CSCE 689: Special Topics in Advanced NLP

Lecture 17: Vision-Language Alignment

Kuan-Hao Huang

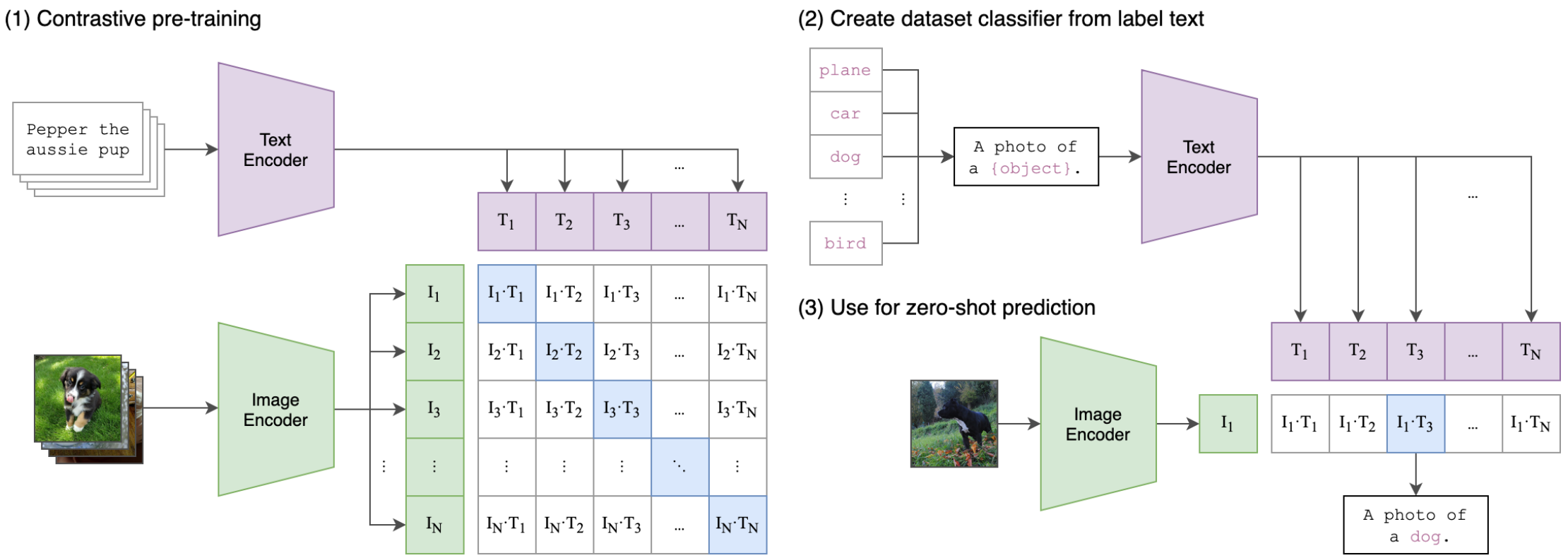
Fall 2026



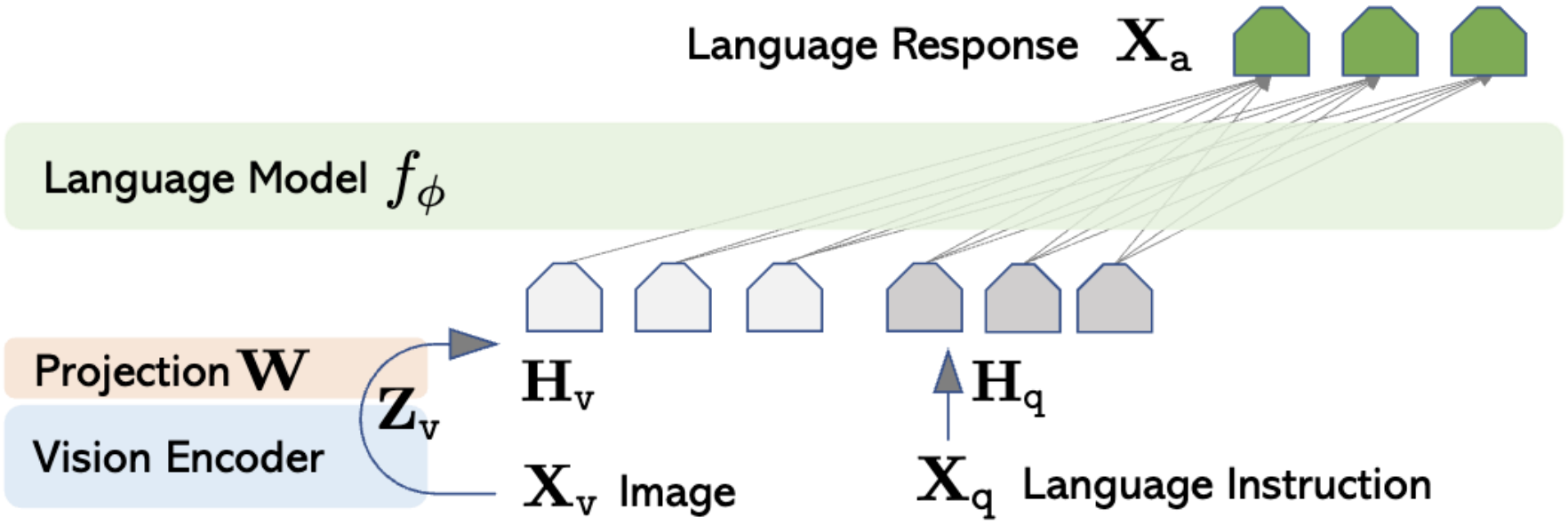
Course Project – Midterm Report

- Due: 11/5
- Page limit: 4 pages (excluding reference)
- Format: ACL style
- The report should include
  - Introduction to the topic you choose and problem definition
  - Related literature and overview of existing progress and challenges
  - Proposed solutions, novelty, and expected contributions
  - Evaluation metrics
  - Planned implementation details, including dataset, models, codebases, etc.
  - Current progress (preliminary results, baseline results, etc.)
  - Next steps and expected timeline

CLIP: Contrastive Language-Image Pre-Training



LLaVA: Large Language and Vision Assistant



When and Why Vision-Language Models Behave like Bags-Of-Words, and What to Do About It?

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, James Zou
Stanford University
Stanford, CA 94305
{merty, fede, pkalluri, jurafsky, jamesz}@stanford.edu

ARO (Attribution, Relation and Order) Benchmark

Visual Genome Relation

Assessing relational understanding (23,937 test cases)



✔ *the horse is eating the grass*
✘ *the grass is eating the horse*

Visual Genome Attribution

Assessing attributive understanding (28,748 test cases)



✔ *the paved road and the white house*
✘ *the white road and the paved house*

ARO (Attribution, Relation and Order) Benchmark

COCO Order and Flickr Order

Assessing sensitivity to order (6,000 test cases)



✔ a brown cat is looking at a gray dog and sitting in a white bathtub

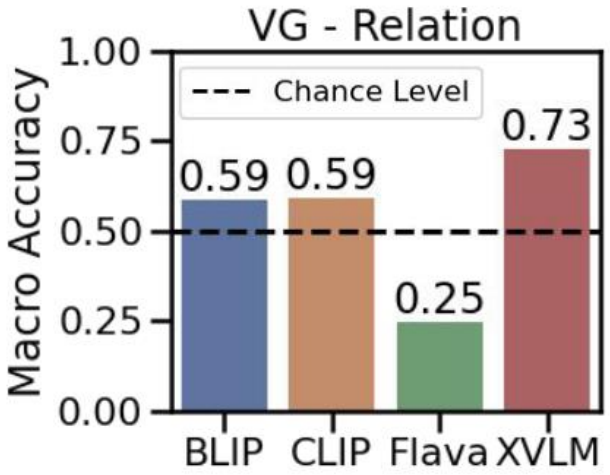
✗ (shuffle adjective/noun) a gray bathtub is looking at a white cat and sitting in a brown dog

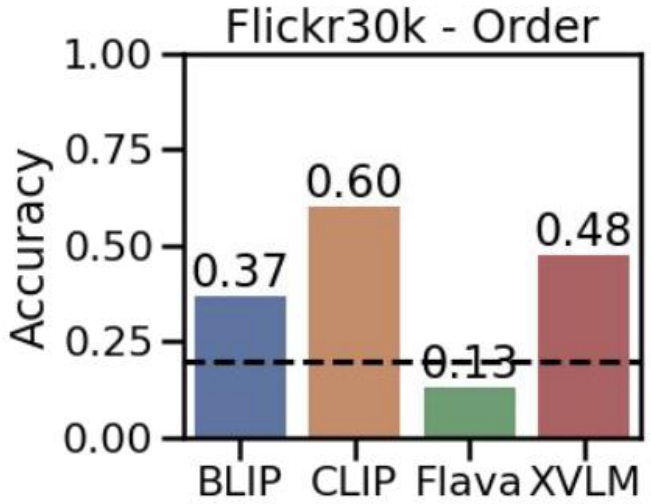
✗ (shuffle all but adjective/noun) at brown cat a in looking a gray dog sitting is and a white bathtub

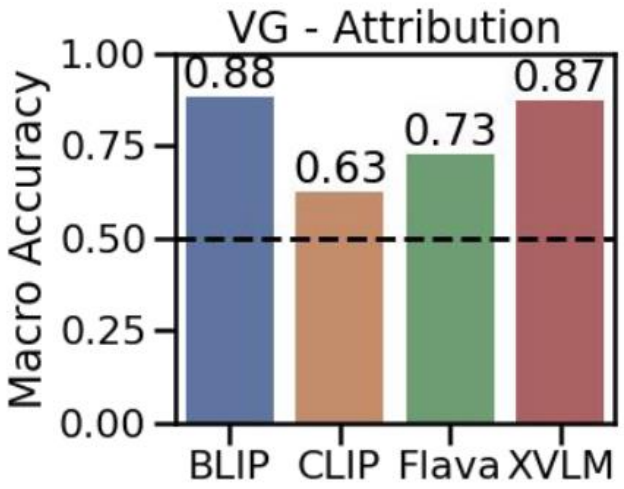
✗ (shuffle words within trigrams) cat brown a at is looking a gray dog in and sitting bathtub a white

