“Lecture” 13: Course Conclusion

Announcements

- **Please make sure to sign up for project advising sessions today and tomorrow** (see Ed post for the sign-up sheet), this is 5% of your project grade. Contact Xiaohan if an alternate time is needed.
- Project milestone grades will be provided before the session time.
- **Project final presentation is during the final exam slot Mon Dec 9** from 3:30-6:30pm. More details will be provided, expect ~7 minutes presentation per project.
- **Project final report is due Wed Dec 11**

- James's office hours will start at 5:30 today instead of 4pm; if this poses difficulty for you, please contact him to arrange an alternate time

Announcements

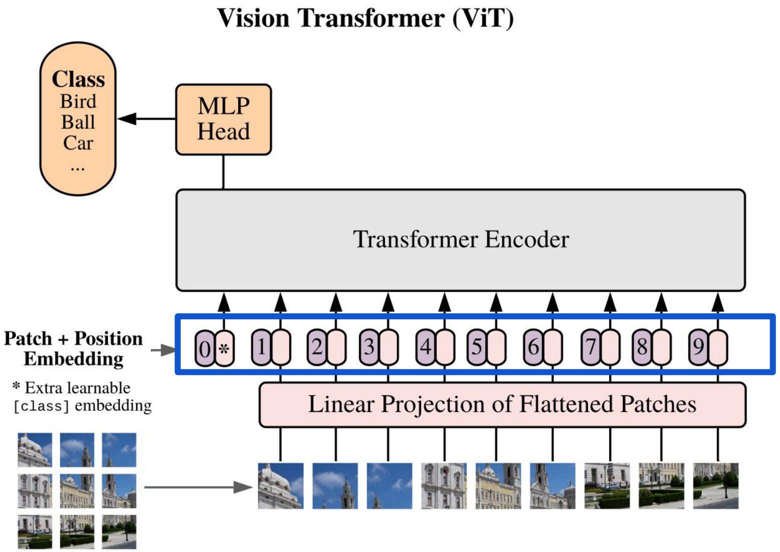
- **Wednesday will be a final guest lecture from Khaled Saab, first-author of the Med-Gemini paper. This lecture will be remote over zoom (link will be provided in Ed), please plan to attend!**

Our goals at the beginning of the course

- Fluency in cutting edge computer vision models and research
    - Focus on vision and vision-language foundation models, from representation learners to diffusion and generative models
- Understanding of existing applications and opportunities for future utility in biomedicine
- Ability to think through considerations for real-world use, including model size and computation, training and inference settings, and training data

First topic: Vision Transformers (ViT)

Key idea: Convert image into sequence of patches. Can then benefit from Transformer architecture and self-attention, which jointly attends over all patches


Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

Converted image sequence ready for input to Transformer encoder (analogous to text sequence)

Dosovitsky et al. 2021

Two major classes of vision foundation models

Representation Learners

Train on huge amounts of data

Objective: learn powerful generalist feature extractors that can be used for downstream computer vision tasks

- Vision Only (e.g. DINOv2)
- Vision-Language (e.g. CLIP)

Generative Models

Train on huge amounts of data

Objective: learn to generate images and/or text

- Text -> Image (e.g. StableDiffusion, DALL-E)
- Text, Image -> Text (e.g. Flamingo, GPT4-V)

Two major classes of vision foundation models

Representation Learners

Train on huge amounts of data

Objective: learn powerful generalist feature extractors that can be used for downstream computer vision tasks

- Vision Only (e.g. DINOv2)
- Vision-Language (e.g. CLIP)

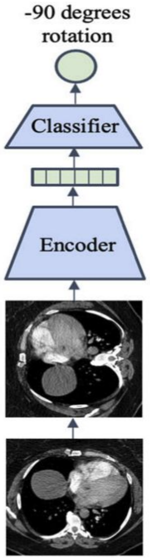
Generative Models

Train on huge amounts of data

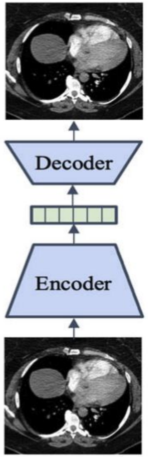
Objective: learn to generate images and/or text

- Text -> Image (e.g. StableDiffusion, DALL-E)
- Text, Image -> Text (e.g. Flamingo, GPT4-V)

Different representation learning paradigms

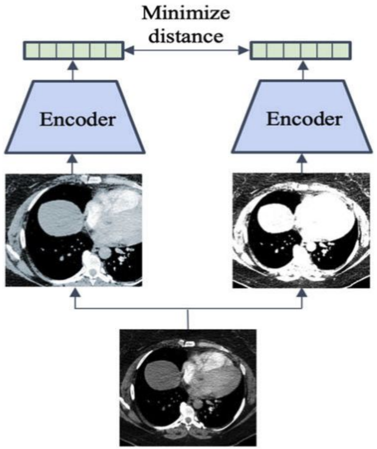


Innate relationship objective
E.g., predict rotation angle (or some other innate property) of an image

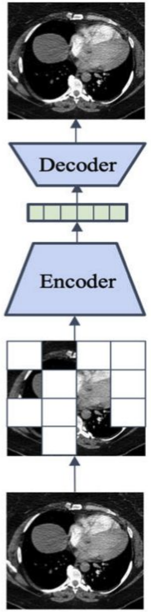


Generative objective
Compress and then reconstruct input image (e.g. autoencoders)

Popular state-of-the-art approaches



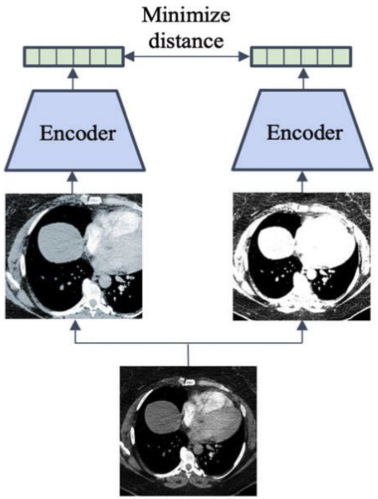
Contrastive objective
Different views of the same input should have more similar representation to each other than with a different input



Self-prediction objective
Mask parts of input data and predict these parts

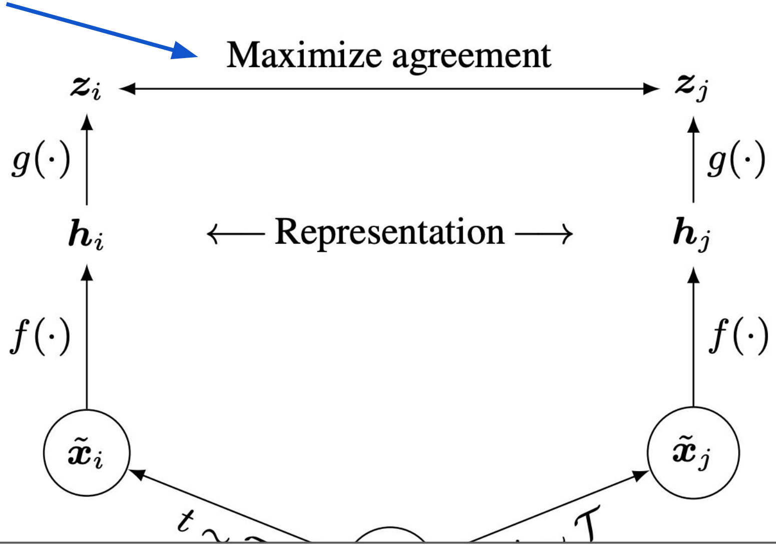
Figure credit: Huang et al. 2023.

SimCLR: a foundational method for contrastive self-supervised learning



Contrastive loss

SimCLR formulation

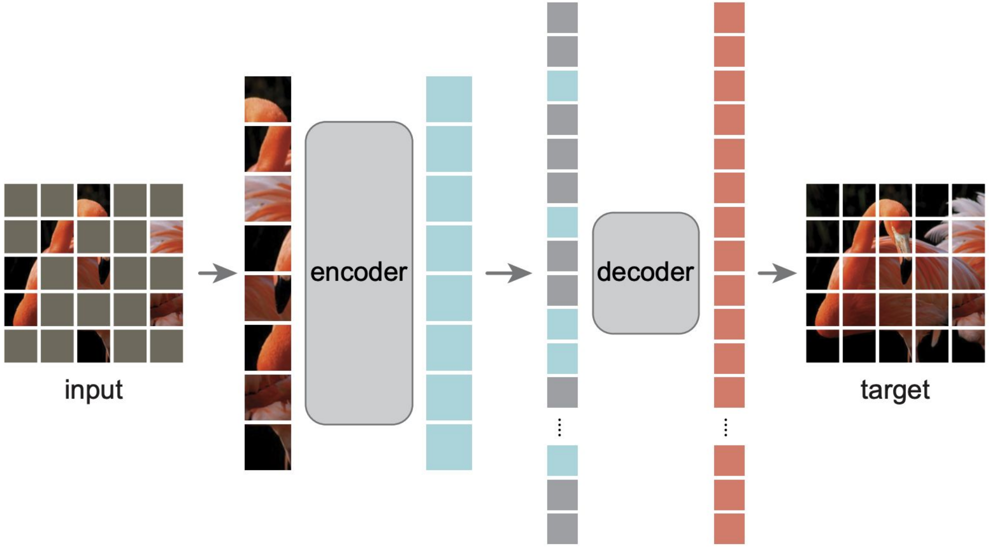


After self-supervised training, can fine-tune the encoder *f* on smaller labeled datasets. Can also directly extract learned representations h for downstream tasks.

Masked Autoencoders (MAE)

- Key idea: mask substantial parts of the input, train the model to reconstruct (predict) these parts

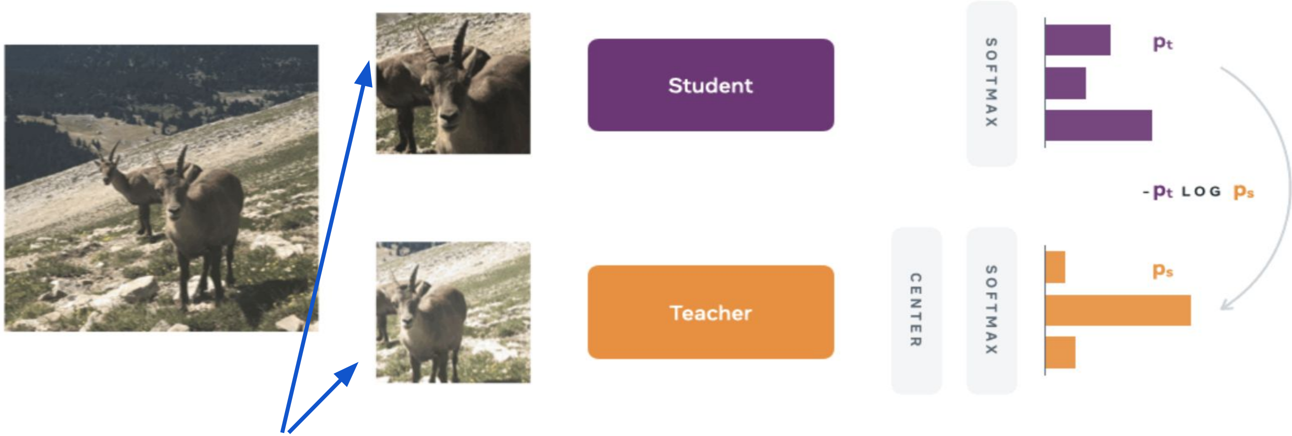
- Inspired by major self-supervised representation learning paradigm in NLP (e.g. BERT), that masks tokens in sentences and trains models to reconstruct them



He et al. 2021

Another SSL approach: DINO (Self-**Di**stillation with **No** Labels)

Not an example of a contrastive learning objective for self-supervised learning, but related! Builds on the notion of matching representations of augmented views of the same image, but no longer uses negative samples. Instead, proposes a **teacher-student framework**.

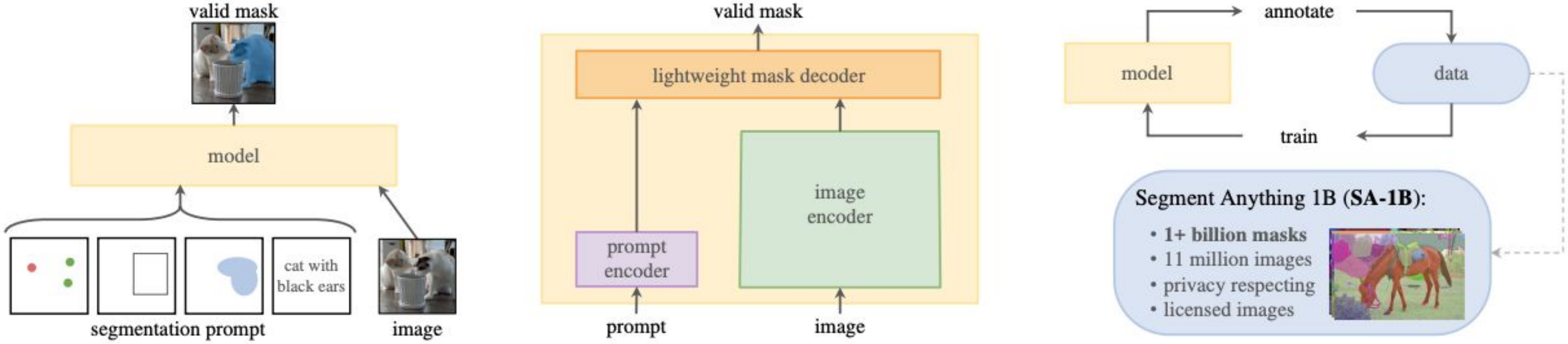


Augmented "views" of the input are passed to both the student and teacher networks.
Student views: more diverse and aggressive augmentations, to learn representations that generalize across augmentations
Teacher views: less aggressive augmentations (including larger crops), to provide a more stable target for the student to follow.

