

Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 12.2: Quantification

Louis-Philippe Morency

** Original course co-developed with Tadas Baltrusaitis. Major revision co-developed with Paul Liang. Co-lecturers included Yonatan Bisk, Daniel Fried and Carlos Busso.*

Administrative Stuff

Upcoming Course Assignments

Week 12 Reading Assignment (due Tuesday 11/17 at 8pm ET)

- Two papers (with a focus on multimodal generation)

Final project presentations (Tuesday 12/1 and Thursday 12/3)

- Read final report instructions first!
- Poster presentations
  - Poster will focus on a subset of the research ideas and results

Final project report (due Sunday 12/6 at 8pm ET)

- N-1 research ideas, exploring new modeling approaches
  - All ideas should be detailed in the report

Attendance using Poll Everywhere



1. Scan QR Code
   or
   https://pe.app/mmml

2. Enter your AndrewID
   (as your “name tag”)

❌ *Do not try to sign-in*

3. Click ”Share location”



Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 12.2: Quantification

Louis-Philippe Morency

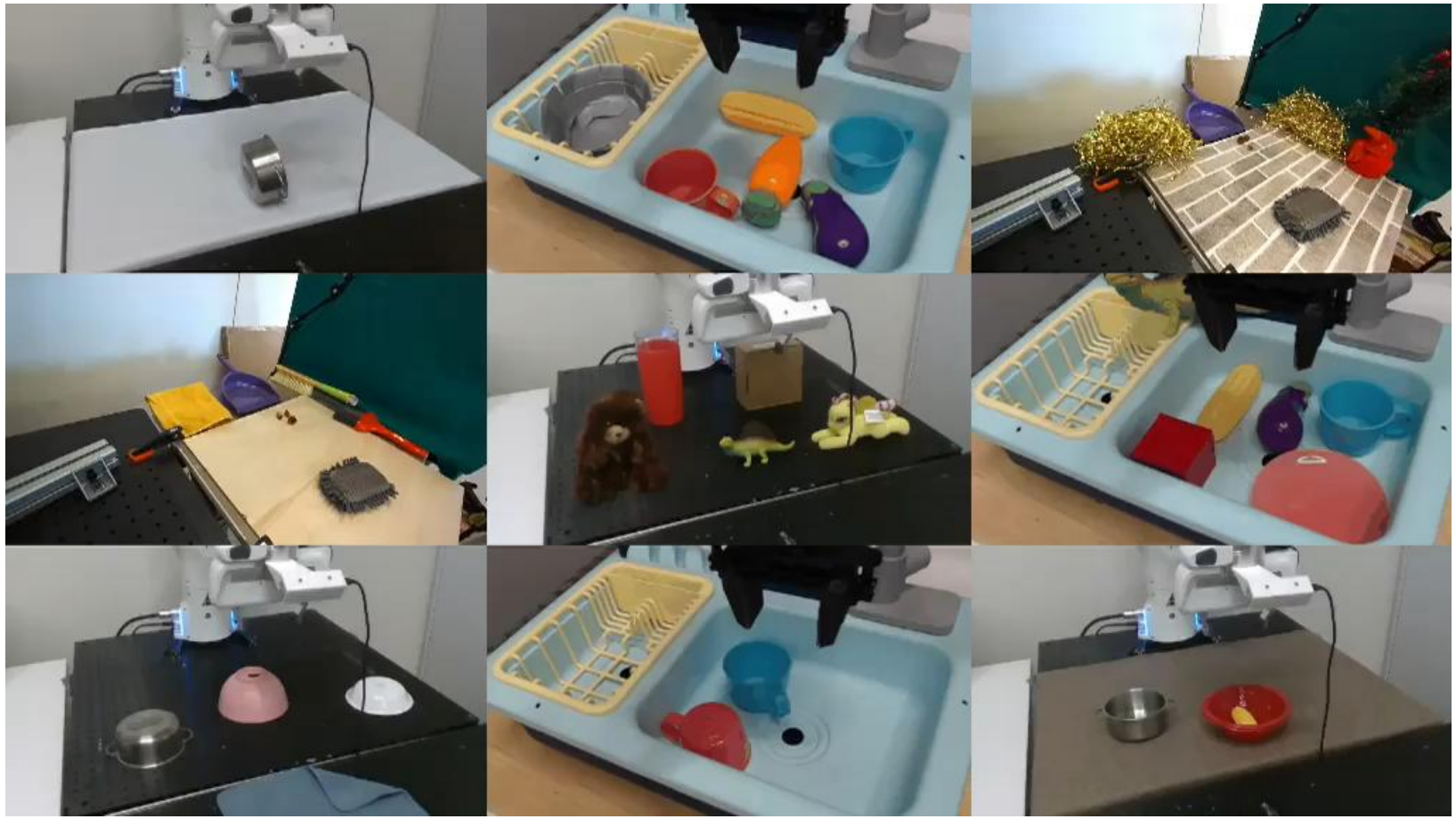
** Original course co-developed with Tadas Baltrusaitis. Major revision co-developed with Paul Liang. Co-lecturers included Yonatan Bisk, Daniel Fried and Carlos Busso.*

Objectives of today's class

- Vision-language-action models and action representations
  - Video generators as policies and simulators
- Quantification: heterogeneity, interactions, learning process
  - Modality biases, modality collapse, and social biases
- Quantifying cross-modal interactions
  - Evaluating quantification: direct and indirect
- Multimodal learning process and OGR

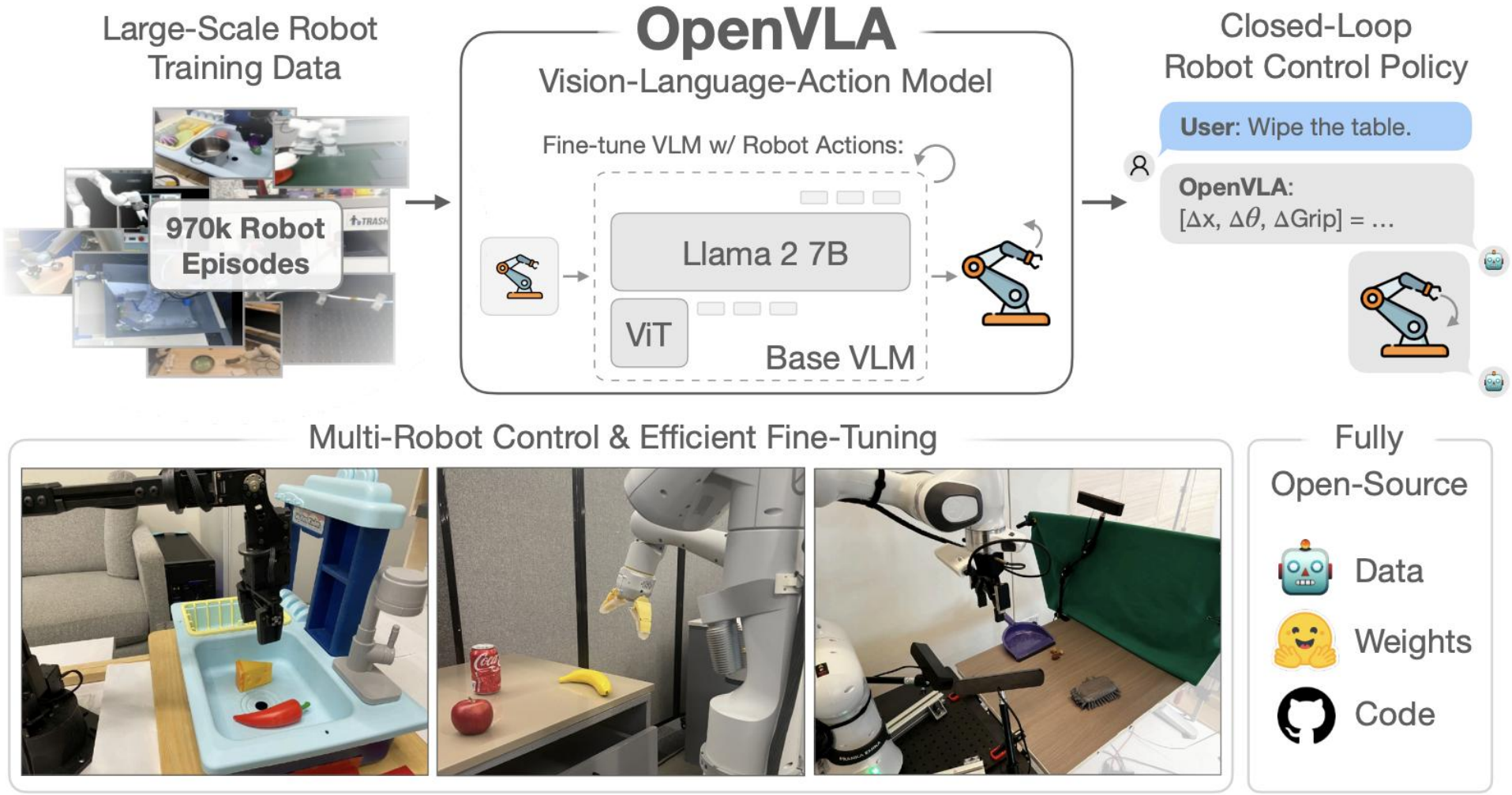
Embodied Agents

Generate precise robotics control directly via trained vision language models.



[Kim et al., OpenVLA: An Open-Source Vision-Language-Action Model. 2024]

Vision Language Action Models



OpenVLA: https://openvla.github.io/

How to represent action output?

Torques on motors?

Speed control on each joint?

Joint position control?

Pose and closeness control of gripper?

Code using Robot Skill APIs?

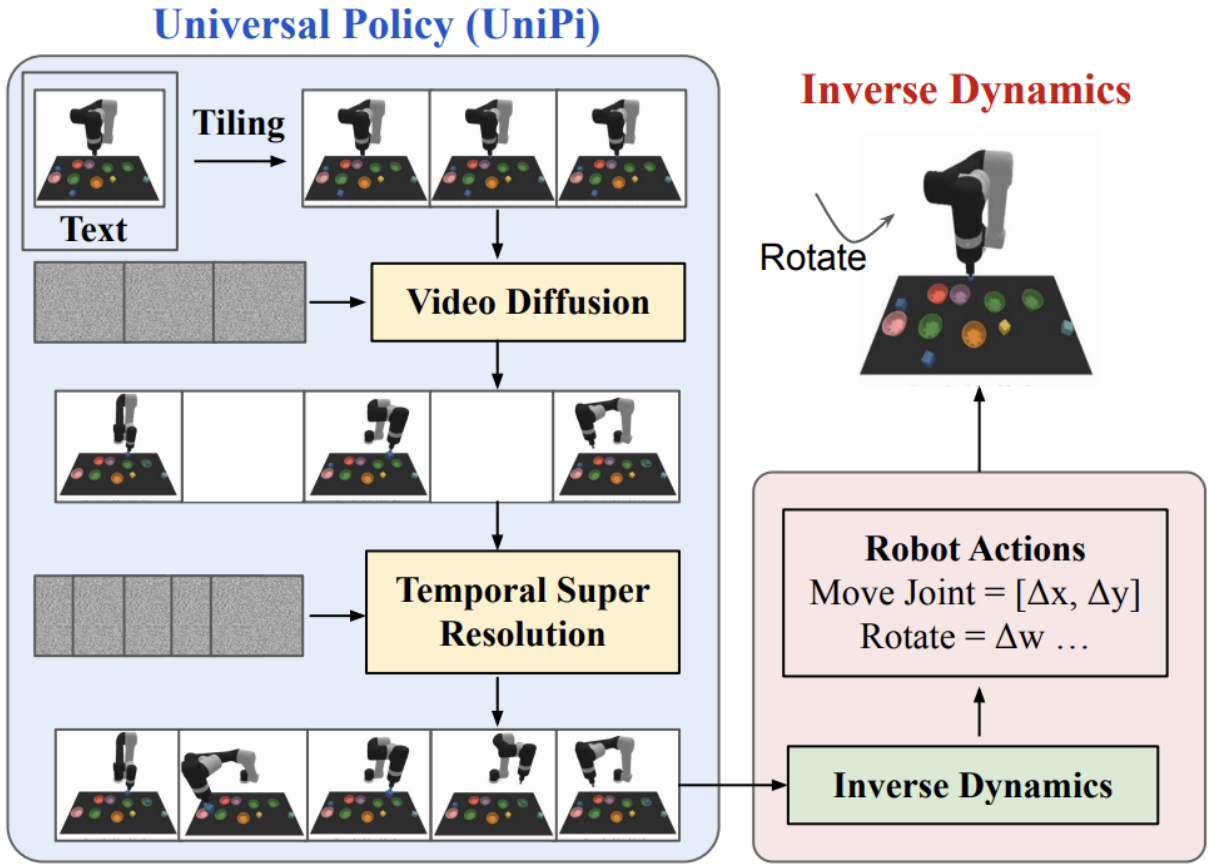
Video of how future looks like?

