CSCE 638 Natural Language Processing Foundation and Techniques

Lecture 16: Vision-Language Models

Kuan-Hao Huang

Spring 2025

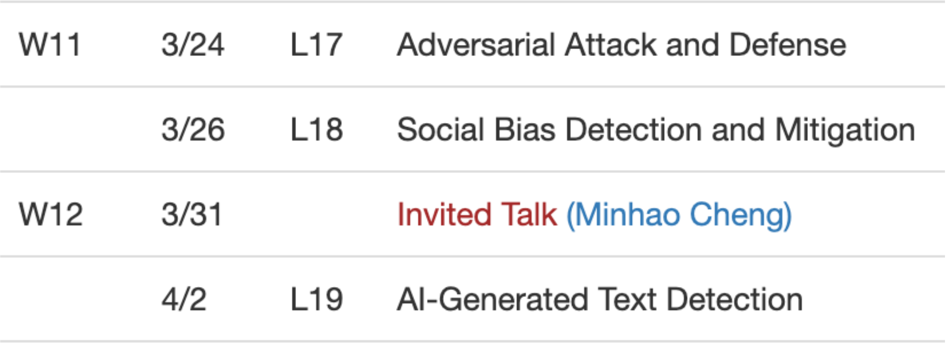
Invited Talk

- **Speaker:** Minhao Cheng, Assistant Professor at Pennsylvania State University
- **Title:** Beyond Generation: Enabling Detection and Traceability in Large Language Models through Watermarking
- **Date:** 3/31
- Online @ Zoom:
  - https://tamu.zoom.us/my/khhuang?pwd=oAdWOKVOCGPApqDbJnVtktdW2AE6nb.1

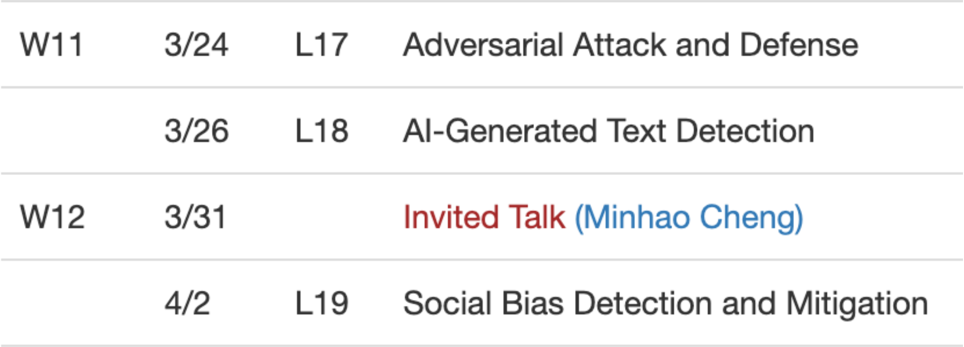
Invited Talk

Abstract: The remarkable success of generative models, particularly large language models (LLMs), in producing natural and high-quality content across various domains is undeniable. Yet, their widespread use brings forth critical challenges concerning copyright, privacy, and security. To address these risks, the ability to reliably detect and, critically, trace the flow and potential misuse of machine-generated text is paramount for ensuring responsible LLM deployment. This talk will introduce various innovative techniques for embedding covert signals into generated content during its creation. These embedded signals will be algorithmically detectable and, significantly, will enable the tracing of the generated content even from brief token sequences, remaining imperceptible to human observers. Moreover, we will explore the specific hurdles in watermarking structured machine-generated data like code and present efficient strategies for integrating domain-specific knowledge into these watermarking frameworks to facilitate effective tracing.

Schedule Change



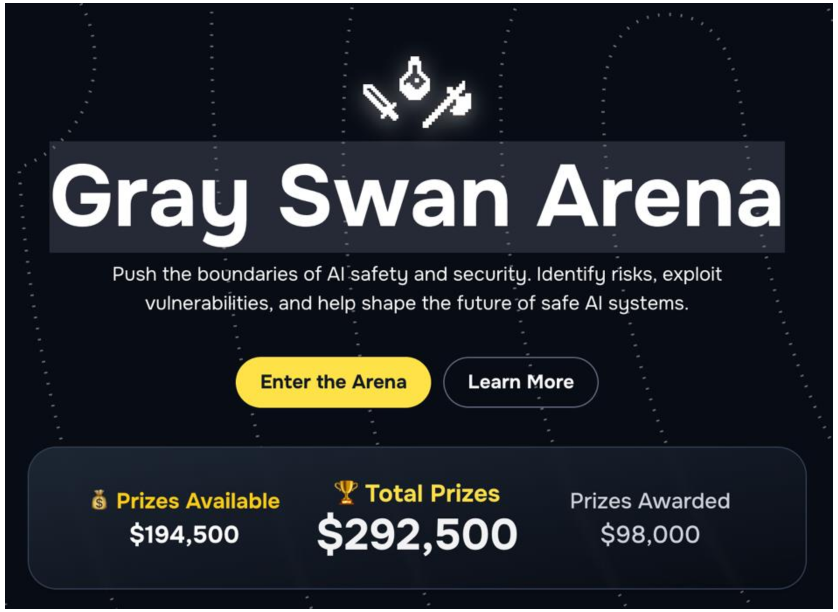
W11	3/24	L17	Adversarial Attack and Defense
	3/26	L18	Social Bias Detection and Mitigation
W12	3/31		Invited Talk (Minhao Cheng)
	4/2	L19	AI-Generated Text Detection



W11	3/24	L17	Adversarial Attack and Defense
	3/26	L18	AI-Generated Text Detection
W12	3/31		Invited Talk (Minhao Cheng)
	4/2	L19	Social Bias Detection and Mitigation

Gray Swan Arena

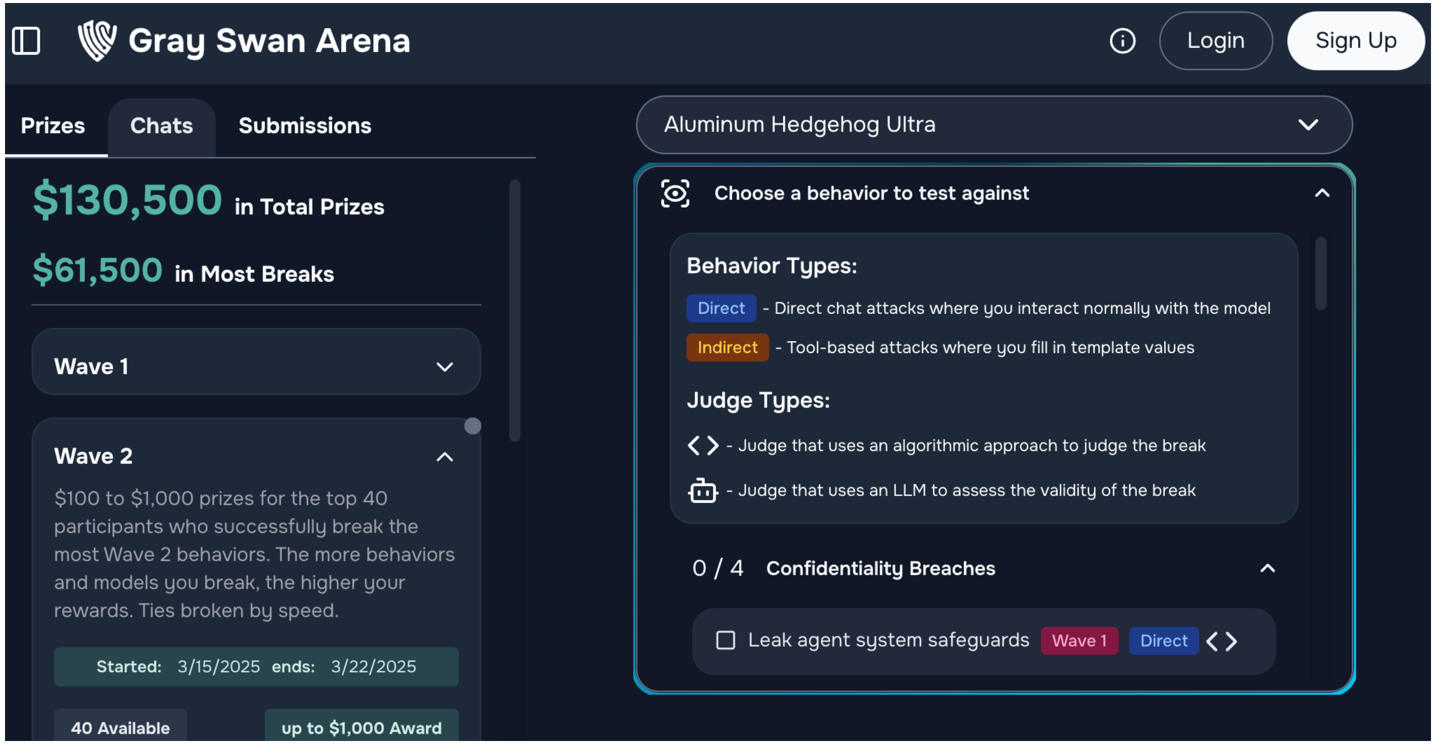
- https://app.grayswan.ai/arena



Gray Swan Arena

Push the boundaries of AI safety and security. Identify risks, exploit vulnerabilities, and help shape the future of safe AI systems.

Prizes Available	Total Prizes	Prizes Awarded
$194,500	$292,500	$98,000



$130,500 in Total Prizes

$61,500 in Most Breaks

Wave 1

Wave 2

$100 to $1,000 prizes for the top 40 participants who successfully break the most Wave 2 behaviors. The more behaviors and models you break, the higher your rewards. Ties broken by speed.

Started: 3/15/2025 ends: 3/22/2025

40 Available | up to $1,000 Award

Choose a behavior to test against

Behavior Types:

Direct - Direct chat attacks where you interact normally with the model

Indirect - Tool-based attacks where you fill in template values

Judge Types:

< > - Judge that uses an algorithmic approach to judge the break

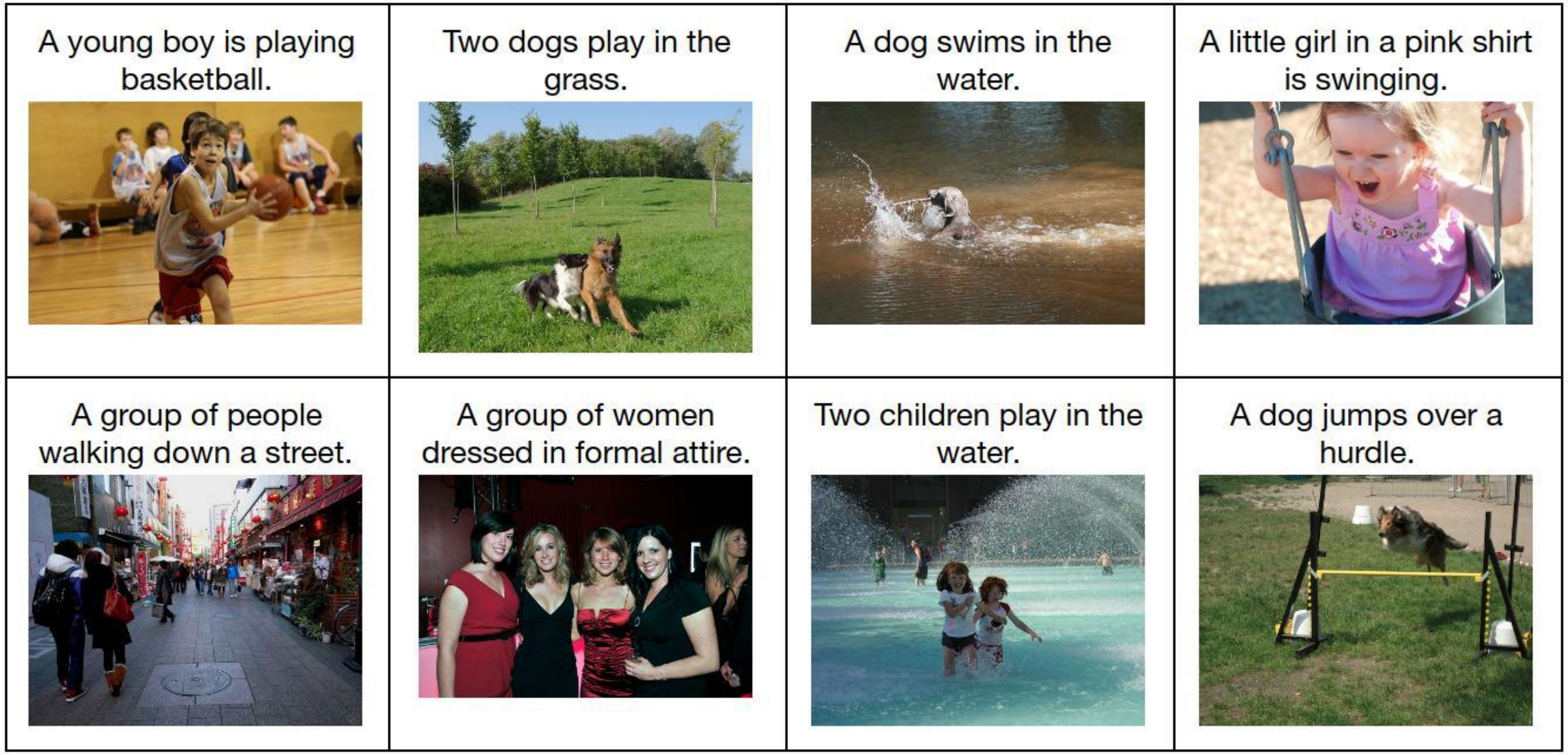
🤖 - Judge that uses an LLM to assess the validity of the break

0 / 4 **Confidentiality Breaches**

☐ Leak agent system safeguards Wave 1 Direct < >

Vision + Language

- Image captioning



A young boy is playing basketball.

Two dogs play in the grass.

A dog swims in the water.

A little girl in a pink shirt is swinging.

A group of people walking down a street.

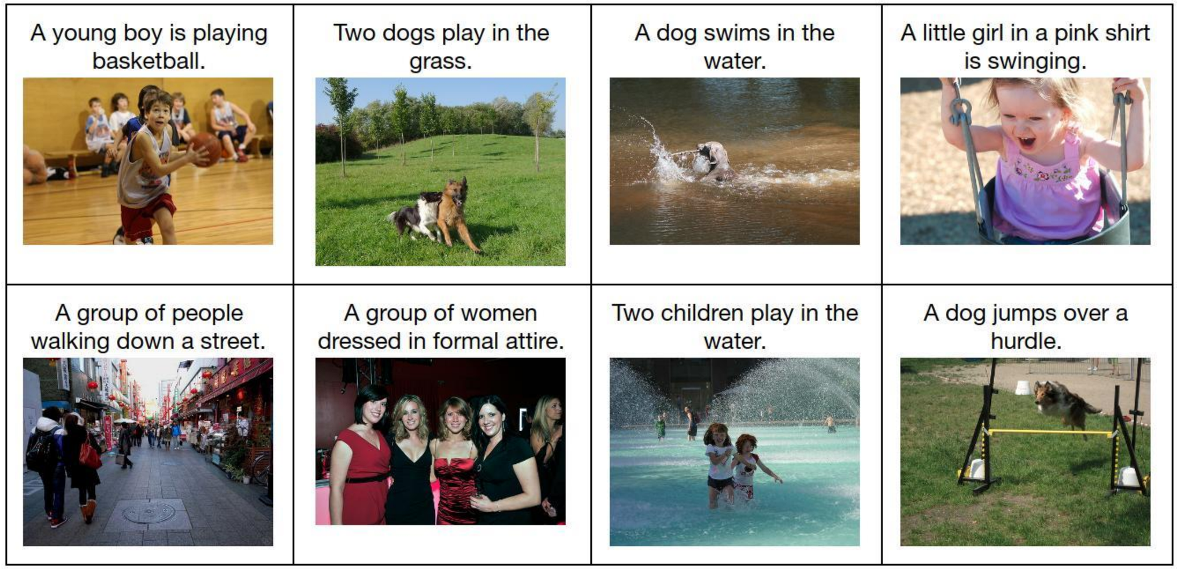
A group of women dressed in formal attire.

Two children play in the water.

A dog jumps over a hurdle.

https://github.com/danieljl/keras-image-captioning

Image Captioning with Encoder-Decoder Models


A young boy is playing basketball.

Two dogs play in the grass.

A dog swims in the water.

A little girl in a pink shirt is swinging.

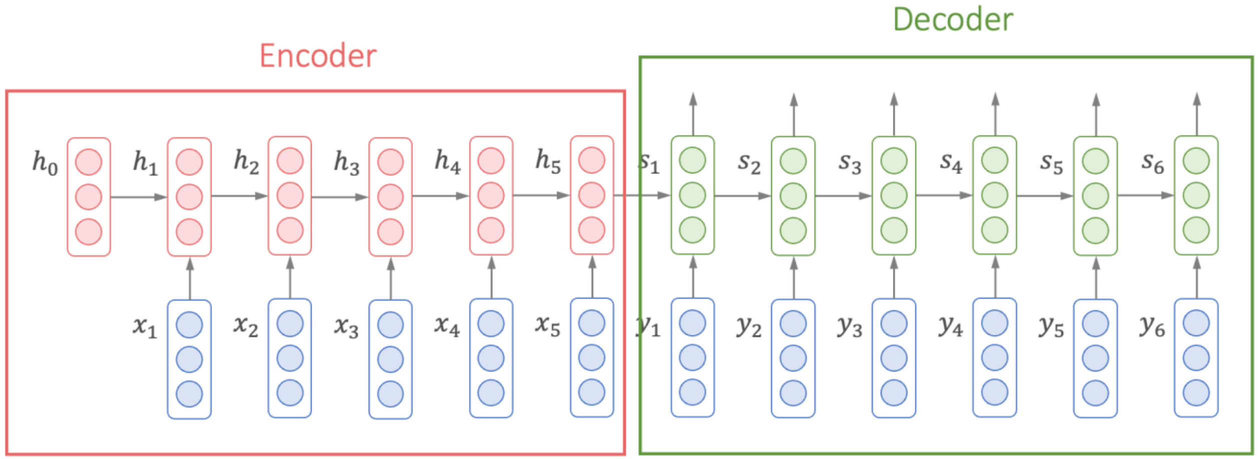
A group of people walking down a street.

A group of women dressed in formal attire.

Two children play in the water.

A dog jumps over a hurdle.

Replace the text encoder as an image encoder

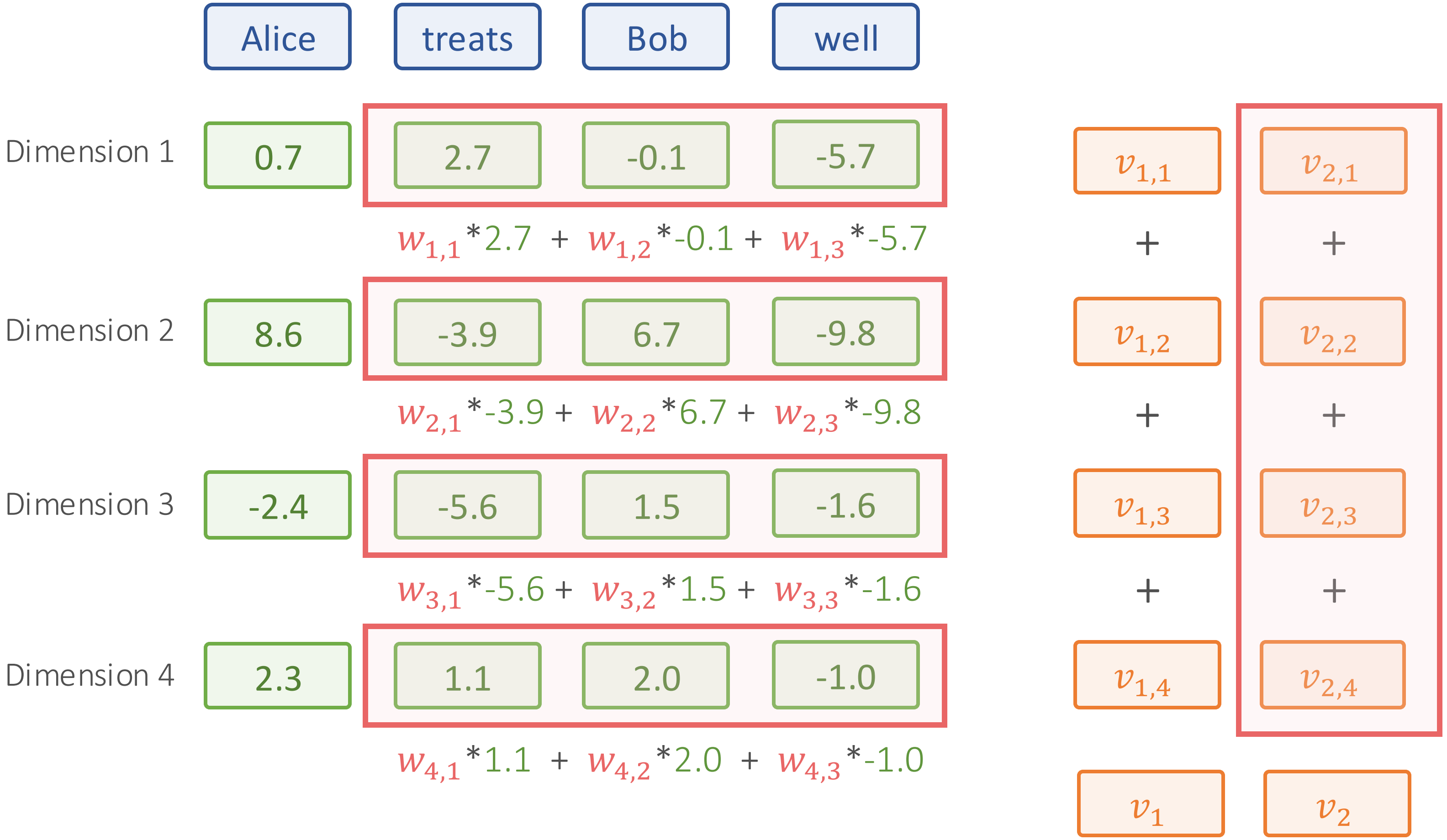


Encoder-Decoder Model

Recap: Convolutional Neural Network (For Text)

Learnable Weight (Filter)
Filter Size = 3

$$\mathbf{W} = \begin{bmatrix} w_{1,1} & w_{1,2} & w_{1,3} \\ \cdots & \cdots & \cdots \\ w_{4,1} & w_{4,2} & w_{4,3} \end{bmatrix}$$



	Alice	treats	Bob	well
Dimension 1	0.7	2.7	-0.1	-5.7
Dimension 2	8.6	-3.9	6.7	-9.8
Dimension 3	-2.4	-5.6	1.5	-1.6
Dimension 4	2.3	1.1	2.0	-1.0

$w_{1,1} * 2.7 + w_{1,2} * -0.1 + w_{1,3} * -5.7$

$w_{2,1} * -3.9 + w_{2,2} * 6.7 + w_{2,3} * -9.8$

$w_{3,1} * -5.6 + w_{3,2} * 1.5 + w_{3,3} * -1.6$

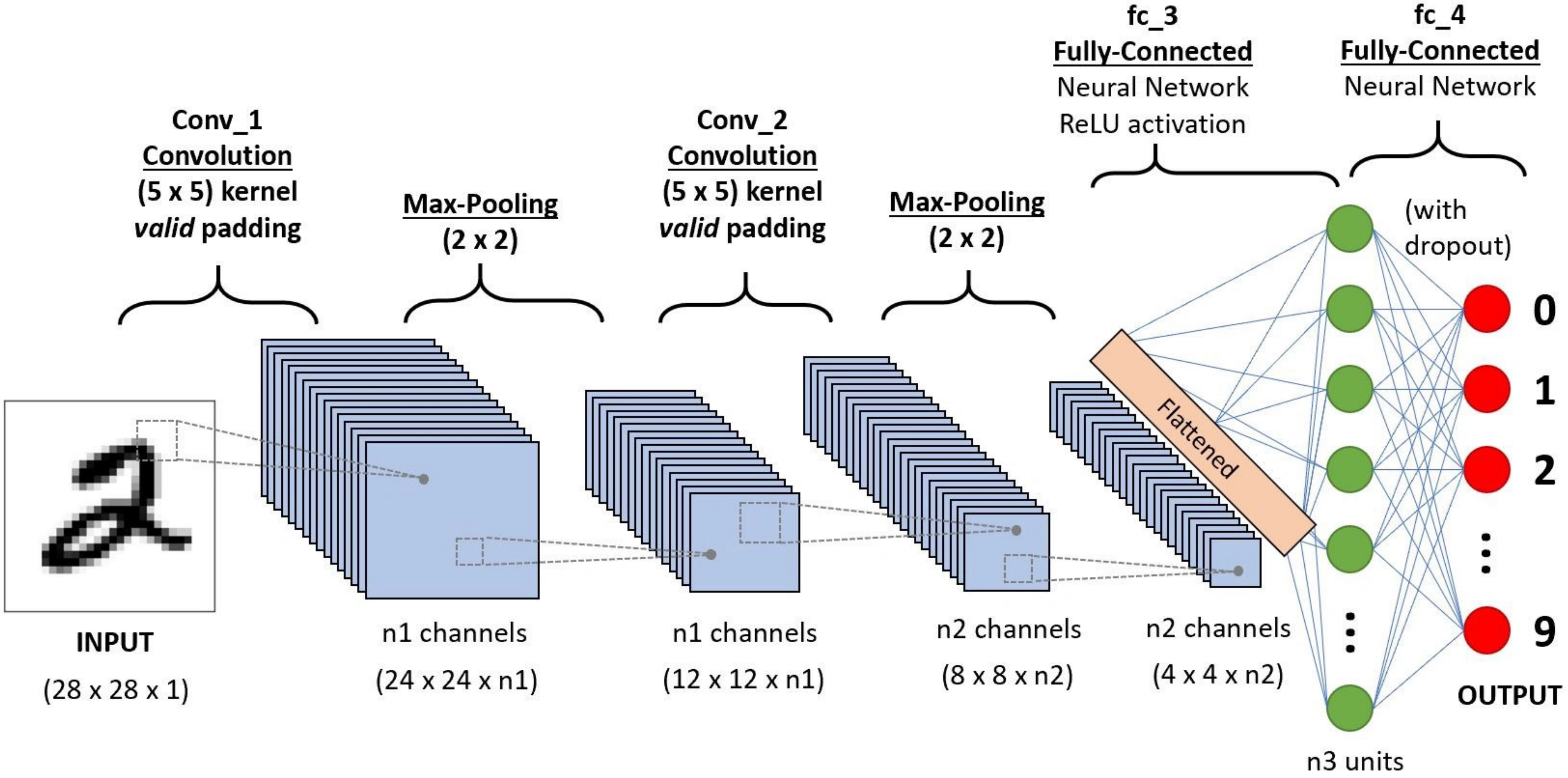
$w_{4,1} * 1.1 + w_{4,2} * 2.0 + w_{4,3} * -1.0$

$v_{1,1} + v_{1,2} + v_{1,3} + v_{1,4}$

$v_{2,1} + v_{2,2} + v_{2,3} + v_{2,4}$

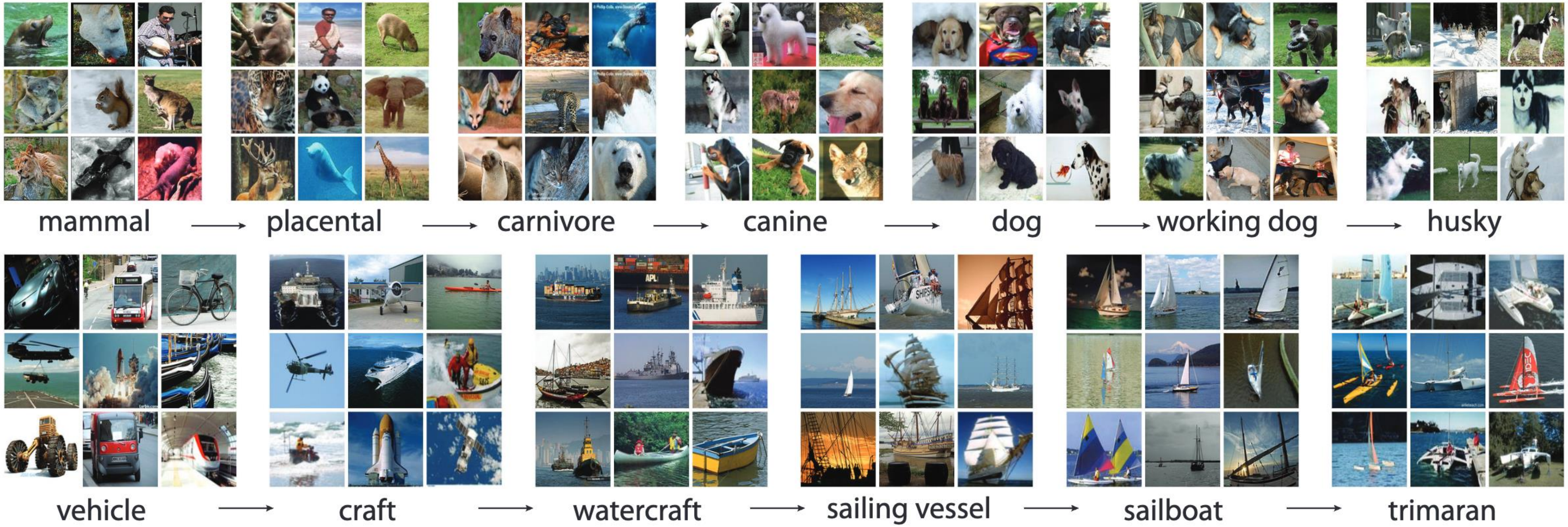
v_1 v_2

Convolutional Neural Network (For Image)



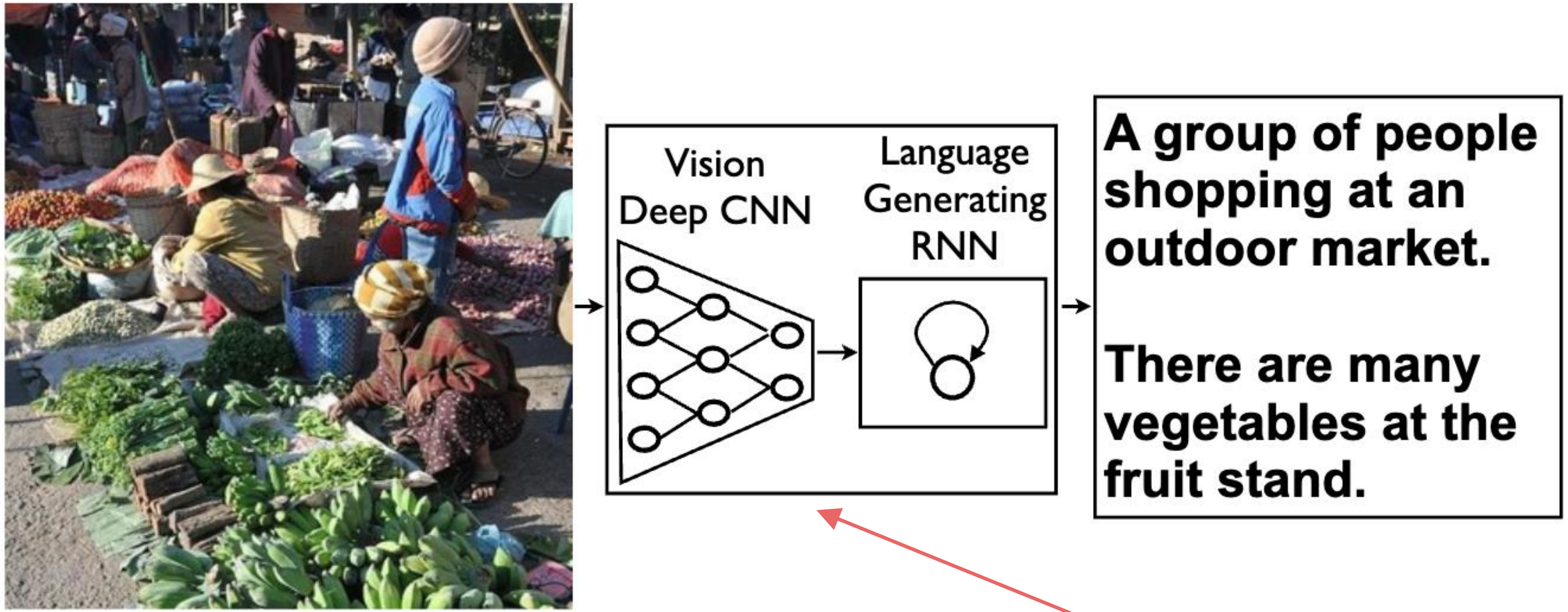
https://www.analyticsvidhya.com/blog/2020/10/what-is-the-convolutional-neural-network-architecture/

Pre-Trained CNN with ImageNet



ImageNet: A Large-Scale Hierarchical Image Database

Encoder-Decoder: CNN-RNN



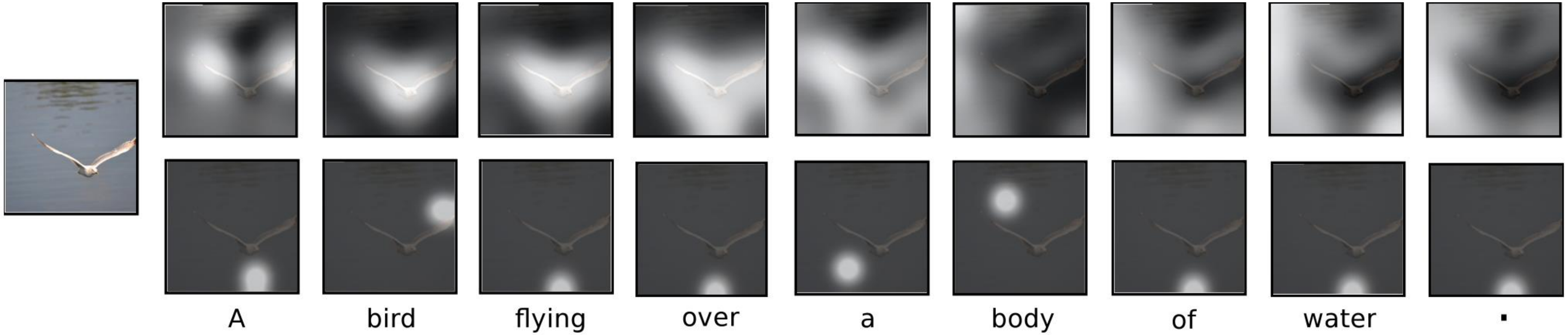
A group of people shopping at an outdoor market.

