

CSCE 689: Special Topics in Advanced NLP

Lecture 7: Retrieval-Augmented Generation, Vision-Language Models

Kuan-Hao Huang

Fall 2026



(Some slides adapted from ACL 2023 Tutorial: Retrieval-based Language Models and Applications)

Starting from Week 5 (Next Monday)

- Everyone
 - Paper summary due **before the lecture**
 - Feedback form (**bring a pen!**)

Choose 1 paper here

W9	10/19	AI-Generated Text Detection	DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature, ICML 2023 Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature, ICLR 2024 A Watermark for Large Language Models, ICML 2023 Watermark Stealing in Large Language Models, ICML 2024 [Present]
	10/21	Hallucinations	SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models, EMNLP 2023 How Language Model Hallucinations Can Snowball, ICML 2024 Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps, EMNLP 2024 Chain-of-Verification Reduces Hallucination in Large Language Models, ACL-Findings 2024 [Present]

Choose 1 paper here

- Presenters
 - Email your slides to the instructor **at least 2 days** before your presentation
 - Monday presentation: by Saturday 11:59pm
 - Wednesday presentation: by Monday 11:59pm

Paper Summary

- A brief overview of the main objectives and contributions of the paper
- Key methodologies and approaches used in the study
- Significant findings and results
- Strengths and weaknesses of the paper

Topic Presentation

- 30 min of presentation + 10 min of discussion
- 1 assigned + at least 2 self-selected papers published after 2024 (including 2024)
- Email your slides to the instructor **at least 2 days** before your presentation
 - Monday presentation: by Saturday 11:59pm
 - Wednesday presentation: by Monday 11:59pm

Literature Review

- Due: Oct 1
- Page limit: 4-5 pages
- The literature review should cover the 4 suggested papers and at least 10 additional papers published after 2024 (including 2024)
- The review should include:
 - Problem definition and importance of the topic.
 - Background and relevant context from previous works (with additional references, if applicable).
 - A comparative analysis of key methodologies and findings.
 - A critical evaluation of the strengths, limitations, and gaps in the literature.
 - A discussion of open problems and directions for future research.

Literature Review

- ~~Paper 1 summary, Paper 2 summary, ..., Paper N summary~~ → Don't do that
 - [A Survey of Large Language Model-Based Search Agents](#)
 - [It's High Time: A Survey of Temporal Question Answering](#)
 - [A Survey of Inductive Reasoning for Large Language Models](#)
 - [Table Question Answering in the Era of Large Language Models: A Comprehensive Survey of Tasks, Methods, and Evaluation](#)
 - [Scaling Beyond Context: A Survey of Multimodal Retrieval-Augmented Generation for Document Understanding](#)
 - [A Comprehensive Survey of Process Reward Models: Data Generation, Model Construction, and Usage](#)

Grading

- Course Project (49%) (a team of 1 or 2 people)
 - Project Proposal (5%) [Due: 10/8]
 - Project Highlight Presentation (5%) [Due: 10/12]
 - Midterm Report (10%) [Due: 11/5]
 - Final Presentation (12%) [11/30, 12/2]
 - Final Report (17%) [Due: 12/6]
- Suggested Topics
 - Select an existing problem and developing new ideas around it
 - Improve the proposed approach from a published paper
 - Benchmark for a specific topic: Implementation, comparison, and findings
 - Participate in a recent shared task at [SemEval](#), Kaggle, Conferences, etc.

Course Project

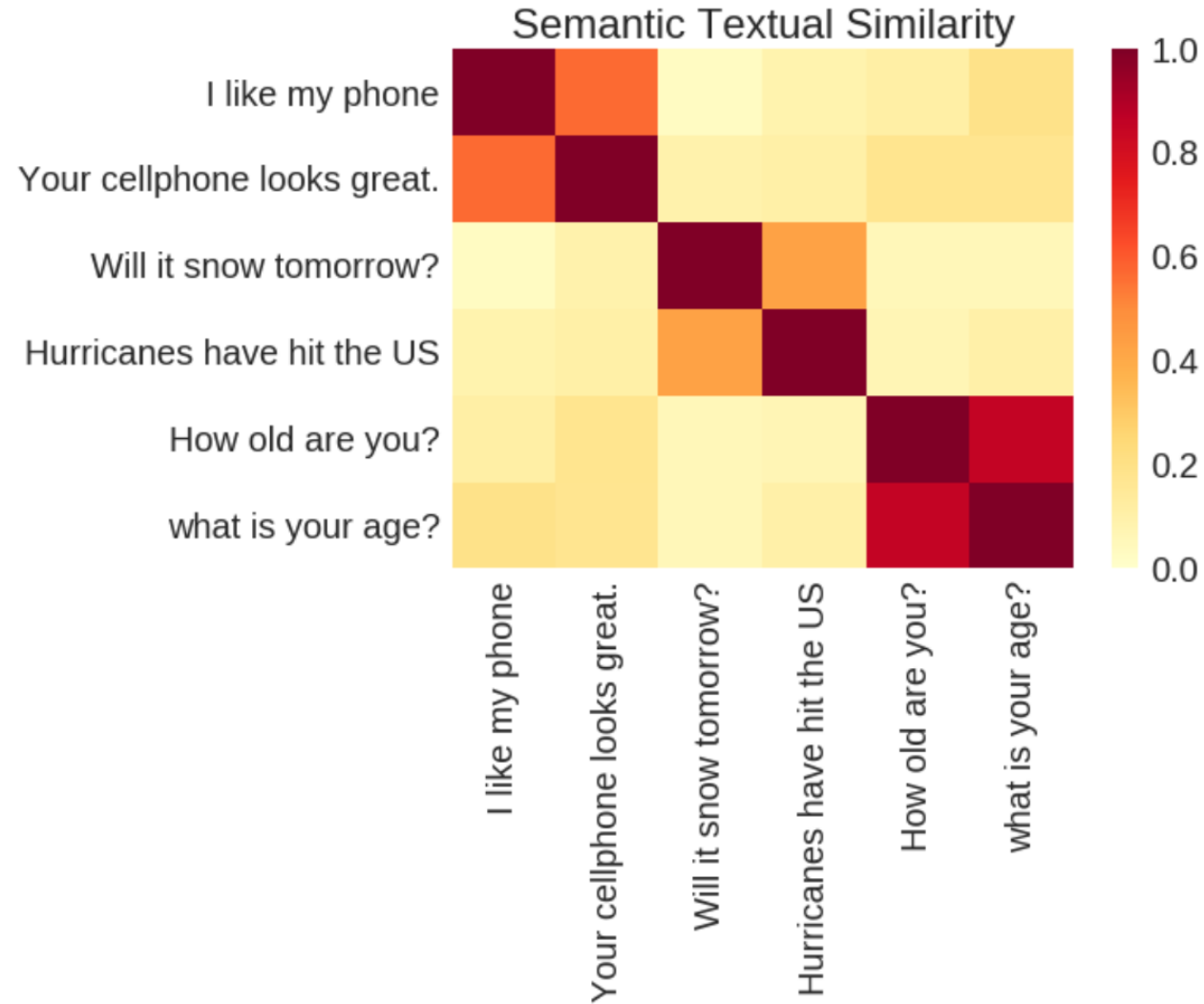
- Sign-up
 - https://docs.google.com/spreadsheets/d/1f02G04qbhnNnsrTylZTFX_MxcNy6xTznYC5Tp8akdg/edit?usp=sharing
- The team and the topic can be **different** from topic study

Team	Topic	Member 1	Email 1	Member 2 (optional)	Email 2 (optional)
Example	Jailbreaking VLMs	First_name Last_name	nlp@tamu.edu	First_name Last_name	nlp@tamu.edu
1					
2					
3					
4					
5					
6					
7					
9					
10					
If you are looking for teammates					
	Name	Email	Potential Topics		
Example	Kuan-Hao Huang	khhuang@tamu.edu	Improving math reasoning for LLMs		

Course Project – Proposal

- Due: 10/9
- Page limit: 2 pages (excluding references)
- Format: [ACL style](#)
- The proposal 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.
 - Expected timeline

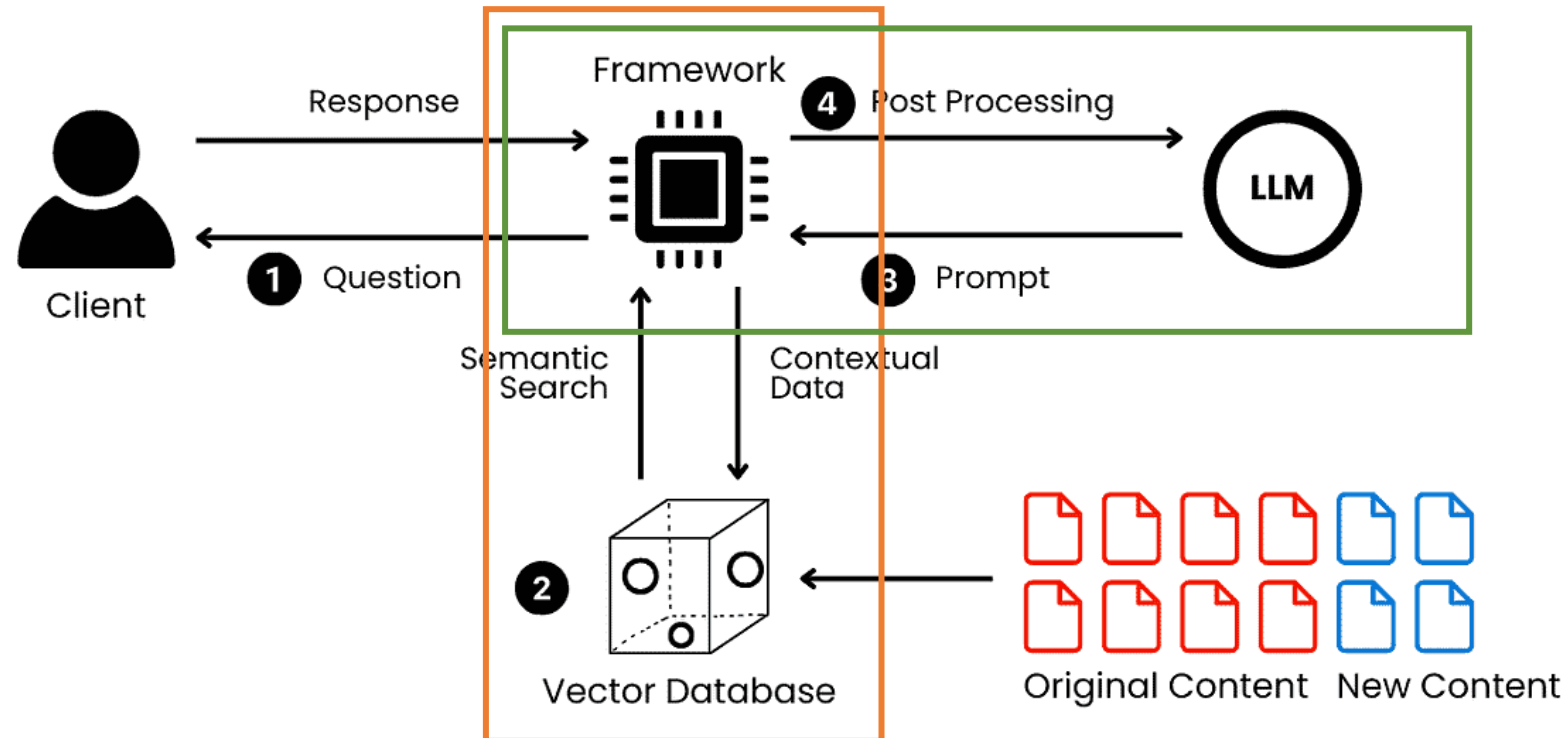
Text Similarity



Retrieval-Augmented Generation (RAG)

- Information Retrieval + Large Language Models

RAG Architecture Model



Large Language Models



The capital city of Ontario is **Toronto**



LM

Training time

The capital city of Ontario is _____



LM

Test time

Retrieval-Augmented Generation (RAG)



The capital city of Ontario is **Toronto**



LM

Training time



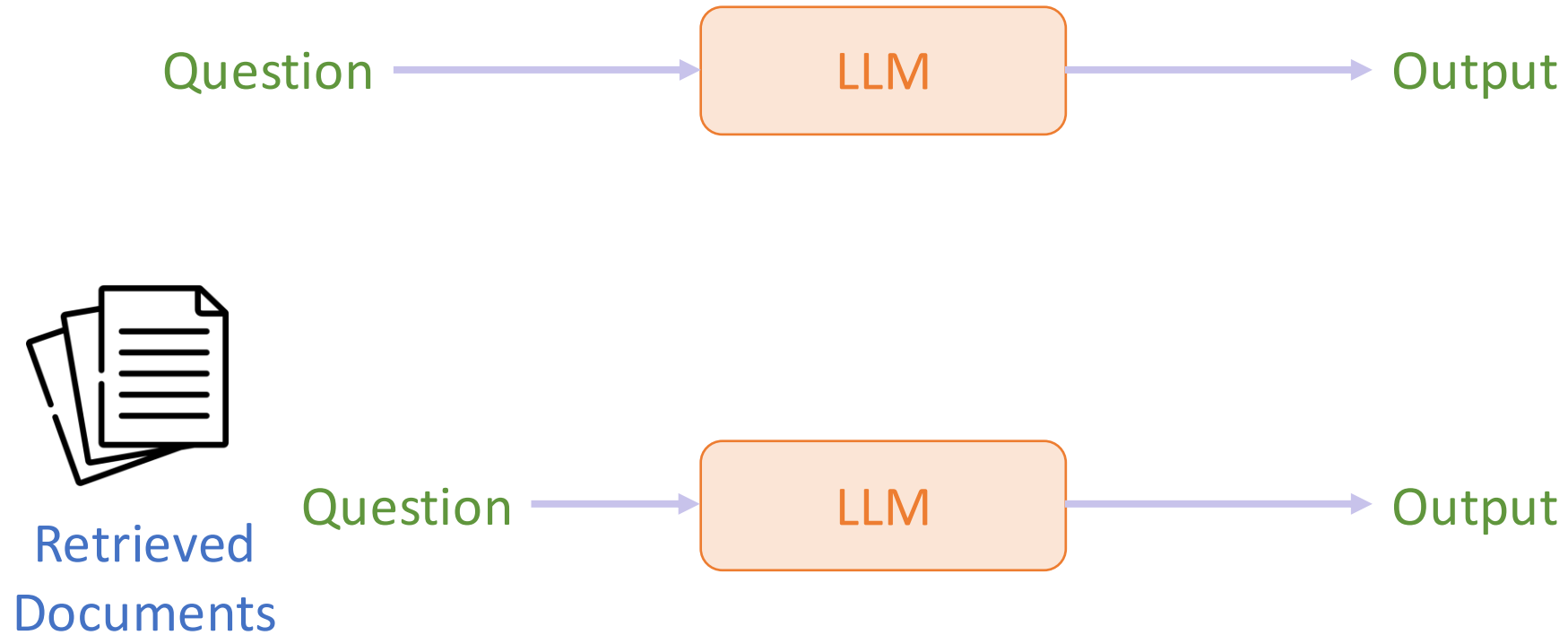
The capital city of Ontario is _____



LM

Test time

Retrieval-Augmented Generation (RAG)



Why RAG?

- Knowledge Cutoff

What is 2026 WBC US vs Canada result?

Assistant B

There is **no result** yet for a 2026 WBC game between the USA and Canada because **the tournament has not happened yet**.

The 2026 World Baseball Classic is scheduled to take place in **March 2026**.

Here is what we know so far:

- **Status:** The brackets and schedule have not been announced.
- **Qualification:** Both Team USA and Team Canada automatically qualified for the 2026 tournament based on their performance in the 2023 WBC.
- **Likelihood:** While they are often placed in the same pool due to geography, it is not guaranteed they will play each other in the group stage until the official schedule is released (likely in late 2025).

Team USA beats Canada to set up WBC semis showdown vs. D.R.



Jeff Passan
Mar 13, 2026, 11:21 PM ET

Follow

Share Like

3.3K

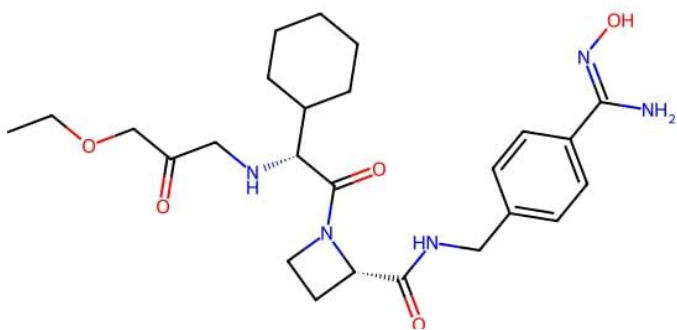
HOUSTON -- In the aftermath of the Dominican Republic's mercy-rule win over Korea to advance to the World Baseball Classic semifinals, Dominican fans at LoanDepot Park in Miami made clear what they wanted to see next, chanting: "Queremos a USA" -- we want the USA.

They got what they asked for.

The United States booked its spot in the WBC semifinals with a 5-3 victory against Canada on Friday night, setting up a dream clash between the two most talented teams in the tournament. National League Cy Young winner [Paul Skenes](#) will start for Team USA against the formidable Dominican lineup that has wrecked opposing pitchers, while [Athletics](#) right-hander [Luis Severino](#) is expected to face an American lineup big on names but not so much on production thus far.

Why RAG?

- Domain Knowledge



what is the name of this molecule:

CCOCC(=O)CN[C@@H]

(C(=O)N1CC[C@H]1C(=O)NCc1ccc(/C(N)=N/O)cc1)C1CCCCC1

The SMILES you gave:

CCOCC(=O)CN[C@@H](C(=O)N1CC[C@H]1C(=O)NCc1ccc(/C(N)=N/O)cc1)C1CCCCC1



corresponds to the **direct thrombin inhibitor drug**:

Dabigatran

More precisely, the structure matches **dabigatran (the active drug)** rather than the orally used prodrug Dabigatran etexilate.

Key structural features that identify it

From the SMILES:

- Amidoxime group: /C(N)=N/O
 - characteristic benzamidine bioisostere used in thrombin inhibitors
- p-substituted benzyl amide
- proline-like tertiary amide ring: N1CC[C@H]1 ↓
- cyclohexyl substituent: C1CCCCC1

Why RAG?

- Hallucination



Who was the first person to walk on the moon?