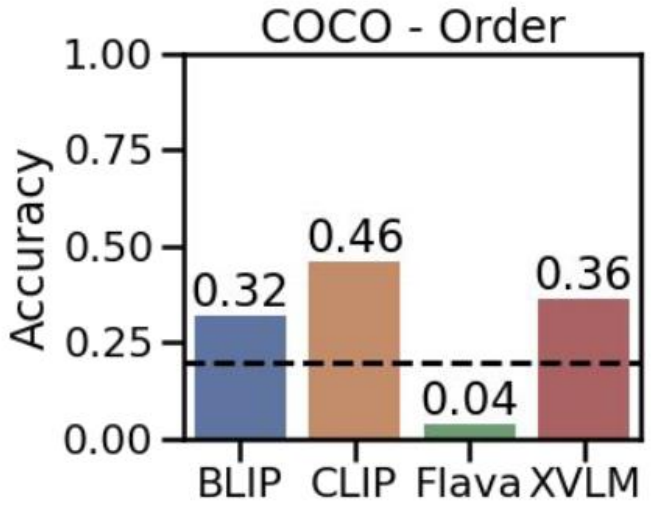
✗ (shuffle trigrams) a brown cat a white bathtub is looking at a gray dog and sitting in

Results

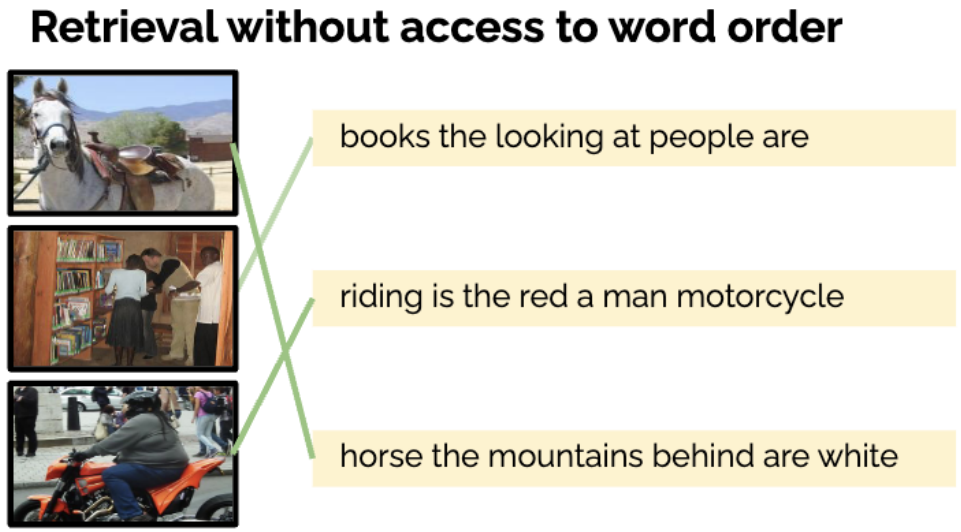

VG - Relation
Macro Accuracy
1.00
0.75
0.50
0.25
0.00
Chance Level
0.59
0.59
0.25
0.73
BLIP
CLIP
Flava
XVLM

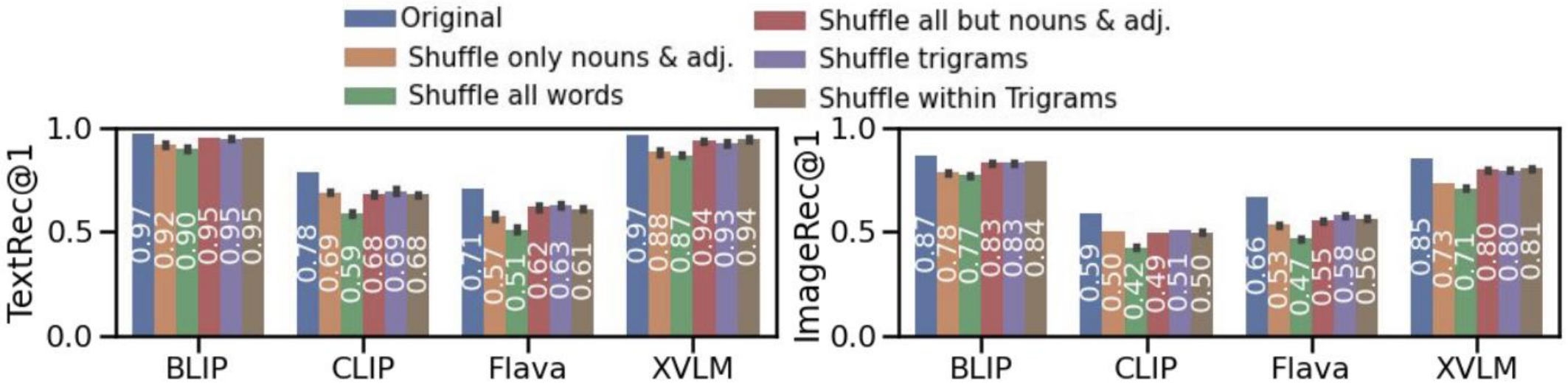

Flickr30k - Order
Accuracy
1.00
0.75
0.50
0.25
0.00
0.37
0.60
0.13
0.48
BLIP
CLIP
Flava
XVLM


VG - Attribution
Macro Accuracy
1.00
0.75
0.50
0.25
0.00
0.88
0.63
0.73
0.87
BLIP
CLIP
Flava
XVLM

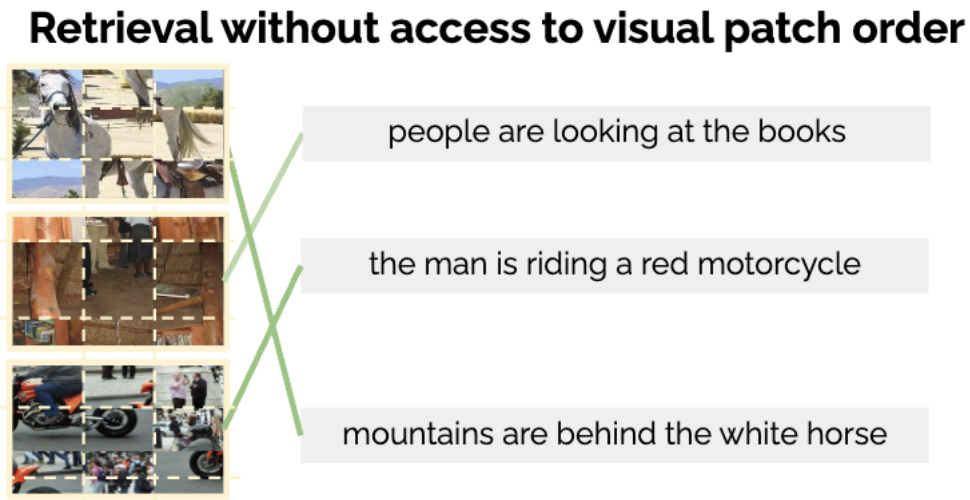

COCO - Order
Accuracy
1.00
0.75
0.50
0.25
0.00
0.32
0.46
0.04
0.36
BLIP
CLIP
Flava
XVLM

