Introduction to Vision-Language Models

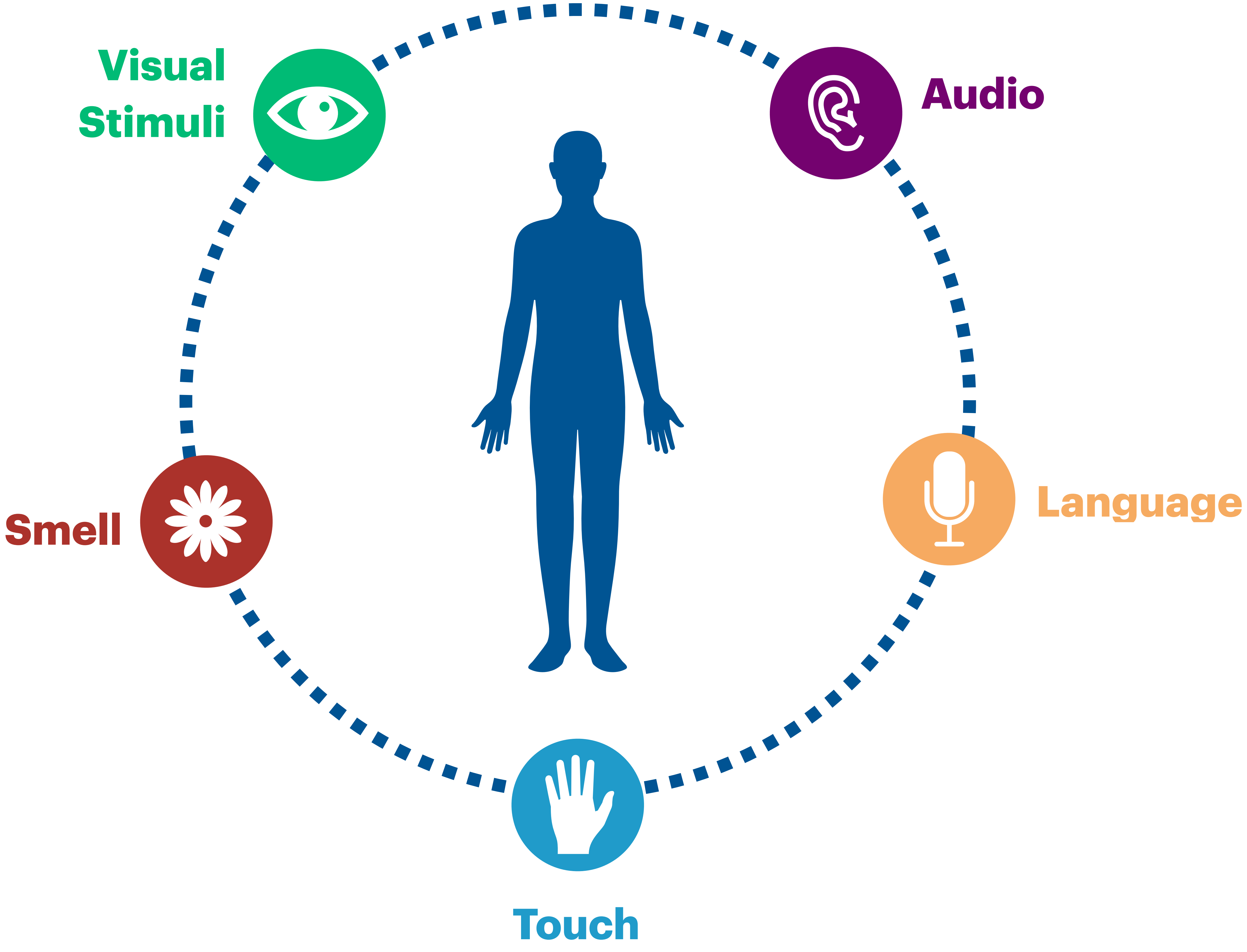
BIODS 271 / CS 277

Maya Varma

Stanford University

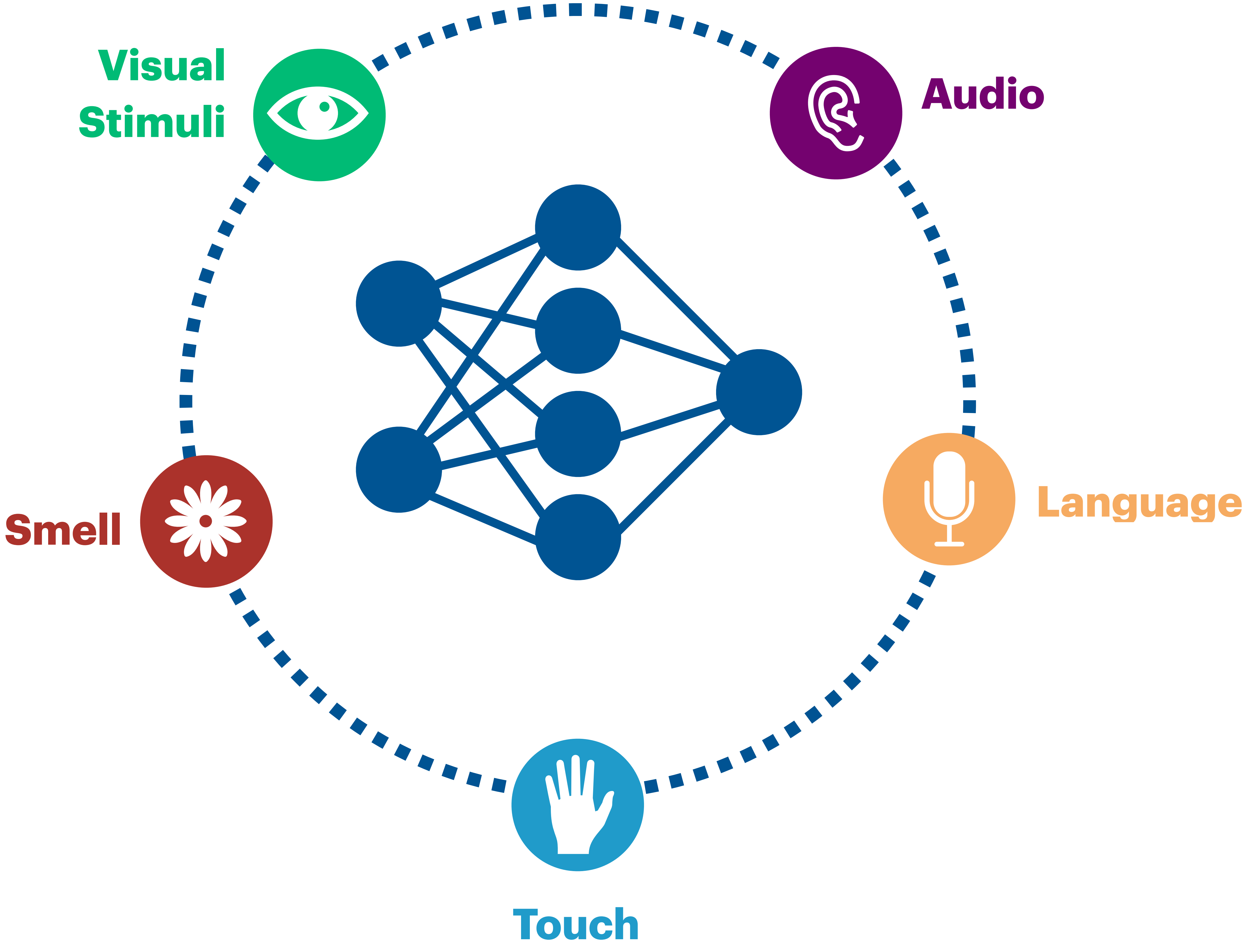
Why do we need VLMs?

The human experience of the world is multimodal.



Why do we need VLMs?

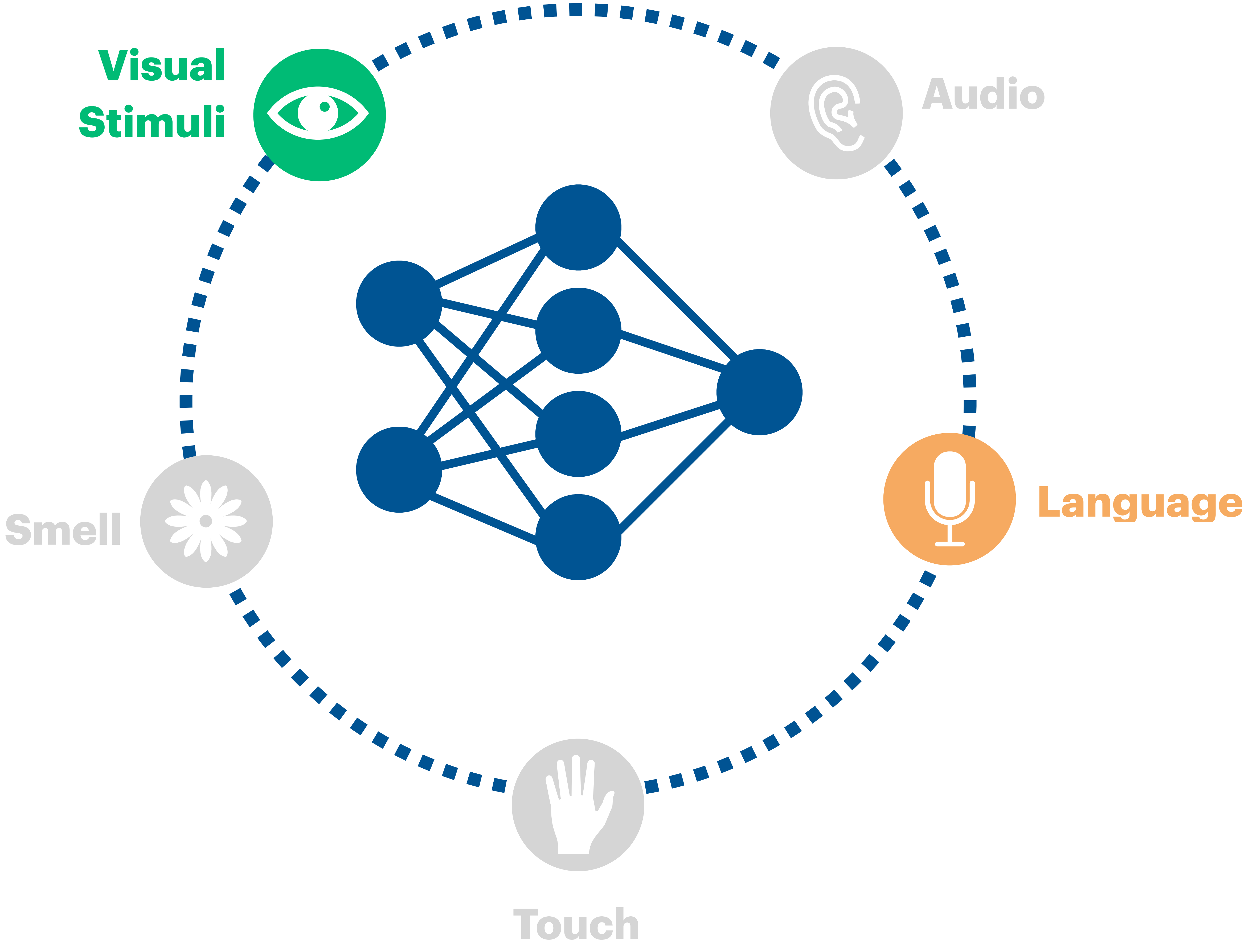
The human experience of the world is multimodal.



We need AI systems capable of simultaneously processing diverse input modalities.

Why do we need VLMs?

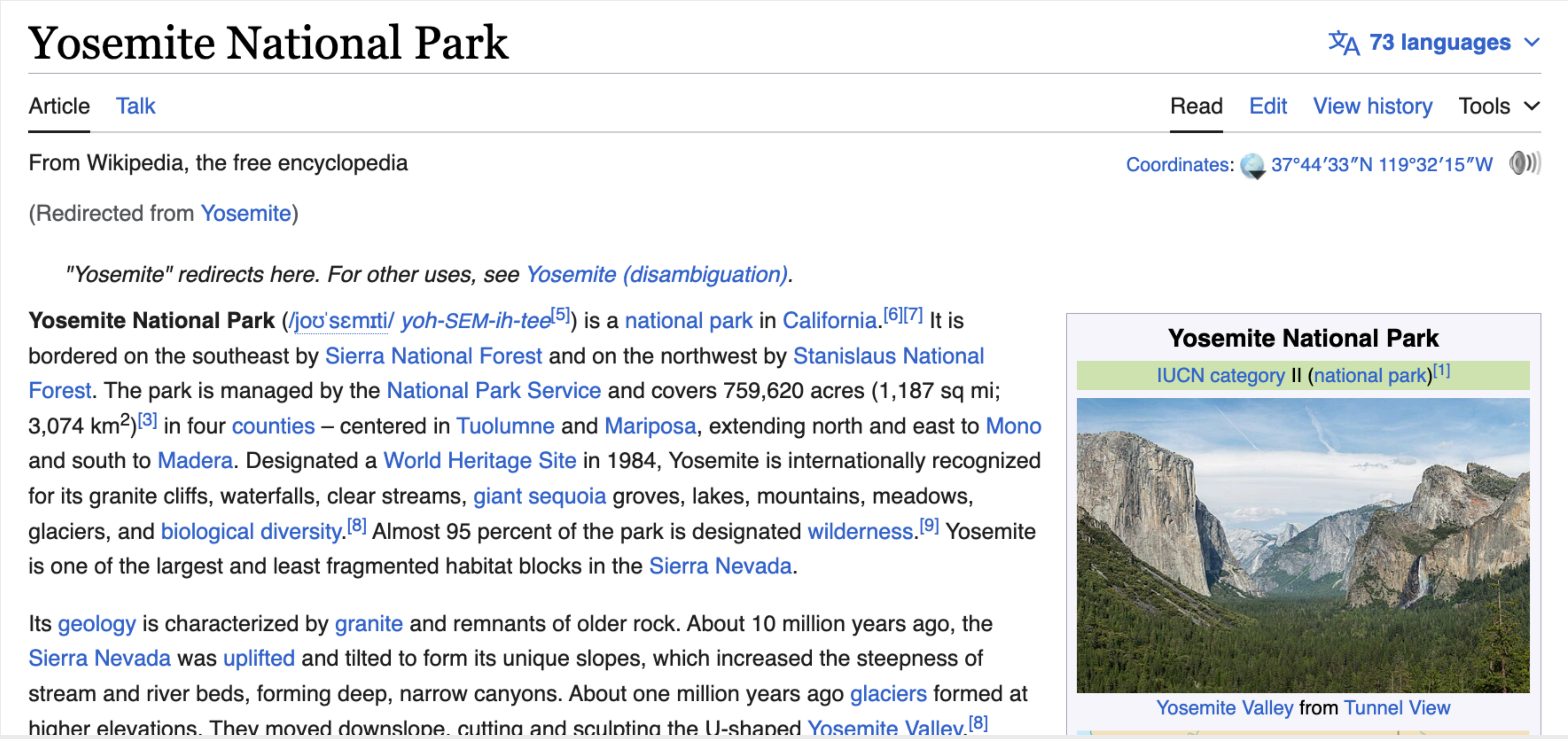
The human experience of the world is multimodal.



We need AI systems capable of simultaneously processing diverse input modalities.

Part 1: Pretraining Methods

Data is often inherently multimodal



Yosemite National Park

From Wikipedia, the free encyclopedia

Coordinates: 37°44′33″N 119°32′15″W

(Redirected from Yosemite)

"Yosemite" redirects here. For other uses, see Yosemite (disambiguation).

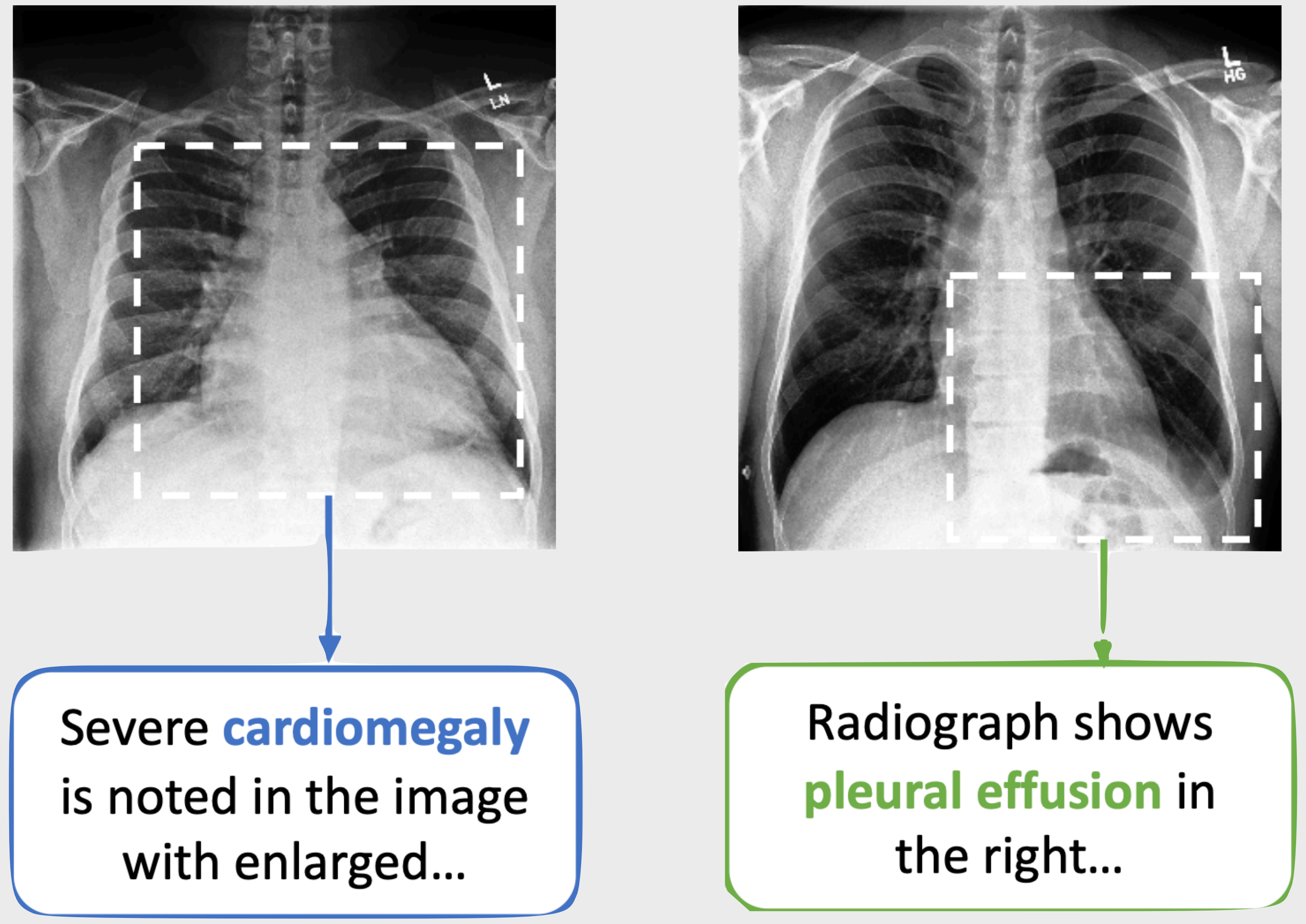
Yosemite National Park (/joʊˈsɛmɪti/ *yoh-SEM-ih-tee*[5]) is a national park in California.[6][7] It is bordered on the southeast by Sierra National Forest and on the northwest by Stanislaus National Forest. The park is managed by the National Park Service and covers 759,620 acres (1,187 sq mi; 3,074 km^2)[3] in four counties – centered in Tuolumne and Mariposa, extending north and east to Mono and south to Madera. Designated a World Heritage Site in 1984, Yosemite is internationally recognized for its granite cliffs, waterfalls, clear streams, giant sequoia groves, lakes, mountains, meadows, glaciers, and biological diversity.[8] Almost 95 percent of the park is designated wilderness.[9] Yosemite is one of the largest and least fragmented habitat blocks in the Sierra Nevada.

Its geology is characterized by granite and remnants of older rock. About 10 million years ago, the Sierra Nevada was uplifted and tilted to form its unique slopes, which increased the steepness of stream and river beds, forming deep, narrow canyons. About one million years ago glaciers formed at higher elevations. They moved downslope, cutting and sculpting the U-shaped Yosemite Valley.[8]

Yosemite National Park

IUCN category II (national park)[1]

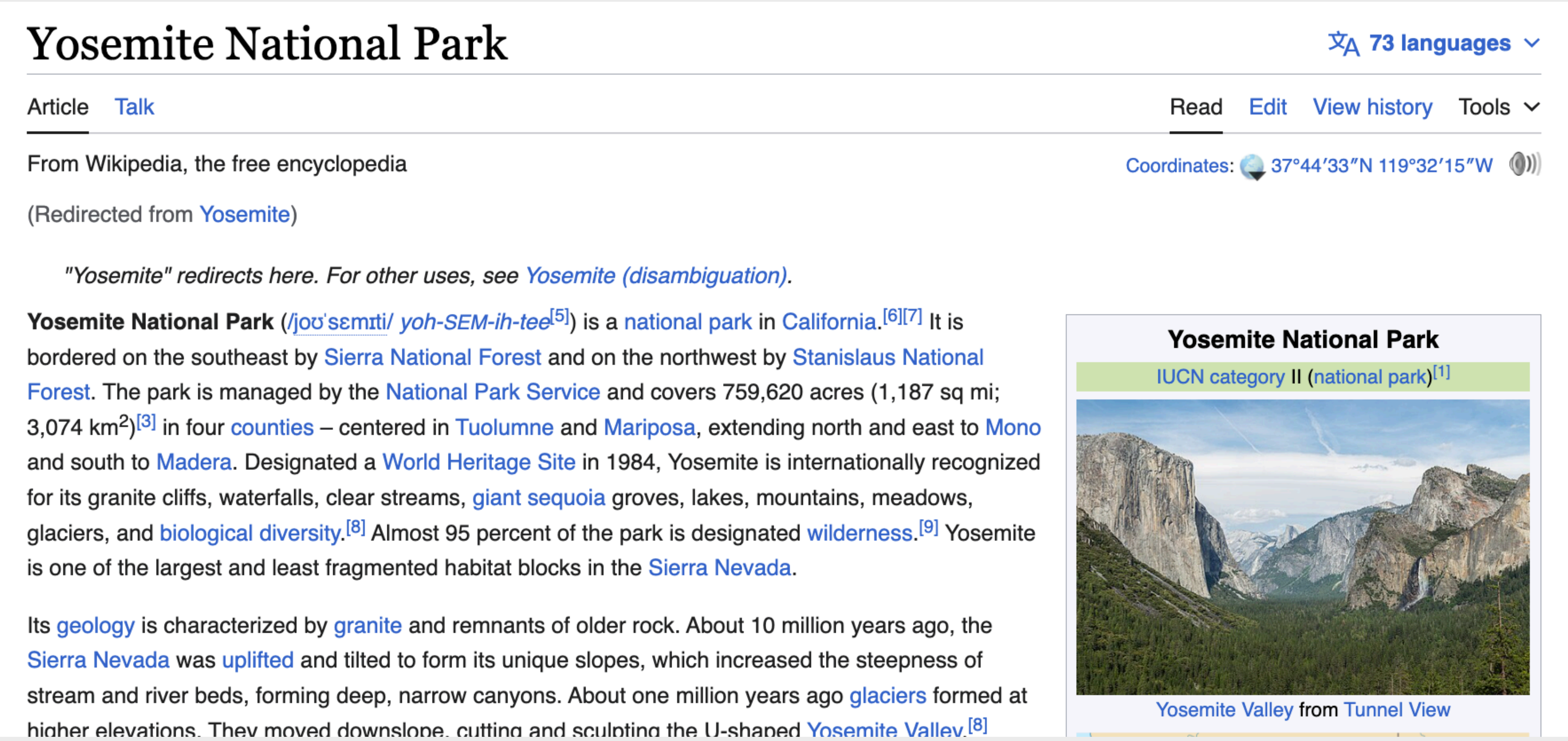
Yosemite Valley from Tunnel View



Severe **cardiomegaly** is noted in the image with enlarged...

Radiograph shows **pleural effusion** in the right...

Data is often inherently multimodal



Yosemite National Park

From Wikipedia, the free encyclopedia

Coordinates: 37°44′33″N 119°32′15″W

(Redirected from Yosemite)

"Yosemite" redirects here. For other uses, see Yosemite (disambiguation).

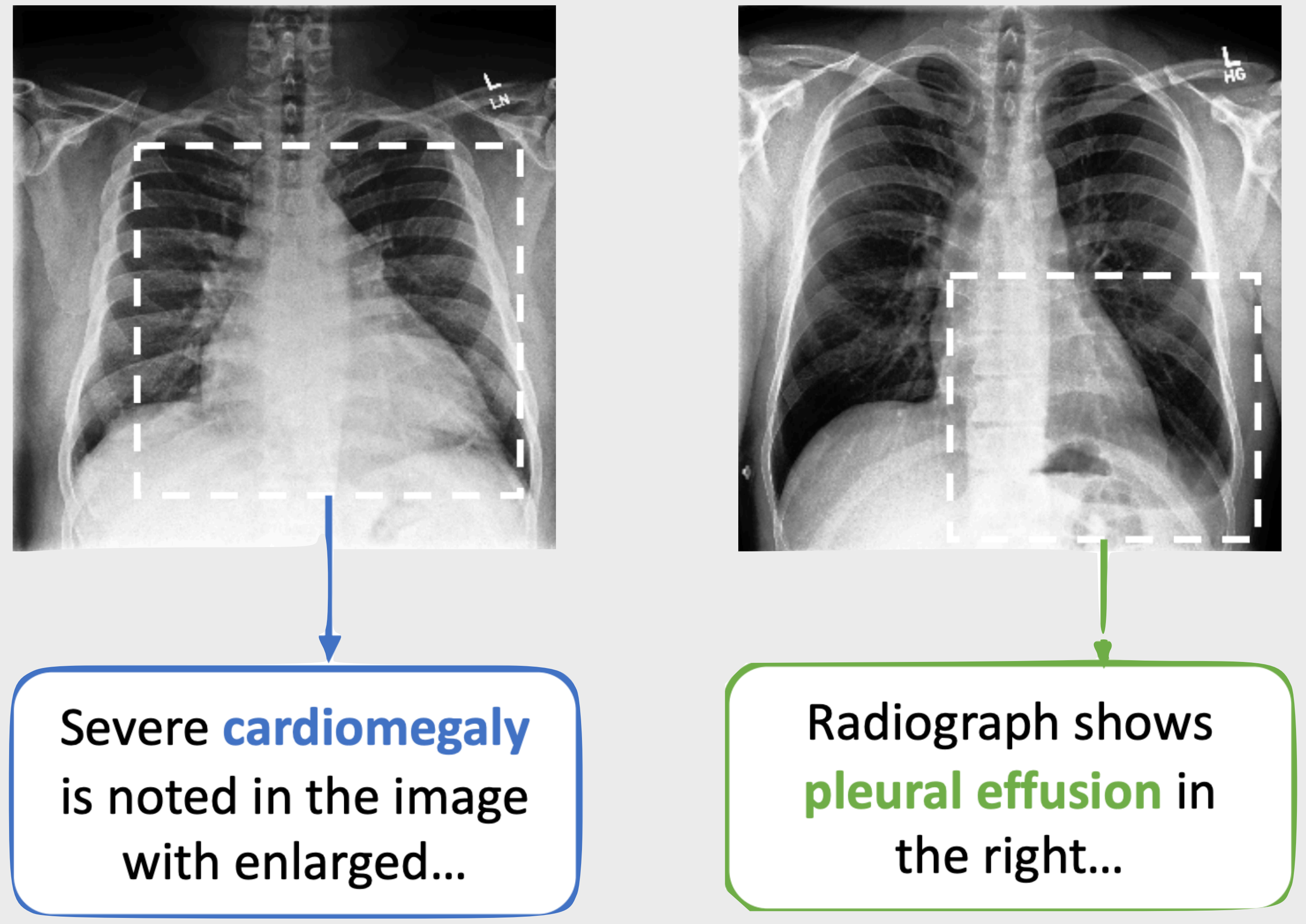
Yosemite National Park (/joʊˈsɛmɪti/ *yoh-SEM-ih-tee*[5]) is a national park in California.[6][7] It is bordered on the southeast by Sierra National Forest and on the northwest by Stanislaus National Forest. The park is managed by the National Park Service and covers 759,620 acres (1,187 sq mi; 3,074 km^2)[3] in four counties – centered in Tuolumne and Mariposa, extending north and east to Mono and south to Madera. Designated a World Heritage Site in 1984, Yosemite is internationally recognized for its granite cliffs, waterfalls, clear streams, giant sequoia groves, lakes, mountains, meadows, glaciers, and biological diversity.[8] Almost 95 percent of the park is designated wilderness.[9] Yosemite is one of the largest and least fragmented habitat blocks in the Sierra Nevada.

Its geology is characterized by granite and remnants of older rock. About 10 million years ago, the Sierra Nevada was uplifted and tilted to form its unique slopes, which increased the steepness of stream and river beds, forming deep, narrow canyons. About one million years ago glaciers formed at higher elevations. They moved downslope, cutting and sculpting the U-shaped Yosemite Valley.[8]

Yosemite National Park

IUCN category II (national park)[1]

Yosemite Valley from Tunnel View



Severe **cardiomegaly** is noted in the image with enlarged...

Radiograph shows **pleural effusion** in the right...

Can we use language to improve visual representation learning?

Pros

- Text is widely available
- Text can provide a form of supervision signal. No need for labels!

