

Lecture 6: Vision-Language Representation Learners in Biomedicine

Announcements

- Detailed project instructions are now posted on the course website
- See the end of the project instructions document for example project ideas
- Project proposal is due Wed Oct 23
- A1 is due Wed Oct 16

Finishing up from last lecture: Vision-Language Representation Learners

Language Supervision is not New

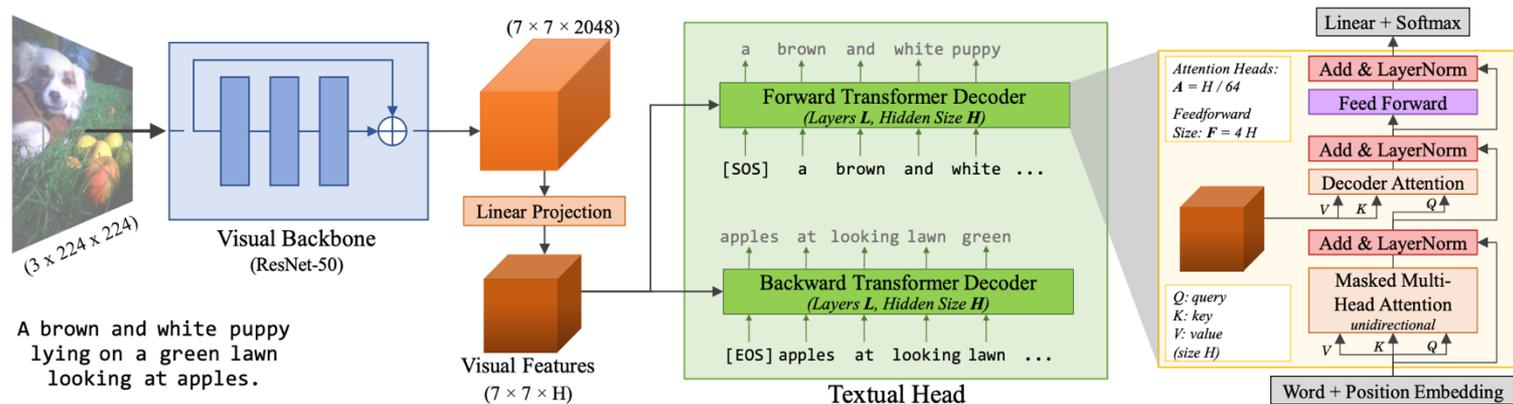
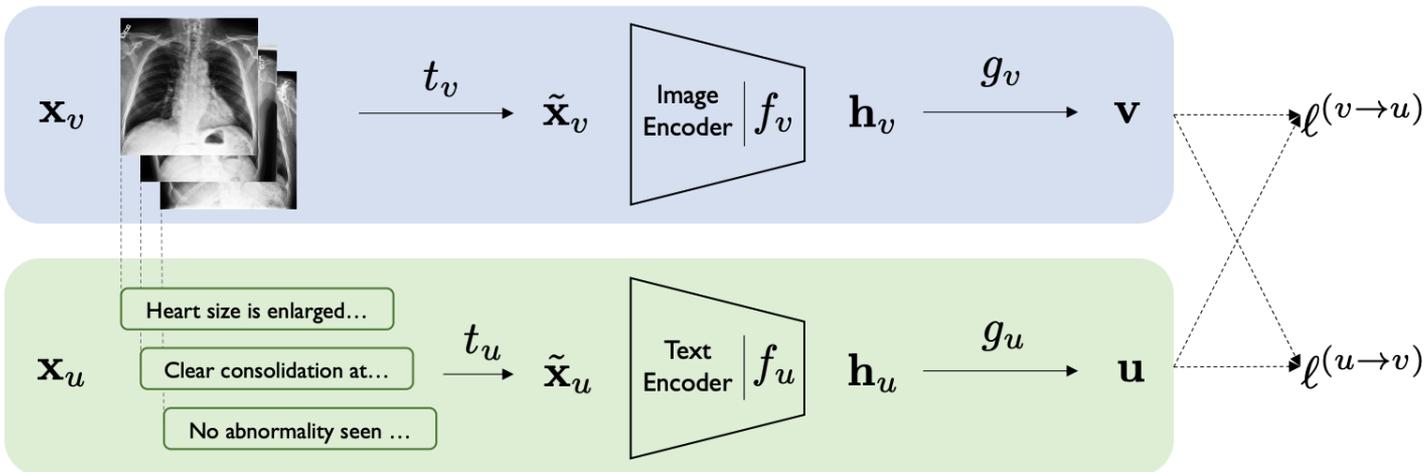


Image-conditioned Bi-directional Next Token Prediction

VirTex (Desai, Johnson 2020)

Serena Yeung-Levy
Xiaohan Wang

Language Supervision is not New



Contrastive Learning of Medical Visual Representations from Paired Images and Text

ConVIRT (Zhang et al. 2020)

Serena Yeung-Levy
Xiaohan Wang

BIODS 276: Adv. Topics in CV and Biomedicine

Lecture 6 - 5

Language Supervision is not New

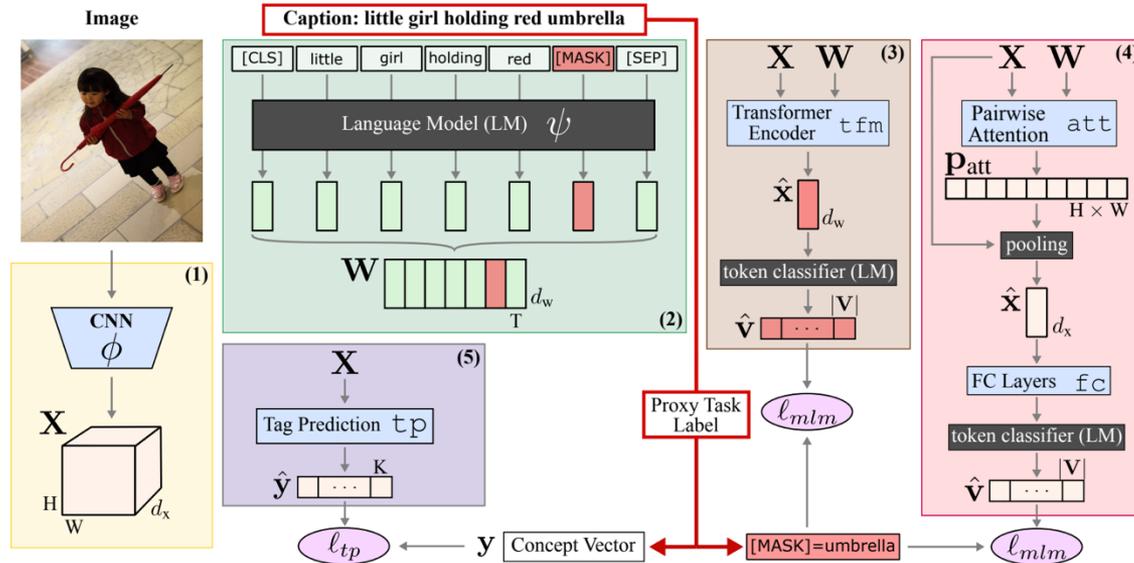
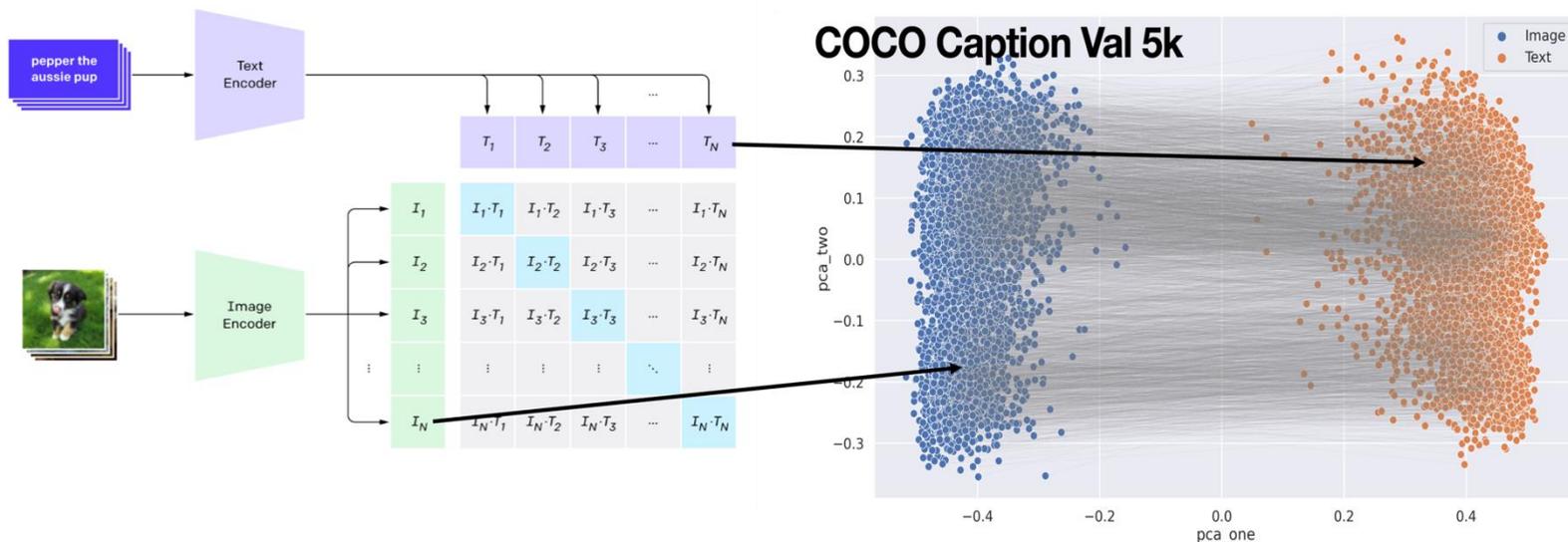


Image-conditioned masked language modeling

Why CLIP Achieves Such Success

- Transferability:
 - Bridging Vision and Language
- Scalability:
 - Large-Scale Pretraining on 400M Data
- Simplicity:
 - Contrastive Learning Enables the Large-Scale Pre-Training

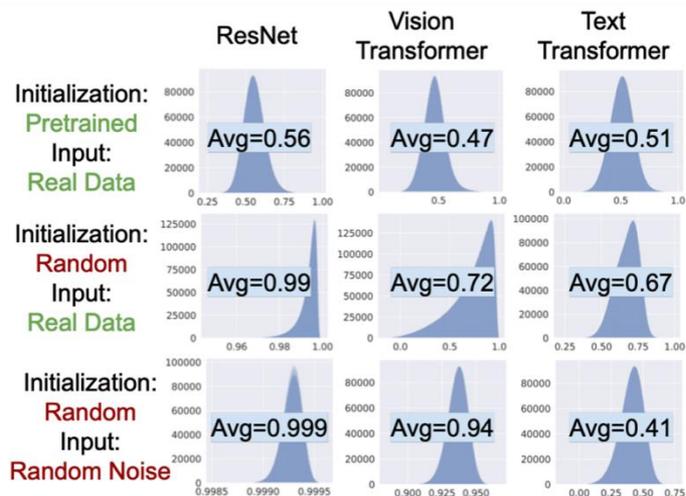
Understanding the Embedding Space of CLIP



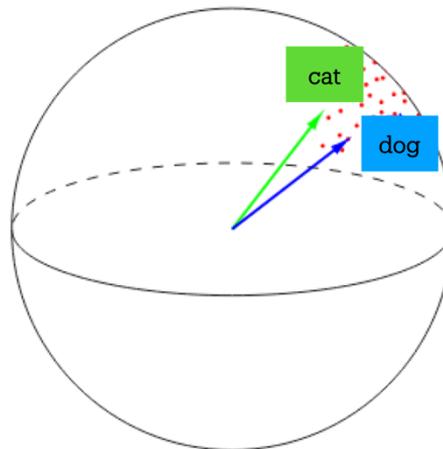
Modality Gap Phenomenon: Paired text embeddings and visual embeddings are not exactly matched

Modality Gap (Zhang et al. 2022)

Understanding the Embedding Space of CLIP

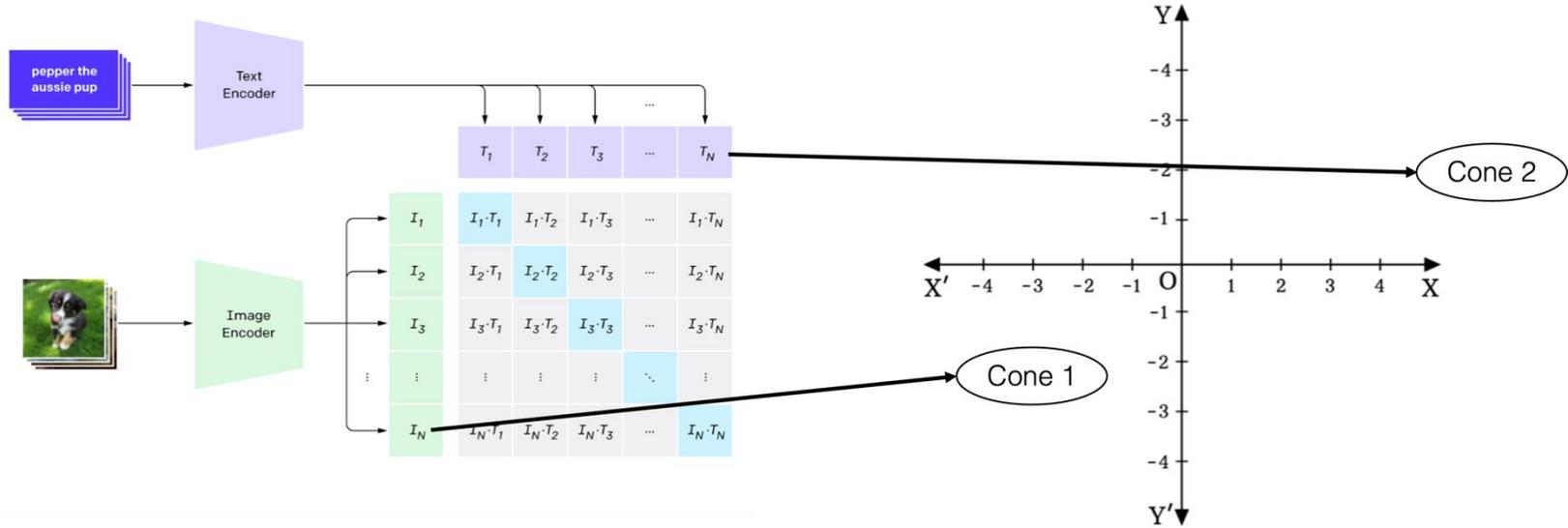


(a) Cosine similarity between random pairs of output features



Cone Effect: General Phenomenon for Any Deep Neural Network

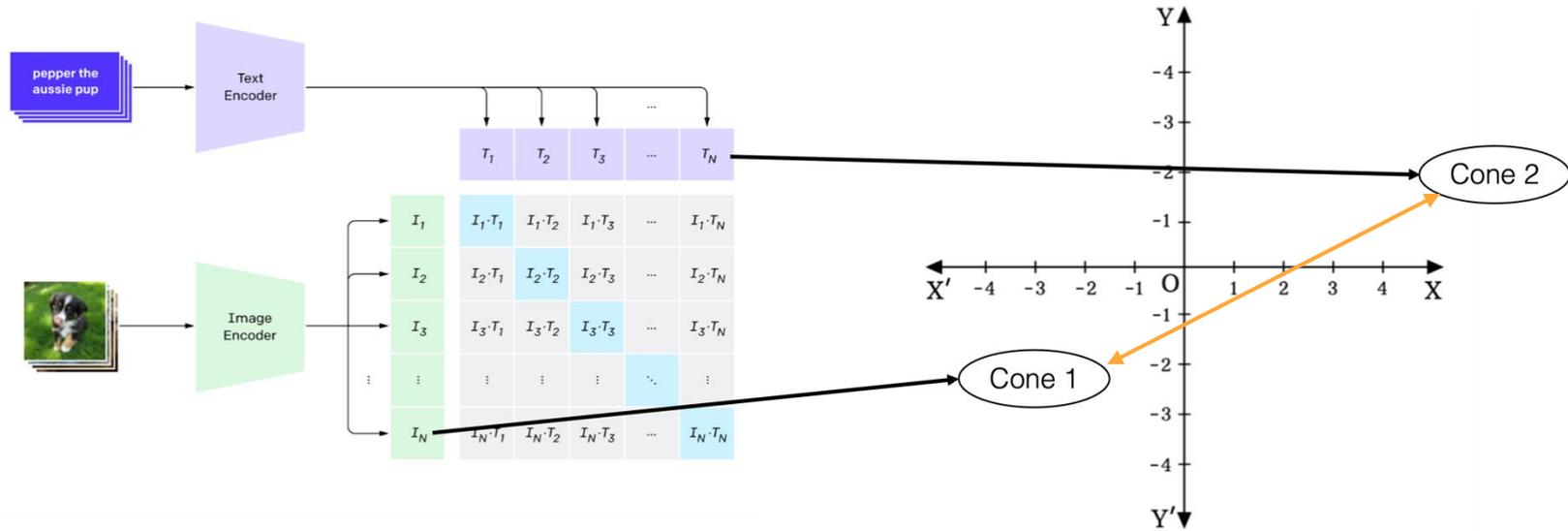
Understanding the Embedding Space of CLIP