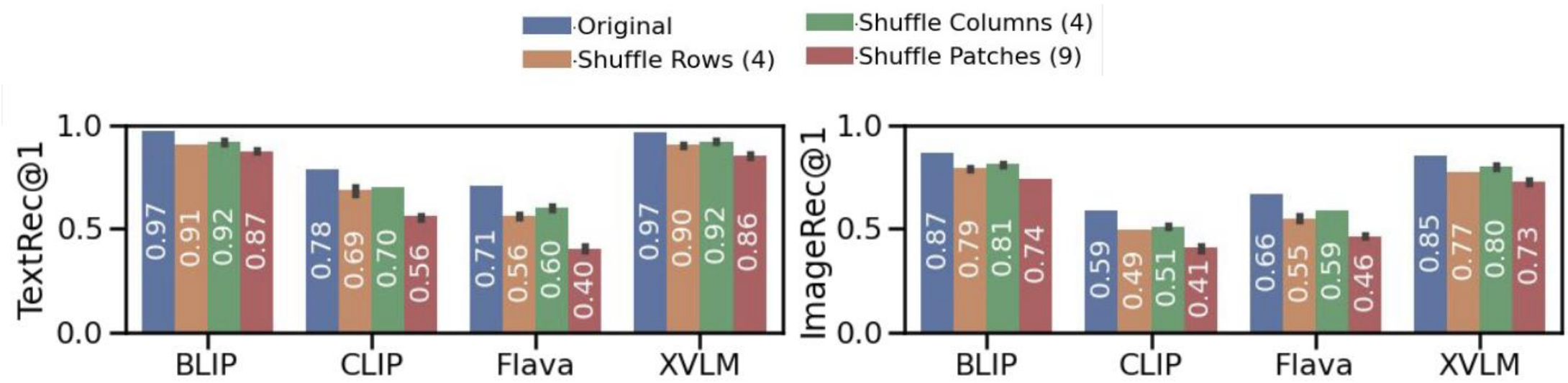
More Analysis on Text-Image Retrieval



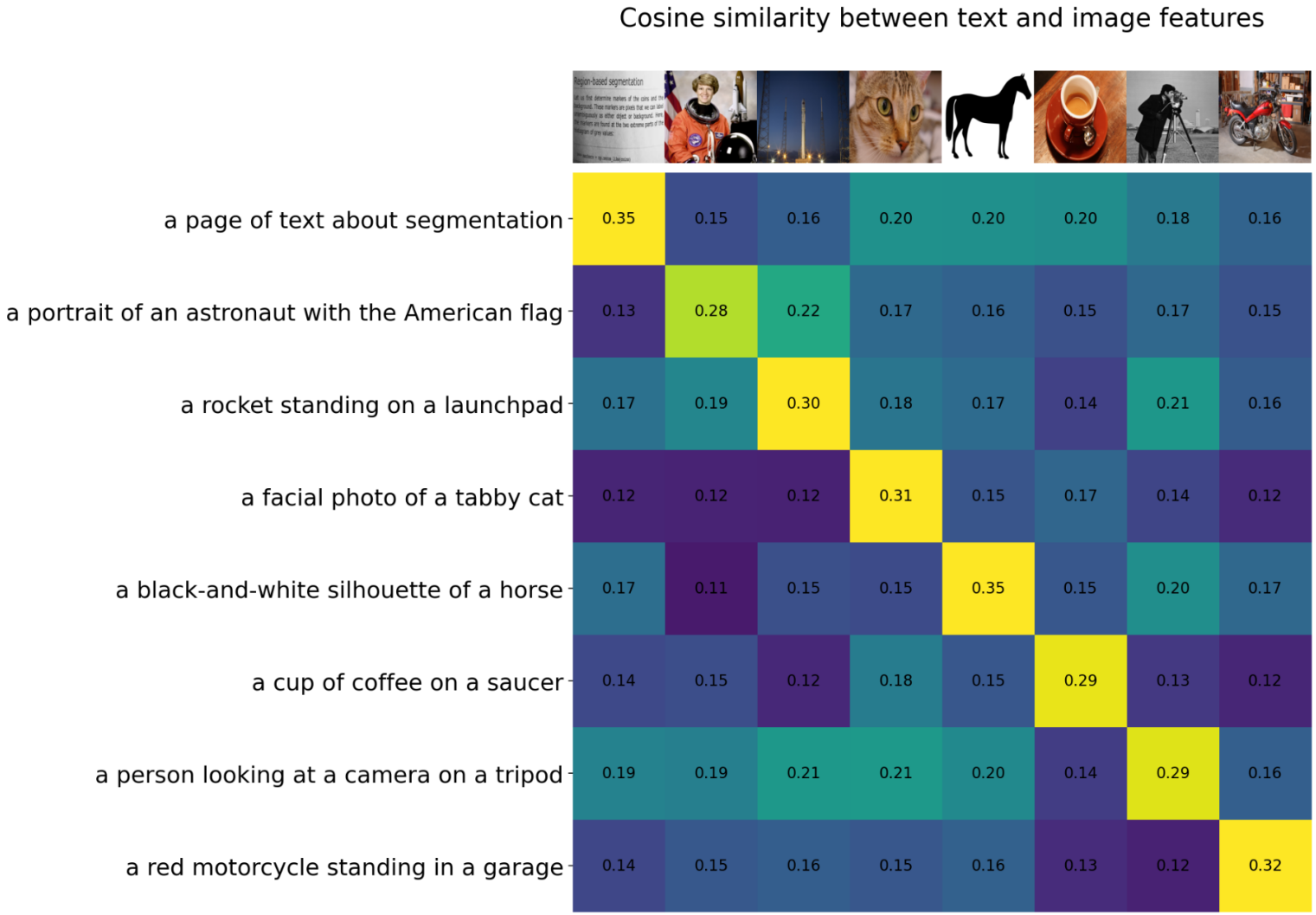


More Analysis on Text-Image Retrieval





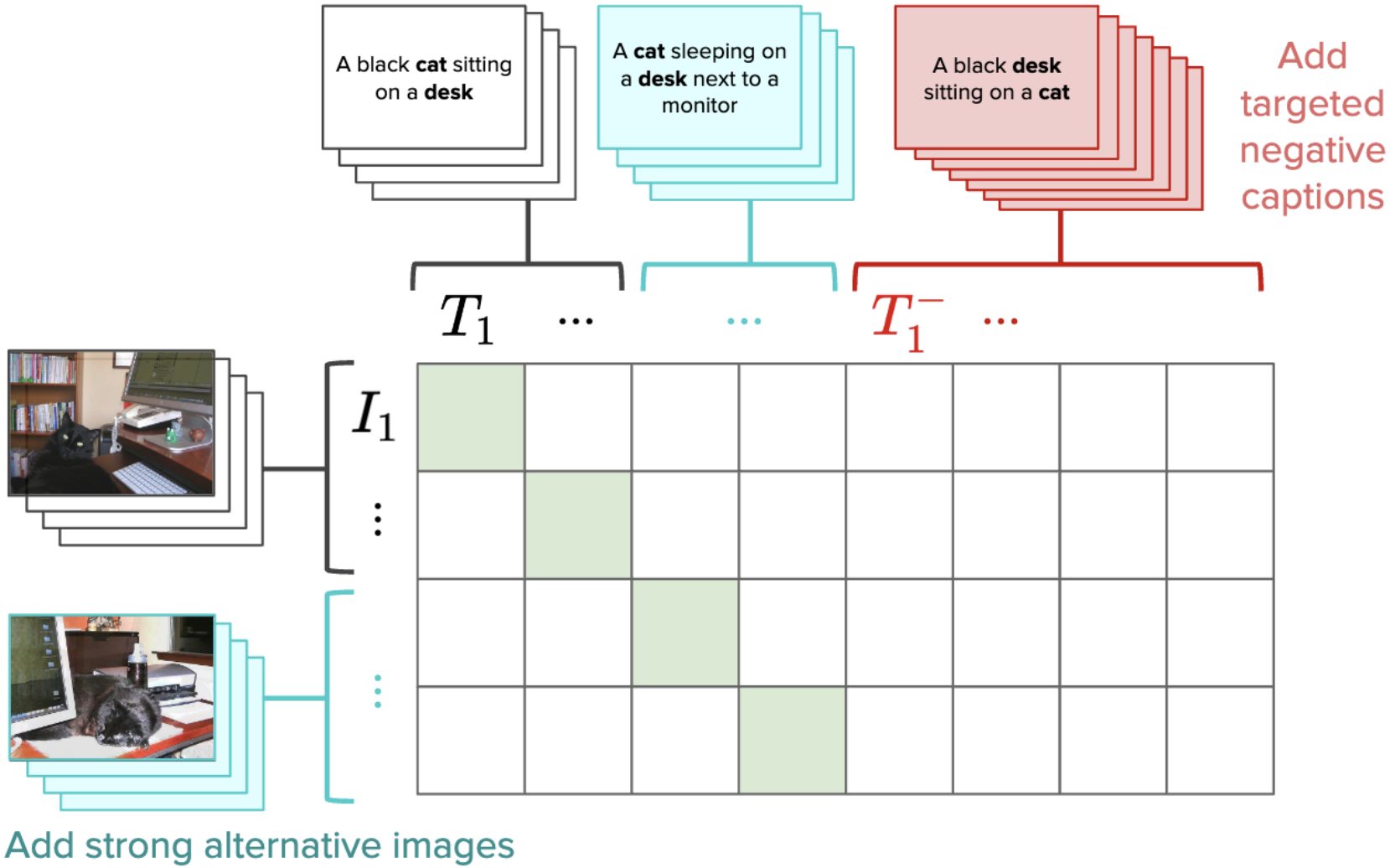
Contrastive Pre-Training



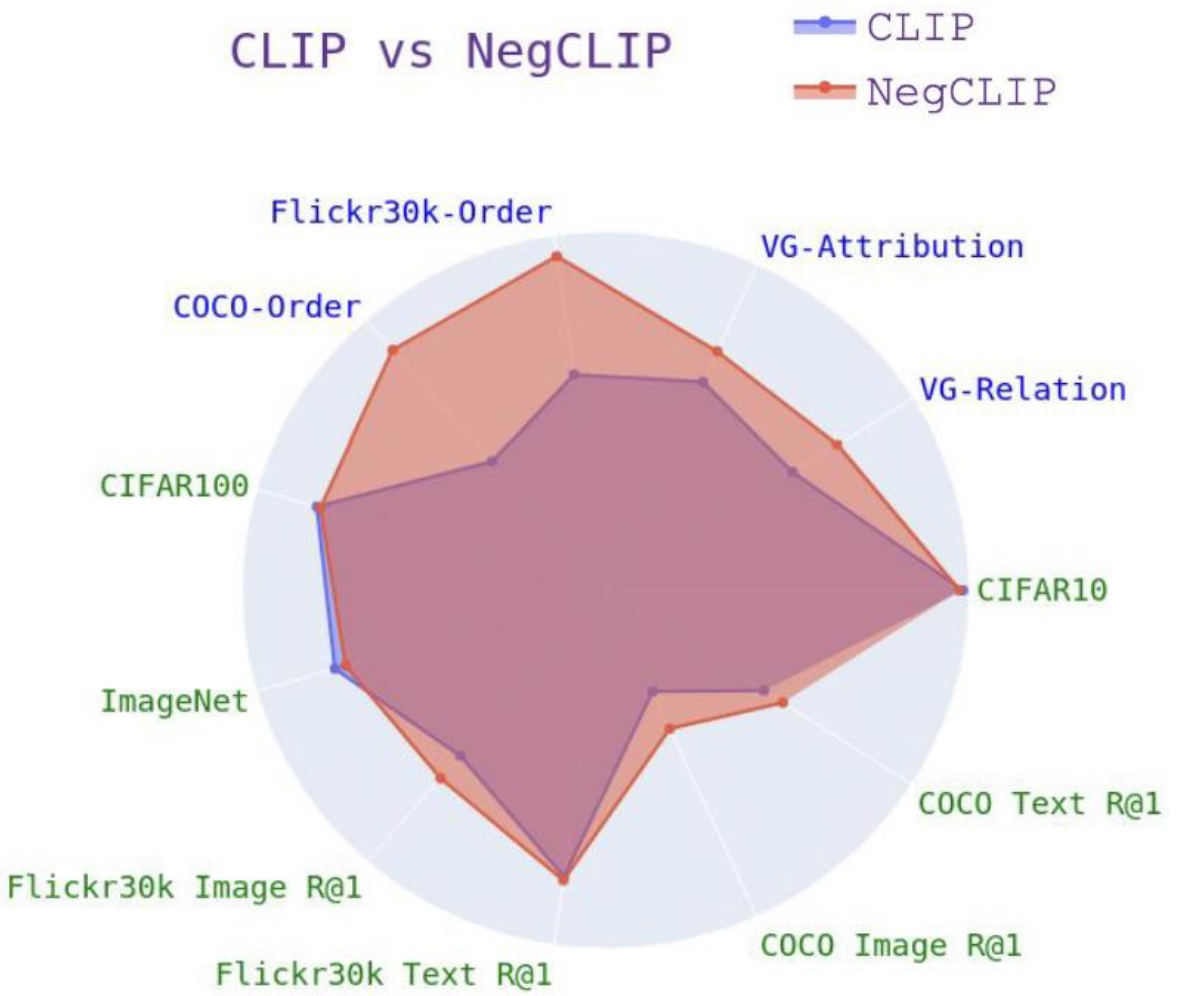
Solution: Composition-Aware Hard Negatives

- Generation of negative captions
  - Generate a negative caption by swapping different linguistic elements: noun phrases, nouns, adjectives, adverbs, verb phrases
- Sampling strong alternative images
  - Sample one of the K nearest neighbors as the strong alternative image

NegCLIP


A black **cat** sitting on a **desk**
A **cat** sleeping on a **desk** next to a monitor
A black **desk** sitting on a **cat**
Add targeted negative captions
T_1 ...
...
T_1^- ...
I_1
Add strong alternative images

