

# Jointly Modeling Inter- & Intra- Modality Dependencies for Multi-modal Learning

Divyam Madaan

Deep Learning Classics and Trends

Joint work with Taro Makino, Sumit Chopra, Kyunghyun Cho



NYU



Genentech



Paper



GitHub



# Background about me





# Map multiple modalities to the target

Input contains **multiple modalities** ( $\mathbf{x}, \mathbf{x}'$ ) and **target label**  $\mathbf{y}$

# Map multiple modalities to the target

Input contains **multiple modalities** ( $\mathbf{x}, \mathbf{x}'$ ) and **target label**  $\mathbf{y}$



Image  $\mathbf{x}$

Text  $\mathbf{x}'$

What sport field is in the background?



# Map multiple modalities to the target

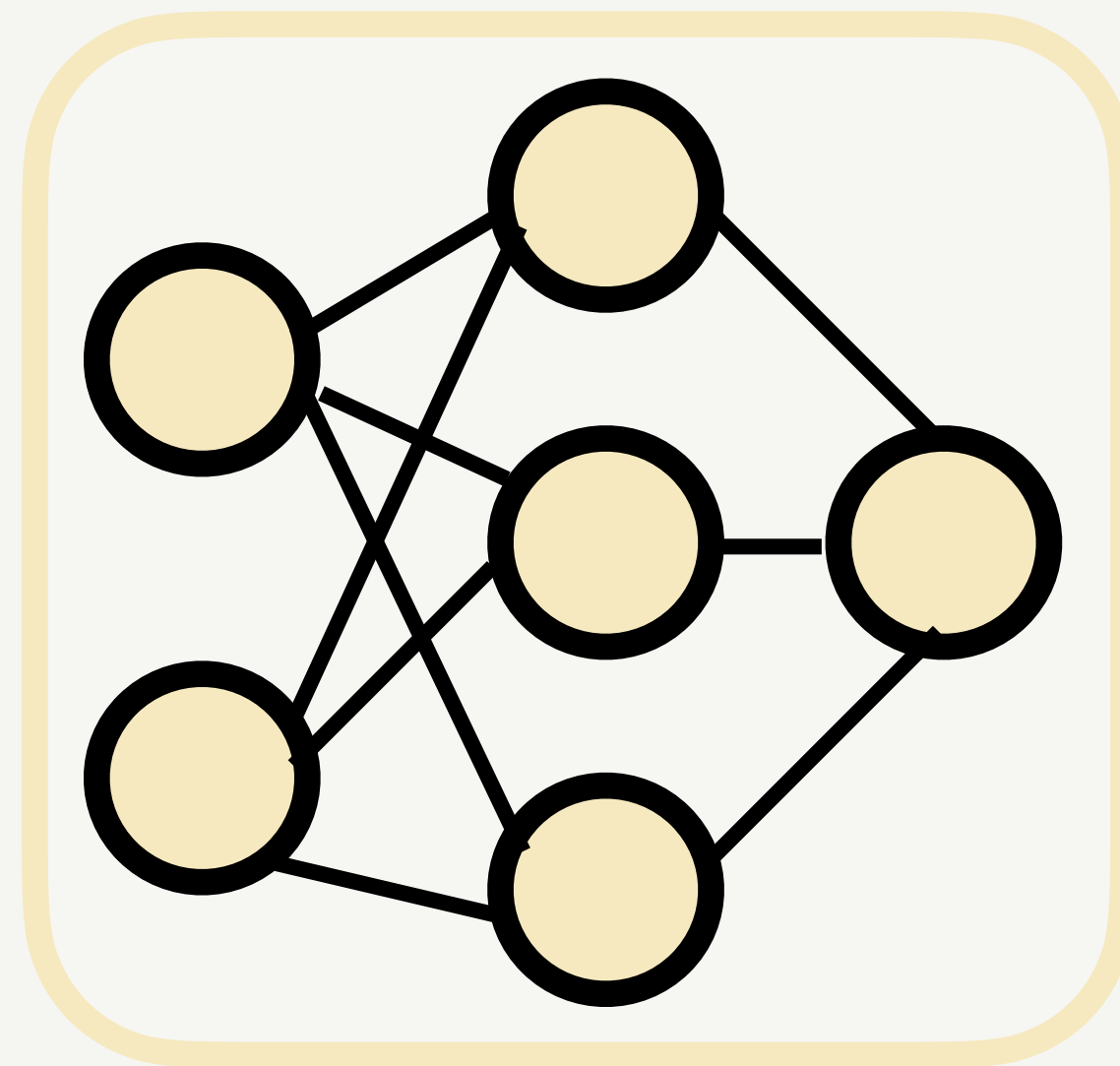
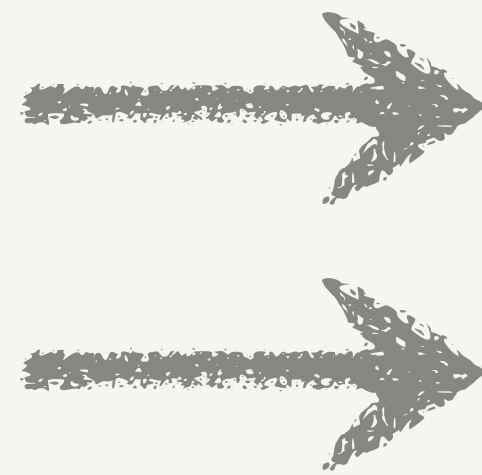
Input contains **multiple modalities** ( $\mathbf{x}$ ,  $\mathbf{x}'$ ) and **target label**  $\mathbf{y}$



Image  $\mathbf{x}$

Text  $\mathbf{x}'$

What sport field is in the background?



# Map multiple modalities to the target

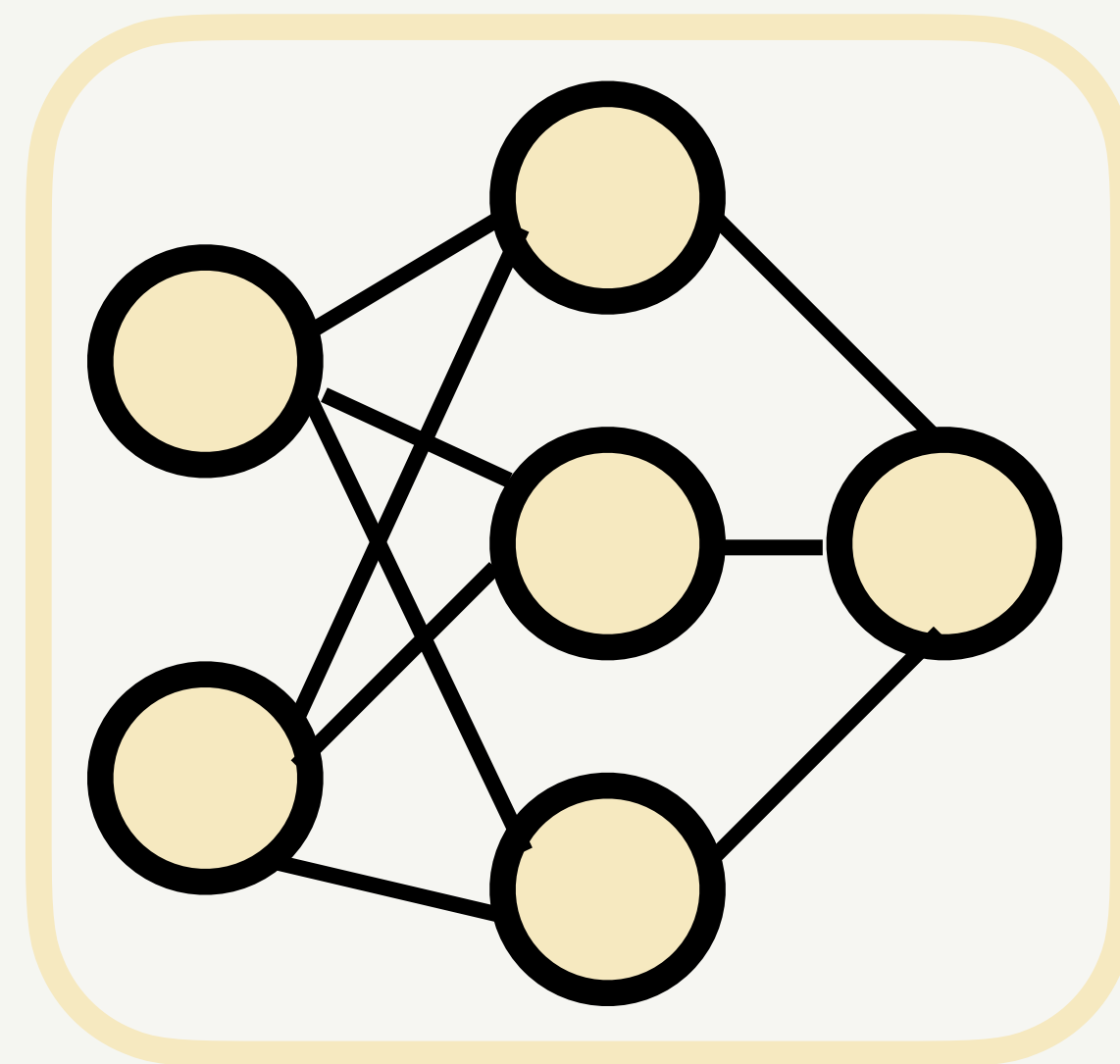
Input contains **multiple modalities** ( $\mathbf{x}$ ,  $\mathbf{x}'$ ) and **target label**  $\mathbf{y}$



Image  $\mathbf{x}$

Text  $\mathbf{x}'$

What sport field is in the background?



Answer,  $\mathbf{y}$

Kiting

Soccer

Tennis

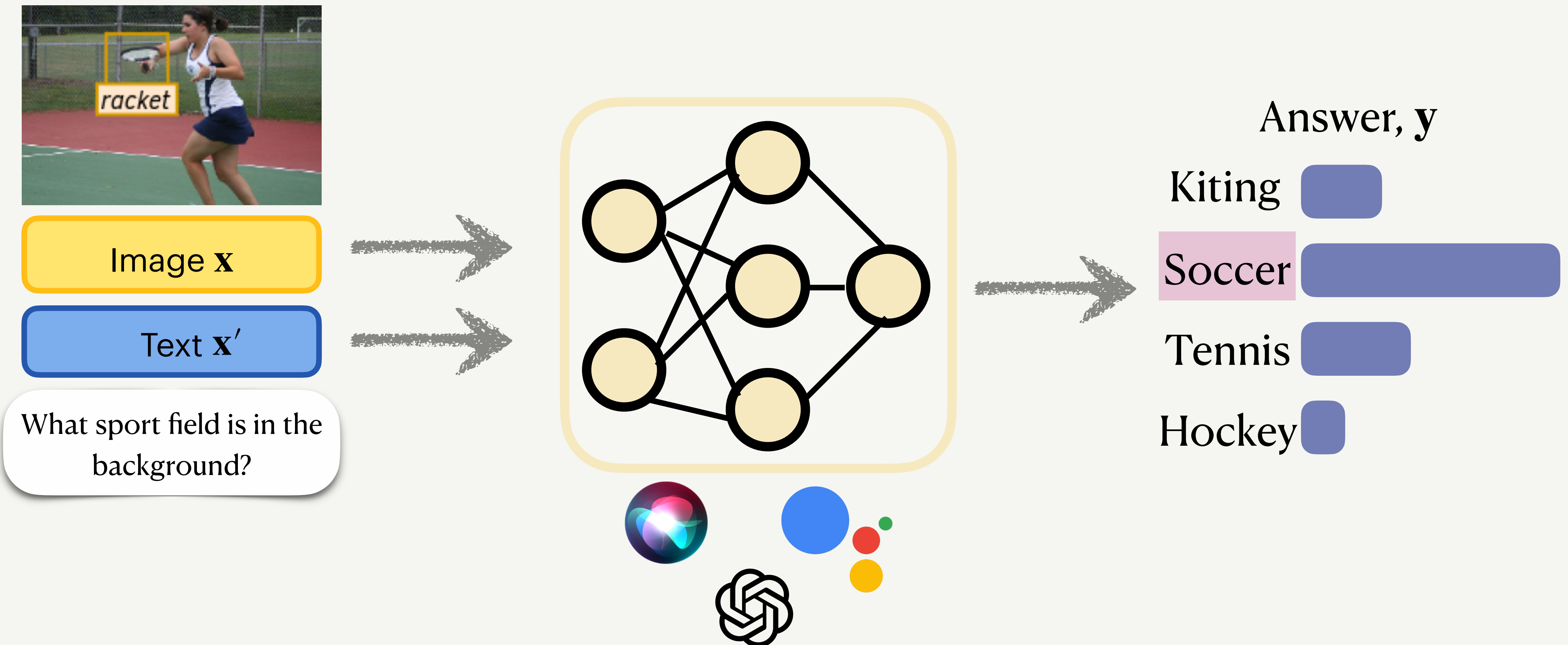
Hockey





# Map multiple modalities to the target

Input contains **multiple modalities** ( $\mathbf{x}$ ,  $\mathbf{x}'$ ) and **target label**  $\mathbf{y}$



# Multi-modal learning: success stories

Usage of multiple modalities has shown **success** in various applications

## LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models

Feng Li<sup>1,2\*</sup>, Renrui Zhang<sup>1,3\*</sup>, Hao Zhang<sup>1,2\*</sup>, Yuanhan Zhang<sup>1,4</sup>, Bo Li<sup>1,4</sup>, Wei Li<sup>1</sup>, Zejun Ma<sup>1</sup>, Chunyuan Li<sup>1</sup>

<sup>1</sup> ByteDance <sup>2</sup> HKUST <sup>3</sup> CUHK <sup>4</sup> NTU \* Core contributor

## Multimodal biomedical AI

Julián N. Acosta<sup>1</sup>, Guido J. Falcone<sup>1</sup>, Pranav Rajpurkar<sup>2,4</sup> and Eric J. Topol<sup>3,4</sup>

*The New ChatGPT Can 'See' and 'Talk.' Here's What It's Like.*

The Next AI Frontier: How Multimodal Systems Are Reshaping Our World

## 4M-21: An Any-to-Any Vision Model for Tens of Tasks and Modalities

Roman Bachmann<sup>1†\*</sup> Oğuzhan Fatih Kar<sup>1\*</sup> David Mizrahi<sup>2†\*</sup> Ali Garjani<sup>1</sup>  
Mingfei Gao<sup>2</sup> David Griffiths<sup>2</sup> Jiaming Hu<sup>2</sup> Afshin Dehghan<sup>2</sup> Amir Zamir<sup>1</sup>

<sup>1</sup>Swiss Federal Institute of Technology Lausanne (EPFL) <sup>2</sup>Apple

# Multimodal: AI's new frontier



# Multi-modal learning: when More Isn't Better

Unimodal models and their combinations obtain **better performance** in many use-cases

## RUBi: Reducing Unimodal Biases for Visual Question Answering

Remi Cadene<sup>1,\*</sup>, Corentin Dancette<sup>1,\*</sup>, Hedi Ben-younes<sup>1</sup>, Matthieu Cord<sup>1</sup>, Devi Parikh<sup>2,3</sup>

## Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs

Shengbang Tong<sup>1</sup>    Zhuang Liu<sup>2</sup>    Yuexiang Zhai<sup>3</sup>  
Yi Ma<sup>3</sup>    Yann LeCun<sup>1</sup>    Saining Xie<sup>1</sup>

<sup>1</sup>New York University    <sup>2</sup>FAIR, Meta    <sup>3</sup>UC Berkeley

## Removing Bias in Multi-modal Classifiers: Regularization by Maximizing Functional Entropies

Itai Gat  
Technion

Idan Schwartz  
Technion

Alexander Schwing  
UIUC

Tamir Hazan  
Technion

## Characterizing and Overcoming the Greedy Nature of Learning in Multi-modal Deep Neural Networks

Nan Wu<sup>1</sup>    Stanisław Jastrzębski<sup>2,1</sup>    Kyunghyun Cho<sup>1,3,4,5</sup>    Krzysztof J. Geras<sup>2,1,3</sup>

## What Makes Training Multi-modal Classification Networks Hard?

Weiyao Wang, Du Tran, Matt Feiszli  
Facebook AI

# When multi-modal learning isn't better 🚫

Multi-modal learning **fails to solve the task** in many scenarios



what sport field is in the background?



# When multi-modal learning isn't better 🚫

Multi-modal learning **fails to solve the task** in many scenarios



what sport field is in the background?



The field in the background is a tennis court. The surface is painted with green and red sections, typical of many tennis court designs, and the player is holding a racket, indicating this is a tennis match or practice session.

# When multi-modal learning isn't better 🚫

Multi-modal learning **fails to solve the task** in many scenarios



what sport field is in the background?

The field in the background is a tennis court. The surface is painted with green and red sections, typical of many tennis court designs, and the player is holding a racket, indicating this is a tennis match or practice session.

do you see any soccer field in the image?