Cons

- Text may not always be available
- Text quality may be highly variable

Contrastive Language-Image Pretraining (CLIP)

Key Idea: Maximize the similarity between true image-text embedding pairs and minimize similarity between mismatched image-text embedding pairs

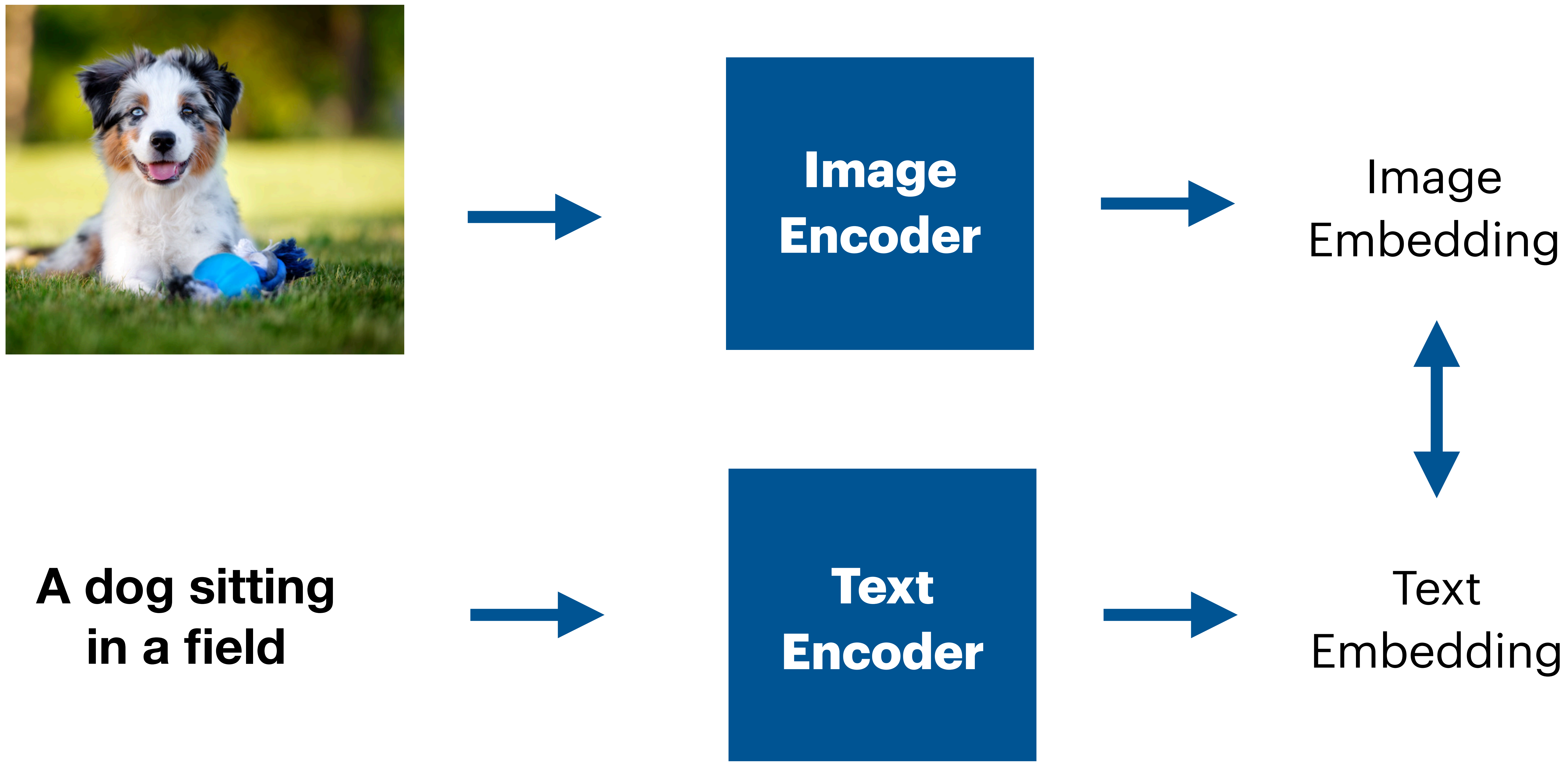


Image Encoder

Image Embedding

A dog sitting in a field

Text Encoder

Text Embedding

Radford et al. "Learning Transferable Visual Models From Natural Language Supervision"

Contrastive Language-Image Pretraining (CLIP)

Key Idea: Maximize the similarity between true image-text embedding pairs and minimize similarity between mismatched image-text embedding pairs

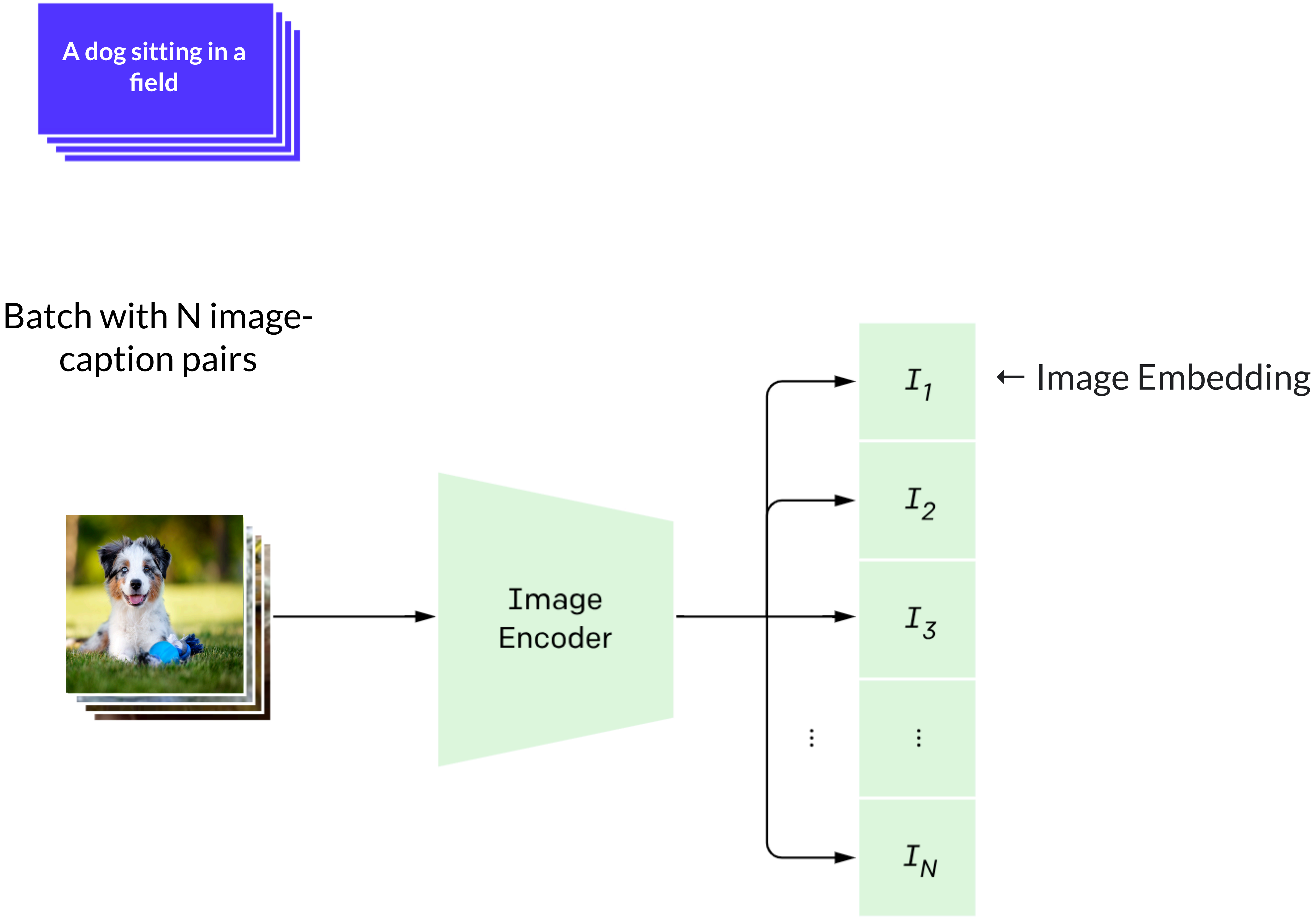
A dog sitting in a field

Batch with N image-caption pairs

Radford et al. "Learning Transferable Visual Models From Natural Language Supervision”

Contrastive Language-Image Pretraining (CLIP)

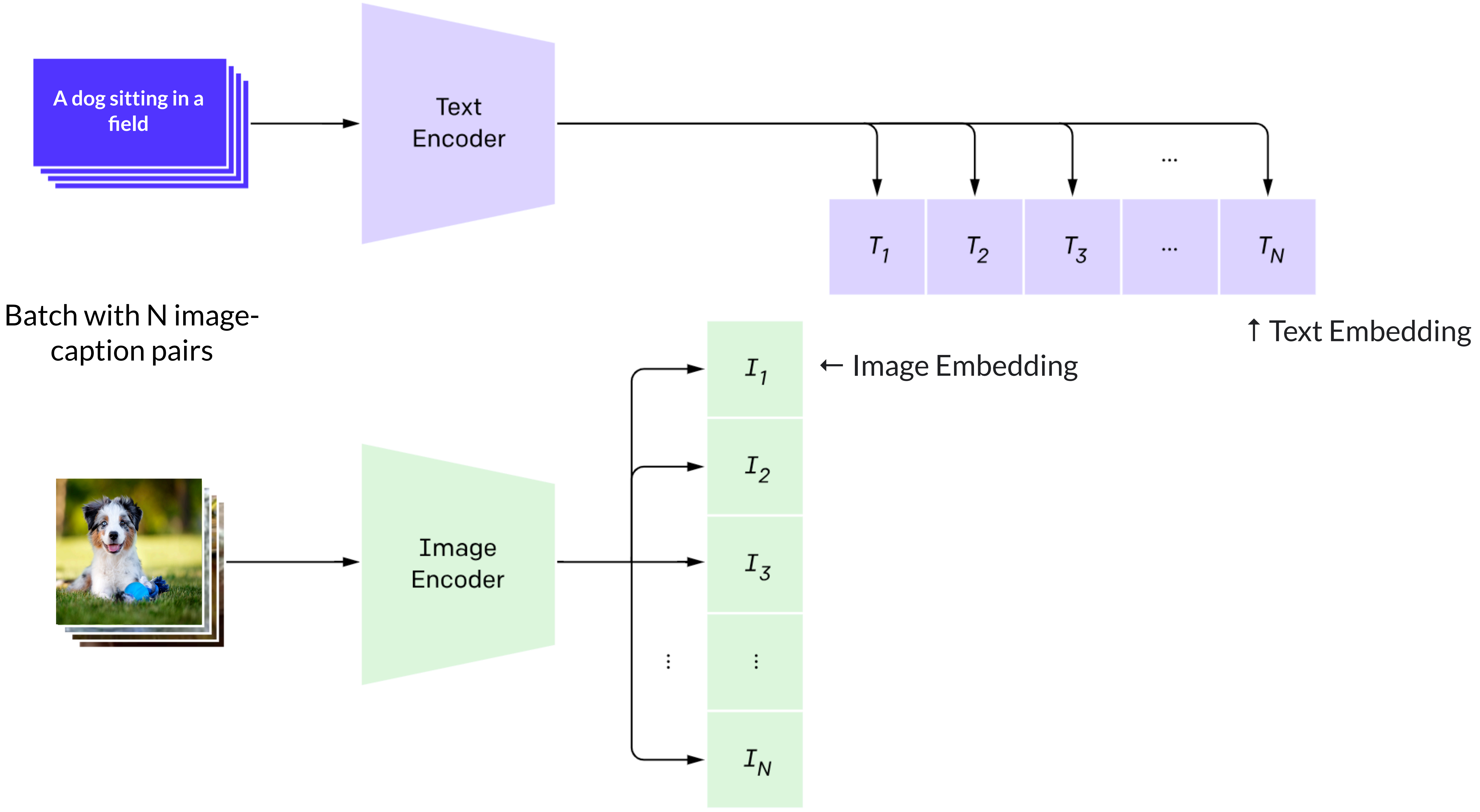
Key Idea: Maximize the similarity between true image-text embedding pairs and minimize similarity between mismatched image-text embedding pairs



Radford et al. "Learning Transferable Visual Models From Natural Language Supervision”

Contrastive Language-Image Pretraining (CLIP)

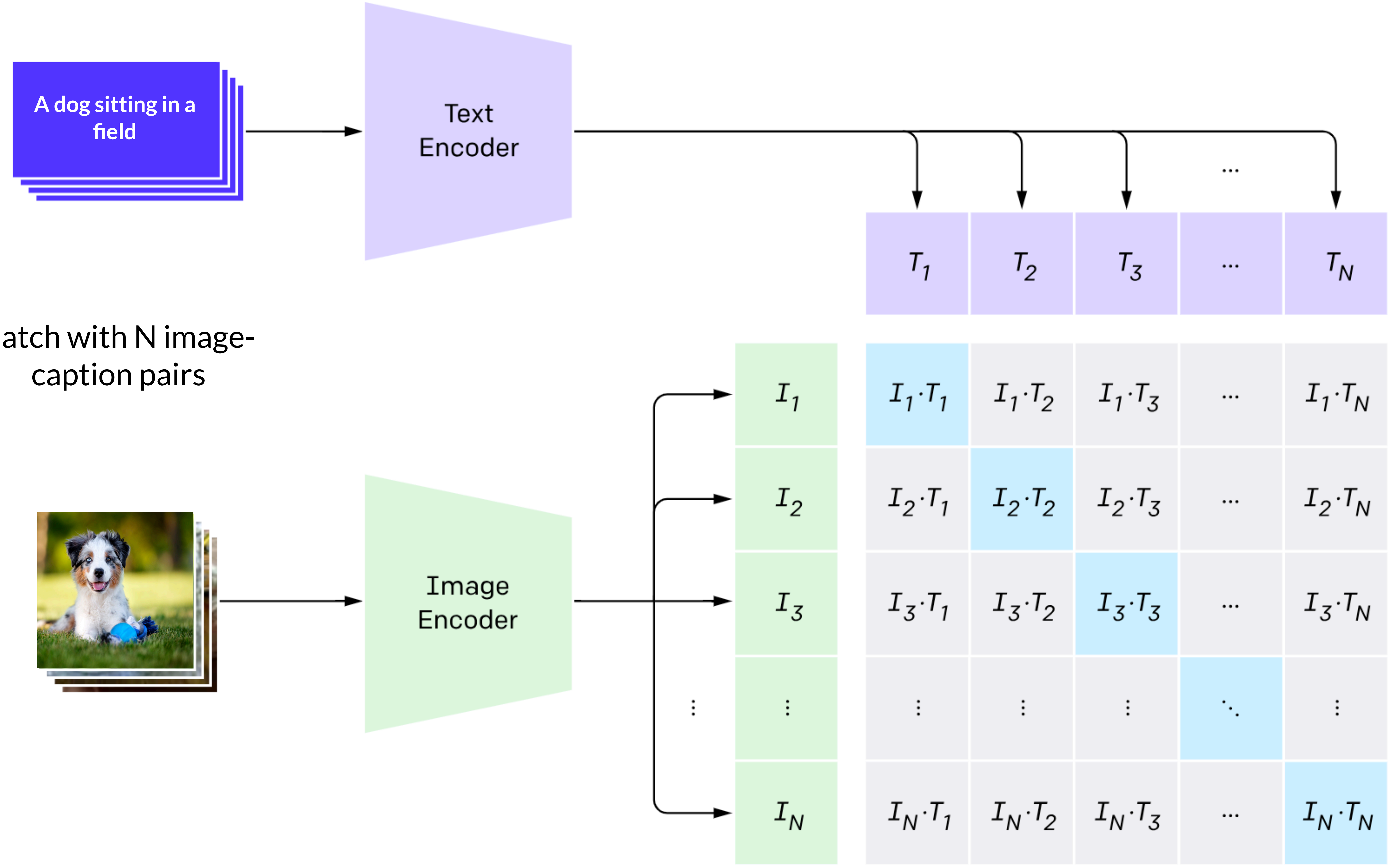
Key Idea: Maximize the similarity between true image-text embedding pairs and minimize similarity between mismatched image-text embedding pairs



Radford et al. "Learning Transferable Visual Models From Natural Language Supervision"

Contrastive Language-Image Pretraining (CLIP)

Key Idea: Maximize the similarity between true image-text embedding pairs and minimize similarity between mismatched image-text embedding pairs



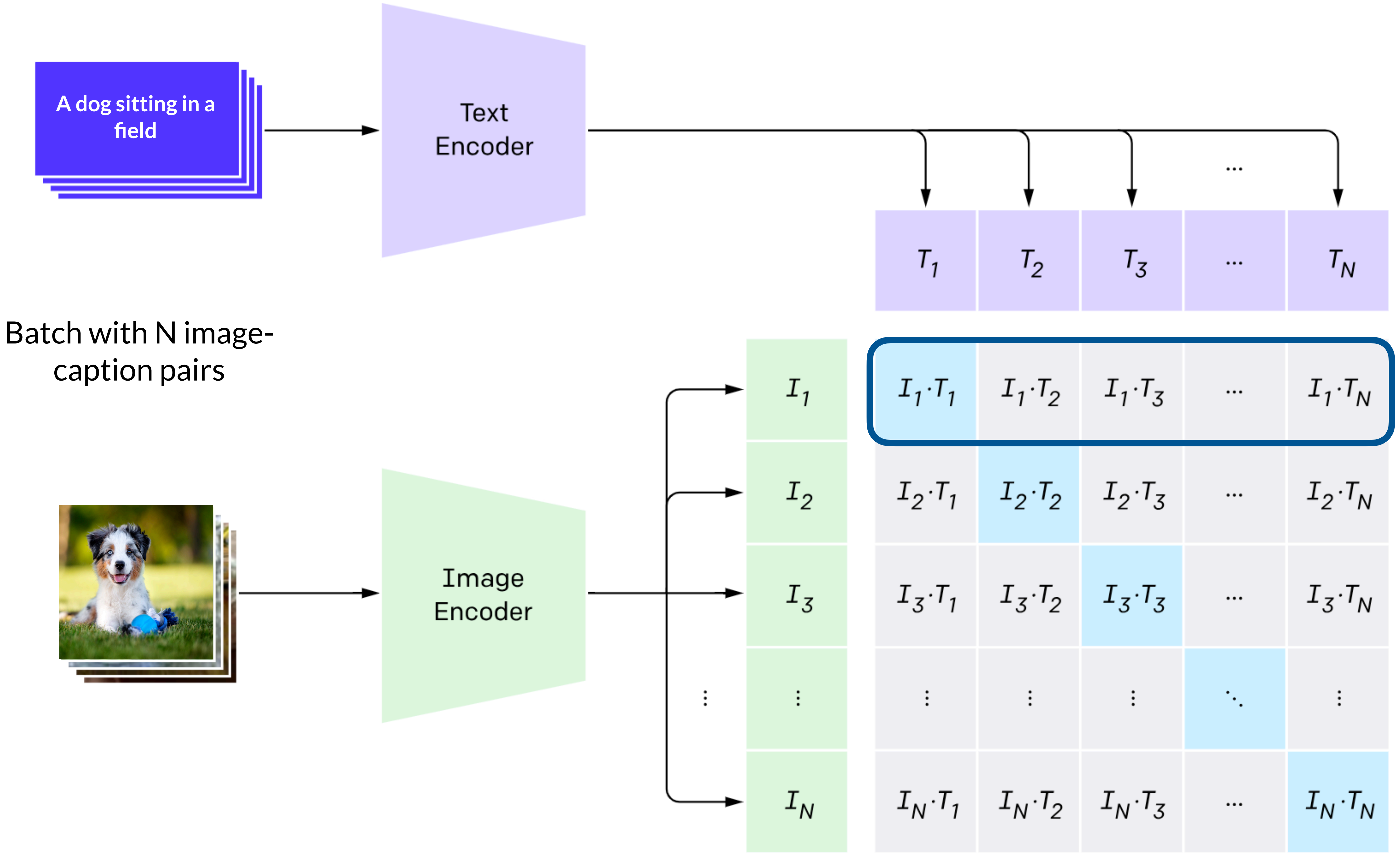
Batch with N image-caption pairs

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

Radford et al. "Learning Transferable Visual Models From Natural Language Supervision"

Contrastive Language-Image Pretraining (CLIP)

Key Idea: Maximize the similarity between true image-text embedding pairs and minimize similarity between mismatched image-text embedding pairs



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

Objective: InfoNCE Loss Function

$$L_{I \to T} = \sum_{k=1}^{N} - \log \frac{exp(I_k \cdot T_k / \tau)}{\sum_{j=1}^{N} exp(I_k \cdot T_j / \tau)}$$

Positive Image-Text Pairs

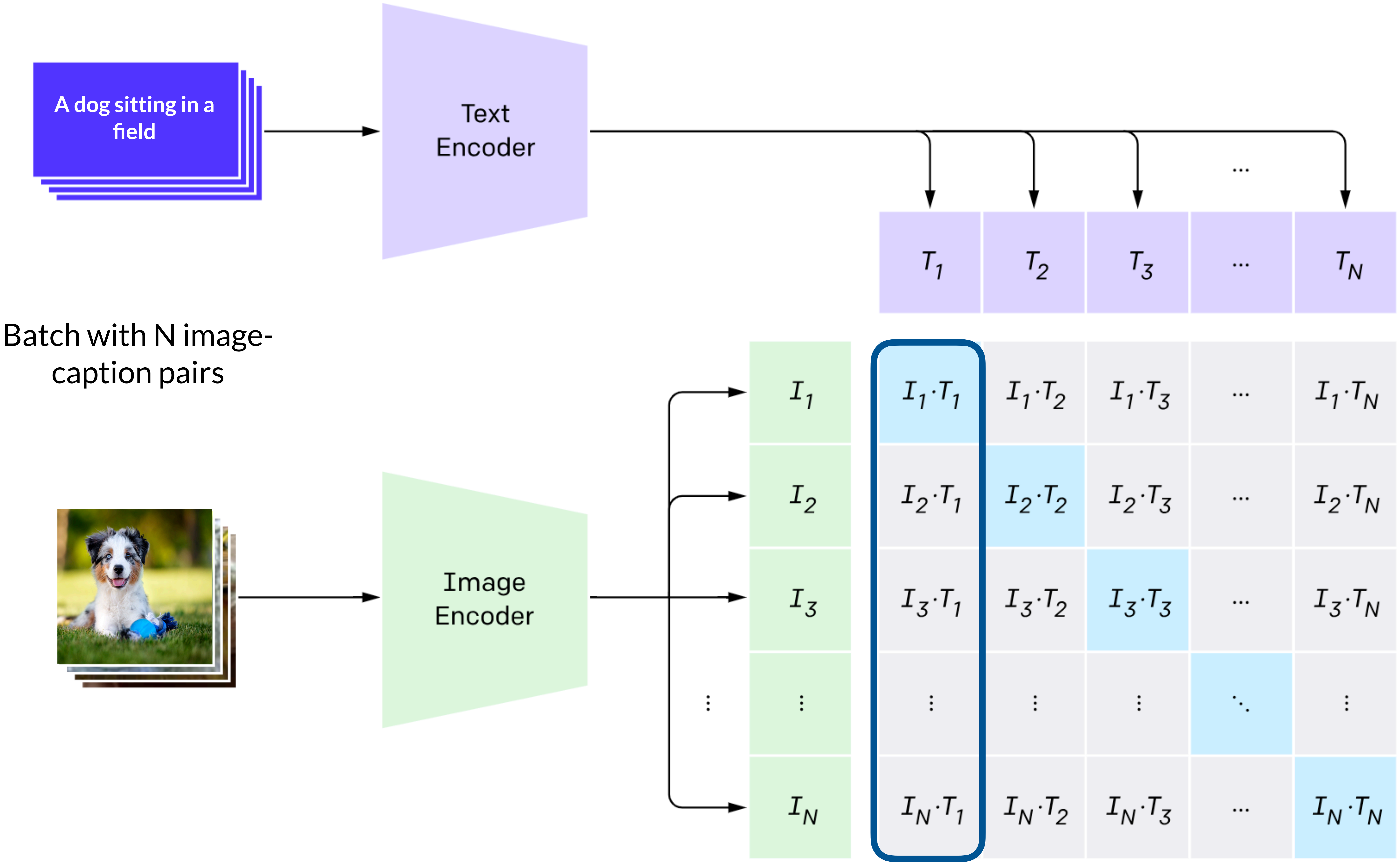
Negative Image-Text Pairs

Softmax Function

Radford et al. "Learning Transferable Visual Models From Natural Language Supervision"

Contrastive Language-Image Pretraining (CLIP)

Key Idea: Maximize the similarity between true image-text embedding pairs and minimize similarity between mismatched image-text embedding pairs



Batch with N image-caption pairs

$$\begin{array}{c|ccccc} & T_1 & T_2 & T_3 & \cdots & T_N \\ \hline I_1 & I_1\cdot T_1 & I_1\cdot T_2 & I_1\cdot T_3 & \cdots & I_1\cdot T_N \\ I_2 & I_2\cdot T_1 & I_2\cdot T_2 & I_2\cdot T_3 & \cdots & I_2\cdot T_N \\ I_3 & I_3\cdot T_1 & I_3\cdot T_2 & I_3\cdot T_3 & \cdots & I_3\cdot T_N \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ I_N & I_N\cdot T_1 & I_N\cdot T_2 & I_N\cdot T_3 & \cdots & I_N\cdot T_N \end{array}$$

Objective: InfoNCE Loss Function

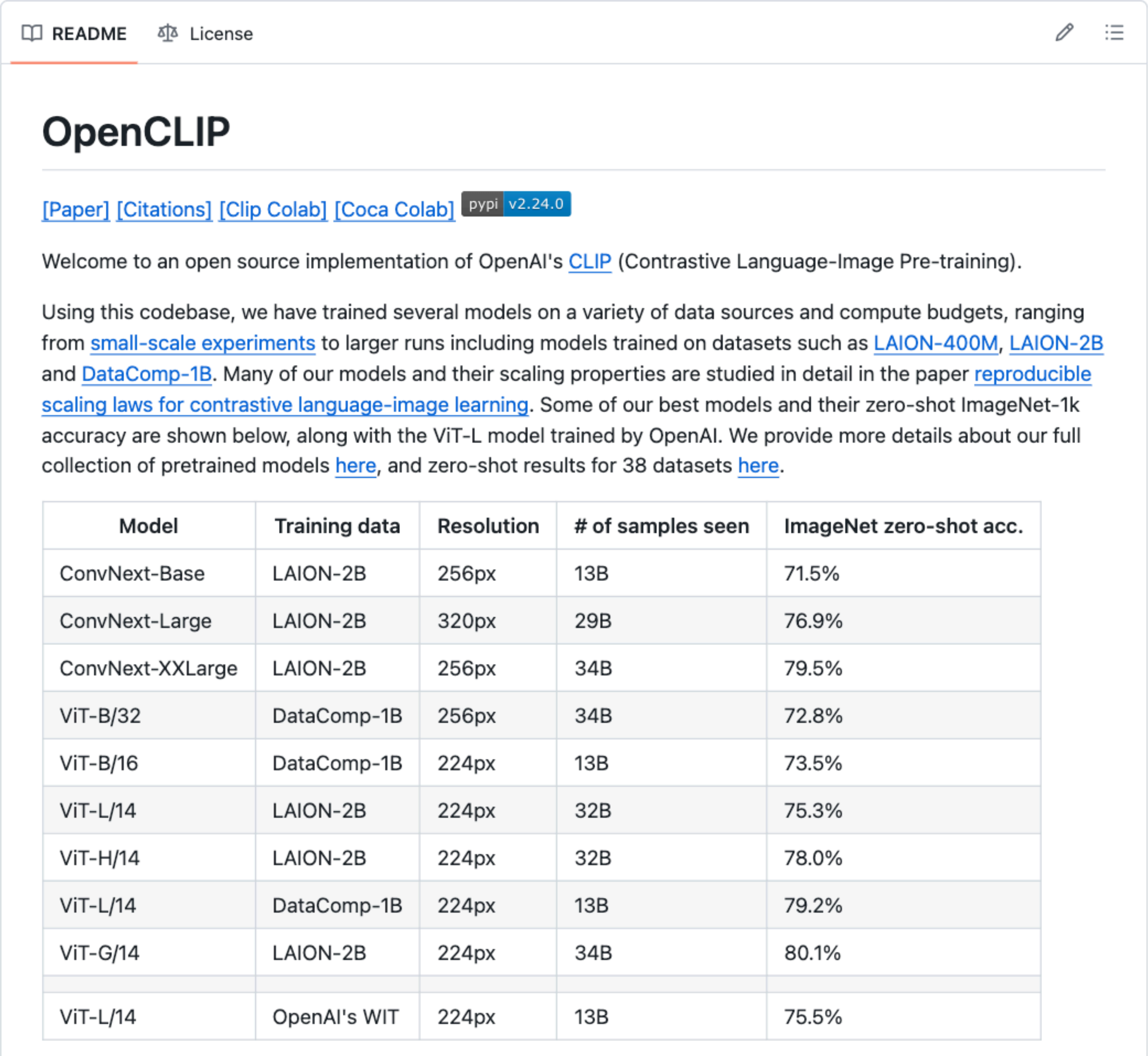
$$L_{I\to T} = \sum_{k=1}^{N} -\log \frac{exp(I_k \cdot T_k/\tau)}{\sum_{j=1}^{N} exp(I_k \cdot T_j/\tau)}$$

$$L_{T\to I} = \sum_{k=1}^{N} -\log \frac{exp(I_k \cdot T_k/\tau)}{\sum_{j=1}^{N} exp(I_j \cdot T_k/\tau)}$$

$$L = L_{T\to I} + L_{I\to T}$$

Radford et al. "Learning Transferable Visual Models From Natural Language Supervision"

OpenCLIP



OpenCLIP

[Paper] [Citations] [Clip Colab] [Coca Colab] pypi v2.24.0

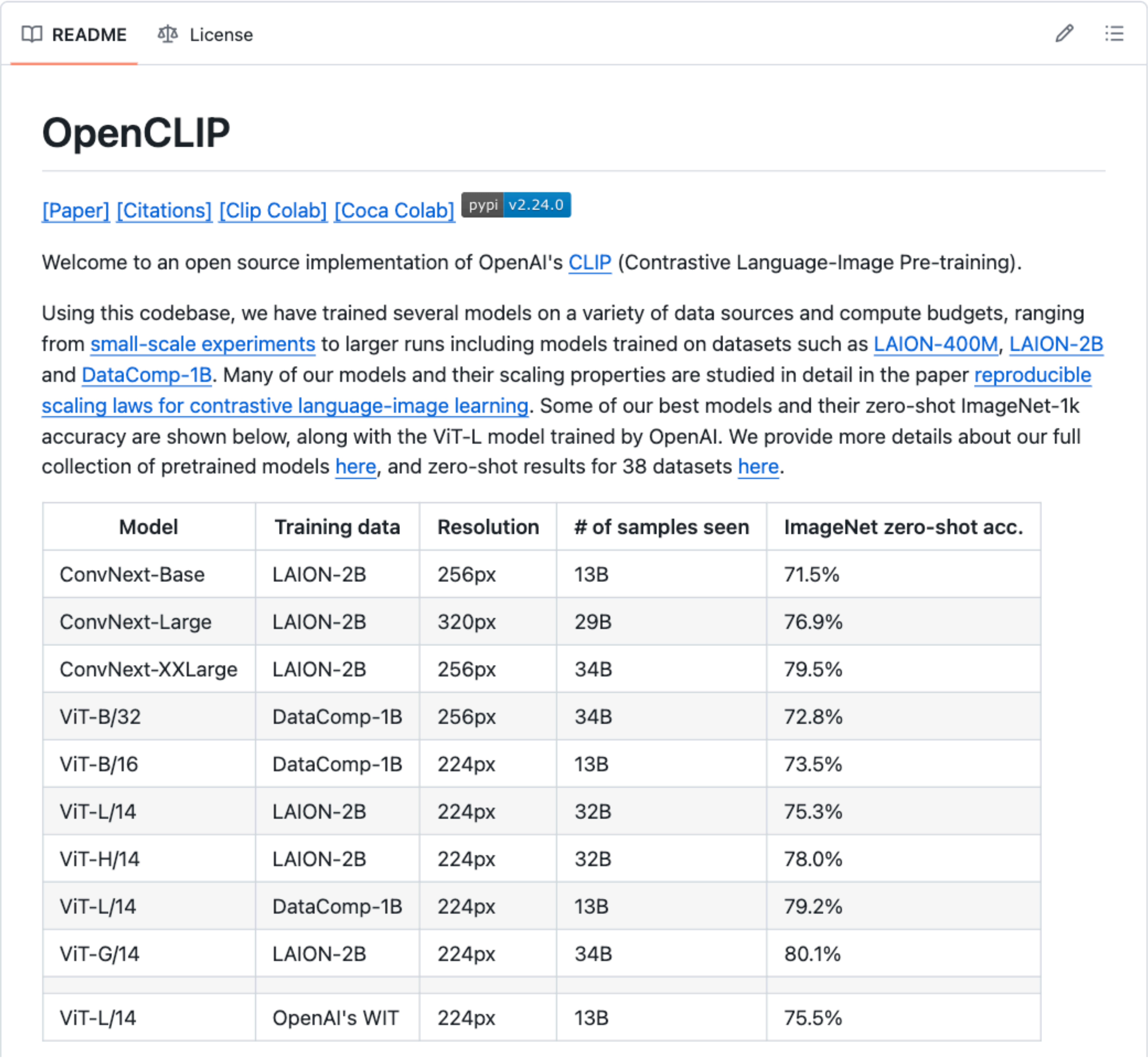
Welcome to an open source implementation of OpenAI's CLIP (Contrastive Language-Image Pre-training).

