Deep Learning for NLP

Module 6: Multimodality

Teaching: Dr. Khyathi Chandu

(Acknowledgements to some slides adapted from James Park)

MultiModal Systems

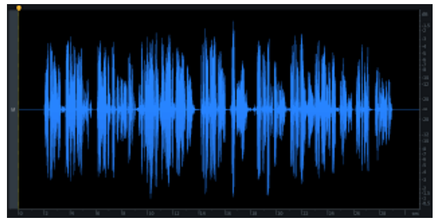
- Dictionary Definition: “characterized by several different modes of activity or occurrence”
- **Multimodal AI**: System that integrates various data types and sensory inputs (images, videos, audio, other sensory information) to create a unified representation or understanding.

A person throwing a frisbee.

Text

Image

Video



Audio

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

Examples of Multimodal Tasks

VQA & Visual Reasoning
Q: What is the dog holding with its paws?
A: Frisbee.

Image Captioning
Caption: A dog is lying on the grass next to a frisbee.

Text-to-Image Retrieval
Query: A dog is lying on the grass next to a frisbee.

Negative Images

Text-to-Video Retrieval
Query: A dog is lying on the grass next to a frisbee, *while shaking its tail*.

Negative Videos

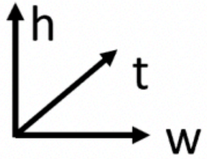


Image Classification
Labels: [dog, grass, frisbee]

Object Detection

dog, grass, frisbee

Segmentation

dog, grass, frisbee

Video Question Answering
Q: Is the dog perfectly still?
A: No.

Video Captioning
Caption: A dog is lying on the grass next to a frisbee, *while shaking its tail*.

Multimodal Language Models

How to train these 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.

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

f_v

x_v

A person throwing a frisbee.

Word2Vec, BERT, …

Text Encoder

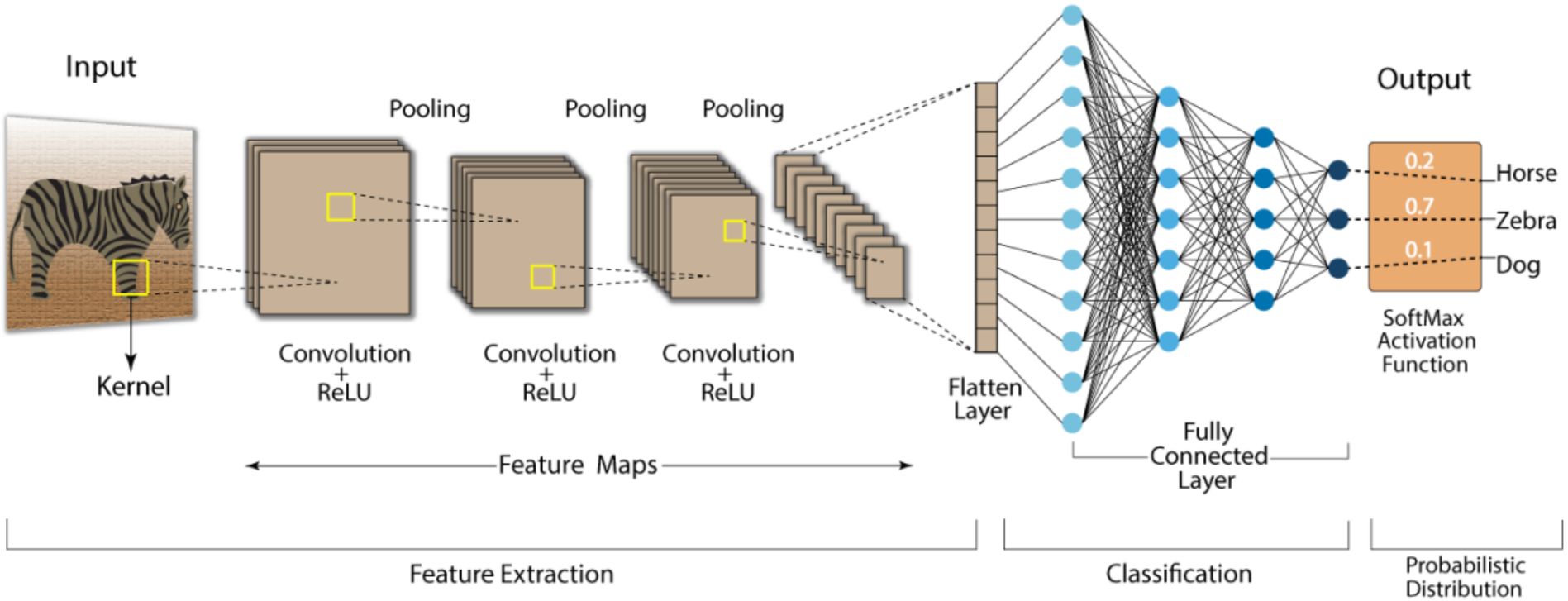
f_t

x_t

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

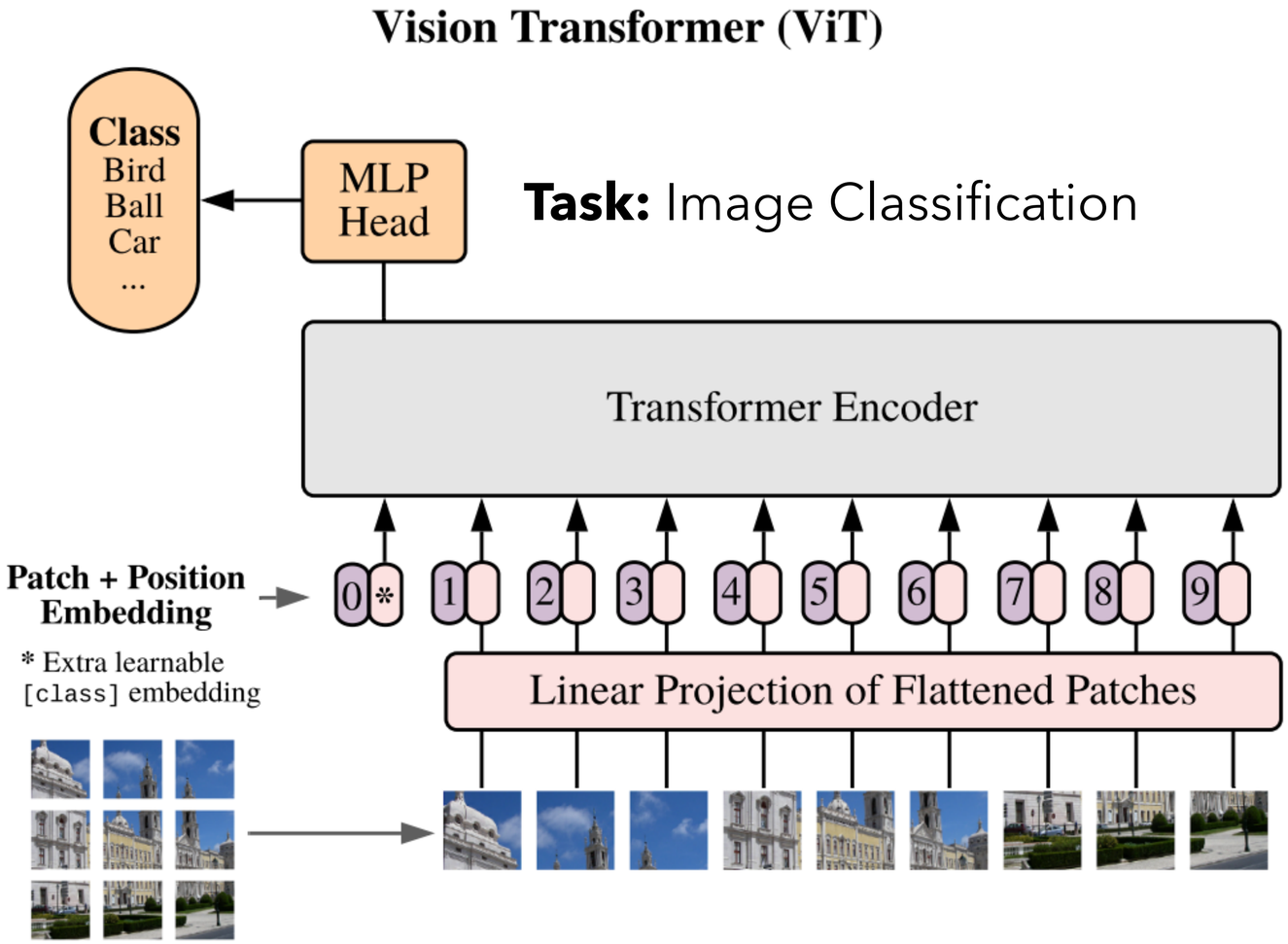
Vision Encoder: Convolutional Neural Networks

- **CNNs**: Extract features that encode spatial and temporal relationships in image with convolution operations.
  - **Pooling**: Reduce dimensionality of the convoluted features for efficient computation
- De facto model for Image Classification



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 encode s translation equivariance and locality.



Steps of Image-Text Alignment

CNNs: ResNet
Transformers: ViT,
...

Image Encoder

f_v

x_v

Fusion

Word2Vec, BERT, ...

A person throwing a frisbee.

Text Encoder

f_t

x_t

- **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.

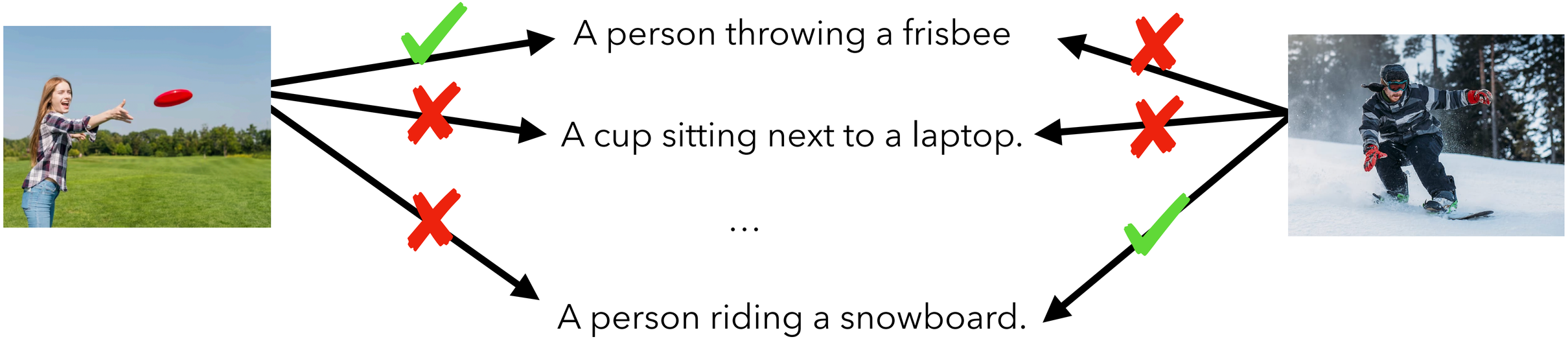
$x_t \in \mathbb{R}^t$

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.



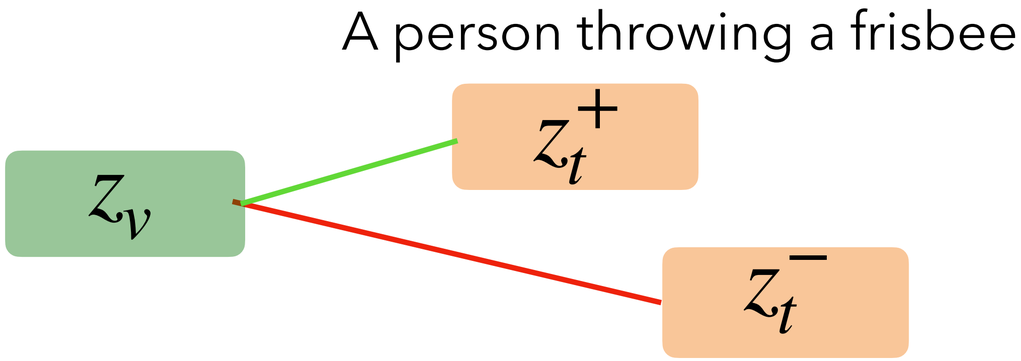
Contrastive Learning

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

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

Why preferred over Euclidean Distance?

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

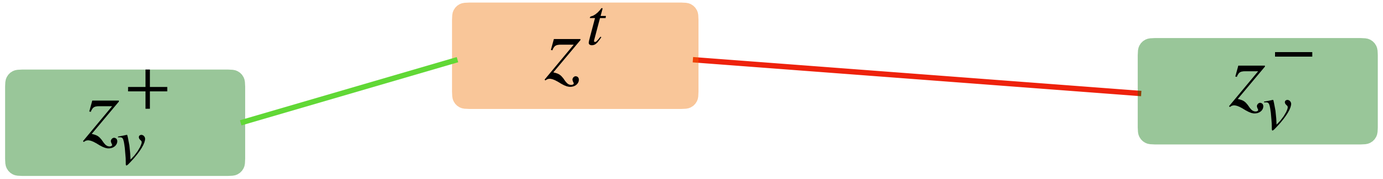


Contrastive Learning

- Adjust similarity of learned embeddings with a distance metric.
  - Euclidean Distance
  - Cosine Similarity $\cos(u, v) = \frac{u \cdot v}{||u||_2 ||v||_2}$

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

A person throwing a frisbee



Contrastive Learning

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

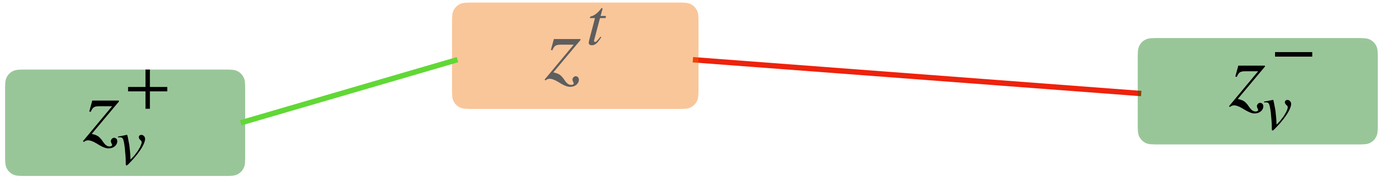
- Adjust similarity of learned embeddings with a distance metric.
  - Euclidean Distance
  - Cosine Similarity

Triplet Loss

$$\max(0, \text{sim}(z_v, z_t^+) - \text{sim}(z_v, z_t^-) + m) +$$
$$\max(0, \text{sim}(z_v^+, z_t) - \text{sim}(z_v^-, z_t) + m)$$

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

A person throwing a frisbee



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

Contrastive Learning

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

```
def compute_distance_matrix(anchor, positive, negative):
    """
    Compute distance matrix between anchor, positive, and negative samples.
    """
    distance_matrix = torch.zeros(anchor.size(0), 3)
    distance_matrix[:, 0] = euclidean_distance(anchor, anchor)
    distance_matrix[:, 1] = euclidean_distance(anchor, positive)
    distance_matrix[:, 2] = euclidean_distance(anchor, negative)
    return distance_matrix
```

ngs with a distance metric.

$$\max(0, \text{sim}(z_v, z_t^+) - \text{sim}(z_v, z_-^+) + m) +$$
$$\max(0, \text{sim}(z_v^+, z_t) - \text{sim}(z_v^-, z_t) + m)$$

Triplet Loss

```
def batch_all_triplet_loss(anchor, positive, negative, margin=0.2):
    """
    Compute triplet loss using the batch all strategy.
    """
    distance_matrix = compute_distance_matrix(anchor, positive, negative)
    loss = torch.max(torch.tensor(0.0), distance_matrix[:, 0] - distance_matrix[:, 1] + margin)
    loss += torch.max(torch.tensor(0.0), distance_matrix[:, 0] - distance_matrix[:, 2] +
margin)
    return torch.mean(loss)
```

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!