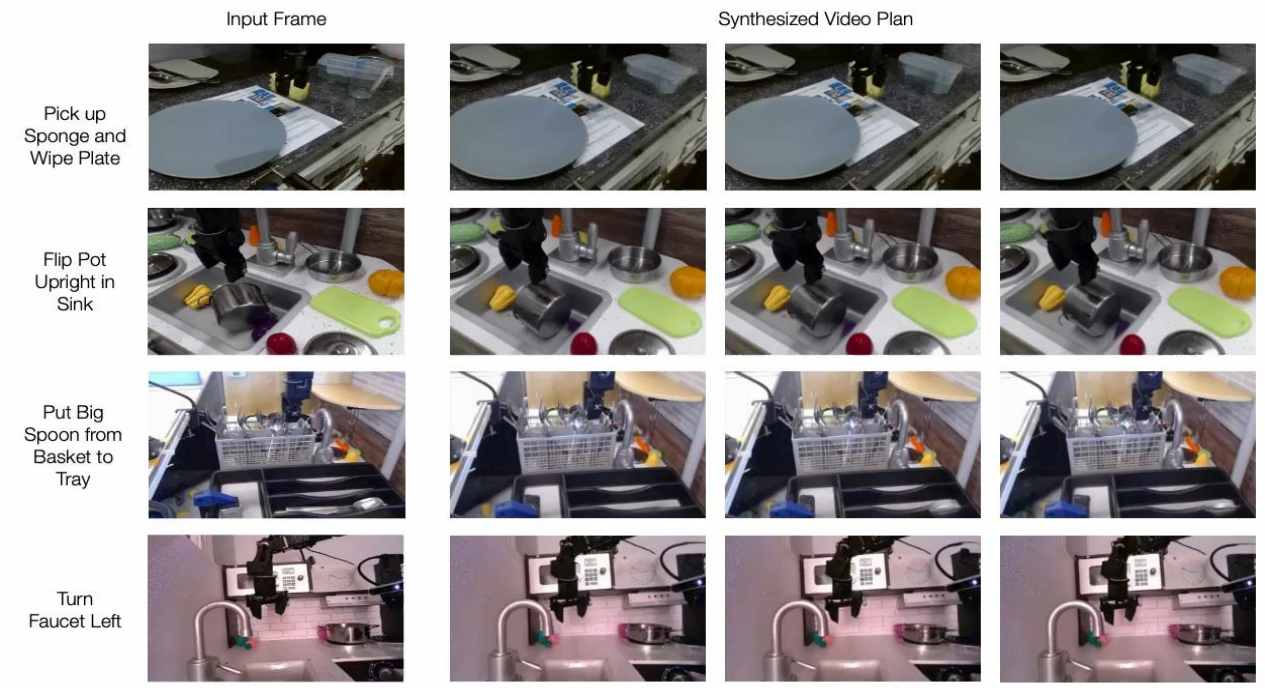
…



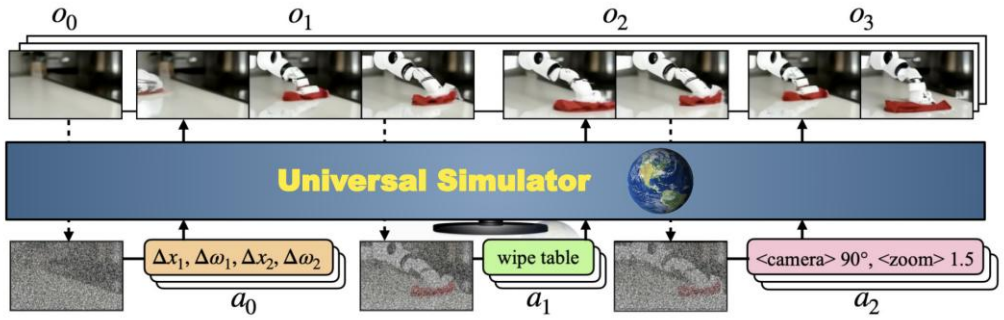


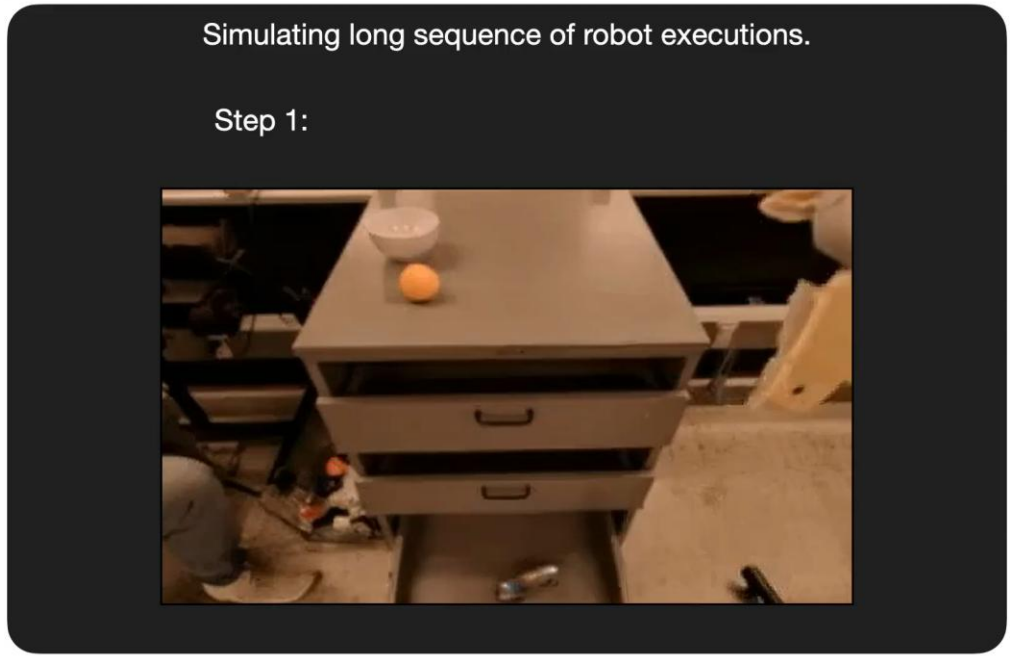
Video generators as Universal Policies





Video generators as Universal Simulators





https://universal-simulator.github.io/unisim/

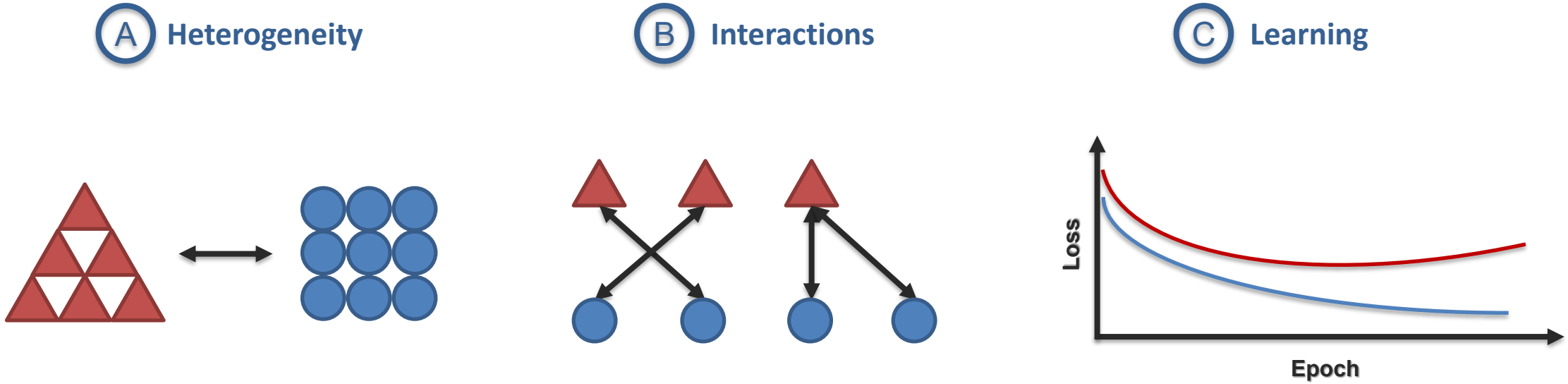
Prompt: The camera follows behind a white vintage SUV with a black roof rack as it speeds up a steep dirt road surrounded by pine trees on a steep mountain slope, dust kicks up from it's tires, the More



https://openai.com/sora#research

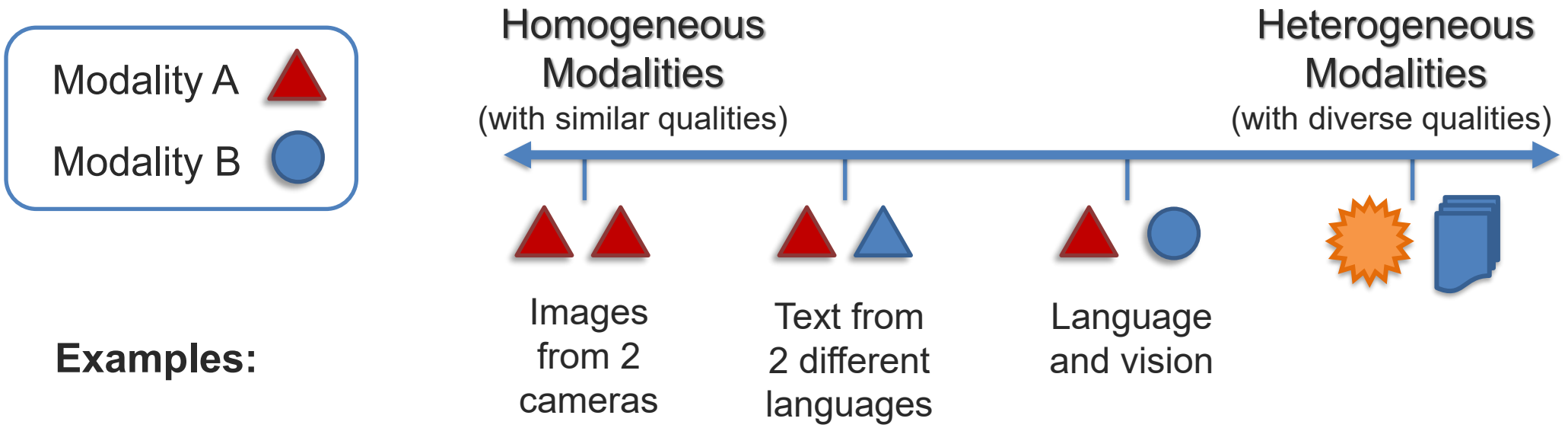
Quantification

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



Sub-Challenge 6a: Heterogeneity

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



Examples:

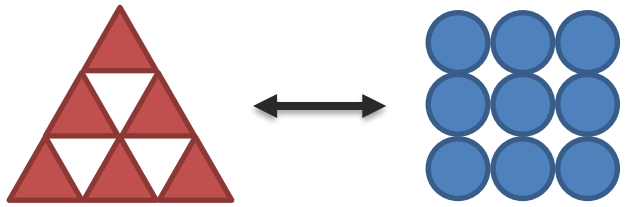
1. **Element representation**
2. **Element distribution**
3. **Structure**
4. **Information**
5. **Noise**
6. **Relevance**

Quantifying Heterogeneity

Open challenges:

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

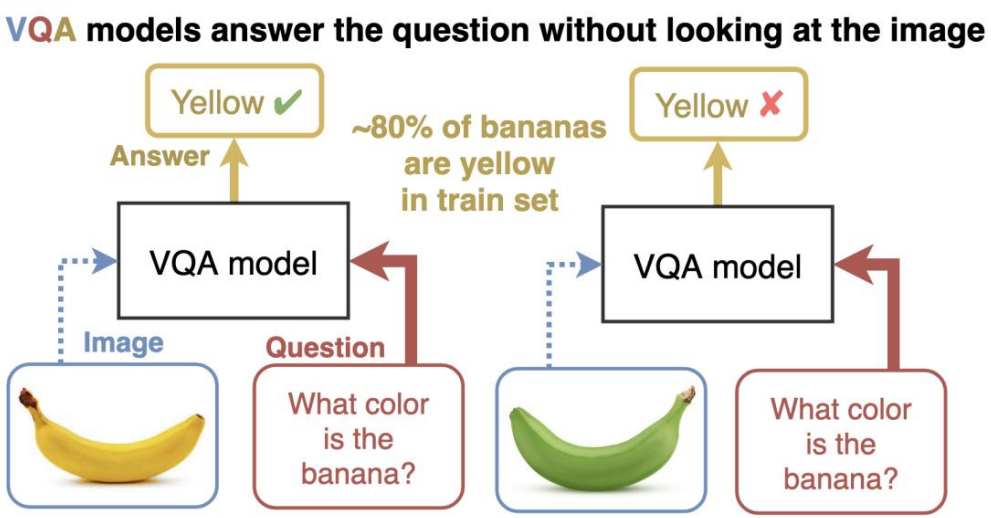
- Related fields: federated learning, active learning, distributed systems, structure & invariances



Modality Biases

Heterogeneity in information and relevance

Unimodal biases and modality collapse

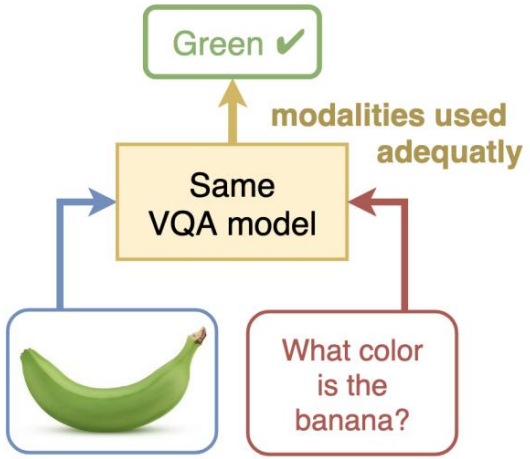


Balancing modalities

Balancing training



Not the case when trained with RUBi



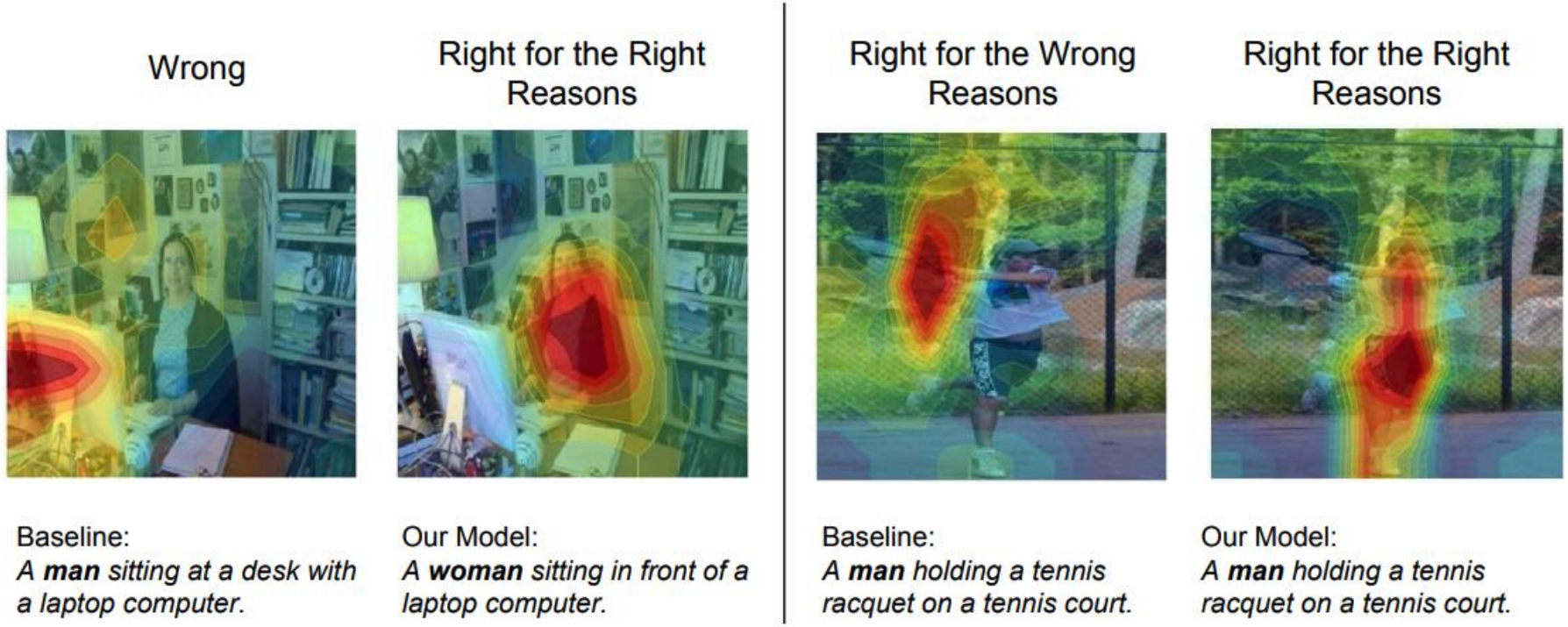
[Wu et al., Characterizing and Overcoming the Greedy Nature of Learning in Multi-modal Deep Neural Networks. ICML 2022]
[Javaloy et al., Mitigating Modality Collapse in Multimodal VAEs via Impartial Optimization. ICML 2022]
[Goyal et al., Making the V in VQA Matter: Elevating the Role of Image Understanding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

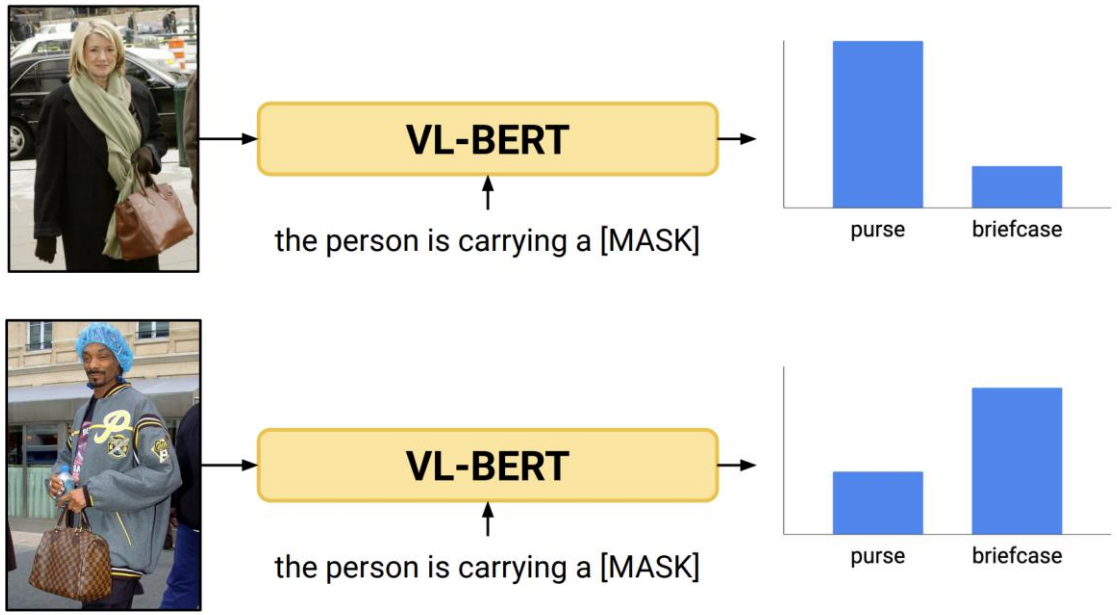


[Hendricks et al., Women also Snowboard: Overcoming Bias in Captioning Models. ECCV 2018]

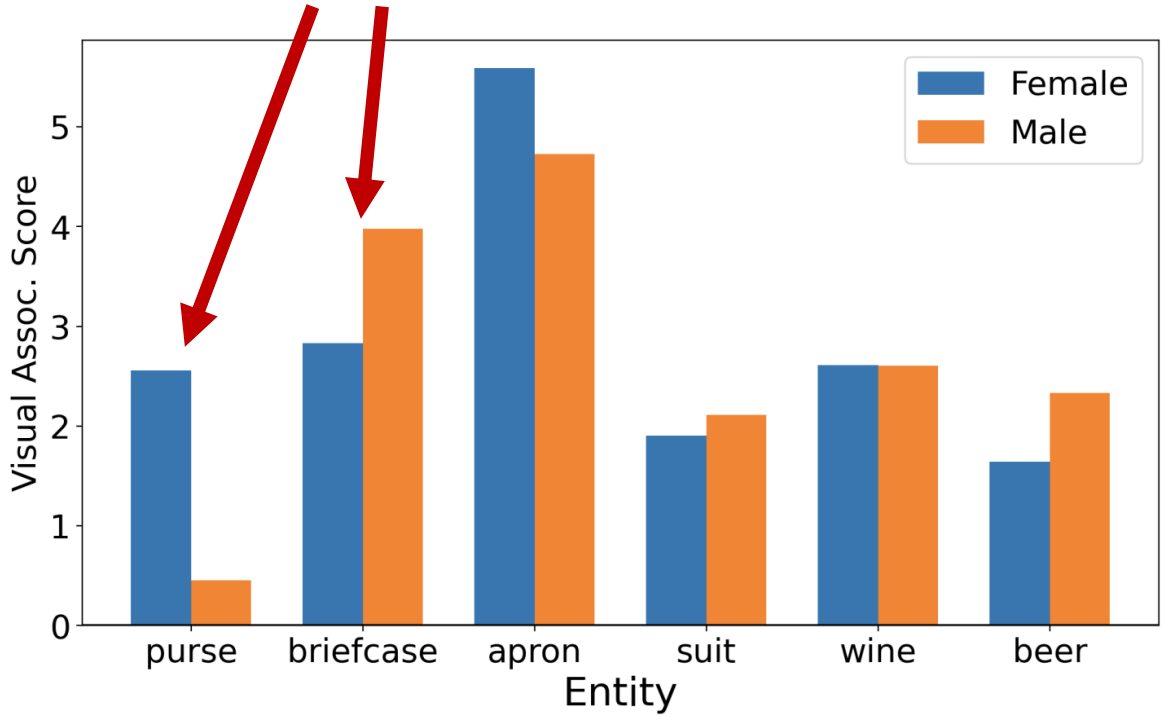
Modality Biases

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



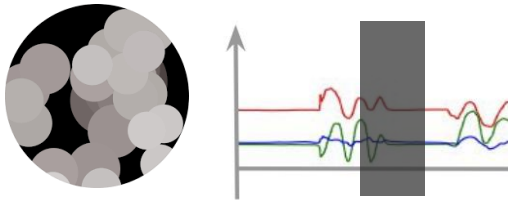
[Srinivasan and Bisk, Worst of Both Worlds: Biases Compound in Pre-trained Vision-and-Language Models. NAACL 2022]