# When multi-modal learning isn't better 🚫

Multi-modal learning **fails to solve the task** in many scenarios

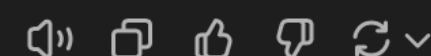


what sport field is in the background?

The field in the background is a tennis court. The surface is painted with green and red sections, typical of many tennis court designs, and the player is holding a racket, indicating this is a tennis match or practice session.

do you see any soccer field in the image?

There isn't a visible soccer field in this image. The court shown in the image has characteristics of a tennis court, as mentioned, with specific colored surfaces and line markings typical for tennis rather than soccer. The fence around the court and the player with a racket further indicate that this is likely a tennis setting, not a soccer field.



what sport is in the background?

image

ⓘ This image has been removed

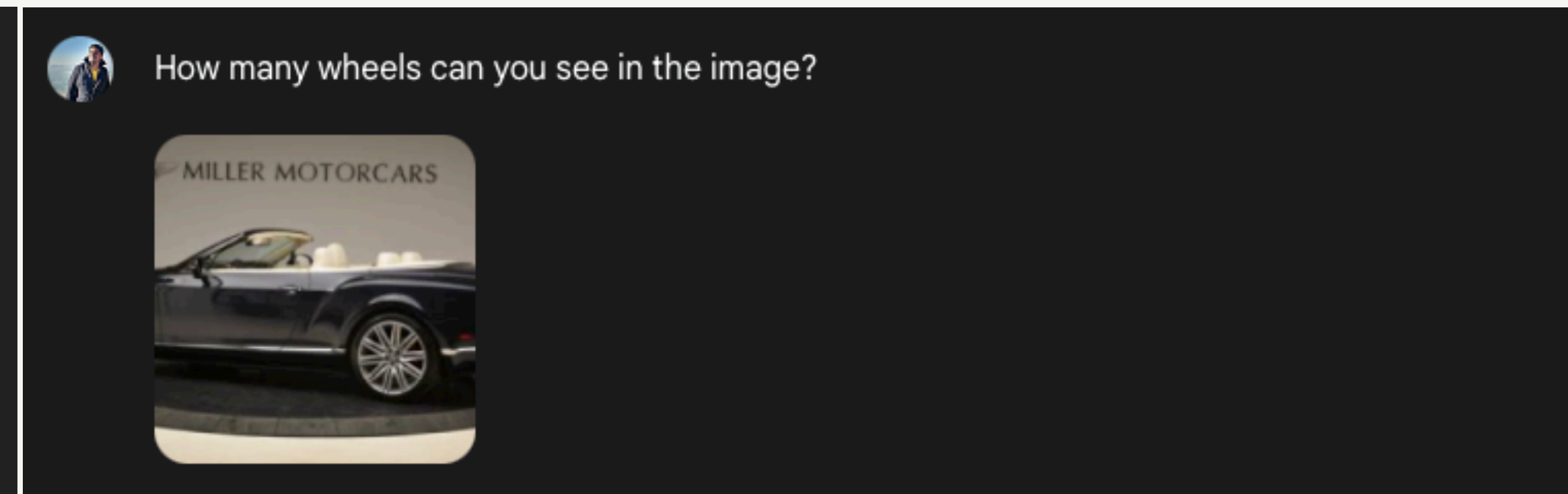
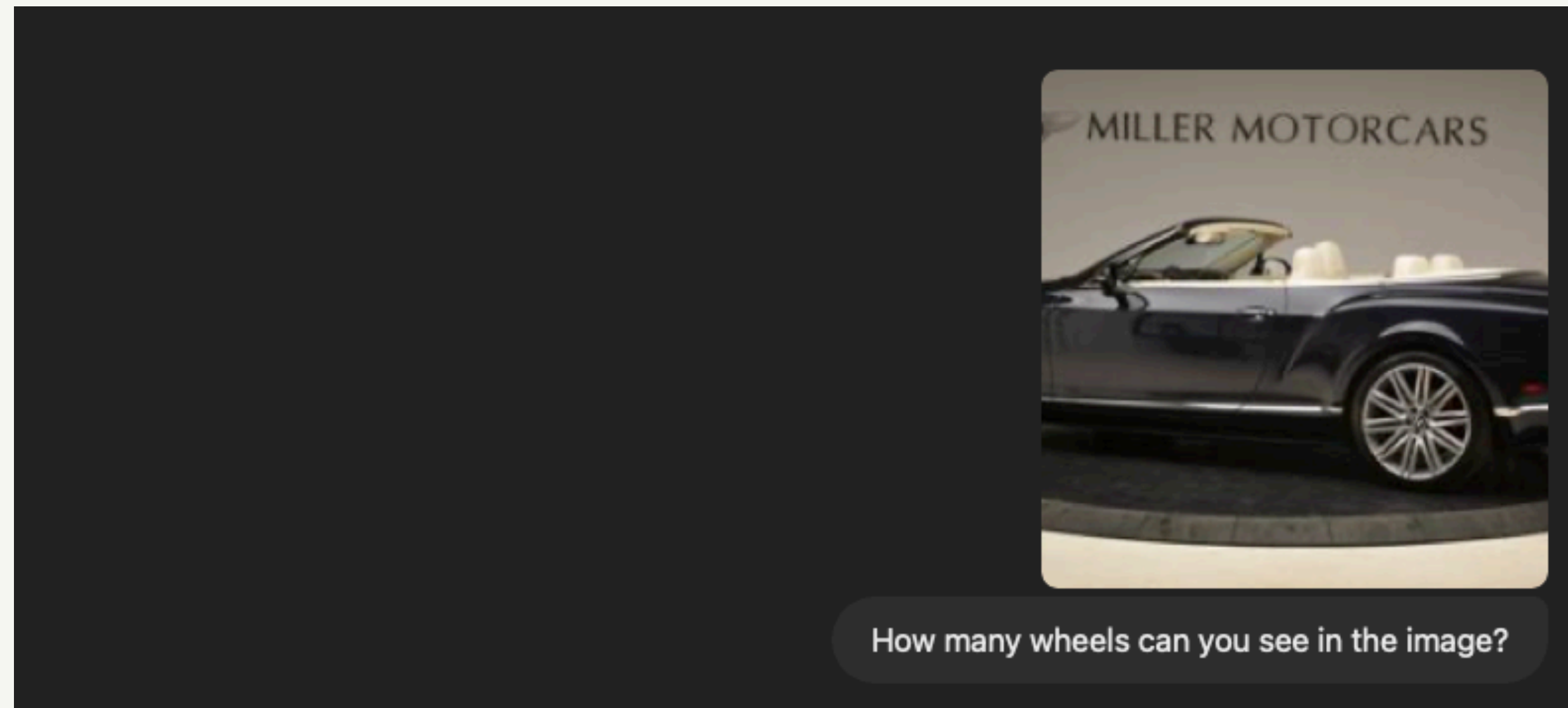


Sorry, I can't help with images of people yet.



# When multi-modal learning isn't better 🚫

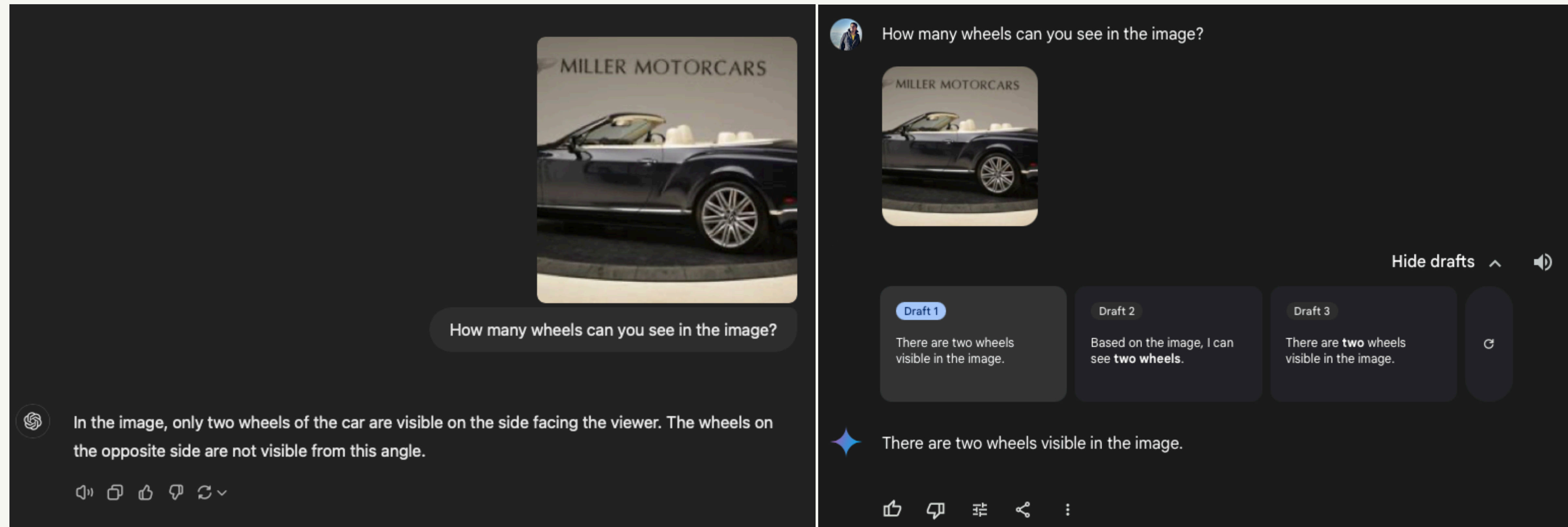
Multi-modal learning **fails to solve the task** in many scenarios





# When multi-modal learning isn't better

Multi-modal learning **fails to solve the task** in many scenarios



The image shows a side-by-side comparison of a multi-modal learning interface. On the left, a user asks "How many wheels can you see in the image?" with an image of a dark blue convertible car. The AI response states that only two wheels are visible from this angle. On the right, the same question is asked, but the interface shows three draft responses, all of which incorrectly state that two wheels are visible. The final AI response is also incorrect, stating that two wheels are visible.

**Left Panel (User Question):**

How many wheels can you see in the image?

**Left Panel (AI Answer):**

In the image, only two wheels of the car are visible on the side facing the viewer. The wheels on the opposite side are not visible from this angle.

**Right Panel (User Question):**

How many wheels can you see in the image?

**Right Panel (Drafts):**

- Draft 1: There are two wheels visible in the image.
- Draft 2: Based on the image, I can see **two wheels**.
- Draft 3: There are **two** wheels visible in the image.

**Right Panel (Final AI Answer):**

There are two wheels visible in the image.

Uncover underlying factor **for**  
**these discrepancies**



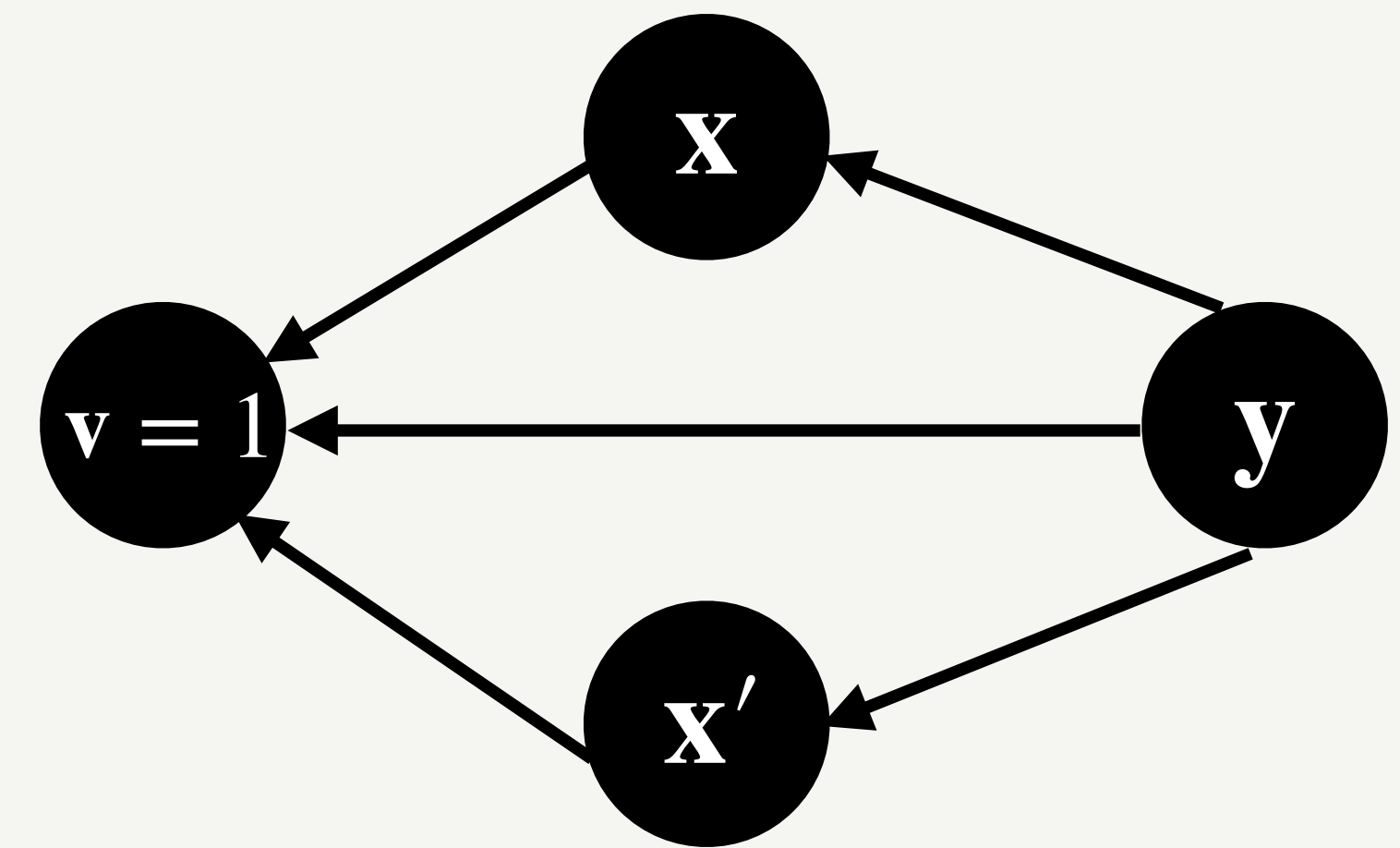
Uncover underlying factor **for**  
**these discrepancies**

Introduce **I2M2: a simple**  
**approach** for multi-modal learning

# What is Multi-modal Learning?

# Multi-modal learning

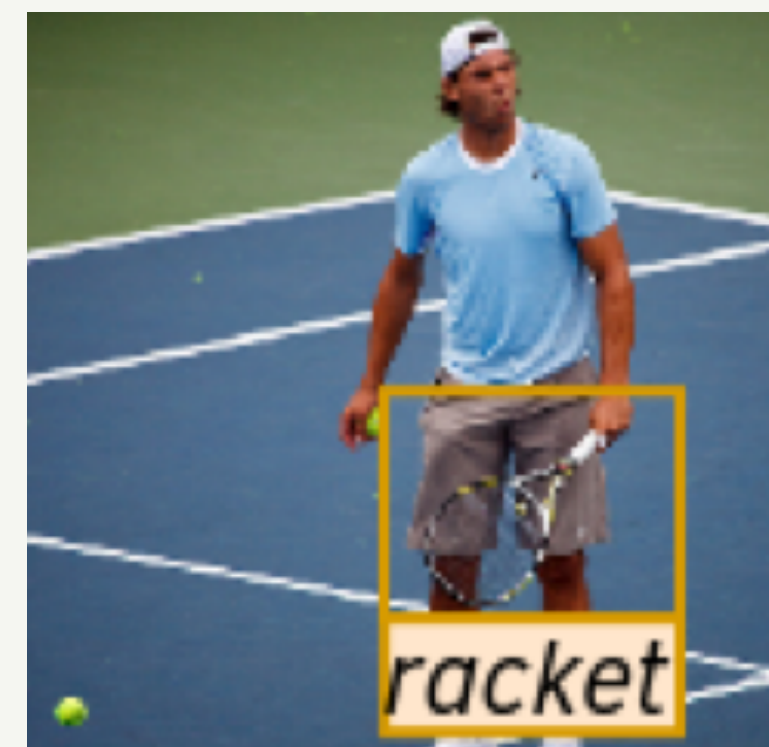
Multi-modal learning contains **inter- and intra-modality dependencies**



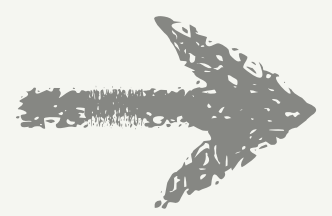


# Multi-modal learning contains **intra-modality dependencies**

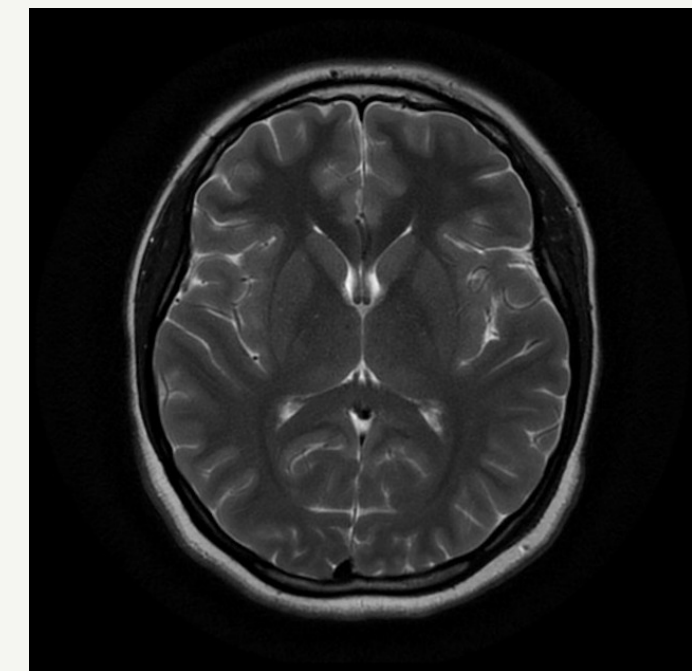
Dependency among the **features within each modality given the label**



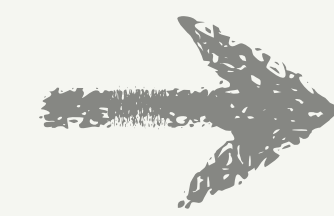
What sport is being played?