Using this codebase, we have trained several models on a variety of data sources and compute budgets, ranging from small-scale experiments to larger runs including models trained on datasets such as LAION-400M, LAION-2B and DataComp-1B. Many of our models and their scaling properties are studied in detail in the paper reproducible scaling laws for contrastive language-image learning. Some of our best models and their zero-shot ImageNet-1k accuracy are shown below, along with the ViT-L model trained by OpenAI. We provide more details about our full collection of pretrained models here, and zero-shot results for 38 datasets here.

Model	Training data	Resolution	# of samples seen	ImageNet zero-shot acc.
ConvNext-Base	LAION-2B	256px	13B	71.5%
ConvNext-Large	LAION-2B	320px	29B	76.9%
ConvNext-XXLarge	LAION-2B	256px	34B	79.5%
ViT-B/32	DataComp-1B	256px	34B	72.8%
ViT-B/16	DataComp-1B	224px	13B	73.5%
ViT-L/14	LAION-2B	224px	32B	75.3%
ViT-H/14	LAION-2B	224px	32B	78.0%
ViT-L/14	DataComp-1B	224px	13B	79.2%
ViT-G/14	LAION-2B	224px	34B	80.1%
ViT-L/14	OpenAI's WIT	224px	13B	75.5%

Ilharco et al. “OpenCLIP”
Cherti et al. “Reproducible scaling laws for contrastive language-image learning”

OpenCLIP



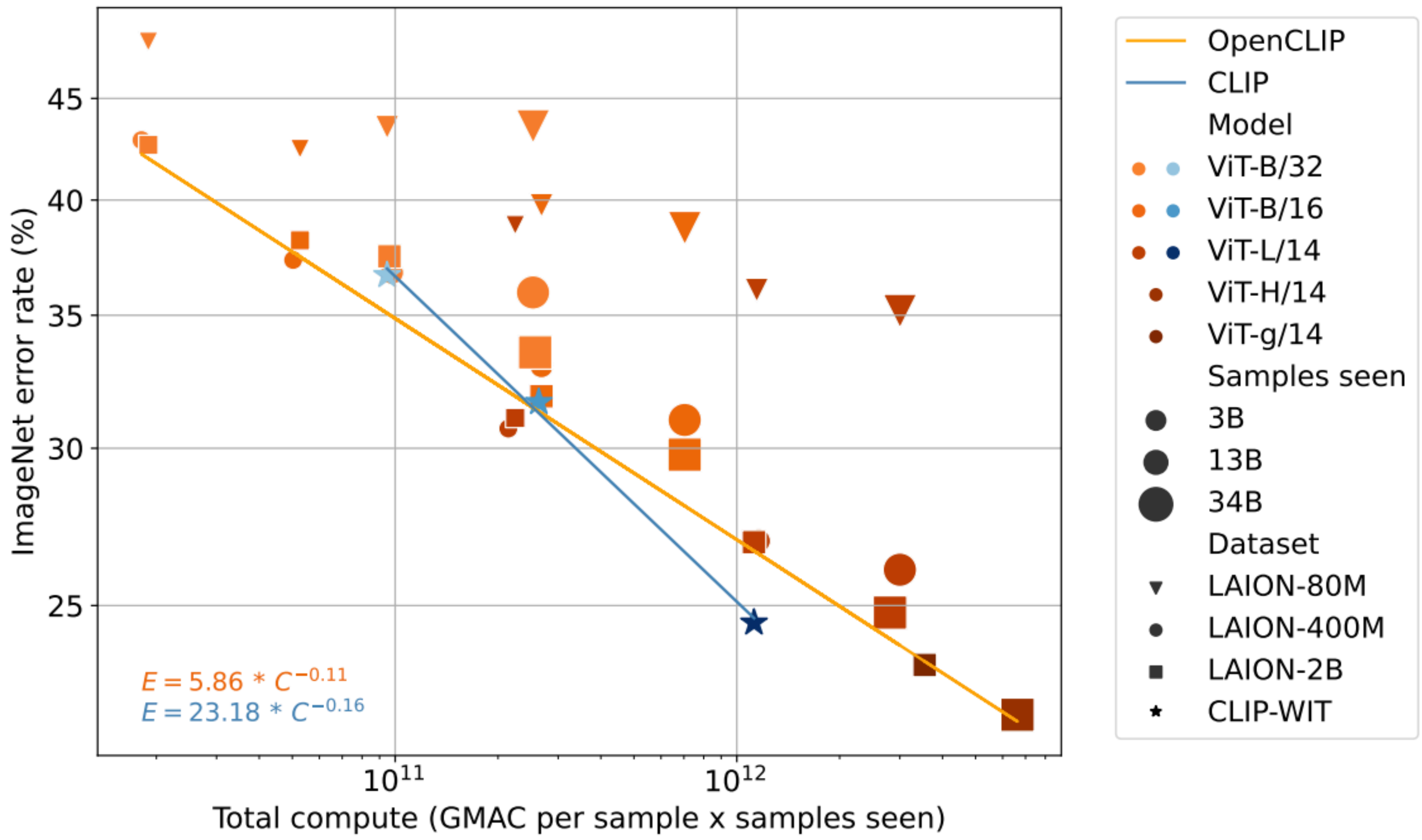
OpenCLIP

[Paper] [Citations] [Clip Colab] [Coca Colab] pypi v2.24.0

Welcome to an open source implementation of OpenAI's CLIP (Contrastive Language-Image Pre-training).

Using this codebase, we have trained several models on a variety of data sources and compute budgets, ranging from small-scale experiments to larger runs including models trained on datasets such as LAION-400M, LAION-2B and DataComp-1B. Many of our models and their scaling properties are studied in detail in the paper reproducible scaling laws for contrastive language-image learning. Some of our best models and their zero-shot ImageNet-1k accuracy are shown below, along with the ViT-L model trained by OpenAI. We provide more details about our full collection of pretrained models here, and zero-shot results for 38 datasets here.

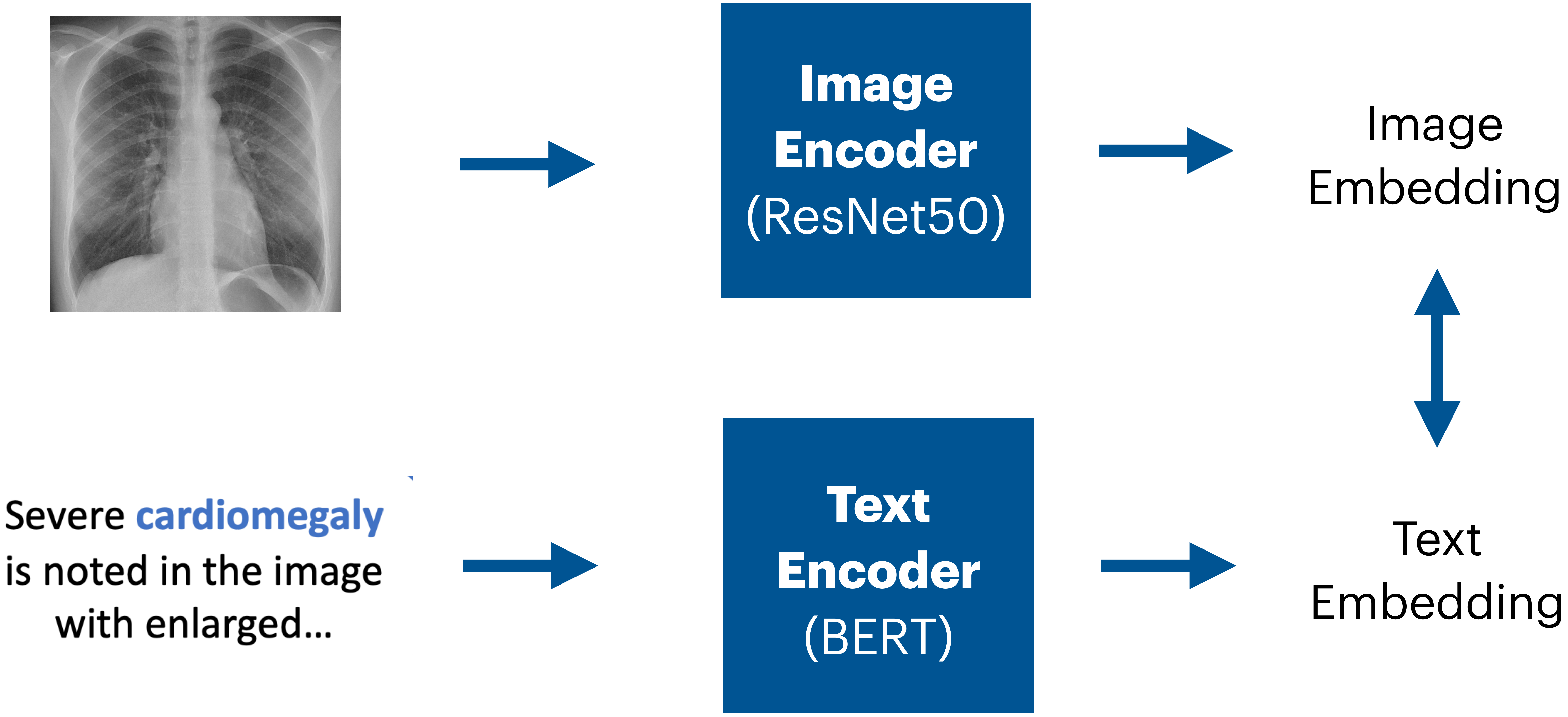
Model	Training data	Resolution	# of samples seen	ImageNet zero-shot acc.
ConvNext-Base	LAION-2B	256px	13B	71.5%
ConvNext-Large	LAION-2B	320px	29B	76.9%
ConvNext-XXLarge	LAION-2B	256px	34B	79.5%
ViT-B/32	DataComp-1B	256px	34B	72.8%
ViT-B/16	DataComp-1B	224px	13B	73.5%
ViT-L/14	LAION-2B	224px	32B	75.3%
ViT-H/14	LAION-2B	224px	32B	78.0%
ViT-L/14	DataComp-1B	224px	13B	79.2%
ViT-G/14	LAION-2B	224px	34B	80.1%
ViT-L/14	OpenAI's WIT	224px	13B	75.5%



Ilharco et al. "OpenCLIP"
Cherti et al. "Reproducible scaling laws for contrastive language-image learning"

ConVIRT

Key Idea: Maximize the similarity between true image-text embedding pairs and minimize similarity between mismatched image-text embedding pairs



Severe **cardiomegaly** is noted in the image with enlarged...

Image Encoder (ResNet50)

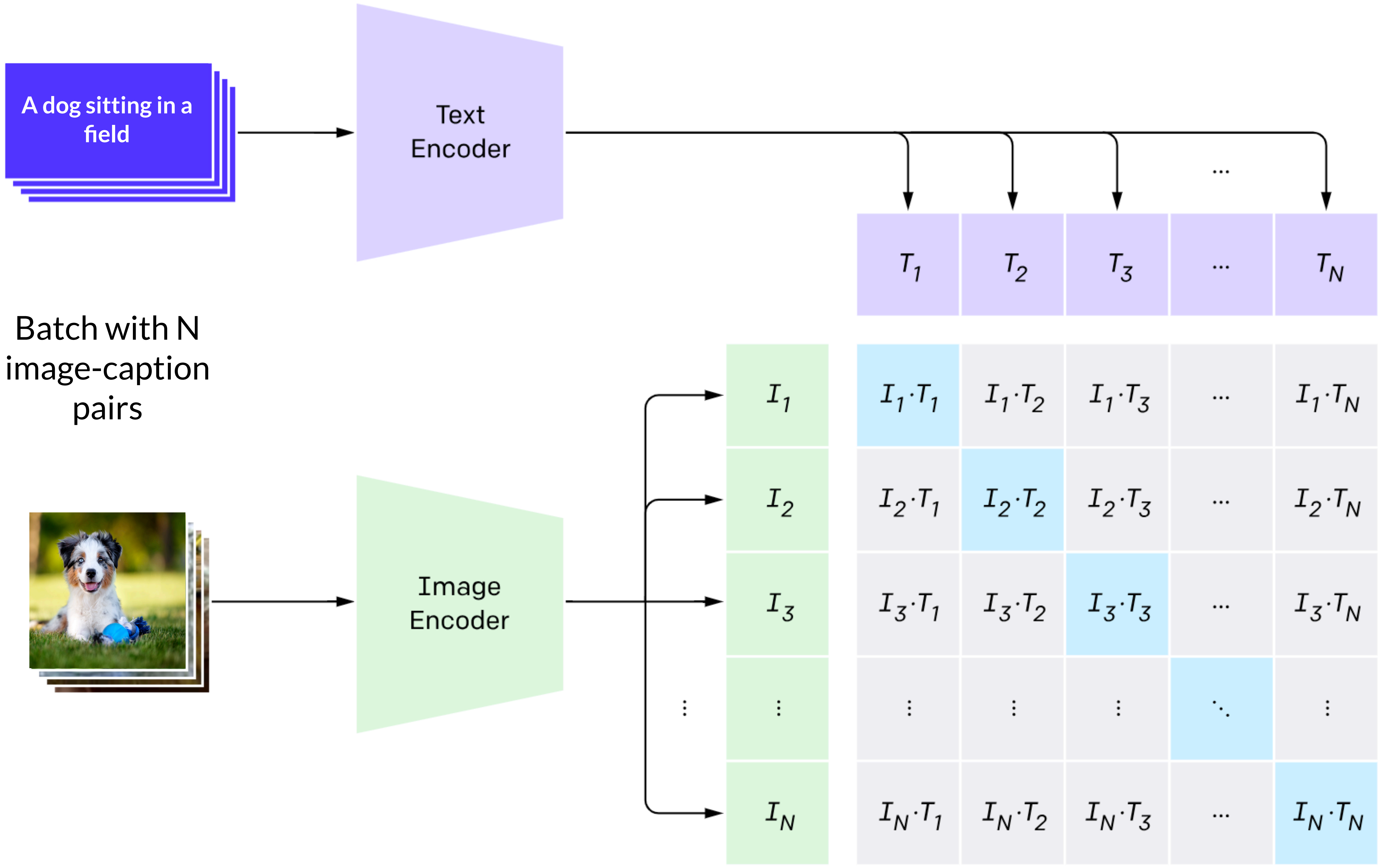
Image Embedding

Text Encoder (BERT)

Text Embedding

Zhang et al. "Contrastive Learning of Medical Visual Representations from Paired Images and Text”

Considerations for CLIP



	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

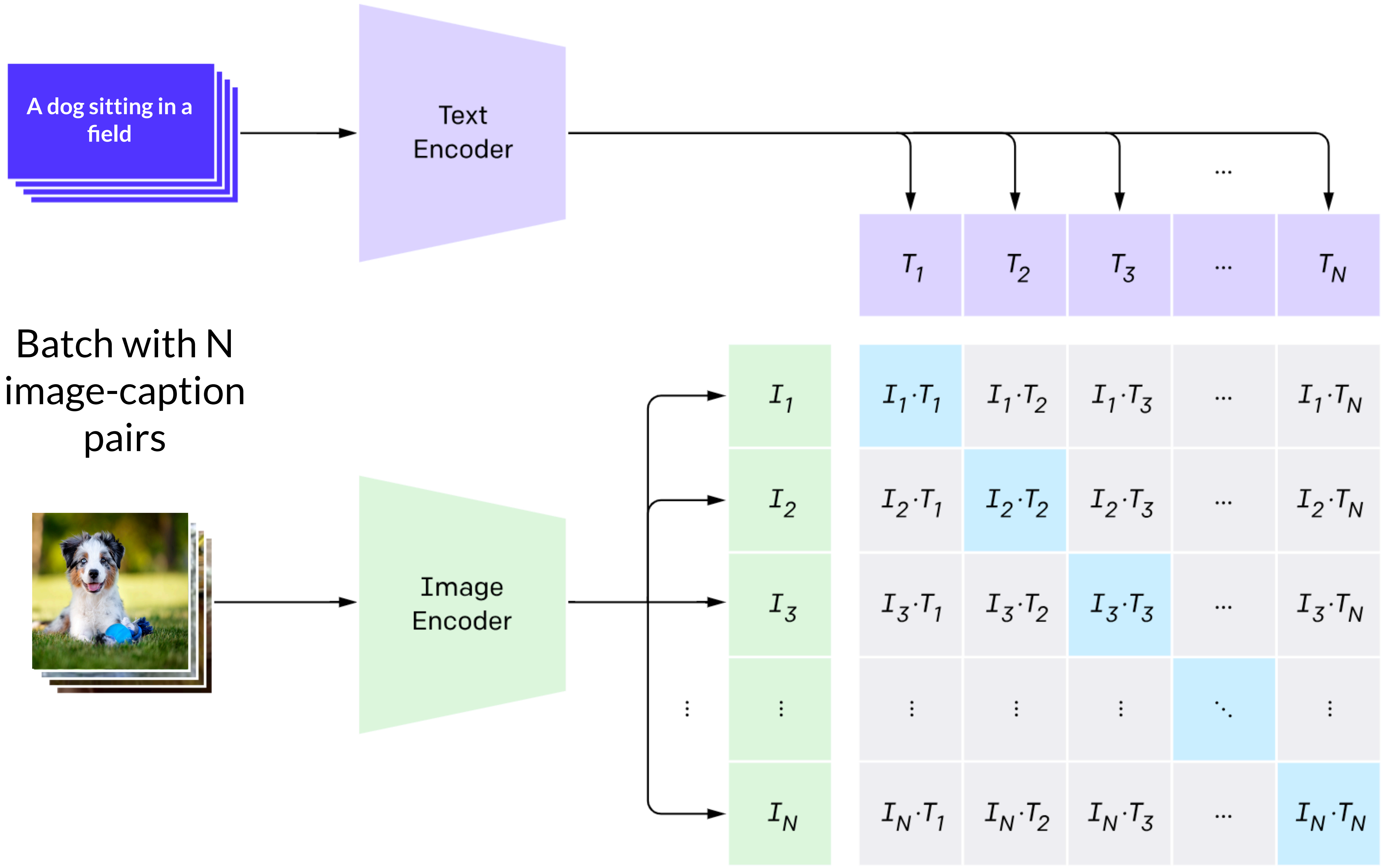
Q: What is the most important hyperparameter when training CLIP?

A: Batch Size!

CLIP uses a batch size of 32,768 (trained on up to 592 V100 GPUs)

Can we reduce the need for massive batch sizes?

SigLIP: Sigmoid Loss for Language Image Pre-Training



Batch with N image-caption pairs

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

CLIP Objective: InfoNCE Loss Function

$$L_{I \to T} = \sum_{k=1}^{N} -\log \frac{exp(I_k \cdot T_k/\tau)}{\sum_{j=1}^{N} exp(I_k \cdot T_j/\tau)}$$

Softmax Function

Positive Image-Text Pairs

Negative Image-Text Pairs

SigLIP Objective: Sigmoid Loss Function

$$-\frac{1}{|\mathcal{B}|}\sum_{k=1}^{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\underbrace{\log\frac{1}{1+e^{z_{ik}(-tI_k \cdot T_j+b)}}}_{\mathcal{L}_{ij}}$$

Sigmoid Function

z_{ik} = 1 for positive image-text pairs (i.e. k=j)

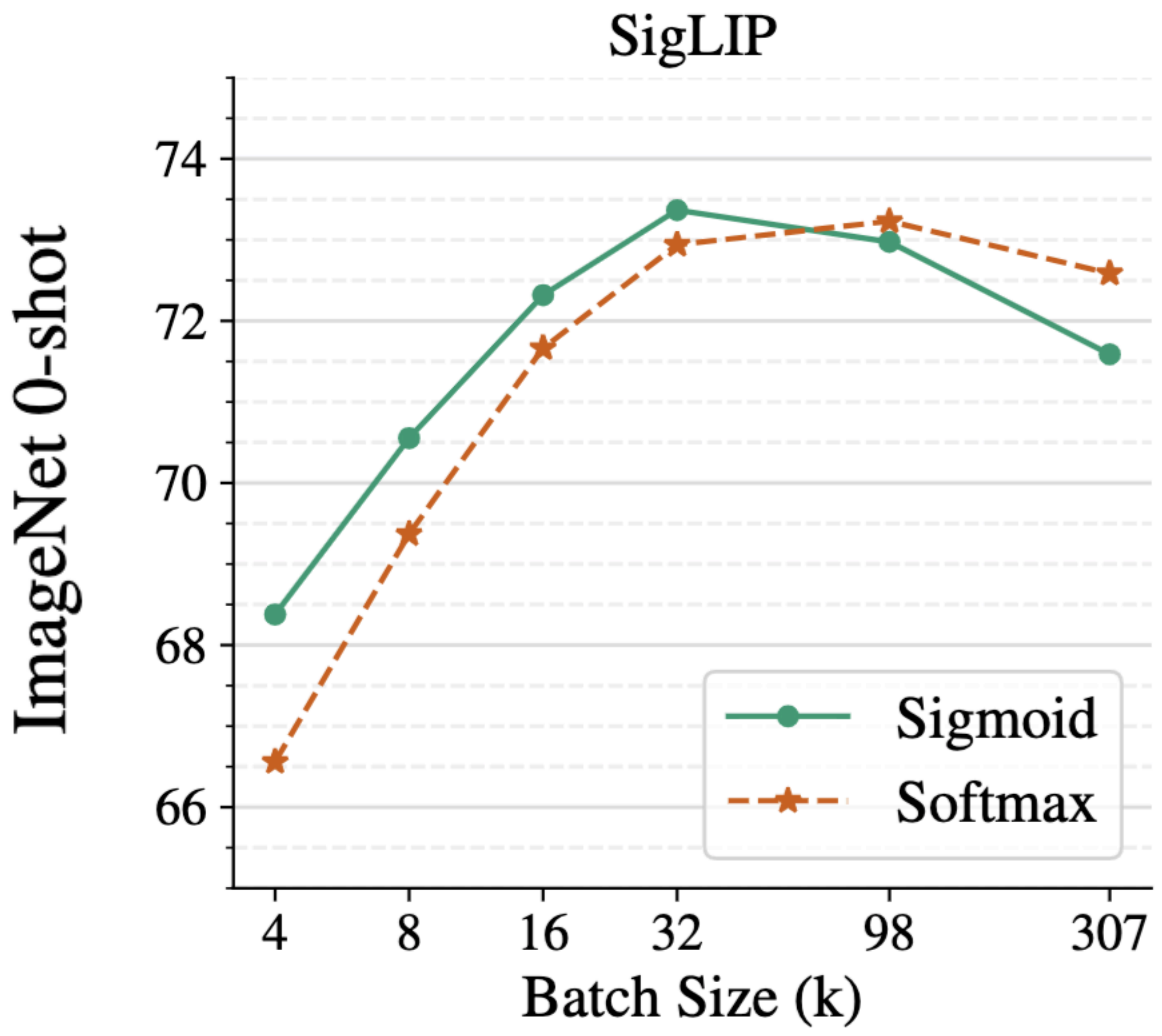
z_{ik} = -1 for negative image-text pairs (i.e. k!=j)

Zhai et al. Sigmoid Loss for Language Image Pre-Training. ICCV 2023.

SigLIP: Sigmoid Loss for Language Image Pre-Training

Advantages

- SigLIP is more memory-efficient than CLIP → avoids materializing a |B| x |B| matrix.
- SigLIP outperforms CLIP at smaller batch sizes



Part 2: Data

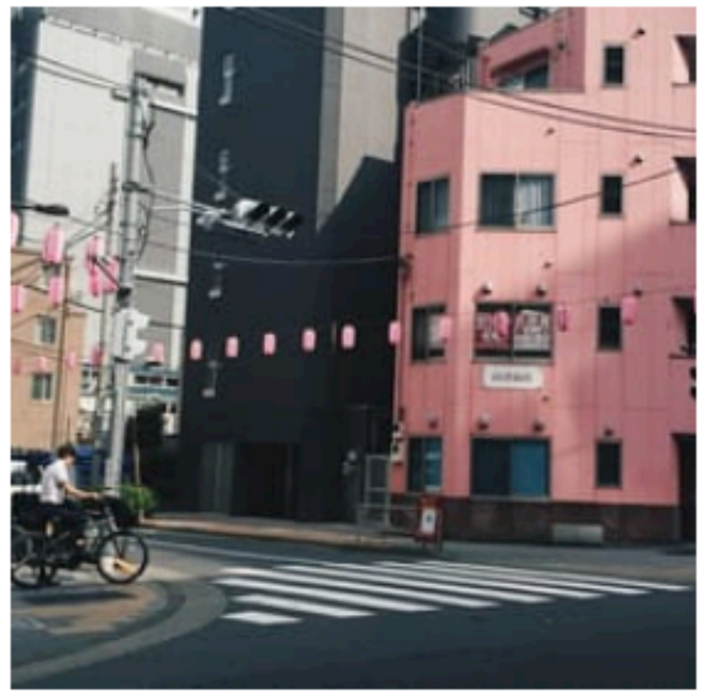
General-Domain Data: LAION-5B

LAION-5B contains 5 billion image-text pairs obtained from CommonCrawl

C: Green Apple Chair

C: sun snow dog

C: Color Palettes

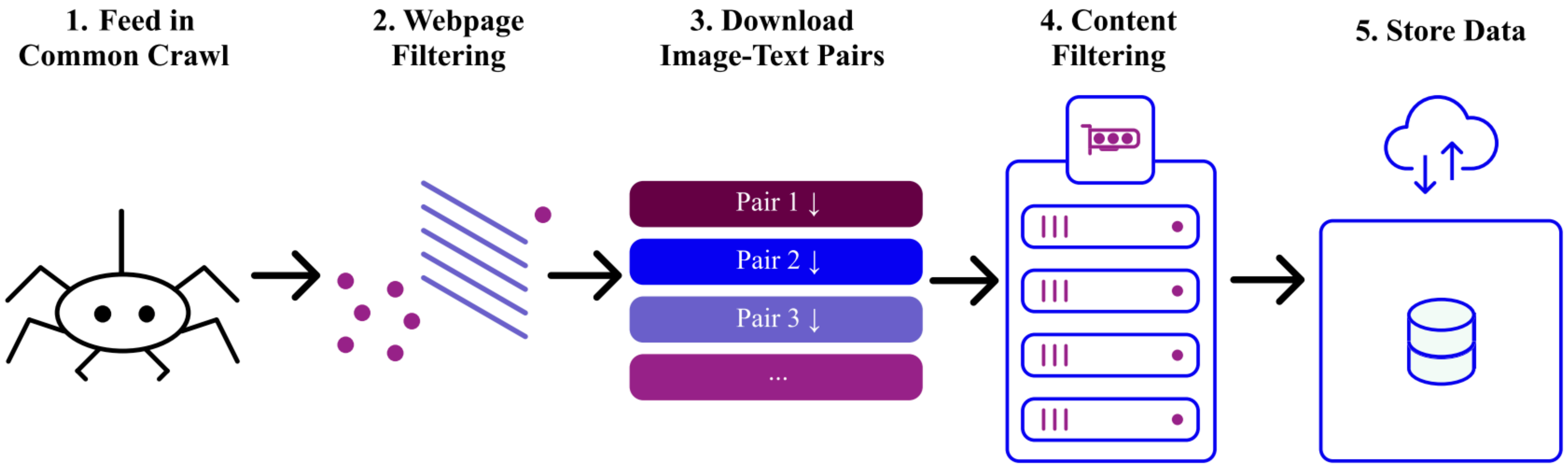


C: pink, japan, aesthetic image

Schumann et al. “LAION-5B: An open large-scale dataset for training next generation image-text models”