Noise Topologies and Robustness

Heterogeneity in noise

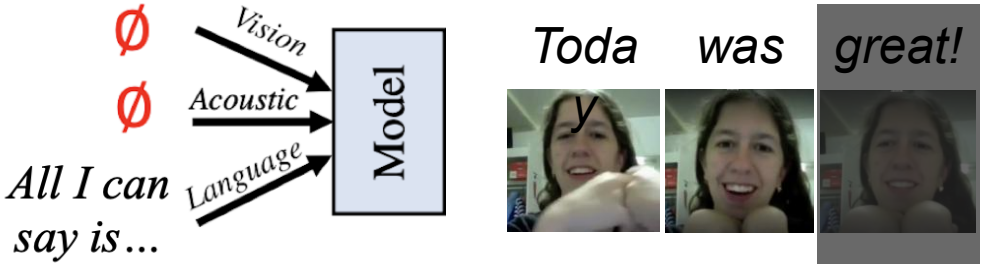
Modality-specific robustness

noise ⟶ nosie



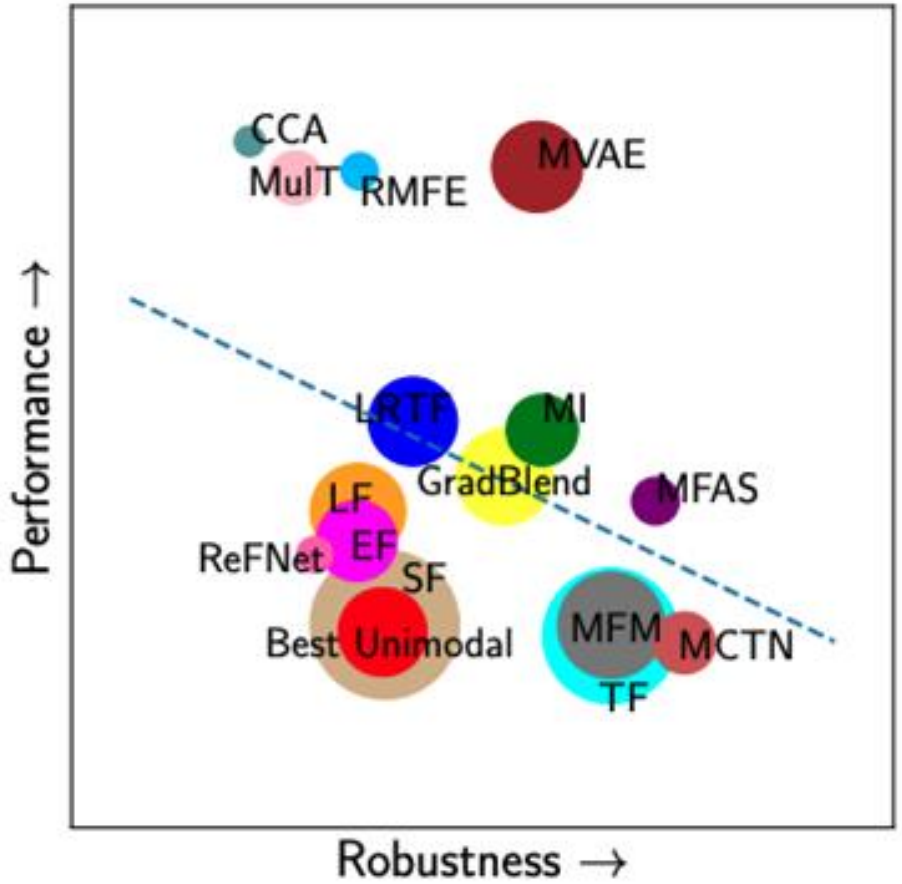
[Belinkov & Bisk, 2018; Subramaniam et al., 2009; Boyat & Joshi, 2015]

Multimodal robustness



[Zadeh et al., 2020]

Strong tradeoffs between performance and robustness

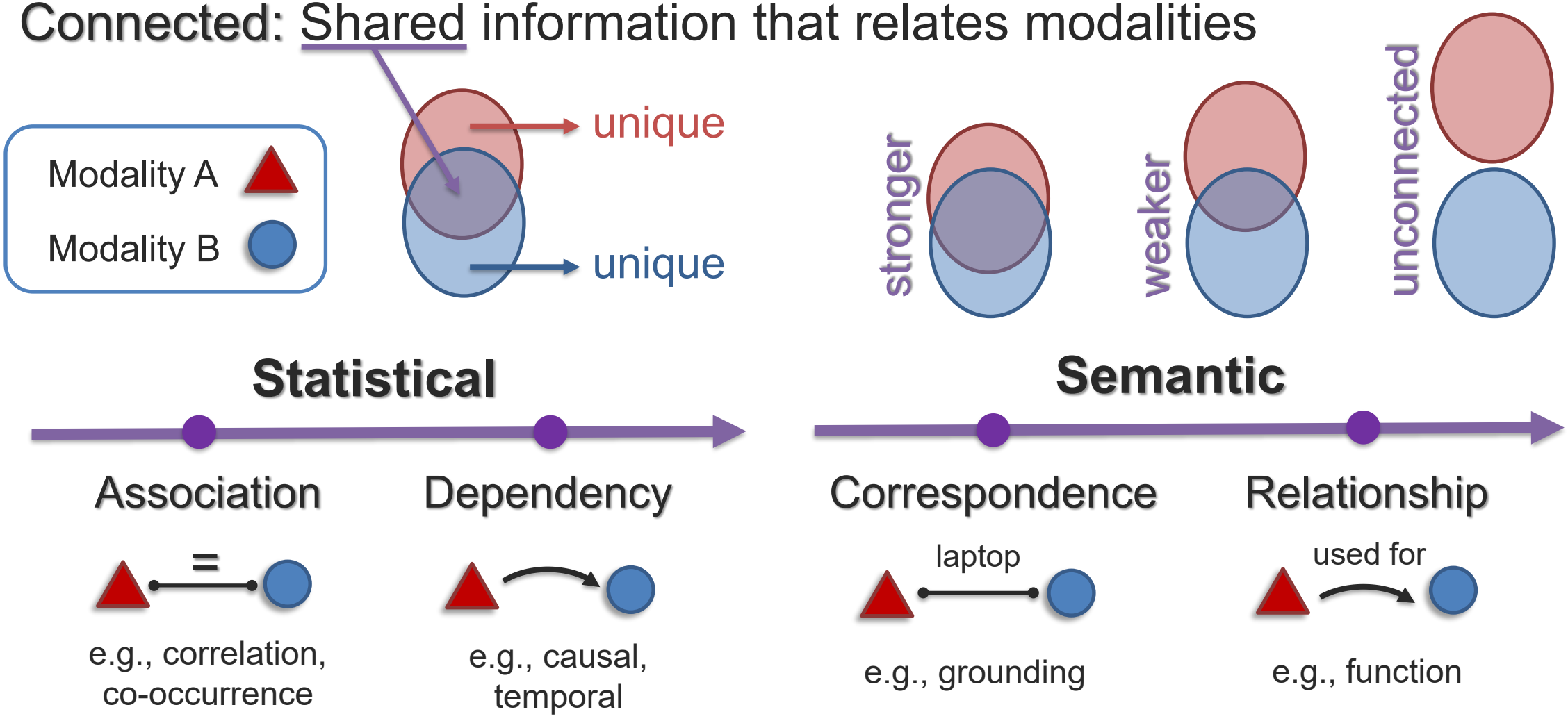


rate of accuracy drops

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

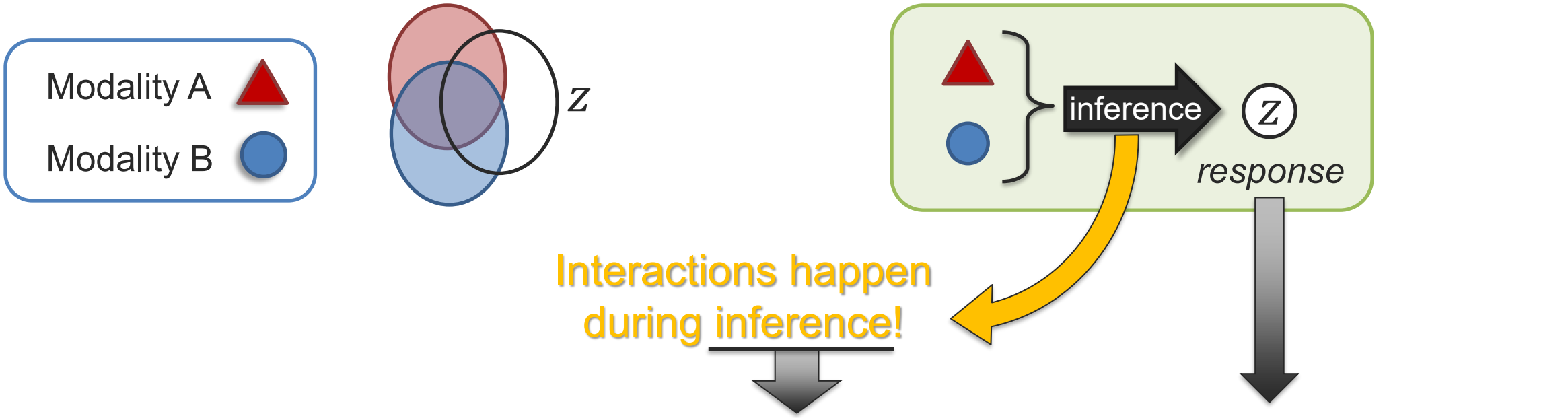
Sub-Challenge 6b: Cross-modal Connections

Connected: Shared information that relates modalities



Sub-Challenge 6b: Cross-modal Interactions

Interacting: process affecting each modality, creating new response



"Inference" examples:

- Representation fusion
- Prediction task
- Modality translation

representation

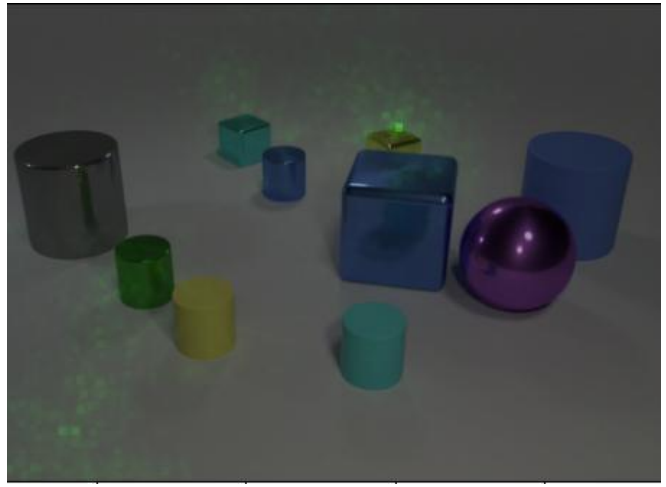
$\hat{y}$ *prediction*

modality C

Quantifying Cross-modal Interactions

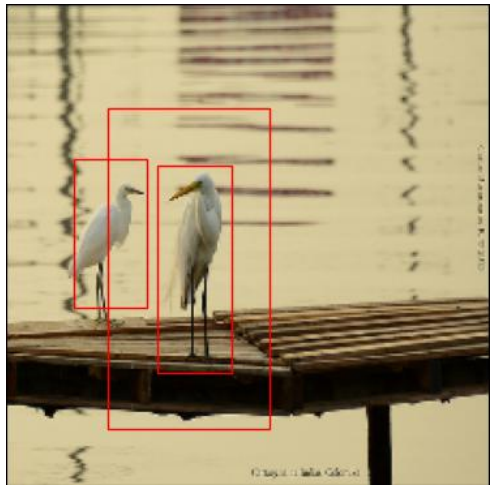
Identifying individual cross-modal interactions

CLEVR



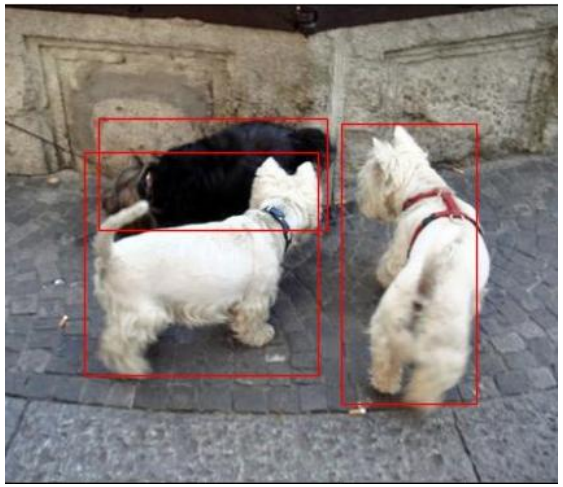
*The other small shiny thing that is the same shape as the **tiny yellow shiny object** is what color?*

VQA 2.0



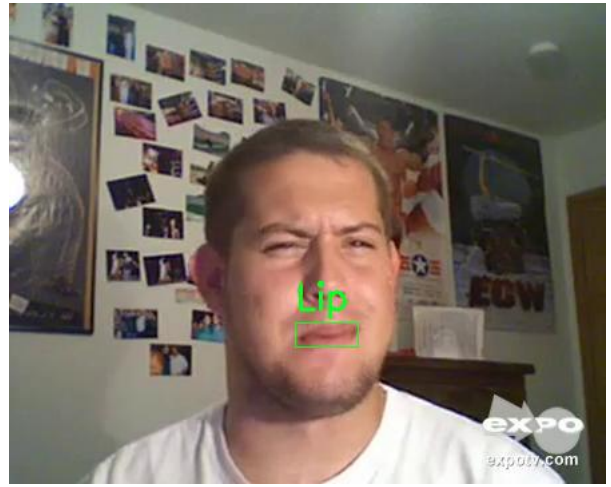
*How many **birds**?*

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**…*

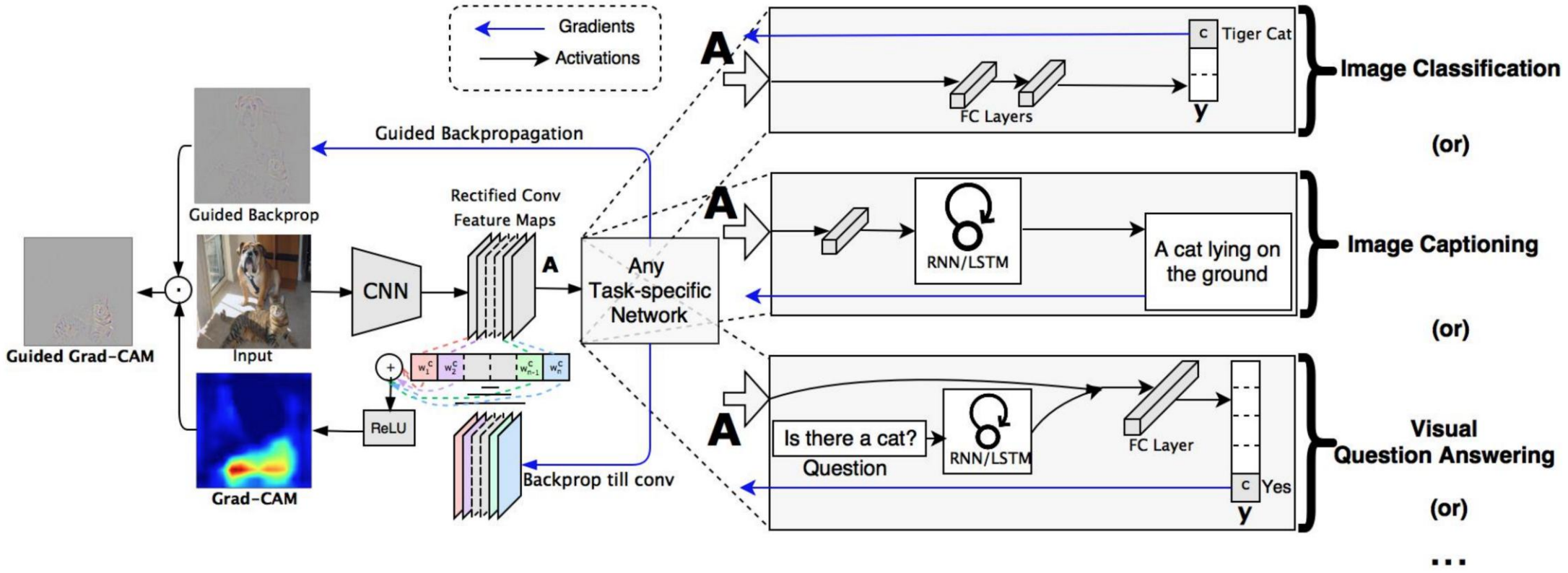
Correspondence

Relationships

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

Quantifying Cross-modal Interactions

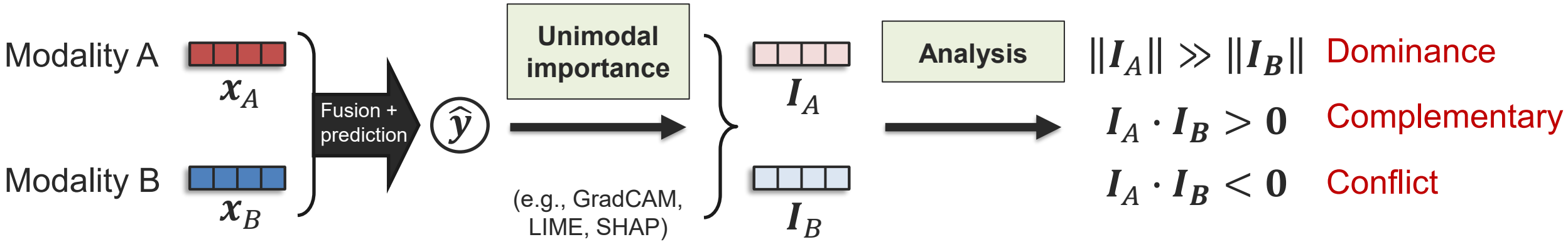
Gradients can compute the influence of the input modalities on the loss

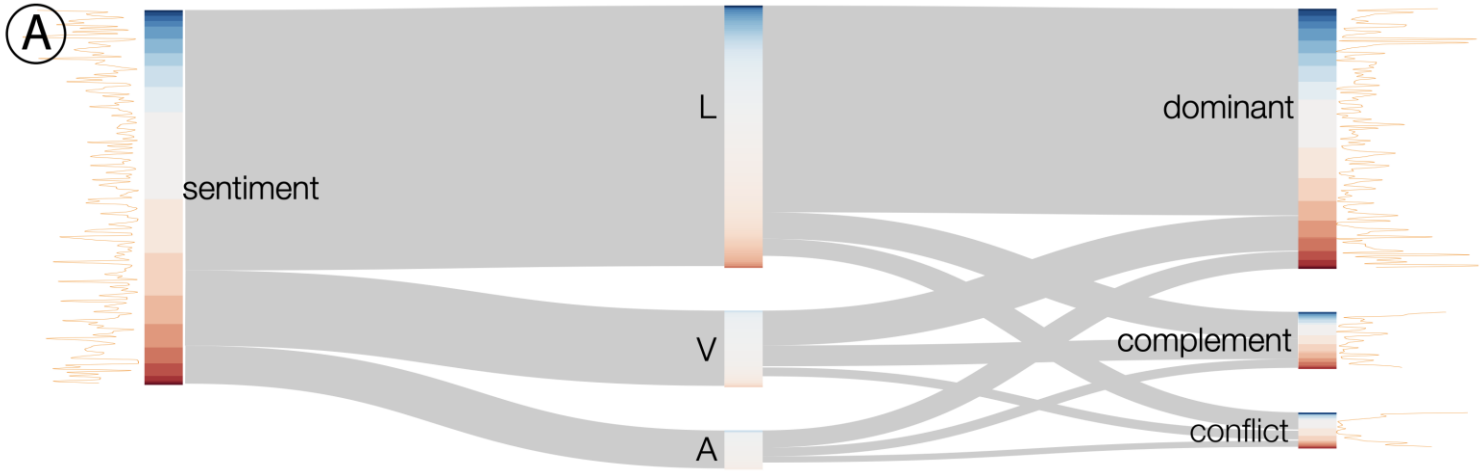


[Selvaraju et al. et al., GradCAM: Visual Explanations from Deep Networks via Gradient-based Localization]

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]

Modeling Multimodal Interactions

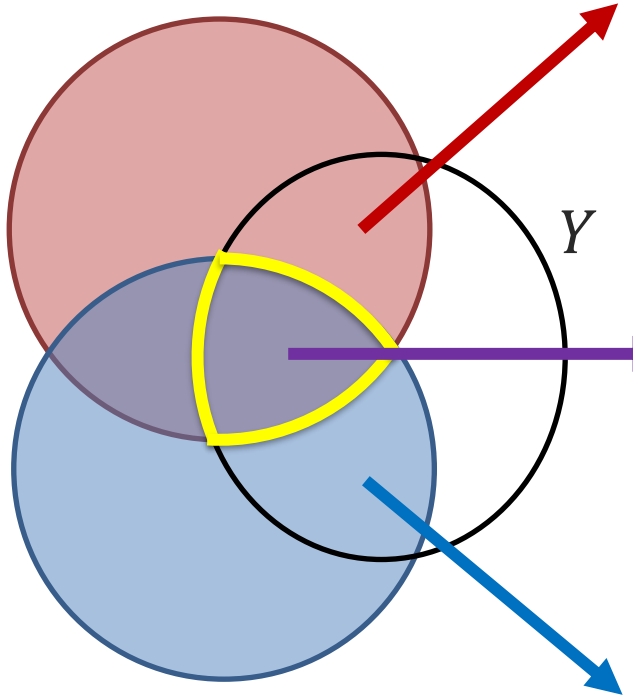
Partial Information Decomposition (PID)

$I(\triangle;\bigcirc;Y) = R - S$

factorizes 3-way mutual information into:

R: redundancy

S: Synergy



Task-relevant **unique** information

$I(\triangle;Y|\bigcirc)$

Task-relevant **shared** information

$I(\triangle;\bigcirc;Y)$

…but how to compute 3-way mutual information?