Tennis



-	-	-
-	-	-
-	-	-
-	-	-
-	-	-

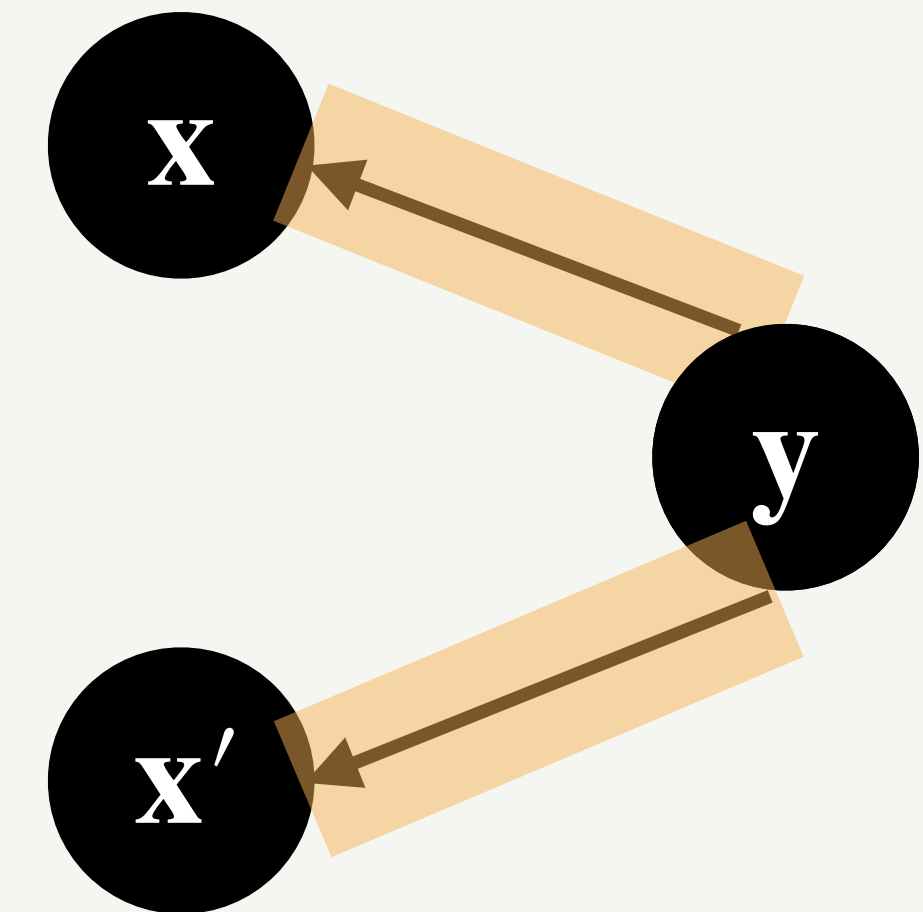


Alzheimer's (G30.9)

# Multi-modal learning contains **intra-modality dependencies**

Dependency among the **features within each modality given the label**

$$p(y, \mathbf{x}, \mathbf{x}') = p(y) p(\mathbf{x} | y) p(\mathbf{x}' | y)$$

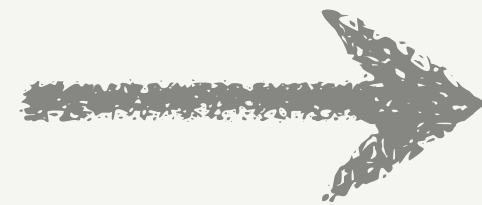


# Multi-modal learning contains **inter-modality dependencies**

Dependency among the **features across modalities given the label**



What sport field is in the background?



Soccer

We have built a smart robot. It understands a lot about images.

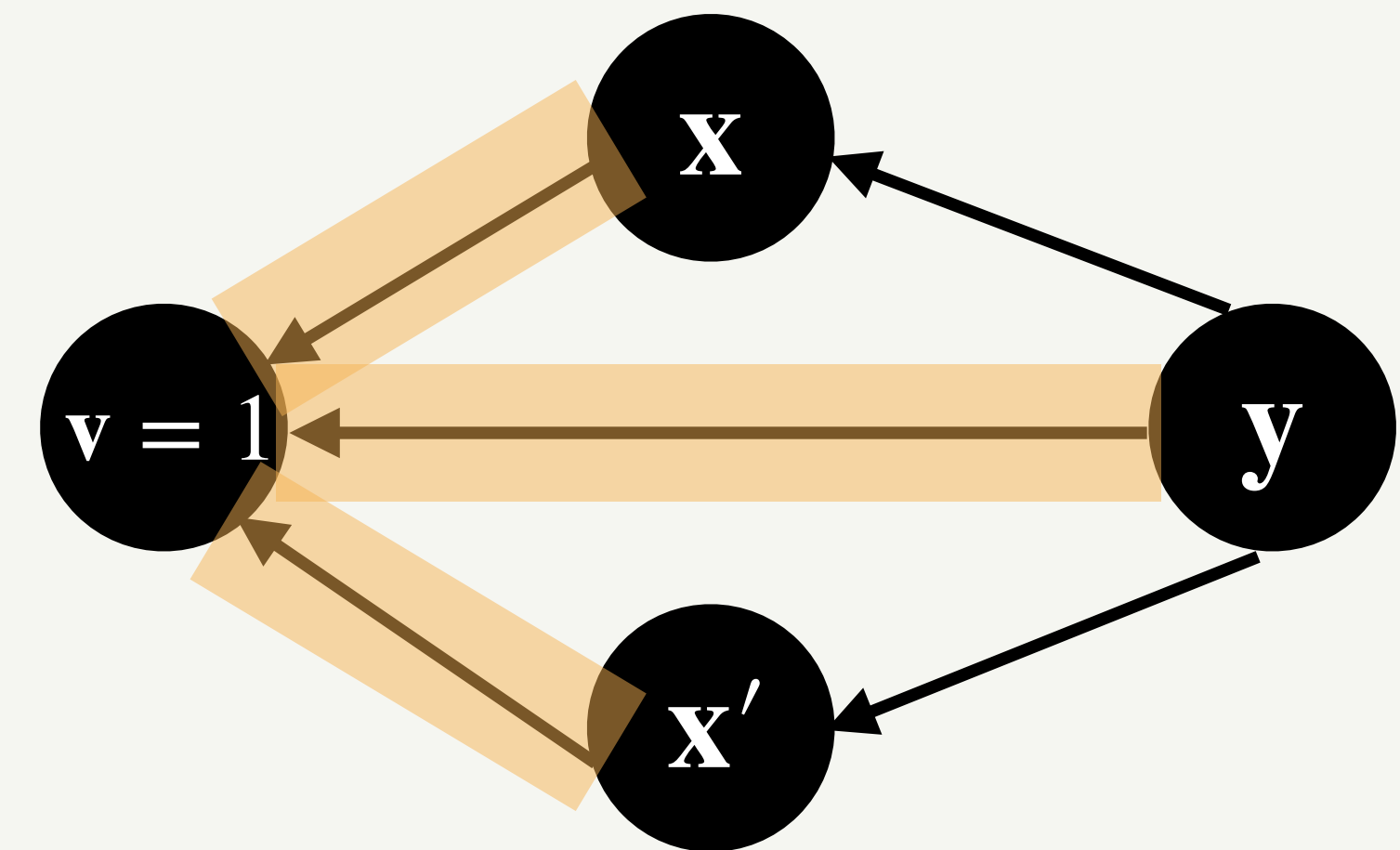
***Your task is to stump the smart robot***



# Multi-modal learning contains **inter-modality dependencies**

Dependency among the **features across modalities given the label**

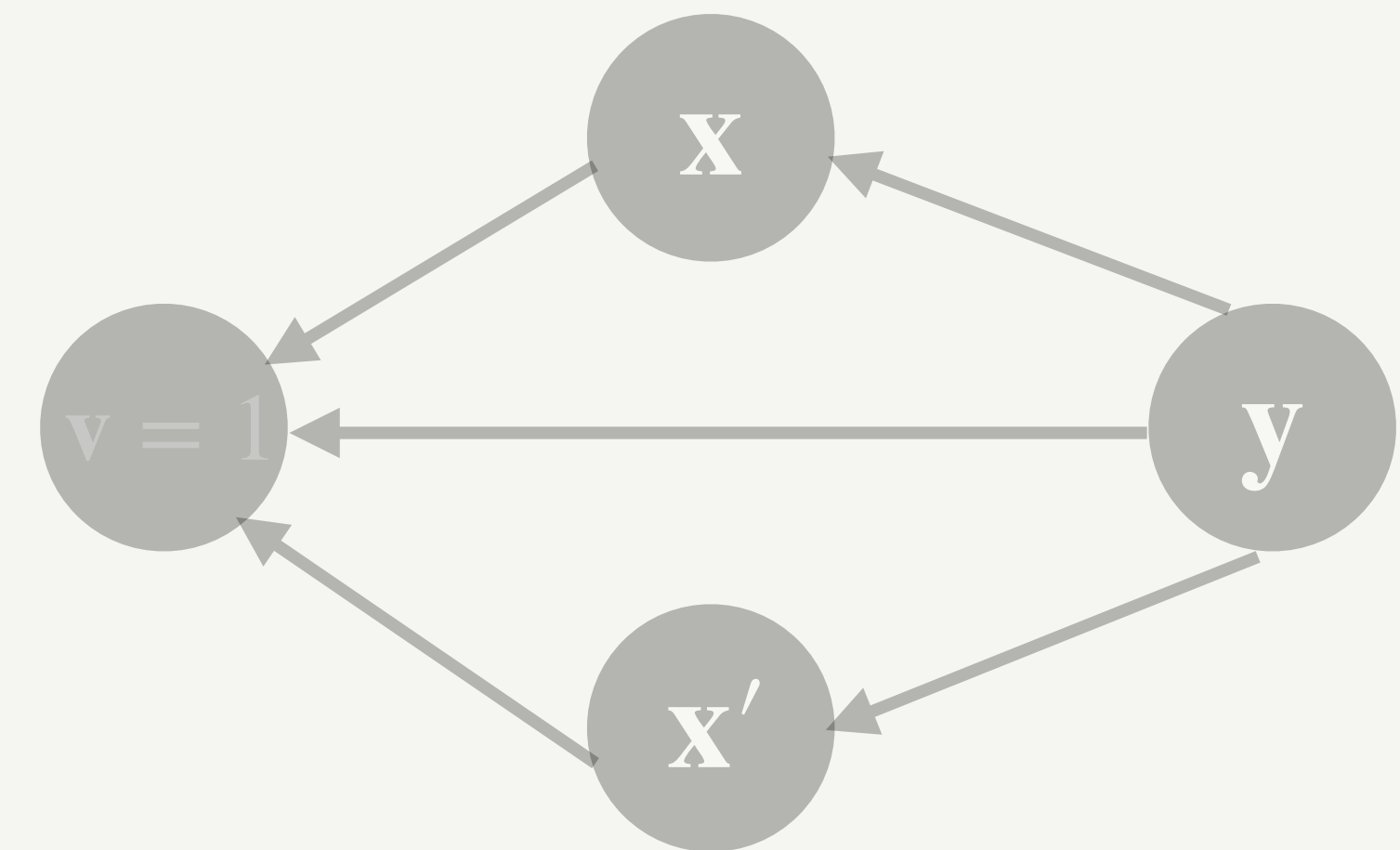
$$p(y, \mathbf{x}, \mathbf{x}', v = 1) = p(y) p(\mathbf{x} | y) p(\mathbf{x}' | y) p(v = 1 | y, \mathbf{x}, \mathbf{x}')$$



# Multi-modal learning

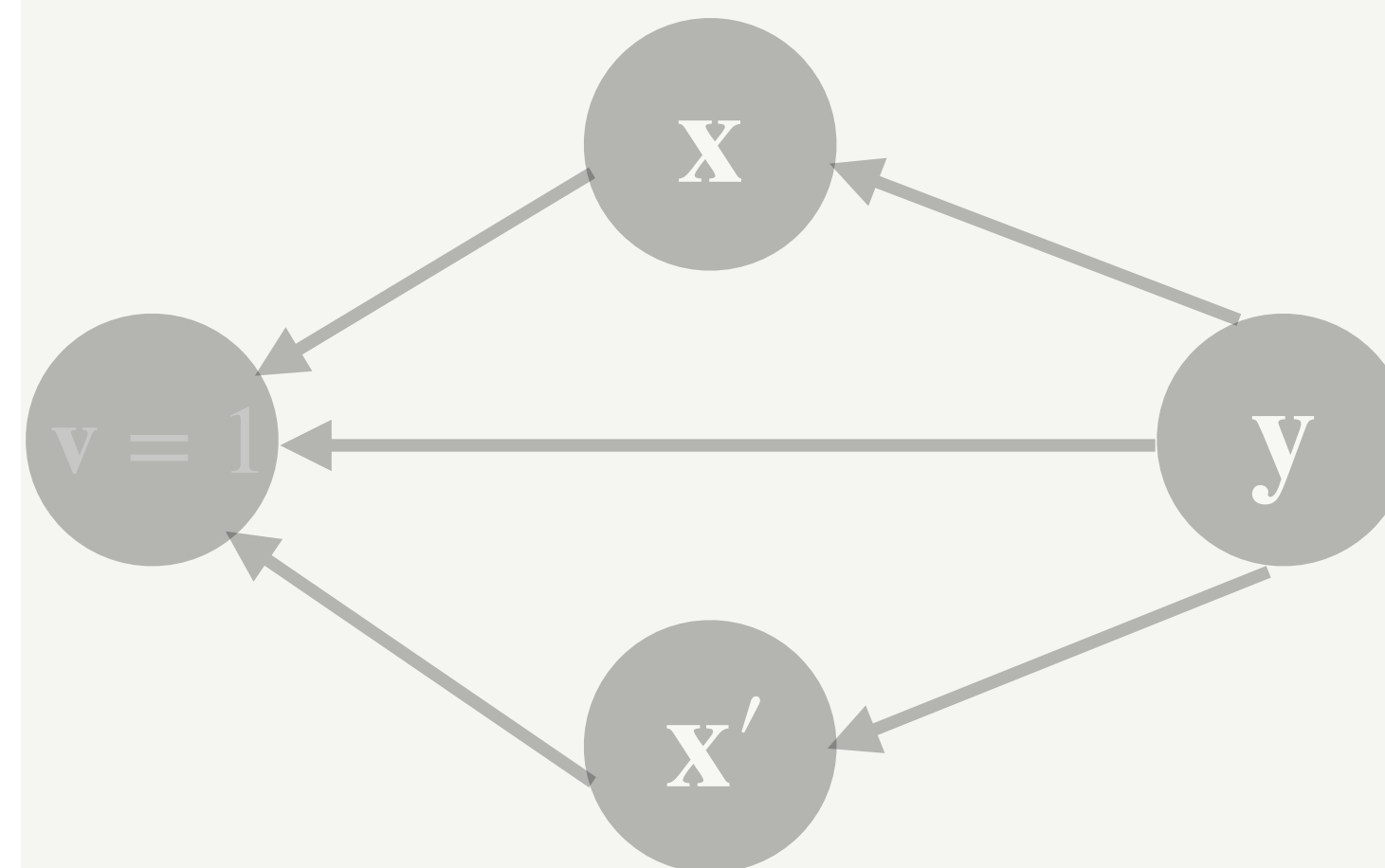
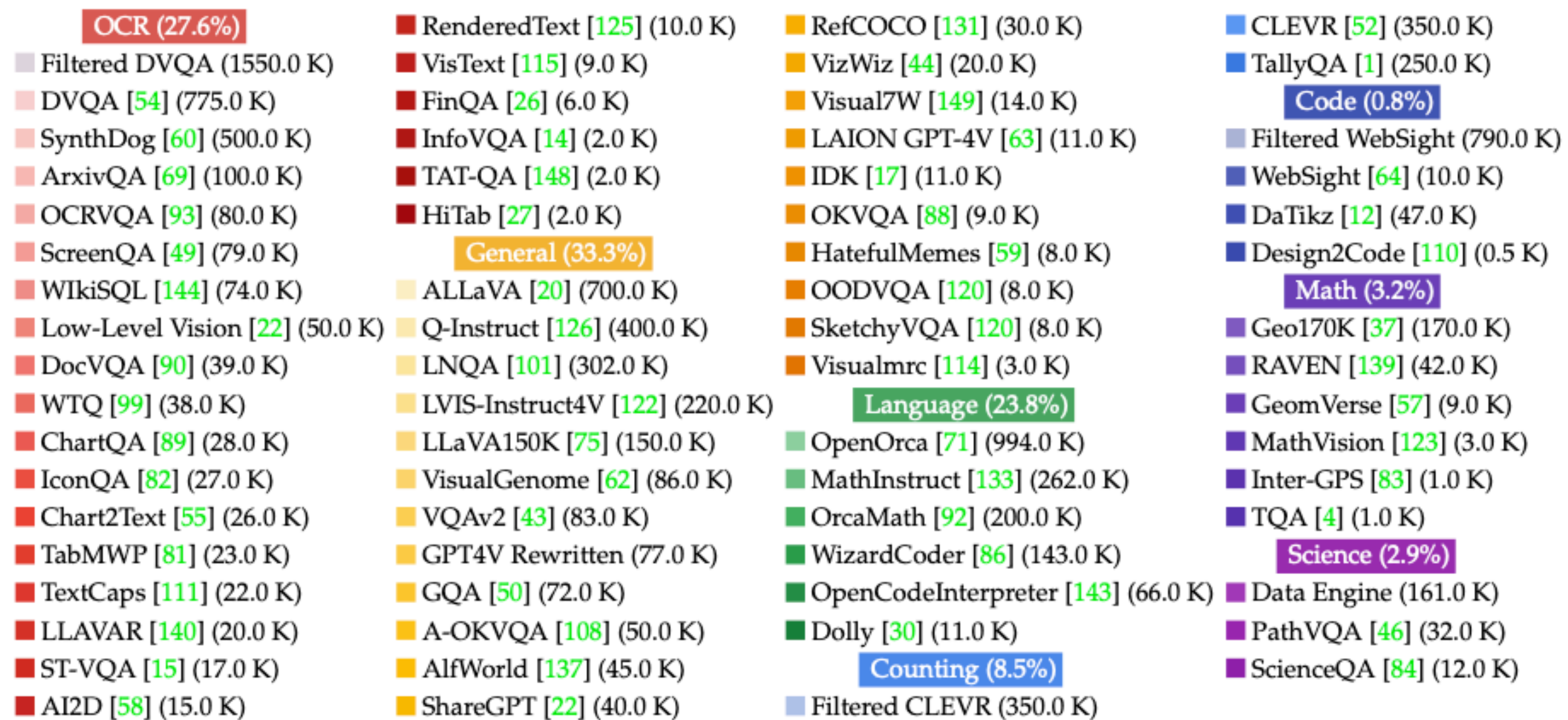
Dependency among the **features across modalities given the label**

