MANUTECH SLEIGHT
Université de Lyon

Foundation models in medical imaging

Olivier Bernard (Professor at INSA, IUF member)

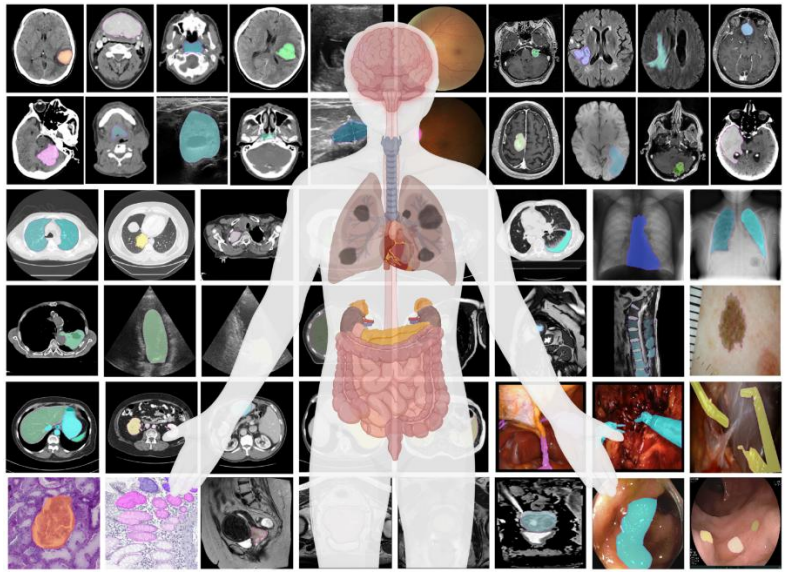
Motivations - Biomedical imaging

High/multi dimensionality

- 2D vs 3D
- 1024x1024 px
- 256x256x256 px

Multimodality

- MR / CT / US / Optics
- EEG / ECG / Genetics
- Report



Annotations

- Interobserver variability
- Quality

Scarcity

- Annotations
- Image accessibility

Domain shifts

- Multicentre
- Multivendor
- Acquisition settings

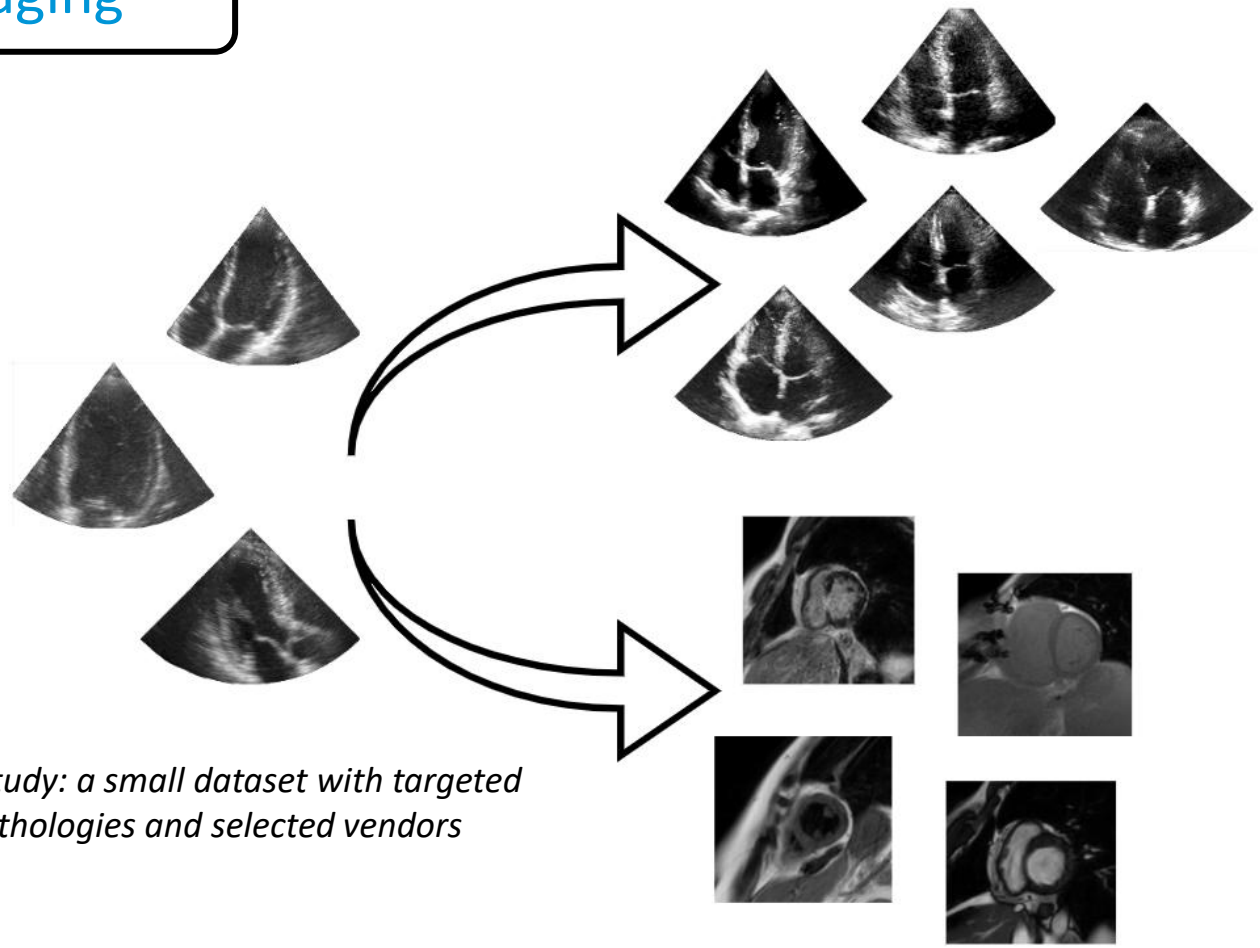
Figure from [Ma, Nature Communications, 2024]

Motivations - Biomedical imaging

Needs for generalization

- On large scale dataset
- Through modalities
- Through pathologies

Large-scaled dataset with varying degrees of pathology and multi-vendor equipment

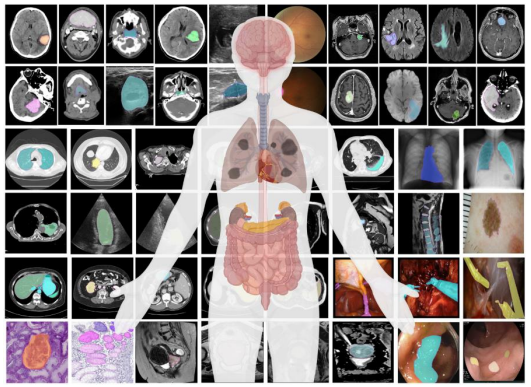


Pilot study: a small dataset with targeted pathologies and selected vendors

Large-scaled dataset with different imaging modalities

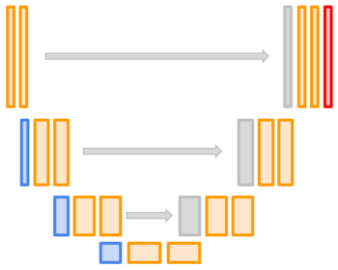
Foundation models - Definition

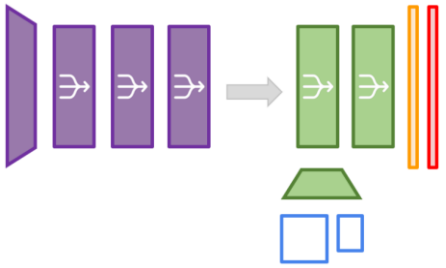
- A large-scale pretrained model
- Adapting to multiple tasks with little or no fine-tuning