Task-relevant **unique** information

$I(\bigcirc;Y|\triangle)$

More details:
Liang et al., Quantifying & Modeling Feature Interactions: An Information Decomposition Framework. Neurips 2023

Quantifying Multimodal Datasets

1. Dataset quantification:

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

Language: *And he I don't think he got mad when hah I don't know maybe.*

Vision: Gaze aversion

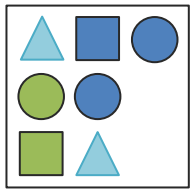


Acoustic: (frustrated voice)

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?

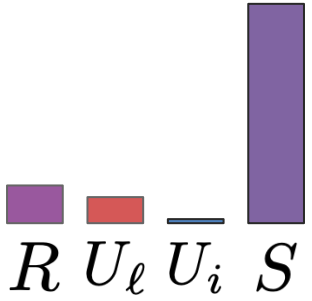
Sentiment



Sarcasm



VQA



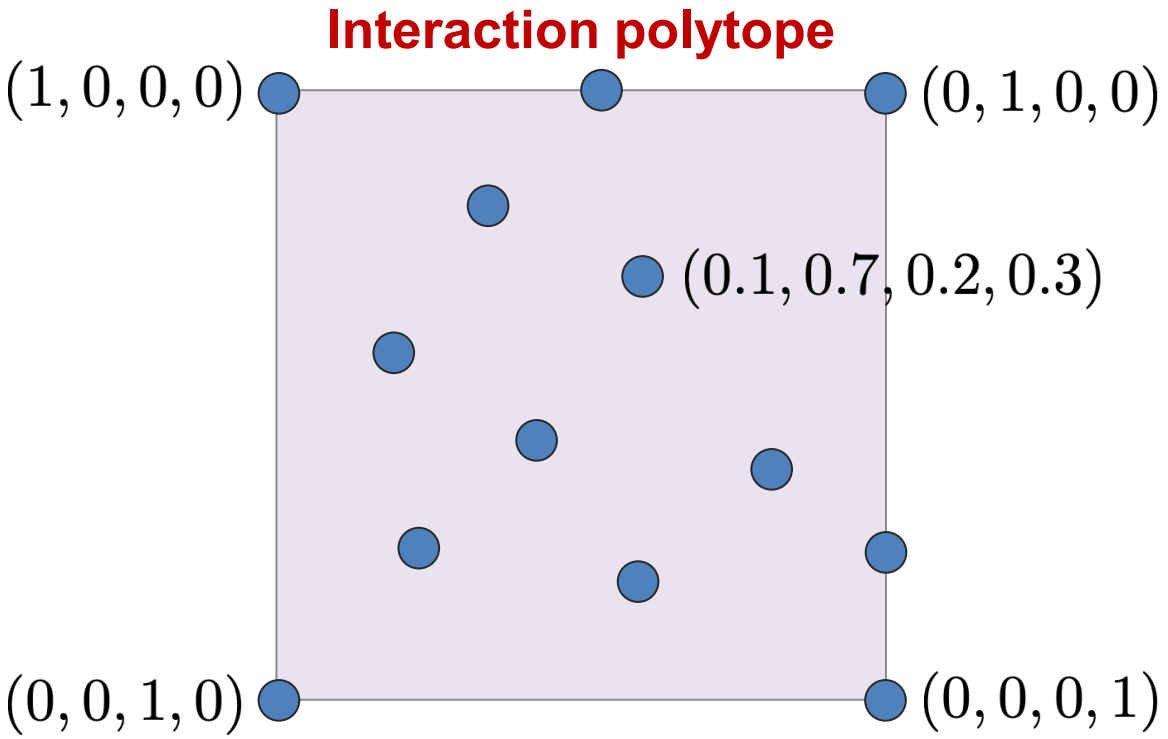
Also matches human judgment of interactions, and other sanity checks on synthetic datasets

[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

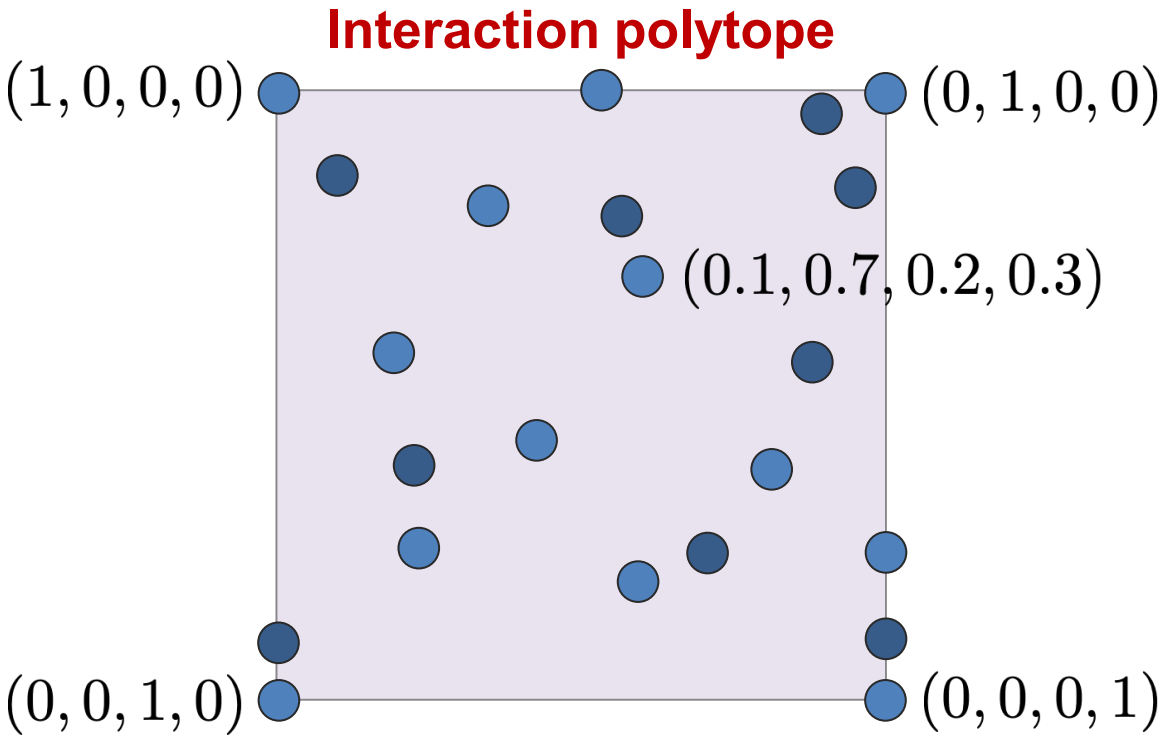
[Liang et al., Quantifying & Modeling Multimodal Interactions: An Information Decomposition Framework. NeurIPS 2023]

Model Selection

2. Model quantification:

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

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



Model families trained on synthetic data

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

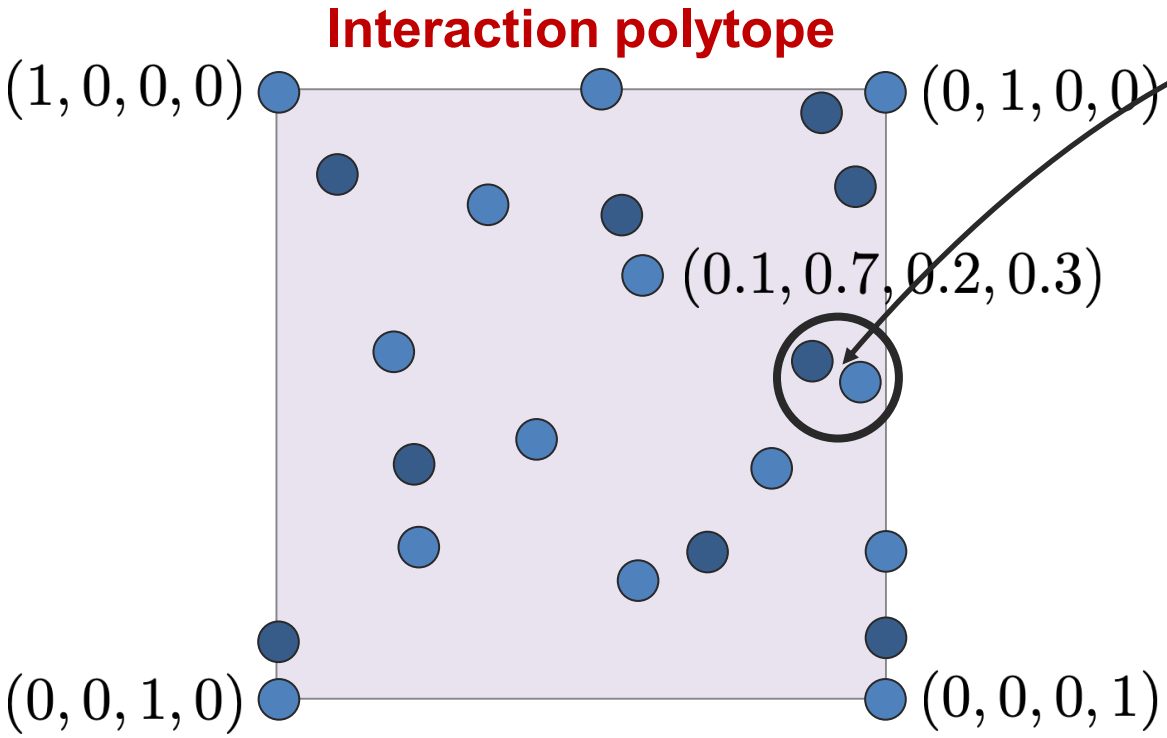
[Liang et al., Quantifying & Modeling Multimodal Interactions: An Information Decomposition Framework. NeurIPS 2023]

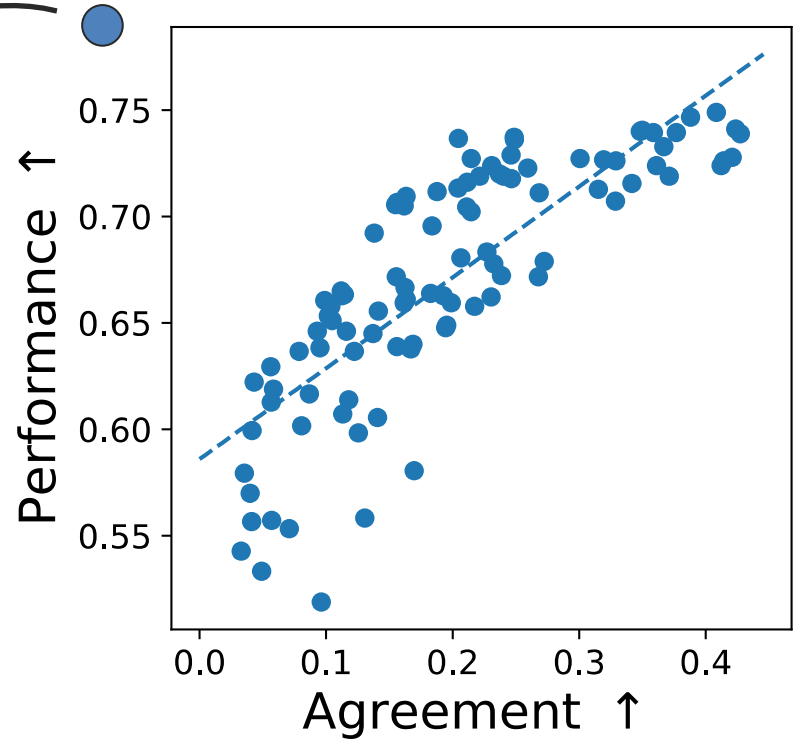
Model Selection

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:

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

Language: *And he I don't think he got mad when hah I don't know maybe.*

Vision: Gaze aversion



Acoustic: (frustrated voice)

Sentiment



Language/Agreement

Sheldon :

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

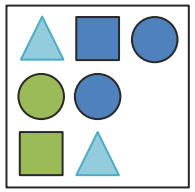
- **Text** : suggests a compliment.
- **Audio** : neutral tone.
- **Video** : straight face.