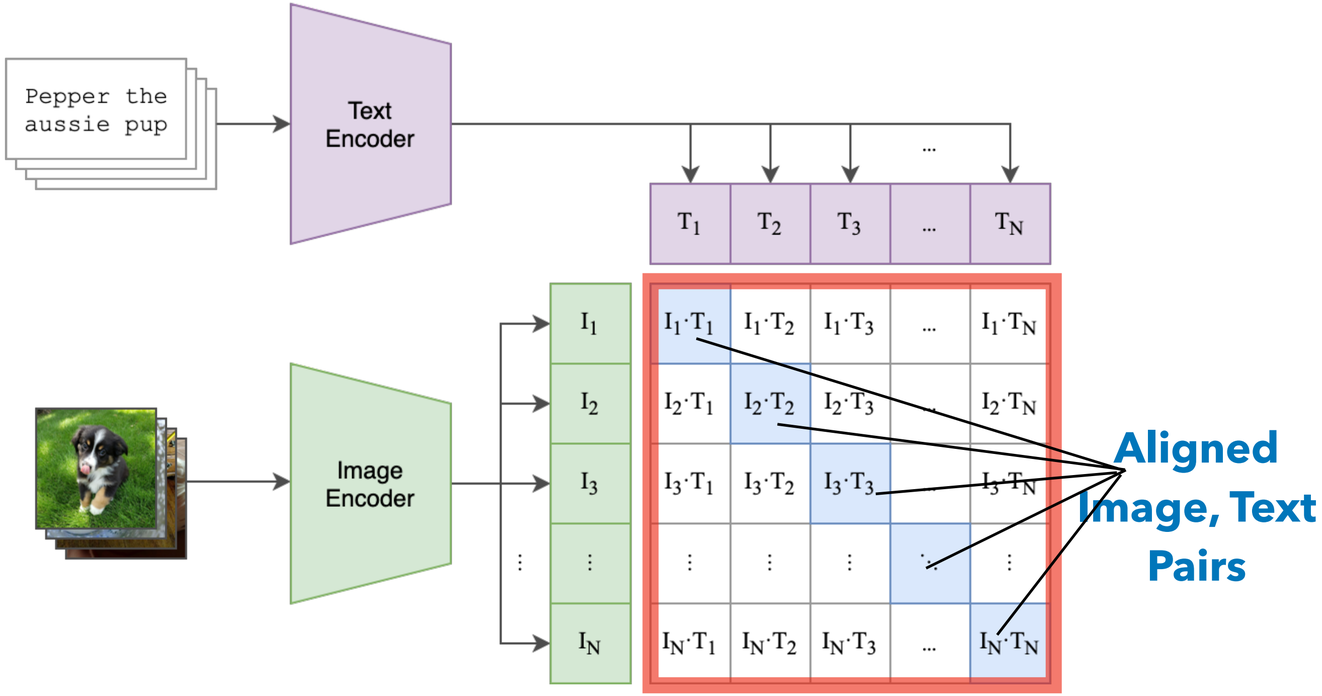
Positive — A person throwing a frisbee

Negative — A cup sitting next to a laptop.

…

Negative — A person riding a snowboard.

CLIP: Contrastive Language-Image Pre-Training

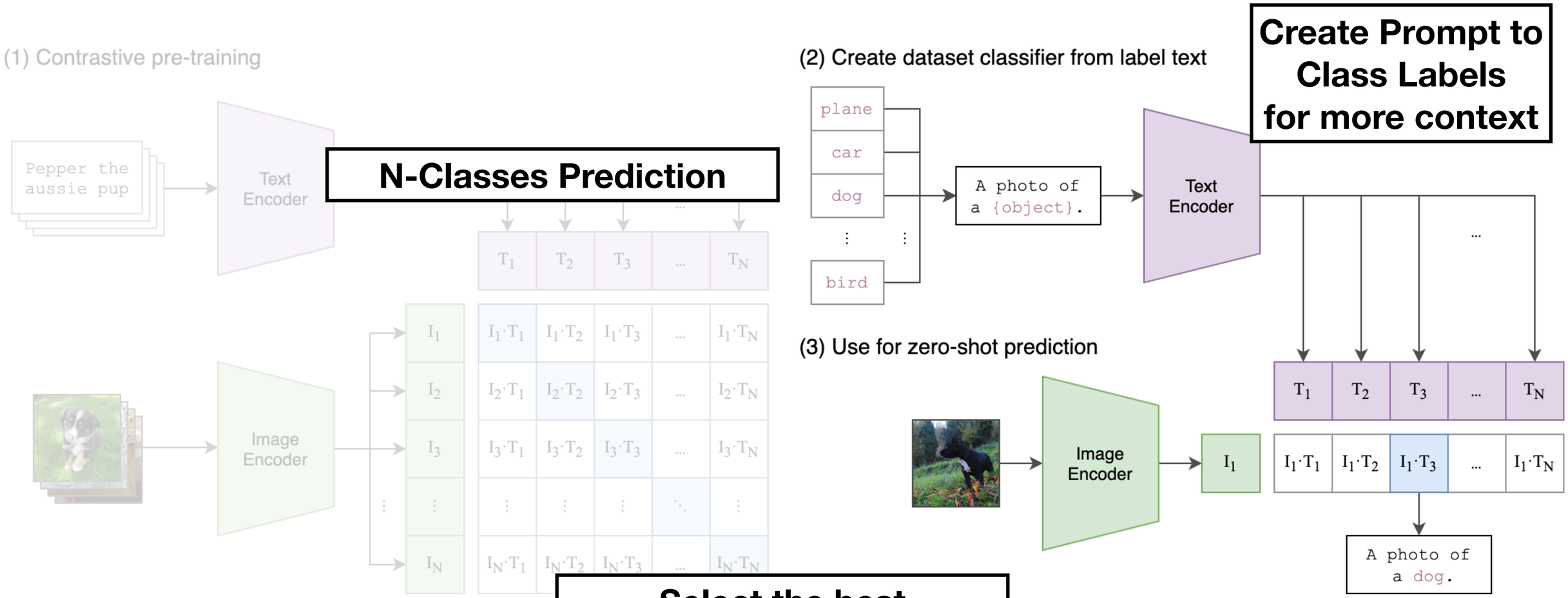


Objective: given a batch of N (image, text) pairs, predict which of the N × N possible (image, text) pairings across a batch actually occurred.

Minimize InfoNCE Loss

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

CLIP: Contrastive Language-Image Pre-Training



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

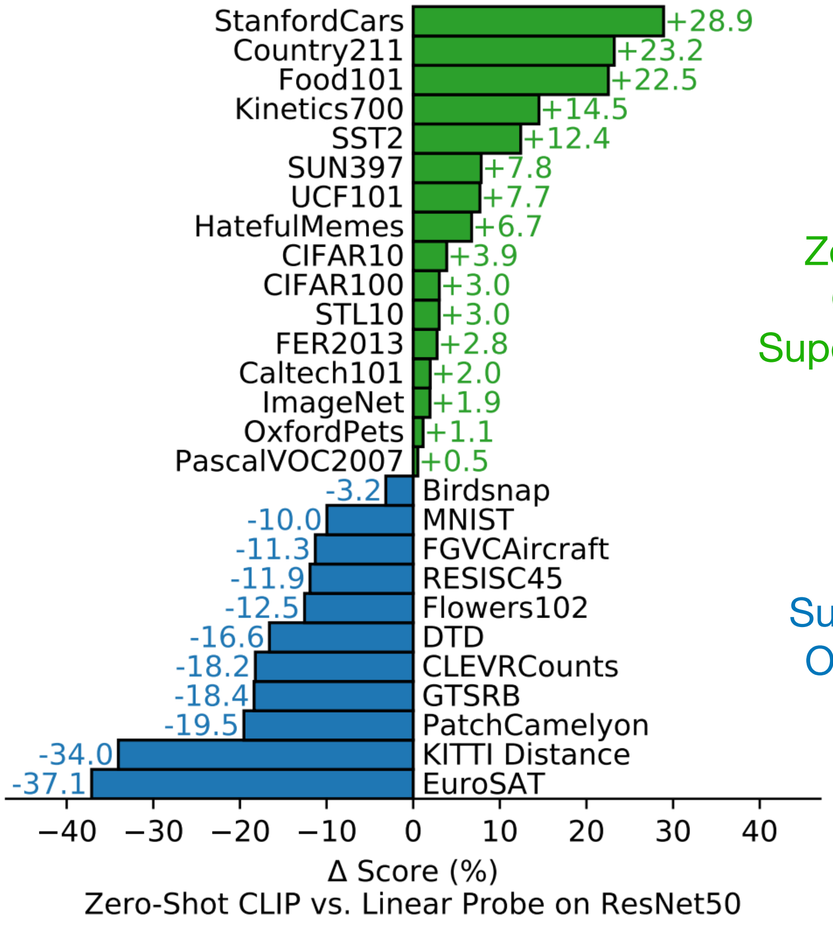
Table 3.2: Statistics of the pre-training datasets used in a typical academic setting.

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

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.

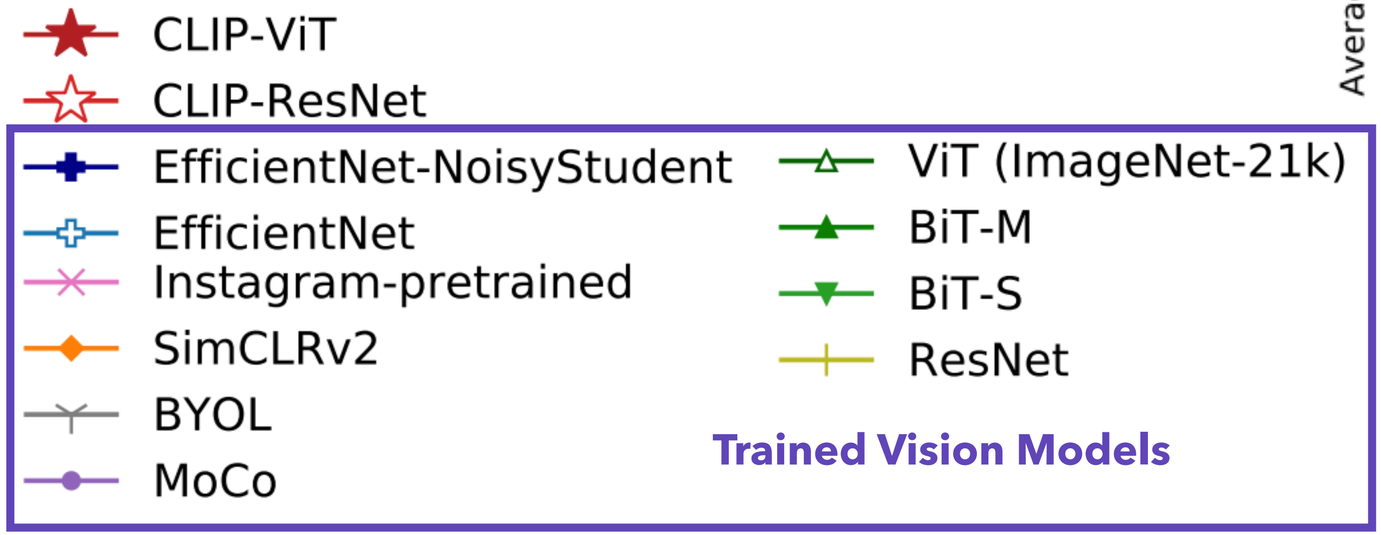


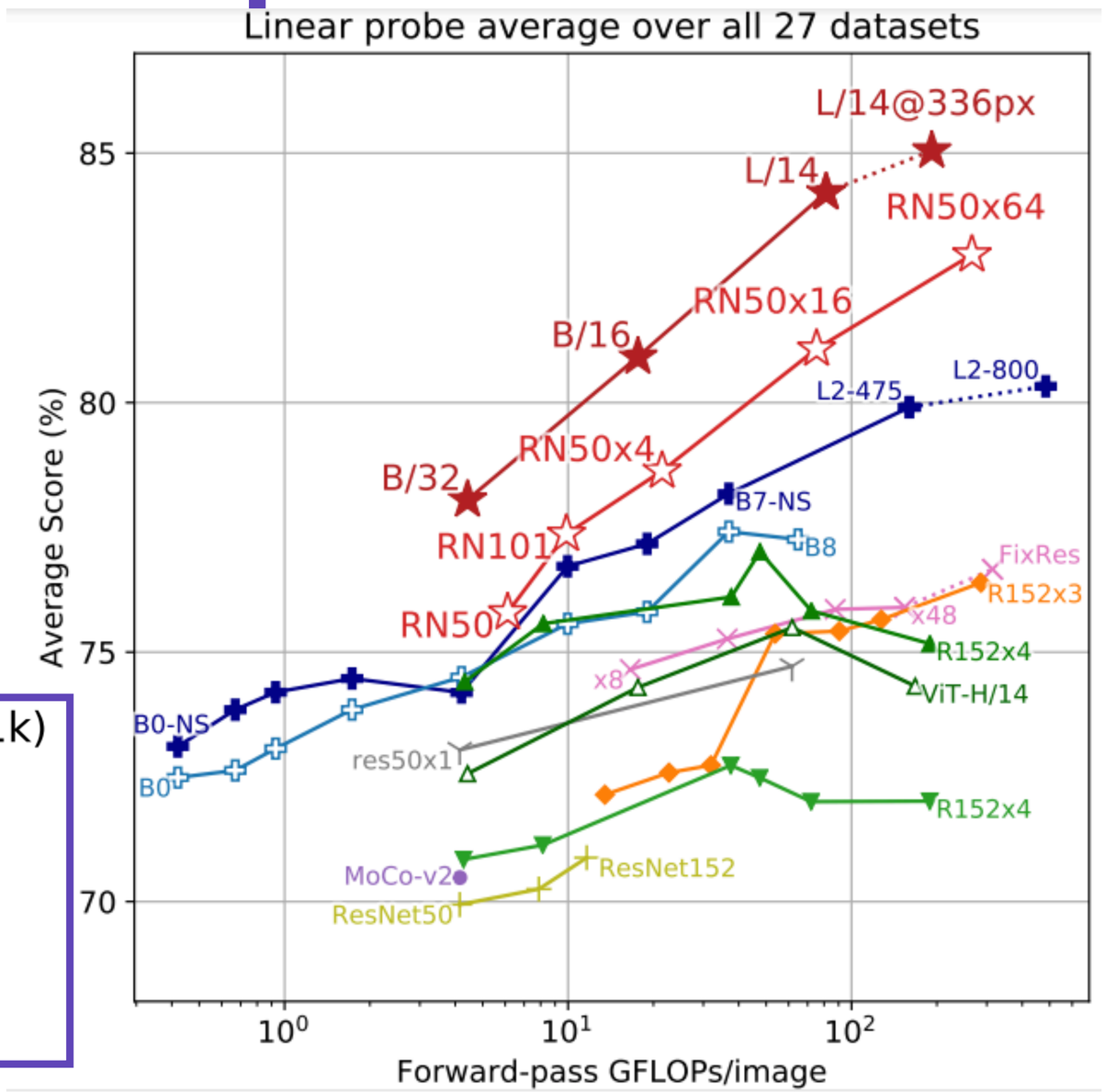
Zero-Shot CLIP Outperforms Supervised baseline.