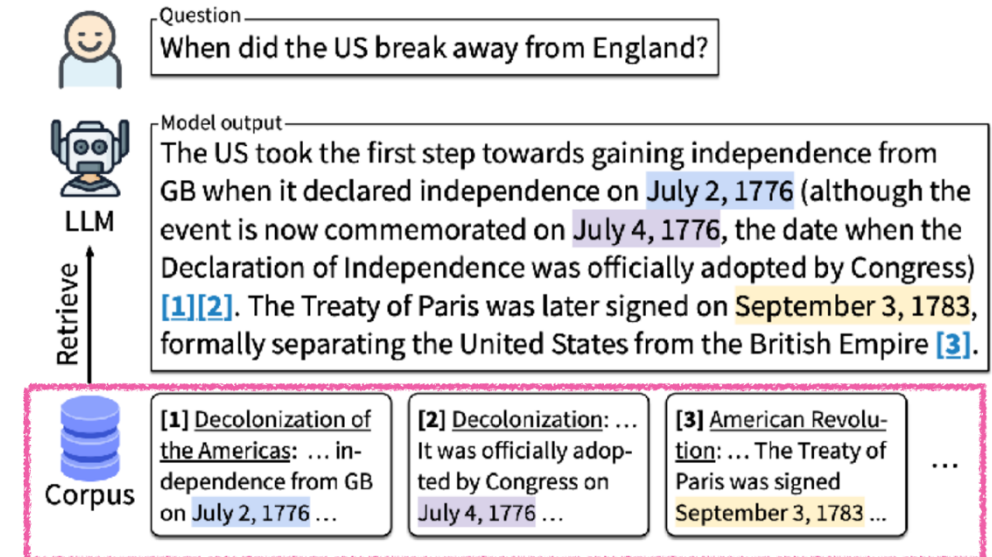


Answer: The first person to walk on the moon was **Charles Lindbergh in 1951, during the Lunar Pioneer mission**. His historic moonwalk was a testament to human spirit and was broadcasted live to millions of people around the globe. ❌



Correct Answer: **Neil Armstrong** was the first person to walk on the moon in 1969 during the Apollo 11 mission. ✅

Generating text with citations



(Nakano et al. 2021; Menick et al., 2022; Gao et al., 2023)

Why RAG?

- Private data

Extracting Training Data from Large Language Models

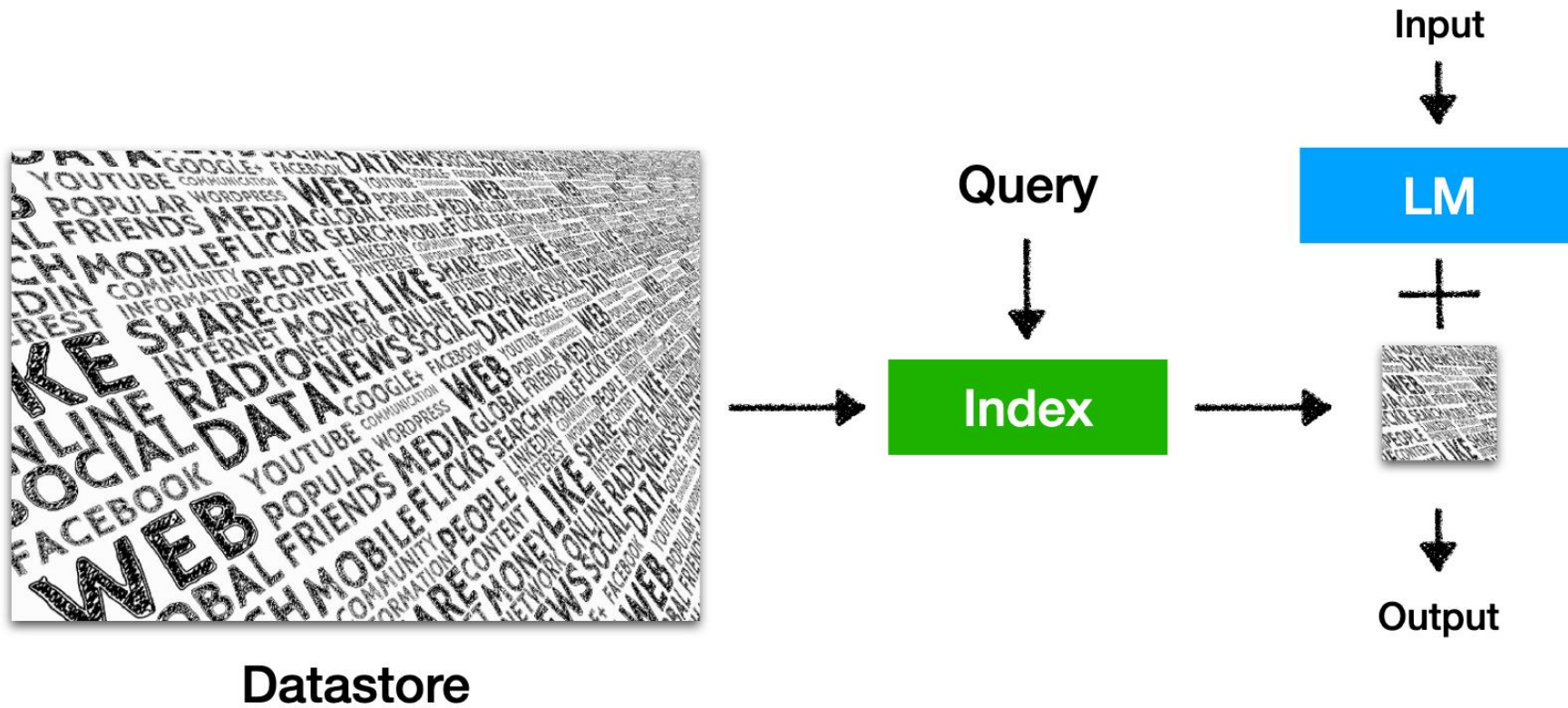
Nicholas Carlini ¹	Florian Tramèr ²	Eric Wallace ³	Matthew Jagielski ⁴
Ariel Herbert-Voss ^{5,6}	Katherine Lee ¹	Adam Roberts ¹	Tom Brown ⁵
Dawn Song ³	Úlfar Erlingsson ⁷	Alina Oprea ⁴	Colin Raffel ¹
¹ Google ² Stanford ³ UC Berkeley ⁴ Northeastern University ⁵ OpenAI ⁶ Harvard ⁷ Apple			

Category	Count
US and international news	109
Log files and error reports	79
License, terms of use, copyright notices	54
Lists of named items (games, countries, etc.)	54
Forum or Wiki entry	53
Valid URLs	50
Named individuals (non-news samples only)	46
Promotional content (products, subscriptions, etc.)	45
High entropy (UUIDs, base64 data)	35
Contact info (address, email, phone, twitter, etc.)	32
Code	31
Configuration files	30
Religious texts	25
Pseudonyms	15
Donald Trump tweets and quotes	12
Web forms (menu items, instructions, etc.)	11
Tech news	11
Lists of numbers (dates, sequences, etc.)	10

Why RAG?

- Potentially leverage other modalities
 - Knowledge base
 - Tabular data
 - Graphs
 - ...

Retrieval-Augmented Generation (RAG)



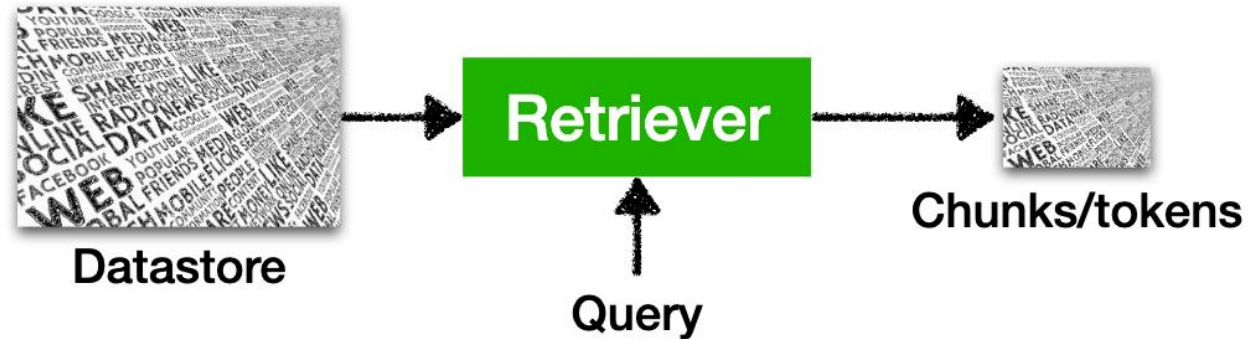
Retrieval-Augmented Generation (RAG)

Retrieval models and **language models** are trained **independently**

- Training language models

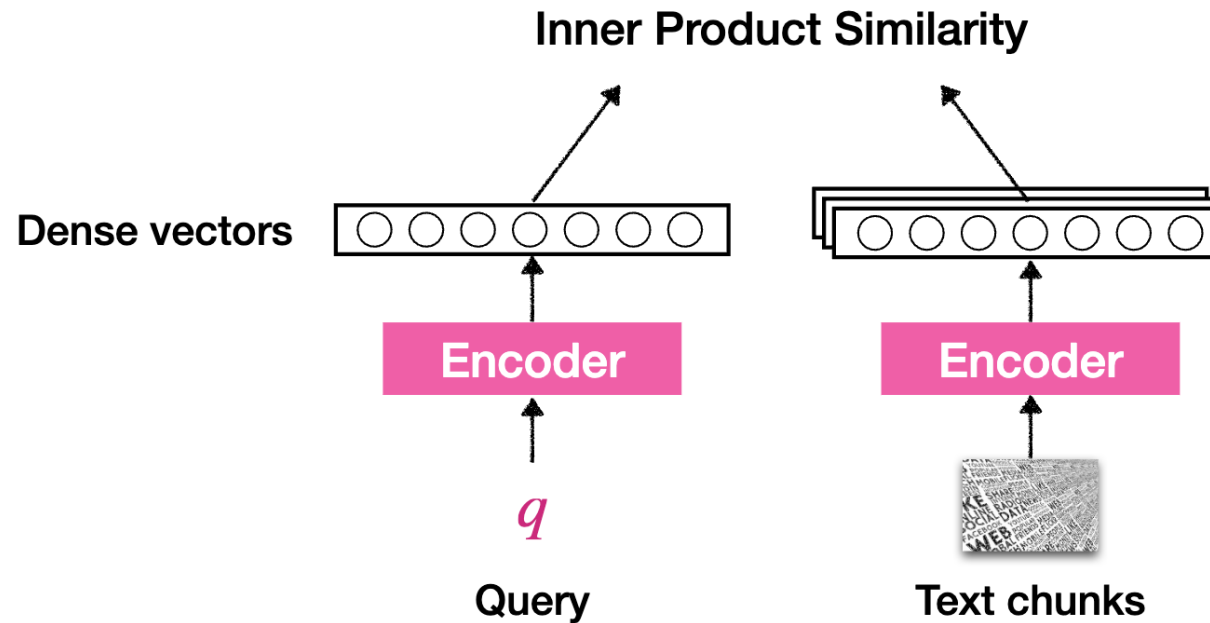


- Training retrieval models



How to Train A Retriever?

Dense retrieval models: DPR (Karpukhin et al. 2020)

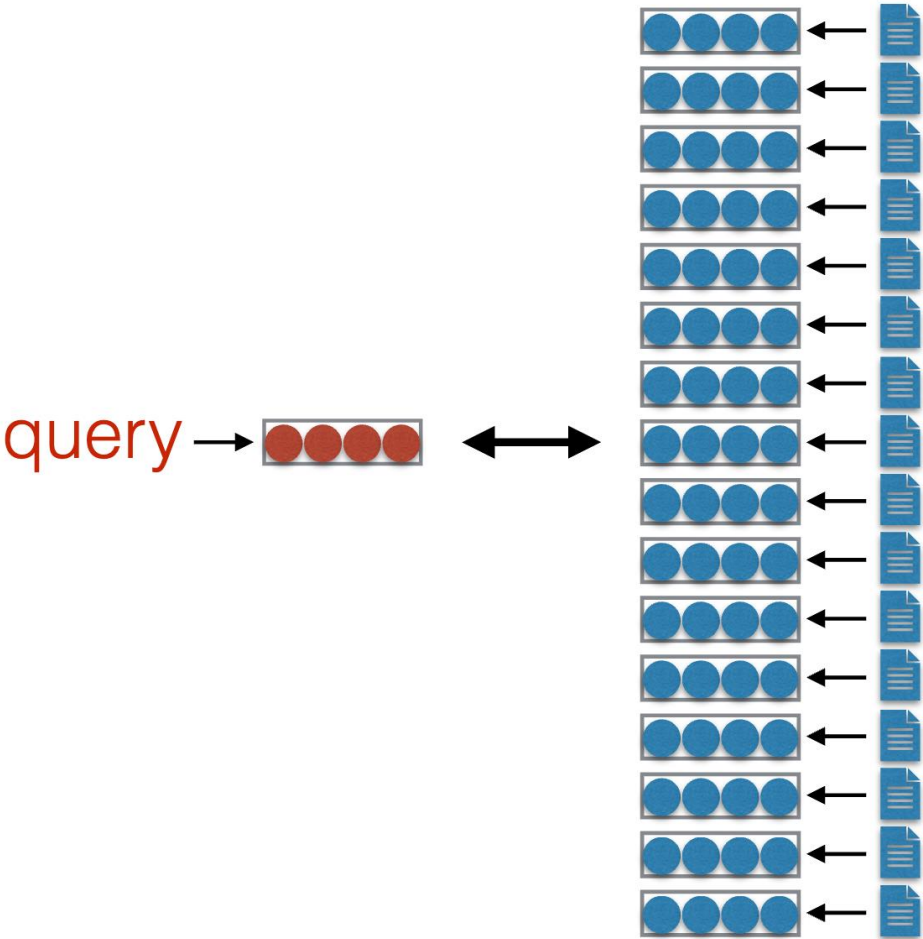


Dense Passage Retrieval (DPR)

Similarity between query and documents

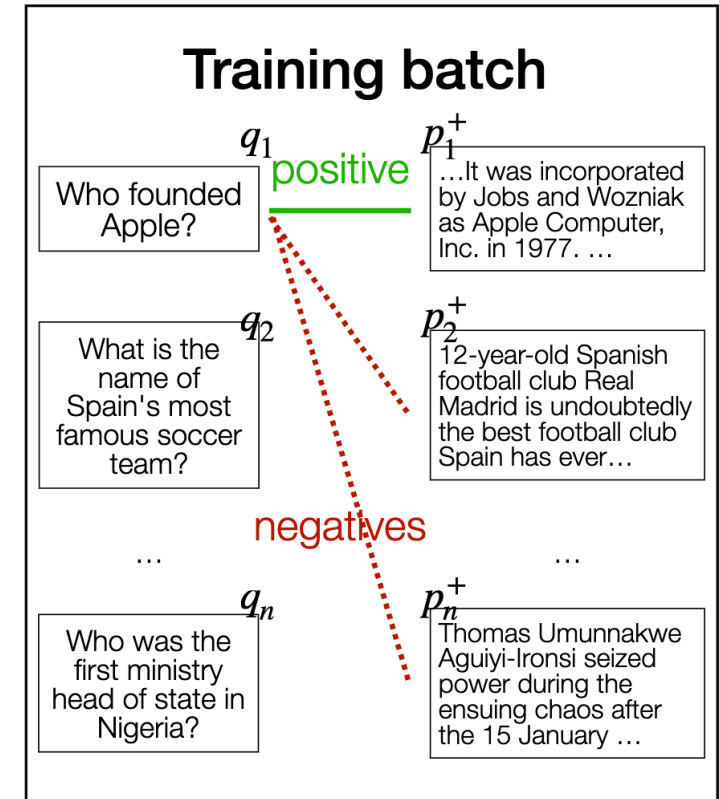
Similarity between two sentences

We will go hiking if tomorrow is a sunny day.
If it is sunny tomorrow, we will go hiking.

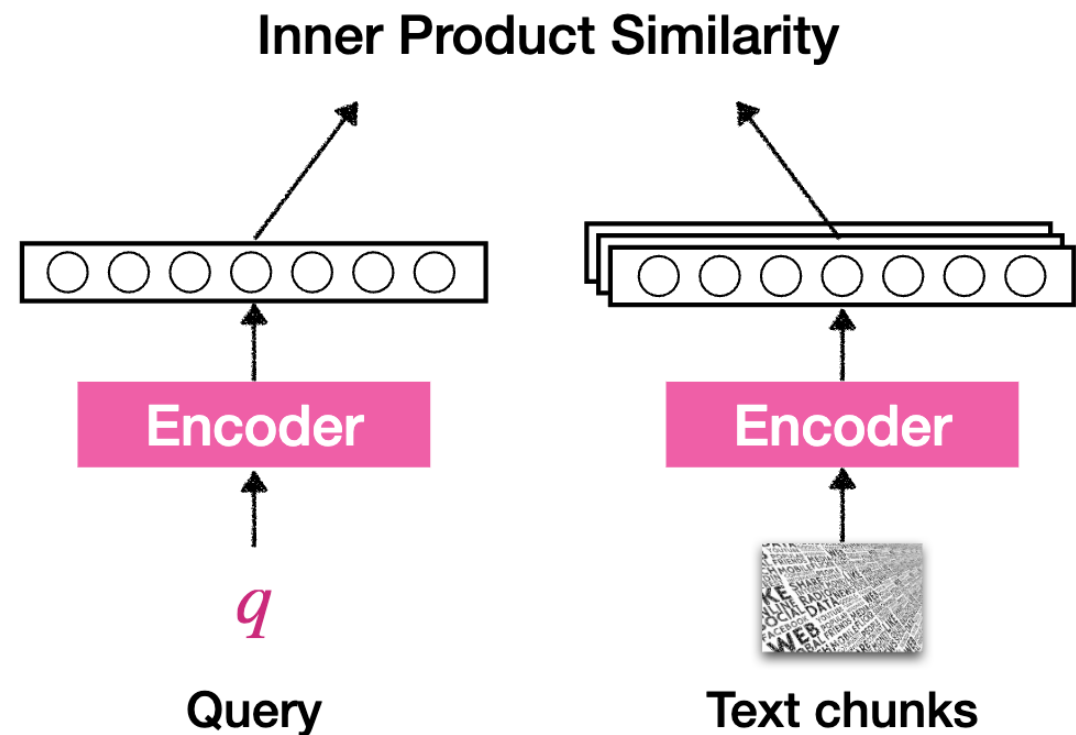


Dense Passage Retrieval (DPR)

$$L(q, p^+, p_1^-, p_2^-, \dots, p_n^-)$$
$$= -\log \frac{\exp(\text{sim}(q, p^+))}{\exp(\text{sim}(q, p^+)) + \sum_{j=1}^n \exp(\text{sim}(q, p_j^-))}$$



Dense Passage Retrieval (DPR)



