The University of Texas at Austin
Cockrell School of Engineering

WNCG CENTER FOR aUTonomy

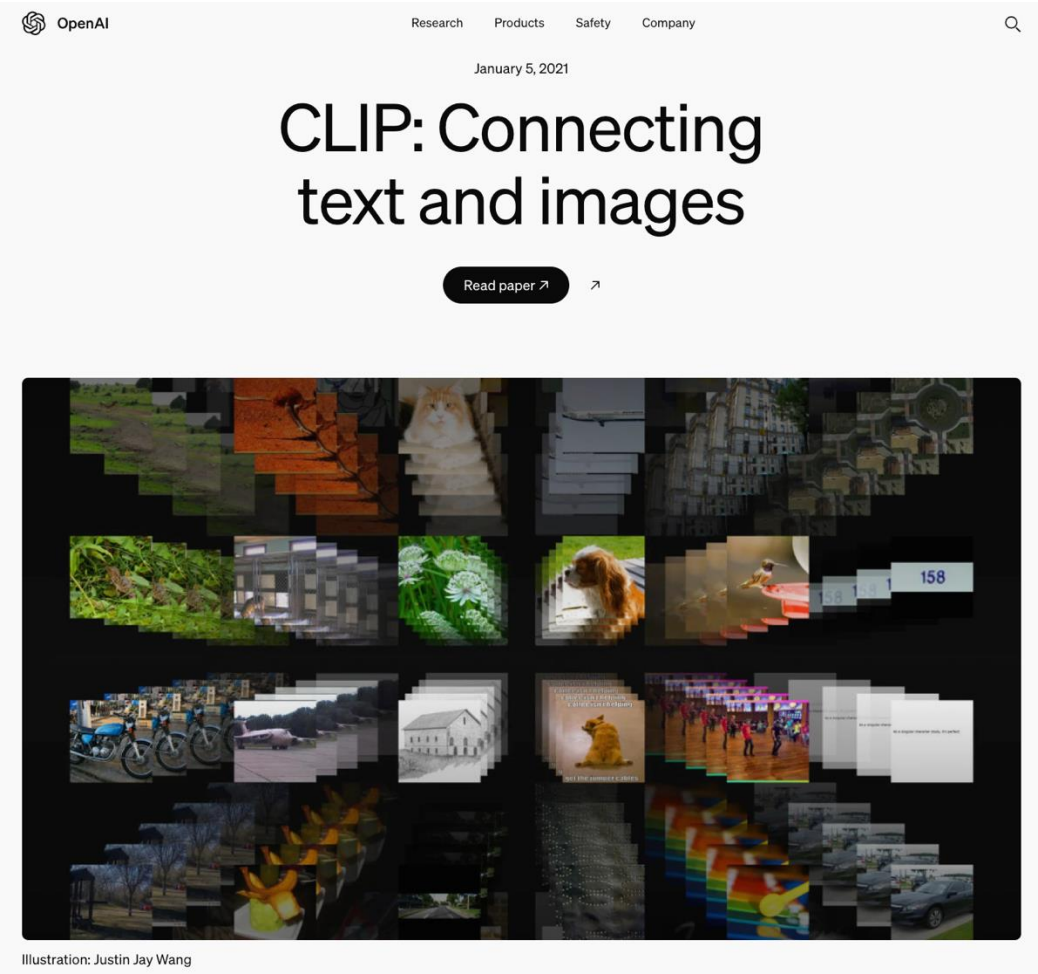
CSA: Data-efficient Mapping of Unimodal Features to Multimodal Features

