CSCE 638 Natural Language Processing Foundation and Techniques

Lecture 18: Vision-Language Models

Kuan-Hao Huang

Spring 2026

Invited Talk

- **Date:** 4/6 online @ Zoom
  - https://tamu.zoom.us/my/khhuang?pwd=oAdWOKVOCGPApqDbJnVtktdW2AE6nb.1
  - **Talk:** Improving Personalization and Consistency of Large Foundation Models
  - **Speaker:** Jindong Wang, Assistant Professor, The College of William & Mary
- **Date:** 4/15 online @ Zoom
  - https://tamu.zoom.us/my/khhuang?pwd=oAdWOKVOCGPApqDbJnVtktdW2AE6nb.1
  - **Talk:** Less Is More: Why Compression May Be the Missing Incentive for LLM Generalization
  - **Speaker:** Ben Zhou, Assistant Professor, Arizona State University

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.

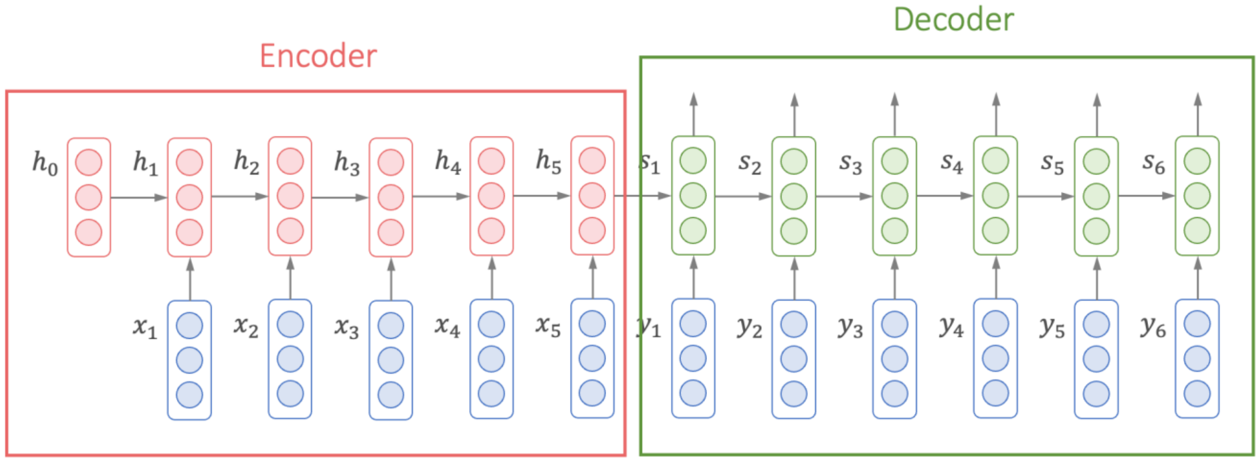
Two children play in the water.

A dog jumps over a hurdle.

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

Image Captioning with Encoder-Decoder Models

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

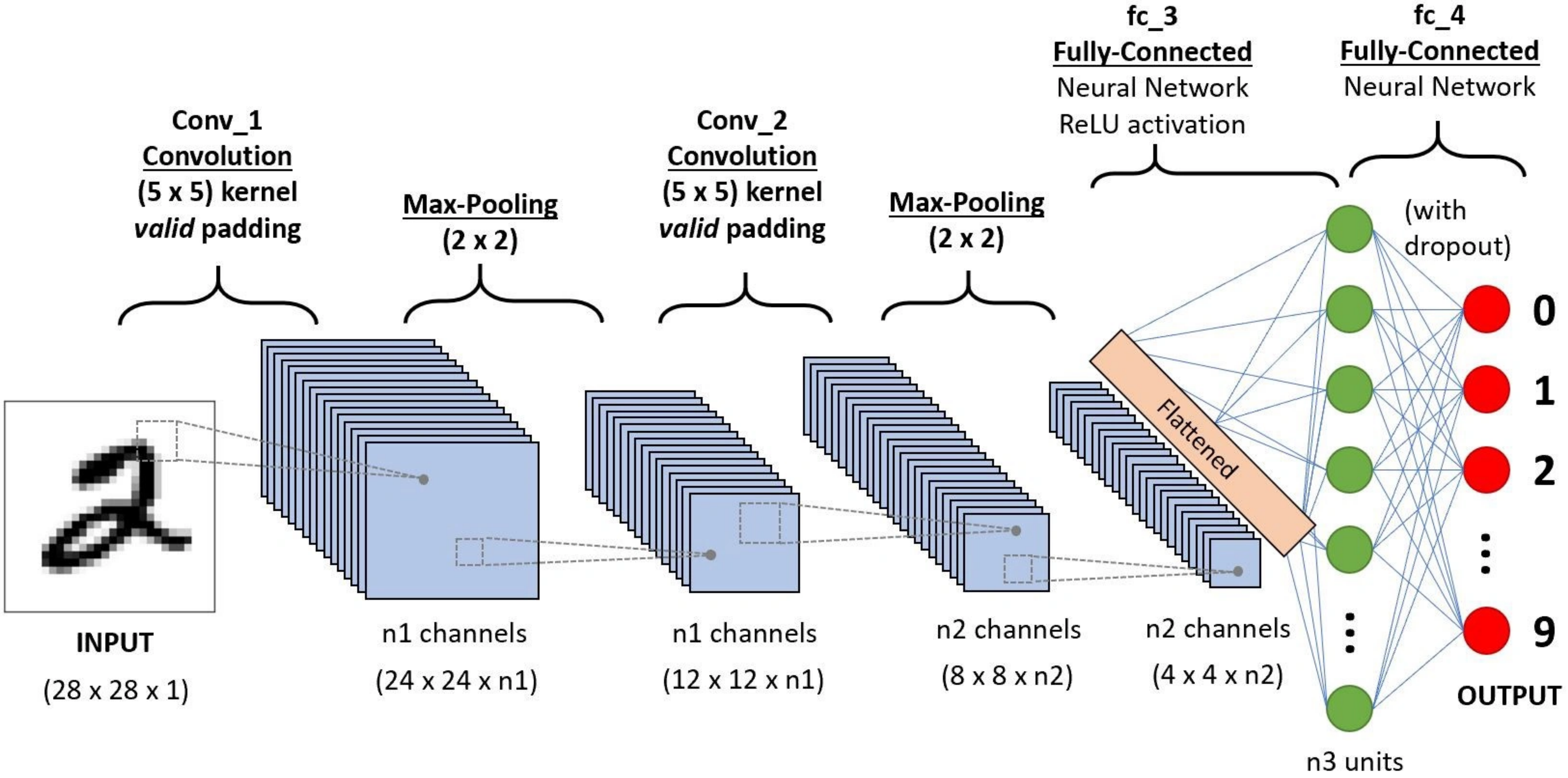
$w_{3,1}$*-5.6 + $w_{3,2}$*1.5 + $w_{3,3}$*-1.6

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

$$v_{1,1} + v_{1,2} + v_{1,3} + v_{1,4} \rightarrow v_1$$

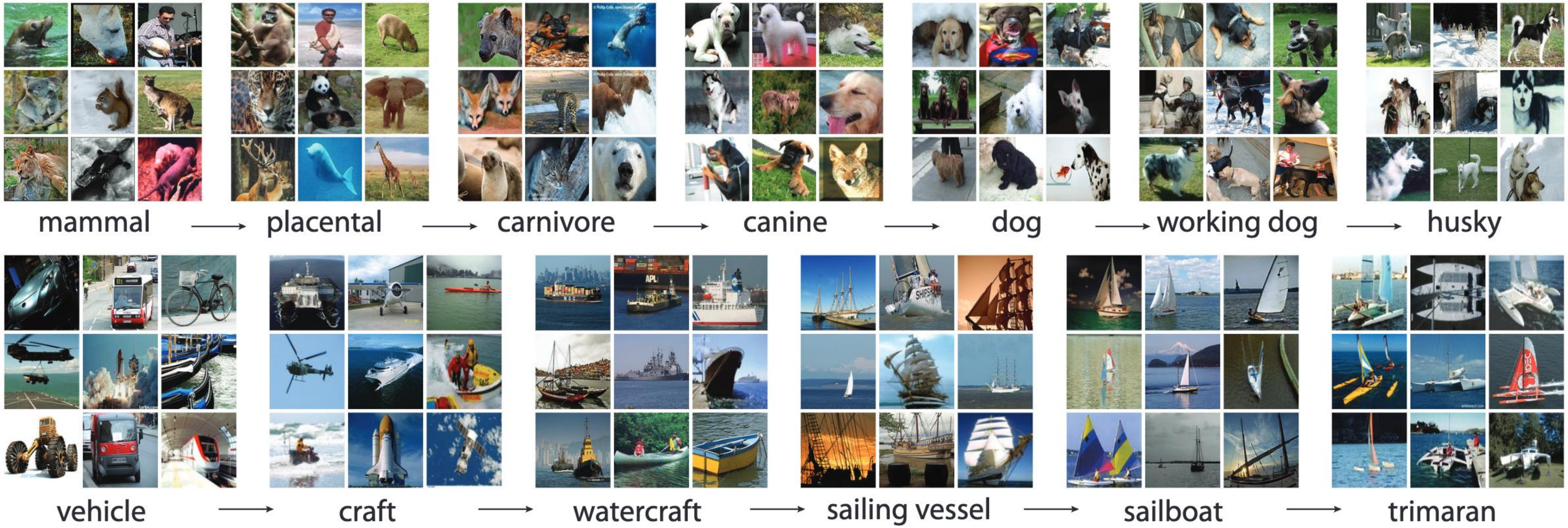
$$v_{2,1} + v_{2,2} + v_{2,3} + v_{2,4} \rightarrow v_2$$

Convolutional Neural Network (For Image)



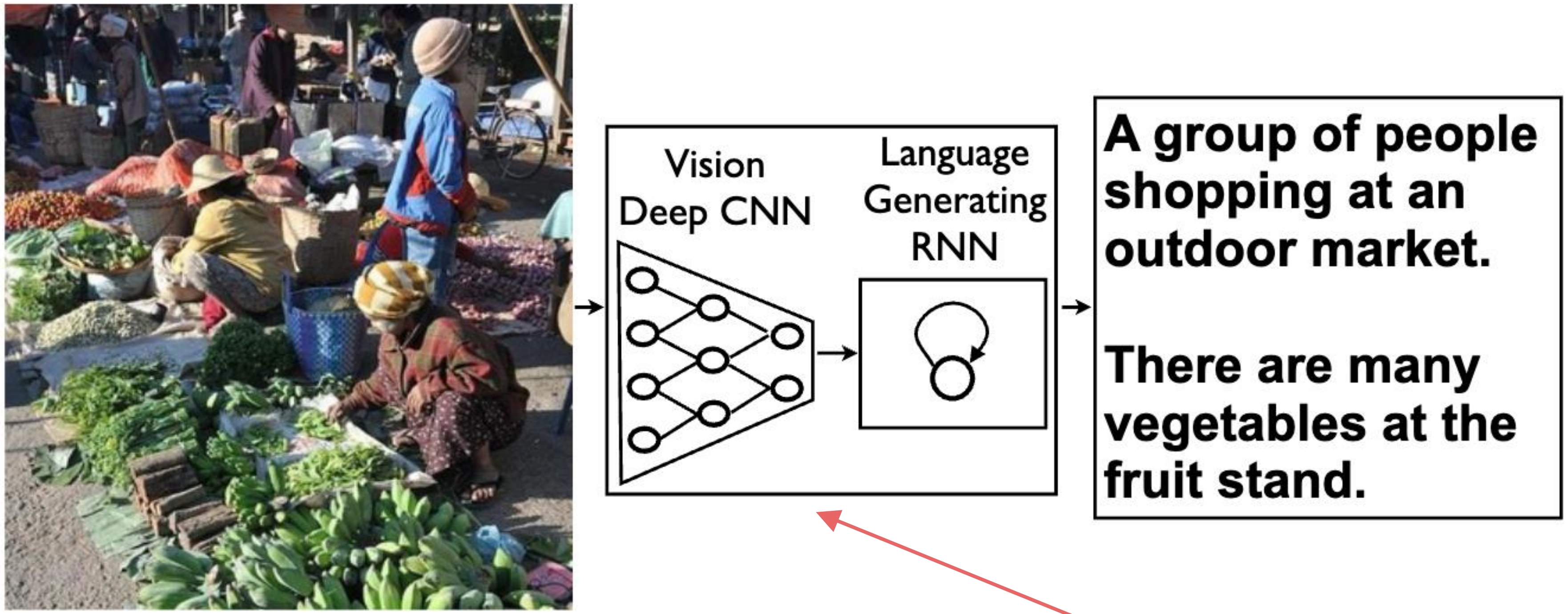
https://www.analyticsvidhya.com/blog/2020/10/what-is-the-convolutional-neural-network-architecture/

Pre-Trained CNN with ImageNet



ImageNet: A Large-Scale Hierarchical Image Database

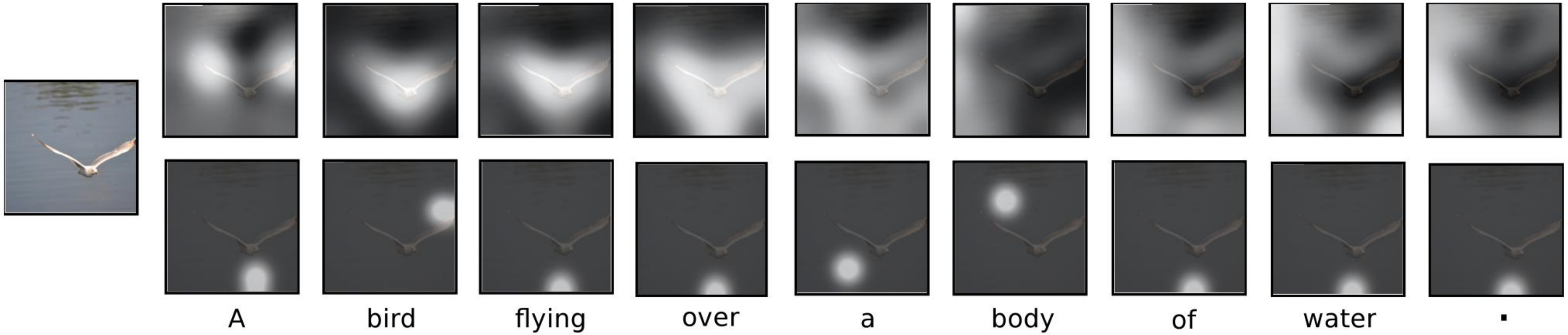
Encoder-Decoder: CNN-RNN

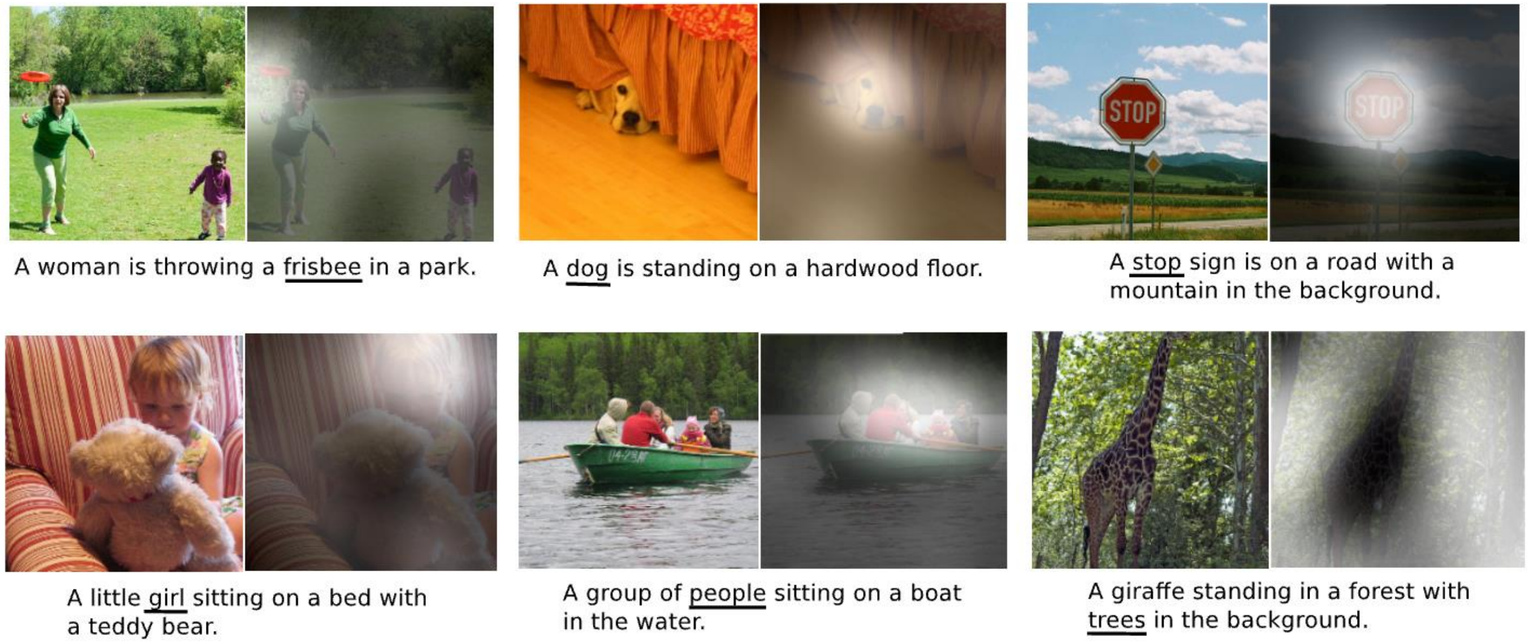


Text embedding space and image embedding space can be aligned!