Supervised baseline Outperforms CLIP.

CLIP vs Unimodal Visual Representations

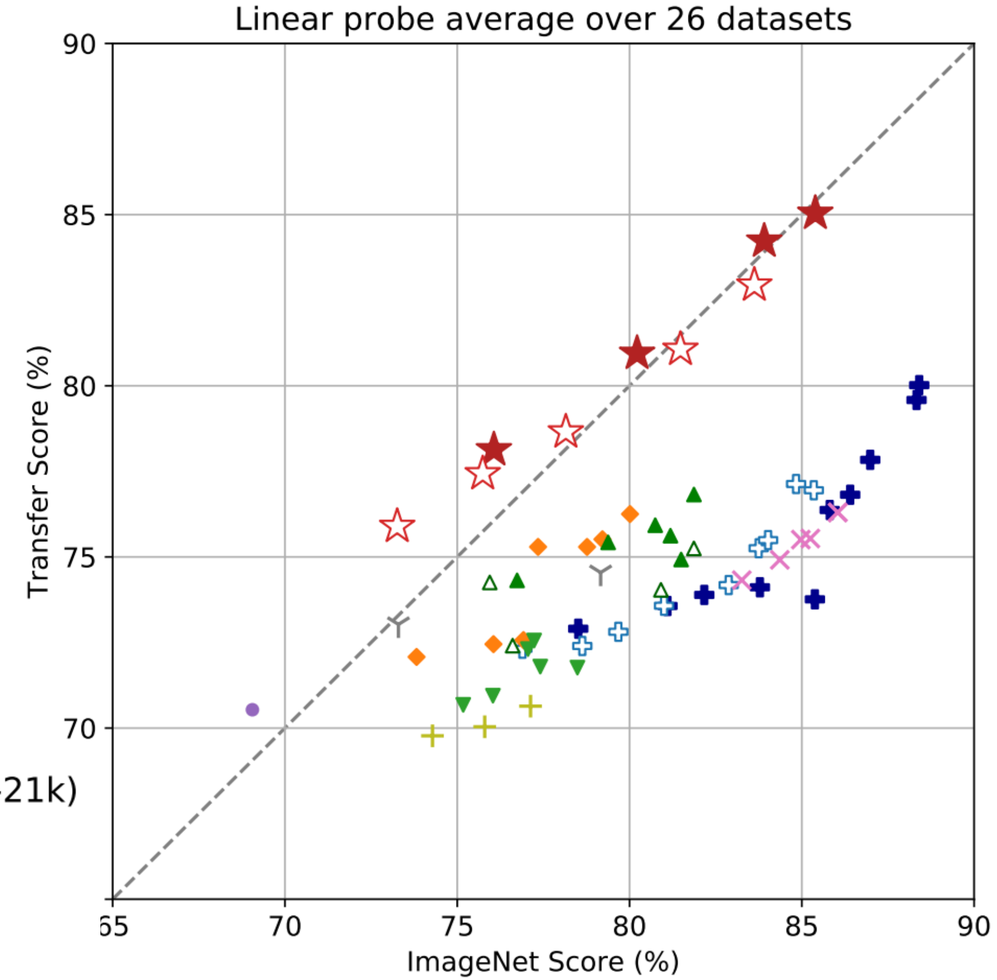
- Linear Probe performance v.s. computer vision models
- CLIP provides visual representations with better transferability





CLIP vs Unimodal Visual Representations

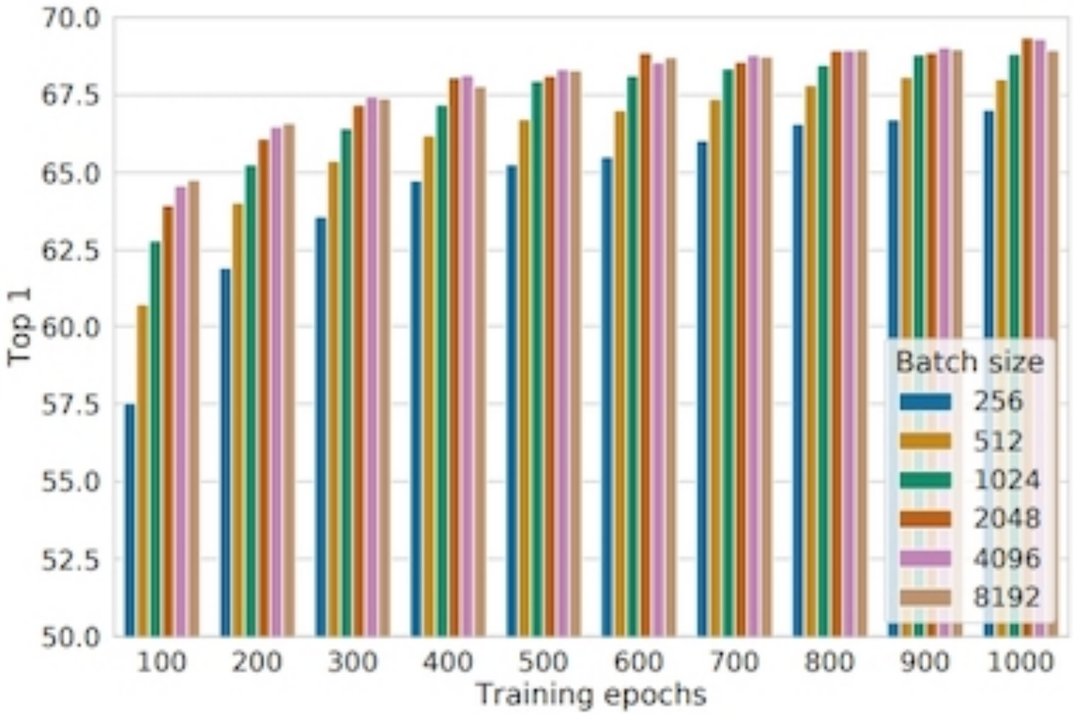
- CLIP features are more **robust to task shift** compared to vision models pre-trained on ImageNet.
- Higher transfer scores of linear probes trained on CLIP over models with similar ImageNet performance.



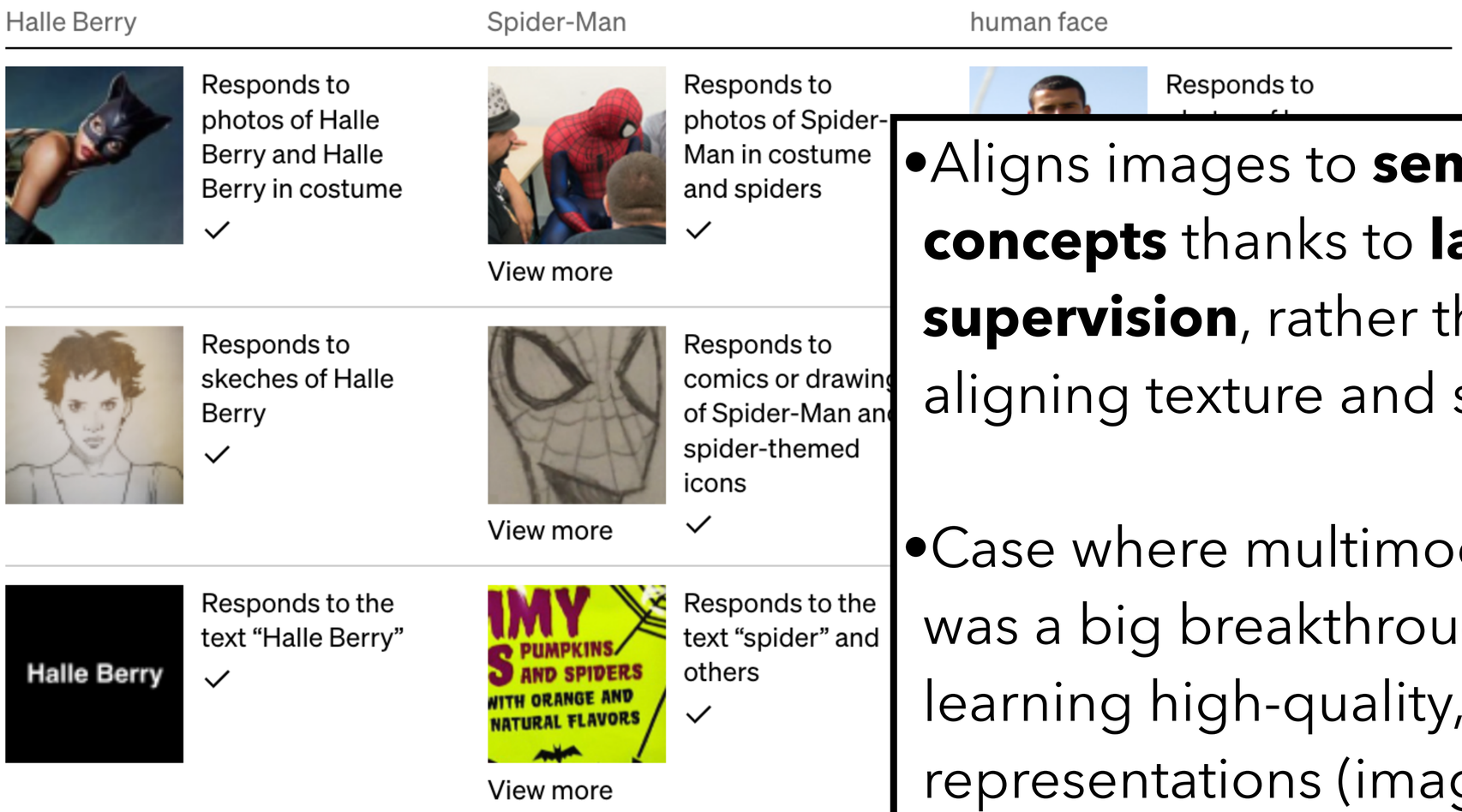
CLIP-ViT
CLIP-ResNet
EfficientNet-NoisyStudent
EfficientNet
Instagram
SimCLRv2
BYOL
MoCo
ViT (ImageNet-21k)
BiT-M
BiT-S
ResNet

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

CLIP, ViT, ResNet

Image Encoder

f_v

x_v

What is the object being thrown?