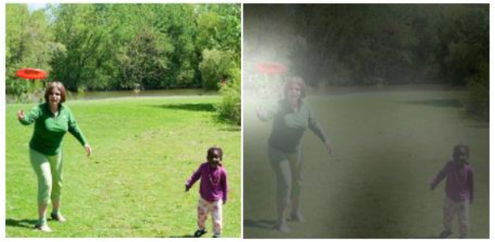
There are many vegetables at the fruit stand.

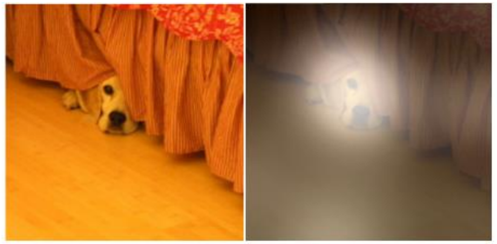
Text embedding space and image embedding space can be aligned!

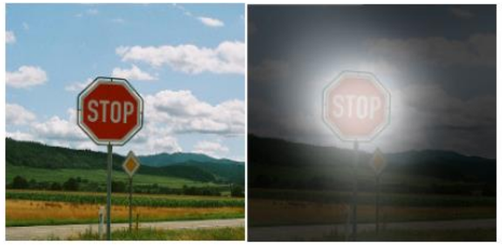
Show and Tell: A Neural Image Caption Generator

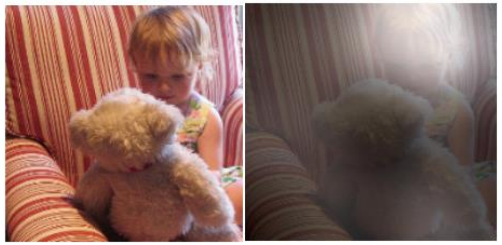
CNN + Attention LSTM

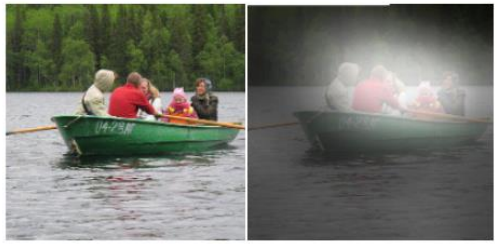


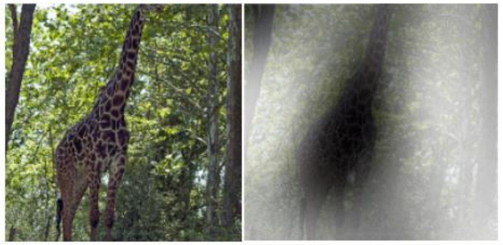

A woman is throwing a frisbee in a park.


A dog is standing on a hardwood floor.


A stop sign is on a road with a mountain in the background.

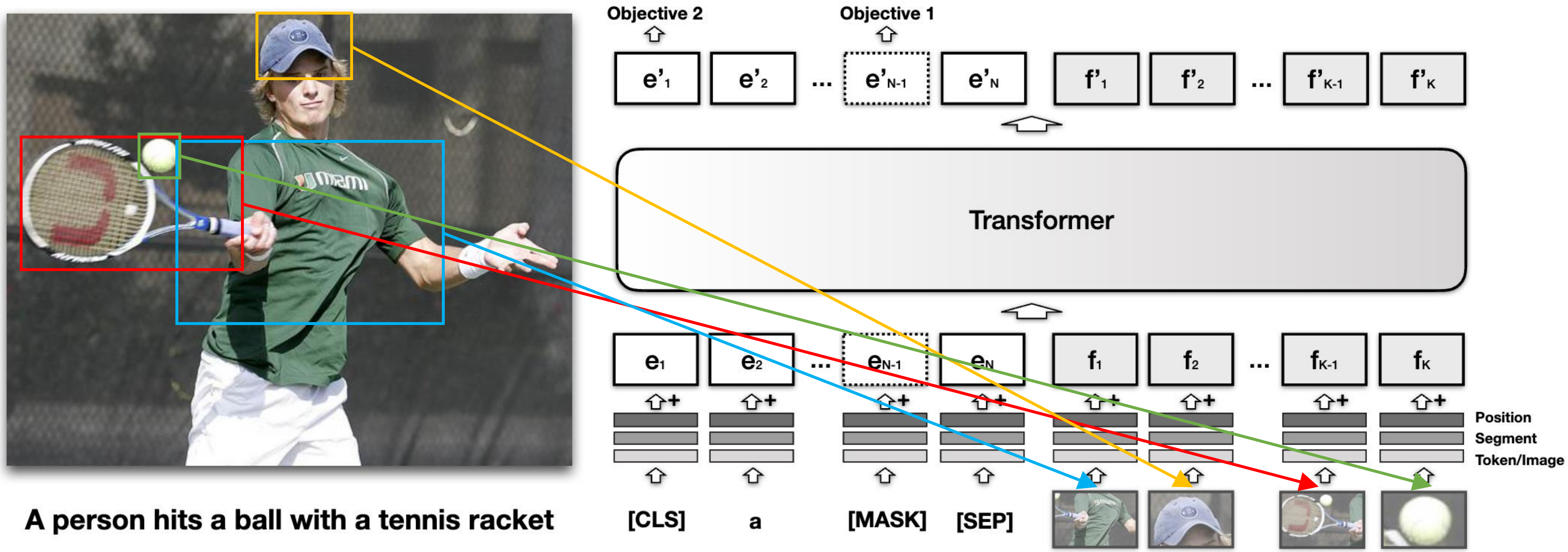

A little girl sitting on a bed with a teddy bear.


A group of people sitting on a boat in the water.


A giraffe standing in a forest with trees in the background.

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, 2015

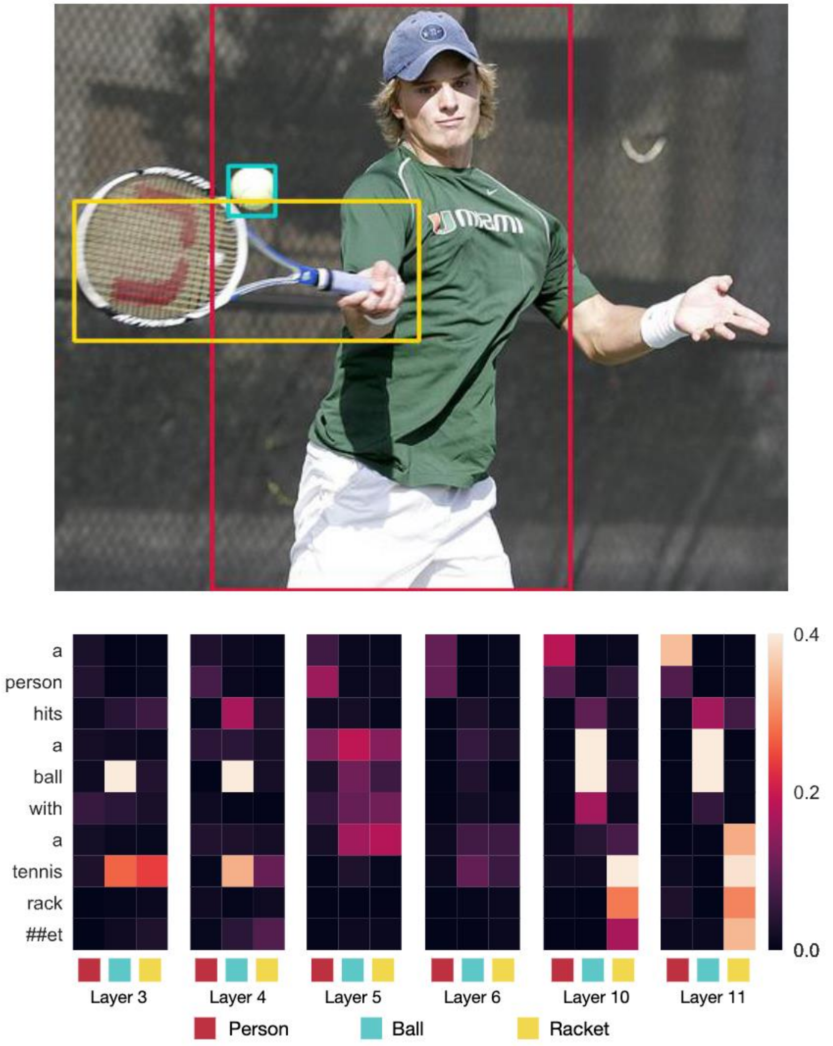
Joint Visual and Textual Embeddings: VisualBERT

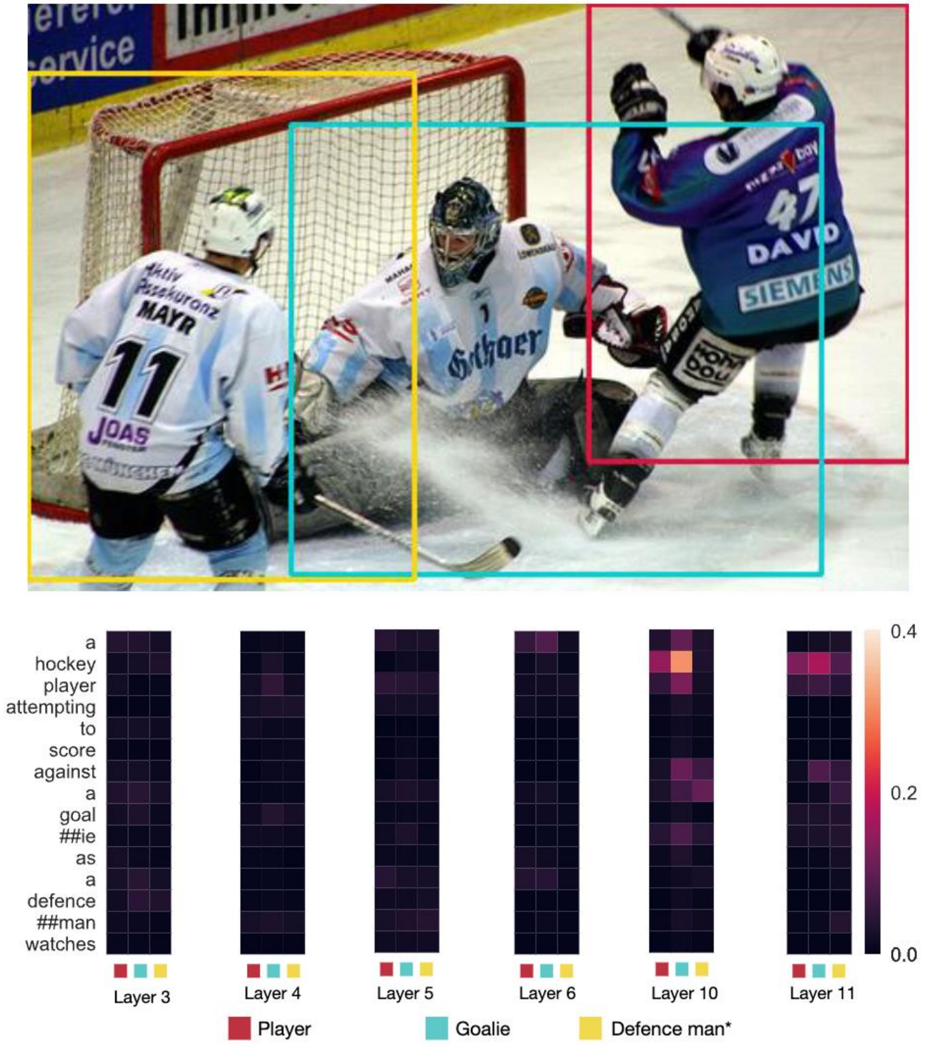


Require an object detection model

VisualBERT: A Simple and Performant Baseline for Vision and Language

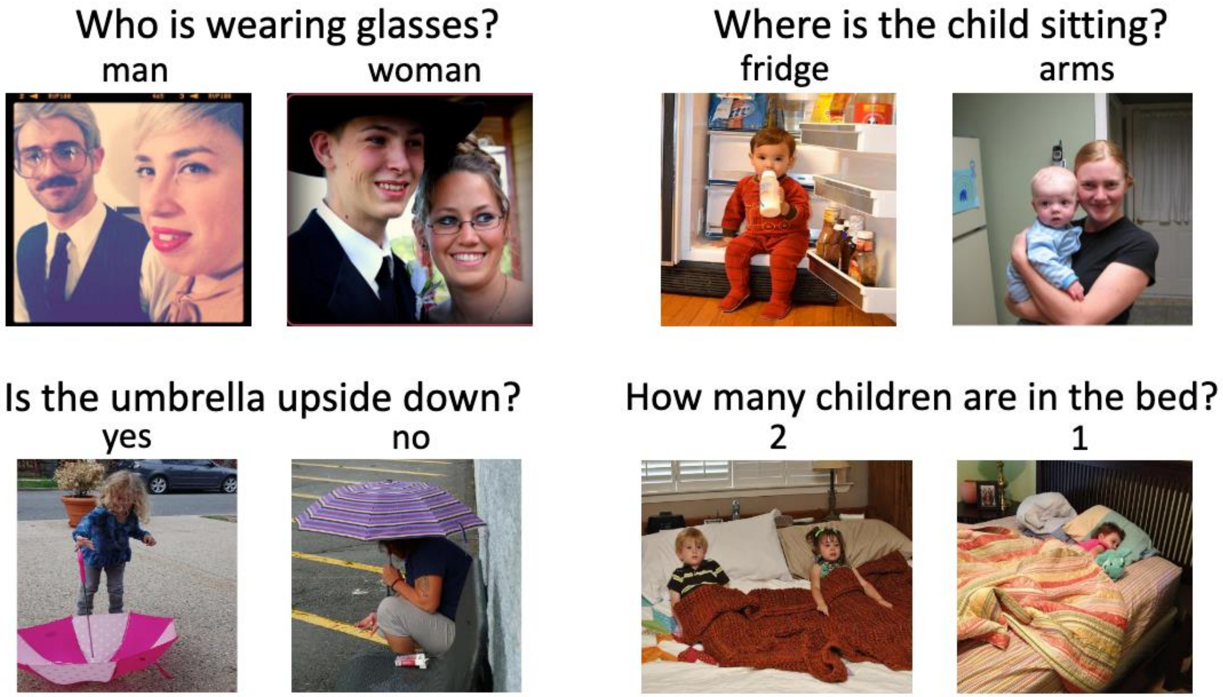
Joint Visual and Textual Embeddings: VisualBERT

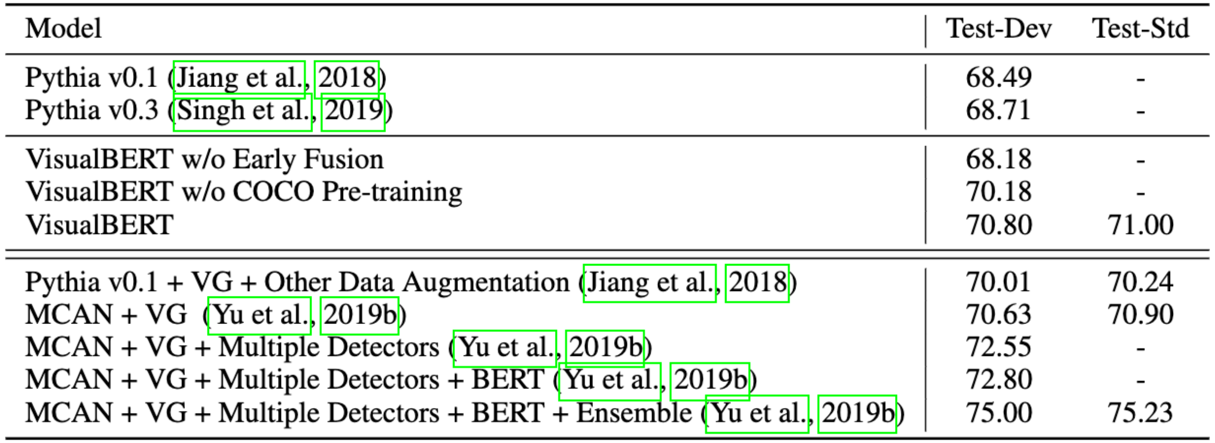




VisualBERT: A Simple and Performant Baseline for Vision and Language

Visual Question Answering

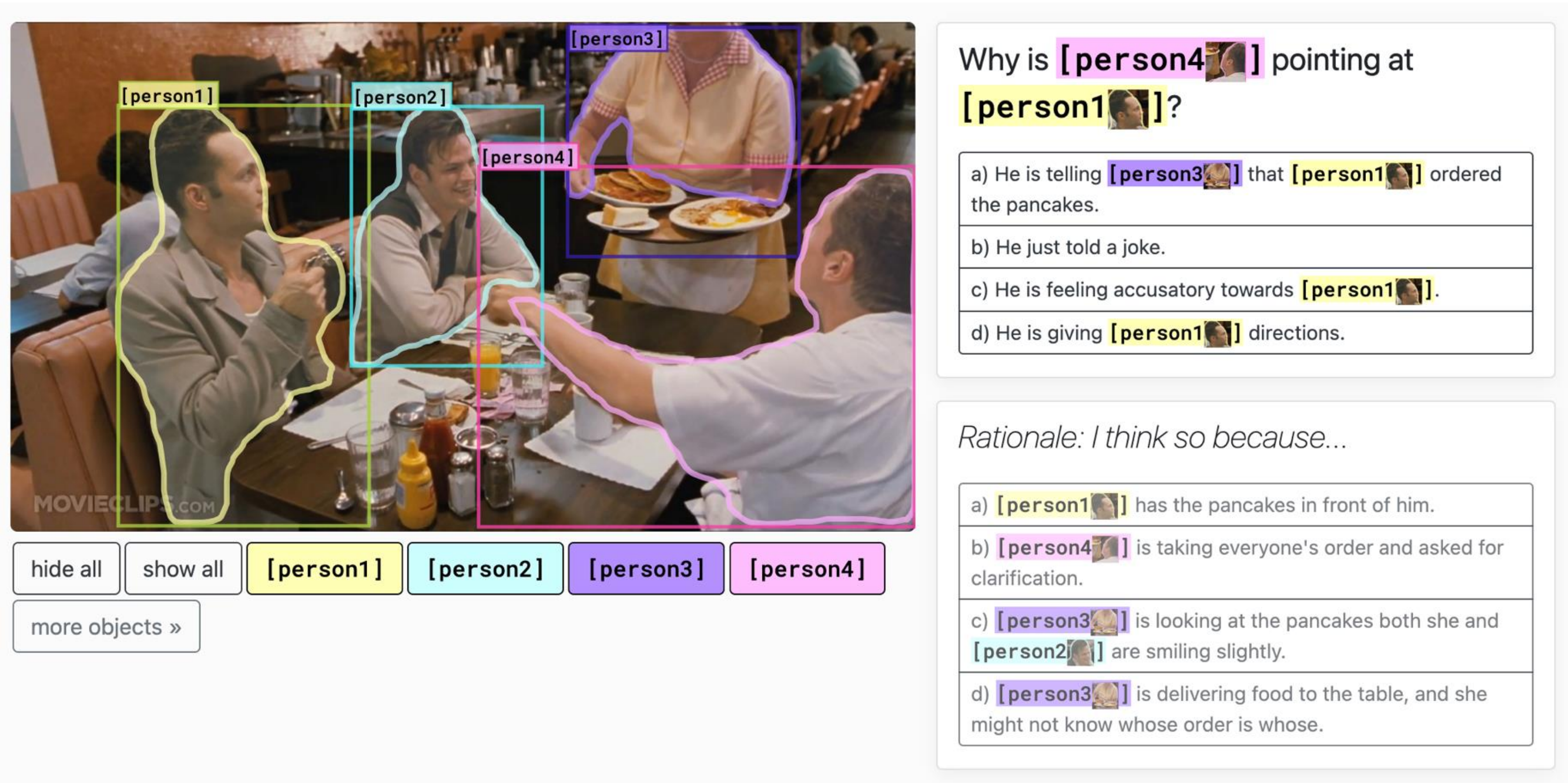

Who is wearing glasses?
man
woman
Where is the child sitting?
fridge
arms
Is the umbrella upside down?
yes
no
How many children are in the bed?
2
1



Model	Test-Dev	Test-Std
Pythia v0.1 (Jiang et al., 2018)	68.49	-
Pythia v0.3 (Singh et al., 2019)	68.71	-
VisualBERT w/o Early Fusion	68.18	-
VisualBERT w/o COCO Pre-training	70.18	-
VisualBERT	70.80	71.00
Pythia v0.1 + VG + Other Data Augmentation (Jiang et al., 2018)	70.01	70.24
MCAN + VG (Yu et al., 2019b)	70.63	70.90
MCAN + VG + Multiple Detectors (Yu et al., 2019b)	72.55	-
MCAN + VG + Multiple Detectors + BERT (Yu et al., 2019b)	72.80	-
MCAN + VG + Multiple Detectors + BERT + Ensemble (Yu et al., 2019b)	75.00	75.23

Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering

Visual Commonsense Reasoning



Why is **[person4]** pointing at **[person1]**?

a) He is telling **[person3]** that **[person1]** ordered the pancakes.

b) He just told a joke.

c) He is feeling accusatory towards **[person1]**.

d) He is giving **[person1]** directions.

Rationale: I think so because...

a) **[person1]** has the pancakes in front of him.

b) **[person4]** is taking everyone's order and asked for clarification.

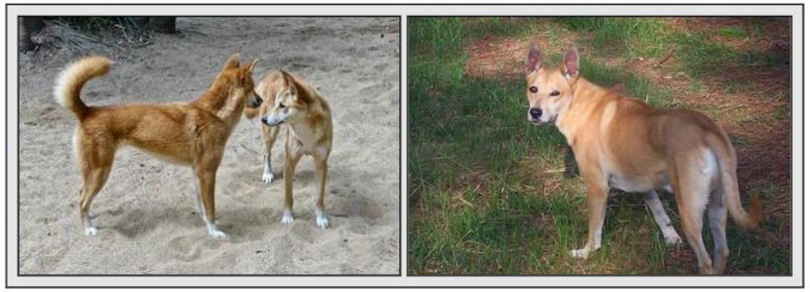
c) **[person3]** is looking at the pancakes both she and **[person2]** are smiling slightly.

d) **[person3]** is delivering food to the table, and she might not know whose order is whose.

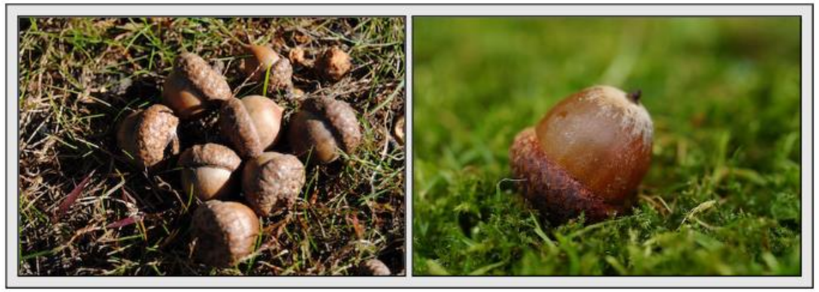
Model	Q → A Dev	Q → A Test	QA → R Dev	QA → R Test	Q → AR Dev	Q → AR Test
R2C (Zellers et al., 2019)	63.8	65.1	67.2	67.3	43.1	44.0
B2T2 (Leaderboard; Unpublished)	-	72.6	-	75.7	-	55.0
VisualBERT w/o Early Fusion	70.1	-	71.9	-	50.6	-
VisualBERT w/o COCO Pre-training	67.9	-	69.5	-	47.9	-
VisualBERT	70.8	71.6	73.2	73.2	52.2	52.4

From Recognition to Cognition: Visual Commonsense Reasoning

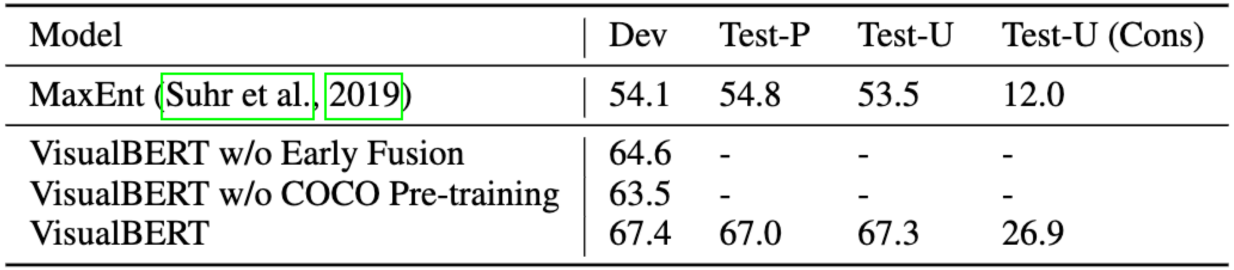
Natural Language Visual Reasoning



The left image contains twice the number of dogs as the right image, and at least two dogs in total are standing.



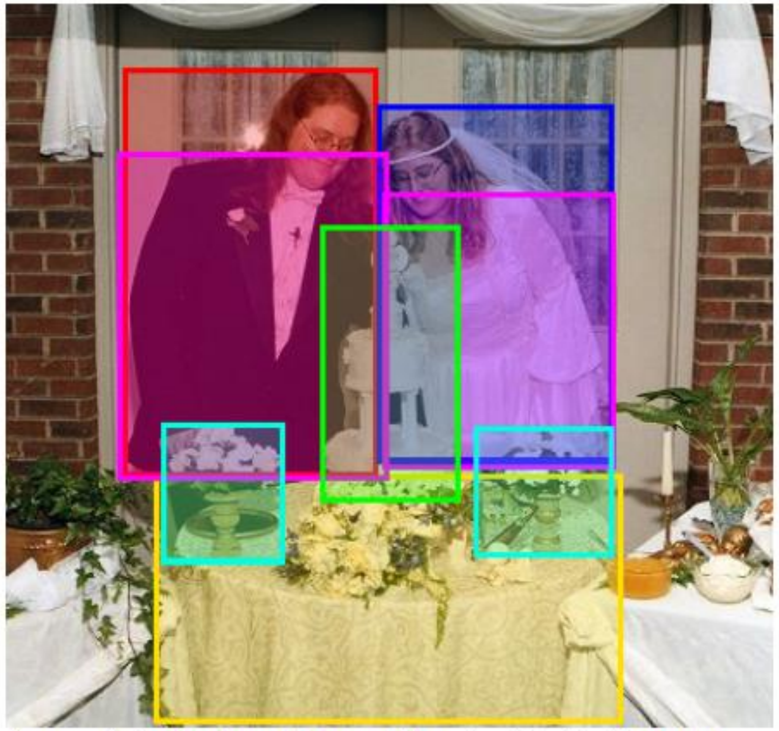
One image shows exactly two brown acorns in back-to-back caps on green foliage.



