Multimodal Models: Text and Image

CS6120: Natural Language Processing
Northeastern University

David Smith
with slides from Yejin Choi

Multimodal Systems

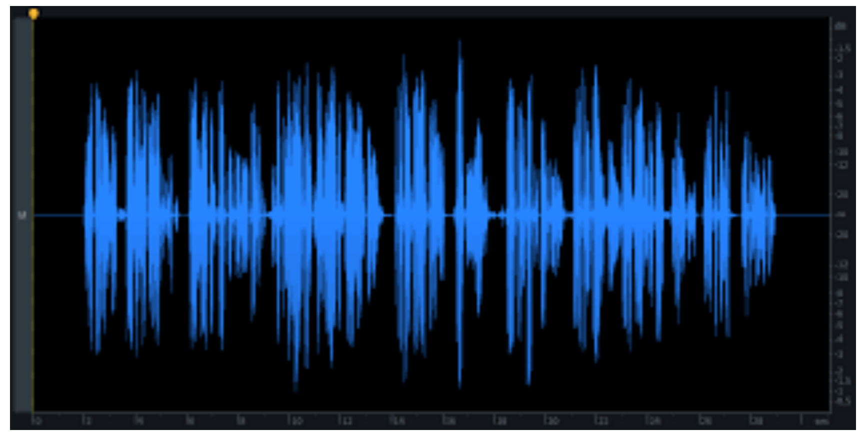
- **Multimodal AI**: System that integrates various data types and sensory inputs (images, videos, audio, other sensory information) to create a unified representation or understanding.

A person throwing a frisbee.

Text

Image

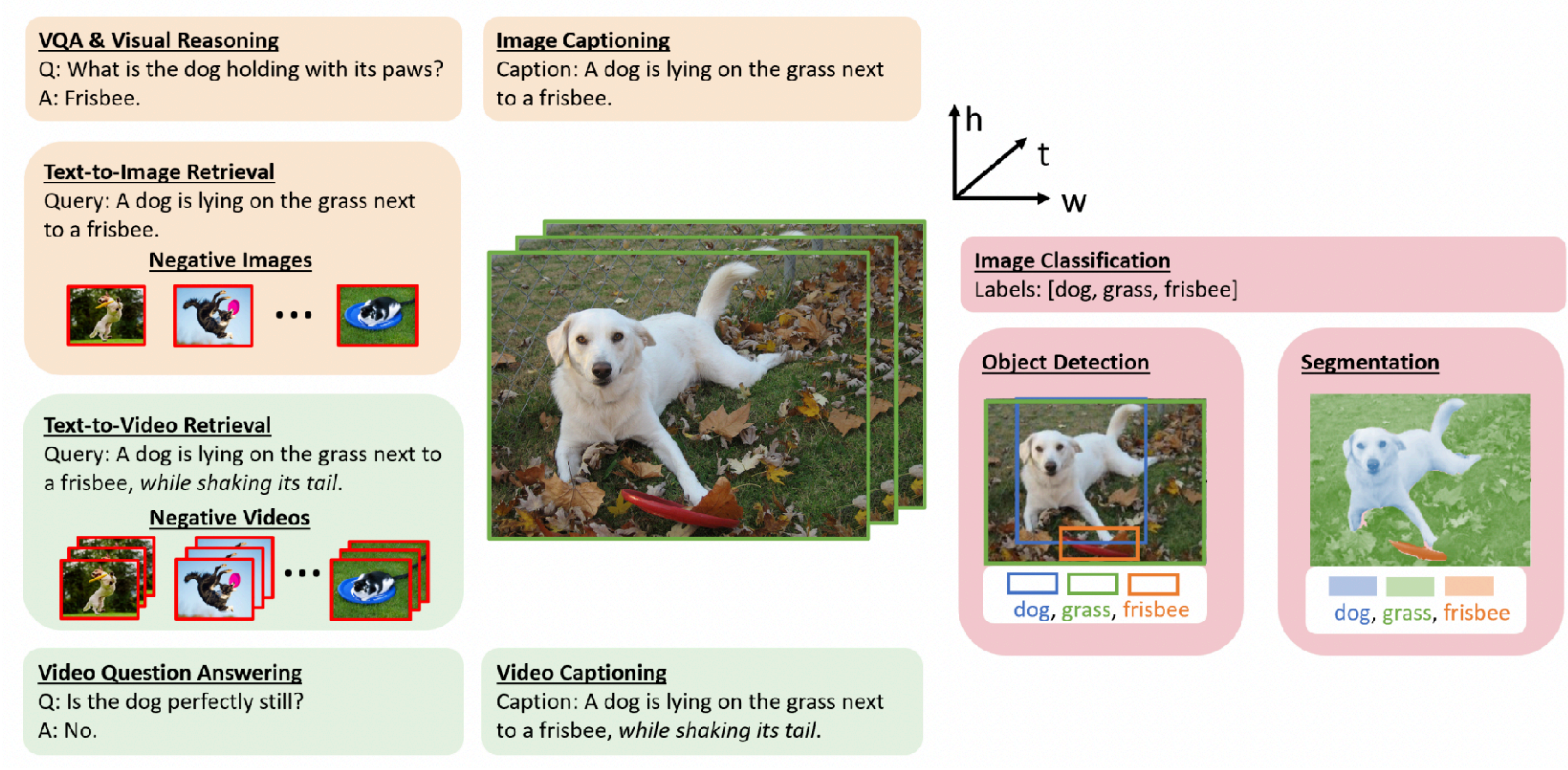
Video



Audio

- This lecture: will focus on **image** & **text** only.

Examples of Multimodal Tasks



Multimodal Language Models

User Can you explain this meme?

Sometimes I just look at pictures of the earth from space and I marvel at how beautiful it all is.

How to train these models?

GPT-4 This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets.
The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the world.
The humor in this meme comes from the unexpected juxtaposition of the text and the image. The text sets up an expectation of a majestic image of the earth, but the image is actually something mundane and silly.

Multimodal Learning (for Image & Text)

Image & Text Alignment

A person throwing
a frisbee.

Image + Text Understanding

What is the object
being thrown?

A frisbee

Text to Image Generation

A person throwing
a frisbee.

Note: For simplicity, we will cover image and text as the two modalities.

Steps of Image-Text Alignment

??

Image Encoder $\rightarrow x_v$

f_v

- **Step1:** Encode different modalities into shared embeddings.
- **Step2**: Bring modalities that encode same meaning into the same space.

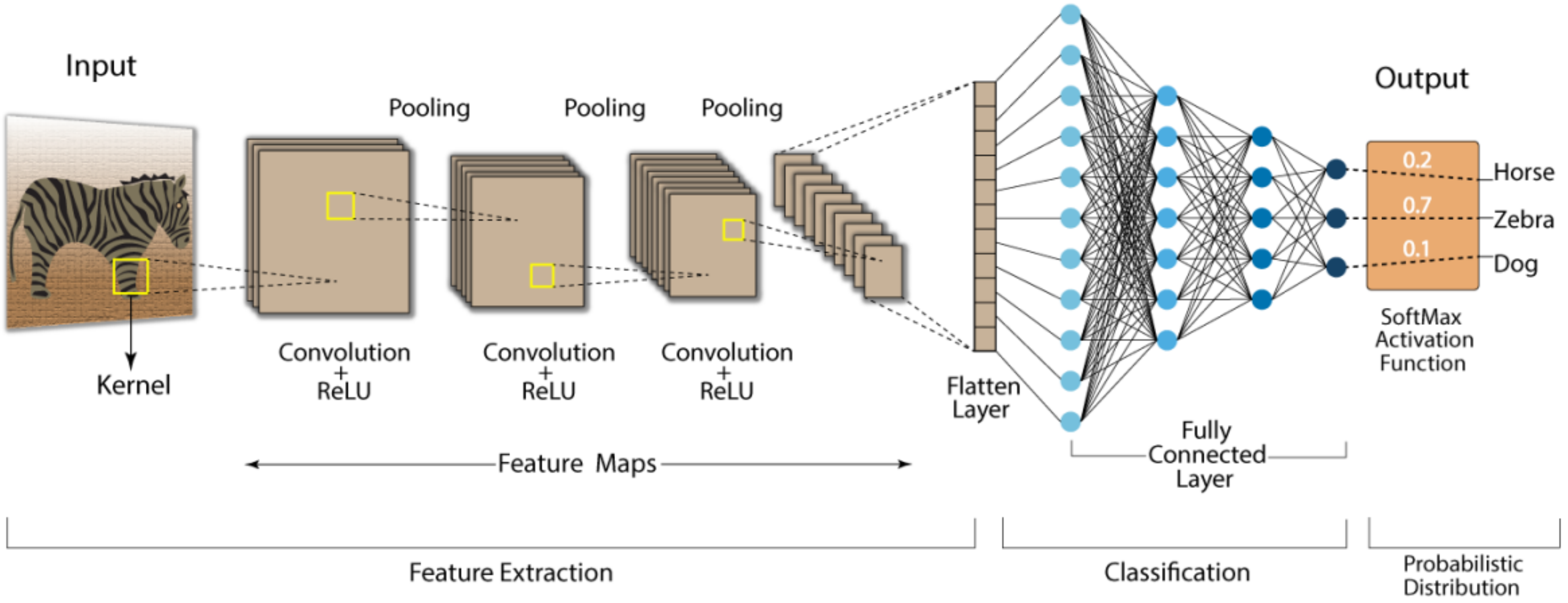
Word2Vec, BERT, ...

A person throwing a frisbee. $\rightarrow$ Text Encoder $\rightarrow x_t$

f_t

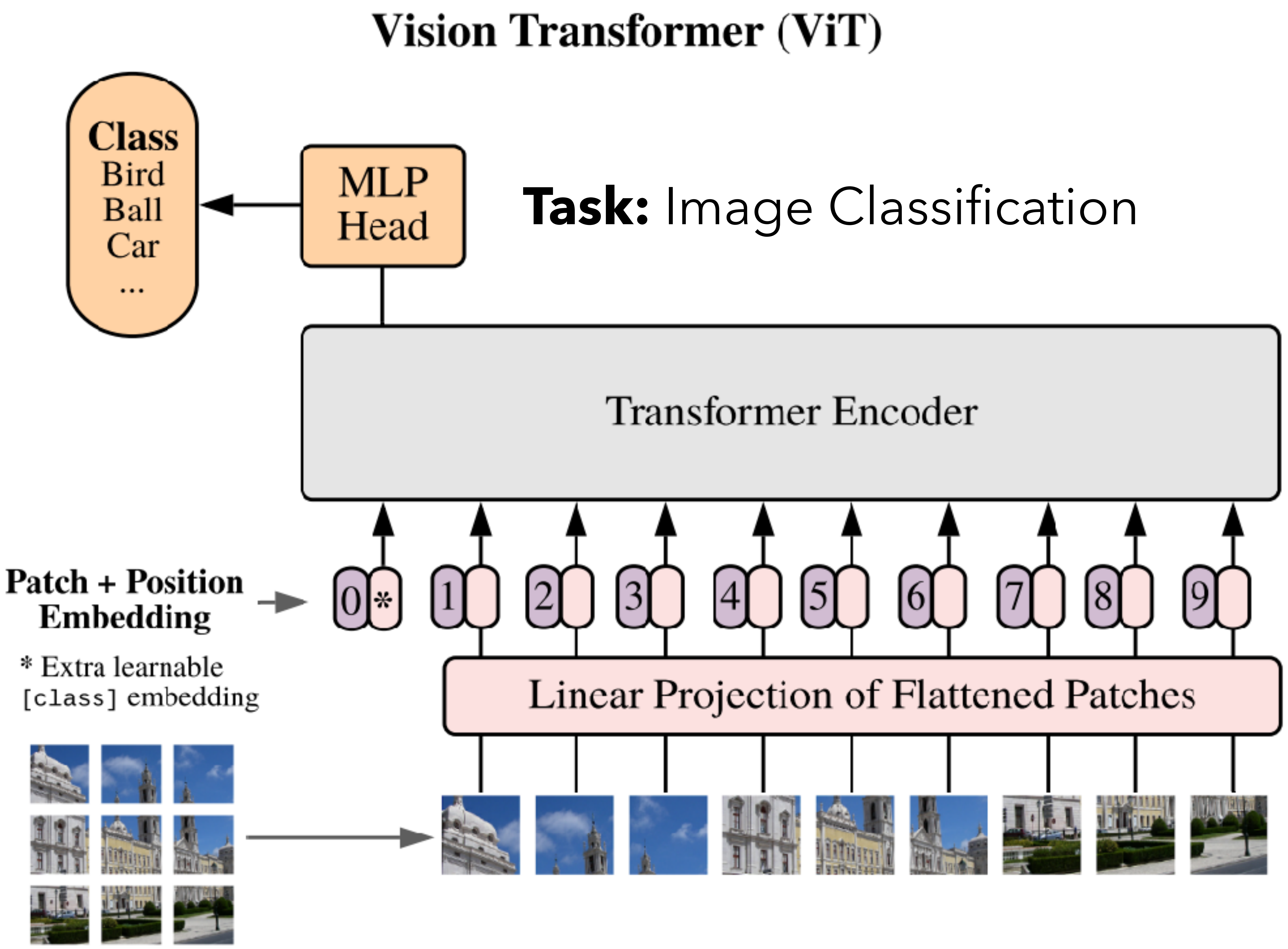
Vision Encoder: Convolutional Neural Networks

- **CNNs**: Extract features that encode spatial and temporal relationships in image with convolution operations
    - **Pooling**: Reduce dimensionality of the convoluted features for efficient computation
- State-of-the-art model for image classification for ~2010s; more in CV course, of course

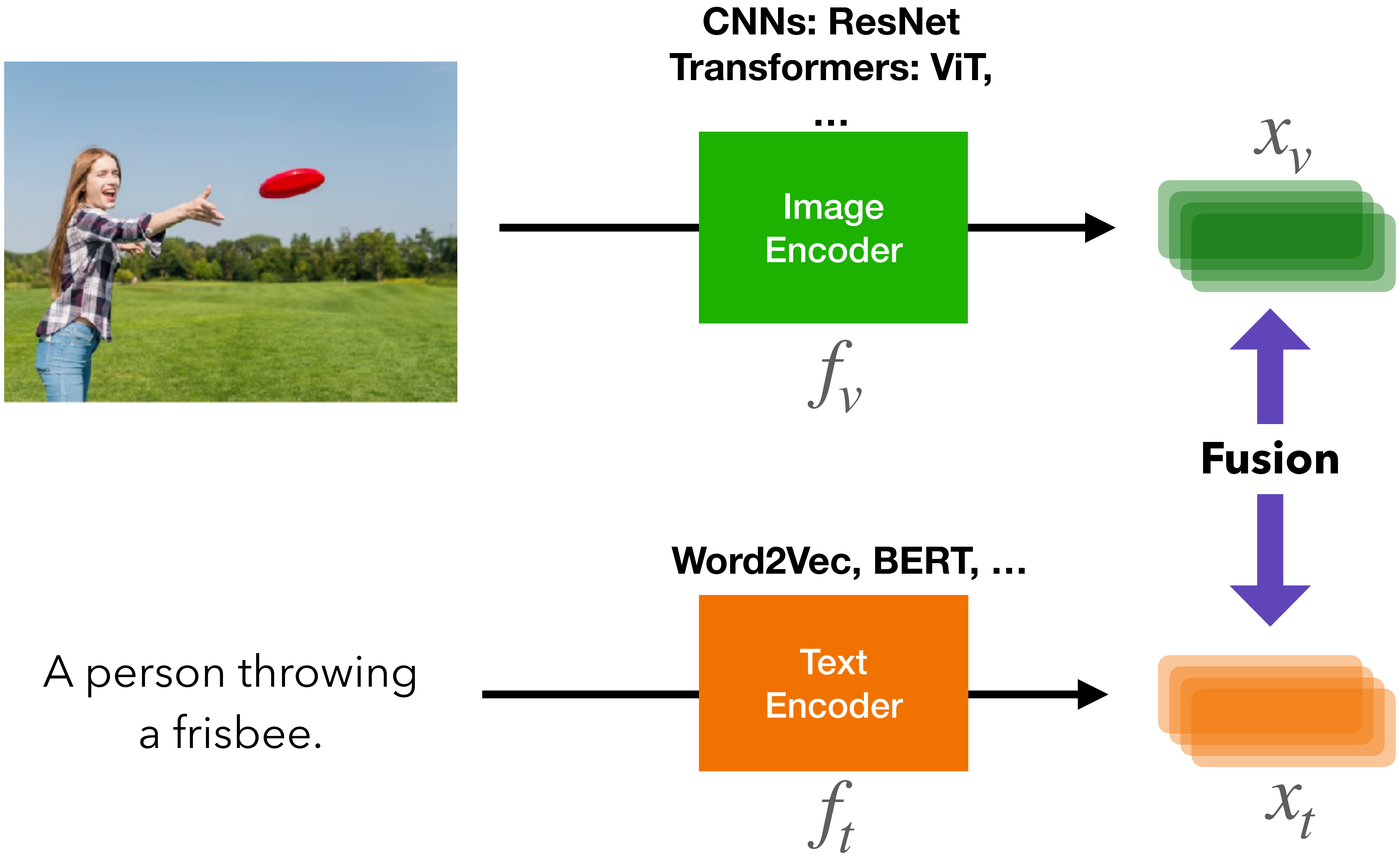


The Vision Transformer: Image Encoding via Patch Tokens

- **Tokenize** images as sequence of "**patches**" of fixed size (e.g. 16 x16 px)
  - Resize images to same size to ensure same number of patches in training.
  - Image Size 224*224px = 14*14 patches
- Use the same transformer encoder architecture in NLP
  - Add [CLS] token for classification tasks.
  - Add positional embedding to be aware of location of patches.
- **Less image-specific inductive bias** than CNNs that encodes translation equivariance and locality.