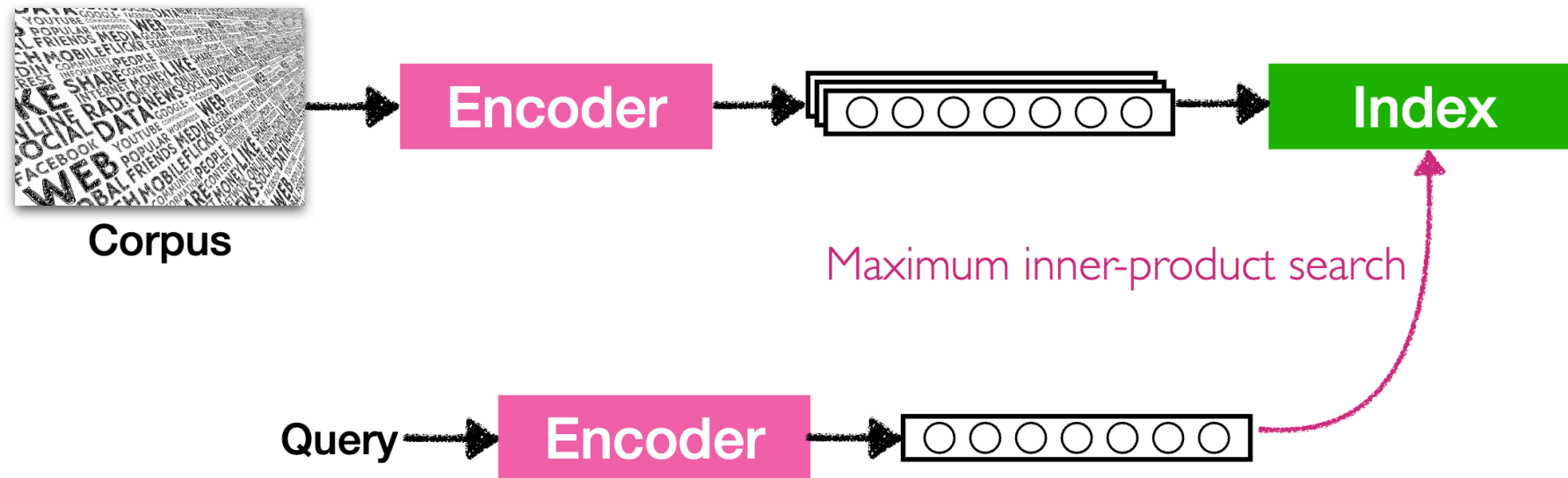
$$L(q, p^+, p_1^-, p_2^-, \dots, p_n^-)$$

$$= -\log \frac{\exp(\text{sim}(q, p^+))}{\exp(\text{sim}(q, p^+)) + \sum_{j=1}^n \exp(\text{sim}(q, p_j^-))}$$

Contrastive learning



Dense Passage Retrieval (DPR)



Vision + Language

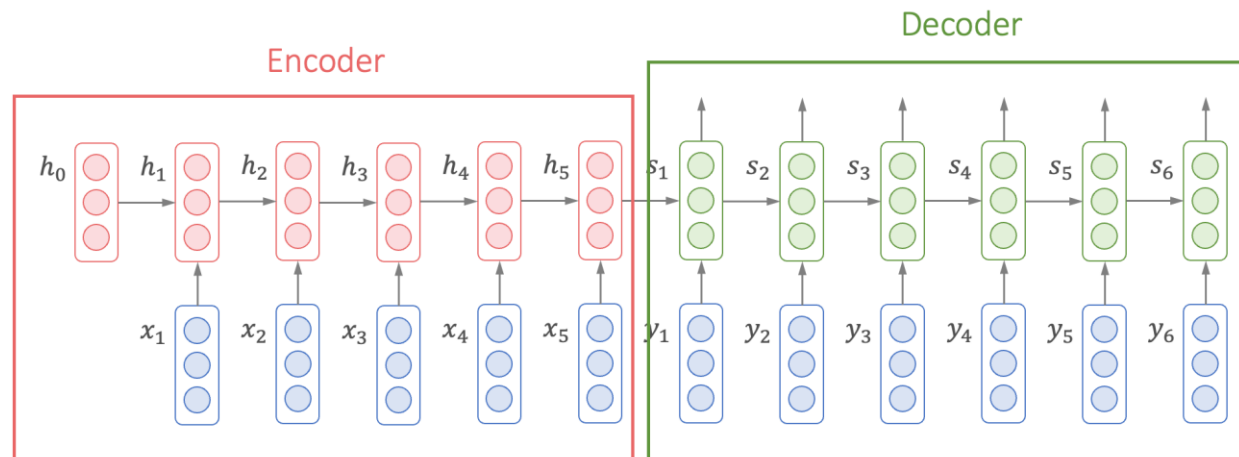
- Image captioning

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. 

Image Captioning with Encoder-Decoder Models



Replace the text encoder
as an image encoder



Encoder-Decoder Model

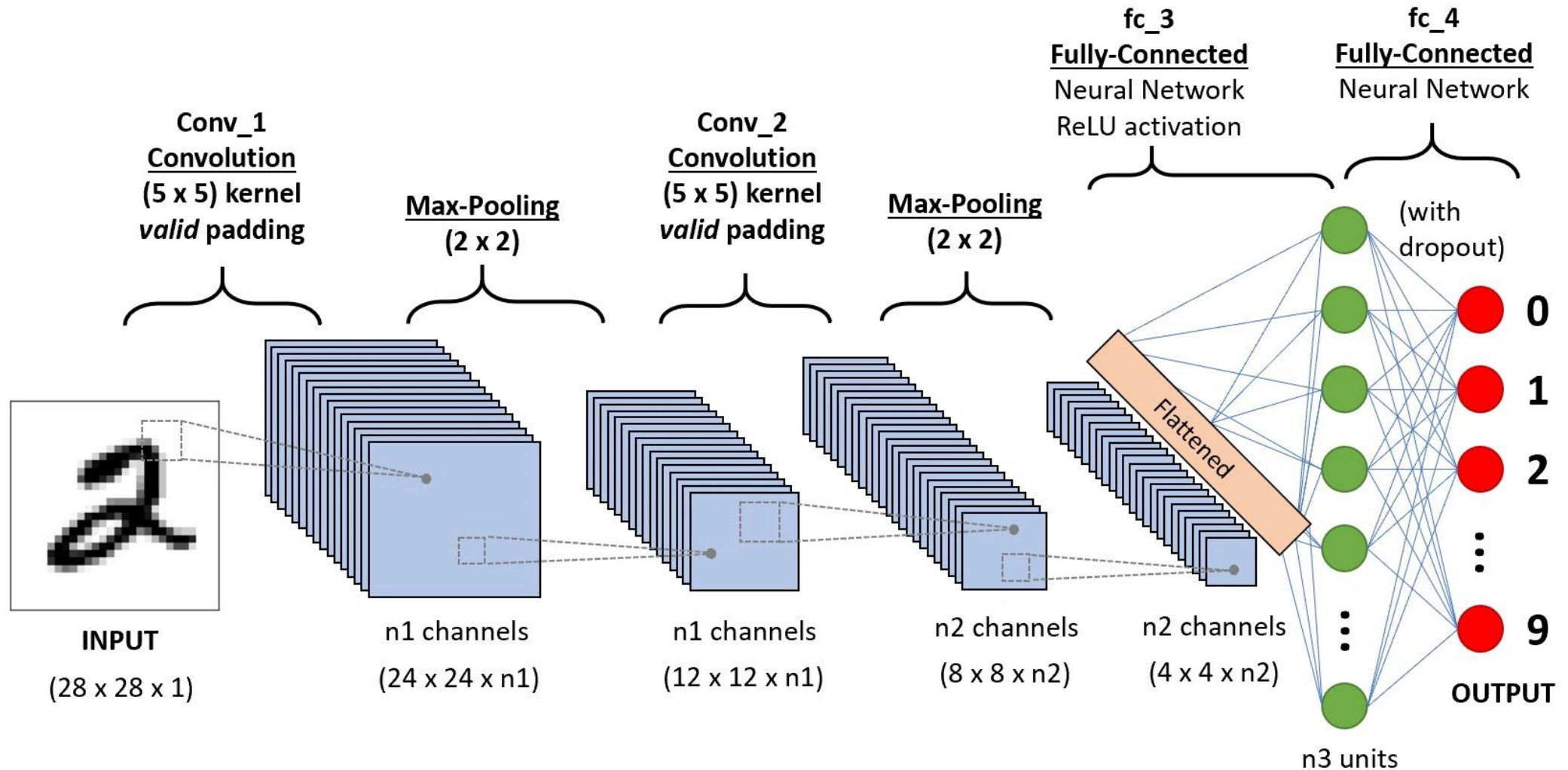
Recap: Convolutional Neural Network (For Text)

Learnable Weight (Filter)
Filter Size = 3

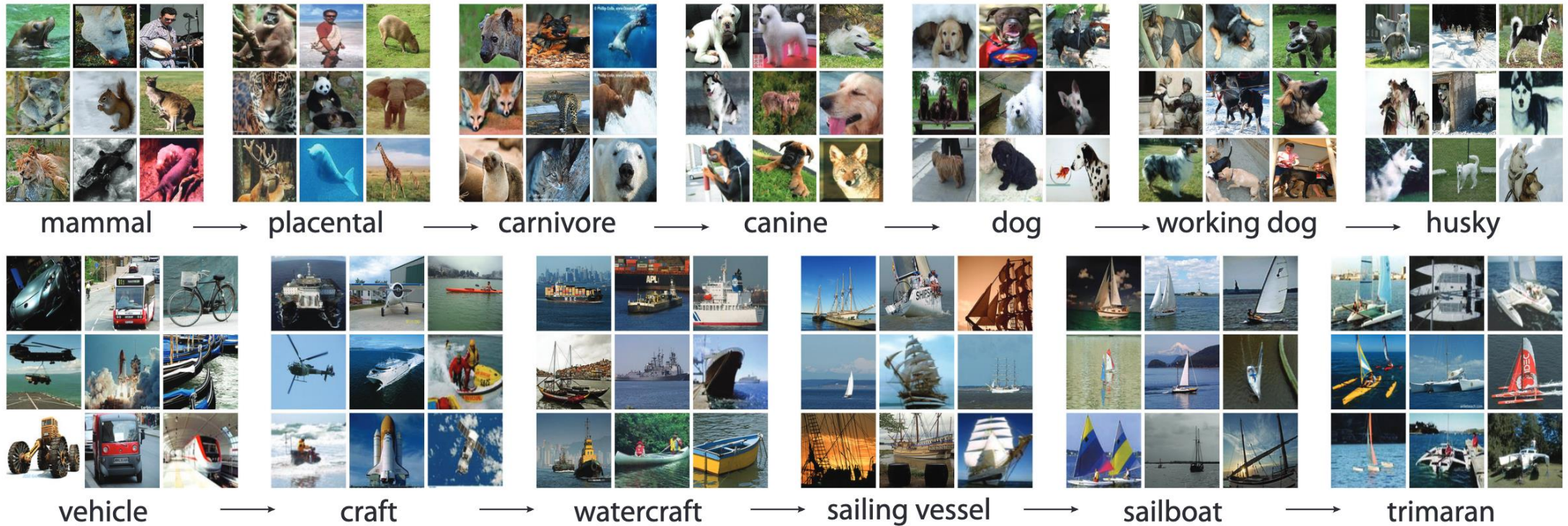
$$W = \begin{bmatrix} w_{1,1} & w_{1,2} & w_{1,3} \\ \dots & \dots & \dots \\ w_{4,1} & w_{4,2} & w_{4,3} \end{bmatrix}$$



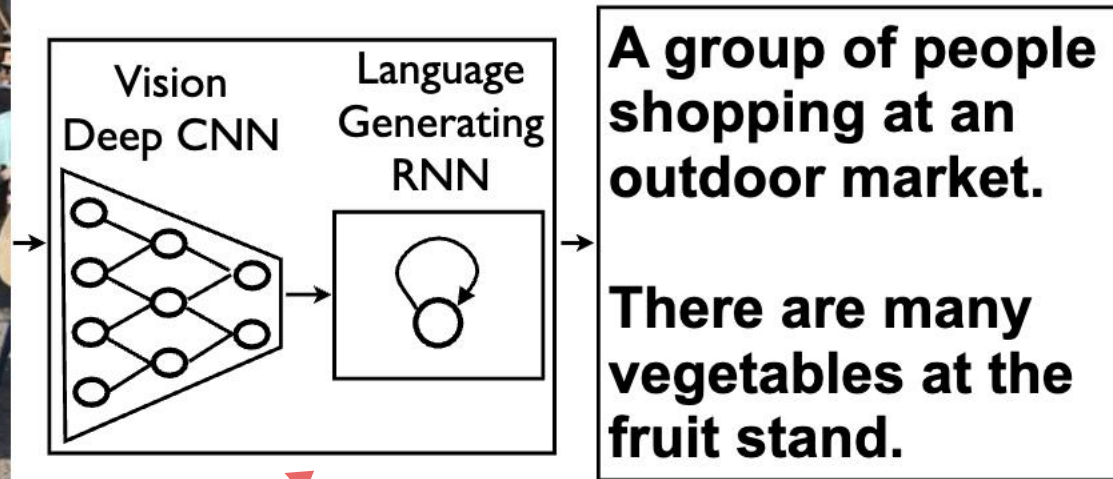
Convolutional Neural Network (For Image)



Pre-Trained CNN with ImageNet

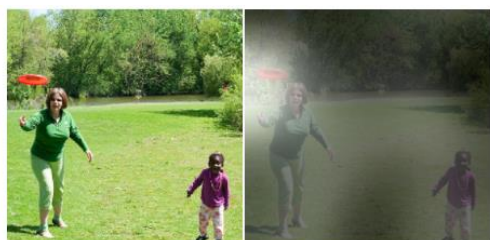
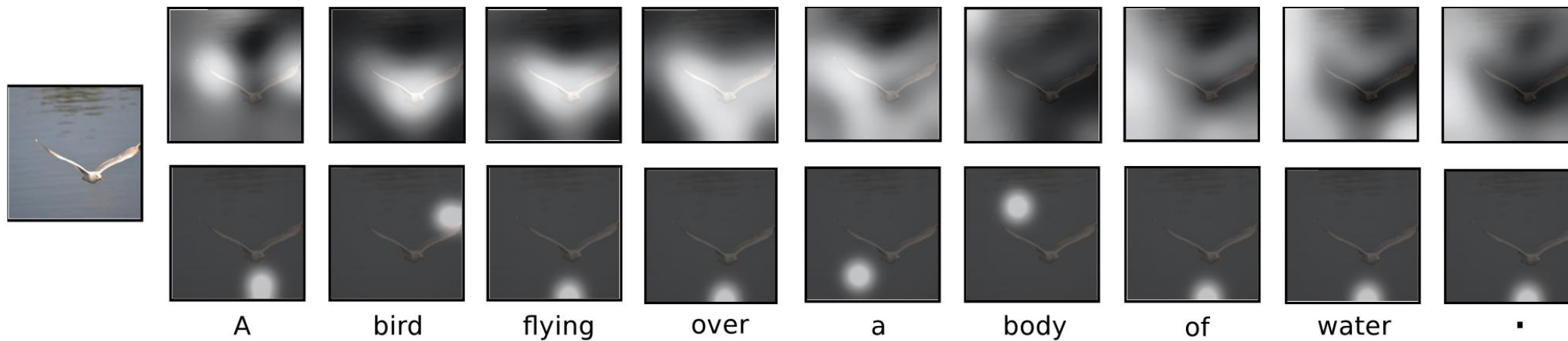


Encoder-Decoder: CNN-RNN

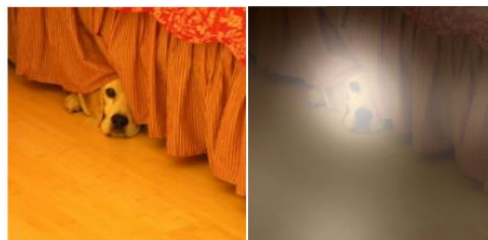


Text embedding space and image embedding space can be aligned!

CNN + Attention LSTM



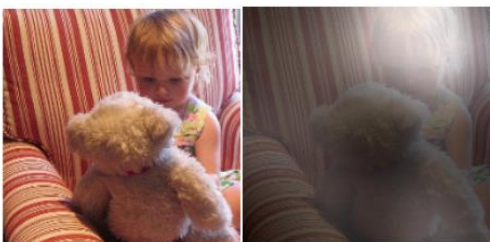
A woman is throwing a frisbee in a park.



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.

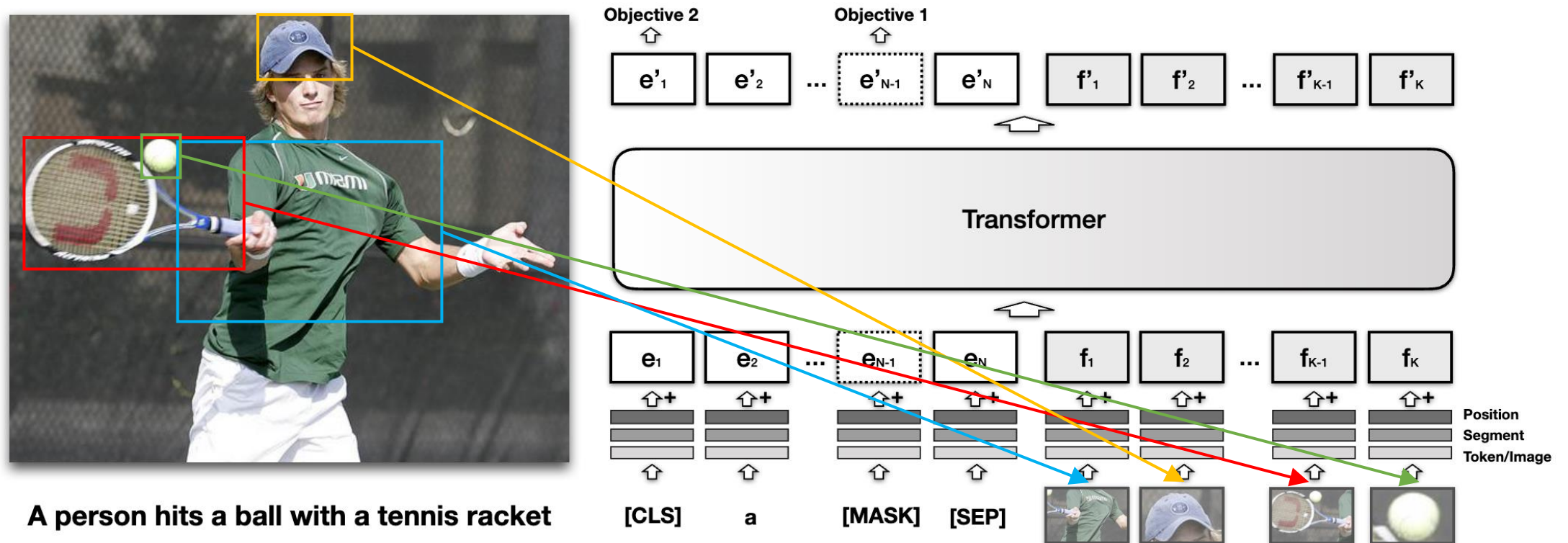


A group of people sitting on a boat in the water.