Two encoders produce two cones

Modality Gap (Zhang et al. 2022)

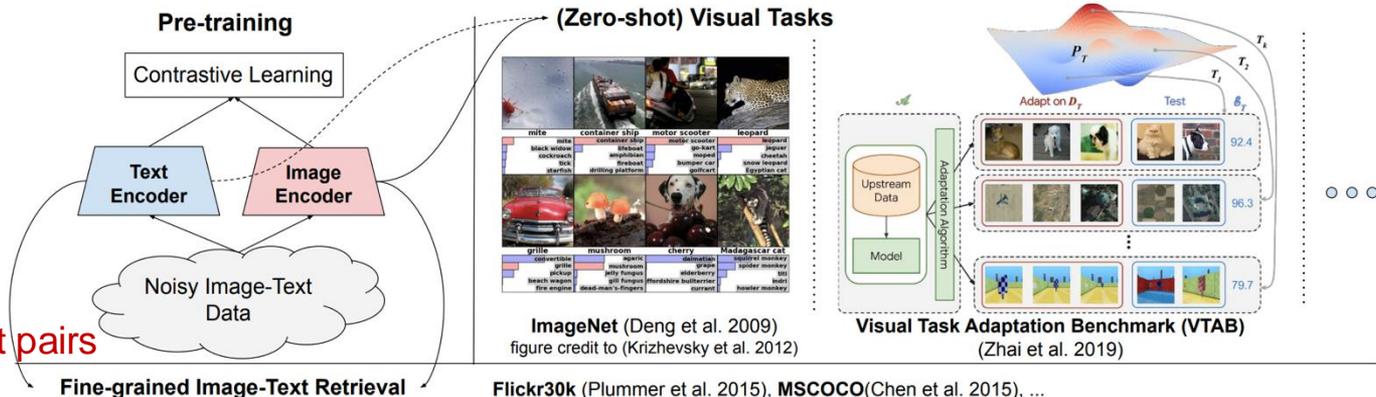
Understanding the Embedding Space of CLIP



Contrastive Learning Preserves the Gap

Modality Gap (Zhang et al. 2022)

Contrastive Learning with Noisy Text Supervision



(A) Text -> Image Retrieval



(B) Image -> Text Retrieval

"original picture of monet haystack"

"monet haystack png"

"haystack series monet art institute of chicago"

...



(C) Image + Text -> Image Retrieval

ALIGN (Jia et al. 2021)

Contrastive Learning with Noisy Text Supervision

Model	ImageNet	ImageNet-R	ImageNet-A	ImageNet-V2
CLIP	76.2	88.9	77.2	70.1
ALIGN	76.4	92.2	75.8	70.1

Comparable Performance with CLIP on ImageNet

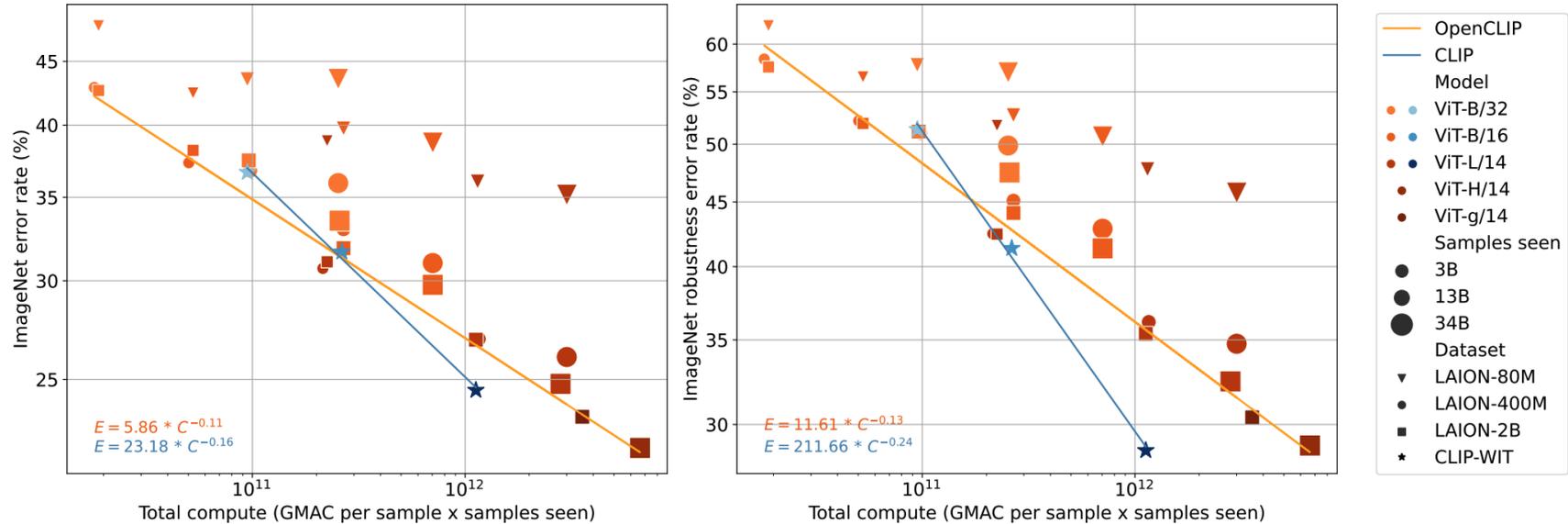
ALIGN (Jia et al. 2021)

How to Reproduce CLIP?

- Data: Open large-scale datasets LAION-400M/2B
- Model:
 - Same transformers as CLIP: ViT-B/32, ViT-B/16, ViT-L/14
 - Even larger transformers: ViT-H/14 and ViT-G/14
- Compute: up to 1520 NVIDIA A100 GPUs

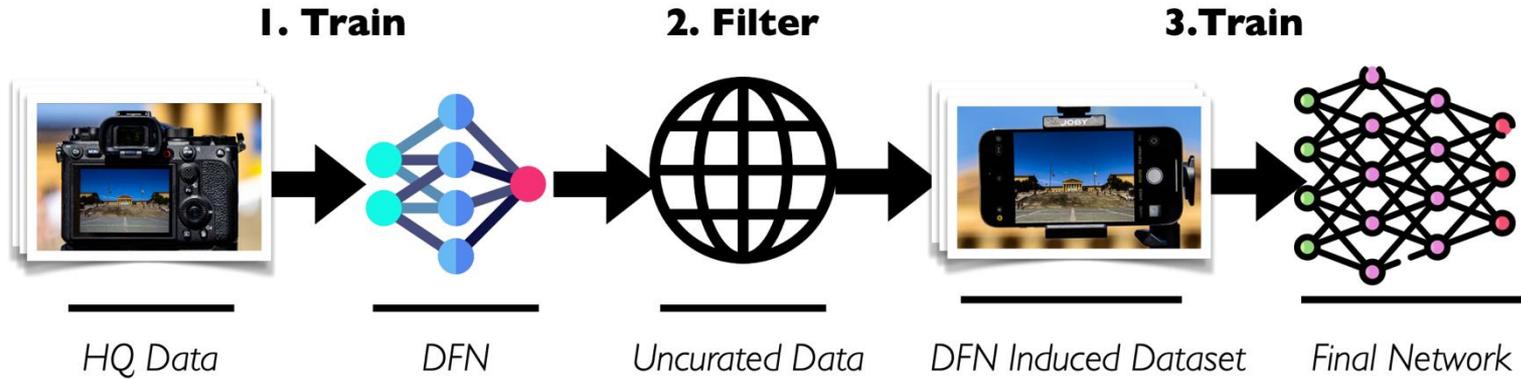
OpenCLIP (Cherti et al. 2022)

How to Reproduce CLIP?



OpenCLIP (Cherti et al. 2022)

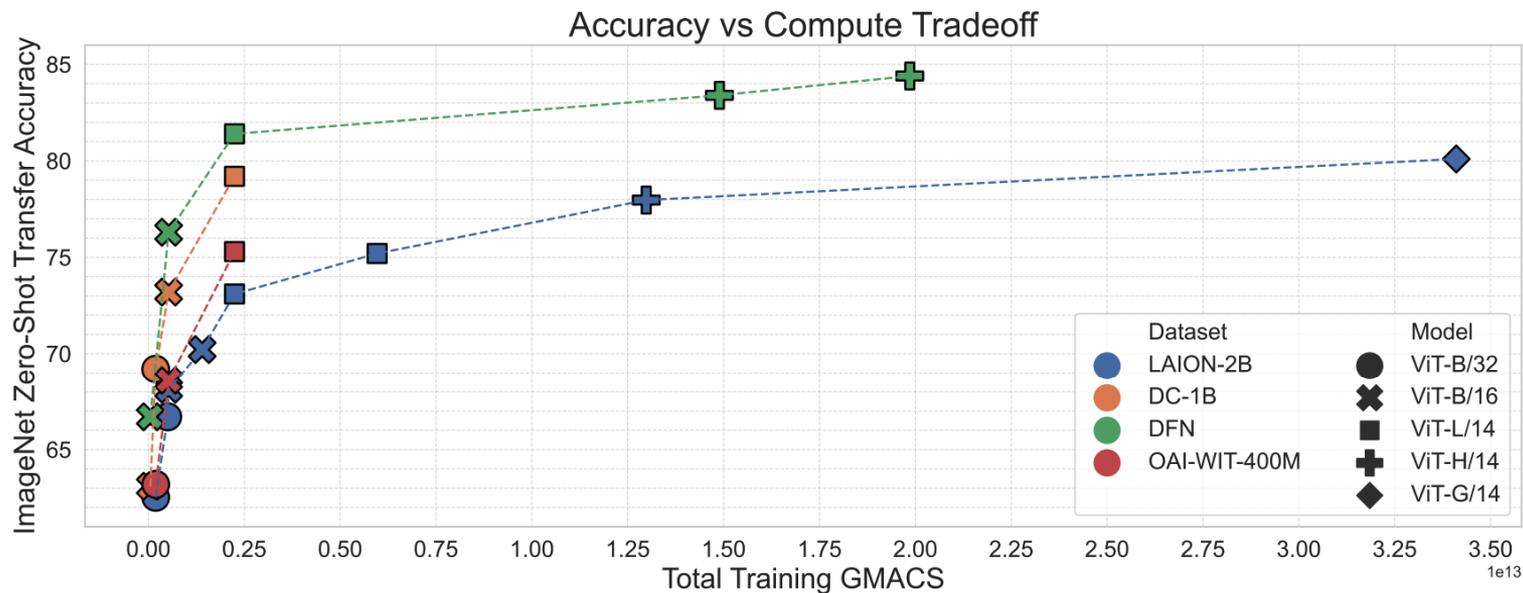
How to Improve CLIP?



Larger and Better Data

DFN (Fang et al. 2023)

How to Improve CLIP?



DFN (Fang et al. 2023)

How to Improve CLIP?

Modified Loss Function to decouples the batch size from the comparison

$$-\frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \sum_{j=1}^{|\mathcal{B}|} \underbrace{\log \frac{1}{1 + e^{z_{ij}(-t\mathbf{x}_i \cdot \mathbf{y}_j + b)}}}_{\mathcal{L}_{ij}}$$

Larger Data: WebLI 10B/12B

Algorithm 1 Sigmoid loss pseudo-implementation.

```
1 # img_emb       : image model embedding [n, dim]
2 # txt_emb       : text model embedding [n, dim]
3 # t_prime, b    : learnable temperature and bias
4 # n             : mini-batch size
5
6 t = exp(t_prime)
7 zimg = l2_normalize(img_emb)
8 ztxt = l2_normalize(txt_emb)
9 logits = dot(zimg, ztxt.T) * t + b
10 labels = 2 * eye(n) - ones(n) # -1 with diagonal 1
11 l = -sum(log_sigmoid(labels * logits)) / n
```

How to Improve CLIP?

Method	Image Encoder		ImageNet-1k				COCO R@1	
	ViT size	# Patches	Validation	v2	ReaL	ObjectNet	I → T	T → I
CLIP	B	196	68.3	61.9	-	55.3	52.4	33.1
OpenCLIP	B	196	70.2	62.3	-	56.0	59.4	42.3
EVA-CLIP	B	196	74.7	67.0	-	62.3	58.7	42.2
SigLIP	B	196	76.2	69.6	82.8	70.7	64.4	47.2
SigLIP	B	256	76.7	70.0	83.1	71.3	65.1	47.4
SigLIP	B	576	78.6	72.1	84.5	73.8	67.5	49.7
SigLIP	B	1024	79.2	73.0	84.9	74.7	67.6	50.4
CLIP	L	256	75.5	69.0	-	69.9	56.3	36.5
OpenCLIP	L	256	74.0	61.1	-	66.4	62.1	46.1
CLIPA-v2	L	256	79.7	72.8	-	71.1	64.1	46.3
EVA-CLIP	L	256	79.8	72.9	-	75.3	63.7	47.5
SigLIP	L	256	80.5	74.2	85.9	77.9	69.5	51.1
CLIP	L	576	76.6	72.0	-	70.9	57.9	37.1
CLIPA-v2	L	576	80.3	73.5	-	73.1	65.5	47.2
EVA-CLIP	L	576	80.4	73.8	-	78.4	64.1	47.9
SigLIP	L	576	82.1	75.9	87.0	81.0	70.6	52.7
OpenCLIP	G (2B)	256	80.1	73.6	-	73.0	67.3	51.4
CLIPA-v2	H (630M)	576	81.8	75.6	-	77.4	67.2	49.2
EVA-CLIP	E (5B)	256	82.0	75.7	-	79.6	68.8	51.1
SigLIP	SO (400M)	729	83.2	77.2	87.5	82.9	70.2	52.0

The best CLIP model

2024/10

SigLIP (Zhai et al. 2023)

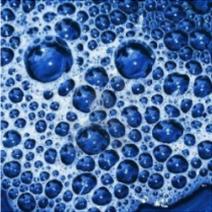
How to Improve CLIP?

Caltech101	Prompt	Accuracy
	a [CLASS].	82.68
	a photo of [CLASS].	80.81
	a photo of a [CLASS].	86.29
	[V]₁ [V]₂ ... [V]_M [CLASS].	91.83

(a)

Flowers102	Prompt	Accuracy
	a photo of a [CLASS].	60.86
	a flower photo of a [CLASS].	65.81
	a photo of a [CLASS], a type of flower.	66.14
	[V]₁ [V]₂ ... [V]_M [CLASS].	94.51

(b)

Describable Textures (DTD)	Prompt	Accuracy
	a photo of a [CLASS].	39.83
	a photo of a [CLASS] texture.	40.25
	[CLASS] texture.	42.32
	[V]₁ [V]₂ ... [V]_M [CLASS].	63.58

(c)

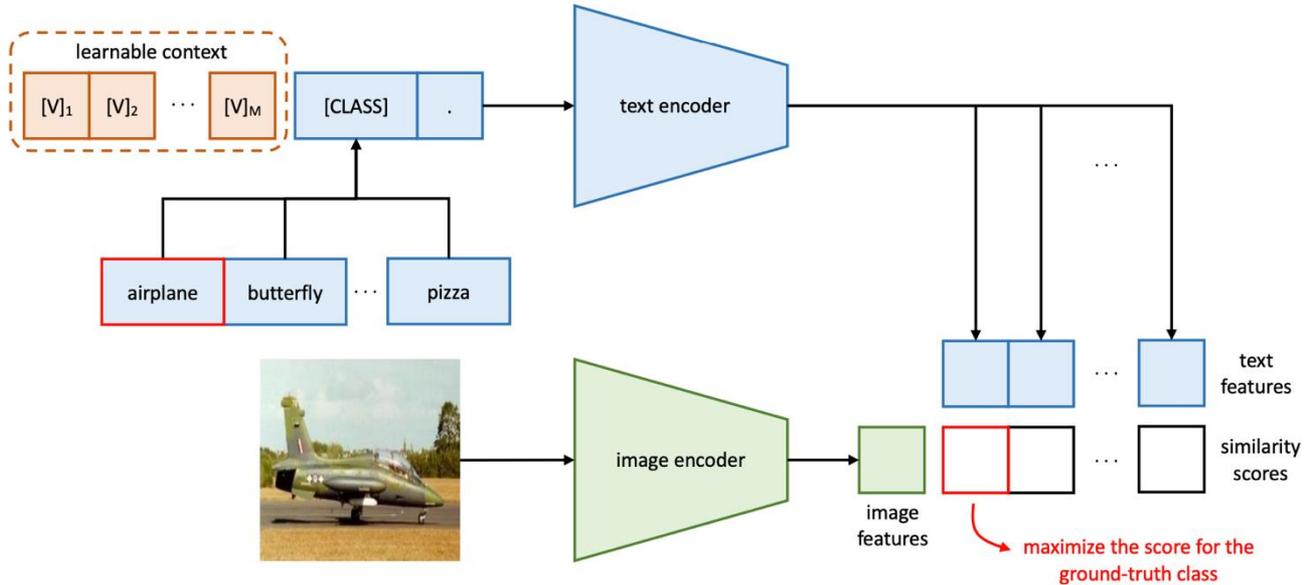
EuroSAT	Prompt	Accuracy
	a photo of a [CLASS].	24.17
	a satellite photo of [CLASS].	37.46
	a centered satellite photo of [CLASS].	37.56
	[V]₁ [V]₂ ... [V]_M [CLASS].	83.53

(d)

Prompt Engineering vs Context Optimization

CoOp (Zhou et al. 2021)

How to Improve CLIP?