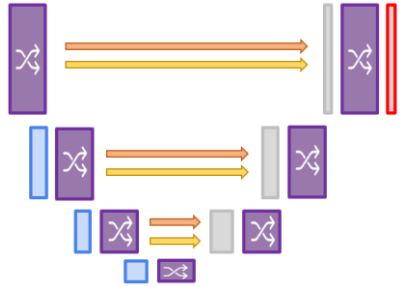


Large-scale dataset : millions of images

CNN *Transformer* *In context*



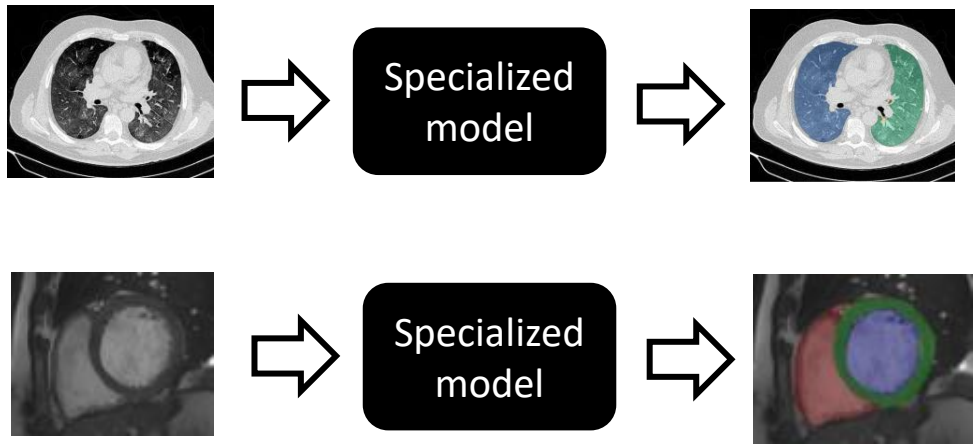




Foundation models

Before

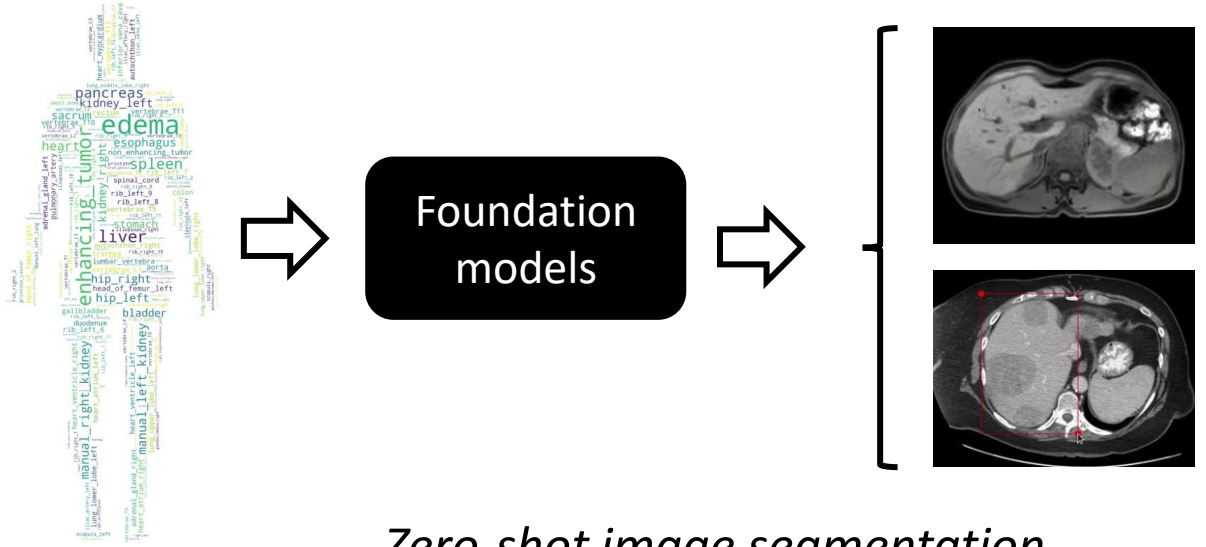
- Specialized model
- Task-specific training
- Annotated dataset



Training / inference scheme

After

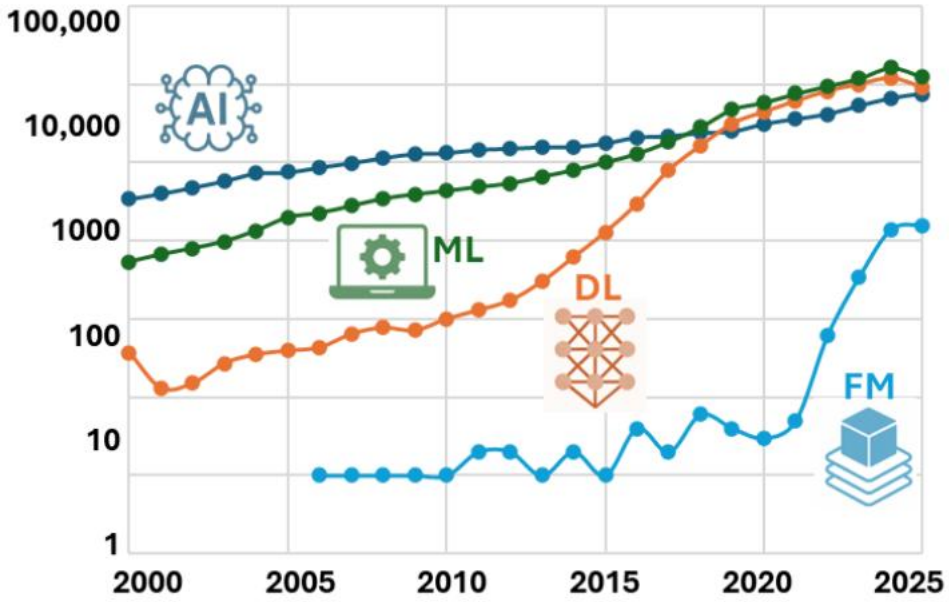
- Generic model
- Zero-shot inference
- Few-shot learning