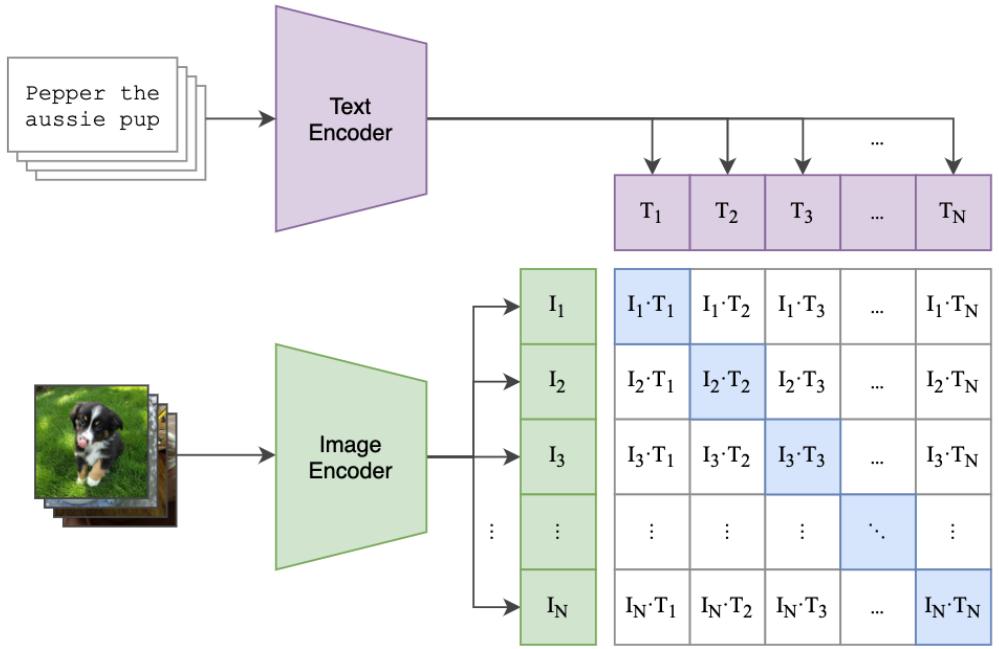
Po-han Li, Sandeep Chinchali , Ufuk Topcu

ICLR 2025

TL;DR: Can we replicate CLIP for all modalities using two unimodal encoders with 50K fewer multimodal data pairs and NO GPU training?

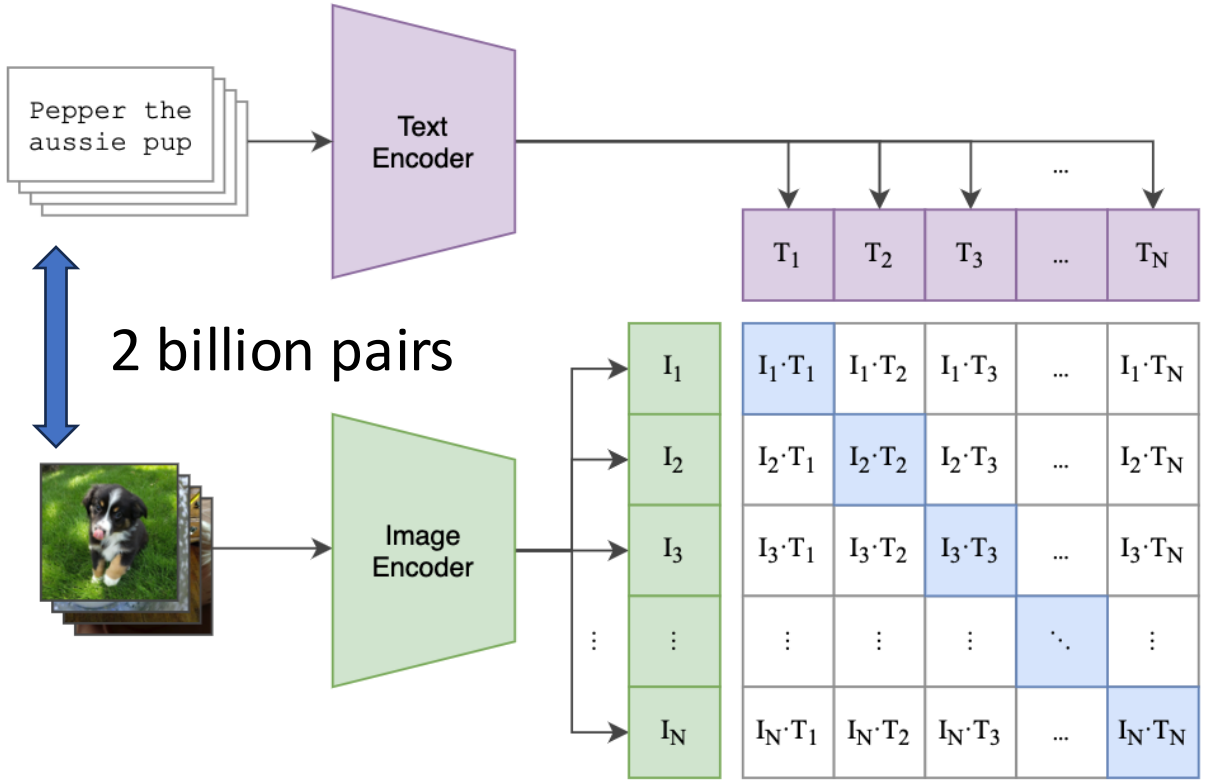
Bridging the Gap–Multimodal Encoders





Radford, Kim, et al., (2021). Learning Transferable Visual Models From Natural Language Supervision. ICML.