General-Domain Data: LAION-5B

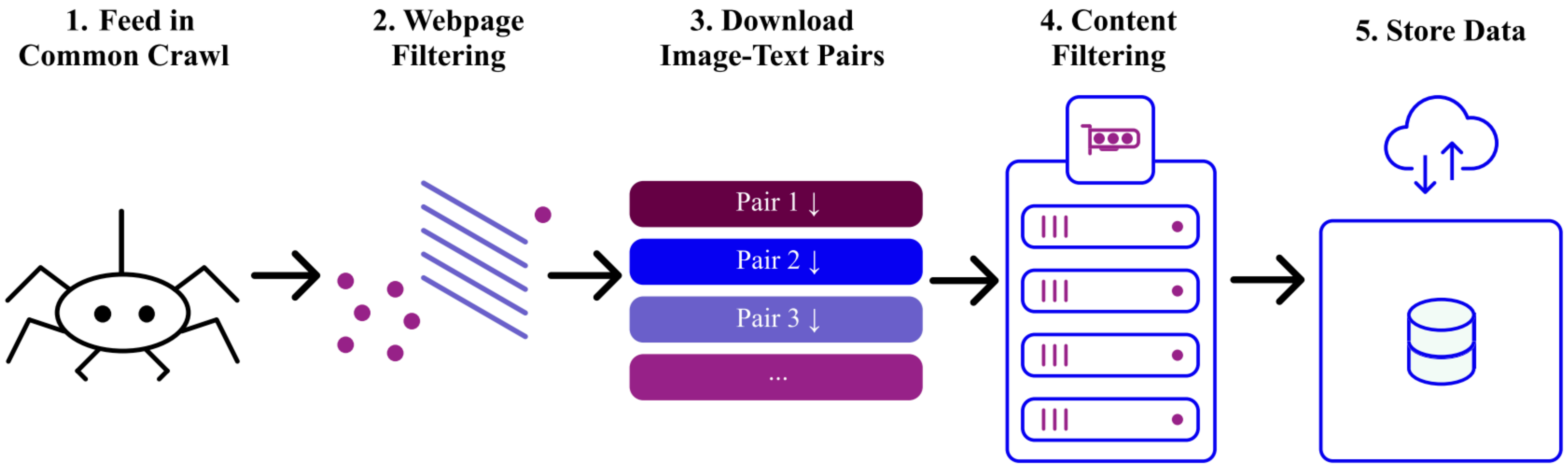
LAION-5B contains 5 billion image-text pairs obtained from CommonCrawl



Schumann et al. "LAION-5B: An open large-scale dataset for training next generation image-text models"

General-Domain Data: LAION-5B

LAION-5B contains 5 billion image-text pairs obtained from CommonCrawl



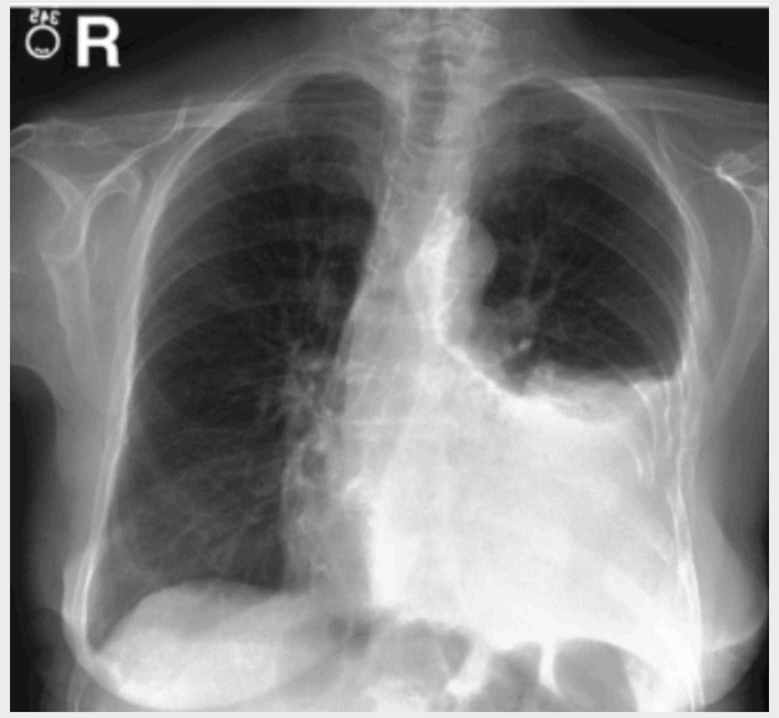
Content filtering is performed using a pre-trained CLIP model
(i.e. by computing cosine similarity between the image and text embeddings)

Schumann et al. “LAION-5B: An open large-scale dataset for training next generation image-text models”

Medical-Domain Data

MIMIC-CXR

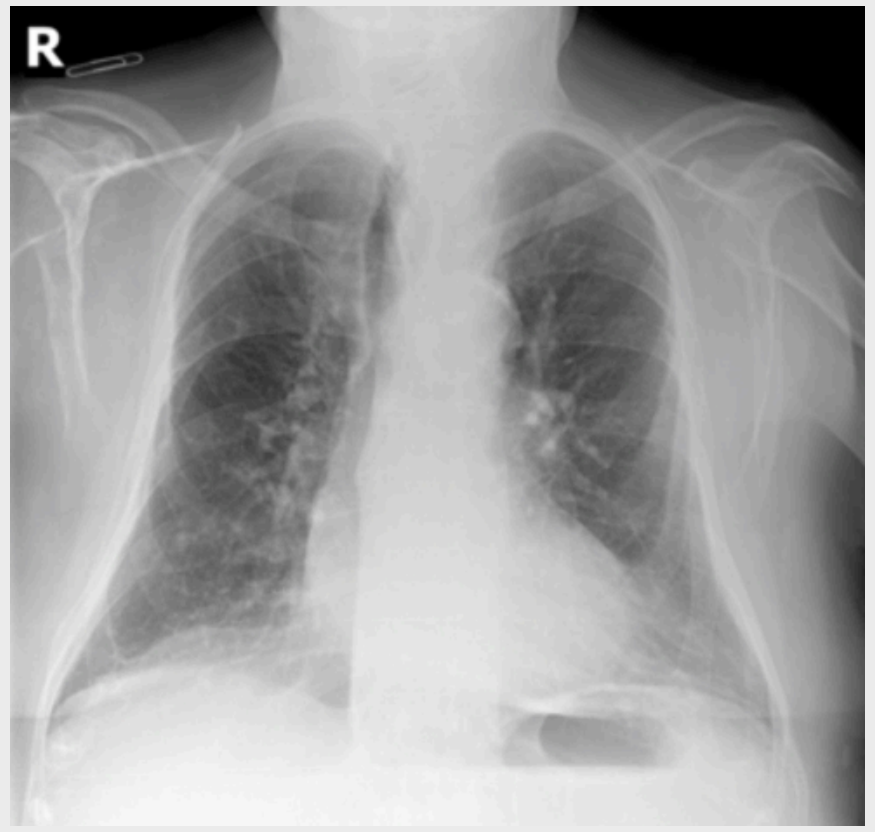
370k chest X-rays with 220k reports



Cardiac size cannot be evaluated. Large left pleural effusion is new. Small right effusion is new. The upper lungs are clear. There is no pneumothorax.

PadChest

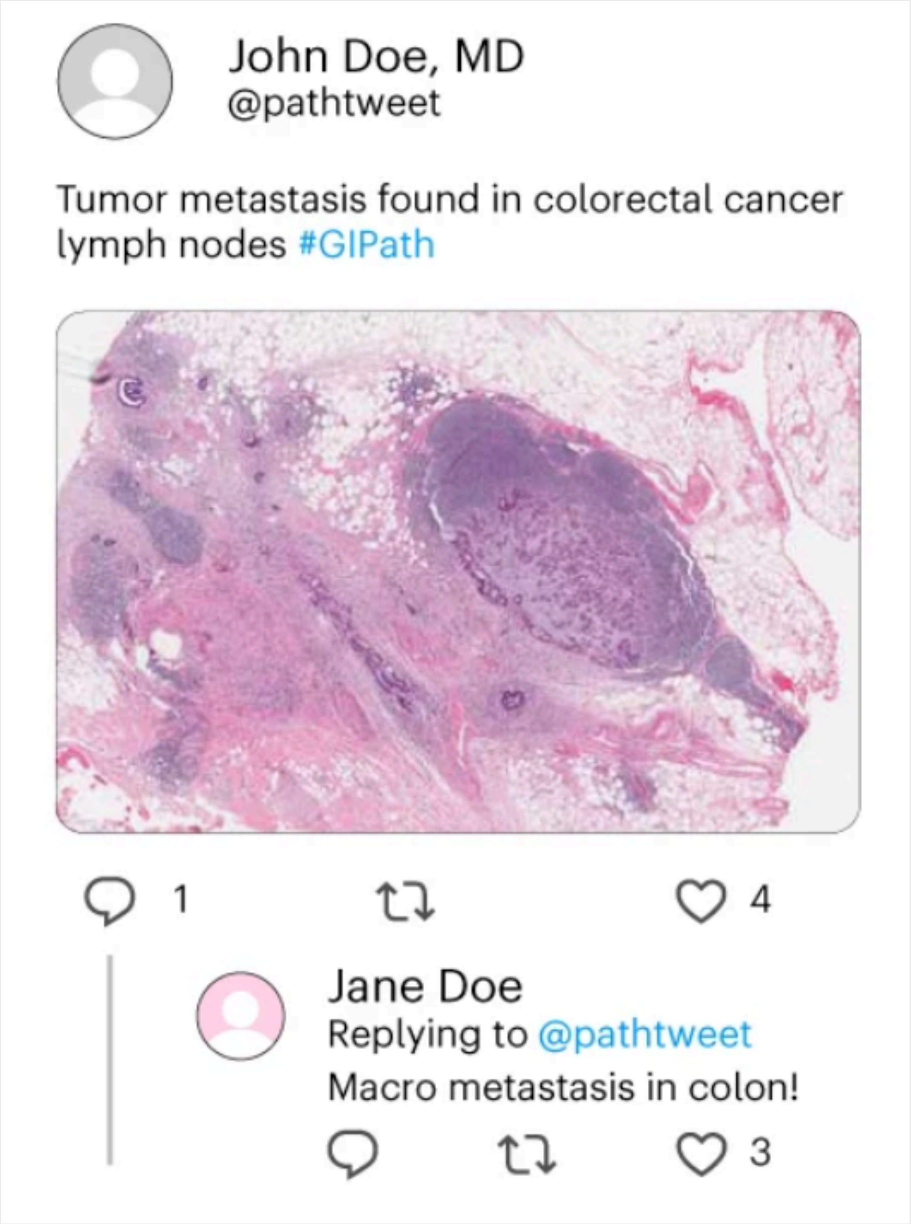
160k chest X-rays with 110k reports (Spanish)



cambi pulmonar cronic sever. sign fibrosis bibasal. sutil infiltr pseudonodul milimetr vidri deslustr localiz bas. cifosis sever

OpenPath

200k histopathology image-text pairs (Twitter)



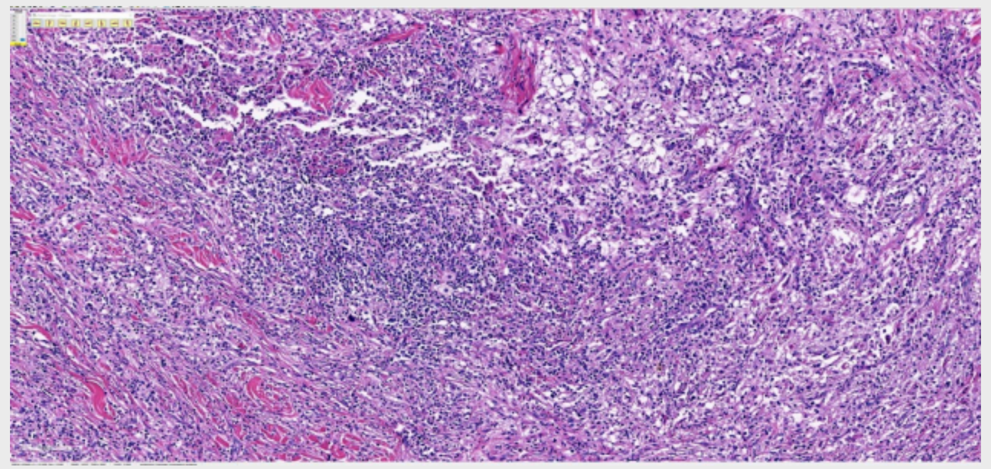
John Doe, MD
@pathtweet

Tumor metastasis found in colorectal cancer lymph nodes #GIPath

Jane Doe
Replying to @pathtweet
Macro metastasis in colon!

Quilt-1M

1M histopathology image-text pairs (Youtube)



Large histiocytes with abundant cytoplasm identified as Rosai-Dorfman histiocytes

Johnson et al. “MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports.”
Bustos et al. “PadChest: A large chest x-ray image dataset with multi-label annotated reports”
Ikezogwo et al. “Quilt-1M: One Million Image-Text Pairs for Histopathology”
Huang et al. “A visual-language foundation model for pathology image analysis using medical Twitter”

Part 3: Evaluation

Evaluating VLMs

Let’s consider a standard classification setup for a vision model

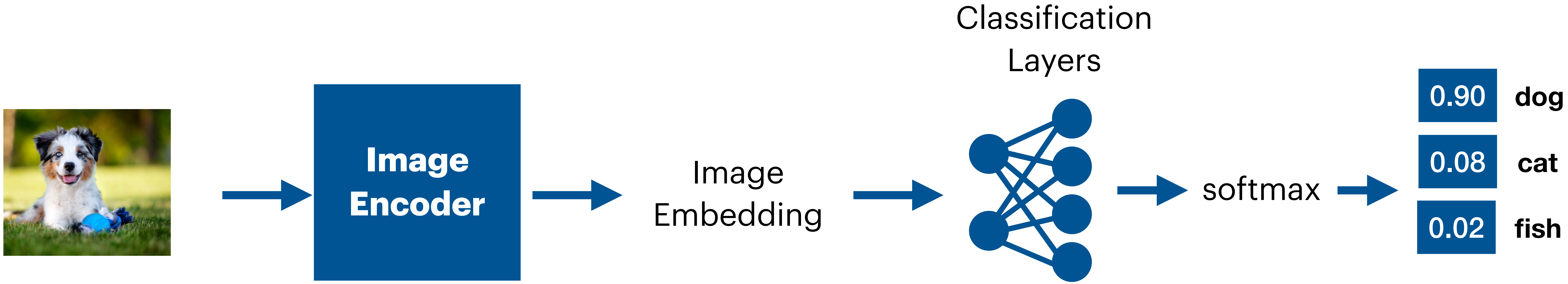


Image Encoder

Image Embedding

Classification Layers

softmax

0.90 dog

0.08 cat

0.02 fish

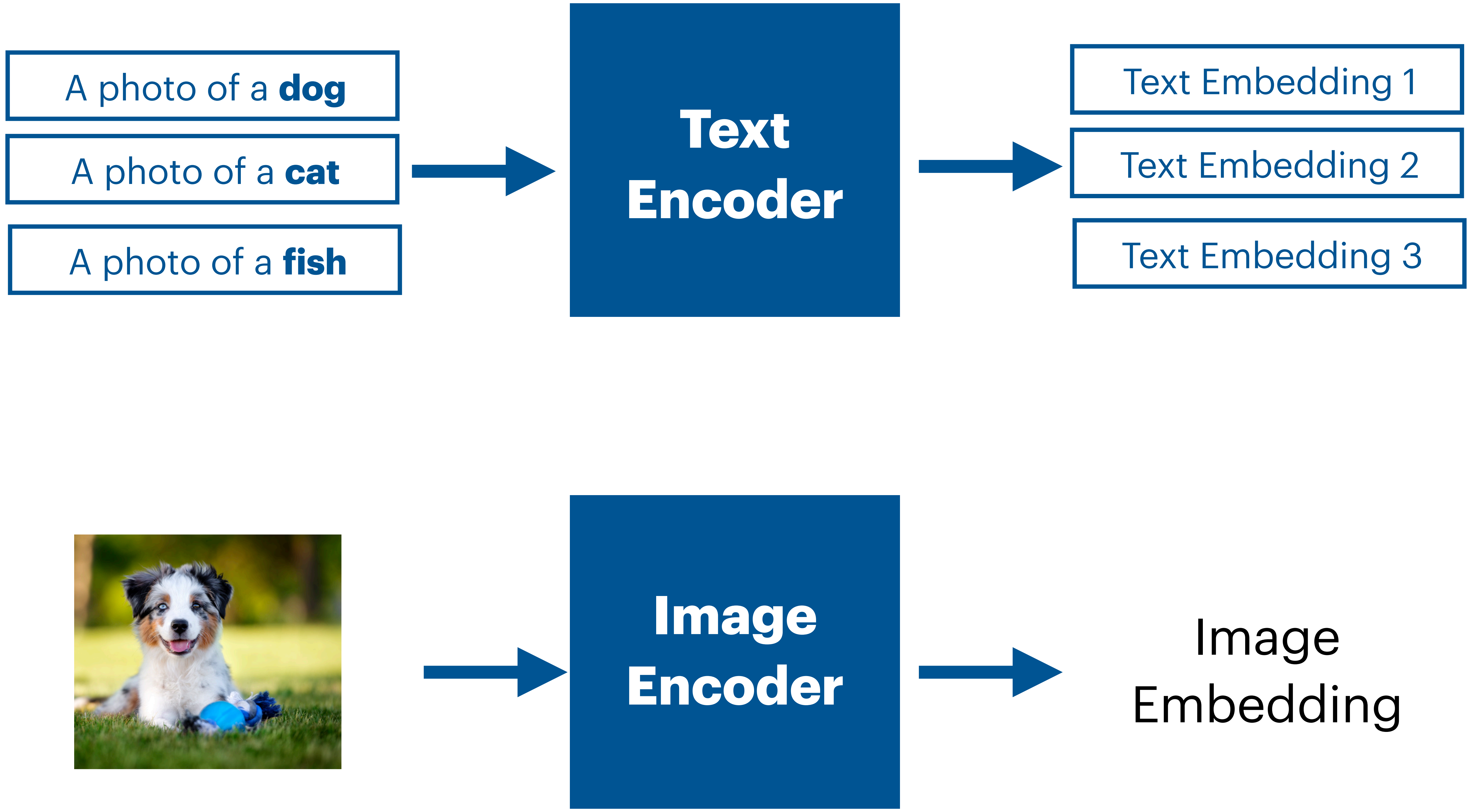
Q: What is undesirable about this approach?

➡ Classification layers need to be trained on an annotated dataset.

➡ Labels are fixed. Changing the labels requires retraining classification layers.

Evaluating VLMs

Zero-Shot Classification



A photo of a **dog**

A photo of a **cat**

A photo of a **fish**

Text Encoder

Text Embedding 1

Text Embedding 2

Text Embedding 3

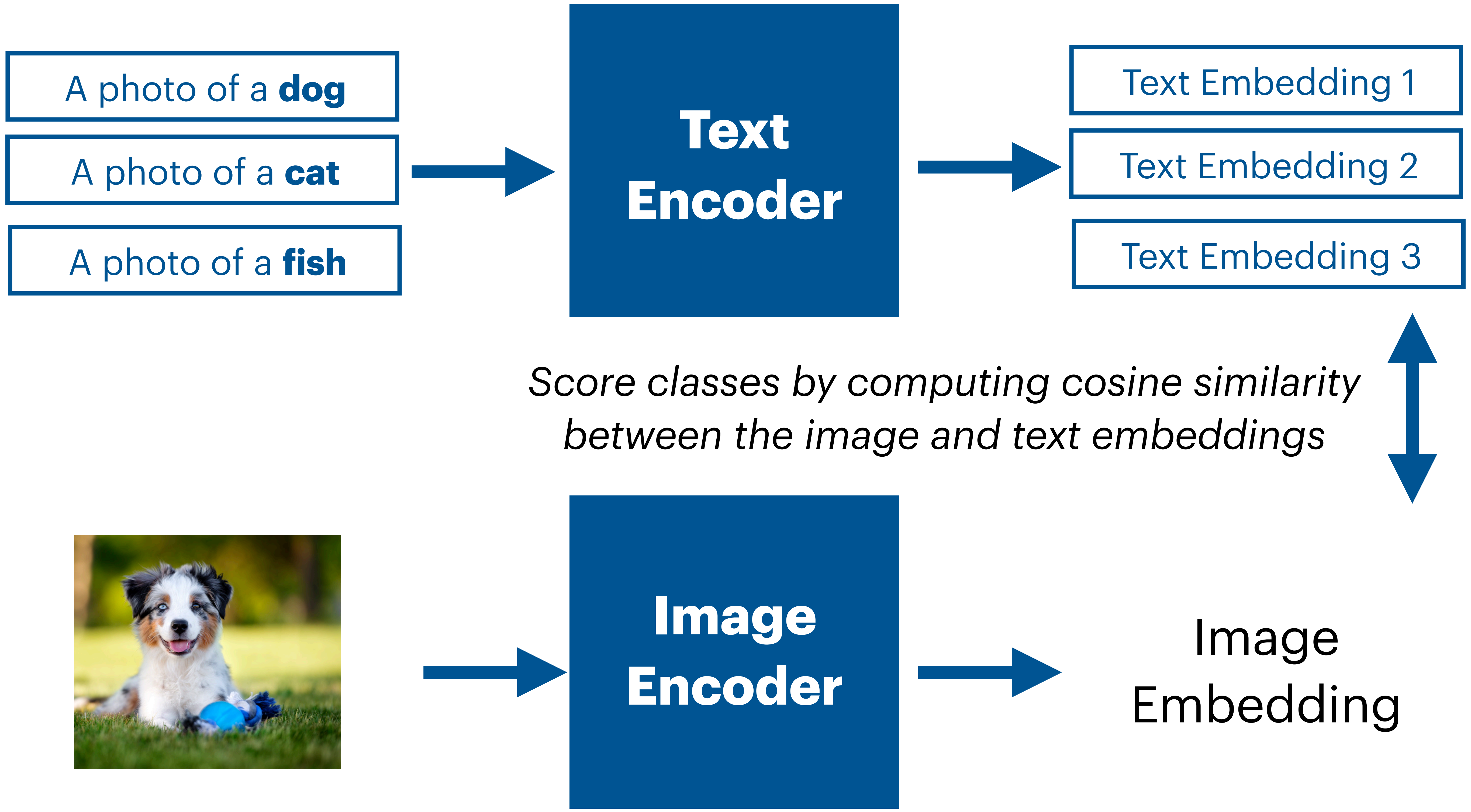
Image Encoder

Image Embedding

Radford et al. "Learning Transferable Visual Models From Natural Language Supervision”

Evaluating VLMs

Zero-Shot Classification



A photo of a **dog**

A photo of a **cat**

A photo of a **fish**

Text Encoder

Text Embedding 1

Text Embedding 2

Text Embedding 3

Score classes by computing cosine similarity between the image and text embeddings

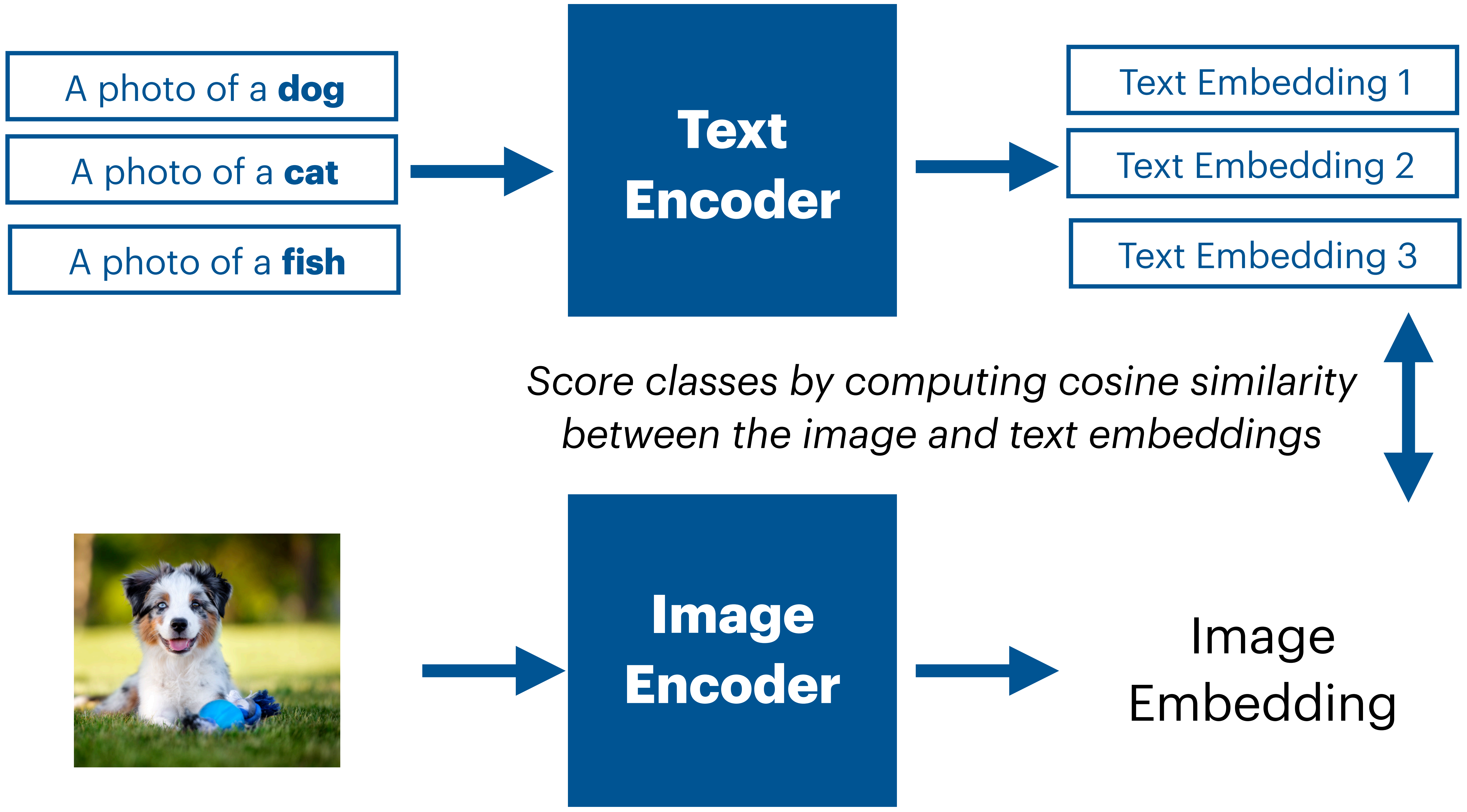
Image Encoder

Image Embedding

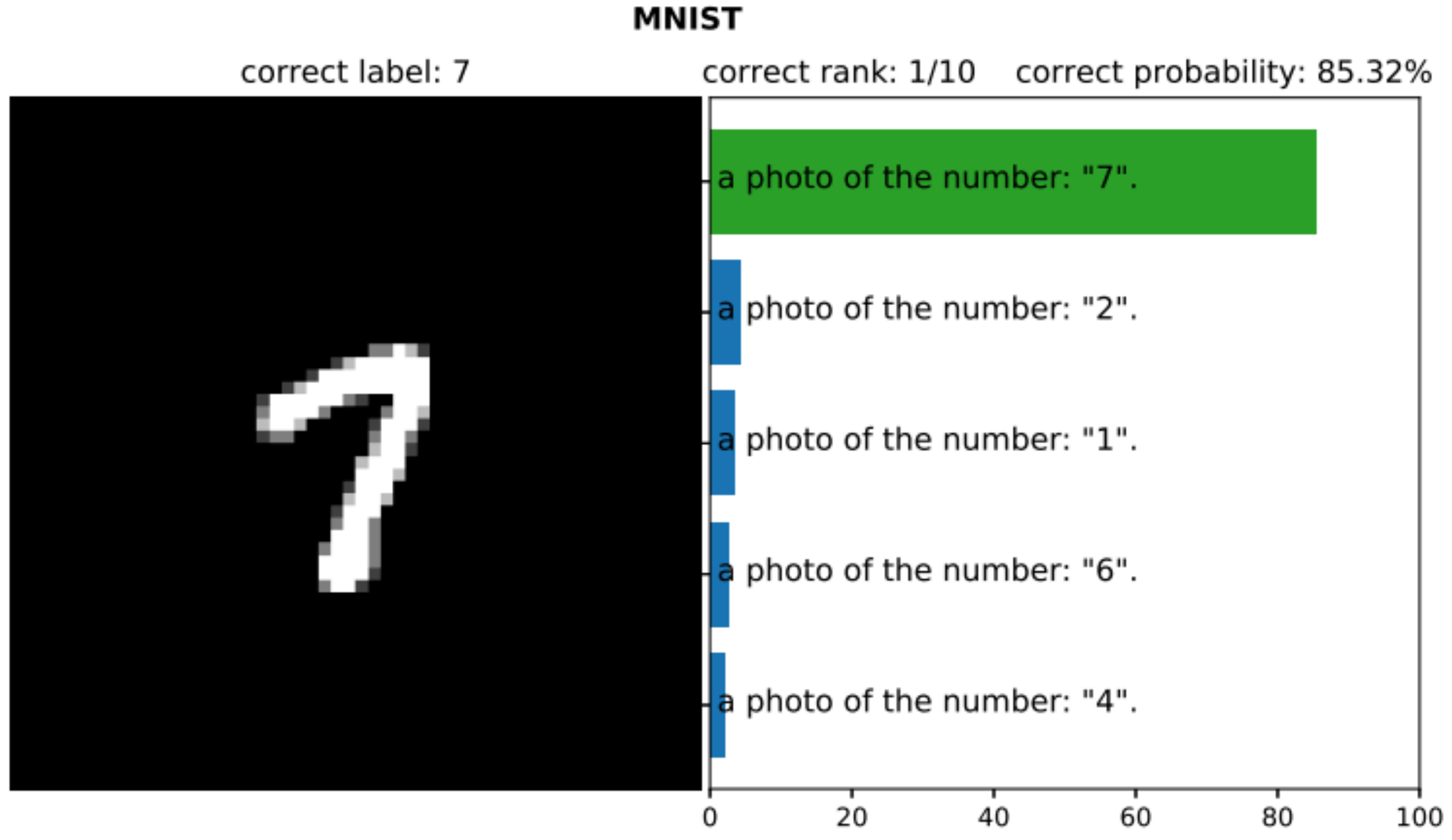
Radford et al. "Learning Transferable Visual Models From Natural Language Supervision”