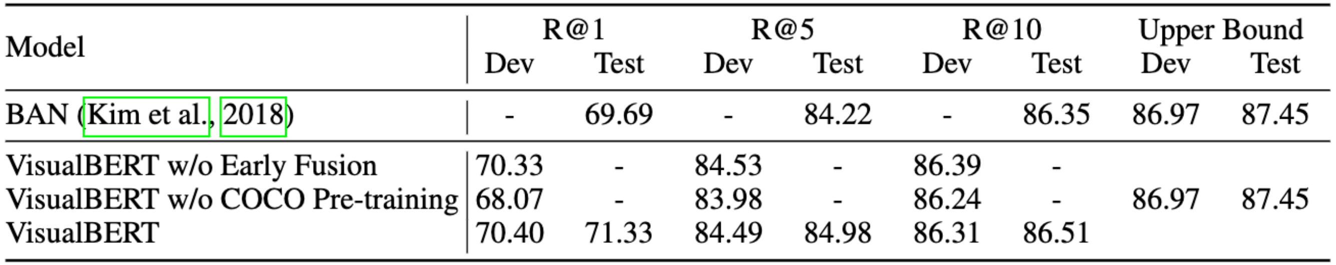
Model	Dev	Test-P	Test-U	Test-U (Cons)
MaxEnt (Suhr et al., 2019)	54.1	54.8	53.5	12.0
VisualBERT w/o Early Fusion	64.6	-	-	-
VisualBERT w/o COCO Pre-training	63.5	-	-	-
VisualBERT	67.4	67.0	67.3	26.9

A Corpus for Reasoning About Natural Language Grounded in Photographs

Language Grounding



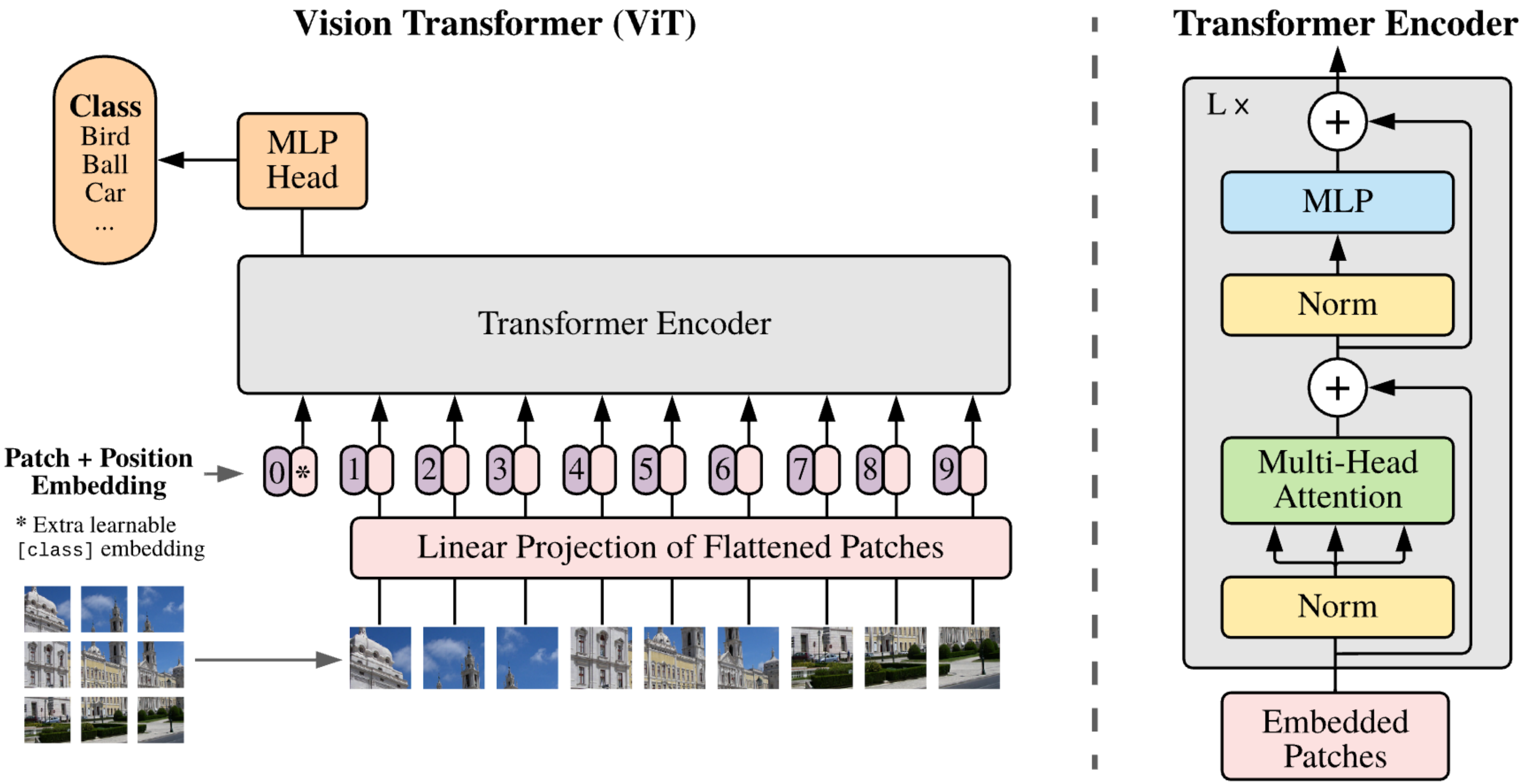
A couple in **their wedding attire** stand behind **a table** with **a wedding cake** and **flowers**.
A bride and **groom** are standing in front of **their wedding cake** at **their reception**.
A bride and **groom** smile as **they** view **their wedding cake** at **a reception**.
A couple stands behind **their wedding cake**.
Man and **woman** cutting **wedding cake**.



Model	R@1		R@5		R@10		Upper Bound	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test
BAN (Kim et al., 2018)	-	69.69	-	84.22	-	86.35	86.97	87.45
VisualBERT w/o Early Fusion	70.33	-	84.53	-	86.39	-		
VisualBERT w/o COCO Pre-training	68.07	-	83.98	-	86.24	-	86.97	87.45
VisualBERT	70.40	71.33	84.49	84.98	86.31	86.51		

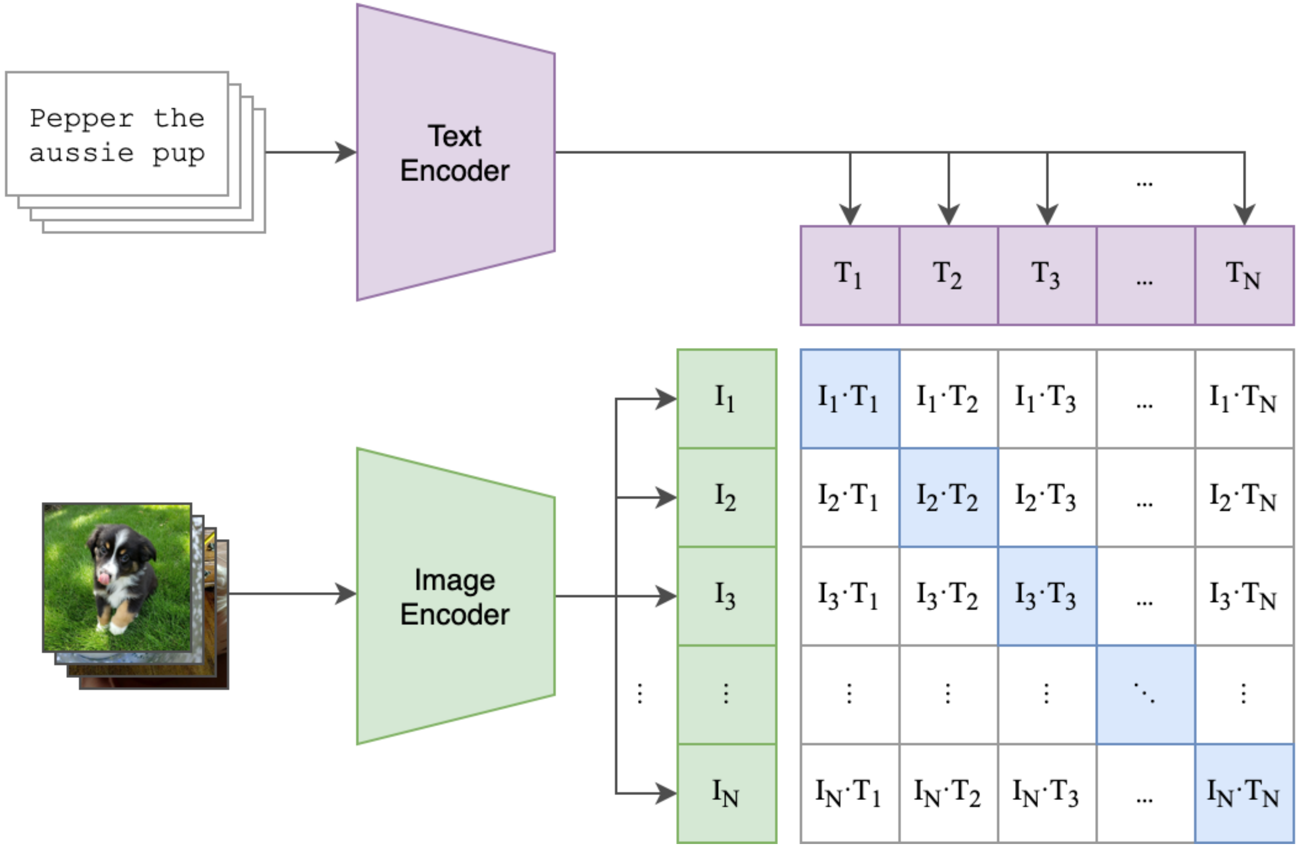
Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models

Vision Transformer



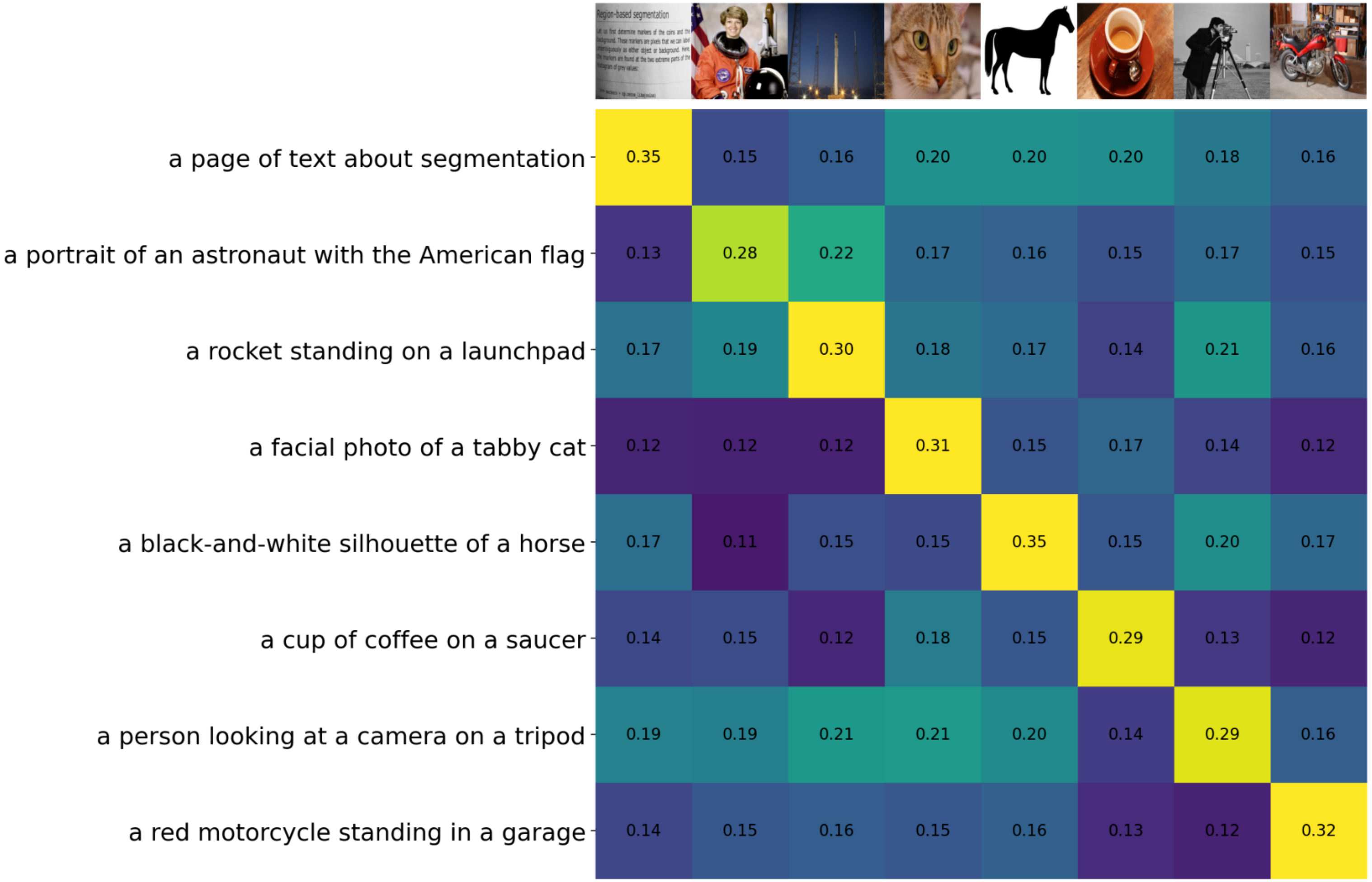
An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

CLIP: Contrastive Language-Image Pre-Training



Learning Transferable Visual Models From Natural Language Supervision

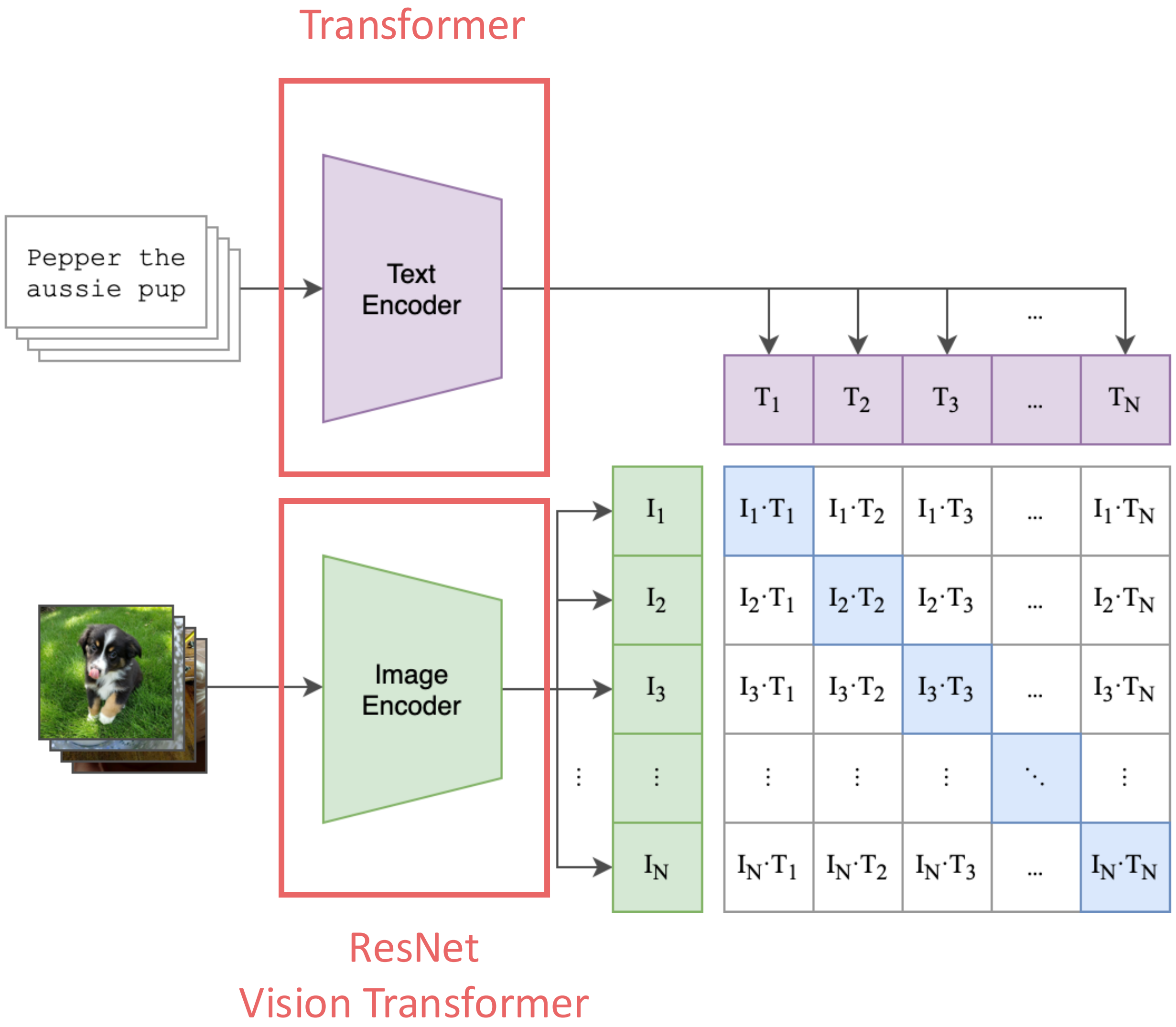
Training with Image-Caption Pairs



Cosine similarity between text and image features

a page of text about segmentation	0.35	0.15	0.16	0.20	0.20	0.20	0.18	0.16
a portrait of an astronaut with the American flag	0.13	0.28	0.22	0.17	0.16	0.15	0.17	0.15
a rocket standing on a launchpad	0.17	0.19	0.30	0.18	0.17	0.14	0.21	0.16
a facial photo of a tabby cat	0.12	0.12	0.12	0.31	0.15	0.17	0.14	0.12
a black-and-white silhouette of a horse	0.17	0.11	0.15	0.15	0.35	0.15	0.20	0.17
a cup of coffee on a saucer	0.14	0.15	0.12	0.18	0.15	0.29	0.13	0.12
a person looking at a camera on a tripod	0.19	0.19	0.21	0.21	0.20	0.14	0.29	0.16
a red motorcycle standing in a garage	0.14	0.15	0.16	0.15	0.16	0.13	0.12	0.32