Show and Tell: A Neural Image Caption Generator

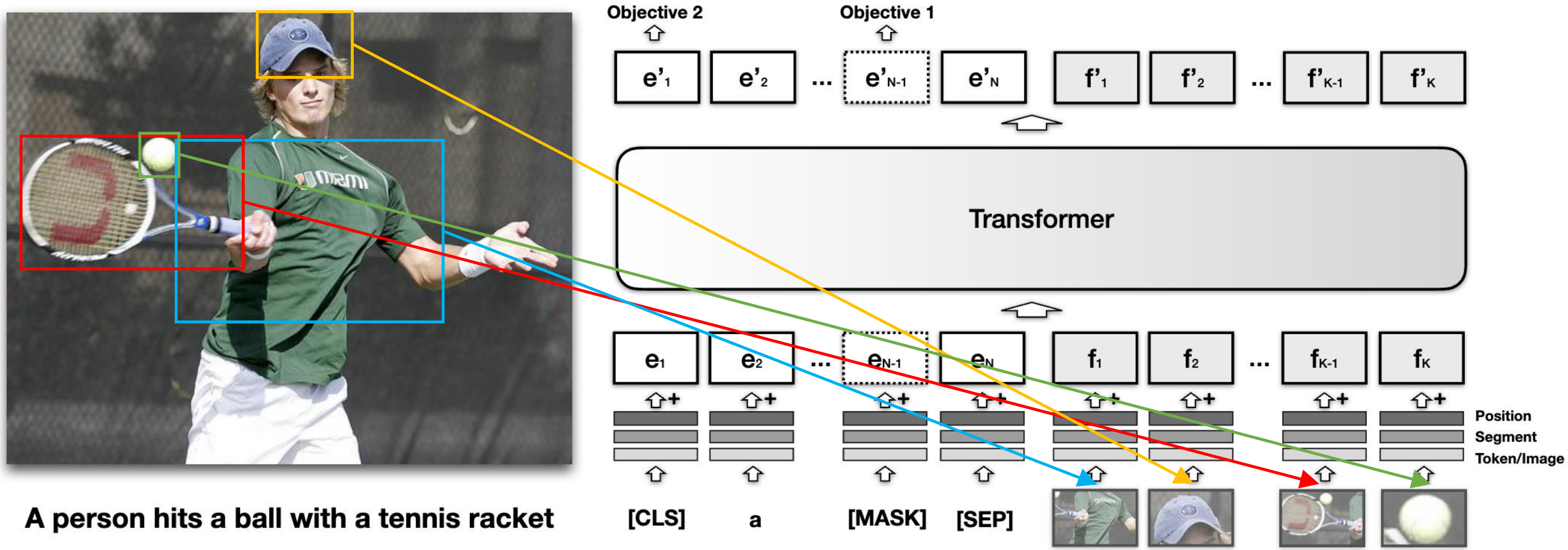
CNN + Attention LSTM





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

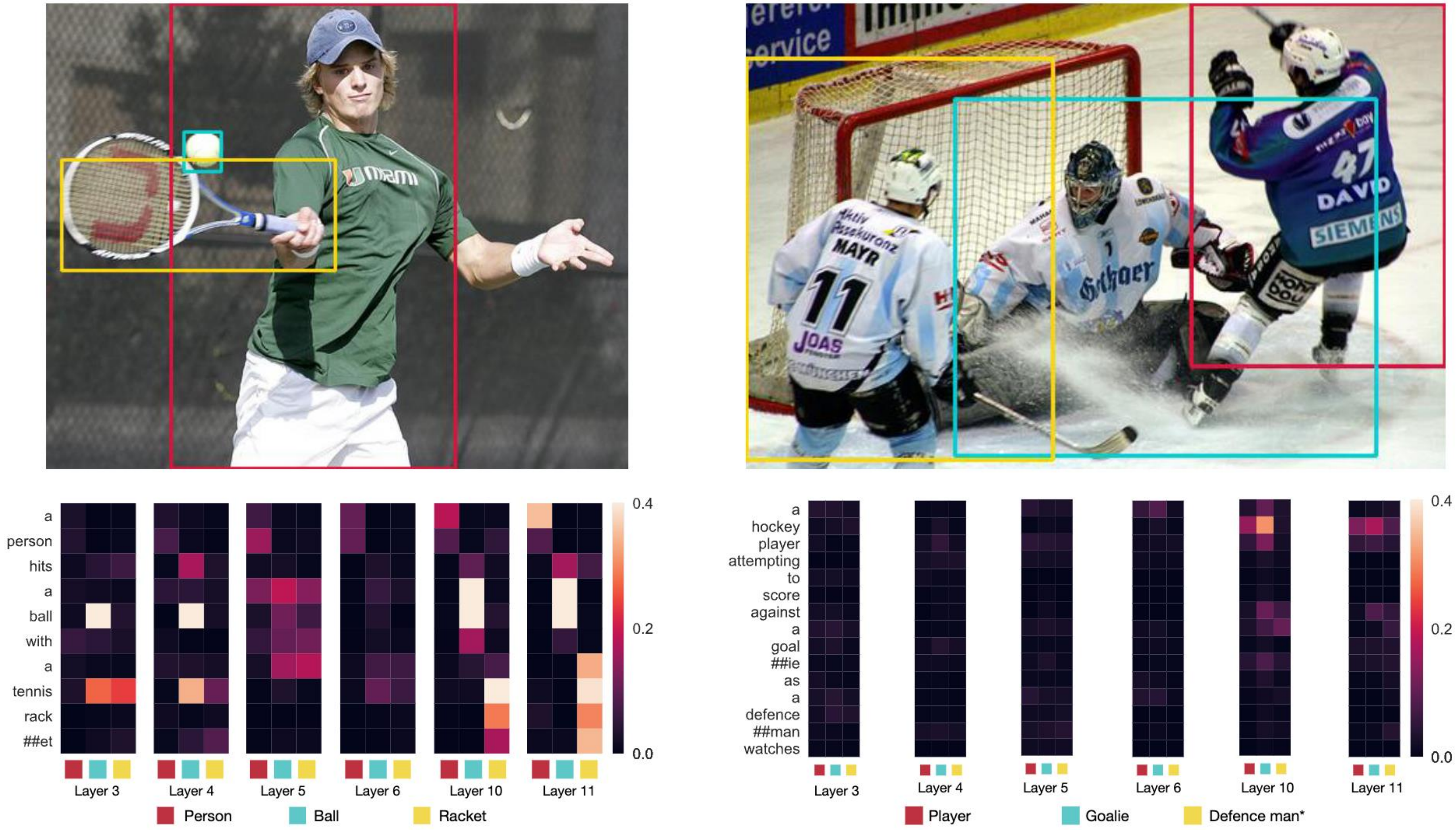
Joint Visual and Textual Embeddings: VisualBERT



Require an object detection model

VisualBERT: A Simple and Performant Baseline for Vision and Language

Joint Visual and Textual Embeddings: VisualBERT



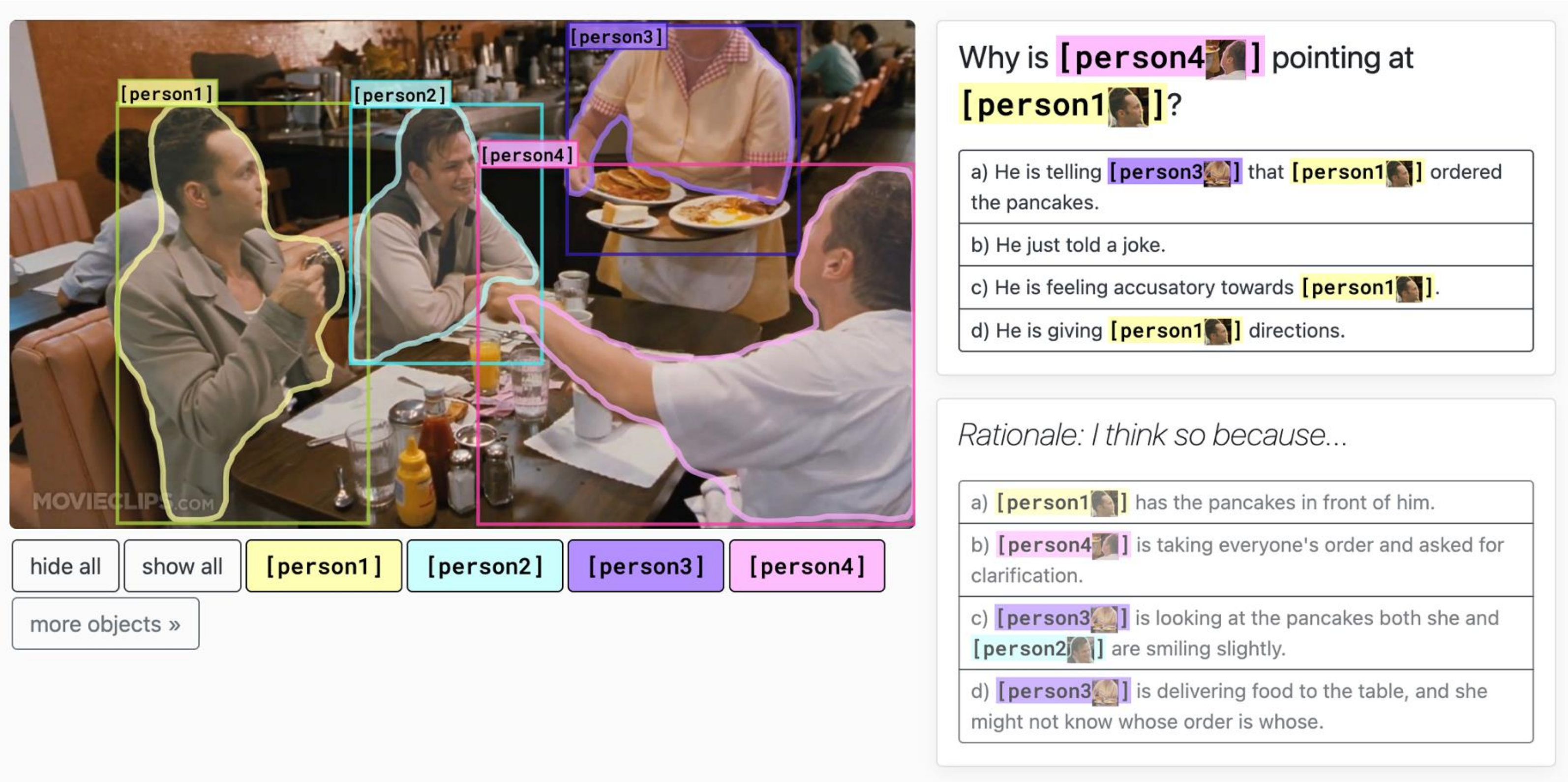
VisualBERT: A Simple and Performant Baseline for Vision and Language

Visual Question Answering

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



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

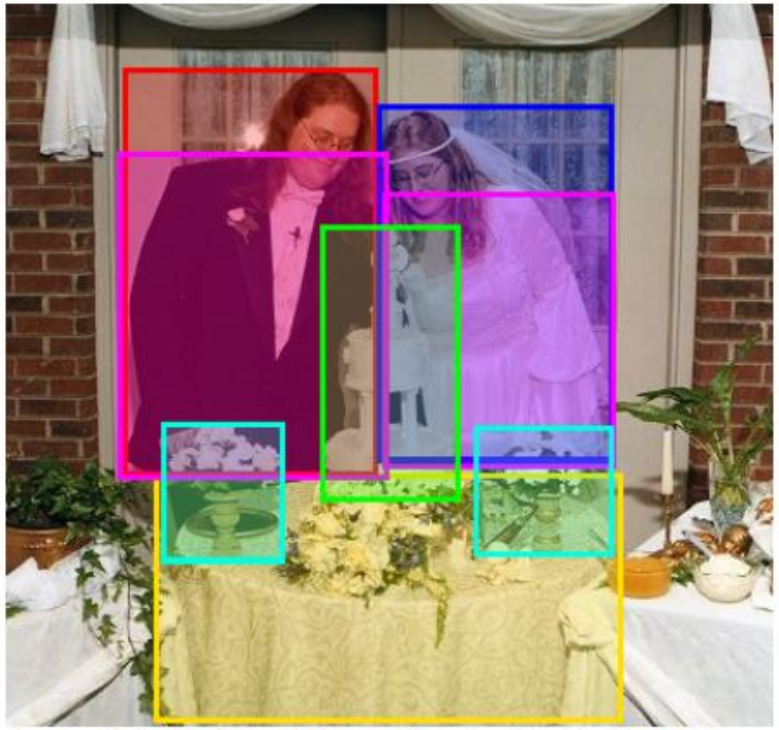
The left image contains twice the number of dogs as the right image, and at least two dogs in total are standing.

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

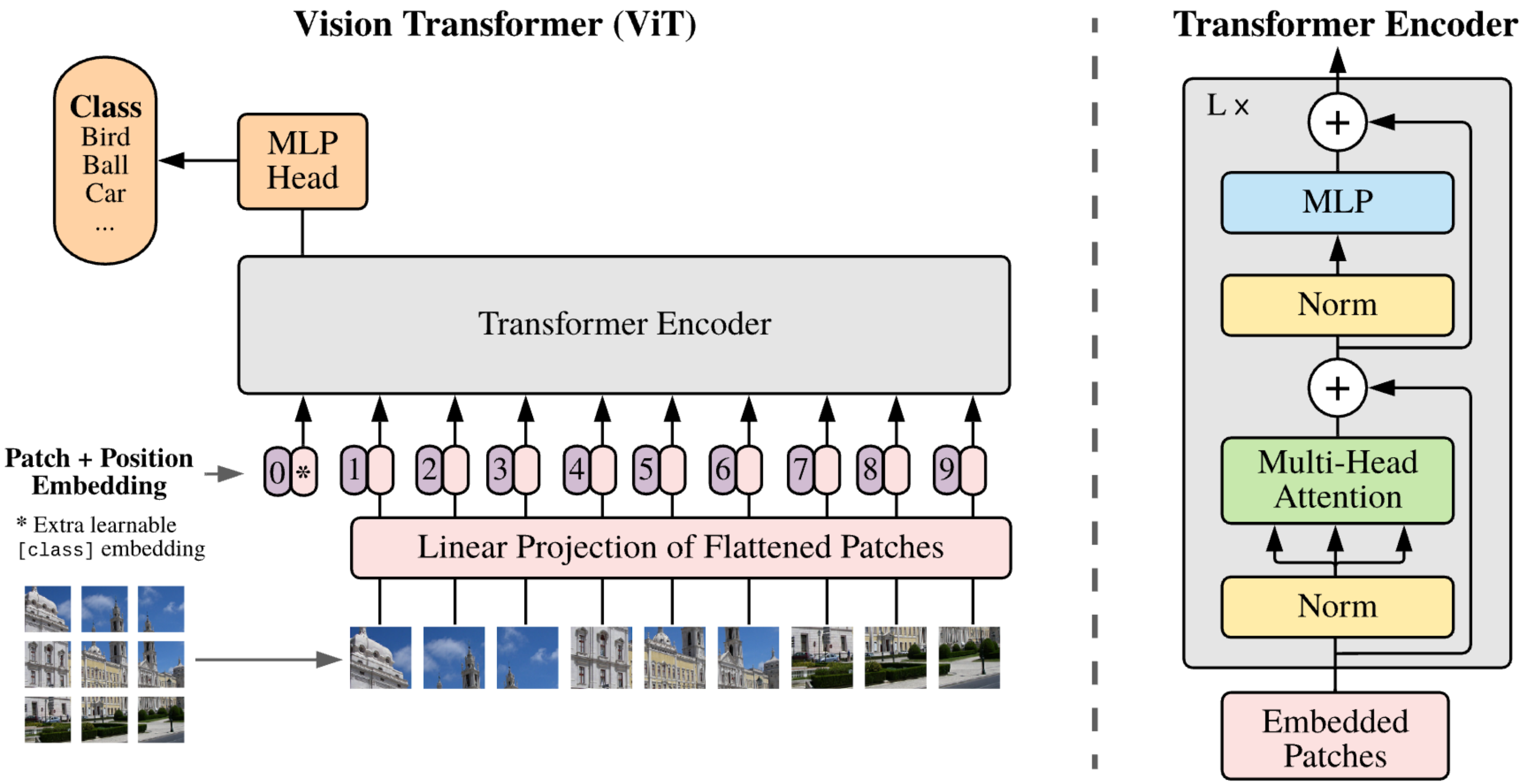


A couple in their wedding attire stand behind a table with a wedding cake and flowers.
A bride and groom are standing in front of their wedding cake at their reception.
A bride and groom smile as they view their wedding cake at a reception.
A couple stands behind their wedding cake.
Man and woman cutting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

Model	R@1 Dev	R@1 Test	R@5 Dev	R@5 Test	R@10 Dev	R@10 Test	Upper Bound Dev	Upper Bound 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		

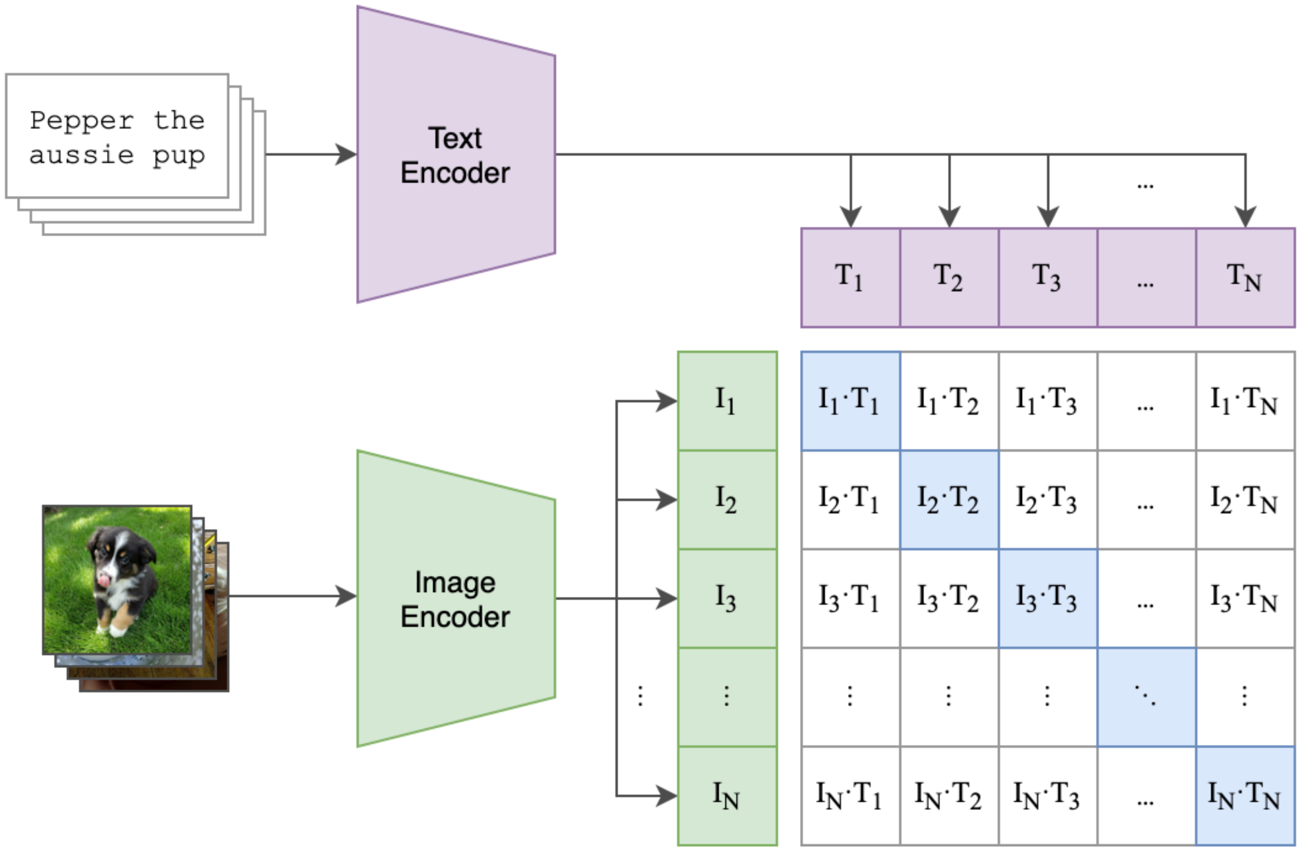
Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models