Steps of Image-Text Alignment



- **Step1:** Encode different modalities into shared embeddings.
- **Step2**: Bring modalities that encode same meaning into the same space.

Step2: Learning to Align Embeddings

$x_v \in \mathbb{R}^v$

Linear Projection

$z_v = W_v x_v^T + b_v^T \in \mathbb{R}^m$

- How to define the **loss function**?

A person throwing a frisbee.

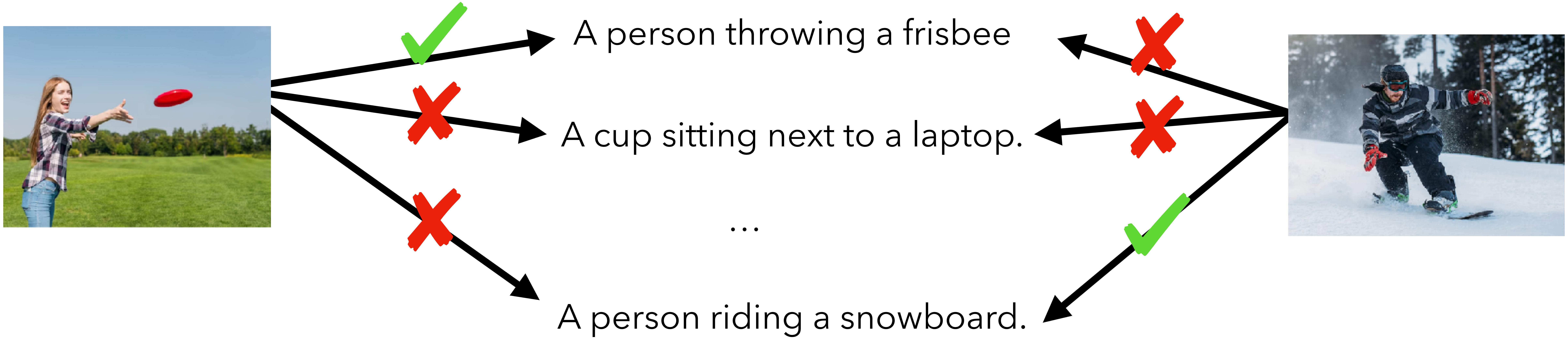
$x_t \in \mathbb{R}^t$

Linear Projection

$z_t = W_t x_t^T + b_t^T \in \mathbb{R}^m$

Contrastive Learning

- **Contrastive Learning**: learn the shared embedding by **contrasting** **positive** and **negative** pairs of instances
  - **Positives**: matched image-text pairs
  - **Negatives**: image-text from mismatched instances
- **Idea: Positive** instances should be closer together in a learned embedding space, while **Negatives** should be farther apart.

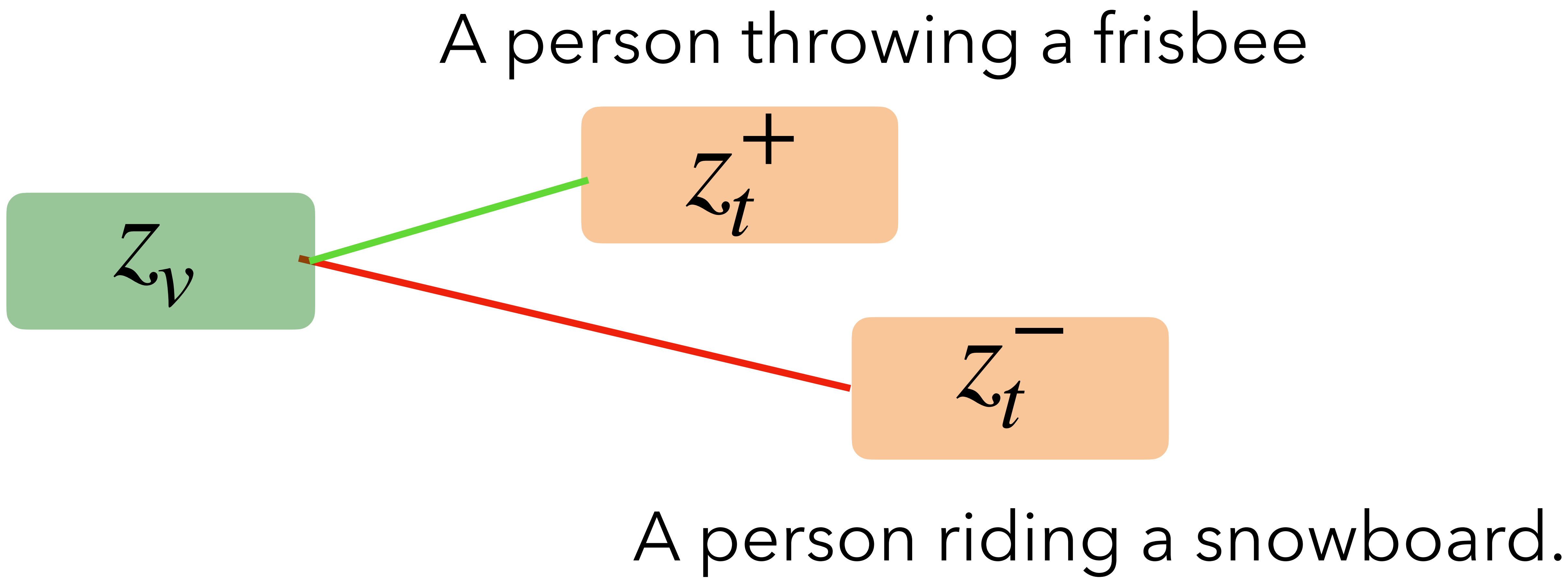


Contrastive Learning

- Adjust similarity of learned embeddings with a distance metric.
  - Euclidean Distance
  - Cosine Similarity

$$\cos(u, v) = \frac{u \cdot v}{||u||_2 ||v||_2}$$

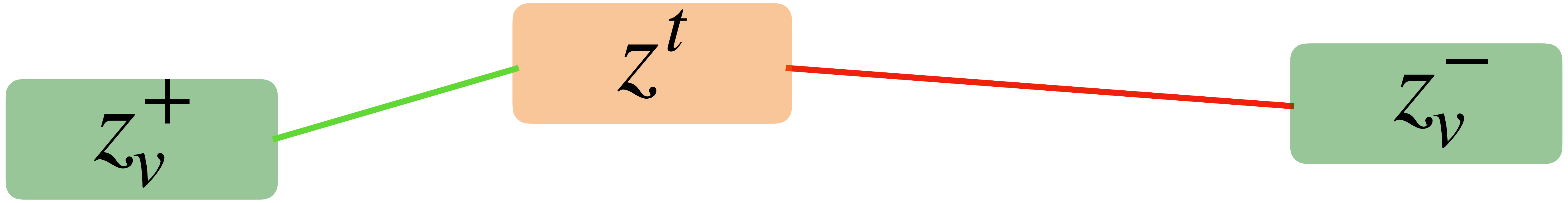
- $\text{sim}(z_v, z_t^+) >> \text{sim}(z_v, z_t^-)$



Contrastive Learning

- Adjust similarity of learned embeddings with a distance metric.
    - Euclidean Distance
    - Cosine Similarity $\cos(u, v) = \dfrac{u \cdot v}{||u||_2 ||v||_2}$
- sim(z_v, z_t^{+}) >> sim(z_v, z_{-}^{+}) + sim(z_v^{+}, z_t) >> sim(z_v^{-}, z_t)

A person throwing a frisbee



Contrastive Learning

margin parameter: min distance b.w. positive and negatives

- Adjust similarity of learned embeddings with a distance metric.
  - Euclidean Distance
  - Cosine Similarity

Triplet Loss

$$\boxed{\begin{aligned}&\max(0, \mathrm{sim}(z_v, z_t^+) - \mathrm{sim}(z_v, z_t^-) + m) + \\ &\max(0, \mathrm{sim}(z_v^+, z_t) - \mathrm{sim}(z_v^-, z_t) + m)\end{aligned}}$$

- $\mathrm{sim}(z_v, z_t^+) >> \mathrm{sim}(z_v, z_t^-)$ + $\mathrm{sim}(z_v^+, z_t) >> \mathrm{sim}(z_v^-, z_t)$

A person throwing a frisbee

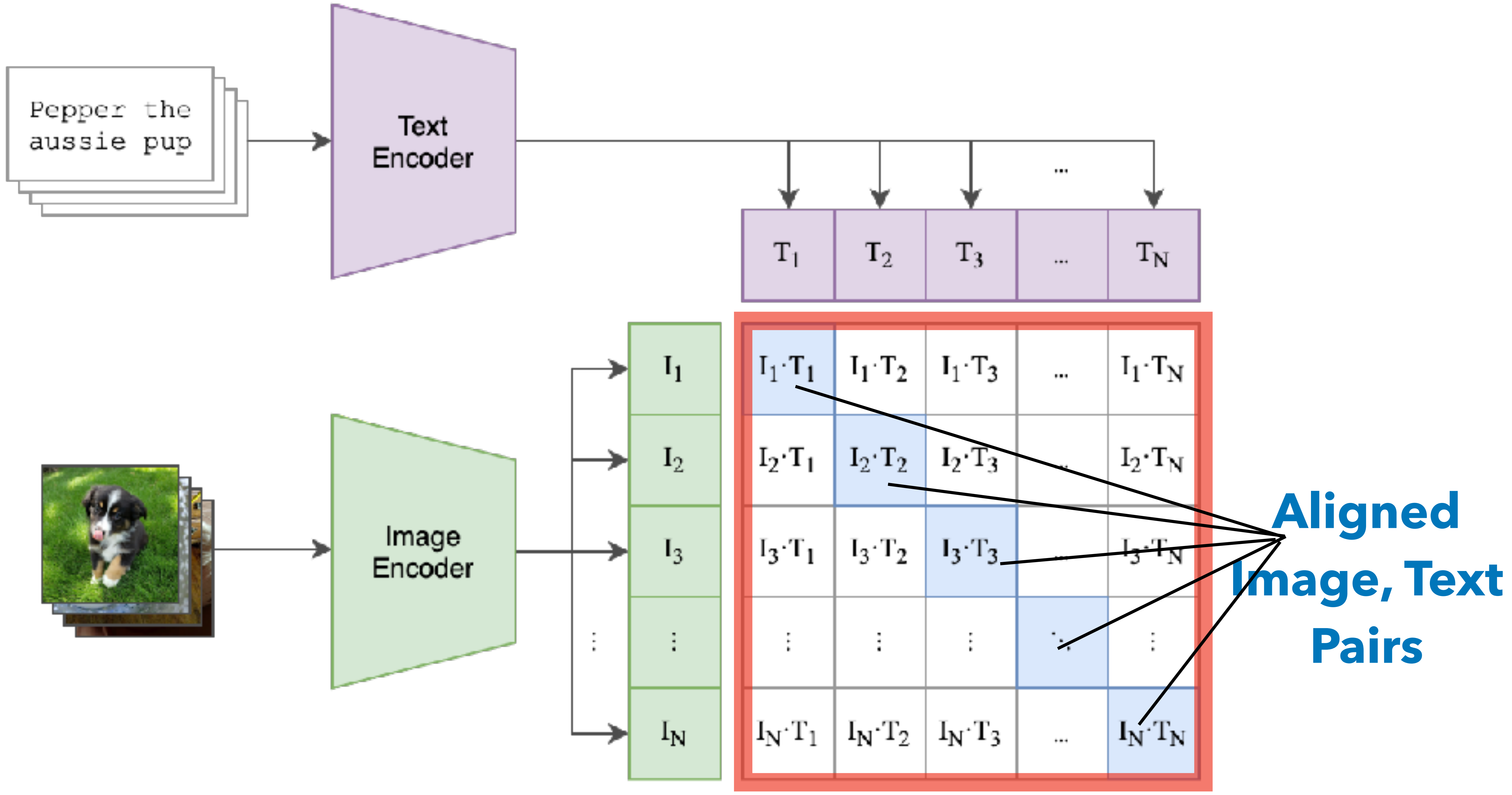
z_v^+ — z^t — z_v^-

https://www.v7labs.com/blog/triplet-loss#:~:text=Triplet%20loss%20is%20a%20way,a%20dissimilar%20item%20(negative).

A Different View of Contrastive Learning

- What does this look like?
- Classification over distance embedding!

CLIP: Contrastive Language-Image Pre-Training



Objective: given a batch of N (image, text) pairs, predict which of the N × N possible (image, text) pairings across a batch actually occurred.

Minimize InfoNCE Loss

$$L_{NCE} = -\log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_j / \tau)}{\sum_{k=0}^{N} \exp(\mathbf{z}_i \cdot \mathbf{z}_k / \tau)}$$

```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# 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
```

Use the [CLS] token for transformers

Contrastive Learning as Binary Classification

- Contrastive Learning as Classification among the Batch Instances?
- Why does this work?
- Inspiration from Information Theory:
  - **Maximize** the **agreement** between the **positive pairs** and **minimize** the **agreement** between the **negative pairs**

Contrastive Learning as Binary Classification

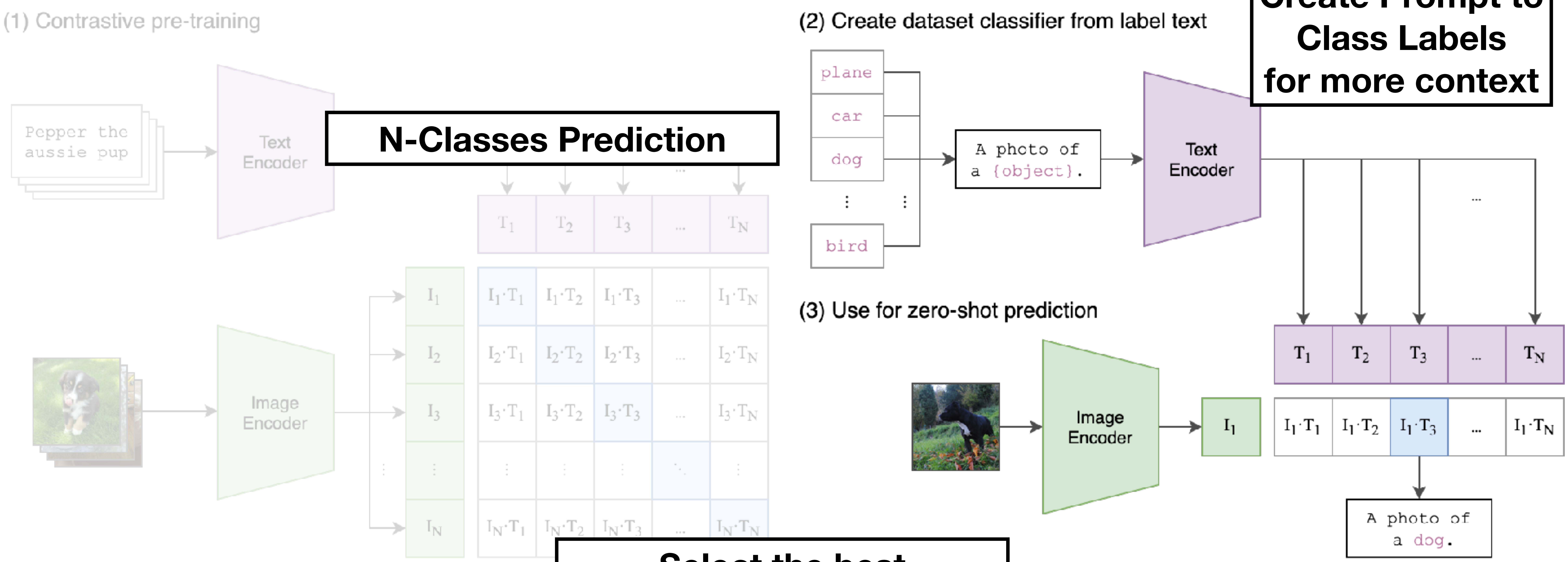
- Specifically, maximize the **mutual information** between the two variables.

$$I(X;Y) = \sum_{(X,Y)} p(X,Y) \log \frac{p(X \mid Y)}{p(X)}$$

$$= \mathrm{KL}\big(p(X,Y) \,||\, p(X)p(Y)\big)$$

- If (X,Y) are independent/not related, information = 0. If I see an image, then I don’t know anything about caption from a different image.
- If (X,Y) agree with each other, information is H(X). Knowing about Y gives me enough information about what X is.
- In our case, we want to maximize the agreement between positive pairs and minimizes the agreement between negative pairs

CLIP: Contrastive Language-Image Pre-Training



N-Classes Prediction

Create Prompt to Class Labels for more context

Select the best text prompt that gives the highest similarity.