Learning to prompt for few-shot classification

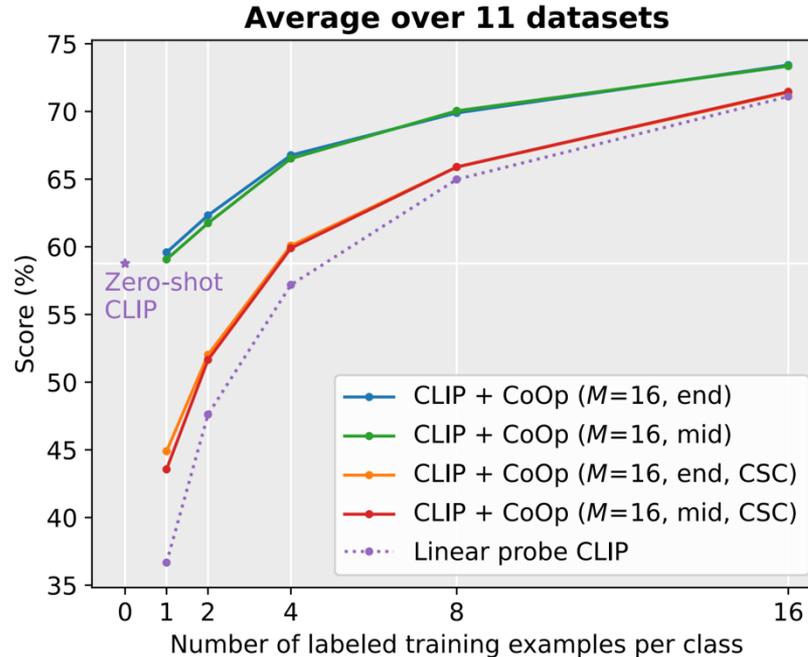
CoOp (Zhou et al. 2021)

Serena Yeung-Levy
Xiaohan Wang

BIODS 276: Adv. Topics in CV and Biomedicine

Lecture 6 - 21

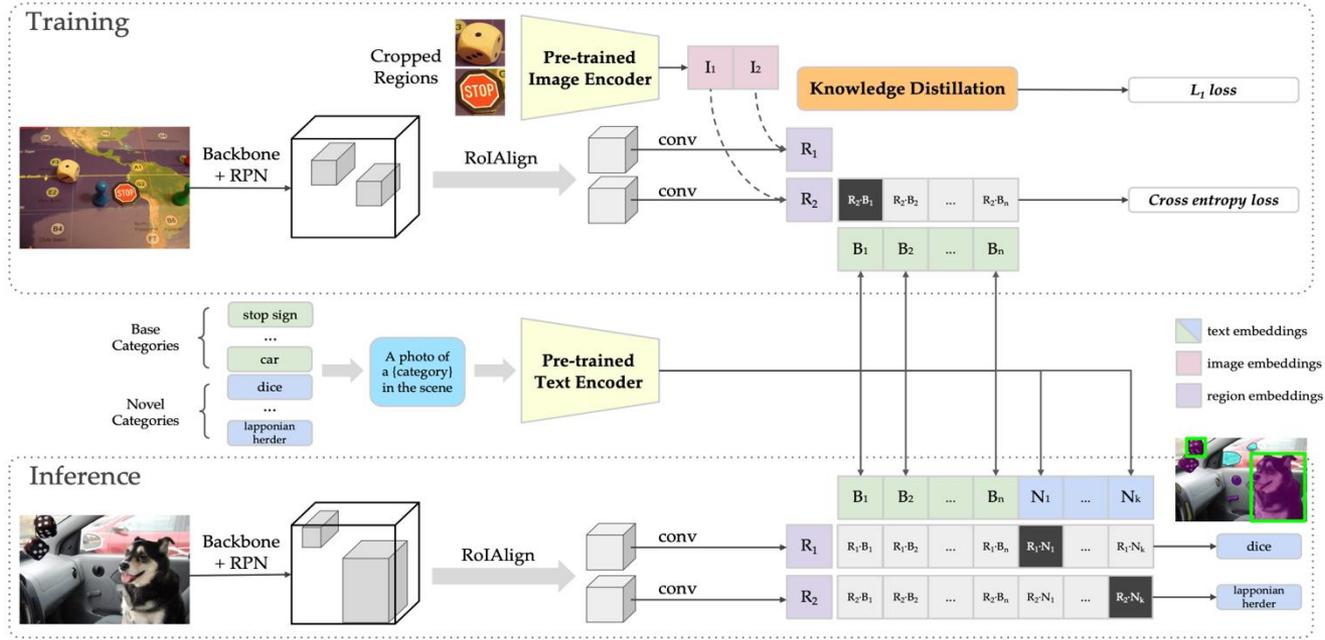
How to Improve CLIP?



M denotes the context length. “end” or “mid” means putting the class token in the end or middle. CSC means class-specific context.

CoOp effectively turns CLIP into a strong few-shot learner

Apply CLIP to Different Tasks



Open-Vocabulary Object Detection

ViLD (Gu et al. 2022)

Serena Yeung-Levy
Xiaohan Wang

BIODS 276: Adv. Topics in CV and Biomedicine

Lecture 6 - 23

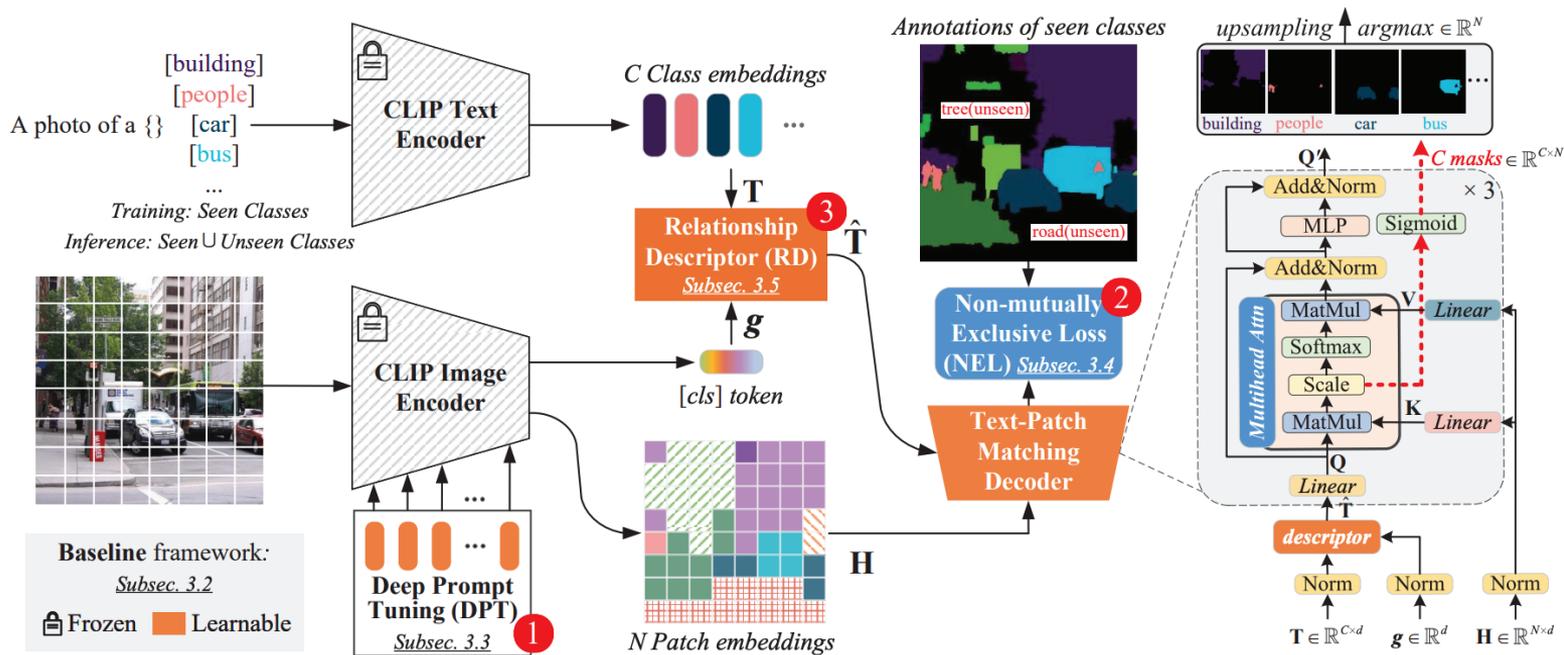
Apply CLIP to Different Tasks

Method	Training source	Novel AP	Base AP	Overall AP
Bilen & Vedaldi (2016)	image-level labels in $C_B \cup C_N$	19.7	19.6	19.6
Ye et al. (2019)		20.3	20.1	20.1
Bansal et al. (2018)	instance-level labels in C_B	0.31	29.2	24.9
Zhu et al. (2020)		3.41	13.8	13.0
Rahman et al. (2020)		4.12	35.9	27.9
Zareian et al. (2021)	image captions in $C_B \cup C_N$ instance-level labels in C_B	22.8	46.0	39.9
CLIP on cropped regions	image-text pairs from Internet (may contain $C_B \cup C_N$)	26.3	28.3	27.8
ViLD-text		5.9	61.8	47.2
ViLD-image		24.1	34.2	31.6
ViLD ($w = 0.5$)		instance-level labels in C_B	27.6	59.5

Open-Vocabulary Object Detection **4.8% ↑**

ViLD (Gu et al. 2022)

Apply CLIP to Different Tasks



CLIP for Zero-shot Semantic Segmentation

ZegCLIP (Zhou et al. 2021)

Serena Yeung-Levy
Xiaohan Wang

BIODS 276: Adv. Topics in CV and Biomedicine

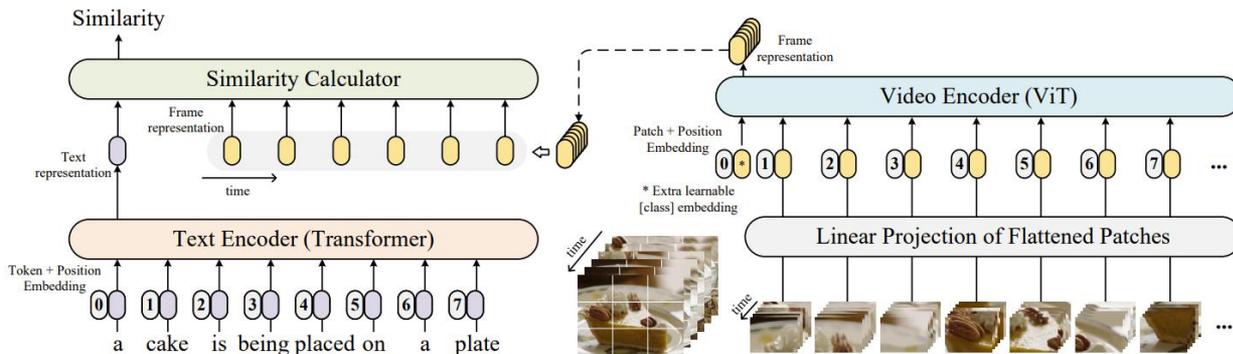
Lecture 6 - 25

Apply CLIP to Different Tasks

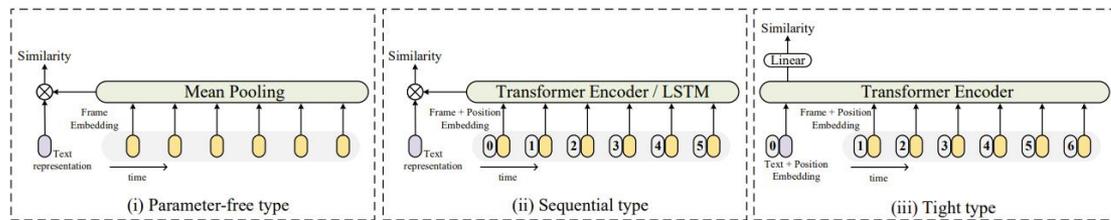
Methods	PASCAL VOC 2012			
	pAcc	mIoU(S)	mIoU(U)	hIoU
<i>Inductive</i>				
SPNet [44]	-	78.0	15.6	26.1
ZS3 [3]	-	77.3	17.7	28.7
CaGNet [17]	80.7	78.4	26.6	39.7
SIGN [10]	-	75.4	28.9	41.7
Joint [1]	-	77.7	32.5	45.9
ZegFormer [12]	-	86.4	63.6	73.3
zsseg [49]	90.0	83.5	72.5	77.5
ZegCLIP (Ours)	94.6	91.9	77.8	84.3

CLIP for Zero-shot Semantic Segmentation **5.3% ↑**

Apply CLIP to Different Tasks



(a) Main structure



(b) Similarity calculator

CLIP for Video Retrieval

Apply CLIP to Different Tasks

Methods	TrainD	E2E	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
C+LSTM+SA ^a	M	✓	4.2	12.9	19.9	55	-
VSE ^b	M	✓	3.8	12.7	17.1	66	-
SNUVL ^c	M	✓	3.5	15.9	23.8	44	-
Kaufman et al. ^d	M	✓	4.7	16.6	24.1	41	-
CT-SAN ^e	M	✓	4.4	16.6	22.3	35	-
JSFusion ^f	M	✓	10.2	31.2	43.2	13	-
HowTo100M ^g	H+M	✓	14.9	40.2	52.8	9	-
ActBERT ^h	H+M		8.6	23.4	33.1	36	-
NoiseE ⁱ	H+M		17.4	41.6	53.6	8	-
UniVL ^j	H+M		21.2	49.6	63.1	6	-
HERO ^k	H+M		16.8	43.4	57.7	-	-
ClipBERT ^l	C+G+M	✓	22.0	46.8	59.9	6	-
(Ours)-meanP	W+M	✓	42.1	71.9	81.4	2	15.7
(Ours)-seqLSTM	W+M	✓	41.7	68.8	78.7	2	16.6
(Ours)-seqTransf	W+M	✓	42.0	68.6	78.7	2	16.2
(Ours)-tightTransf	W+M	✓	37.8	68.4	78.4	2	17.2

(a) Training on Training-7K

Methods	TrainD	E2E	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
MIL-NCE ^m	H	✓	9.9	24.0	32.4	29.5	-
CLIP-straight ⁿ	W	✓	31.2	53.7	64.2	4	-

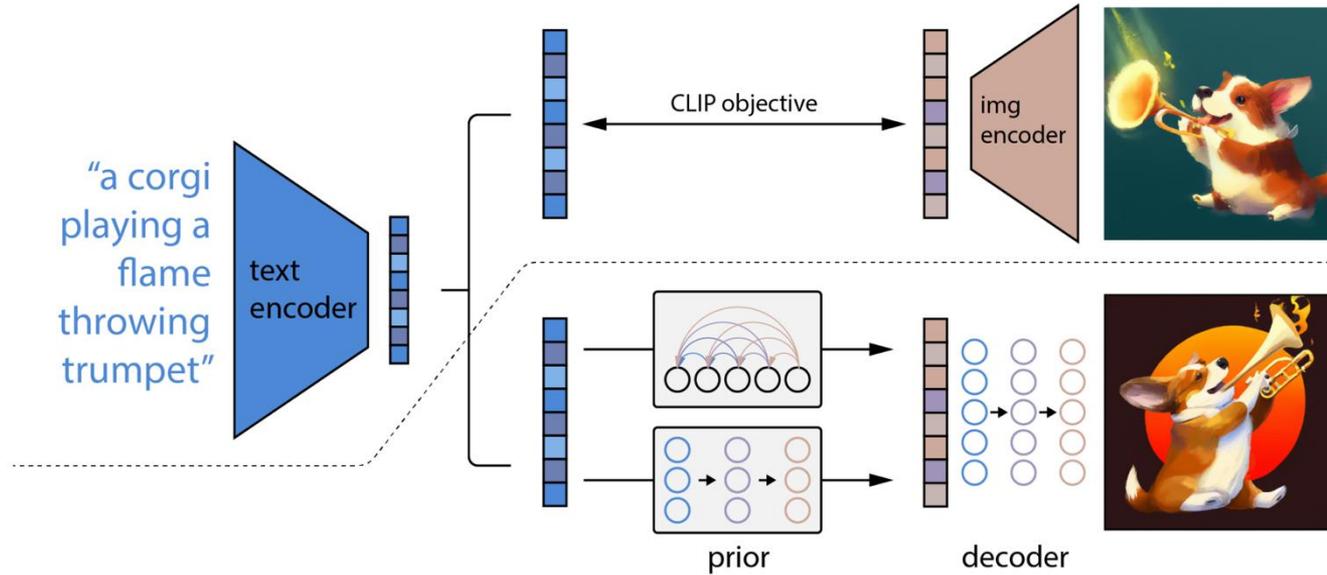
(b) Zero-shot

Methods	TrainD	E2E	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
CE ^o	M		20.9	48.8	62.4	6	28.2
MMT ^p	H+M		26.6	57.1	69.6	4	24.0
AVLnet ^q	H+M		27.1	55.6	66.6	4	-
SSB ^r	H+M		30.1	58.5	69.3	3	-
MDMMT ^s	MD+M		38.9	69.0	79.7	2	16.5
Frozen ^t	CW+M	✓	31.0	59.5	70.5	3	-
HiT ^u	H+M		30.7	60.9	73.2	2.6	-
TT-CE+ ^v	M		29.6	61.6	74.2	3	-
(Ours)-meanP	W+M	✓	43.1	70.4	80.8	2	16.2
(Ours)-seqLSTM	W+M	✓	42.5	70.8	80.7	2	16.7
(Ours)-seqTransf	W+M	✓	44.5	71.4	81.6	2	15.3
(Ours)-tightTransf	W+M	✓	40.2	71.5	80.5	2	13.4