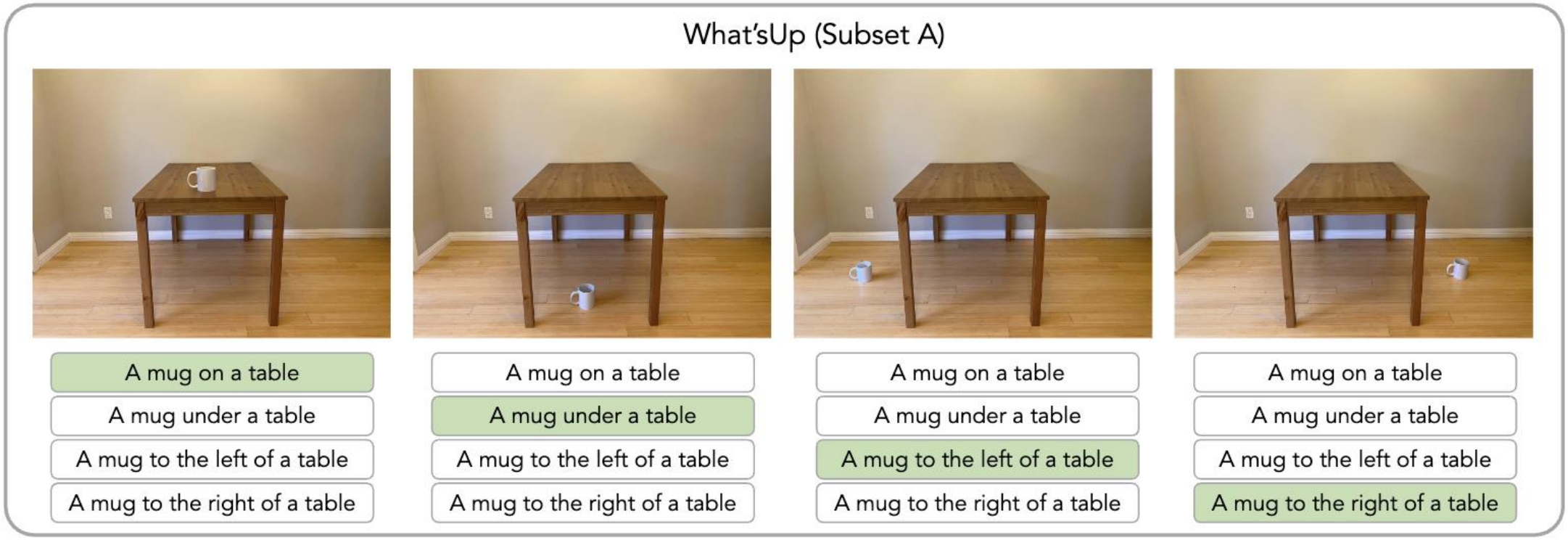
Results


CLIP vs NegCLIP
CLIP
NegCLIP
Flickr30k-Order
VG-Attribution
COCO-Order
VG-Relation
CIFAR100
CIFAR10
ImageNet
COCO Text R@1
Flickr30k Image R@1
COCO Image R@1
Flickr30k Text R@1

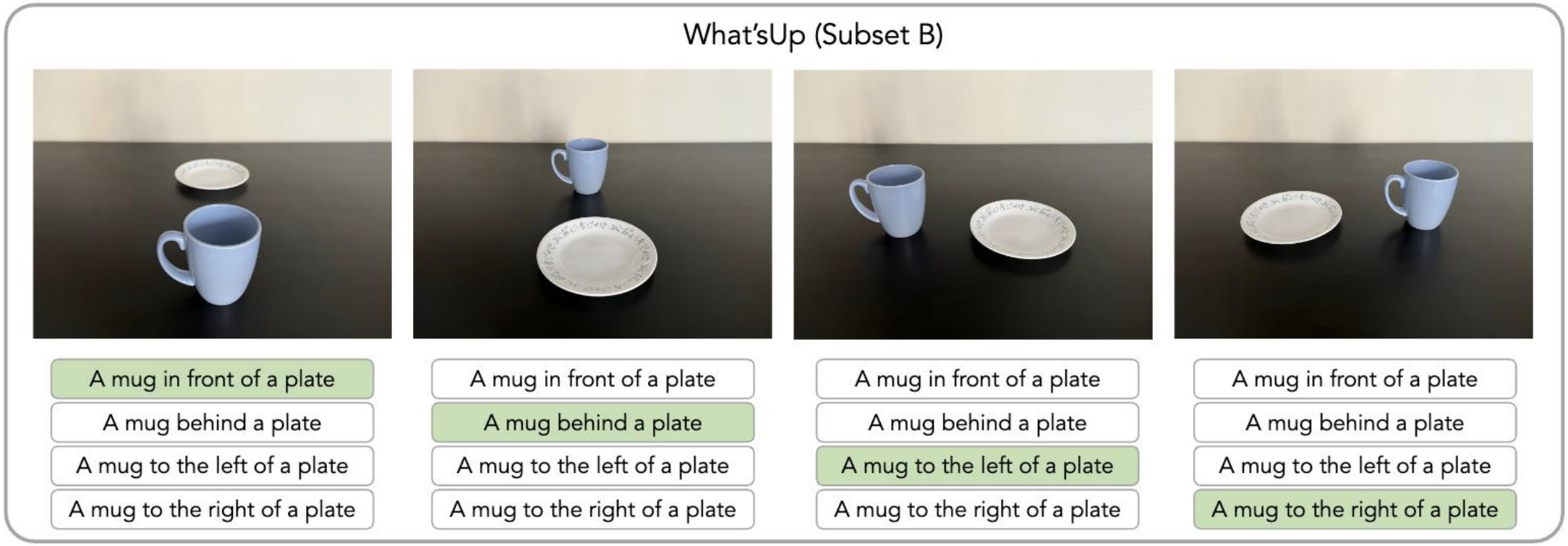
What's "up" with vision-language models? Investigating their struggle with spatial reasoning

Amita Kamath[1] **Jack Hessel**[2] **Kai-Wei Chang**[1]
[1] University of California, Los Angeles
[2] Allen Institute for AI
{kamatha, kwchang}@cs.ucla.edu, jackh@allenai.org

What's Up: Benchmarks with Spatial Reasoning Questions



What's Up: Benchmarks with Spatial Reasoning Questions



Results

Model	Whats-Up	COCO-spatial	GQA-spatial	Avg
CLIP ViT-B/32	31.0	47.4	46.9	41.8
CLIP ViT-L/14	26.1	49.5	47.3	41.0
NegCLIP	34.4	46.9	46.0	42.4
RoBERTaCLIP	25.1	50.0	49.8	41.6
CoCa	29.4	46.7	47.1	41.0
XVLM 4M	31.5	61.7	**58.7**	50.6
XVLM 16M	**41.9**	**65.0**	58.2	**55.0**
BLIP 14M	38.5	54.0	49.8	47.5
BLIP 129M	30.4	49.3	49.0	42.9
BLIP2-ITM	37.6	53.0	49.8	46.8
BLIP2-ITC	29.0	53.7	51.0	44.6
FLAVA	30.5	52.6	51.7	44.9
CoCa-Caption	24.1	48.6	49.5	40.8
XVLM-Flickr30K	44.3	65.2	61.4	56.9
XVLM-COCO	42.1	**71.0**	**68.1**	**60.4**
BLIP-Flickr30K	33.8	54.2	48.9	45.6
BLIP-COCO	32.8	51.4	51.4	45.2
BLIP-VQA	**47.8**	62.0	58.4	56.0
Random / Text-only	25.0	50.0	50.0	41.7
Human Estimate	100.0	97.3	99.0	98.8

Visual Analogies



$$I(\text{red mug}) - I(\text{yellow mug}) + I(\text{yellow bowl})$$

$$I(\text{red bowl}) \ ?$$

61%

Visual Analogies

$$I(\text{mug on table}) - I(\text{mug under table}) + I(\text{bowl under table})$$

$$I(\text{bowl on table}) \ ?$$

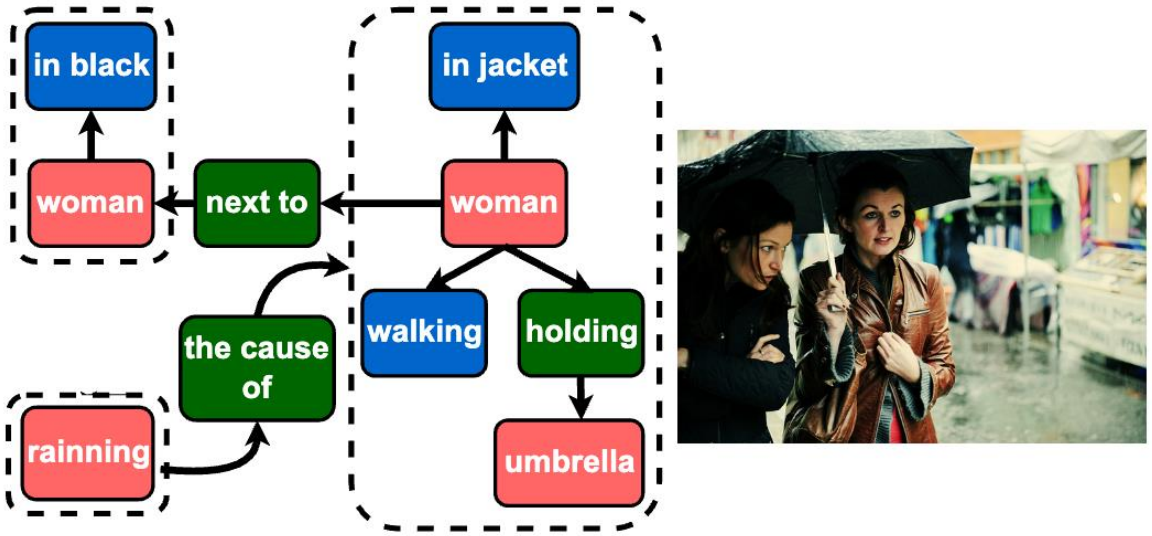
9%

What's Missing in Vision-Language Models? Probing Their Struggles with Causal Order Reasoning

Zhaotian Weng[1], Haoxuan Li[1], Xin Eric Wang[3], Kuan-Hao Huang[2], and Jieyu Zhao[1]

[1]University of Southern California
[2]Texas A&M University
[3]University of California, Santa Barbara
{wengzhao, lihaoxua, jieyuz}@usc.edu, ericxwang@ucsb.edu, khhuang@tamu.edu

Multimodal Models Struggle with Causal Reasoning



VQA-Causal:

The woman's holding an umbrella is caused by the rain. ✓

The rain is caused by the woman holding an umbrella. ✗

Object and Activity Understanding Test:

The woman is holding an umbrella. ✓

The woman is running. ✗

It is raining. ✓

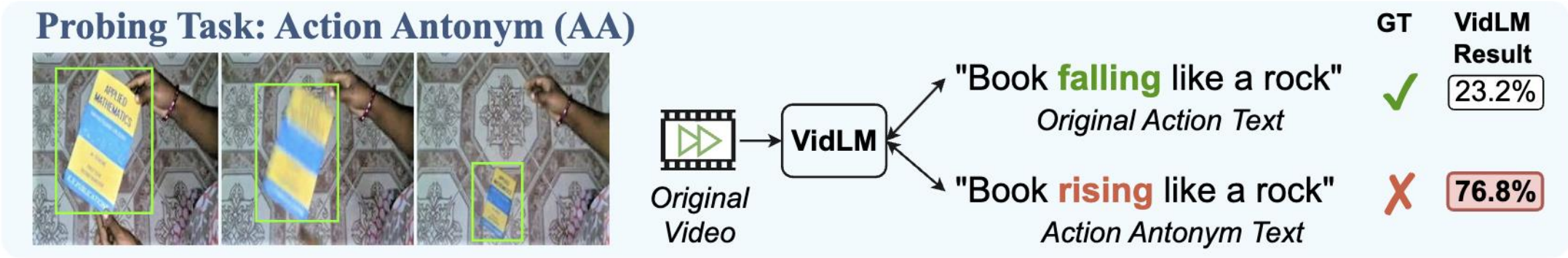
It is sunny. ✗

Model	VQA-Causal	VCR-Causal
	Score-Based Models	
BLIP ITM Base	48.94	50.66
BLIP ITM Large	48.68	47.99
BLIP2 ITM	50.76	49.95
BLIP2 FE	51.51	50.76
CLIP ViT B/32	51.62	50.35
CLIP ViT L/14	50.74	51.66
NegCLIP	50.89	51.30
RobustCLIP	50.66	53.68
FLAVA	48.52	49.99
	Text-Generation Models	
Vicuna 1.5	50.86	56.03
LLaVA 1.5	53.19	52.12
LLaVA 1.6	58.11	57.24
Qwen2.5-VL	50.66	47.26
Qwen3-VL	42.01	43.73
Random	50.00	50.00
Human Estimate	99.17	98.17

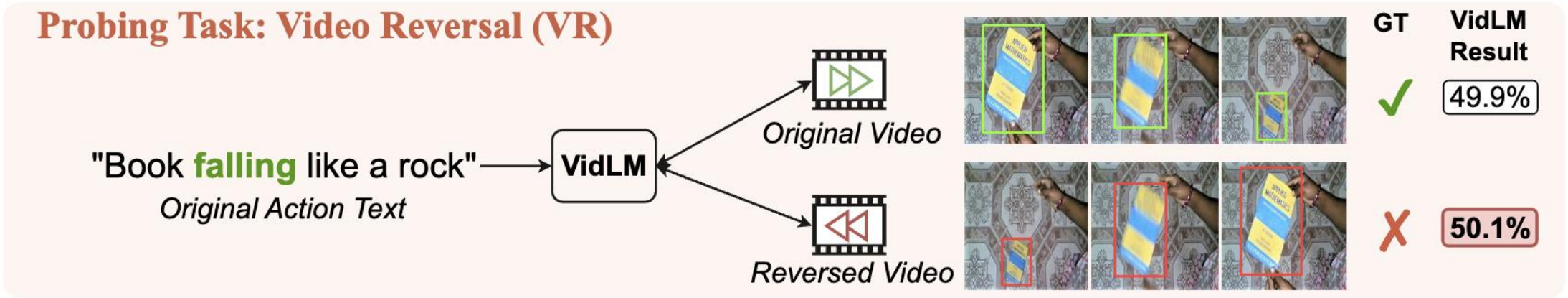
PAXION: Patching Action Knowledge in Video-Language Foundation Models

Zhenhailong Wang[1], **Ansel Blume**[1], **Sha Li**[1], **Genglin Liu**[1],
Jaemin Cho[2], **Zineng Tang**[2], **Mohit Bansal**[2], **Heng Ji**[1]
[1]UIUC [2]UNC
{wangz3,hengji}@illinois.edu

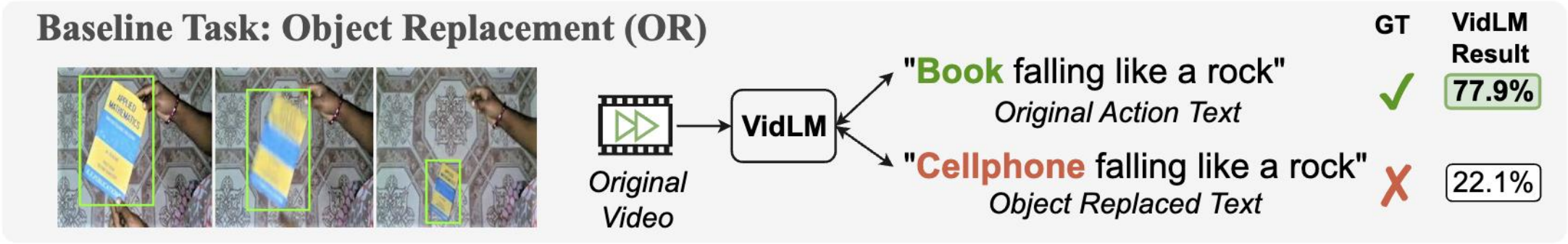
Probing Task: Action Antonym



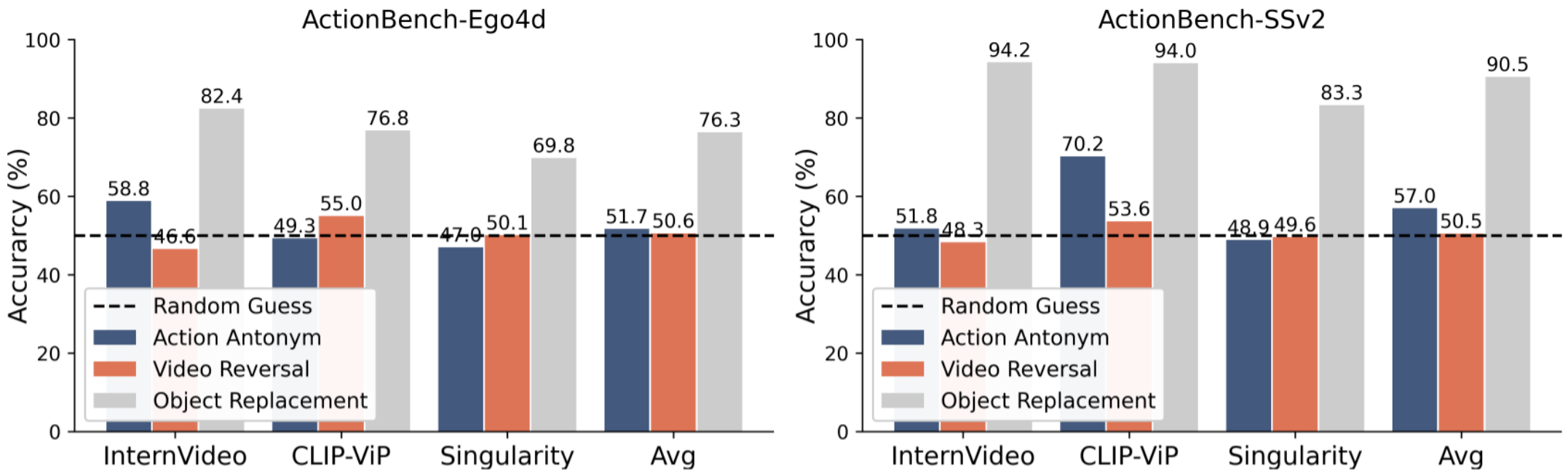
Probing Task: Video Reversal



Baseline Task: Object Replacement



Results


ActionBench-Ego4d
Accuracy (%)
100
80
60
40
20
0
58.8
46.6
82.4
49.3
55.0
76.8
47.0
50.1
69.8
51.7
50.6
76.3
Random Guess
Action Antonym
Video Reversal
Object Replacement
InternVideo
CLIP-ViP
Singularity
Avg
ActionBench-SSv2
Accuracy (%)
100
80
60
40
20
0
51.8
48.3
94.2
70.2
53.6
94.0
48.9
49.6
83.3
57.0
50.5
90.5
Random Guess
Action Antonym
Video Reversal
Object Replacement
InternVideo
CLIP-ViP
Singularity
Avg