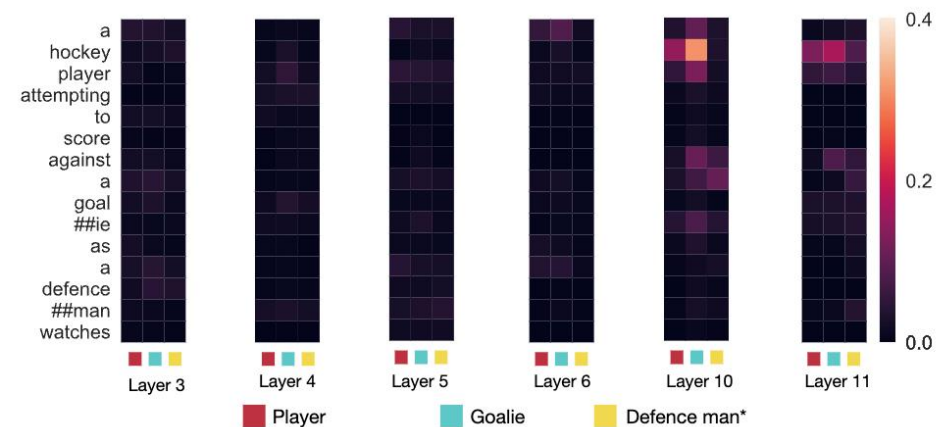
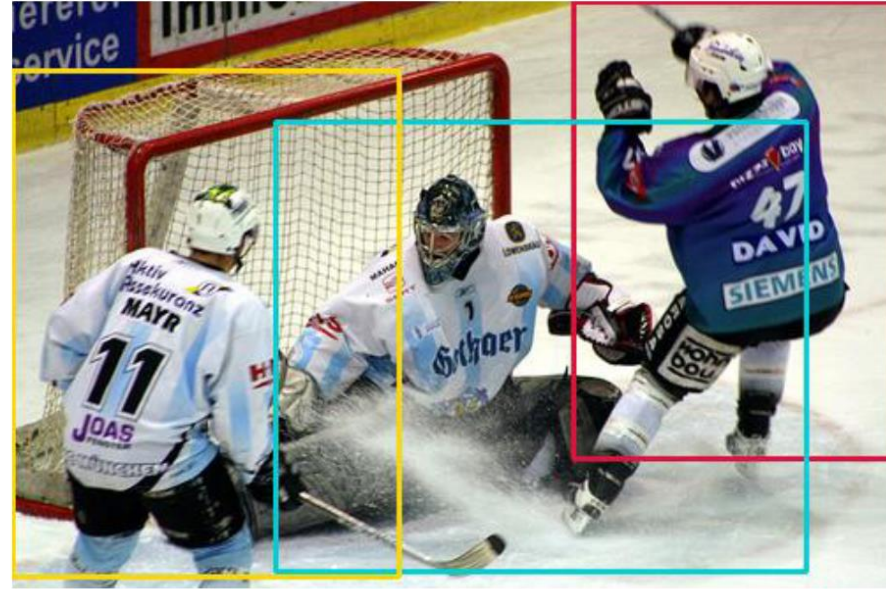
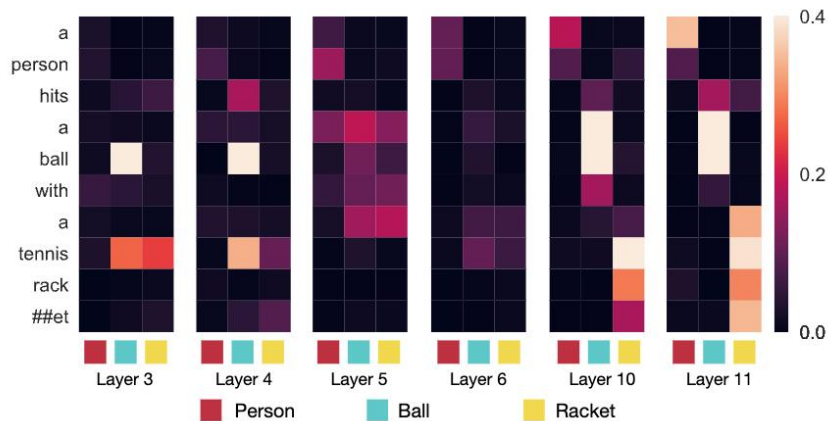
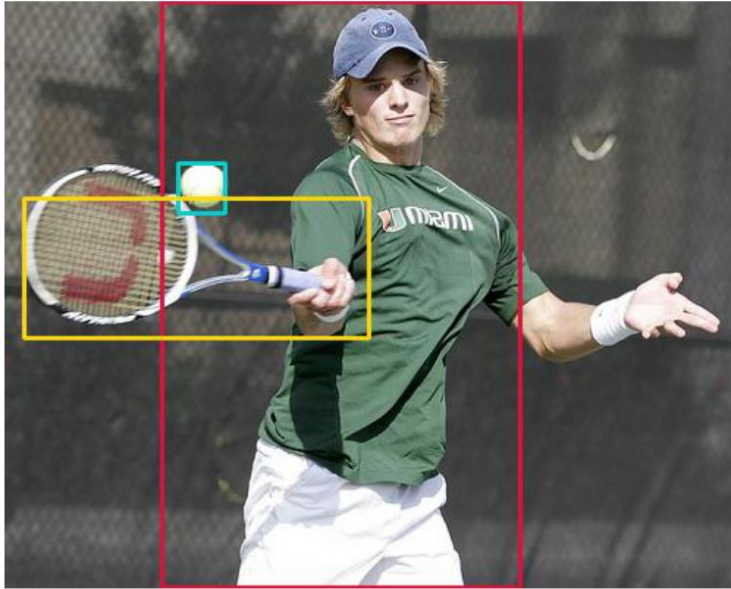
A giraffe standing in a forest with trees in the background.

Joint Visual and Textual Embeddings: VisualBERT



Require an object detection model

Joint Visual and Textual Embeddings: VisualBERT



Visual Question Answering

Who is wearing glasses?

man



woman

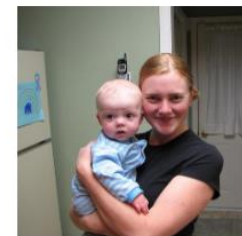


Where is the child sitting?

fridge



arms



Is the umbrella upside down?

yes



no



How many children are in the bed?

2

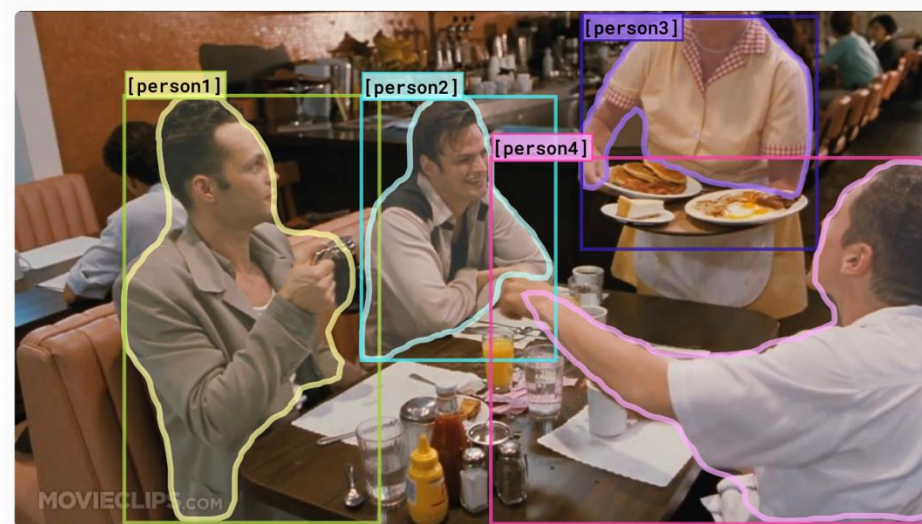


1



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

Visual Commonsense Reasoning



hide all

show all

[person1]

[person2]

[person3]

[person4]

more objects »

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		QA → R		Q → AR	
	Dev	Test	Dev	Test	Dev	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

Natural Language Visual Reasoning



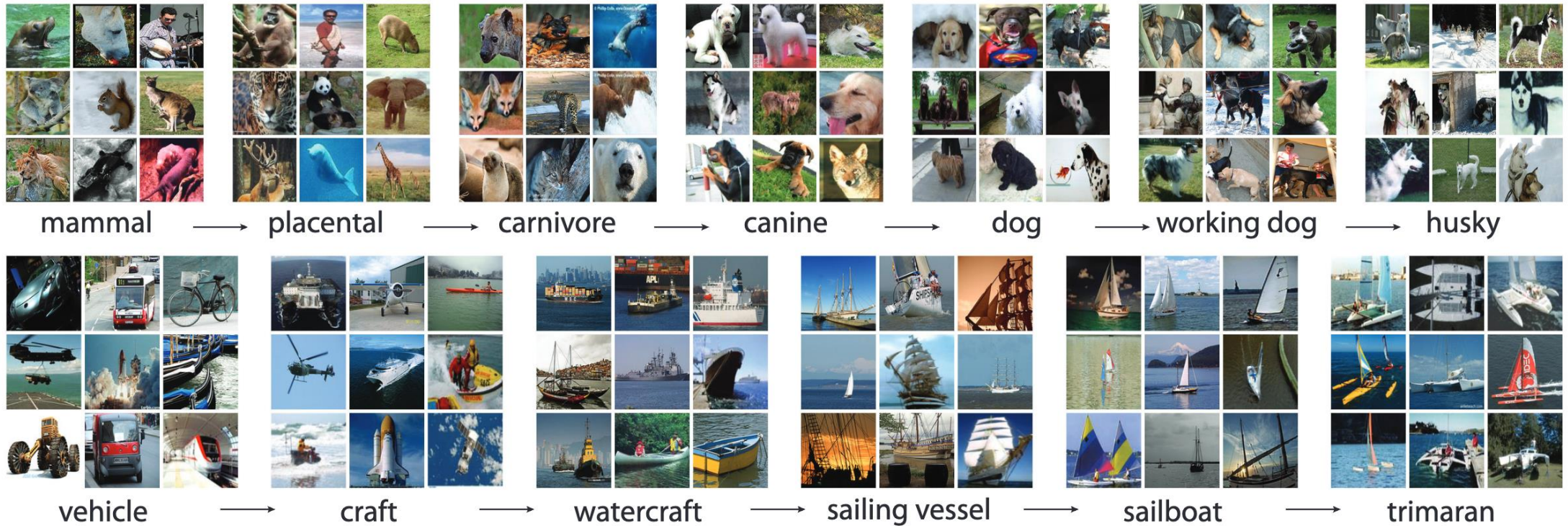
The left image contains twice the number of dogs as the right image, and at least two dogs in total are standing.



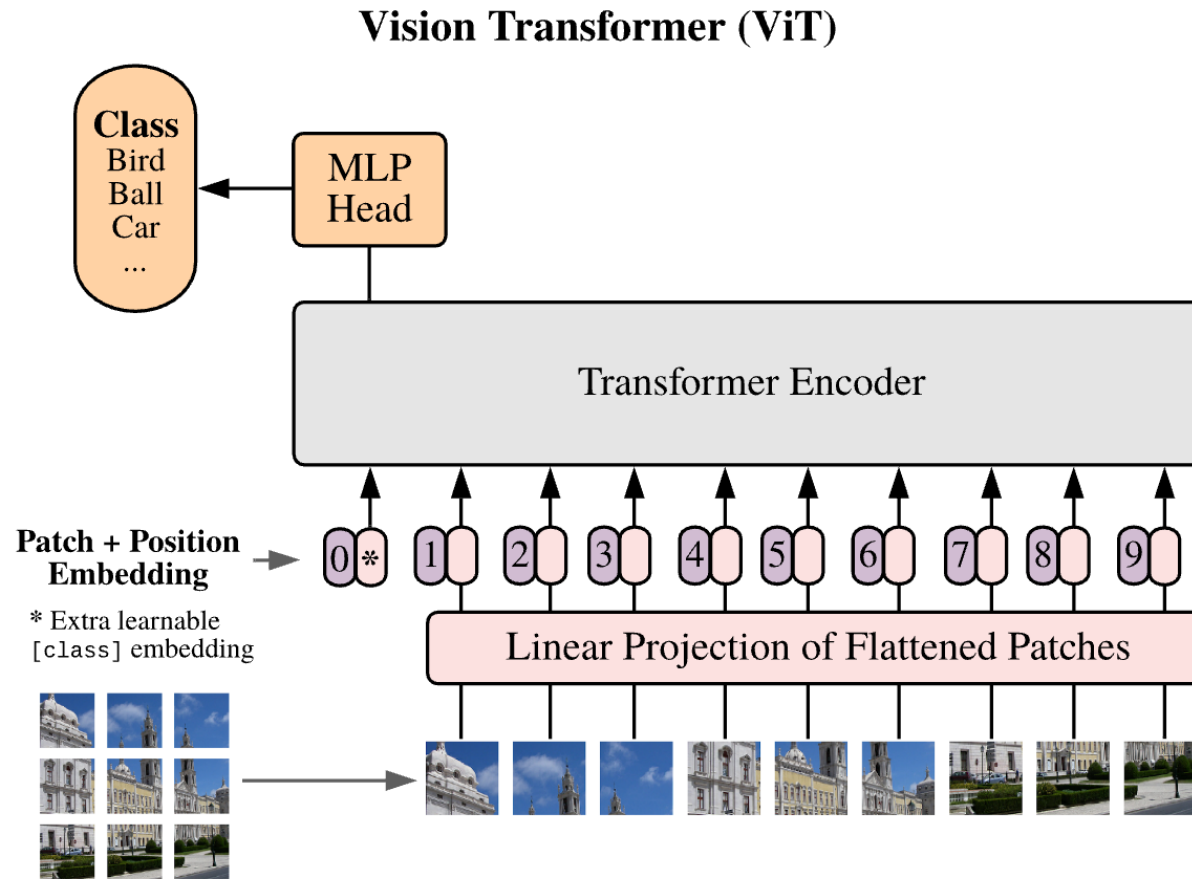
One image shows exactly two brown acorns in back-to-back caps on green foliage.