$$p(y, \mathbf{x}, \mathbf{x}', v = 1) = p(y) p(\mathbf{x} | y) p(\mathbf{x}' | y) p(v = 1 | y, \mathbf{x}, \mathbf{x}')$$



Relative strength of these dependencies is unknown

# Multi-modal learning



Relative strength of these dependencies is unknown

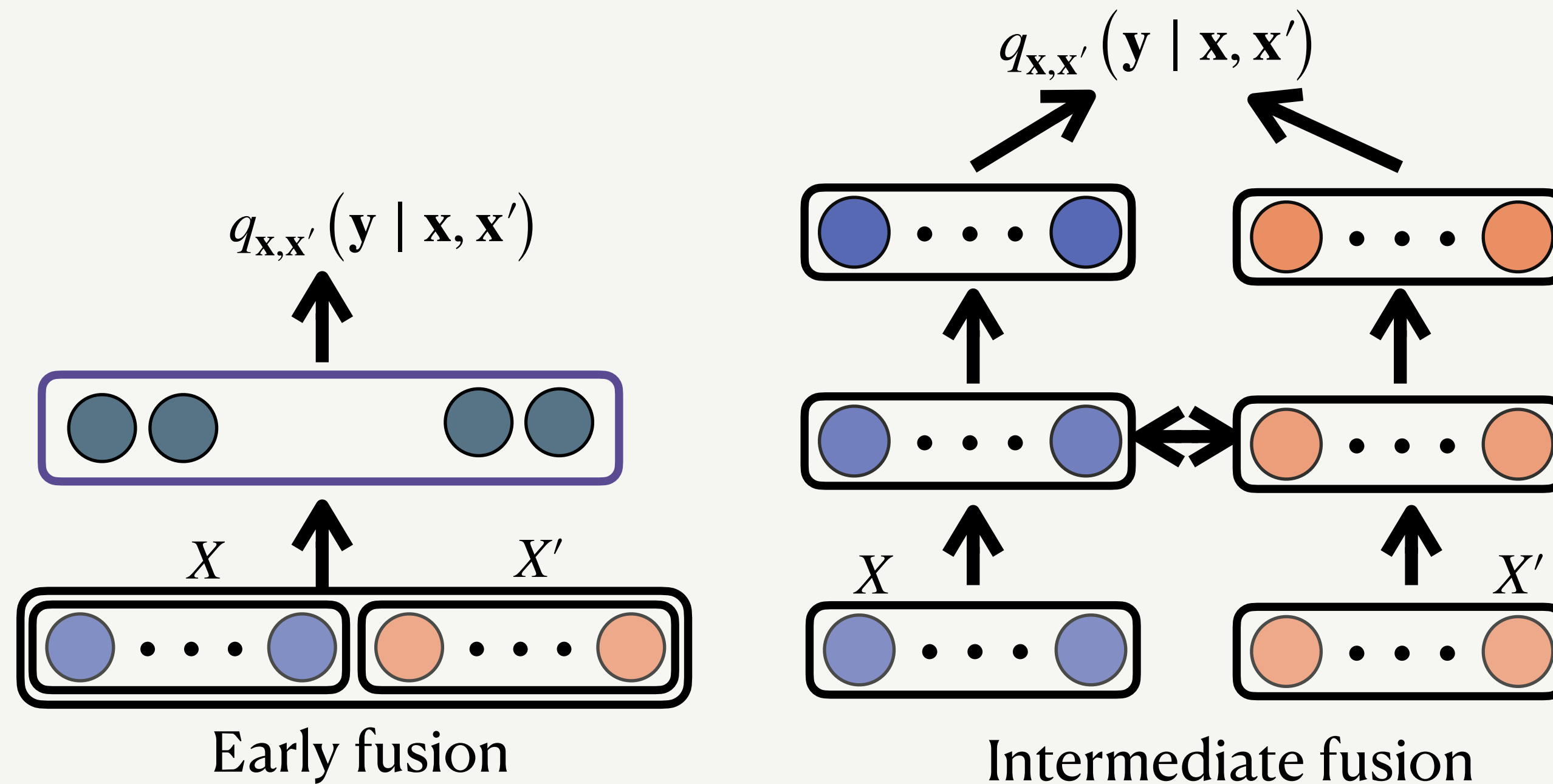


Conventional methods capture  
**either** inter- or intra-modality  
dependencies

Conventional methods capture **either** inter- or intra-modality dependencies, but **not** both

# Categorization of architectural strategies

For input modalities  $(\mathbf{x}, \mathbf{x}')$  and label  $\mathbf{y}$



What makes training multi-modal classification networks hard? (Wang et al., 2020)

Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks (Wu et al., 2022)

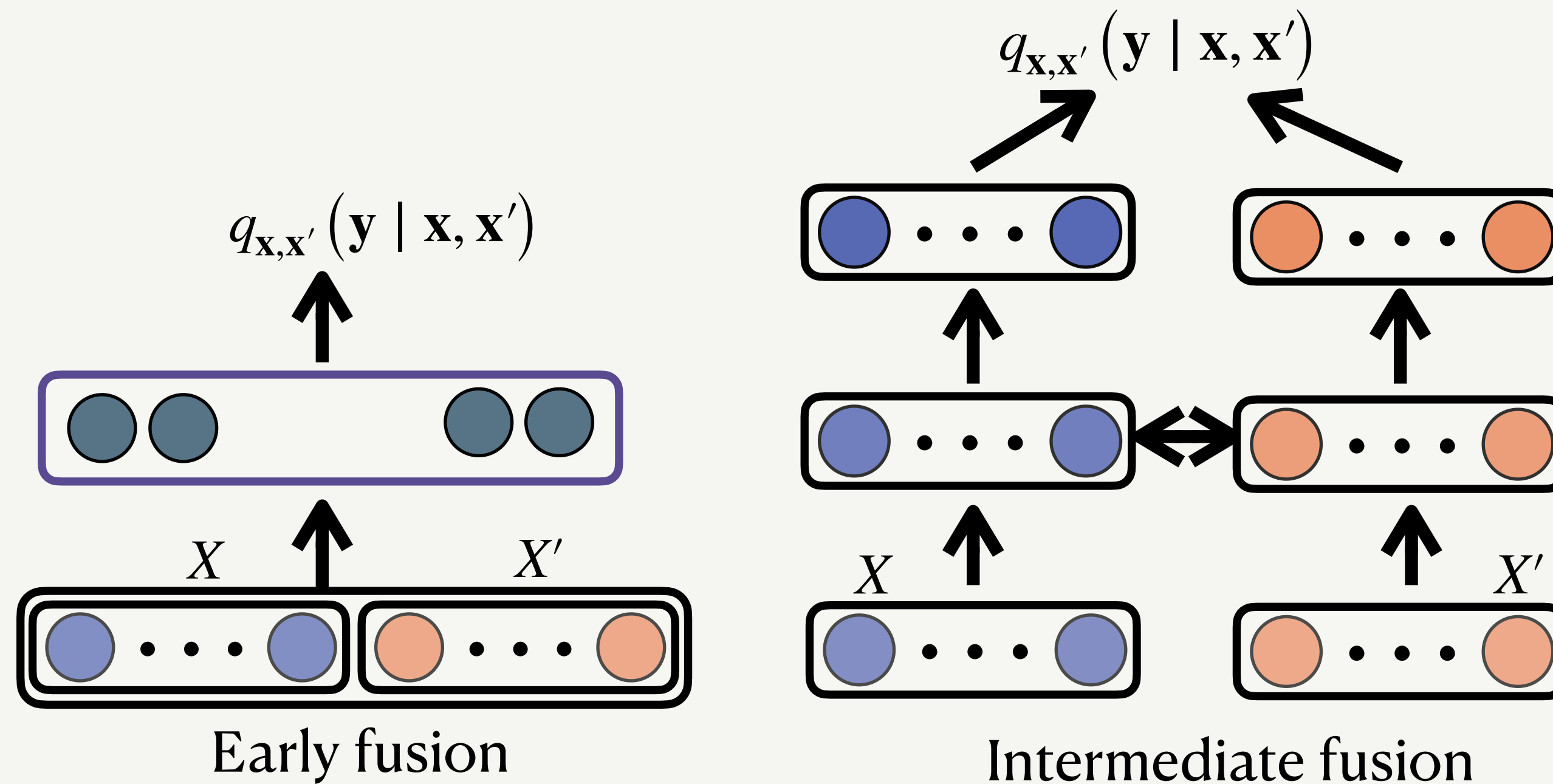
Balanced Multimodal Learning via On-the-fly Gradient Modulation (Peng et al., 2022)

Coarse-to-Fine Vision-Language Pre-training with Fusion in the Backbone. (Dou et al., 2022)



# Categorization of architectural strategies

For input modalities  $(\mathbf{x}, \mathbf{x}')$  and label  $\mathbf{y}$



**Inter-modality modeling**

What makes training multi-modal classification networks hard? (Wang et al., 2020)

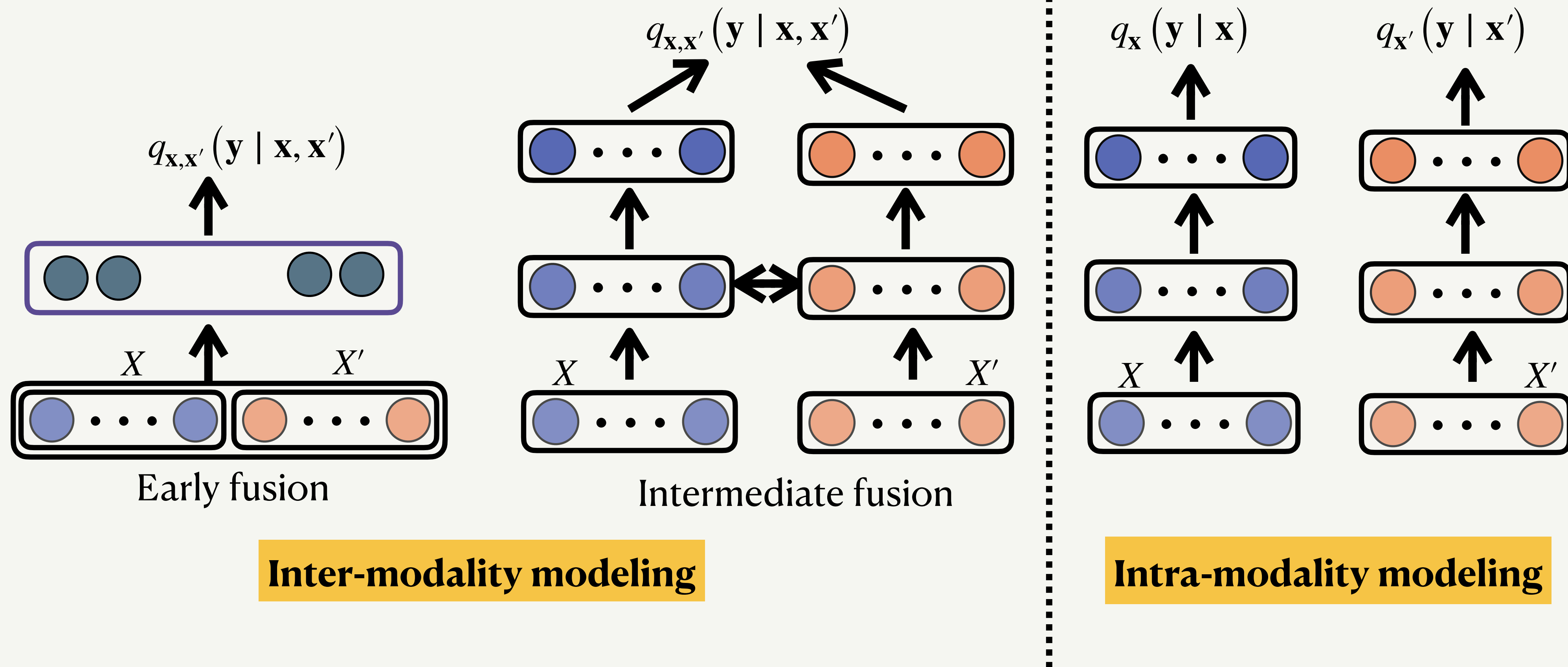
Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks (Wu et al., 2022)

Balanced Multimodal Learning via On-the-fly Gradient Modulation (Peng et al., 2022)

Coarse-to-Fine Vision-Language Pre-training with Fusion in the Backbone. (Dou et al., 2022)

# Categorization of architectural strategies

For input modalities  $(\mathbf{x}, \mathbf{x}')$  and label  $\mathbf{y}$



Majority Vote of Diverse Classifiers for Late Fusion (Morvant et al., 2014)

On integrating a language model into neural machine translation. (Gulcehre et al., 2017)

FLAVA: A Foundational Language And Vision Alignment Model (Singh et al., 2022)