(c) Training on Training-9K

CLIP for Video Retrieval **20.1% ↑**

Apply CLIP to Different Tasks

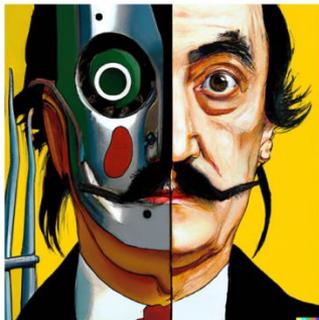


Hierarchical Text-Conditional Image Generation with CLIP Latents

DALL-E 2 (Ramesh et al. 2022)

Apply CLIP to Different Tasks

Lecture 7



vibrant portrait painting of Salvador Dalí with a robotic half face



a shiba inu wearing a beret and black turtleneck



a close up of a handpalm with leaves growing from it



an espresso machine that makes coffee from human souls, artstation



panda mad scientist mixing sparkling chemicals, artstation



a corgi's head depicted as an explosion of a nebula

DALL-E 2 (Ramesh et al. 2022) Hierarchical Text-Conditional Image Generation with CLIP Latents

Apply CLIP to Different Tasks

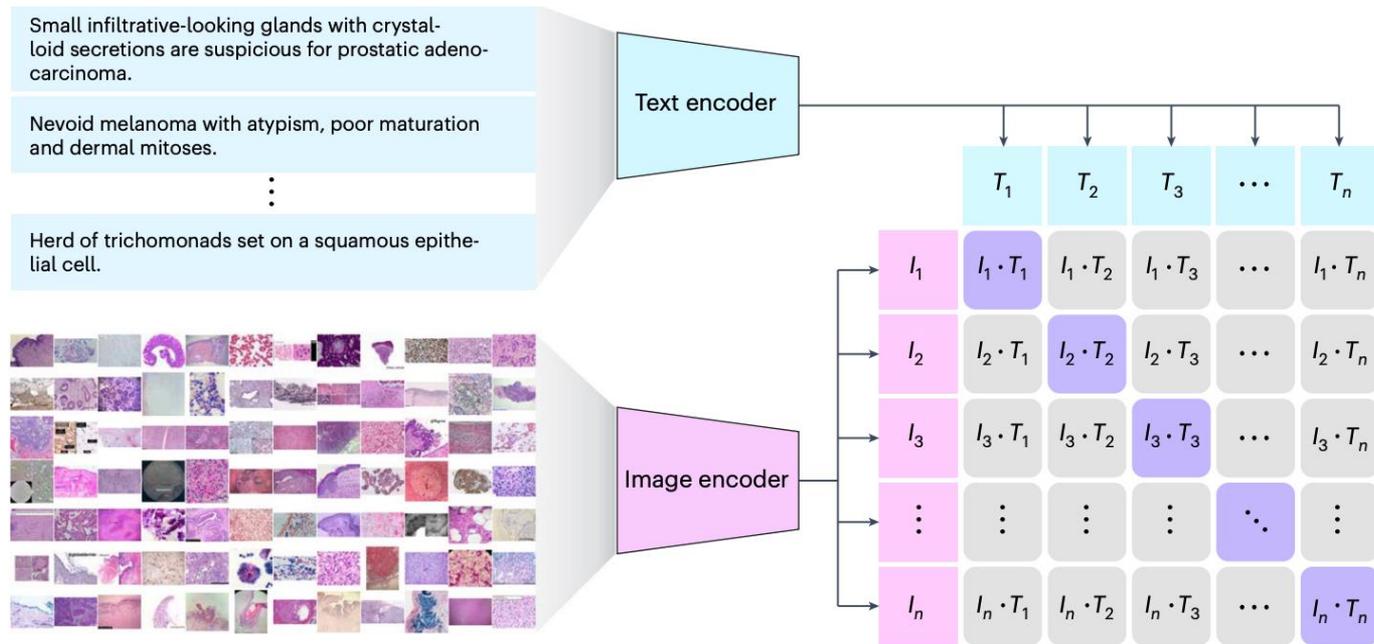


Meta MovieGen 2024

Text input summary: A red-faced monkey with white fur is bathing in a natural hot spring. The monkey is playing in the water with a miniature sail ship in front of it, made of wood with a white sail and a small rudder. The hot spring is surrounded by lush greenery, with rocks and trees.

Apply CLIP to Different Tasks

Lecture 6

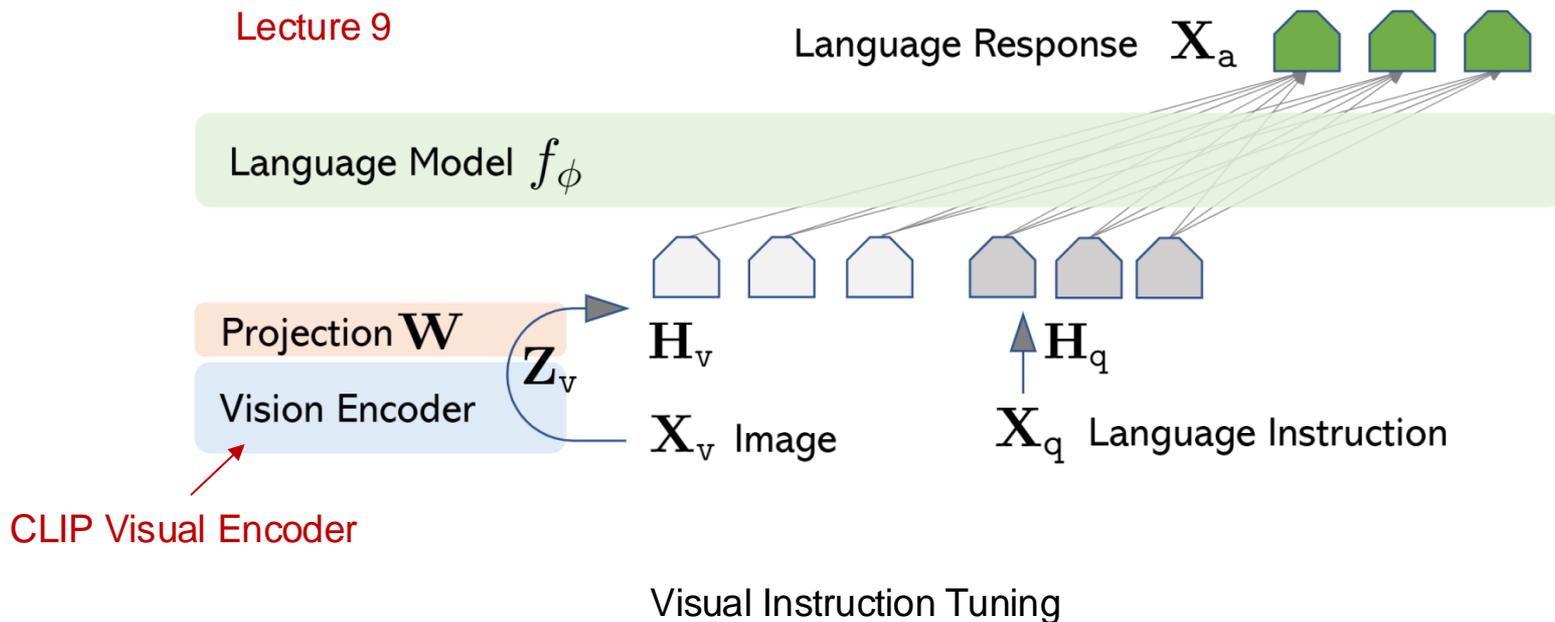


Fine-tuning CLIP on Pathology Image-Text Pairs

PILP (Huang et al. 2023)

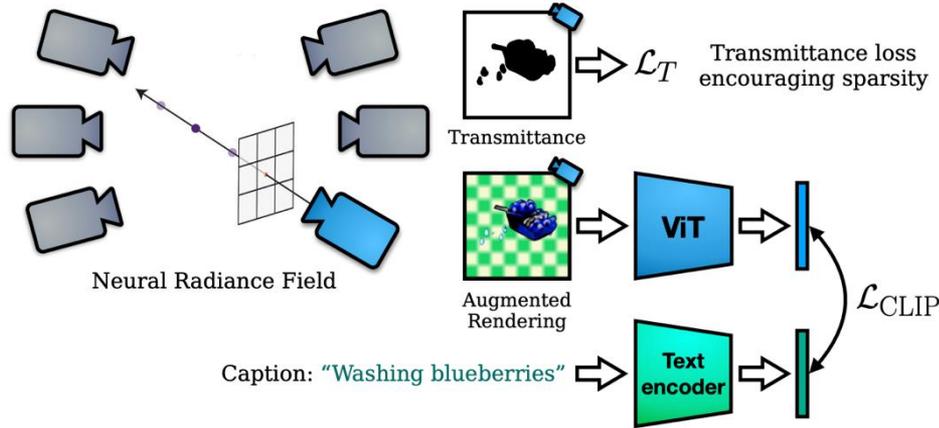
Apply CLIP to Different Tasks

Lecture 9



CLIP Visual Encoder

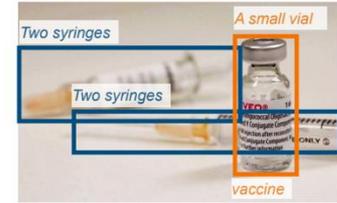
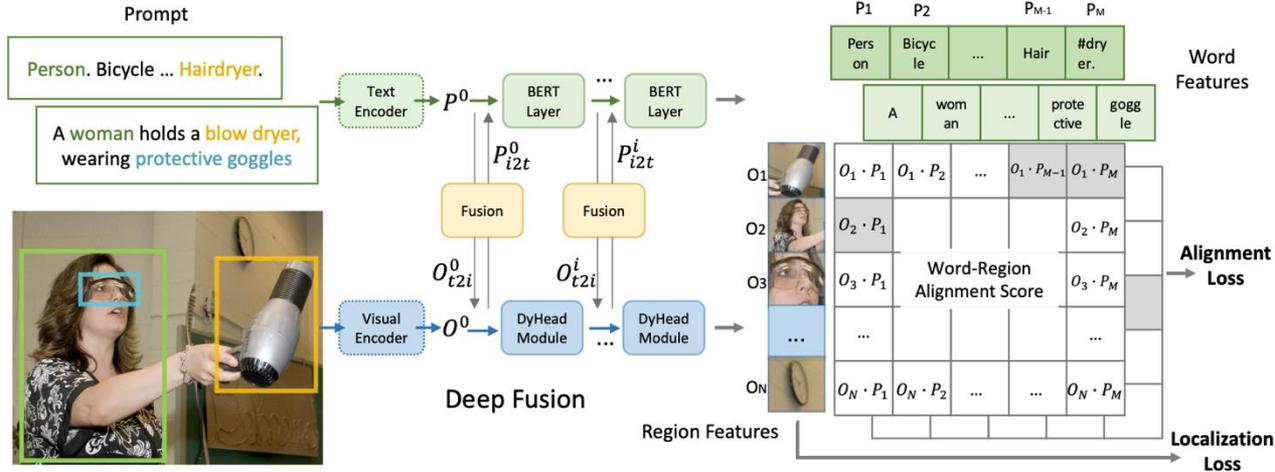
Apply CLIP to Different Tasks



CLIP for Zero-shot 3D Object Generation

DreamField (Jain et al. 2022)

Region-based CLIP

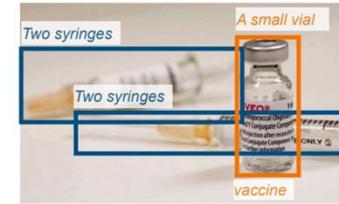
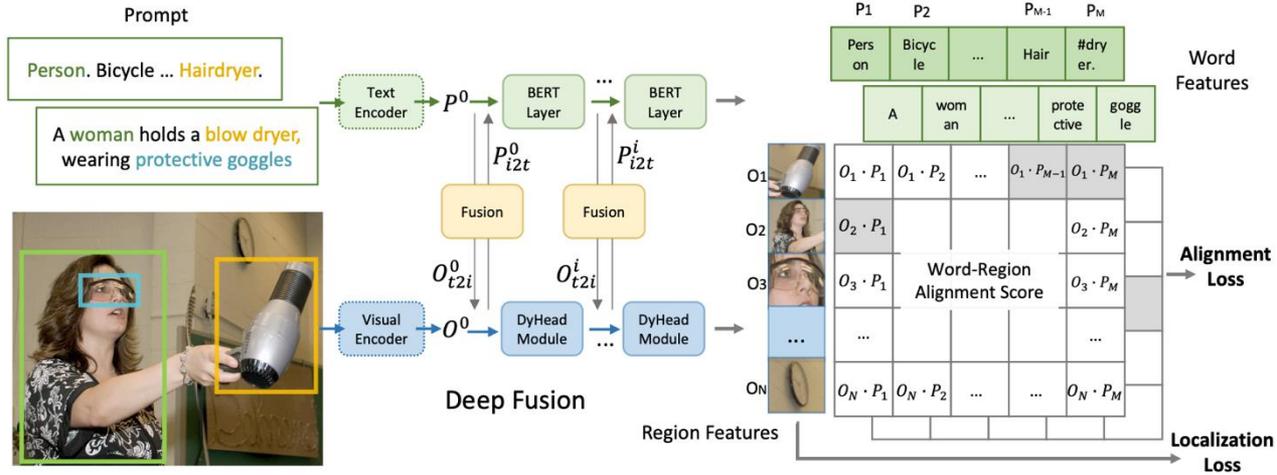


Two syringes and a small vial of vaccine.



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

Region-based CLIP

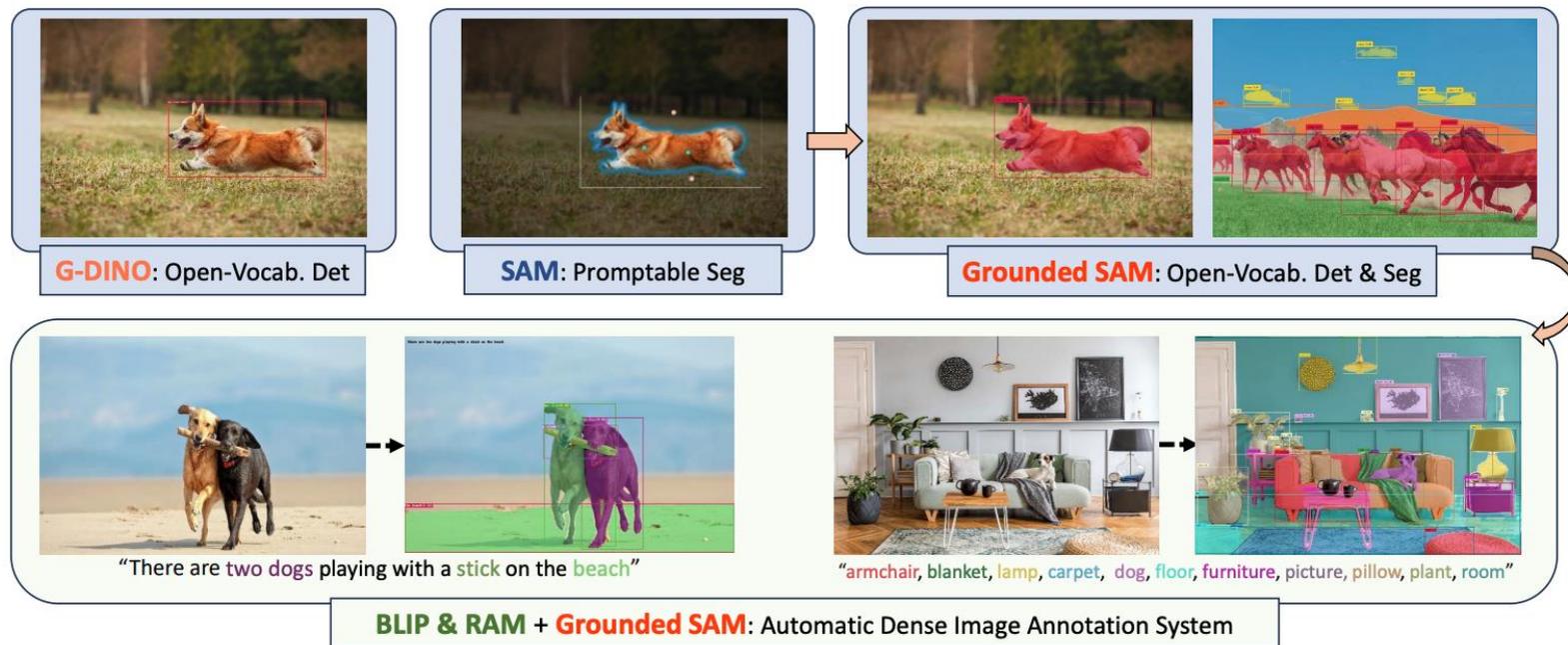


Two syringes and a small vial of vaccine.