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

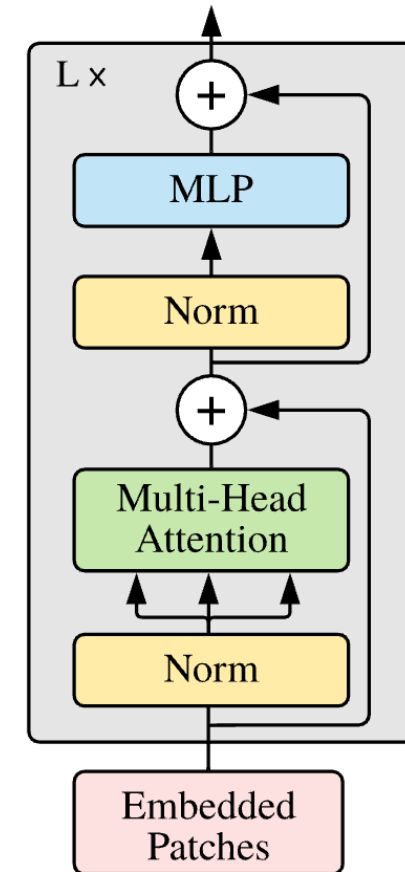
Pre-Trained CNN with ImageNet



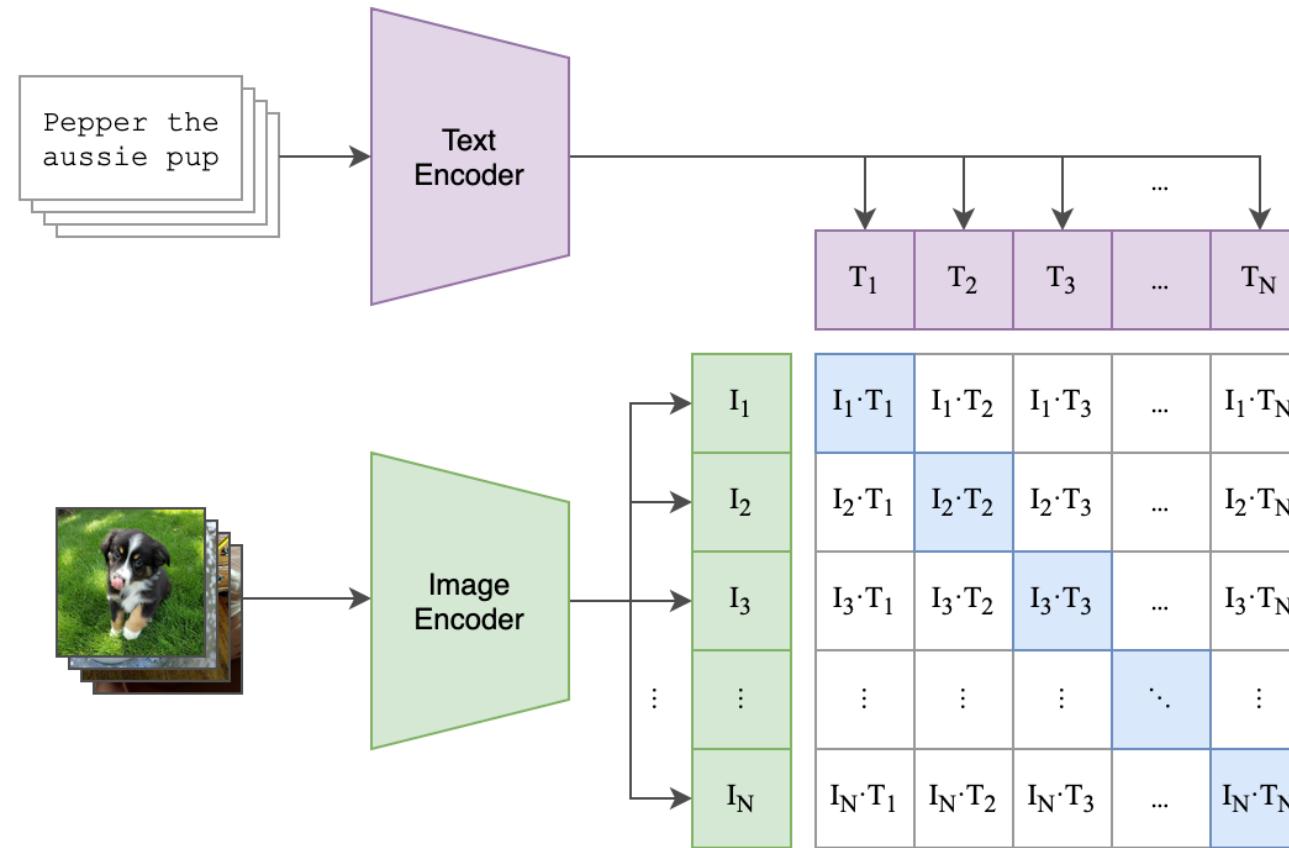
Vision Transformer



Transformer Encoder

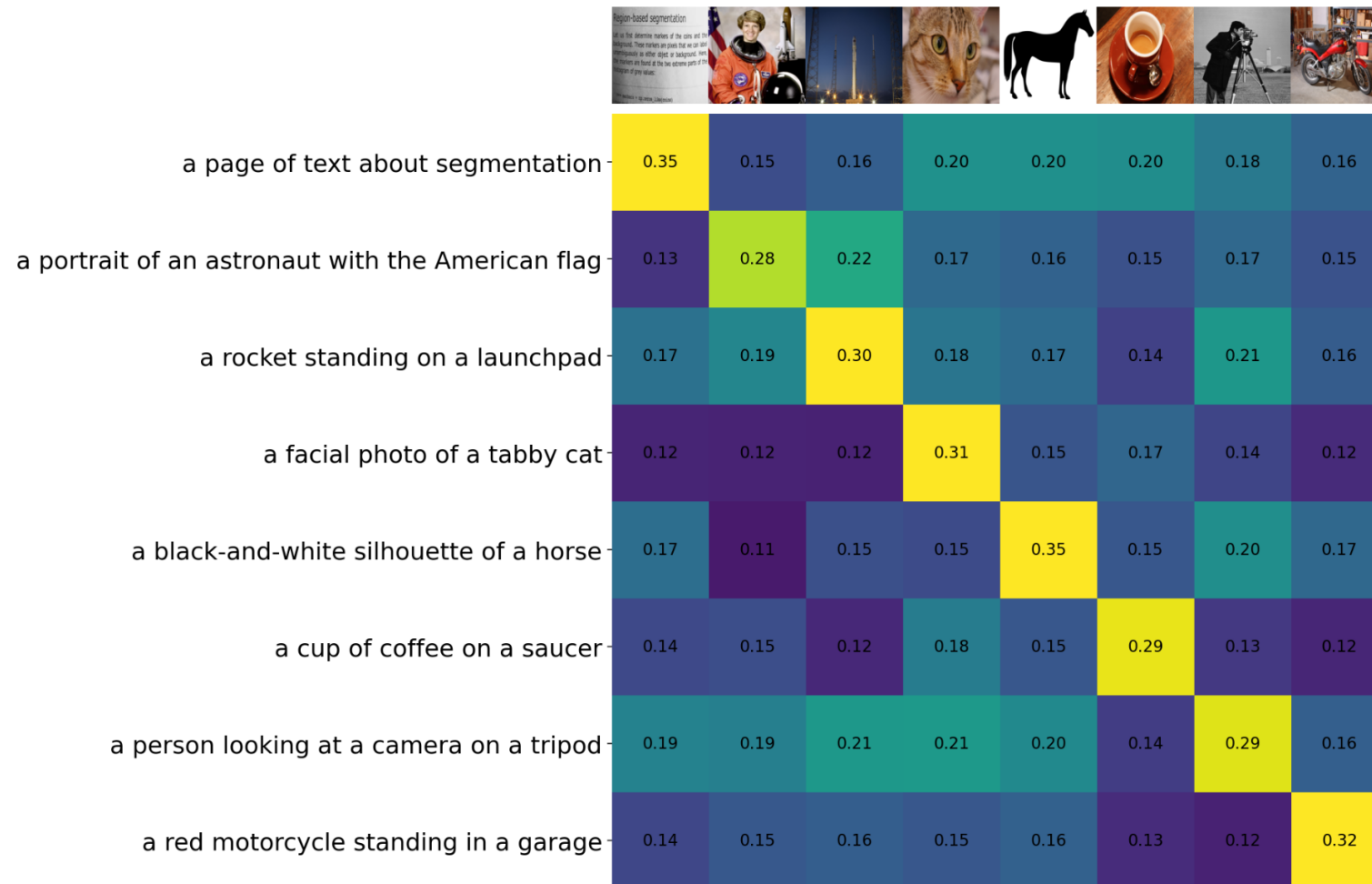


CLIP: Contrastive Language-Image Pre-Training

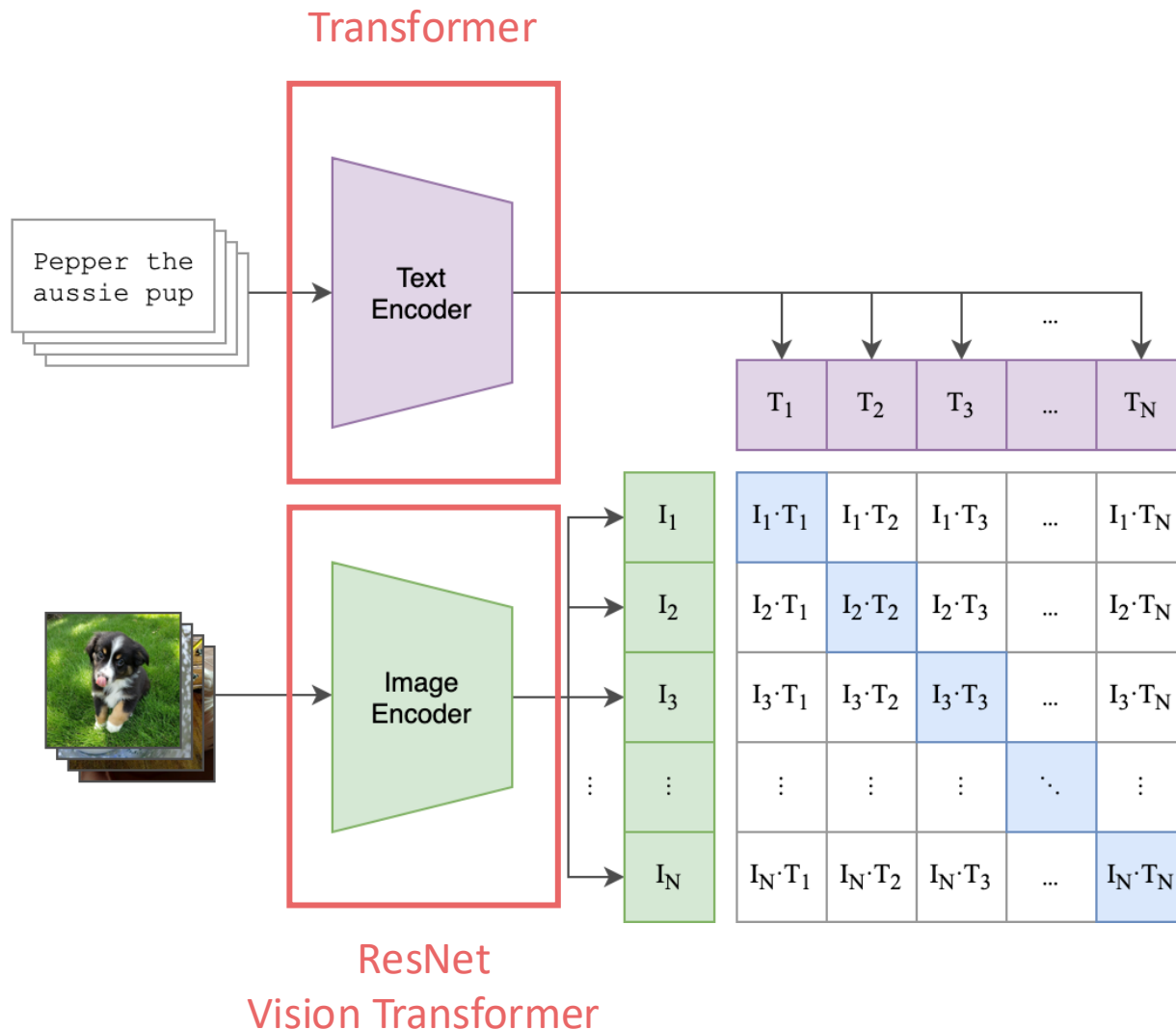


Training with Image-Caption Pairs

Cosine similarity between text and image features



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)
```

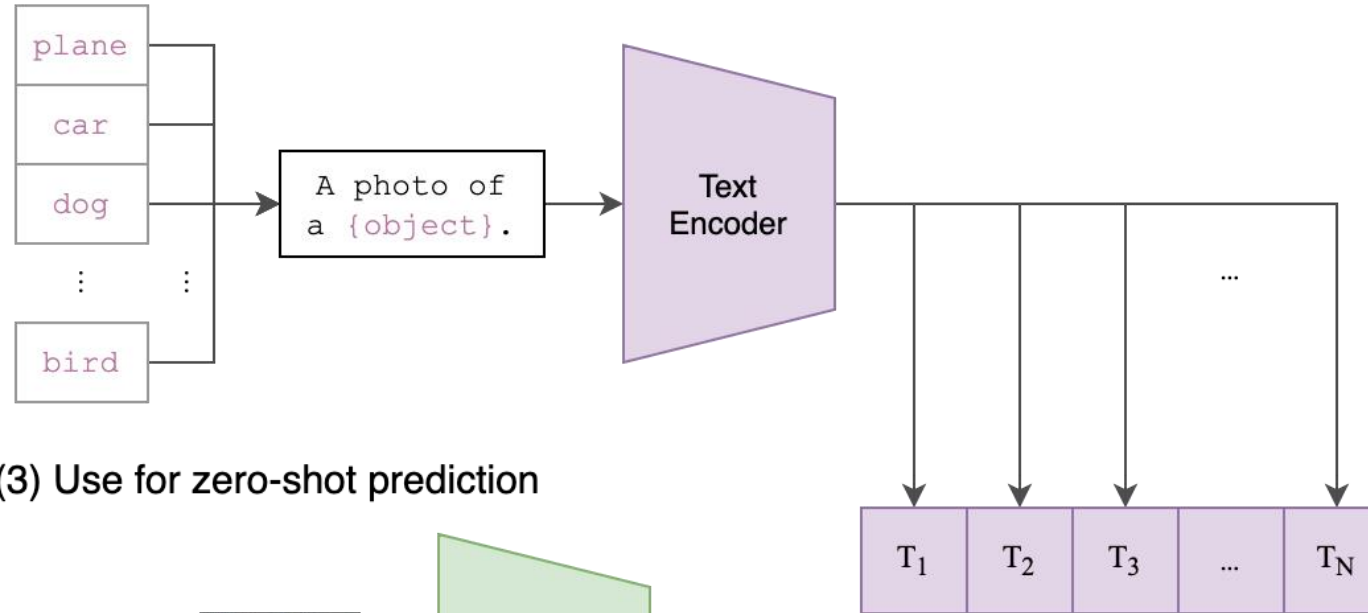
```
# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)
```