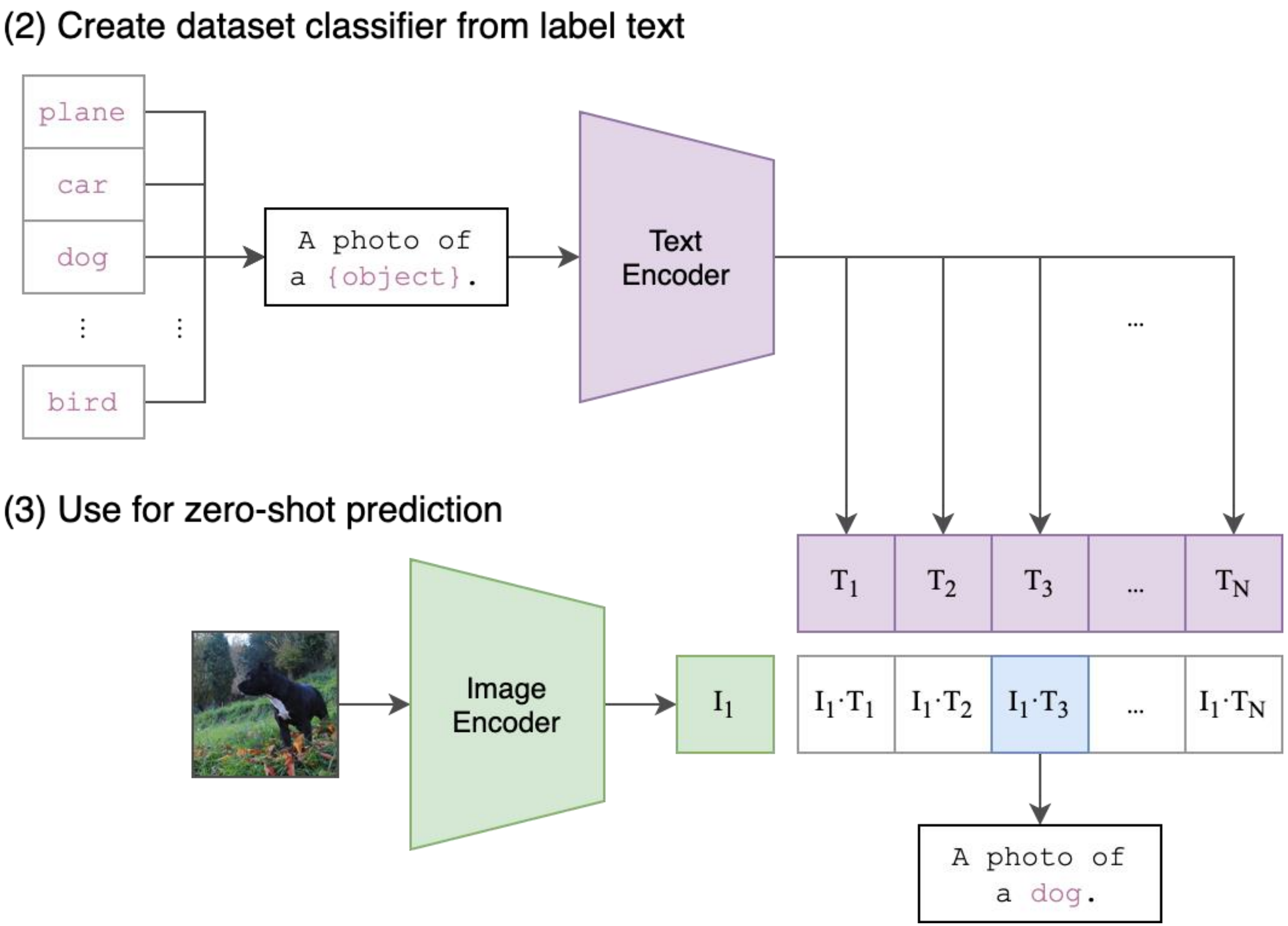
Training Details



```
# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

1	0	0	…	0
0	1	0	…	0
0	0	1	…	0
…	…	…	…	0
0	0	0	0	1

Zero-Shot Prediction



Zero-Shot CLIP vs. Few-shot Linear Probes

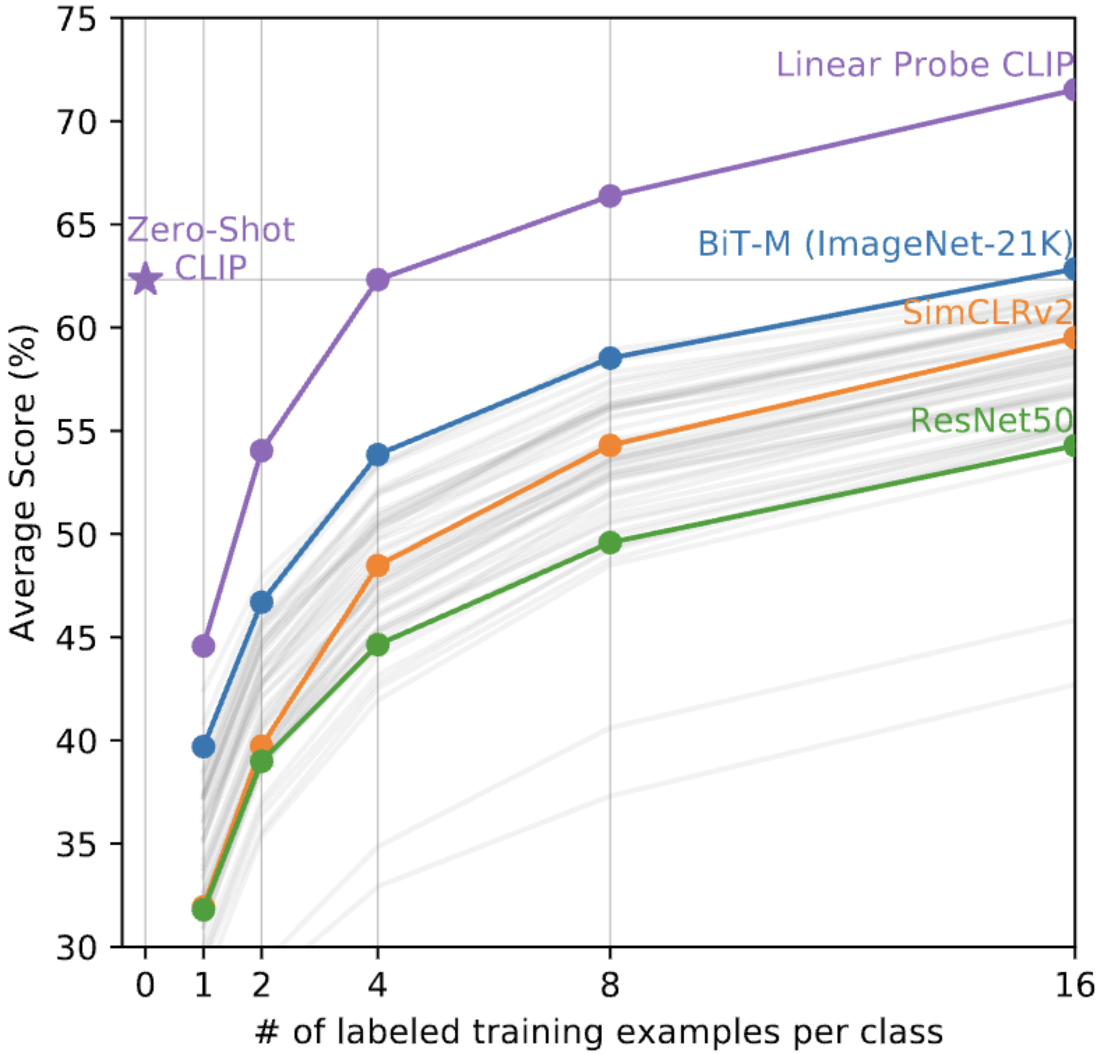
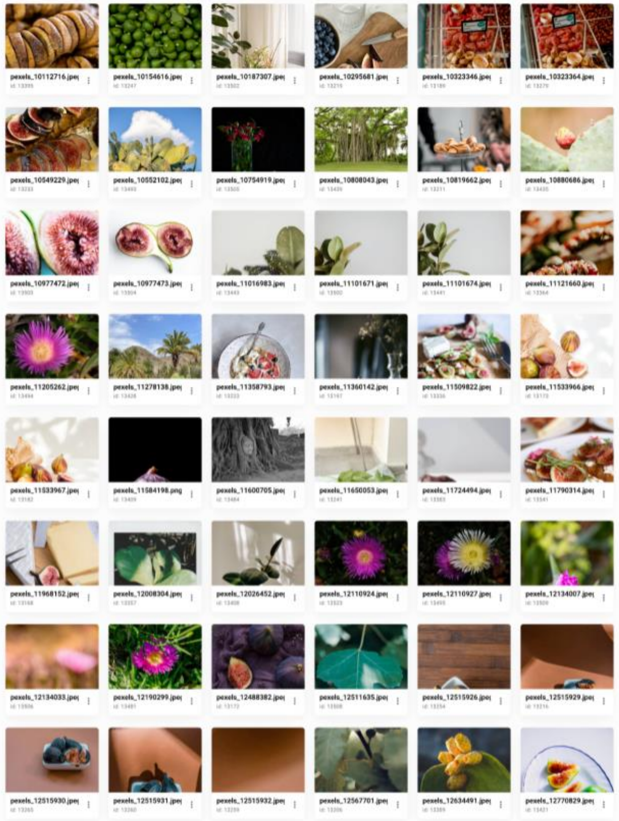


Image Retrieval with Text Query

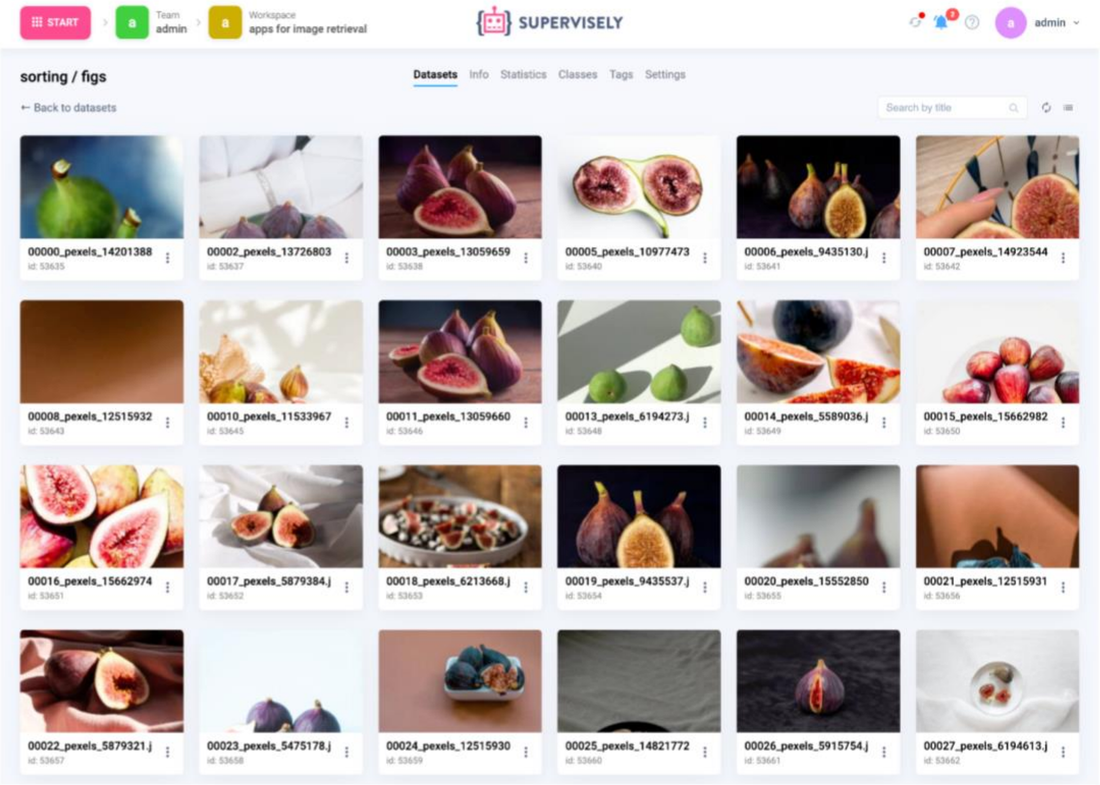
Large image database



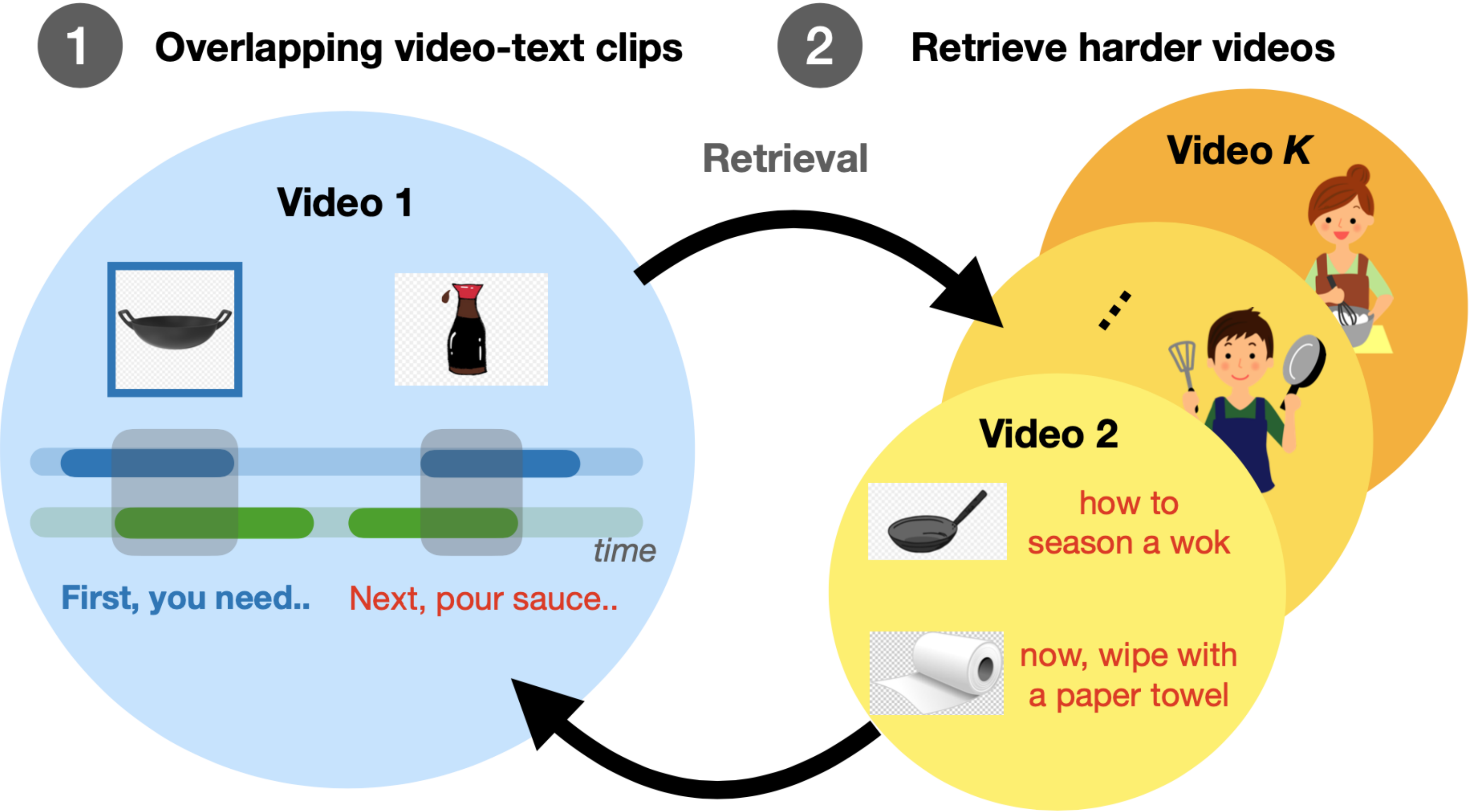
Text prompt (search query):

`images with fig fruits without flowers`

Semantic search results:



VideoCLIP

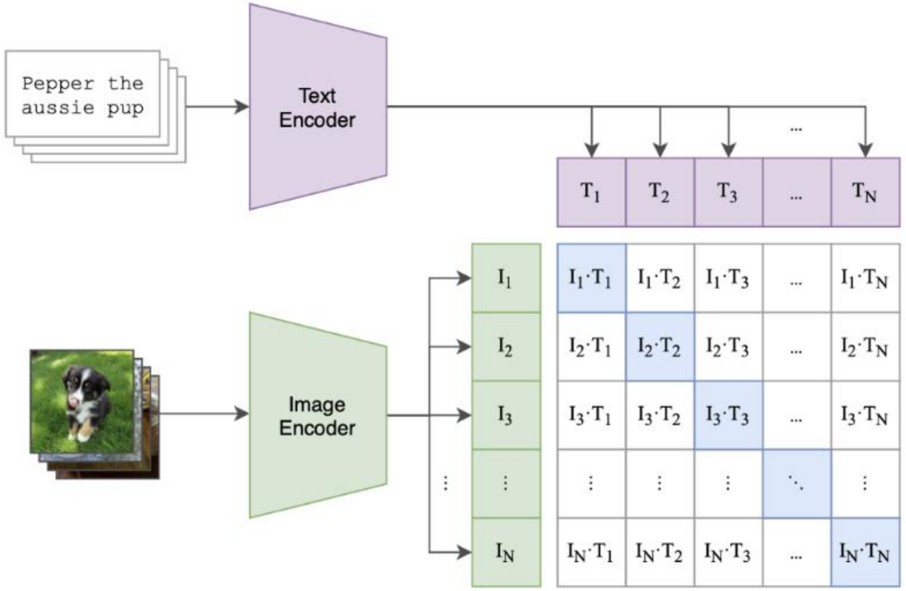


VideoCLIP: Contrastive learning with hard-retrieved negatives and overlapping positives for video-text pre-training.

VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding

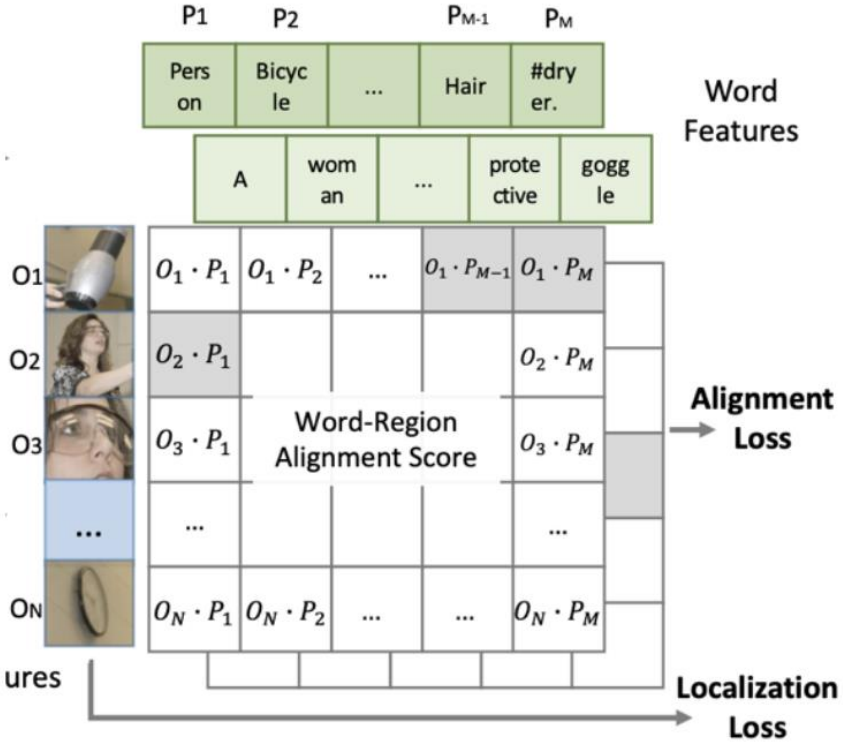
GLIP: Grounded Language-Image Pre-training

CLIP: capture information for whole image



	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

GLIP: capture information more for objects/entities

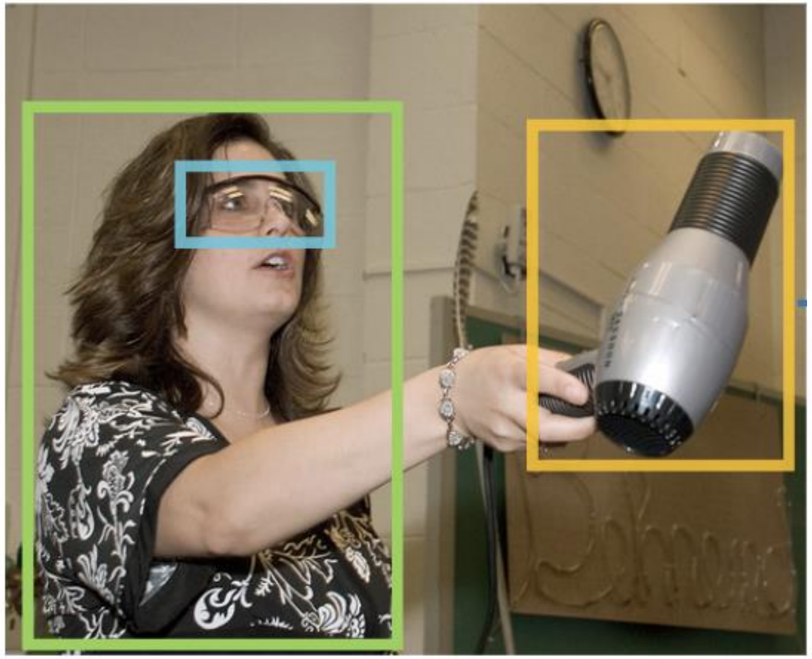


	P1	P2	...	PM-1	PM
O1	$O_1 \cdot P_1$	$O_1 \cdot P_2$	...	$O_1 \cdot P_{M-1}$	$O_1 \cdot P_M$
O2	$O_2 \cdot P_1$				$O_2 \cdot P_M$
O3	$O_3 \cdot P_1$	Word-Region Alignment Score			$O_3 \cdot P_M$
...	...				...
ON	$O_N \cdot P_1$	$O_N \cdot P_2$	...	...	$O_N \cdot P_M$

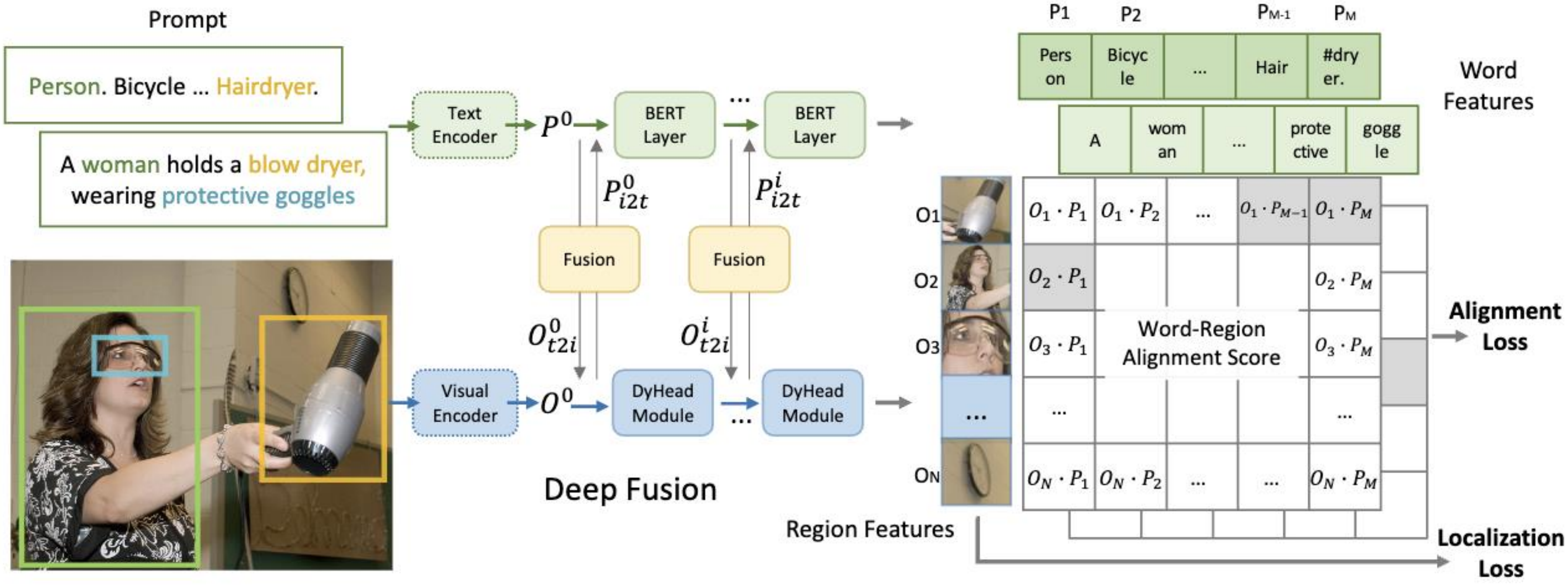
Object Detection and Text Grounding

Person. Bicycle … Hairdryer.

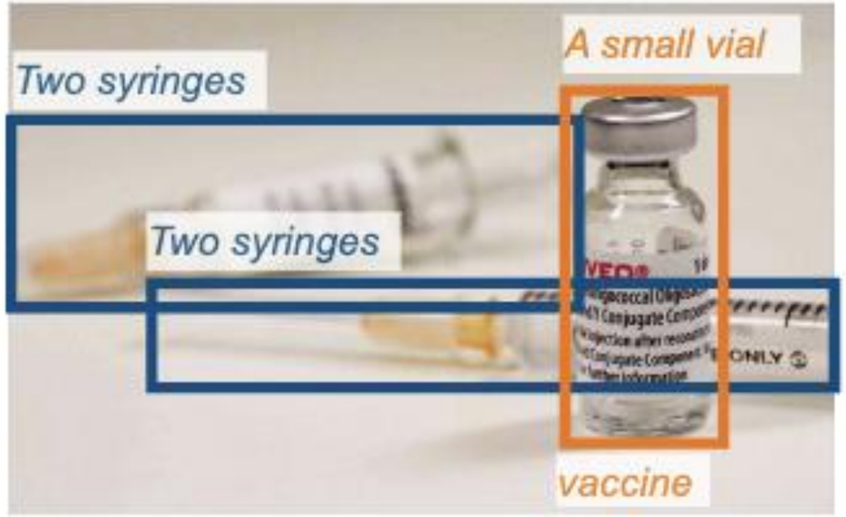
A woman holds a blow dryer, wearing protective goggles



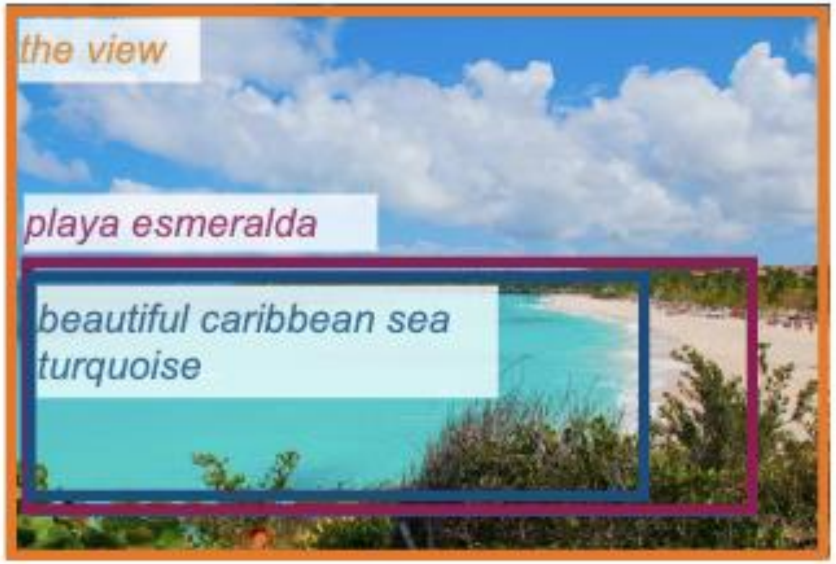
GLIP: Grounded Language-Image Pre-training



Grounding Results

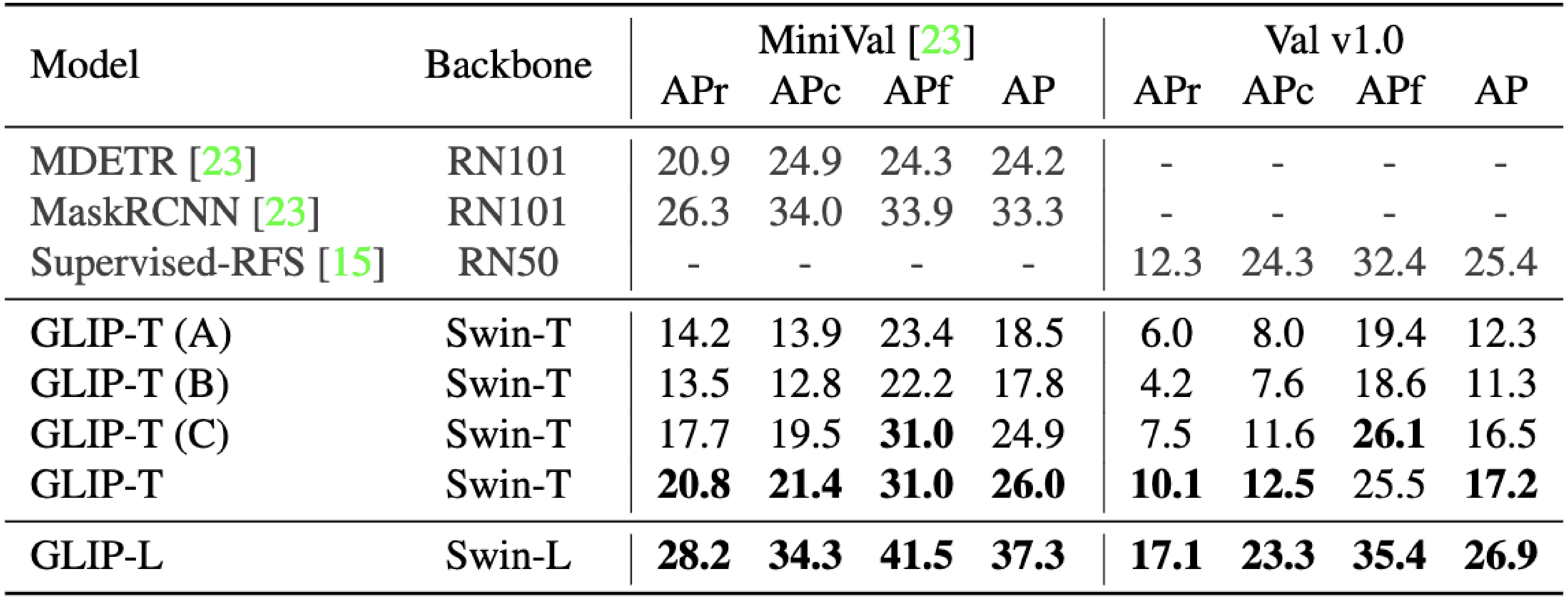


Two syringes and a small vial of vaccine.



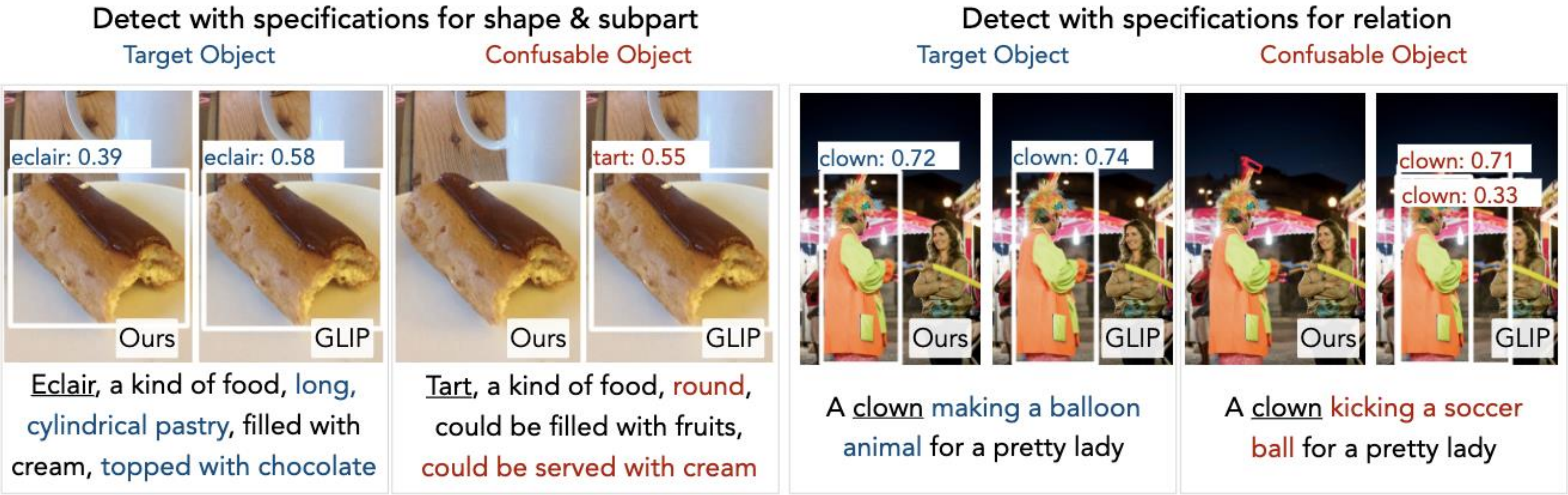
playa esmeralda in holguin, cuba. the view from the top of the beach. beautiful caribbean sea turquoise

Zero-Shot Grounding



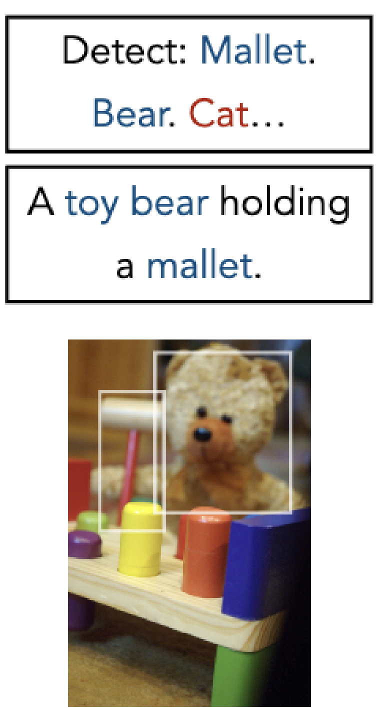
Model	Backbone	MiniVal [23]				Val v1.0			
		APr	APc	APf	AP	APr	APc	APf	AP
MDETR [23]	RN101	20.9	24.9	24.3	24.2	-	-	-	-
MaskRCNN [23]	RN101	26.3	34.0	33.9	33.3	-	-	-	-
Supervised-RFS [15]	RN50	-	-	-	-	12.3	24.3	32.4	25.4
GLIP-T (A)	Swin-T	14.2	13.9	23.4	18.5	6.0	8.0	19.4	12.3
GLIP-T (B)	Swin-T	13.5	12.8	22.2	17.8	4.2	7.6	18.6	11.3
GLIP-T (C)	Swin-T	17.7	19.5	**31.0**	24.9	7.5	11.6	**26.1**	16.5
GLIP-T	Swin-T	**20.8**	**21.4**	**31.0**	**26.0**	**10.1**	**12.5**	25.5	**17.2**
GLIP-L	Swin-L	**28.2**	**34.3**	**41.5**	**37.3**	**17.1**	**23.3**	**35.4**	**26.9**

DesCo: Object Recognition with Language Description


Detect with specifications for shape & subpart
Target Object
Confusable Object
eclair: 0.39
Ours
eclair: 0.58
GLIP
Ours
tart: 0.55
GLIP
Eclair, a kind of food, long, cylindrical pastry, filled with cream, topped with chocolate
Tart, a kind of food, round, could be filled with fruits, could be served with cream
Detect with specifications for relation
Target Object
Confusable Object
clown: 0.72
Ours
clown: 0.74
GLIP
A clown making a balloon animal for a pretty lady
Ours
clown: 0.71
clown: 0.33
GLIP
A clown kicking a soccer ball for a pretty lady

DesCo: Learning Object Recognition with Rich Language Descriptions

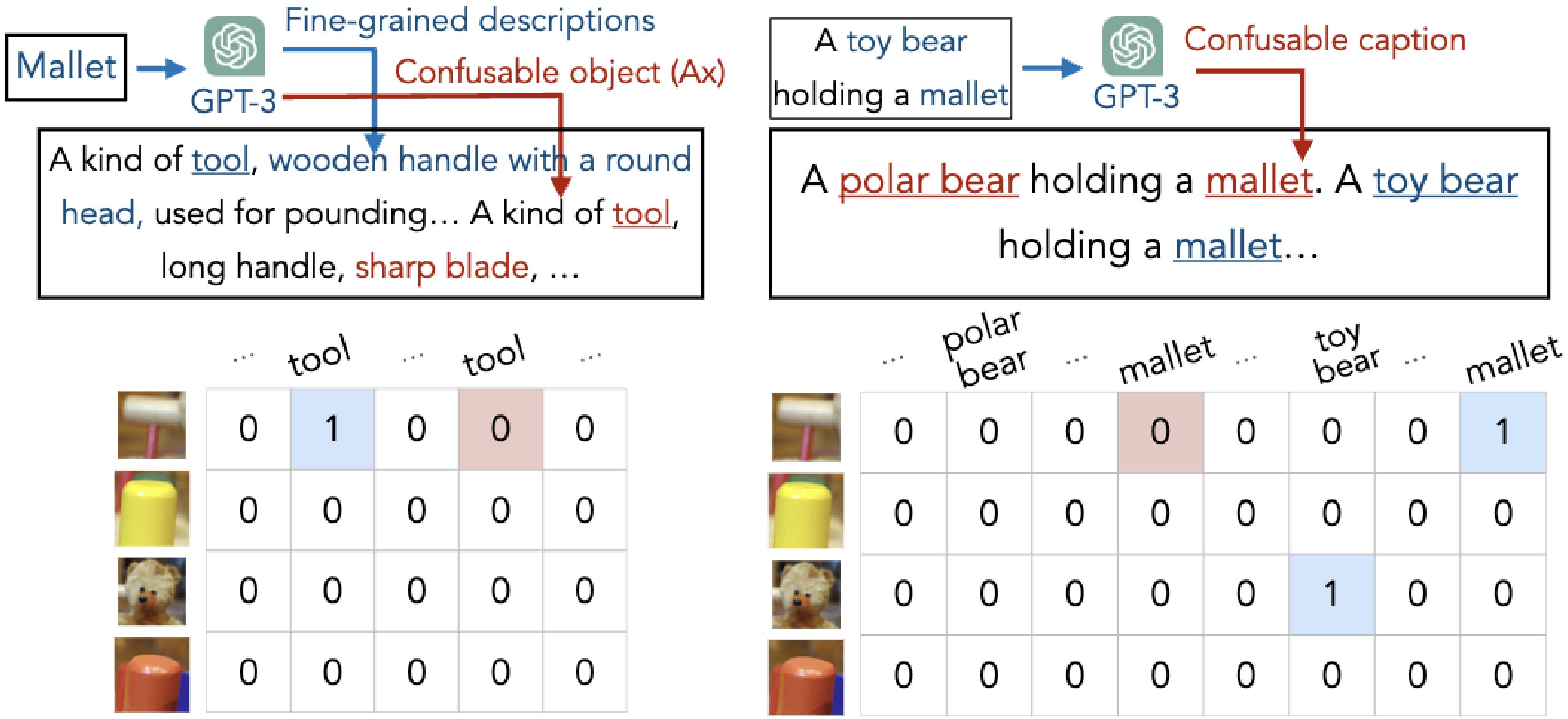
DesCo: Description-Conditioned



Detect: Mallet. Bear. Cat…

A toy bear holding a mallet.

Original training data for GLIP



Mallet → GPT-3

Fine-grained descriptions

Confusable object (Ax)

A kind of tool, wooden handle with a round head, used for pounding… A kind of tool, long handle, sharp blade, …

	…	tool	…	tool	…
	0	1	0	0	0
	0	0	0	0	0
	0	0	0	0	0
	0	0	0	0	0

A toy bear holding a mallet → GPT-3

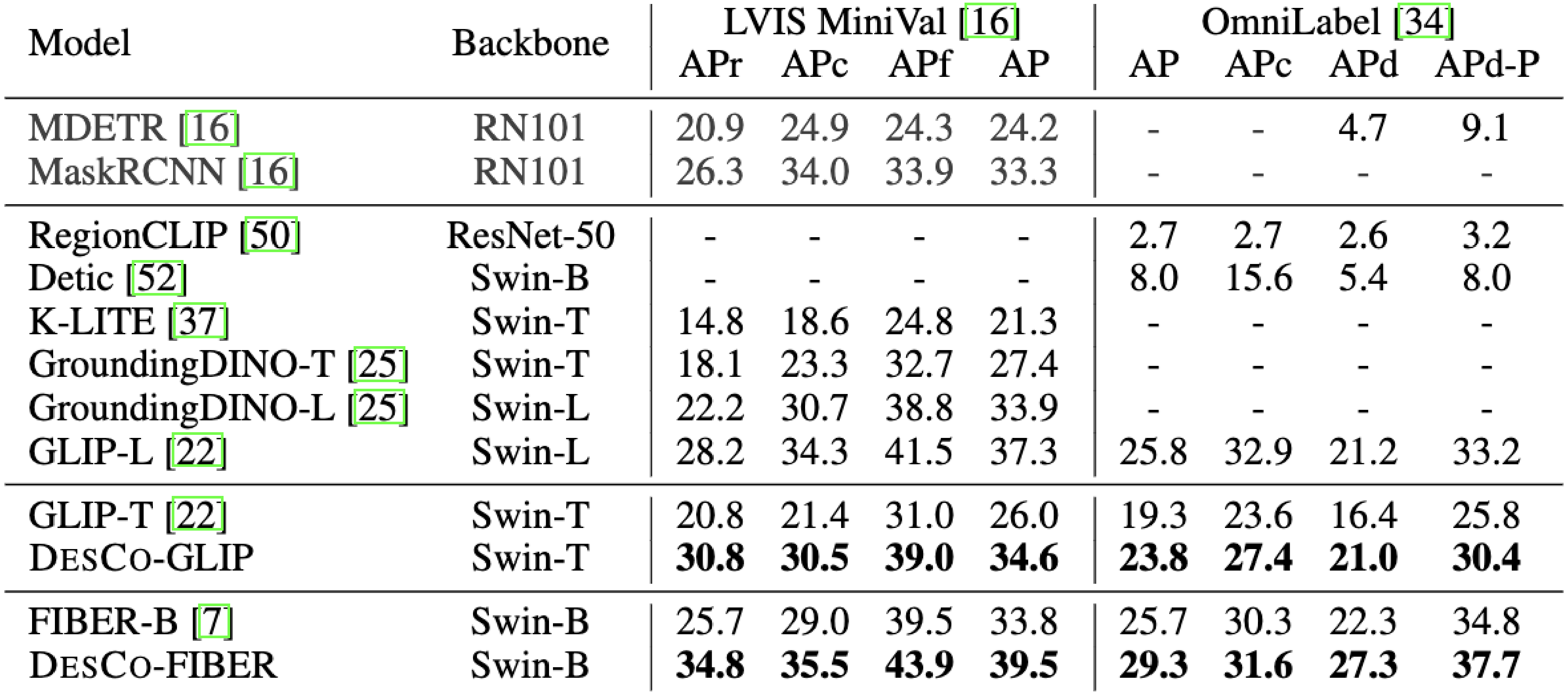
Confusable caption

A polar bear holding a mallet. A toy bear holding a mallet…

	…	polar bear	…	mallet	…	toy bear	…	mallet
	0	0	0	0	0	0	0	1
	0	0	0	0	0	0	0	0
	0	0	0	0	0	1	0	0
	0	0	0	0	0	0	0	0

Description-rich and context-sensitive data for DesCo-GLIP

Zero-Shot Grounding

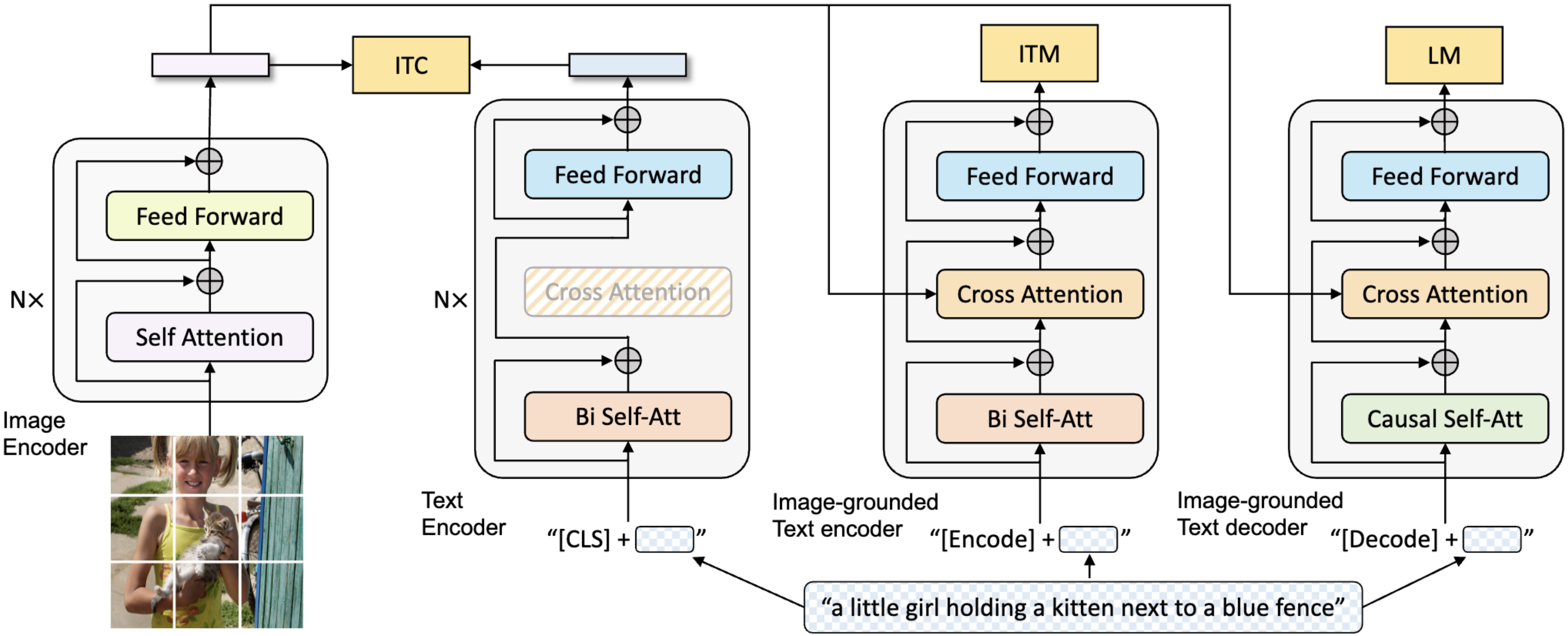


Model	Backbone	LVIS MiniVal [16]				OmniLabel [34]			
		APr	APc	APf	AP	AP	APc	APd	APd-P
MDETR [16]	RN101	20.9	24.9	24.3	24.2	-	-	4.7	9.1
