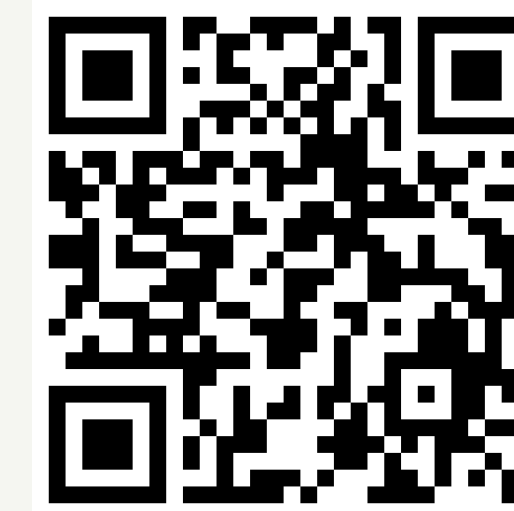




[Paper](#)



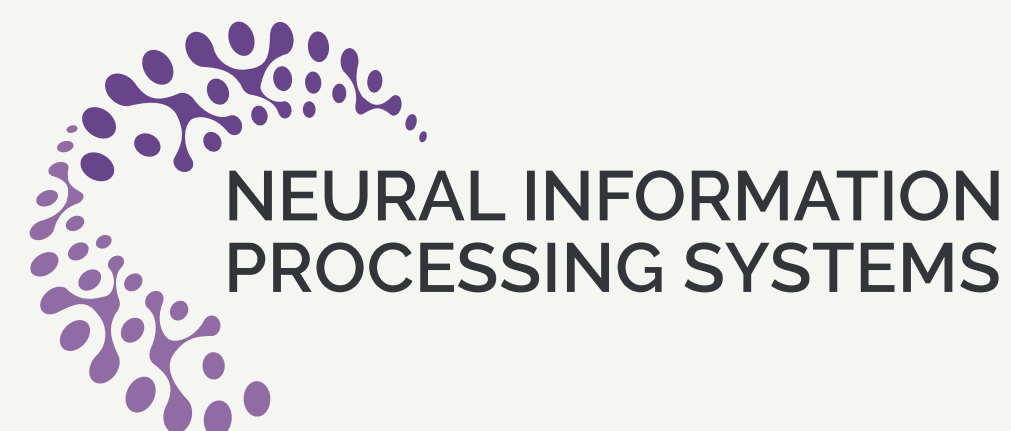
[GitHub](#)