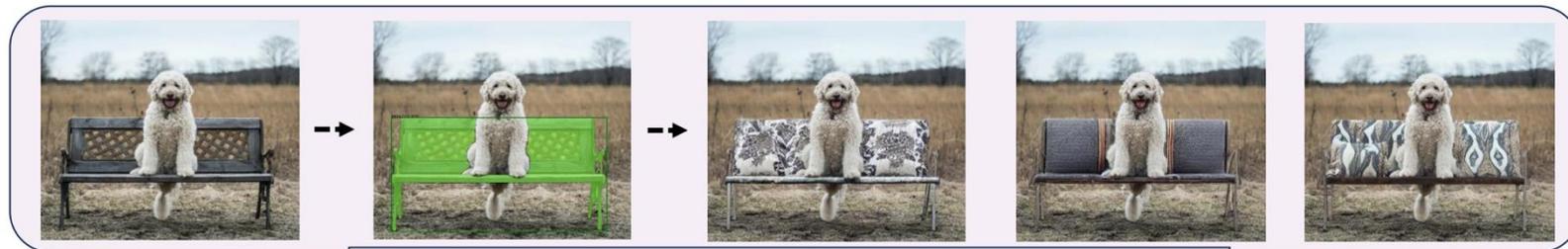
playa esmeralda in holguin, cuba. the view from the top of the beach. beautiful caribbean sea turquoise

Assembling Open-World Models for Diverse Visual Tasks



Grounded SAM (Ren et al. 2022)

Assembling Open-World Models for Diverse Visual Tasks



Grounded SAM + Stable-Diffusion: Highly Controllable Image Editing



Grounded SAM + OSX: Promptable Human Motion Analysis

Grounded SAM (Ren et al. 2022)

Serena Yeung-Levy
Xiaohan Wang

BIODS 276: Adv. Topics in CV and Biomedicine

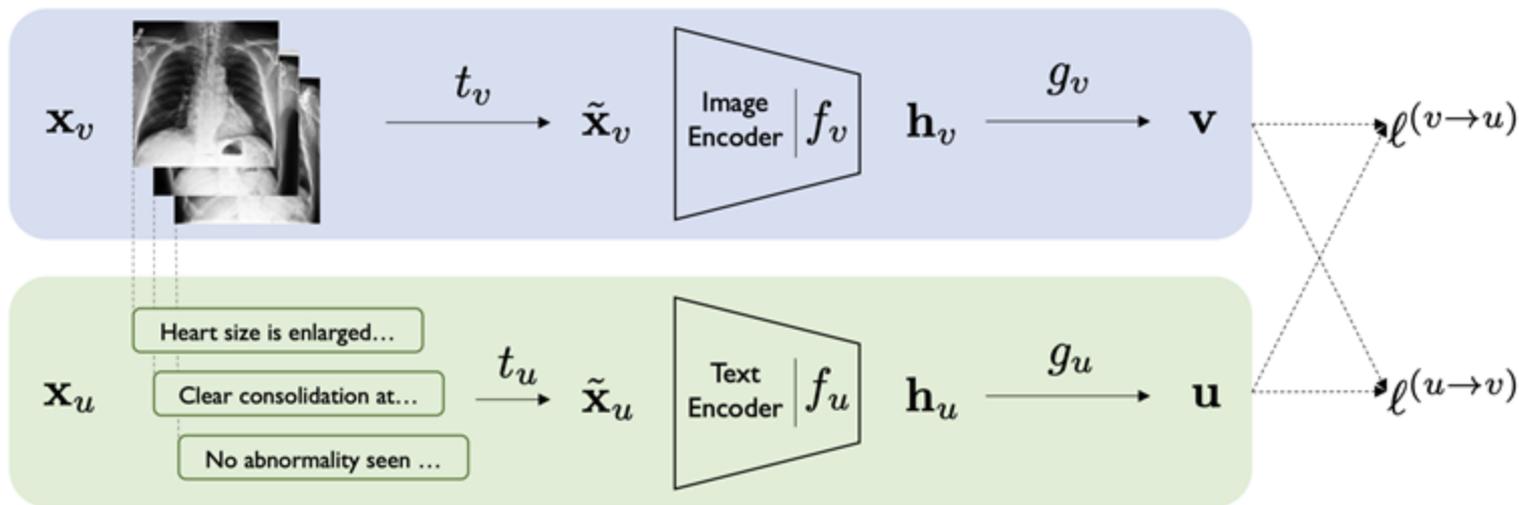
Lecture 6 - 38

Next:

Vision-Language Representation Learners in Biomedicine

Early work on vision-language representation learning in radiology

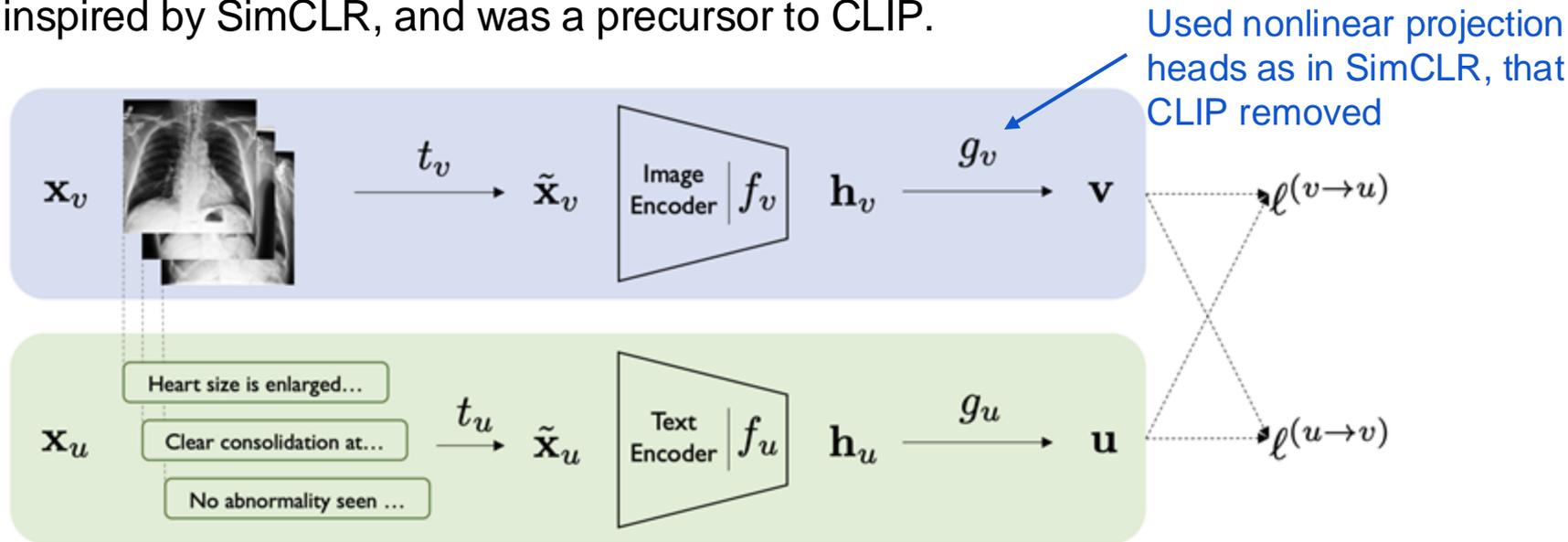
- ConVIRT performed contrastive learning on radiology image-report pairs. It was inspired by SimCLR, and was a precursor to CLIP.



Zhang et al. Contrastive Learning of Medical Visual Representations from Paired Images and Text. MLHC 2022.

Early work on vision-language representation learning in radiology

- ConVIRT performed contrastive learning on radiology image-report pairs. It was inspired by SimCLR, and was a precursor to CLIP.



Zhang et al. Contrastive Learning of Medical Visual Representations from Paired Images and Text. MLHC 2022.

Early work on vision-language representation learning in radiology

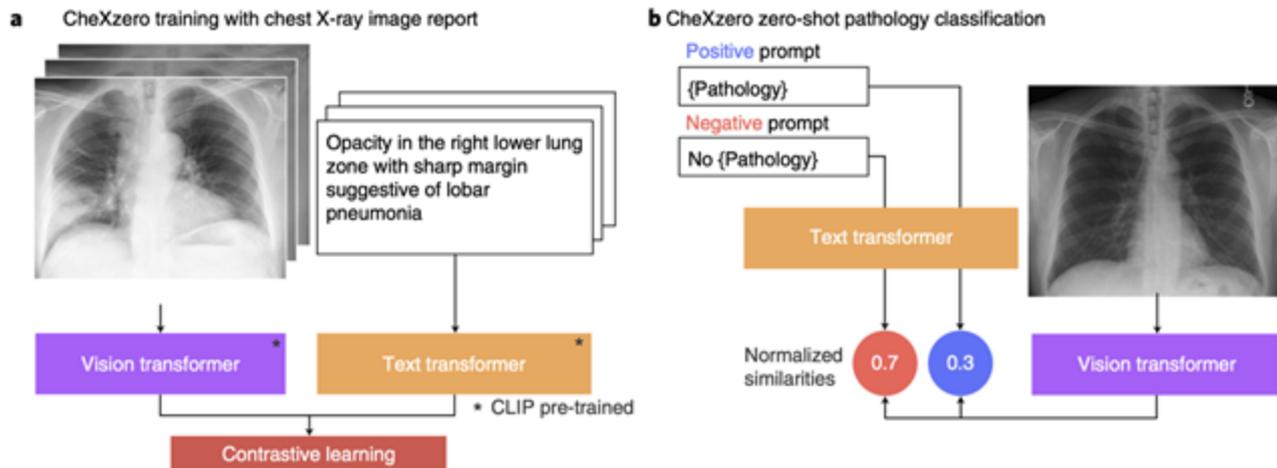
- ConVIRT trained on MIMIC-CXR (~217K image-text pairs), with random sampling of sentences from radiology reports. Outperformed image-only self-supervised learning from MIMIC-CXR for downstream classification and retrieval tasks.

Method	RSNA (Linear, 1%)	CheXpert (Linear, 1%)	Image-Image (Prec@10)
ImageNet	82.8	75.7	14.4
SimCLR (Chen et al., 2020a)	86.3	77.4	17.6
MoCo v2 (Chen et al., 2020b)	86.6	81.3	20.6
ConVIRT	90.7	85.9	42.9

Zhang et al. Contrastive Learning of Medical Visual Representations from Paired Images and Text. MLHC 2022.

CheXzero: leveraged CLIP to improve on ConVIRT and demonstrate zero-shot capabilities

- Compared to ConVIRT, updated to the same architecture as CLIP (better encoders, no nonlinear projection) as well as CLIP-pretrained weights.
- Also used the “impressions” section of the radiology report instead of ConVIRT sampling.



Tiu et al. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. Nature Biomedical Engineering, 2022.

CheXzero: leveraged CLIP to improve on ConVIRT and demonstrate zero-shot capabilities

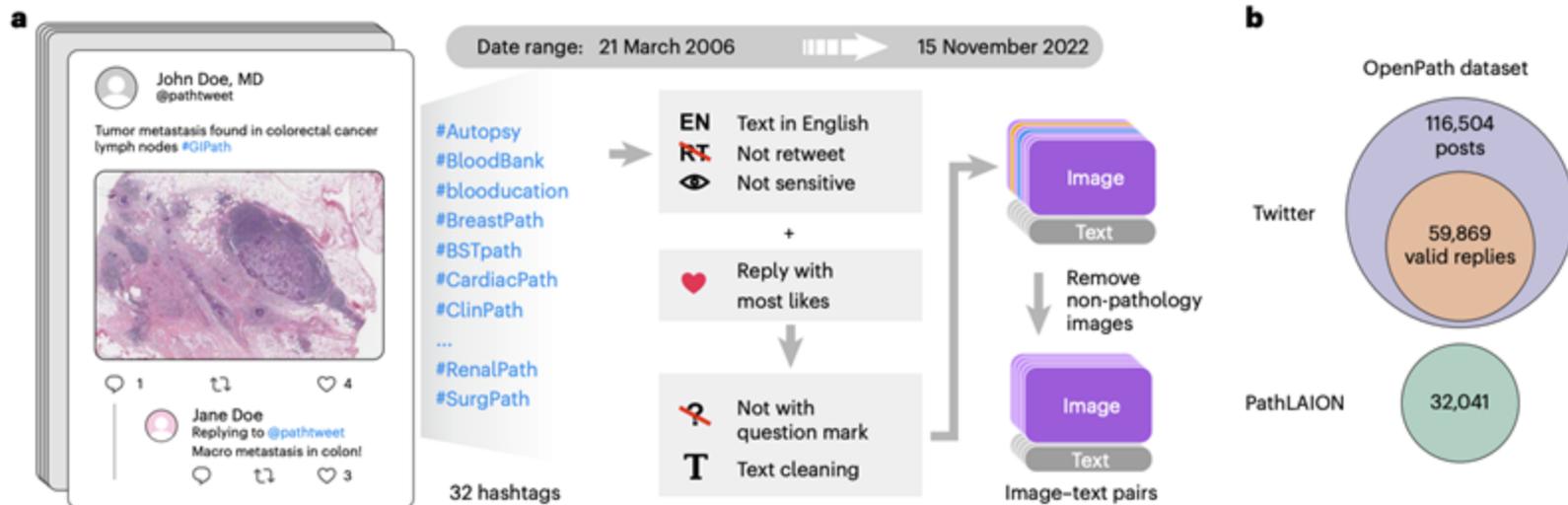
Comparison with other supervised and self-supervised approaches on the CheXpert test dataset

	Model	Mean AUC
Supervised	DAM	0.931
	DenseNet-121	0.902
Self-supervised	GLoRIA ^a	0.534
	ConVIRT-ResNet-50—1%	0.870
	ConVIRT-ResNet-50—10%	0.881
	ConVIRT-ResNet-50—100%	0.881
	ConVIRT-ViT—1% ^b	0.725
	ConVIRT-ViT—10% ^b	0.809
	ConVIRT-ViT—100% ^b	0.856
	MedAug—1%	0.810
	MoCo-CXR—1%	0.802
	MoCo-CXR—10%	0.850
	MoCo-CXR—100%	0.884
	CheXzero—0%	0.889

Tiu et al. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. Nature Biomedical Engineering, 2022.