Enables Open Vocabulary Classification class labels.

Image-Text Training Dataset

- Previous Image-Text Pre-Training Dataset
  - Leverage filtered, carefully annotated dataset for academic research
  - 10M was considered as "large-scale" pre-training

	COCO	VG	SBU	CC3M	Total
#Images	113K	108K	875K	3.1M	4.2M
#Captions	567K	5.4M	875K	3.1M	10M

Table 3.2: Statistics of the pre-training datasets used in a typical academic setting.

Image-Text Training Dataset

- Previous Image-Text Pre-Training Dataset
  - Leverage filtered, carefully annotated dataset for academic research
  - 10M was considered as “large-scale” pre-training
- **CLIP: 400M** Image-Text pairs crawled from web
  - Unfiltered, highly varied, and highly noisy data
  - Covers much more diverse concepts and images

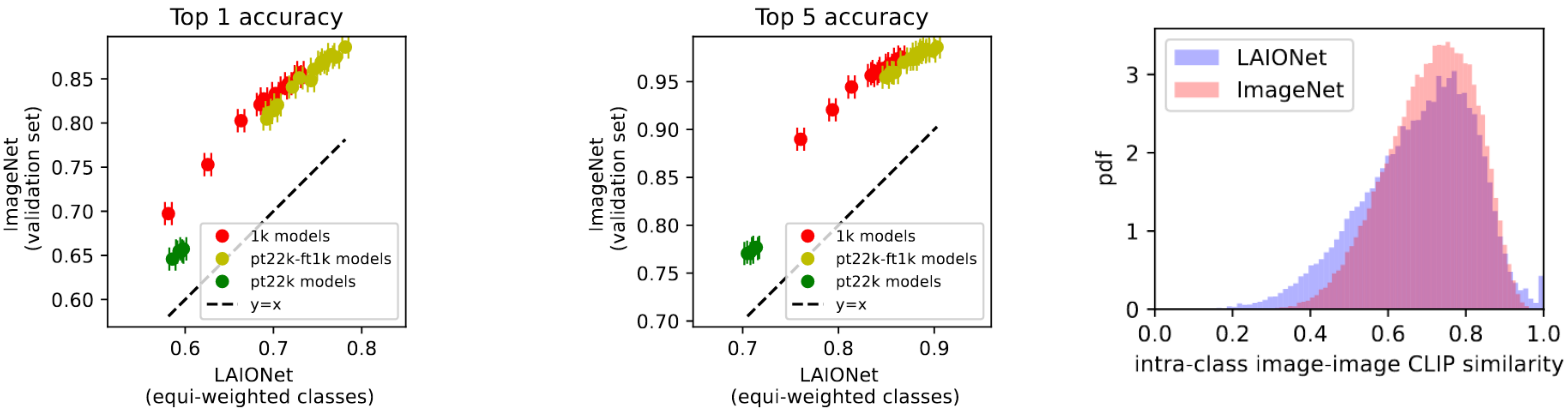
Image-Text Training Dataset

- **CLIP Training Data:** 400M Image-Text Pairs crawled from the web
  - Wasn't open to public for training
- **LAION Dataset**: 400M/5B Image and alt-text attributes

Dataset	Year	Num. of Image-Text Pairs	Language	Public
SBU Caption [92] [link]	2011	1M	English	✓
COCO Caption [93] [link]	2016	1.5M	English	✓
Yahoo Flickr Creative Commons 100 Million (YFCC100M) [94] [link]	2016	100M	English	✓
Visual Genome (VG) [95] [link]	2017	5.4 M	English	✓
Conceptual Captions (CC3M) [96] [link]	2018	3.3M	English	✓
Localized Narratives (LN) [97] [link]	2020	0.87M	English	✓
Conceptual 12M (CC12M) [98] [link]	2021	12M	English	✓
Wikipedia-based Image Tex (WIT) [99] [link]	2021	37.6M	108 Languages	✓
Red Caps (RC) [100] [link]	2021	12M	English	✓
CLIP [14]	2021	400M	English	✗
LAION400M [28] [link]	2021	400M	English	✓
LAION5B [27] [link]	2022	5B	Over 100 Languages	✓

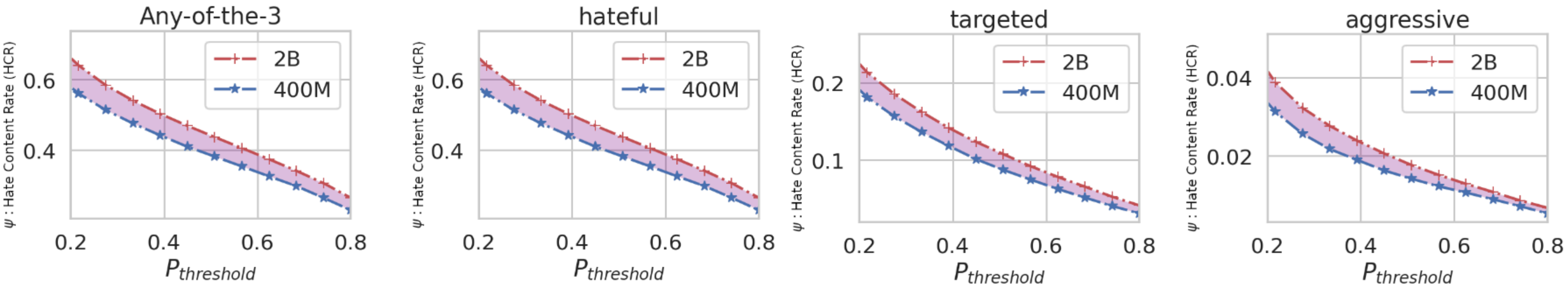
Scale isn't the only difference

- When they recreate ImageNet by querying the open LAION dataset, Shirali and Hardt (2023) find that concepts ("synsets" in ImageNet) are more diverse.



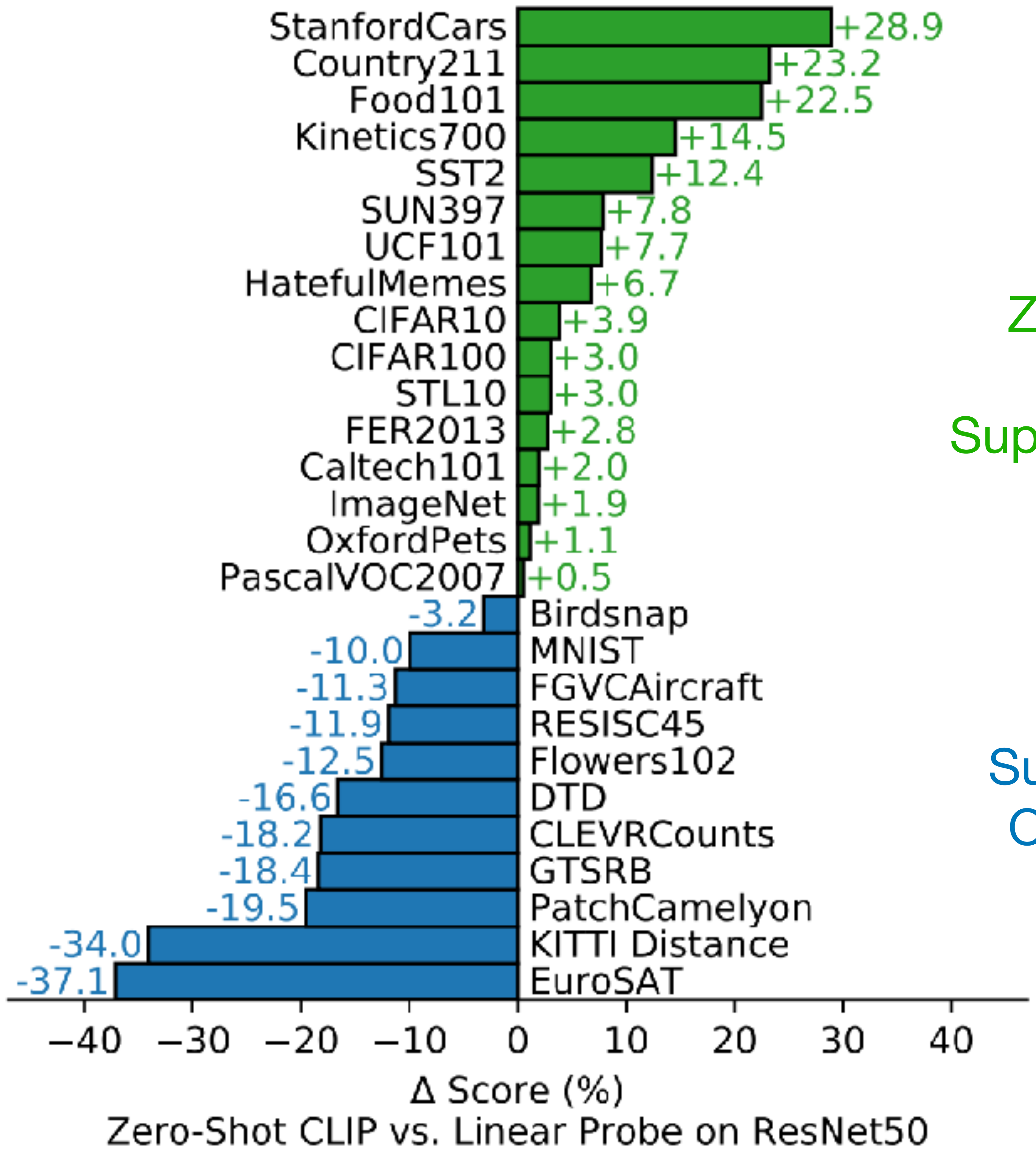
Undesirable content can increase super-linearly

- Birhane et al. (2023) analyze scaling up of LAION to show increasing amounts of undesirable content in captions that evade image-only filters.



Text Supervision Enables Strong Zero-Shot Performance in Vision Tasks

- Large-Scale Training on Noisy Image-Text Data -> Great Zero-Shot Performance
- **Zero-Shot CLIP** is **competitive with fully supervised** Resnet50 in Image Classification
    - *Linear Probe*: Train linear layer on top of fixed, pre-trained embeddings.

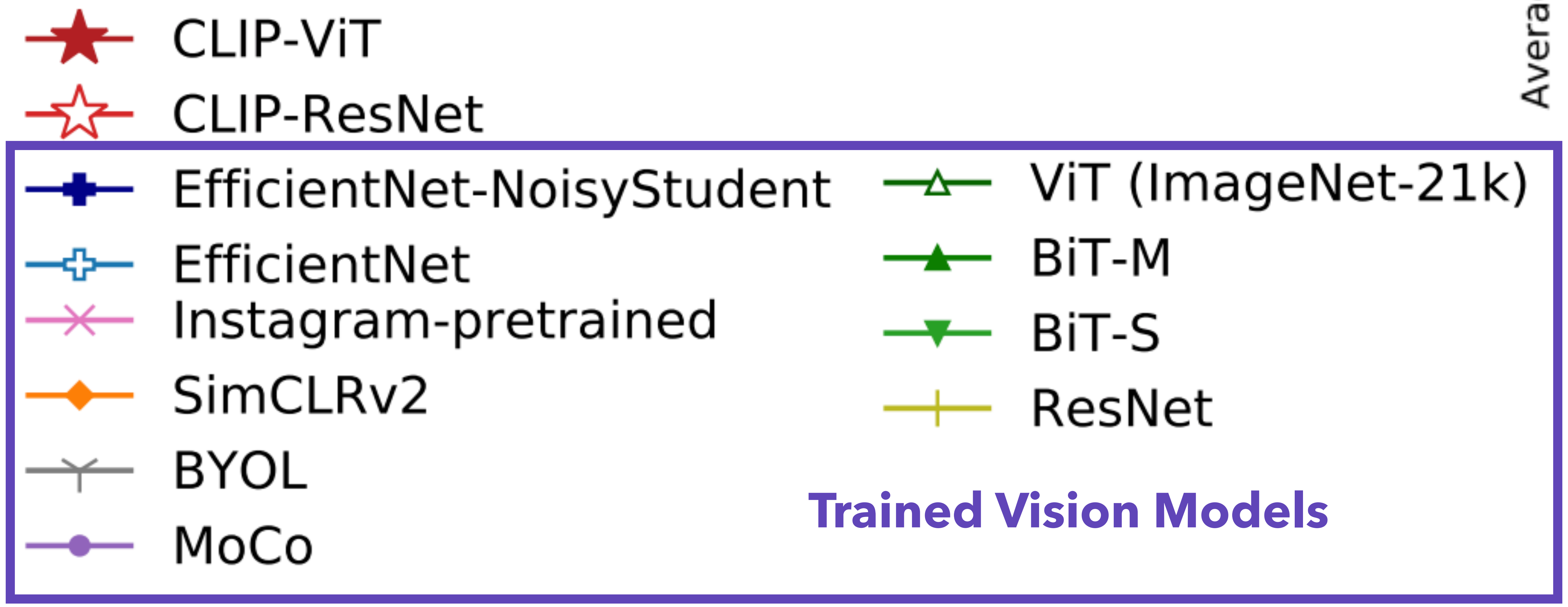


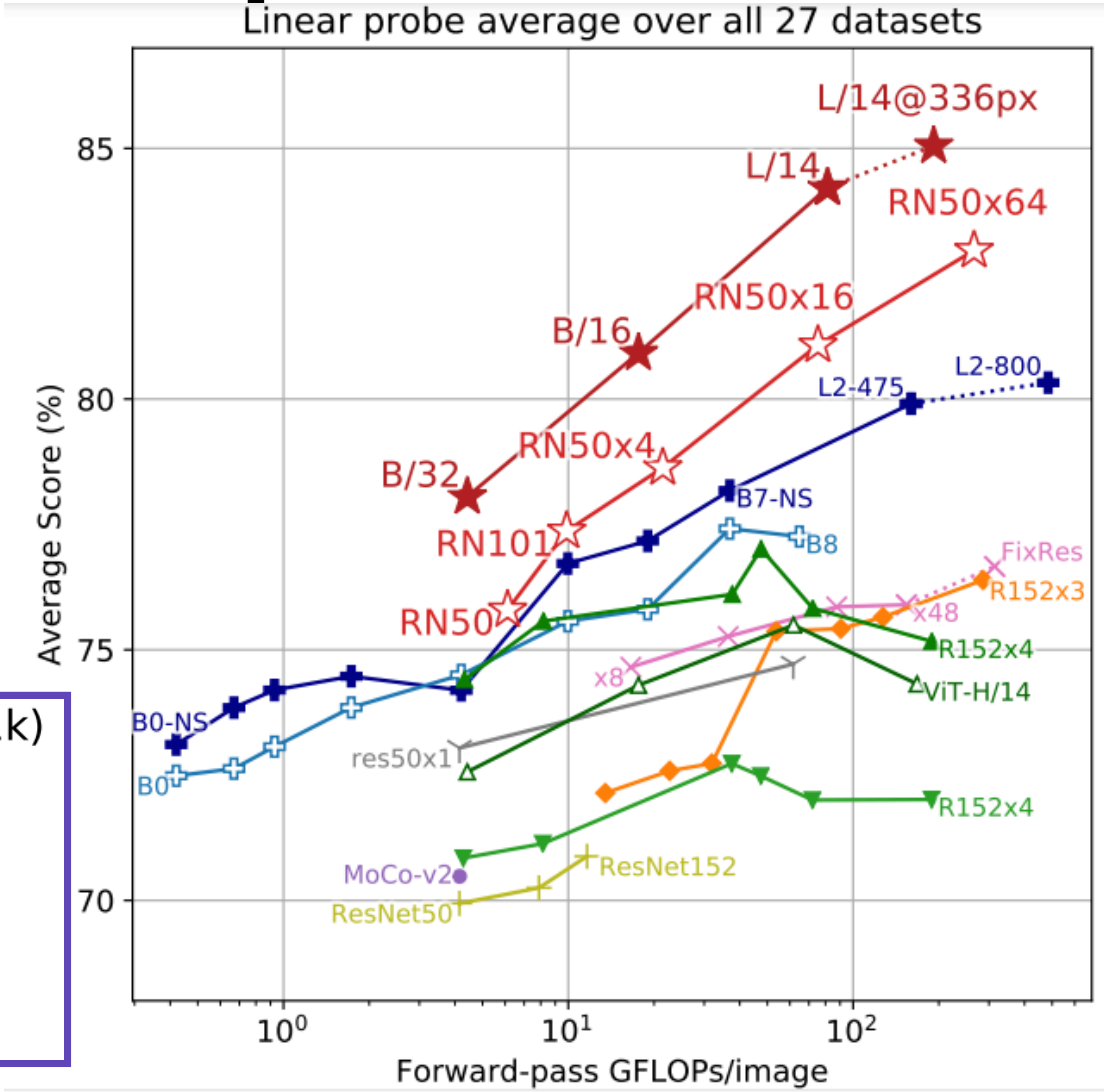
Zero-Shot CLIP Outperforms Supervised baseline.