Caron et al. 2021
Oquab et al. 2024

Segment Anything Model (SAM): A foundation model targeted for segmentation

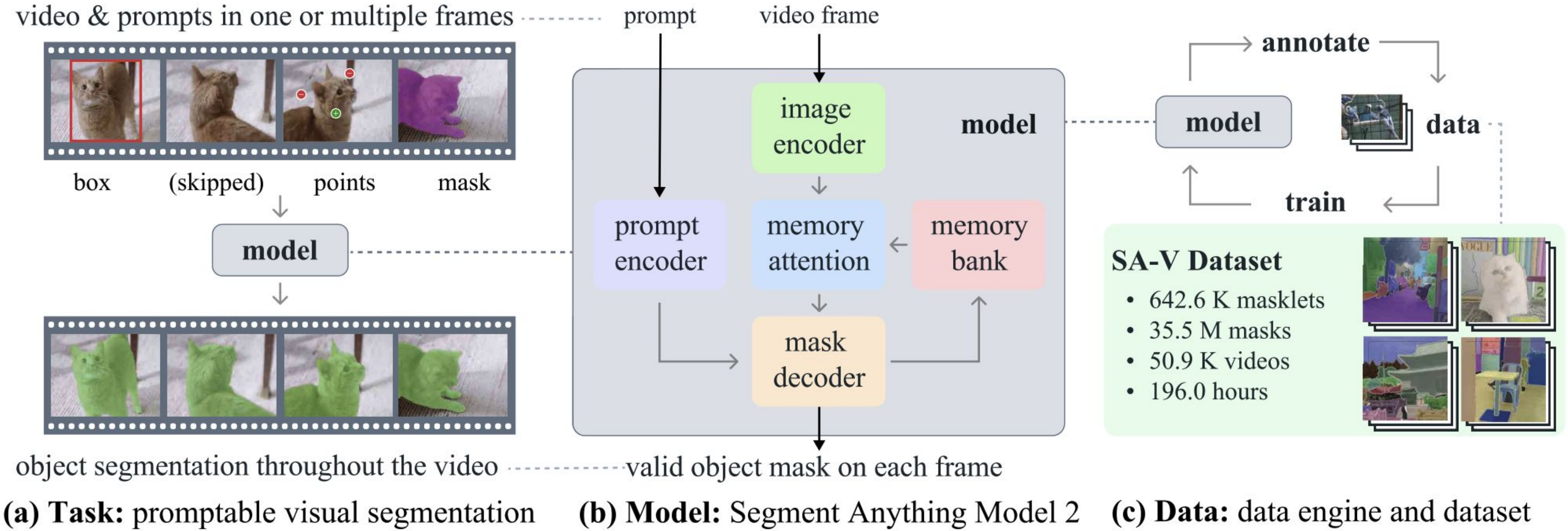
- Foundation model for promptable segmentation, based on a Transformer encoder-decoder architecture. Generalizes to many segmentation tasks.
- Not all representation learning needs to be through self-supervised learning. Here, a supervised paradigm that also achieves powerful representation learning.
- Trained on 1 billion masks from 11 million images, using the model in a data collection loop.



(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

Kirrilov et al. 2023.

Segment Anything Model 2 (SAM 2): Extending to videos

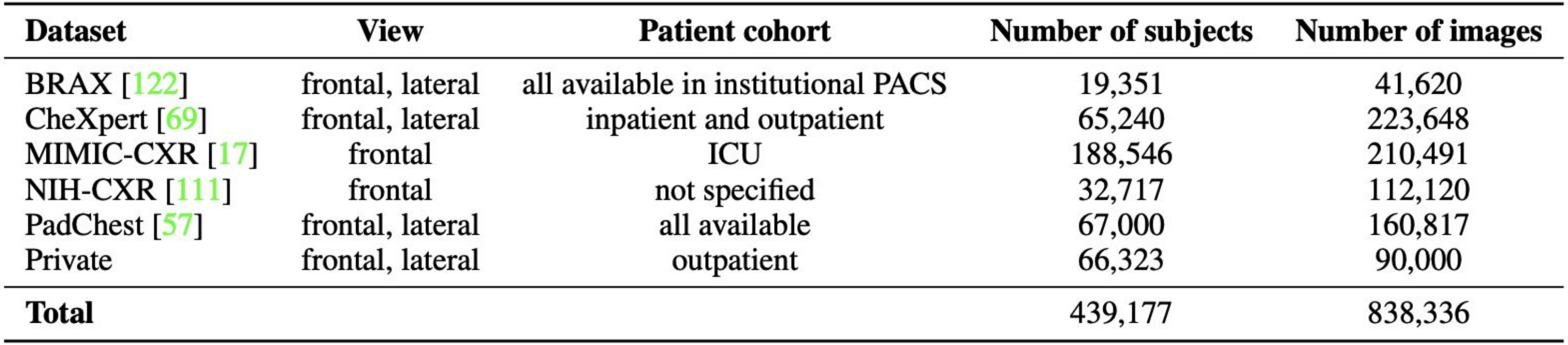


(a) Task: promptable visual segmentation **(b) Model:** Segment Anything Model 2 **(c) Data:** data engine and dataset

Ravi et al. 2024

RAD-DINO: Improving scale and model for chest X-ray pre-training

- Used the DINOv2 self-supervised method, with ViT architecture. Started from public DINOv2 ViT-B/14 model and continued training with CXR images.
- Trained on an extended CXR dataset, Multi-CXR, aggregating multiple datasets including CheXpert

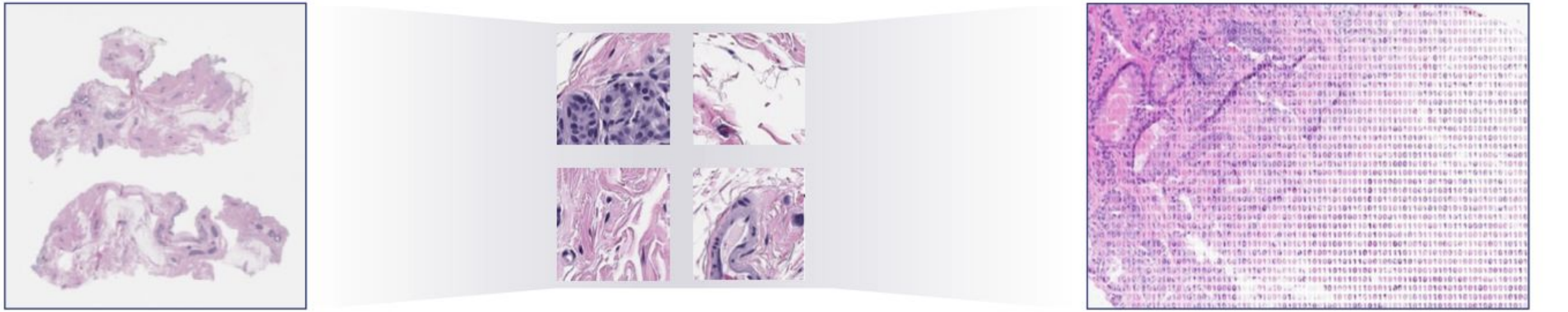


Dataset	View	Patient cohort	Number of subjects	Number of images
BRAX [122]	frontal, lateral	all available in institutional PACS	19,351	41,620
CheXpert [69]	frontal, lateral	inpatient and outpatient	65,240	223,648
MIMIC-CXR [17]	frontal	ICU	188,546	210,491
NIH-CXR [111]	frontal	not specified	32,717	112,120
PadChest [57]	frontal, lateral	all available	67,000	160,817
Private	frontal, lateral	outpatient	66,323	90,000
Total			439,177	838,336

Perez-Garcia et al. RAD-DINO: Exploring scalable medical image encoders beyond text supervision. 2024.

Virchow: Large-scale self-supervised learning on digital pathology images

- Trained a DINO v2 based model on 1.5 million H&E stained whole slide images (WSIs) from Memorial Sloan Kettering Cancer Center and external consults, corresponding to ~100,000 patients



H&E slide

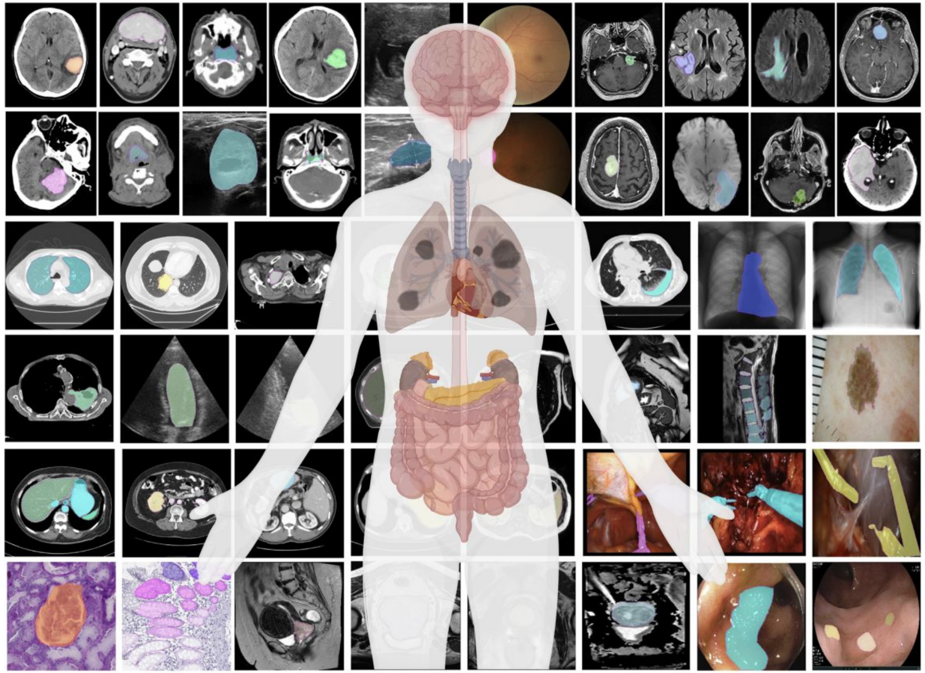
Tissue tiles
224 × 224 pixel crops crops from tissue regions in the slide

Virchow
Foundation model with ViT-H architecture (632 million parameters) trained using DINOv2 framework

Vorontsov et al. A foundation model for clinical-grade computational pathology and rare cancers detection. Nature Medicine, 2024.

MedSAM: Segment anything in medical images

- Tackles universal medical image segmentation across many imaging modalities and tasks
- Trained on a large-scale medical image dataset with 1,570,263 image-mask pairs, covering 10 imaging modalities and over 30 cancer types
- Evaluated on 86 internal validation and 60 external validation tasks



Ma et al. Segment anything in medical images. Nature Communications, 2024.

Two major classes of vision foundation models

Representation Learners

Train on huge amounts of data

Objective: learn powerful generalist feature extractors that can be used for downstream computer vision tasks

- Vision Only (e.g. DINOv2)
- Vision-Language (e.g. CLIP)

Generative Models

Train on huge amounts of data

Objective: learn to generate images and/or text

- Text -> Image (e.g. StableDiffusion, DALL-E)
- Text, Image -> Text (e.g. Flamingo, GPT4-V)

Two major classes of vision foundation models

Representation Learners

Train on huge amounts of data

Objective: learn powerful generalist feature extractors that can be used for downstream computer vision tasks

- Vision Only (e.g. DINOv2)
- Vision-Language (e.g. CLIP)

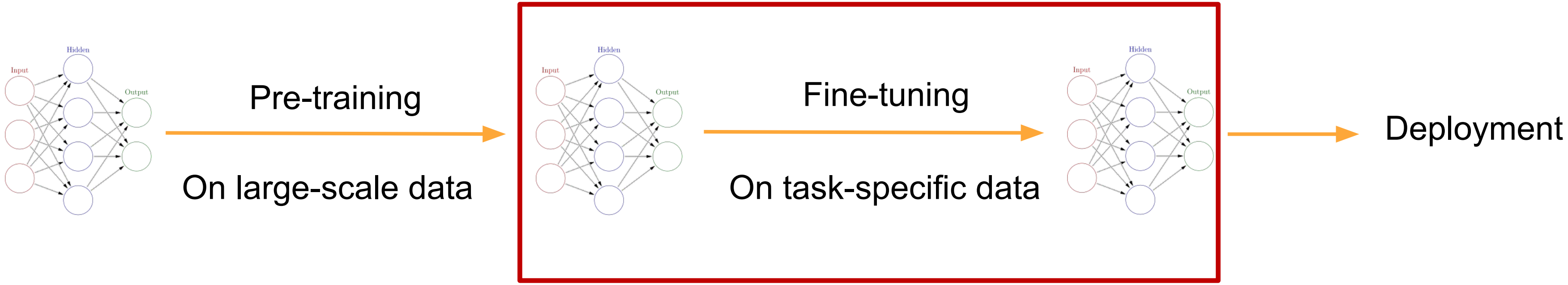
Generative Models

Train on huge amounts of data

Objective: learn to generate images and/or text

- Text -> Image (e.g. StableDiffusion, DALL-E)
- Text, Image -> Text (e.g. Flamingo, GPT4-V)

Adapt to Downstream Task



Pre-training

On large-scale data

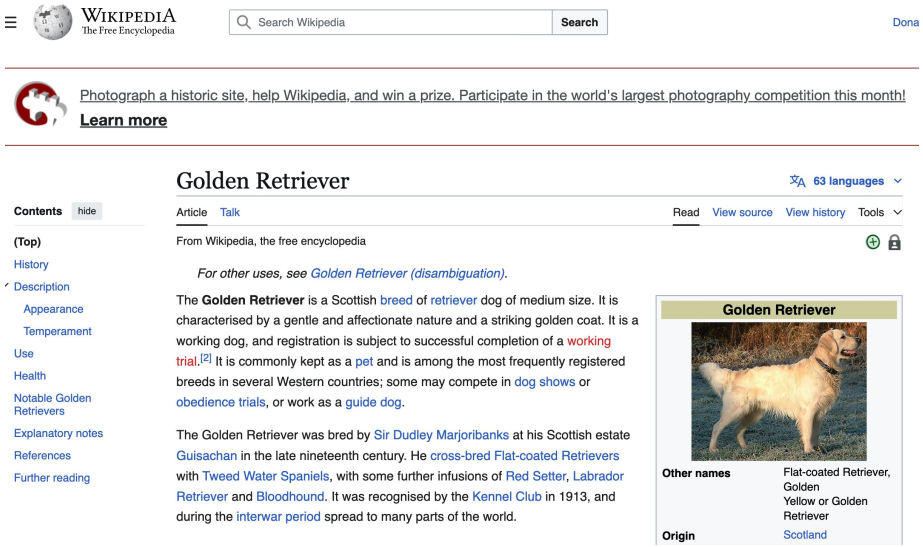
Fine-tuning

On task-specific data

Deployment

How to skip this step to enable zero-shot generalization?