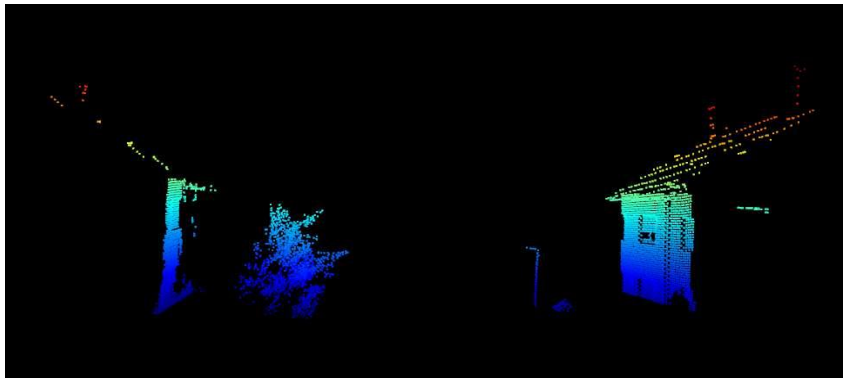
CLIP is a Data-hungry Monster!



Ilharco, Wortsman, et al., (2021). OpenCLIP. arXiv.

Src: irasutoya.com

Bridging the Gap–Emerging Pairs of Modalities

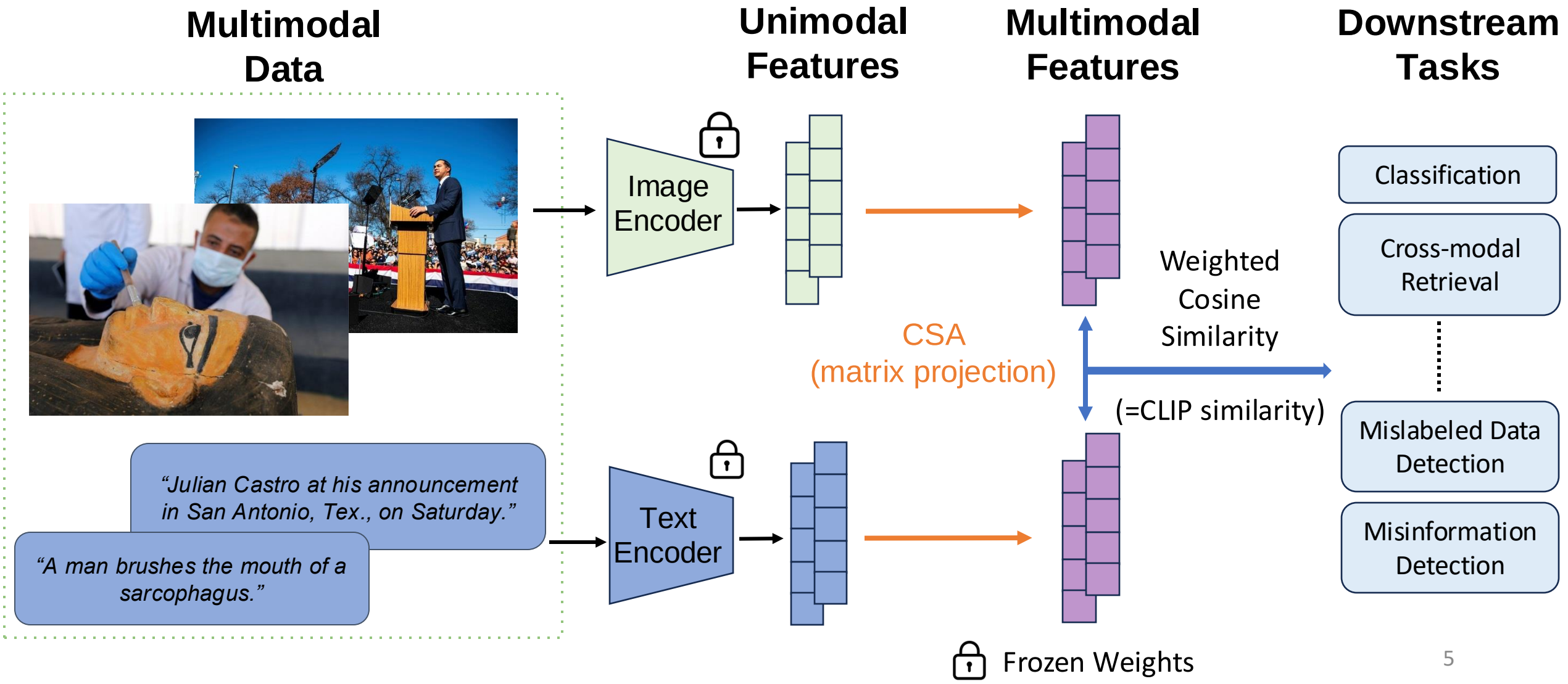