Supervised baseline Outperforms CLIP.

CLIP vs Unimodal Visual Representations

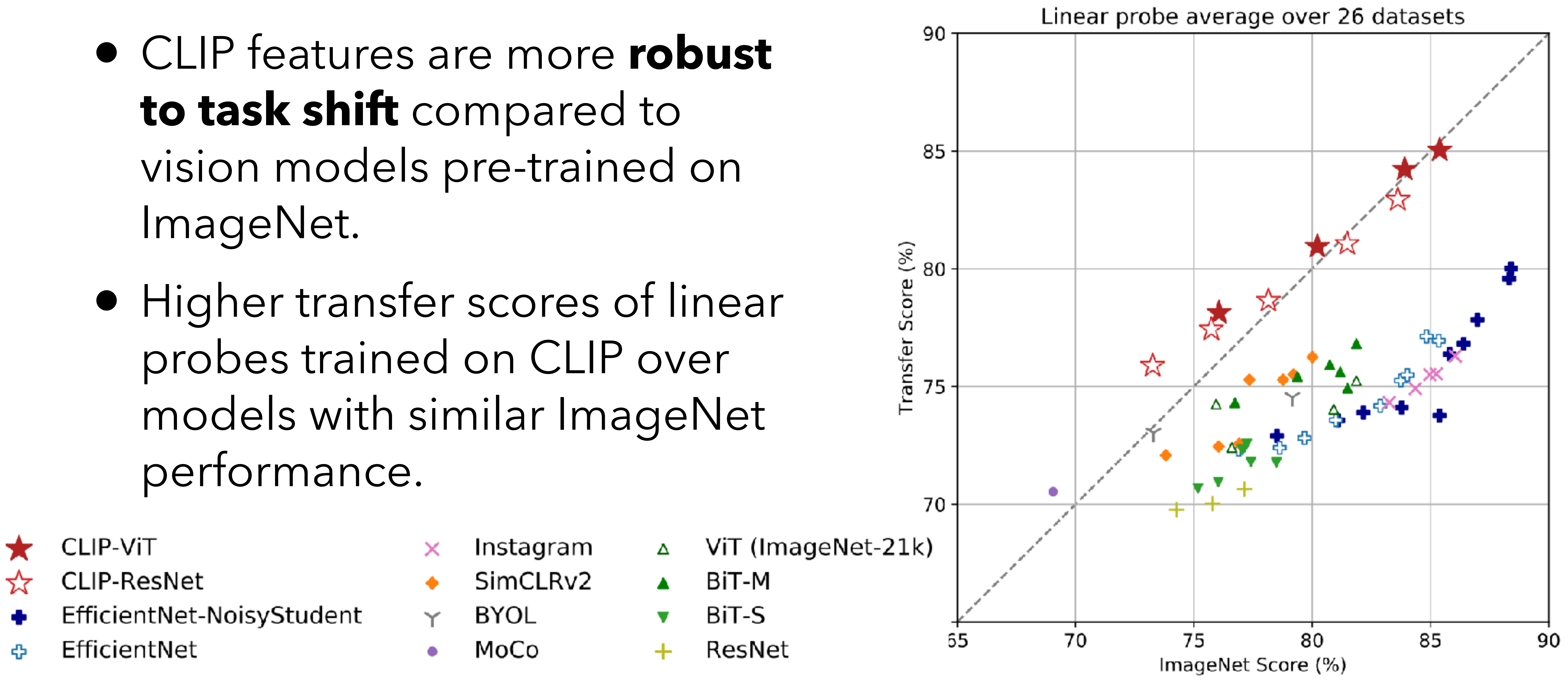
- Linear Probe performance vs. computer vision models
- CLIP provides visual representations with better transferability





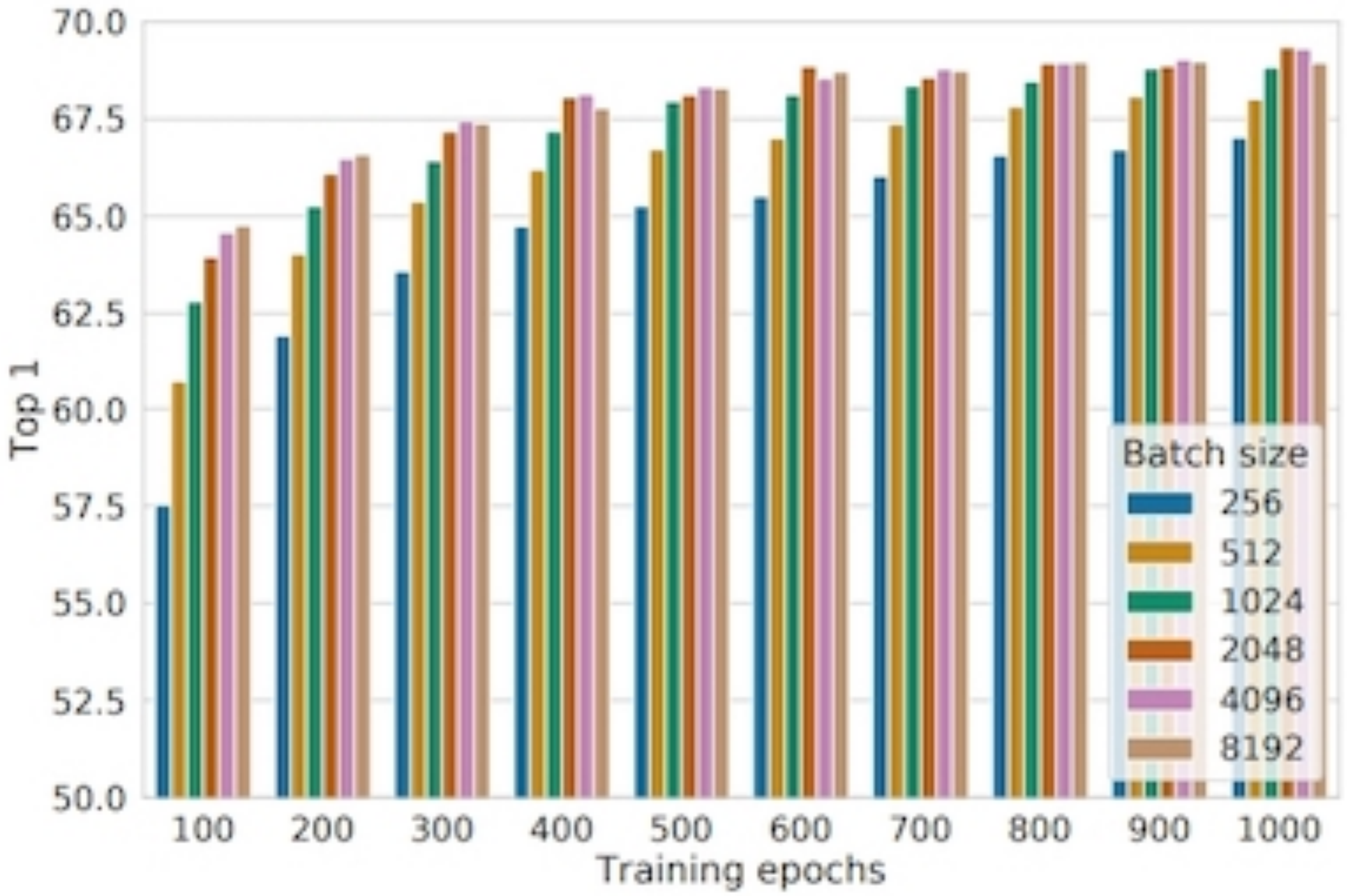
CLIP vs Unimodal Visual Representations

- CLIP features are more **robust to task shift** compared to vision models pre-trained on ImageNet.
- Higher transfer scores of linear probes trained on CLIP over models with similar ImageNet performance.

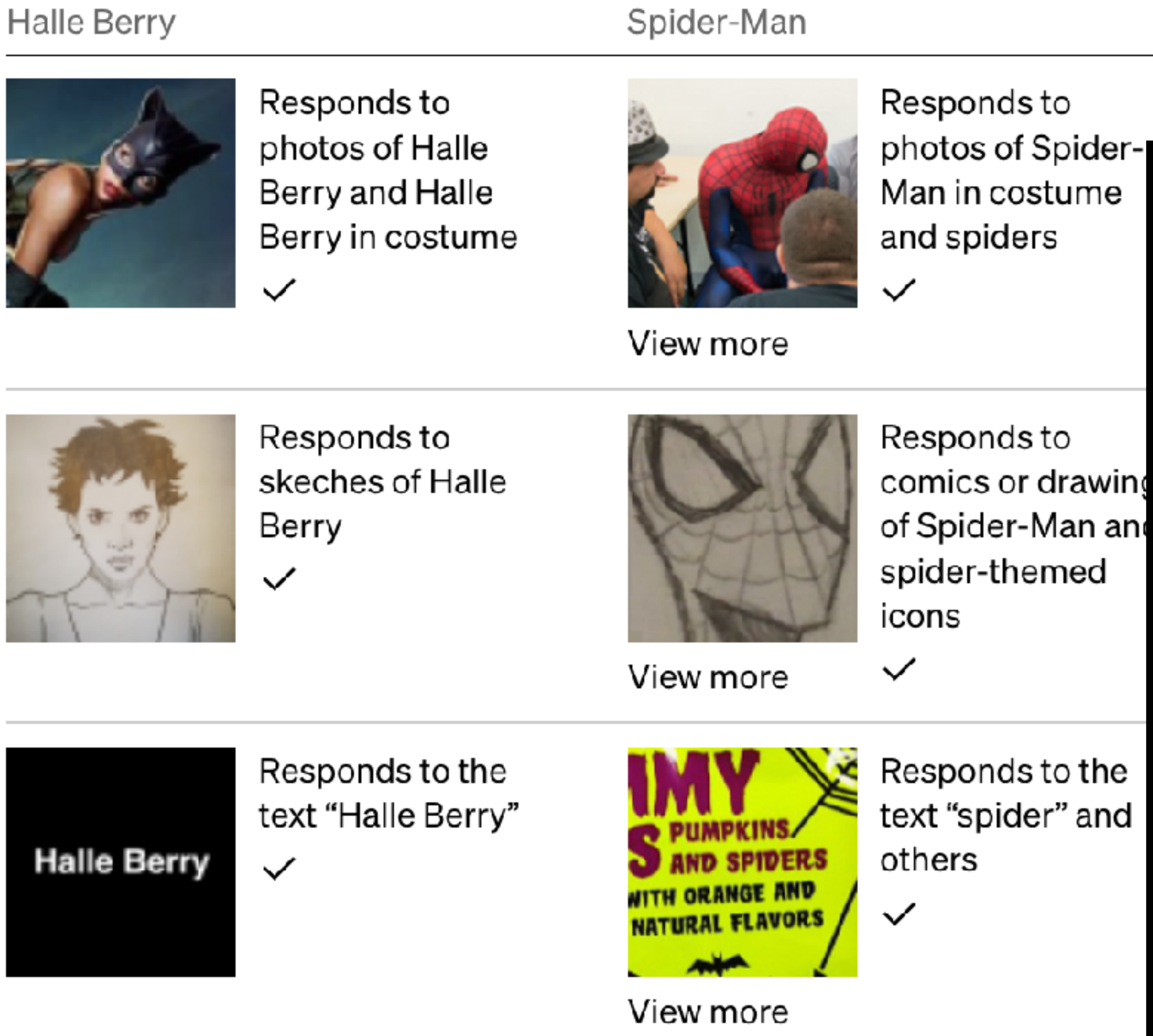


Why is CLIP so good?

- Learning **visual representation** with **language supervision**: learns visual concepts much more efficiently.
- Exploited Scalability benefits:
    - 256 GPUS + 4096 batch size with 2 weeks of training
    - Large batch size in Contrastive Learning
        - More negatives to compare against.
        - More challenging task to distinguish the negatives, requiring fine-grained visual recognition.



Understanding Multimodal Capabilites of CLIP



- Aligns images to **semantic concepts** thanks to **language supervision**, rather than just aligning texture and shapes.
- Case where multimodal learning was a big breakthrough for learning high-quality, unimodal representations (image)

Vision and Language Systems

Image & Text Alignment

A person throwing
a frisbee.

Image to Text Understanding

What is the object
being thrown?

A frisbee

Text to Image Generation

A person throwing
a frisbee.

Note: For simplicity, we will cover image and text as the two modalities.

CLIP for Visual Reasoning?

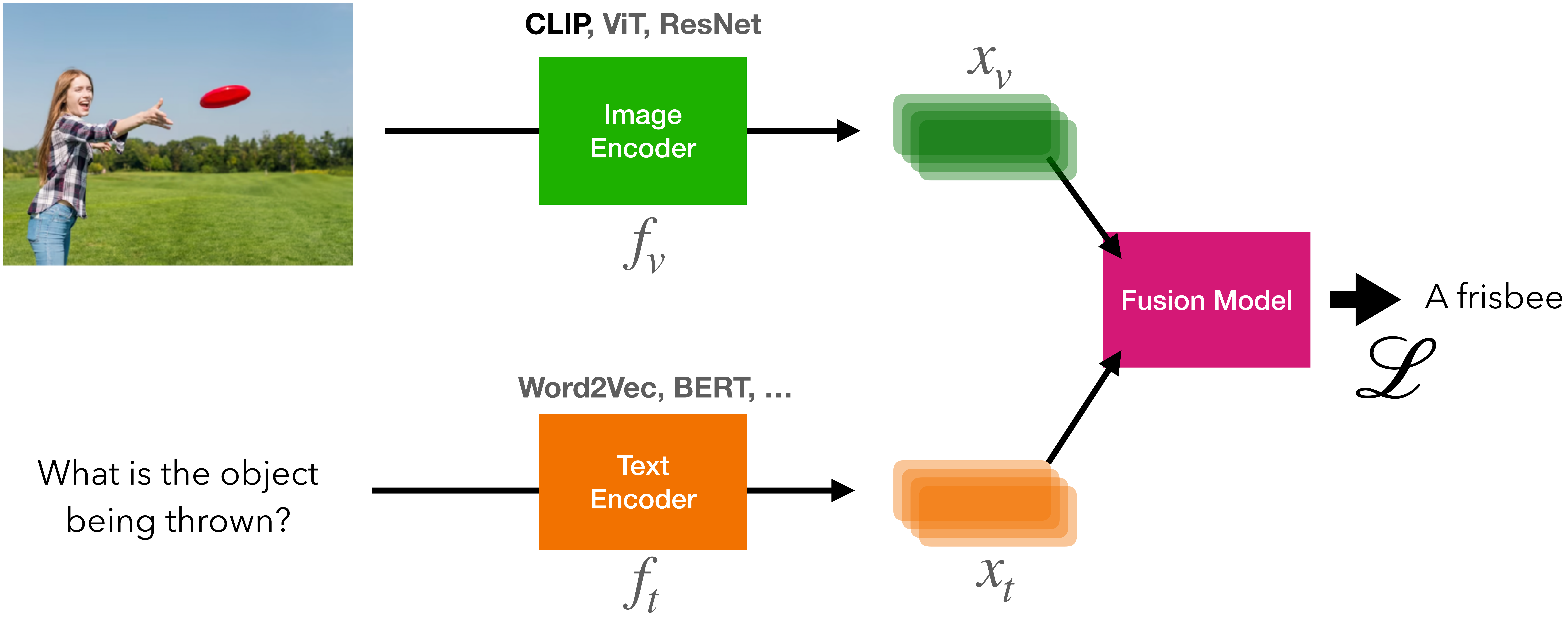
- Supports retrieval but not capable of generation
- **VQA Prompt**: *"question: [question text] answer: [answer text]"*
- Note: CLIP is trained to align images with alt-text captions
  - Not suitable for reasoning tasks such as question answering.

Model	VQA Question Type		
	yes/no	number	other
CLIP-Res50	0.037	0.057	0.0
CLIP-ViT-B_{PE}	0.019	0.0	0.0
CLIP-$Res50_{PE}$	0.055	0.057	0.0
CLIP-$Res101_{PE}$	0.260	0.0	0.0
CLIP-$Res50x4_{PE}$	0.446	0.118	0.034

Table 7: Zero-shot performance of CLIP on VQA v2.0 `mini-eval`, "PE" denotes we follow similar prompt engineering as suggested in CLIP paper.