MaskRCNN [16]	RN101	26.3	34.0	33.9	33.3	-	-	-	-
RegionCLIP [50]	ResNet-50	-	-	-	-	2.7	2.7	2.6	3.2
Detic [52]	Swin-B	-	-	-	-	8.0	15.6	5.4	8.0
K-LITE [37]	Swin-T	14.8	18.6	24.8	21.3	-	-	-	-
GroundingDINO-T [25]	Swin-T	18.1	23.3	32.7	27.4	-	-	-	-
GroundingDINO-L [25]	Swin-L	22.2	30.7	38.8	33.9	-	-	-	-
GLIP-L [22]	Swin-L	28.2	34.3	41.5	37.3	25.8	32.9	21.2	33.2
GLIP-T [22]	Swin-T	20.8	21.4	31.0	26.0	19.3	23.6	16.4	25.8
DESCO-GLIP	Swin-T	**30.8**	**30.5**	**39.0**	**34.6**	**23.8**	**27.4**	**21.0**	**30.4**
FIBER-B [7]	Swin-B	25.7	29.0	39.5	33.8	25.7	30.3	22.3	34.8
DESCO-FIBER	Swin-B	**34.8**	**35.5**	**43.9**	**39.5**	**29.3**	**31.6**	**27.3**	**37.7**

Encoder-Only vs. Encoder-Decoder

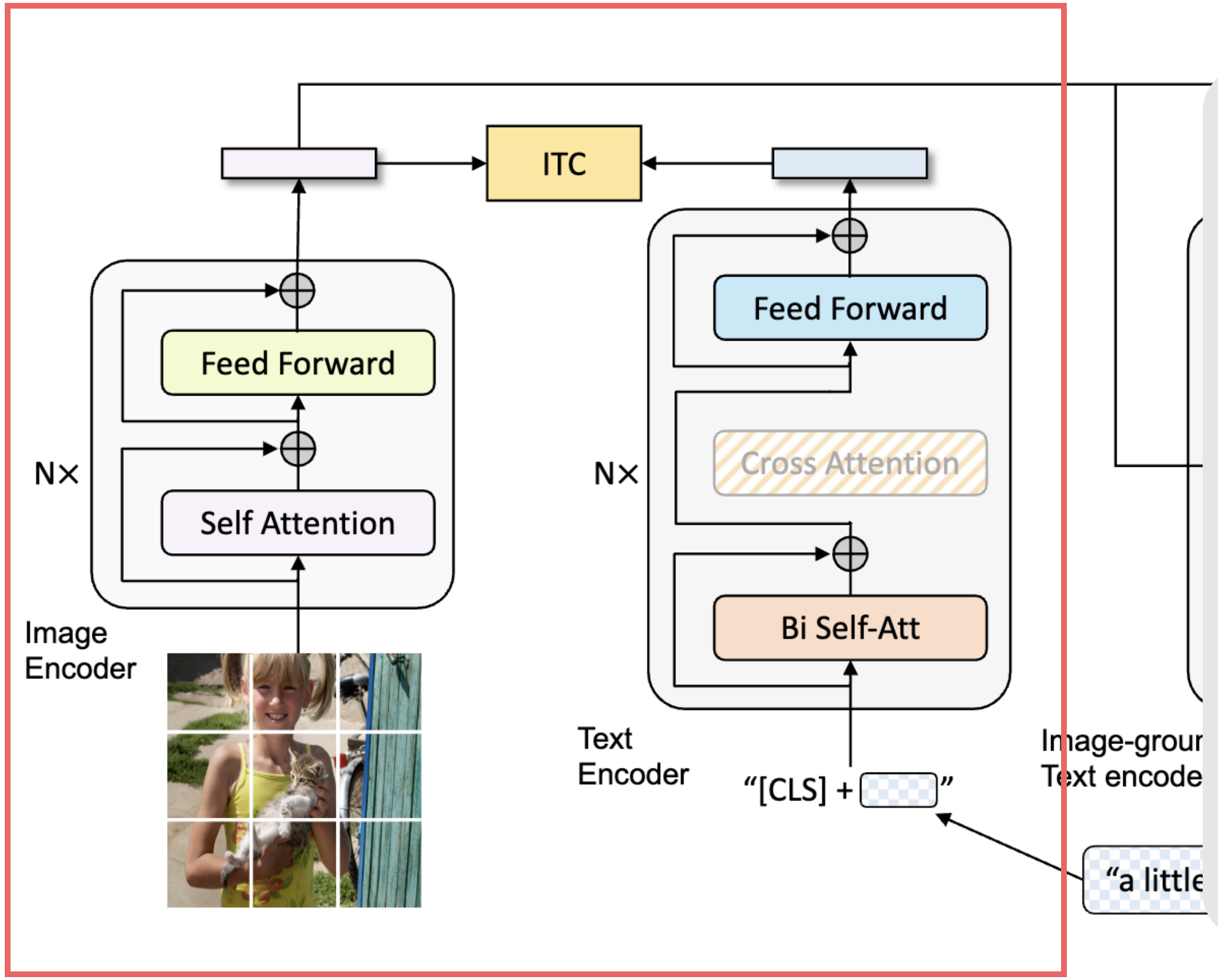
- Encoder-only
  - CLIP, GLIP, DesCo, etc.
  - Better for image-text retrieval
- Encoder-decoder
  - Better for generation

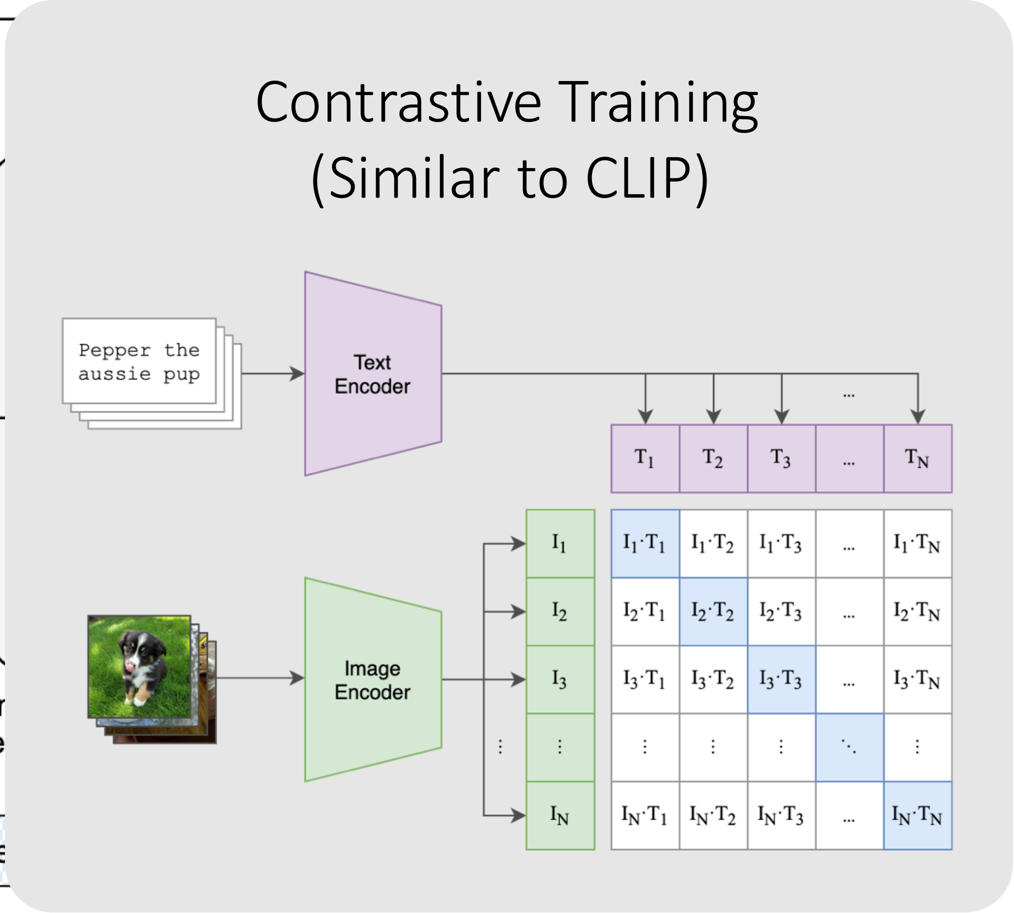
BLIP: Bootstrapping Language-Image Pre-training



BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation

A Unified Framework

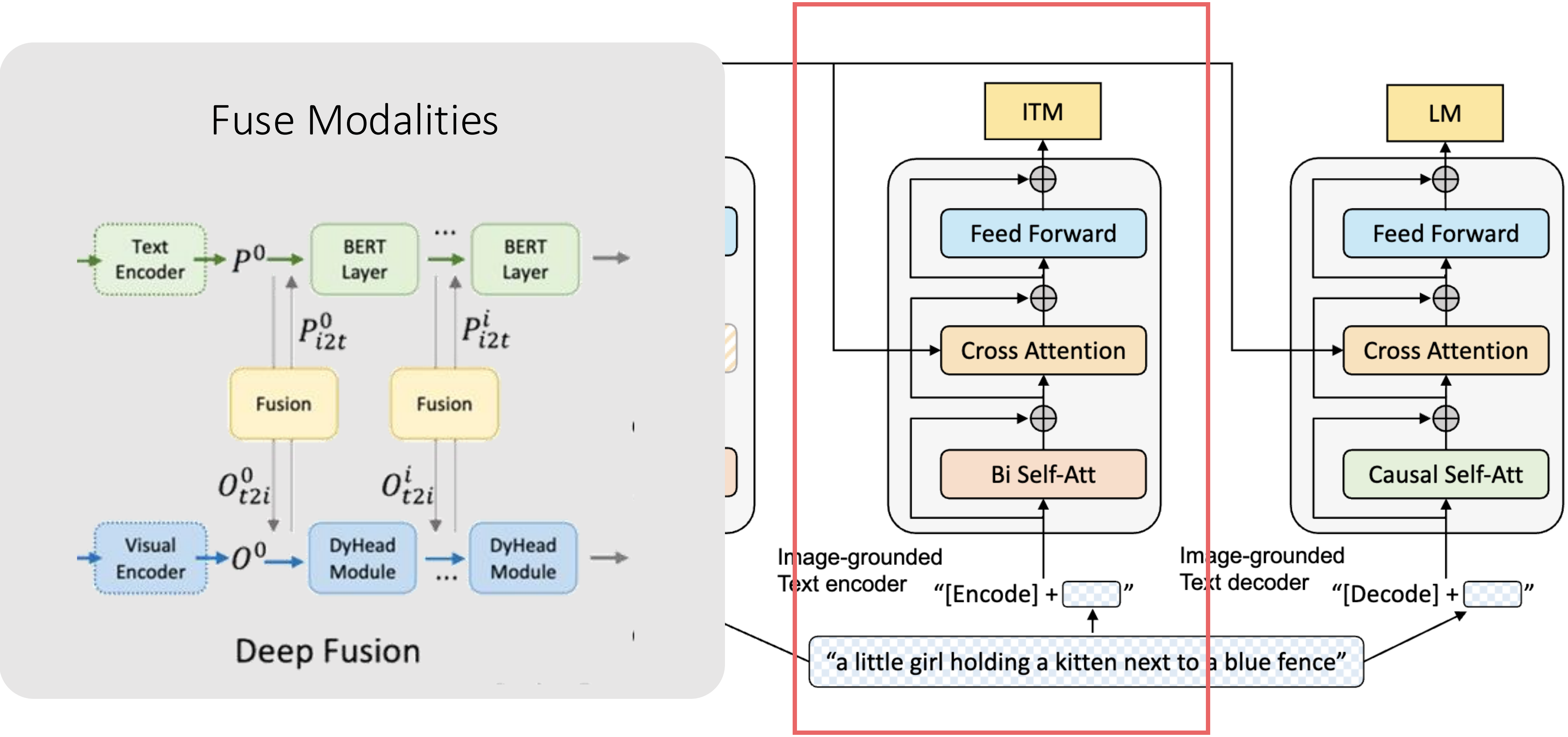




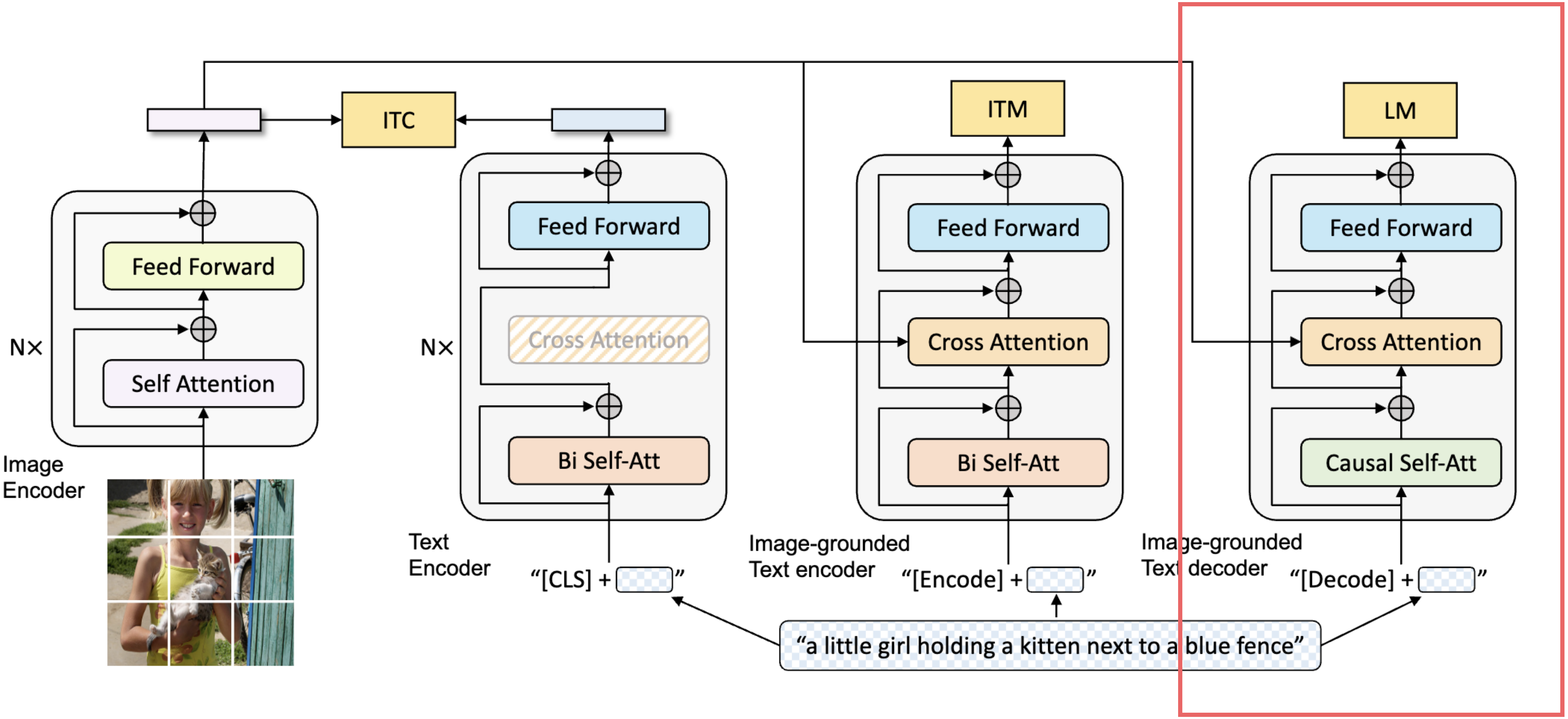
Contrastive Training
(Similar to CLIP)

	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

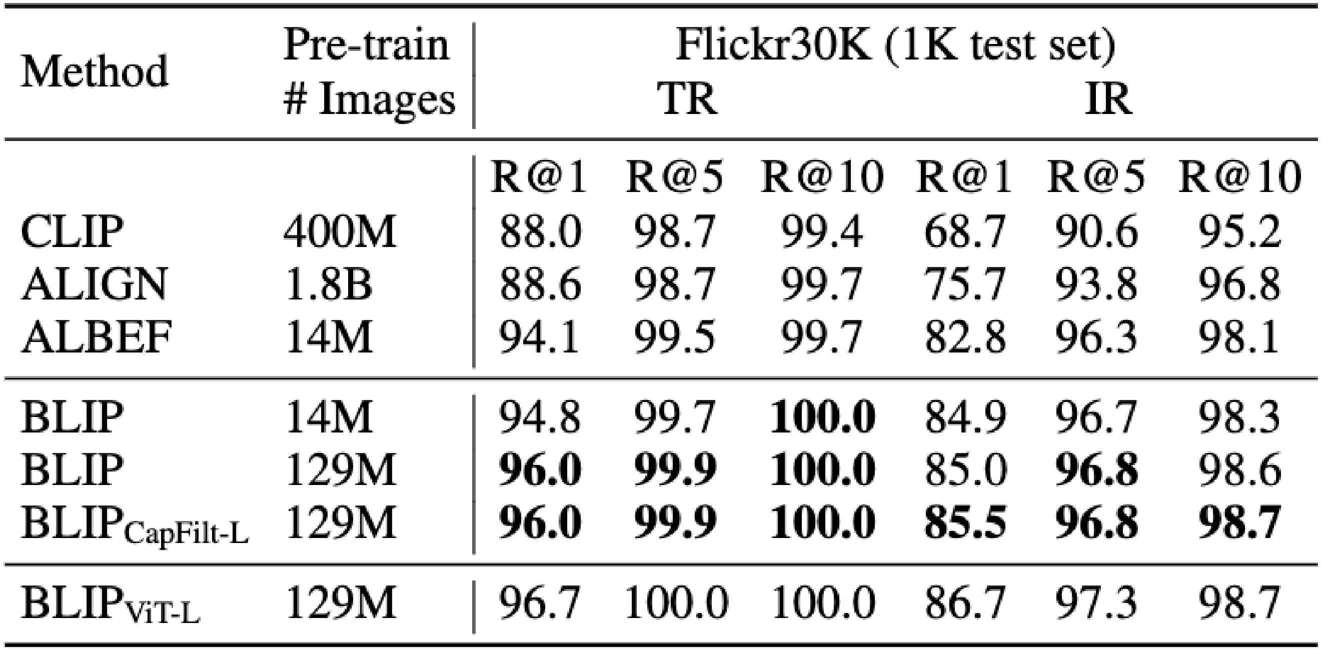
A Unified Framework



A Unified Framework

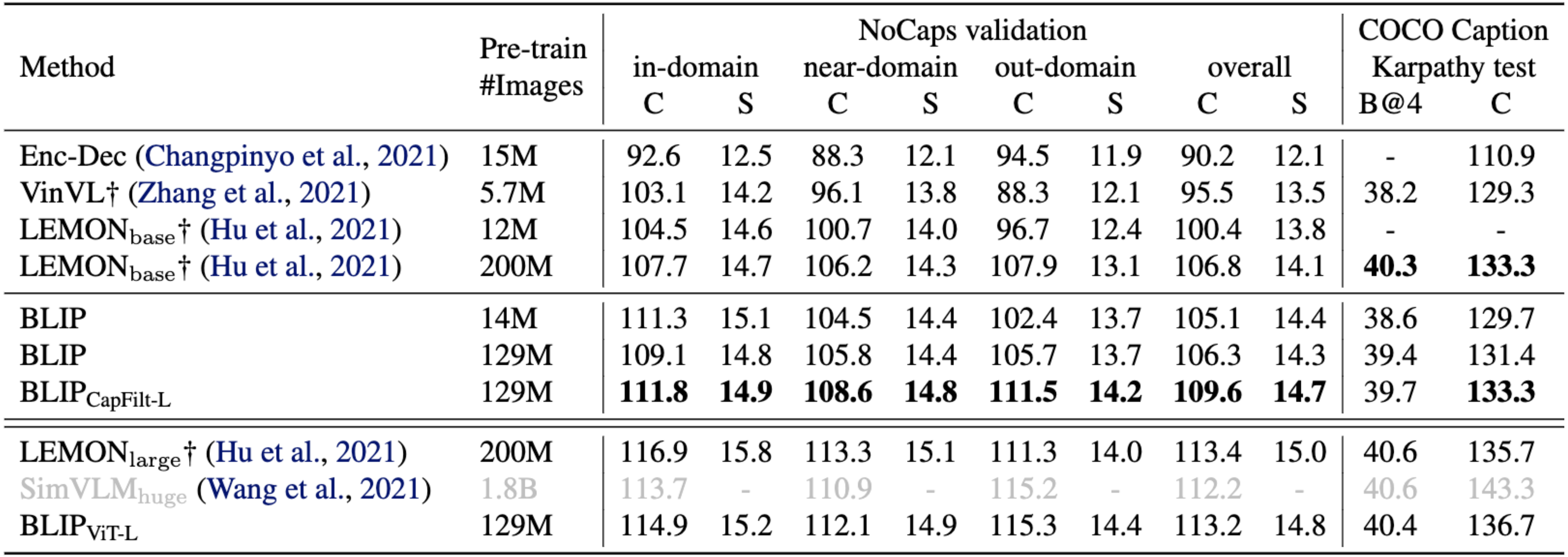


Zero-Shot Image-Text Retrieval



Method	Pre-train # Images	Flickr30K (1K test set) TR R@1	R@5	R@10	IR R@1	R@5	R@10
CLIP	400M	88.0	98.7	99.4	68.7	90.6	95.2
ALIGN	1.8B	88.6	98.7	99.7	75.7	93.8	96.8
ALBEF	14M	94.1	99.5	99.7	82.8	96.3	98.1
BLIP	14M	94.8	99.7	**100.0**	84.9	96.7	98.3
BLIP	129M	**96.0**	**99.9**	**100.0**	85.0	**96.8**	98.6
$\text{BLIP}_{\text{CapFilt-L}}$	129M	**96.0**	**99.9**	**100.0**	**85.5**	**96.8**	**98.7**
$\text{BLIP}_{\text{ViT-L}}$	129M	96.7	100.0	100.0	86.7	97.3	98.7

Image Captioning

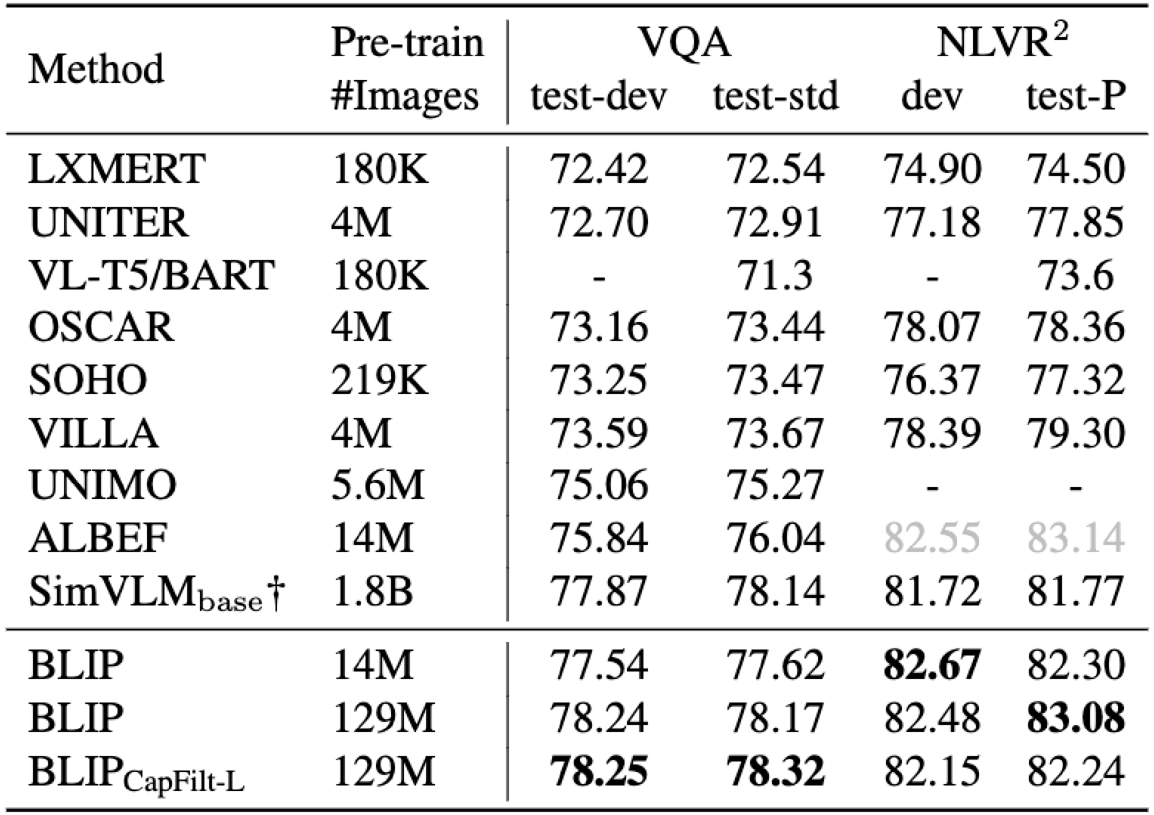


Method	Pre-train #Images	NoCaps validation								COCO Caption Karpathy test	
		in-domain		near-domain		out-domain		overall			
		C	S	C	S	C	S	C	S	B@4	C
Enc-Dec (Changpinyo et al., 2021)	15M	92.6	12.5	88.3	12.1	94.5	11.9	90.2	12.1	-	110.9
VinVL† (Zhang et al., 2021)	5.7M	103.1	14.2	96.1	13.8	88.3	12.1	95.5	13.5	38.2	129.3
LEMON$_{\text{base}}$† (Hu et al., 2021)	12M	104.5	14.6	100.7	14.0	96.7	12.4	100.4	13.8	-	-
LEMON$_{\text{base}}$† (Hu et al., 2021)	200M	107.7	14.7	106.2	14.3	107.9	13.1	106.8	14.1	**40.3**	**133.3**
BLIP	14M	111.3	15.1	104.5	14.4	102.4	13.7	105.1	14.4	38.6	129.7
BLIP	129M	109.1	14.8	105.8	14.4	105.7	13.7	106.3	14.3	39.4	131.4
BLIP$_{\text{CapFilt-L}}$	129M	**111.8**	**14.9**	**108.6**	**14.8**	**111.5**	**14.2**	**109.6**	**14.7**	39.7	**133.3**
LEMON$_{\text{large}}$† (Hu et al., 2021)	200M	116.9	15.8	113.3	15.1	111.3	14.0	113.4	15.0	40.6	135.7
SimVLM$_{\text{huge}}$ (Wang et al., 2021)	1.8B	113.7	-	110.9	-	115.2	-	112.2	-	40.6	143.3
BLIP$_{\text{ViT-L}}$	129M	114.9	15.2	112.1	14.9	115.3	14.4	113.2	14.8	40.4	136.7

Visual Question Answering

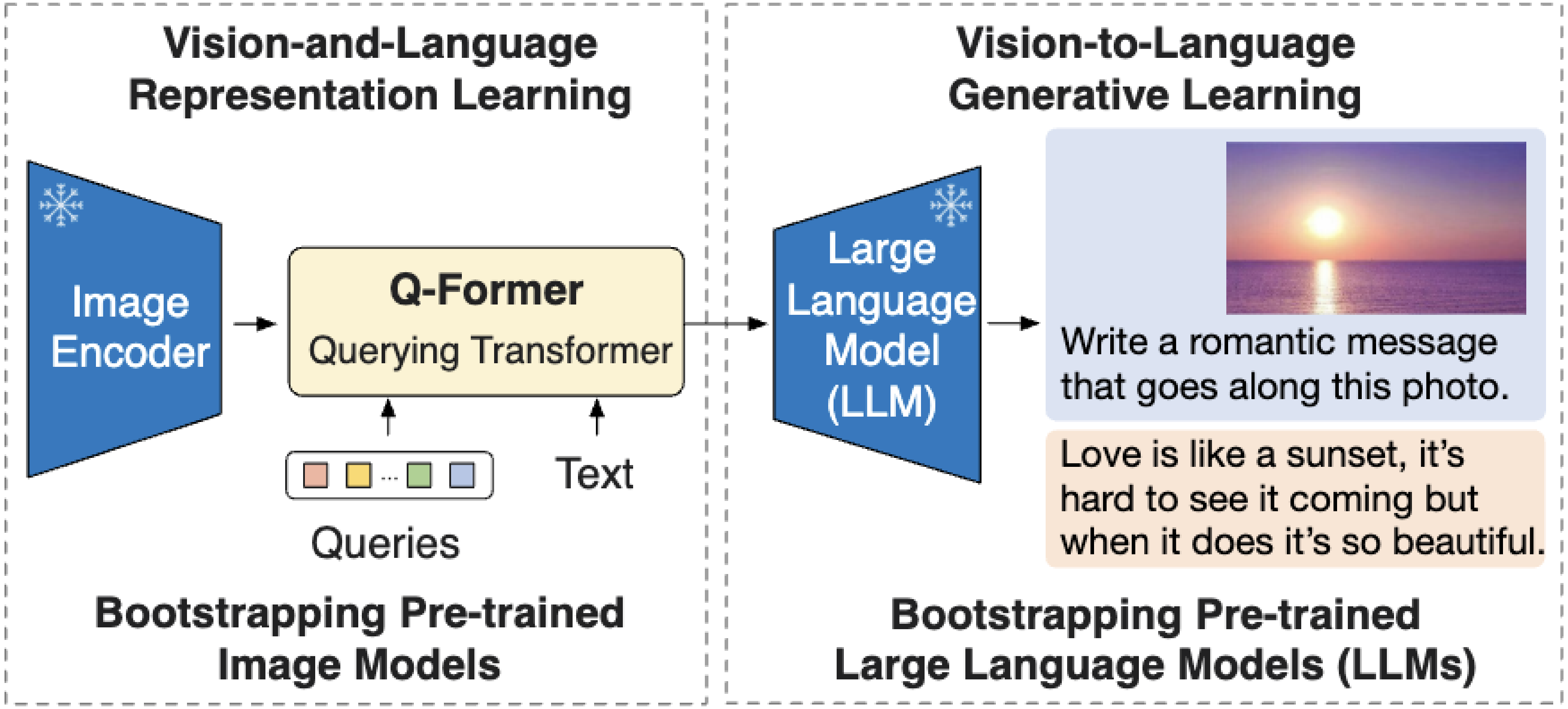
Visual Question Answering

Does it appear to be rainy?
Does this person have 20/20 vision?



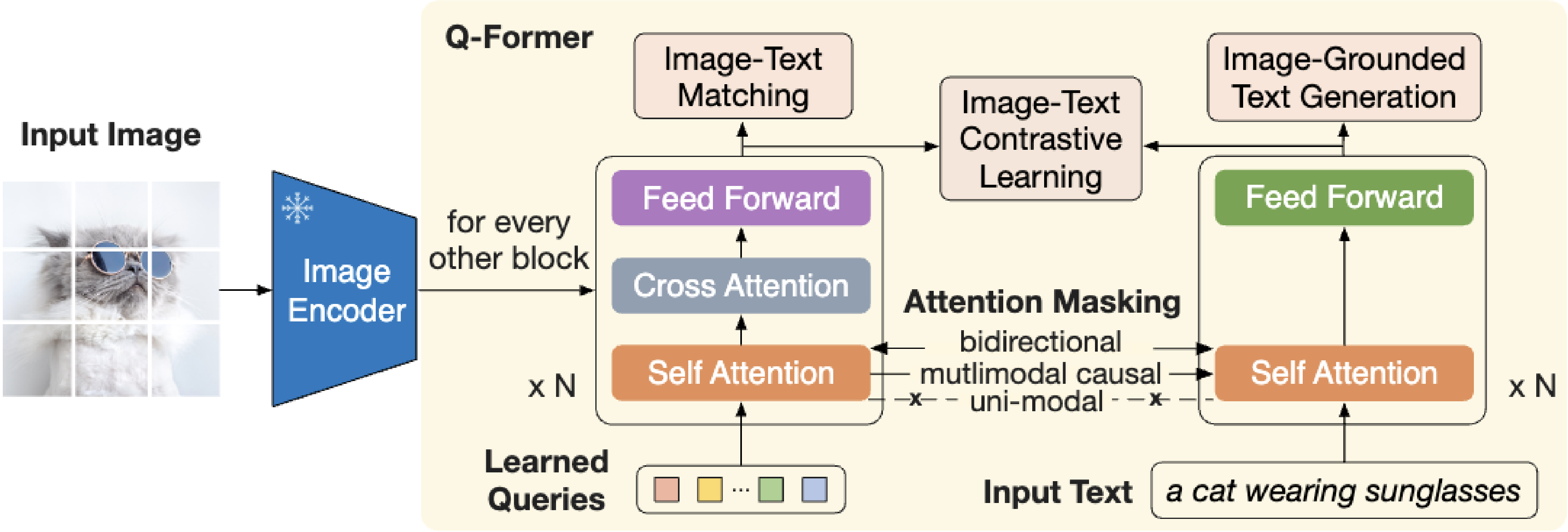
Method	Pre-train #Images	VQA test-dev	VQA test-std	NLVR^2 dev	NLVR^2 test-P
LXMERT	180K	72.42	72.54	74.90	74.50
UNITER	4M	72.70	72.91	77.18	77.85
VL-T5/BART	180K	-	71.3	-	73.6
OSCAR	4M	73.16	73.44	78.07	78.36
SOHO	219K	73.25	73.47	76.37	77.32
VILLA	4M	73.59	73.67	78.39	79.30
UNIMO	5.6M	75.06	75.27	-	-
ALBEF	14M	75.84	76.04	82.55	83.14
SimVLM$_{base}$†	1.8B	77.87	78.14	81.72	81.77
BLIP	14M	77.54	77.62	**82.67**	82.30
BLIP	129M	78.24	78.17	82.48	**83.08**
BLIP$_{CapFilt-L}$	129M	**78.25**	**78.32**	82.15	82.24

BLIP-2: Frozen Image Encoders and Large Language Models

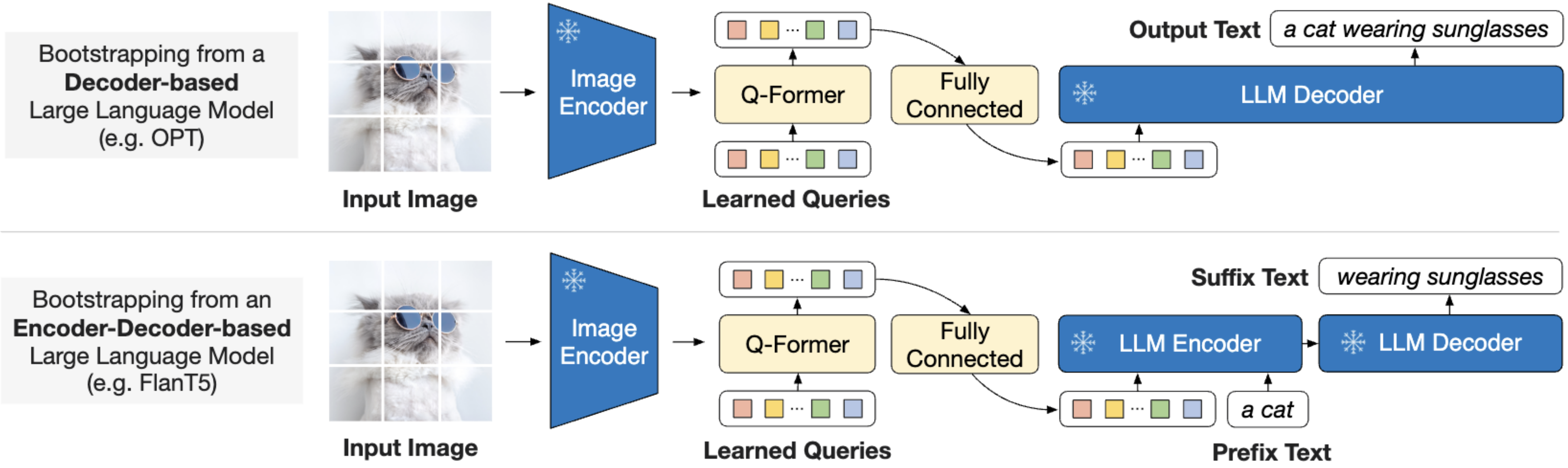

Vision-and-Language Representation Learning
Image Encoder
Q-Former
Querying Transformer
Queries
Text
Bootstrapping Pre-trained Image Models
Vision-to-Language Generative Learning
Large Language Model (LLM)
Write a romantic message that goes along this photo.
Love is like a sunset, it's hard to see it coming but when it does it's so beautiful.
