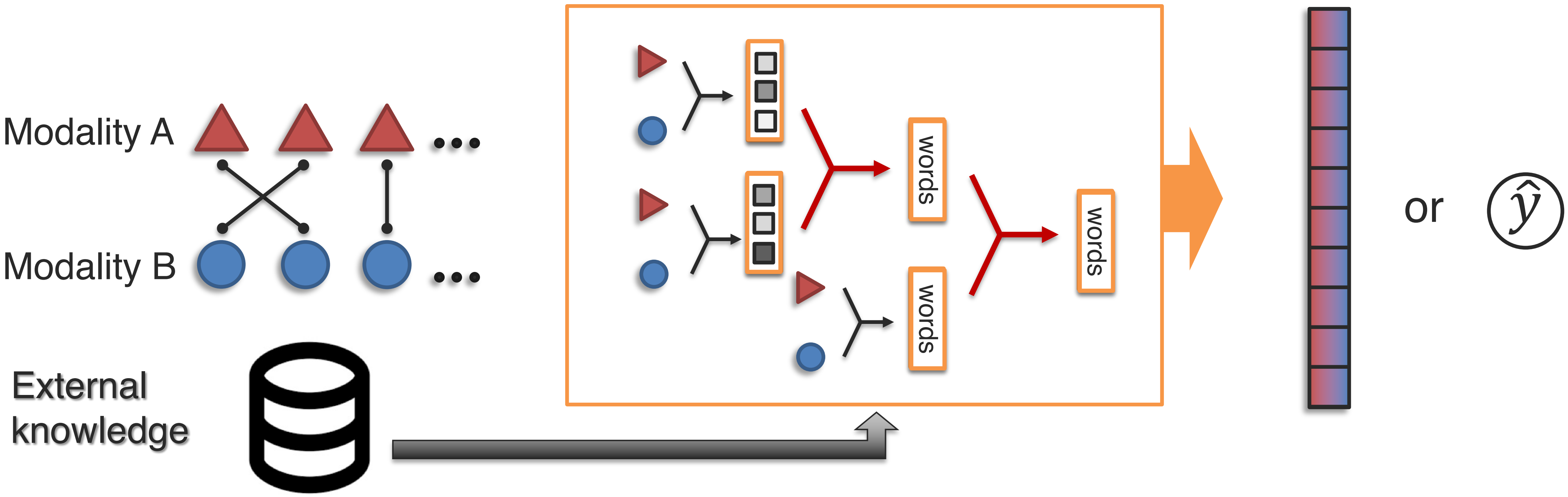
# Challenge 3: Reasoning

**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

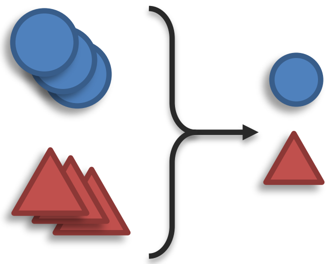


# Challenge 4: Generation

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

## Sub-challenges:

### Summarization



Reduction

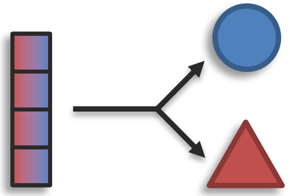
□ > □

### Translation

Maintenance

□ = □

### Creation



Expansion