Hidden in plain sight:
VLMs overlook their visual representations

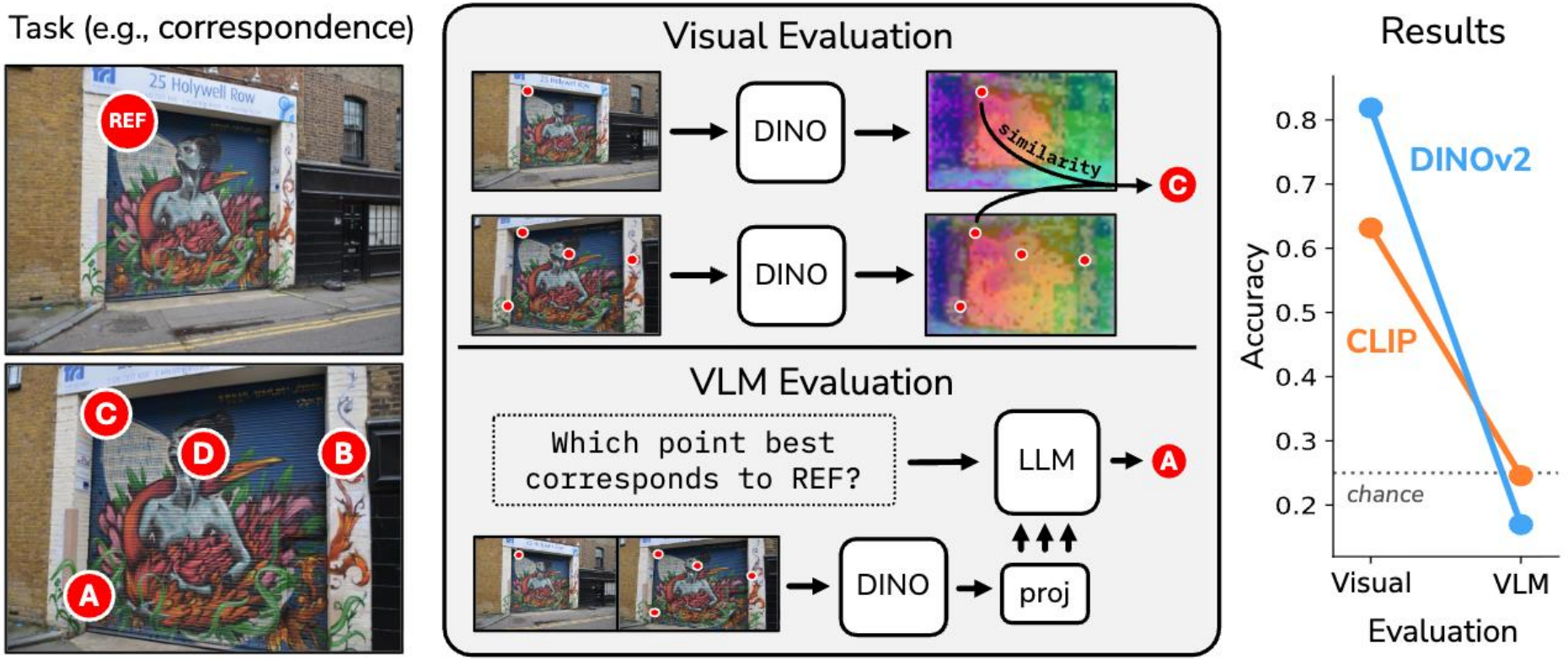
Stephanie Fu
UC Berkeley

Tyler Bonnen
UC Berkeley

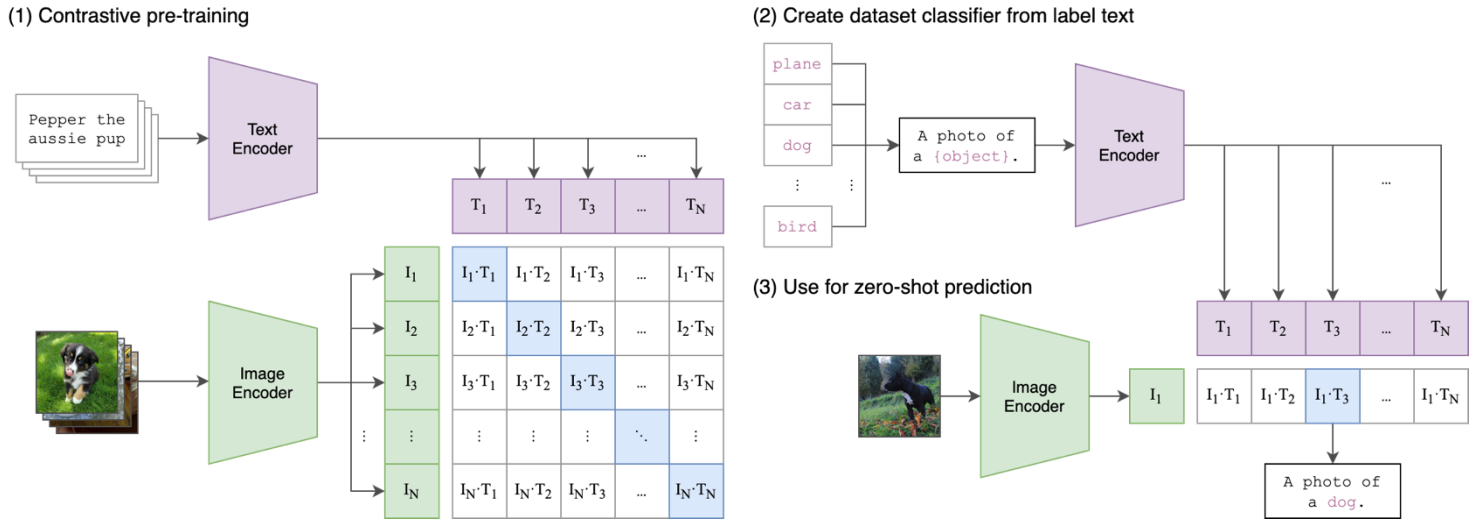
Devin Guillory
UC Berkeley

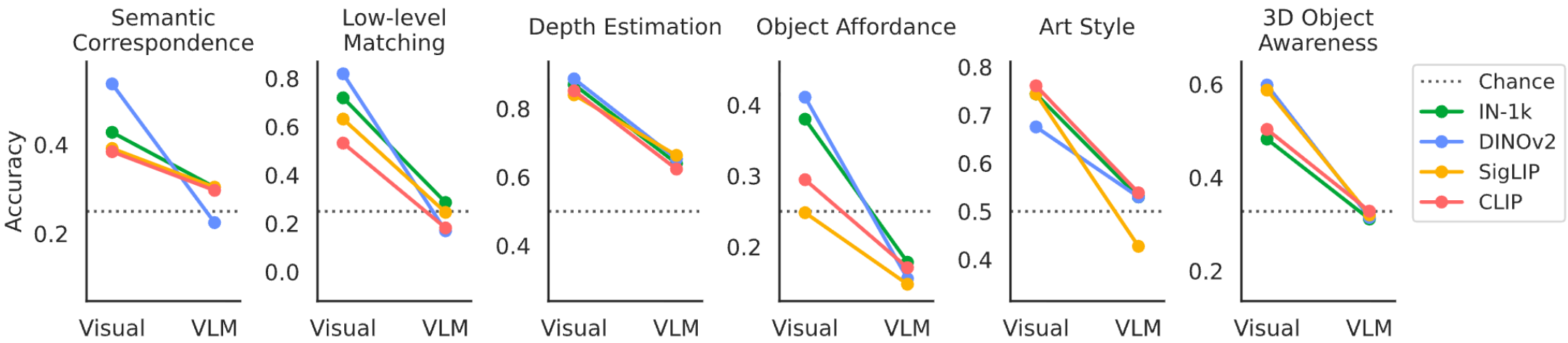
Trevor Darrell
UC Berkeley

Vision Encoders Reveals Failure to Utilize Visual Information

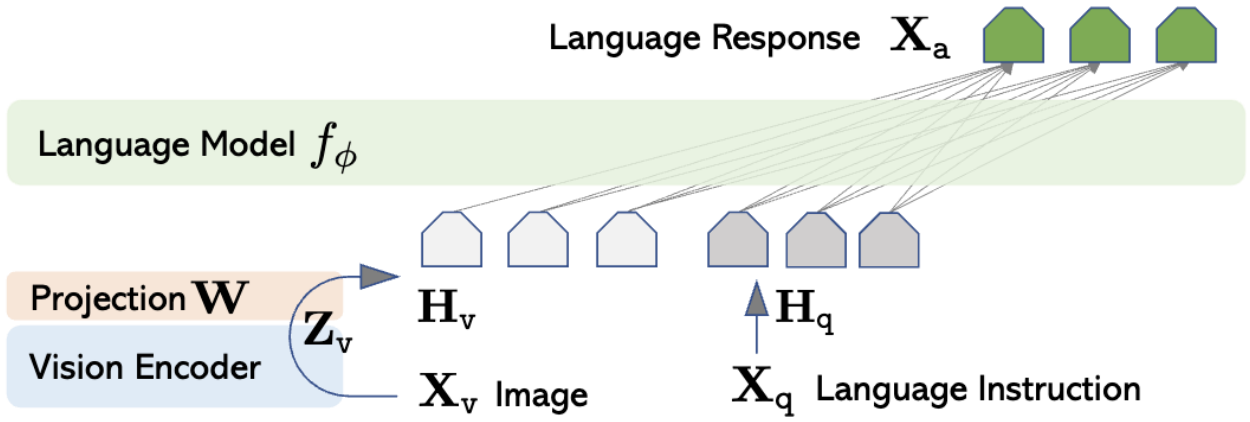


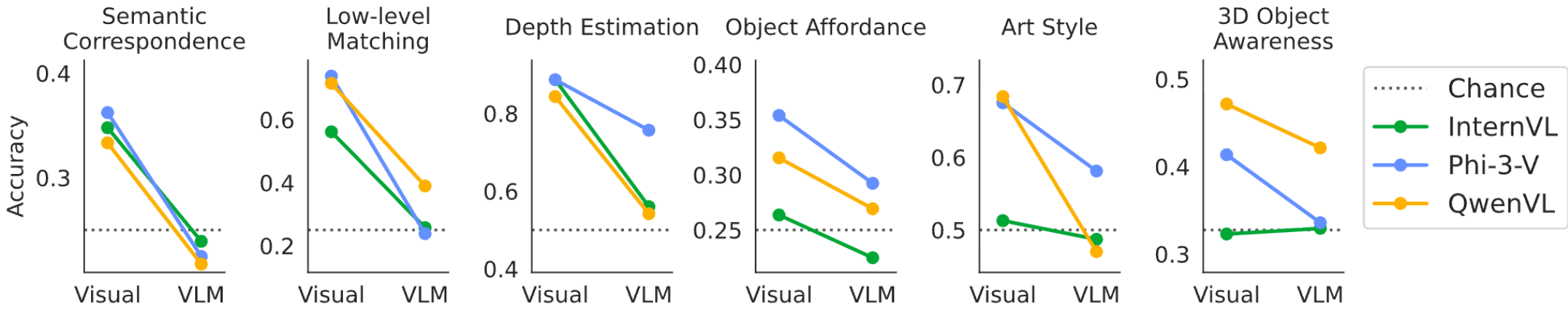
Vision Encoders Reveals Failure to Utilize Visual Information





Vision Encoders Reveals Failure to Utilize Visual Information





Is CLIP Cross-Eyed? Revealing and Mitigating Center Bias in the CLIP Family

Oscar Chew[1], **Hsiao-Ying Huang**[2], **Kunal Jain**[1], **Tai-I Chen**[2,3], **Khoa D. Doan**[4], **Kuan-Hao Huang**[1]
[1]Texas A&M University, [2]National Taiwan University, [3]ASUS, [4]VinUniversity
oscarchew@tamu.edu