Zero-shot image segmentation

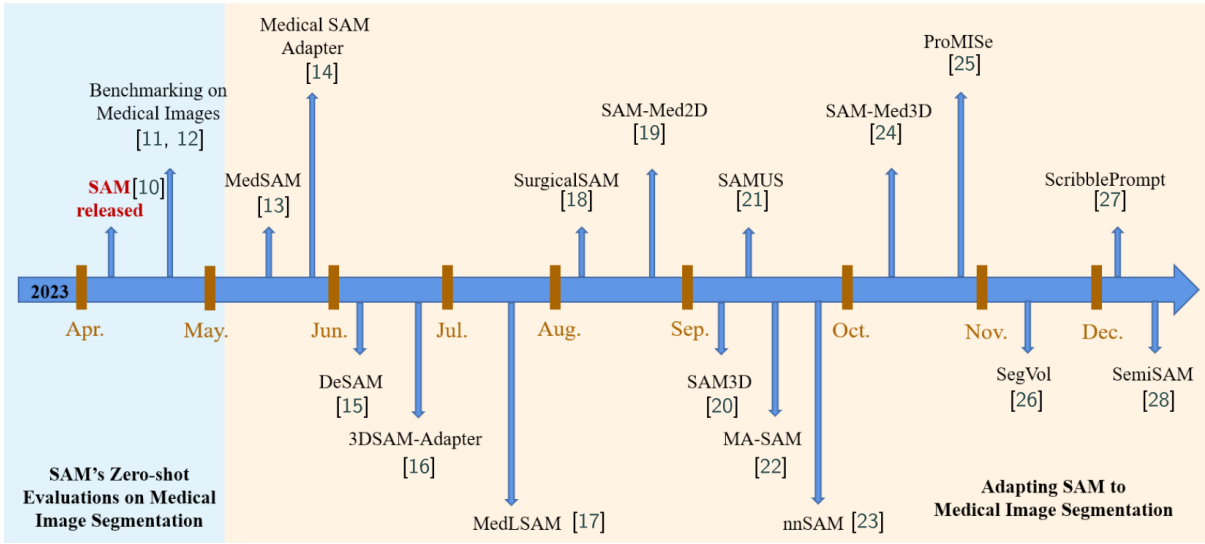
Motivations

Publication trends



Source: Scopus, September 2025

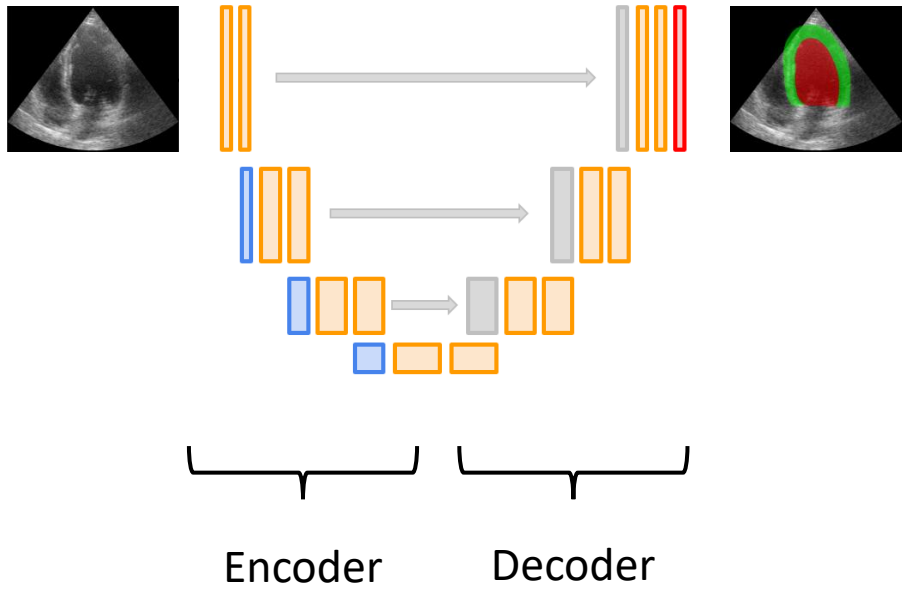
8 months of publications between 2023 and2024



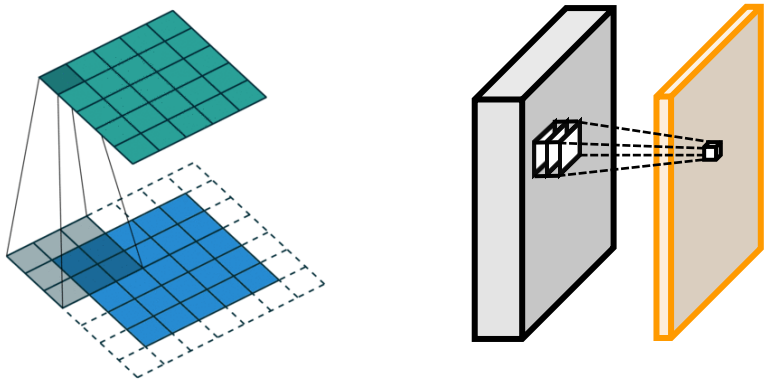
Source: [Zhang et al., CIBM, 2024]

Before foundation models - CNN-based approaches

- Create relevant information called feature map (convolution + non linear function)
- Encoder / decoder architecture



Convolution operation

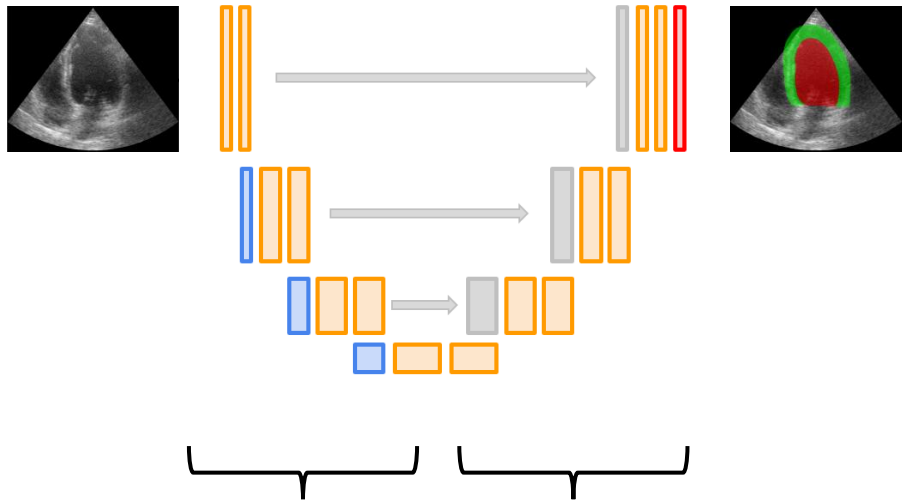


Filter of size 3×3

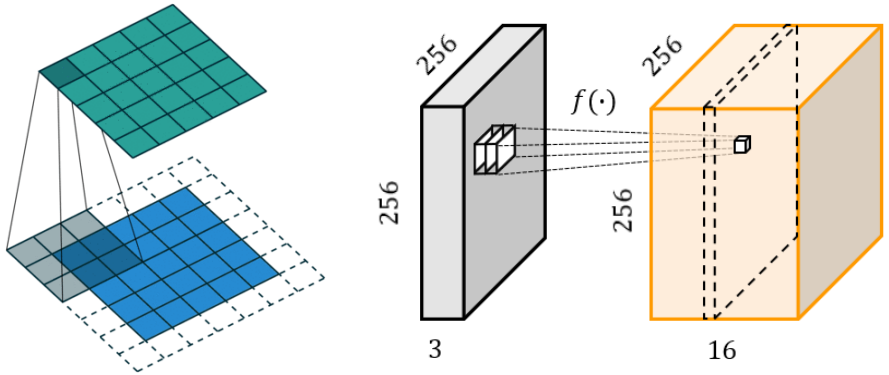
$$\#\ param = 3 \times 3 \times 3 + 1 = 28$$

Before foundation models - CNN-based approaches

- Create relevant information called feature map (convolution + non linear function)
- Encoder / decoder architecture