**I2M2**

**Jointly modeling  
Inter-& Intra-modality  
dependencies —**

**Modality & architecture agnostic**



# Both Inter-and Intra-modality are important

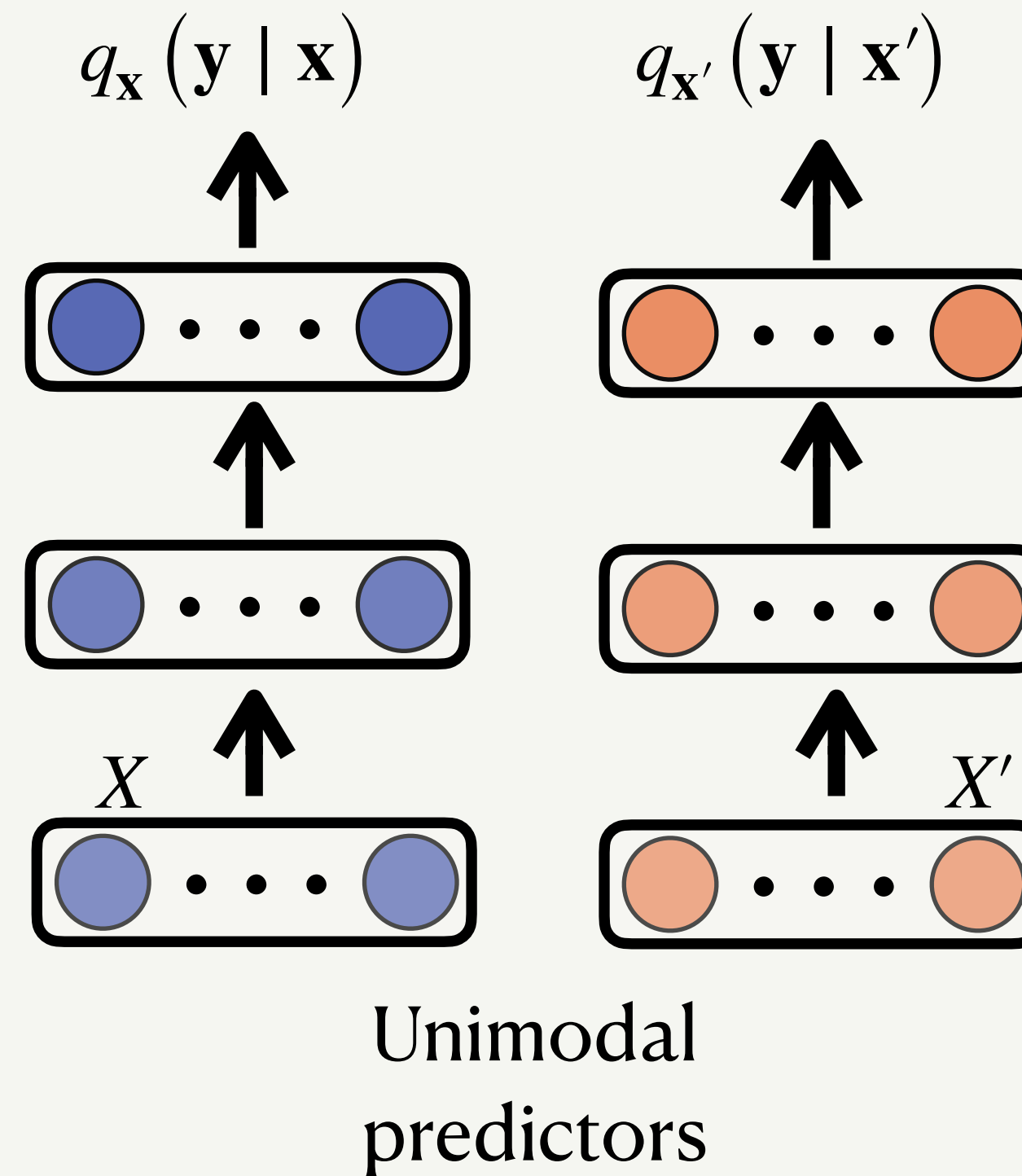
Based on the **multi-modal generative model**, we need to model the **individual modalities** and **their interaction** to predict the **target**

# Capture inter- & intra-modality dependencies

To build a classifier, we need to consider the dependencies between  $\mathbf{x}$ ,  $\mathbf{x}'$ , and  $\mathbf{y}$

$$p(\mathbf{y} | \mathbf{x}, \mathbf{x}', \mathbf{v} = 1) \propto p(\mathbf{y})$$

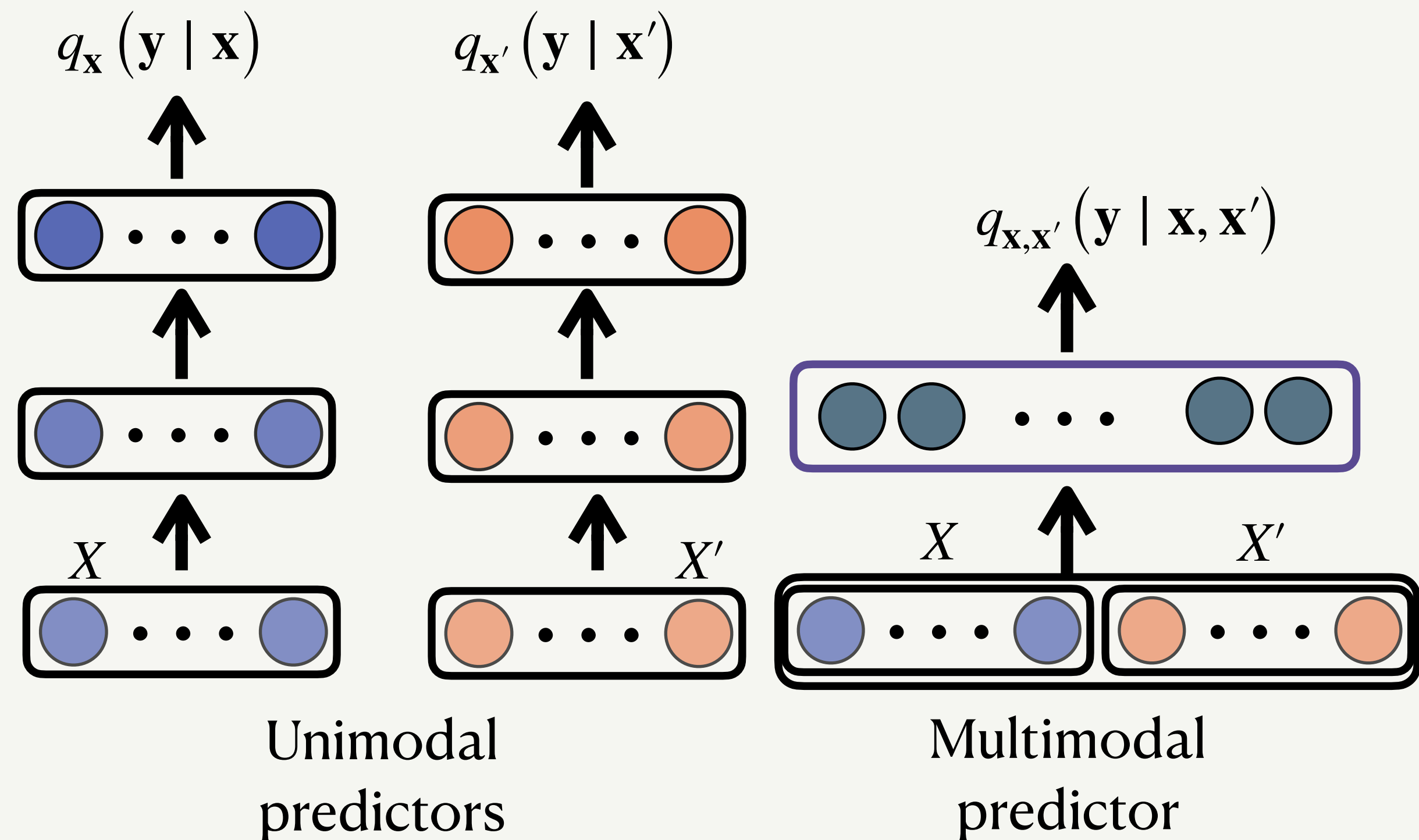
$$\times \underbrace{q_{\mathbf{x}}(\mathbf{y} | \mathbf{x}) \times q_{\mathbf{x}'}(\mathbf{y} | \mathbf{x}')}_{\text{Unimodal predictors}}$$



# Capture inter- & intra-modality dependencies

To build a classifier, we need to consider the dependencies between  $\mathbf{x}$ ,  $\mathbf{x}'$ , and  $\mathbf{y}$

$$p(\mathbf{y} | \mathbf{x}, \mathbf{x}', \mathbf{v} = 1) \propto p(\mathbf{y})$$
$$\times \underbrace{q_{\mathbf{x}}(\mathbf{y} | \mathbf{x}) \times q_{\mathbf{x}'}(\mathbf{y} | \mathbf{x}')}_{\text{Unimodal predictors}}$$
$$\times \underbrace{q_{\mathbf{x}, \mathbf{x}'}(\mathbf{y} | \mathbf{x}, \mathbf{x}')}_{\text{Multimodal predictor}}$$

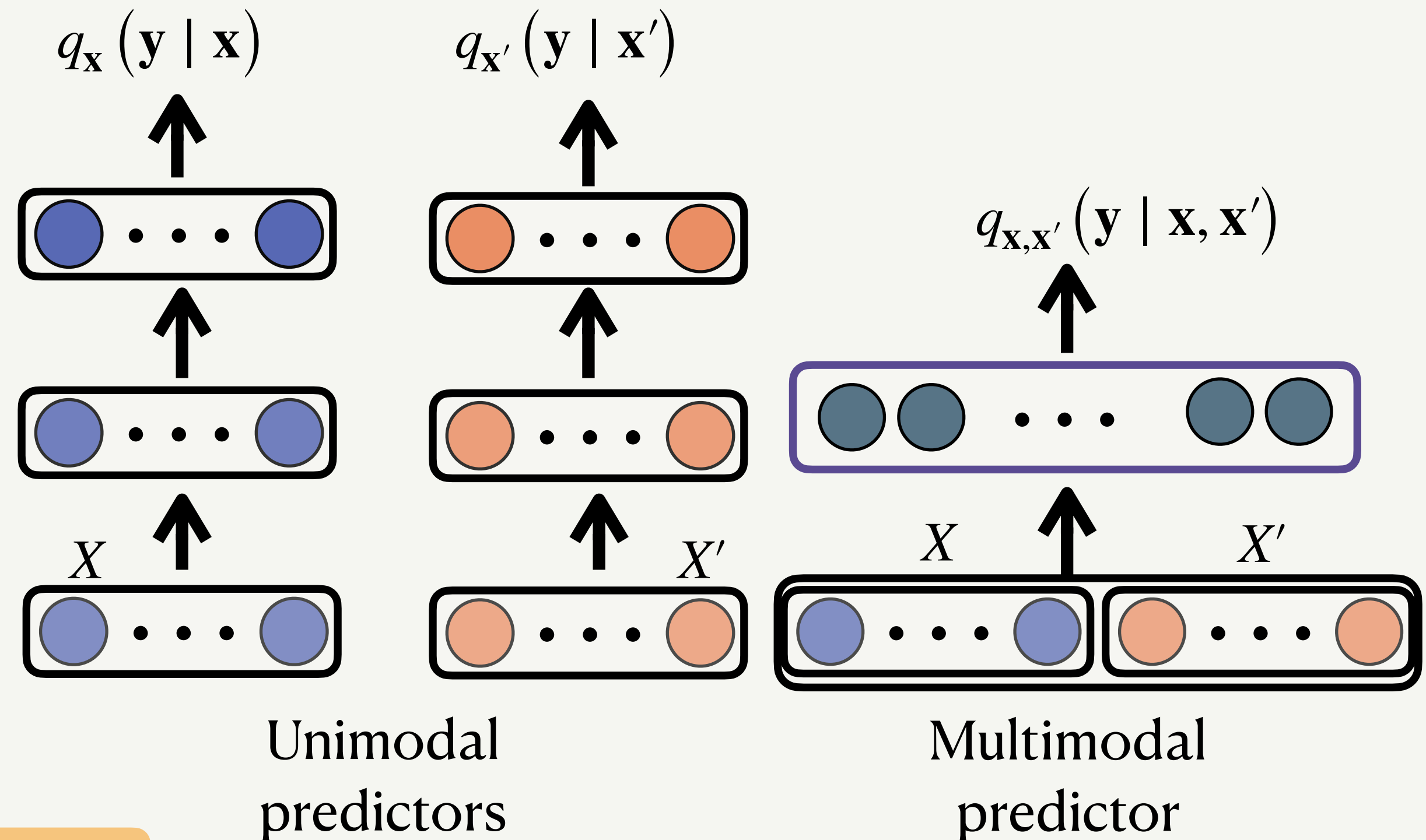




# Capture inter- & intra-modality dependencies

To build a classifier, we need to consider the dependencies between  $\mathbf{x}$ ,  $\mathbf{x}'$ , and  $\mathbf{y}$

$$\begin{aligned}
 p(\mathbf{y} \mid \mathbf{x}, \mathbf{x}', \mathbf{v} = 1) &\propto p(\mathbf{y}) \\
 &\times \underbrace{q_{\mathbf{x}}(\mathbf{y} \mid \mathbf{x}) \times q_{\mathbf{x}'}(\mathbf{y} \mid \mathbf{x}')}_{\text{Unimodal predictors}} \\
 &\times \underbrace{q_{\mathbf{x}, \mathbf{x}'}(\mathbf{y} \mid \mathbf{x}, \mathbf{x}')}_{\text{Multimodal predictor}}
 \end{aligned}$$



## Training Recipe

- Build modality-specific classifiers and a classifier that captures their interaction
- Combine them by building a product of experts

Results on **healthcare, vision**  
**and language tasks**  
with **different strengths** of  
inter- & intra-modality  
dependencies

# Digit recognition with audio and vision modalities

Vision modality includes **images of handwritten digits**, while the audio modality contains **recordings of digits** spoken by humans



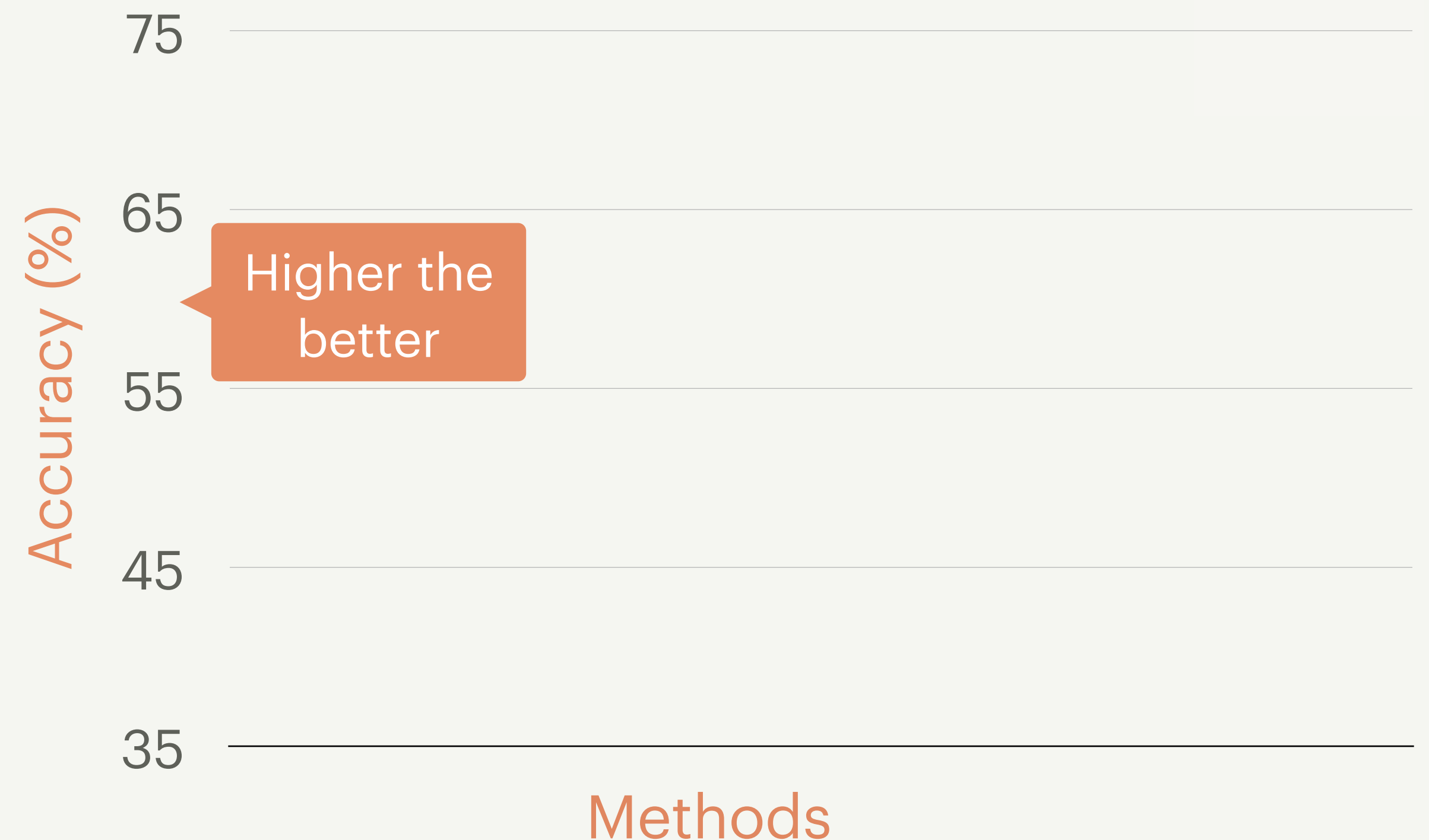
Audio-Vision MNIST

# Digit recognition with audio and vision modalities

Vision modality includes **images of handwritten digits**, while the audio modality contains **recordings of digits** spoken by humans