Natural Language Supervision



Photograph a historic site, help Wikipedia, and win a prize. Participate in the world's largest photography competition this month! **Learn more**

Golden Retriever

From Wikipedia, the free encyclopedia

For other uses, see Golden Retriever (disambiguation).

The **Golden Retriever** is a Scottish breed of retriever dog of medium size. It is characterised by a gentle and affectionate nature and a striking golden coat. It is a working dog, and registration is subject to successful completion of a working trial.[2] It is commonly kept as a pet and is among the most frequently registered breeds in several Western countries; some may compete in dog shows or obedience trials, or work as a guide dog.

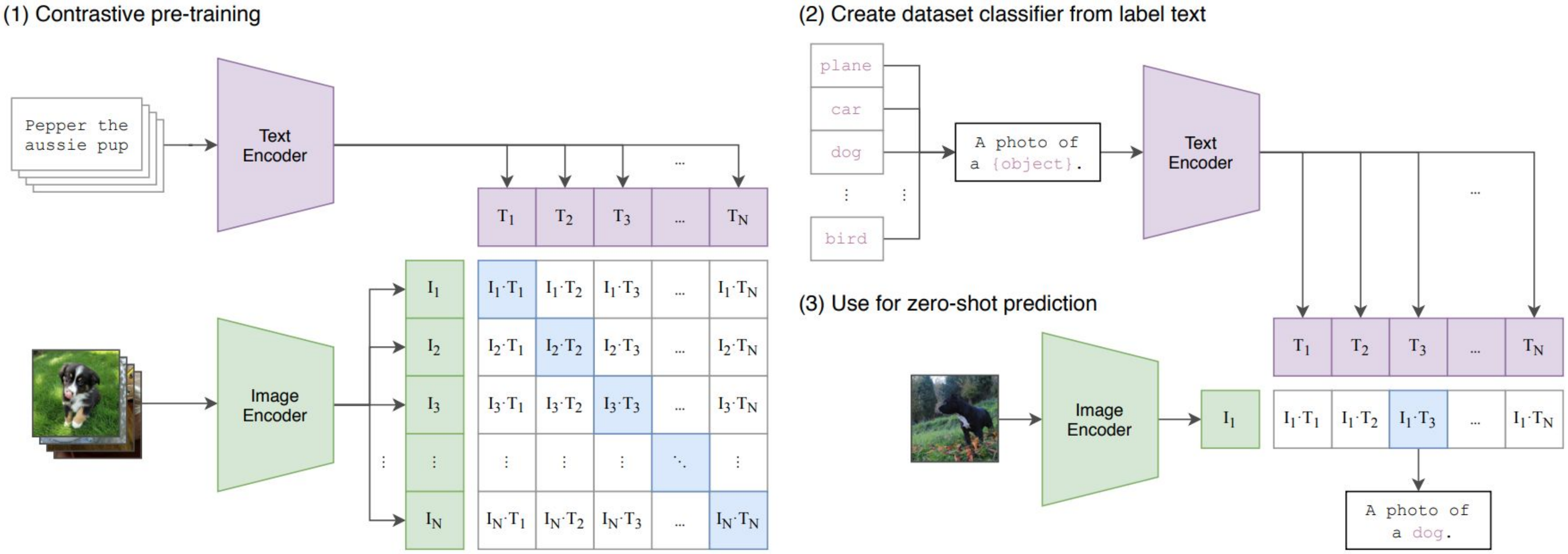
The Golden Retriever was bred by Sir Dudley Marjoribanks at his Scottish estate Guisachan in the late nineteenth century. He cross-bred Flat-coated Retrievers with Tweed Water Spaniels, with some further infusions of Red Setter, Labrador Retriever and Bloodhound. It was recognised by the Kennel Club in 1913, and during the interwar period spread to many parts of the world.

Golden Retriever	
Other names	Flat-coated Retriever, Golden Yellow or Golden Retriever
Origin	Scotland

- ❖ Unlimited Data: vast amount of text-image pairs on the internet

- ❖ Flexible zero-shot transfer: learn a representation that connects visual content to language

CLIP learns a joint representation space for images and text



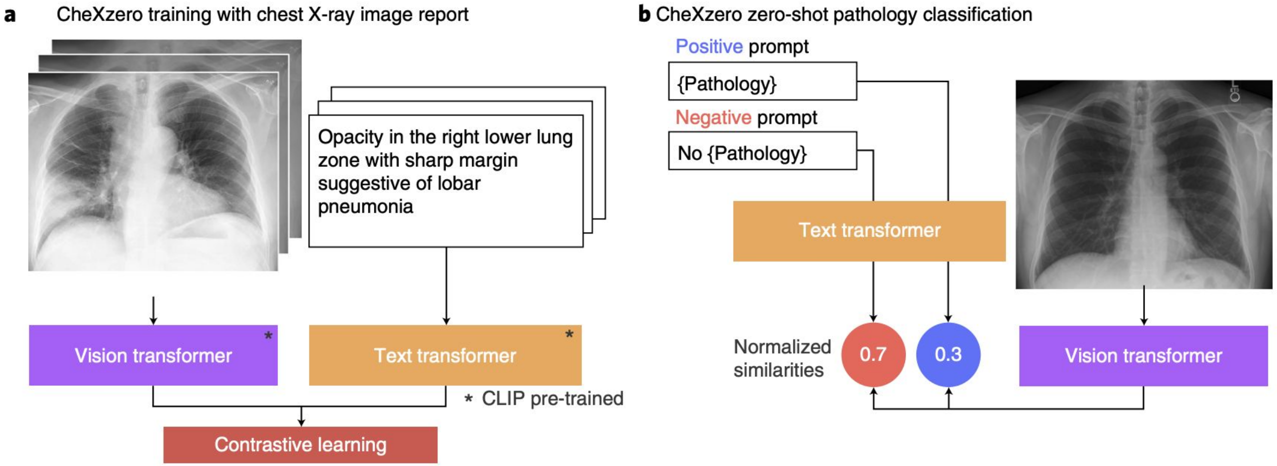
	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...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

	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$

Radford et al. 2021.

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

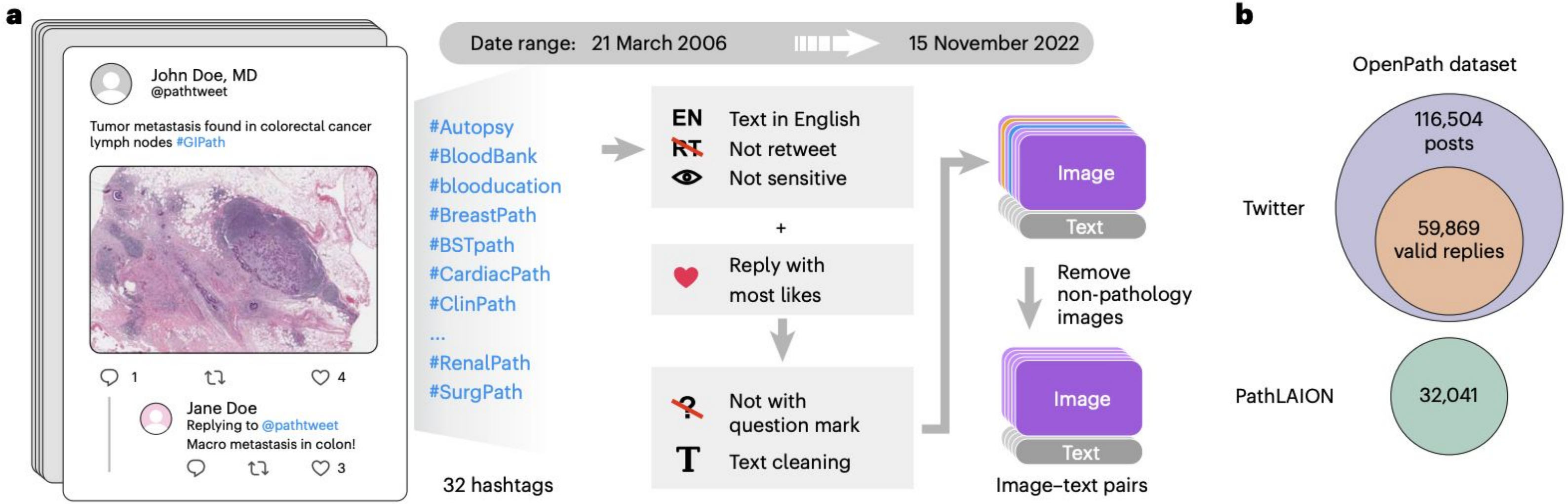
- Compared to ConVIRT, updated to the same architecture as CLIP (better encoders, no nonlinear projection) as well as CLIP-pretrained weights.
- Also used the “impressions” section of the radiology report instead of ConVIRT sampling.



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

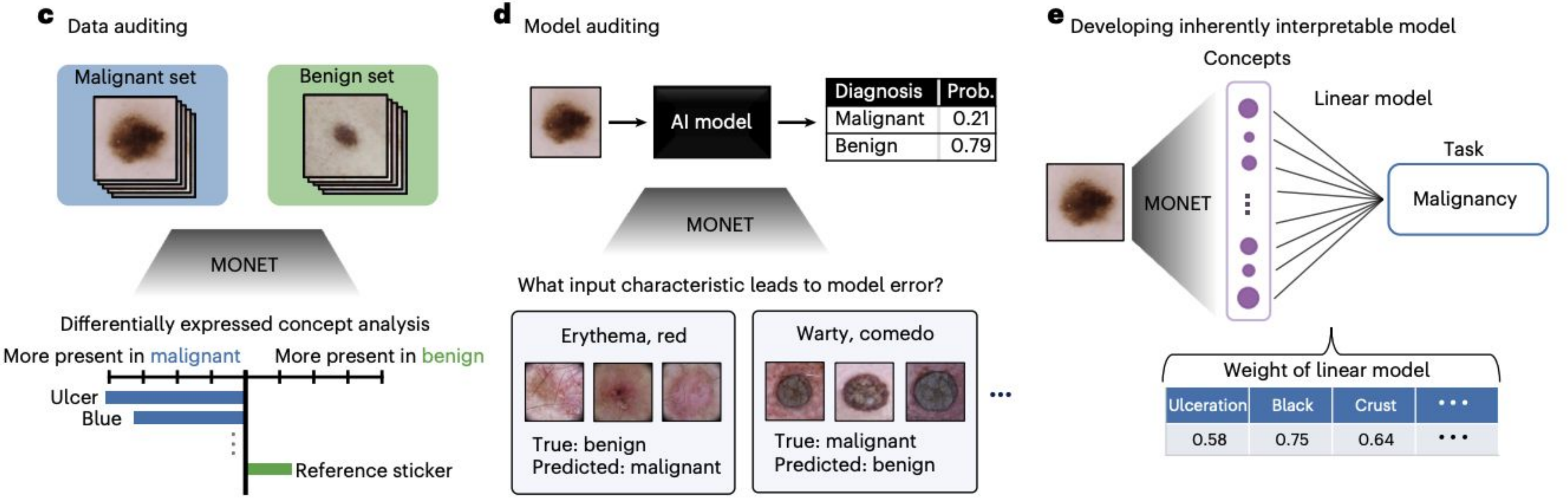
PLIP: vision-language foundation model for pathology trained from Twitter

- 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.