```
# 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
```

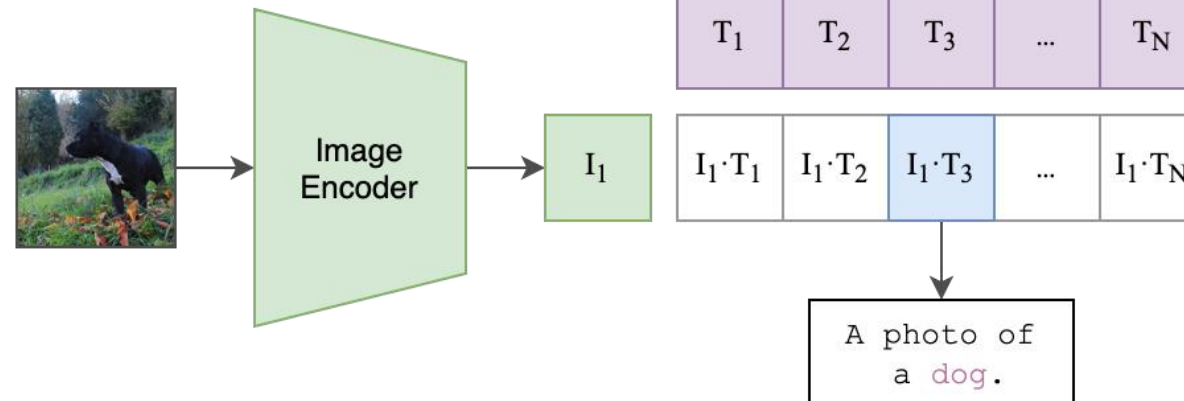
1	0	0	...	0
0	1	0	...	0
0	0	1	...	0
0	0	0	...	0
...	...	...	...	0
0	0	0	0	1

Zero-Shot Prediction

(2) Create dataset classifier from label text



(3) Use for zero-shot prediction



Zero-Shot CLIP vs. Few-shot Linear Probes

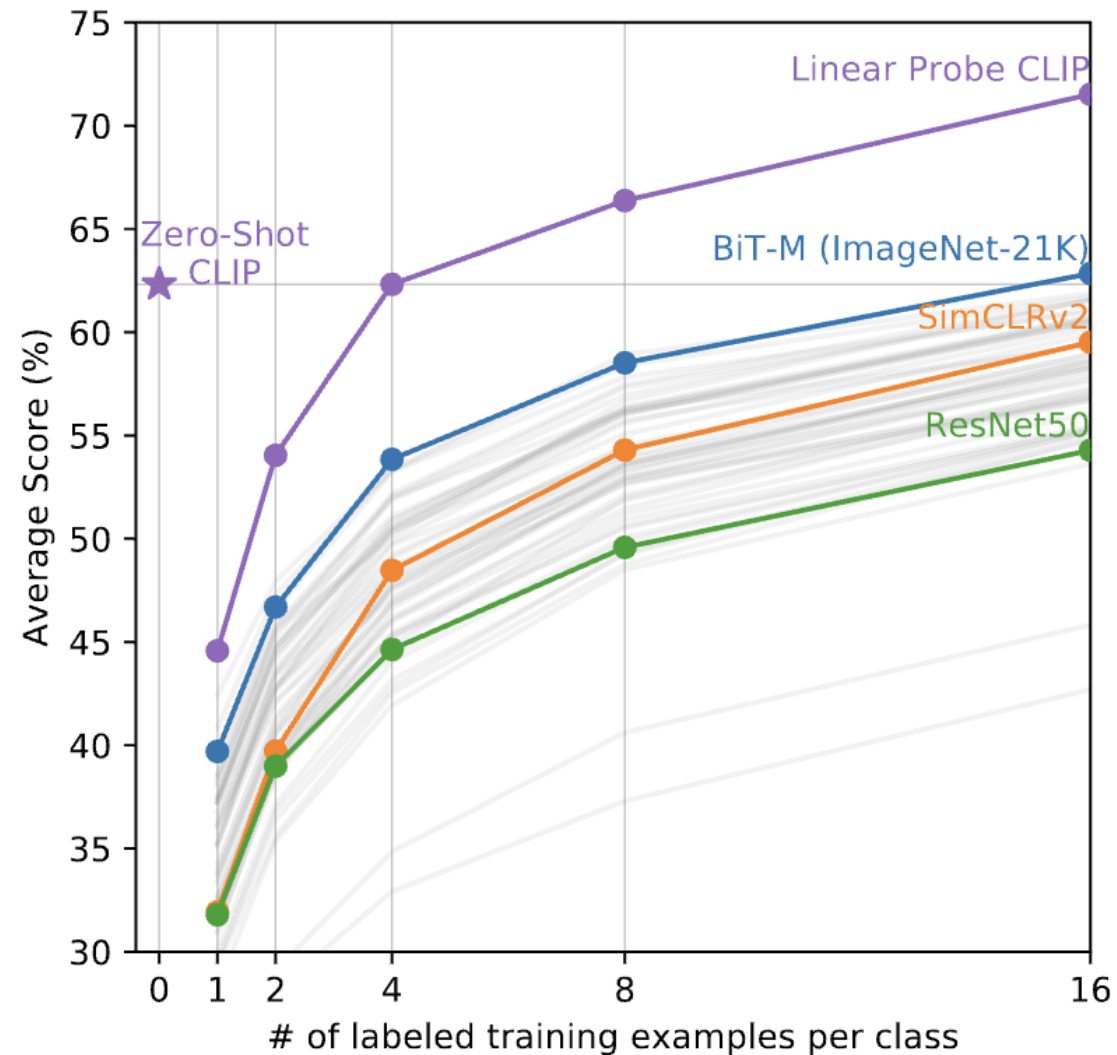
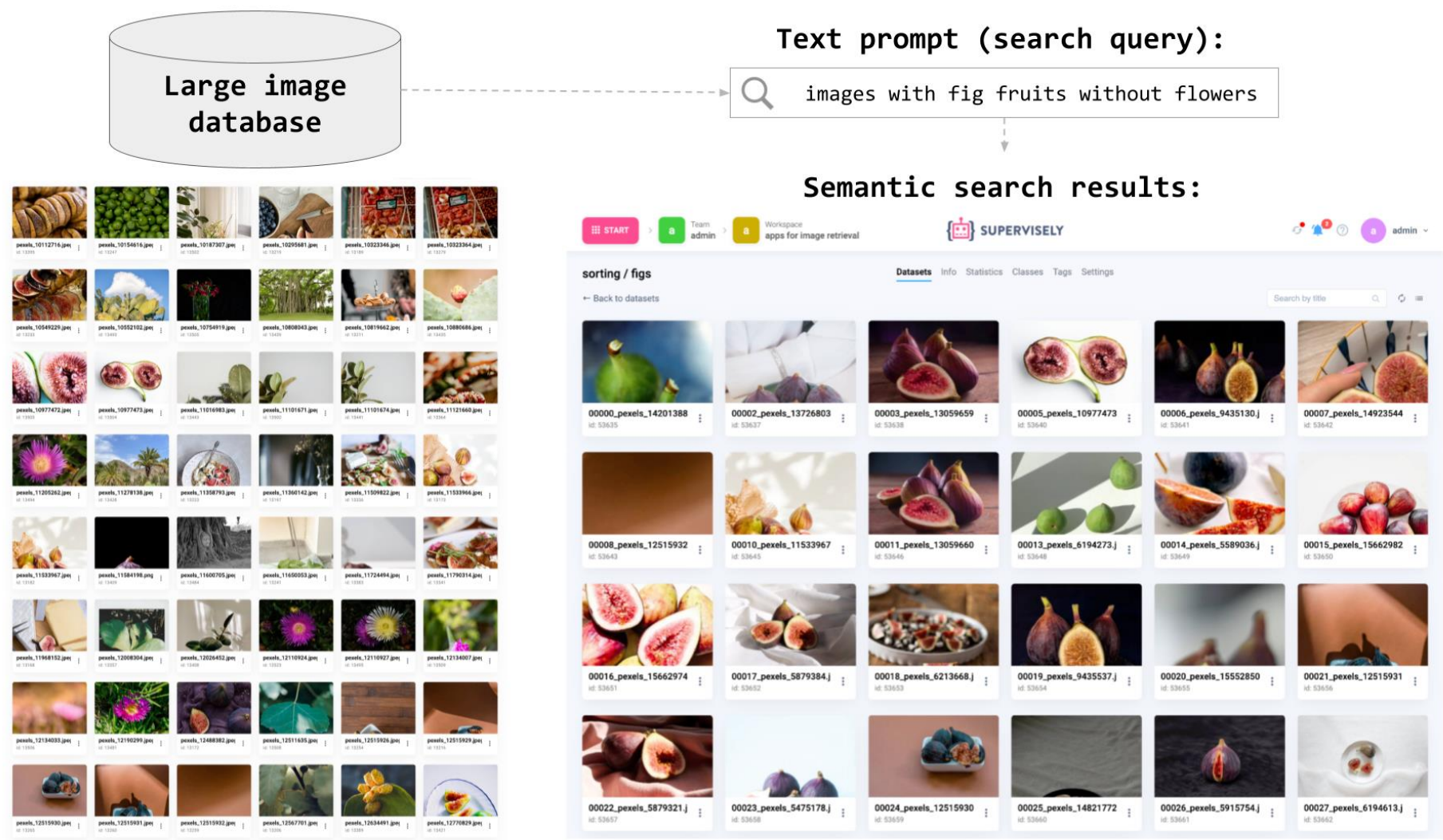
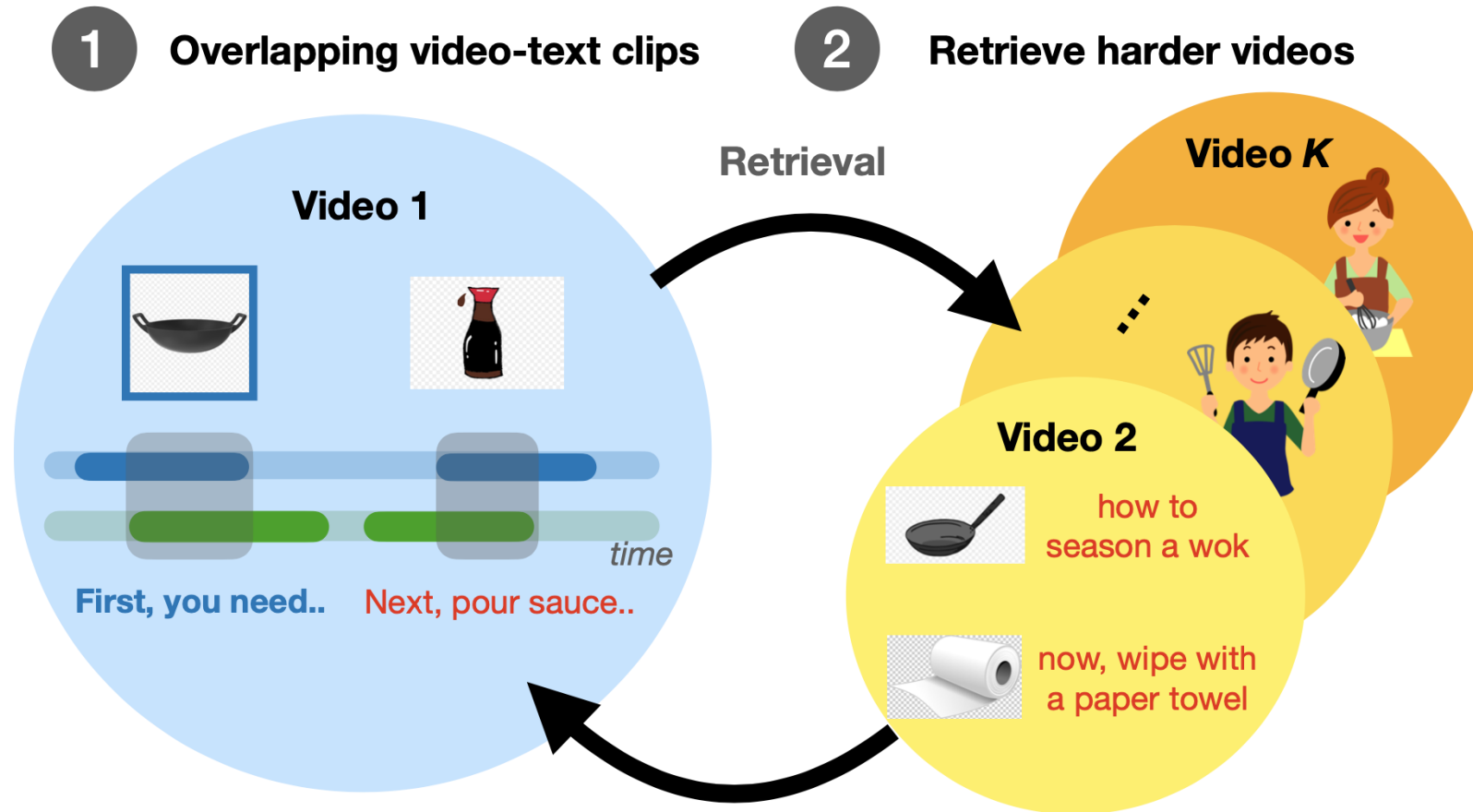


Image Retrieval with Text Query



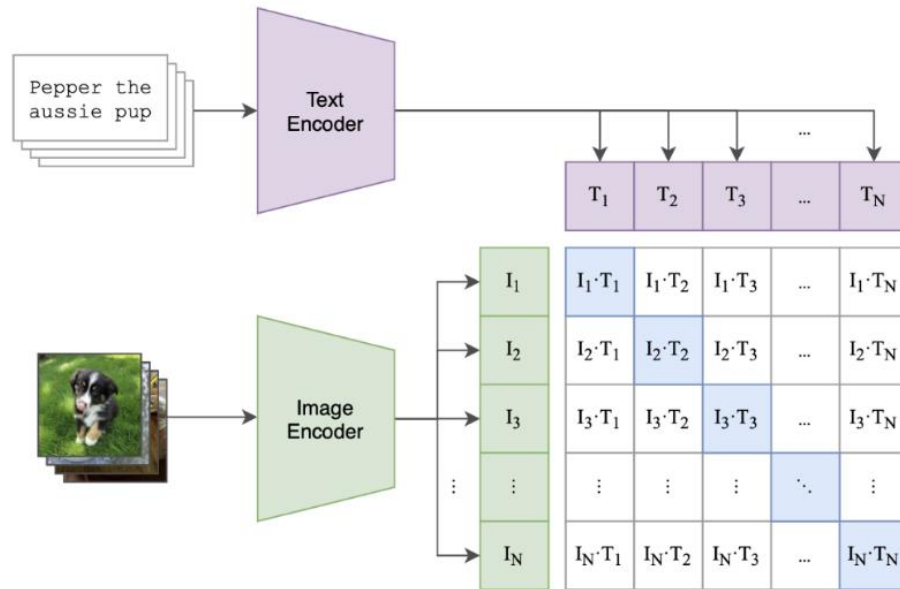
VideoCLIP



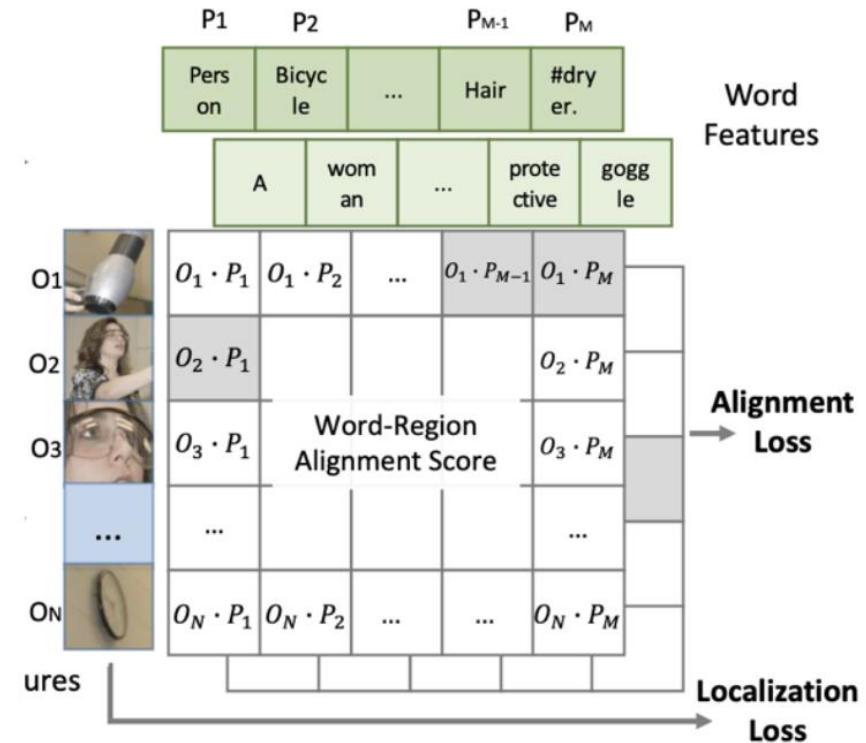
VideoCLIP: Contrastive learning with **hard-retrieved negatives** and **overlapping positives** for video-text pre-training.

GLIP: Grounded Language-Image Pre-training

CLIP: capture information for whole image



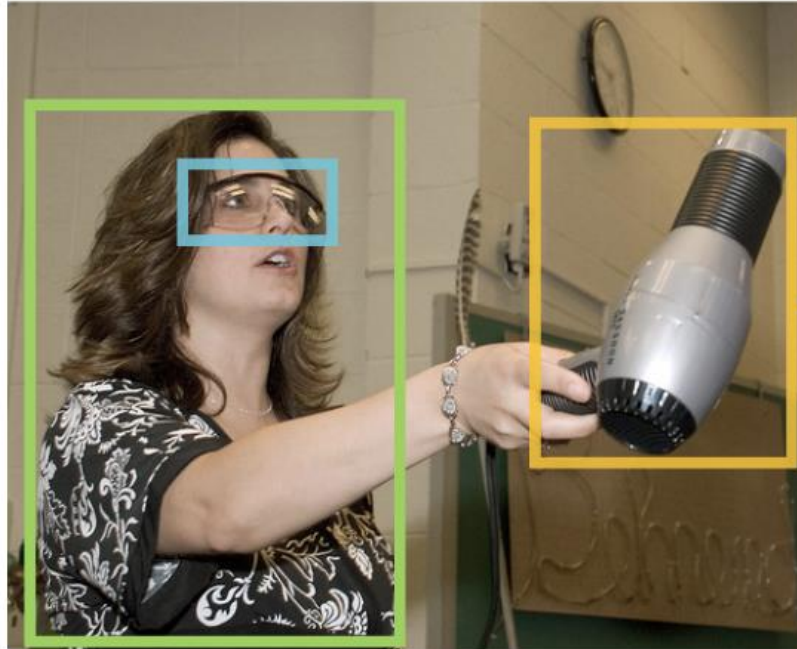
GLIP: capture information more for objects/entities



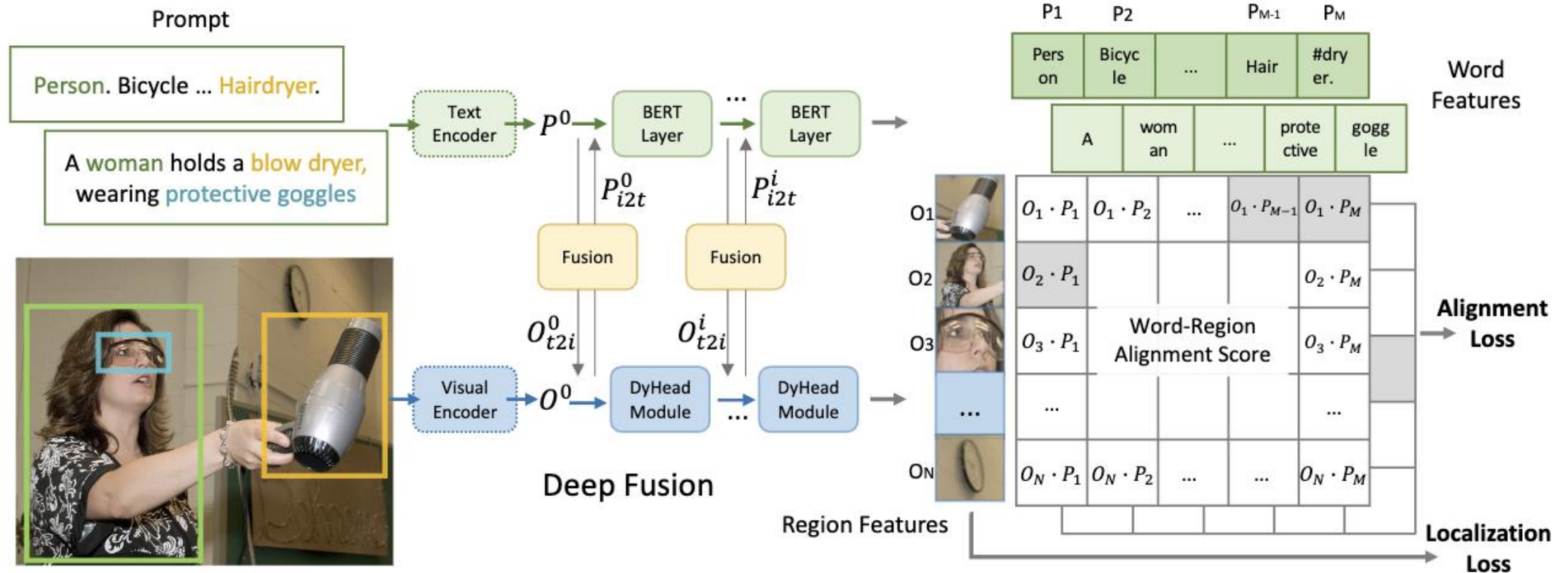
Object Detection and Text Grounding

Person. Bicycle ... Hairdryer.

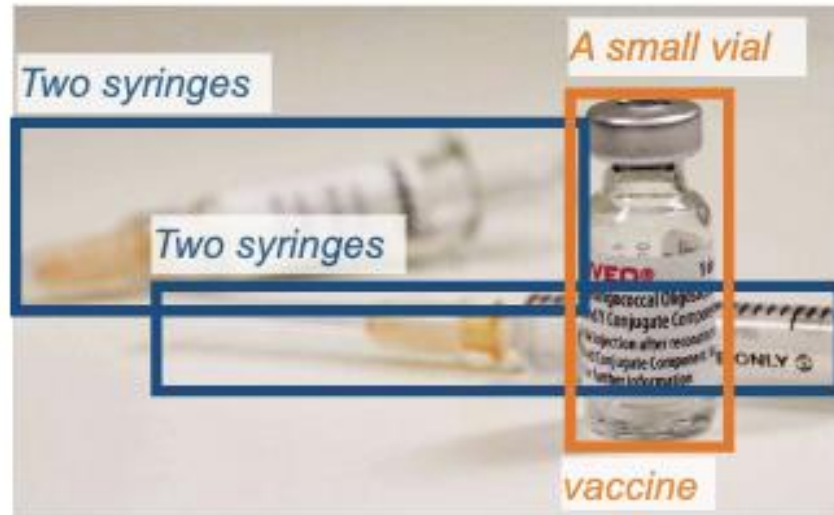
A woman holds a blow dryer,
wearing protective goggles



GLIP: Grounded Language-Image Pre-training



Grounding Results



Two syringes and a small vial of vaccine.



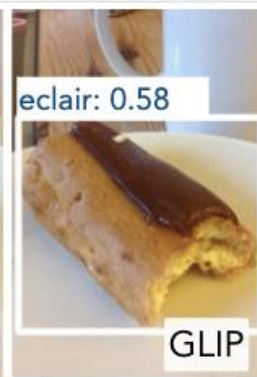
playa esmeralda in holguin, cuba. the view from the top of the beach. beautiful caribbean sea turquoise

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

Detect with specifications for shape & subpart
Target Object Confusable Object

 eclair: 0.39 Ours	 eclair: 0.58 GLIP	 tart: 0.55 Ours	 tart: 0.55 GLIP
<u>Eclair</u>, a kind of food, long, cylindrical pastry, filled with cream, topped with chocolate		<u>Tart</u>, 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	 clown: 0.71 Ours	 clown: 0.33 GLIP
A <u>clown</u> making a balloon animal for a pretty lady		A <u>clown</u> kicking a soccer ball for a pretty lady	