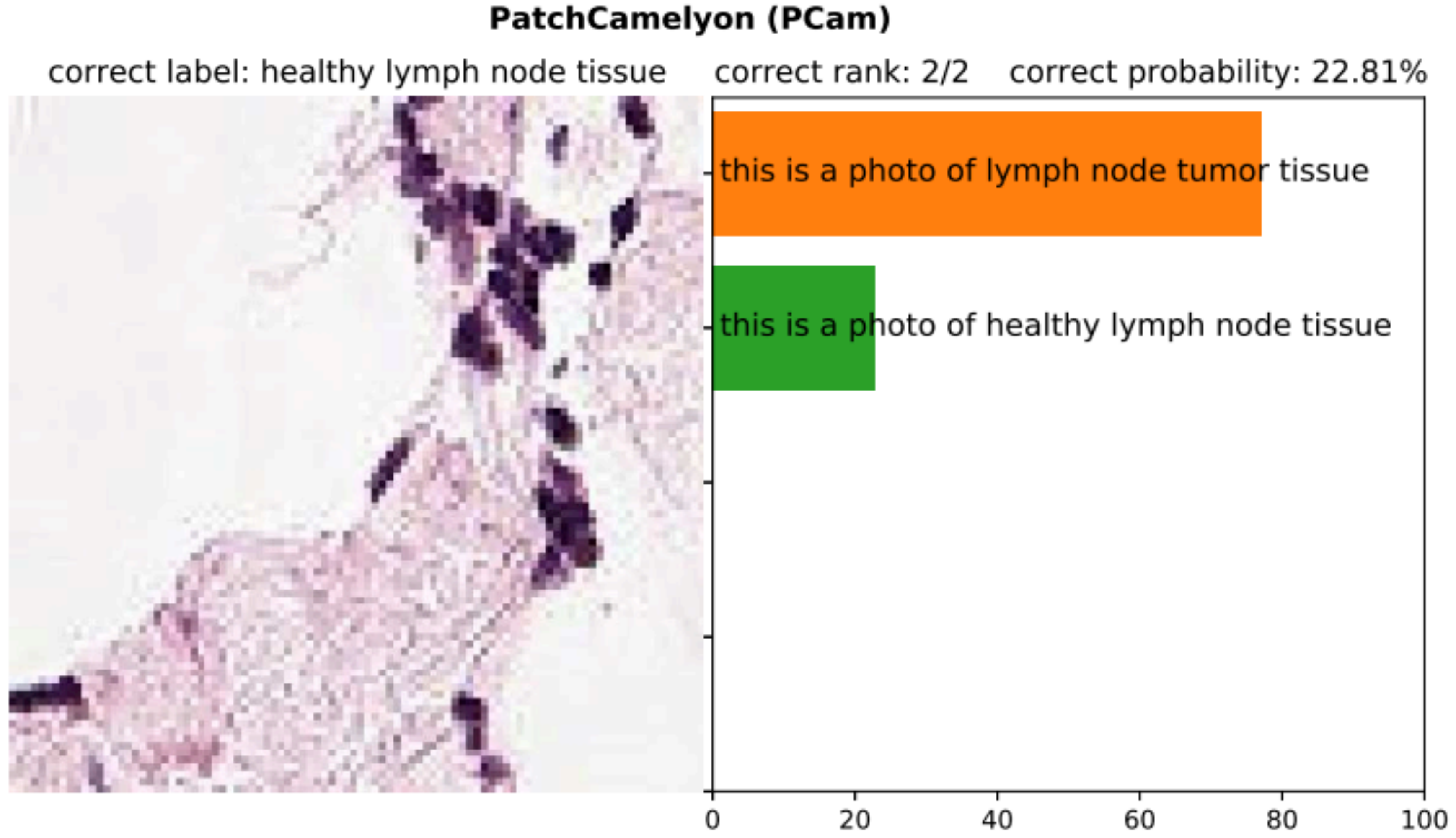
Evaluating VLMs

Zero-Shot Classification



Score classes by computing cosine similarity between the image and text embeddings

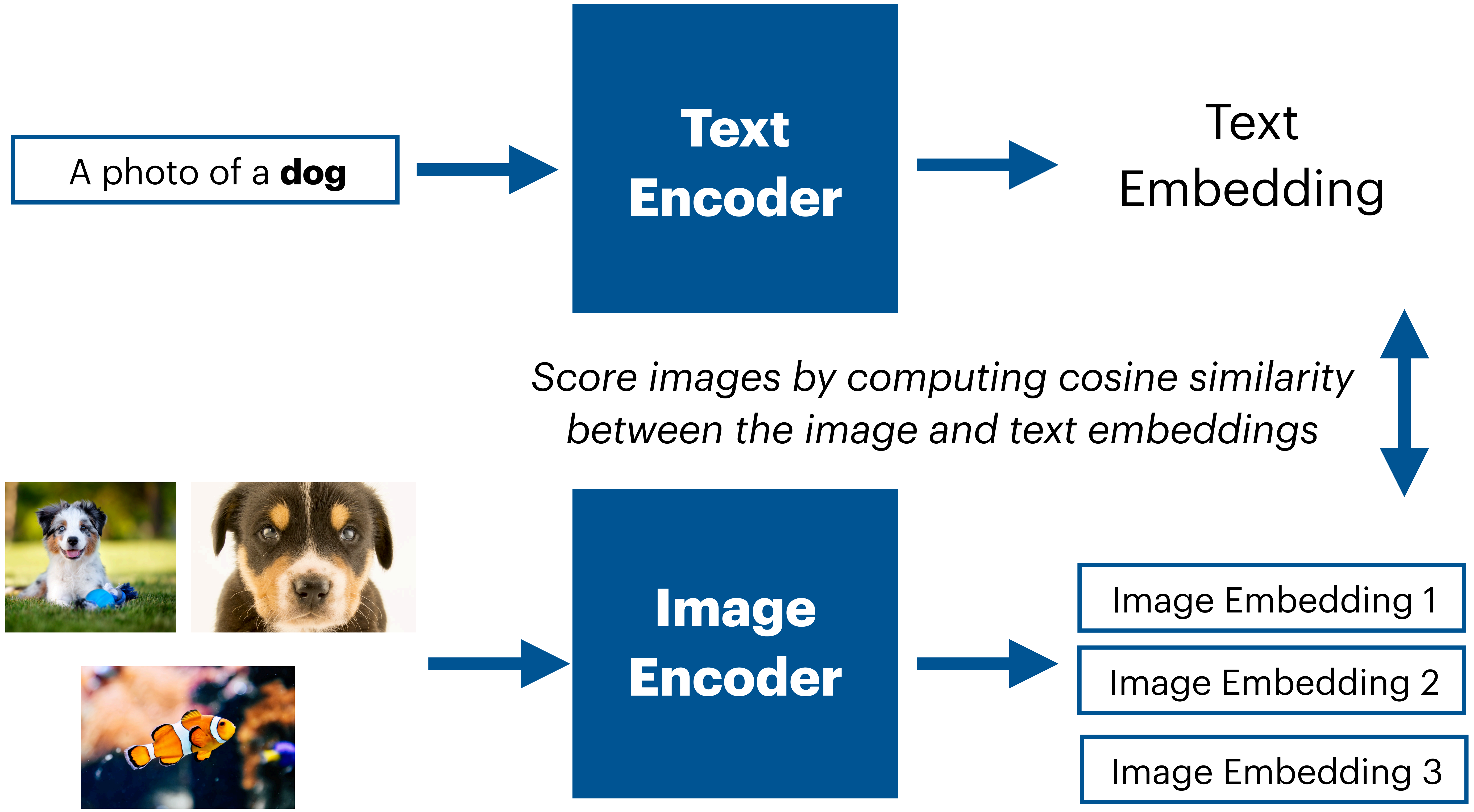




Radford et al. "Learning Transferable Visual Models From Natural Language Supervision”

Evaluating VLMs

Text → Image Retrieval



A photo of a **dog**

Text Encoder

Text Embedding

Score images by computing cosine similarity between the image and text embeddings

Image Encoder

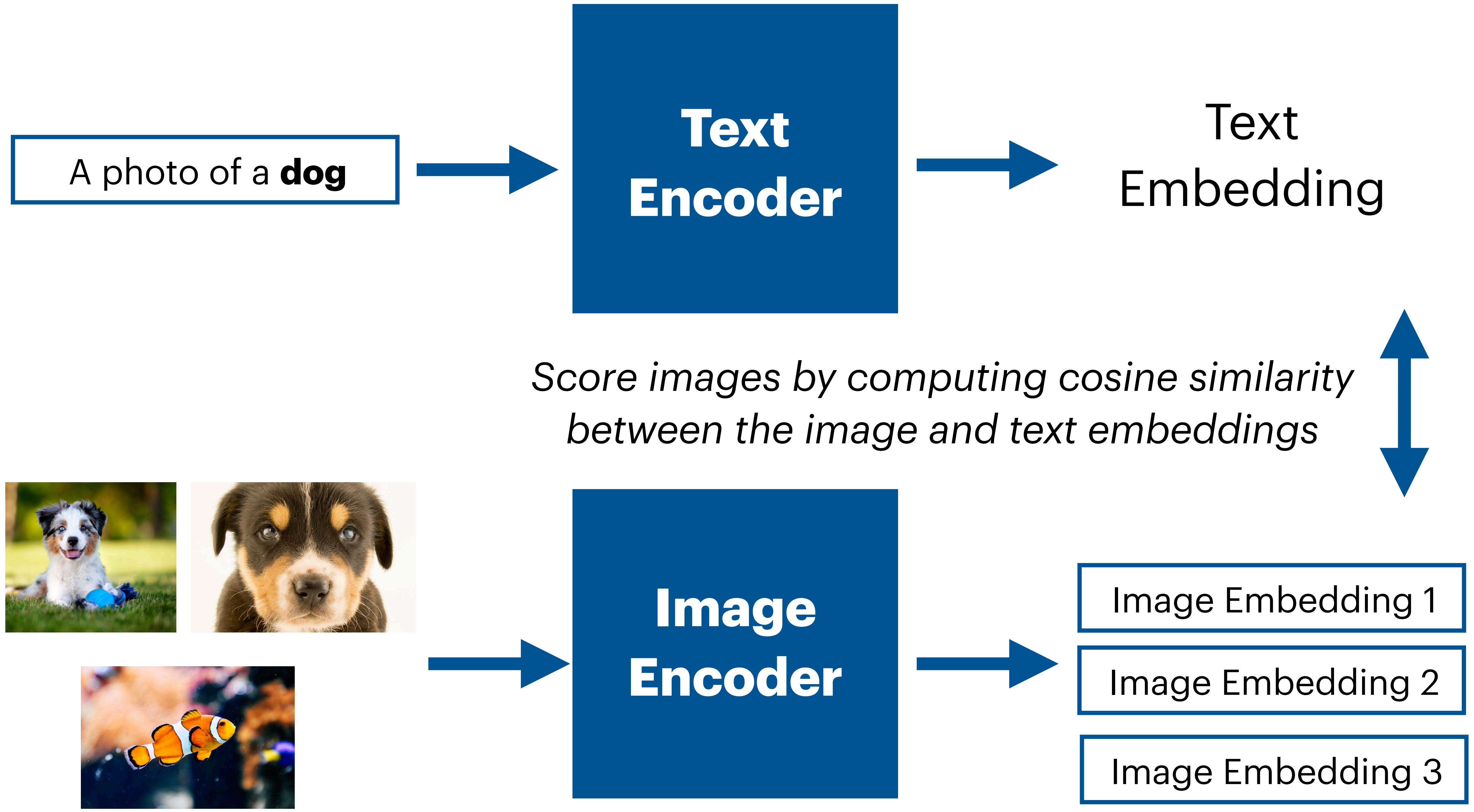
Image Embedding 1

Image Embedding 2

Image Embedding 3

Evaluating VLMs

Text → Image Retrieval



A photo of a **dog**

Text Encoder

Text Embedding

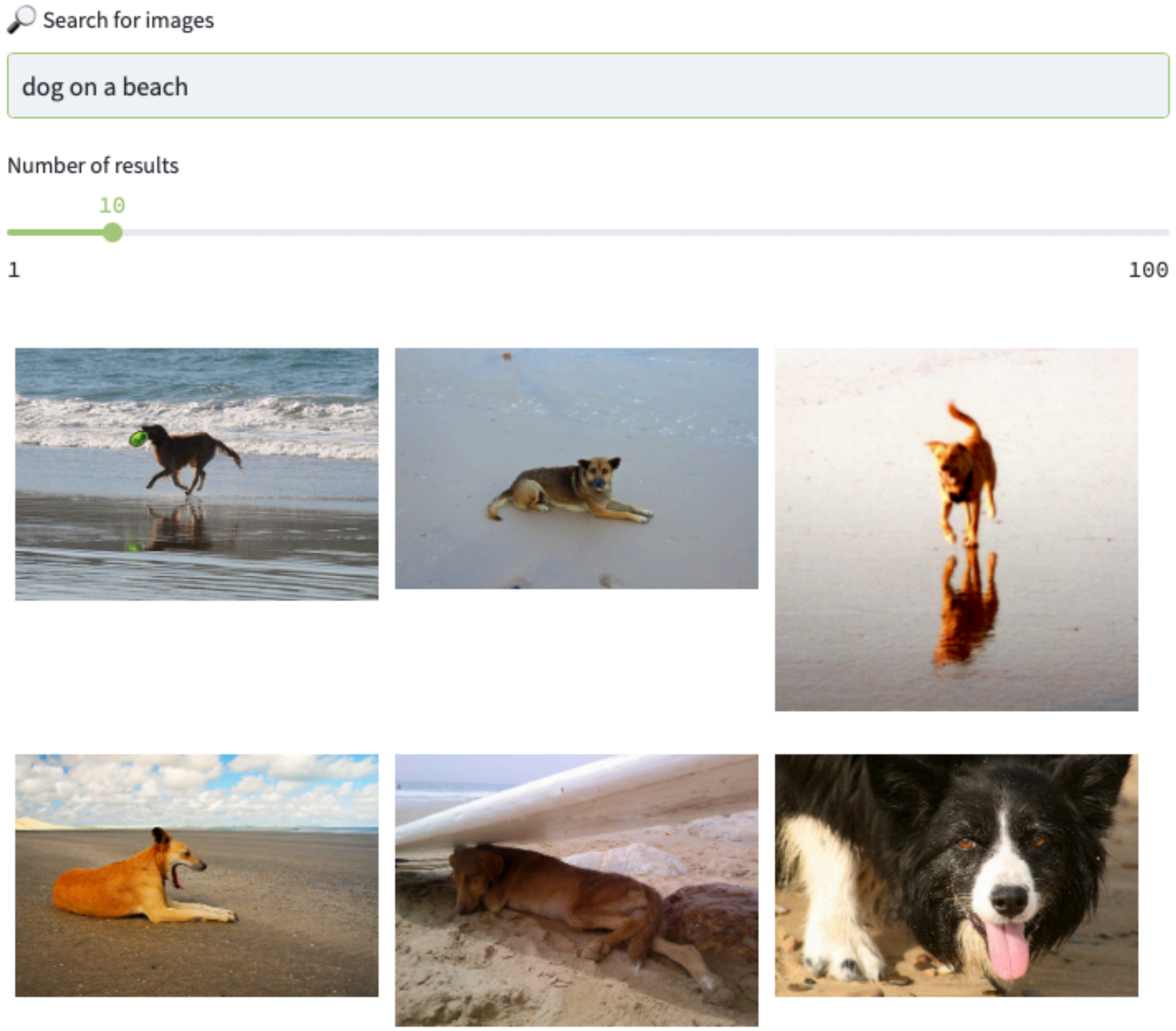
Score images by computing cosine similarity between the image and text embeddings

Image Encoder

Image Embedding 1

Image Embedding 2

Image Embedding 3



Search for images

dog on a beach

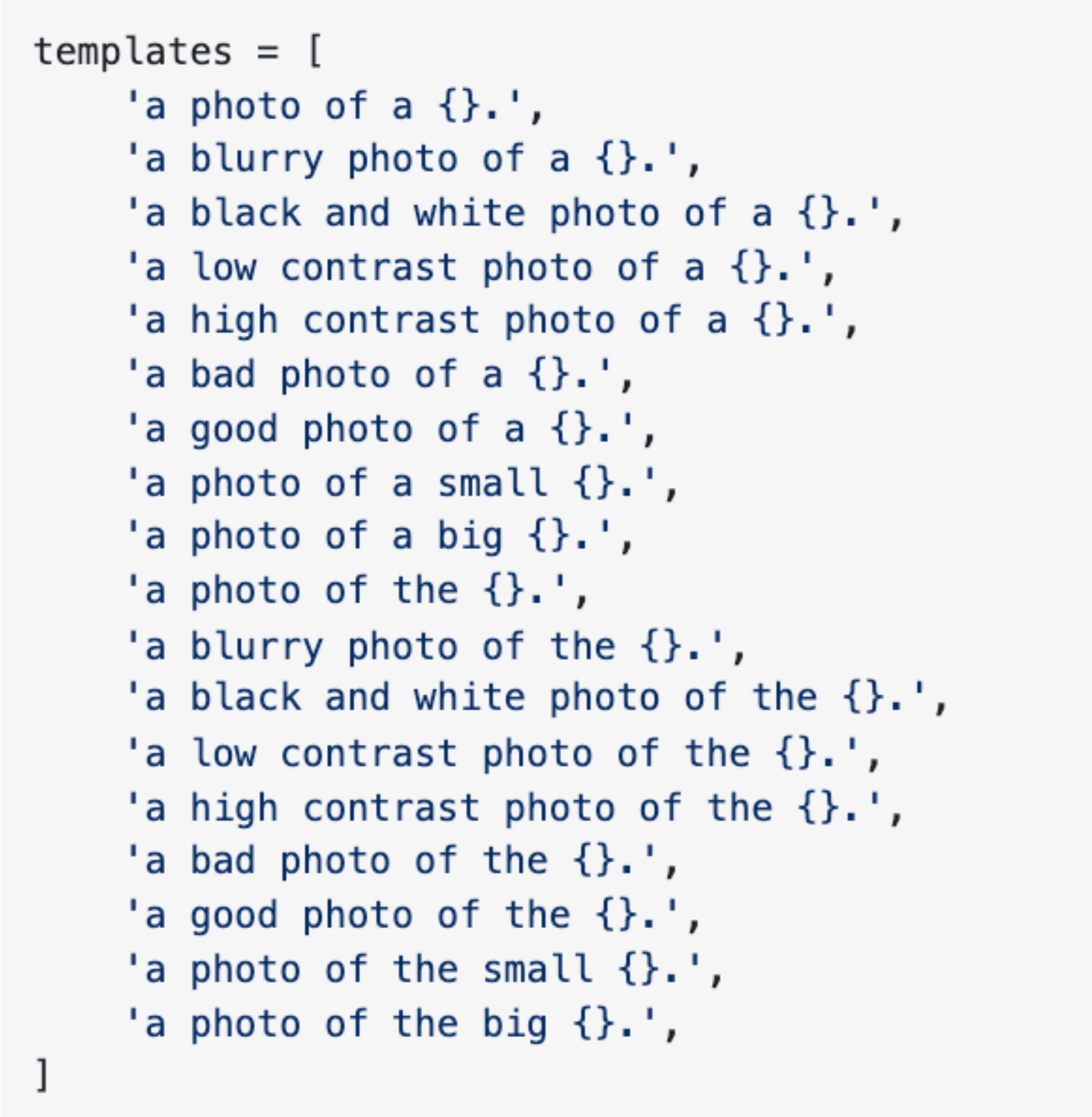
Number of results

10

Prompting VLMs

Textual Prompts

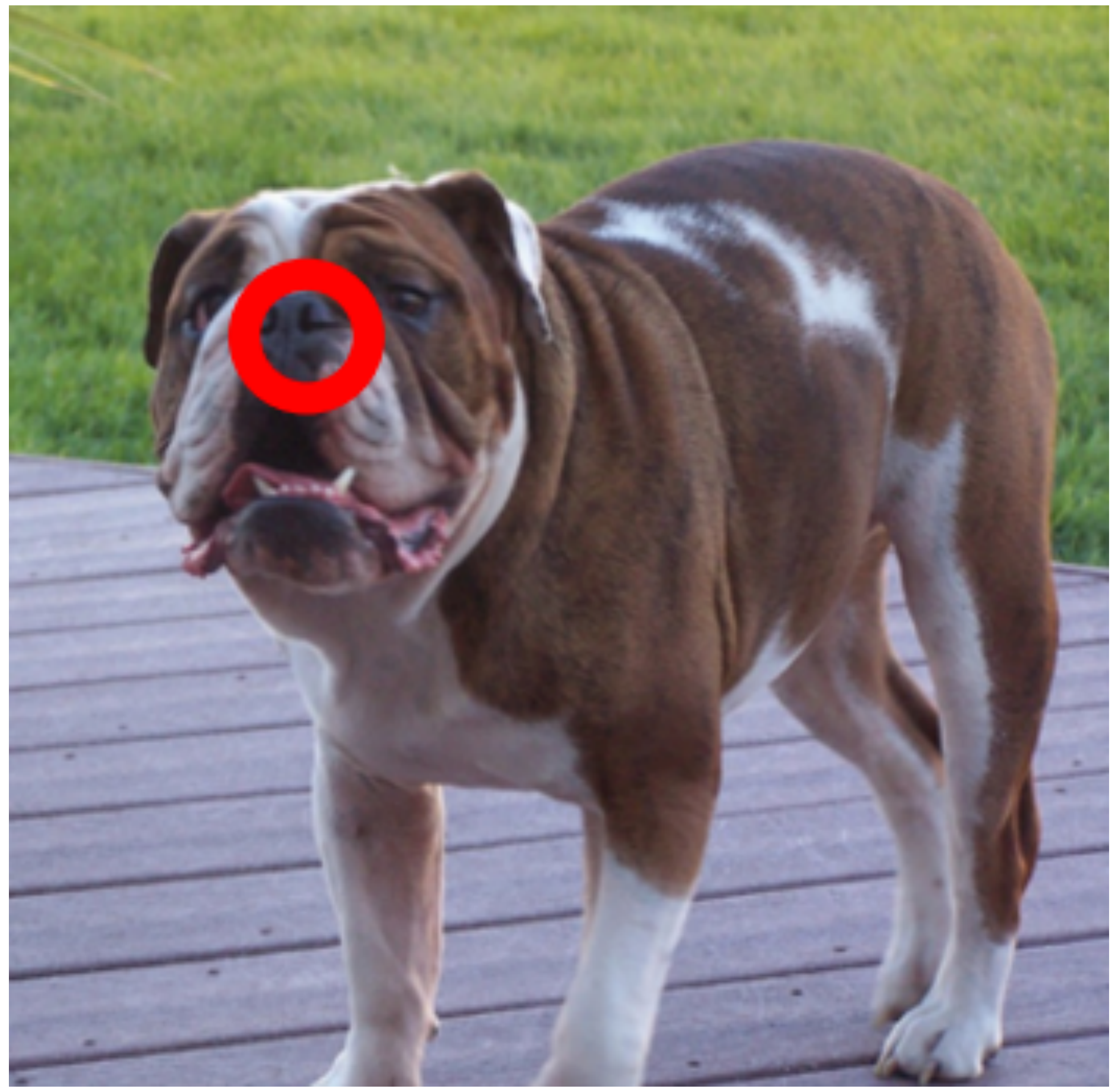
Example text prompts used by CLIP for zero-shot classification on CIFAR-10



```
templates = [
    'a photo of a {}.',
    'a blurry photo of a {}.',
    'a black and white photo of a {}.',
    'a low contrast photo of a {}.',
    'a high contrast photo of a {}.',
    'a bad photo of a {}.',
    'a good photo of a {}.',
    'a photo of a small {}.',
    'a photo of a big {}.',
    'a photo of the {}.',
    'a blurry photo of the {}.',
    'a black and white photo of the {}.',
    'a low contrast photo of the {}.',
    'a high contrast photo of the {}.',
    'a bad photo of the {}.',
    'a good photo of the {}.',
    'a photo of the small {}.',
    'a photo of the big {}.',
]
```

Visual Prompts

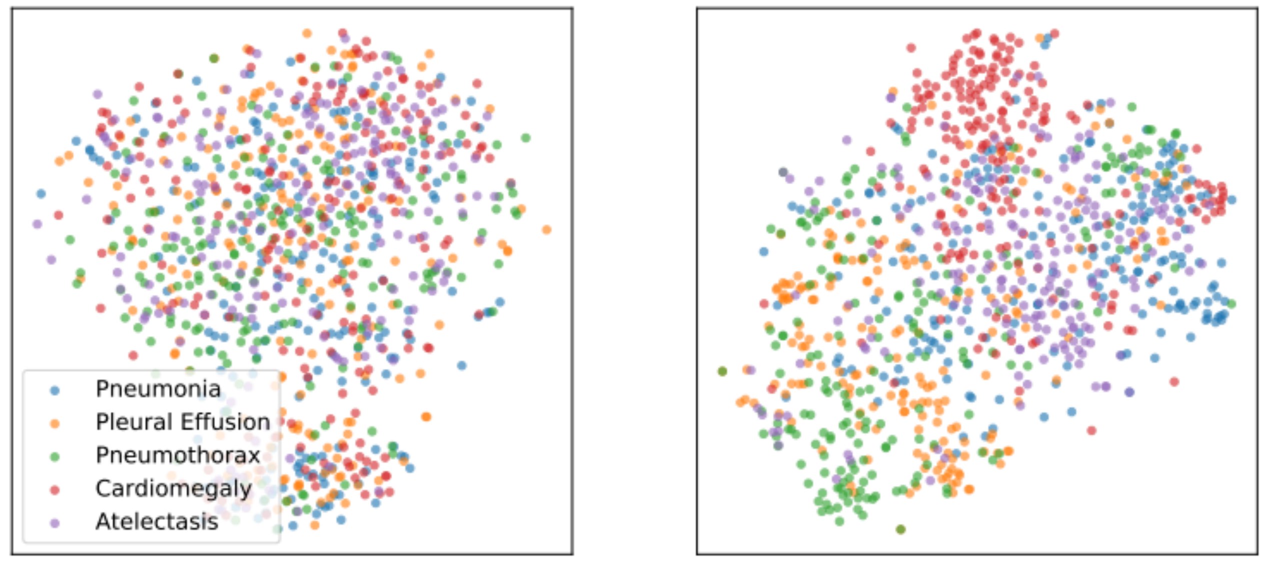
Adding visual signal to images can help with targeted retrieval and classification



Radford et al. "Learning Transferable Visual Models From Natural Language Supervision”
Shtedritski et al. “What does CLIP know about a red circle? Visual prompt engineering for VLMs”