Vision Transformer



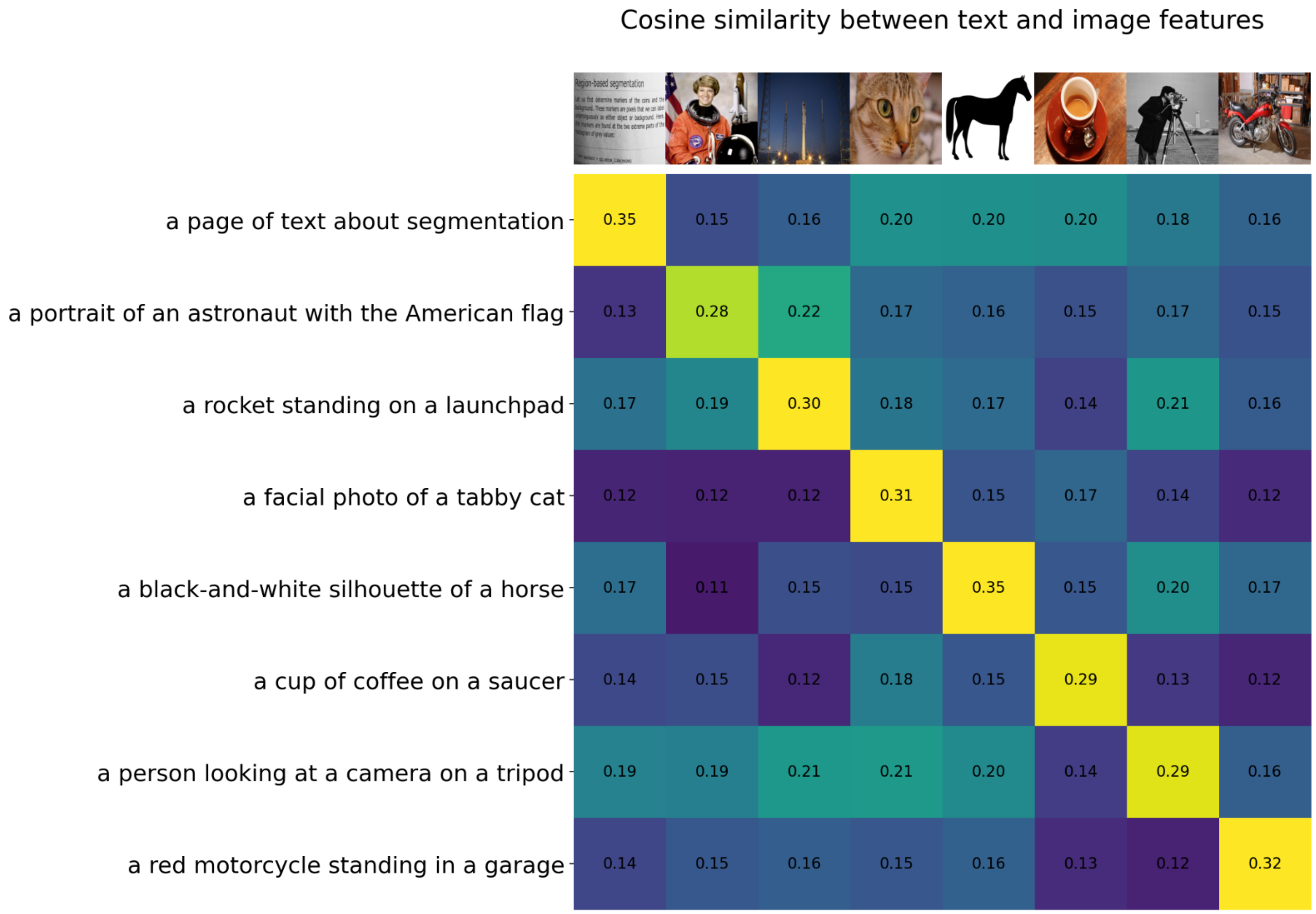
An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