Feature maps

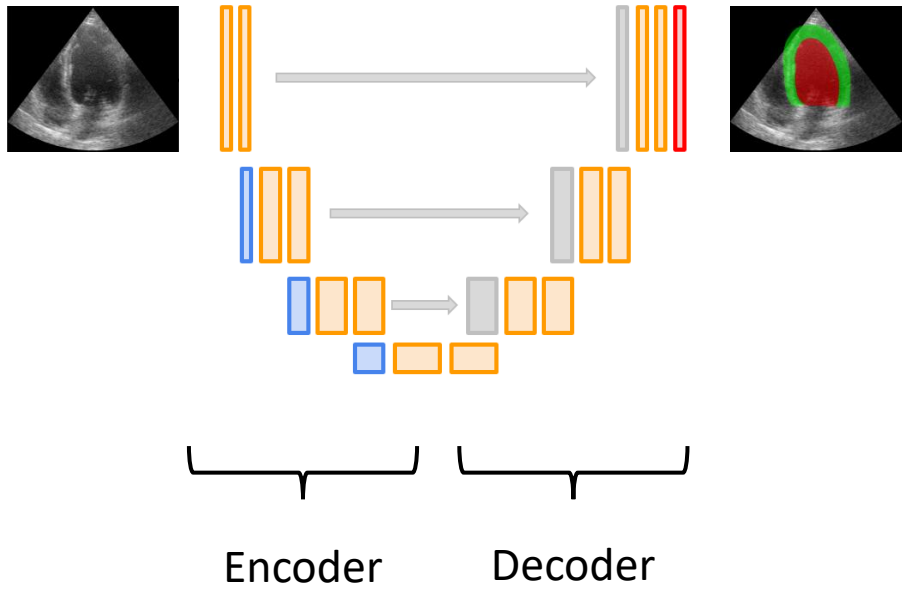


Filter of size 3×3

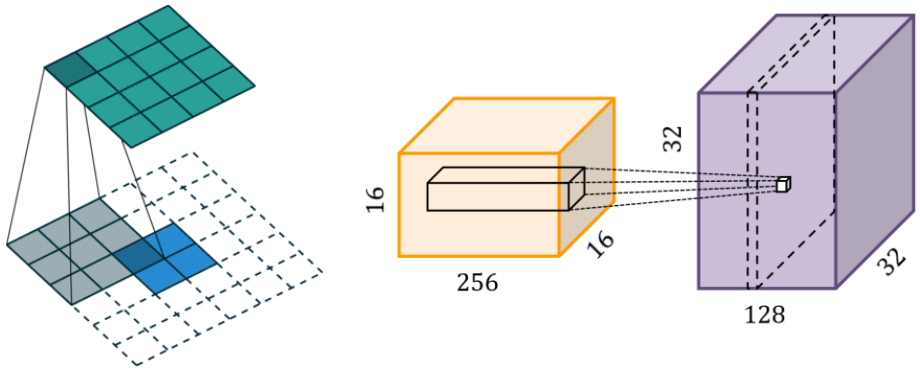
$$\#\, param = 16 \times (3 \times 3 \times 3 + 1) = 448$$

Before foundation models - CNN-based approaches

- Create relevant information called feature map (convolution + non linear function)
- Encoder / decoder architecture



Deconvolution operation

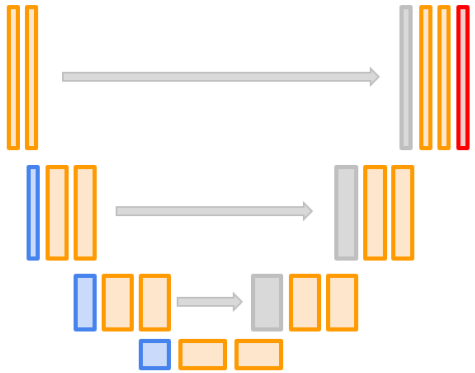


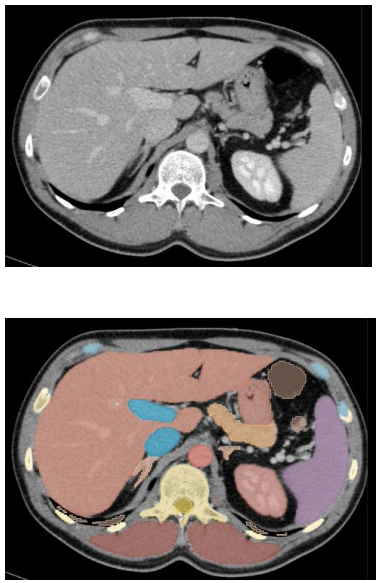
Filter of size 3×3

$$\#\ param = 128 \times (3 \times 3 \times 256 + 1) = 295{,}040$$

Before foundation models - TotalSegmentor

- 40 M of parameters (depending on the GPU)
- 3D images from publicly available medical datasets
- Based on nnU-Net algorithm

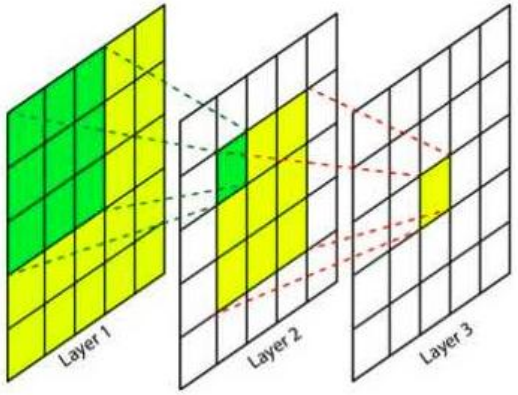




- Multimodal model: CT, MRI
- CT model trained on:
  - 1204 CT
  - 104 anatomical structures
  - 27 organs, 59 bones, 10 muscles, 8 vessels
- CT/MRI model trained on:
  - 527 CT and 616 MRI
  - 80 anatomical structures

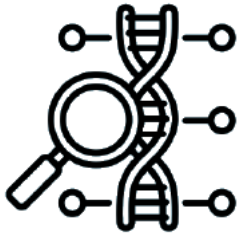
Limitations of CNN-based approaches

- Receptive field
  - Different parts of an image communicate in deep layers



- How to integrate images with others modalities?