??? pairs

A street with several cars parked along the side, and two buses on the road.

Src: irasutoya.com, KITTI-360 dataset

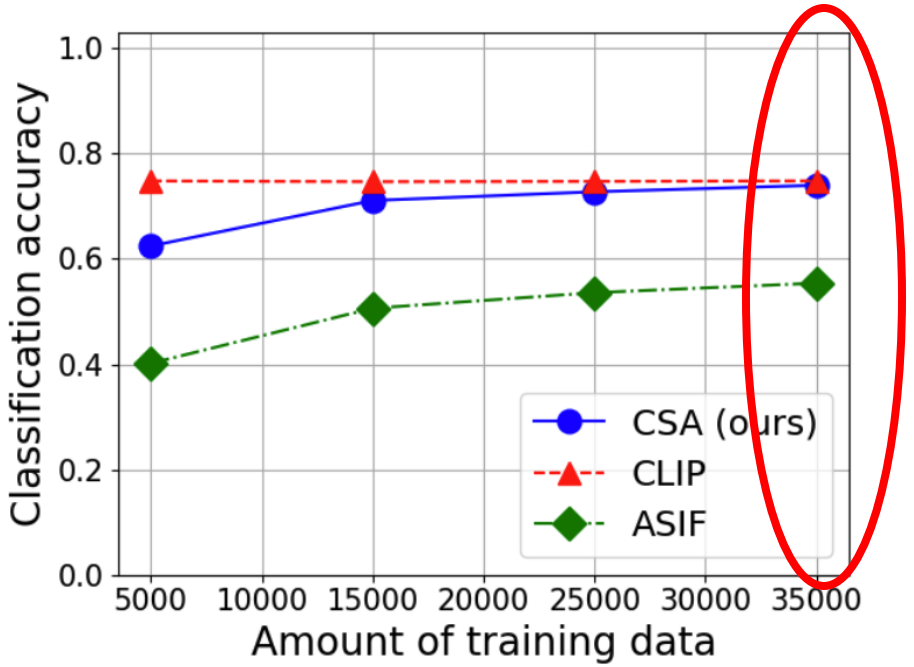
Canonical Similarity Analysis (CSA)