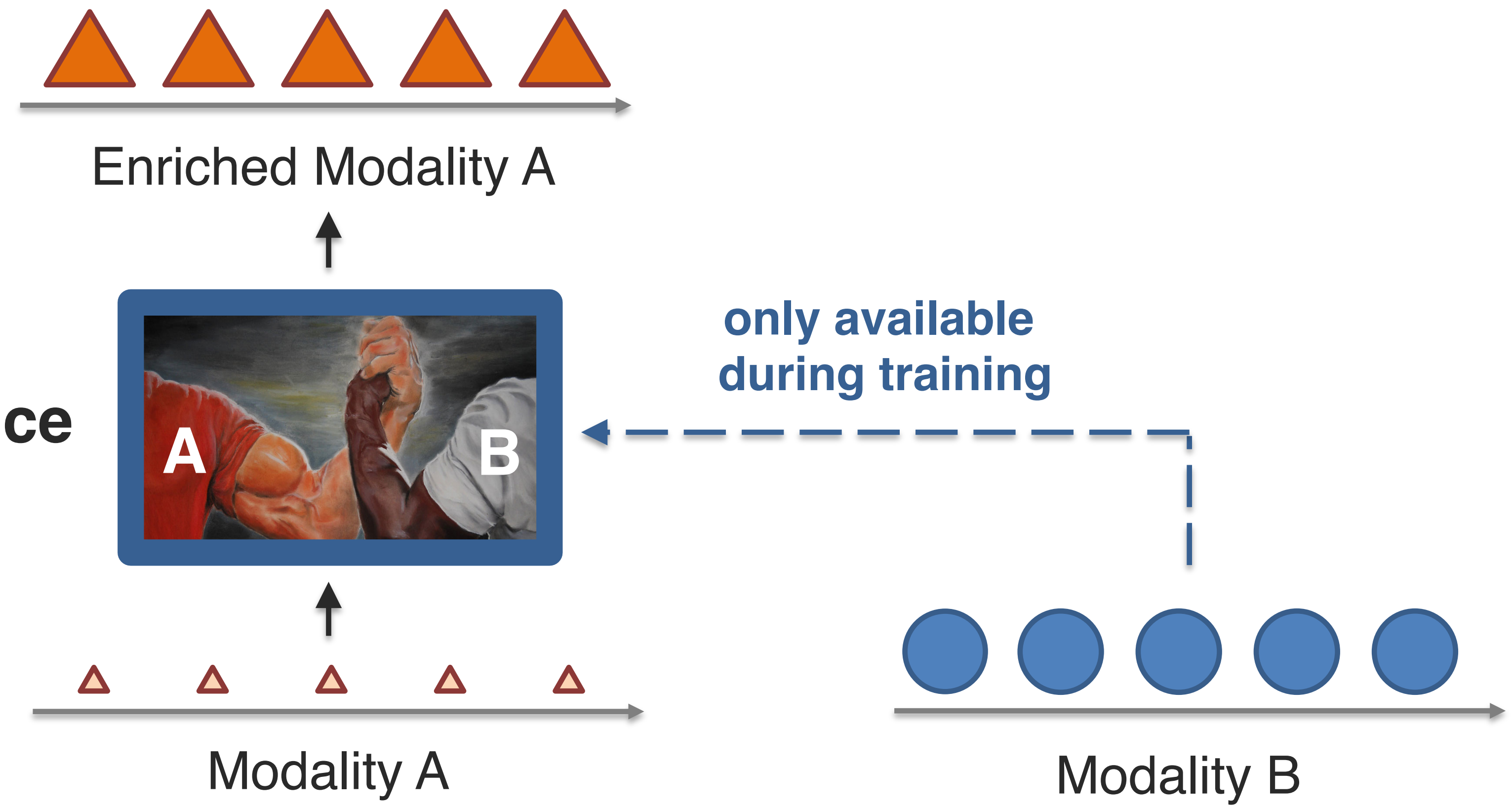
□ < □

**Information:** (content)

# Challenge 5: Transference

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

**Transference**

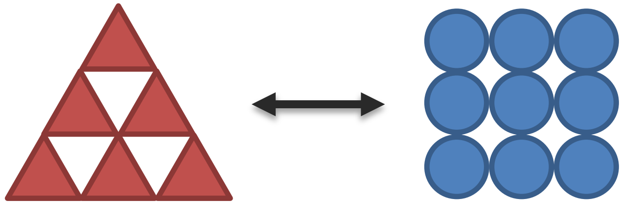


# Challenge 6: Quantification

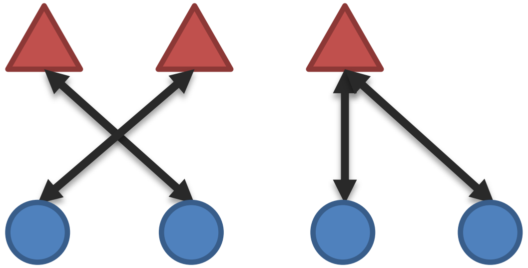
**Definition:** Empirical and theoretical study to better understand heterogeneity, cross-modal interactions and the multimodal learning process