Word2Vec, BERT, …

Text Encoder

f_t

x_t

Fusion Model

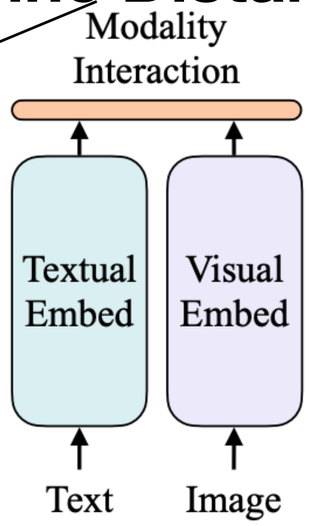
A frisbee

$\mathcal{L}$

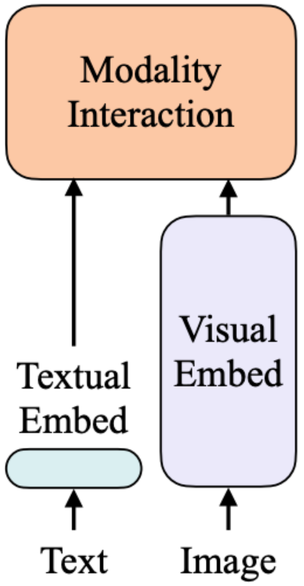
Embedding vs Fusion Trade Offs

Cosine Distance

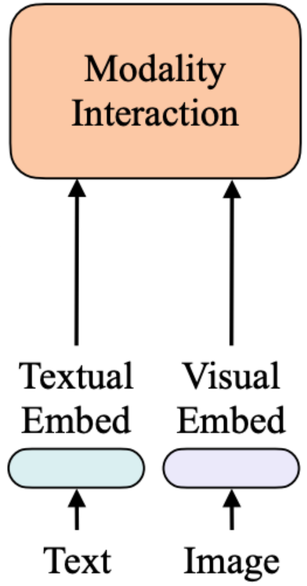
Enough for image-text matching



CLIP

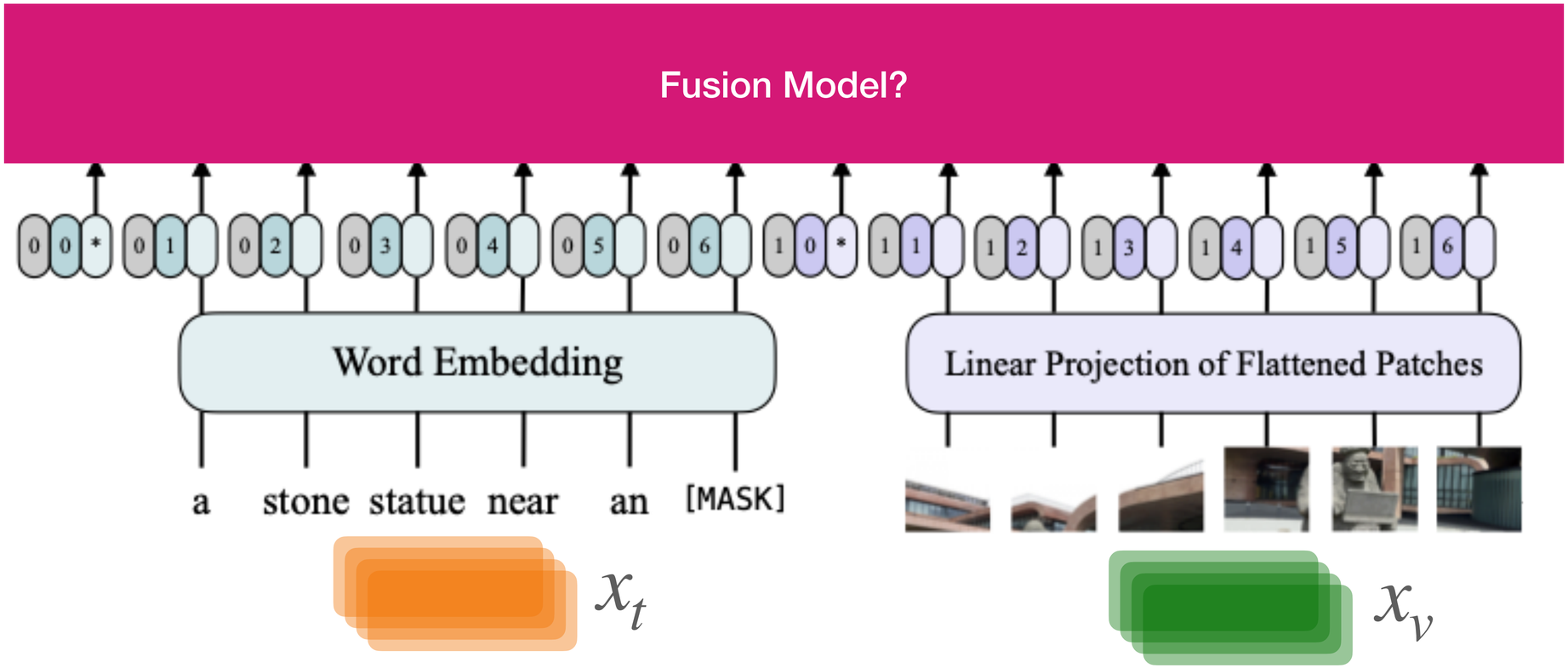


Perhaps, need stronger fusion mechanism for complex reasoning tasks

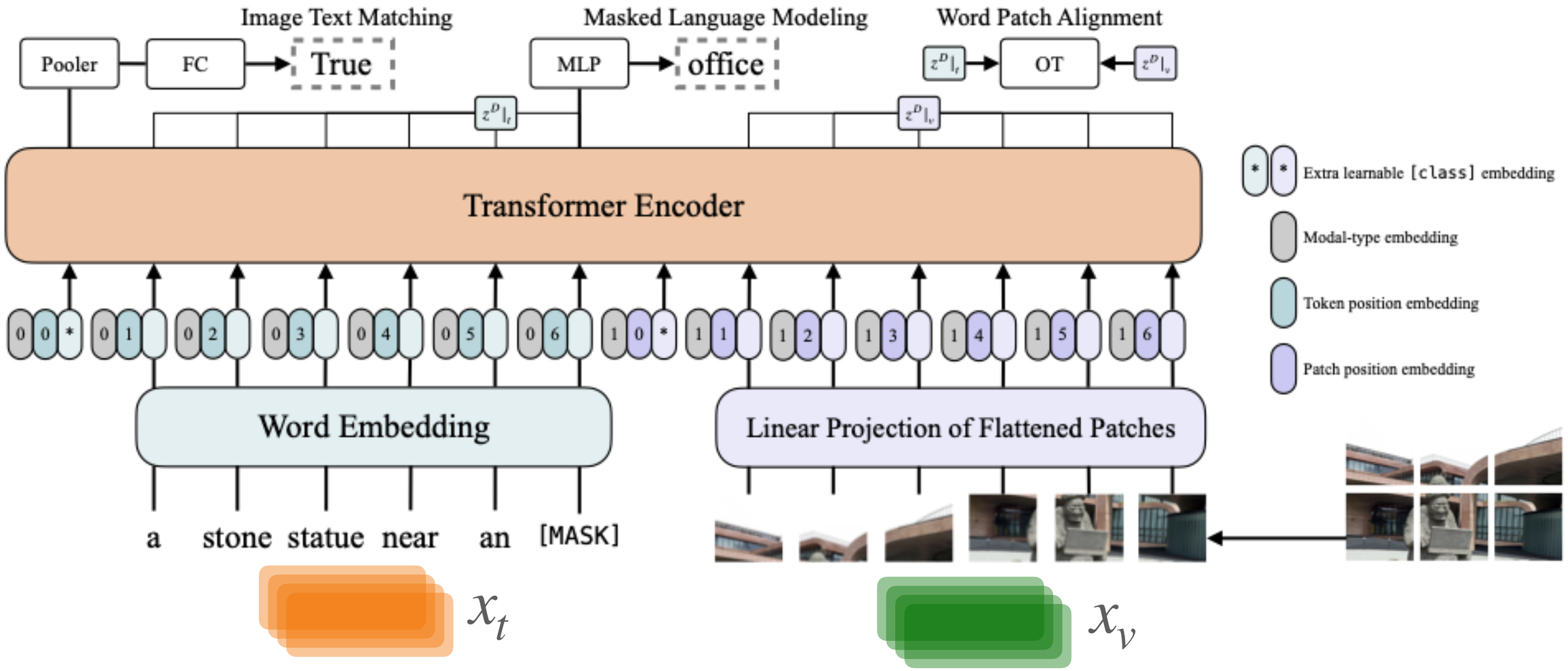


Vision and Language Fusion

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



VILT: The Vision-Language Transformer



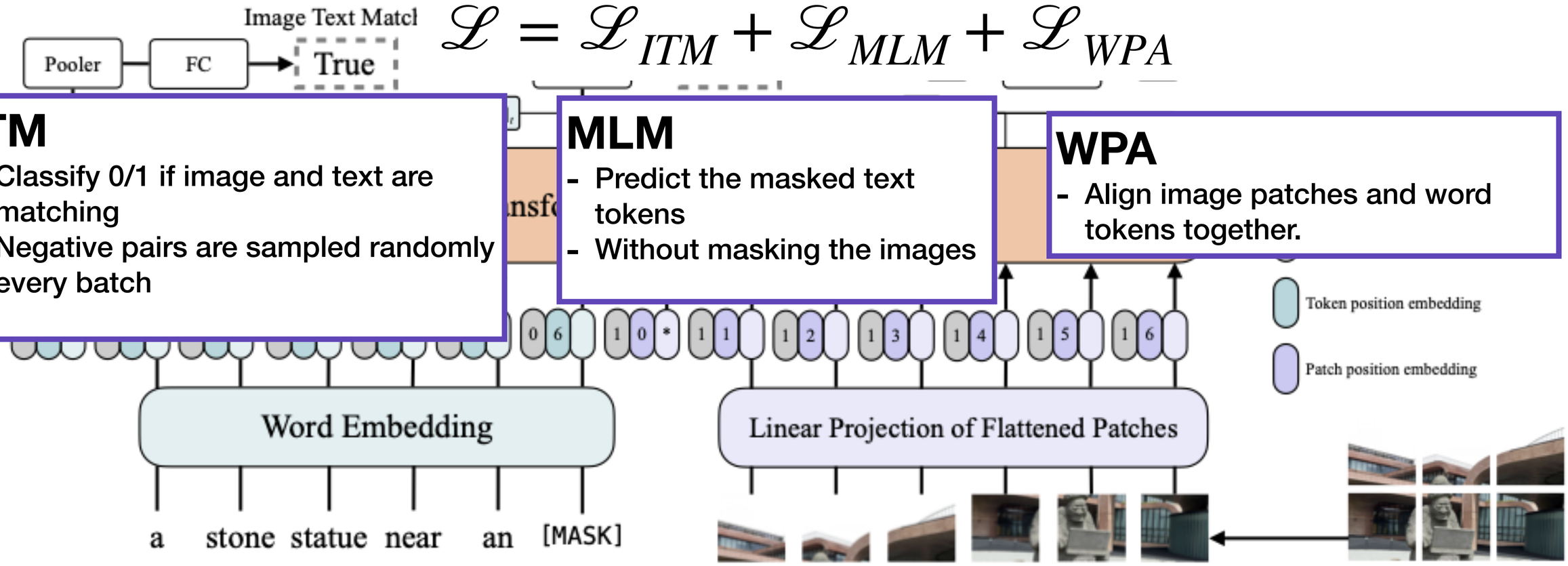
VILT: The Vision-Language Transformer

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

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

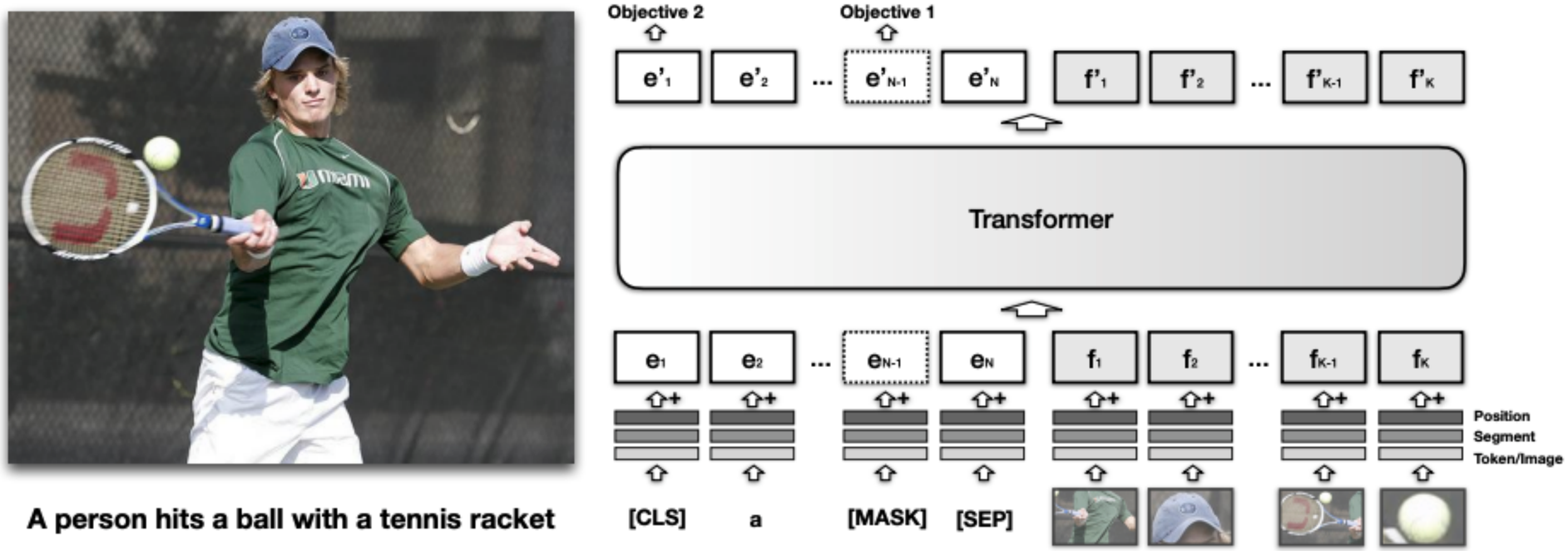
WPA
- Align image patches and word tokens together.



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

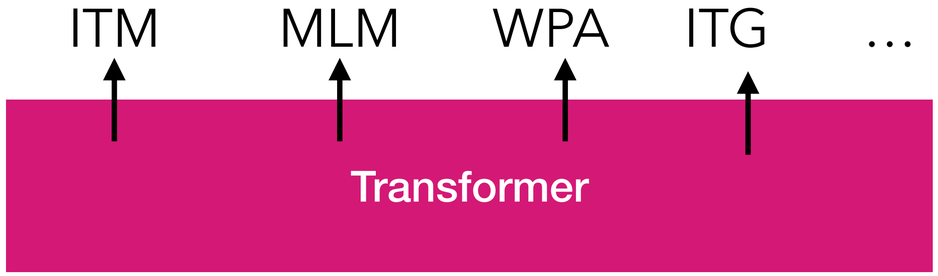
- **Masked Language Modeling** (MLM): Predict labels of masked text tokens.
- **Image-Text Matching** (ITM): Classify if image-text pairs are aligned
- **Word Region/Patch Alignment** (WPA): Align image regions/patches with text tokens

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.



A person throwing a frisbee.

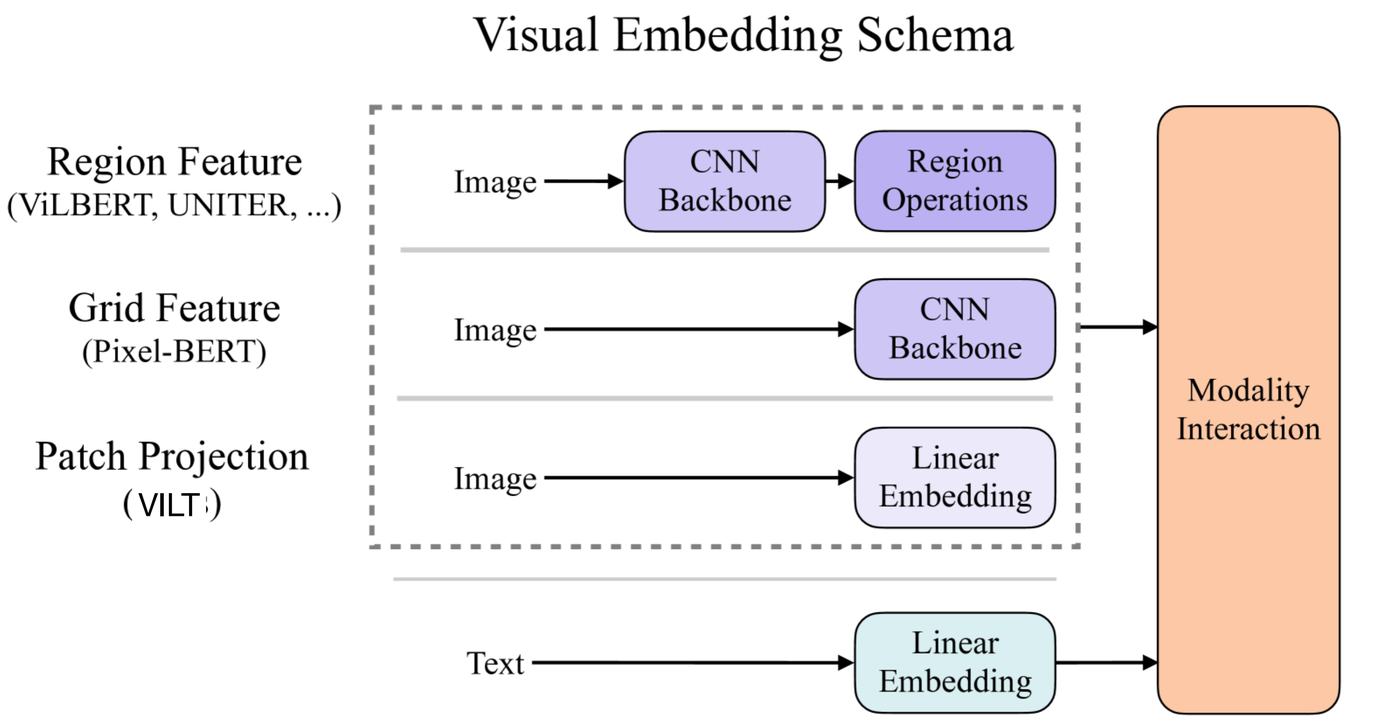
Pre-Training to Learn Image-Text Alignment

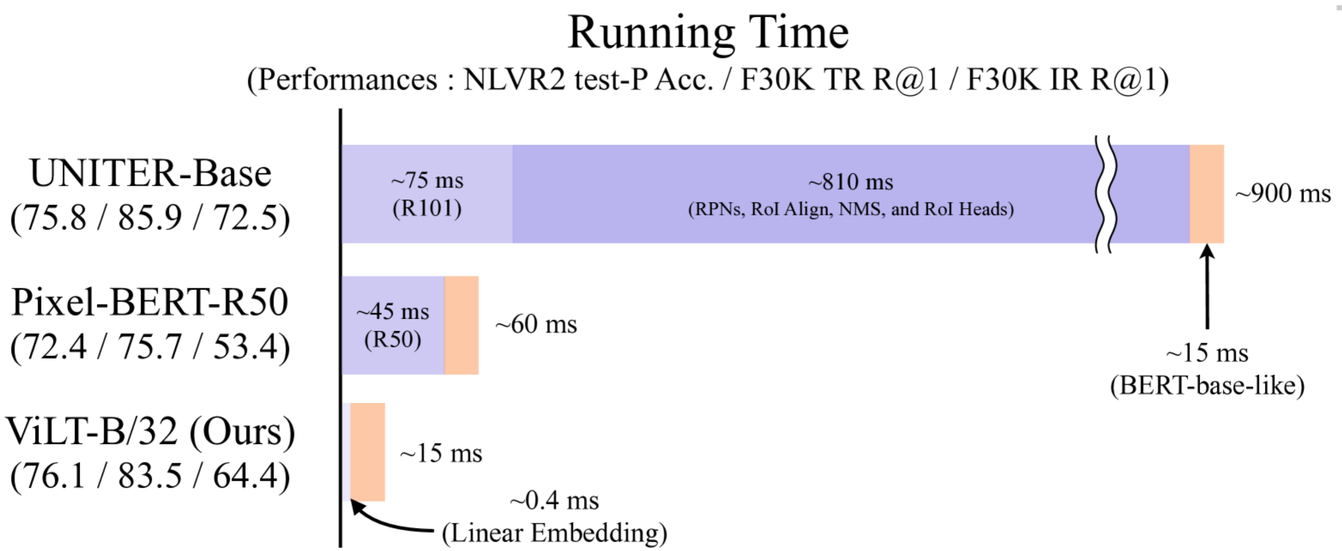
What is the object being thrown?