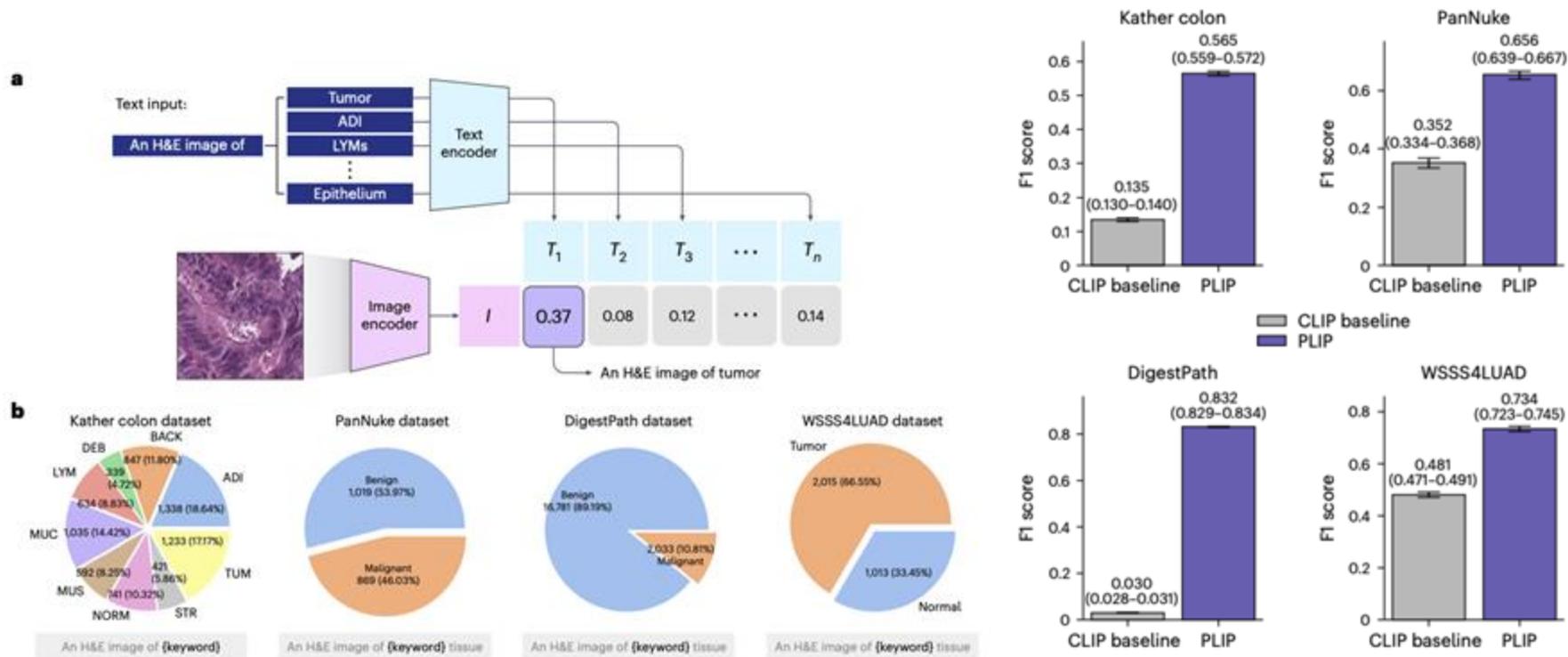
PLIP: vision-language foundation model for pathology trained from Twitter

- Curated and trained on OpenPath: 208,414 image-text pairs scraped from the Internet, mostly from Twitter



Huang et al. A visual–language foundation model for pathology image analysis using medical Twitter. Nature Medicine, 2023.

Zero-shot classification using PLIP



Huang et al. A visual-language foundation model for pathology image analysis using medical Twitter. Nature Medicine, 2023.

Text-to-image and image-to-image retrieval

Text-to-image retrieval:

Breast tumor surrounded by fat

Most relevant image (similarity = 0.2994)

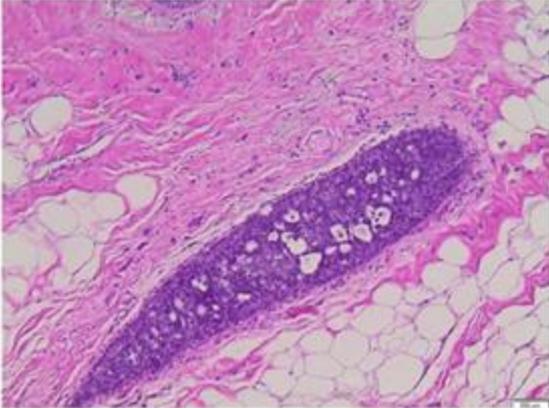
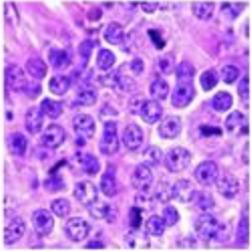
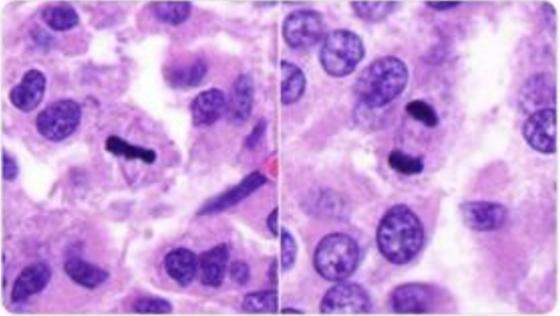
A histological image showing a breast tumor (dark purple) surrounded by adipose tissue (yellowish). The tumor is located in the lower right quadrant of the image.

Image-to-image retrieval:

Drop file here

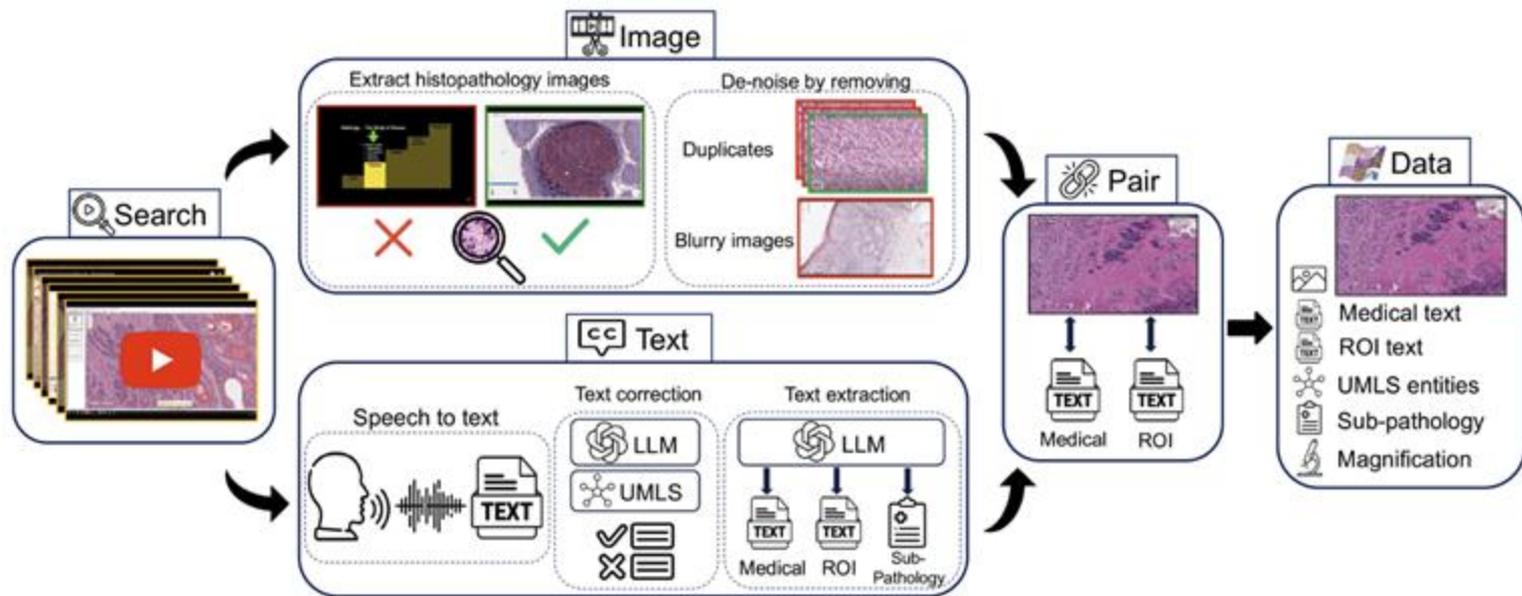
A histological image showing a mitotic figure (purple) in a cell. The image is labeled "Input image: mitotic figure".

Most relevant image (similarity = 0.9091)

A histological image showing a mitotic figure (purple) in a cell, similar to the input image. The image is labeled "Most relevant image (similarity = 0.9091)".

Huang et al. A visual–language foundation model for pathology image analysis using medical Twitter. Nature Medicine, 2023.

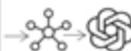
Quilt-1M: expanding pathology vision-language training data using YouTube



Ikezogwo et al. Quilt-1M: One Million Image-Text Pairs for Histopathology. NeurIPS 2023 Datasets & Benchmarks.

An LLM was used to process noisy speech transcription into corrected and structured texts for CLIP training

System Prompt:	You are an Automatic speech recognition's noisy medical text correction engine	System Prompt:	Extract medically relevant information from the following text, including any physical descriptions or attention to specific regions or concepts.
User Prompt:	Thinking step by step, Acting as a medical/histopathology ontology/glossary and a medical search-engine return best replacements for words in incorrect list provided. the text is from an ASR so take that into consideration, output the right phrases in context to the previous statements and medical factuality.	User Prompt:	Please extract the key medical information from the following text, including any descriptions of physical characteristics or specific regions/concepts that are mentioned. Think step by step, extract medical/histopathology content from the following text, do not add new words. Also, for ROI extract only medically relevant substrings where the narrator is physically describing/pointing attention to a region/concept/image and using words like 'here', 'you can see', 'this area/region'.
Few-shot examples:	<p>"The most common ...is the radicular cyst. We've already discussed in some detail that ...the periapex from the epithelial risks of malacid. As a result of inflammation_ periaepical race, the resulting in the development. Conditioning words: ['periaepical race']":</p> <p>{("conditioned_output": {"periaepical race": "periapical radix", "periapex": "periapex"}, "unconditioned_output": {"epithelial risks of malacid": "epithelial cell of malassez"})}</p> <p>INPUT: "So pigment and perperic dermatosis, and they, some people... neutrophils. They don't have lupusidoclastic vasculitis. That's a pigment and perperic dermatosis... And although the virus arcomadus DFSPs. Conditioning words: ['perperic dermatosis', 'lupusidoclastic vasculitis']":</p> <p>OUTPUT: {"conditioned_output": {"pigment and perperic dermatosis": "pigmented purpuric dermatoses", "lupusidoclastic vasculitis": "leukocytoclastic vasculitis"}, "unconditioned_output": {"virus arcomadus": "Fibrosarcomatous"}}</p>	Few-shot examples:	<p>"The most common by quite some distance of these is the radicular cyst. We've already discussed in some detail that arises within the periodontal ligament space, particularly the periapex from the epithelial cell of malassez. As a result of inflammation following the death of the pulp extending into the periapical radix, the resulting in the development of cysts.":</p> <p>{("MED": ["Radicular cyst arises within the periodontal ligament space, particularly the periapex from the epithelial cell of malassez.", "These radicular cysts are caused by inflammation following the death of the pulp extending into the periapical radix."], "ROI": ["Radicular cyst within the periodontal ligament space.", "Inflammation following the death of the pulp Extending into the periapical radix"])}:</p> <p>INPUT: "So pigmented purpuric dermatosis, and they, some people think at least some of the cases are a lymphocytic vasculitis that the lymphocytes are damaging the vessels. Okay, fine. Maybe so, because there is hemorrhage, but they don't usually have neutrophils. They don't have leukocytoclastic vasculitis. That's a pigmented purpuric dermatosis. elsewhere, but if I had a biopsy, we'd just one area, like that, I'd probably say DFSP, and with a comment that there's an area that started to get particular, and I'm concerned it could be fibrosarcomatous, compare it with the excision specimen. The good news, they're going to treat it the same way. They take it out with a margin, and then we look at the whole excision specimen. And although the fibrosarcomatous DFSPs."</p> <p>OUTPUT: {"MED": ["Pigmented purpuric dermatosis may be a lymphocytic vasculitis, with lymphocytes damaging the vessels and causing hemorrhage.", "Neutrophils are usually not present in this condition.", "Leukocytoclastic vasculitis is not seen in pigmented purpuric dermatosis.", "Biopsy suggests Dermatofibrosarcoma Protuberans (DFSP)."], "ROI": ["Pigment and purpuric dermatosis.", "Fibrosarcomatous DFSPs"]}</p>



Ikezogwo et al. Quilt-1M: One Million Image-Text Pairs for Histopathology. NeurIPS 2023 Datasets & Benchmarks.

Example of input narration with corresponding medical and ROI text



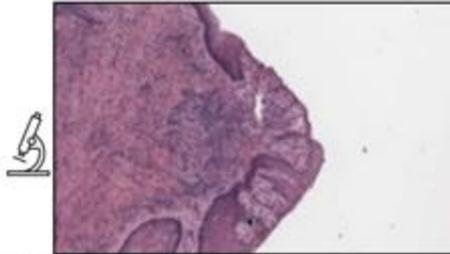
INPUT: "...so here we have a huge lumen and we have the lining epithelium and then we have the underlying connective tissue capsule. let's see the features in detail one by one. if you see this lumen lumen has sparse amount of keratin. so definitely this is a keratinous cyst. so we have a cyst that is actually producing keratin. let's move into the epithelium and the epithelium if you carefully observe most of the areas the epithelial appears uniform thickness right from here.

OUTPUT:

{MED:} "The cyst is producing keratin, indicating it is a keratinous cyst."
"The epithelial lining appears to have uniform thickness"]

{ROI:}

"Sparse amount of keratin in the lumen.",
"uniform thickness of the epithelial lining"]



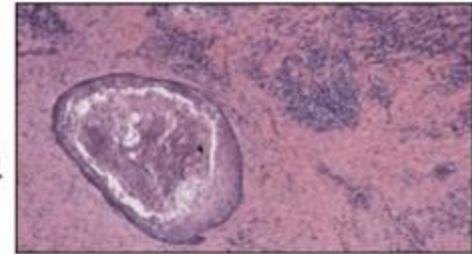
INPUT: "...tend to see some inflammation here yes. when you notice this there is a dense amount of inflammation here yeah you see this tiny blue dots these are the lymphocytes and these cells are your inflammatory cells and they are kind of having an effect on the epithelium. if you see this epithelium because of the inflammation the epithelium starts "

OUTPUT:

{MED:}

"Dense inflammation is present in the region with lymphocytes and other inflammatory cells having an effect on the epithelial cells"]

{ROI:} "dense inflammation here",
"lymphocytes",
"inflammatory cells"]



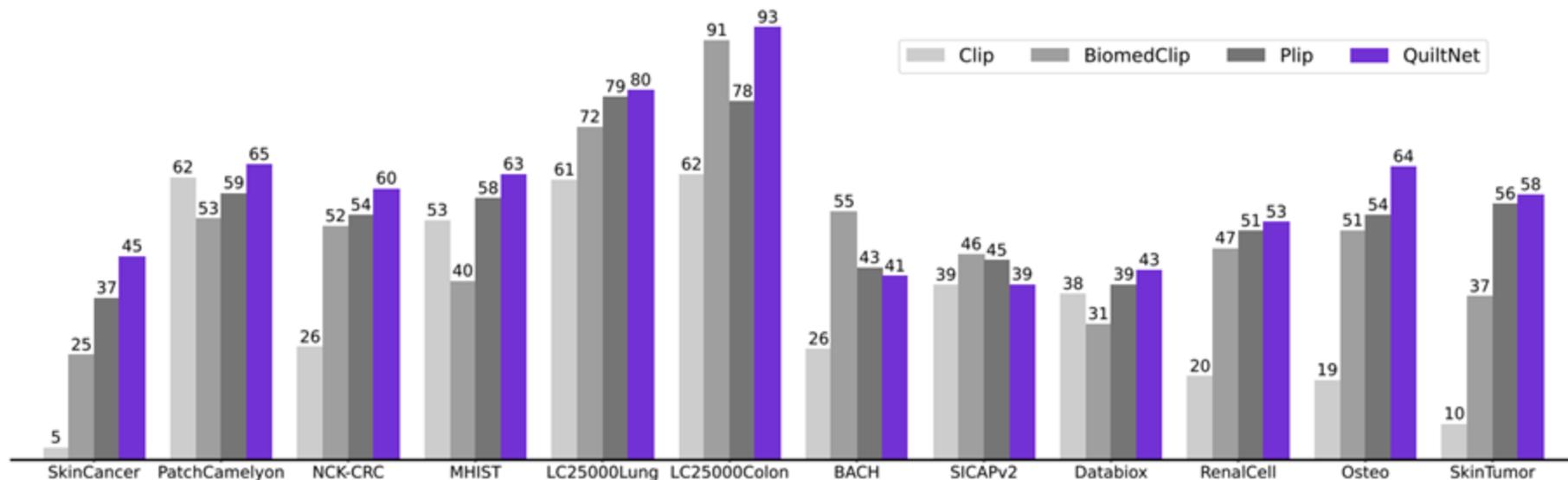
INPUT: "...of a daughter cyst where you see so this was an epithelial island with showing central keratinization the cells producing lot of keratin and at the same time you have the epithelium with the lumen. so this forms a tiny daughter cyst and this is one of the most important trees in why this keratocyst has a high amount of recurrence rate. one more important feature of odontogenic keratocyst is lies within the epithelium."