Medical images


Genetics

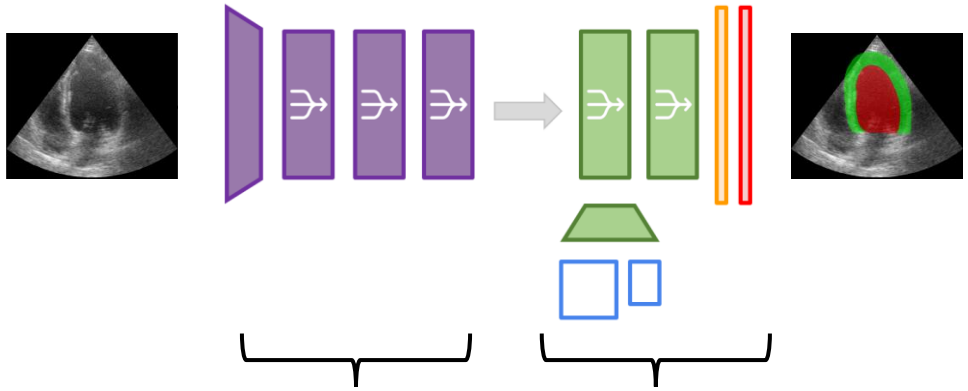

Reports


Lab tests

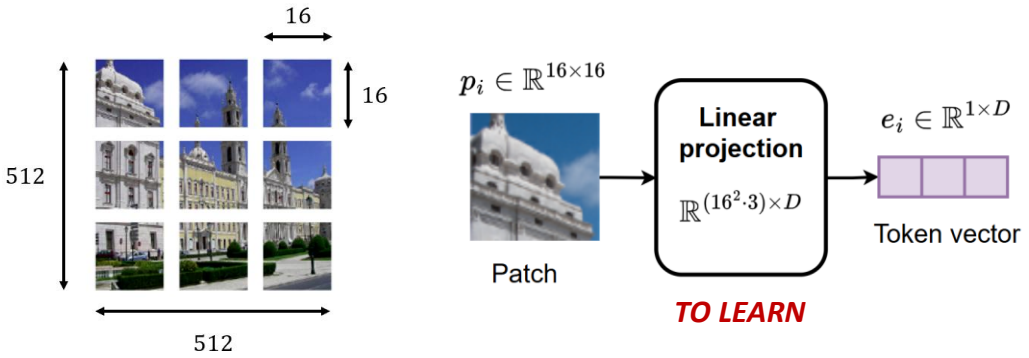

EHR

Foundation models - transformer-based approaches

- Create relevant information called tokens (tokenization + attention + non linear function)
- Encoder / decoder architecture



Tokenization procedure

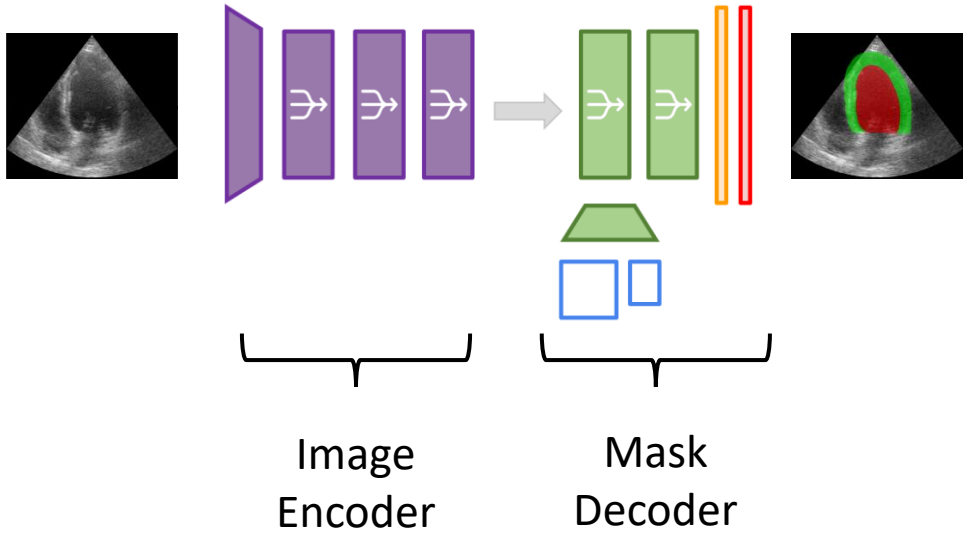


Token size $D = 768$

$\#\ param = 16 \times 16 \times 3 \times D = 589{,}824$

Foundation models - transformer-based approaches

- Create relevant information called tokens (tokenization + attention + non linear function)
- Encoder / decoder architecture



Tokenization procedure

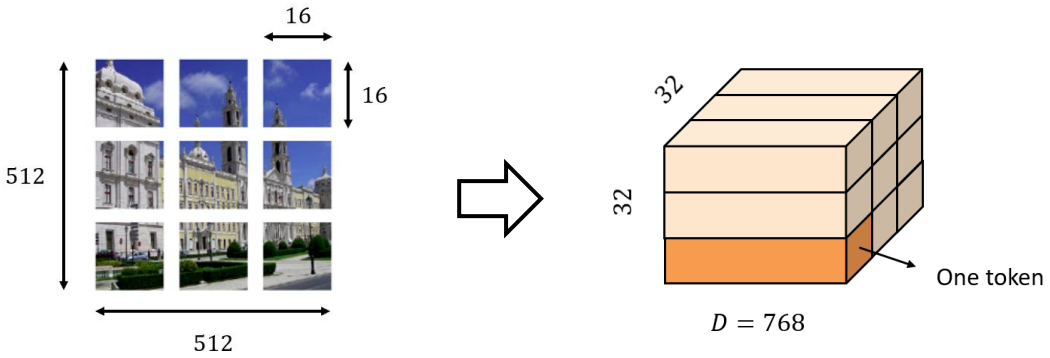
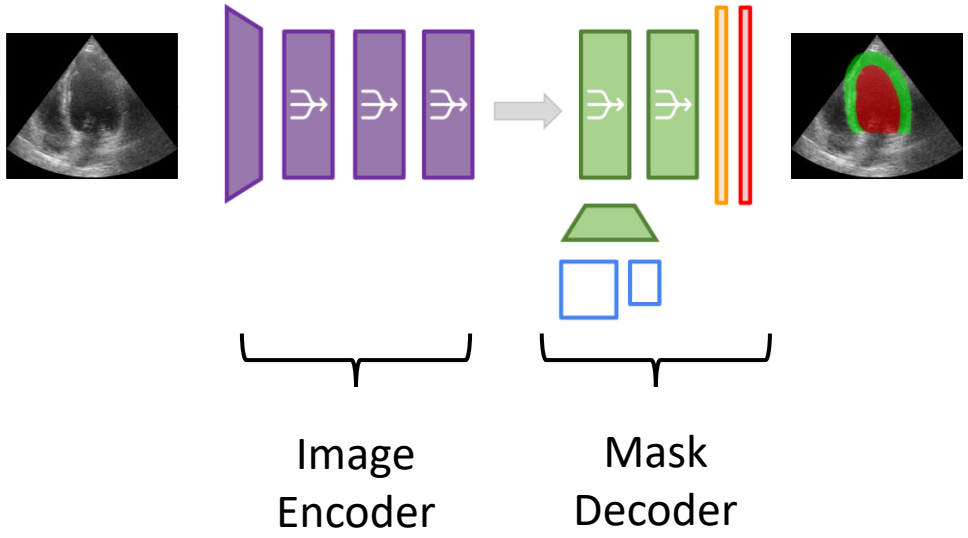


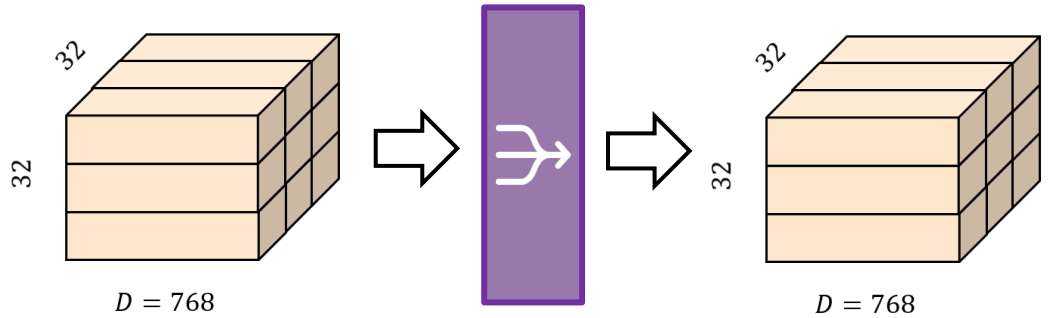
Image encoding
From $512 \times 512 \times 3$ px to $32 \times 32 \times 768$ px