Finetune to Downstream Tasks (e.g. VQA)

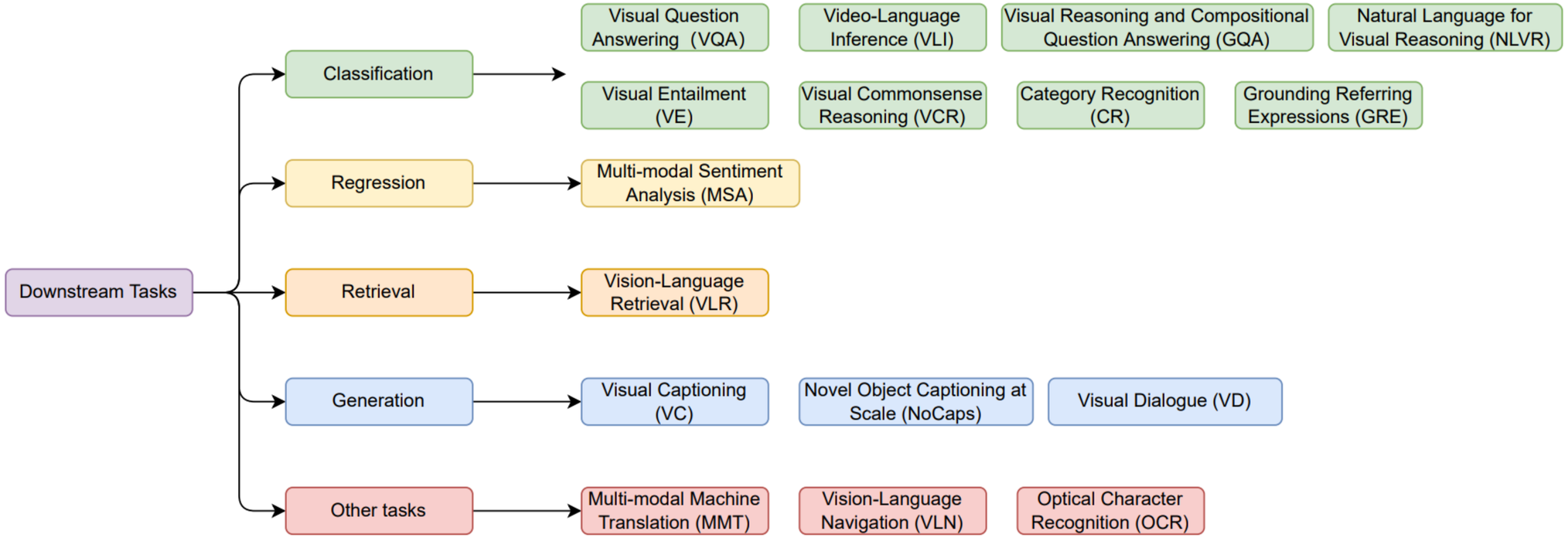
Which model is better?

- **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





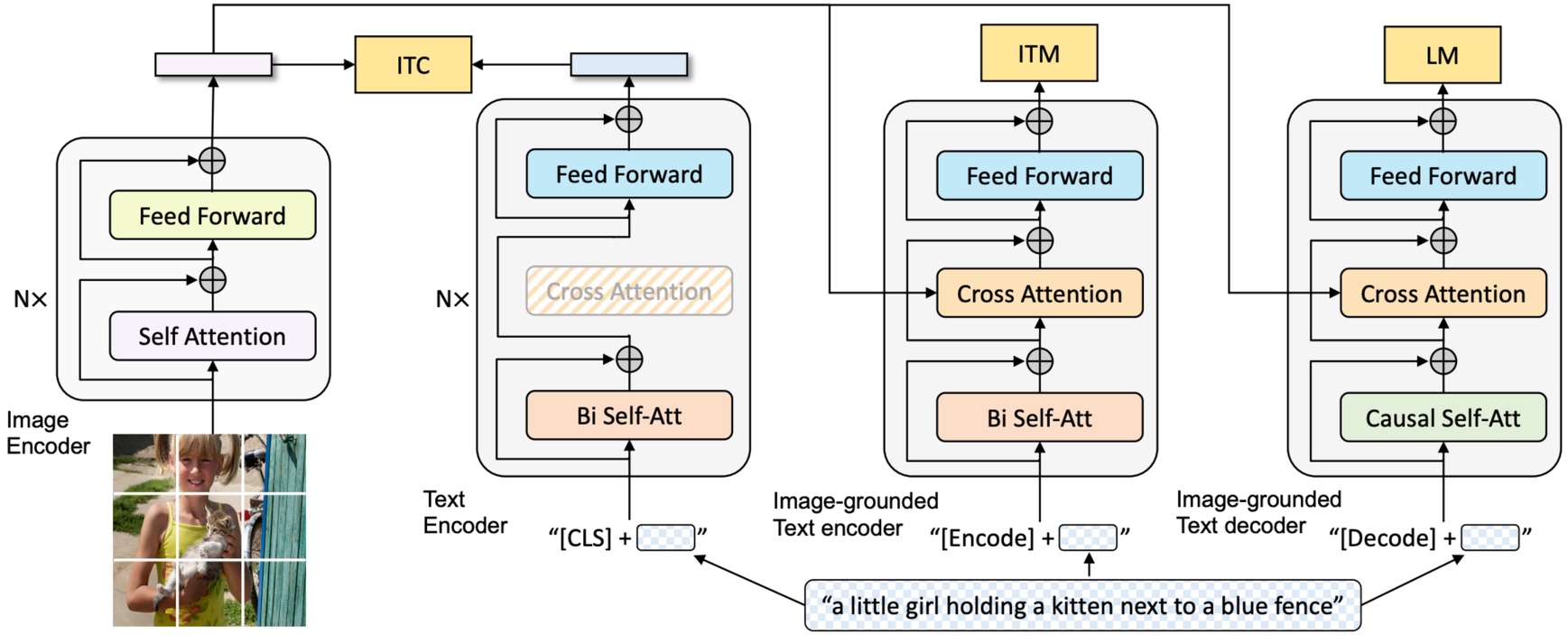
Available Downstream VL Tasks



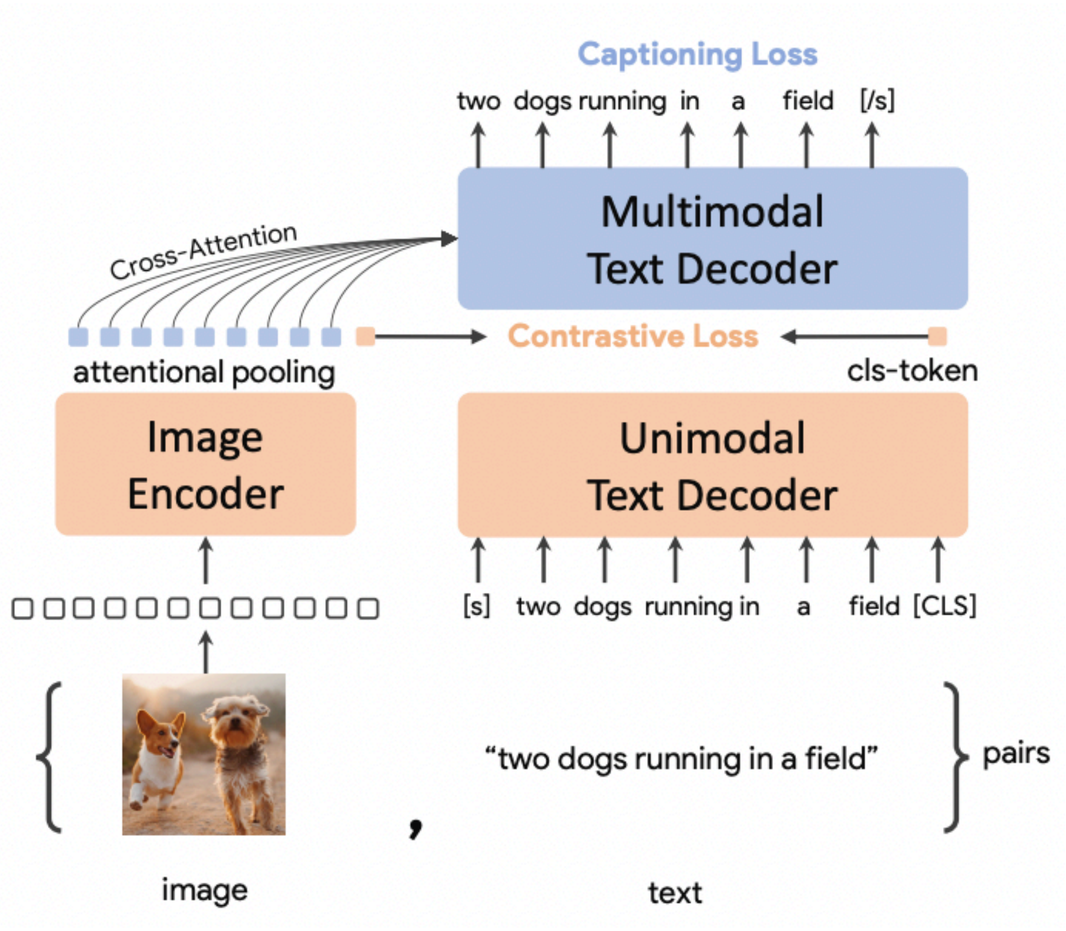
Effective Pre-Training Losses?

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

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



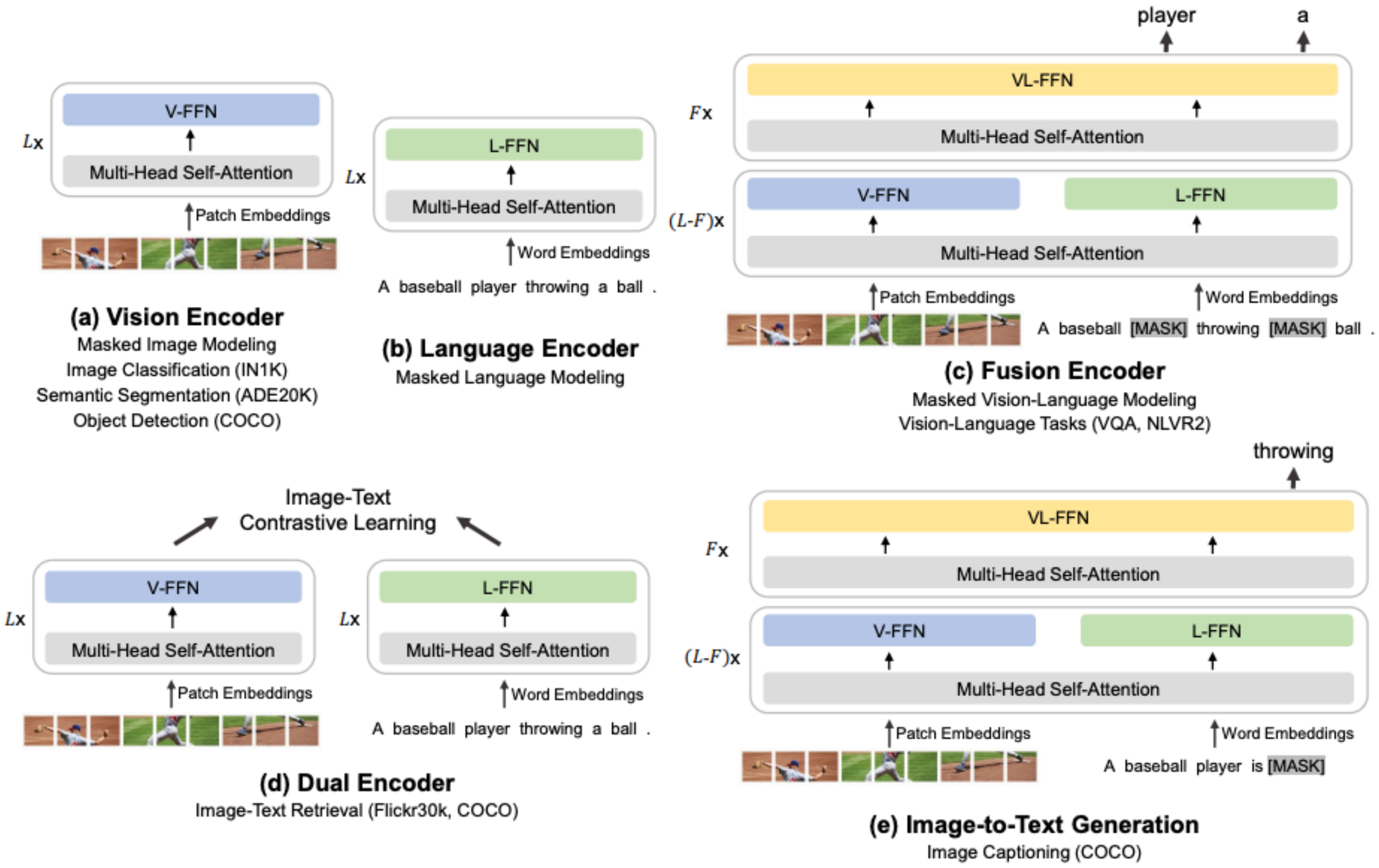
CoCA: Contrastive Captioning



Model	VQA test-dev	VQA test-std	SNLI-VE dev	SNLI-VE test	NLVR2 dev	NLVR2 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†	90.2†	-	-
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



Efficient Training of VLMs?