## Sub-challenges:

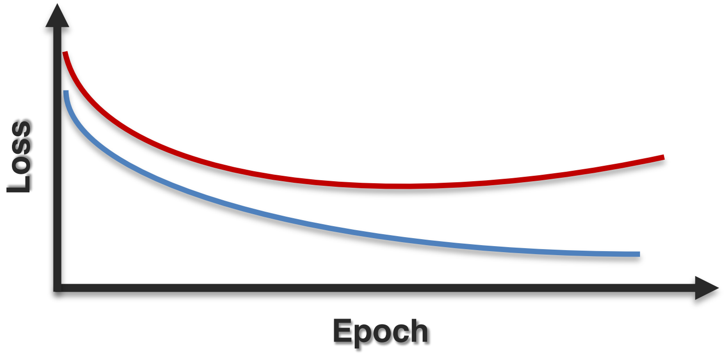
**Heterogeneity**



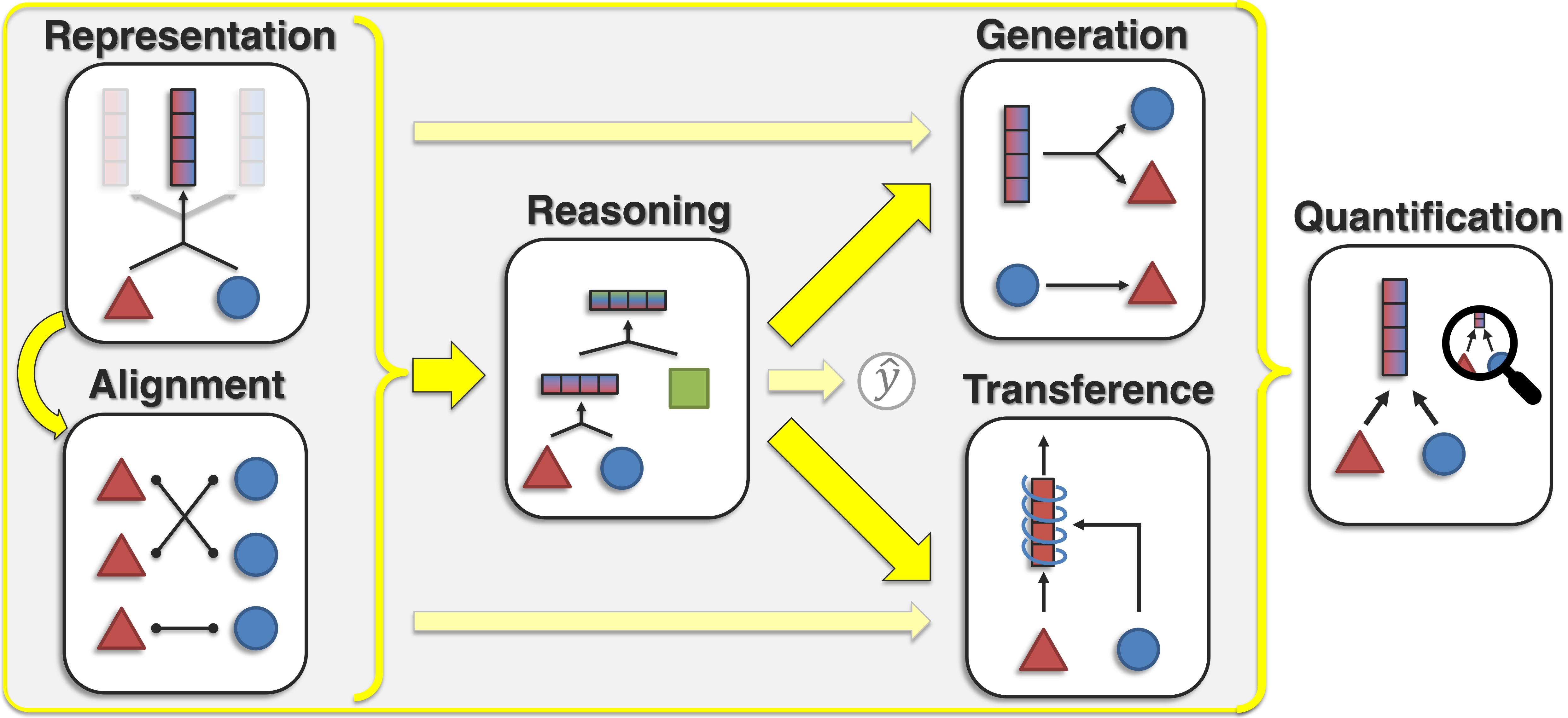
**Connections & Interactions**



**Learning**

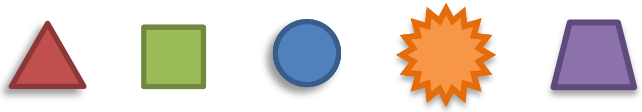


# Core Multimodal Challenges



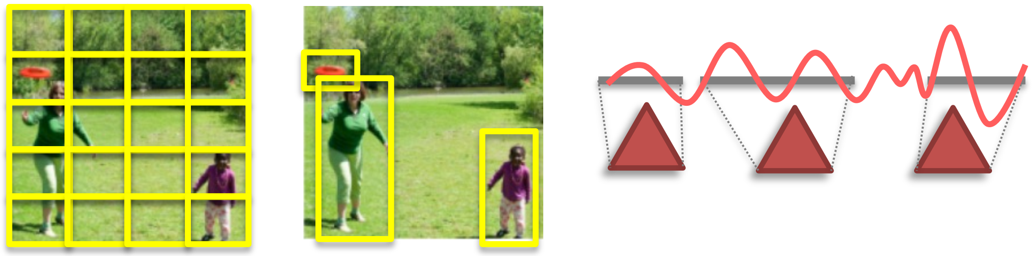
# Future Direction: Heterogeneity

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