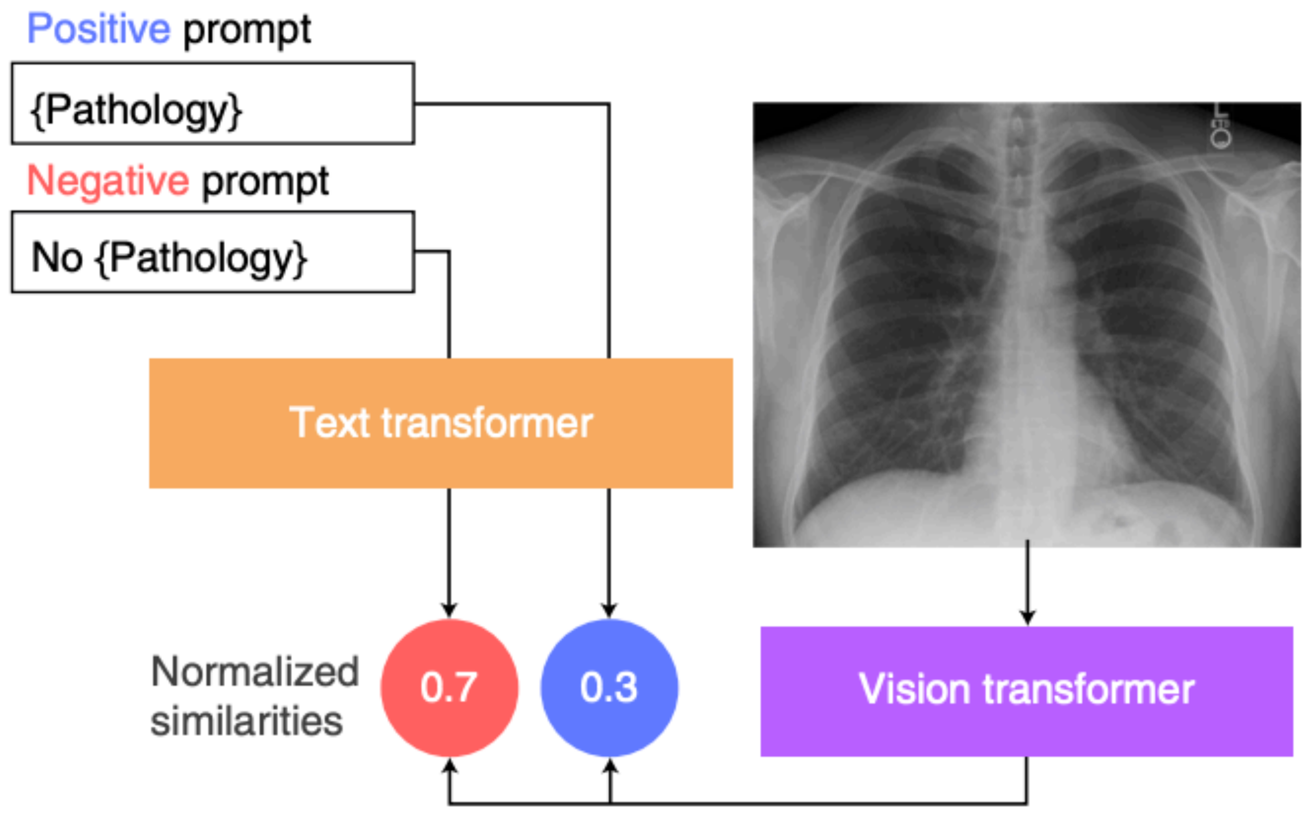
Evaluating Medical VLMs

Classification

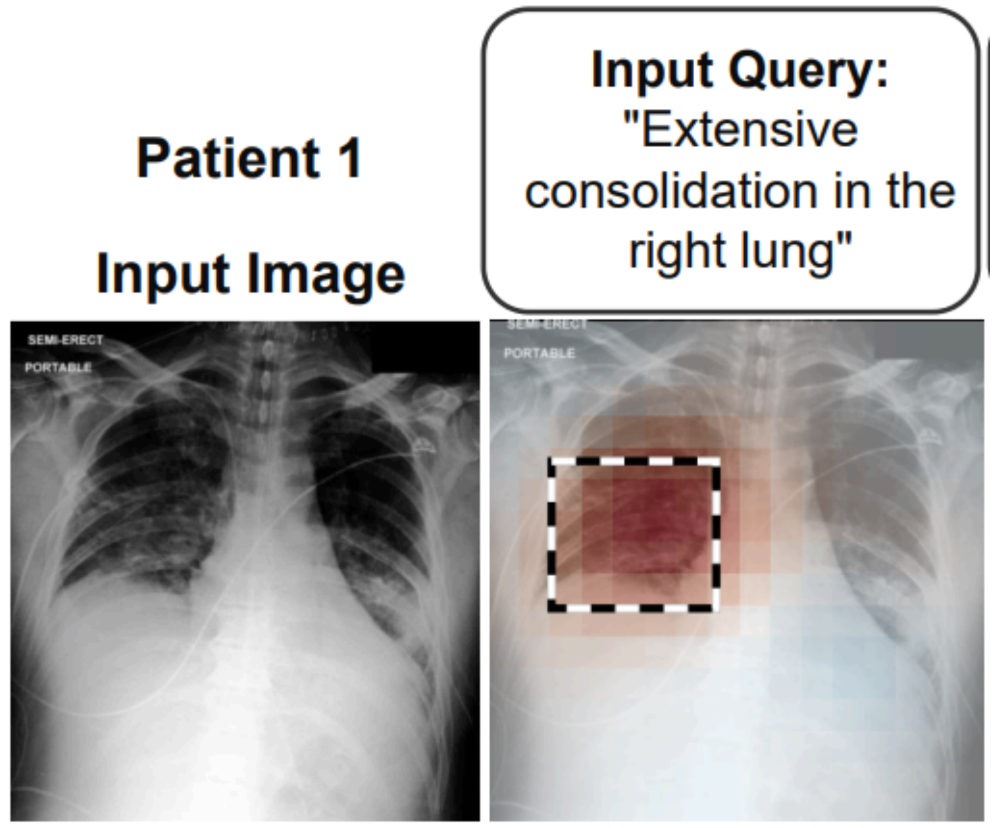


(*a*) ImageNet Pretraining (*b*) ConVIRT Pretraining

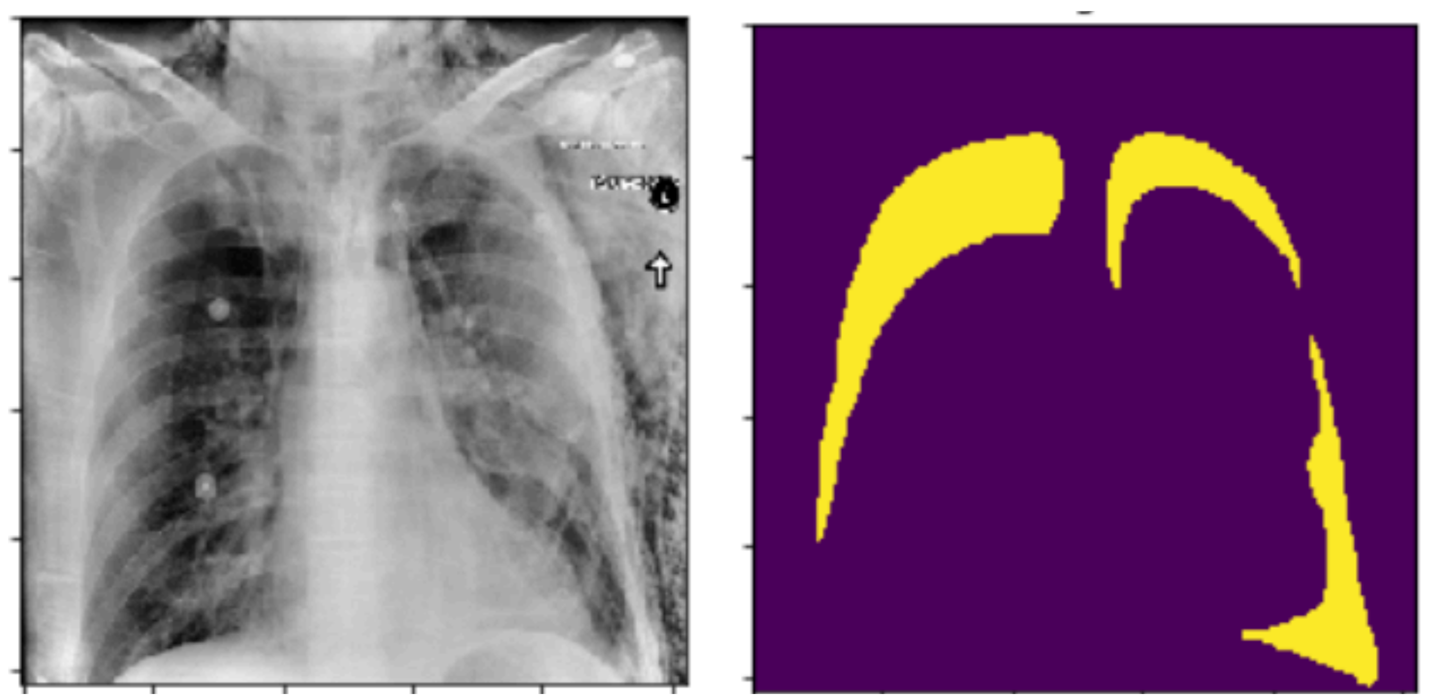
Zero-Shot Classification



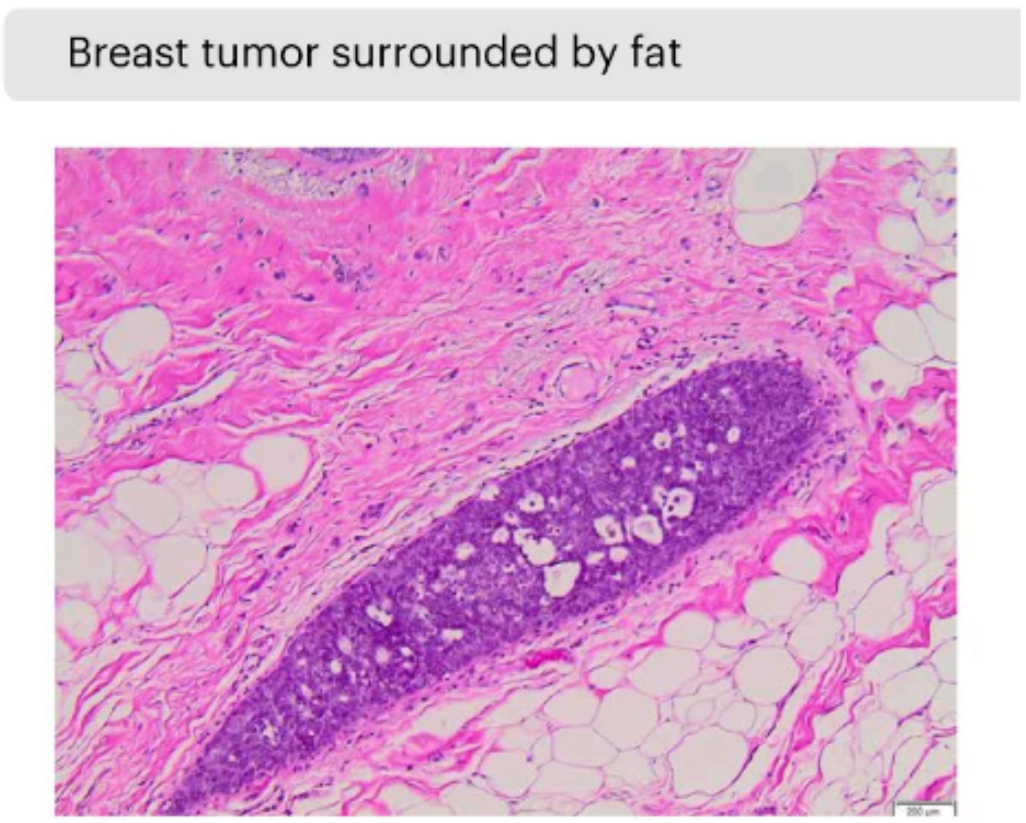
Visual Grounding



Segmentation



Text to Image Retrieval



Breast tumor surrounded by fat

Natural Language Inference

Sentence 1:

No pneumothorax is seen

Sentence 2:

Previously-seen pneumothorax is no longer visualized

Type: ***Entailment***

Miura et al. “RadNLI: A natural language inference dataset for the radiology domain”
Zhang et al. "Contrastive Learning of Medical Visual Representations from Paired Images and Text”
Huang et al. “A visual-language foundation model for pathology image analysis using medical Twitter”

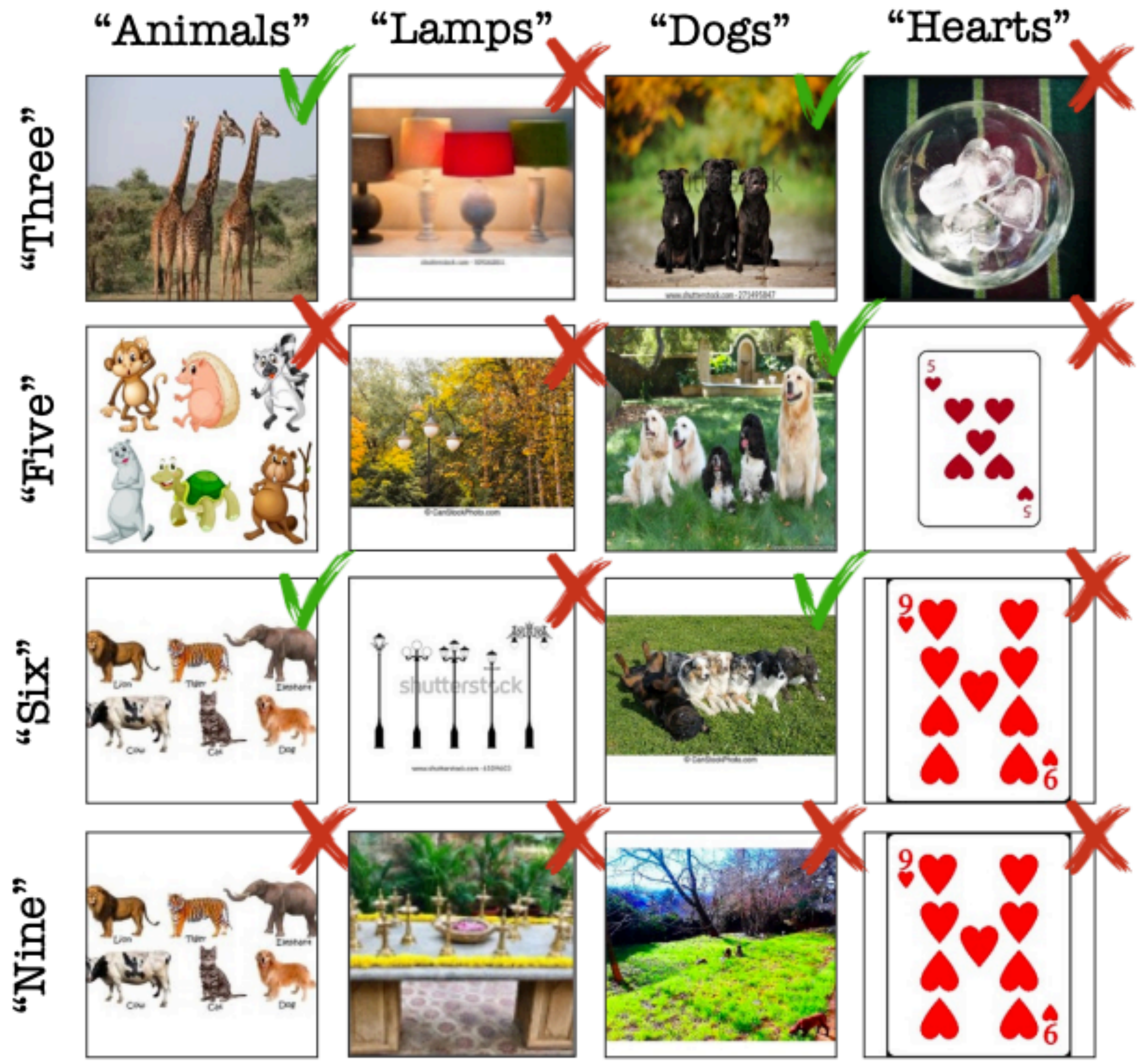
Tiu et al. “Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning”
Boecking et al. “Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing”

Part 4: Limitations

Limitations: Contrastive Training

Complex Patterns (e.g. counting)

Image Retrieval Results:



Original CLIP Model

Relational Understanding

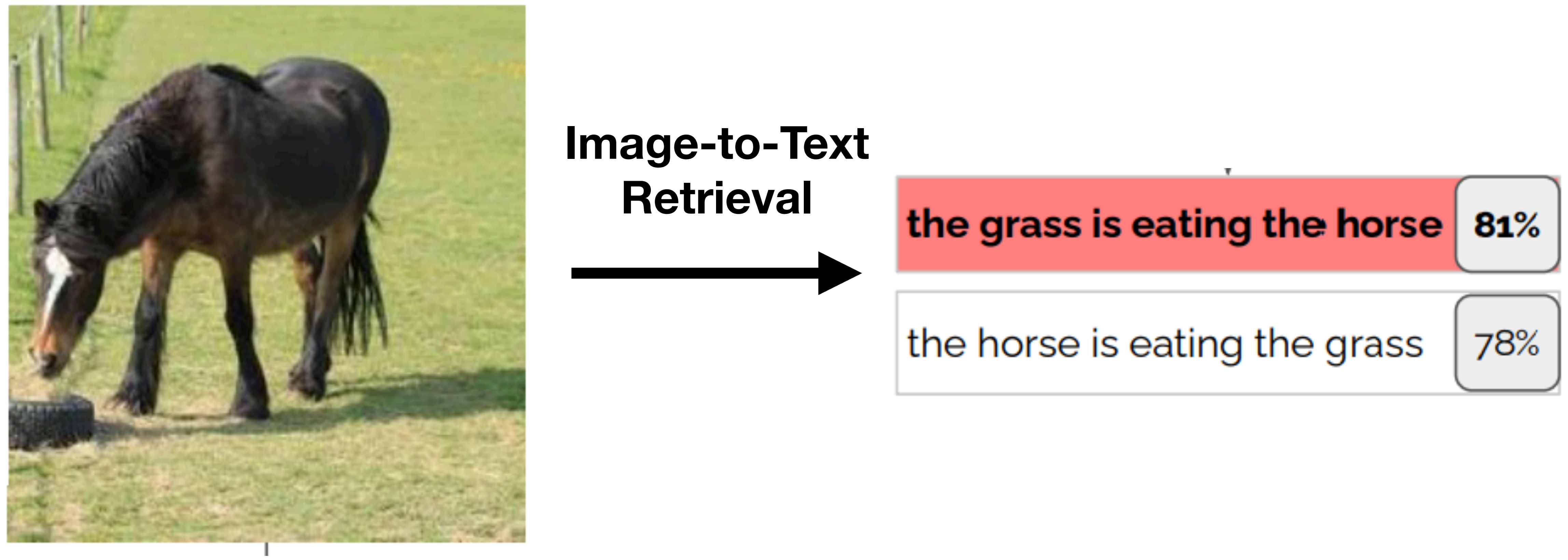
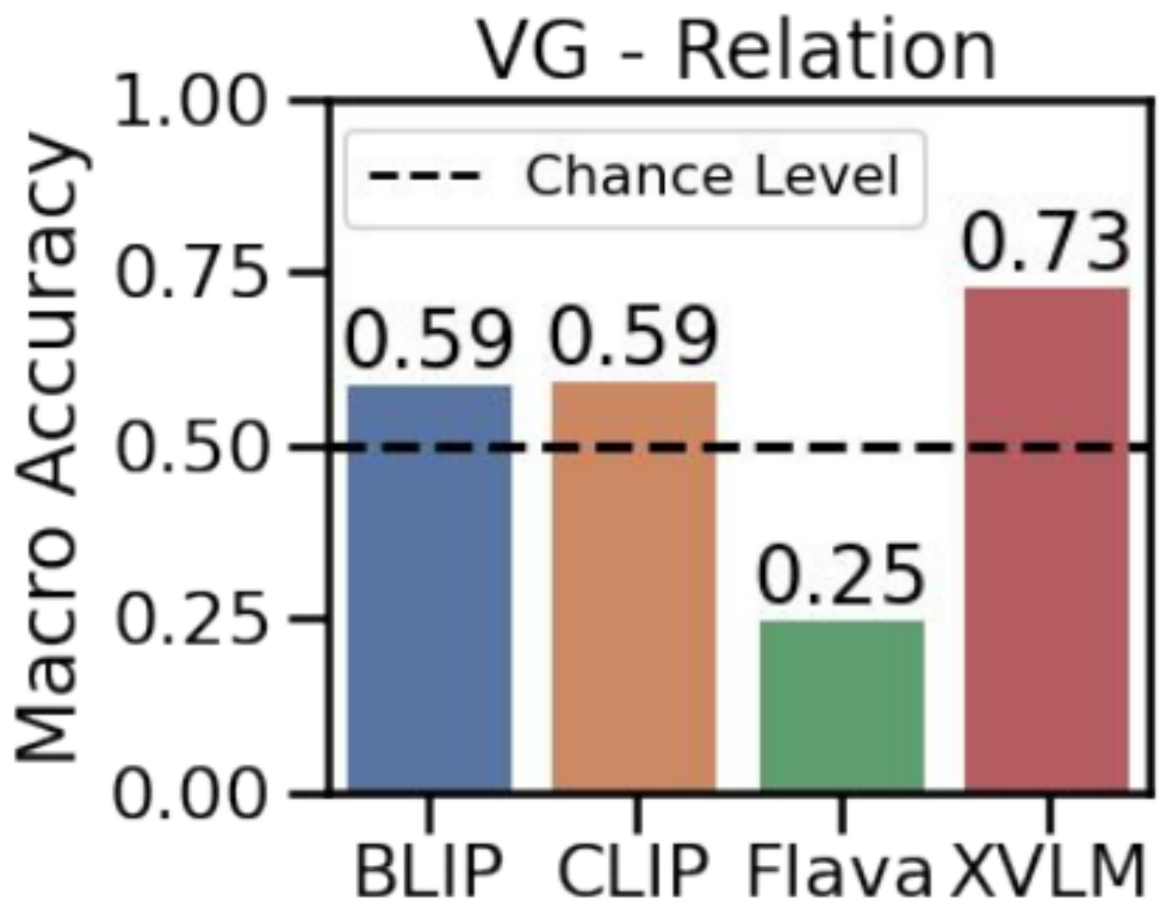


Image-to-Text Retrieval

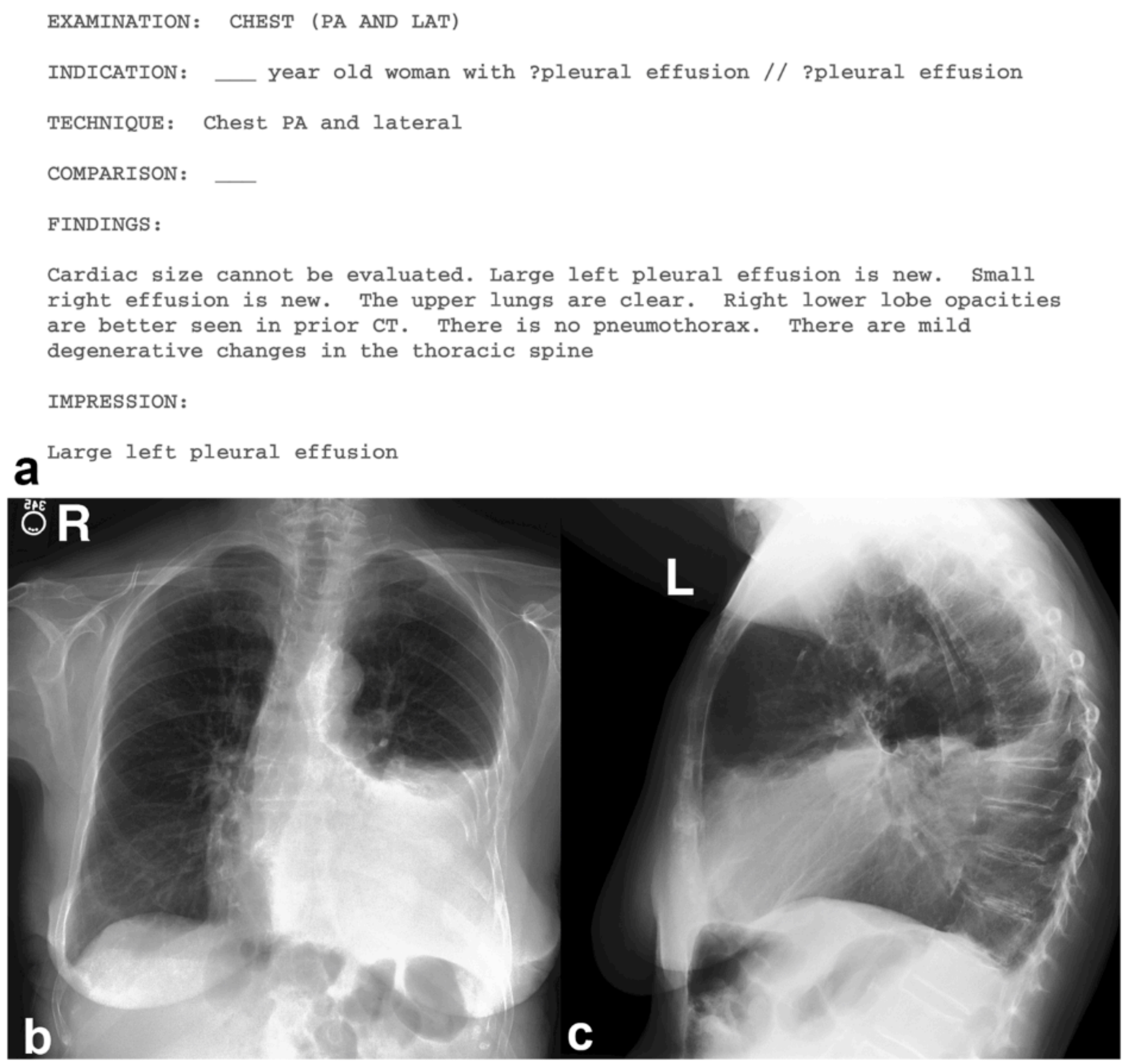
the grass is eating the horse	**81%**
the horse is eating the grass	78%



Paiss et al. "Teaching CLIP to Count to Ten"
Yuksekgonul et al. "When and Why Vision-Language Models Behave Like Bags-of-Words and What to Do About it?"

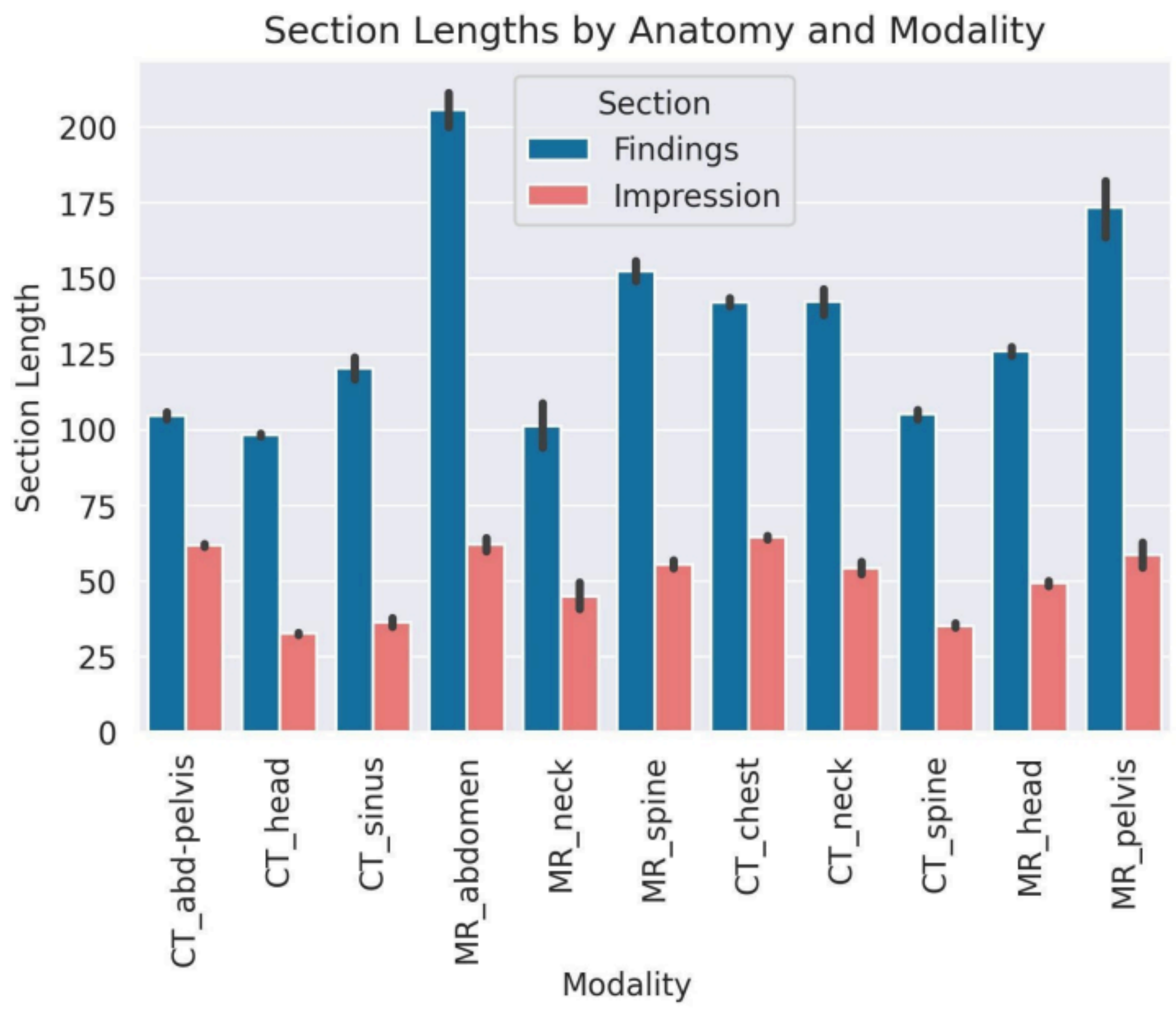
Limitations: Domain-Specific Challenges

Fine-Grained Visual Information



```
EXAMINATION:  CHEST (PA AND LAT)

INDICATION:  ___ year old woman with ?pleural effusion // ?pleural effusion

TECHNIQUE:  Chest PA and lateral

COMPARISON:  ___

FINDINGS:

Cardiac size cannot be evaluated. Large left pleural effusion is new.  Small
right effusion is new.  The upper lungs are clear.  Right lower lobe opacities
are better seen in prior CT.  There is no pneumothorax.  There are mild
degenerative changes in the thoracic spine

IMPRESSION:

Large left pleural effusion
```

Lengthy and Complex Text

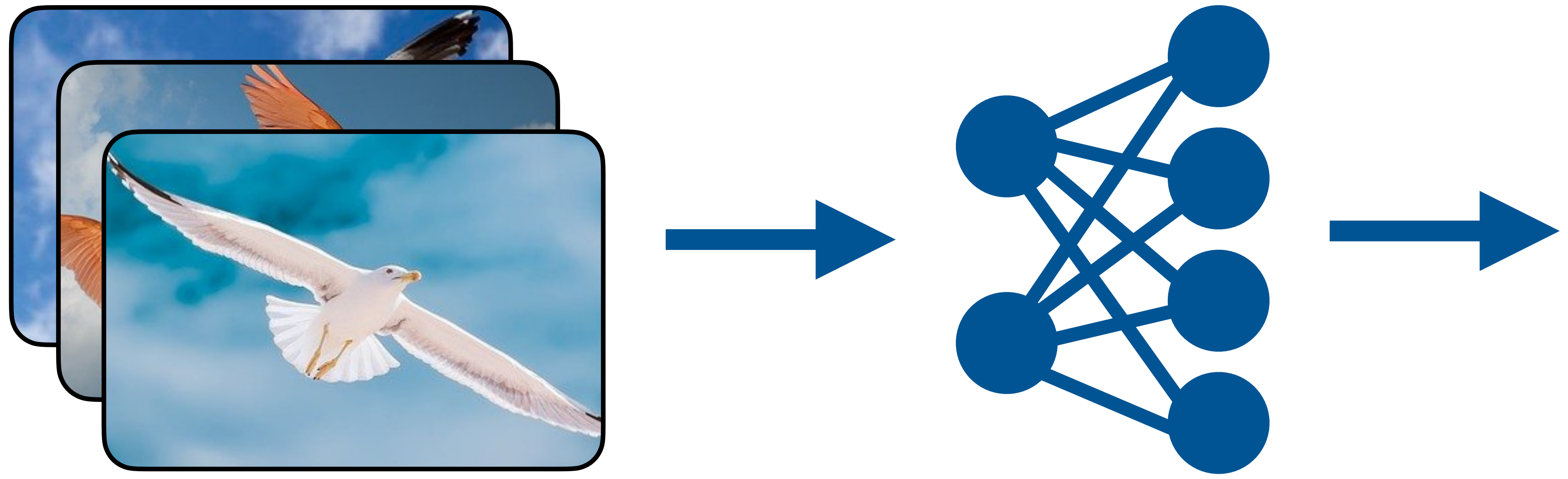


Chen*, Varma*, et al. “Toward Expanding the Scope of Radiology Report Summarization to Multiple Anatomies and Modalities"

Part 5: Applications

Application 1: Discovering Systematic Errors

Computer vision models often demonstrate high overall performance…



Classification Performance on Class “`bird`”: 0.95

…yet make systematic errors on specific data subgroups

Classification Performance on subgroup with blue skies:

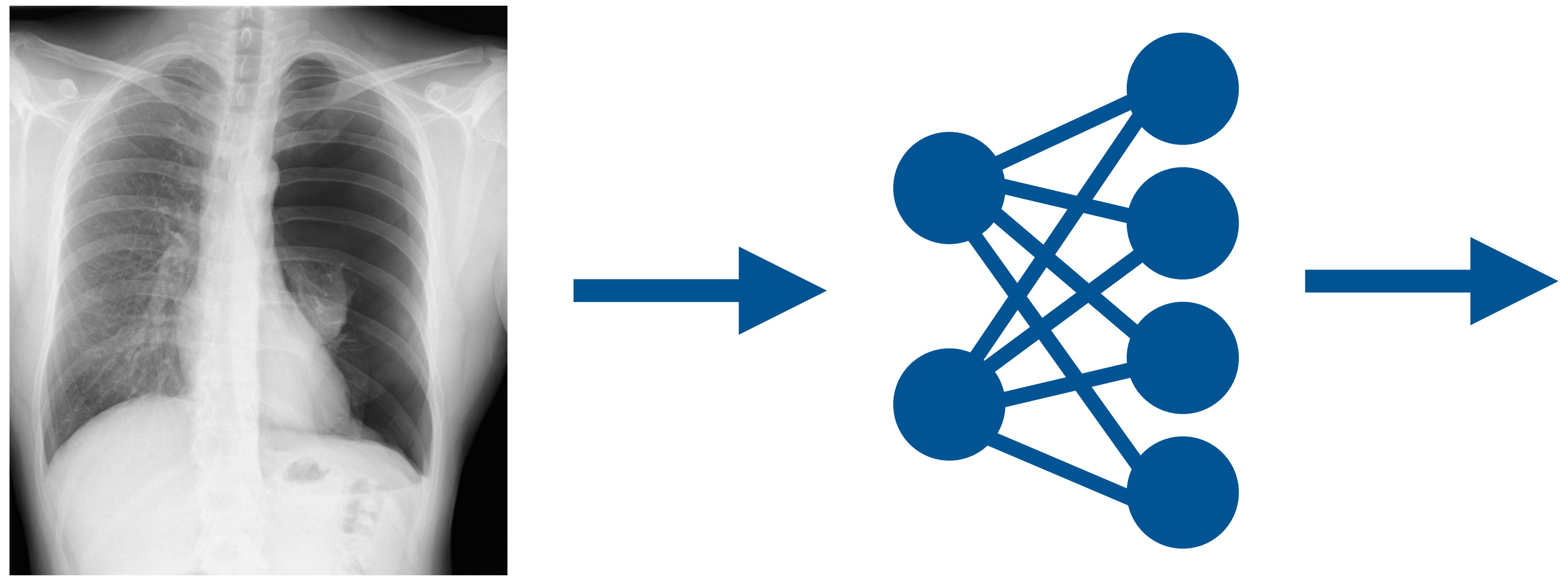
0.98

Classification Performance on subgroup with forest backgrounds:

0.32

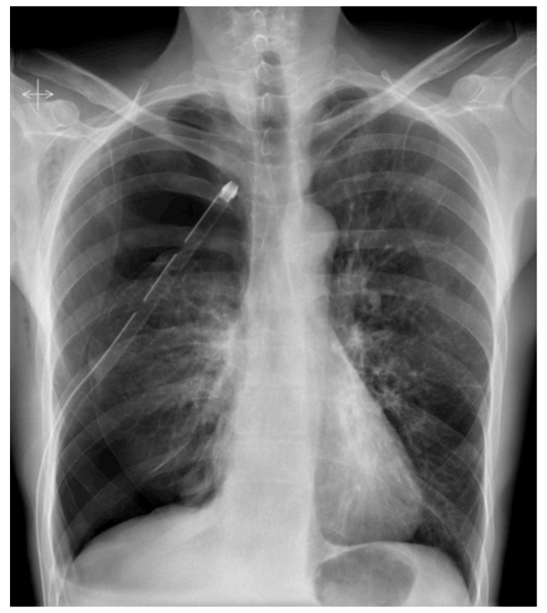
Application 1: Discovering Systematic Errors

Computer vision models often demonstrate high overall performance…



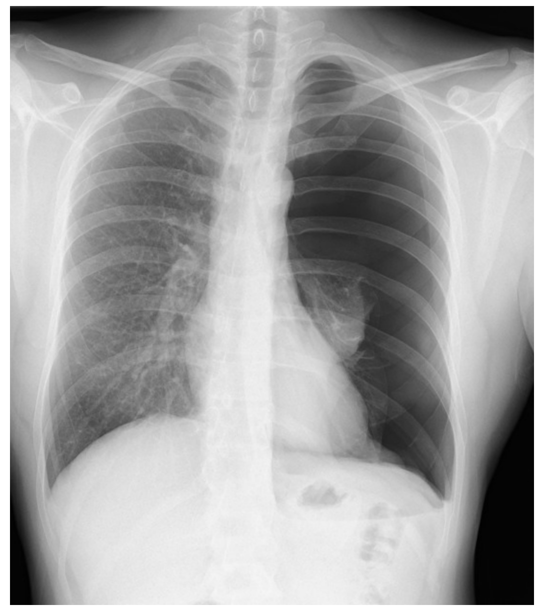
Classification Performance: 0.87

…yet make systematic errors on specific data subgroups



Classification Performance on subgroup with chest tubes:

0.94



Classification Performance on subgroup without chest tubes:

0.77

Key Challenge: Subgroups are not labeled!