OUTPUT:

{MED:} "The presence of daughter cysts is one of the reasons for the high recurrence rate of keratocysts",
"Epithelial islands with central keratinization and high keratin production are seen in keratocysts"]

{ROI:} "epithelial island",
"central keratinization", "daughter cyst"]

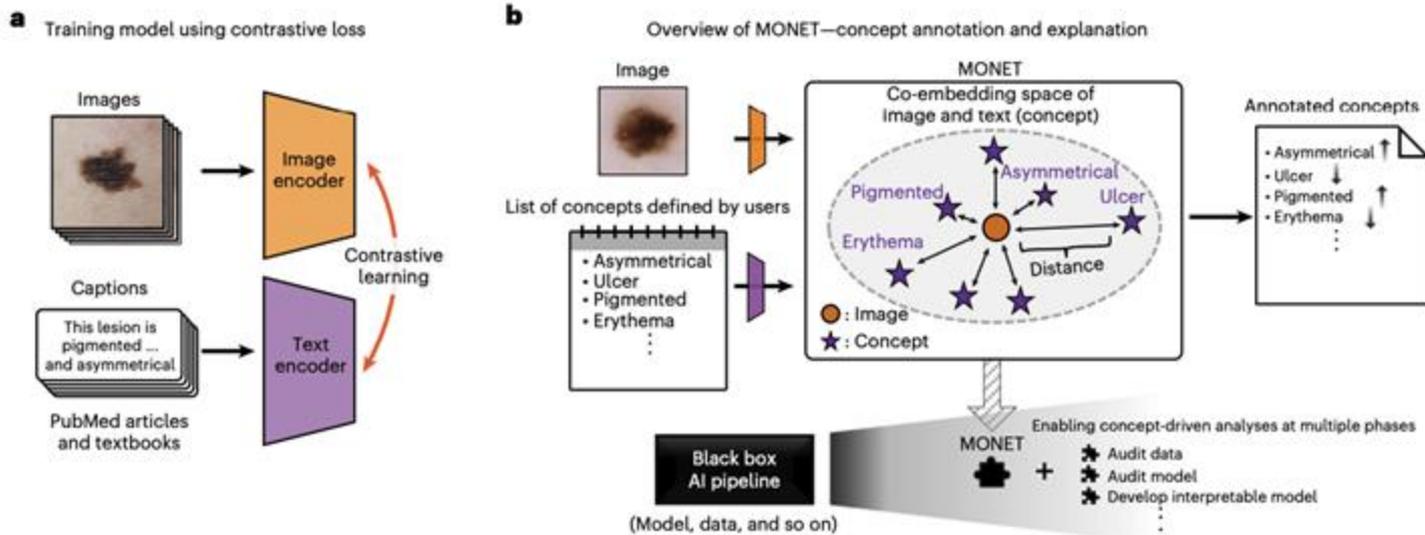
Performance of QuiltNet vs baselines on zero-shot classification tasks



Ikezogwo et al. Quilt-1M: One Million Image-Text Pairs for Histopathology. NeurIPS 2023 Datasets & Benchmarks.

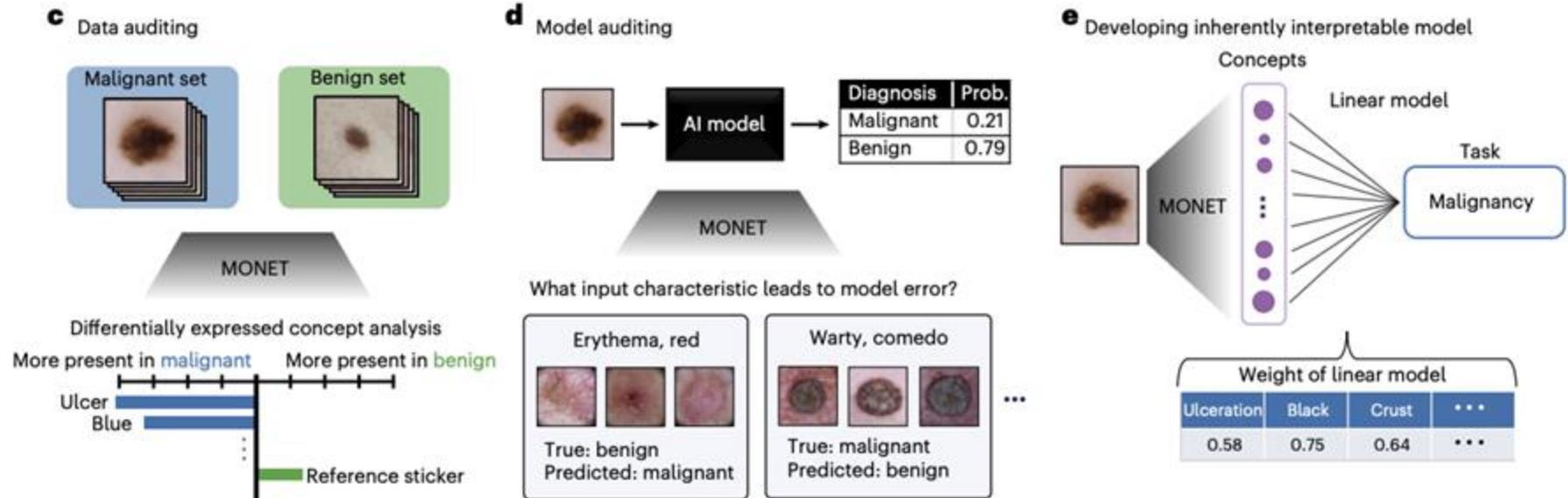
MONET: Leveraging a contrastively trained model to perform dataset and model auditing

- Trained on 105,550 dermatology image-text pairs from PubMed articles and medical textbooks



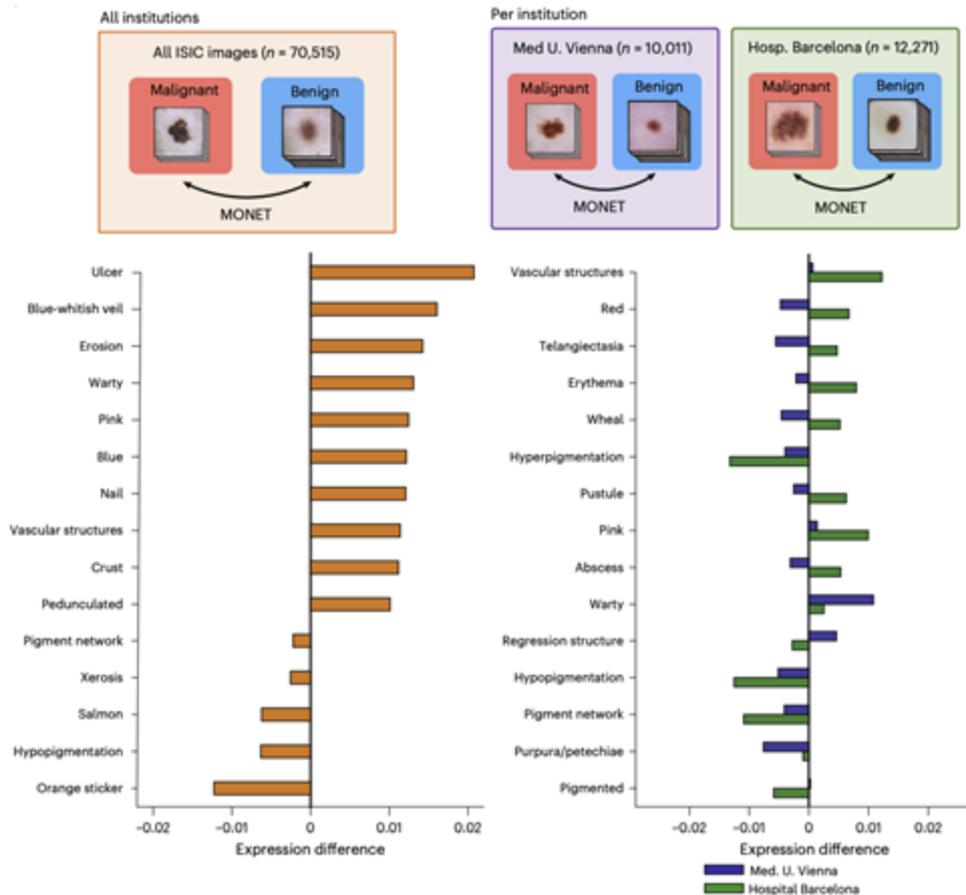
Kim et al. Transparent medical image AI via an image–text foundation model grounded in medical literature. Nature Medicine, 2024.

MONET: Leveraging a contrastively trained model to perform dataset and model auditing



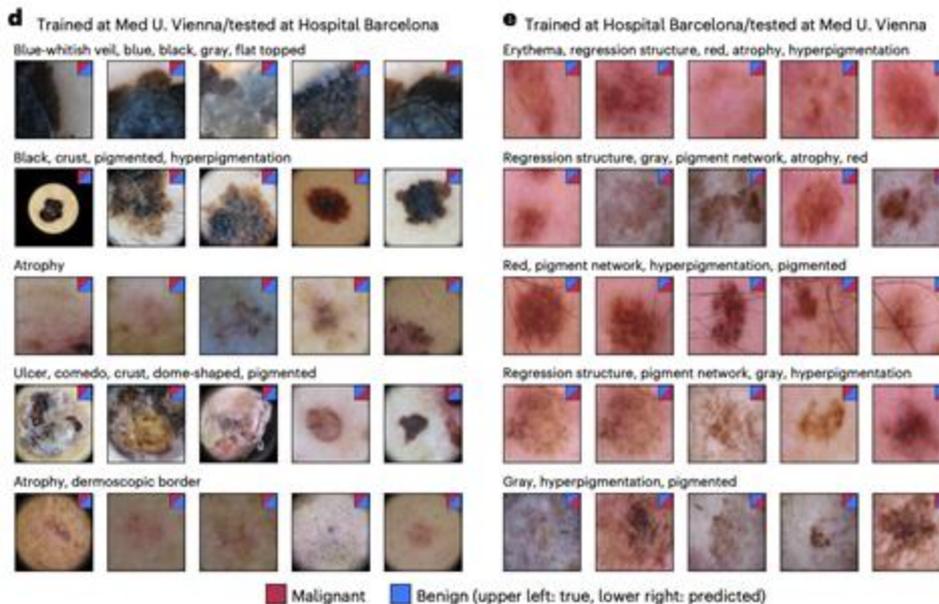
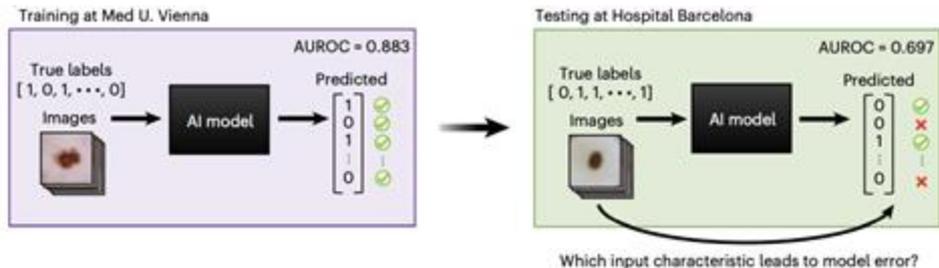
Kim et al. Transparent medical image AI via an image–text foundation model grounded in medical literature. Nature Medicine, 2024.

Concept differences identified by MONET during data auditing



Kim et al. Transparent medical image AI via an image–text foundation model grounded in medical literature. Nature Medicine, 2024.

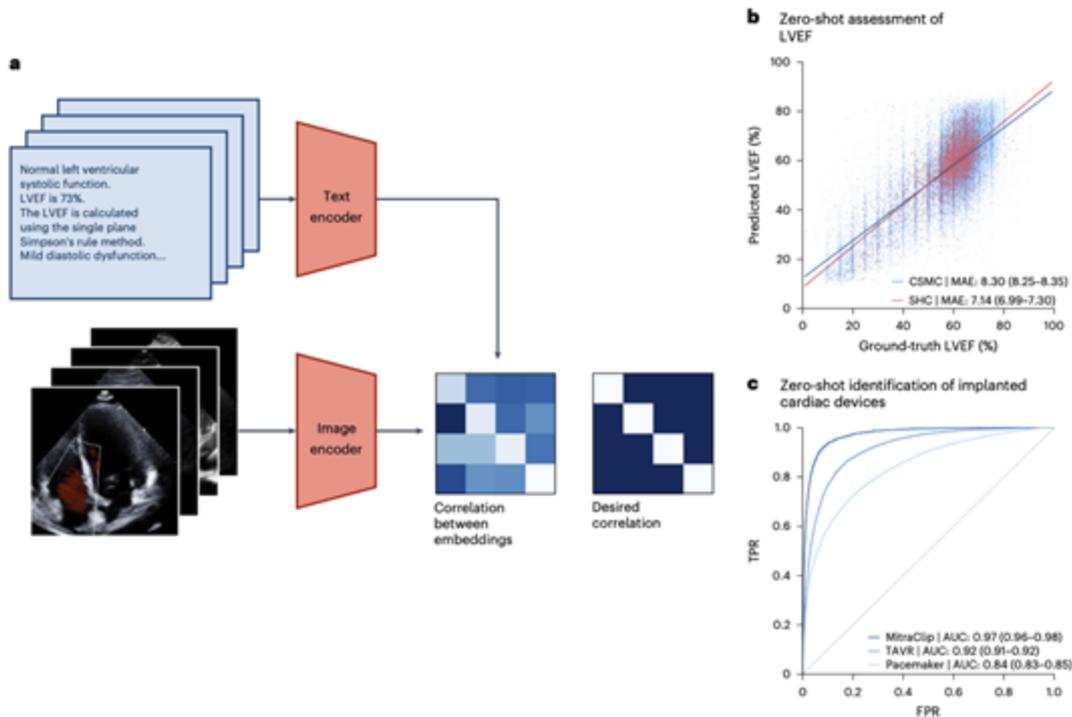
Concept differences identified by MONET during model auditing



Kim et al. Transparent medical image AI via an image–text foundation model grounded in medical literature. Nature Medicine, 2024.

Some more examples of CLIP-based foundation models...

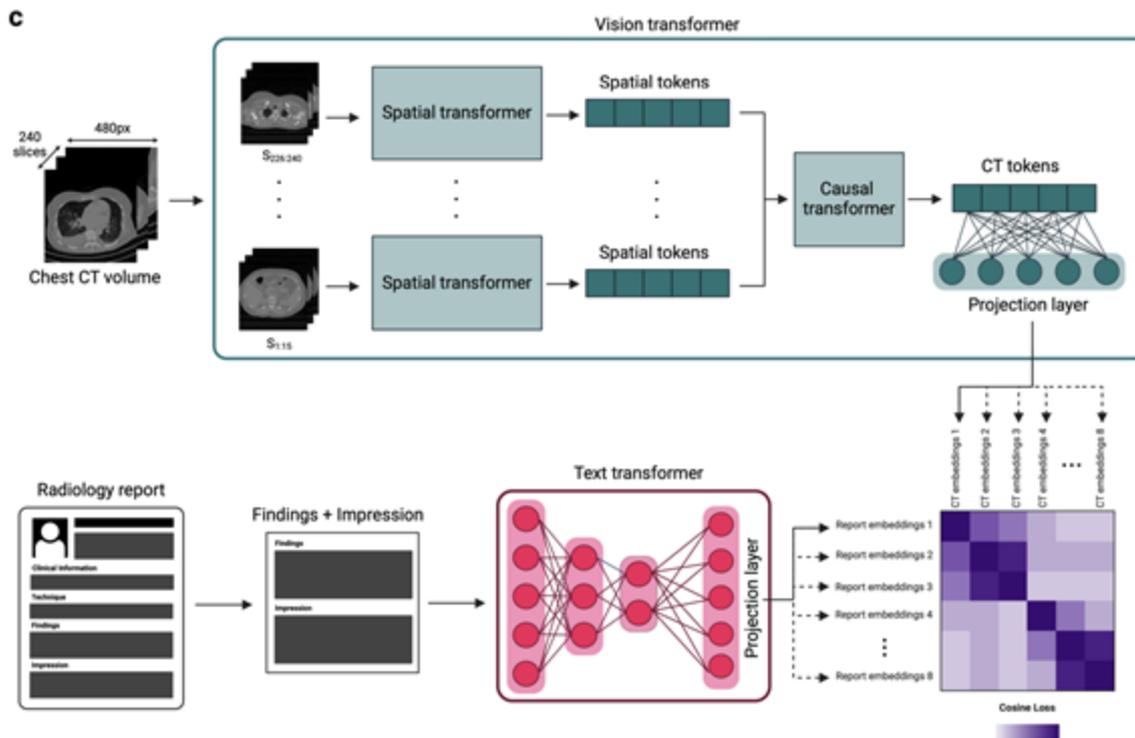
EchoCLIP: Based on 1,032,975 video-text pairs (but trained on images) for echocardiogram interpretation



Christensen et al. Vision-language foundation model for echocardiogram interpretation. Nature Medicine, 2024.

Some more examples of CLIP-based foundation models...