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

**Divyam Madaan, Taro Makino, Sumit Chopra, Kyunghyun Cho**

NeurIPS 2024

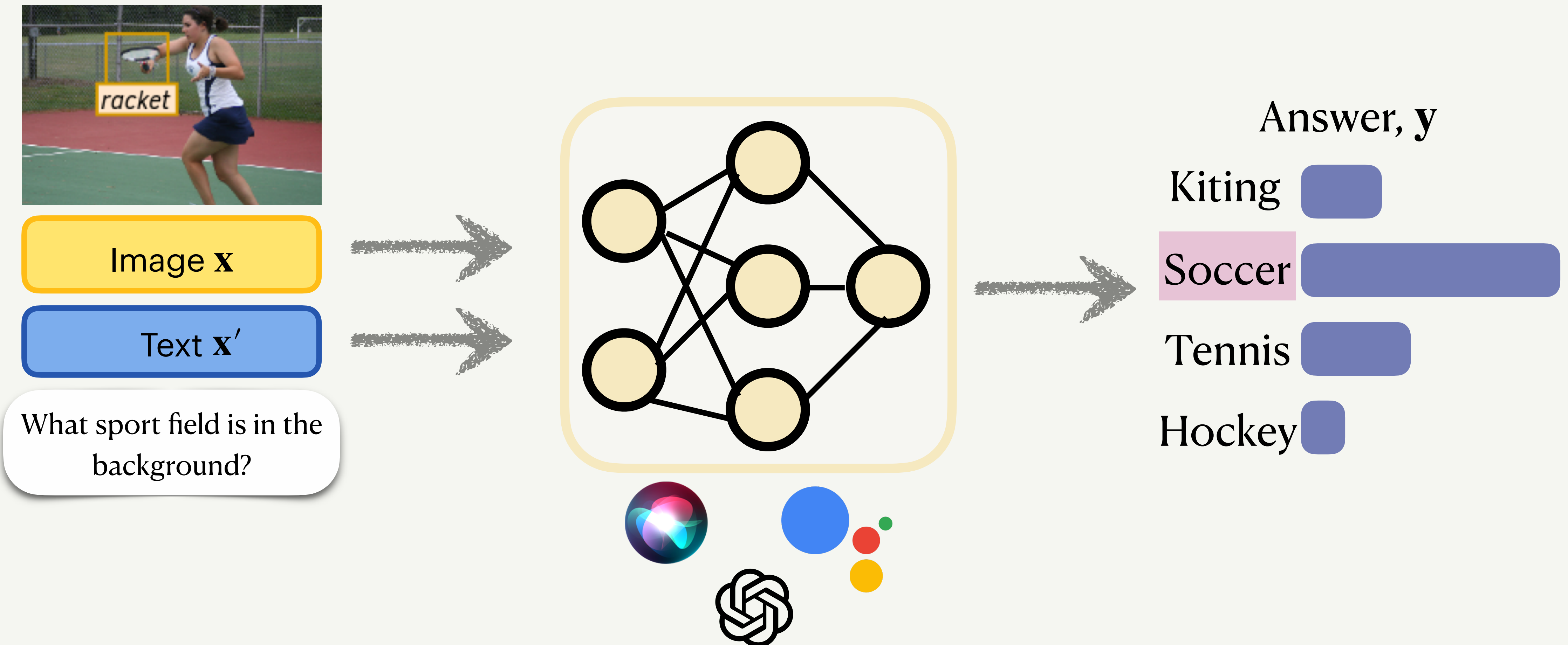
Wed 11 Dec 1 p.m. CST — 4 p.m. CST



**Genentech**

# Map multiple modalities to the target

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



# When multi-modal learning isn't better 🚫

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

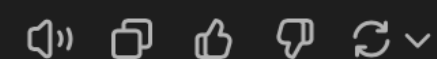


what sport field is in the background?

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.