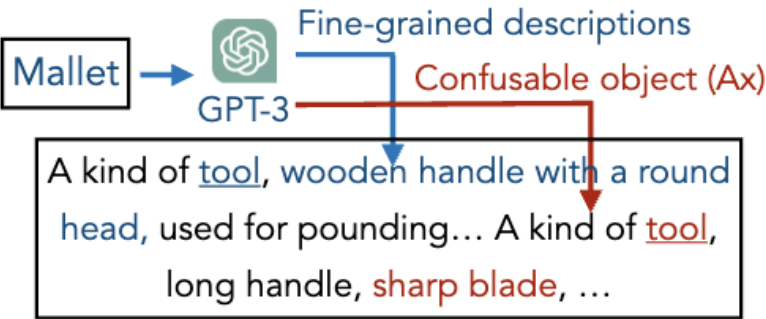
DesCo: Description-Conditioned

Detect: Mallet.
Bear. Cat...

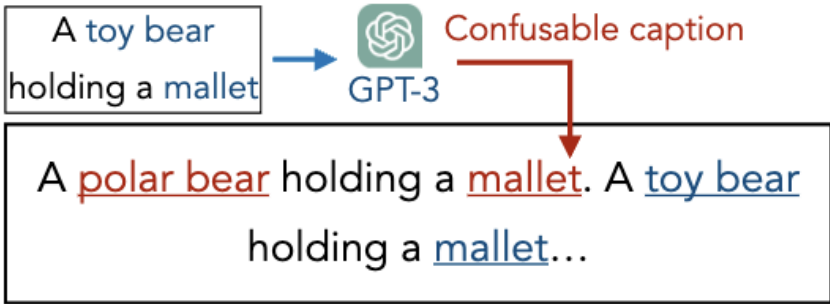
A toy bear holding
a mallet.



Original training
data for GLIP



	...	tool	...	tool	...
	0	1	0	0	0
	0	0	0	0	0
	0	0	0	0	0
	0	0	0	0	0



	...	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

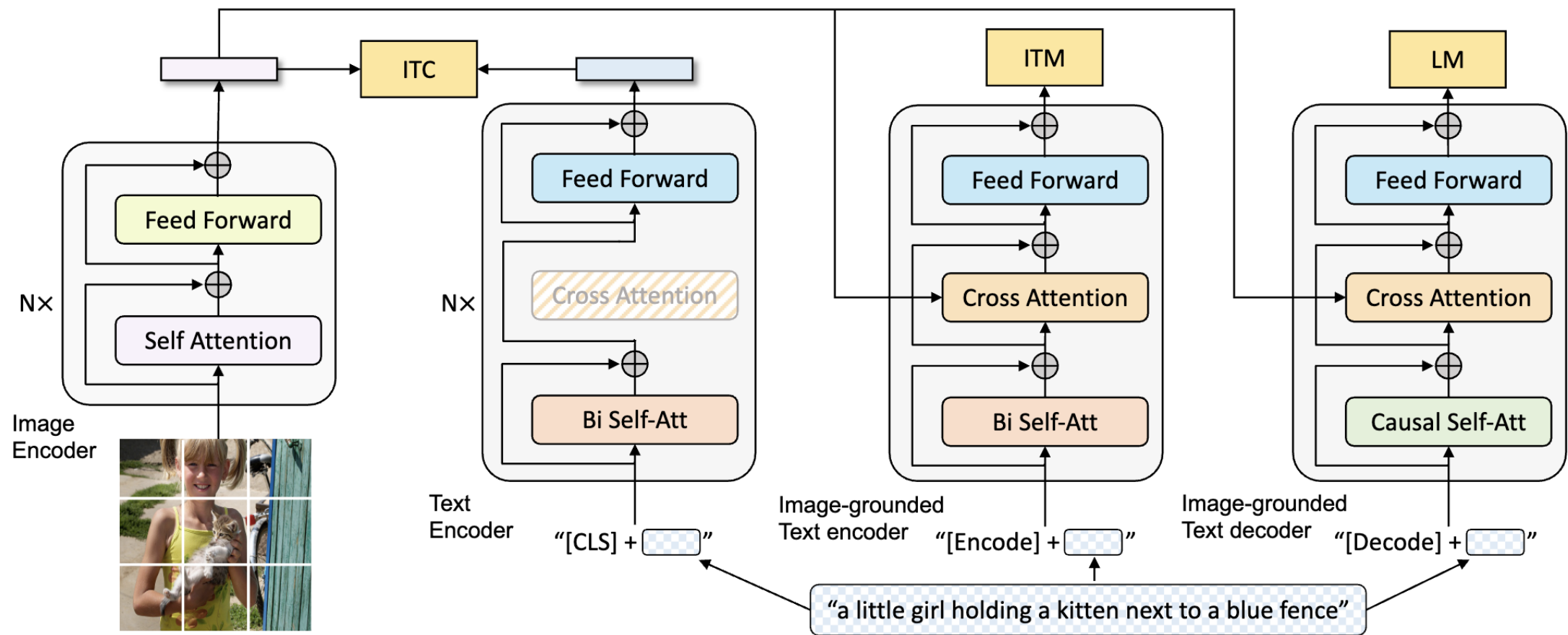
Zero-Shot Grounding

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

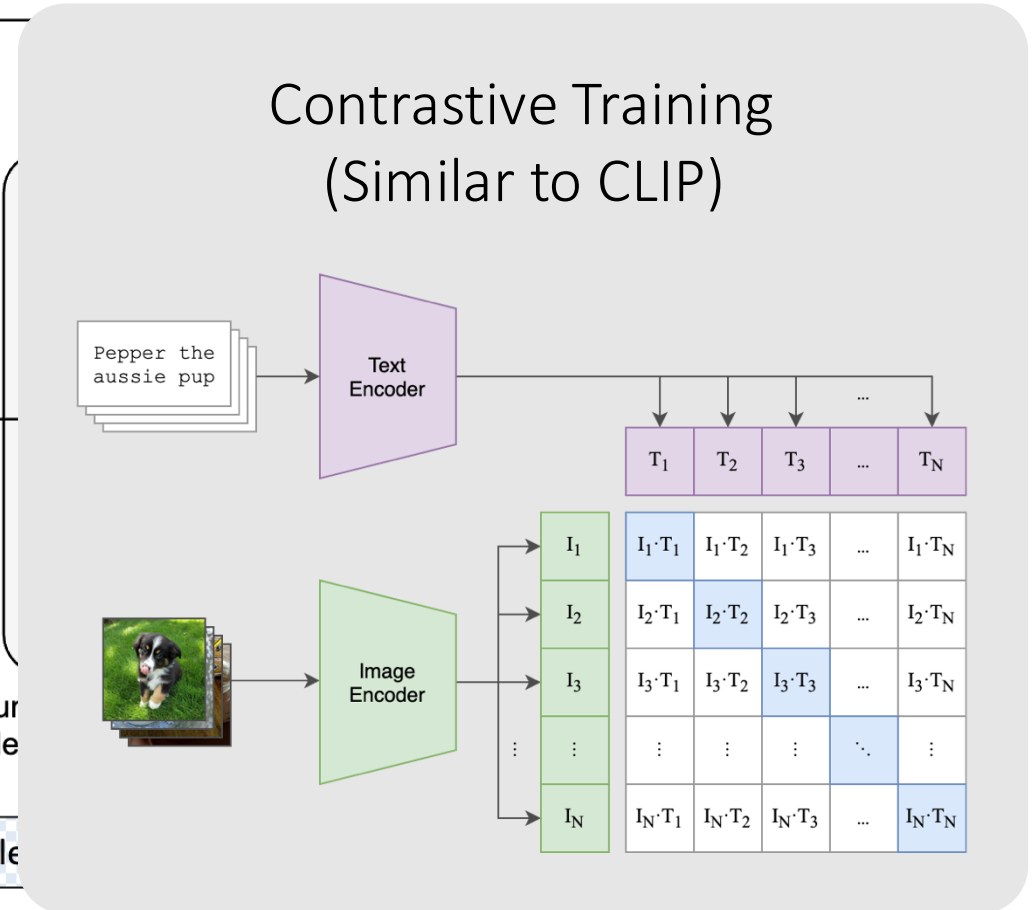
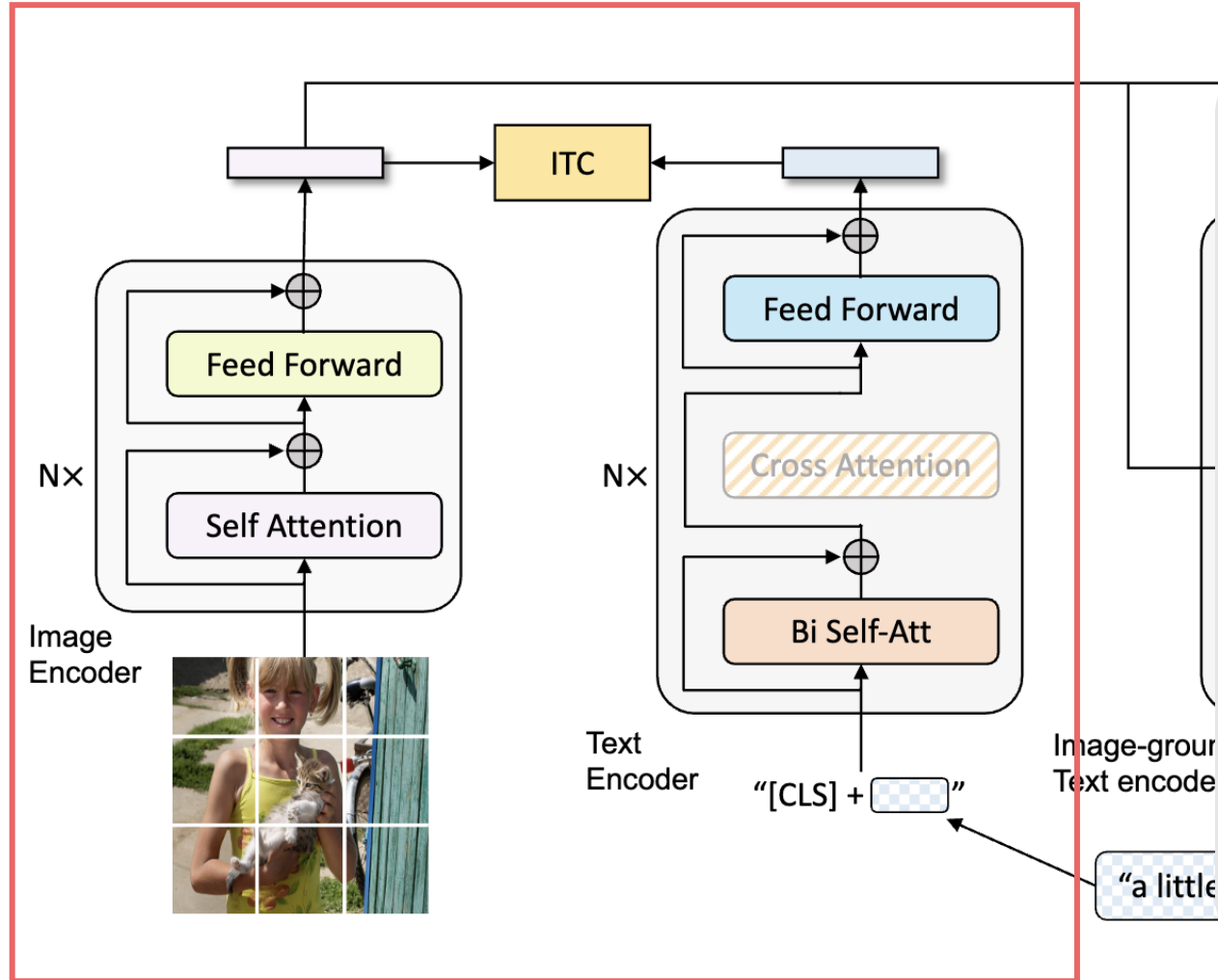
Encoder-Only vs. Encoder-Decoder

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

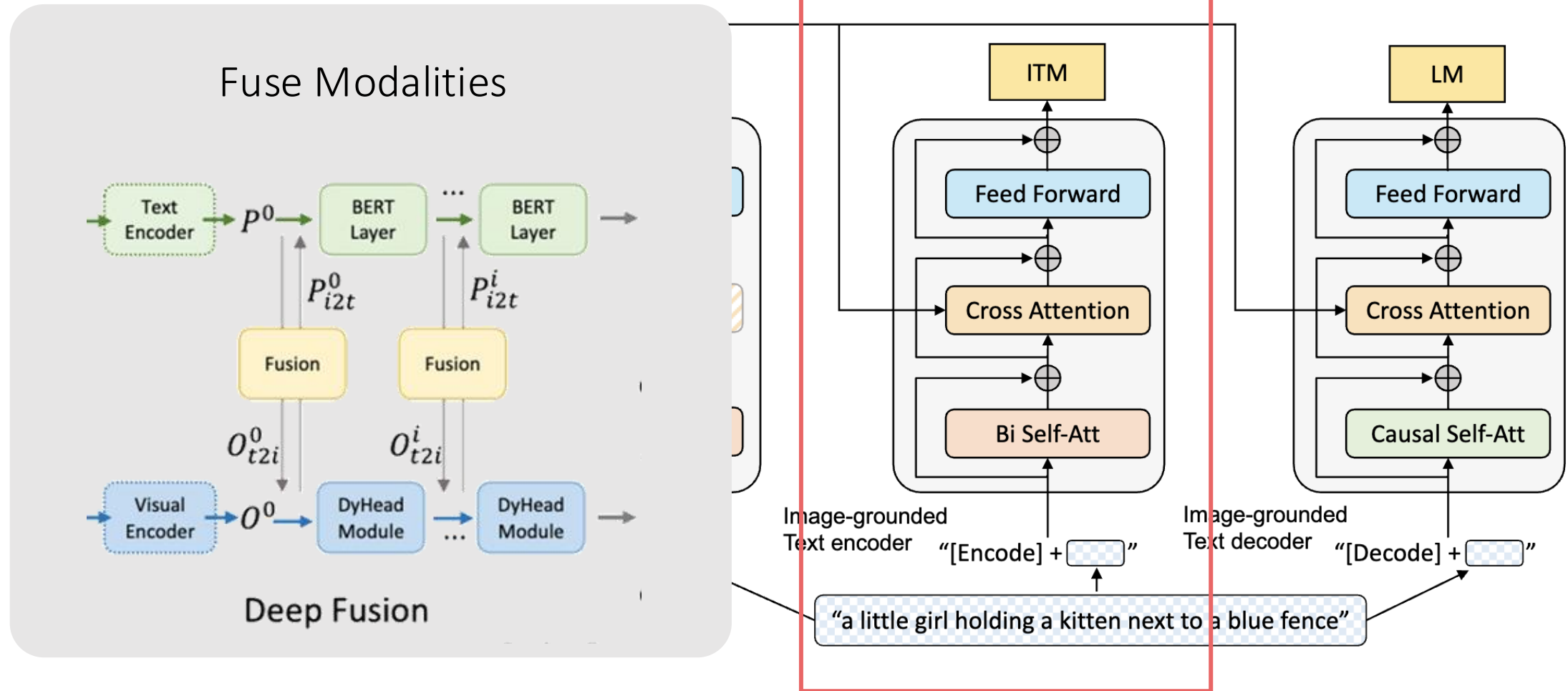
BLIP: Bootstrapping Language-Image Pre-training



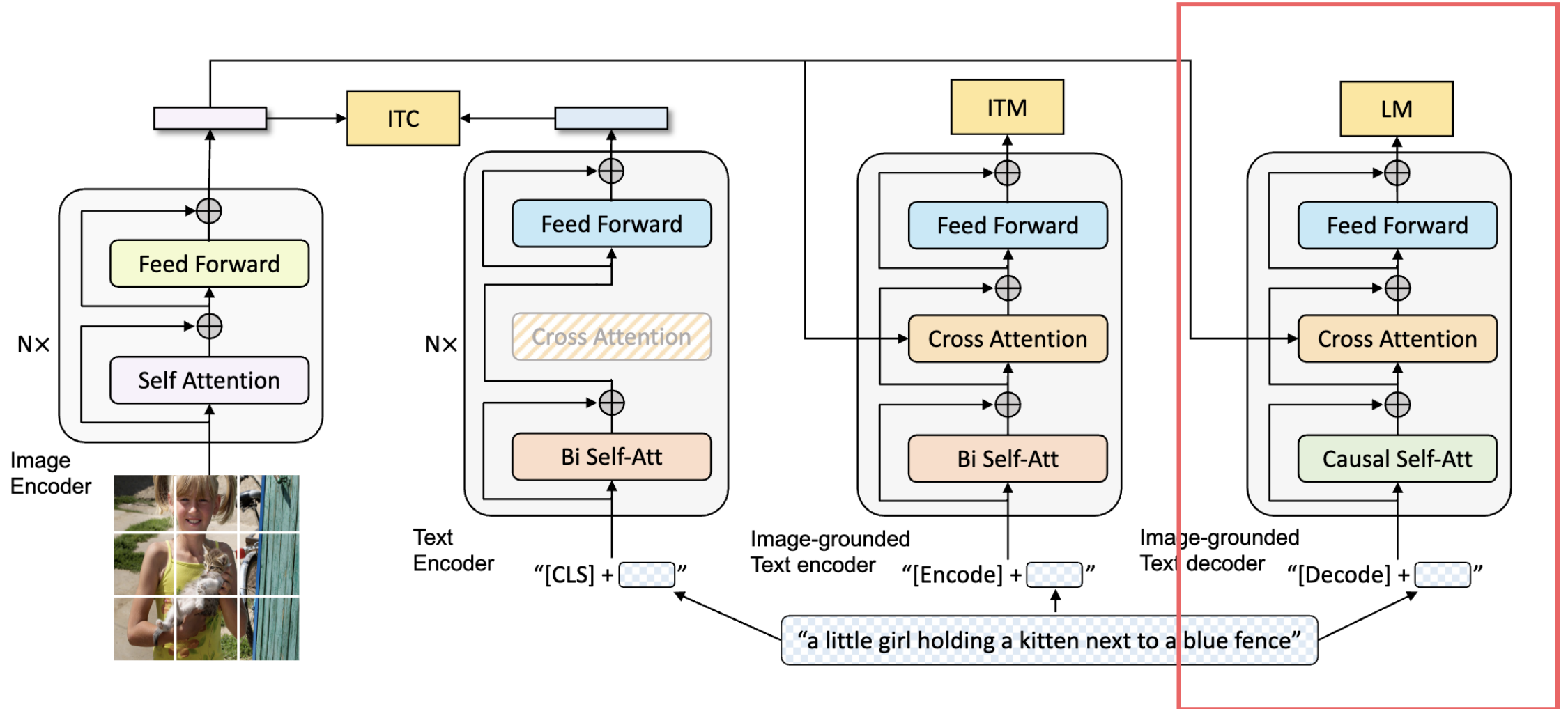
A Unified Framework



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
BLIP _{CapFilt-L}	129M	96.0	99.9	100.0	85.5	96.8	98.7
BLIP _{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		B@4	C
		C	S	C	S	C	S	C	S		
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 _{base} [†] (Hu et al., 2021)	12M	104.5	14.6	100.7	14.0	96.7	12.4	100.4	13.8	-	-
LEMON _{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 _{CapFilt-L}	129M	111.8	14.9	108.6	14.8	111.5	14.2	109.6	14.7	39.7	133.3
LEMON _{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 _{huge} (Wang et al., 2021)	1.8B	113.7	-	110.9	-	115.2	-	112.2	-	40.6	143.3
BLIP _{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

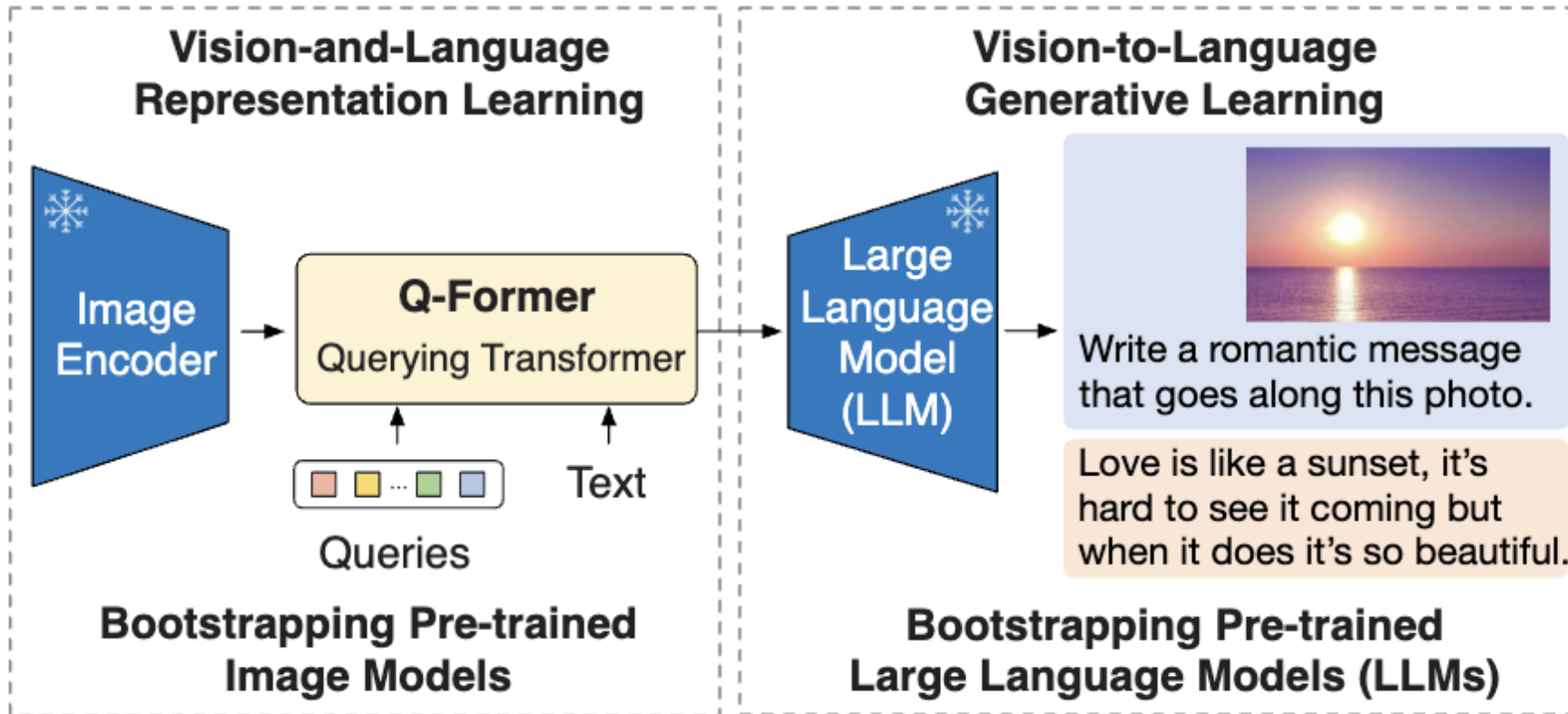


Does it appear to be rainy?

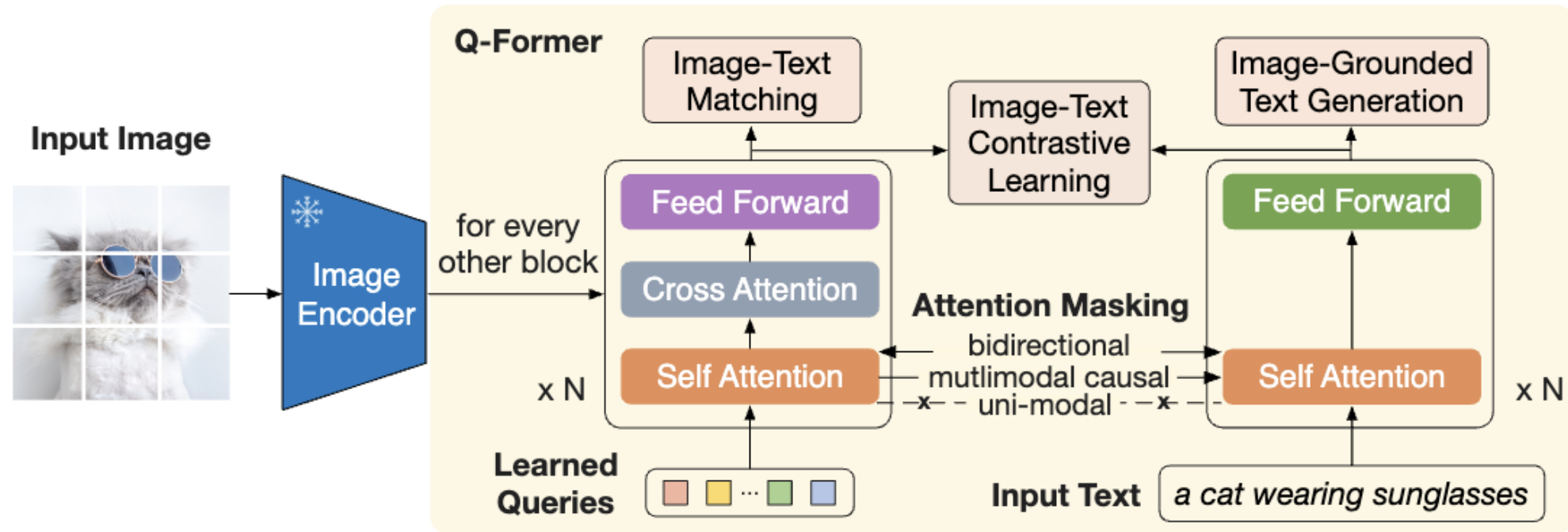
Does this person have 20/20 vision?

Method	Pre-train #Images	VQA		NLVR ²	
		test-dev	test-std	dev	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

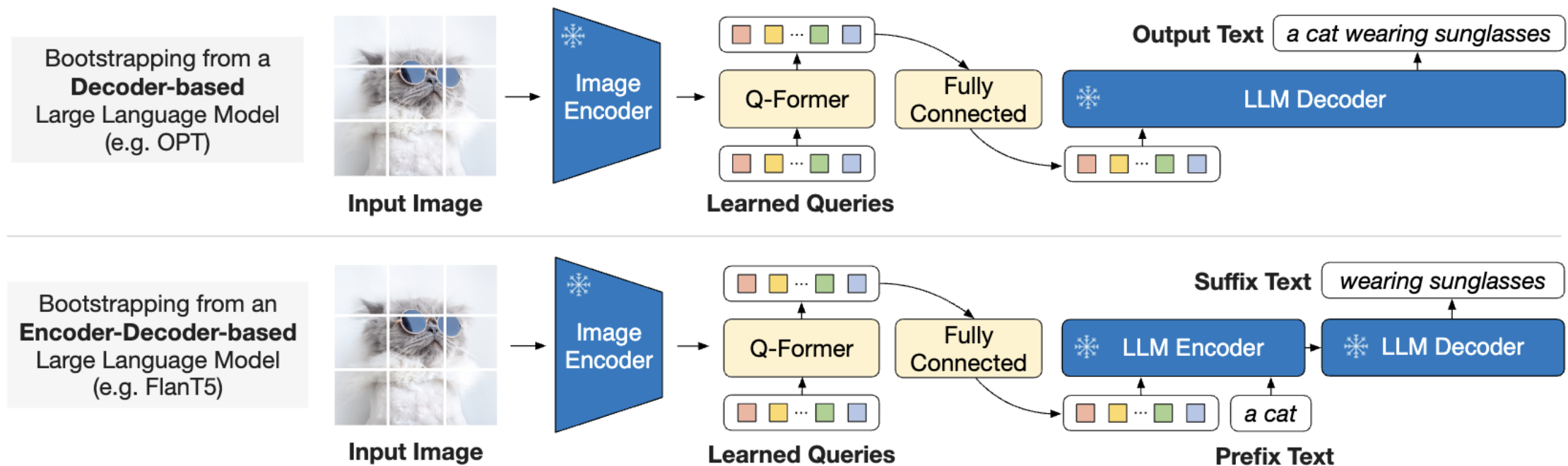
BLIP-2: Frozen Image Encoders and Large Language Models



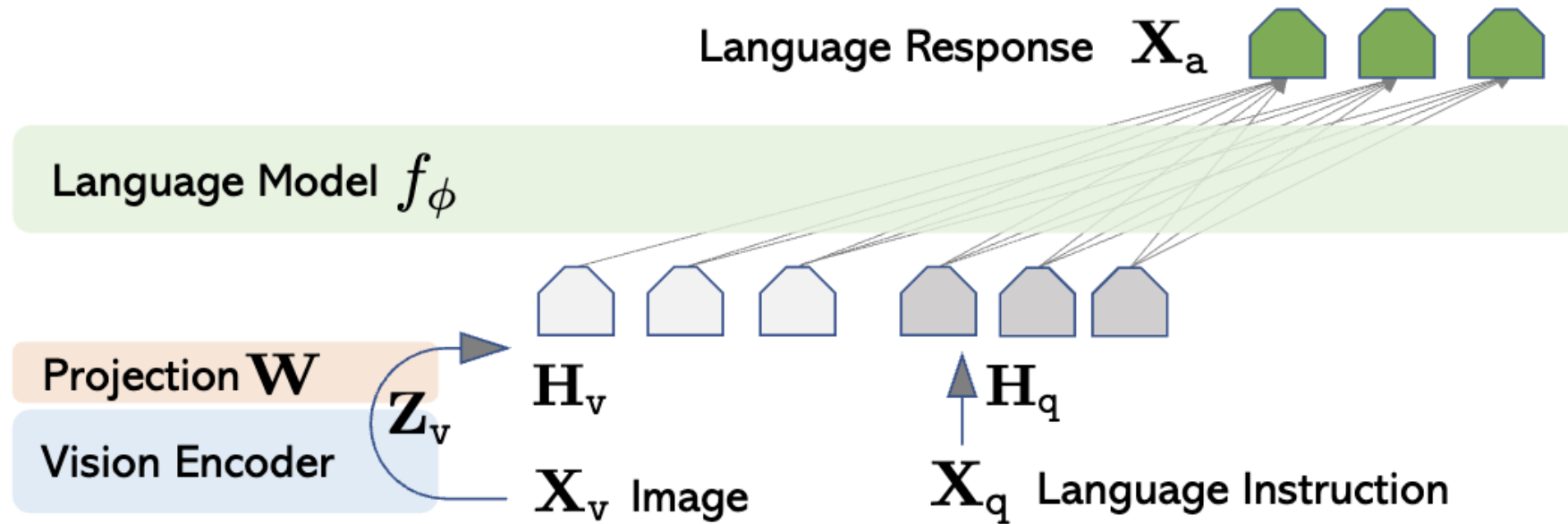