Vision-Language Models are Bad at Spatial Reasoning



A dog on a table

A dog right of a table



A dog left of a table

A dog right of a table



A dog right of a table

A dog right of a table

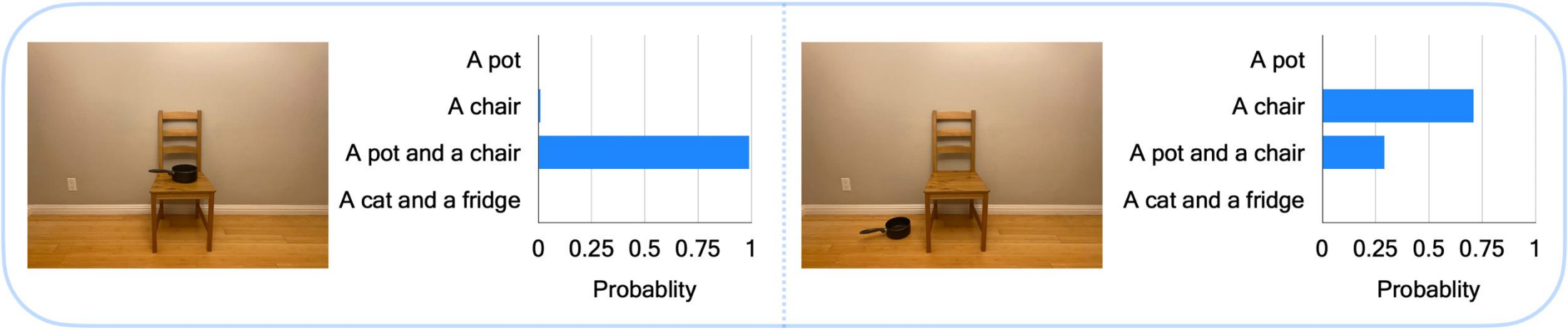


A dog under a table

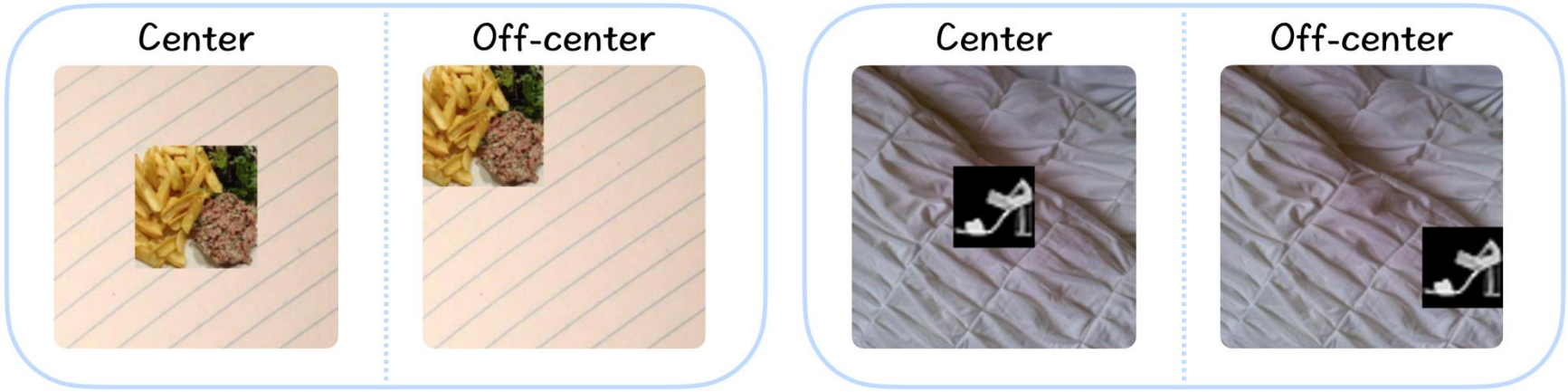
A dog right of a table

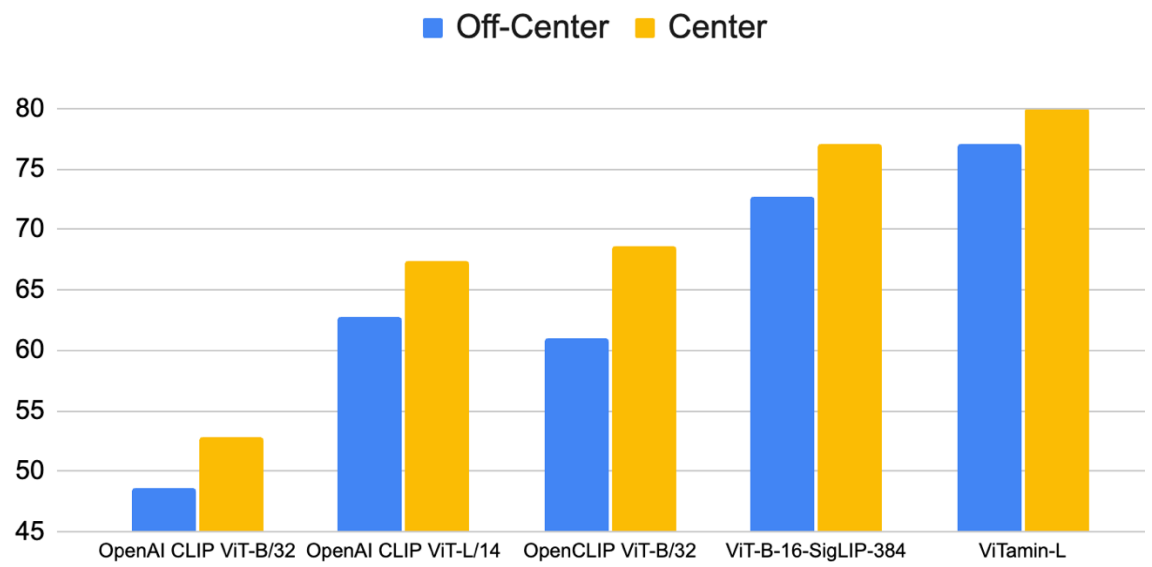
What's "up" with vision-language models? Investigating their struggle with spatial reasoning, 2023