CT-CLIP: trained on 50,188 3D chest CT volumes with accompanying reports, using a previously developed 3D image encoder

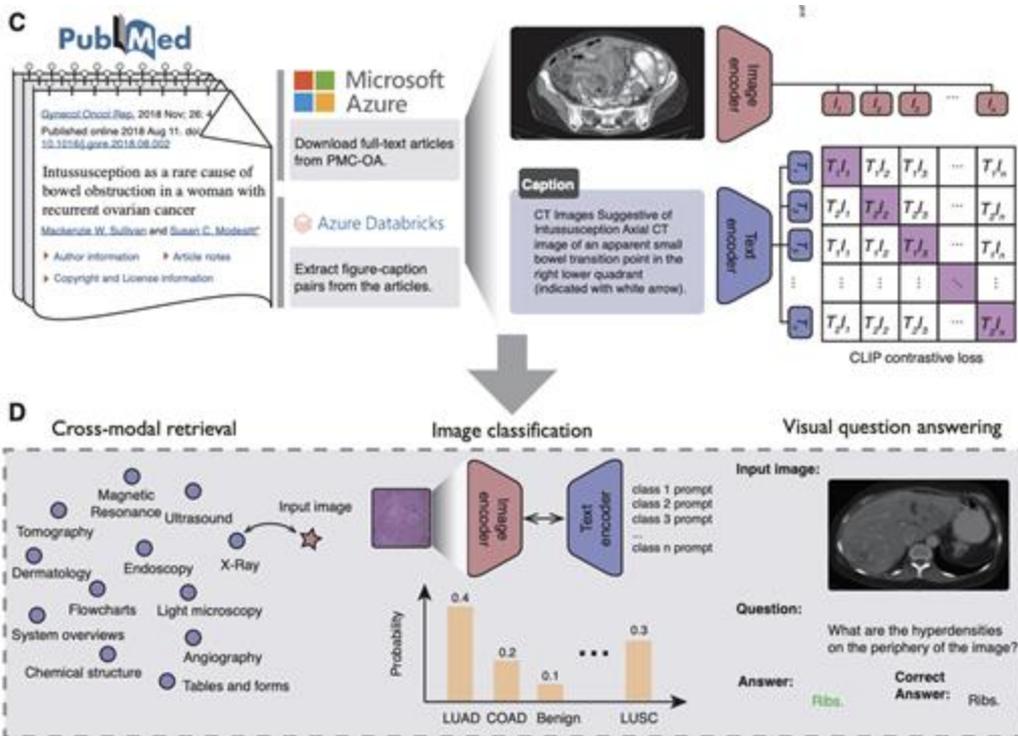


Hamamci et al. A foundation model utilizing chest CT volumes and radiology reports for supervised-level zero-shot detection of abnormalities. arXiv, 2024.

BiomedCLIP: Generalist foundation model trained on PubMed

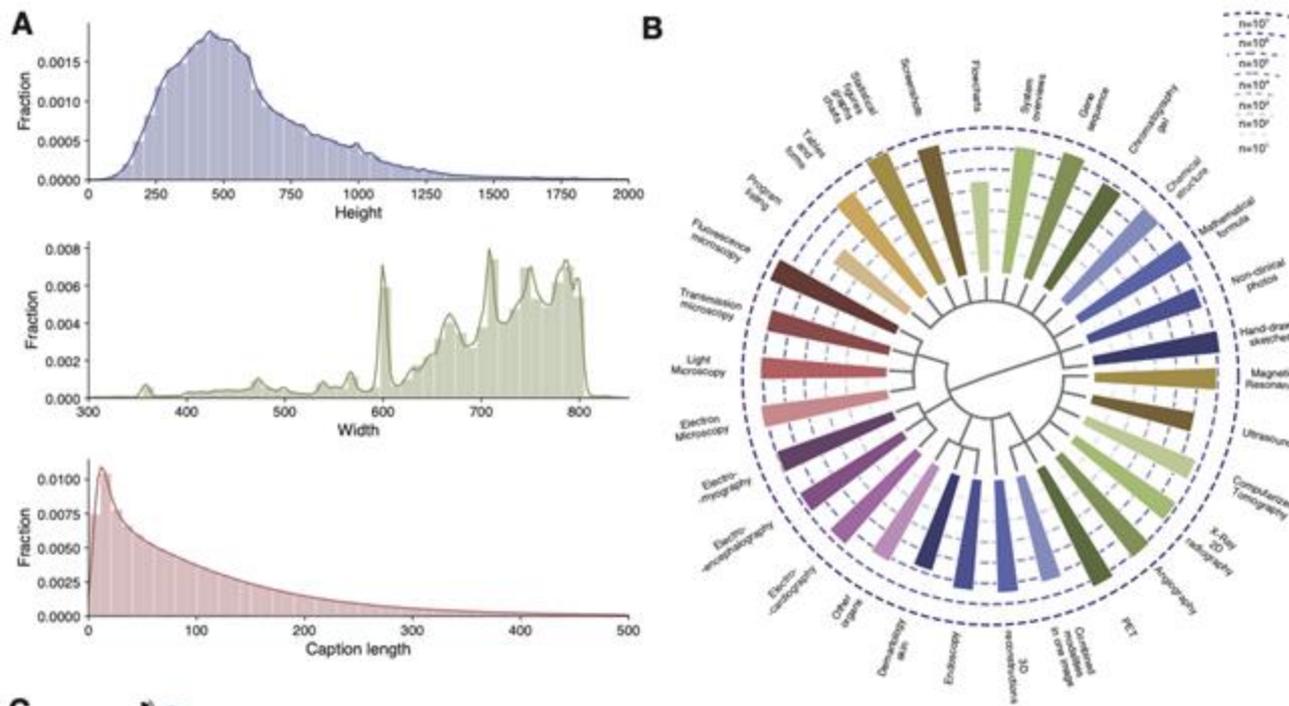
PMC-15M: 15 million image-caption pairs from 4.4 million publicly available full-text articles in PubMed Central

Complete article packages are downloaded, and figure files and matching captions are extracted



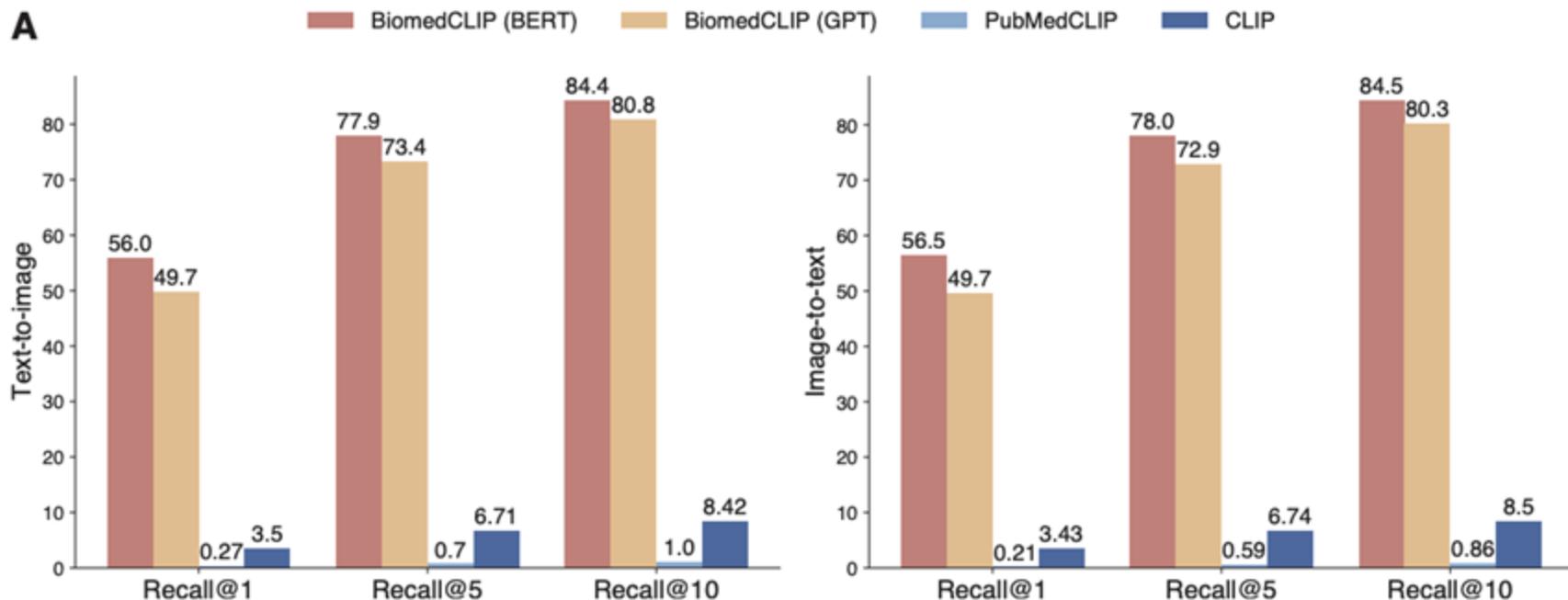
Zhang et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv, 2024.

PMC-15M statistics of image sizes and caption lengths



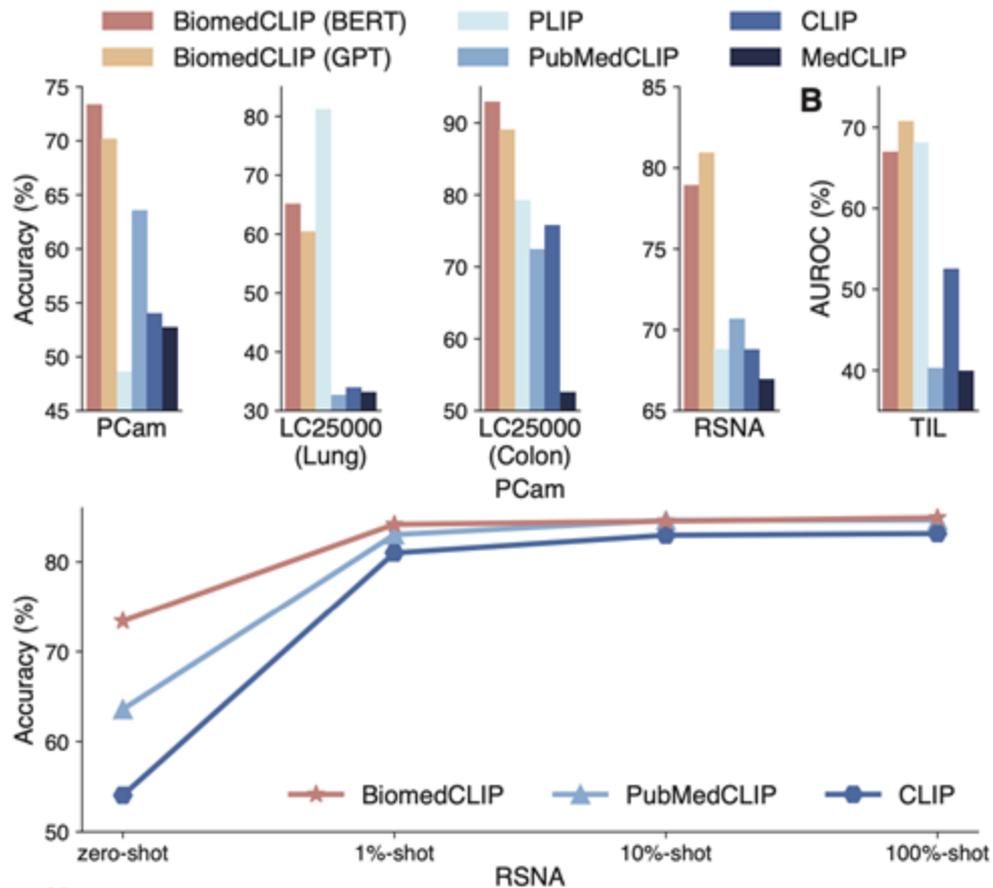
Zhang et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv, 2024.

BiomedCLIP comparison with baseline methods on cross-modal retrieval



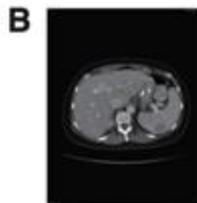
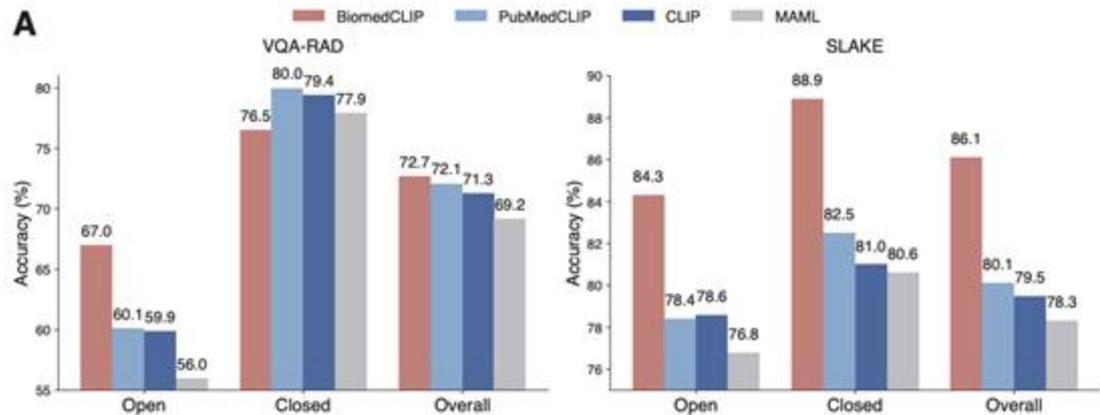
Zhang et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv, 2024.

BiomedCLIP comparison with baseline methods on zero-shot classification and linear probing



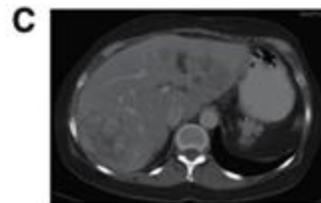
Zhang et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv, 2024.

BiomedCLIP comparison with baseline methods on medical visual question answering (VQA)



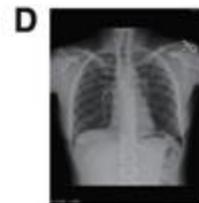
Question: Are there multiple or just 1 metastatic focus?

Answer: **one**
 MEVF: right chest X
 QCR: no X
 PubMedCLIP: yes X
 BiomedCLIP: right lobe of liver X



Question: What are the hyperdensities on the periphery of the image?

Answer: **ribs**
 storage of urine X
 intestine X
 spinal cord X
 ribs ✓



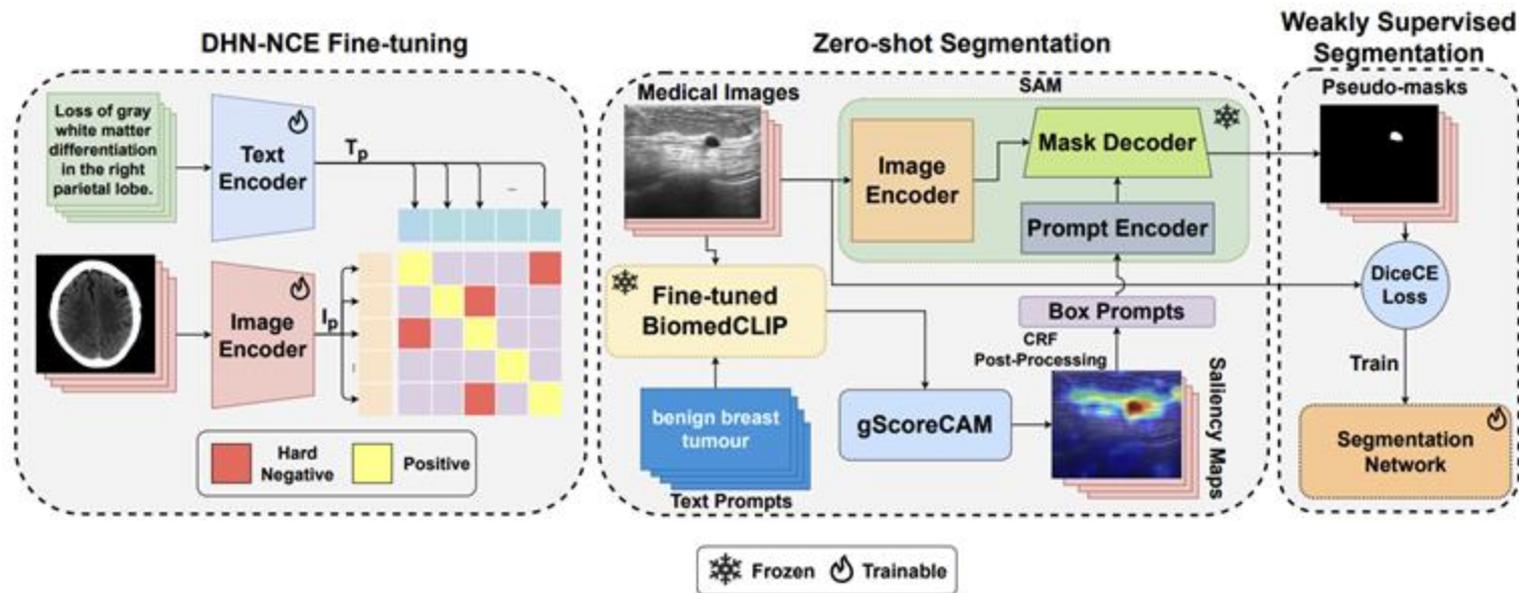
Question: What is the biological sex of the patient?

Answer: **female**
 inflammation ... X
 treat brain diseases ... X
 nodule X
 female ✓

Zhang et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv, 2024.

Combining biomedical CLIP models with SAM

- Approaches such as MedCLIP-SAM can perform zero-shot and weakly supervised segmentation



Koleilat et al. MedCLIP-SAM: Bridging Text and Image Towards Universal Medical Image Segmentation. MICCAI, 2024.

Next time

- Vision Diffusion and Generative Models