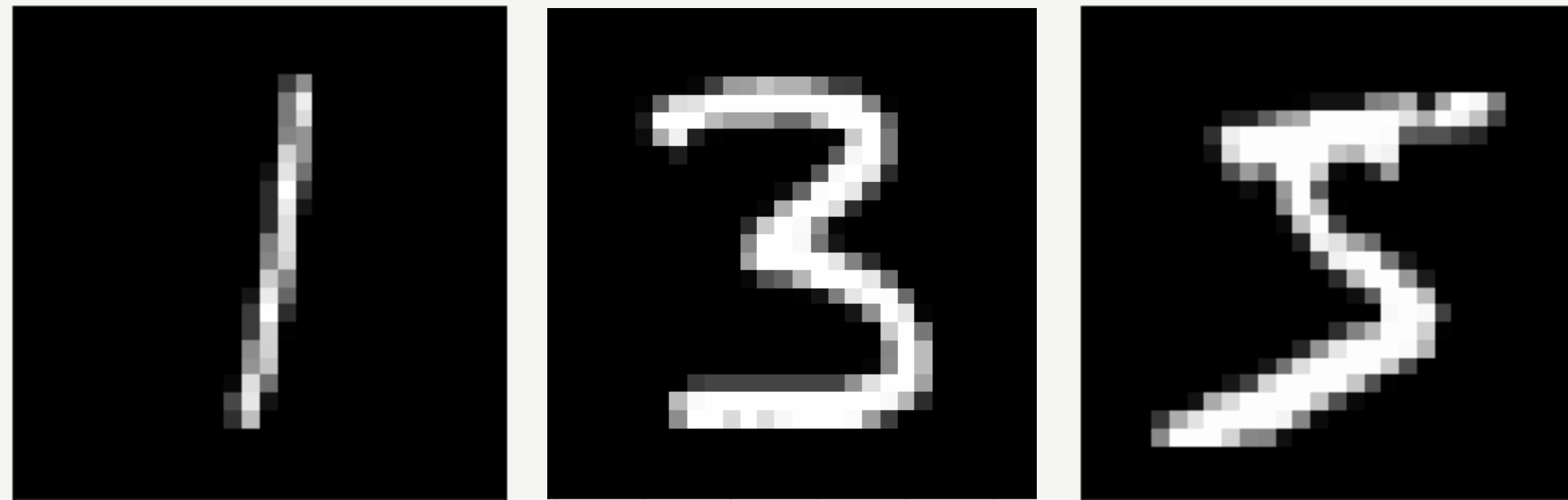
Audio-Vision MNIST



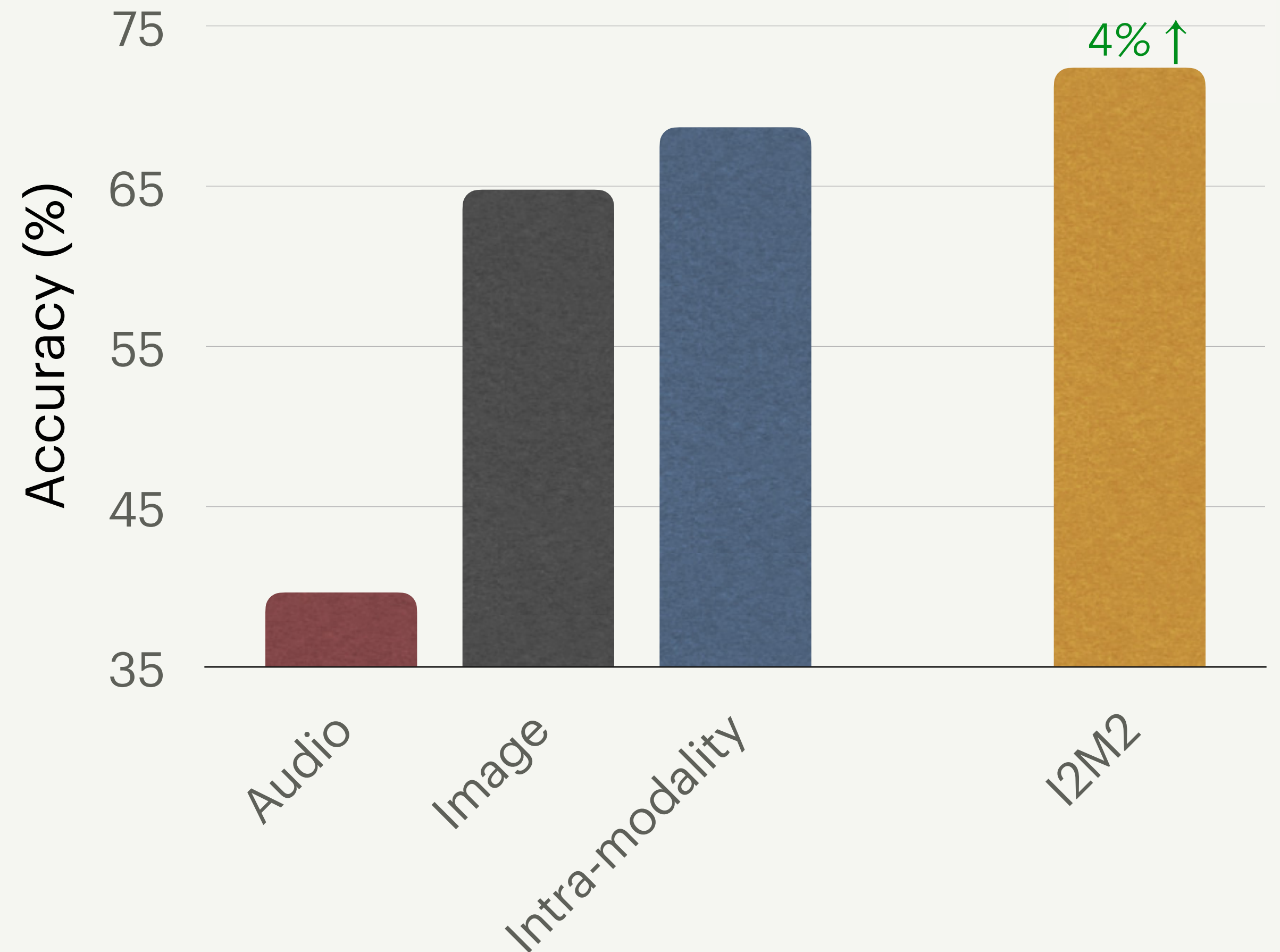


# Inter-modality interactions are essential

**Inter-modality interactions** are **essential** for this task as I2M2 **outperforms** both the unimodal and intra-modality models.



Audio-Vision MNIST

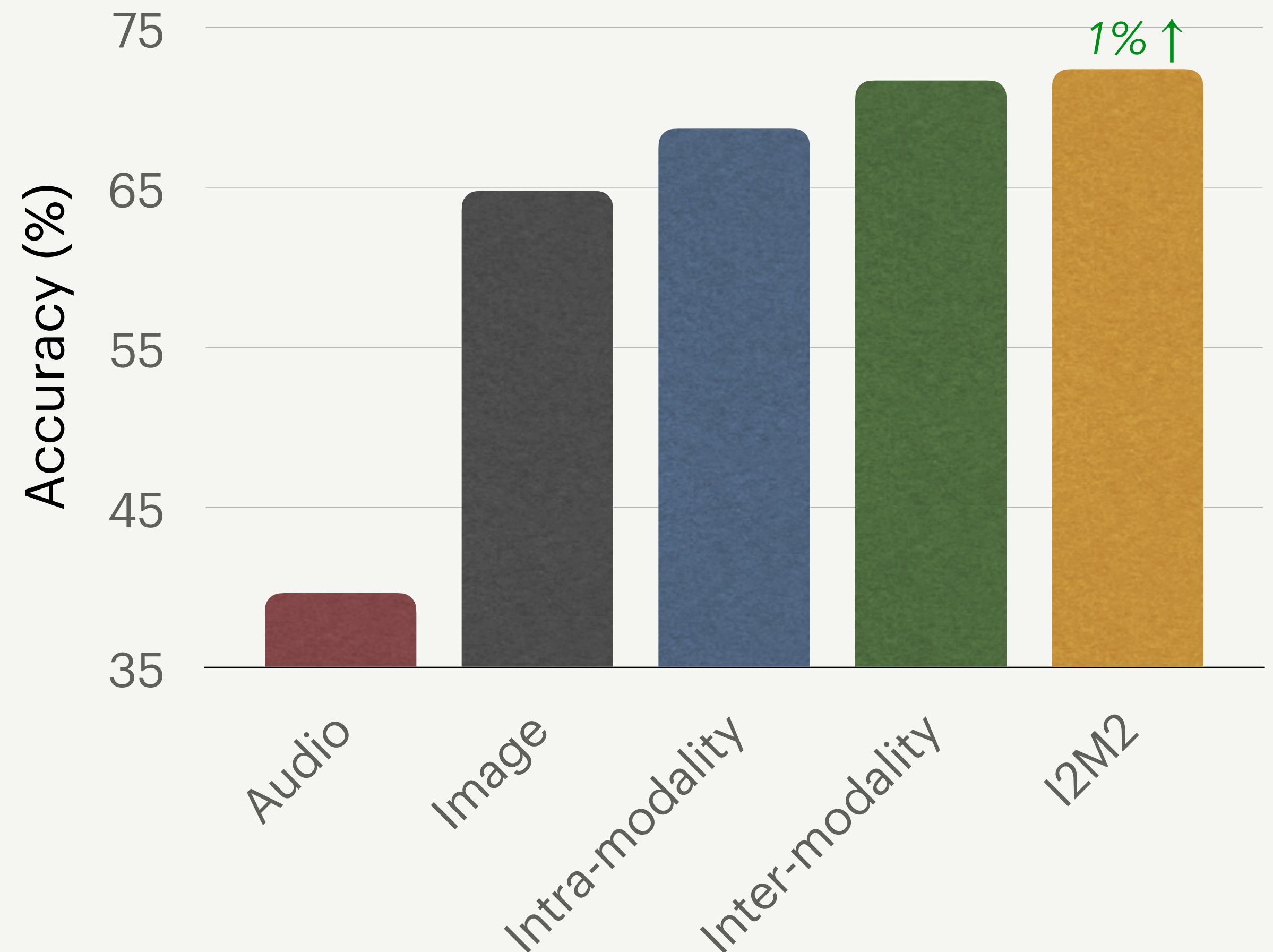


# Enhanced flexibility and effectiveness

I2M2 **eliminates** the **need to pre-determine** which dependencies should be modeled