Near Chance Performance

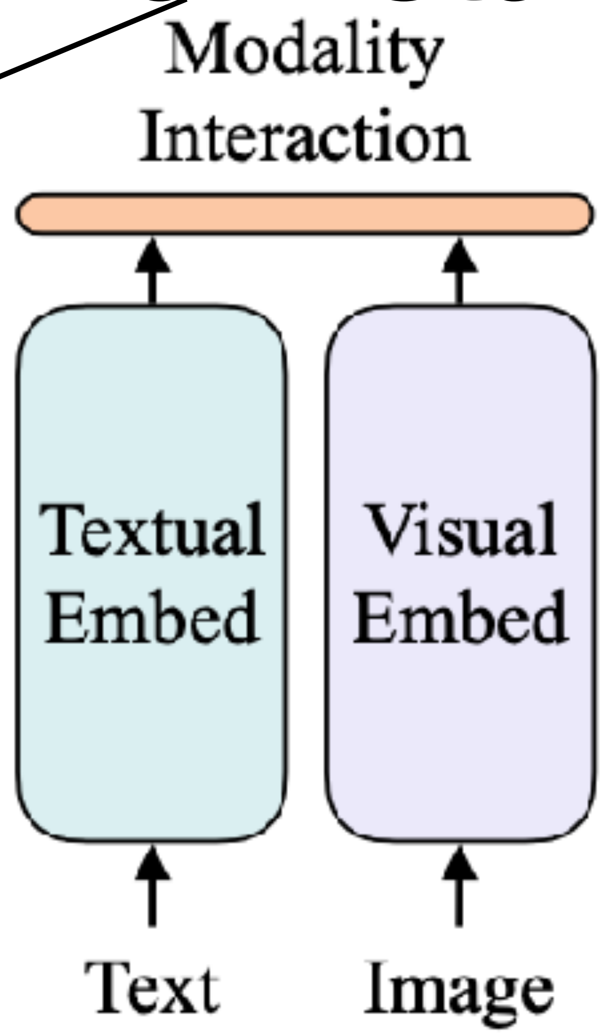
Image and Text Understanding



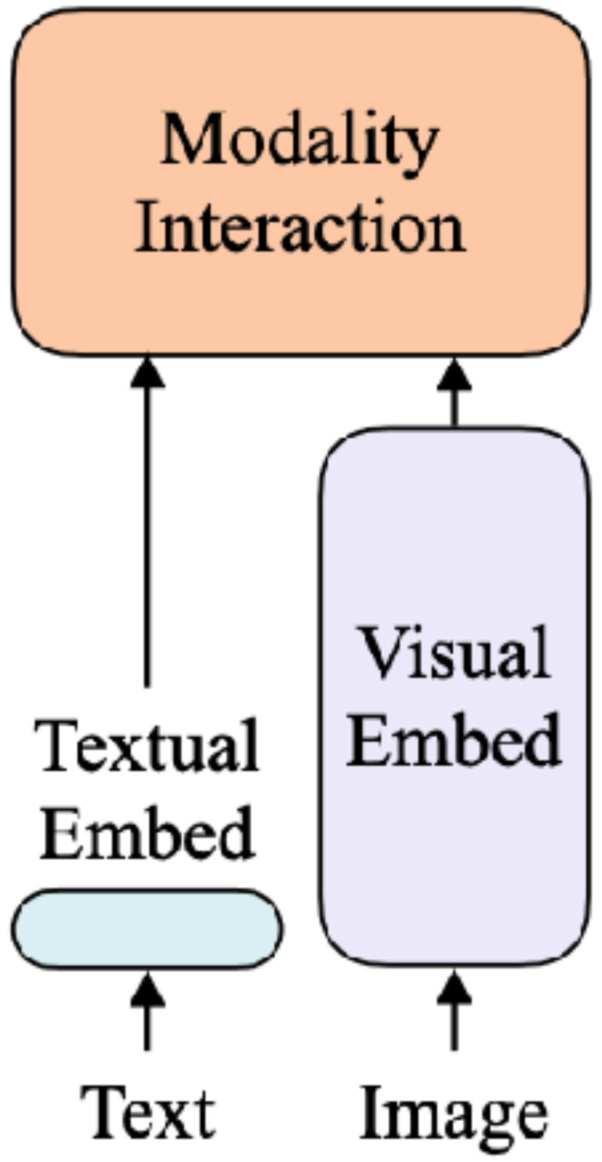
Embedding vs Fusion Trade Offs

Cosine Distance

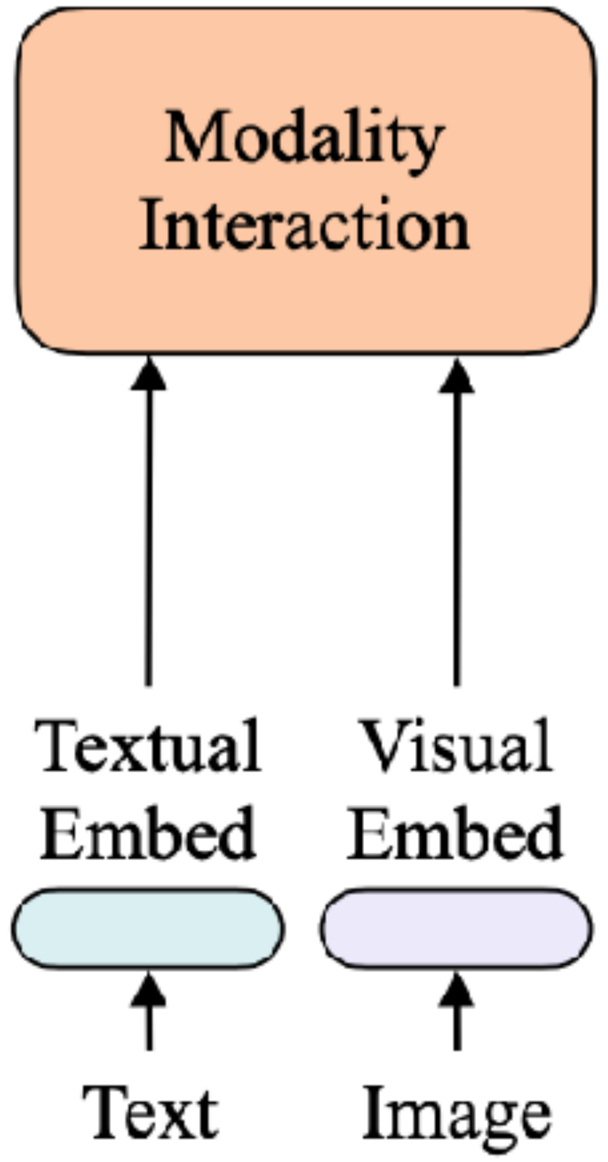
Enough for image-text matching



CLIP

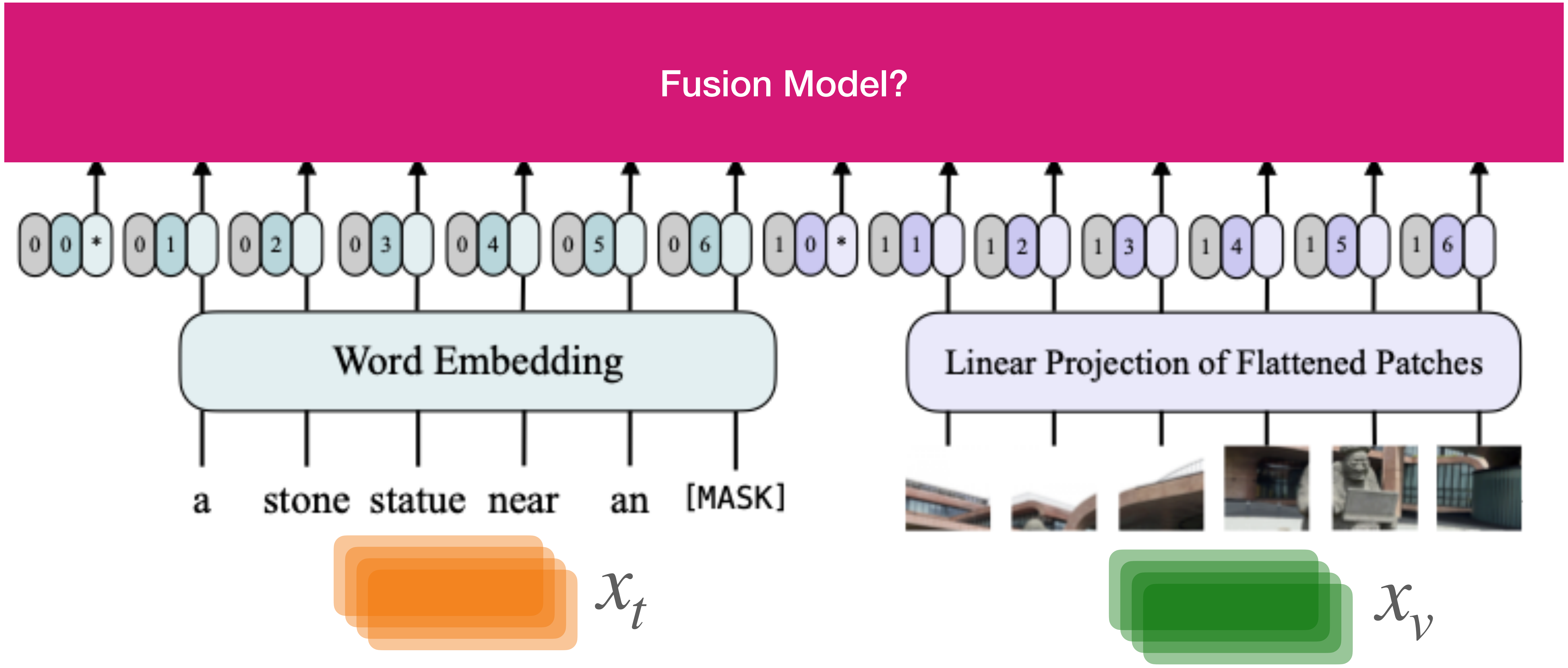


Perhaps, need stronger fusion mechanism for complex reasoning tasks

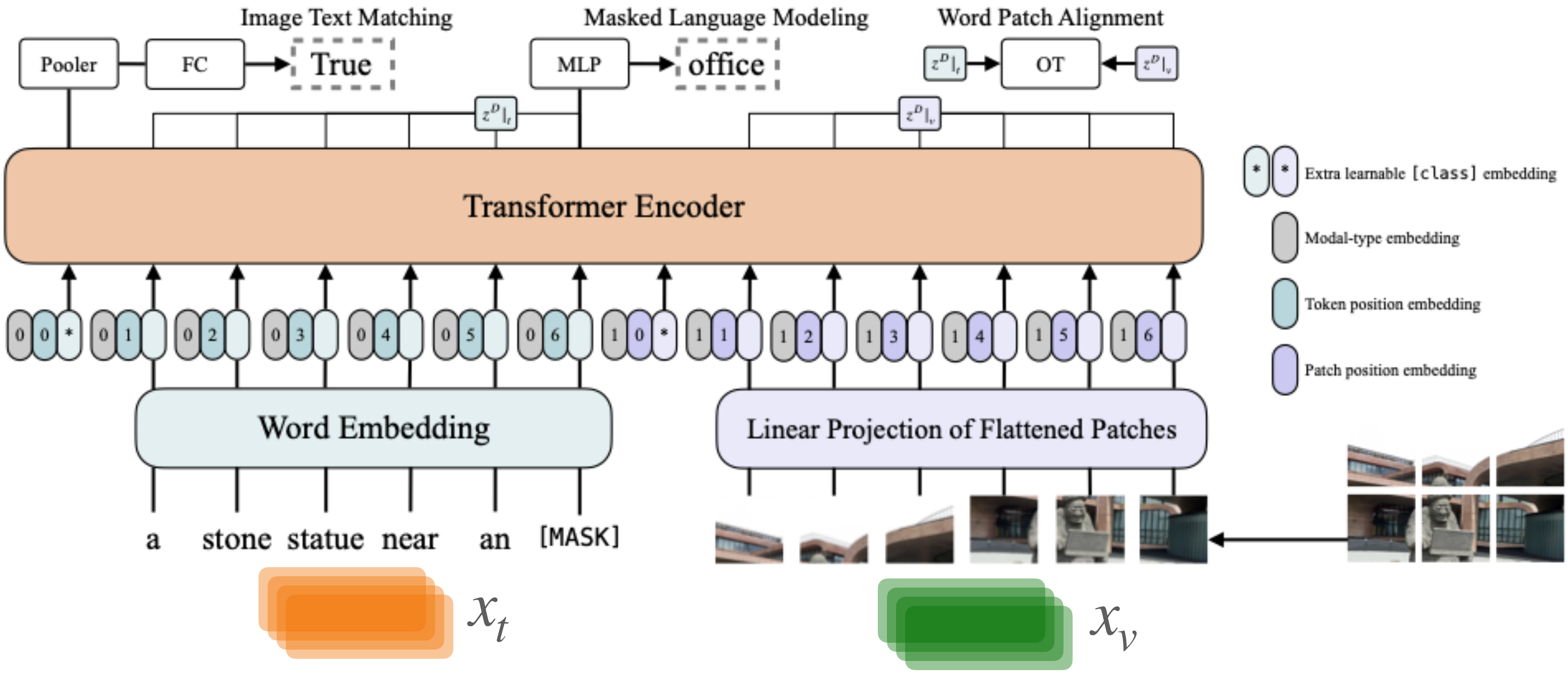


Vision and Language Fusion

- Is there a good model that can efficiently encode interactions among the sequence?
- **Hint:** What models have been covered in this class?



VILT: The Vision-Language Transformer



VILT: The Vision-Language Transformer

$$\mathscr{L} = \mathscr{L}_{ITM} + \mathscr{L}_{MLM} + \mathscr{L}_{WPA}$$

ITM

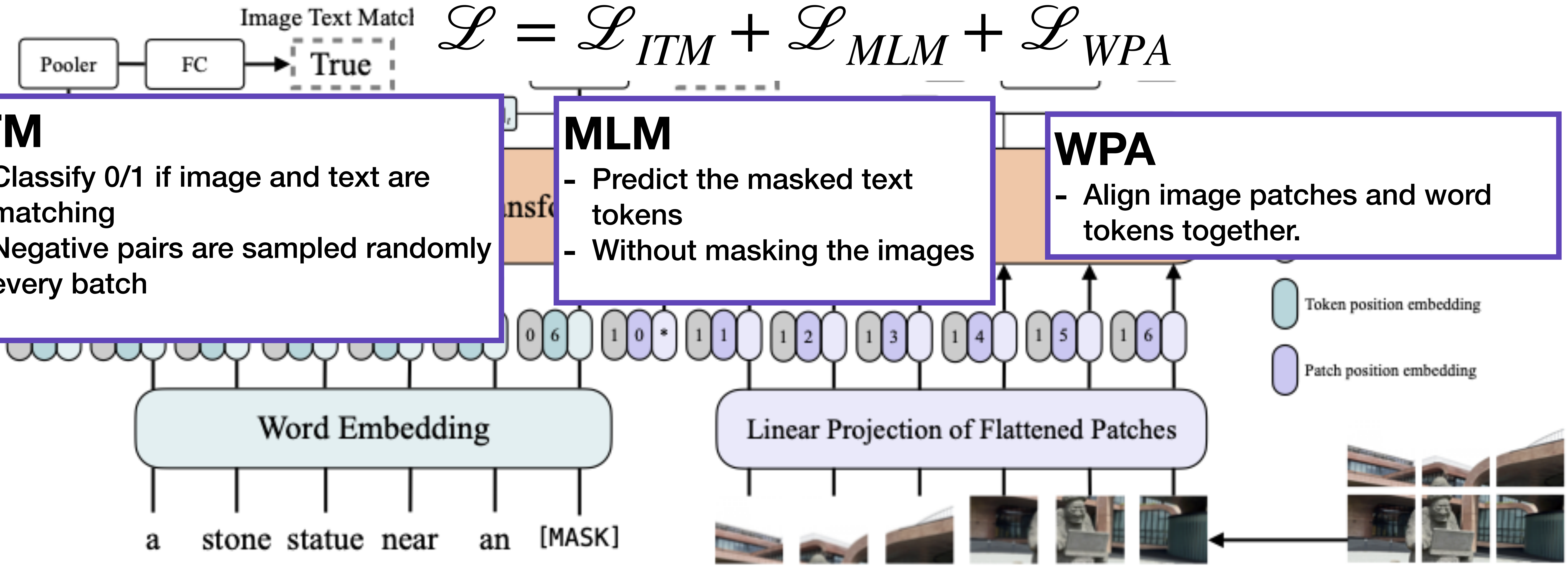
- Classify 0/1 if image and text are matching
- Negative pairs are sampled randomly every batch

MLM

- Predict the masked text tokens
- Without masking the images

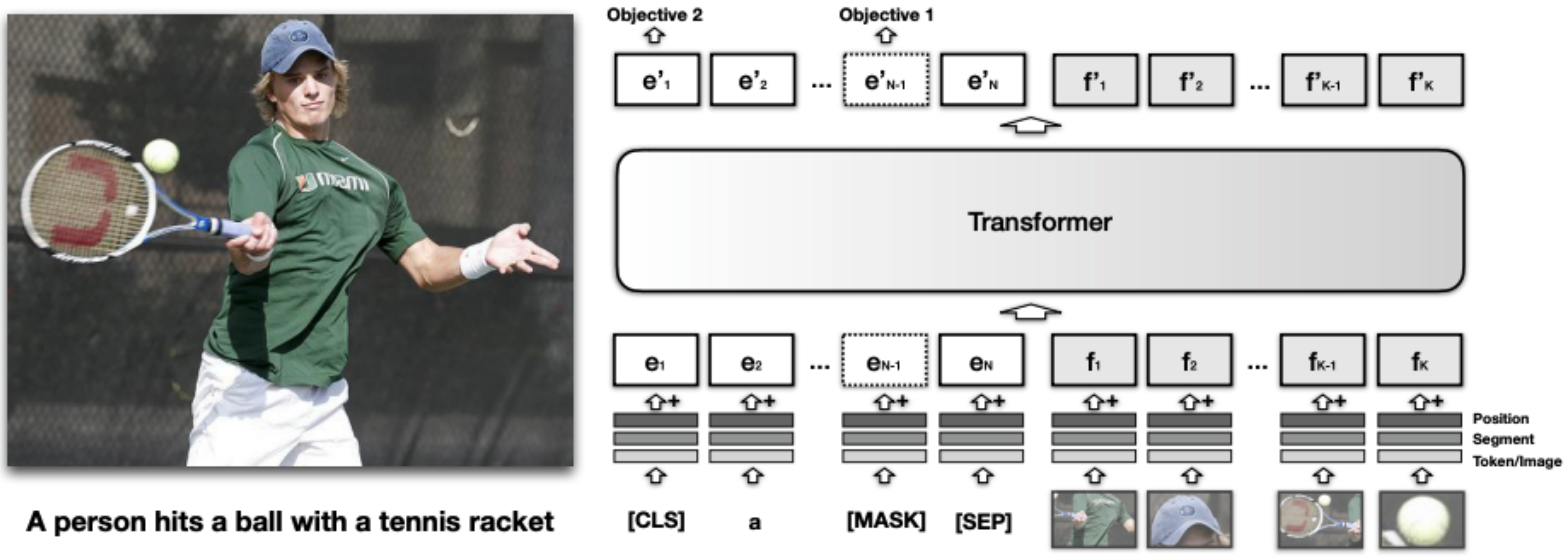
WPA

- Align image patches and word tokens together.