CLIP: Contrastive Language-Image Pre-Training

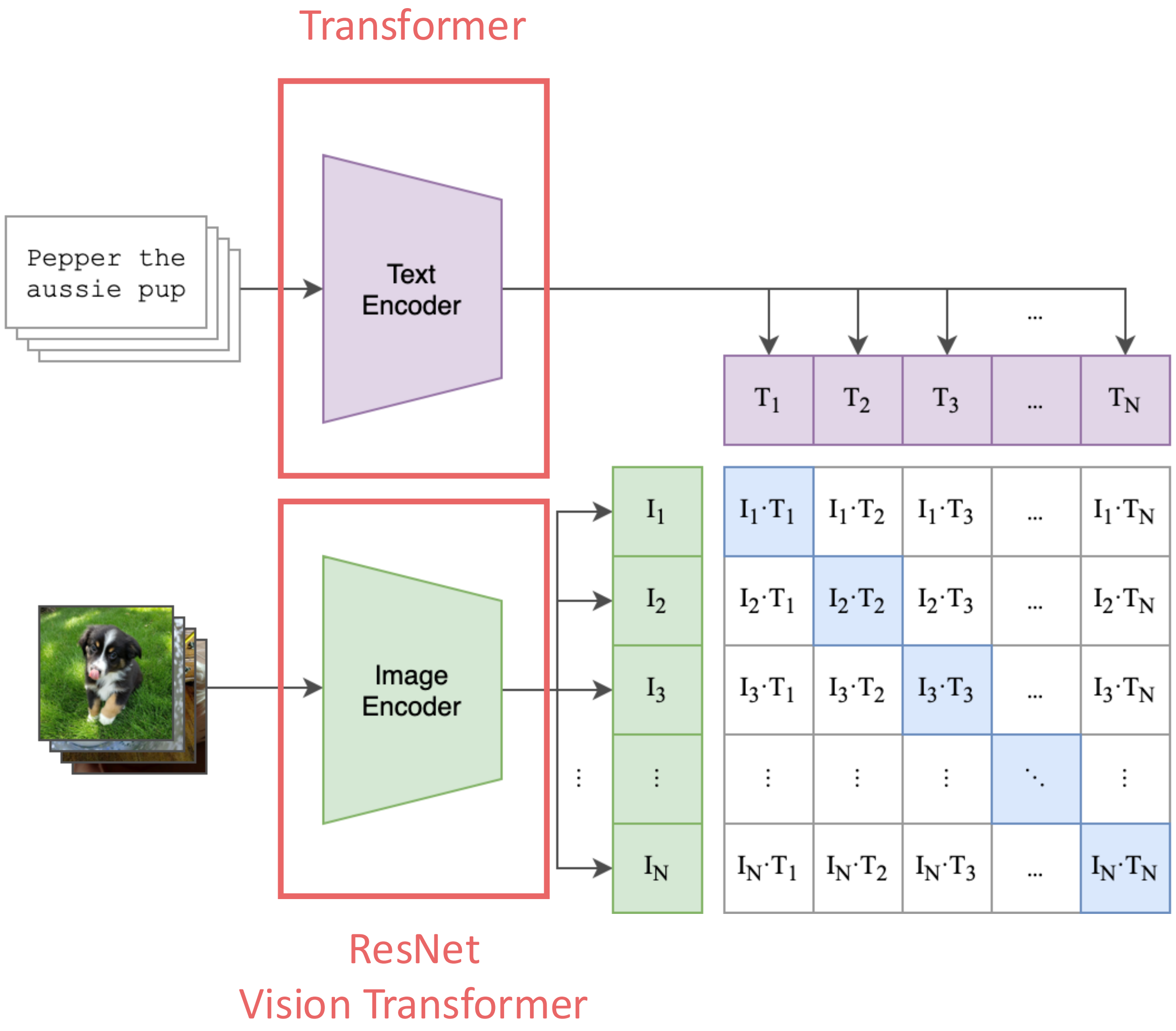


Learning Transferable Visual Models From Natural Language Supervision

Training with Image-Caption Pairs



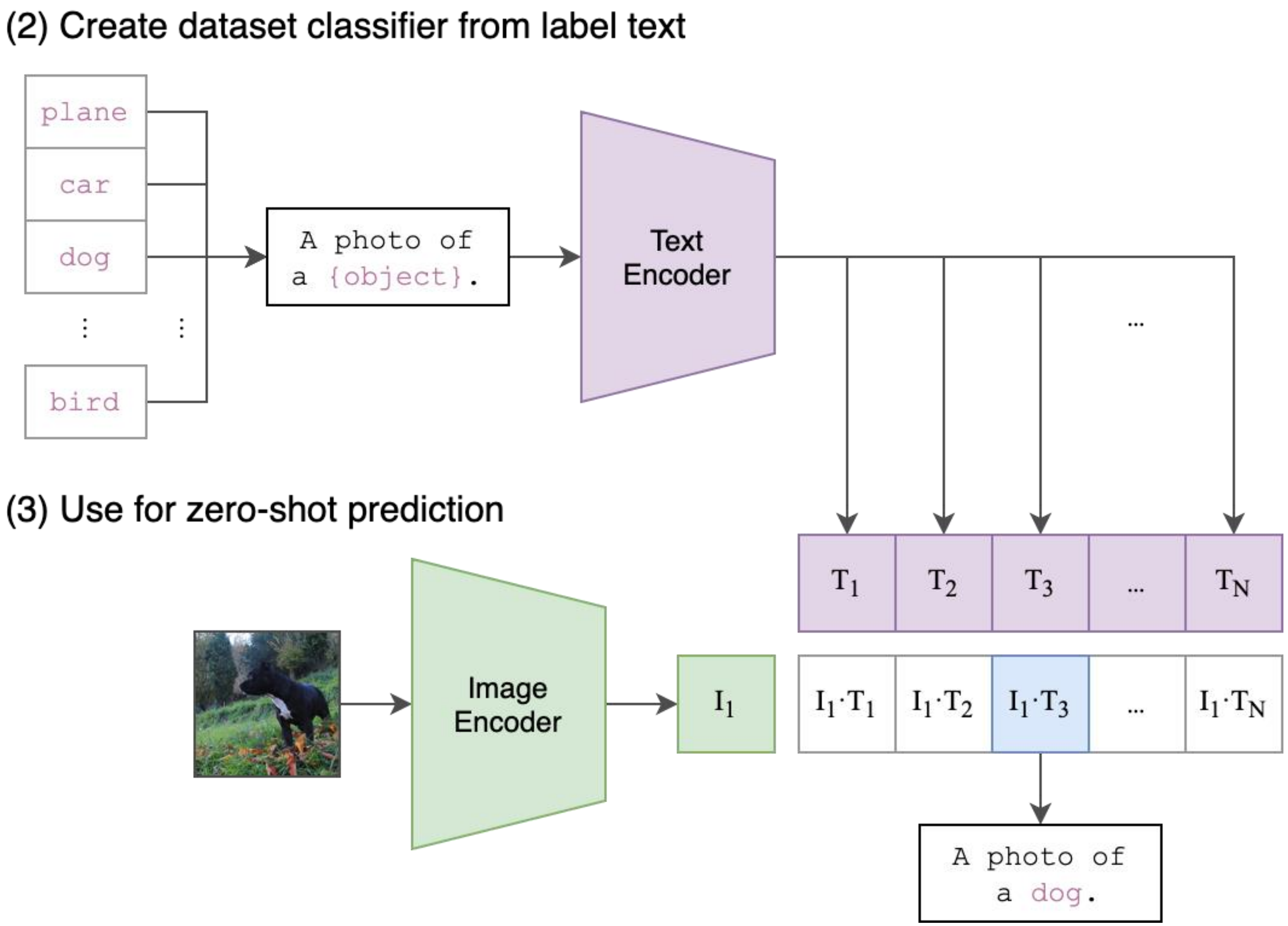
Training Details




```
# 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

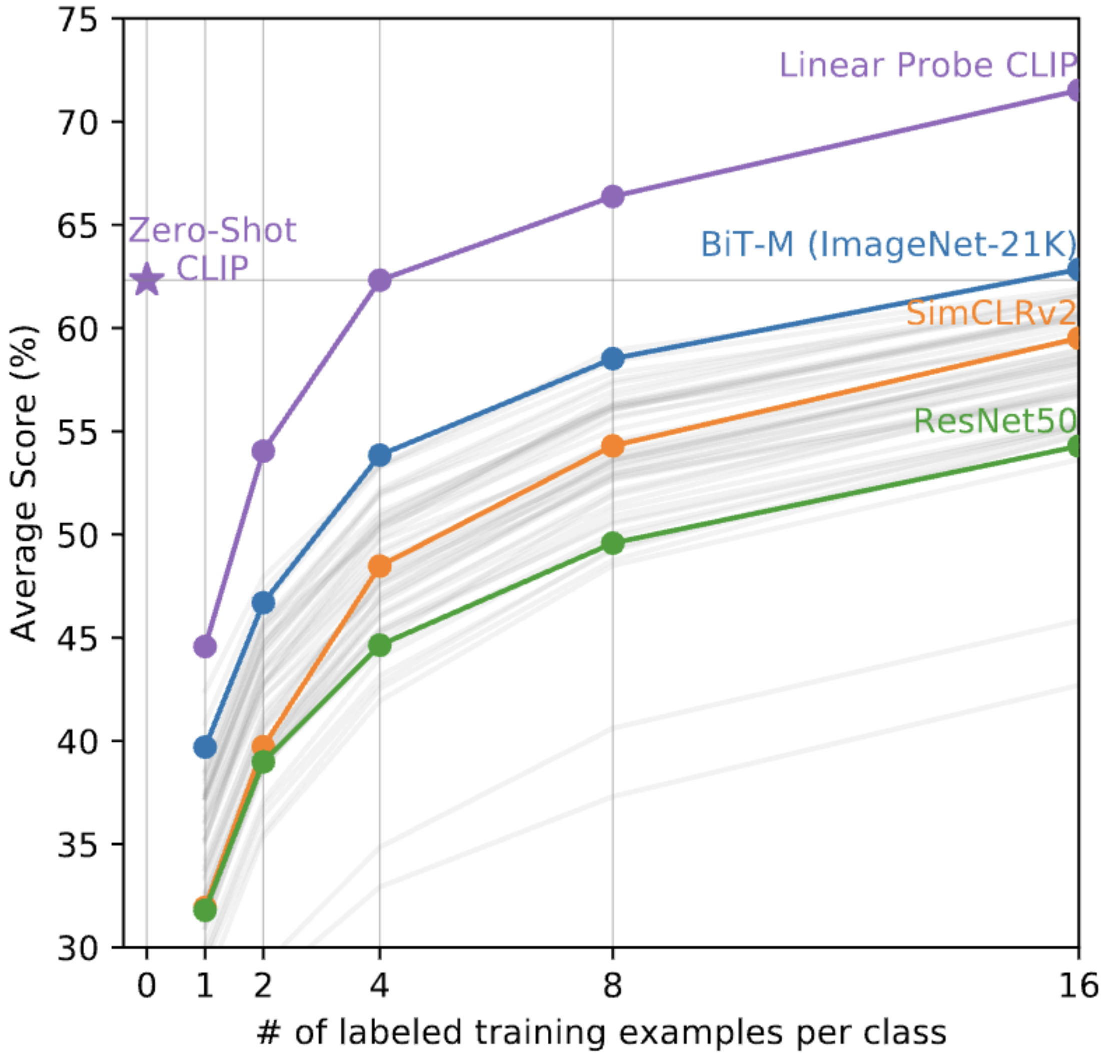
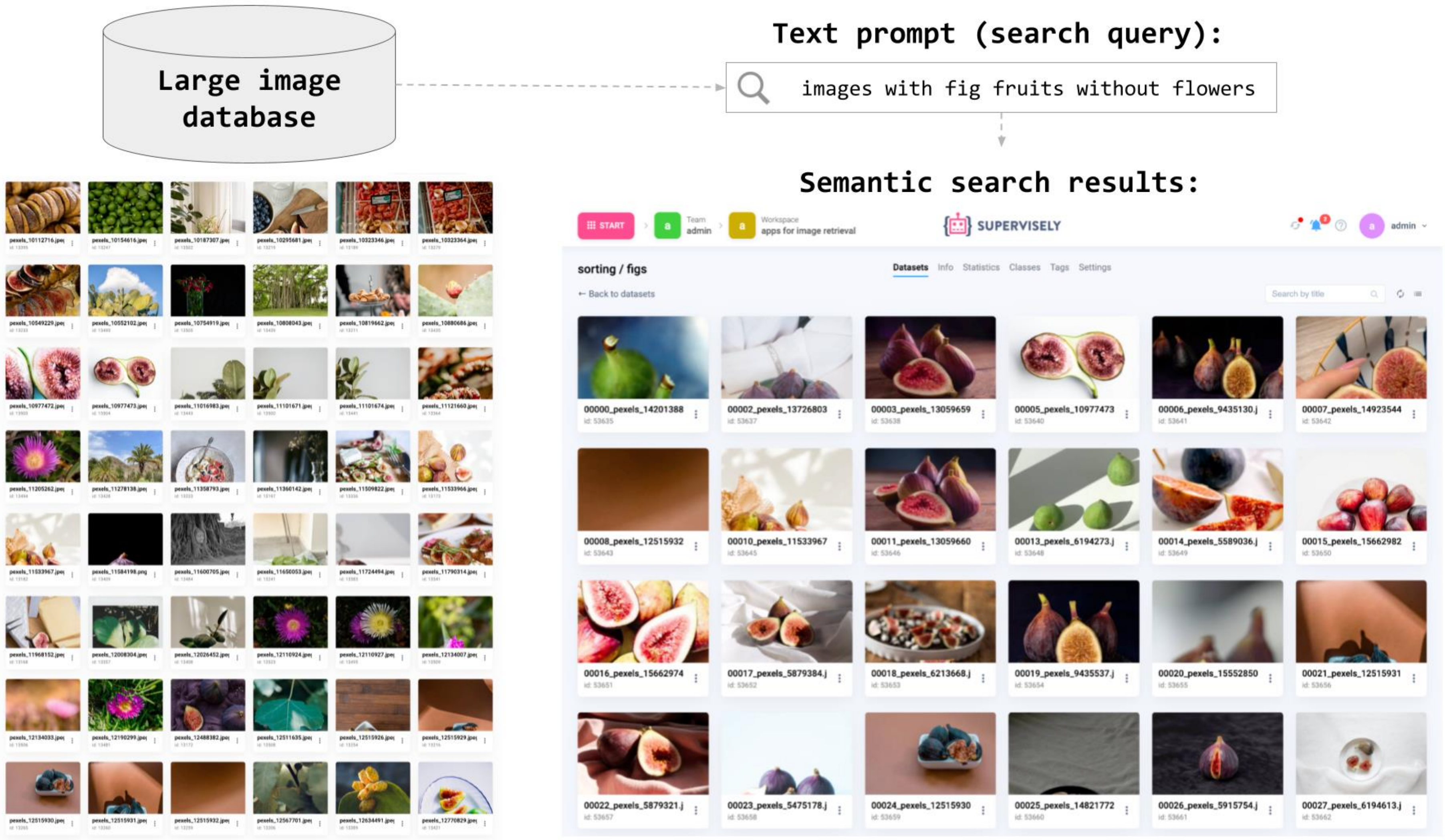
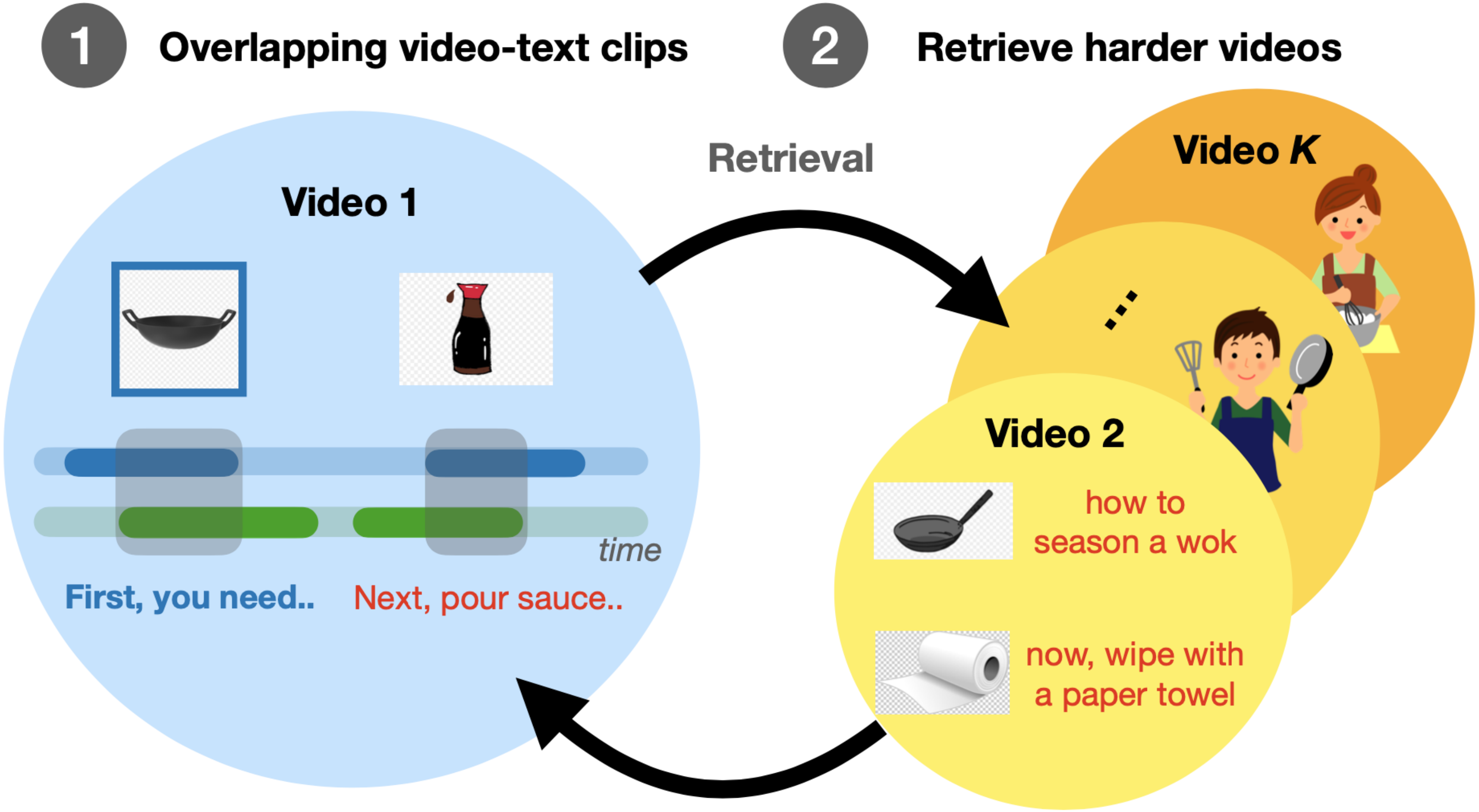


Image Retrieval with Text Query



VideoCLIP

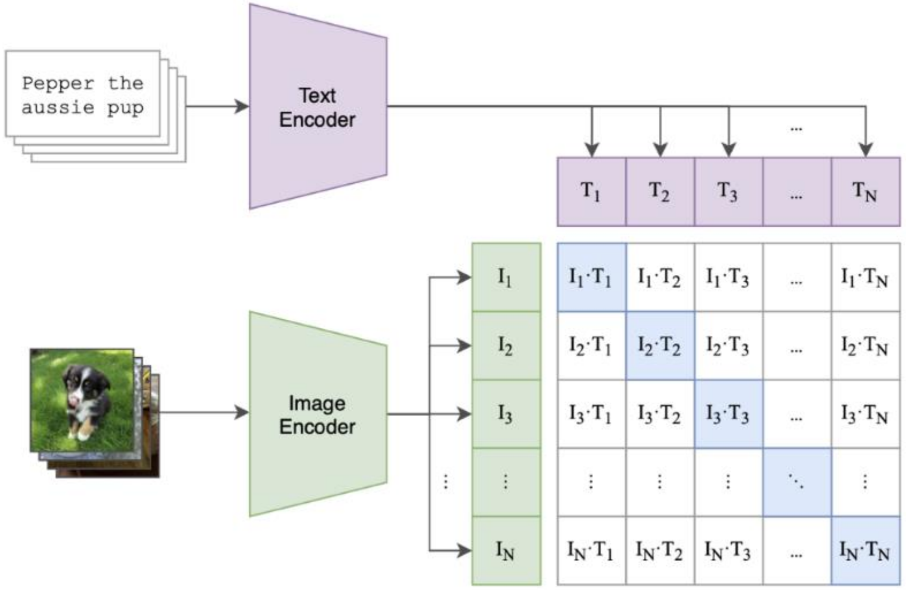


VideoCLIP: Contrastive learning with hard-retrieved negatives and overlapping positives for video-text pre-training.

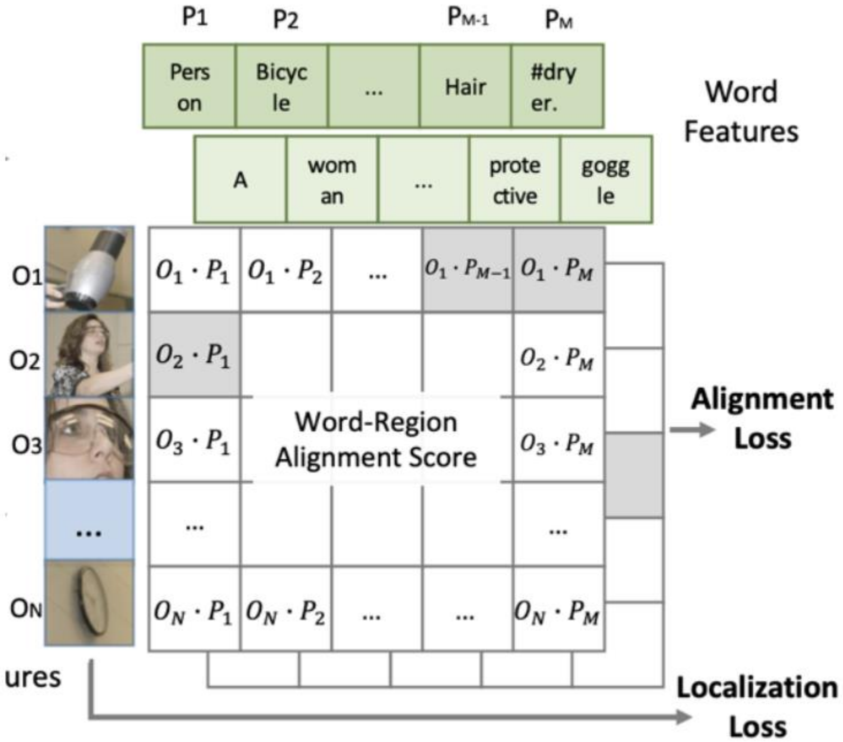
VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding

GLIP: Grounded Language-Image Pre-training

CLIP: capture information for whole image



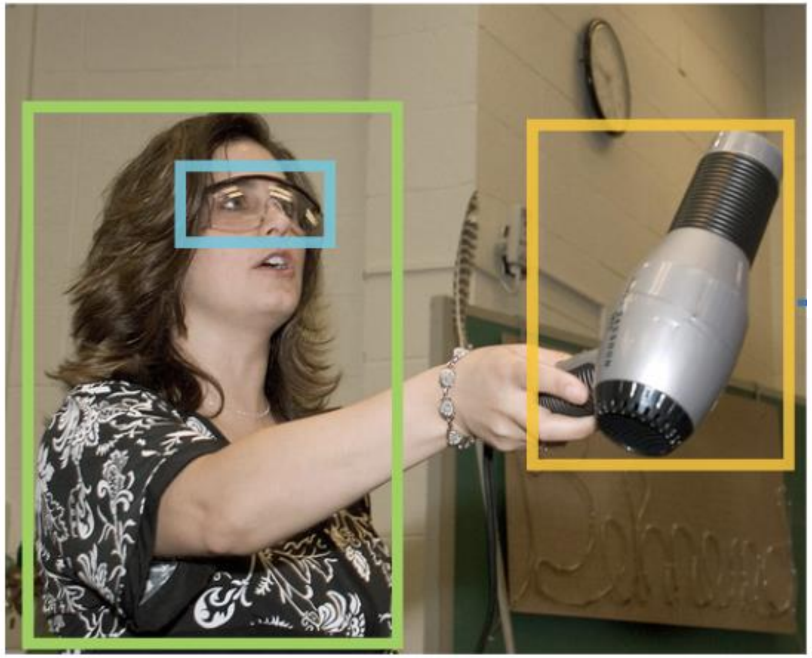
GLIP: capture information more for objects/entities



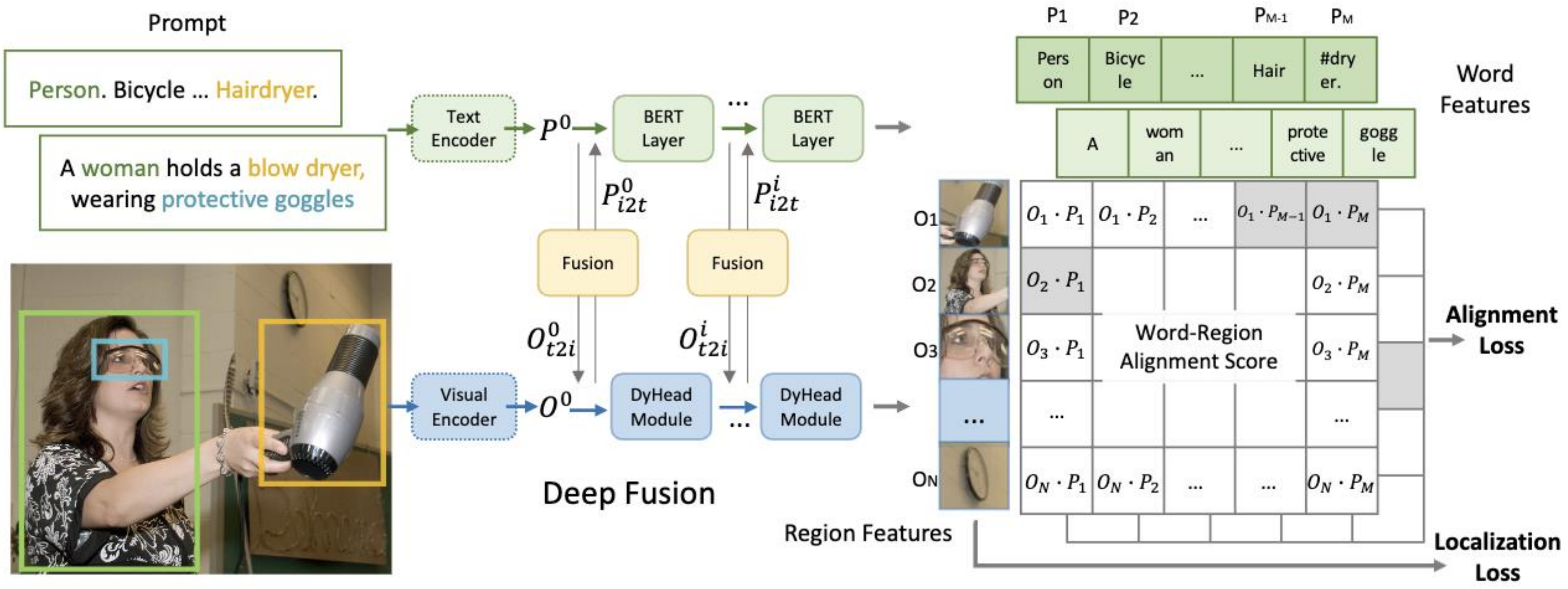
Object Detection and Text Grounding

Person. Bicycle … Hairdryer.

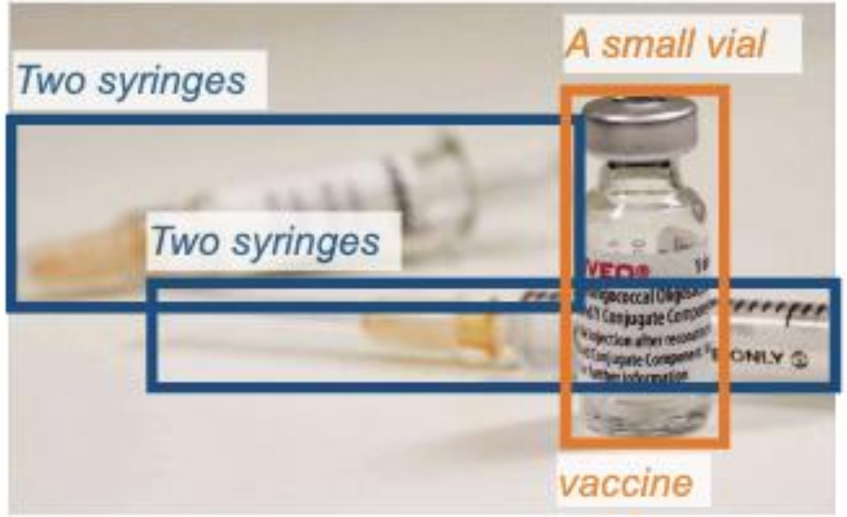
A woman holds a blow dryer, wearing protective goggles



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

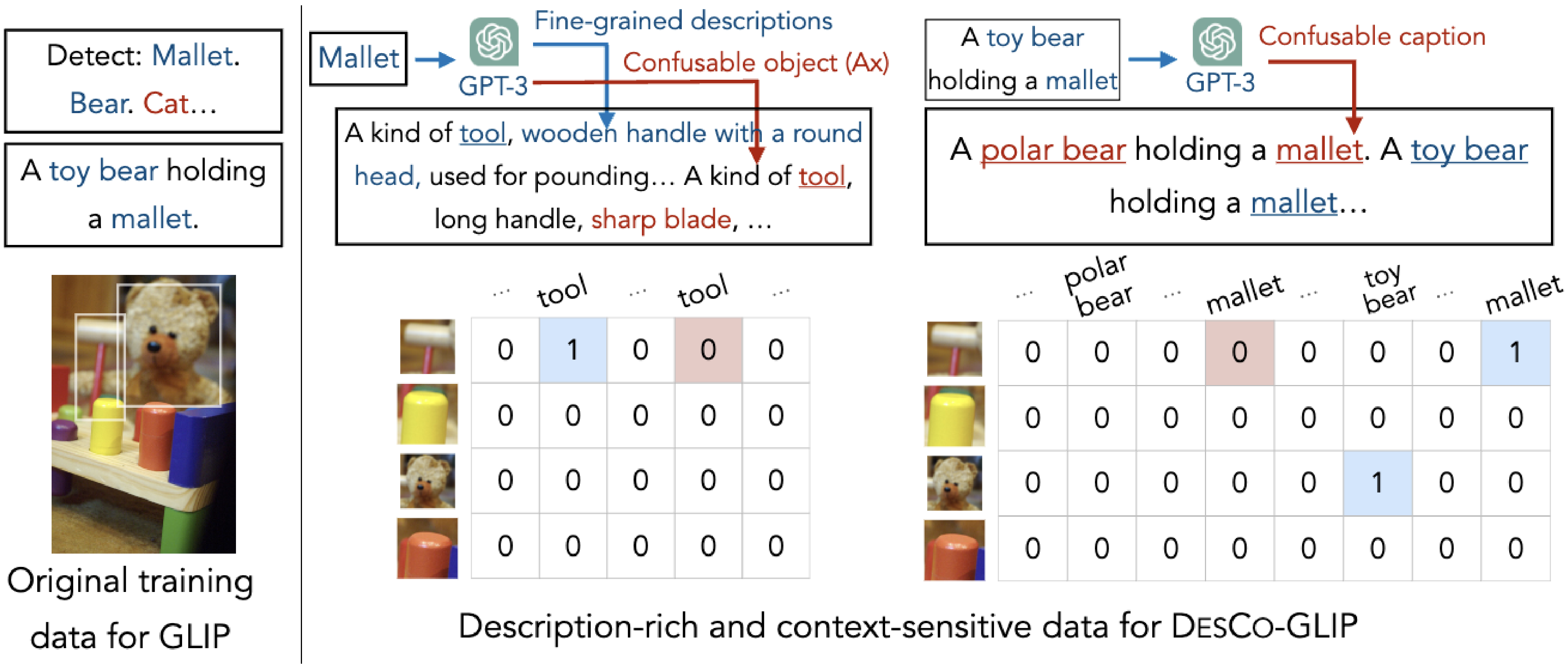
Zero-Shot Grounding

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

DesCo: Learning Object Recognition with Rich Language Descriptions

DesCo: Description-Conditioned



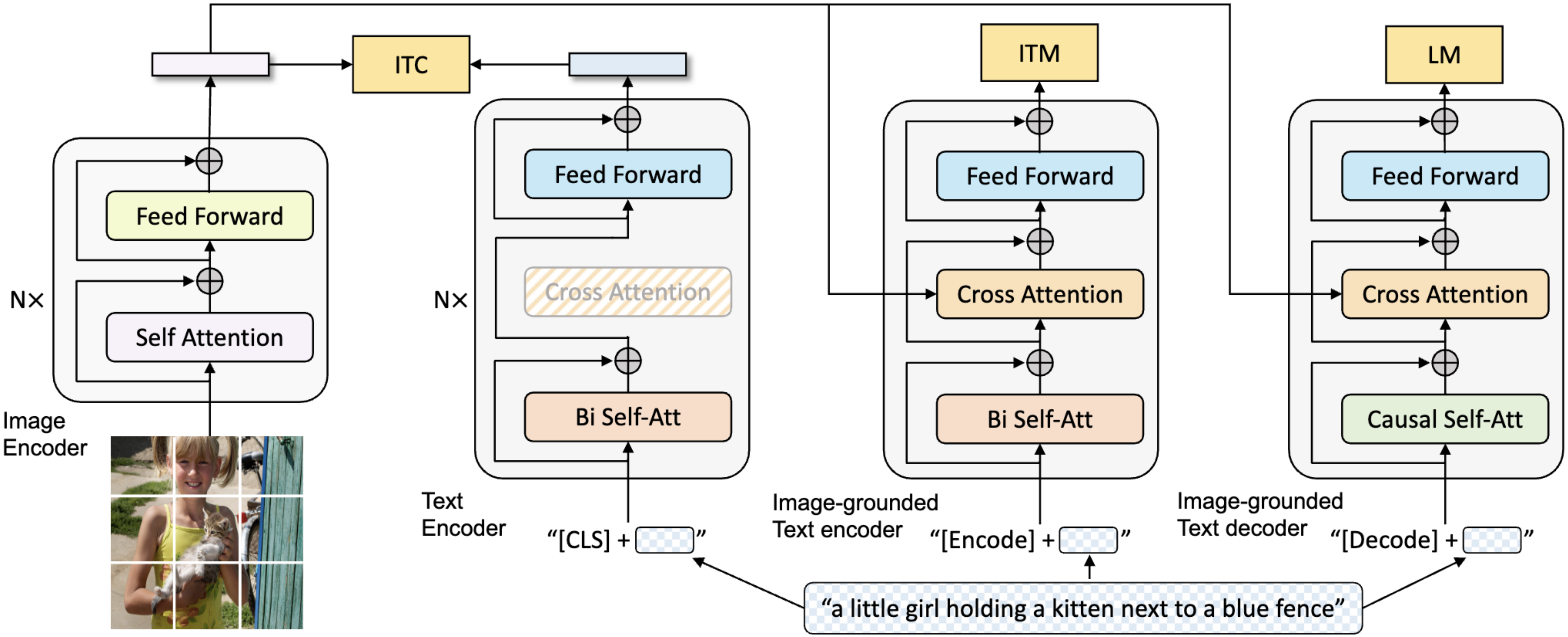
Zero-Shot Grounding

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

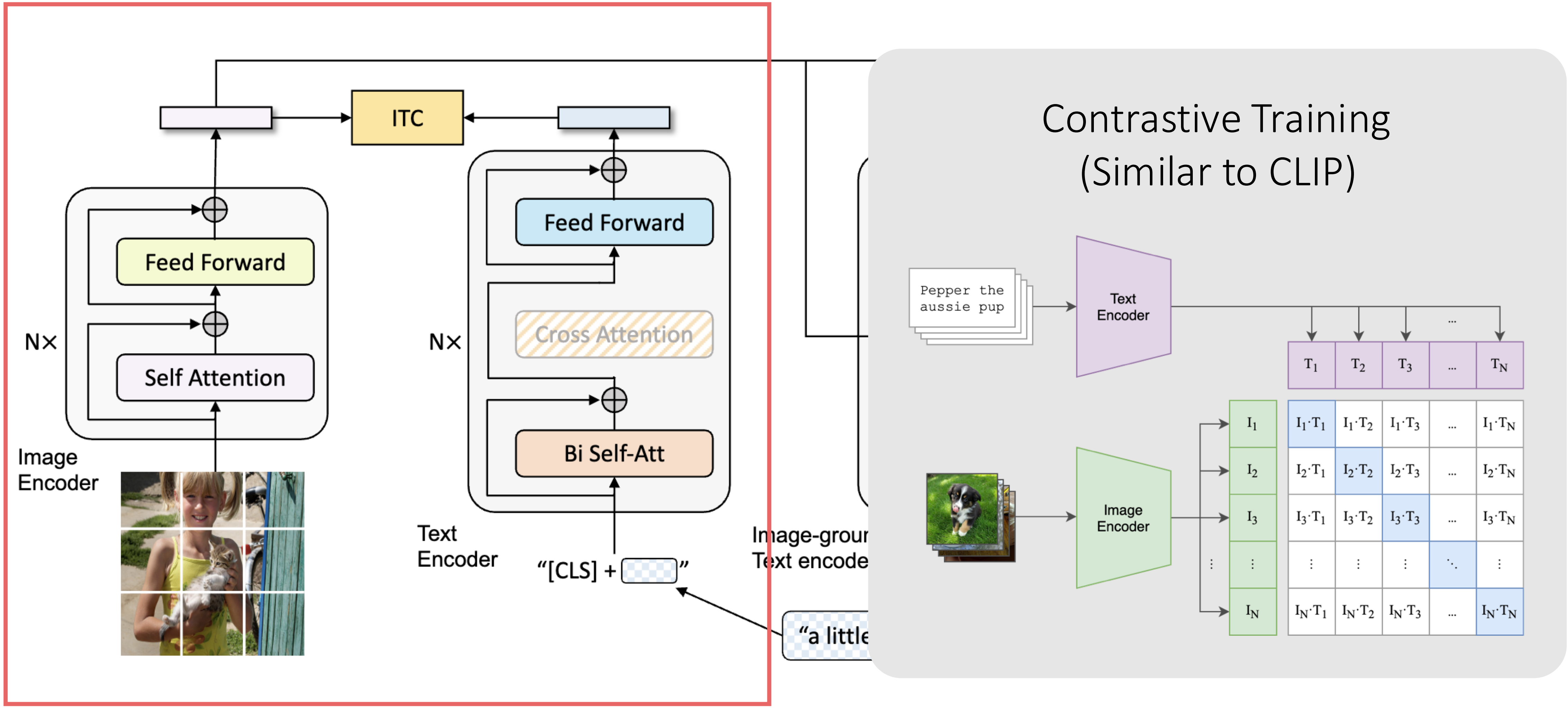
- Encoder-only
  - CLIP, GLIP, DesCo, etc.
  - Better for image-text retrieval
- Encoder-decoder
  - Better for generation

BLIP: Bootstrapping Language-Image Pre-training

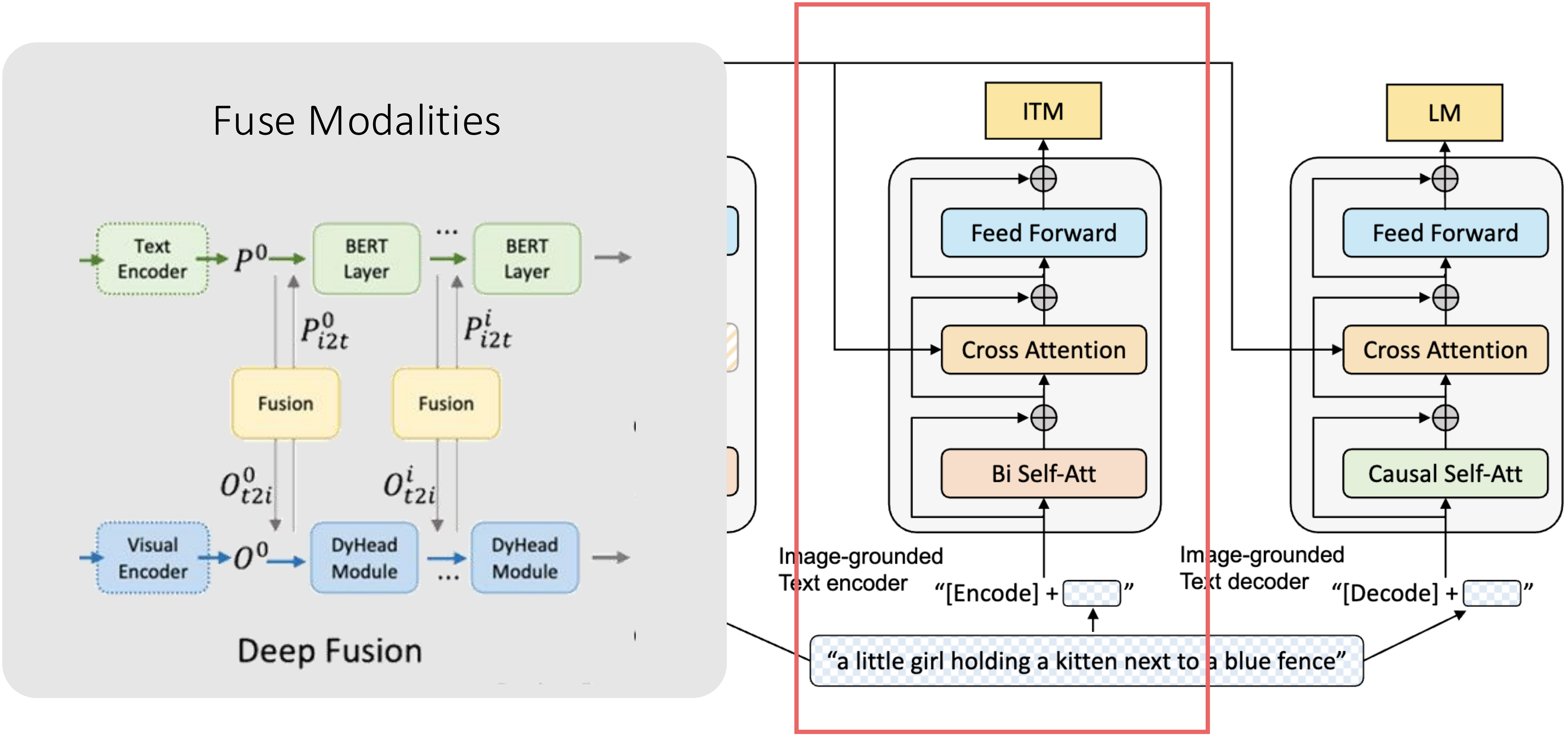


BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation

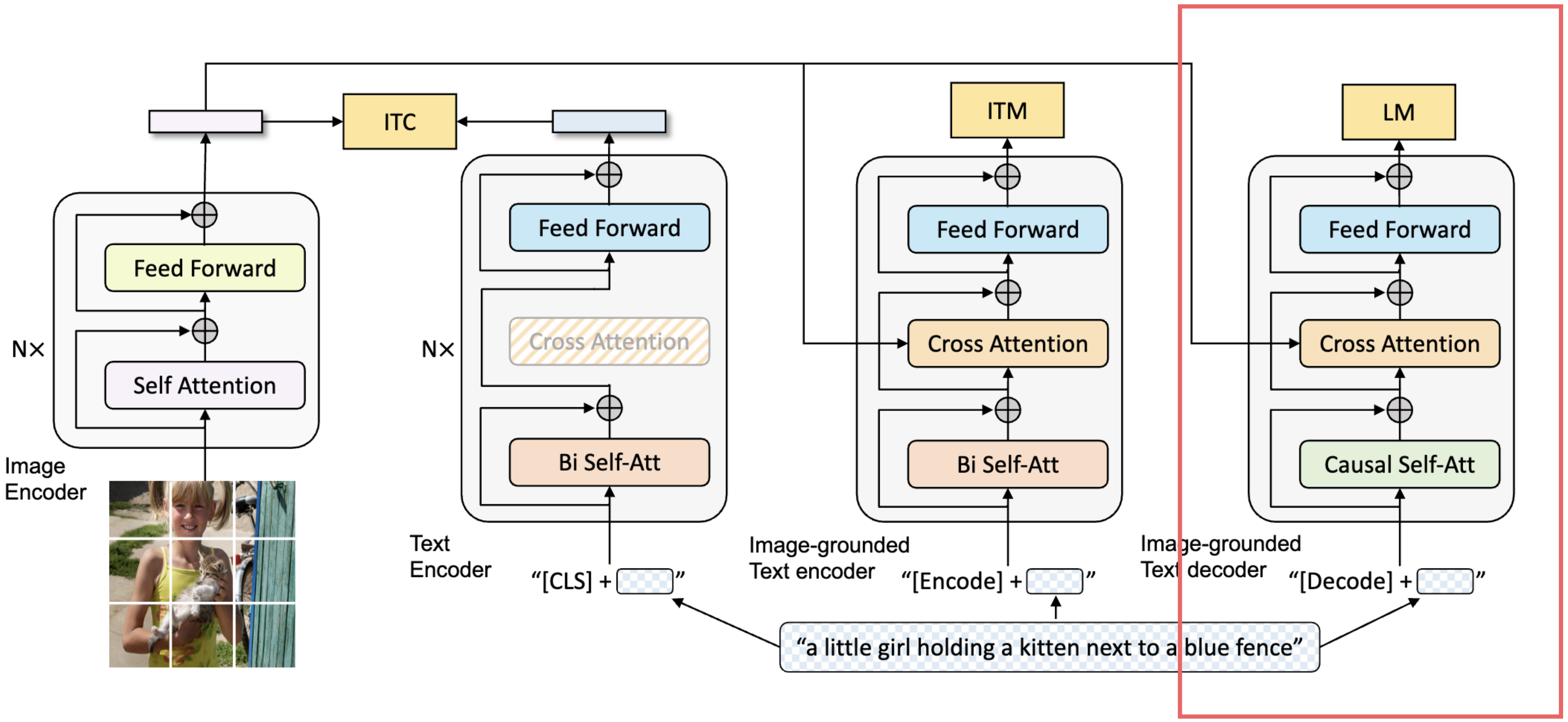
A Unified Framework



A Unified Framework



A Unified Framework



Zero-Shot Image-Text Retrieval

Method	Pre-train # Images	Flickr30K (1K test set) TR			IR		
		R@1	R@5	R@10	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
$\text{LEMON}_{\text{base}}$† (Hu et al., 2021)	12M	104.5	14.6	100.7	14.0	96.7	12.4	100.4	13.8	-	-
$\text{LEMON}_{\text{base}}$† (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
$\text{BLIP}_{\text{CapFilt-L}}$	129M	**111.8**	**14.9**	**108.6**	**14.8**	**111.5**	**14.2**	**109.6**	**14.7**	39.7	**133.3**
$\text{LEMON}_{\text{large}}$† (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
$\text{SimVLM}_{\text{huge}}$ (Wang et al., 2021)	1.8B	113.7	-	110.9	-	115.2	-	112.2	-	40.6	143.3
$\text{BLIP}_{\text{ViT-L}}$	129M	114.9	15.2	112.1	14.9	115.3	14.4	113.2	14.8	40.4	136.7

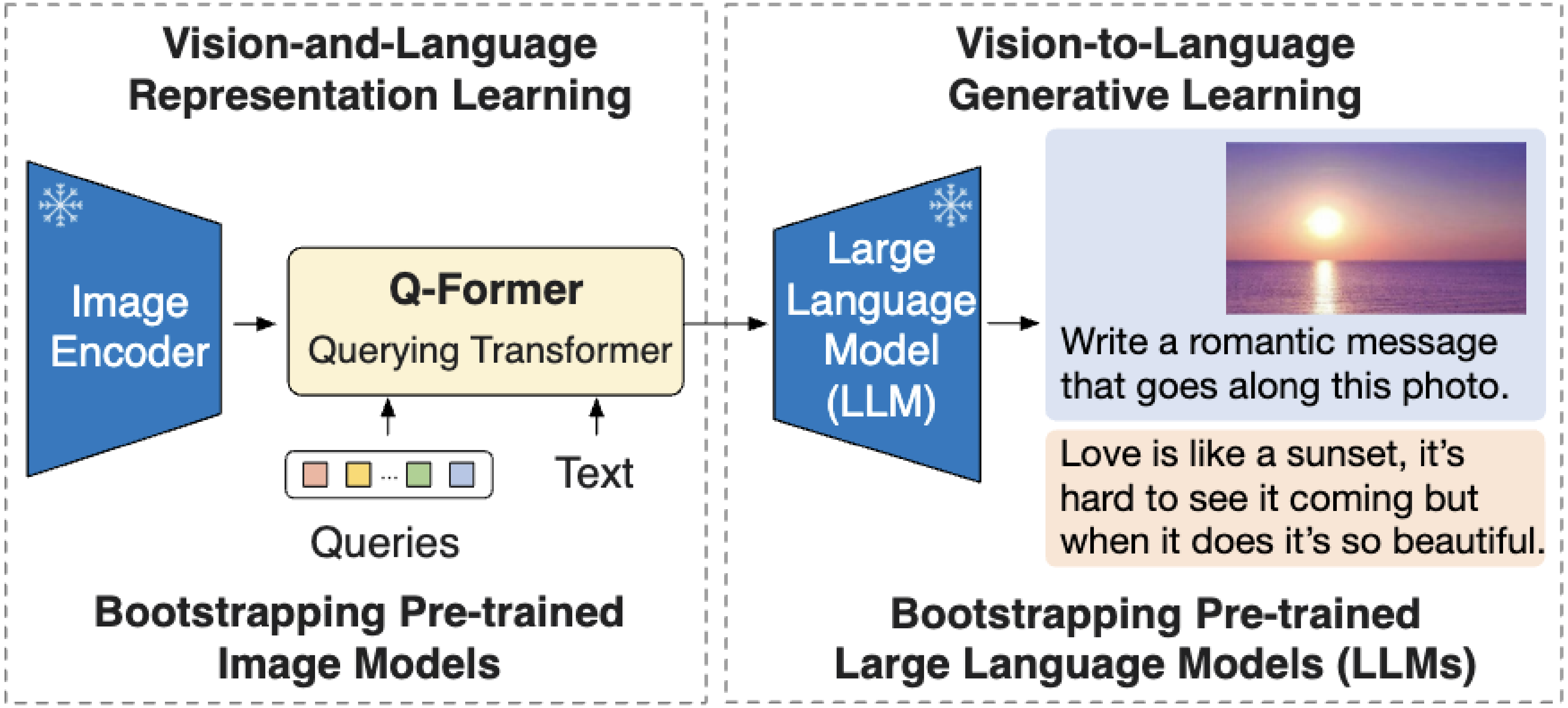
Visual Question Answering

Visual Question Answering

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

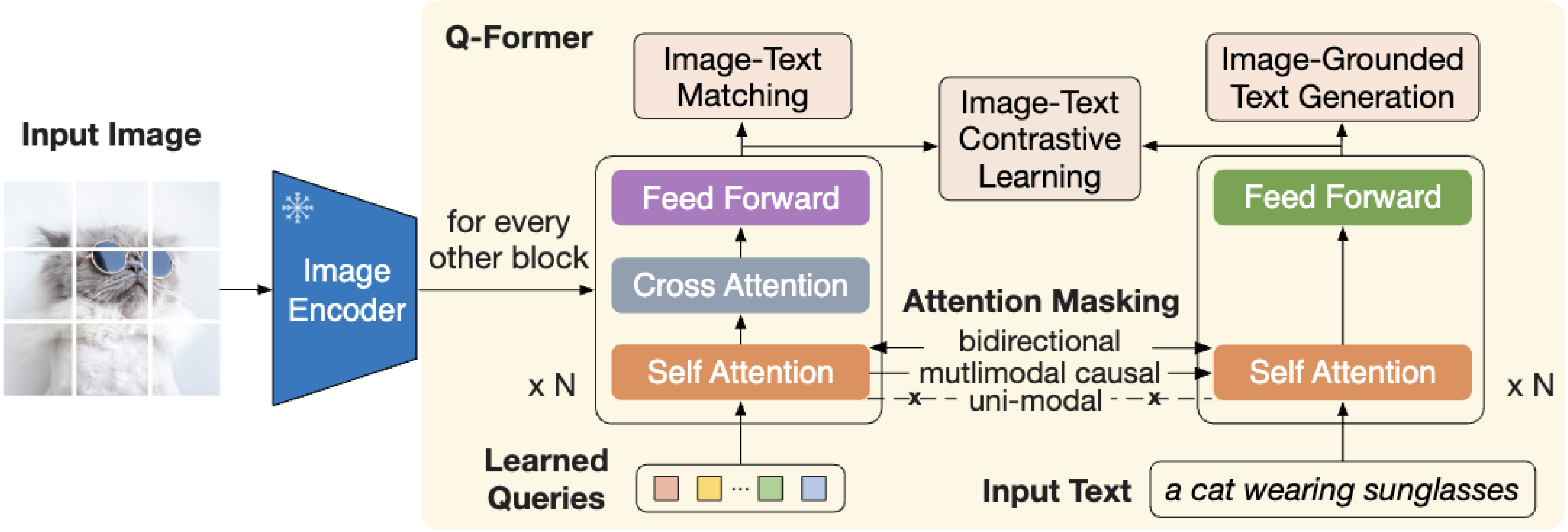
Method	Pre-train #Images	VQA test-dev	VQA test-std	$NLVR^2$ dev	$NLVR^2$ test-P
LXMERT	180K	72.42	72.54	74.90	74.50
UNITER	4M	72.70	72.91	77.18	77.85
VL-T5/BART	180K	-	71.3	-	73.6
OSCAR	4M	73.16	73.44	78.07	78.36
SOHO	219K	73.25	73.47	76.37	77.32
VILLA	4M	73.59	73.67	78.39	79.30
UNIMO	5.6M	75.06	75.27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

BLIP-2: Frozen Image Encoders and Large Language Models

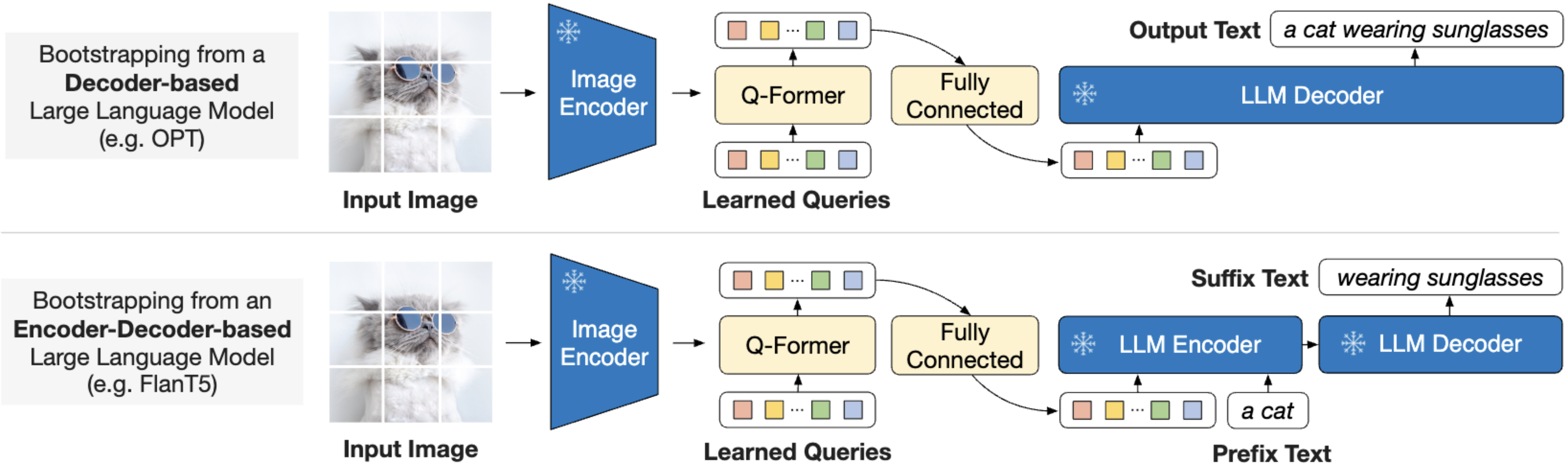


BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

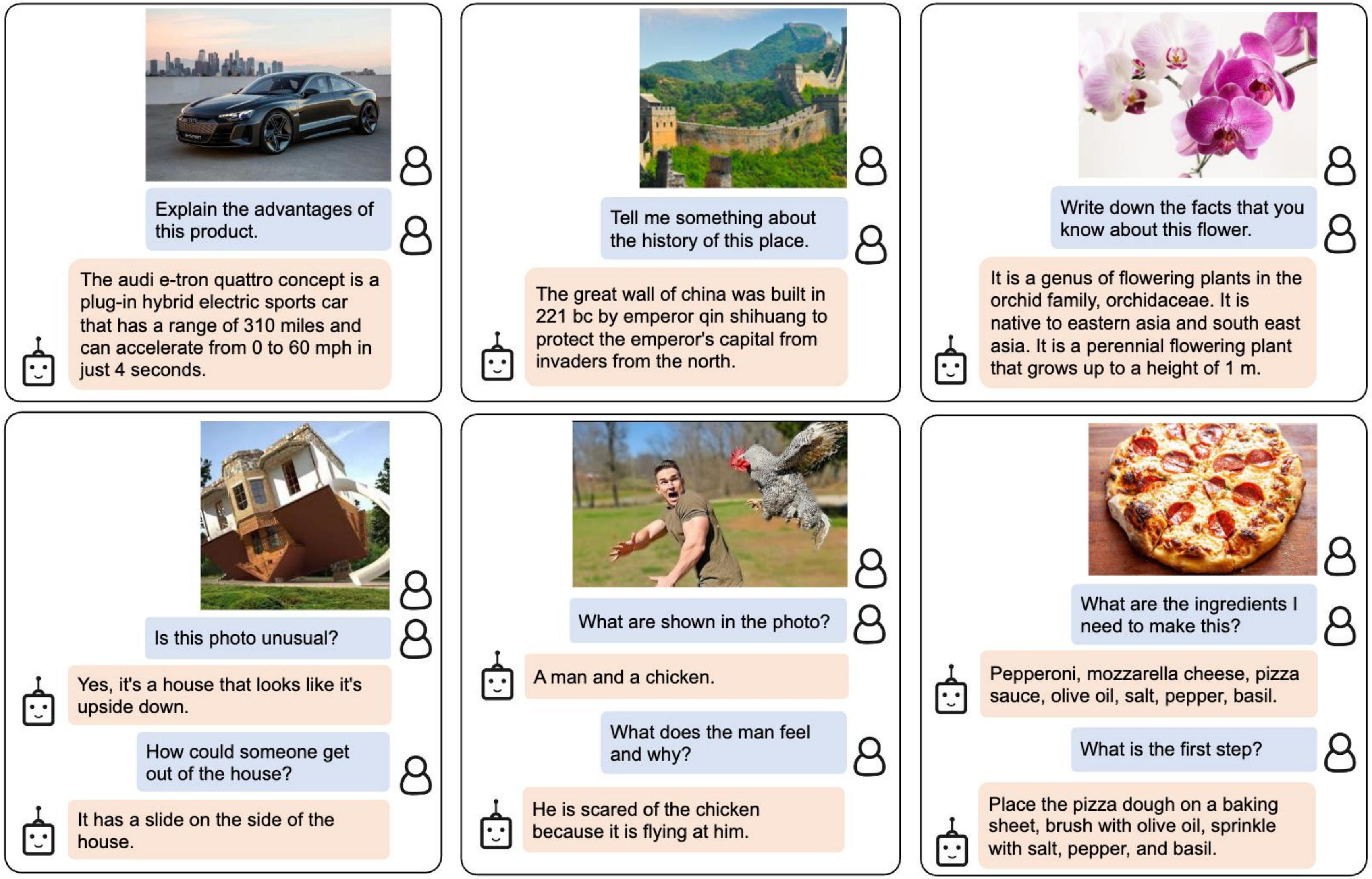
Vision-Language Representation Learning



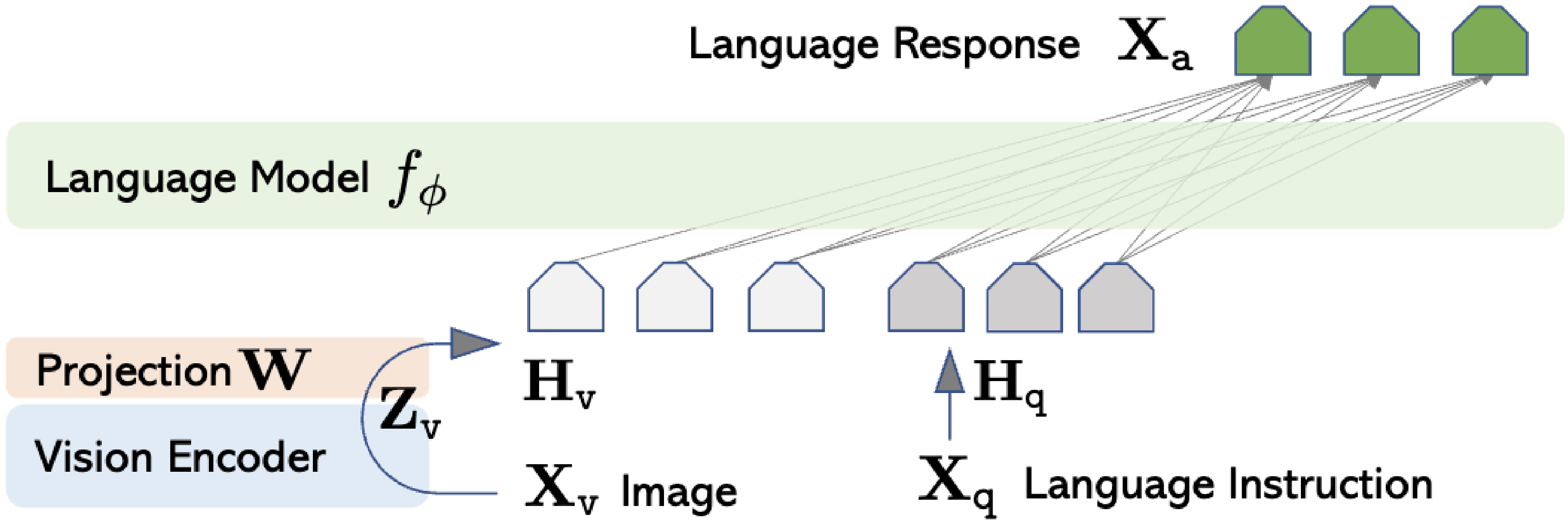
Vision-to-Language Generative Pre-Training



Instructed Zero-Shot Image-to-Text Generation



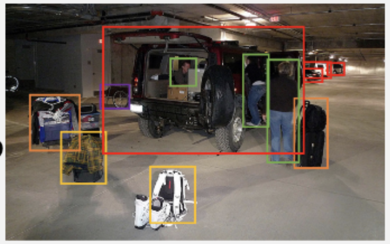