Concept Vanishing



VLMs Have Center Bias: Grid Experiments





VLMs Have Center Bias: Real Image Experiments

Center



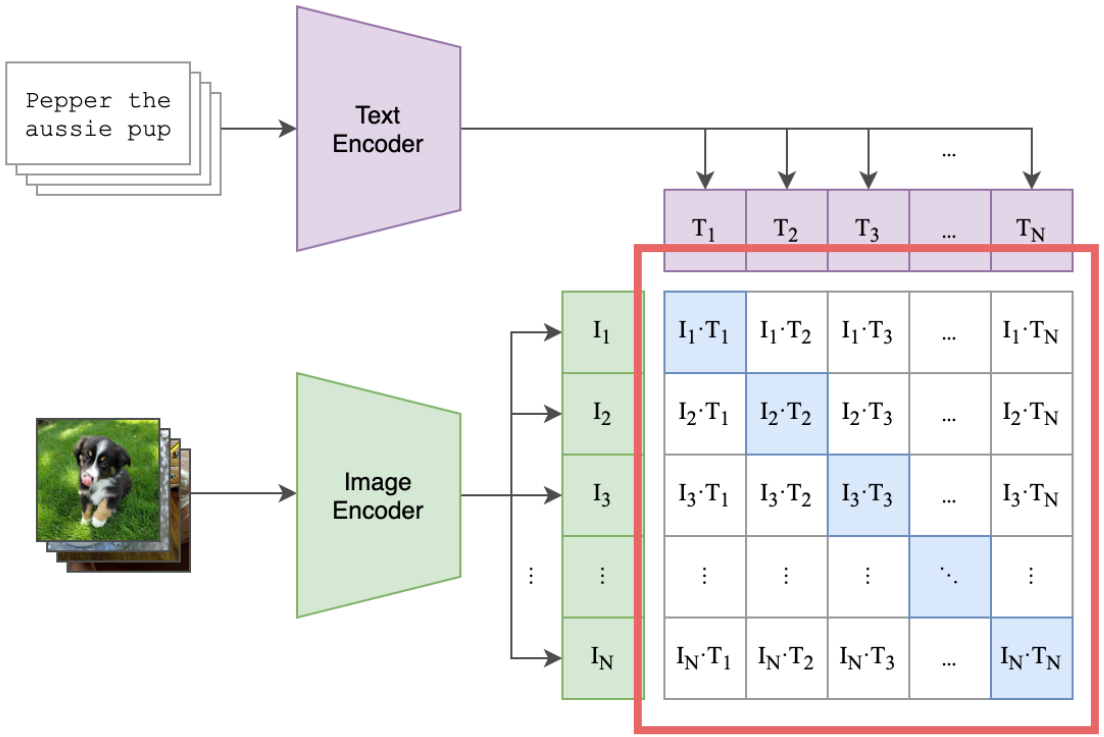
Off-Center



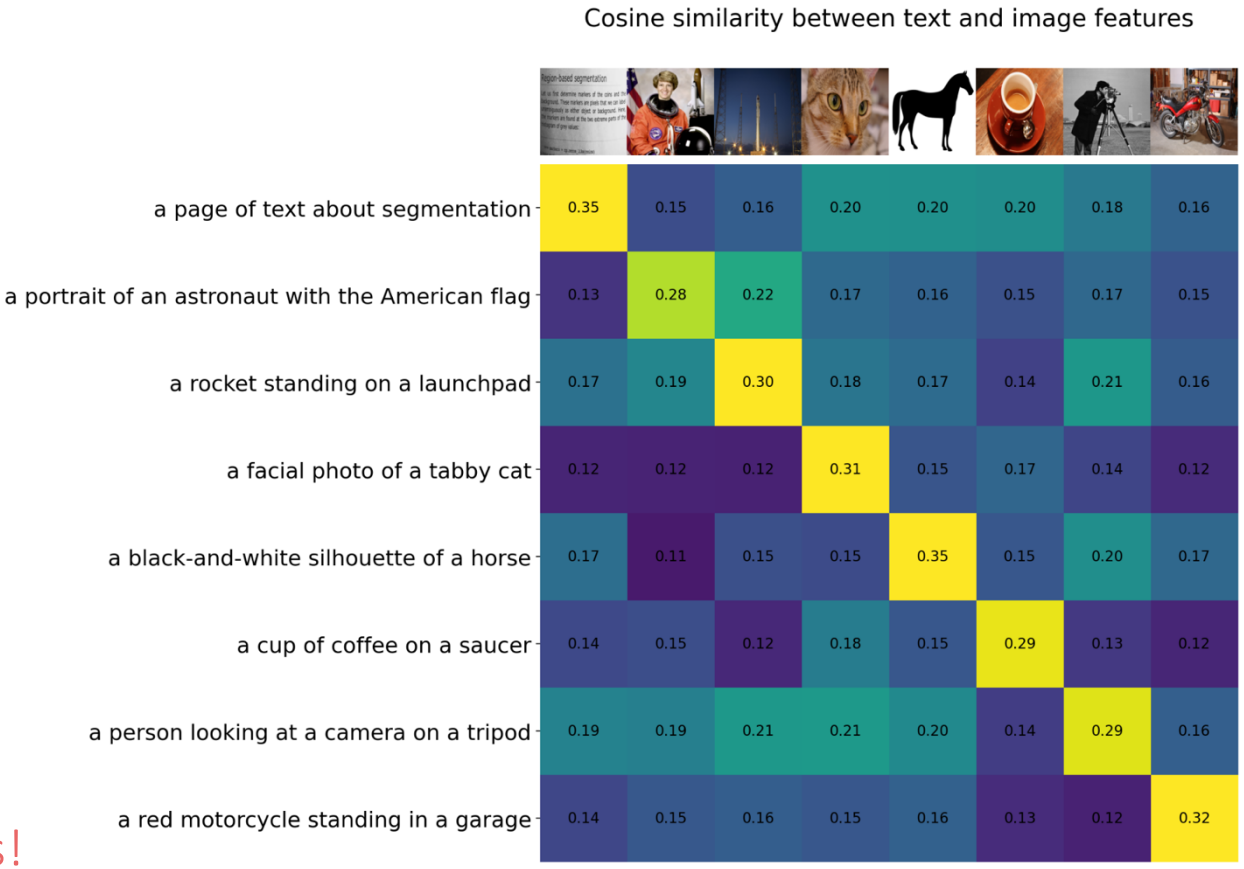
Model	center (↑)	off-center (↑)	center bias (↓)
OpenAI CLIP ViT-B/32	62.9	31.9	31.0
OpenAI CLIP ViT-L/14	94.2	**76.7**	**17.5**
OpenCLIP ViT-B/32	65.7	10.5	55.2
OpenCLIP ViT-L/14	87.6	46.0	41.6
RoBERTaCLIP ViT-B/32	64.8	15.7	49.1
CoCa ViT-B/32	39.0	6.70	32.3
EVA02 ViT-B/16	56.2	13.7	42.5
EVA02 ViT-L/14	88.6	35.1	53.5
EVA02 ViT-L/14 (336)	89.5	33.5	56.0
SigLIP (512)	**97.1**	69.6	27.5
ViTamin-B	66.7	34.5	32.2
ViTamin-L	92.3	57.8	34.5
ViTamin-L (336)	**97.1**	67.1	30.0
NegCLIP ViT-B-32	50.4	22.6	27.8
DetailCLIP	28.6	6.40	22.2
SuperCLIP ViT-B/16	50.5	22.4	28.1
SuperCLIP ViT-L/16	60.9	11.8	49.1

CLIP-Based Vision-Language Models

Contrastive Language-Image Pre-Training

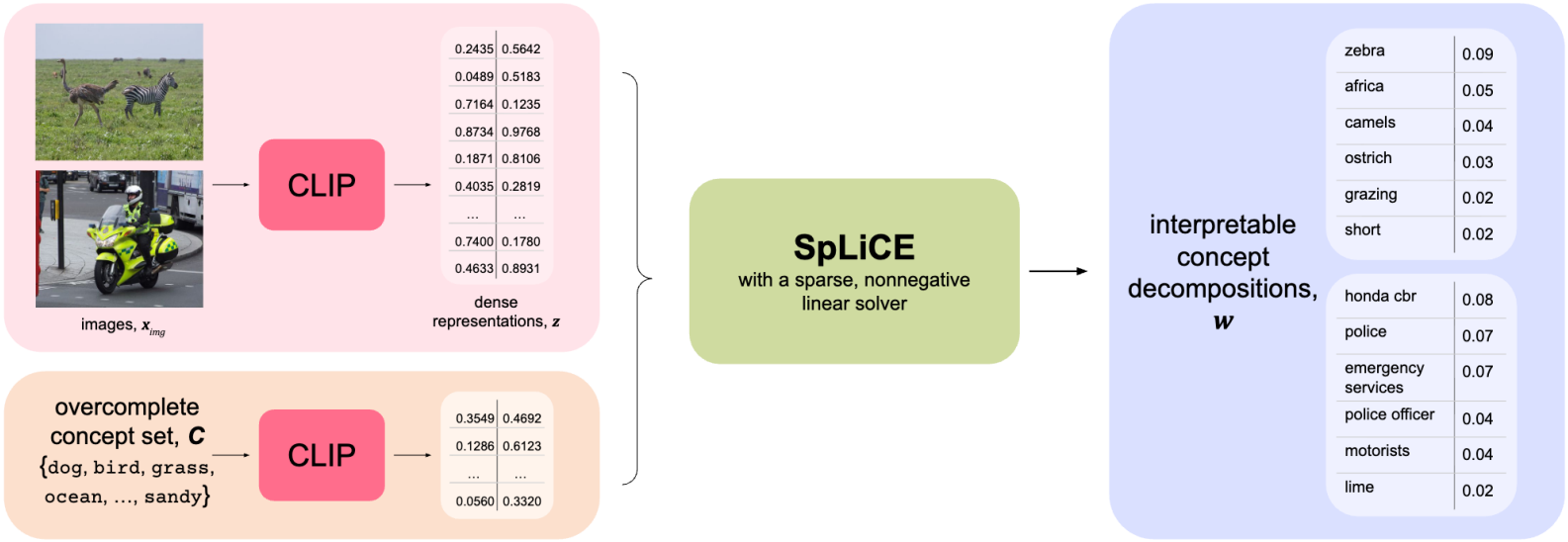


Alignment has problems!



Learning Transferable Visual Models From Natural Language Supervision, 2021

SpLiCE: Sparse Linear Concept Embedding



Text Concepts

Visual Representations

$$\min_{\mathbf{w}\in\mathbb{R}_+^c} \|\mathbf{C}\mathbf{w} - \mathbf{z}\|_2^2 + 2\lambda\|\mathbf{w}\|_1$$

Learnable Weight

Interpreting CLIP with Sparse Linear Concept Embeddings (SpLiCE), 2024

Concept Vanishing in Image Embeddings



Top weighted concepts	
table	0.1209
desk	0.0349
dog	0.0335
purchased	0.0289
finished	0.0237
tabletop	0.0227
puppy	0.0182
chewy	0.0144



Top weighted concepts	
table	0.1450
finished	0.0551
purchased	0.0276
woodworking	0.0240
dining	0.0207
desk	0.0155
before	0.0127
furniture	0.0072



armchair	0.1063
upholstery	0.0683
teal	0.0297
novels	0.0261
listed	0.0255
chair	0.0197
furnished	0.0179
bookshelf	0.0169



upholstery	0.1010
armchair	0.0703
chair	0.0638
furnished	0.0323
blue	0.0225
teal	0.0216
purchased	0.0131
received	0.0098



chair	0.1261
oils	0.0392
cider	0.0206
essentials	0.0196
furniture	0.0196
upcycled	0.0142
oak	0.0140
raised	0.0117



chair	0.1602
woodworking	0.0236
finished	0.0192
seated	0.0180
upcycled	0.0176
toddler	0.0172
bookshelf	0.0155
listed	0.0139

Off-center concepts disappear in image embeddings

When Does Concepts Disappear?

Image tokes: Index 1 for [CLS], other index for patches



1 [CLS]



4



10



20



30



48

Our conclusion: visual and text encoders are fine.

The problem is the moment when [CLS] aggregate information

Post-Hoc Solutions

Grounding + Visual Prompting



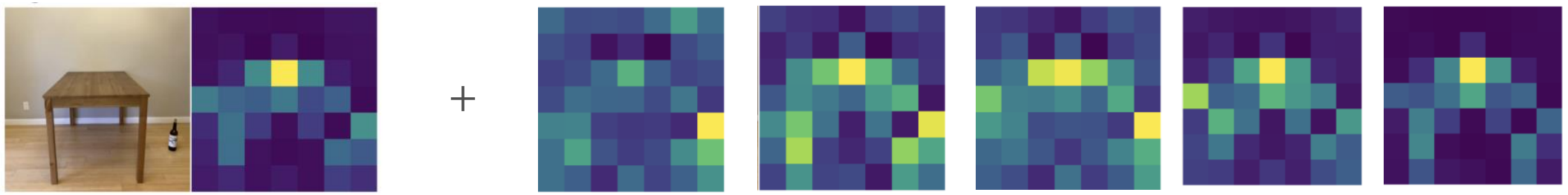
Top weighted concepts	
table	0.1450
finished	0.0551
purchased	0.0276
woodworking	0.0240
dining	0.0207
desk	0.0155
before	0.0127
furniture	0.0072



Top weighted concepts	
table	0.1562
before	0.0578
finished	0.0401
pets	0.0261
woodworking	0.0211
ikea	0.0160
toddler	0.0139
dog	0.0053

Attention Redistribution

α $\qquad\qquad$ $1-\alpha$



[CLS]

Post-Hoc Solutions Reduce Center Bias

Model	center (↑)	off-center (↑)	center bias (↓)	improv. off-center (↑)
Visual Prompting				
OpenAI CLIP ViT-B/32	54.2	21.8	32.4	-10.1
OpenAI CLIP ViT-L/14	87.6	76.7	**10.9**	0.00
OpenCLIP ViT-B/32	59.0	23.3	35.7	+12.8
OpenCLIP ViT-L/14	81.0	56.5	24.5	+10.5
RoBERTaCLIP ViT-B/32	52.3	21.7	30.6	+6.00
CoCa ViT-B/32	23.8	5.10	18.7	-1.60
EVA02 ViT-B/16	65.7	33.8	31.9	+20.1
EVA02 ViT-L/14	86.6	69.0	17.6	+33.9
EVA02 ViT-L/14 (336)	91.4	73.5	17.9	+40.0
SigLIP (512)	88.5	70.9	17.6	+1.30
ViTamin-B	56.1	38.3	17.8	+3.80
ViTamin-L	79.0	56.8	22.2	-1.00
ViTamin-L (336)	88.5	70.9	17.6	+3.80
NegCLIP ViT-B-32	37.1	14.1	23.0	-8.50
DetailCLIP	20.0	10.9	9.10	+4.50
SuperCLIP ViT-B/32	52.4	30.0	22.4	+7.60
SuperCLIP ViT-L/16	50.5	20.1	30.4	+8.30
mean performance	67.9	45.2	22.7	+7.90
Attention Redistribution				
OpenAI CLIP ViT-B/32	76.2	49.2	27.0	+17.3
OpenAI CLIP ViT-L/14	**94.3**	**77.3**	17.0	+0.60
OpenCLIP ViT-B/32	68.6	25.2	43.4	+14.7
OpenCLIP ViT-L/14	89.5	48.9	40.6	+2.90
mean performance	82.2	50.2	32.0	+8.90

Questions?