The “Vision-Language“ BERTs

- **Before Vision Transformer**: Tokenize images with **detected objects** and **region-wise ConvNet features** instead of raw image patches.
- Intuition: We understand images based on interaction among objects, so let’s directly encode this inductive bias to the model.



Models:
LXMERT
ViLBERT,
VLBERT,
UNITER
OSCAR

Potential Pre-Training Objectives

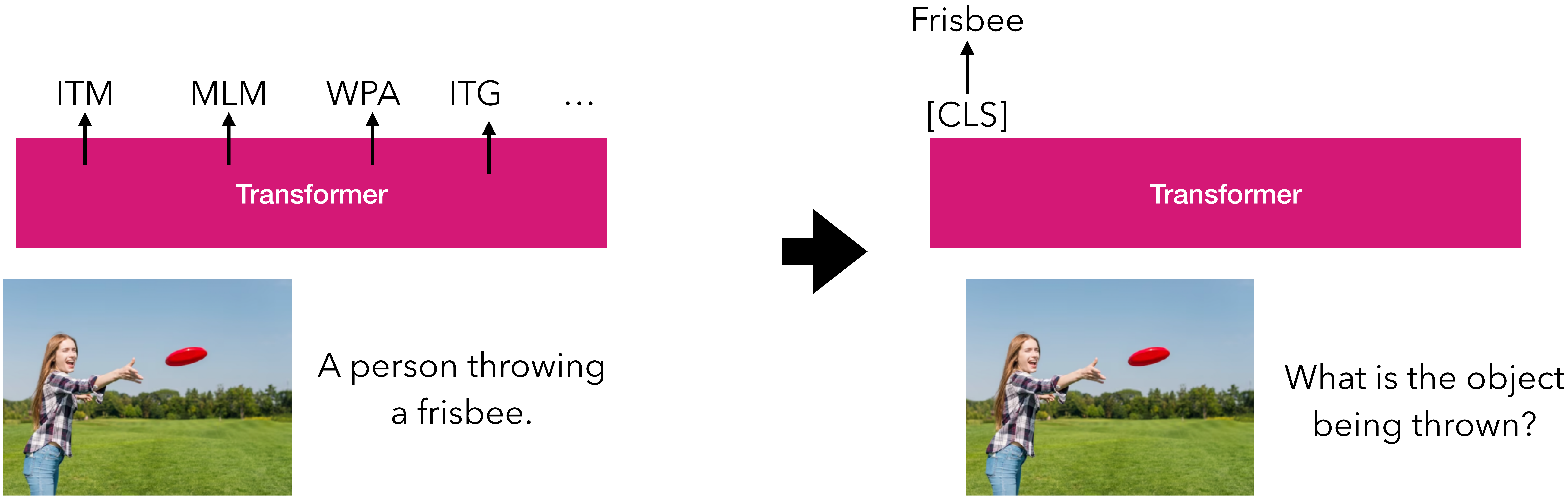
- **Masked Language Modeling** (MLM): Predict labels of masked text tokens.
- **Image-Text Matching** (ITM): Classify if image-text pairs are aligned
- **Word Region/Patch Alignment** (WPA): Align image regions/patches with text tokens

Potential Pre-Training Objectives

- **Masked Language Modeling** (MLM): Predict labels of masked text tokens.
- **Image-Text Matching** (ITM): Classify if image-text pairs are aligned
- **Word Region/Patch Alignment** (WPA): Align image regions/patches with text tokens
- **Image to Text Generation** (ITG): Generate the next text tokens.
- **Masked Image Modeling** (MIM): Predict/Regress masked image patches
- **Region Prediction**: Predict object labels of provided regions.
- Many more….

Pre-Training to Downstream VL Tasks

- Similar to BERT, finetune [CLS] token for classification tasks.

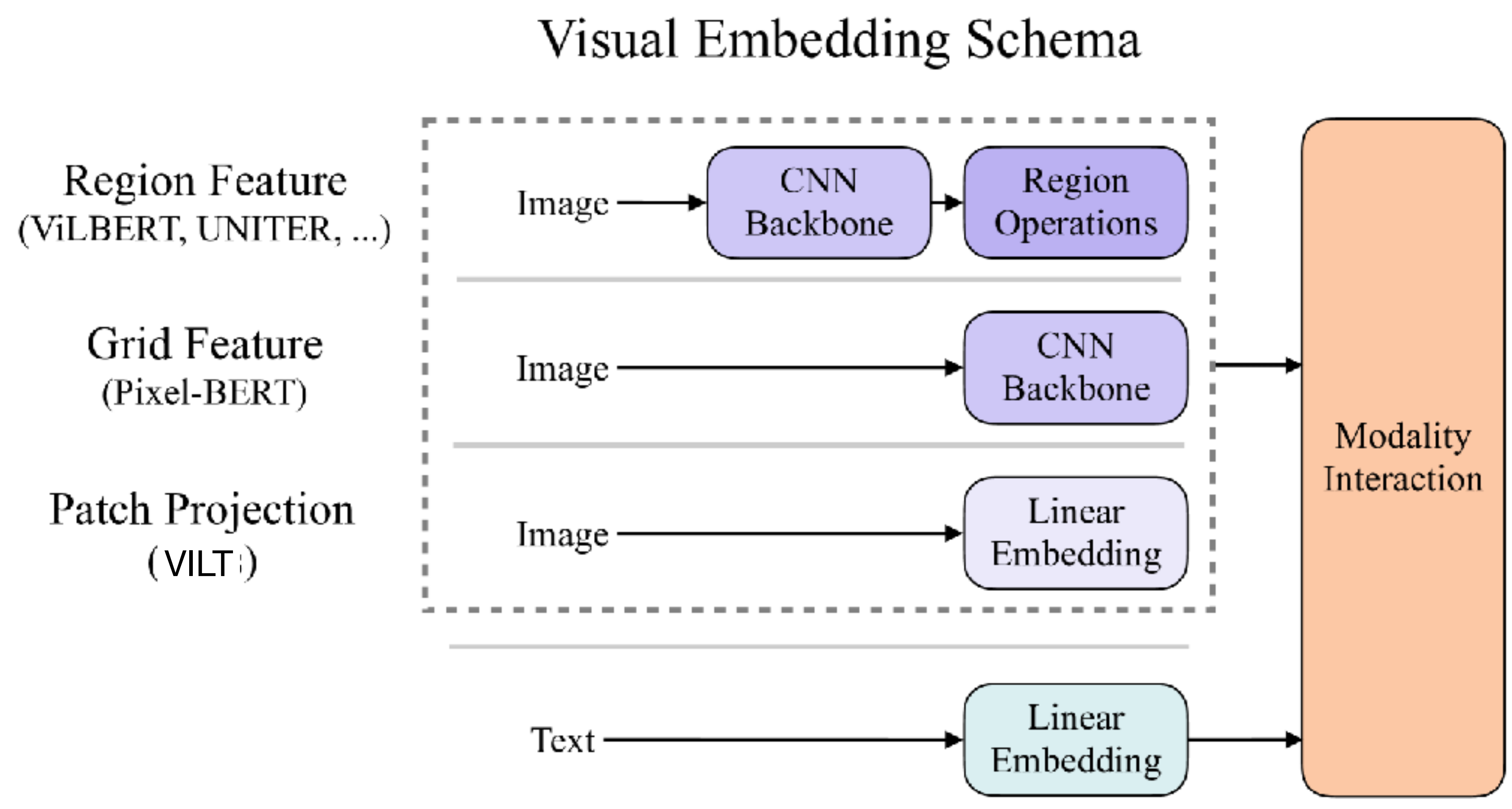


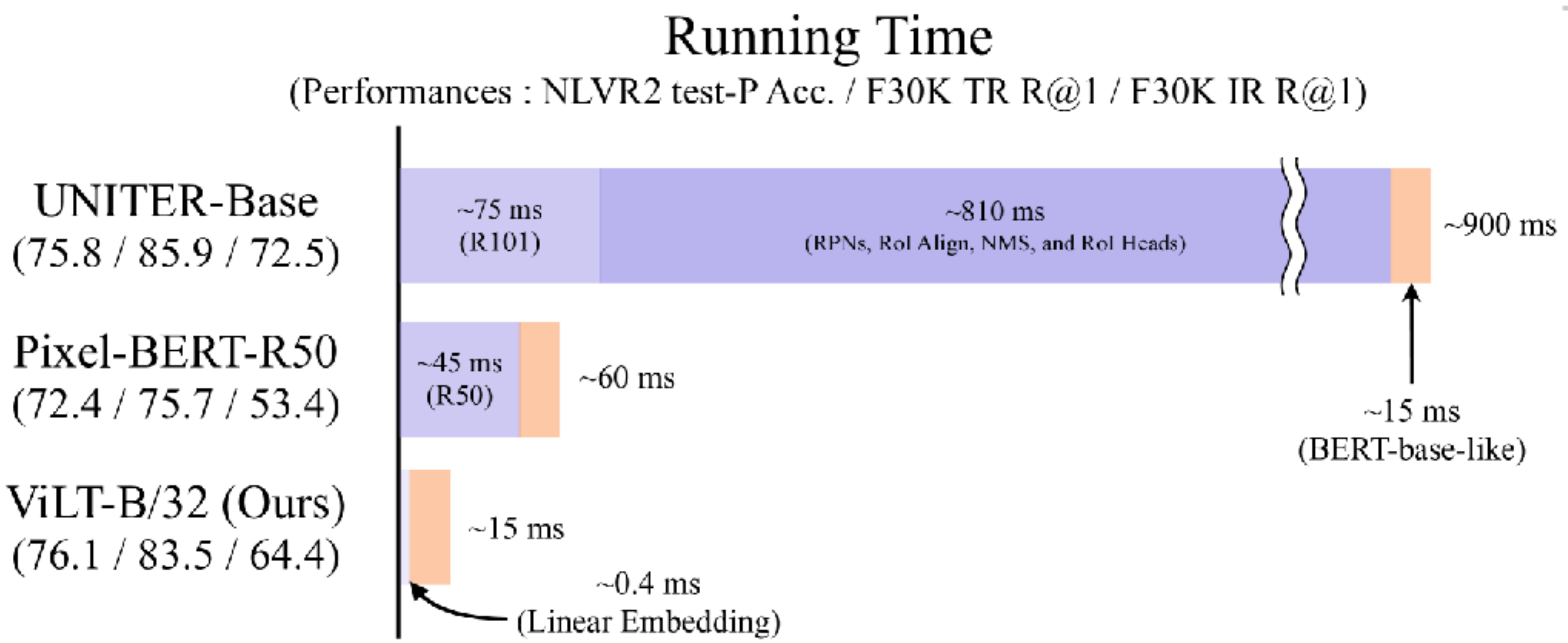
Pre-Training to Learn Image-Text Alignment

Finetune to Downstream Tasks (e.g. VQA)

Which model is better?

- **Region** based visual models
  - Slightly higher performance than patch based
- **Patch** based visual models: More efficient training and running time without losing much accuracy
  - Not reliant on object detections
  - Easily scalable

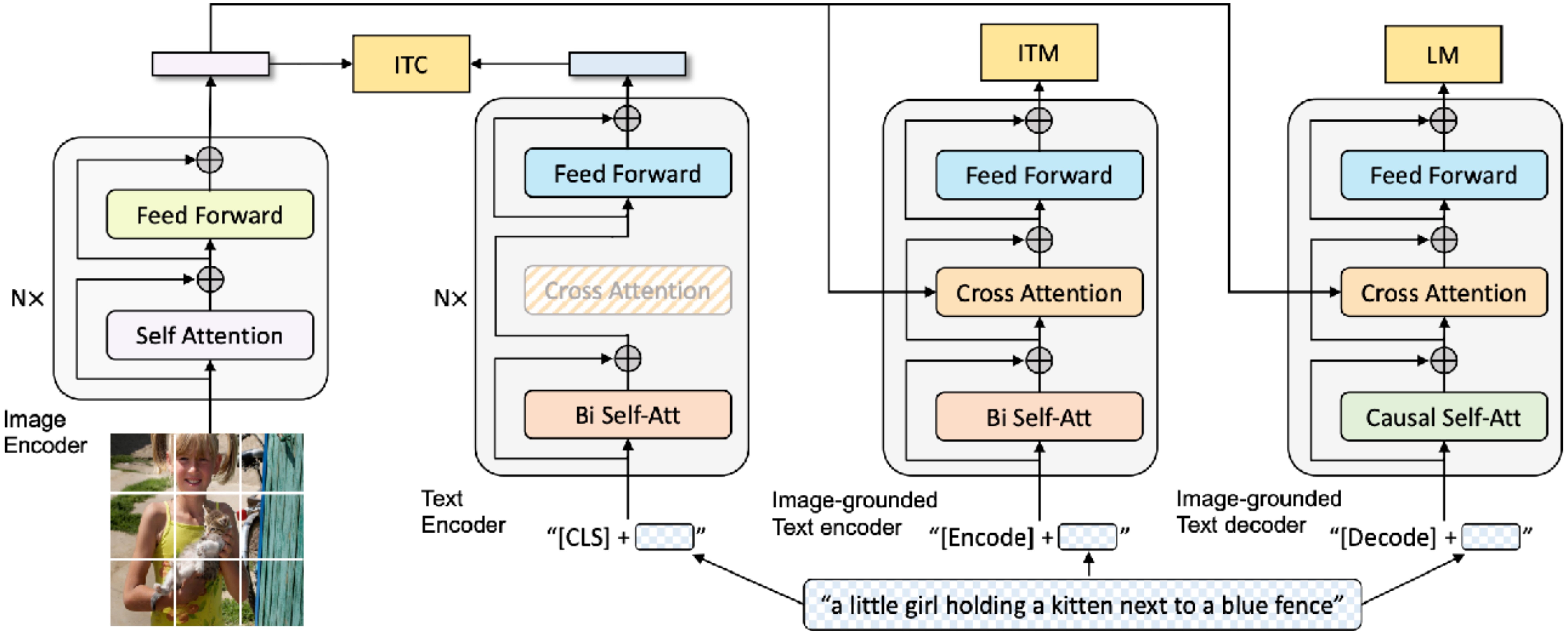




Effective Pre-Training Losses?