Foundation models - transformer-based approaches

- Create relevant information called tokens (tokenization + attention + non linear function)
- Encoder / decoder architecture



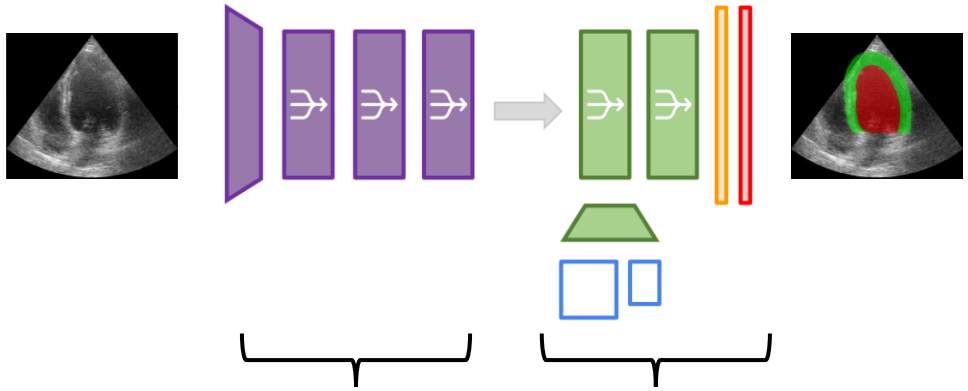
Transformer blocs / layers



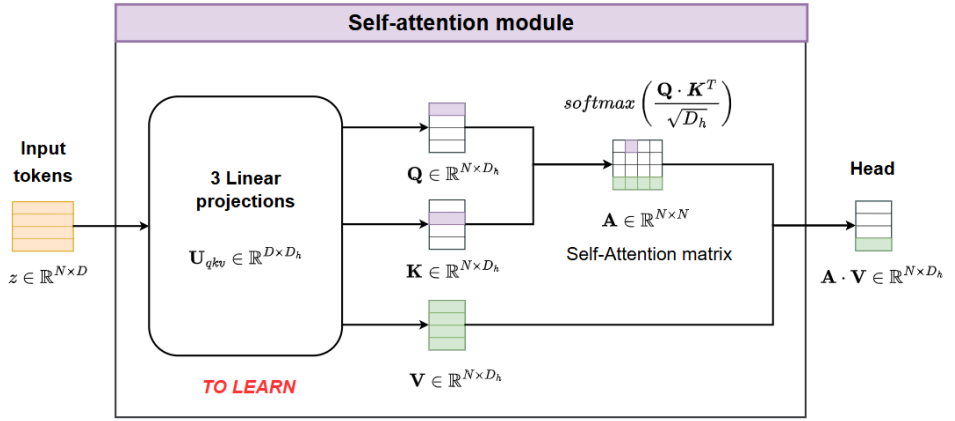
Multi-head attention module + non linearity

Foundation models - transformer-based approaches

- Create relevant information called tokens (tokenization + attention + non linear function)
- Encoder / decoder architecture



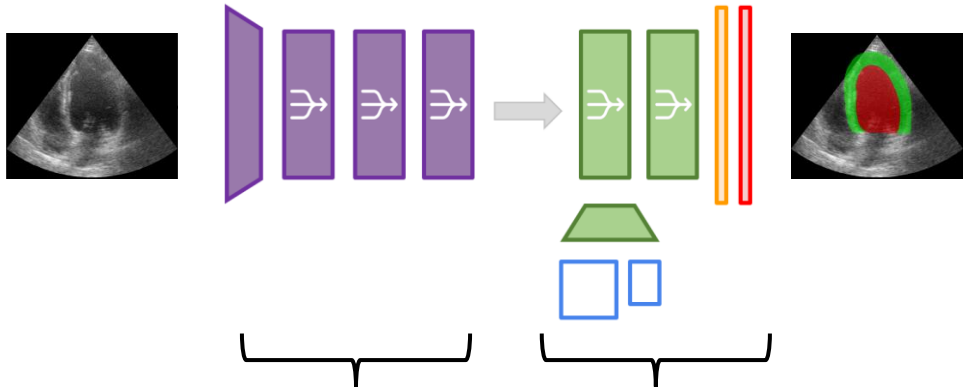
Self-attention module



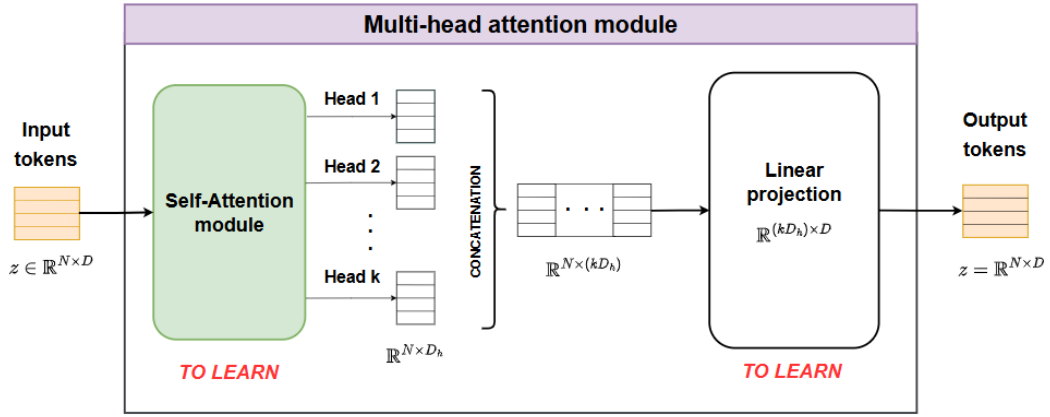
$$D = 768, D_h = 64$$
$$\#\, param = 3 \times 768 \times 64 = 147{,}456$$

Foundation models - transformer-based approaches

- Create relevant information called tokens (tokenization + attention + non linear function)
- Encoder / decoder architecture