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

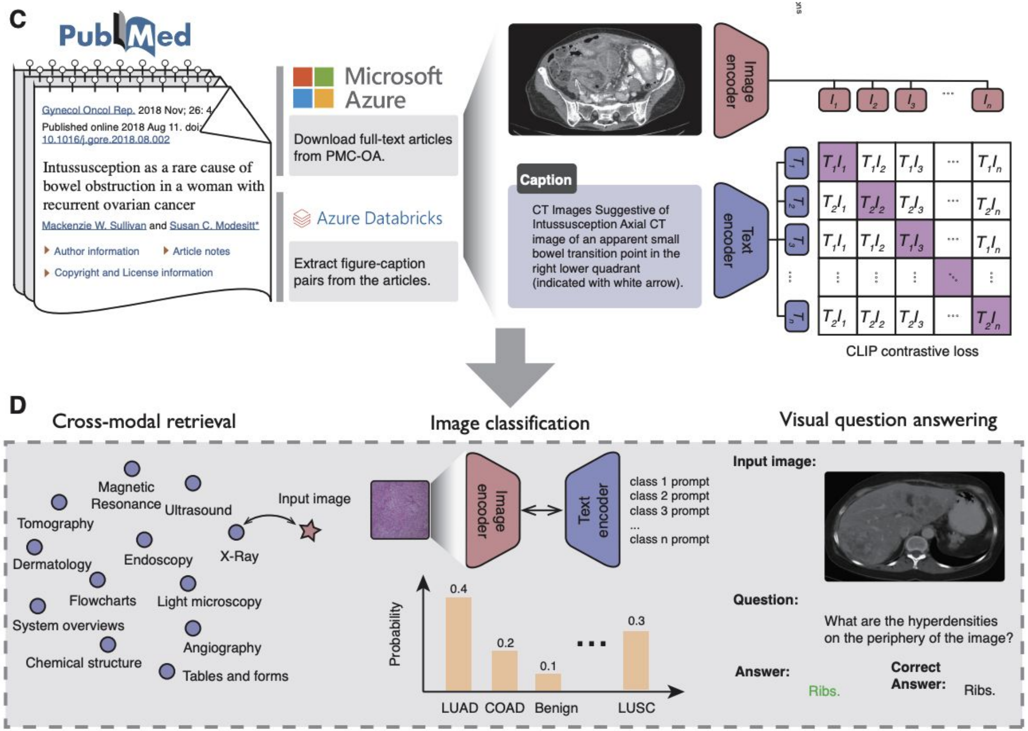


Kim et al. Transparent medical image AI via an image–text foundation model grounded in medical literature. Nature Medicine, 2024.

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


C
PubMed
Gynecol Oncol Rep. 2018 Nov; 26: 4
Published online 2018 Aug 11. doi:
10.1016/j.gore.2018.08.002
Intussusception as a rare cause of bowel obstruction in a woman with recurrent ovarian cancer
Mackenzie W. Sullivan and Susan C. Modesitt*
Author information
Article notes
Copyright and License information
Microsoft Azure
Download full-text articles from PMC-OA.
Azure Databricks
Extract figure-caption pairs from the articles.
Caption
CT Images Suggestive of Intussusception Axial CT image of an apparent small bowel transition point in the right lower quadrant (indicated with white arrow).
Image encoder
I_1 I_2 I_3 ... I_n
Text encoder
T_1 T_2 T_3 ⋮ T_n
T_1I_1 T_1I_2 T_1I_3 ... T_1I_n
T_2I_1 T_2I_2 T_2I_3 ... T_2I_n
T_1I_1 T_1I_2 T_1I_3 ... T_1I_n
T_2I_1 T_2I_2 T_2I_3 ... T_2I_n
CLIP contrastive loss
D
Cross-modal retrieval
Magnetic Resonance
Ultrasound
Tomography
Input image
Endoscopy
X-Ray
Dermatology
Flowcharts
Light microscopy
System overviews
Angiography
Chemical structure
Tables and forms
Image classification
Image encoder
Text encoder
class 1 prompt
class 2 prompt
class 3 prompt
...
class n prompt
Probability
0.4
0.2
0.1
0.3
LUAD COAD Benign LUSC
Visual question answering
Input image:
Question:
What are the hyperdensities on the periphery of the image?
Answer:
Ribs.
Correct Answer:
Ribs.

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

Two major classes of vision foundation models

Representation Learners

Train on huge amounts of data

Objective: learn powerful generalist feature extractors that can be used for downstream computer vision tasks

- Vision Only (e.g. DINOv2)
- Vision-Language (e.g. CLIP)

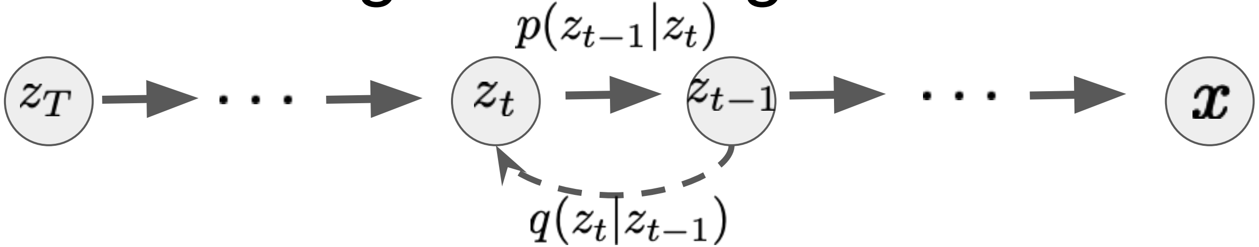
Generative Models

Train on huge amounts of data

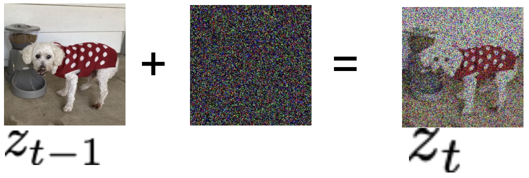
Objective: learn to generate images and/or text

- Text -> Image (e.g. StableDiffusion, DALL-E)
- Text, Image -> Text (e.g. Flamingo, GPT4-V)

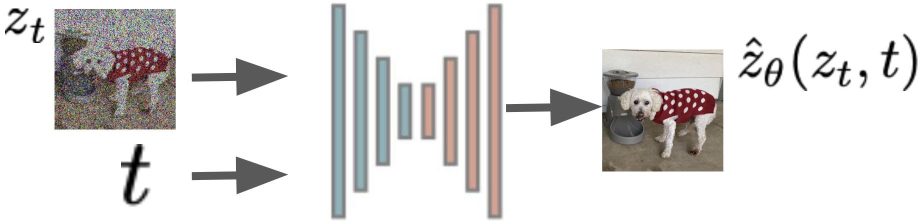
Diffusion training as learning to reverse noising



Sample t,
Apply noise: $q(z_t|z_{t-1})$



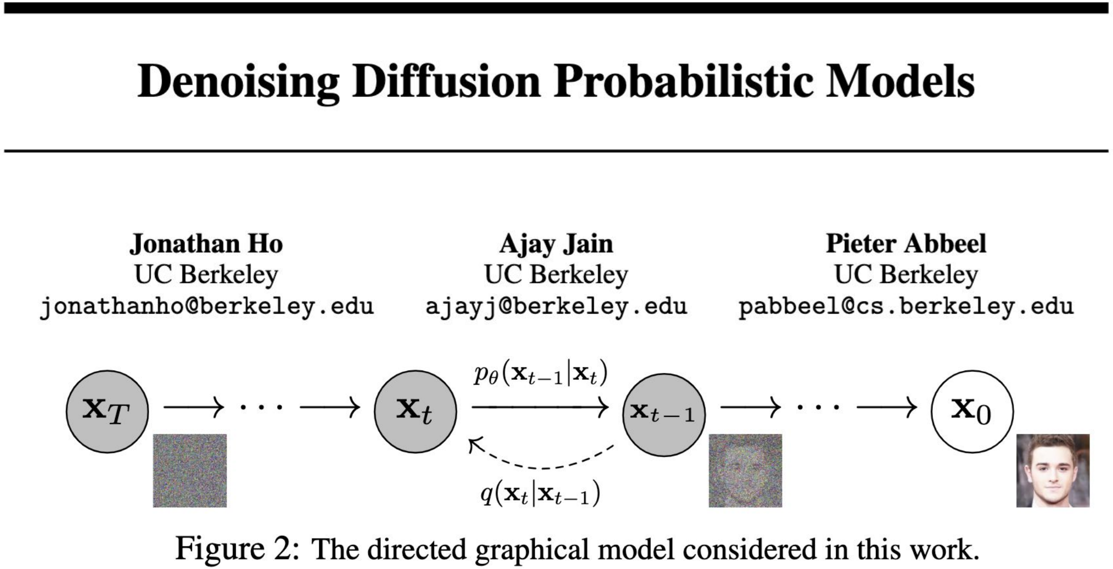
Use Unet to reverse the noise



Reconstruction loss

$$\mathbb{E}_{t\sim U(2,T)}\left[c(\alpha_t)\|\hat{z}_\theta(z_t,t)-z_{t-1}\|\right]$$

DDPM: the original pixel-level diffusion model



Denoising Diffusion Probabilistic Models

Jonathan Ho
UC Berkeley
jonathanho@berkeley.edu

Ajay Jain
UC Berkeley
ajayj@berkeley.edu

Pieter Abbeel
UC Berkeley
pabbeel@cs.berkeley.edu

$\mathbf{x}_T$
$\mathbf{x}_t$
$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$
$\mathbf{x}_{t-1}$
$q(\mathbf{x}_t|\mathbf{x}_{t-1})$
$\mathbf{x}_0$

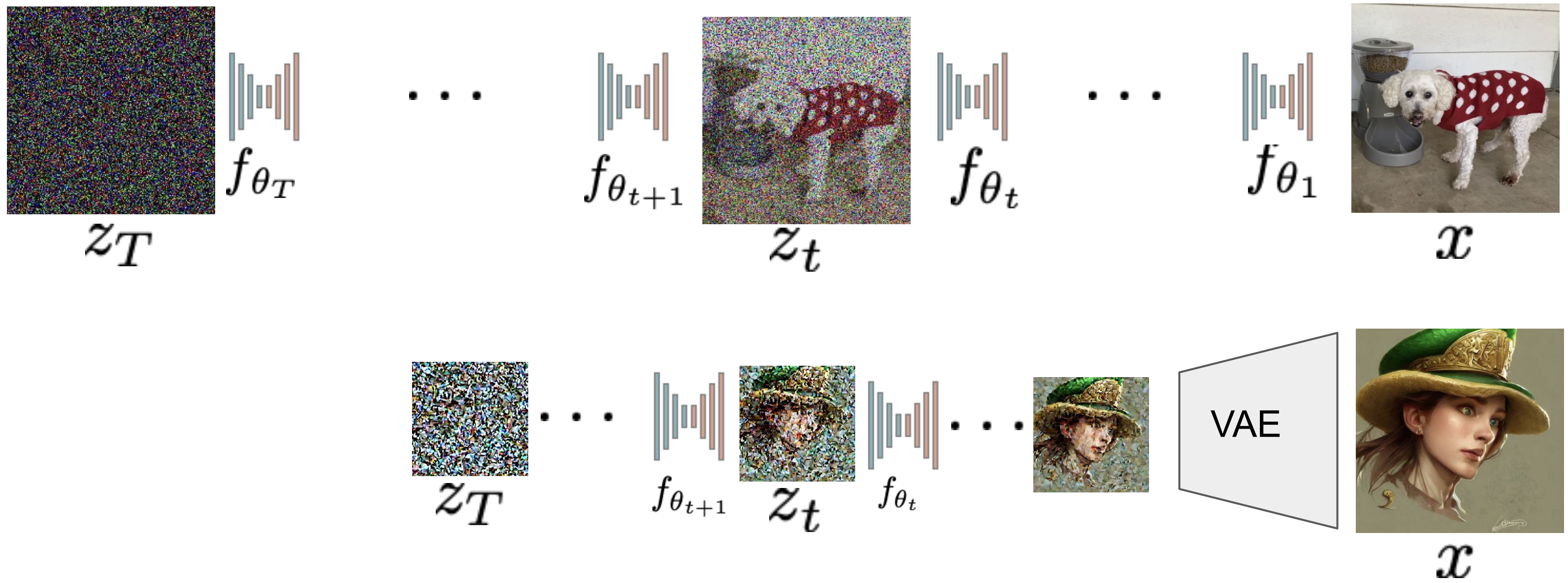
Figure 2: The directed graphical model considered in this work.

Stable Diffusion or Latent Variable Models

Idea: instead of diffusion models on *pixels*, do diffusion in a *compressed latent space*

Diffusion (DDPM)

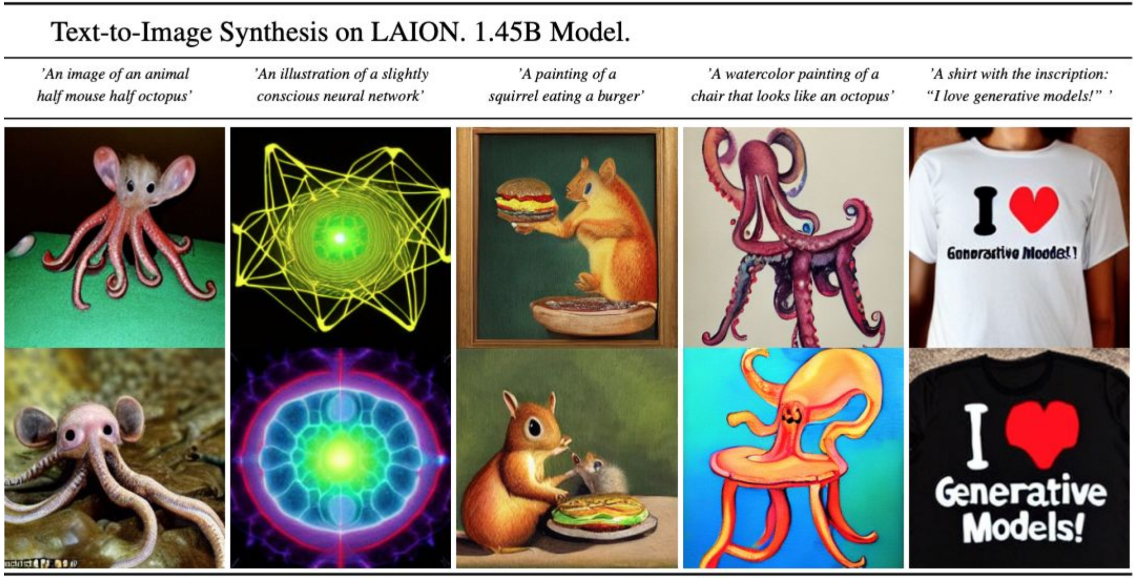
Latent diffusion (Stable Diffusion)



“High-Resolution Image Synthesis with Latent Diffusion Models”
“Score-based Generative Modeling in Latent Space”

Stable Diffusion: example of text -> image generative model (+ other types of conditioning!)