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

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



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

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

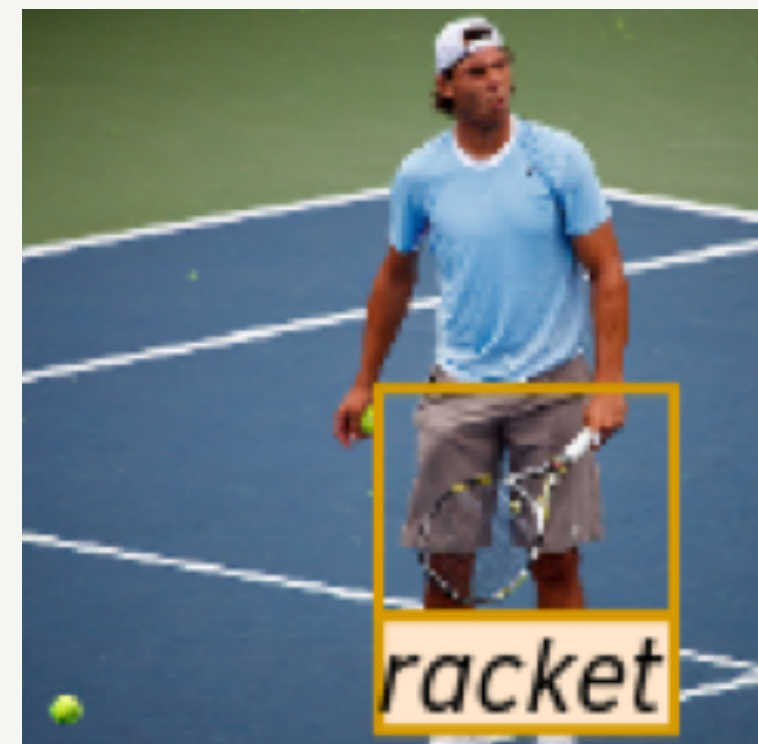


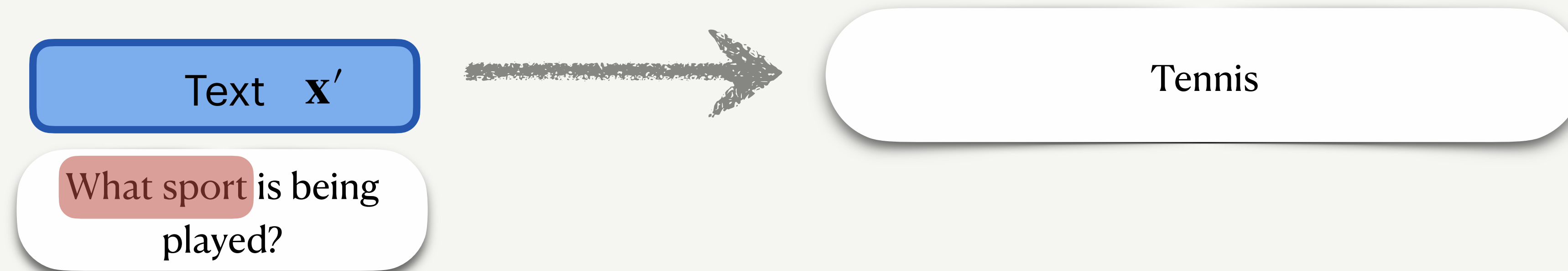
Image  $x$



Tennis

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

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



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

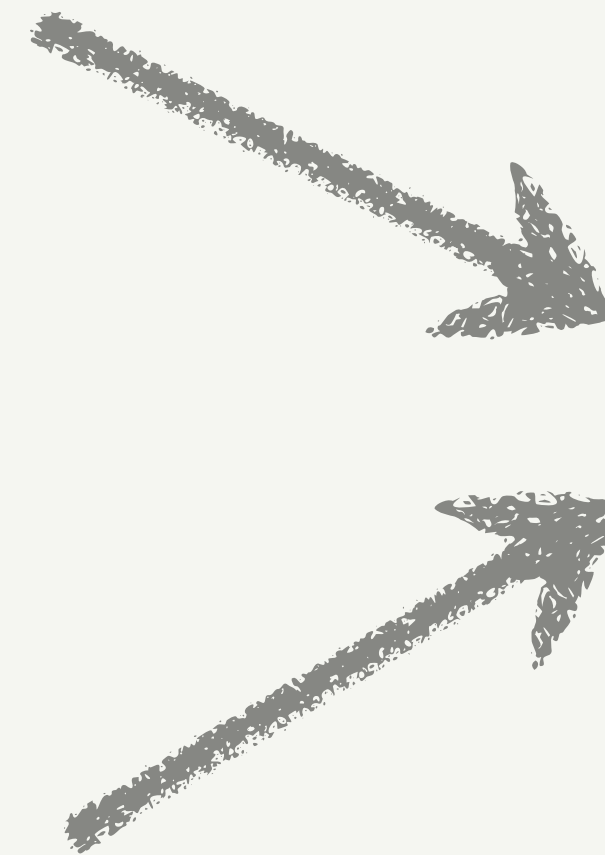
Dependency among the **features across modalities given the label**



Image  $x$

Text  $x'$

What sport field is in the background?