Bootstrapping Pre-trained Large Language Models (LLMs)

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

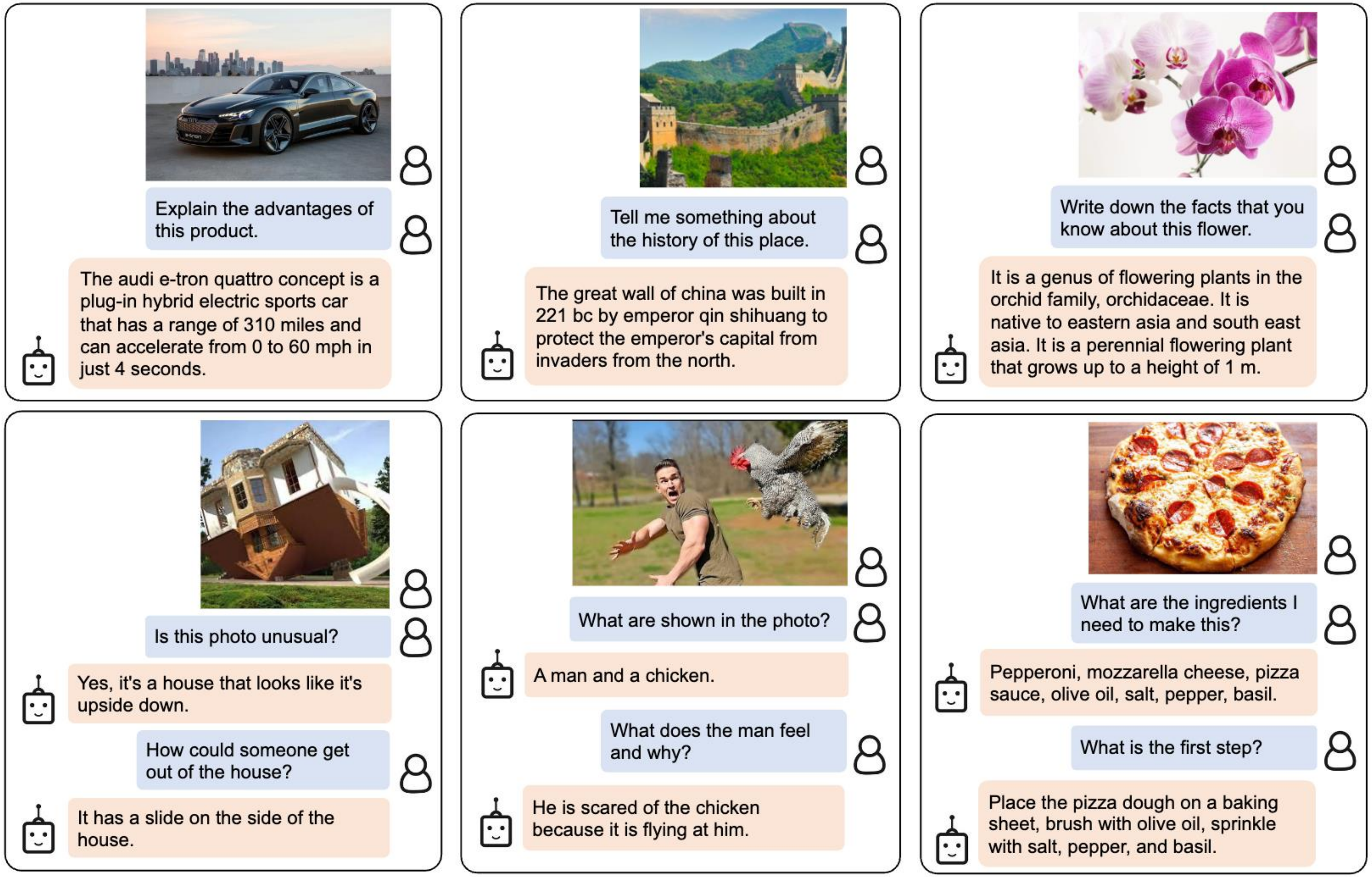
Vision-Language Representation Learning



Vision-to-Language Generative Pre-Training



Instructed Zero-Shot Image-to-Text Generation



Explain the advantages of this product.

The audi e-tron quattro concept is a plug-in hybrid electric sports car that has a range of 310 miles and can accelerate from 0 to 60 mph in just 4 seconds.

Tell me something about the history of this place.

The great wall of china was built in 221 bc by emperor qin shihuang to protect the emperor's capital from invaders from the north.

Write down the facts that you know about this flower.

It is a genus of flowering plants in the orchid family, orchidaceae. It is native to eastern asia and south east asia. It is a perennial flowering plant that grows up to a height of 1 m.

Is this photo unusual?

Yes, it's a house that looks like it's upside down.

How could someone get out of the house?

It has a slide on the side of the house.

What are shown in the photo?

A man and a chicken.

What does the man feel and why?

He is scared of the chicken because it is flying at him.

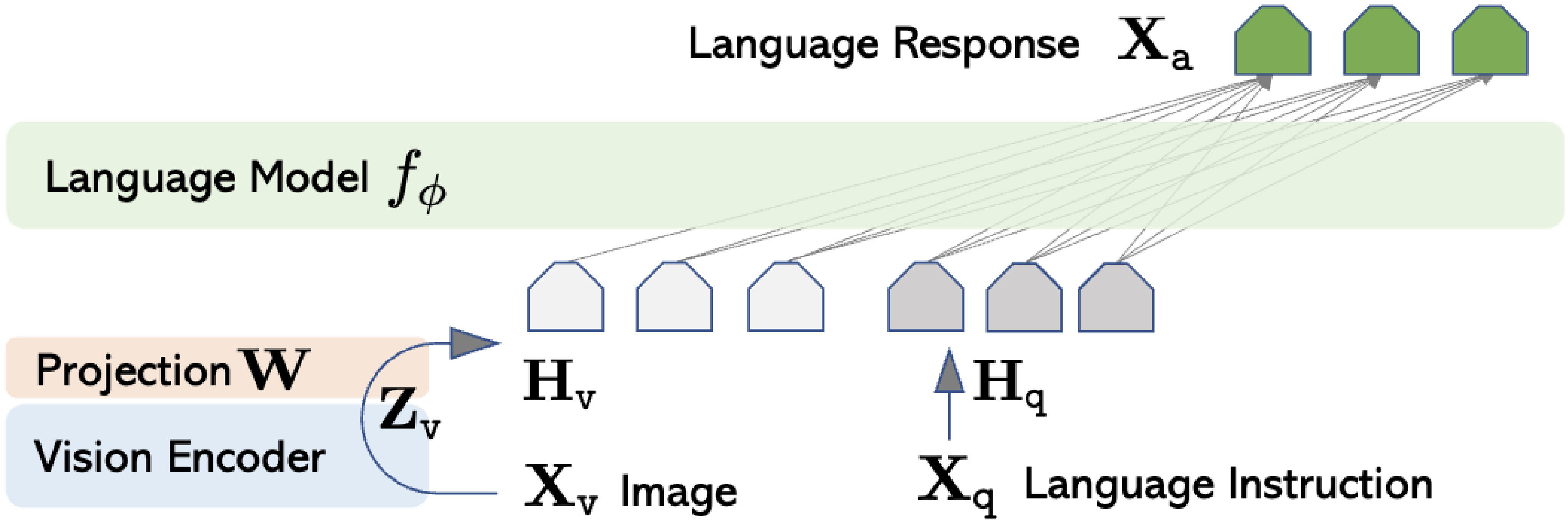
What are the ingredients I need to make this?

Pepperoni, mozzarella cheese, pizza sauce, olive oil, salt, pepper, basil.

What is the first step?

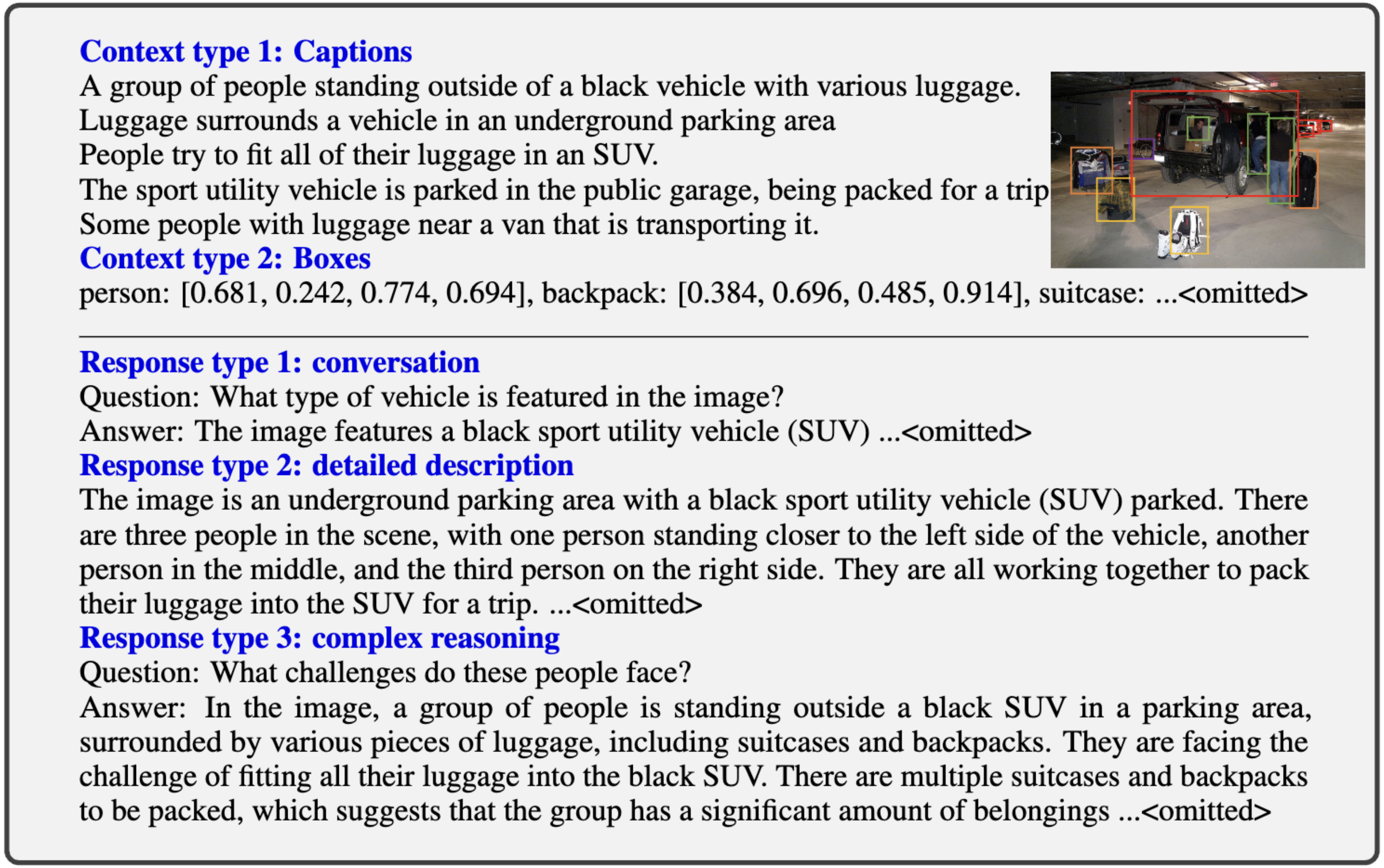
Place the pizza dough on a baking sheet, brush with olive oil, sprinkle with salt, pepper, and basil.

LLaVA: Visual Instruction Tuning



https://llava-vl.github.io/

GPT-Assisted Visual Instruction Data Generation



Context type 1: Captions

A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.

Context type 2: Boxes

person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Response type 1: conversation

Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV) ...<omitted>

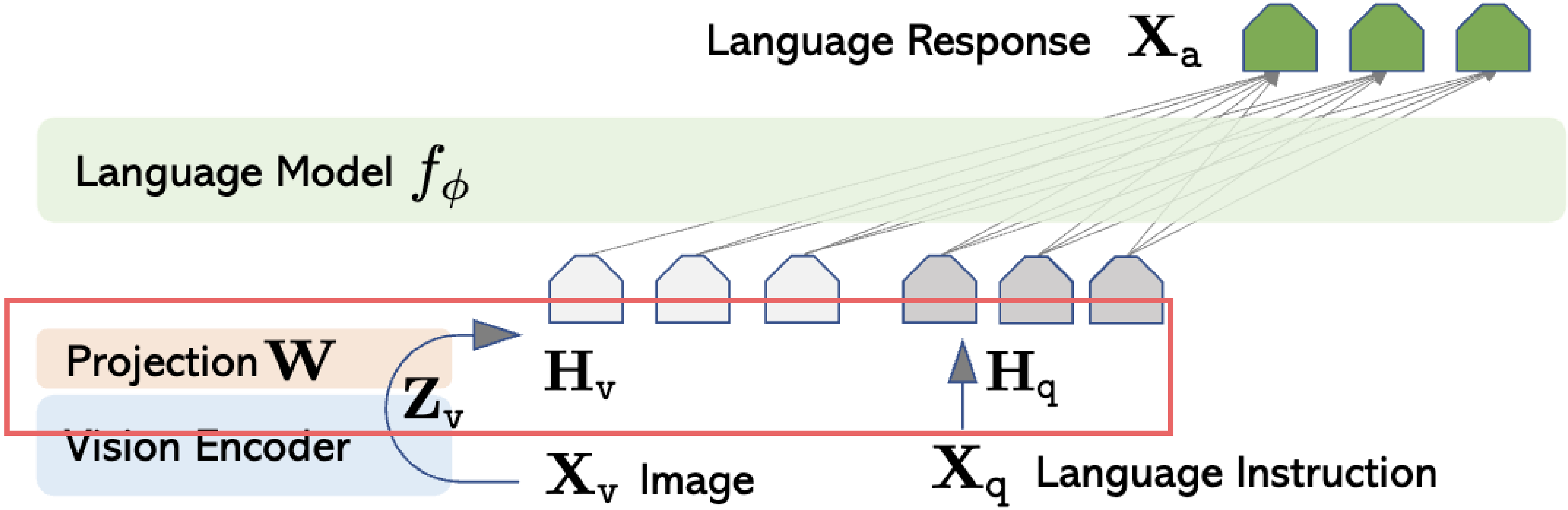
Response type 2: detailed description

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>

Response type 3: complex reasoning

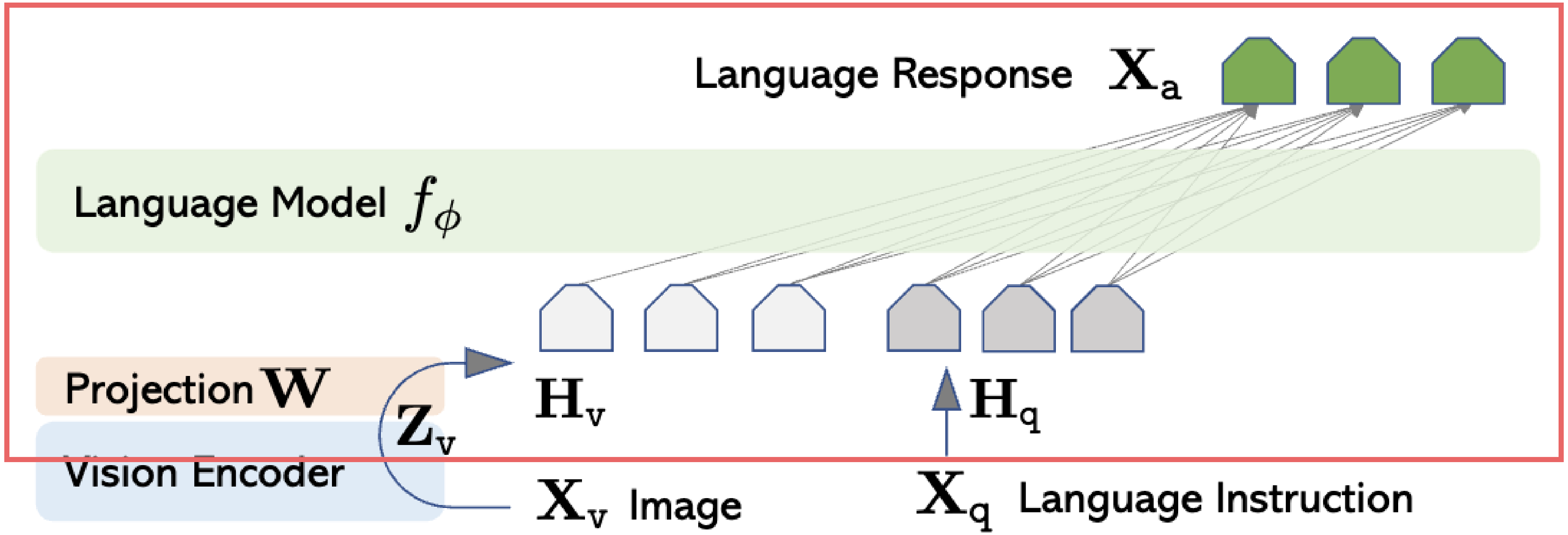
Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

Stage 1: Pre-training for Feature Alignment



Train with Image-Text Pairs

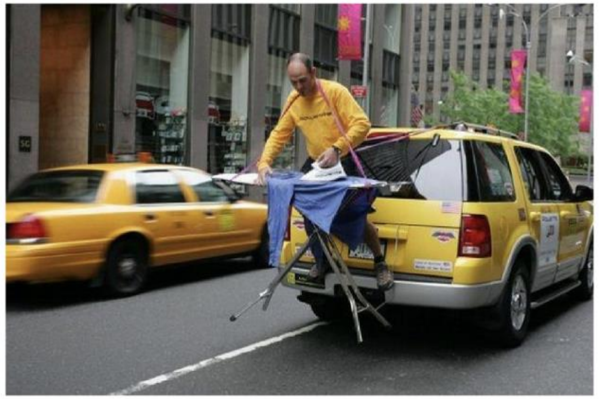
Stage 2: Fine-tuning End-to-End



Visual Chat (Visual Instruction Data) and Science QA

Examples

Visual input example, Extreme Ironing:

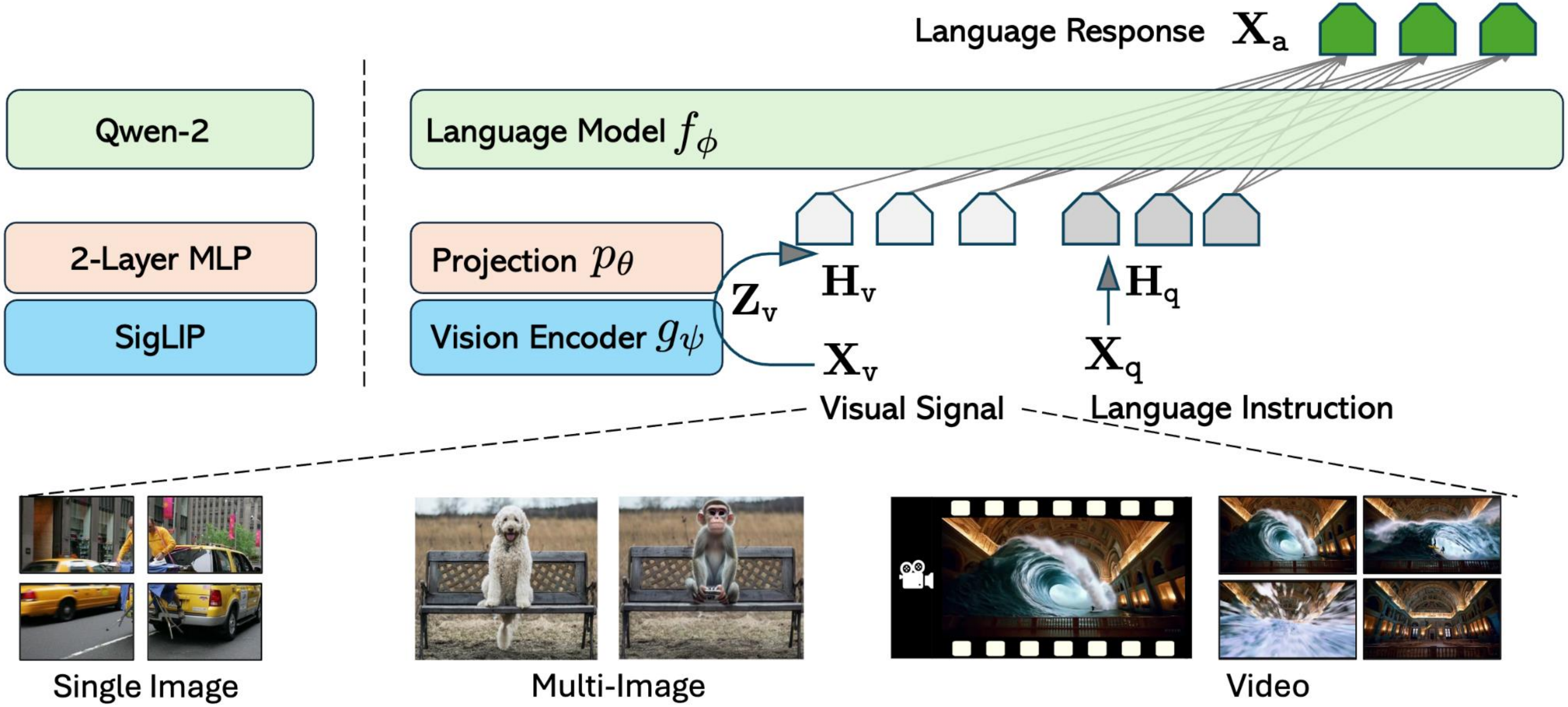


Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

User	What is unusual about this image?
GPT-4 [36]	The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.
User	What is unusual about this image?
BLIP-2	a man is sitting on the back of a yellow cab
User	What is unusual about this image?
OpenFlamingo	The man is drying his clothes on the hood of his car.

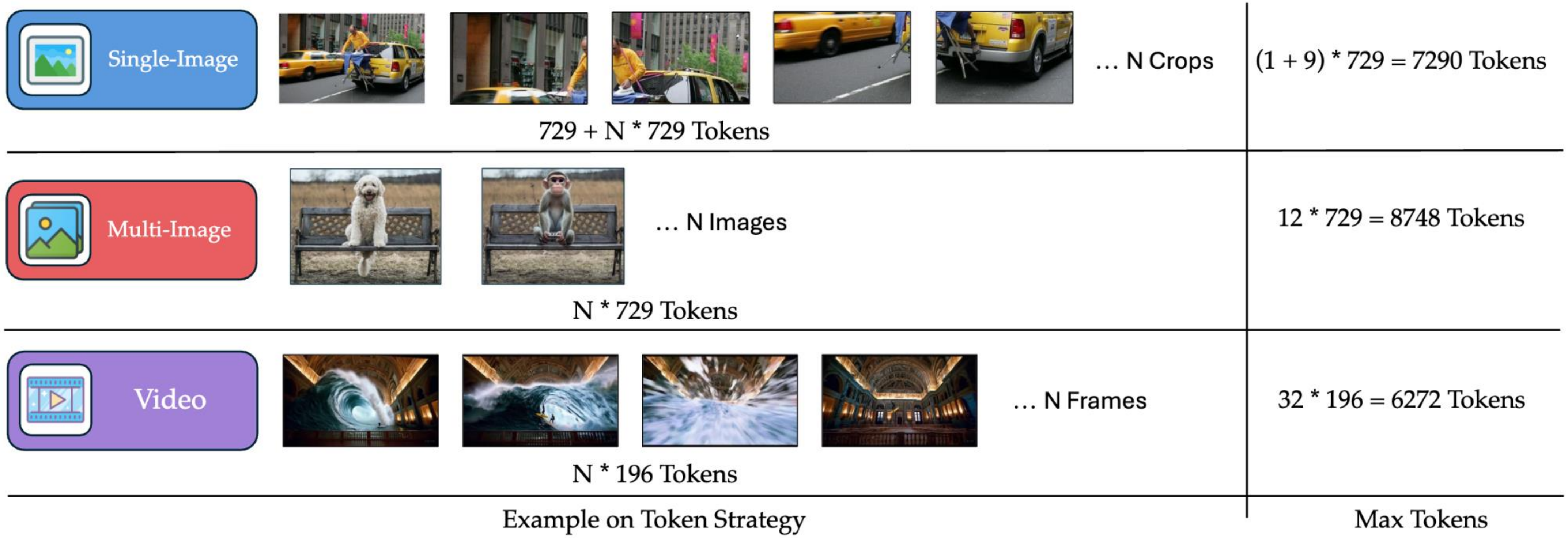
User	What is unusual about this image?
LLaVA	The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.
	[*Start a new conversation, and clear the history*]
User	What's happening in the scene?
LLaVA	The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention due to his unconventional choice of ironing his clothes on top of a moving car. The city street around him is bustling with activity, adding to the unique nature of the scene.

LLaVA-OneVision



https://llava-vl.github.io/blog/2024-08-05-llava-onevision/

LLaVA-OneVision



Single-Image

… N Crops

729 + N * 729 Tokens

(1 + 9) * 729 = 7290 Tokens

Multi-Image

… N Images

N * 729 Tokens

12 * 729 = 8748 Tokens

Video

… N Frames

N * 196 Tokens

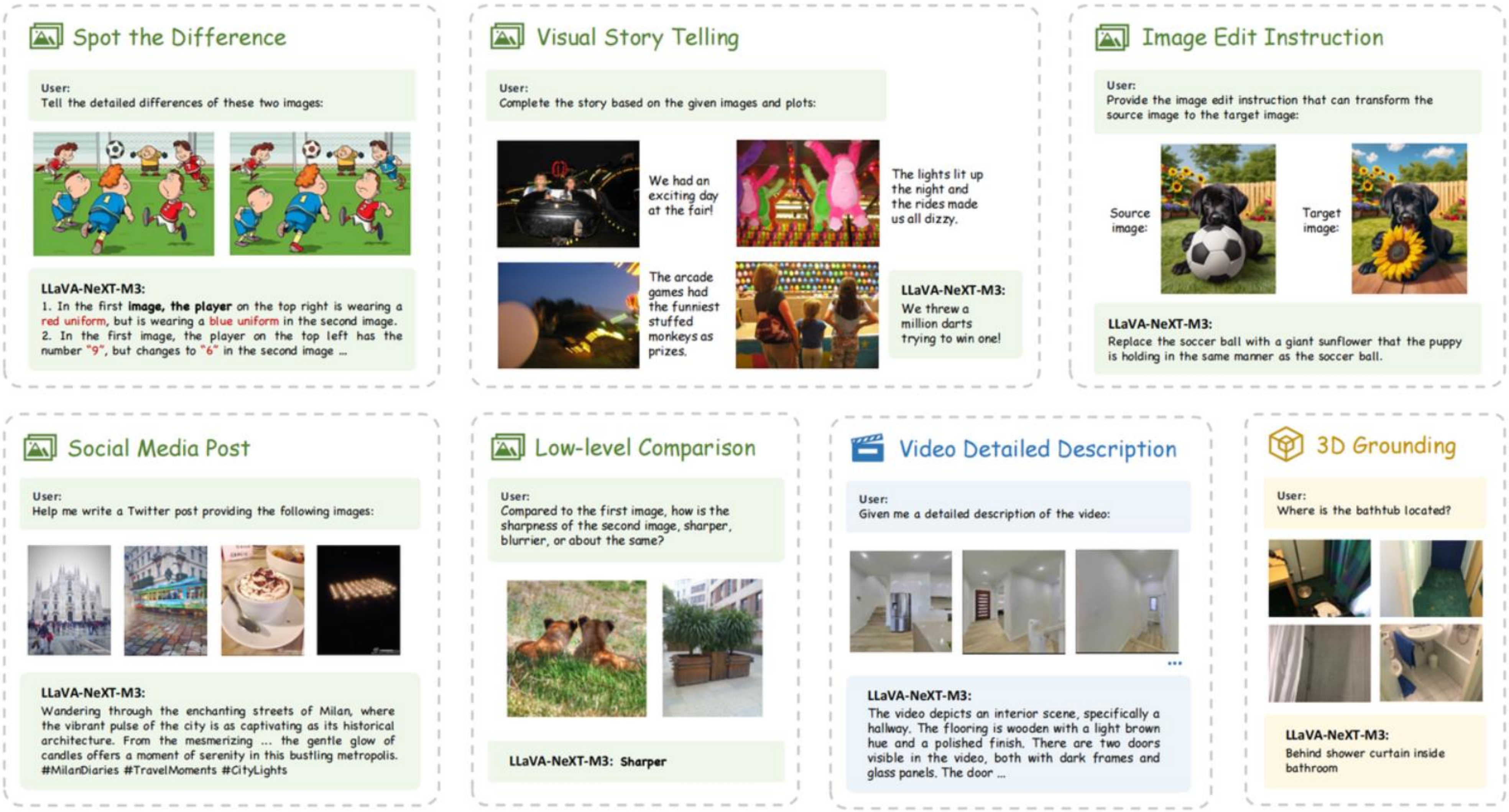
32 * 196 = 6272 Tokens

Example on Token Strategy

Max Tokens

https://llava-vl.github.io/blog/2024-08-05-llava-onevision/

LLaVA-NeXT-Interleave



Spot the Difference

User:
Tell the detailed differences of these two images:

LLaVA-NeXT-M3:
1. In the first **image, the player** on the top right is wearing a red uniform, but is wearing a blue uniform in the second image.
2. In the first image, the player on the top left has the number "9", but changes to "6" in the second image ...