Audio-Vision MNIST





# Automated Knee Pathology Diagnosis

We leverage **emulated single coil *k*-space data** for automated diagnosis

ACL



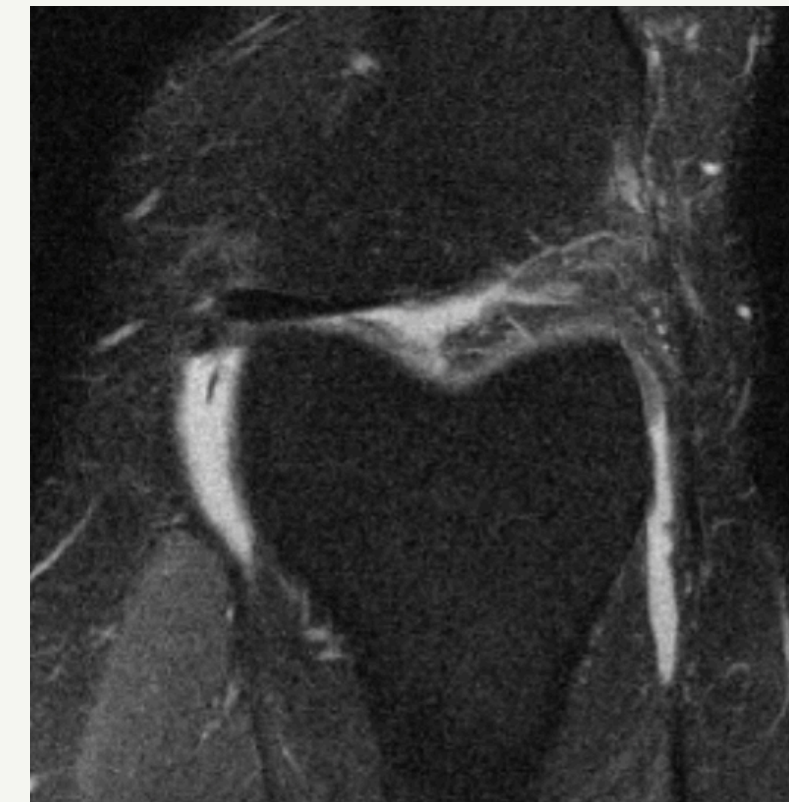
Meniscus Tear



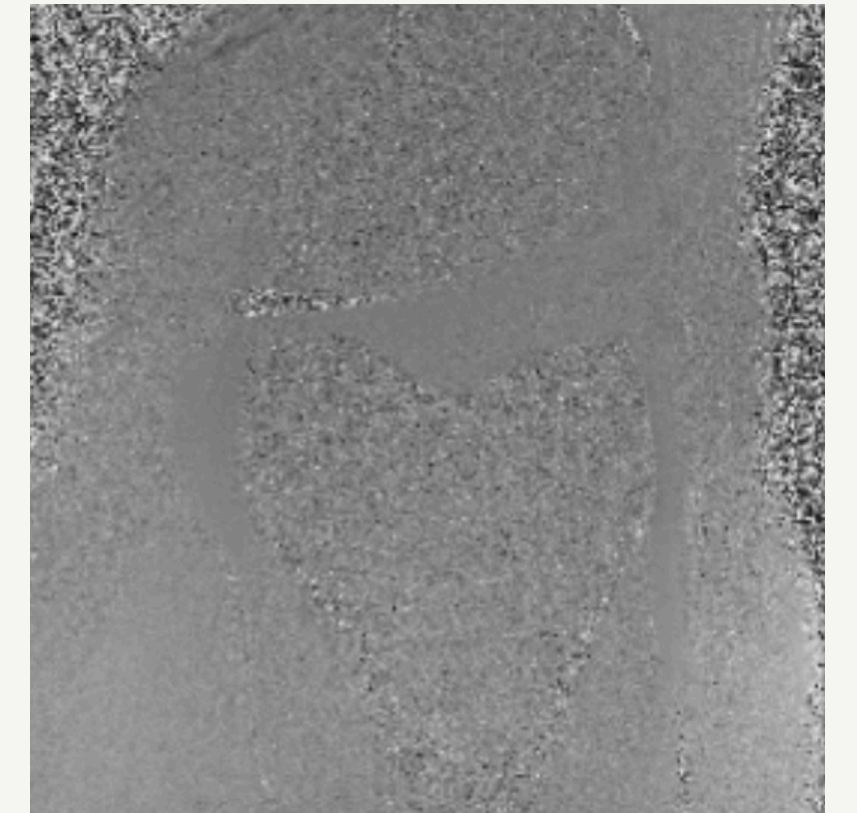
Cartilage



Magnitude



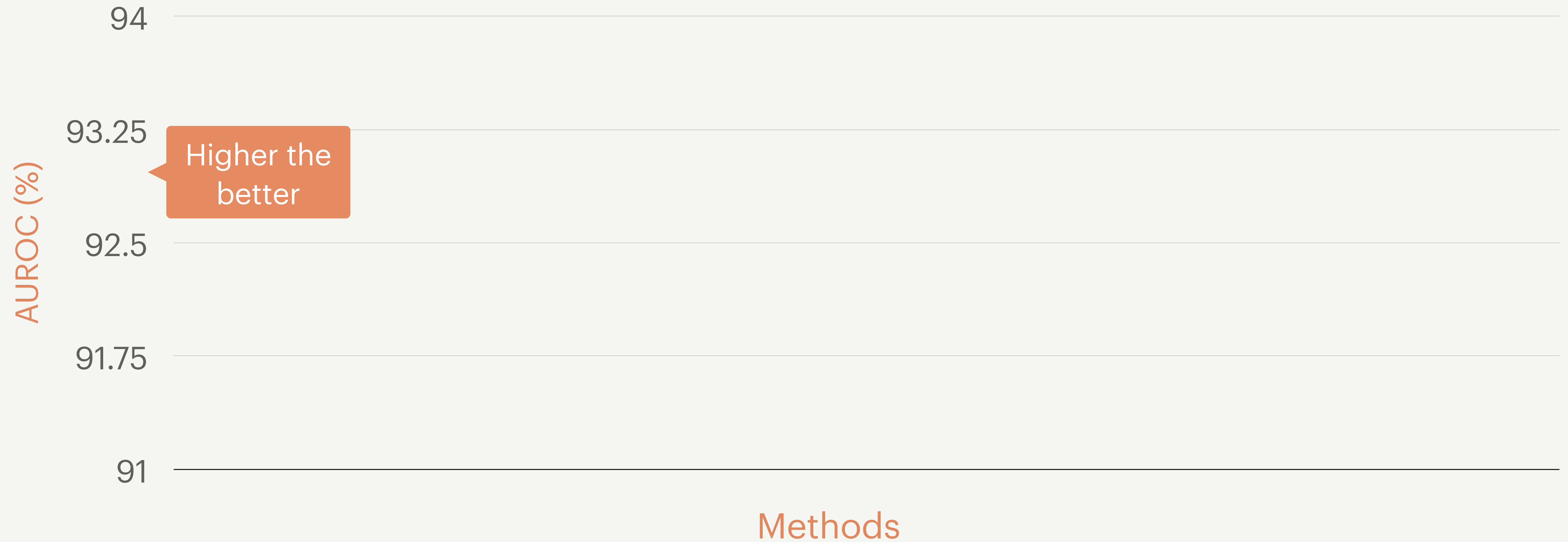
Phase



Target labels

Input modalities

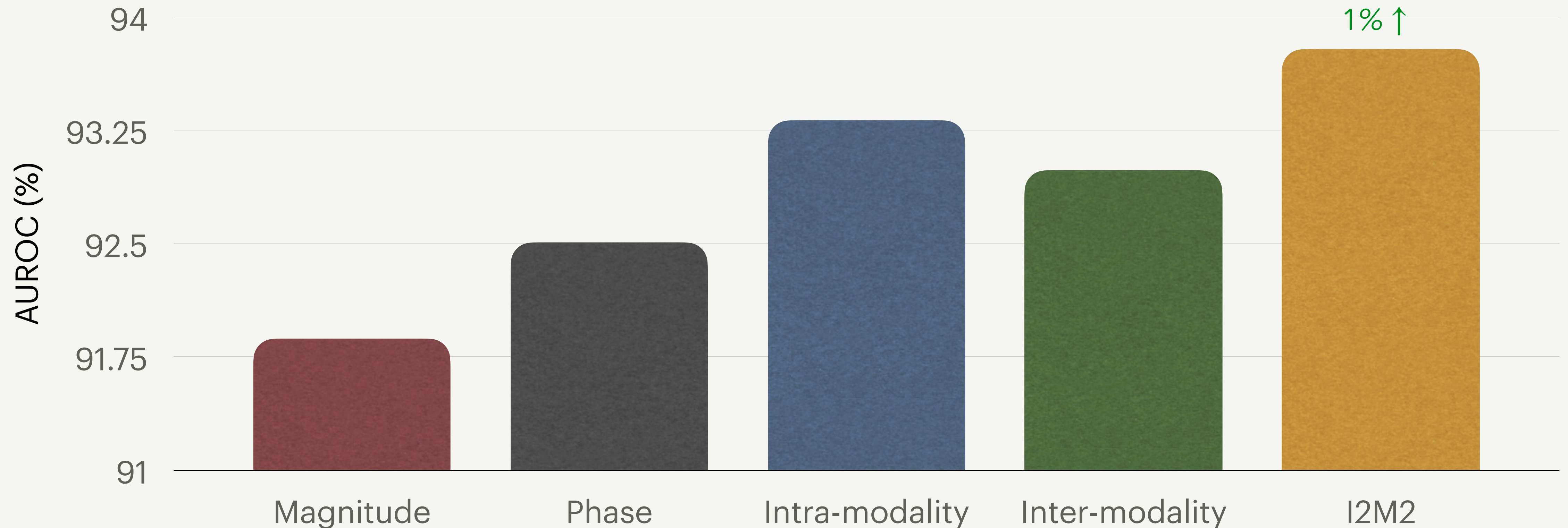
# Automated Knee Pathology Diagnosis





# Navigating Inter-& Intra-Modality Dependencies

When Intra-modality dependencies matter i.e., one dependency type is missing, I2M2 *effectively captures the other*





# Natural Language Visual Reasoning

Ascertain if the sentence **accurately (True)** or **inaccurately (False)** the image pair

Image Modality



Question

There are more birds in the image on the left than in the image on the right.

There are more birds in the image on the left than in the image on the right.

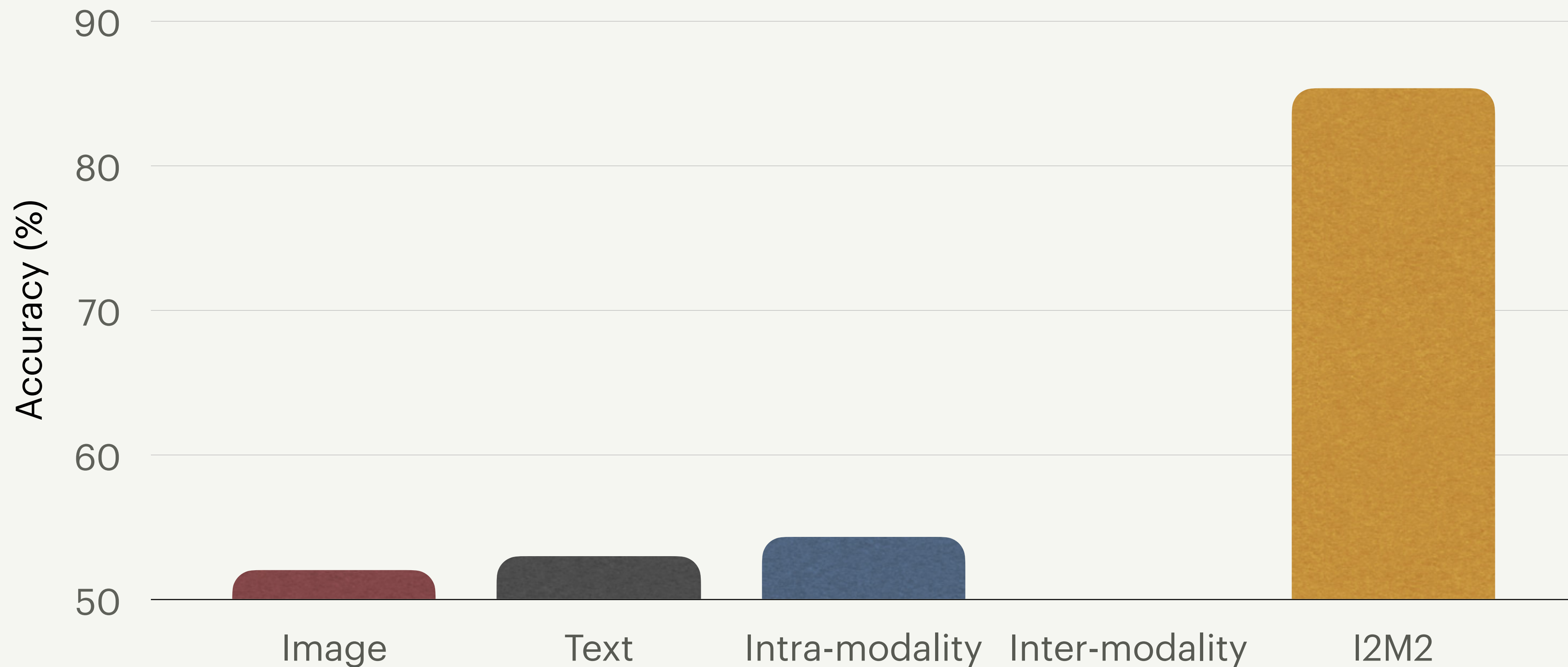
Answer

True

False

# Navigating Inter-& Intra-Modality Dependencies

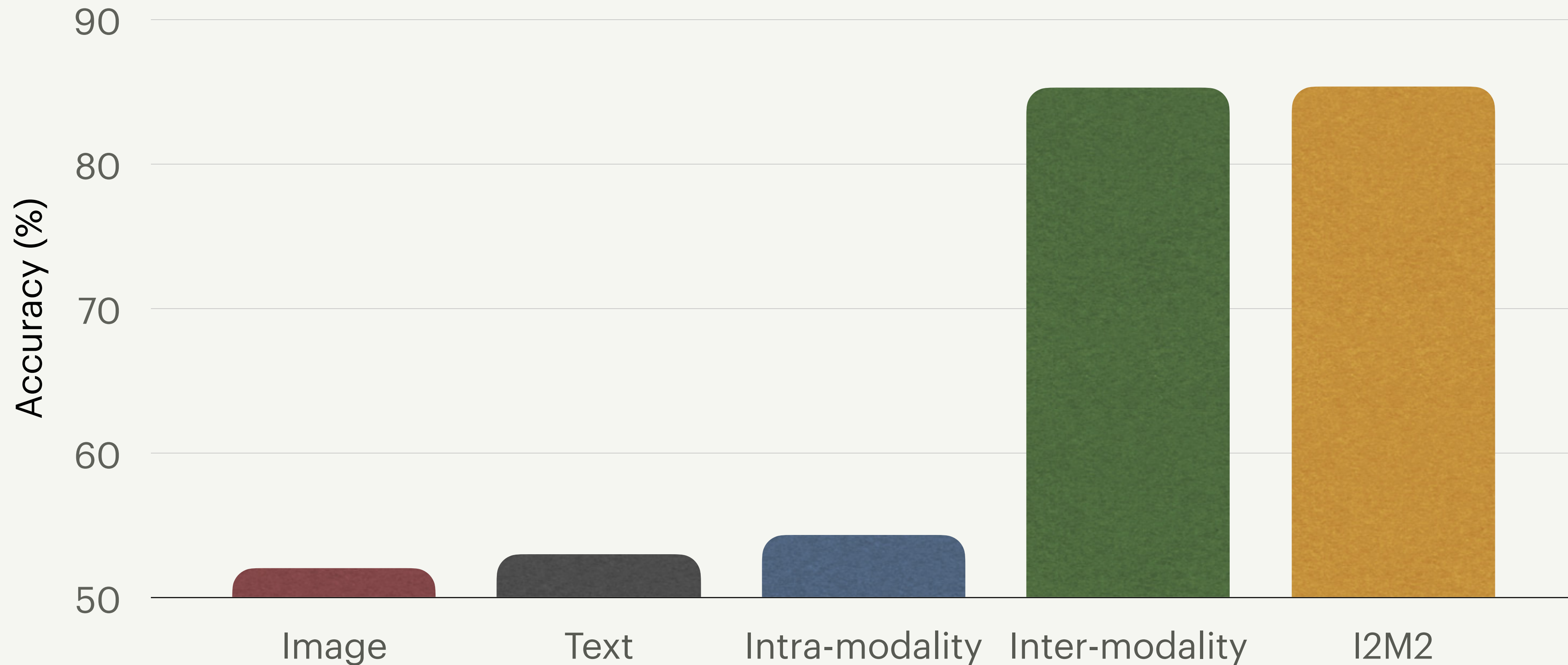
When Inter-modality dependencies matter i.e, one dependency type is missing, I2M2 *effectively captures the other*





# Navigating Inter-& Intra-Modality Dependencies

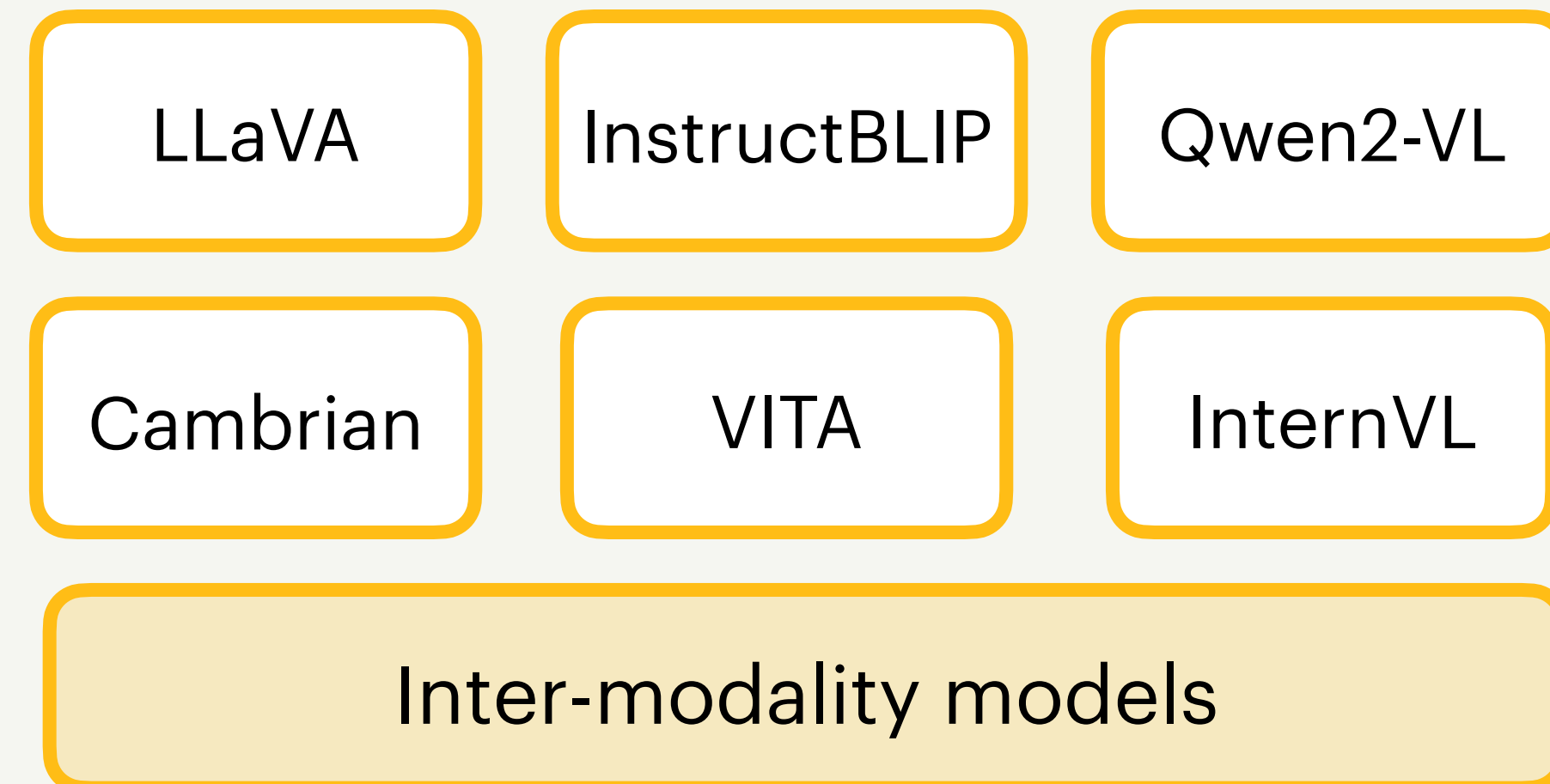
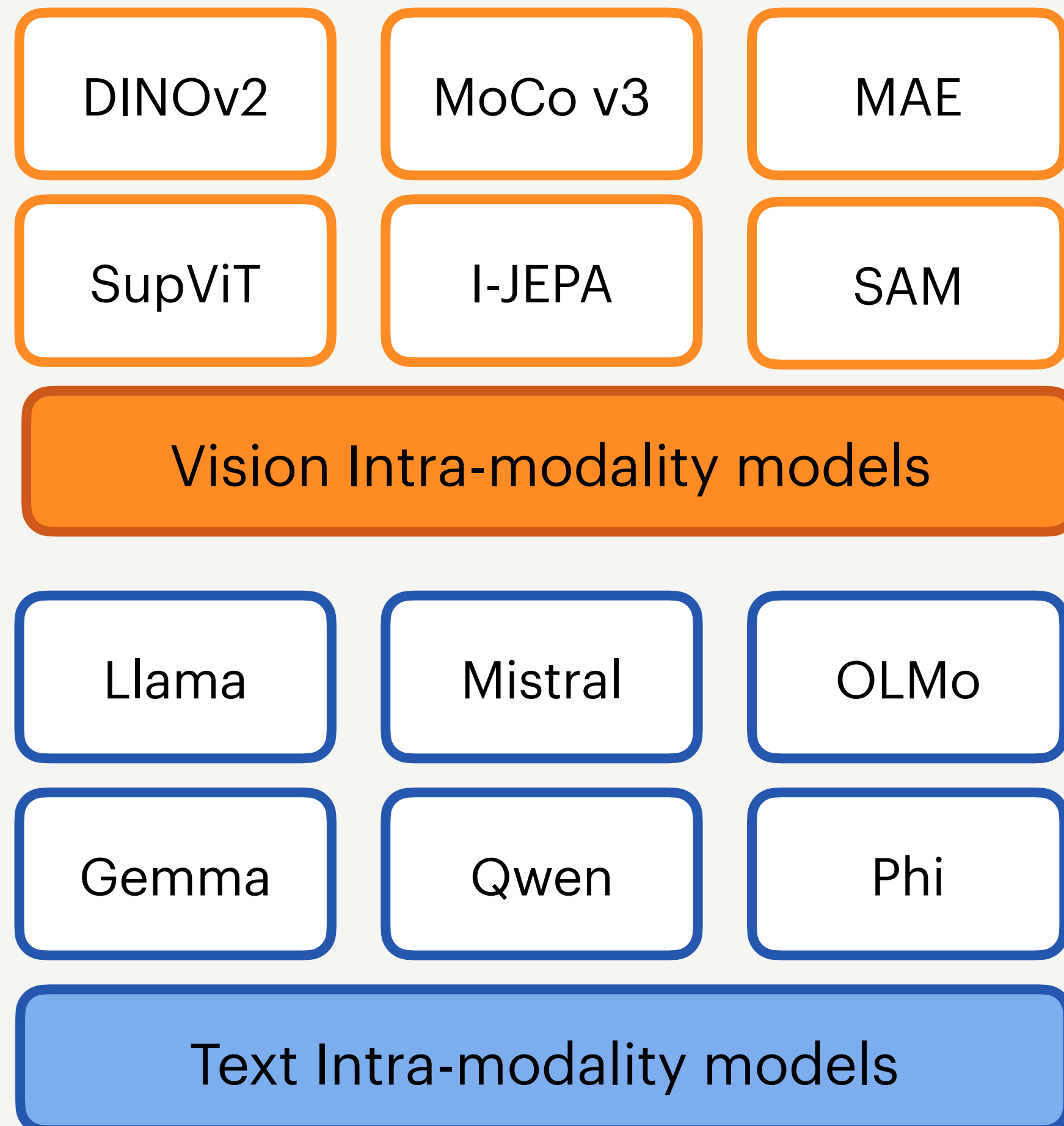
When Inter-modality dependencies matter i.e, one dependency type is missing, I2M2 *effectively captures the other*