Text -> image model trained on LAION: open-source dataset of 400M image-text pairs



Text-to-Image Synthesis on LAION. 1.45B Model.

'An image of an animal half mouse half octopus'

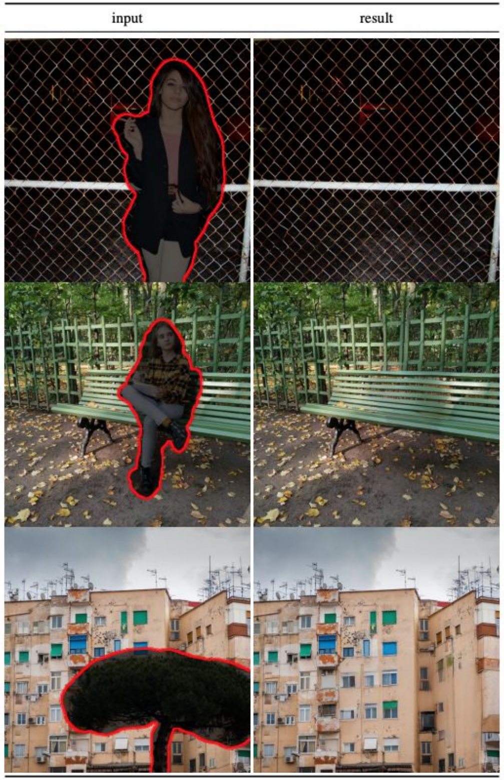
'An illustration of a slightly conscious neural network'

'A painting of a squirrel eating a burger'

'A watercolor painting of a chair that looks like an octopus'

'A shirt with the inscription: "I love generative models!" '

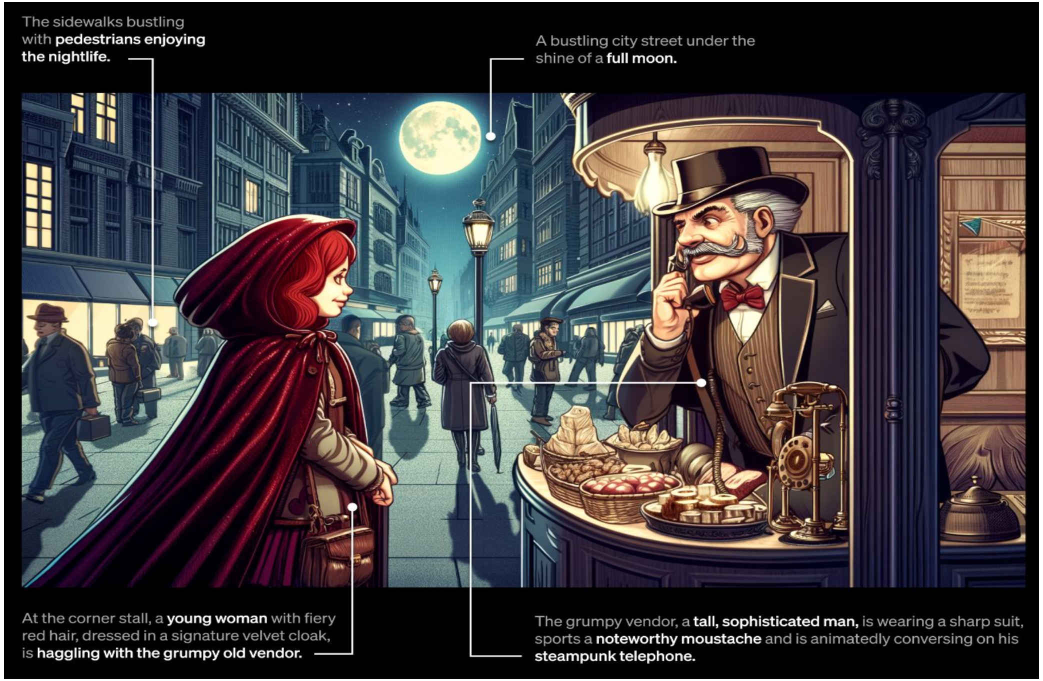
Condition on provided input images to remove objects



Rombach et al. 2022.

DALL-E: related family of OpenAI models that focus on high quality text-conditioned image generation

DALL-E 3 is trained on large amounts of detailed, synthetically generated captions to enable high-fidelity and high-detail generation



The sidewalks bustling with **pedestrians enjoying the nightlife.**

A bustling city street under the shine of a **full moon.**

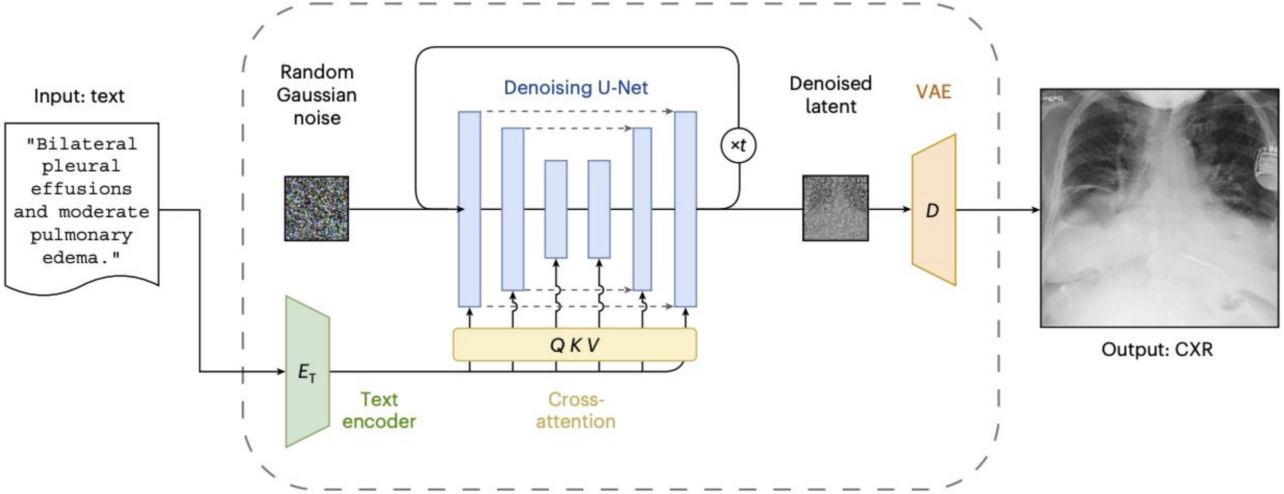
At the corner stall, a **young woman** with fiery red hair, dressed in a signature velvet cloak, is **haggling with the grumpy old vendor.**

The grumpy vendor, a **tall, sophisticated man,** is wearing a sharp suit, sports a **noteworthy moustache** and is animatedly conversing on his **steampunk telephone.**

https://openai.com/dall-e-3

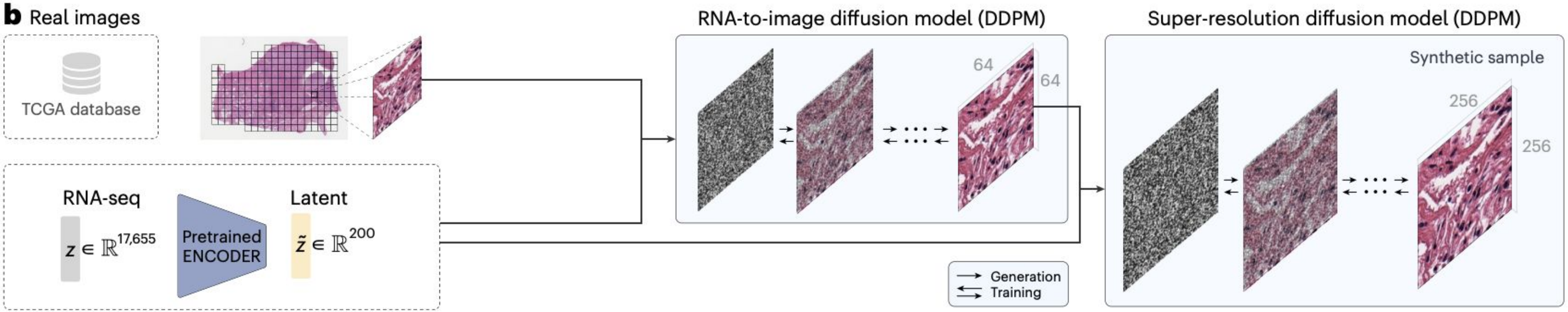
One widely considered application area of vision generative models: augmenting training data

- RoentGen fine-tunes the Stable Diffusion model on the MIMIC-CXR dataset of chest x-ray (CXR) images and corresponding text reports (about ~175k images for training)



Bluethgen et al. A vision–language foundation model for the generation of realistic chest X-ray images. Nature Biomedical Engineering 2024.

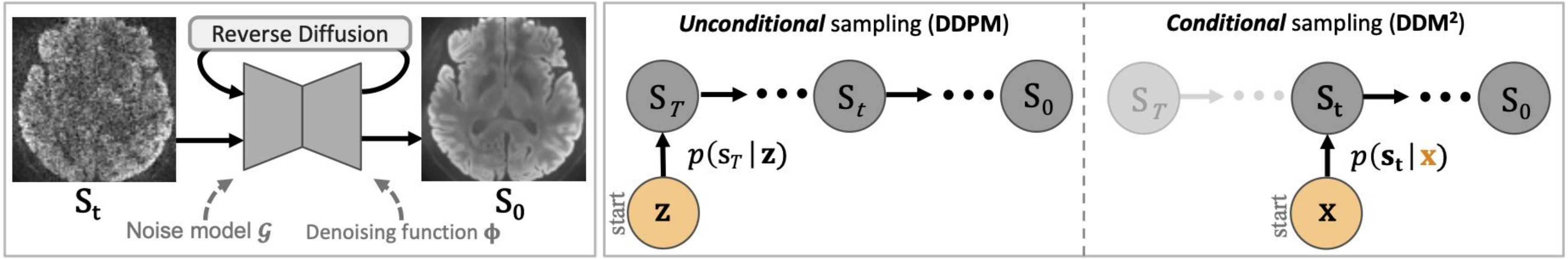
Preview to discussion paper presentations: generating pathology whole-slide image titles conditioned on RNA-sequencing


b
Real images
TCGA database
RNA-seq
$z \in \mathbb{R}^{17,655}$
Pretrained ENCODER
Latent
$\tilde{z} \in \mathbb{R}^{200}$
RNA-to-image diffusion model (DDPM)
64
64
Super-resolution diffusion model (DDPM)
Synthetic sample
256
256
Generation
Training

Carrillo-Perez et al. Generation of synthetic whole-slide image tiles of tumours from RNA-sequencing data via cascaded diffusion models. Nature Biomedical Engineering 2024.

Diffusion models for MRI denoising: DDM^2

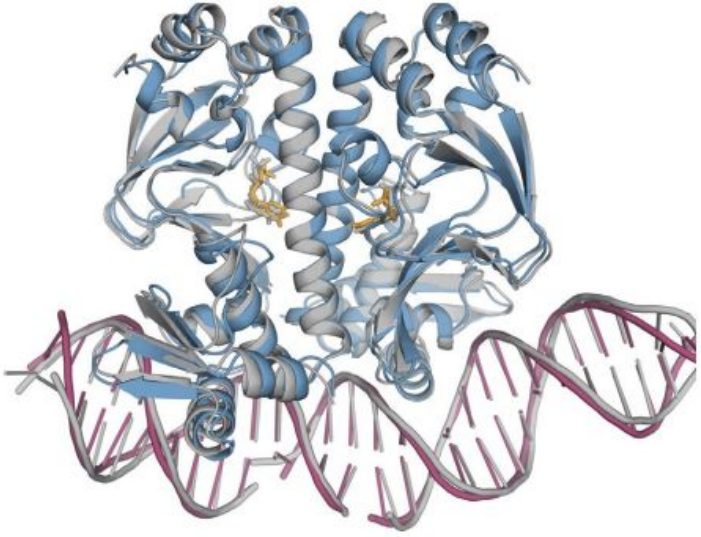
- MRI denoising is a key problem in modern MRI research due to the trade-off between achieving high signal-to-noise MRI scans and requiring long scan time (increased cost and discomfort, harder to accommodate overall patient demand)
- DDM^2 (Denoising Diffusion Models for Denoising Diffusion MRI) uses diffusion models to achieve this denoising. It conditionally samples an image generation based on a noisy image as condition, and matches to an intermediate timestep in the diffusion process.

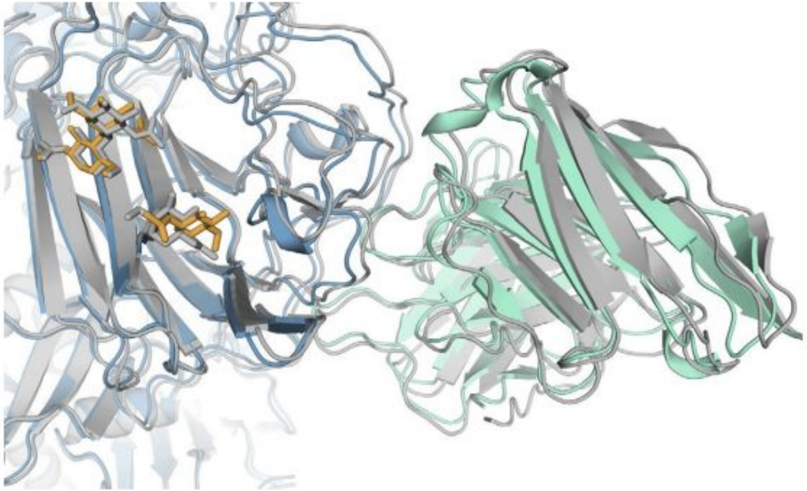

Reverse Diffusion
S_t
S_0
Noise model $\mathcal{G}$
Denoising function ϕ
Unconditional sampling (DDPM)
S_T
S_t
S_0
$p(s_T | \mathbf{z})$
z
start
Conditional sampling (DDM²)
S_T
S_t
S_0
$p(\mathbf{s}_t | \mathbf{x})$
x
start

Xiang et al. DDM2: Self-Supervised Diffusion MRI Denoising with Generative Diffusion Models. ICLR 2023.