Sarcasm

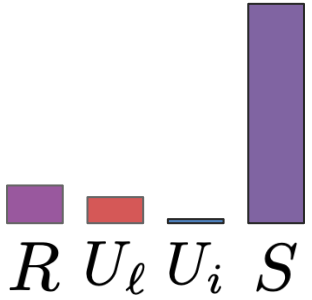


Multimodal Transformer



Is there a red shape above a circle?

VQA



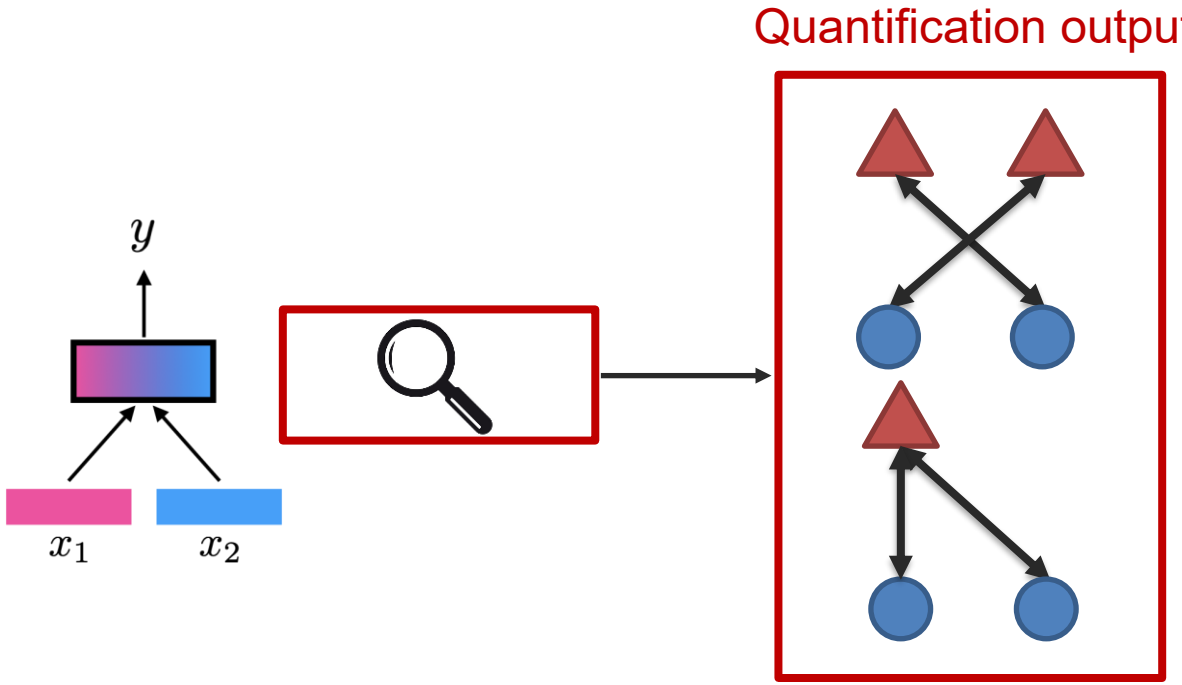
Multiplicative/Transformer

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

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!

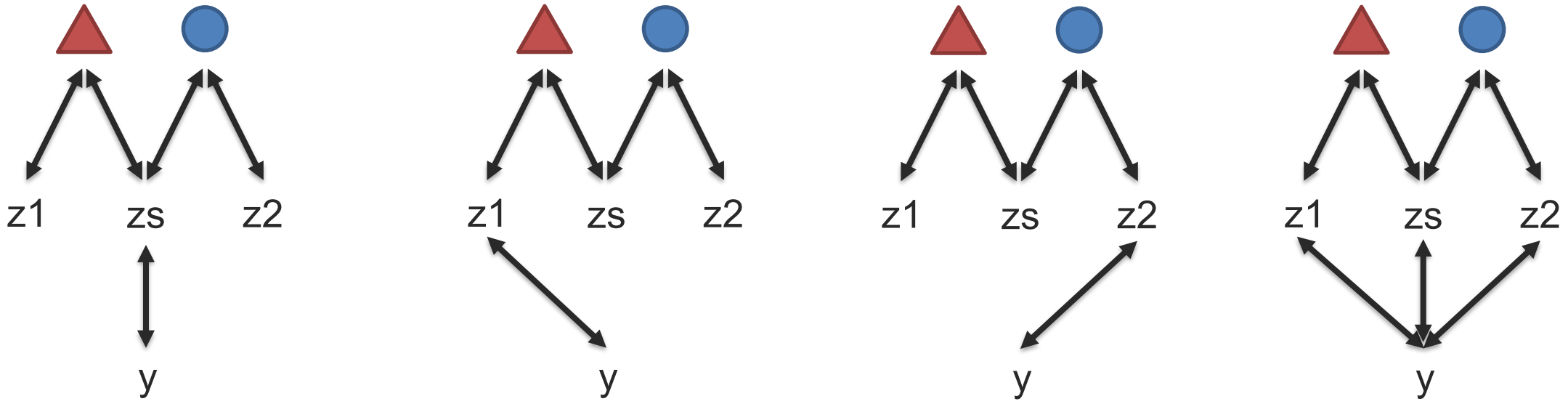


[Liang et al., MultiViz: Towards Visualizing and Understanding Multimodal Models. ICLR 2023]

Directly Evaluating Quantification

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

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



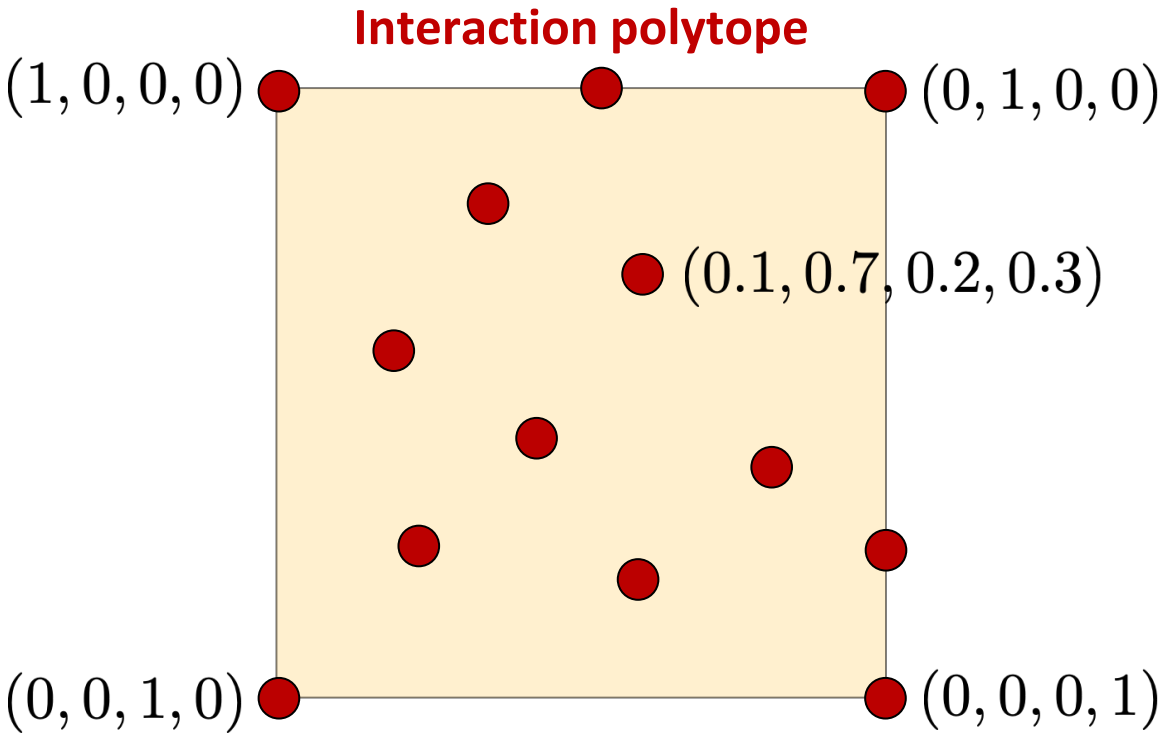
Redundancy | Unique 1 | Unique 2 | Synergy

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

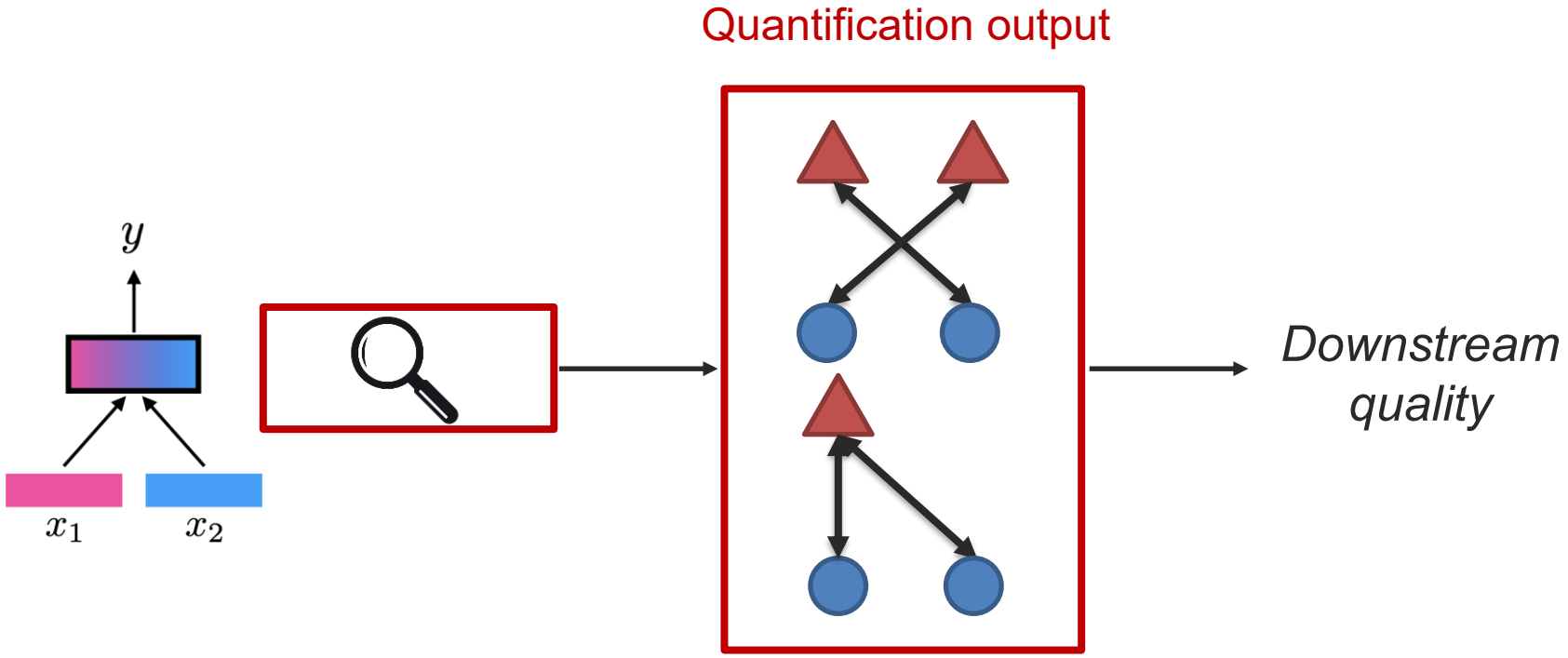


[Liang et al., MultiViz: Towards Visualizing and Understanding Multimodal Models. ICLR 2023]