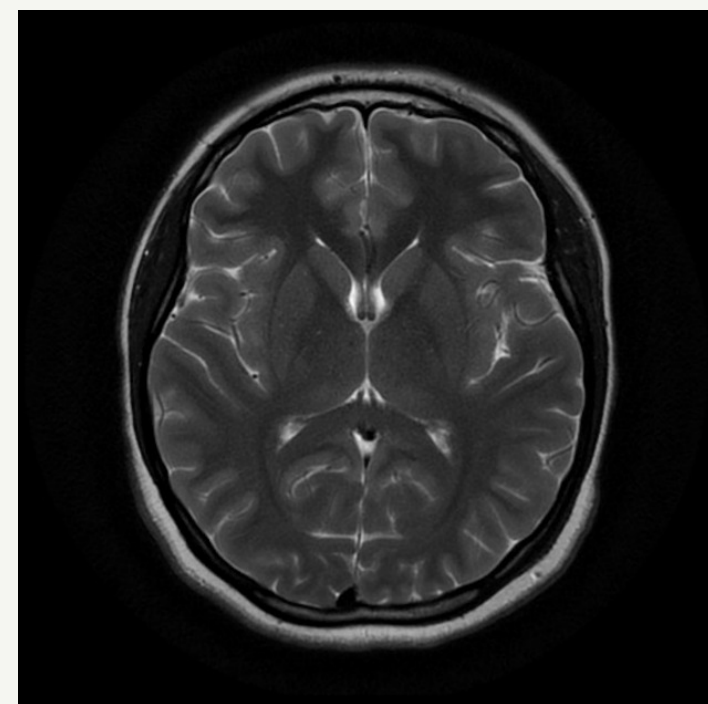
# What's next

Harnessing the capabilities of **recent inter- and intra-modality models**



# What's next

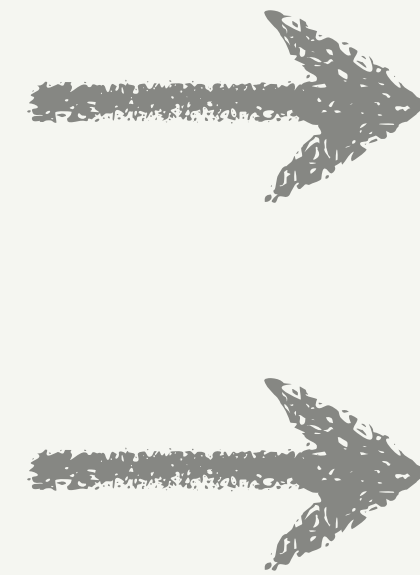
Harnessing the capabilities of **multi-modal healthcare data**



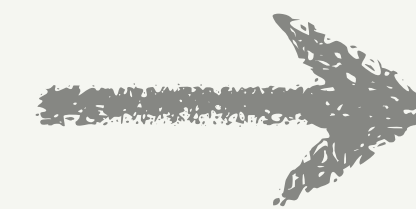
Medical Images

Tabular data

-	-	-
-	-	-
-	-	-
-	-	-
-	-	-



I2M2

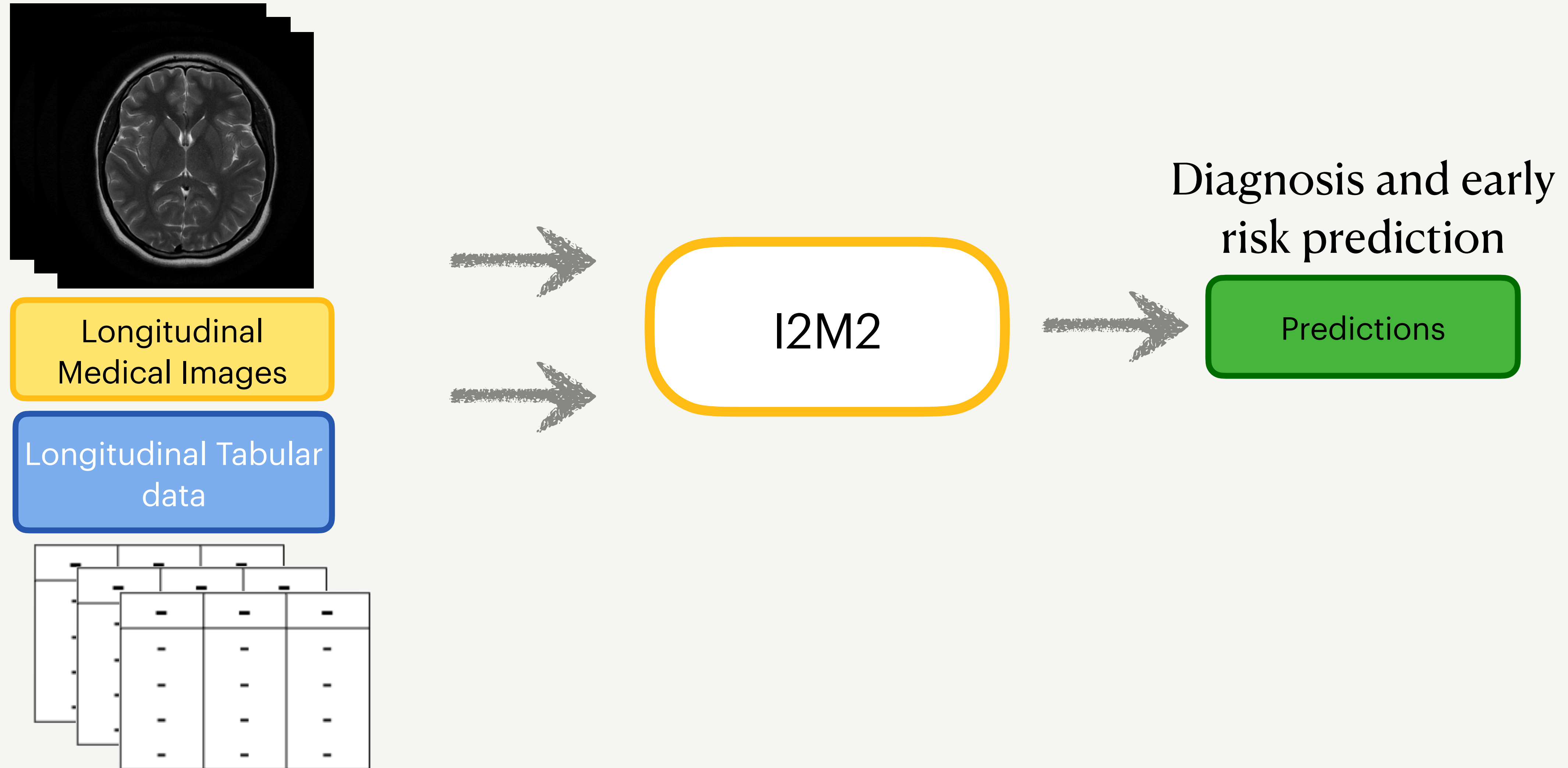


Diagnosis and early  
risk prediction

Predictions

# What's next

Harnessing the capabilities of **multi-modal healthcare data**



# Takeaway



Have multiple  
**"modalities"**?



**Paper**



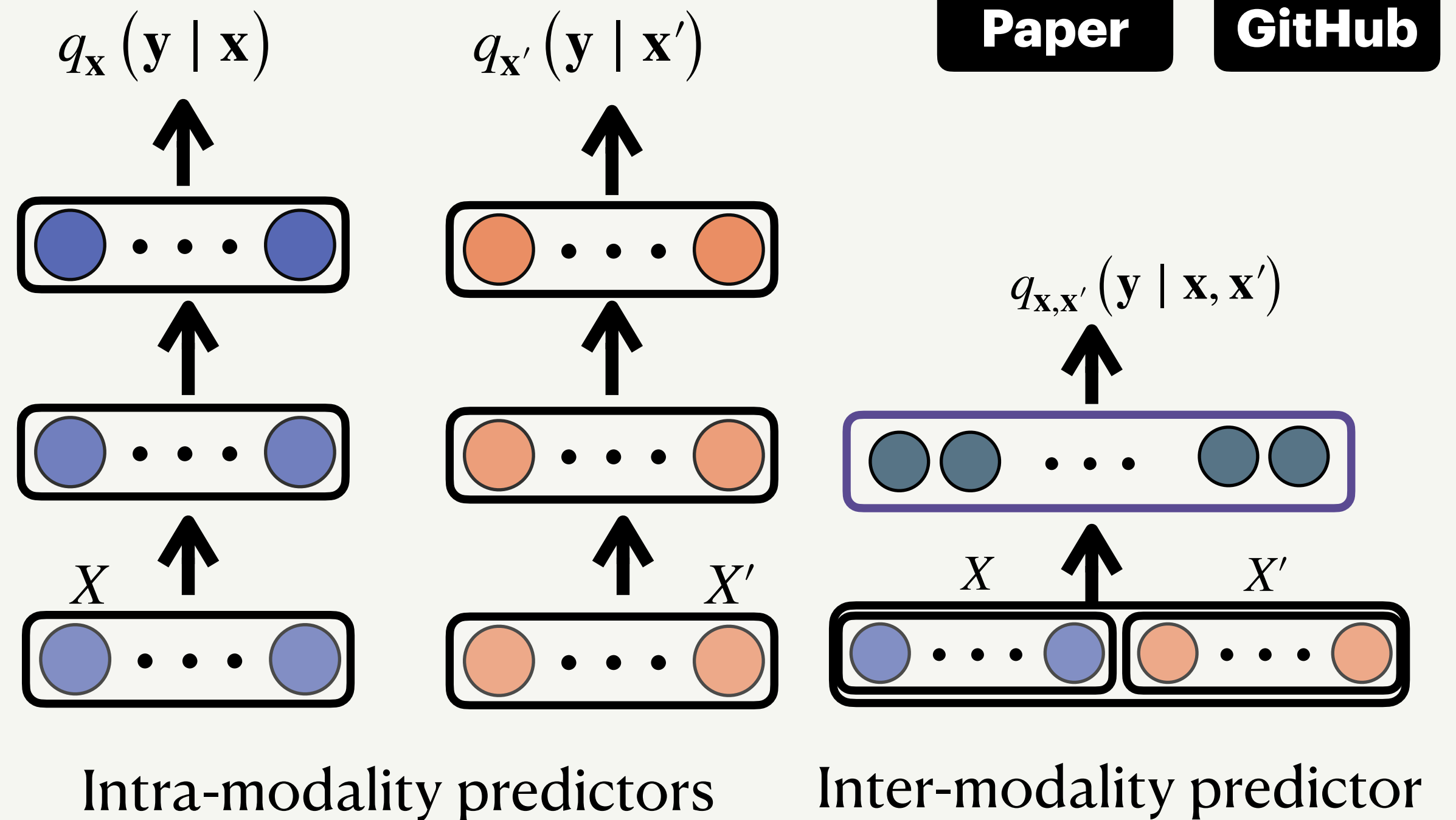
**GitHub**



# Takeaway



Have multiple  
"modalities"?



Paper

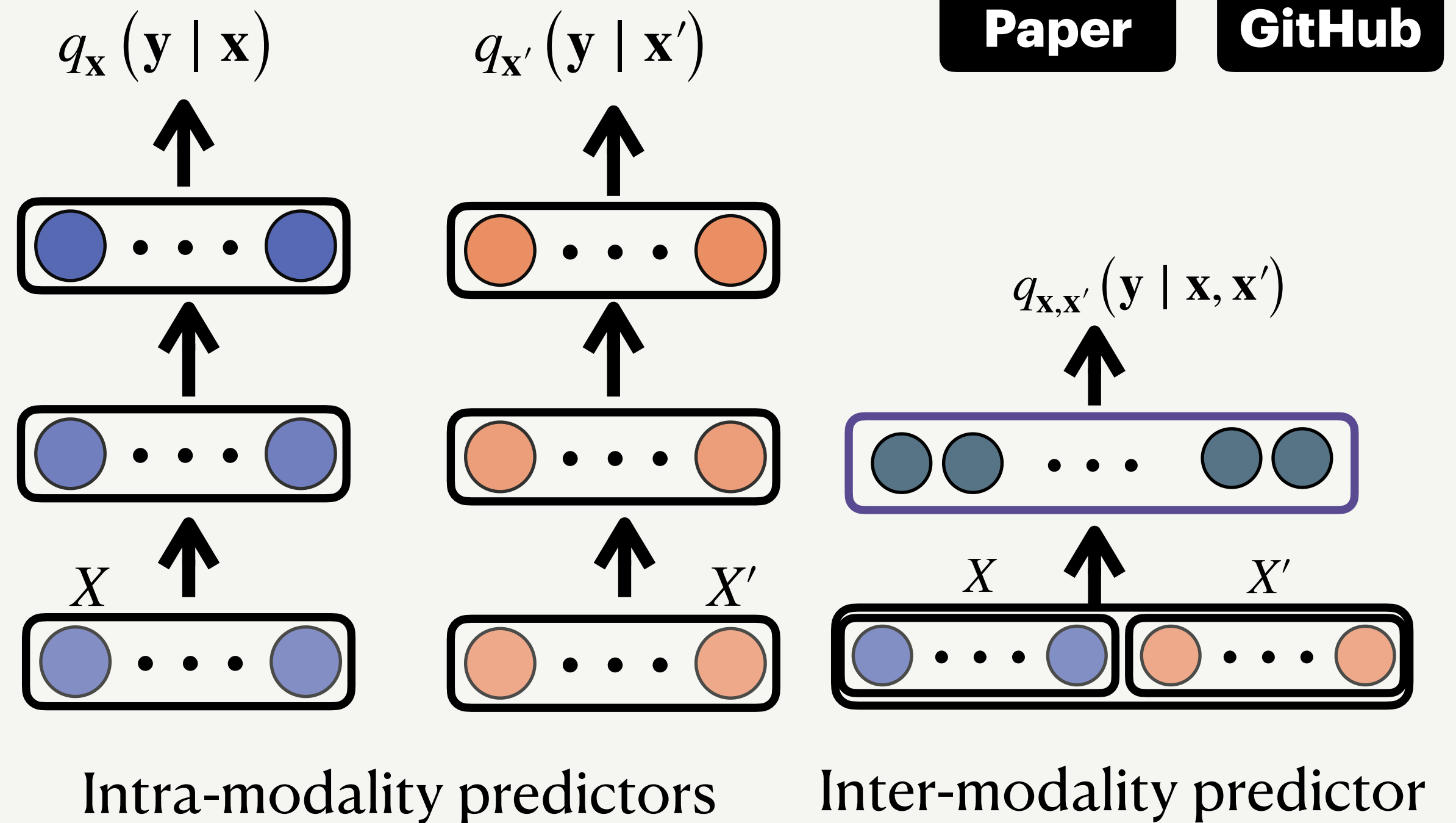
GitHub

**Inter- & Intra-modality modeling (I2M2)**

# Takeaway



Have multiple **“modalities”**?



## Inter- & Intra-modality modeling (I2M2)

Integrate **your dataset** and **inter-modality model** in **our framework** and share the amazing results with the community! 😊

[dmadaan.com/](https://dmadaan.com/)

[dmadaan\\_](https://twitter.com/dmadaan_)

[divyam.madaan@nyu.edu](mailto:divyam.madaan@nyu.edu)