AlphaFold 3

- Diffusion-based architecture that predicts the joint 3D structure of complexes including proteins, nucleic acids, small molecules, etc. from sequences





Abramson et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 2024.

Two major classes of vision foundation models

Representation Learners

Train on huge amounts of data

Objective: learn powerful generalist feature extractors that can be used for downstream computer vision tasks

- Vision Only (e.g. DINOv2)
- Vision-Language (e.g. CLIP)

Generative Models

Train on huge amounts of data

Objective: learn to generate images and/or text

- Text -> Image (e.g. StableDiffusion, DALL-E)
- Text, Image -> Text (e.g. Flamingo, GPT4-V)

Two major classes of vision foundation models

Representation Learners

Train on huge amounts of data

Objective: learn powerful generalist feature extractors that can be used for downstream computer vision tasks

- Vision Only (e.g. DINOv2)
- Vision-Language (e.g. CLIP)

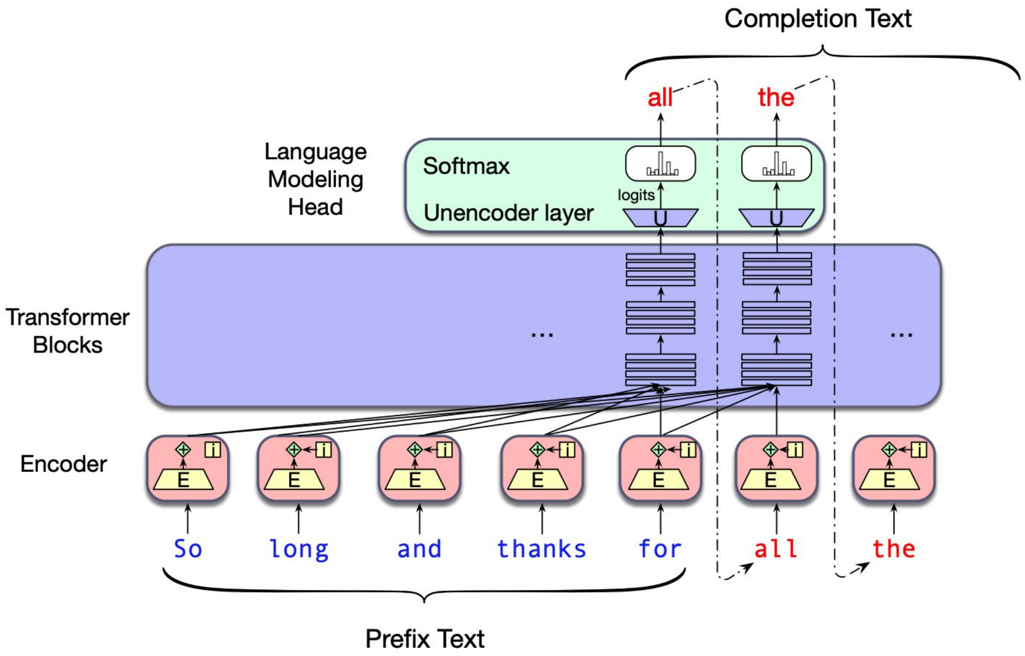
Generative Models

Train on huge amounts of data

Objective: learn to generate images and/or text

- Text -> Image (e.g. StableDiffusion, DALL-E)
- Text, Image -> Text (e.g. Flamingo, GPT4-V)

Autoregressive Language Models



LARGE LANGUAGE MODELS WITH TRANSFORMERS (Daniel Jurafsky & James H. Martin 2024)

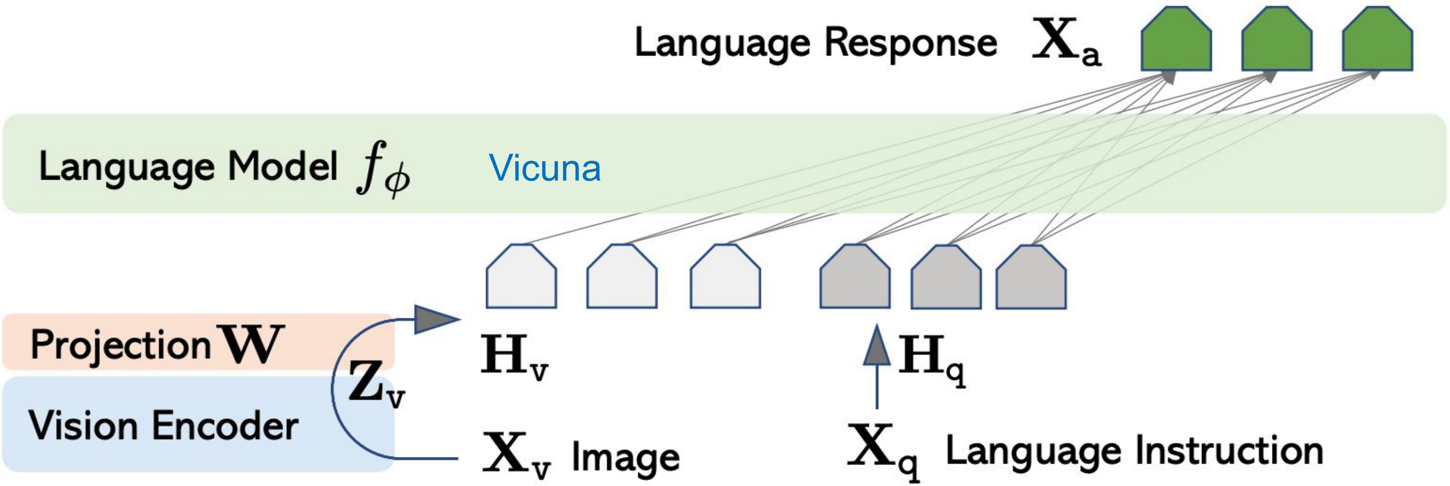
How to Link Vision to LLMs?

Internal Linkage → Vision-Language Models

- ❖ Architecture
  - ○ Integrate Visual Features into Intermediate Layers
  - ○ Integrate Visual Features into Input Layer
  - ○ Integrate Visual Patches into Input Layer

External Linkage → Vision-Language Agents

LLaVA: Integrate Visual Features into Input Layer



MLP

Vicuna

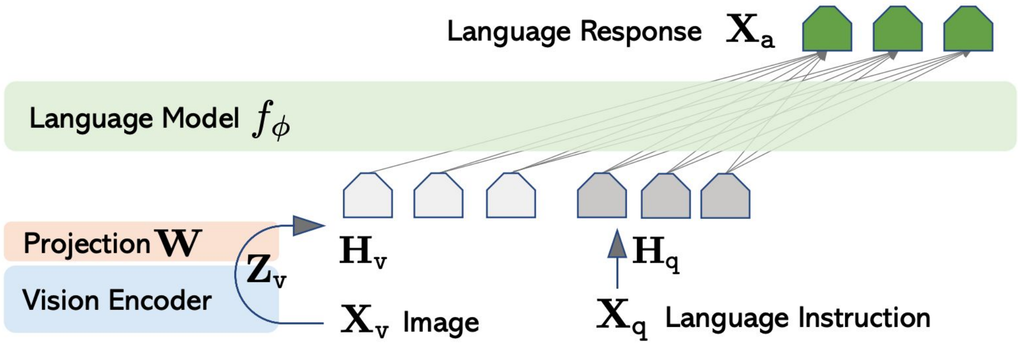
CLIP Vision Encoder
ViT-L/14

Stage 1: Pre-training for Feature Alignment
Stage 2: Fine-tuning End-to-End (W *and* Ø)

LLaVA (Liu et al. 2023)

LLaVA Training Recipe, Stage 1: Pre-training

Goal: Align visual features to LLM's word embedding space



Data: Converted from image captioning data

- 595K image-text pairs filtered from CC3M
- convert to instruction-following format

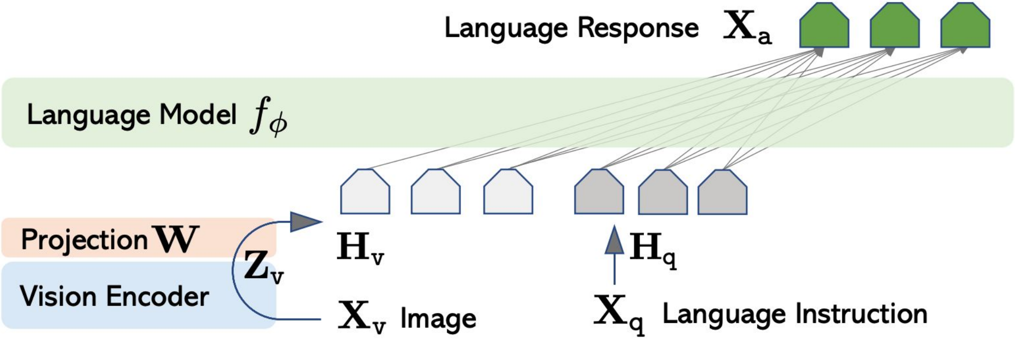
Trainable Parameters

- Only W

LLaVA (Liu et al. 2023)

LLaVA Training Recipe, Stage 2: Instruction Tuning (Supervised Finetuning)

Goal: Visual Captioner → Visual Assistant (Follow user instructions)



Data: leverage ChatGPT/GPT-4 for multimodal instruction-following data collection

- 158K language-image instruction-following data
- 3 Types: Conversation (Multi-Turn), Detailed description, Complex reasoning

Trainable Parameters

- $\{W, \emptyset\}$

LLaVA (Liu et al. 2023)

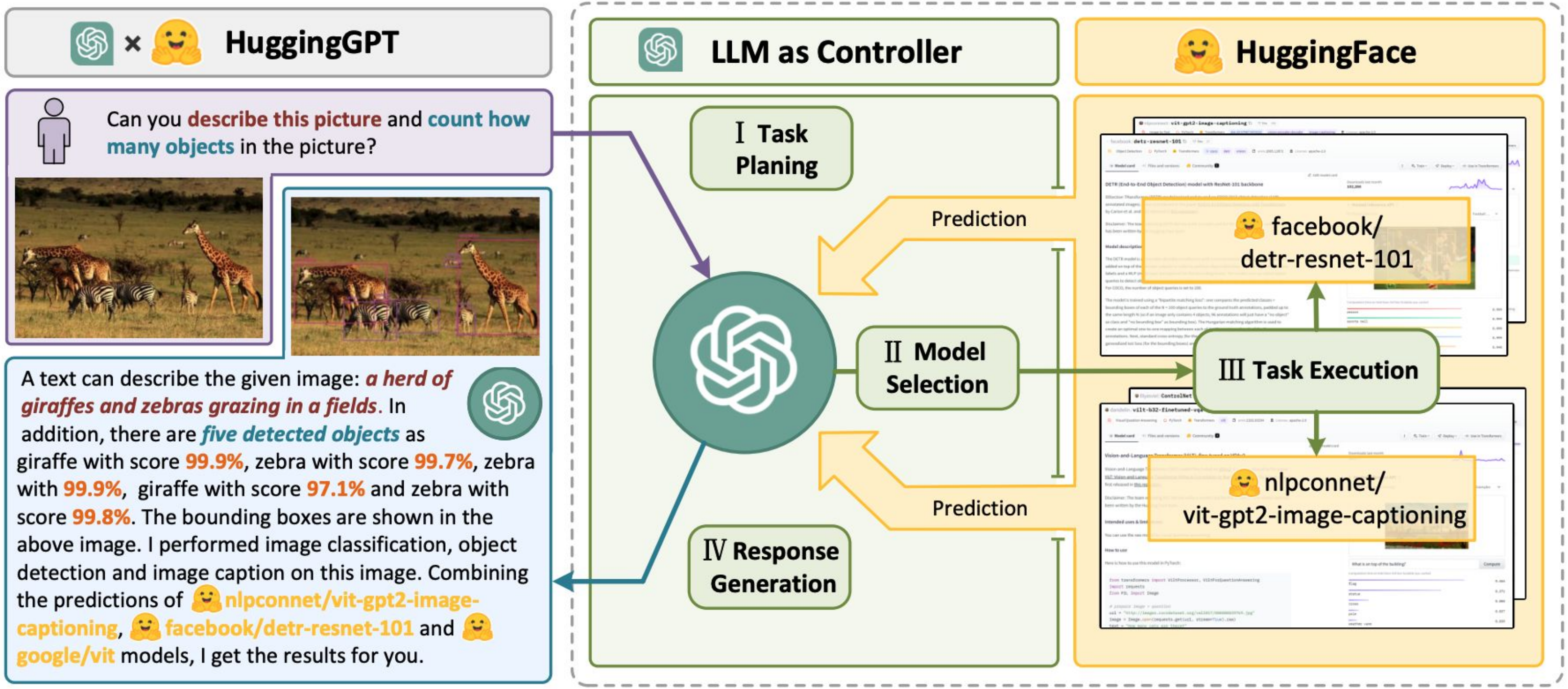
How to Link Vision to LLMs?