LLaVA: Visual Instruction Tuning



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.



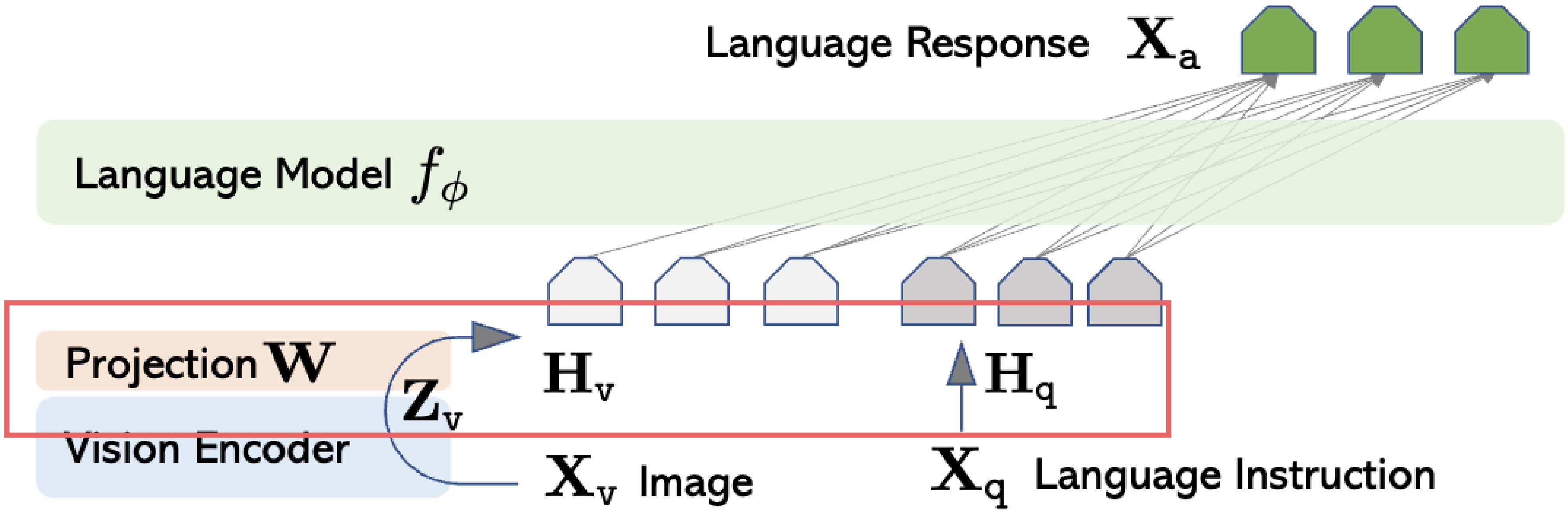
Context type 2: Boxes
person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

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>

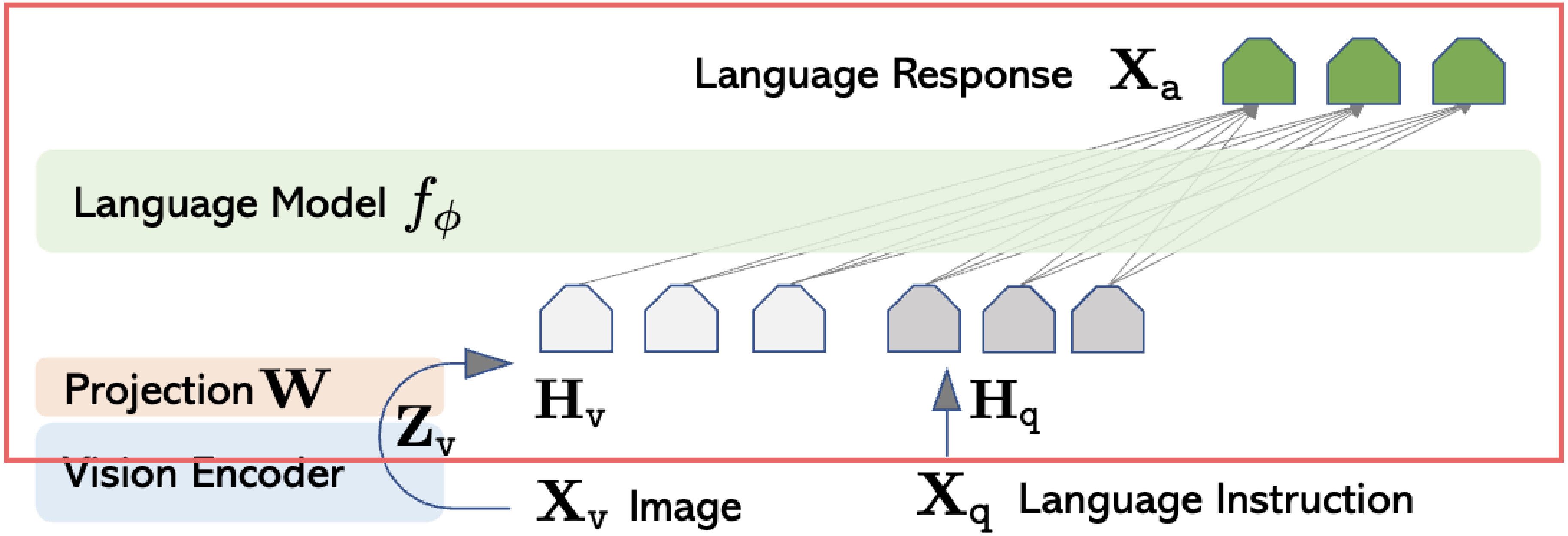
Response type 3: complex reasoning
Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their 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

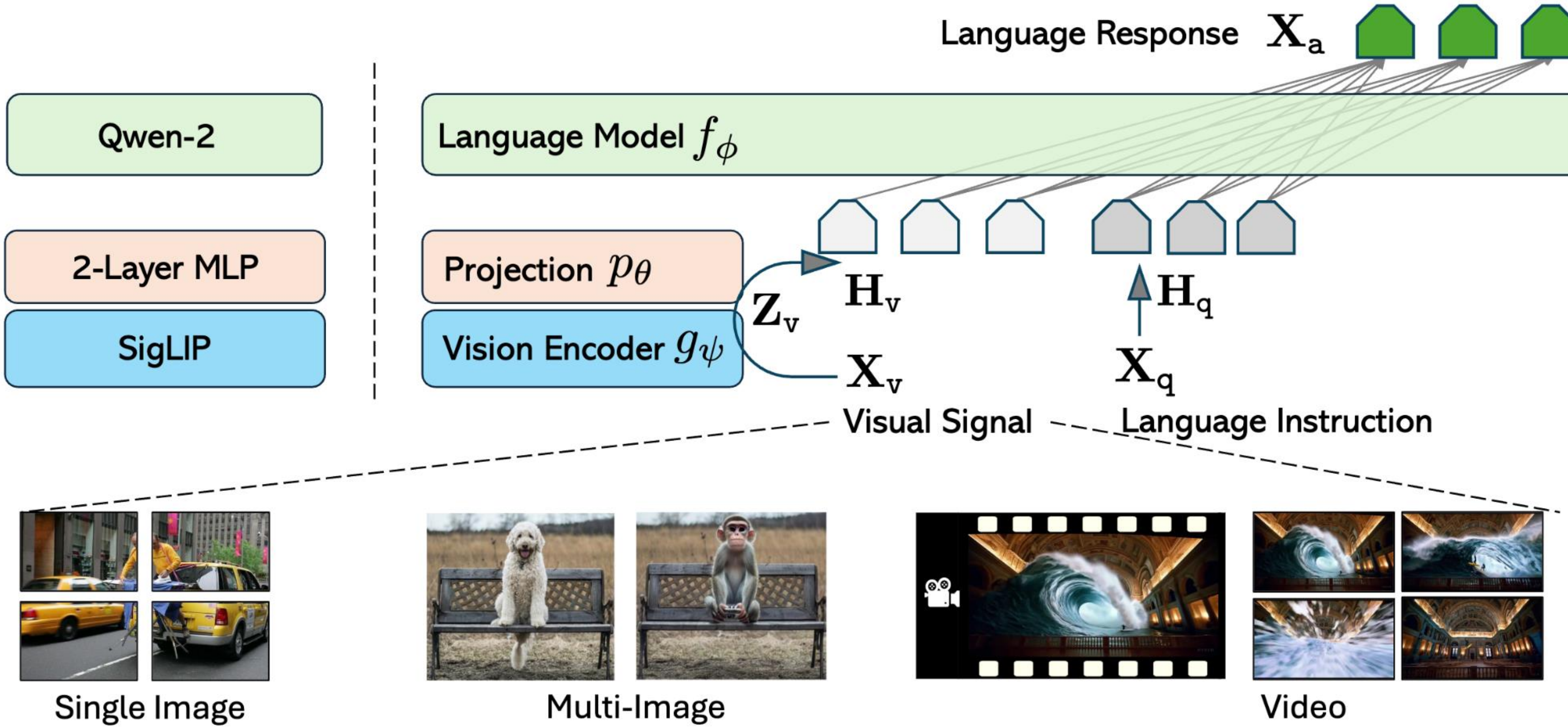
Visual input example, Extreme Ironing:

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

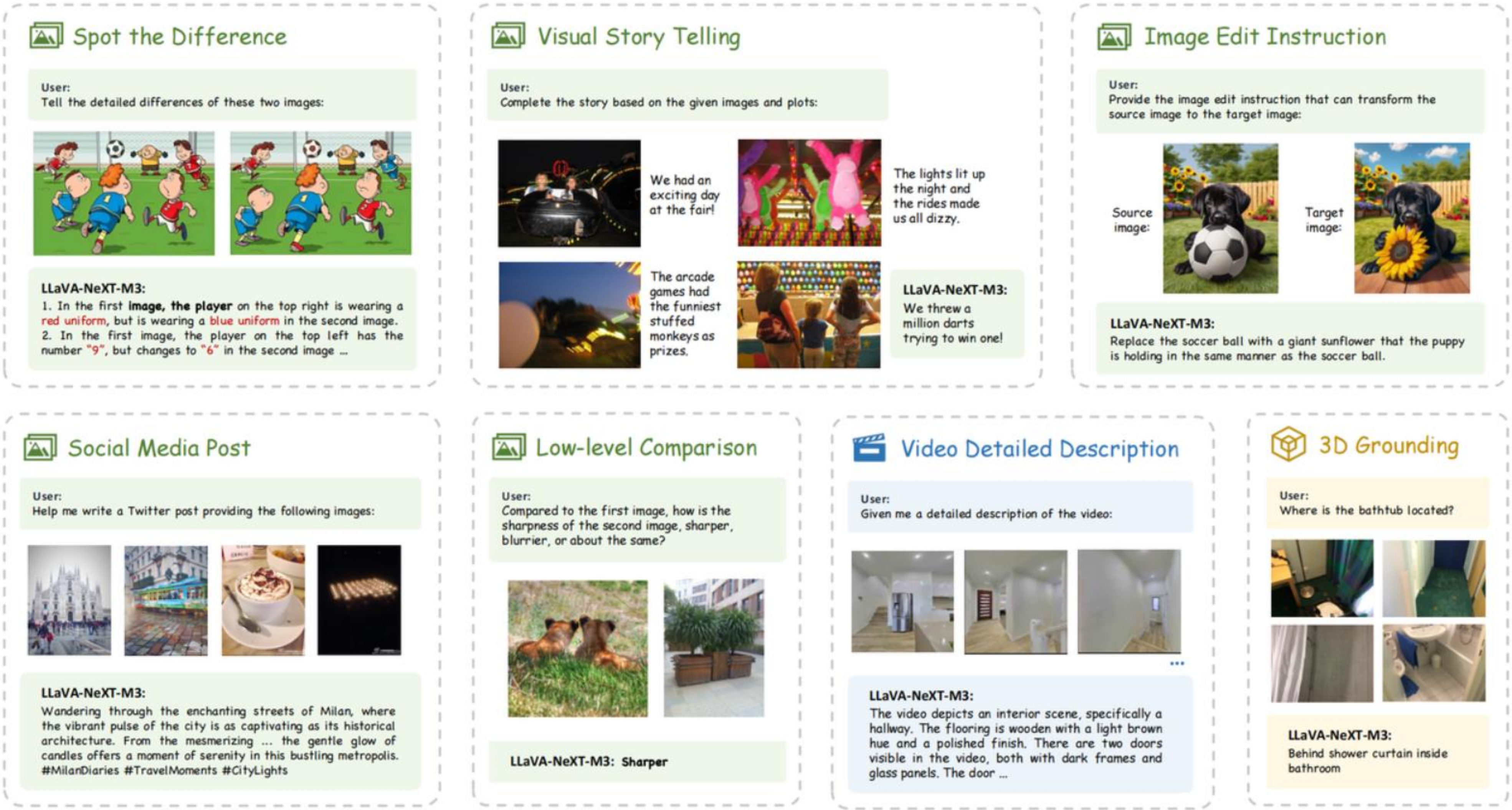


https://llava-vl.github.io/blog/2024-08-05-llava-onevision/

LLaVA-OneVision

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

LLaVA-NeXT-Interleave



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*

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

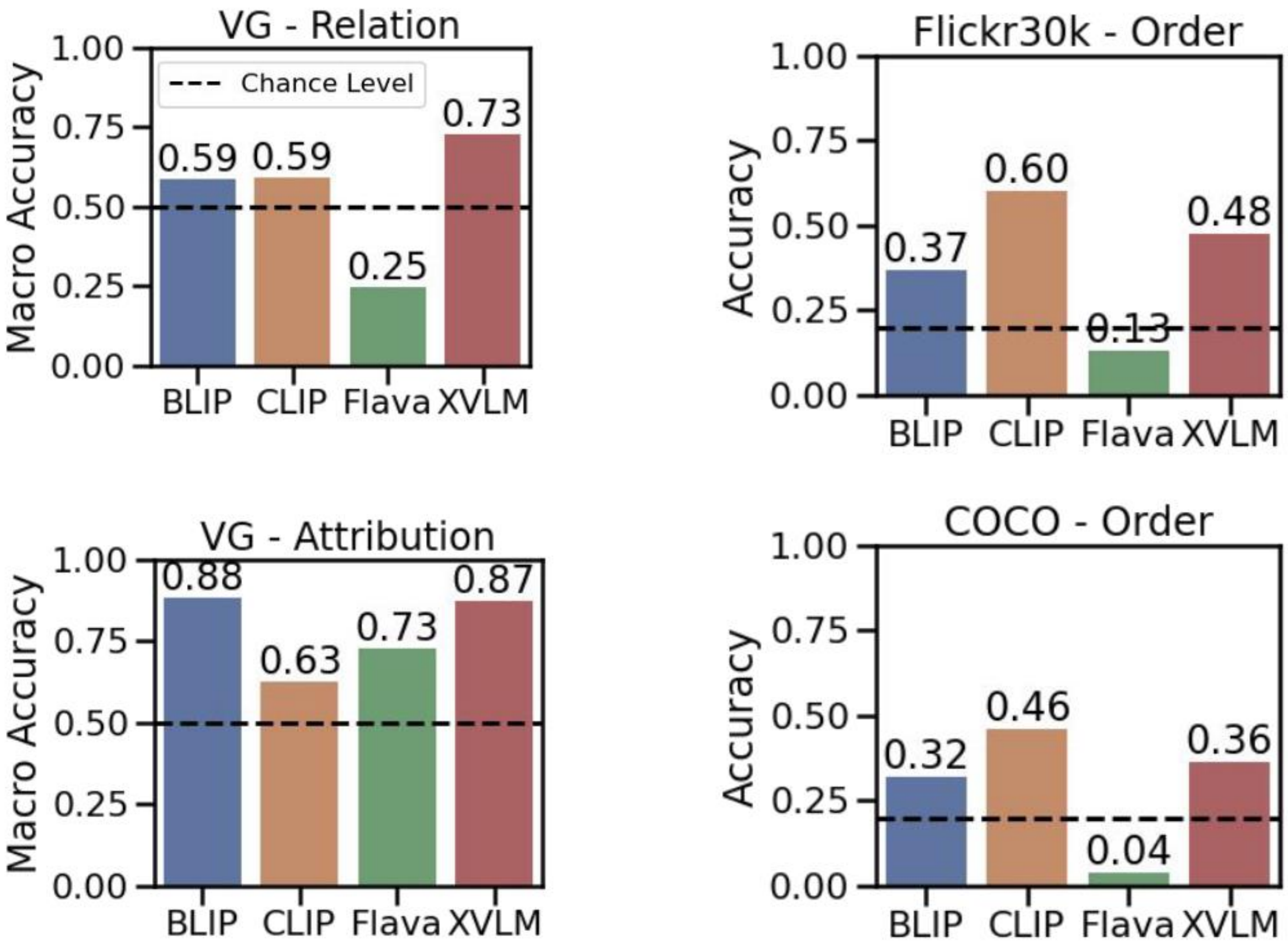
X (shuffle adjective/noun) a gray bathtub is looking at a white cat and sitting in a brown dog

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

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

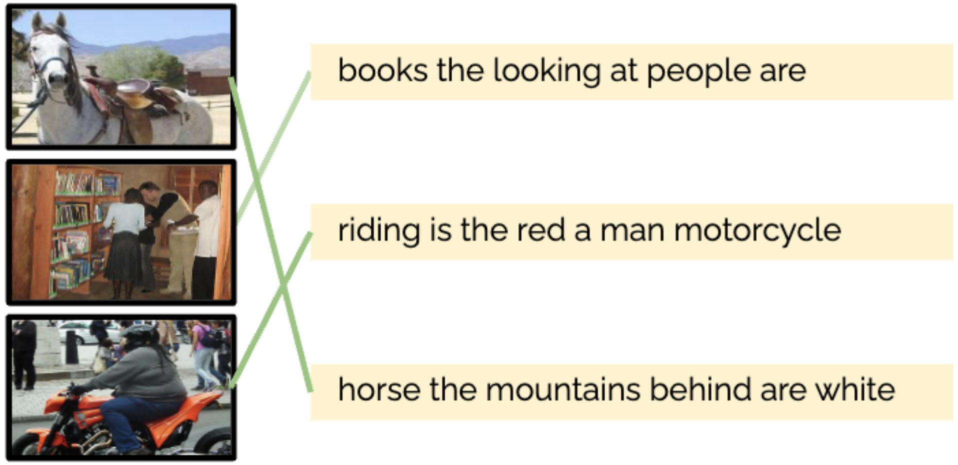
X (shuffle trigrams) a brown cat a white bathtub is looking at a gray dog and sitting in

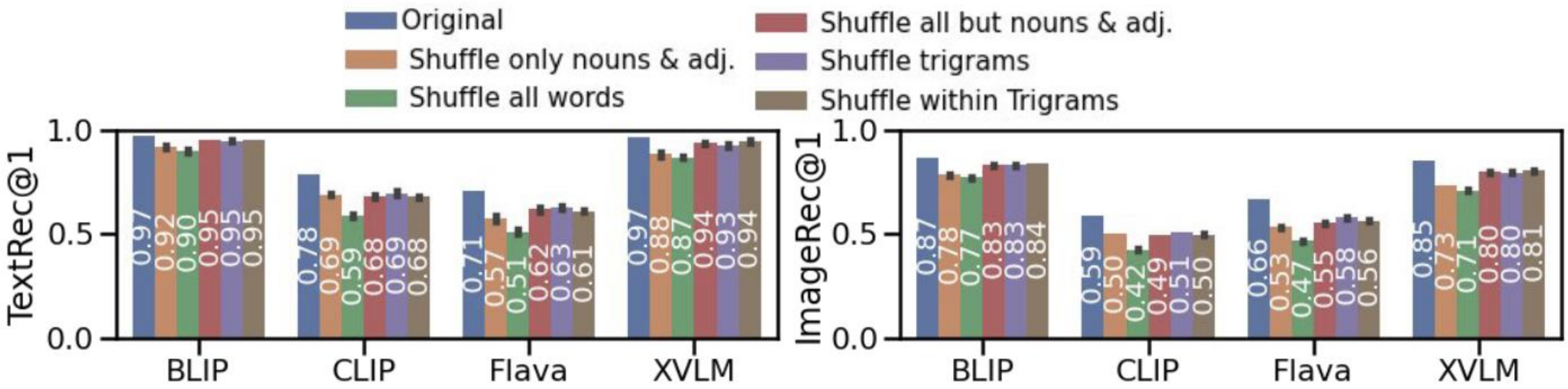
Results


VG - Relation
Macro Accuracy
Chance Level
0.59
0.59
0.25
0.73
1.00
0.75
0.50
0.25
0.00
BLIP
CLIP
Flava
XVLM
Flickr30k - Order
Accuracy
0.37
0.60
0.13
0.48
COCO - Order
0.32
0.46
0.04
0.36
VG - Attribution
0.88
0.63
0.73
0.87

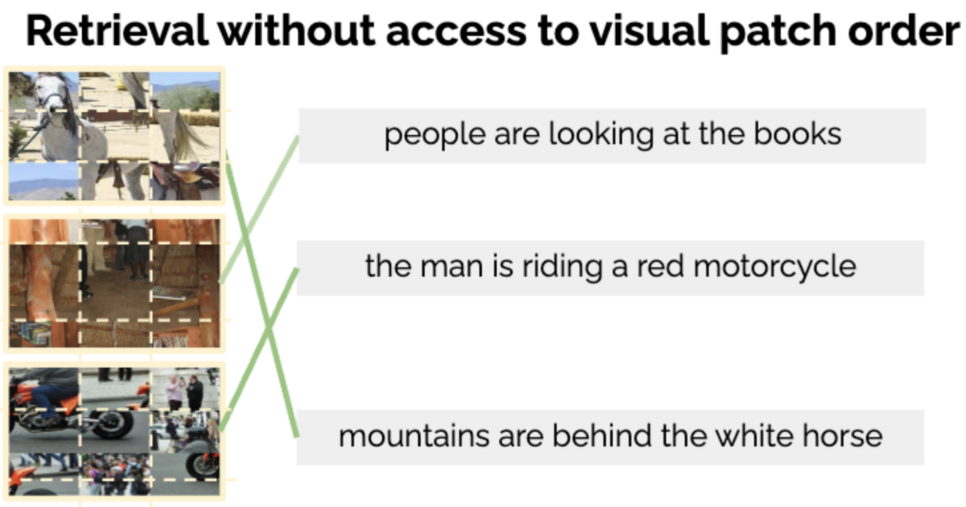
More Analysis on Text-Image Retrieval

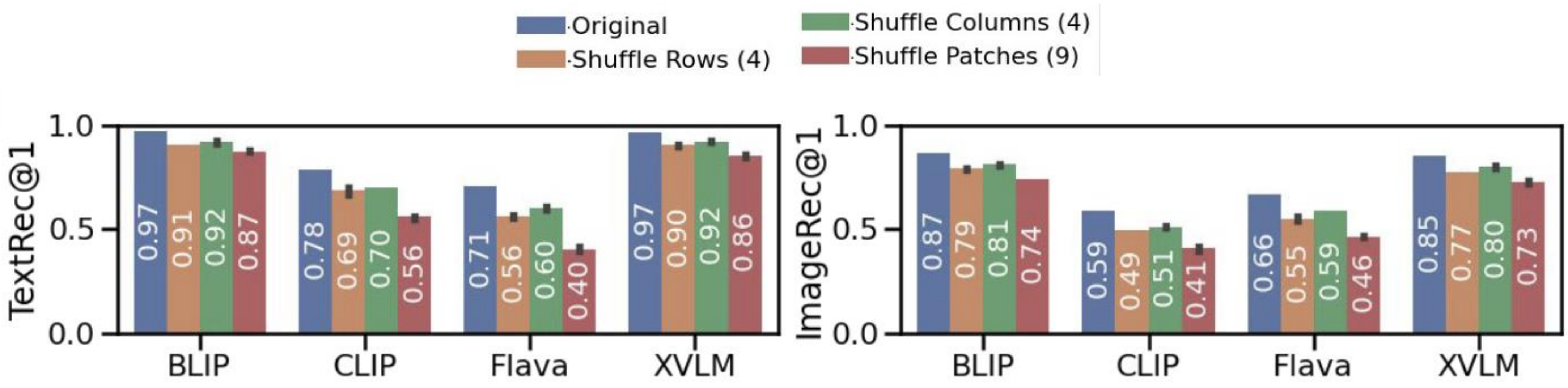
Retrieval without access to word order



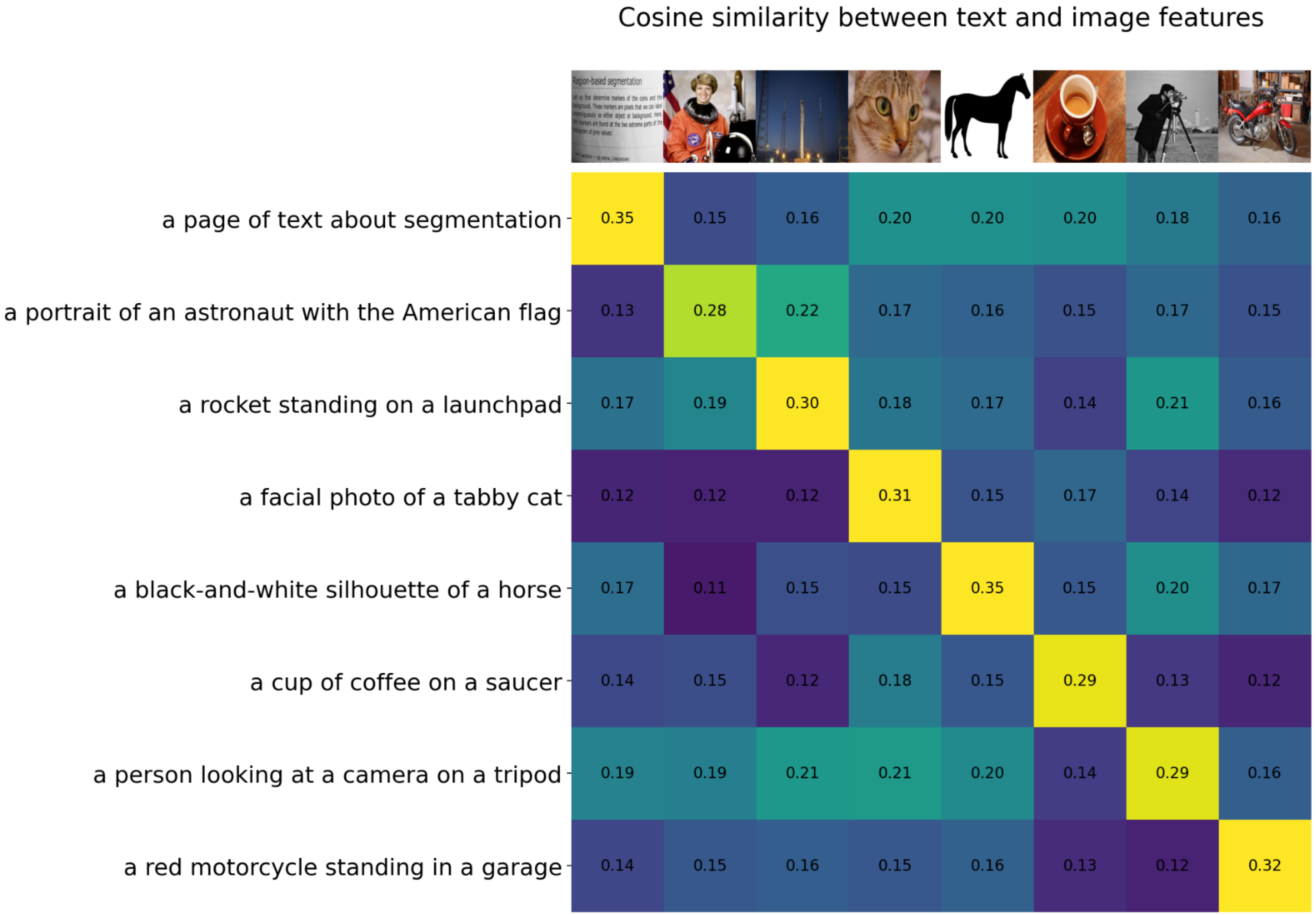


More Analysis on Text-Image Retrieval