Indirectly Evaluating Quantification

Indirect evaluation

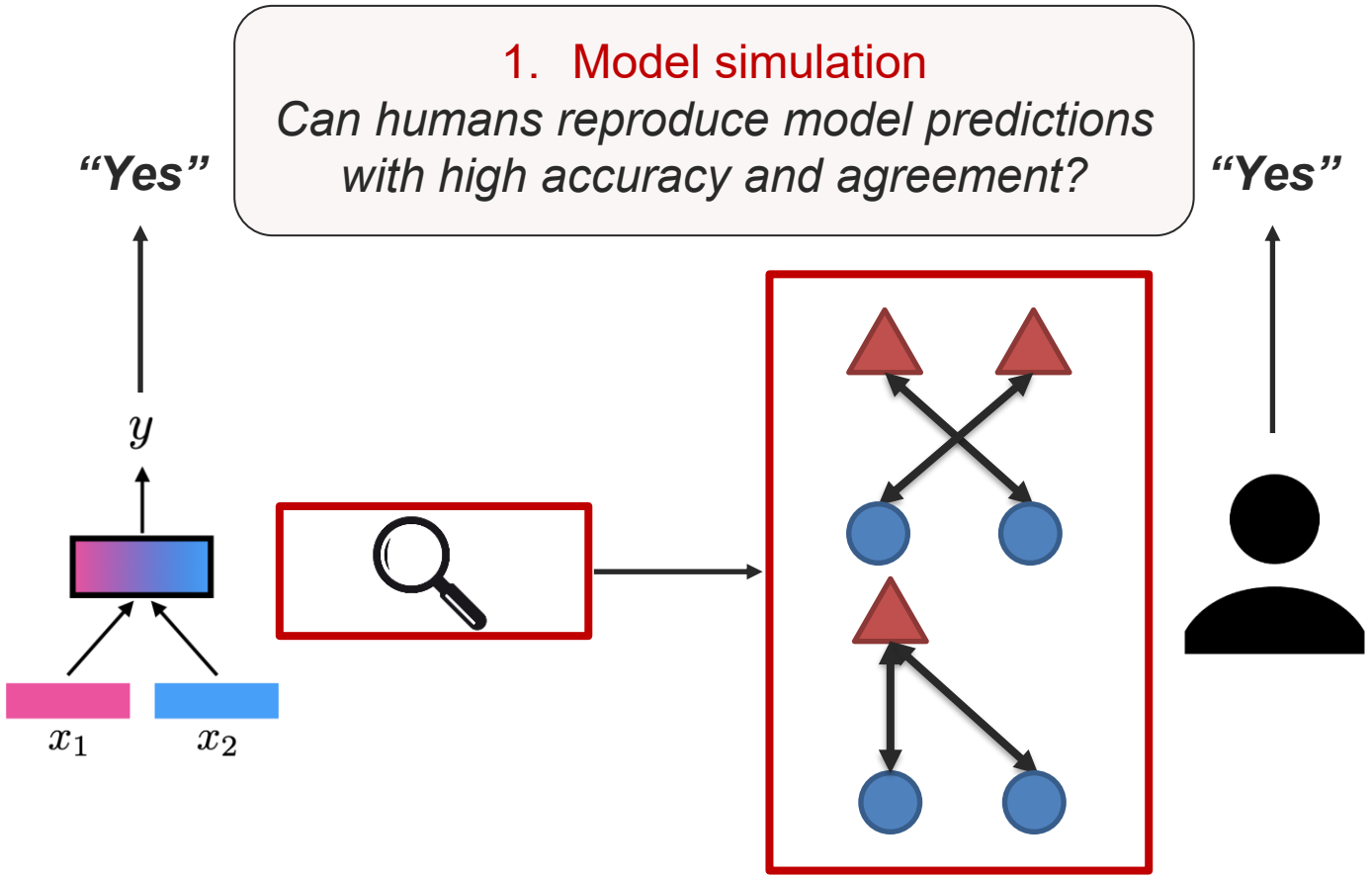
Find some downstream quality that practitioners find useful and can be easily evaluated.



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

Indirectly Evaluating Quantification

Indirect evaluation: Model simulation

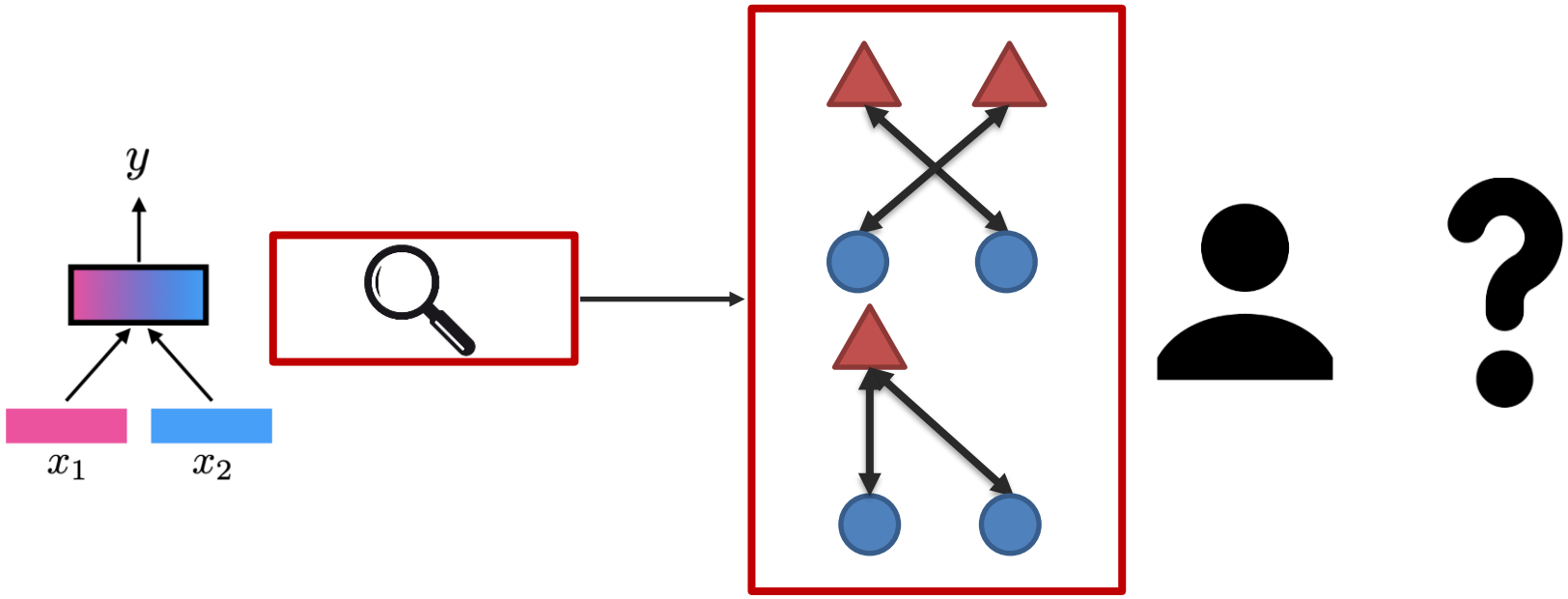


[Liang et al., MultiViz: Towards Visualizing and Understanding Multimodal Models. ICLR 2023]

Challenges: Quantifying Multimodal Interactions

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



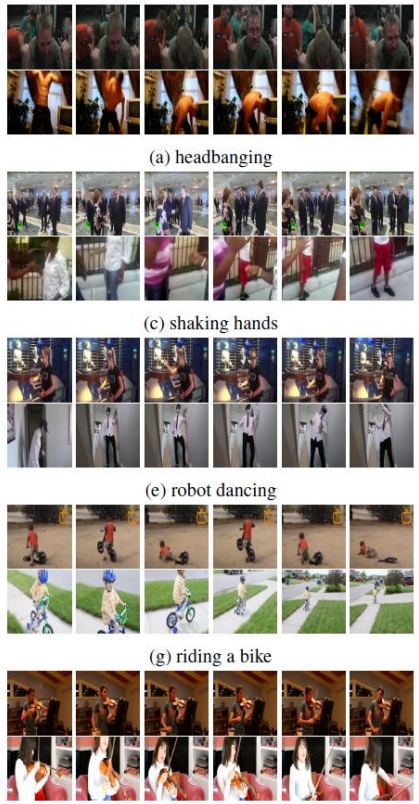
[Chandrasekaran et al., Do explanations make VQA models more predictable to a human? EMNLP 2018]

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

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?**

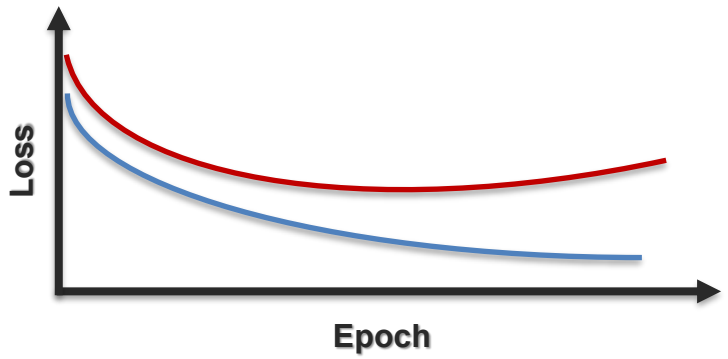
[Wang et al., What Makes Training Multi-modal Classification Networks Hard? CVPR 2020]

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

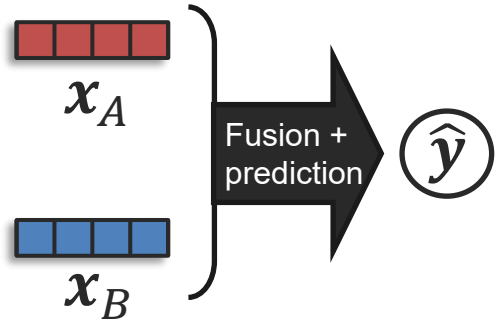
OGR wrt each modality tells us how much to train that modality

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

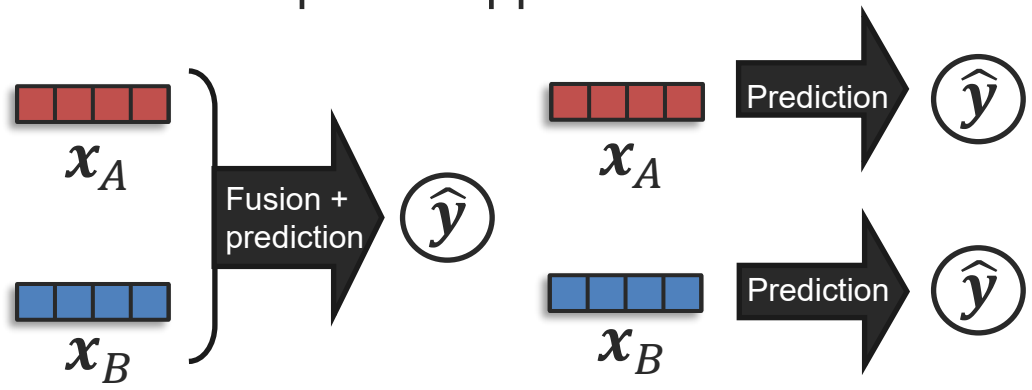
Optimization challenges

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]

Core Multimodal Challenges