Vision-Language Representation Learning



Vision-to-Language Generative Pre-Training



LLaVA: Visual Instruction Tuning



GPT-Assisted Visual Instruction Data Generation

Context type 1: Captions

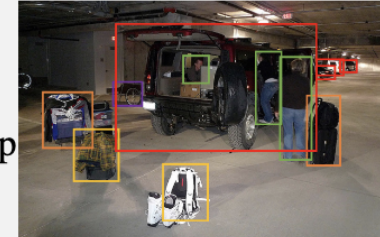
A group of people standing outside of a black vehicle with various luggage.

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

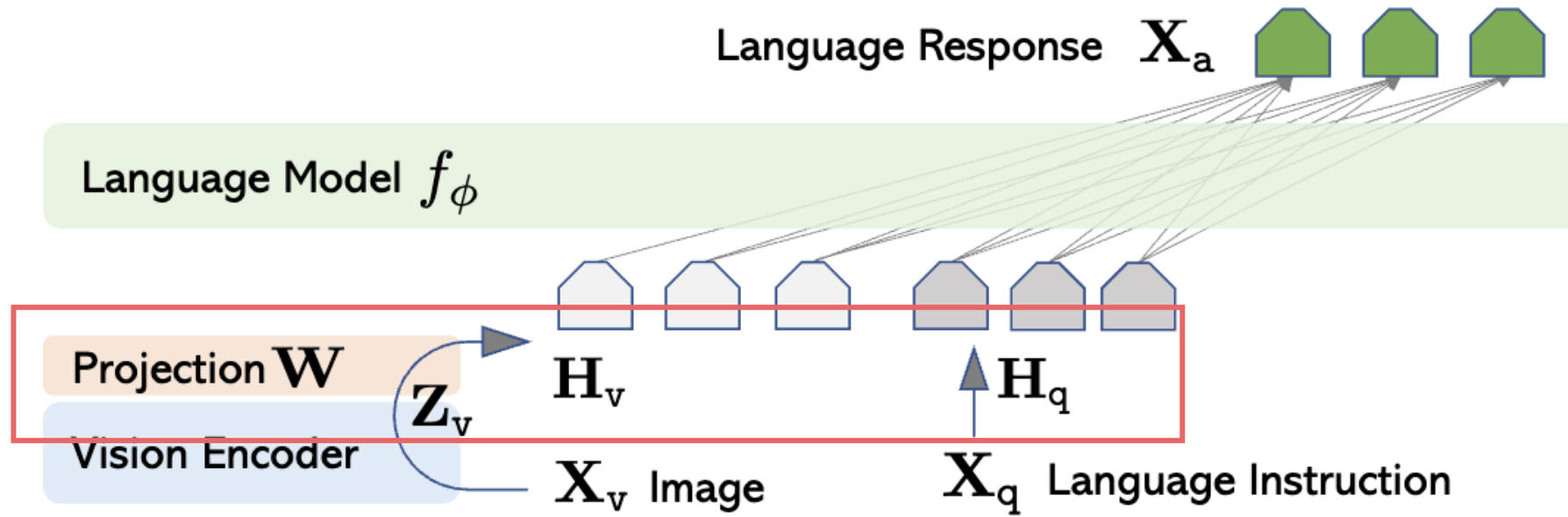
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their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

Response type 3: complex reasoning

Question: What challenges do these people face?

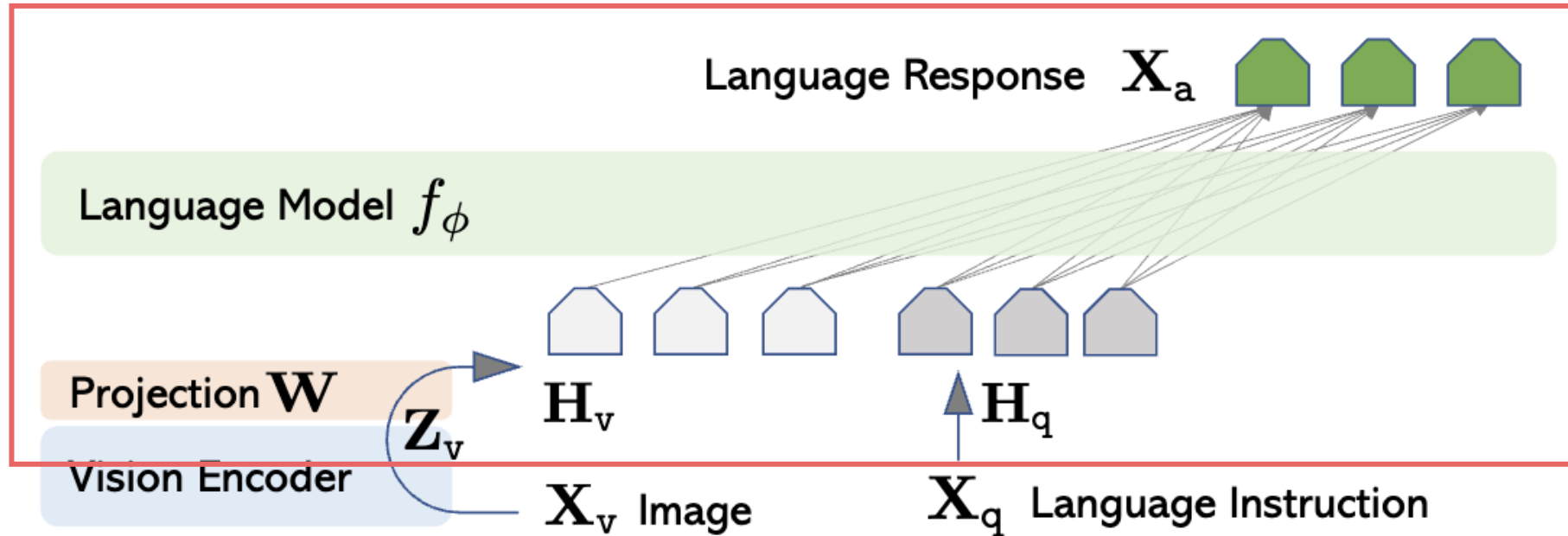
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, 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:

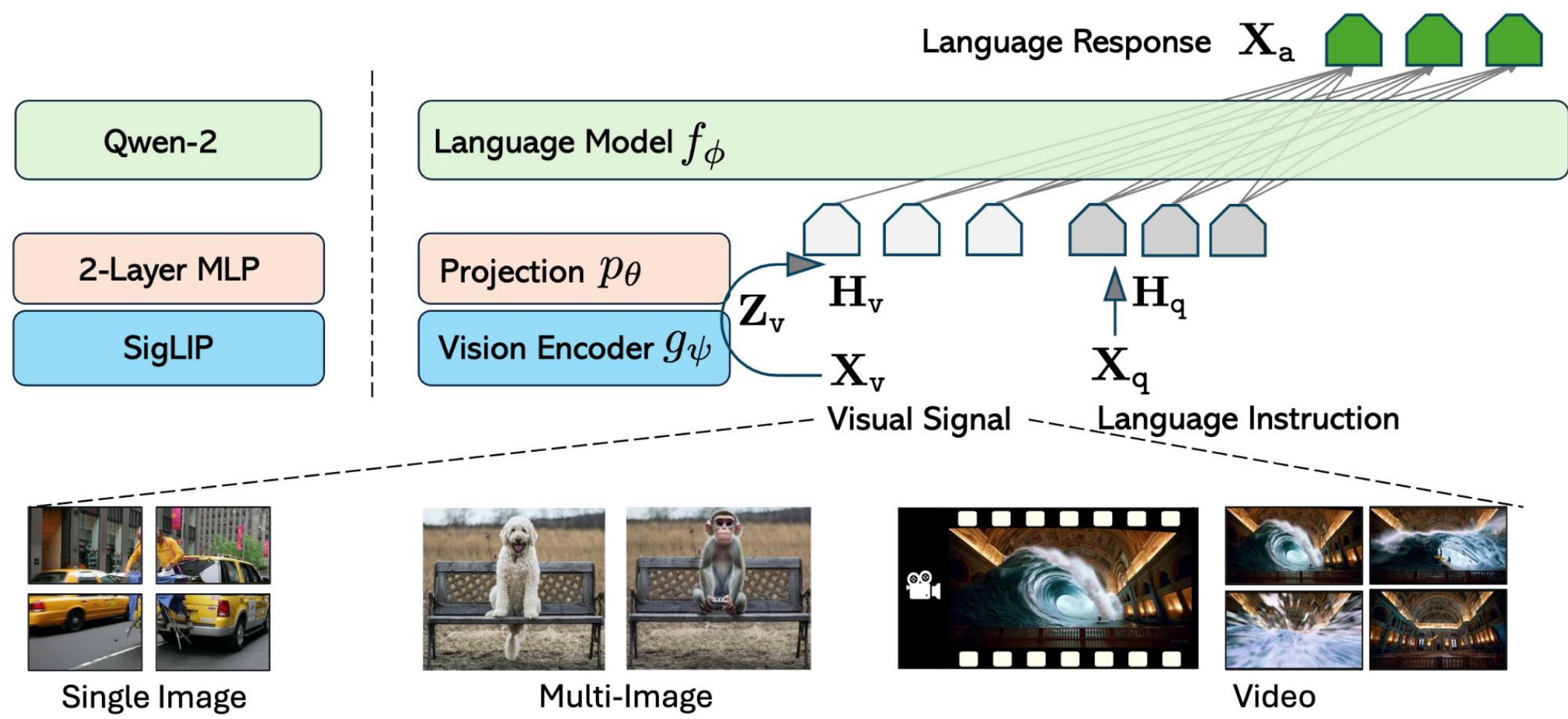


Source: <https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg>

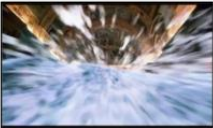
User	What is unusual about this image?
GPT-4 [36]	The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.
User	What is unusual about this image?
BLIP-2	a man is sitting on the back of a yellow cab
User	What is unusual about this image?
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.
User	[Start a new conversation, and clear the history]
LLaVA	What’s happening in the scene? 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



LLaVA-OneVision

 Single-Image	     ... N Crops	$(1 + 9) * 729 = 7290$ Tokens
$729 + N * 729$ Tokens		
 Multi-Image	  ... N Images	$12 * 729 = 8748$ Tokens
$N * 729$ Tokens		
 Video	    ... N Frames	$32 * 196 = 6272$ Tokens
$N * 196$ Tokens		
Example on Token Strategy		Max Tokens

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:



Source image:



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

Qwen Series



<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/VLM right now

Questions?