Bridging the Gap–Data-efficiency

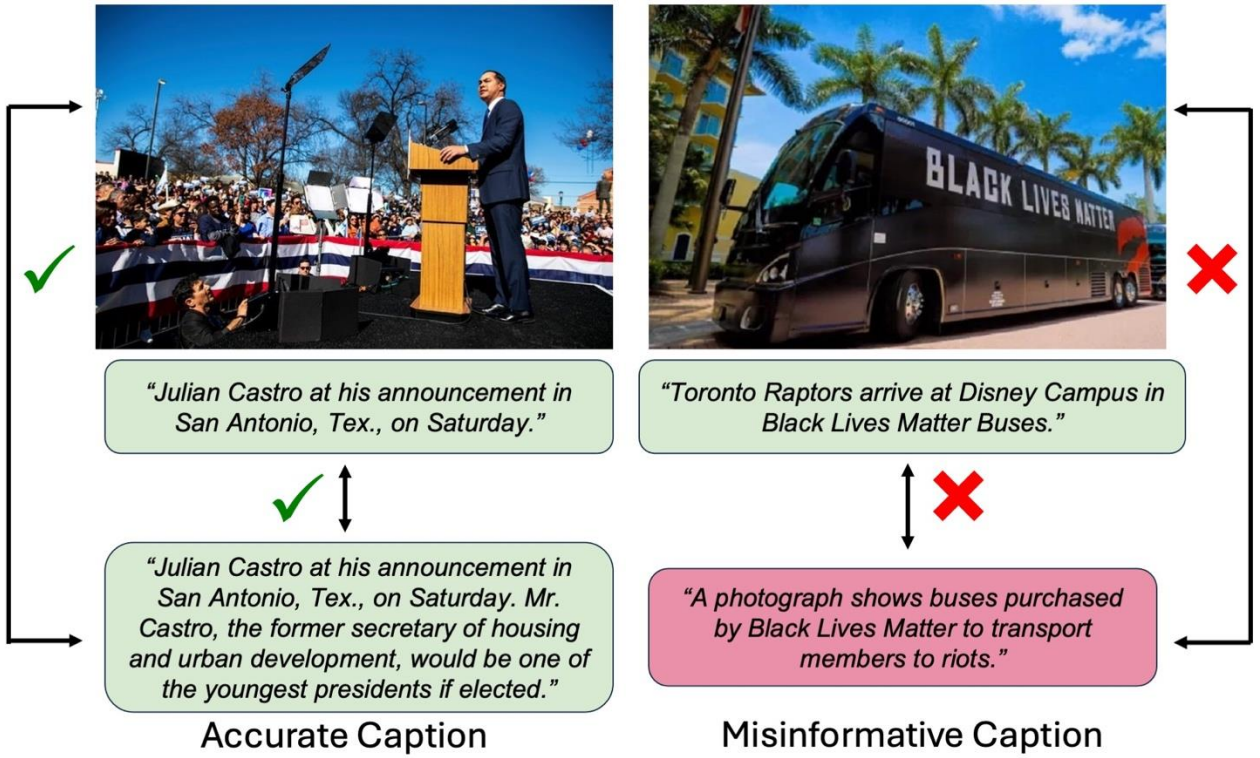
50,000x

Method	Train images	Train text	Parameter
CLIP	2B	2B	1.3B
GTR	✗	2B	335M
DINOv2	142M	✗	1.1B
CSA (ours on ImageNet)	35k	1k	✗
CSA (ours on Leafy Spurge)	800	2	✗
CSA (ours on Flickr30k)	5k	25k	✗
CSA (ours on COSMOS)	41k	41k	✗

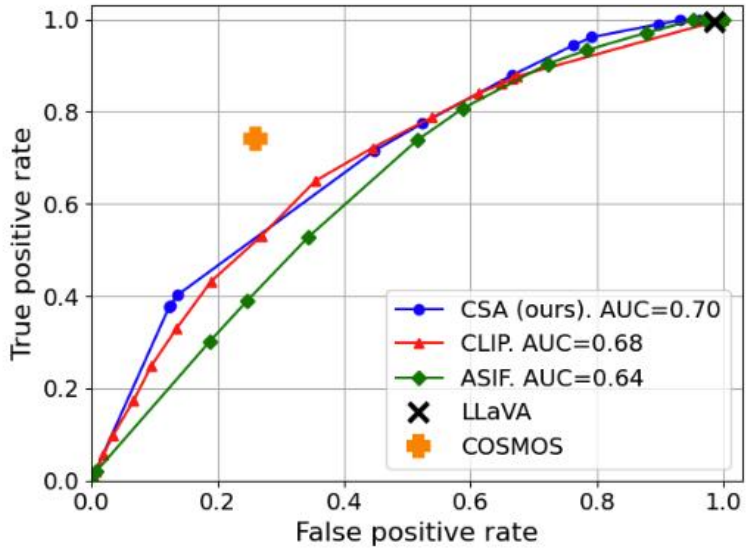


ImageNet Classification

Misinformative News Caption Detection (COSMOS)



(a) Illustrative examples.



(b) Results.

Multimodal Time Series Classification

Method	Modality 1	Modality 2	AUC (one-vs-rest)
ASIF	time series	image	0.50
	time series	text	0.50
CSA	time series	image	0.58
	time series	text	0.54

Table 5: **Multimodal time series classification:** We further demonstrated CSA's performance on multimodal time series classification, which is first in the community.

THANK YOU!

QUESTIONS?

- Po-han Li – pohanli@utexas.edu