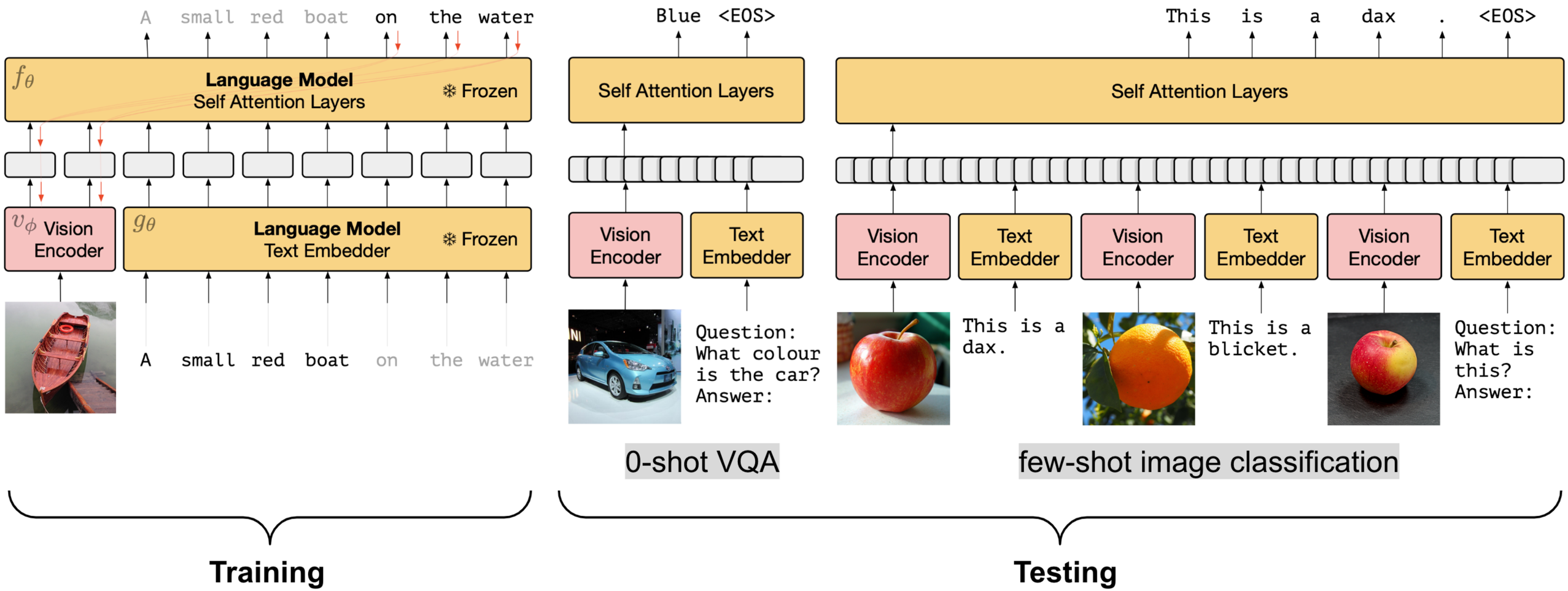
- **Potential Room for Improvement?**: Get even larger amount of data, get $$$ GPUs, with more objectives, and train for long time to outperform the existing models.
- **Need for efficient training:**
  - Can we exploit the already pre trained representations?
    - Do we need to train the whole model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

Multimodality

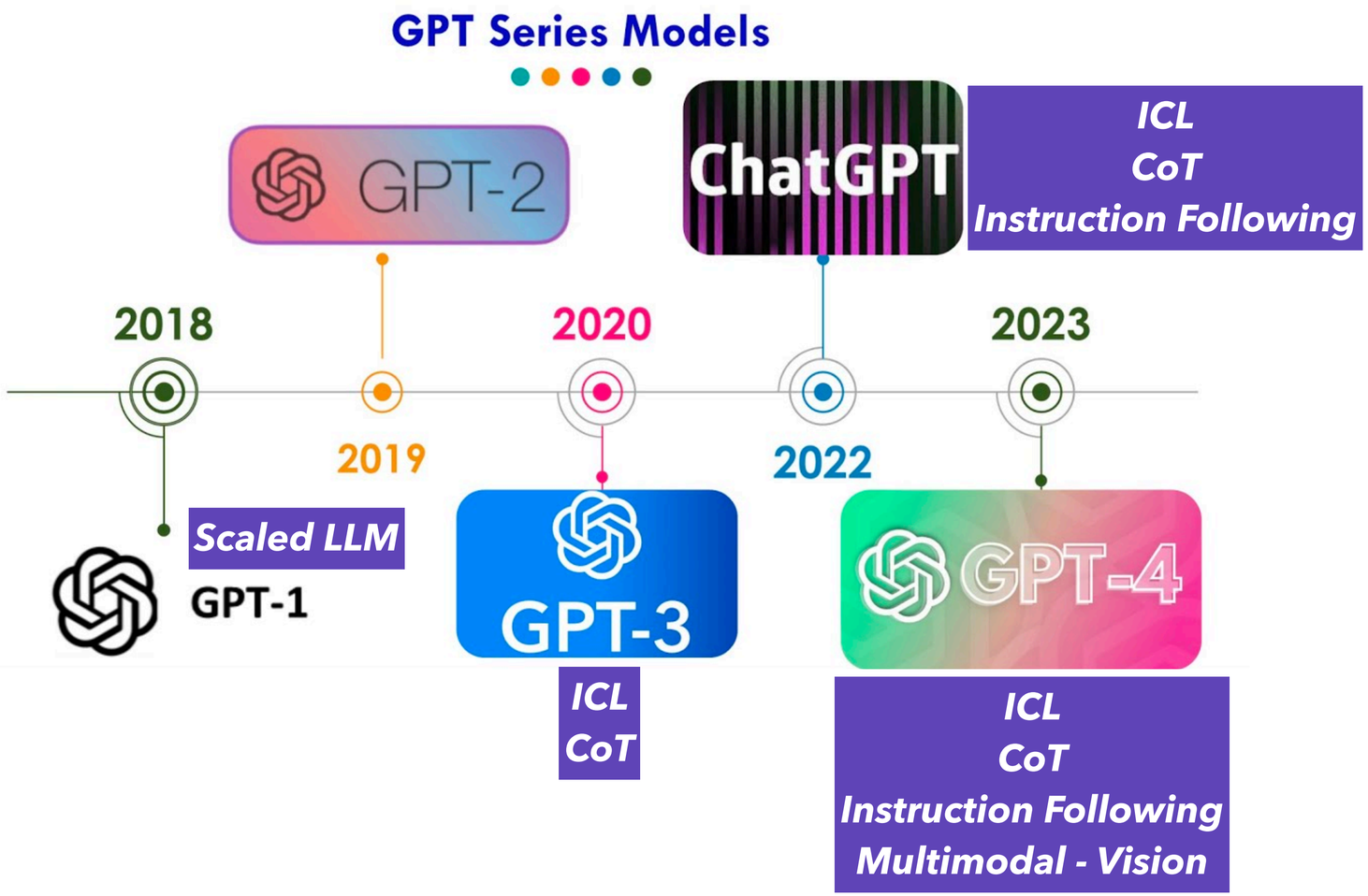
Instruction tuning,

In-context Learning,

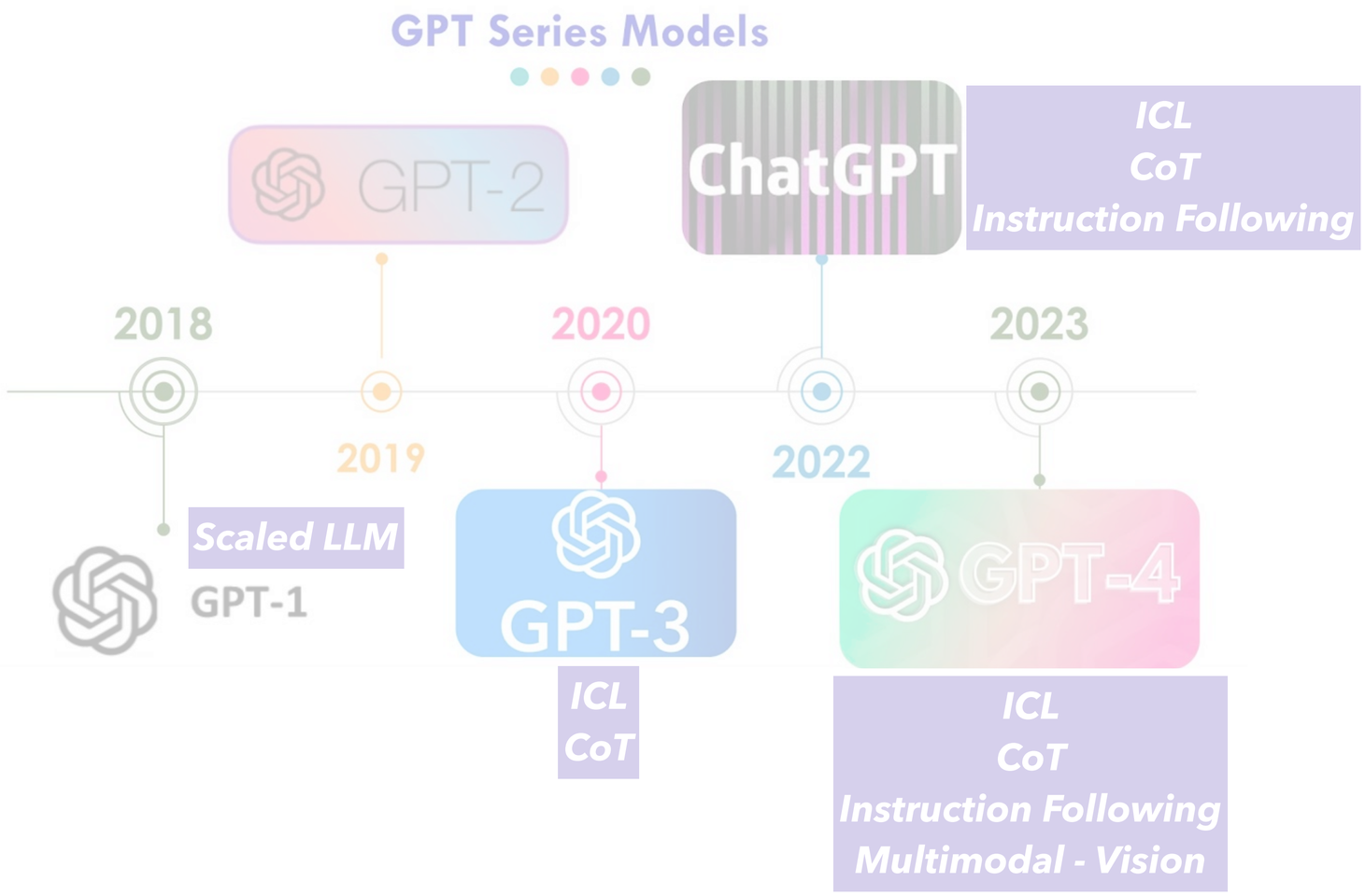
Frozen: Prefix Tuning of Image Embeddings

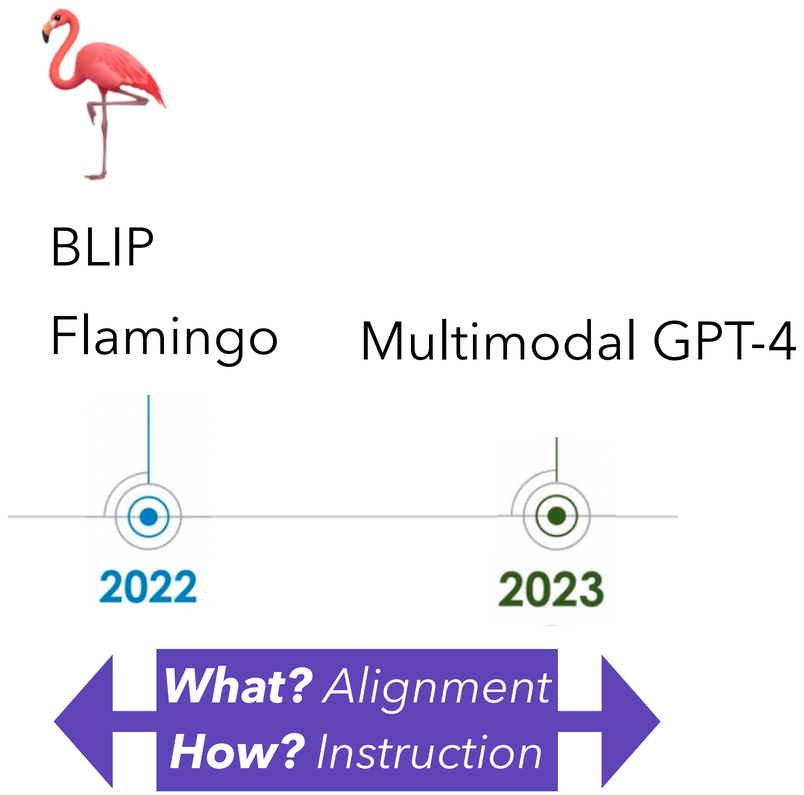


Gaps in Multimodal LMs



Gaps in Multimodal LMs





Instruction Tuning in LLMs

- Easily 0-shot **generalizable** (previously **hard** to generalize)
- Task instructions are given **explicitly** (previously **implicit**)
- Multi-tasking with specific **instructions** (previously **objectives**)
    - Easy to mix tasks with instructions

Instruction | Input | Output

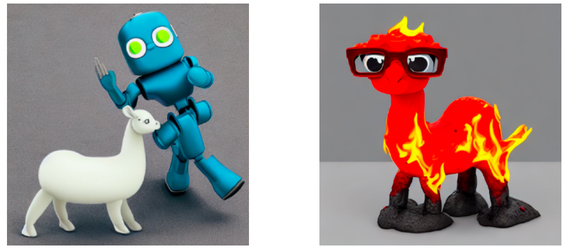
How about MM Instruction-tuning?

- Data?
- Model?
- Improvements?
- Open Questions?

Self-Instruct – strong (V)LM teachers

- LLM Teachers:
  - GPT 3.5
  - ShareGPT
- Resulting IF models

- Challenge for VLMs:
  - VLM Teachers?
  - Can we use LLM teachers?