Soccer

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



**I2M2**

**Inter- & Intra-modality modeling**

**Modality & architecture agnostic**

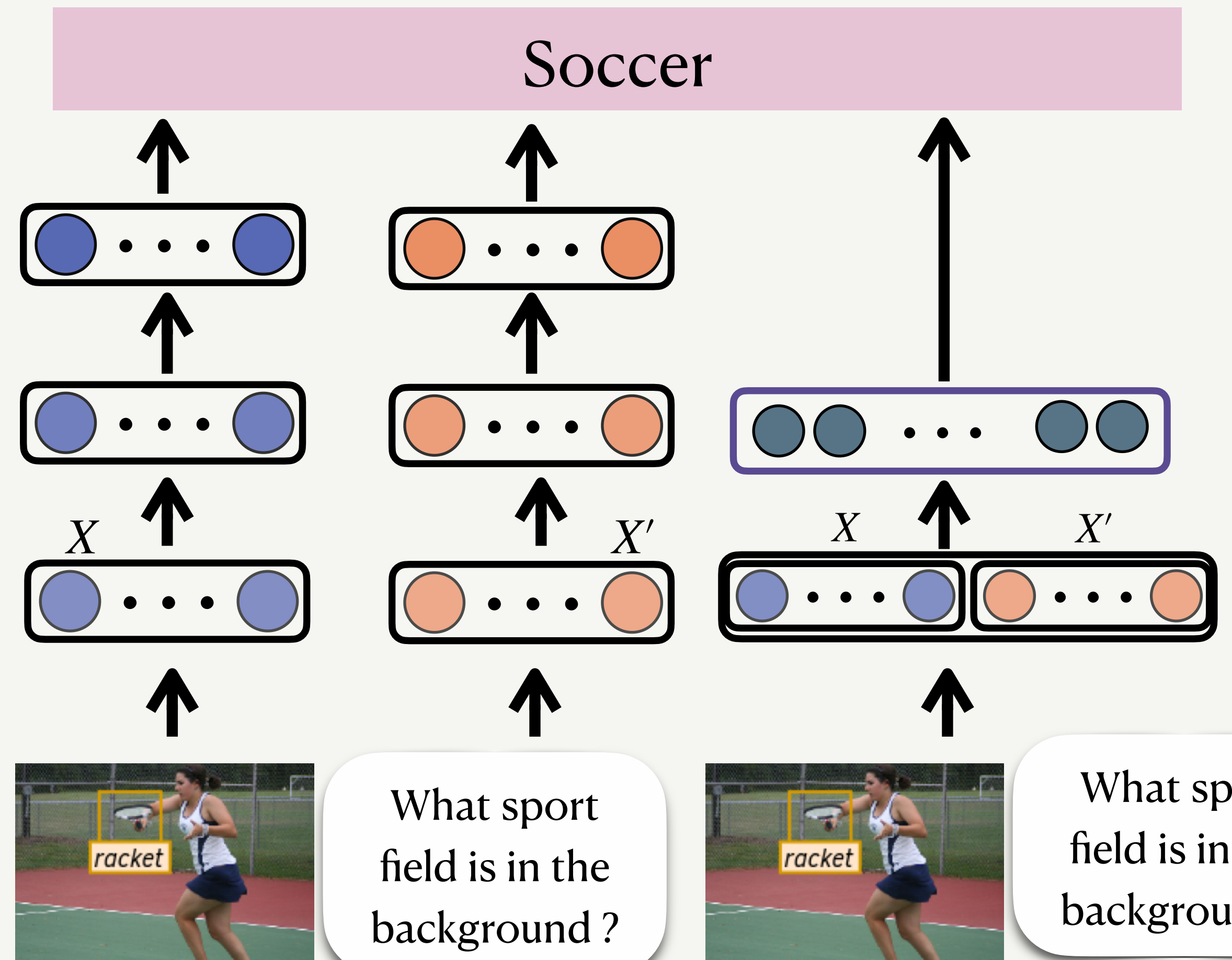
# Inter- & Intra- Modality Modeling (I2M2)



Paper

GitHub

Results on multiple healthcare, vision and language tasks 😊



What sport field is in the background?



What sport field is in the background?

Intra-modality predictors

Inter-modality predictor

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

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

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