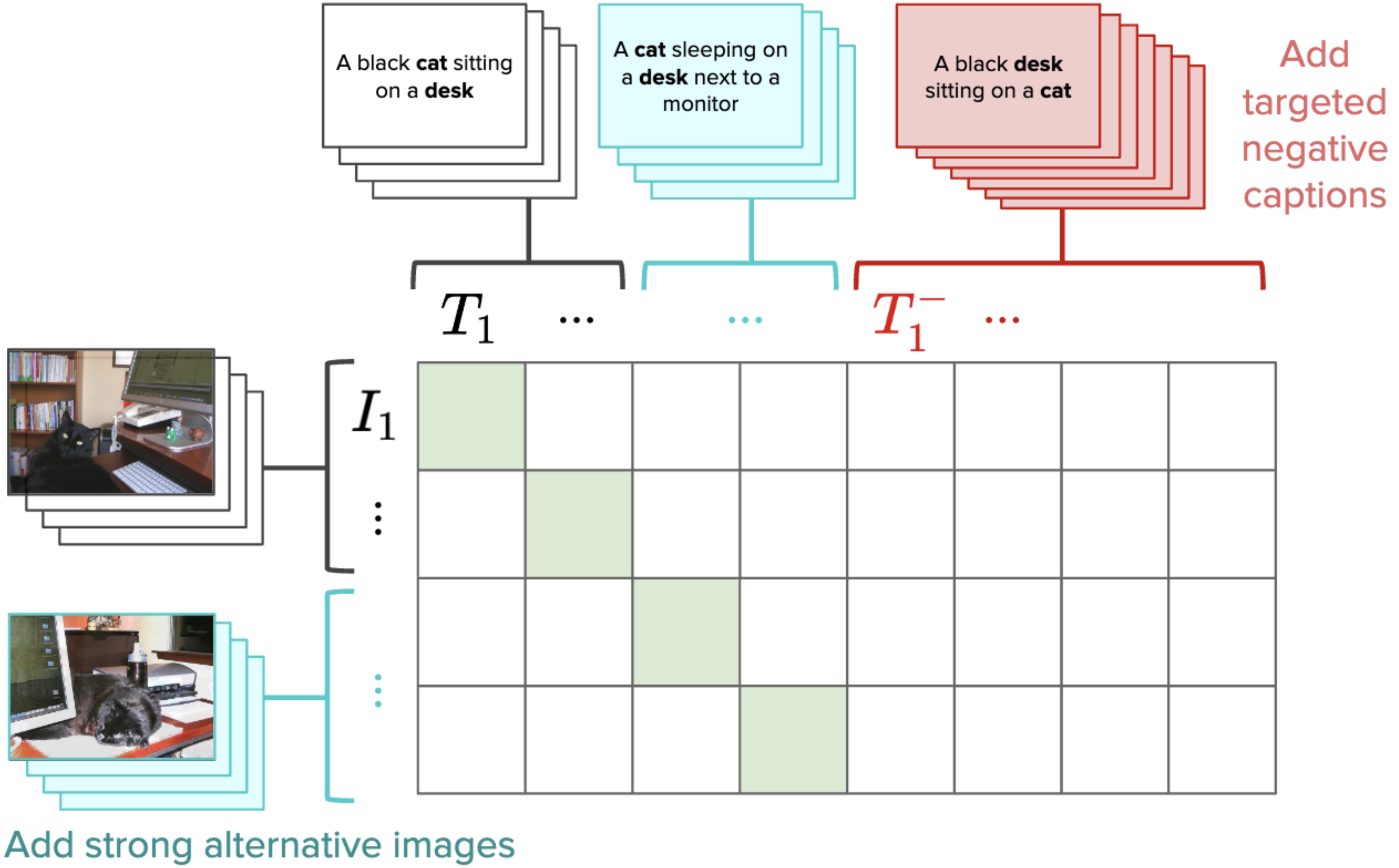
Contrastive Pre-Training



Solution: Composition-Aware Hard Negatives

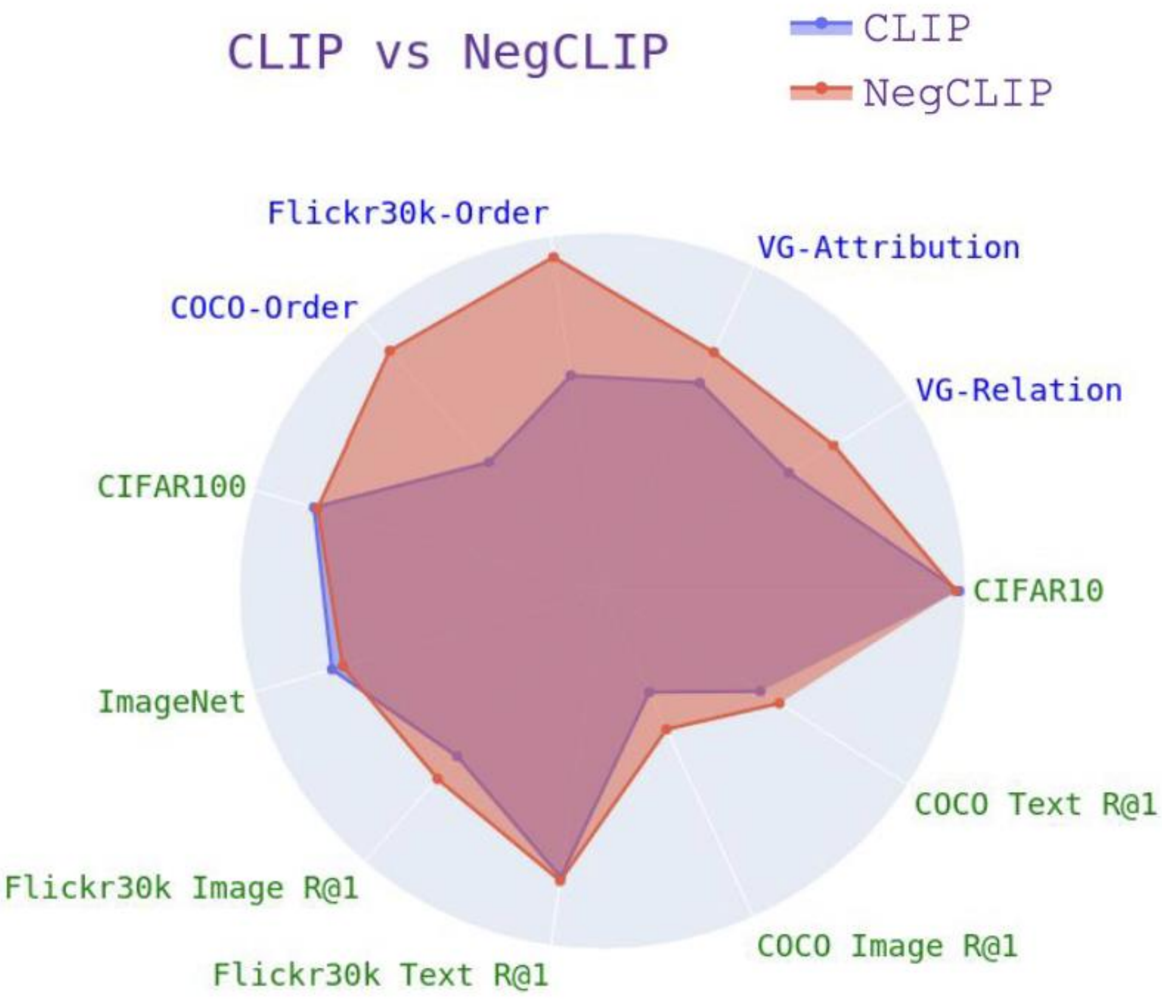
- Generation of negative captions
  - Generate a negative caption by swapping different linguistic elements: noun phrases, nouns, adjectives, adverbs, verb phrases
- Sampling strong alternative images
  - Sample one of the K nearest neighbors as the strong alternative image

NegCLIP

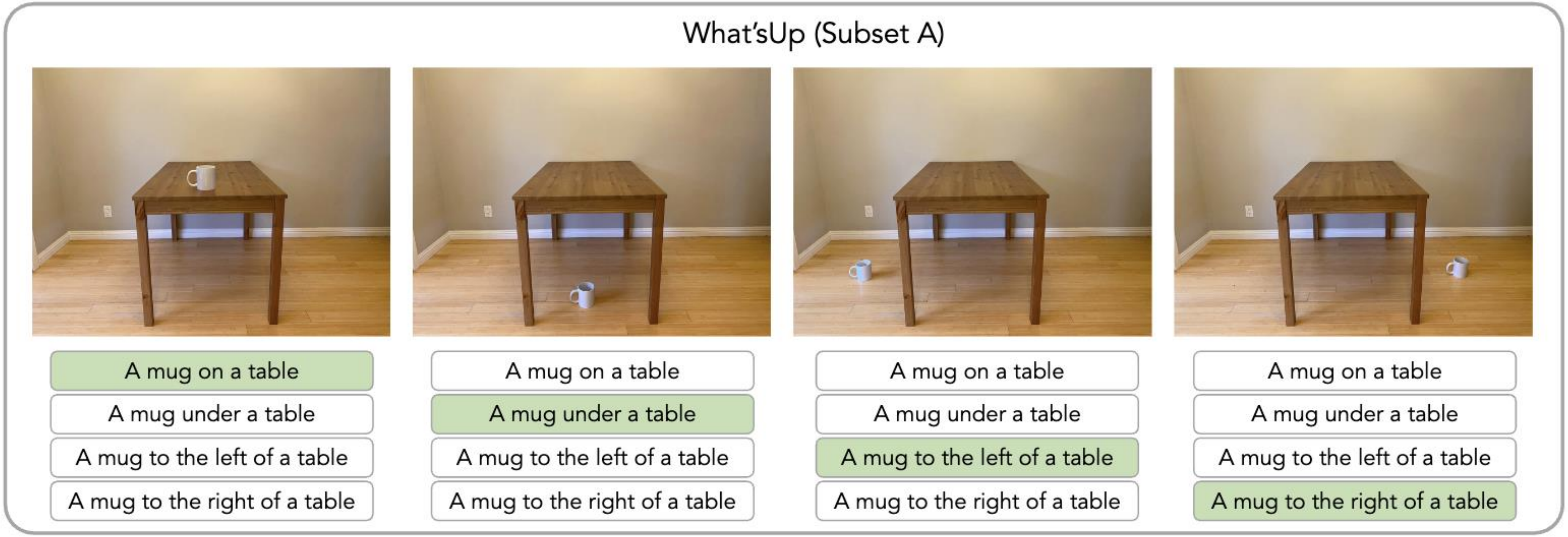


A black **cat** sitting on a **desk**
A **cat** sleeping on a **desk** next to a monitor
A black **desk** sitting on a **cat**
Add targeted negative captions
T_1 …
…
T_1^- …
I_1
⋮
⋮
Add strong alternative images

Results

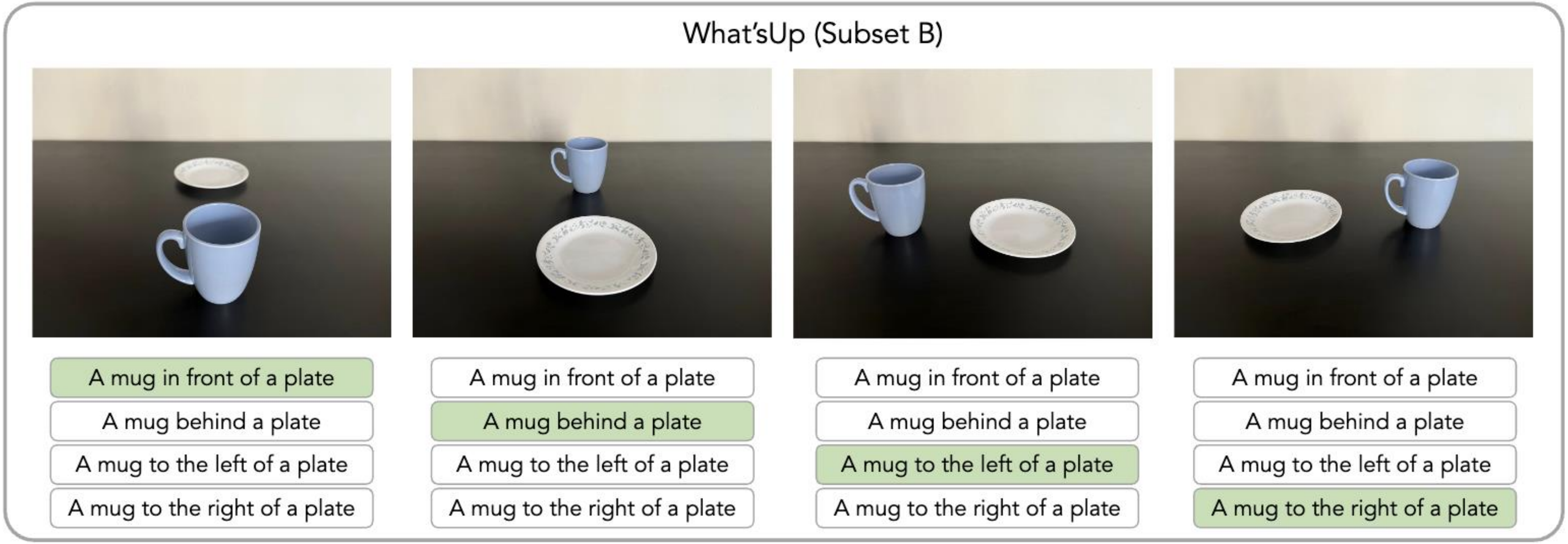

CLIP vs NegCLIP
CLIP
NegCLIP
Flickr30k-Order
VG-Attribution
COCO-Order
VG-Relation
CIFAR100
CIFAR10
ImageNet
COCO Text R@1
Flickr30k Image R@1
Flickr30k Text R@1
COCO Image R@1

What’s Up: Benchmarks with Spatial Reasoning Questions



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

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%

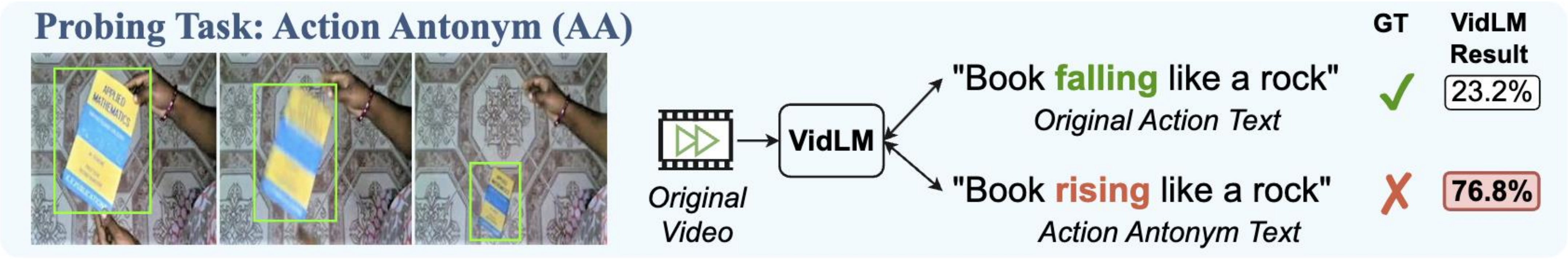
Visual Analogies

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

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

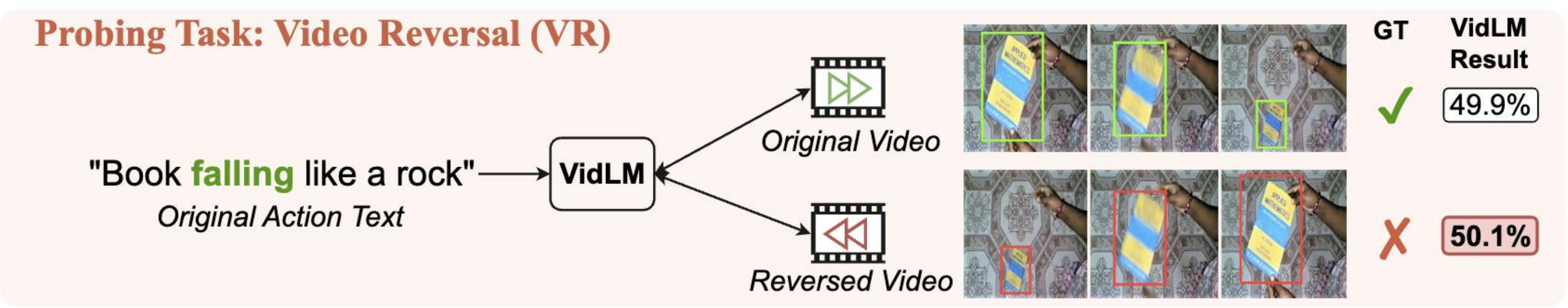
9%

Paxion: Action Robustness for Video-Language Models



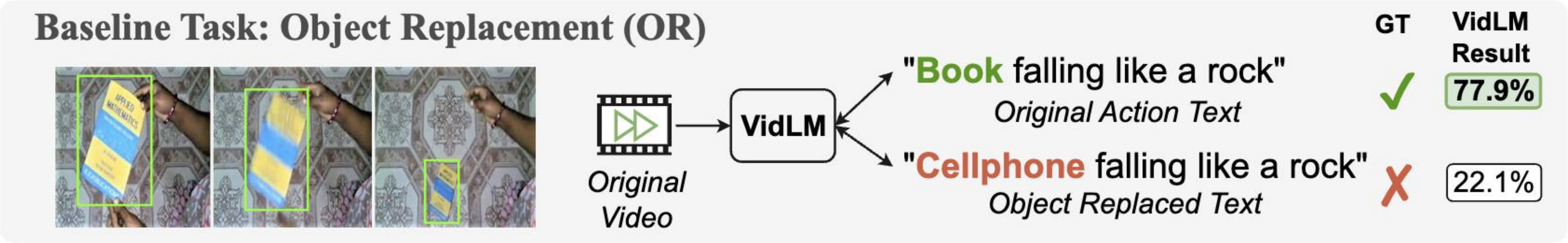
Paxion: Patching Action Knowledge in Video-Language Foundation Models

Paxion: Action Robustness for Video-Language Models



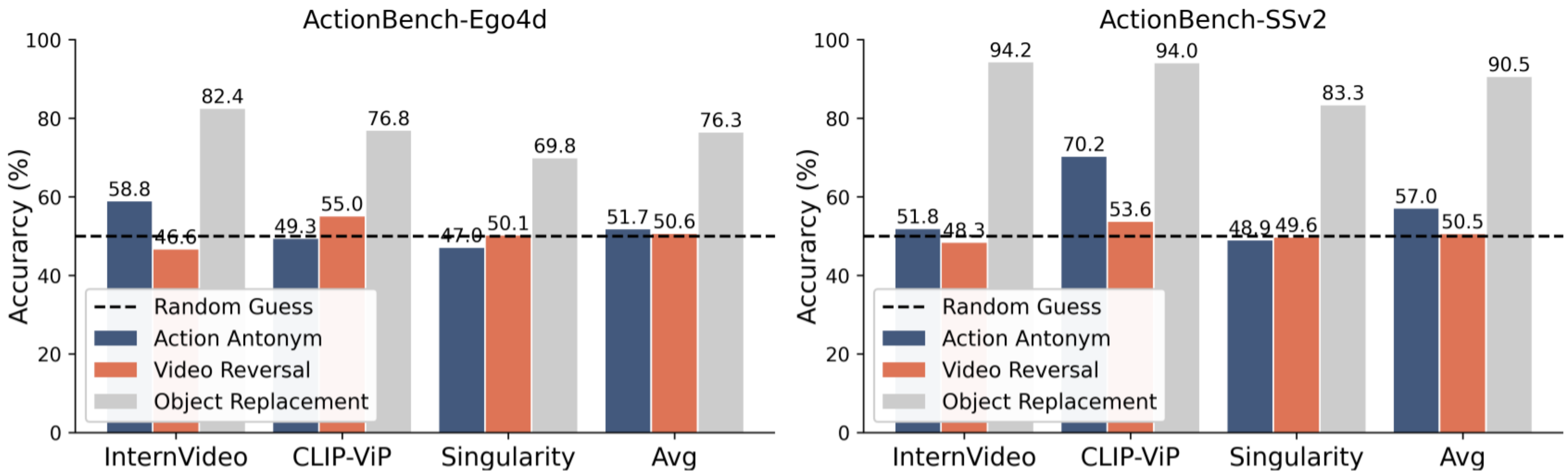
Paxion: Patching Action Knowledge in Video-Language Foundation Models

Paxion: Action Robustness for Video-Language Models



Paxion: Patching Action Knowledge in Video-Language Foundation Models

Results



ActionBench-Ego4d
Accuracy (%)
100
80
60
40
20
0