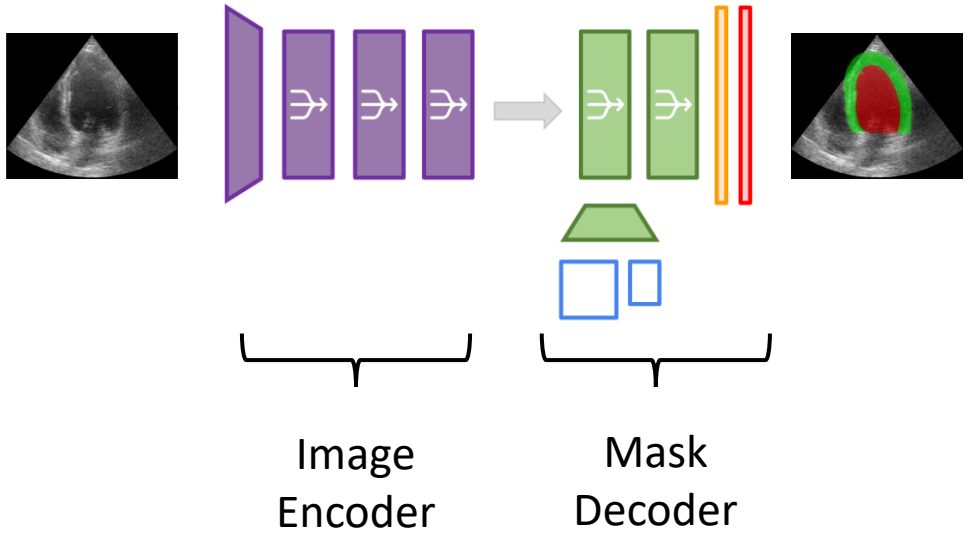
Multi-head attention module



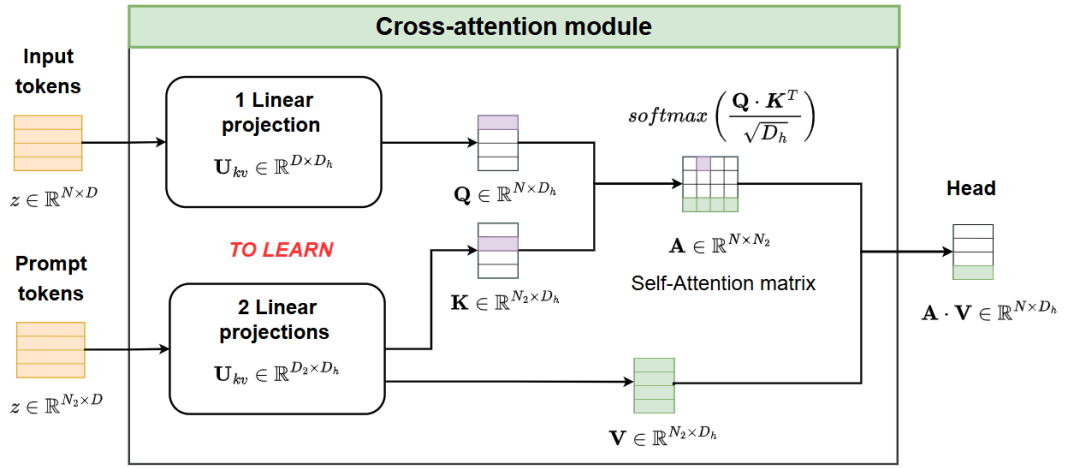
$$D = 768, D_h = 64, k = 12$$
$$\#\, param = 12 \times 3 \times 768 \times 64 + 768 \times 768 = 2{,}359{,}296$$

Foundation models - transformer-based approaches

- Create relevant information called tokens (tokenization + attention + non linear function)
- Encoder / decoder architecture

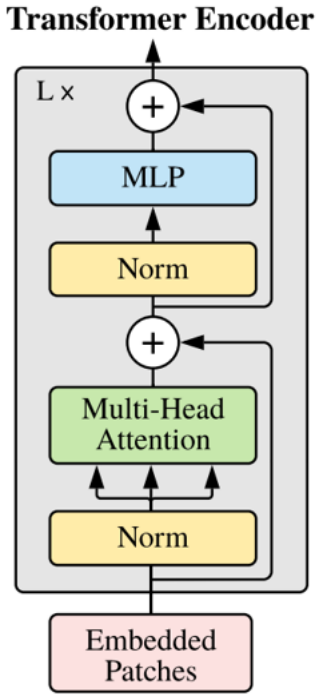


Cross-attention module



Foundation models - transformer-based approaches

- ViT: Visual Transformer
- Reference model to encode images

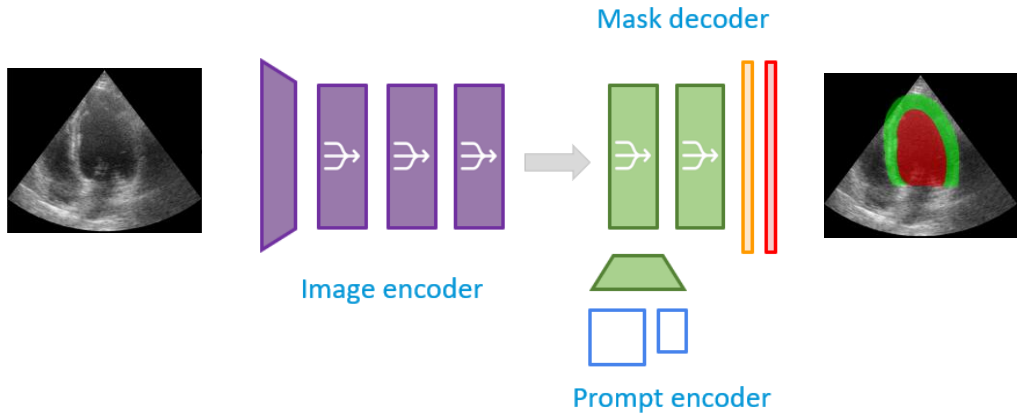


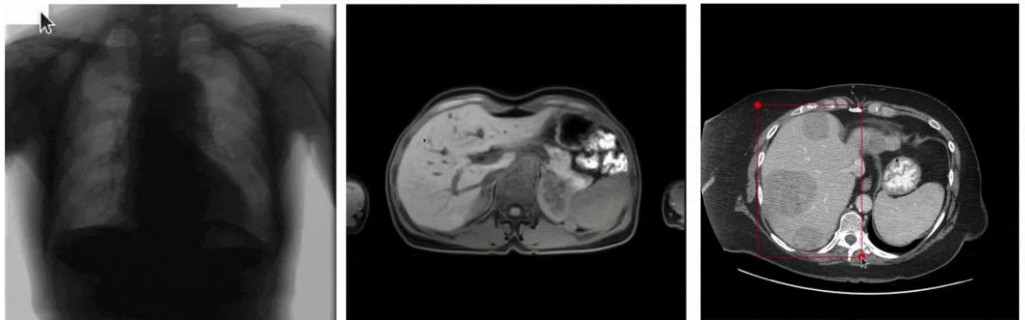
Models	Nb layers	Token size D	Nb heads	Nb parameters
ViT-Base	12	768	12	86 M
ViT-Large	24	1024	16	307 M
ViT-Huge	32	1280	16	632 M

Foundation models - MedSAM

- 91 M of parameters
- Image encoder - Vit-B
- 2D images
- From publicly available medical datasets