Oakden-Rayner et al. "Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging"

Application 1: Discovering Systematic Errors

Domino uses vision-language embeddings to identify and describe systematic prediction errors.

Key Assumption: **Access to validation dataset with predictions and ground-truth labels.**

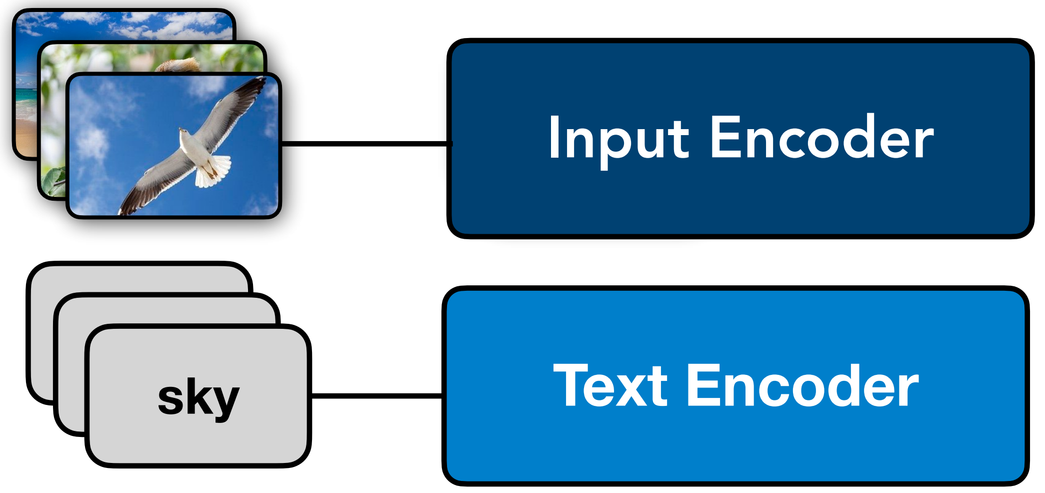
Eyuboglu*, Varma*, Saab*, et al. “Domino: Discovering Systematic Errors with Cross-Modal Embeddings”

Application 1: Discovering Systematic Errors

Domino uses vision-language embeddings to identify and describe systematic prediction errors.

Key Assumption: Access to validation dataset with predictions and ground-truth labels.

1 **Embed inputs with vision-language embeddings**



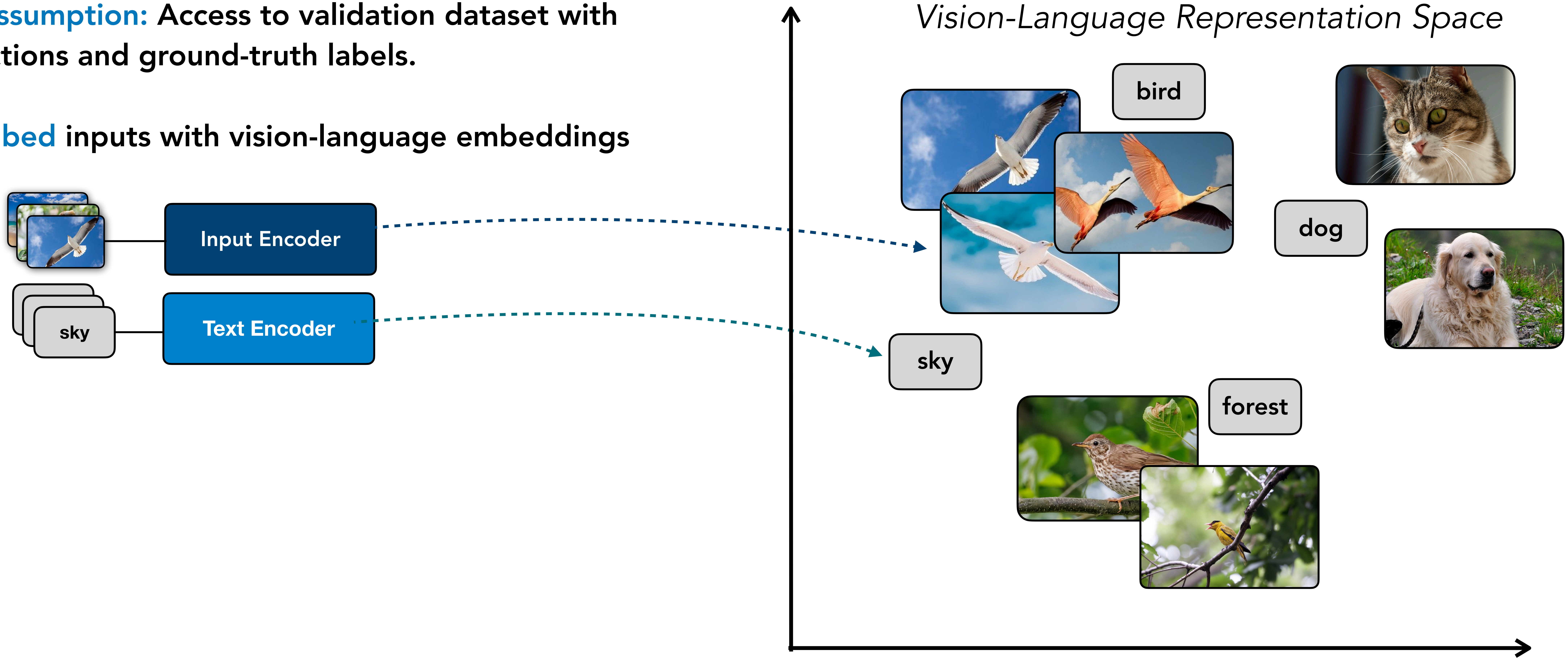
Eyuboglu*, Varma*, Saab*, et al. "Domino: Discovering Systematic Errors with Cross-Modal Embeddings"

Application 1: Discovering Systematic Errors

Domino uses vision-language embeddings to identify and describe systematic prediction errors.

Key Assumption: Access to validation dataset with predictions and ground-truth labels.

1 **Embed inputs with vision-language embeddings**



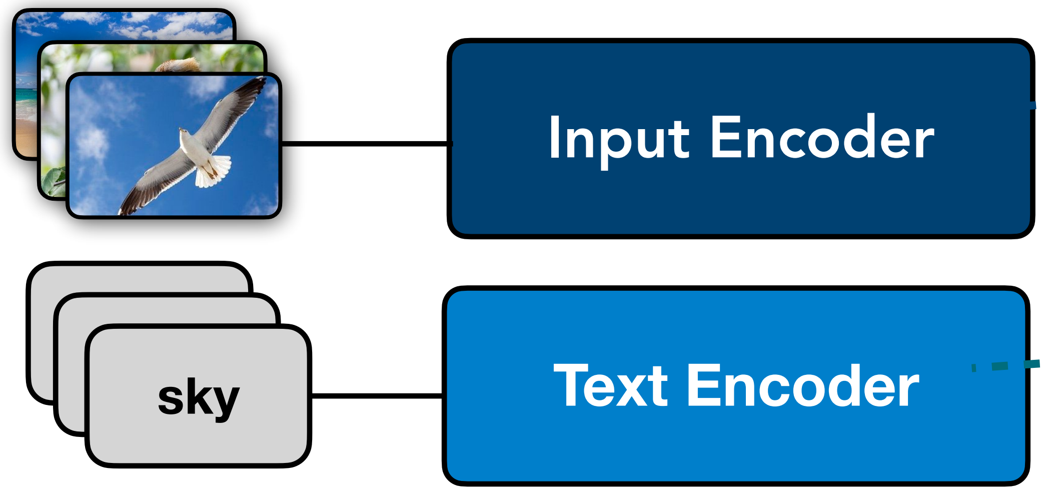
Eyuboglu*, Varma*, Saab*, et al. "Domino: Discovering Systematic Errors with Cross-Modal Embeddings"

Application 1: Discovering Systematic Errors

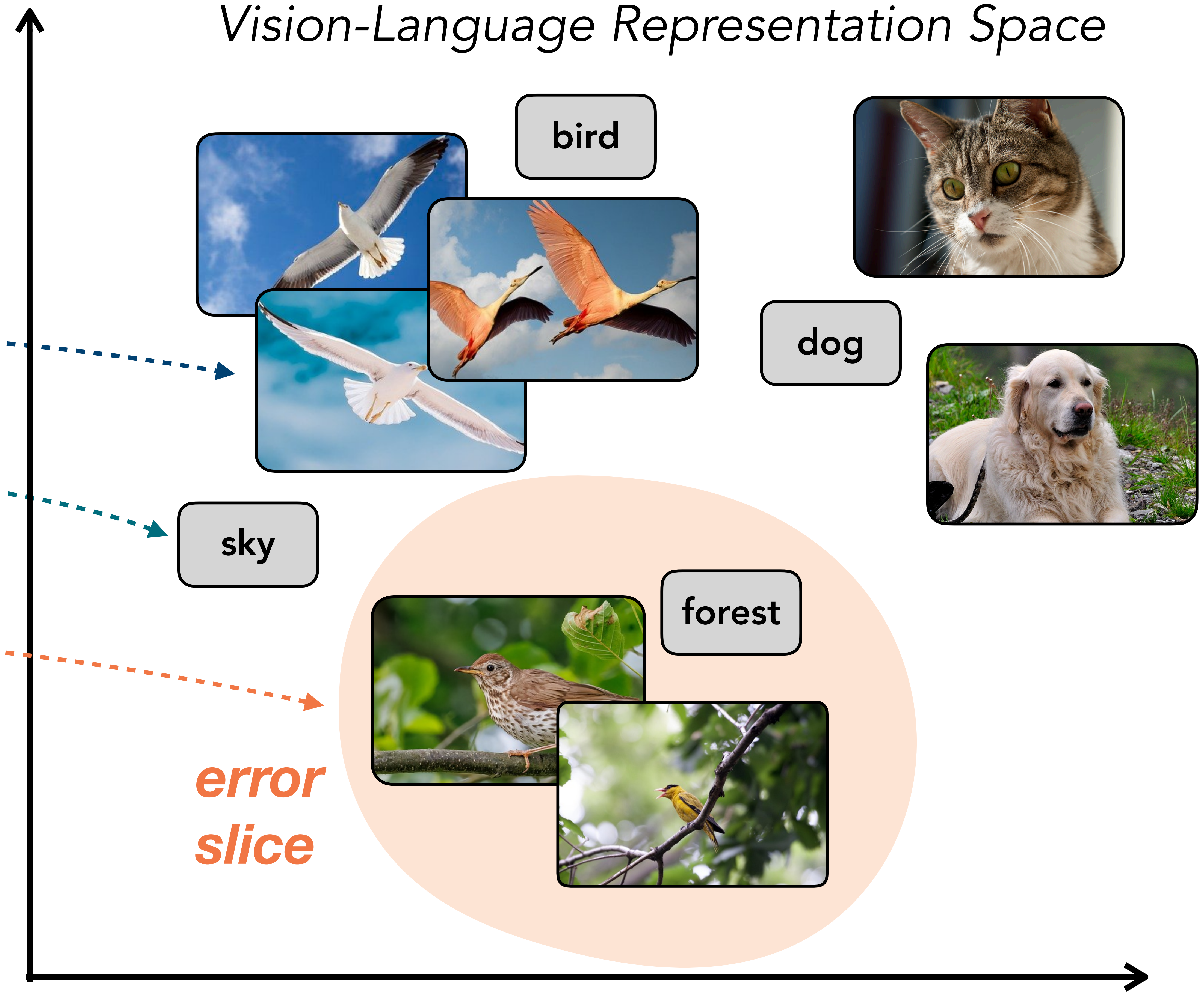
Domino uses vision-language embeddings to identify and describe systematic prediction errors.

Key Assumption: **Access to validation dataset with predictions and ground-truth labels.**

1 **Embed inputs with vision-language embeddings**



2 **Slice to identify high-error regions**



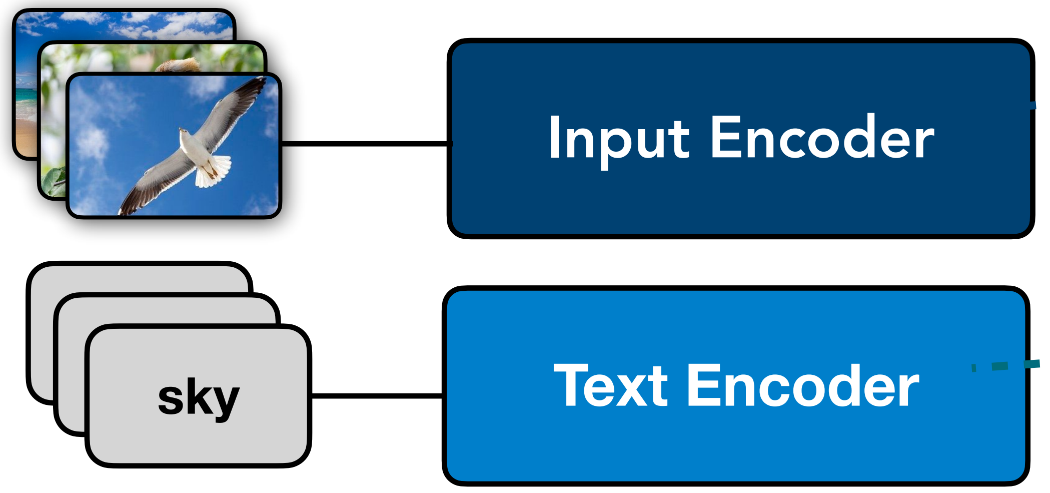
Eyuboglu*, Varma*, Saab*, et al. "Domino: Discovering Systematic Errors with Cross-Modal Embeddings"

Application 1: Discovering Systematic Errors

Domino uses vision-language embeddings to identify and describe systematic prediction errors.

Key Assumption: **Access to validation dataset with predictions and ground-truth labels.**

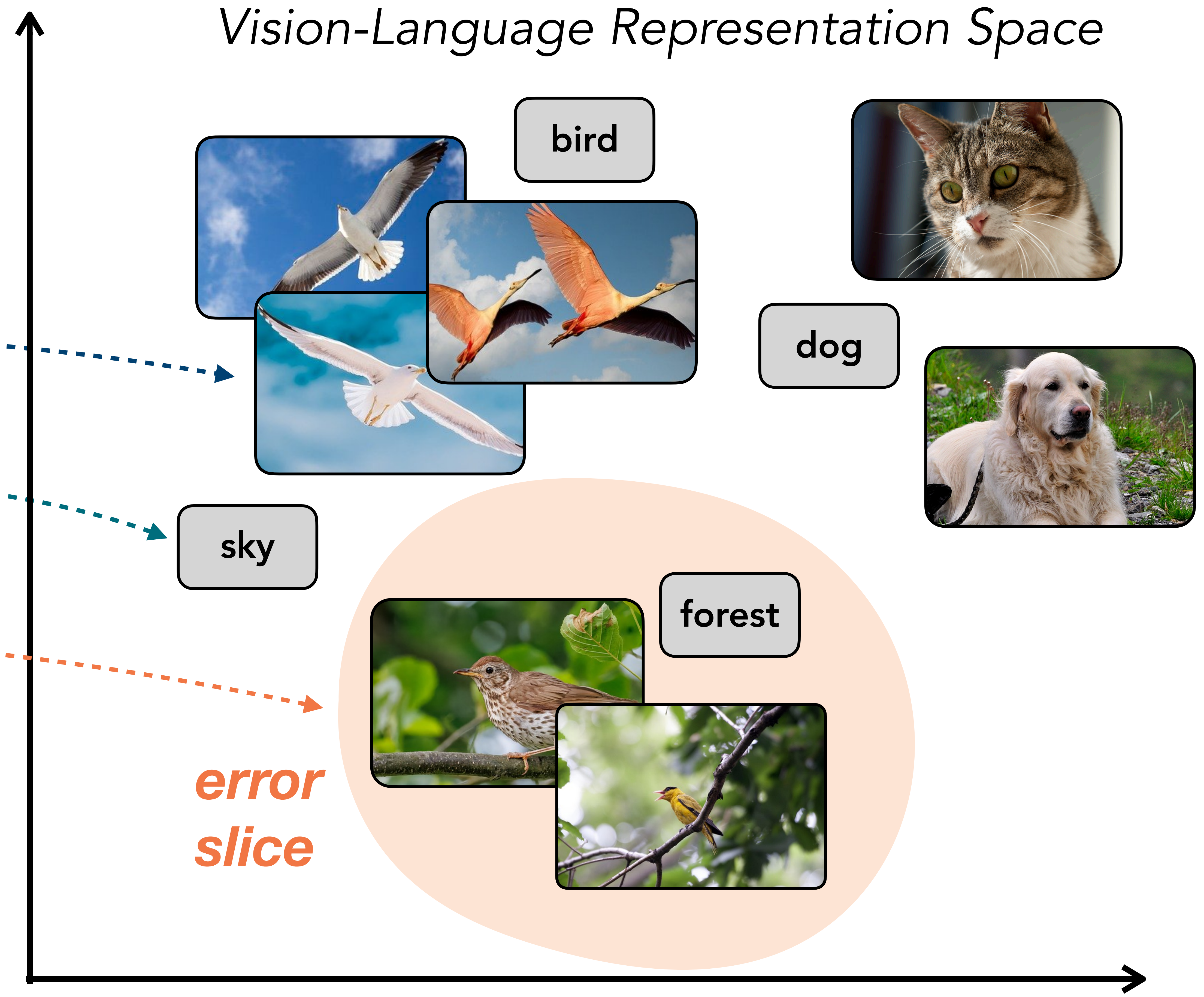
1 Embed **inputs with vision-language embeddings**



2 Slice **to identify high-error regions**

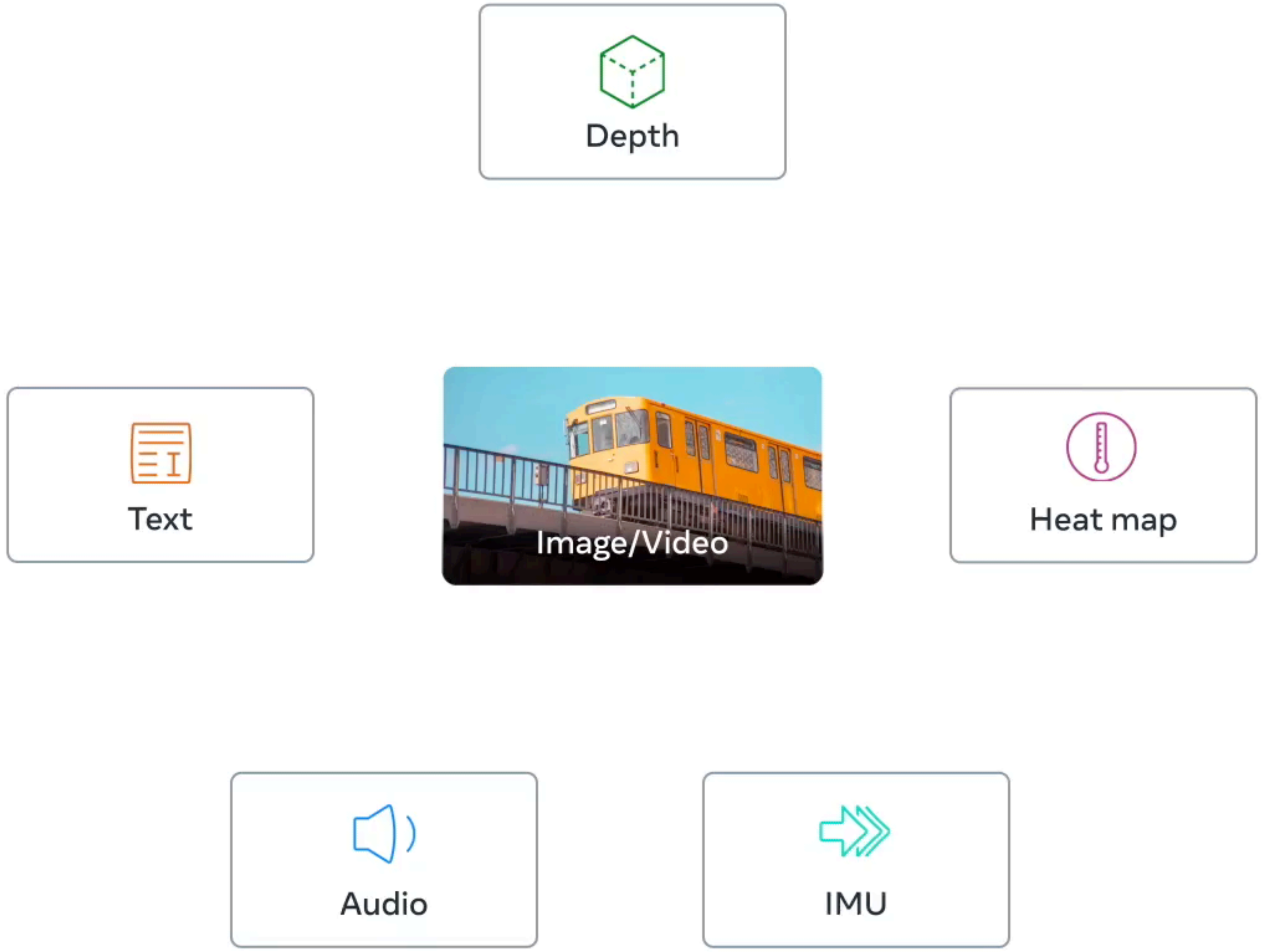
3 Describe **errors with natural language**

birds in forests

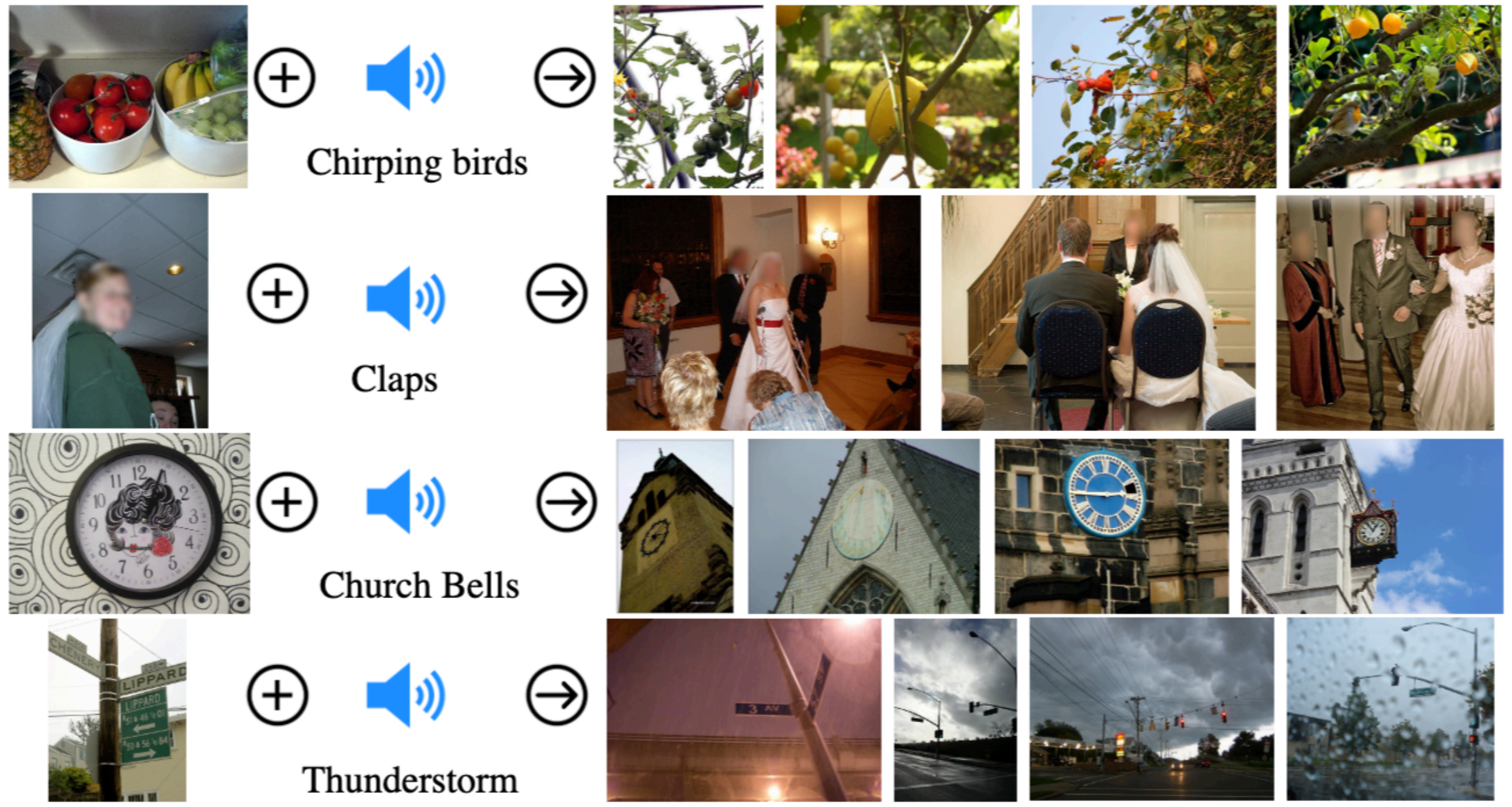


Eyuboglu*, Varma*, Saab*, et al. "Domino: Discovering Systematic Errors with Cross-Modal Embeddings"