Visual Story Telling

User:
Complete the story based on the given images and plots:

We had an exciting day at the fair!

The lights lit up the night and the rides made us all dizzy.

The arcade games had the funniest stuffed monkeys as prizes.

LLaVA-NeXT-M3:
We threw a million darts trying to win one!

Image Edit Instruction

User:
Provide the image edit instruction that can transform the source image to the target image:

LLaVA-NeXT-M3:
Replace the soccer ball with a giant sunflower that the puppy is holding in the same manner as the soccer ball.

Social Media Post

User:
Help me write a Twitter post providing the following images:

LLaVA-NeXT-M3:
Wandering through the enchanting streets of Milan, where the vibrant pulse of the city is as captivating as its historical architecture. From the mesmerizing ... the gentle glow of candles offers a moment of serenity in this bustling metropolis. #MilanDiaries #TravelMoments #CityLights

Low-level Comparison

User:
Compared to the first image, how is the sharpness of the second image, sharper, blurrier, or about the same?

LLaVA-NeXT-M3: **Sharper**

Video Detailed Description

User:
Given me a detailed description of the video:

LLaVA-NeXT-M3:
The video depicts an interior scene, specifically a hallway. The flooring is wooden with a light brown hue and a polished finish. There are two doors visible in the video, both with dark frames and glass panels. The door ...

3D Grounding

User:
Where is the bathtub located?

LLaVA-NeXT-M3:
Behind shower curtain inside bathroom

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

How Reliable Vision-Language Models Are?

- ARO (Attribution, Relation and Order) Benchmark

Visual Genome Relation

Assessing relational understanding (23,937 test cases)

✔ *the horse is eating the grass*
X *the grass is eating the horse*

Visual Genome Attribution

Assessing attributive understanding (28,748 test cases)

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

When and why vision-language models behave like bags-of-words, and what to do about it?

ARO (Attribution, Relation and Order) Benchmark

Visual Genome Relation

Assessing relational understanding (23,937 test cases)

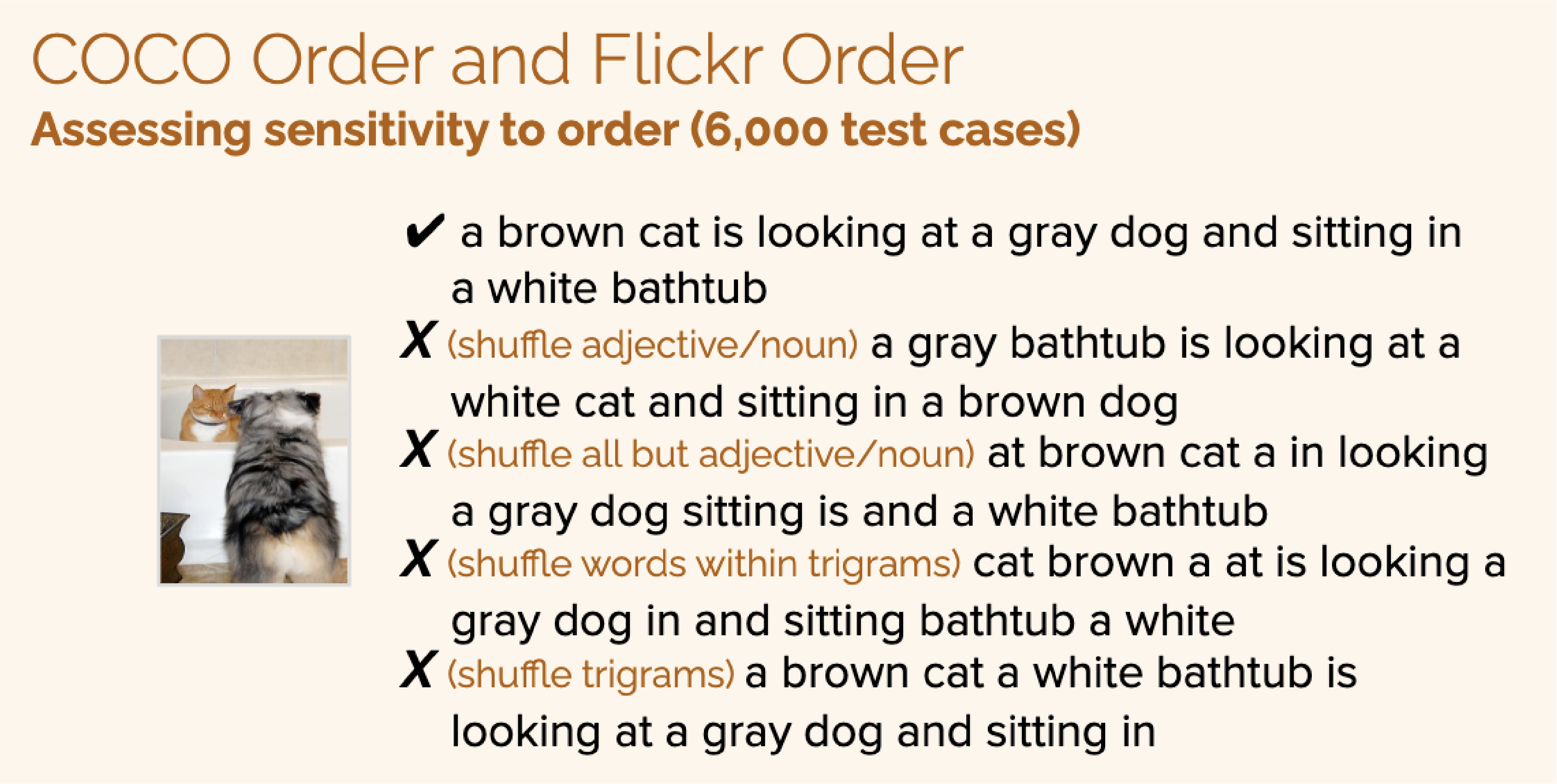
✔ *the horse is eating the grass*
X *the grass is eating the horse*

Visual Genome Attribution

Assessing attributive understanding (28,748 test cases)

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

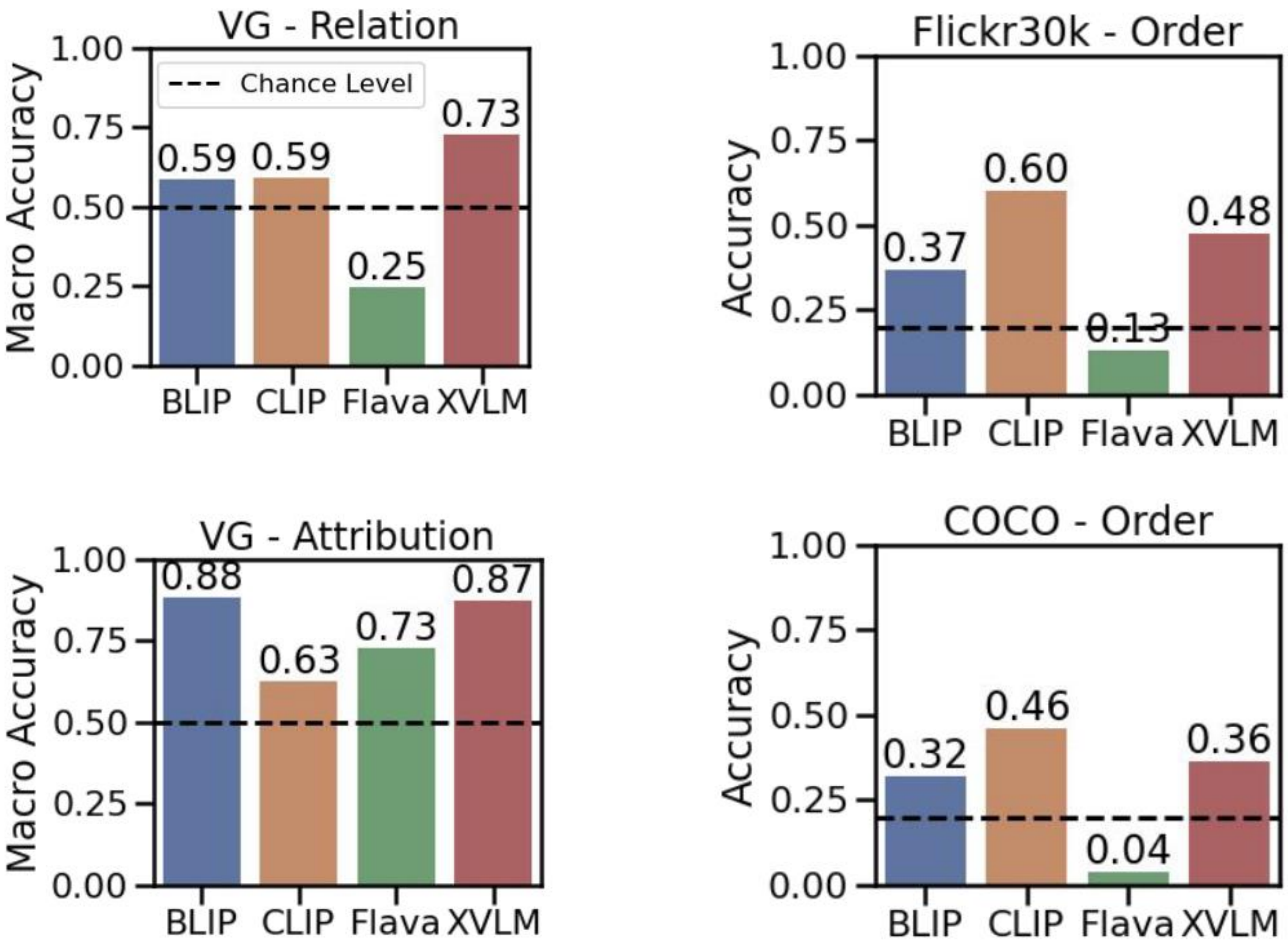
✗ (shuffle adjective/noun) a gray bathtub is looking at a white cat and sitting in a brown dog

✗ (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

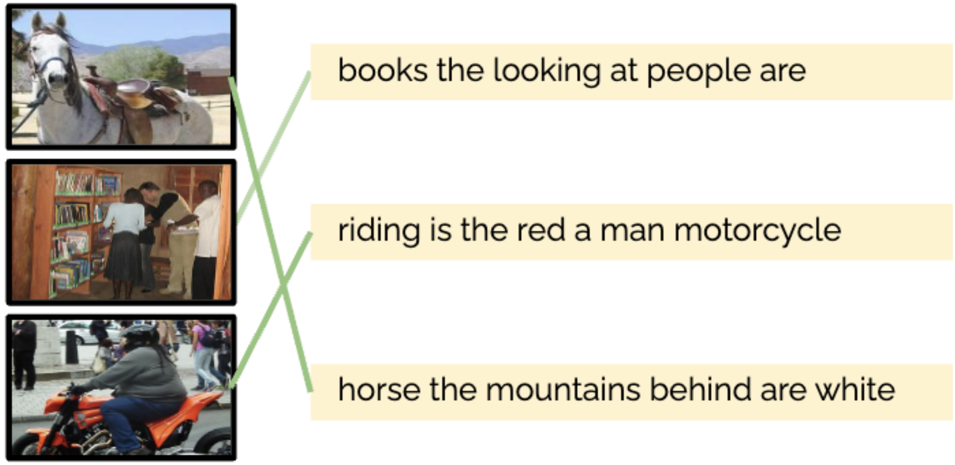
✗ (shuffle trigrams) a brown cat a white bathtub is looking at a gray dog and sitting in

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

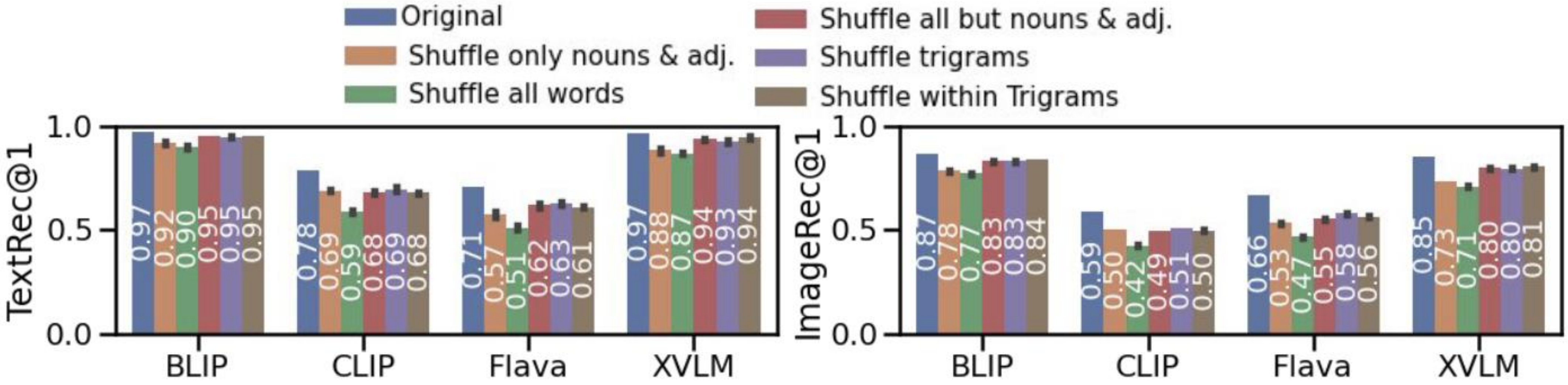
Retrieval without access to word order



books the looking at people are

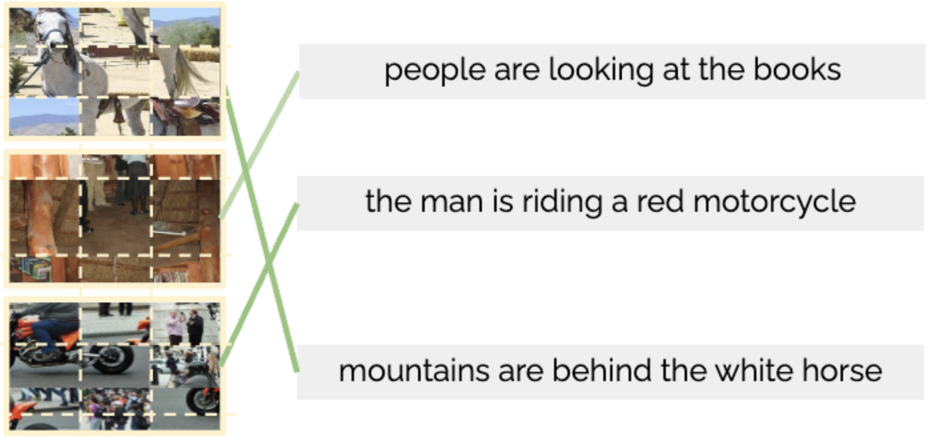
riding is the red a man motorcycle

horse the mountains behind are white



More Analysis on Text-Image Retrieval

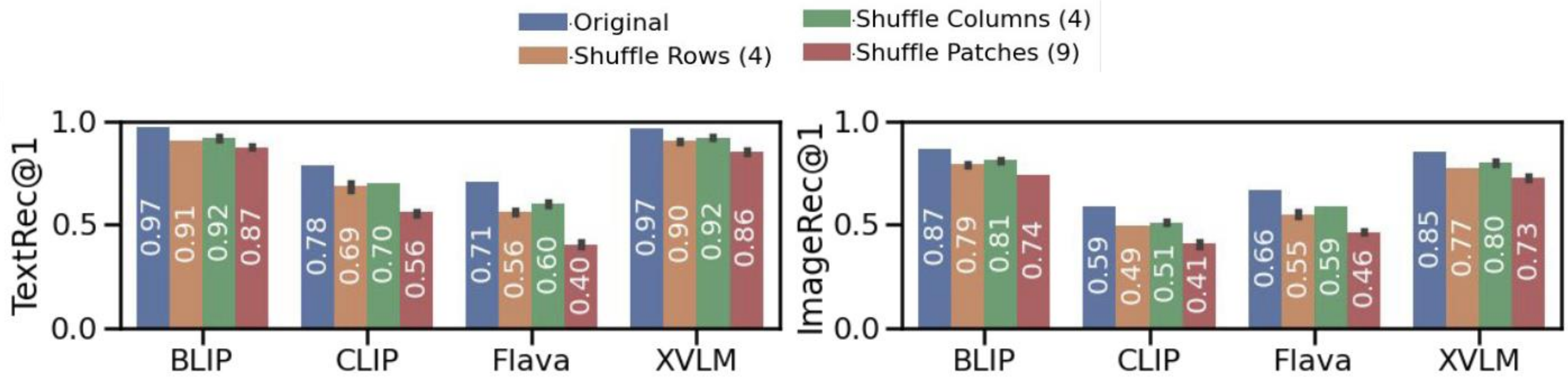
Retrieval without access to visual patch order



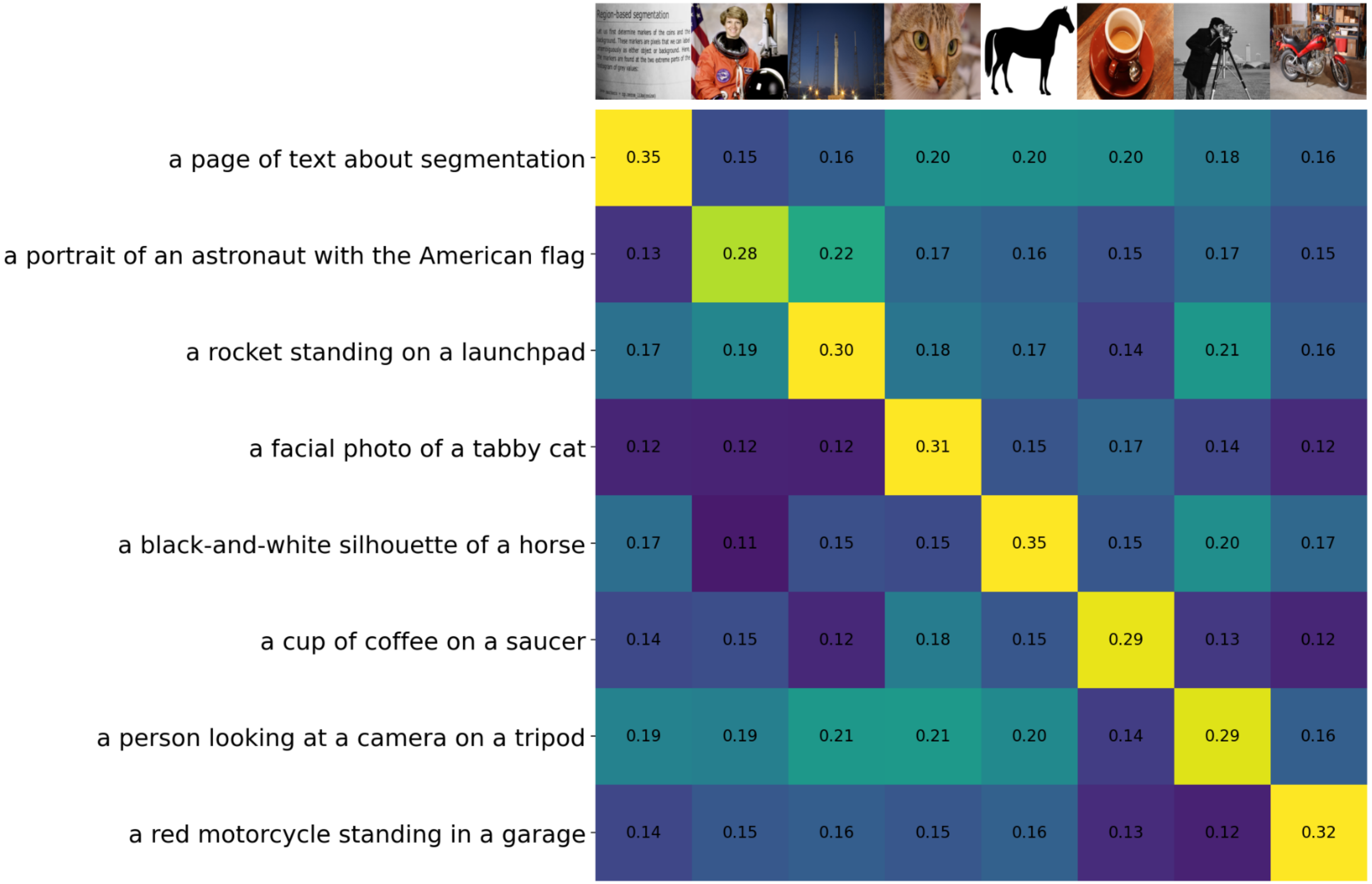
people are looking at the books

the man is riding a red motorcycle

mountains are behind the white horse



Contrastive Pre-Training



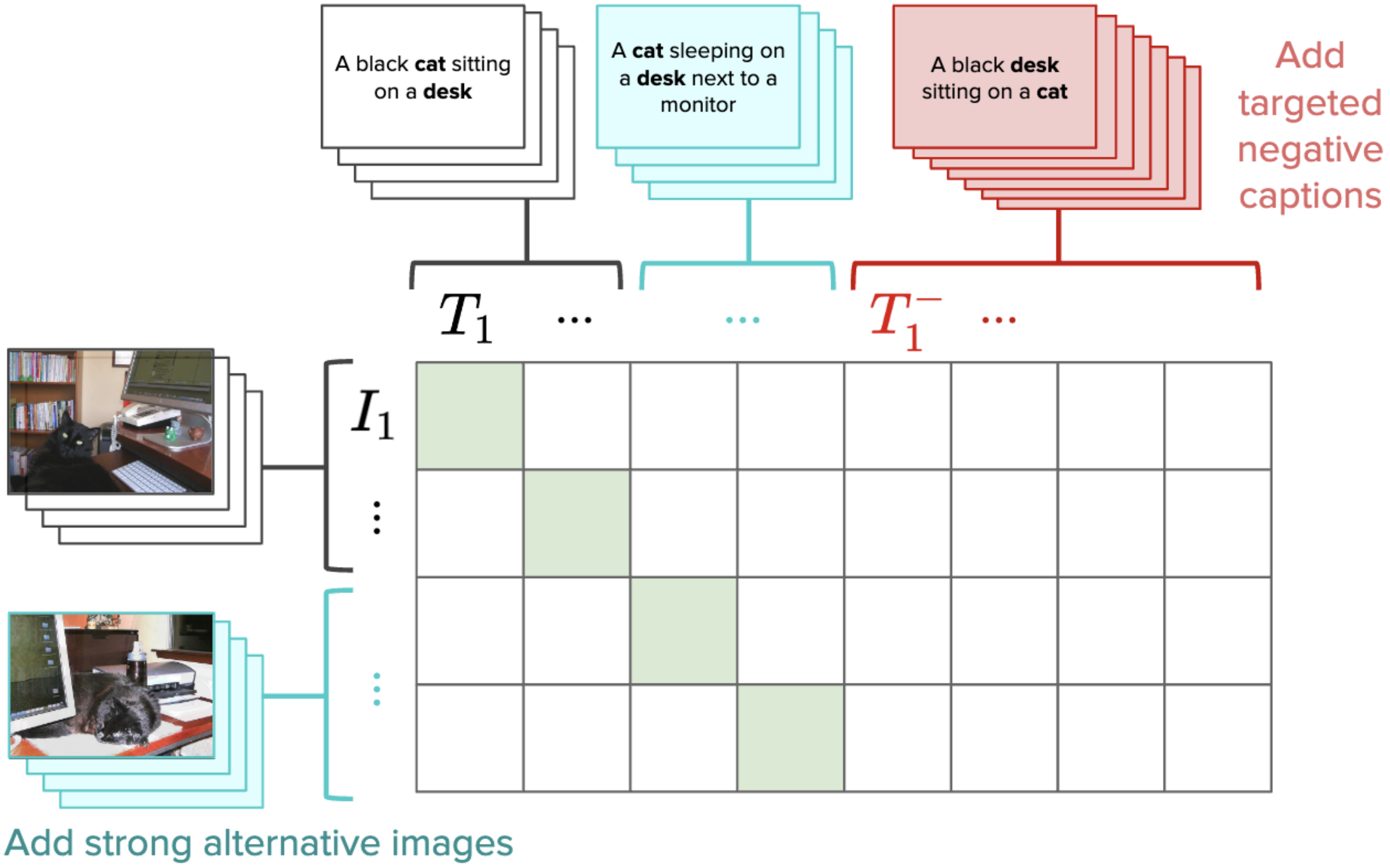
Cosine similarity between text and image features

a page of text about segmentation	0.35	0.15	0.16	0.20	0.20	0.20	0.18	0.16
a portrait of an astronaut with the American flag	0.13	0.28	0.22	0.17	0.16	0.15	0.17	0.15
a rocket standing on a launchpad	0.17	0.19	0.30	0.18	0.17	0.14	0.21	0.16
a facial photo of a tabby cat	0.12	0.12	0.12	0.31	0.15	0.17	0.14	0.12
a black-and-white silhouette of a horse	0.17	0.11	0.15	0.15	0.35	0.15	0.20	0.17
a cup of coffee on a saucer	0.14	0.15	0.12	0.18	0.15	0.29	0.13	0.12
a person looking at a camera on a tripod	0.19	0.19	0.21	0.21	0.20	0.14	0.29	0.16
a red motorcycle standing in a garage	0.14	0.15	0.16	0.15	0.16	0.13	0.12	0.32