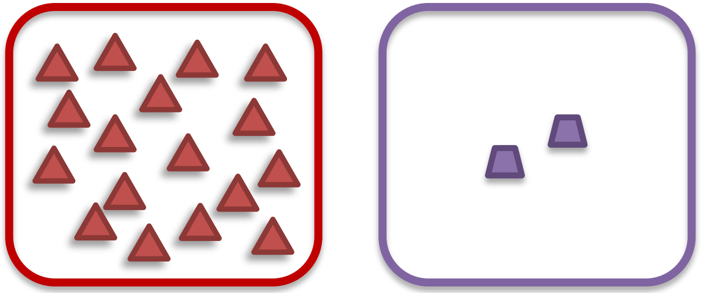
Few modalities → High-modality


Language Vision Audio Graphs Control LIDAR Sensors Set Table Financial Medical

**Examples:**

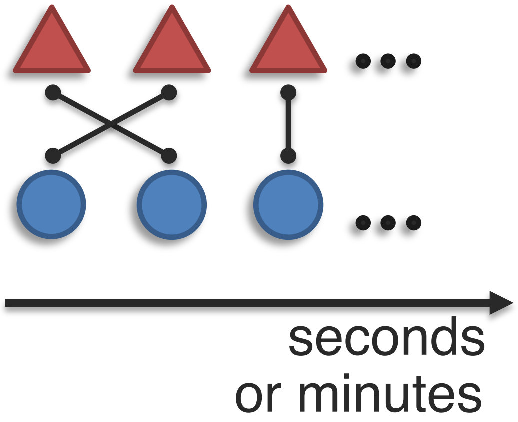
### Non-parallel learning


### Limited resources



# Future Direction: Long-term

## Short-term




## Long-term

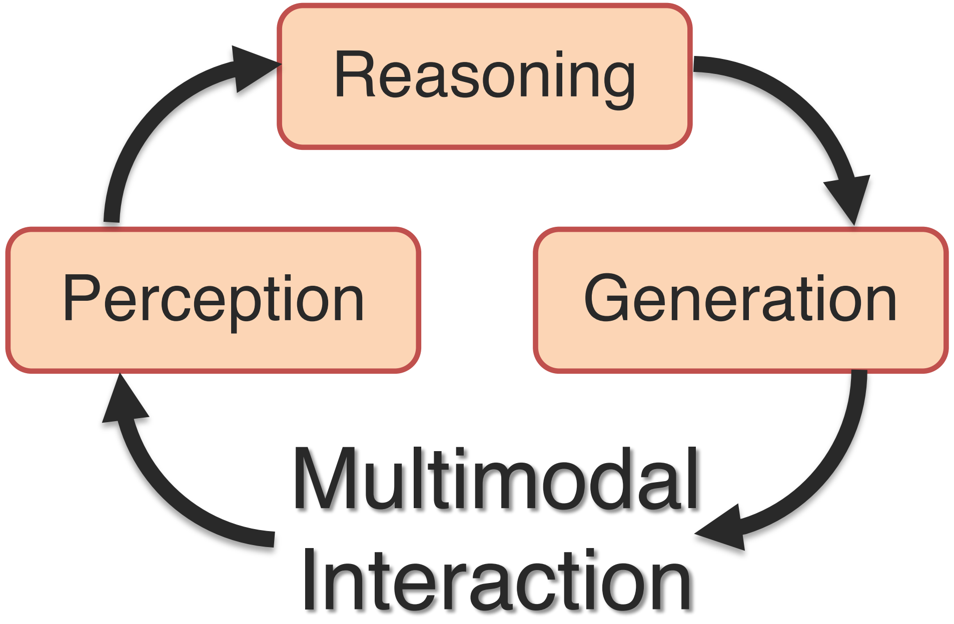
## Examples:

**Compositionality** **Memory** **Personalization**

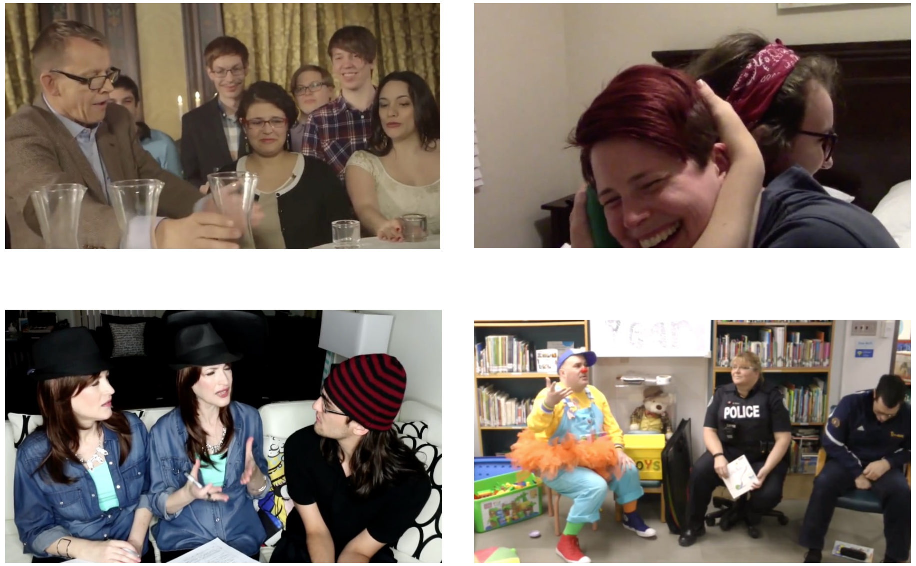
# Future Direction: Interaction

## Social-IQ

https://www.thesocialiq.com/




Social Intelligence



## Examples:

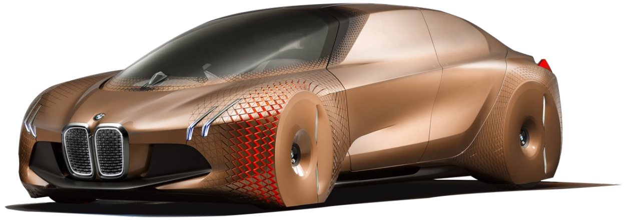
**Multi-Party** **Causality** **Ethical**

## MultiViz

https://github.com/pliang279/MultiViz

# Future Direction: Real-world

Healthcare Decision Support



Intelligent Interfaces and Vehicles

Online Learning and Education

## Examples:

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

## What is Multimodal?

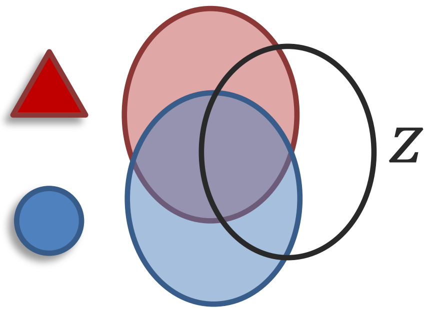
Heterogeneous

+

Connected

+

Interacting




## Why is it hard?

Representation

Alignment

Reasoning

Generation

Transference

Quantification

## What is next?

Heterogeneity

High-modality

Long-term

Interaction

Real-world