Internal Linkage → Vision-Language Models

- ❖ Architecture
  - ○ Integrate Visual Features into Intermediate Layers
  - ○ Integrate Visual Features into Input Layer
  - ○ Integrate Visual Patches into Input Layer

External Linkage → Vision-Language Agents

Language to Connect Vision Models

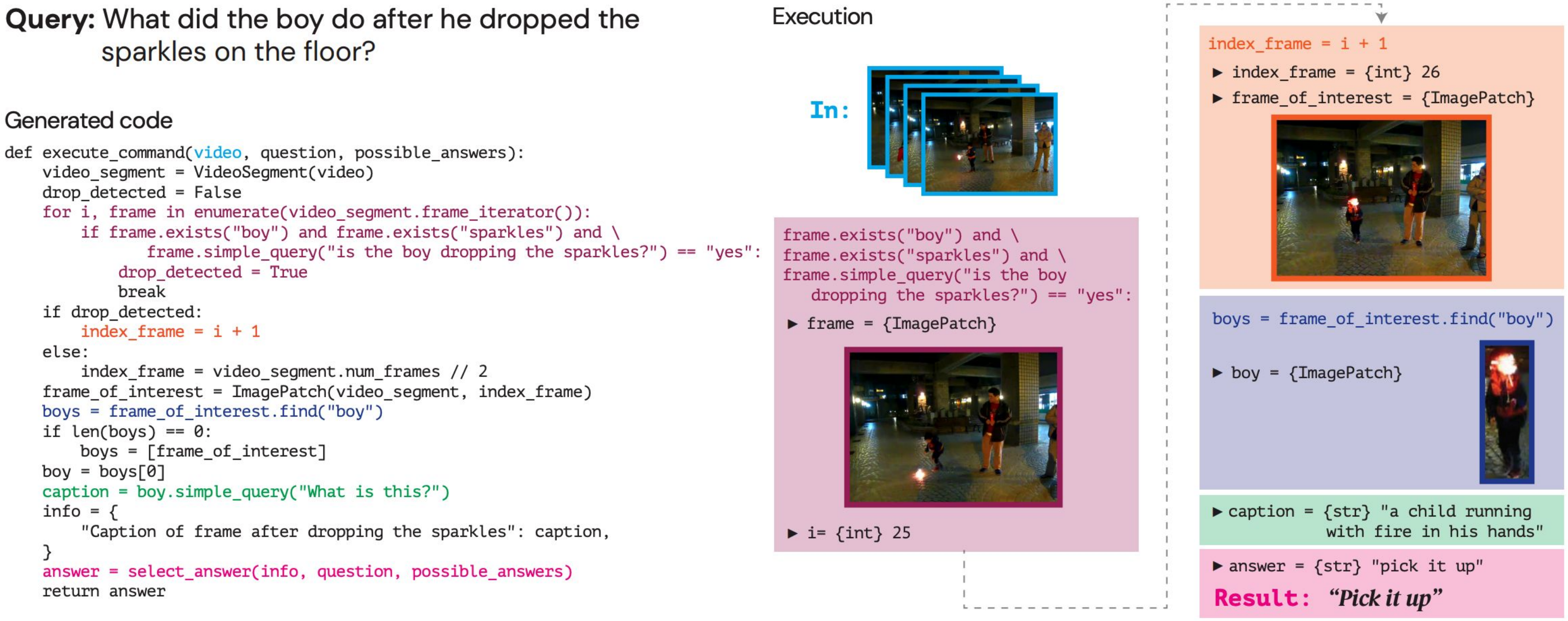


× HuggingGPT
LLM as Controller
HuggingFace
Can you describe this picture and count how many objects in the picture?
Ⅰ Task Planing
Prediction
facebook/ detr-resnet-101
Ⅱ Model Selection
Ⅲ Task Execution
nlpconnet/ vit-gpt2-image-captioning
Prediction
Ⅳ Response Generation

A text can describe the given image: *a herd of giraffes and zebras grazing in a fields*. In addition, there are *five detected objects* as giraffe with score 99.9%, zebra with score 99.7%, zebra with 99.9%, giraffe with score 97.1% and zebra with score 99.8%. The bounding boxes are shown in the above image. I performed image classification, object detection and image caption on this image. Combining the predictions of nlpconnet/vit-gpt2-image-captioning, facebook/detr-resnet-101 and google/vit models, I get the results for you.

HuggingGPT (Shen et al. 2023)

Code to Connect Vision Models



Query: What did the boy do after he dropped the sparkles on the floor?

Generated code

```
def execute_command(video, question, possible_answers):
    video_segment = VideoSegment(video)
    drop_detected = False
    for i, frame in enumerate(video_segment.frame_iterator()):
        if frame.exists("boy") and frame.exists("sparkles") and \
                frame.simple_query("is the boy dropping the sparkles?") == "yes":
            drop_detected = True
            break
    if drop_detected:
        index_frame = i + 1
    else:
        index_frame = video_segment.num_frames // 2
    frame_of_interest = ImagePatch(video_segment, index_frame)
    boys = frame_of_interest.find("boy")
    if len(boys) == 0:
        boys = [frame_of_interest]
    boy = boys[0]
    caption = boy.simple_query("What is this?")
    info = {
        "Caption of frame after dropping the sparkles": caption,
    }
    answer = select_answer(info, question, possible_answers)
    return answer
```

Execution

In:

```
frame.exists("boy") and \
frame.exists("sparkles") and \
frame.simple_query("is the boy
   dropping the sparkles?") == "yes":
▶ frame = {ImagePatch}
▶ i= {int} 25
```

```
index_frame = i + 1
▶ index_frame = {int} 26
▶ frame_of_interest = {ImagePatch}
```

```
boys = frame_of_interest.find("boy")
▶ boy = {ImagePatch}
```

```
▶ caption = {str} "a child running
              with fire in his hands"
```

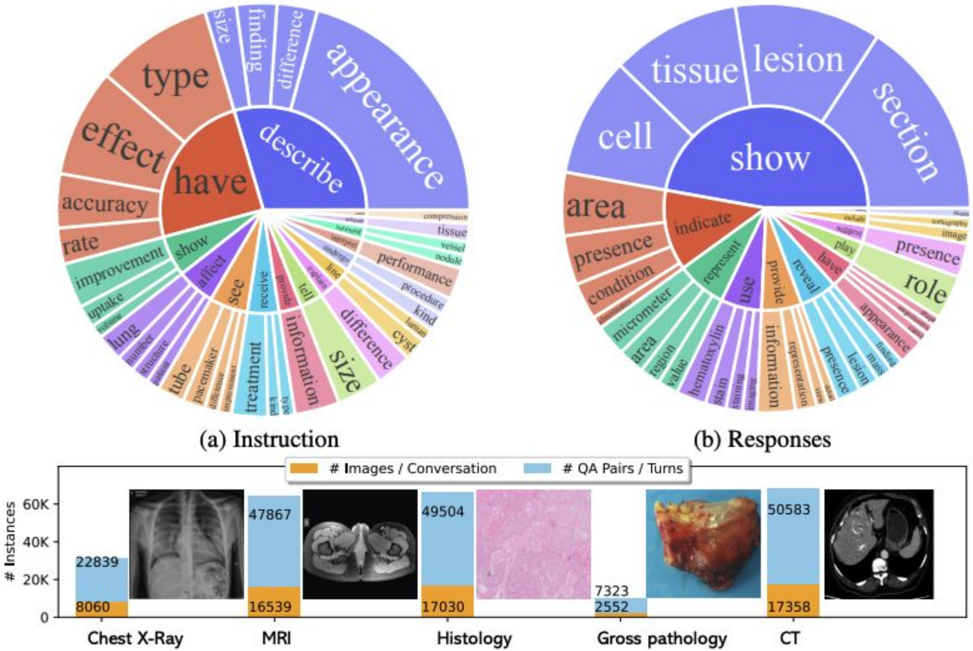
```
▶ answer = {str} "pick it up"
```

Result: *"Pick it up"*

ViperGPT (Surís et al. 2023)

LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day

- Extends LLaVA to better answer biomedical questions, using a new biomedical instruction tuning dataset
- Instruction tuning dataset leverages PMC-15M (PubMed Central figure-caption dataset) and covers diverse domains
- Efficiently trained in < 15 hours using eight A100s



(a) Instruction

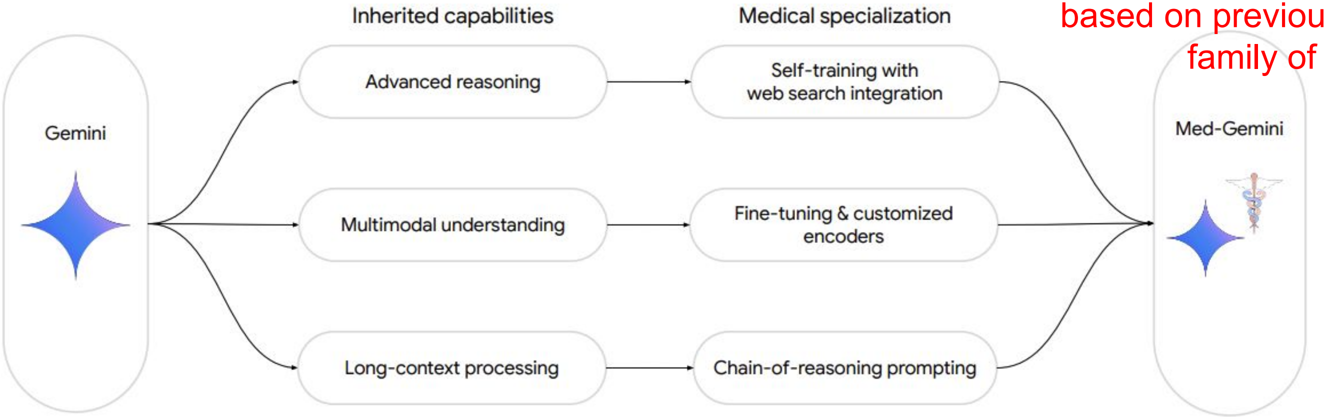
(b) Responses

Li et al. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. NeurIPS Datasets and Benchmarks 2023.

Med-Gemini: state-of-the-art generalist biomedical VLM

- Also a generalist (broad domains) biomedical VLM like LLaVA-Med, but extends from the much more powerful Gemini models (and built by internal Google DeepMind team)
- SoTA (state-of-the-art) due to Gemini foundation and additional techniques for medical specialization

Successor to Med-PaLM model based on previous Google PaLM family of models

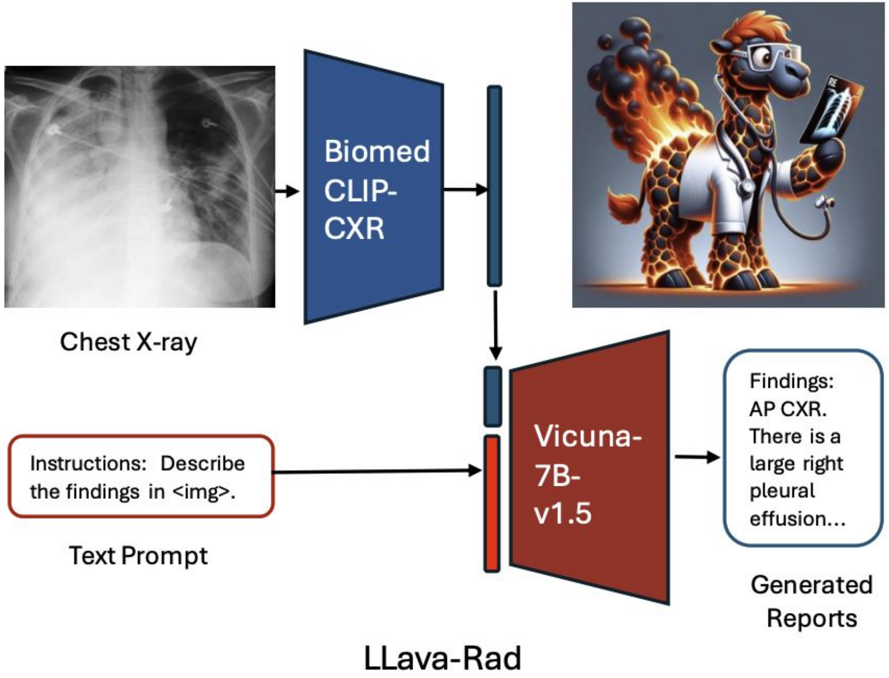


Saab et al. Capabilities of Gemini Models in Medicine. arXiv 2024.

LLaVA-Rad: specialized CXR VLM that improves over CheXpert

- Based on LLaVA (similar to LLaVA-Med), but trains from scratch (instead of fine-tuning LLaVA) since it uses a biomedical encoder (Biomed CLIP-CXR) instead of LLaVA’s CLIP encoder.