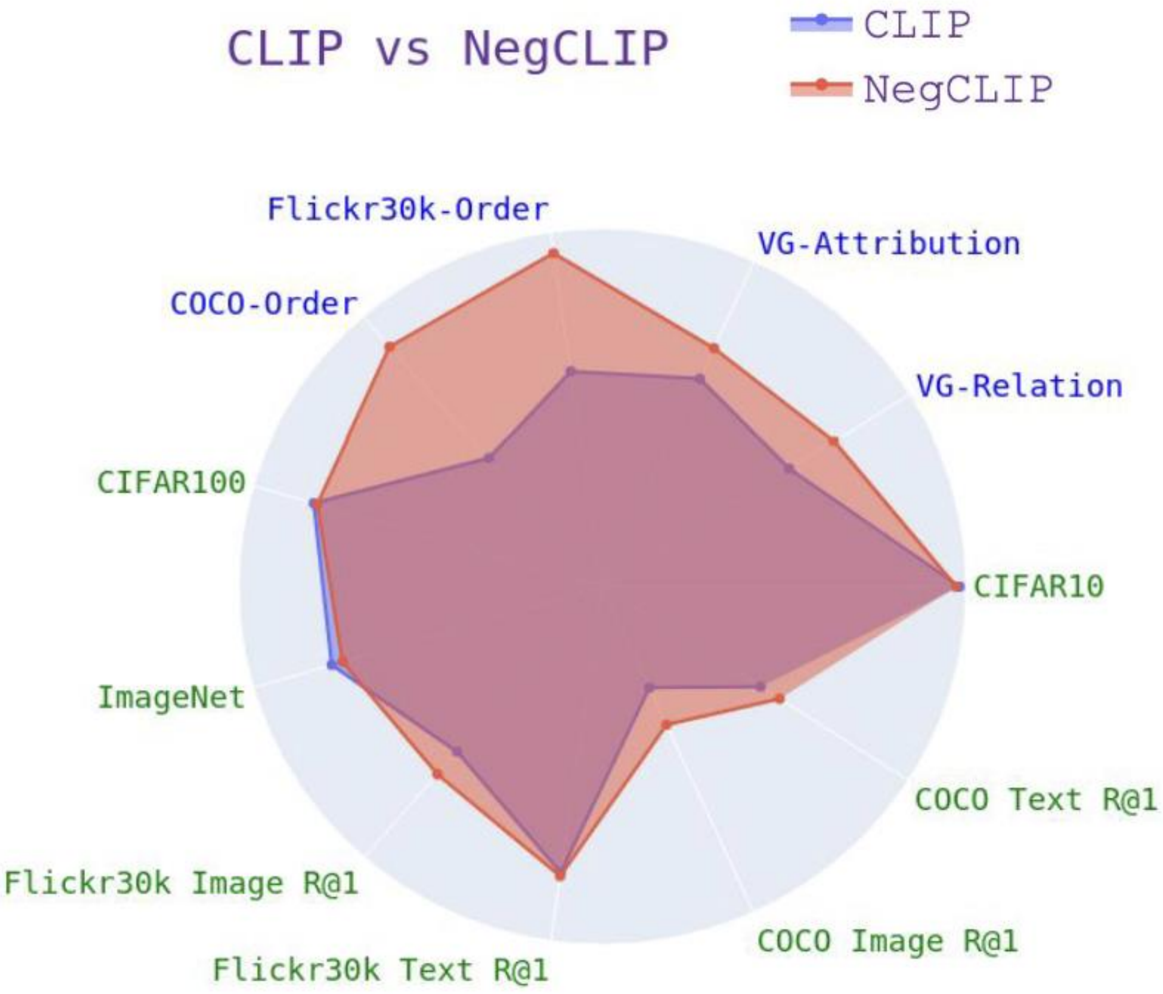
Solution: Composition-Aware Hard Negatives

- 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

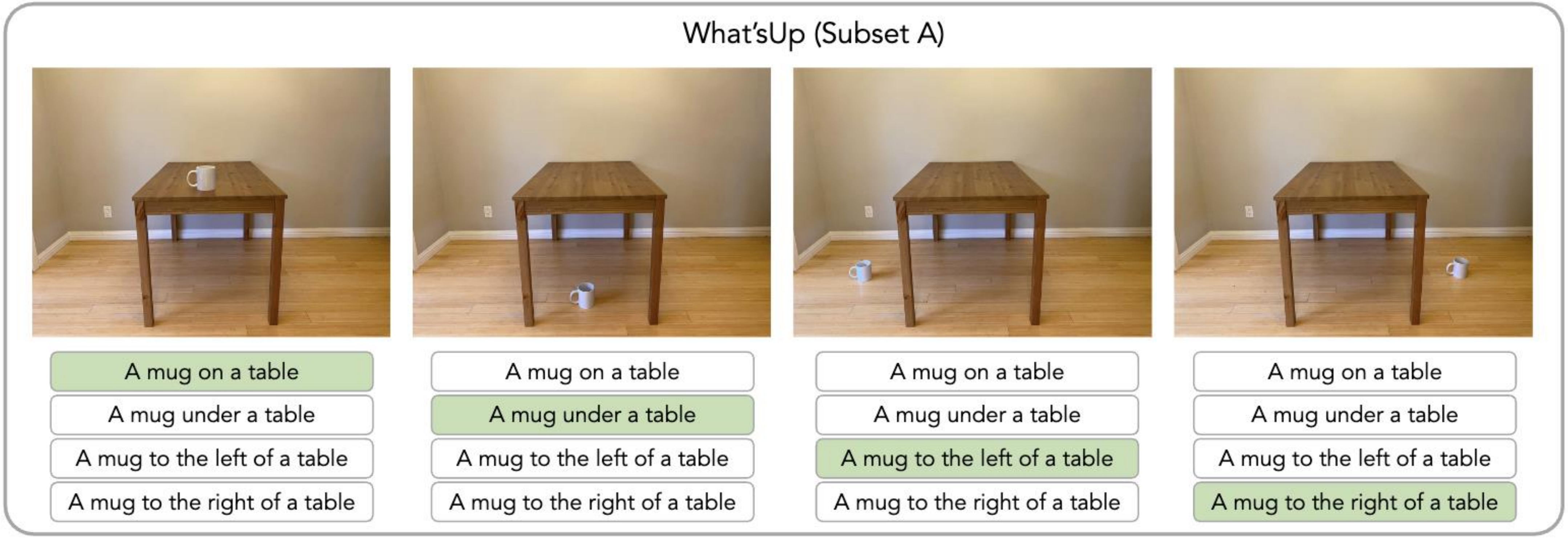
NegCLIP


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

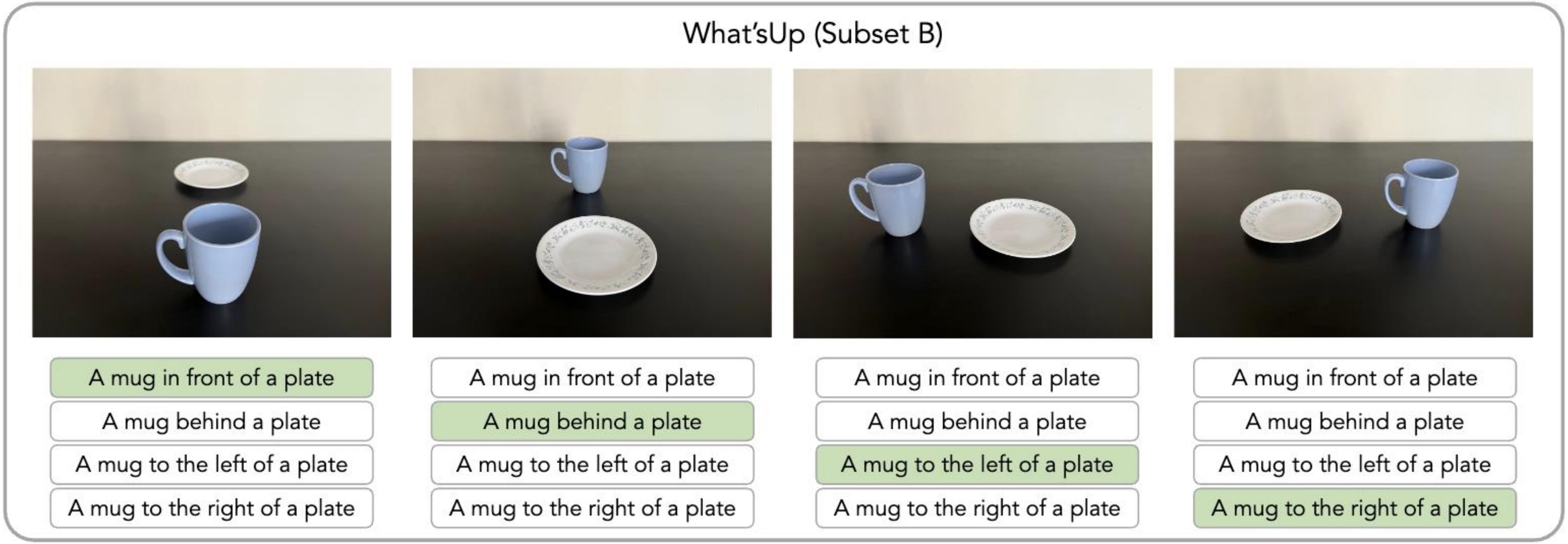

CLIP vs NegCLIP
CLIP
NegCLIP
Flickr30k-Order
VG-Attribution
VG-Relation
COCO-Order
CIFAR100
CIFAR10
ImageNet
COCO Text R@1
Flickr30k Image R@1
Flickr30k Text R@1
COCO Image R@1

What’s Up: Benchmarks with Spatial Reasoning Questions


What'sUp (Subset A)
A mug on a table
A mug under a table
A mug to the left of a table
A mug to the right of a table
A mug on a table
A mug under a table
A mug to the left of a table
A mug to the right of a table
A mug on a table
A mug under a table
A mug to the left of a table
A mug to the right of a table
A mug on a table
A mug under a table
A mug to the left of a table
A mug to the right of a table

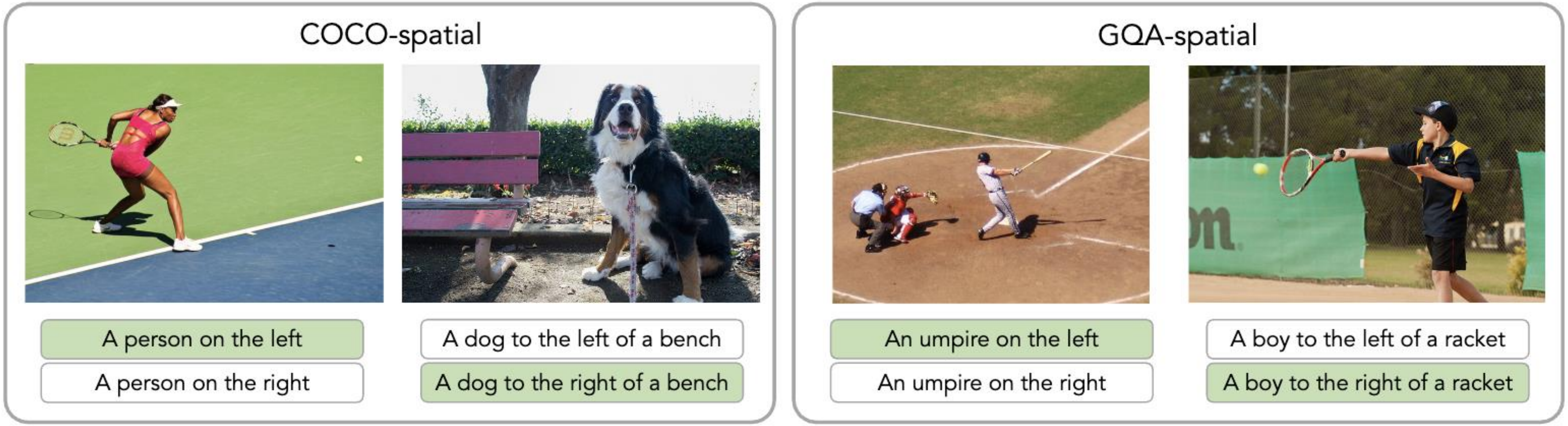
What's "up" with vision-language models? Investigating their struggle with spatial reasoning

What’s Up: Benchmarks with Spatial Reasoning Questions

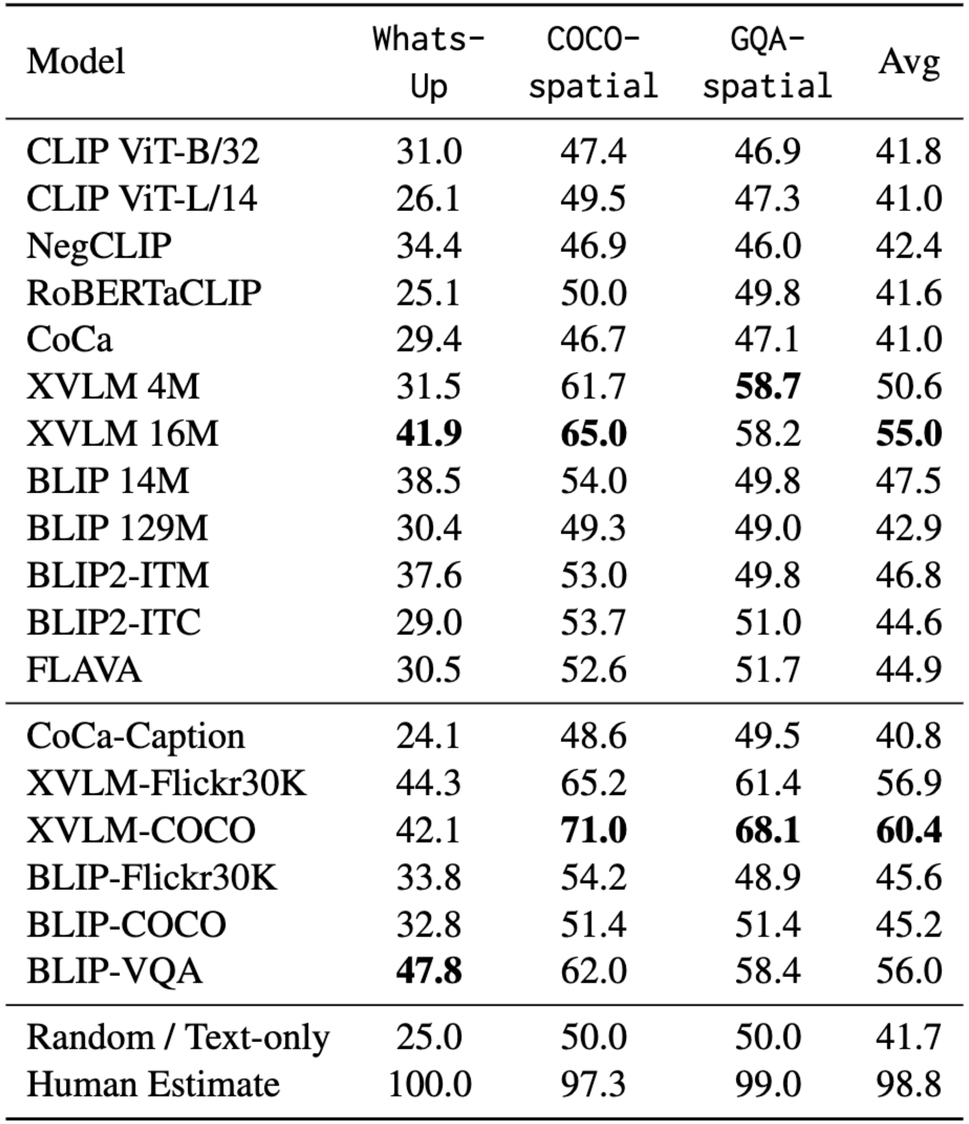

What’sUp (Subset B)

A mug in front of a plate
A mug behind a plate
A mug to the left of a plate
A mug to the right of a plate
A mug in front of a plate
A mug behind a plate
A mug to the left of a plate
A mug to the right of a plate
A mug in front of a plate
A mug behind a plate
A mug to the left of a plate
A mug to the right of a plate
A mug in front of a plate
A mug behind a plate
A mug to the left of a plate
A mug to the right of a plate

COCO-Spatial and GQA-Spatial


COCO-spatial
A person on the left
A person on the right
A dog to the left of a bench
A dog to the right of a bench
GQA-spatial
An umpire on the left
An umpire on the right
A boy to the left of a racket
A boy to the right of a racket

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%

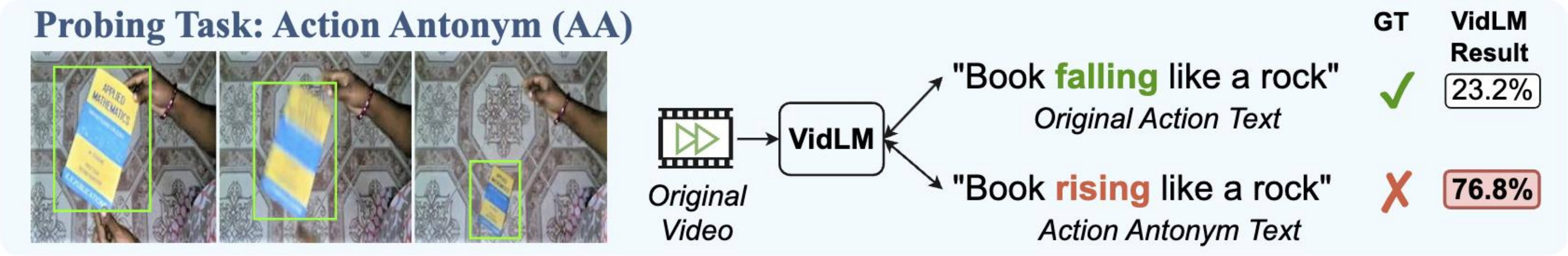
Visual Analogies

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

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

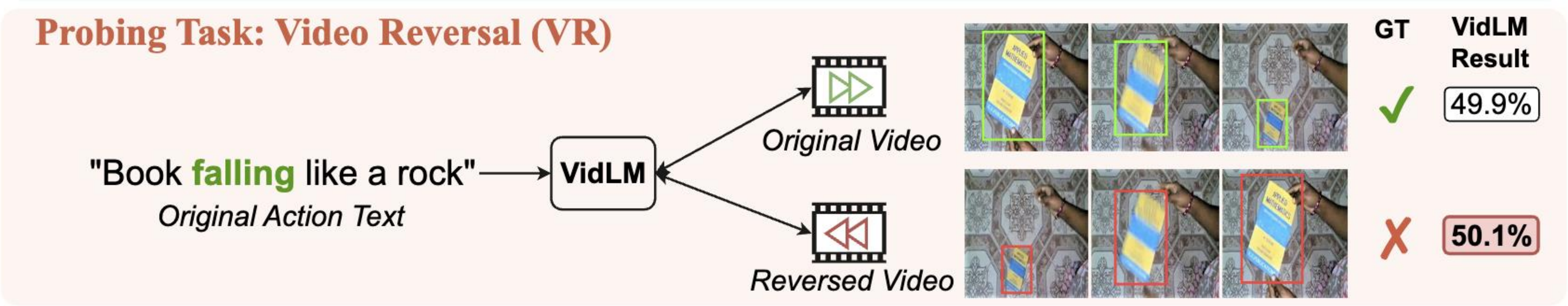
9%

Paxion: Action Robustness for Video-Language Models


Probing Task: Action Antonym (AA)
Original Video
VidLM
"Book falling like a rock"
Original Action Text
"Book rising like a rock"
Action Antonym Text
GT
VidLM Result
✓
23.2%
✗
76.8%

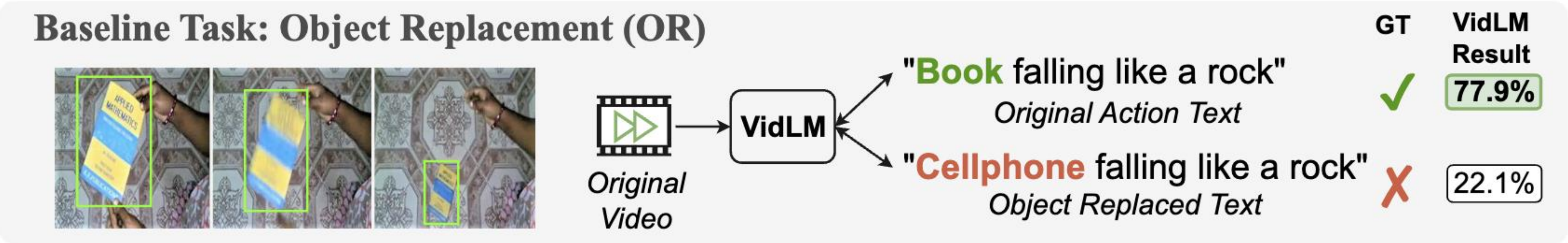
Paxion: Patching Action Knowledge in Video-Language Foundation Models

Paxion: Action Robustness for Video-Language Models



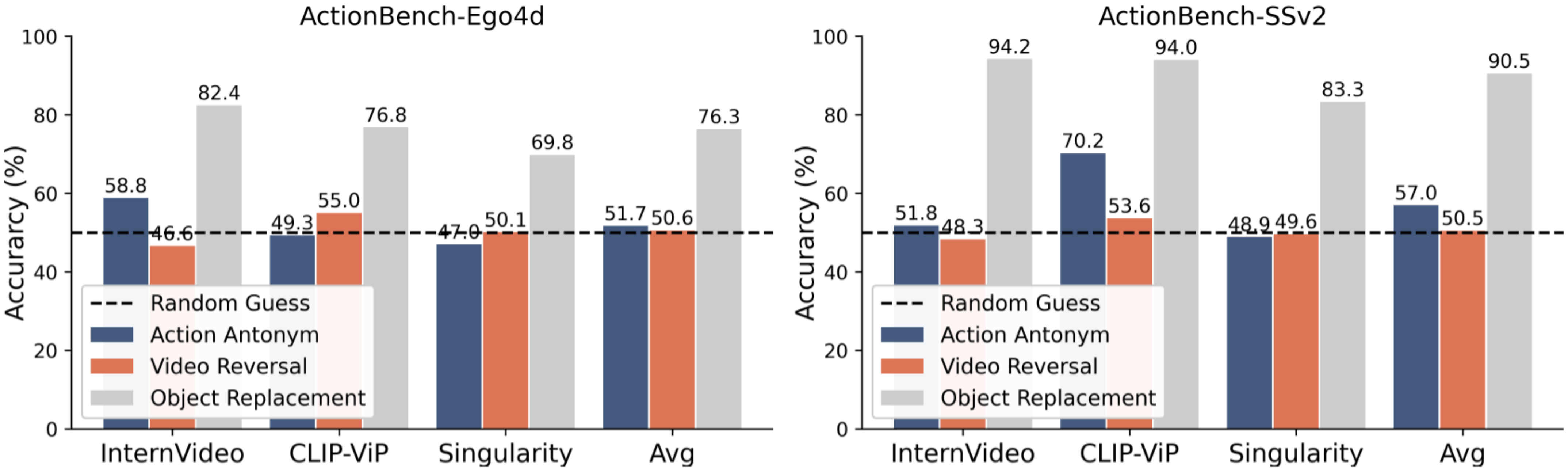
Paxion: Patching Action Knowledge in Video-Language Foundation Models

Paxion: Action Robustness for Video-Language Models

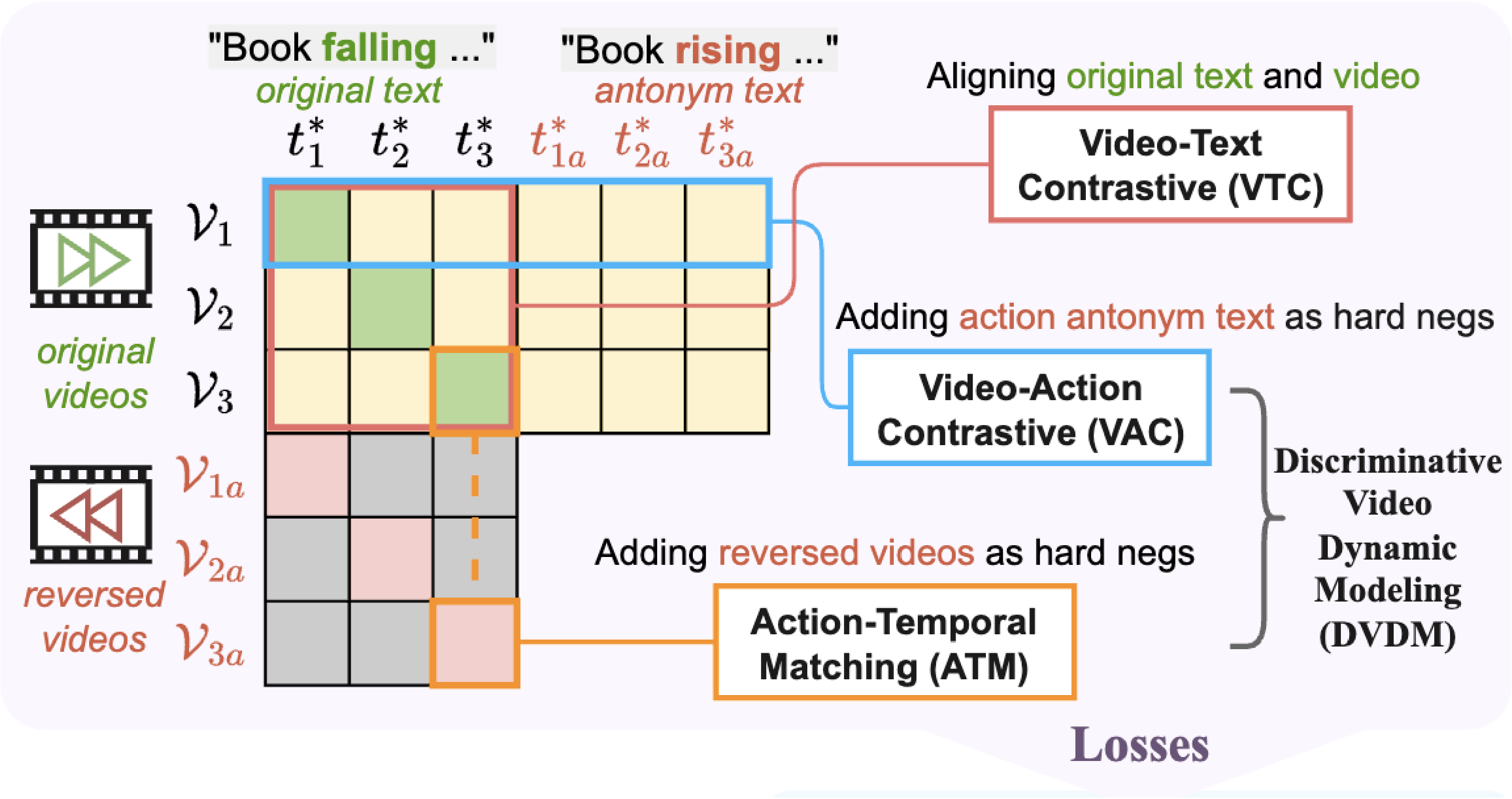

Baseline Task: Object Replacement (OR)
Original Video
VidLM
"Book falling like a rock"
Original Action Text
"Cellphone falling like a rock"
Object Replaced Text
GT
VidLM Result
✓
77.9%
✗
22.1%

Paxion: Patching Action Knowledge in Video-Language Foundation Models

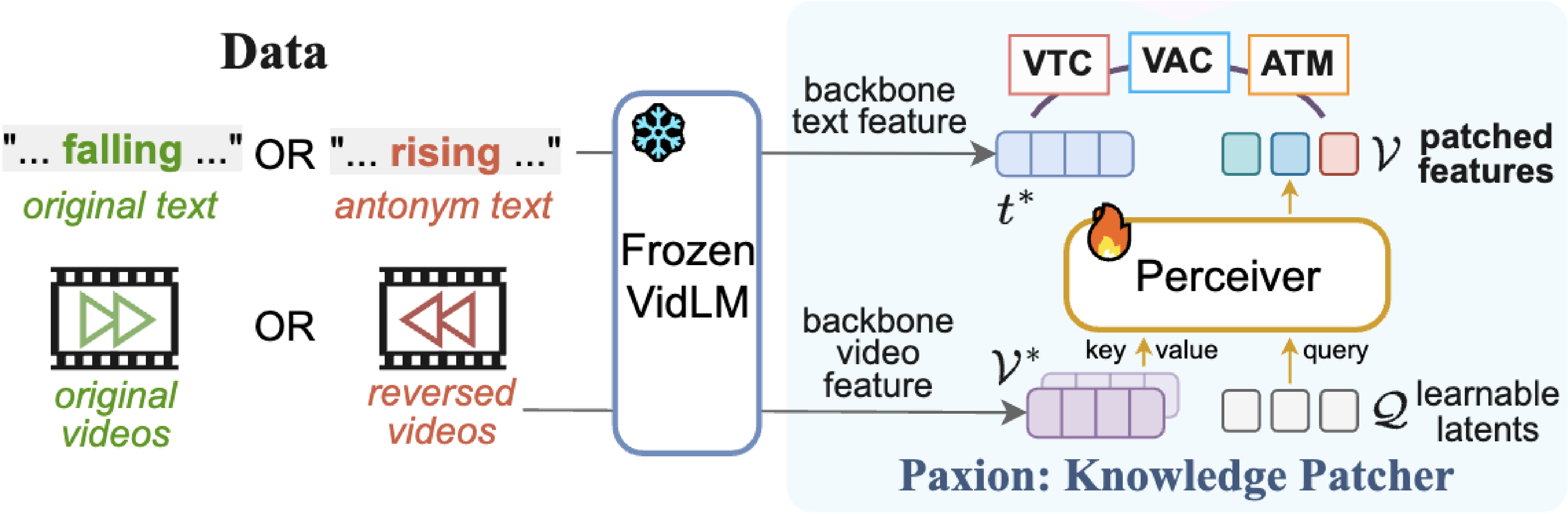
Results


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

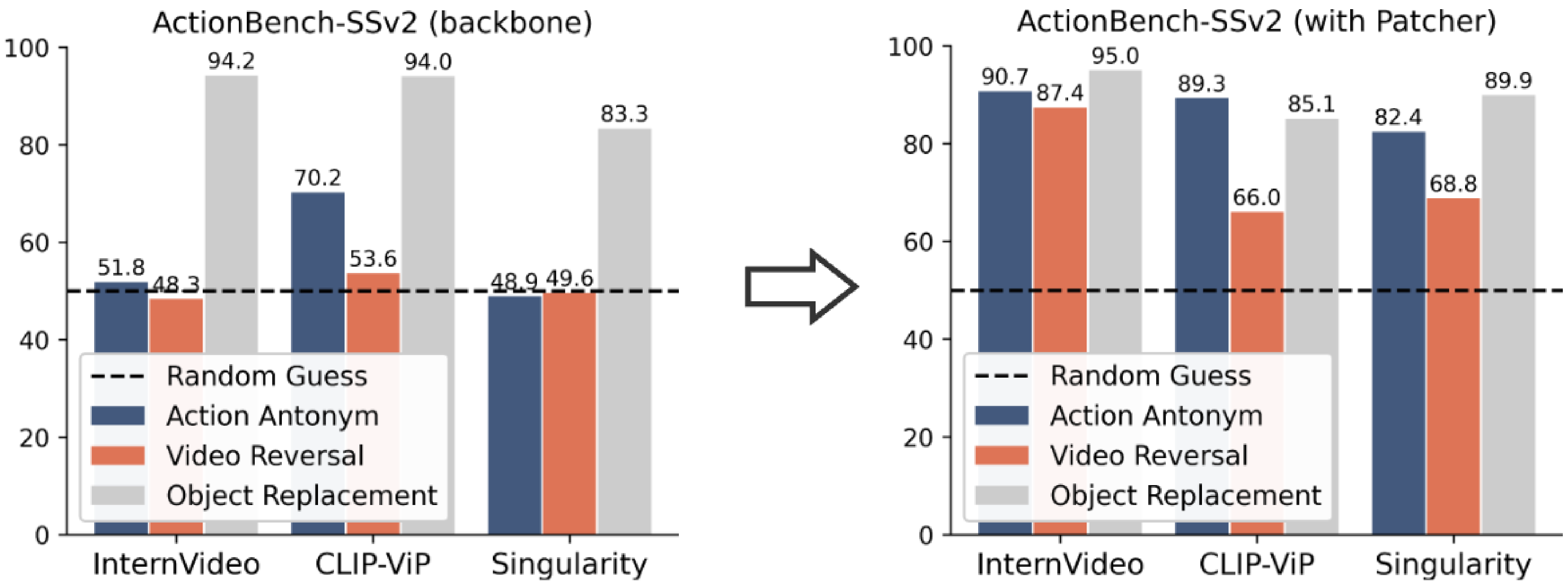
Paxion: Training Objective


"Book falling ..."
original text
"Book rising ..."
antonym text
t_1^* t_2^* t_3^* t_{1a}^* t_{2a}^* t_{3a}^*
$\mathcal{V}_1$
$\mathcal{V}_2$
$\mathcal{V}_3$
$\mathcal{V}_{1a}$
$\mathcal{V}_{2a}$
$\mathcal{V}_{3a}$
original videos
reversed videos
Aligning original text and video
Video-Text Contrastive (VTC)
Adding action antonym text as hard negs
Video-Action Contrastive (VAC)
Adding reversed videos as hard negs
Action-Temporal Matching (ATM)
Discriminative Video Dynamic Modeling (DVDM)
Losses

Paxion: Knowledge Patcher



Results


ActionBench-SSv2 (backbone)
100
80
60
40
20
0
94.2
94.0
83.3
70.2
51.8
48.3
53.6
48.9
49.6
Random Guess
Action Antonym
Video Reversal
Object Replacement
InternVideo
CLIP-ViP
Singularity
ActionBench-SSv2 (with Patcher)
100
80
60
40
20
0
90.7
87.4
95.0
89.3
66.0
85.1
82.4
68.8
89.9
Random Guess
Action Antonym
Video Reversal
Object Replacement
InternVideo
CLIP-ViP
Singularity