- **ALBEF/BLIP**
  - Masked Language Modeling (MLM)
  - Image-Text Matching (ITM)
  - Image-Text Contrastive Learning (ITC)

$$\mathcal{L} = \mathcal{L}_{\mathrm{itc}} + \mathcal{L}_{\mathrm{mlm}} + \mathcal{L}_{\mathrm{itm}}$$



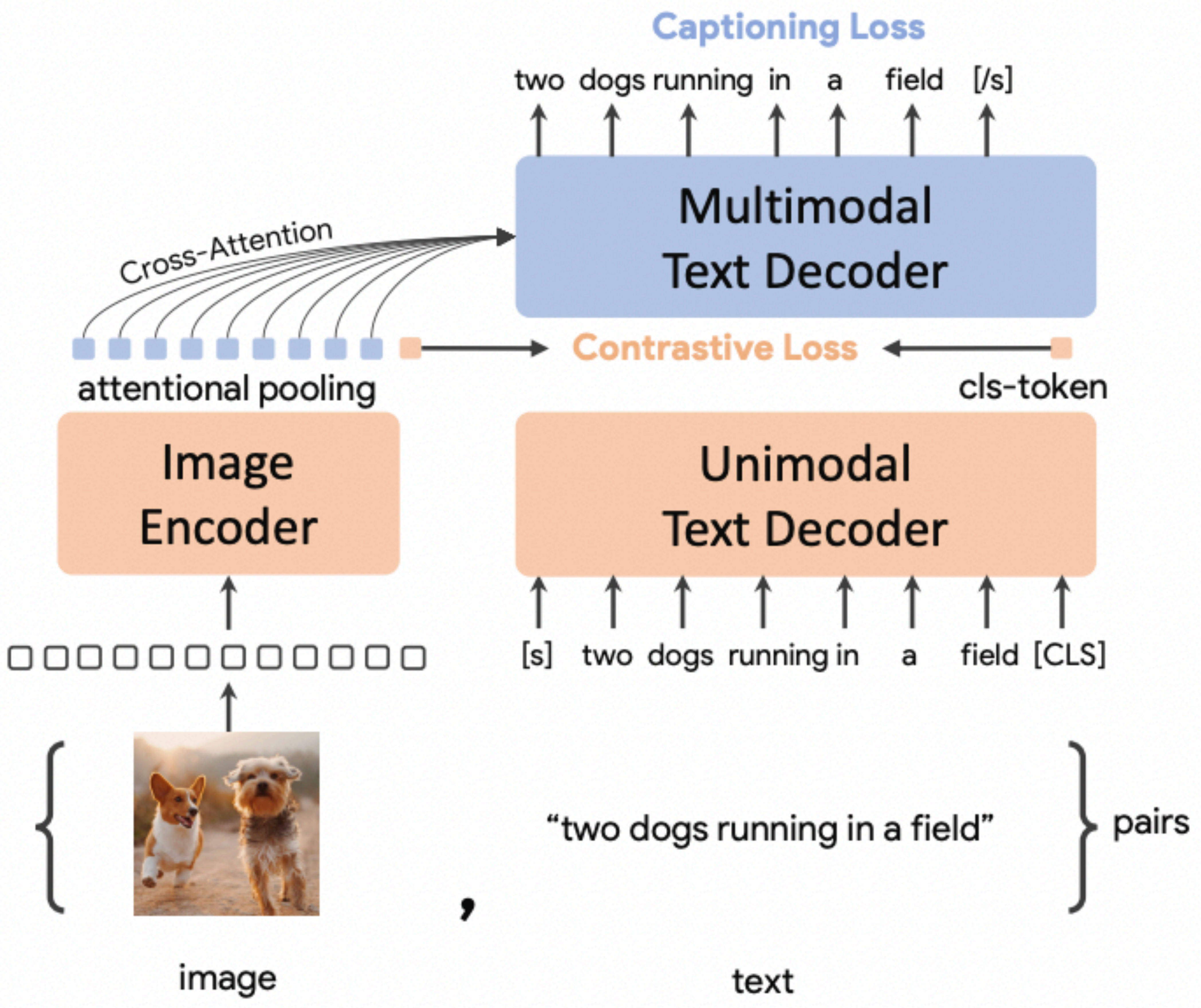
ALBEF: Downstream Task Results

Method	# Pre-train Images	Flickr30K (1K test set)					
		TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10
UNITER [2]	4M	83.6	95.7	97.7	68.7	89.2	93.9
CLIP [6]	400M	88.0	98.7	99.4	68.7	90.6	95.2
ALIGN [7]	1.2B	88.6	98.7	99.7	75.7	93.8	96.8
ALBEF	4M	90.5	98.8	99.7	76.8	93.7	96.7
ALBEF	14M	**94.1**	**99.5**	**99.7**	**82.8**	**96.3**	**98.1**

Table 3: Zero-shot image-text retrieval results on Flickr30K.

Method	VQA		$NLVR^2$		SNLI-VE	
	test-dev	test-std	dev	test-P	val	test
VisualBERT [13]	70.80	71.00	67.40	67.00	-	-
VL-BERT [10]	71.16	-	-	-	-	-
LXMERT [1]	72.42	72.54	74.90	74.50	-	-
12-in-1 [12]	73.15	-	-	78.87	-	76.95
UNITER [2]	72.70	72.91	77.18	77.85	78.59	78.28
VL-BART/T5 [54]	-	71.3	-	73.6	-	-
ViLT [21]	70.94	-	75.24	76.21	-	-
OSCAR [3]	73.16	73.44	78.07	78.36	-	-
VILLA [8]	73.59	73.67	78.39	79.30	79.47	79.03
ALBEF (4M)	74.54	74.70	80.24	80.50	80.14	80.30
ALBEF (14M)	**75.84**	**76.04**	**82.55**	**83.14**	**80.80**	**80.91**

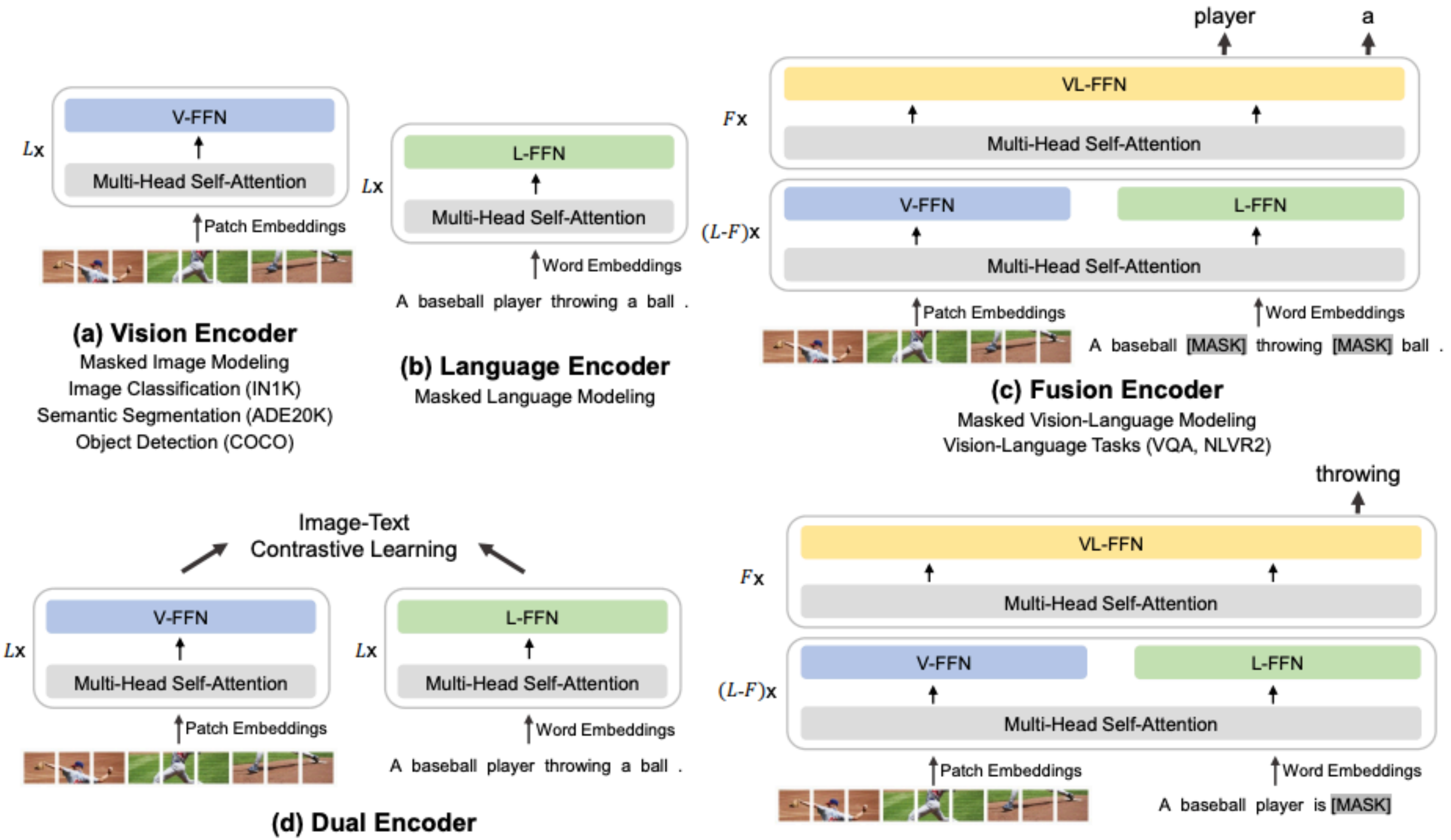
CoCA: Contrastive Captioning



Model	VQA		SNLI-VE		NLVR2	
	test-dev	test-std	dev	test	dev	test-p
UNITER [26]	73.8	74.0	79.4	79.4	79.1	80.0
VinVL [27]	76.6	76.6	-	-	82.7	84.0
CLIP-ViL [73]	76.5	76.7	80.6	80.2	-	-
ALBEF [36]	75.8	76.0	80.8	80.9	82.6	83.1
BLIP [37]	78.3	78.3	-	-	82.2	82.2
OFA [17]	79.9	80.0	$90.3^{\dagger}$	$90.2^{\dagger}$	-	-
VLMo [30]	79.9	80.0	-	-	85.6	86.9
SimVLM [16]	80.0	80.3	86.2	86.3	84.5	85.2
Florence [14]	80.2	80.4	-	-	-	-
METER [74]	80.3	80.5	-	-	-	-
CoCa	**82.3**	**82.3**	**87.0**	**87.1**	**86.1**	**87.0**

$$\mathcal{L}_{\text{CoCa}} = \lambda_{\text{Con}} \cdot \mathcal{L}_{\text{Con}} + \lambda_{\text{Cap}} \cdot \mathcal{L}_{\text{Cap}},$$

BEIT3: VL Masked Modeling Objectives



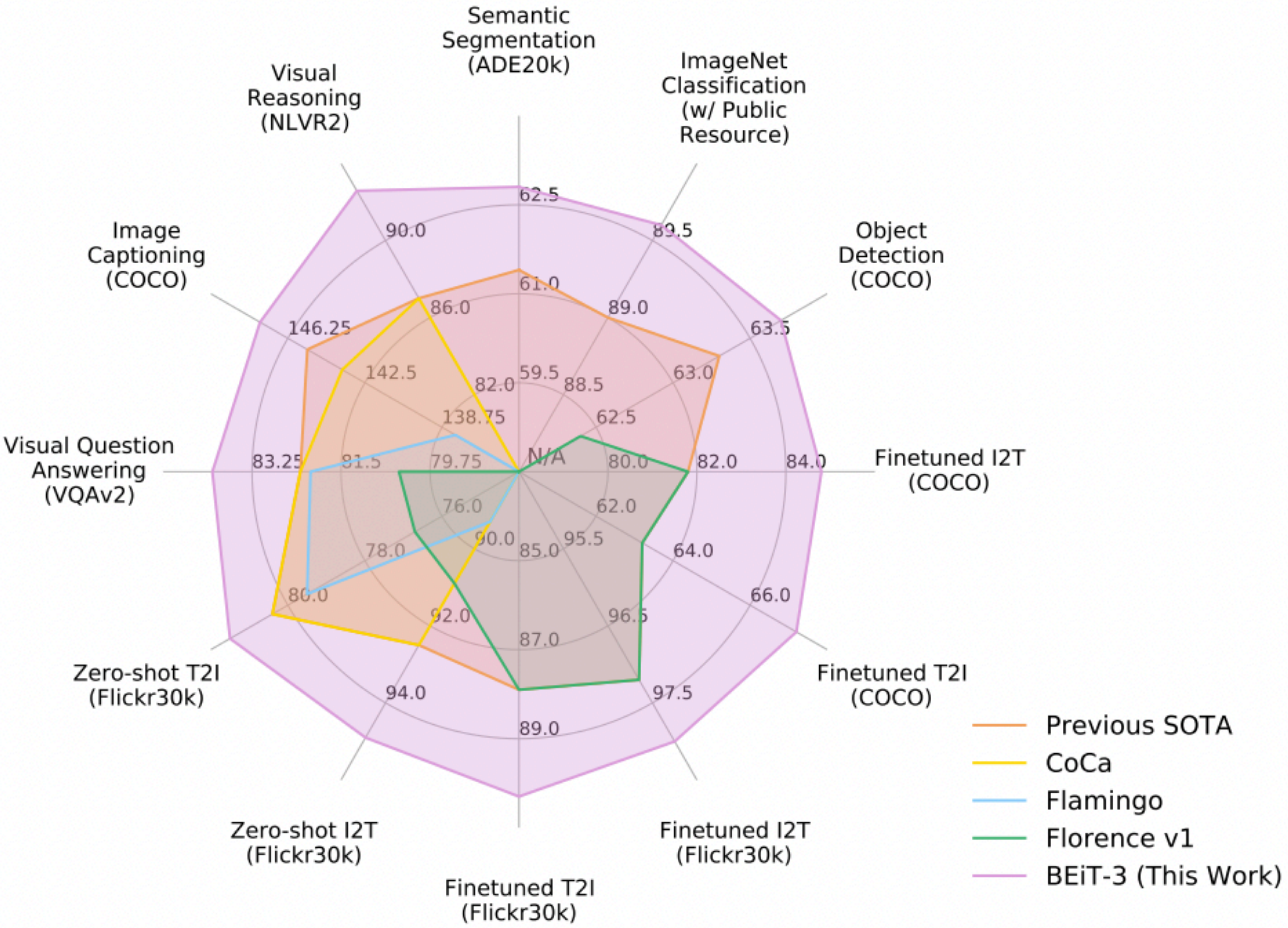
(a) Vision Encoder
Masked Image Modeling
Image Classification (IN1K)
Semantic Segmentation (ADE20K)
Object Detection (COCO)

(b) Language Encoder
Masked Language Modeling

(c) Fusion Encoder
Masked Vision-Language Modeling
Vision-Language Tasks (VQA, NLVR2)

(d) Dual Encoder
Image-Text Retrieval (Flickr30k, COCO)

(e) Image-to-Text Generation
Image Captioning (COCO)



Semantic Segmentation (ADE20k)
ImageNet Classification (w/ Public Resource)
Object Detection (COCO)
Finetuned I2T (COCO)
Finetuned T2I (COCO)
Finetuned I2T (Flickr30k)
Finetuned T2I (Flickr30k)
Zero-shot I2T (Flickr30k)
Zero-shot T2I (Flickr30k)
Visual Question Answering (VQAv2)
Image Captioning (COCO)
Visual Reasoning (NLVR2)
N/A
62.5
61.0
59.5
89.5
89.0
88.5
63.5
63.0
62.5
84.0
82.0
80.0
66.0
64.0
62.0
97.5
96.5
95.5
89.0
87.0
85.0
94.0
92.0
90.0
80.0
78.0
76.0
83.25
81.5
79.75
146.25
142.5
138.75
90.0
86.0
82.0
Previous SOTA
CoCa
Flamingo
Florence v1
BEiT-3 (This Work)

Trends of VL Models

- Race of Scaling Model Size / Dataset / # of Tasks?

Model	Model Size				PT dataset size	PT Tasks
	Image Enc.	Text Enc.†	Fusion†	Total		
CLIP ViT-L/14 (Radford et al., 2021)	302M	123M	0	425M	400M	ITC
ALIGN (Jia et al., 2021)	480M	340M	0	820M	1.8B	ITC
Florence (Yuan et al., 2021)	637M	256M	0	893M	900M	ITC
SimVLM-huge (Wang et al., 2022k)	300M	39M	600M	939M	1.8B	PrefixLM
METER-huge (Dou et al., 2022b)	637M	125M	220M	982M	900M+20M[1]	MLM+ITM
LEMON (Hu et al., 2022)	147M[2]	39M	636M	822M	200M	MLM
Flamingo (Alayrac et al., 2022)	200M	70B	10B	80.2B	2.1B+27M[3]	LM
GIT (Wang et al., 2022d)	637M	40M	70M	747M	800M	LM
GIT2 (Wang et al., 2022d)	4.8B	40M	260M	5.1B	12.9B	LM
CoCa (Yu et al., 2022a)	1B	477M	623M	2.1B	1.8B+3B[4]	ITC+LM
BEiT-3 (Wang et al., 2022g)	692M[5]	692M[5]	52M[5]	1.9B	21M+14M[6]	MIM+MLM +MVLM
PaLI (Chen et al., 2022e)	3.9B	40M	13B	16.9B	1.6B	LM+VQA[7] +OCR+OD

Efficient Training of VLMs?