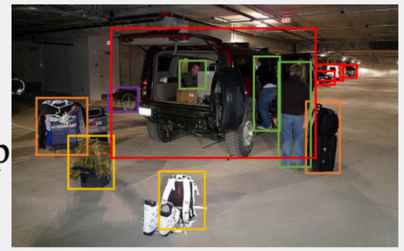


Use text-only teachers to bootstrap

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

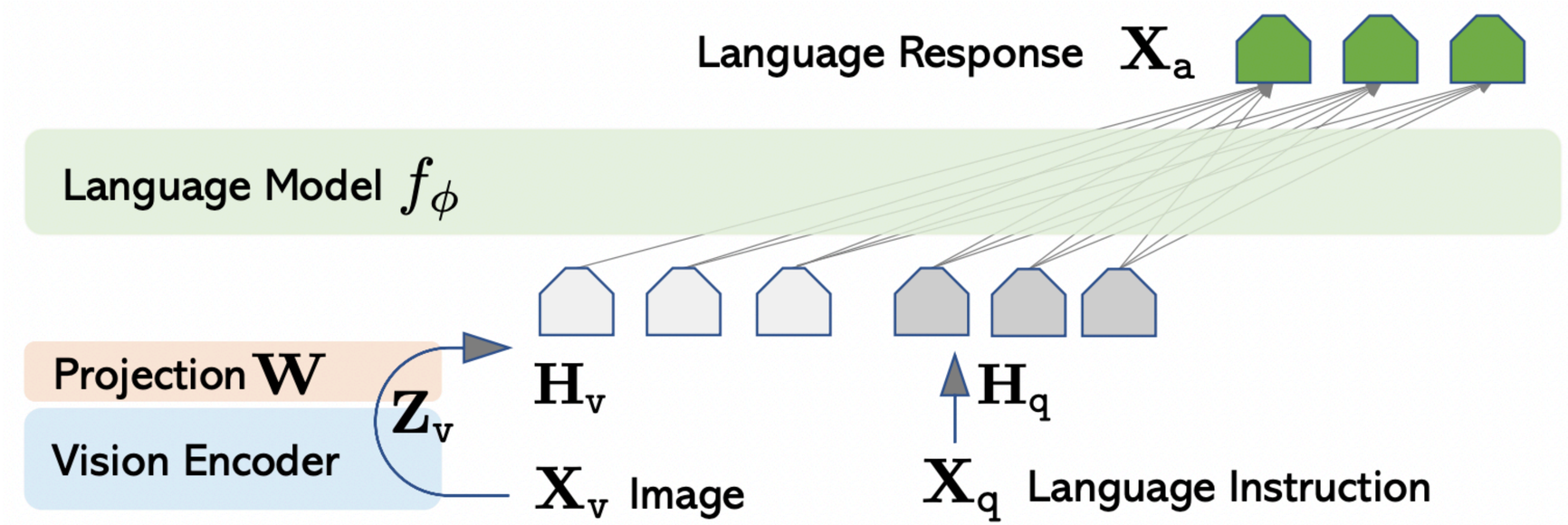
- 158k instruction following samples
  - 58k conversations
  - 23k detailed descriptions
  - 77k complex reasoning

LLaVA: Large Lang and Vis Assistant

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

LLAMA

CLIP VE
VIT-L/14



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

$$\mathbf{X}^t_{\mathtt{instruct}} = \begin{cases} \text{Randomly choose } [\mathbf{X}^1_{\mathrm{q}}, \mathbf{X}_{\mathrm{v}}] \text{ or } [\mathbf{X}_{\mathrm{v}}, \mathbf{X}^1_{\mathrm{q}}], & \text{the first turn } t = 1 \\ \mathbf{X}^t_{\mathrm{q}}, & \text{the remaining turns } t > 1 \end{cases}$$

Dual stage training

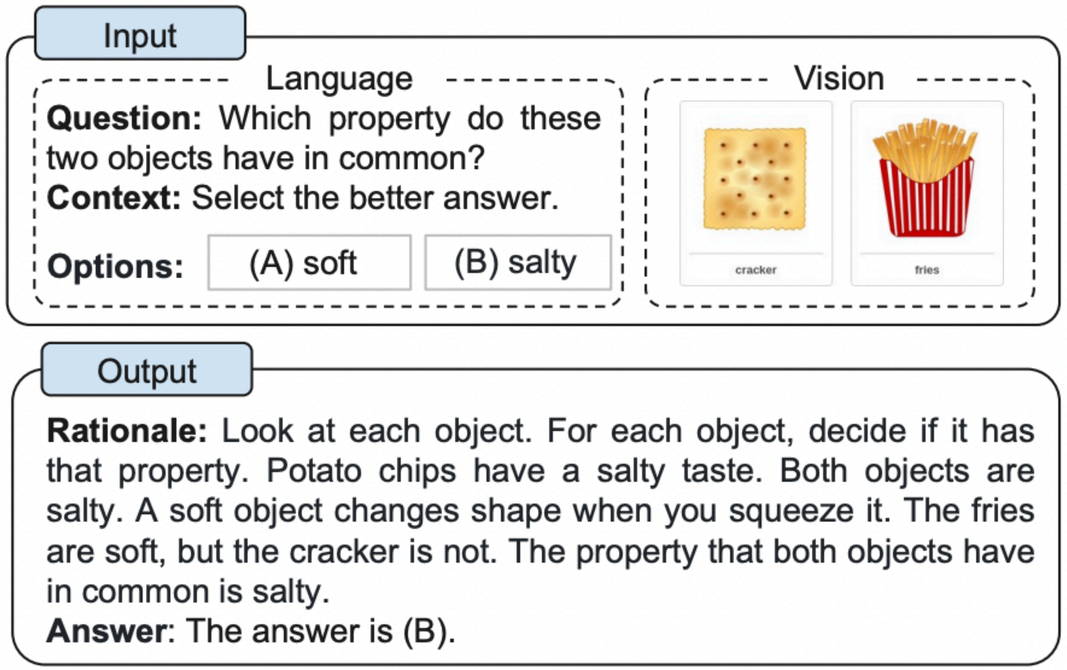
- Stage 1: **Pre-training for feature alignment**
  - Only projection matrix is updated
  - Trained on a subset of CC3M (595k IT pairs)
- Stage 2: **Fine-tuning for user and task orientation**
  - Both projection matrix and LLM are updated
  - Tuned on Visual chat (user chat-like orientation 158k) &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

Open Questions

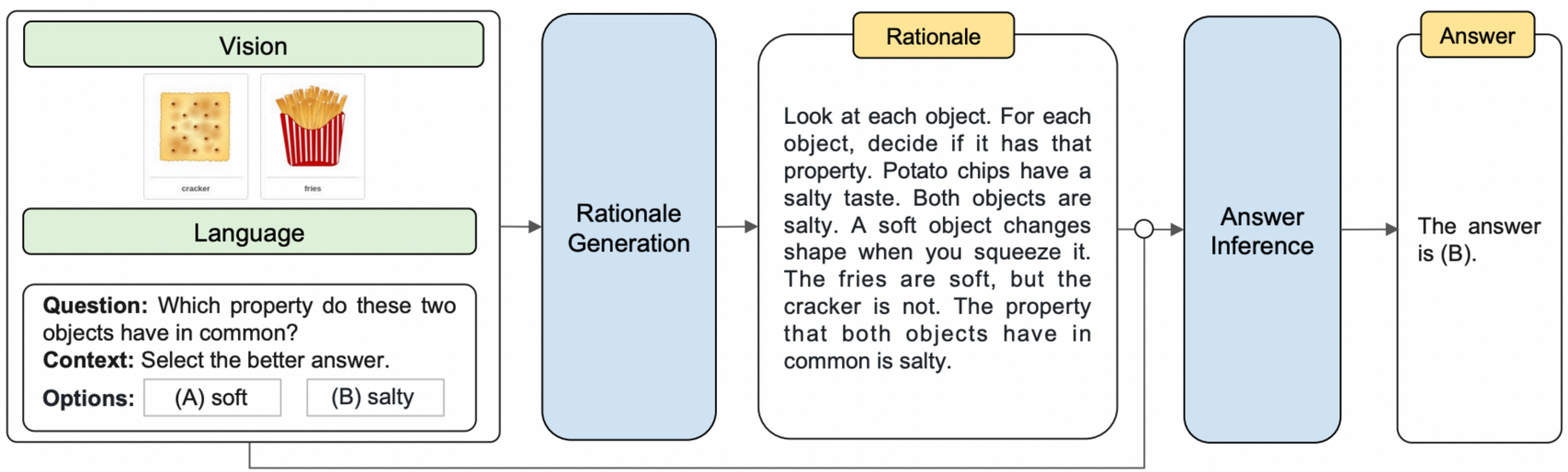
- Human preference in Alignment
- Retrieval Augmentation for In-Context Learning

Multimodal Chain of Thought

- Plug and Play style



Multimodal Chain of Thought



T5

$$H_{\text{language}} = \text{LanguageEncoder}(X_{\text{language}}),$$

DETR (patch rep); projection layer

$$H_{\text{vision}} = W_h \cdot \text{VisionExtractor}(X_{\text{vision}}),$$

Self-attention

$$H^{\text{attn}}_{\text{vision}} = \text{Softmax}(\frac{\mathcal{Q}\mathcal{K}^\top}{\sqrt{d_k}})\mathcal{V},$$

Gating

$$\lambda = \text{Sigmoid}(W_l H_{\text{language}} + W_v H^{\text{attn}}_{\text{vision}}),$$

$$H_{\text{fuse}} = (1-\lambda) \cdot H_{\text{language}} + \lambda \cdot H^{\text{attn}}_{\text{vision}},$$

Open Questions

Method	(i) QCM→ R	(ii) QCMR→ A
Two-Stage Framework	91.76	70.53
w/ Captions	91.85	71.12
w/ Vision Features	96.97	84.91

Answer	CoT Category	Percentage (%)
Correct	CoT is correct	90
	CoT is incorrect	10
Incorrect	Commonsense Mistake	82
	Logical Mistake	12
	CoT is correct	6

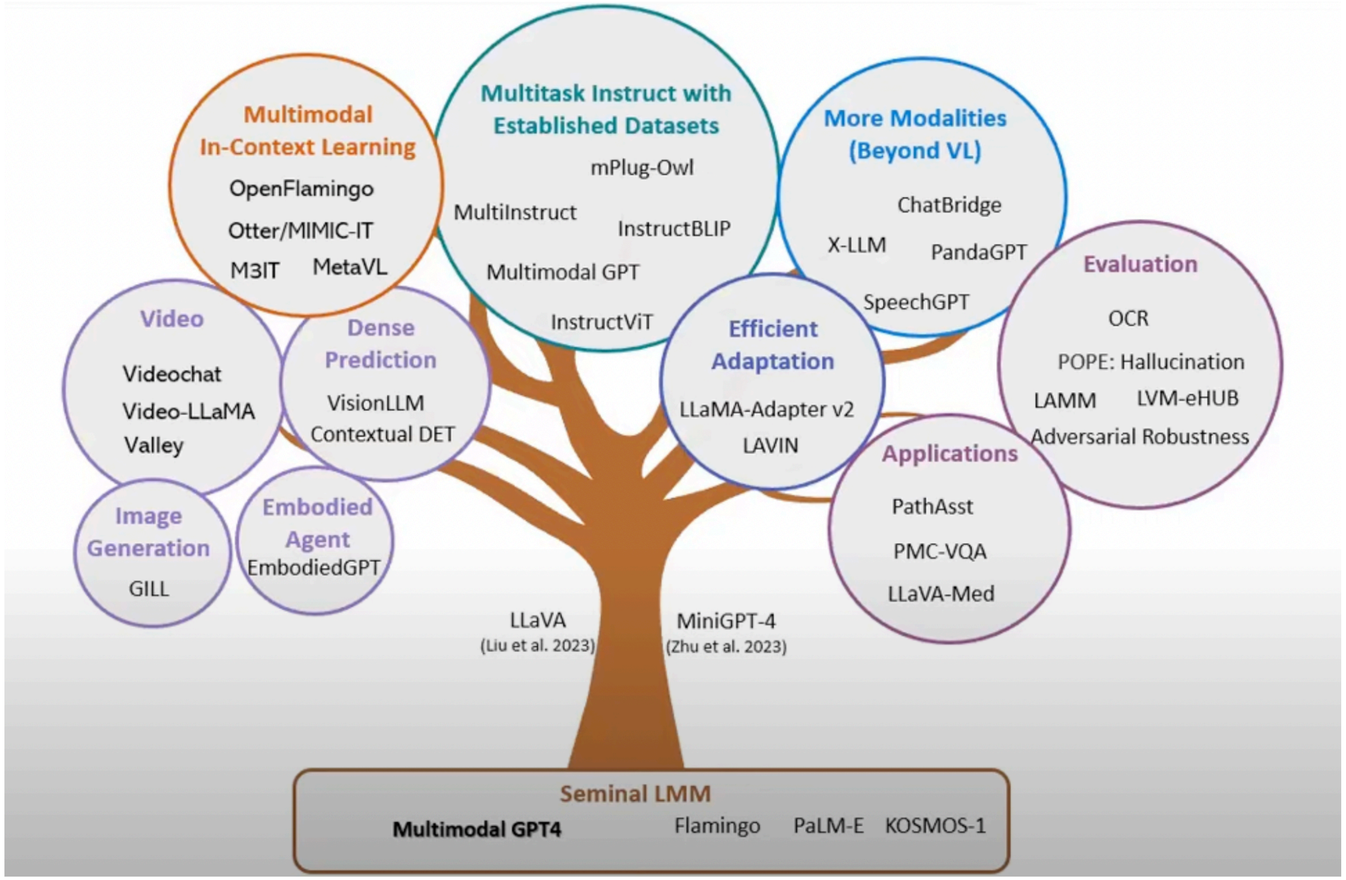
Including Images help

Gaps in dense captions?

Inconsistency between

CoT and answer?

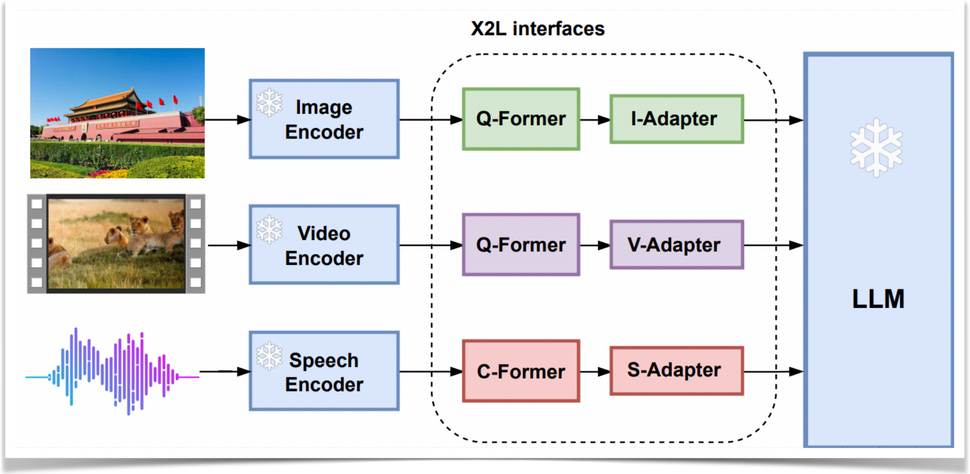
Map of other topics



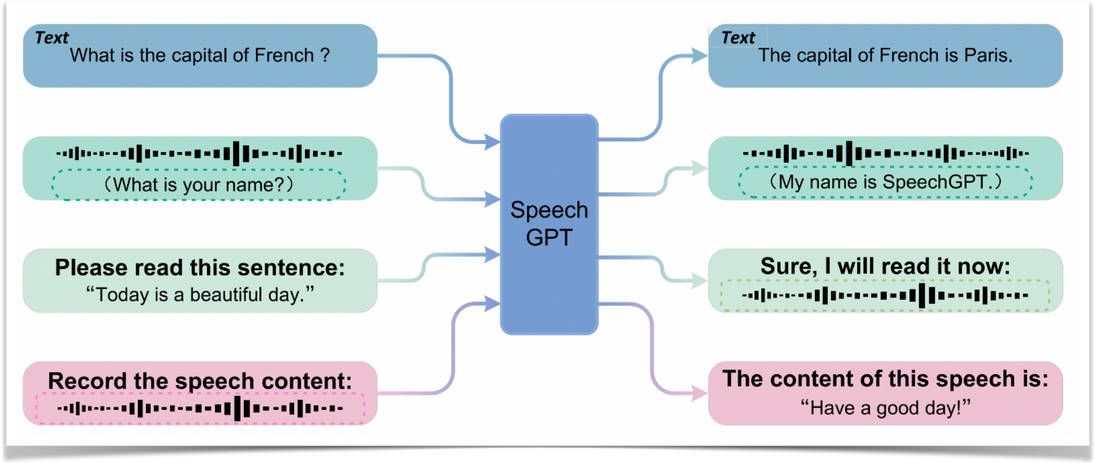
- Credit: VLP Tutorial @CVPR'23

More Modalities (Beyond VL)