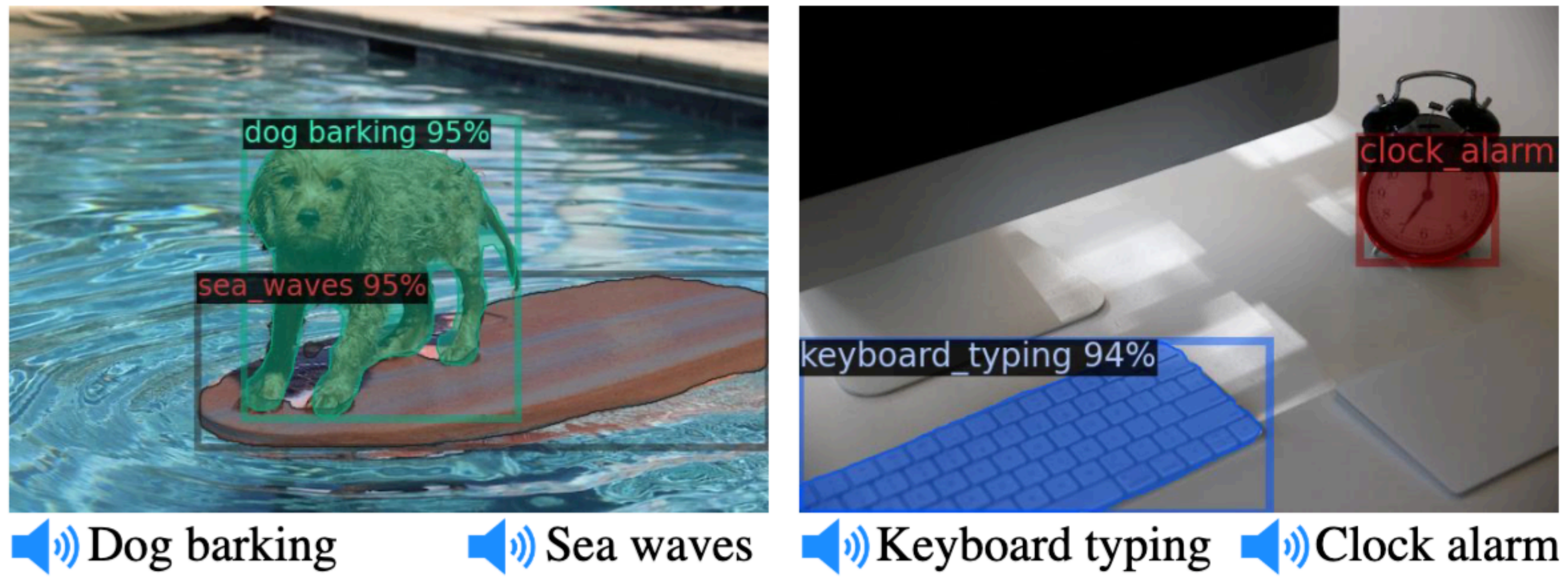
Application 2: Beyond Vision-Language



Embedding Arithmetic with Images and Audio



Object Detection with Audio

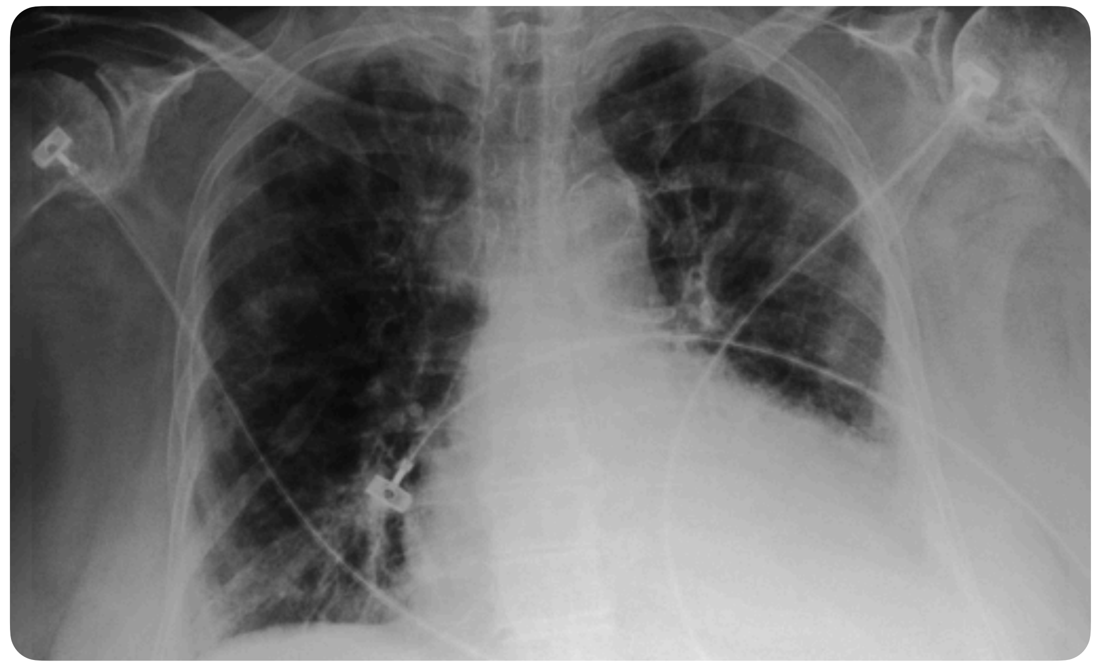


Girdhar et al. “ImageBind: One Embedding Space to Bind them All”

Animation from https://imagebind.metademolab.com/

Application 3: Improving Fine-Grained Reasoning

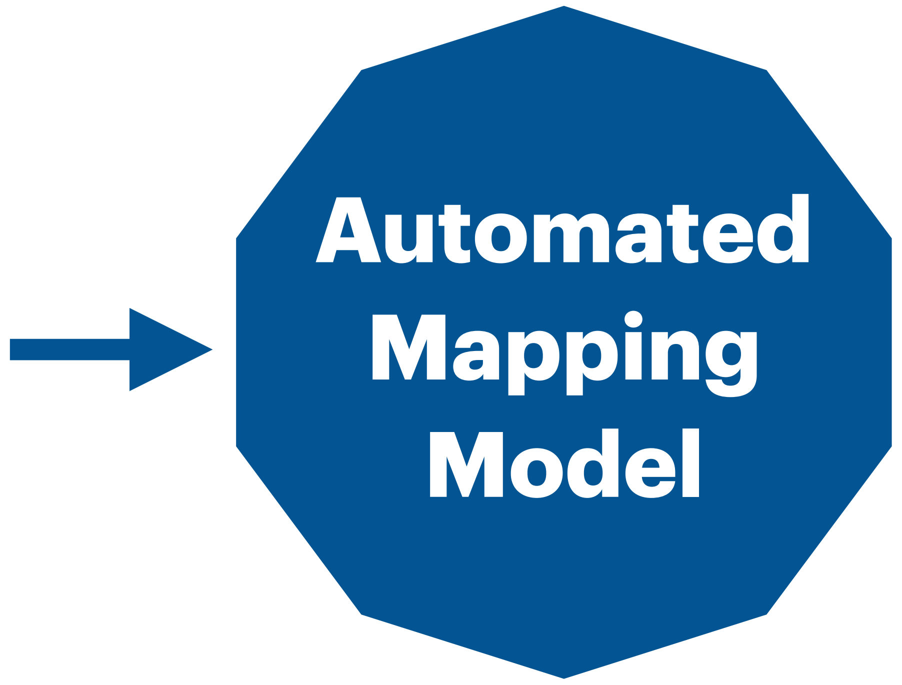
Input



Portable AP chest radiograph. Cardio-mediastinal contours are stable. On the left, there are unchanged areas of basal atelectasis and a moderate left pleural effusion. There is improvement in the pulmonary edema of mid right lung.

Training Stage 1

Goal: Given candidate regions and textual attributes, learn a mapping between these sets.

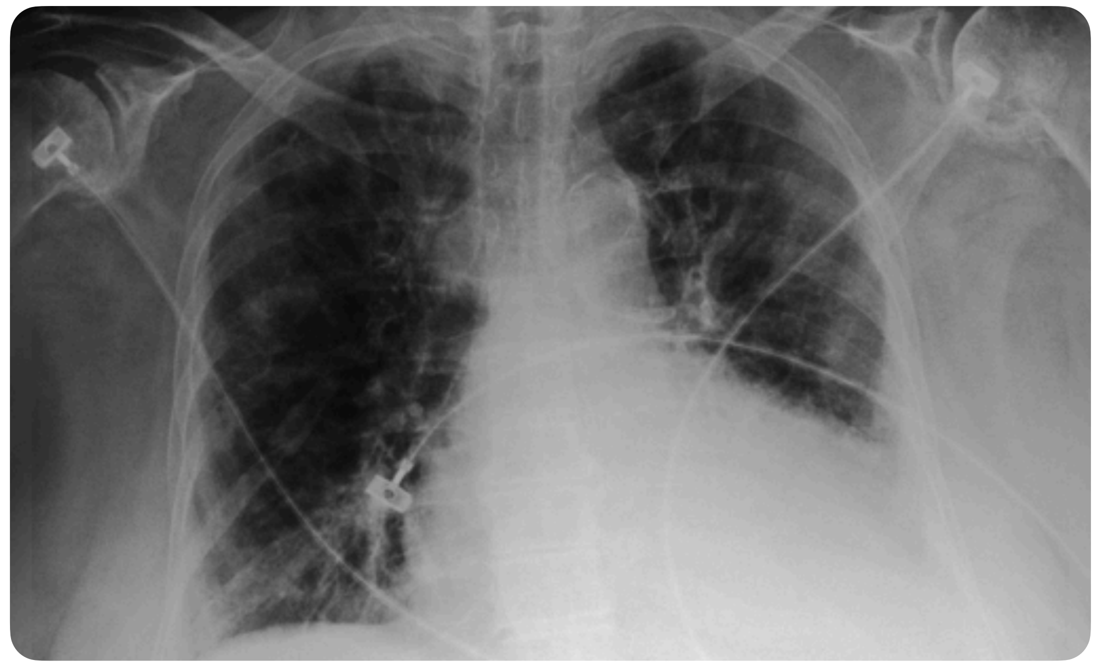


Automated Mapping Model

Varma, et al. "ViLLA: Fine-Grained Vision-Language Representation Learning from Real-World Data"

Application 3: Improving Fine-Grained Reasoning

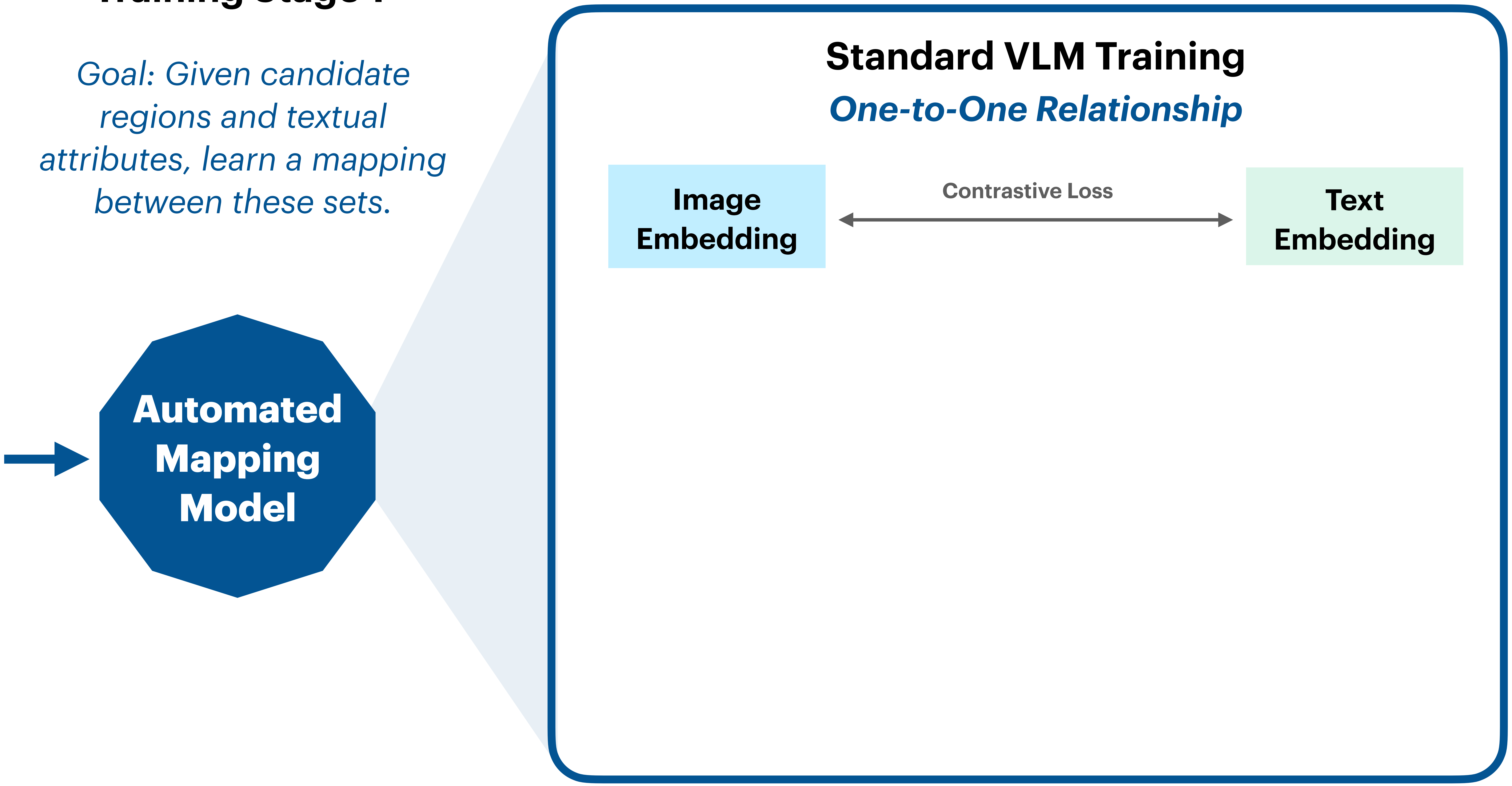
Input



Portable AP chest radiograph. Cardio-mediastinal contours are stable. On the left, there are unchanged areas of basal atelectasis and a moderate left pleural effusion. There is improvement in the pulmonary edema of mid right lung.

Training Stage 1

Goal: Given candidate regions and textual attributes, learn a mapping between these sets.



Varma, et al. "ViLLA: Fine-Grained Vision-Language Representation Learning from Real-World Data"

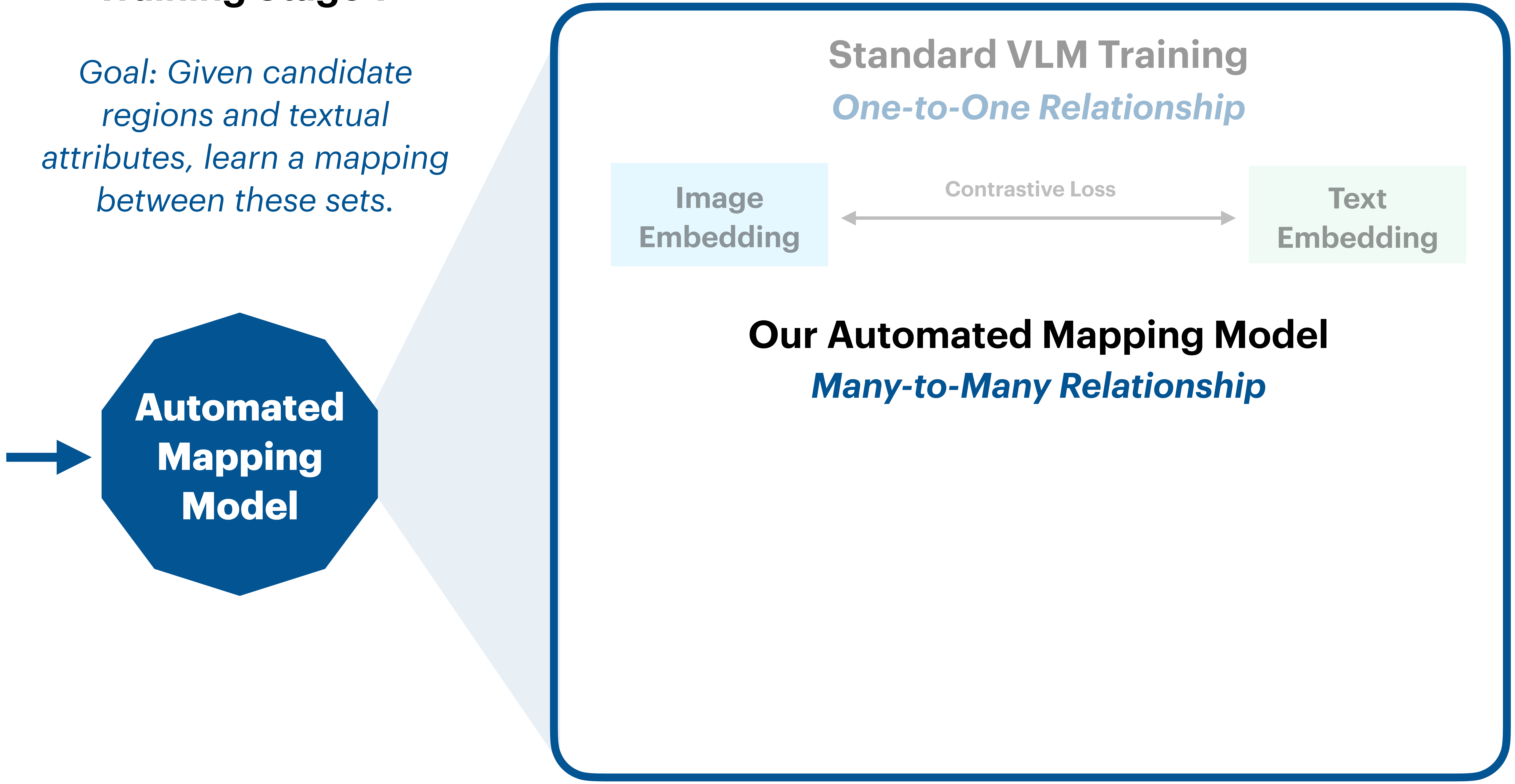
Application 3: Improving Fine-Grained Reasoning

Input

Portable AP chest radiograph. Cardio-mediastinal contours are stable. On the left, there are unchanged areas of basal atelectasis and a moderate left pleural effusion. There is improvement in the pulmonary edema of mid right lung.

Training Stage 1

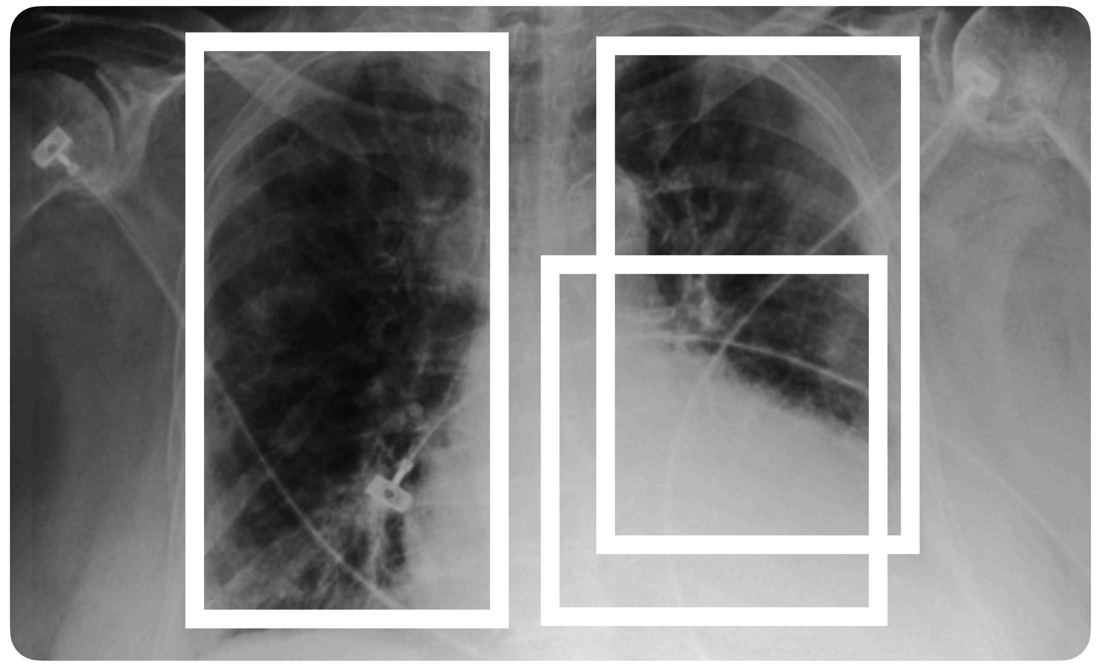
Goal: Given candidate regions and textual attributes, learn a mapping between these sets.



Varma, et al. "ViLLA: Fine-Grained Vision-Language Representation Learning from Real-World Data"

Application 3: Improving Fine-Grained Reasoning

Input

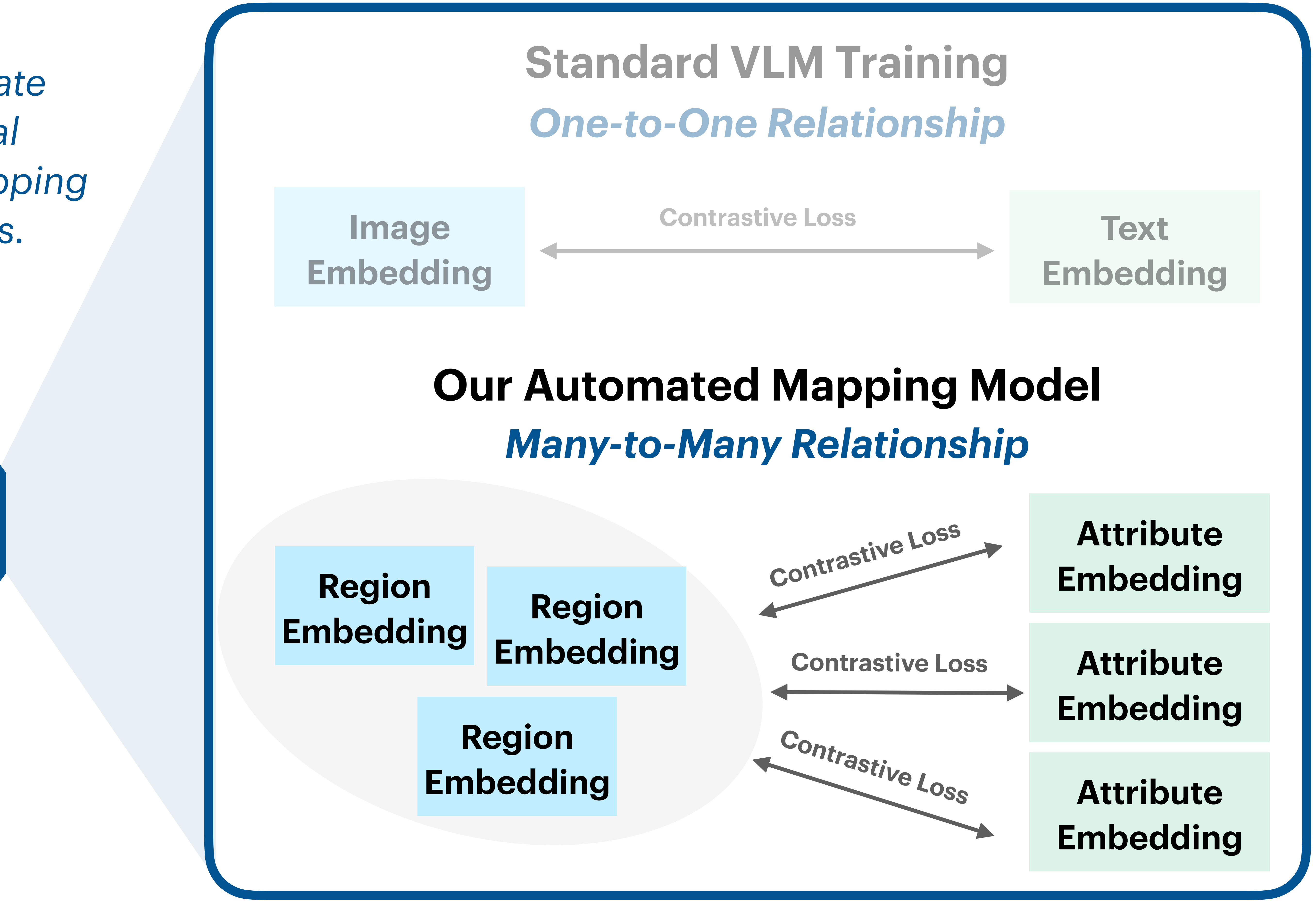


Portable AP chest radiograph. ***Cardio-mediastinal contours are stable****. On the left, there are unchanged* ***areas of basal atelectasis*** *and a* ***moderate left pleural effusion****. There is improvement in the* ***pulmonary edema of mid right lung****.*

Training Stage 1

Goal: Given candidate regions and textual attributes, learn a mapping between these sets.

Automated Mapping Model



Standard VLM Training

One-to-One Relationship

Image Embedding ⟷ Contrastive Loss ⟷ Text Embedding

Our Automated Mapping Model

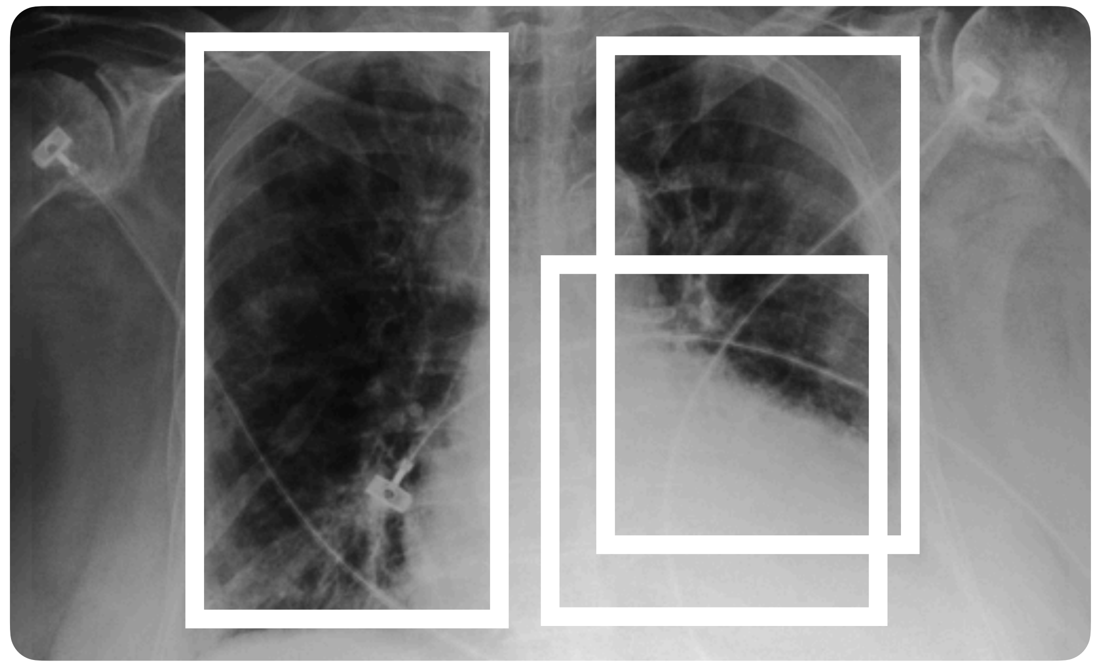
Many-to-Many Relationship

Intuition: *Maximum pairwise similarity between region embeddings and each attribute embedding should be* ***high for positive pairs*** *and* ***low for negative pairs****.*

Varma, et al. "ViLLA: Fine-Grained Vision-Language Representation Learning from Real-World Data"

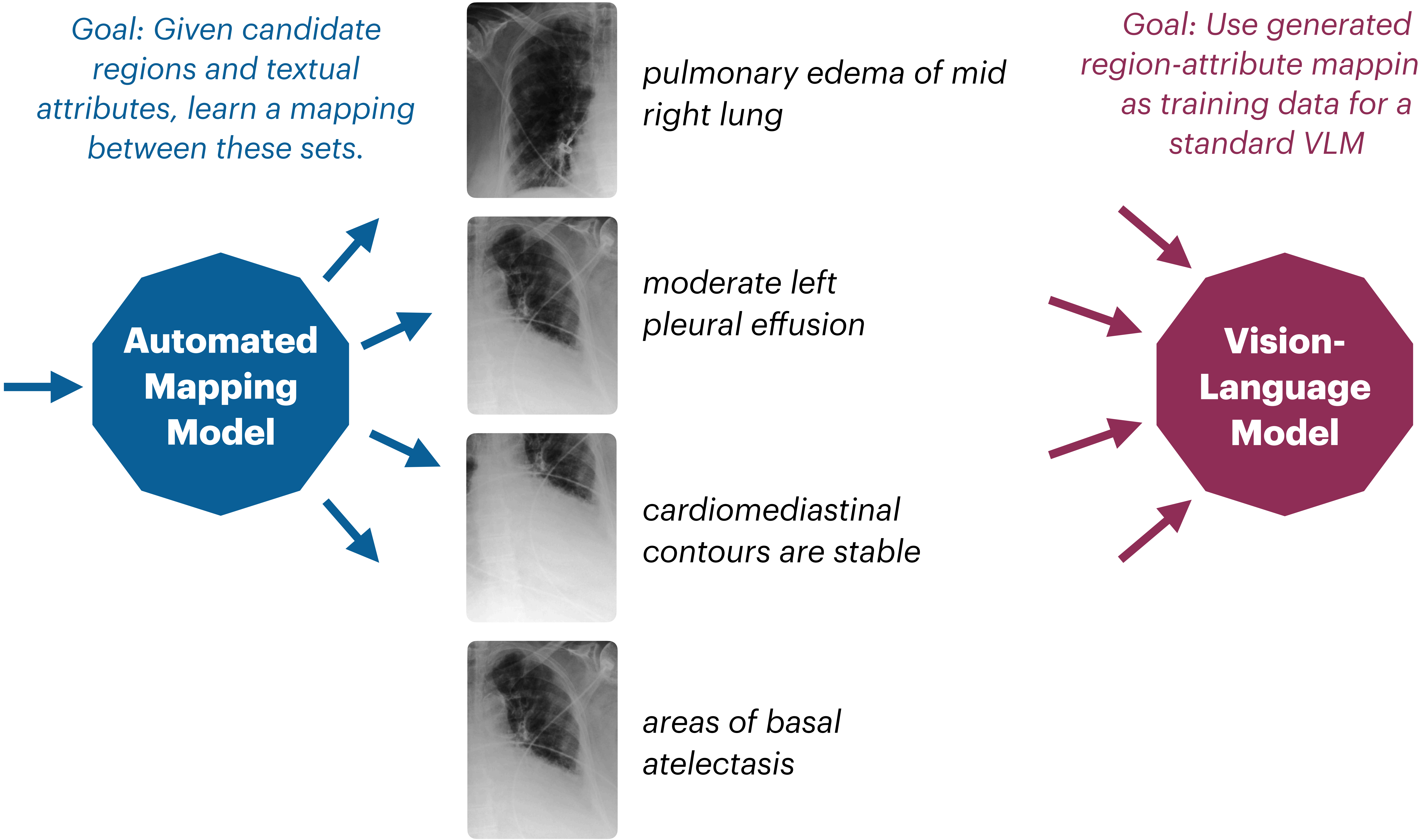
Application 3: Improving Fine-Grained Reasoning

Input



Portable AP chest radiograph. ***Cardio-mediastinal contours are stable****. On the left, there are unchanged* ***areas of basal atelectasis*** *and a* ***moderate left pleural effusion****. There is improvement in the* ***pulmonary edema of mid right lung****.*

Training Stage 1 | **Fine-Grained Region-Attribute Pairs** | **Training Stage 2**



Goal: Given candidate regions and textual attributes, learn a mapping between these sets.

pulmonary edema of mid right lung

moderate left pleural effusion

cardiomediastinal contours are stable

areas of basal atelectasis

Goal: Use generated region-attribute mappings as training data for a standard VLM

Varma, et al. "ViLLA: Fine-Grained Vision-Language Representation Learning from Real-World Data"

Further Reading

3D Vision-Language Representation Learning on Abdominal CTs

Blankemeier et al. “Merlin: A Vision Language Foundation Model for 3D Computed Tomography.” (https://arxiv.org/abs/2406.06512)

Extending CLIP to multiple modalities

Saporta et al. “Contrasting with Symile: Simple Model-Agnostic Representation Learning for Unlimited Modalities.” (https://arxiv.org/abs/2411.01053)

Adapting CLIP for biomedical data

Zhang et al. “BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs.” (https://arxiv.org/abs/2303.00915)

Questions?