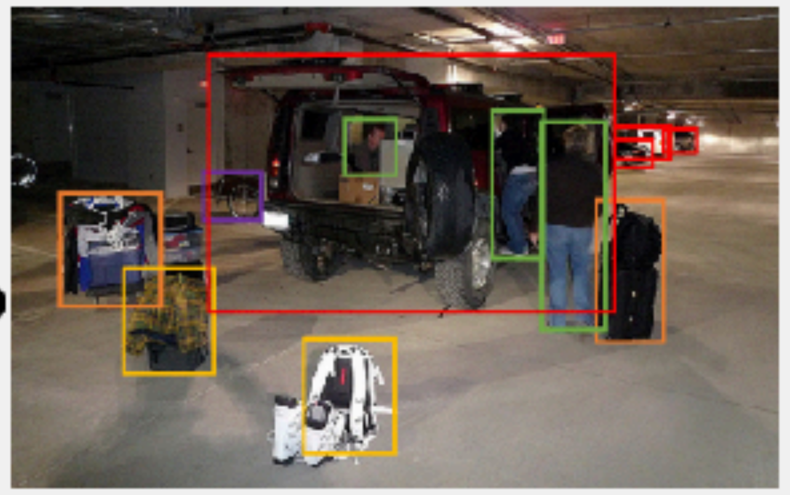
- **Potential Room for Improvement?**: Get even larger amount of data, get $$$ GPUs, with more objectives, and train for long time to outperform the existing models.
- **Need for efficient training:**
    - Can we exploit the already pre trained representations?
        - Do we need to train the whole model from scratch?
        - Do we always find success with a large amount of data?

Data for Visual Instruction-tuning

- Symbolic representations of images from GPT4
  - Captions
  - Bounding boxes
- GPT-assisted self-instruct tuning example 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>

Data for Visual Instruction-tuning

- 3 types of instruction-following questions and responses *(text-only with symbolic repr.)*
  - Conversation (sequential QA pairs)
  - Detailed Description
  - Complex Reasoning (very important)

```
messages = [ {"role":"system", "content": f"""You are an AI visual assistant, and you are seeing a single image. What you see are provided with five sentences, describing the same image you are looking at. Answer all questions as you are seeing the image.

Design a conversation between you and a person asking about this photo. The answers should be in a tone that a visual AI assistant is seeing the image and answering the question. Ask diverse questions and give corresponding answers.

Include questions asking about the visual content of the image, including the object types, counting the objects, object actions, object locations, relative positions between objects, etc. Only include questions that have definite answers:
(1) one can see the content in the image that the question asks about and can answer confidently;
(2) one can determine confidently from the image that it is not in the image. Do not ask any question that cannot be answered confidently.

Also include complex questions that are relevant to the content in the image, for example, asking about background knowledge of the objects in the image, asking to discuss about events happening in the image, etc. Again, do not ask about uncertain details. Provide detailed answers when answering complex questions. For example, give detailed examples or reasoning steps to make the content more convincing and well-organized. You can include multiple paragraphs if necessary."""}
]
for sample in fewshot_samples:
    messages.append({"role":"user", "content":sample['context']})
    messages.append({"role":"assistant", "content":sample['response']} )
messages.append({"role":"user", "content":'\n'.join(query)})
```

Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV).
Question: Where is the vehicle parked?
Answer: The vehicle is parked in an underground parking area, likely in a public garage.
Question: What are the people in the image doing?
Answer: The people are trying to fit all of their luggage into the SUV, likely preparing for a trip.

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.
In the parking area, various luggage items are scattered around the vehicle. There are two backpacks, one located near the left rear wheel and the other closer to the right side of the vehicle. Additionally, there are two suitcases, one on the right side of the car and another further away near the center of the parking area. A bicycle can also be seen on the left side of the vehicle.
Other cars are parked around the main SUV, with one car positioned behind it and slightly to the left, another behind and slightly to the right, and the third car further behind on the right side.

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 to accommodate. They might have to strategize and arrange the luggage efficiently to ensure that everything fits properly into the vehicle. Additionally, they need to consider the comfort of the passengers and visibility while driving, so the placement of the luggage must not obstruct the driver's view or make the passengers uncomfortable during the trip.

Data for Visual Instruction-tuning

- 3 types of instruction-following questions and responses *(text-only with symbolic repr.)*
  - Conversation (sequential QA pairs)
  - Detailed Description
  - Complex Reasoning (very important)

- Use COCO images and captions
- GPT-4 Language only model to prompt
- Few-shot prompting with manual examples

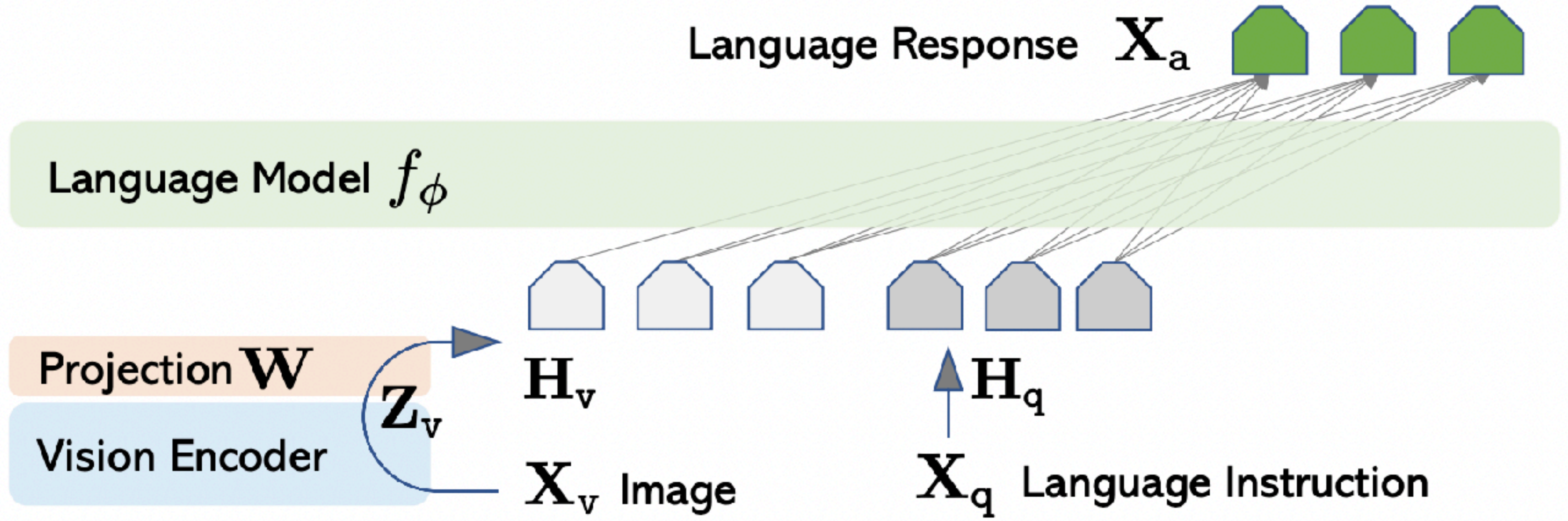
- 158k instruction following samples
  - 58k conversations
  - 23k detailed descriptions
  - 77k complex reasoning

LLaVA: Large Lang and Vis Assistant

$$p(\mathbf{X}_a|\mathbf{X}_v, \mathbf{X}_{\text{instruct}}) = \prod_{i=1}^{L} p_{\theta}(\boldsymbol{x}_i|\mathbf{X}_v, \mathbf{X}_{\text{instruct},<i}, \mathbf{X}_{a,<i}),$$

LLAMA

CLIP VE
VIT-L/14



$\mathbf{X}_{\text{system-message}}$ <STOP>
Human : $\mathbf{X}^1_{\text{instruct}}$ <STOP> Assistant: $\mathbf{X}^1_a$ <STOP>
Human : $\mathbf{X}^2_{\text{instruct}}$ <STOP> Assistant: $\mathbf{X}^2_a$ <STOP> $\cdots$

$$\mathbf{X}^t_{\text{instruct}} = \begin{cases} \text{Randomly choose } [\mathbf{X}^1_q, \mathbf{X}_v] \text{ or } [\mathbf{X}_v, \mathbf{X}^1_q], & \text{the first turn } t = 1 \\ \mathbf{X}^t_q, & \text{the remaining turns } t > 1 \end{cases}$$

Dual stage training

- Stage 1: **Pre-training for feature alignment**
  - Only projection matrix is updated
  - Trained on a subset of CC3M (595k IT pairs)
- Stage 2: **Fine-tuning for user and task orientation**
  - Both projection matrix and LLM are updated
  - Tuned on Visual chat (user chat-like orientation 158k) & Science QA (complex science reasoning)

Evaluations

Evaluating Object Hallucination in Large Vision-Language Models

Yifan Li[1,3*], Yifan Du[1,3*], Kun Zhou[2*], Jinpeng Wang [4], Wayne Xin Zhao[2,3†] and Ji-Rong Wen[1,2,3]

CHEF: A COMPREHENSIVE EVALUATION FRAMEWORK FOR STANDARDIZED ASSESSMENT OF MULTI-MODAL LARGE LANGUAGE MODELS

Zhelun Shi[*,1,2] Zhipin Wang[*,1] Hongxing Fan[*,1] Zhenfei Yin[2]
Lu Sheng[†,1] Yu Qiao[2] Jing Shao[†,2]
[1]Beihang University [2]Shanghai AI Laboratory
[*] Equal Contribution [†] Corresponding Author

LAMM: Language-Assisted Multi-Modal Instruction-Tuning Dataset, Framework, and Benchmark

Zhenfei Yin[*,1,3] Jiong Wang[*,1,4] Jianjian Cao[*,1,4] Zhelun Shi[*,1,2] Dingning Liu[1,5] Mukai Li[1]
Xiaoshui Huang[1] Zhiyong Wang[3] Lu Sheng[2] Lei Bai[†,1] Jing Shao[†,1] Wanli Ouyang[1]
[1]Shanghai Artificial Intelligence Laboratory [2]Beihang University [3]The University of Sydney
[4]Fudan University [5]Dalian University of Technology
[*] Equal Contribution [†] Corresponding Authors

ON THE HIDDEN MYSTERY OF OCR IN LARGE MULTIMODAL MODELS

Yuliang Liu[1], Zhang Li[1], Biao Yang[1], Chunyuan Li[2], Xu-Cheng Yin[3], Cheng-Lin Liu[4], Lianwen Jin[5], Xiang Bai[1*]

Applications to Domains/Tasks

Medical:

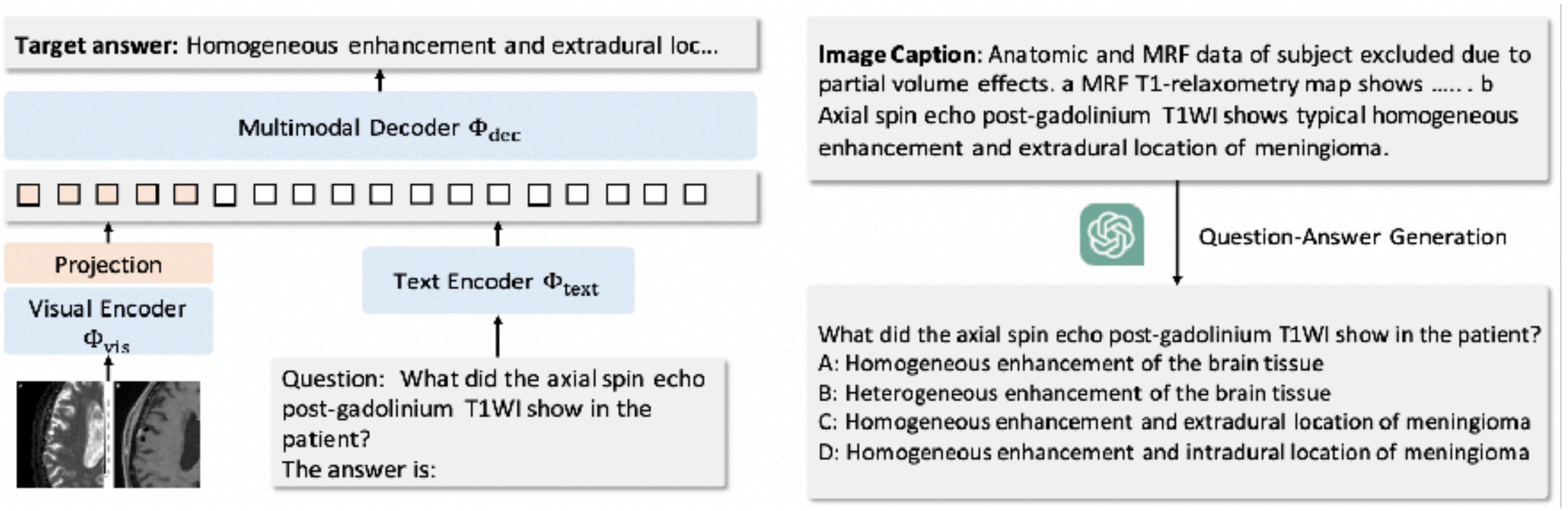
Med-LLaVA

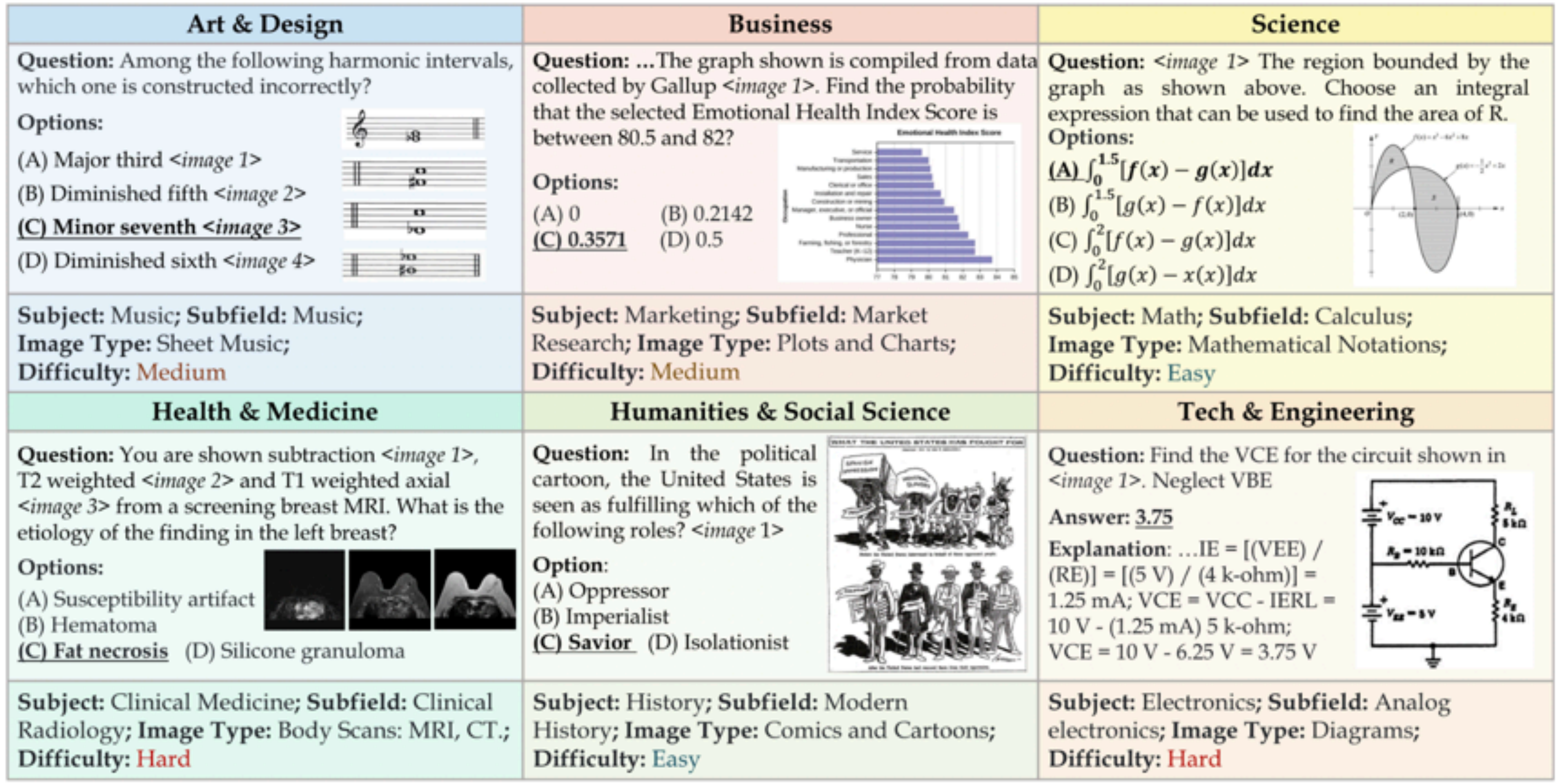
PMC-VQA

Domains - pathology, geometry, art and design

Image types - diagrams, tables, plots, chemical structures

Expert skill - Mathematical equations, science formula





mmmu-benchmark.github.io/

BIG Gap remains – Open Directions

- Encoding high resolution images
  - Including page images with complex layout
- Encoding long sequences
  - Video understanding
- Integrating domain knowledge
  - Structural symmetry
  - Data generating processes