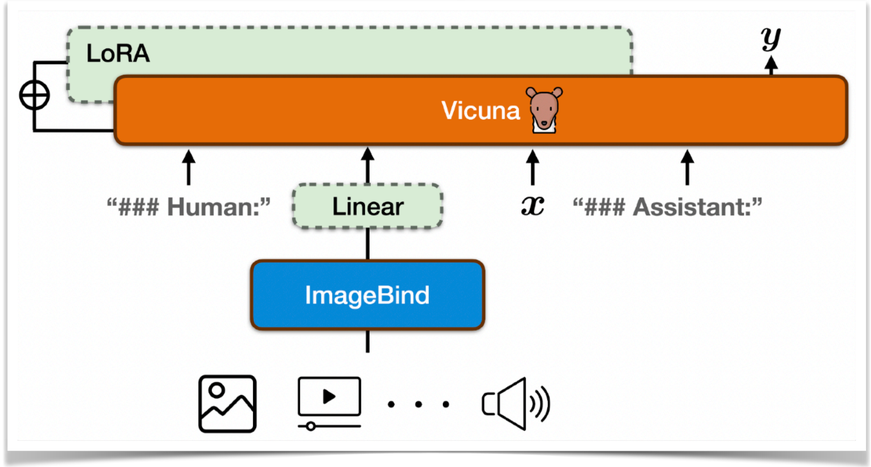
X-LLM



Speech-GPT



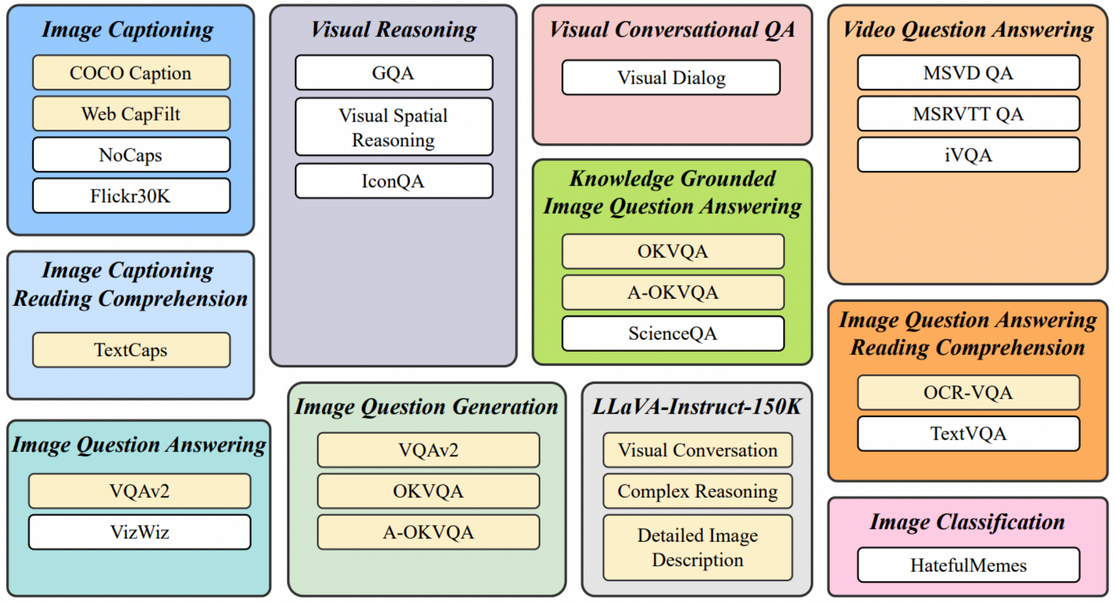
Panda-GPT



Instruct-BLIP

- Task and datapoint sampling
- Convert academic datasets into instructions

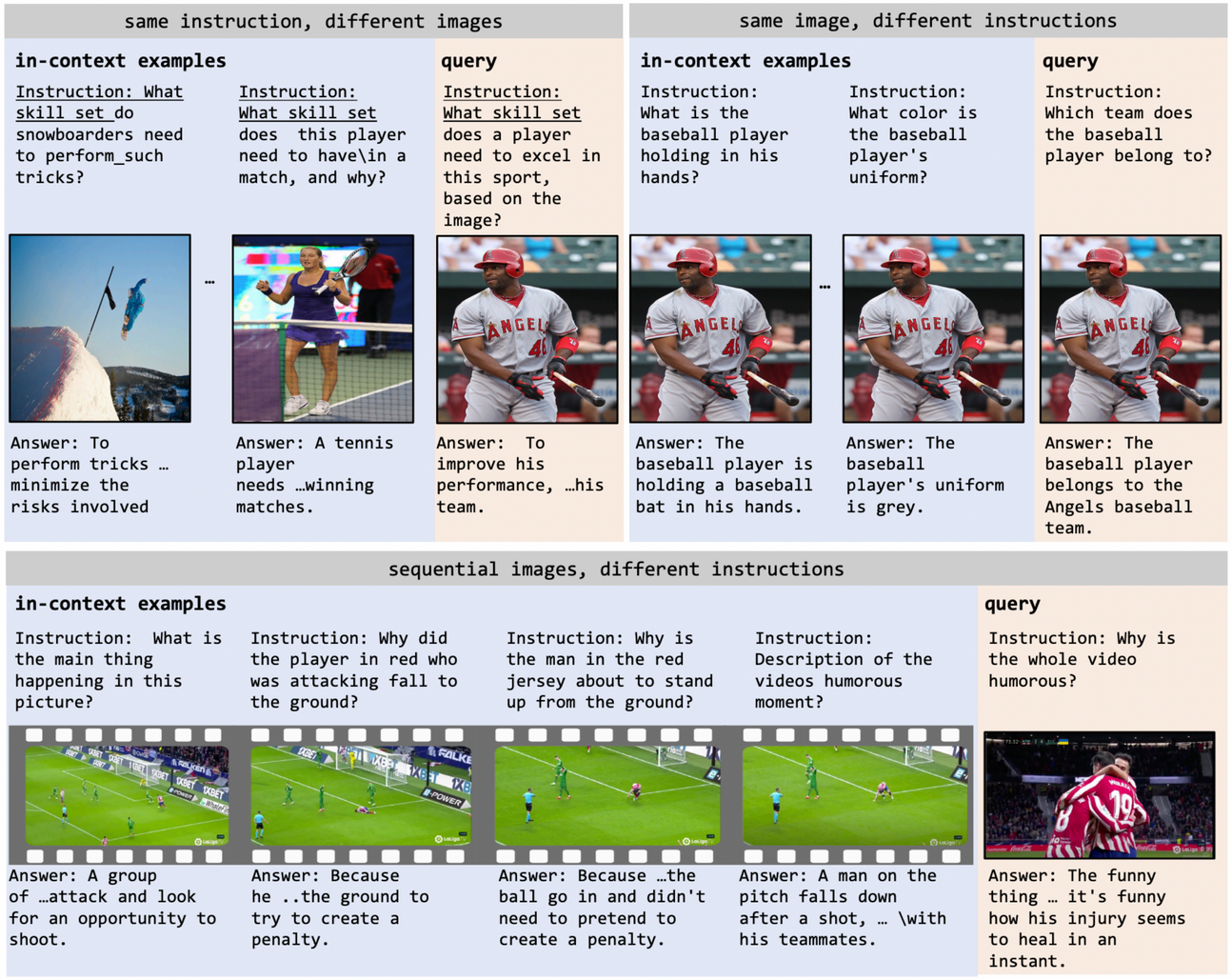
$$p_d = \frac{\sqrt{S_d}}{\sum_{i=1}^{D} \sqrt{S_i}}$$



Multimodal In-context learning

- Otter:
  - Different Images, Same Ins
  - Same Images, Different Ins
  - Seq Images, Different Ins
- M3IT:
  - Also multilingual
- MetaVL:
  - Train LLM to in-context learn
  - Append a Visual Encoder
- IDEFICS:
  - Interleaved VL in arbitrary orders

Note: Multiple images



(b) Multi-Modal In-Context Instruction Tuning (MIMIC-IT)

Applications to Domains/Tasks

Medical:

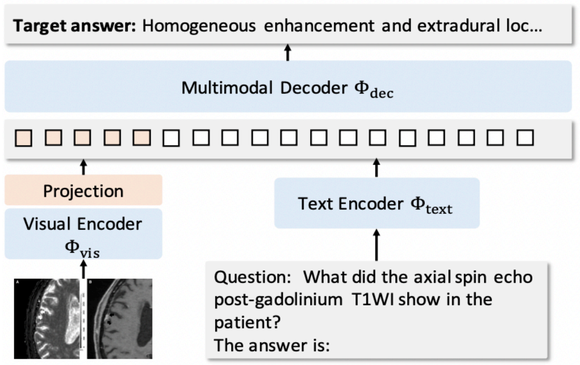
Med-LLaVA

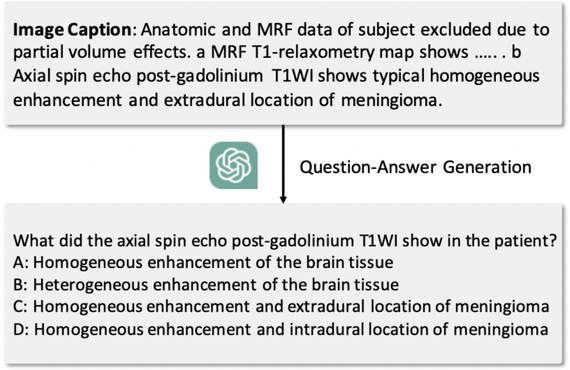
PMC-VQA

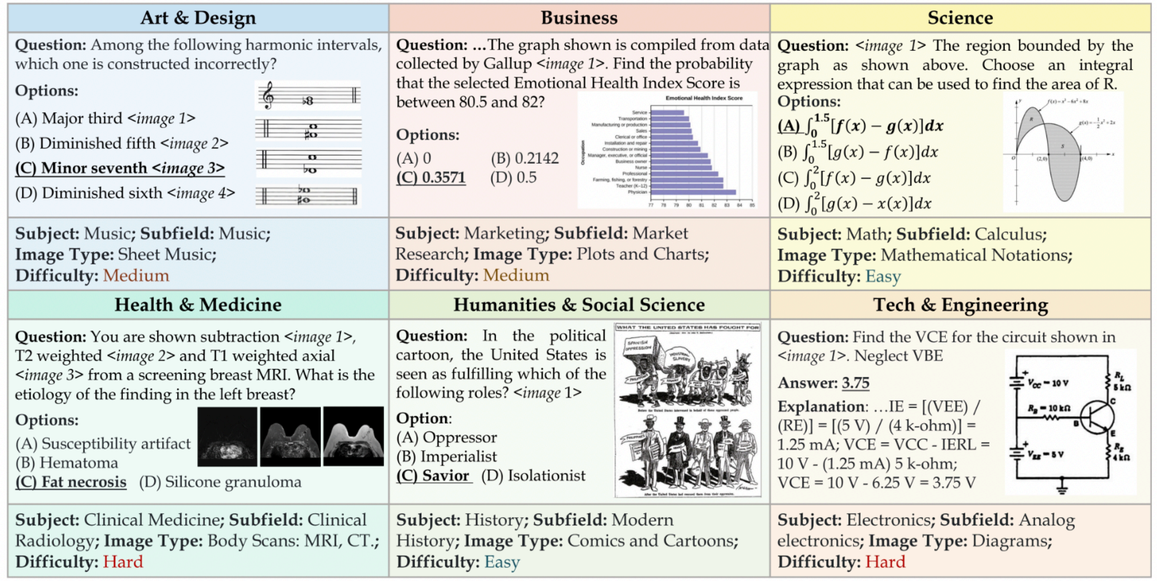
Domains - pathology, geometry, art and design

Image types - diagrams, tables, plots, chemical structures

Expert skill - Mathematical equations, science formula







Art & Design	Business	Science
Question: Among the following harmonic intervals, which one is constructed incorrectly? **Options:** (A) Major third <*image 1*> (B) Diminished fifth <*image 2*> **(C) Minor seventh <*image 3*>** (D) Diminished sixth <*image 4*>	**Question:** ...The graph shown is compiled from data collected by Gallup <*image 1*>. Find the probability that the selected Emotional Health Index Score is between 80.5 and 82? **Options:** (A) 0 (B) 0.2142 **(C) 0.3571** (D) 0.5	**Question:** <*image 1*> The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R. **Options:** **(A)** $\int_0^{1.5}[f(x)-g(x)]dx$ (B) $\int_0^{1.5}[g(x)-f(x)]dx$ (C) $\int_0^{2}[f(x)-g(x)]dx$ (D) $\int_0^{2}[g(x)-x(x)]dx$
Subject: Music; **Subfield:** Music; **Image Type:** Sheet Music; **Difficulty:** Medium	**Subject:** Marketing; **Subfield:** Market Research; **Image Type:** Plots and Charts; **Difficulty:** Medium	**Subject:** Math; **Subfield:** Calculus; **Image Type:** Mathematical Notations; **Difficulty:** Easy
Health & Medicine	**Humanities & Social Science**	**Tech & Engineering**
Question: You are shown subtraction <*image 1*>, T2 weighted <*image 2*> and T1 weighted axial <*image 3*> from a screening breast MRI. What is the etiology of the finding in the left breast? **Options:** (A) Susceptibility artifact (B) Hematoma **(C) Fat necrosis** (D) Silicone granuloma	**Question:** In the political cartoon, the United States is seen as fulfilling which of the following roles? <*image 1*> **Option:** (A) Oppressor (B) Imperialist **(C) Savior** (D) Isolationist	**Question:** Find the VCE for the circuit shown in <*image 1*>. Neglect VBE **Answer:** 3.75 **Explanation:** ...IE = [(VEE) / (RE)] = [(5 V) / (4 k-ohm)] = 1.25 mA; VCE = VCC - IERL = 10 V - (1.25 mA) 5 k-ohm; VCE = 10 V - 6.25 V = 3.75 V
Subject: Clinical Medicine; **Subfield:** Clinical Radiology; **Image Type:** Body Scans: MRI, CT.; **Difficulty:** Hard	**Subject:** History; **Subfield:** Modern History; **Image Type:** Comics and Cartoons; **Difficulty:** Easy	**Subject:** Electronics; **Subfield:** Analog electronics; **Image Type:** Diagrams; **Difficulty:** Hard

mmmu-benchmark.github.io/

BIG Gap remains – Open Directions

- Encoding high resolution images
- Encoding long sequences
- Integrating domain knowledge