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 (%)
100
80
60
40
20
0
94.2
94.0
90.5
83.3
70.2
51.8
48.3
53.6
48.9
49.6
57.0
50.5
Random Guess
Action Antonym
Video Reversal
Object Replacement
InternVideo
CLIP-ViP
Singularity
Avg

Qwen Series

Alibaba Qwen

https://qwen.ai/

Qwen3.5: Towards Native Multimodal Agents

Qwen3.5: Towards Native Multimodal Agents

We are delighted to announce the official release of Qwen3.5, introducing the open-weight of the first model in the Qwen3.5 series, namely Qwen3.5-397B-A17B. As a native vision-language model,...

Open-Source | 2026/02/15

Qwen-Image-2.0: Professional infographics, exquisite photorealism

Qwen-Image-2.0: Professional infographics, exquisite p...

We are launching Qwen-Image-2.0, a next-generation foundational image generation model. The key highlights of Qwen-Image-2.0 include: Professional Typography Rendering: Supports 1k-token instructions for...

Open-Source | 2026/02/09

Qwen3-Coder-Next: Pushing Small Hybrid Models on Agentic Coding

Qwen3-Coder-Next: Pushing Small Hybrid Models on A...

--- We introduce Qwen3-Coder-Next, an open-weight language model designed specifically for coding agents and local development. Built on top of Qwen3-Next-80B-A3B-Base, which adopts a novel architecture...

Open-Source | 2026/02/02

Qwen3-ASR & Qwen3-ForcedAligner is Now Open Sourced: Robust, Streaming and Multilingual!

Qwen3-ASR & Qwen3-ForcedAligner is Now Open Sour...

Qwen3-ASR family includes two powerful all-in-one speech recognition models and a novel non-autoregressive speech forced alignment model. Qwen3-ASR-1.7B and Qwen3-ASR-0.6B are ASR models that support...

Open-Source | 2026/01/28

Widely used in academic research

Multilingual and multimodal support

My personal take: Probably the strongest open-weight LLM right not

https://qwen.ai/research