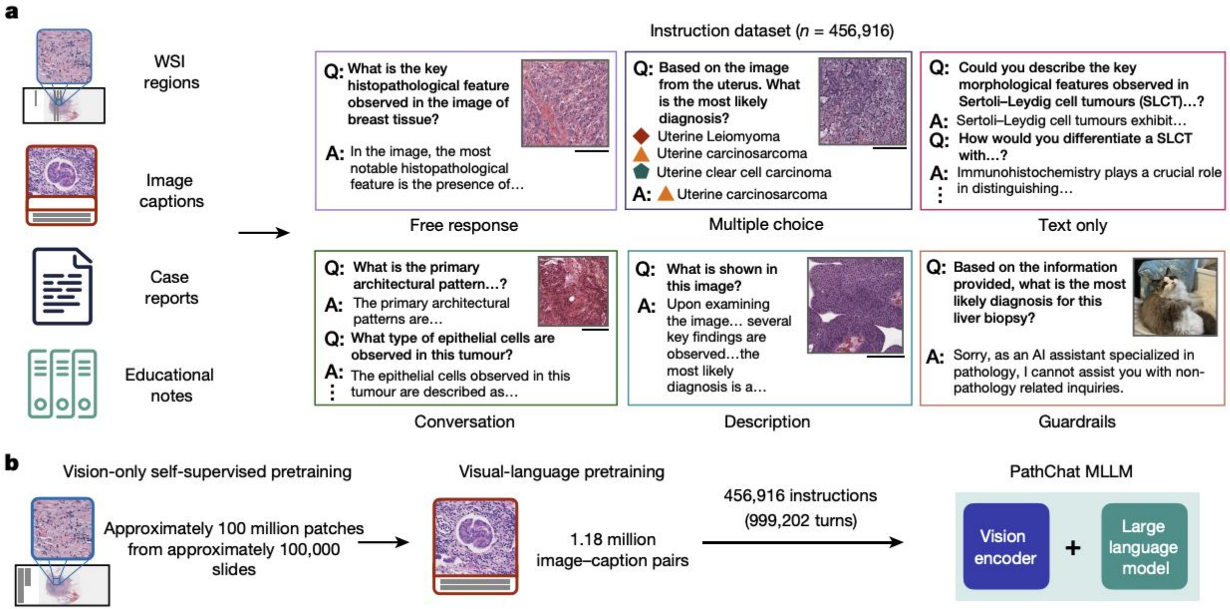
- Trains using 697K image-report pairs from a collection of datasets


Chest X-ray
Biomed CLIP-CXR
Instructions: Describe the findings in <img>.
Text Prompt
Vicuna-7B-v1.5
Findings: AP CXR. There is a large right pleural effusion...
Generated Reports
LLava-Rad

Chaves et al. Towards a clinically accessible radiology multimodal model: open-access and lightweight, with automatic evaluation. arXiv 2024.

PathChat: specialized VLM for pathology

Trained on over 450K instructions from pathology. Will cover further in the discussion presentations!


a
WSI regions
Image captions
Case reports
Educational notes
Instruction dataset (*n* = 456,916)
Free response
Multiple choice
Text only
Conversation
Description
Guardrails
b
Vision-only self-supervised pretraining
Approximately 100 million patches from approximately 100,000 slides
Visual-language pretraining
1.18 million image–caption pairs
456,916 instructions (999,202 turns)
PathChat MLLM
Vision encoder
+
Large language model

Q: What is the key histopathological feature observed in the image of breast tissue?
A: In the image, the most notable histopathological feature is the presence of…

Q: Based on the image from the uterus. What is the most likely diagnosis?
◆ Uterine Leiomyoma
▲ Uterine carcinosarcoma
⬟ Uterine clear cell carcinoma
A: ▲ Uterine carcinosarcoma

Q: Could you describe the key morphological features observed in Sertoli–Leydig cell tumours (SLCT)…?
A: Sertoli–Leydig cell tumours exhibit…
Q: How would you differentiate a SLCT with…?
A: Immunohistochemistry plays a crucial role in distinguishing…
⋮

Q: What is the primary architectural pattern…?
A: The primary architectural patterns are…
Q: What type of epithelial cells are observed in this tumour?
A: The epithelial cells observed in this tumour are described as…
⋮

Q: What is shown in this image?
A: Upon examining the image… several key findings are observed…the most likely diagnosis is a…

Q: Based on the information provided, what is the most likely diagnosis for this liver biopsy?
A: Sorry, as an AI assistant specialized in pathology, I cannot assist you with non-pathology related inquiries.

Lu et al. A multimodal generative AI copilot for human pathology. Nature 2024.

Two major classes of vision foundation models

Representation Learners

Train on huge amounts of data

Objective: learn powerful generalist feature extractors that can be used for downstream computer vision tasks

- Vision Only (e.g. DINOv2)
- Vision-Language (e.g. CLIP)

Generative Models

Train on huge amounts of data

Objective: learn to generate images and/or text

- Text -> Image (e.g. StableDiffusion, DALL-E)
- Text, Image -> Text (e.g. Flamingo, GPT4-V)

GPU Requirement for Models

VLM LLAVA-1.5

- train all models with 8×A100s. 6 hours of pretraining and 20 hours of visual instruction tuning

Vision FM DINO

- A batch size of 1024, distributed over 16 GPUs. Trained for 3 days.

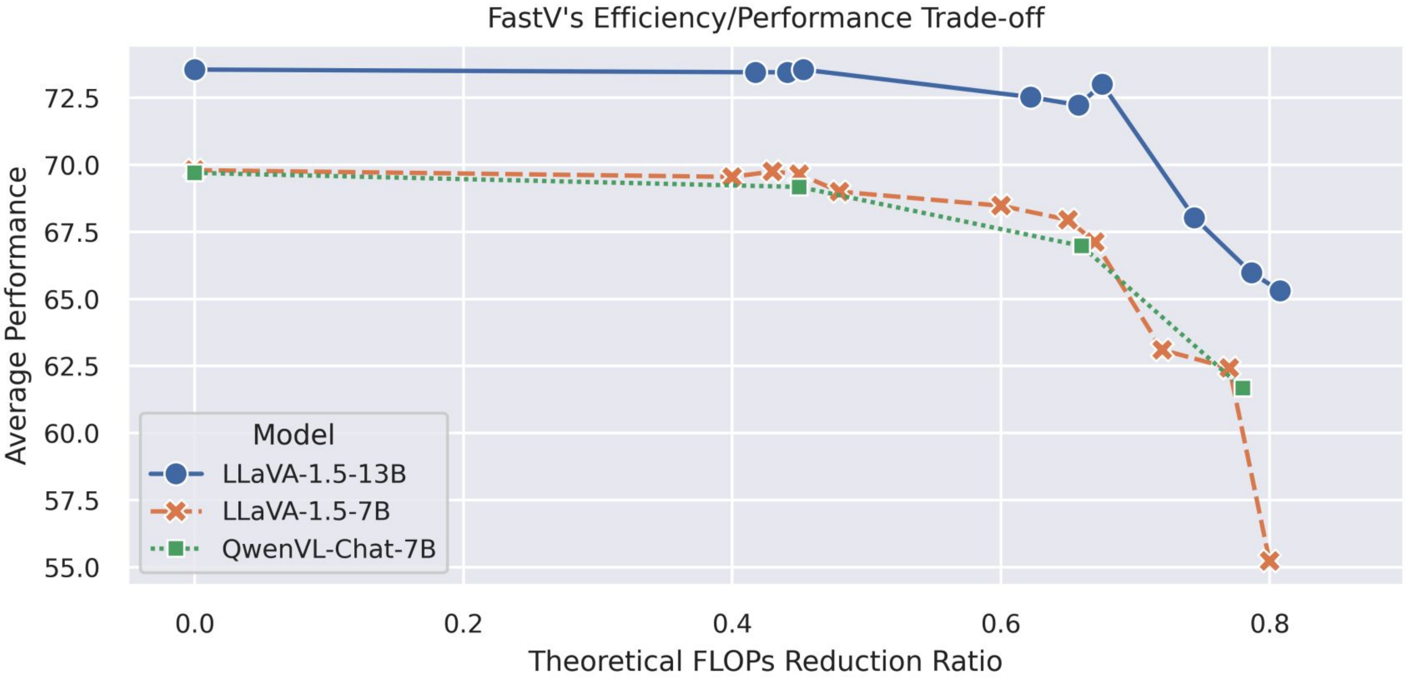
CLIP OPEN-CLIP

- ViT-L/14, LAINON-2B data, 384 A100 GPUs for 319 Hours

Vision Generative Model Stable Diffusion

- using 256 Nvidia A100 GPUs on AWS for a total of 150,000 GPU-hours

Token Pruning



FastV could achieve about 45% FLOPs reduction for different LVLMs without sacrificing the performance

An Image is Worth 1/2 Tokens After Layer 2 (Chen et al. 2024)

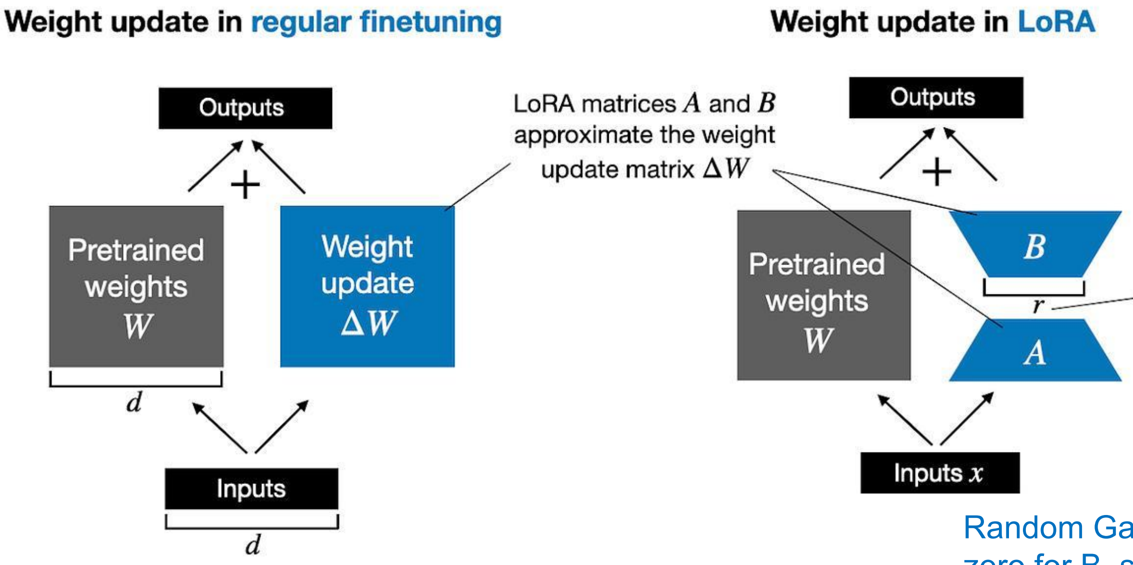
Parameter-Efficient Fine-Tuning

- Full fine-tuning require more computational cost and becomes infeasible to train on consumer hardware.
- Full fine-tuning leads to catastrophic forgetting in the low-data regimes.
- Storing and deploying fine-tuned models independently for each downstream task becomes very expensive

Parameter-Efficient Fine-tuning (PEFT) approaches are meant to address both problems!

- Prefix Tuning / Prompt Tuning
- Adapter Tuning
- LoRA Tuning

LoRA: Low-Rank Adaptation


Weight update in regular finetuning
Weight update in LoRA
Outputs
+
Pretrained weights W
Weight update ΔW
d
Inputs
d
LoRA matrices A and B approximate the weight update matrix ΔW
Outputs
+
Pretrained weights W
B
r
A
The inner dimension r is a hyperparameter
Inputs x

r can be very small, like 4/8

Random Gaussian initialization for A and zero for B, so ∆W = BA is zero at the beginning of training

$$h = W_0 x + \Delta W x = W_0 x + BAx$$

LoRA (Hu et al. 2021)

https://magazine.sebastianraschka.com/p/practical-tips-for-finetuning-llms

The last few class sections: Paper discussions

Discussion: Recent Advances in Transformers and Vision Representation Learners

Transformer architecture analysis and improvements

1. Rotary Position Embedding for Vision Transformer
2. Vision Transformers Need Registers

Extending SAM vision representation learner to videos

3. SAM 2: Segment Anything in Images and Videos

Recent example of large-scale vision representation learning for pathology

4. Virchow2: Scaling Self-Supervised Mixed Magnification Models in Pathology

The last few class sections: Paper discussions

Discussion: Recent Advances in Vision-Language Representation Learners

Techniques for interpretability

1. Interpreting CLIP's Image Representation via Text-Based Decomposition
2. Visual Explanations of Image-Text Representations via Multi-Modal Information Bottleneck Attribution

Extending CLIP to multiple modalities

3. ImageBind: One Embedding Space To Bind Them All

Leveraging CLIP for a unified CT segmentation model across datasets

4. CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection

The last few class sections: Paper discussions

Discussion: Recent Advances Vision Generative Models

Advancing diffusion model control capabilities

1. Adding Conditional Control to Text-to-Image Diffusion Models
2. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion

Extending to video

3. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets

An application connecting RNA sequencing data with image generation

4. Generation of synthetic whole-slide image tiles of tumours from RNA-sequencing data via cascaded diffusion models

The last few class sections: Paper discussions

Discussion: Recent Advances in Vision-Language Generative Models

Generating both discrete and continuous modalities

1. Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model

Agentic system for long-form video understanding

2. VideoAgent: Long-form Video Understanding with Large Language Model as Agent

Instruction tuning on large-scale pathology data

3. A Multimodal Generative AI Copilot for Human Pathology

The last few class sections: Paper discussions

Discussion: Recent Advances in Computing Efficiency

Token pruning

1. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models

IO-aware attention computation

2. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness

Parameter-efficient transfer learning from image to video

3. ST-Adapter: Parameter-Efficient Image-to-Video Transfer Learning

Decomposing medical foundation models into lighter weight expert models

4. LoRKD: Low-Rank Knowledge Decomposition for Medical Foundation Models

Where to go from here?

- We hope this class has equipped you with a lay of the land in cutting edge vision and vision-language models, and a foundation for future independent investigation
- Avenues for future exploration may include:
  - **Product-focused projects** for real-world use cases, building more hands-on experience with these models
  - **Research projects** investigating and pushing capabilities of what is possible, building on lines of investigation and insights from recent papers such as those we have covered
  - **Additional courses** on specific model classes (e.g. CS 236 Deep Generative Models), foundation models in healthcare more broadly (less depth on vision, e.g. BIODS 271 Foundation Models for Healthcare), or even broader courses on foundation models (e.g. CS 324 Advances in Foundation Models)

AMA with the course staff