- 1.5 M 2D image-mask pairs
- 10 imaging modalities
- 30 cancer types

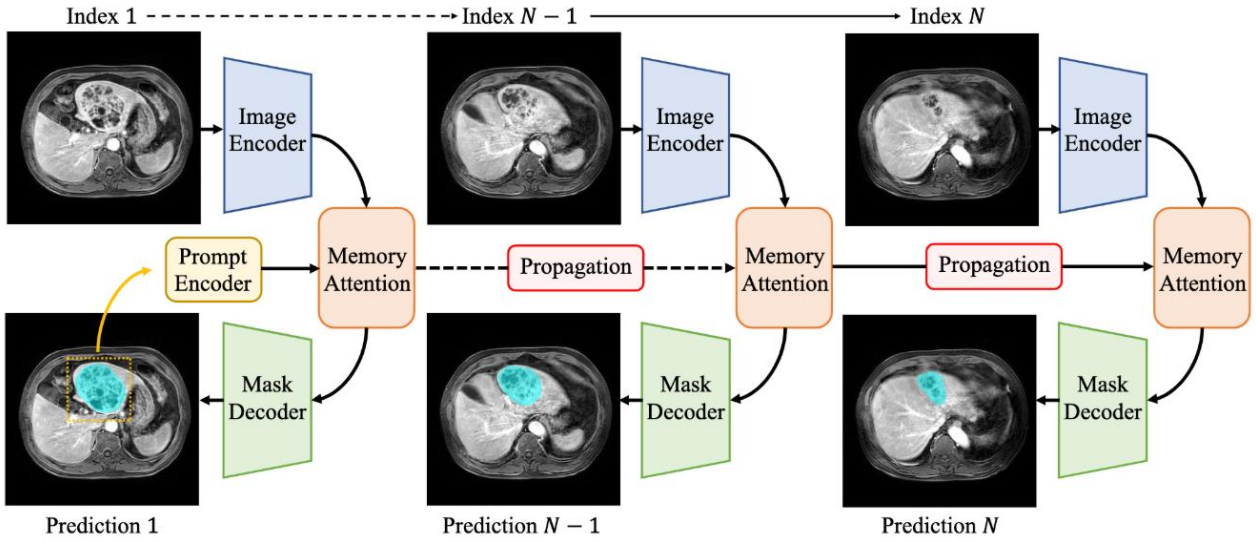


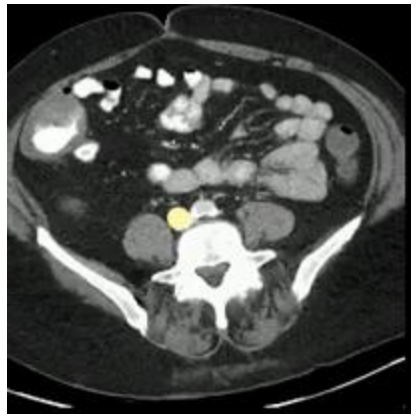


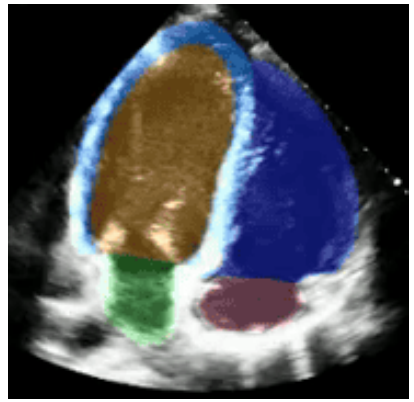
Foundation models - MedSAM2

- 40 M of parameters
- Image encoder – Hiera (Hierarchical ViT)
- 3D images and 2D+t sequences
- From publicly available medical datasets

- 455 K 3D image-mask pairs
- 76 K annotated video frames
- Multimodal: CT, PET, MRI, US, endoscopy





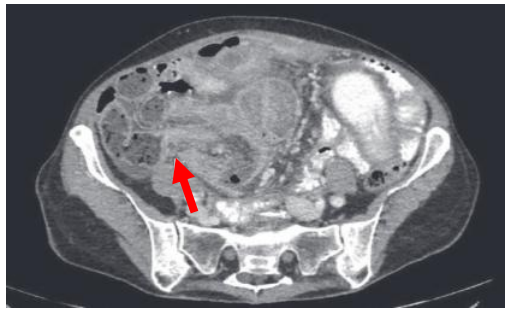


Multimodal foundation models - BiomedClip

- 200 M of parameters
- Biomedical vision-language processing
- Align medical images with text

- 4.4 M scientific reports from PubMed
- 15 M image-text pairs from fig./caption
- Large diversity of image types

- Targeted downstream tasks:
  - Image classification
  - Medical visual question answering



⇔

CT images suggestive of intussusception Axial CT image of an apparent small bowel transition point in the right lower quadrant (indicated with red arrow

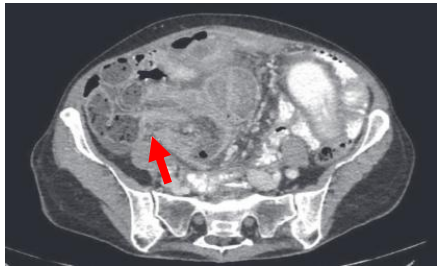
Multimodal foundation models - BiomedClip

Image encoder

- ✓ ViT-B/16
- ✓ Patch size: 16x16
- ✓ 86 M parameters

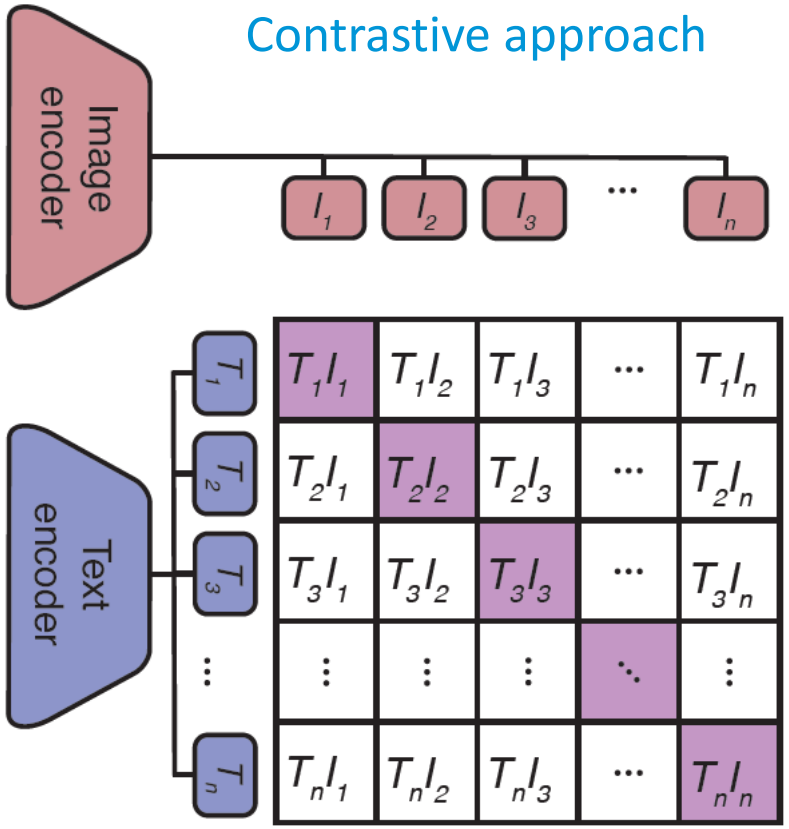
Text encoder

- ✓ PubMedBERT
- ✓ Max. length 256 tokens
- ✓ 110 M parameters

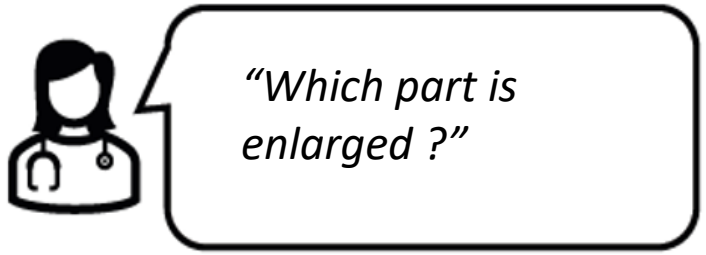


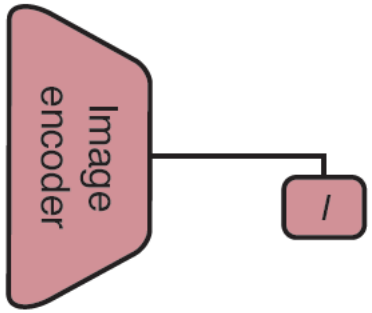
CT images suggestive of intussusception Axial CT image of an apparent small bowel transition point in the right lower quadrant (indicated with red arrow

CLIP-based method
Contrastive approach

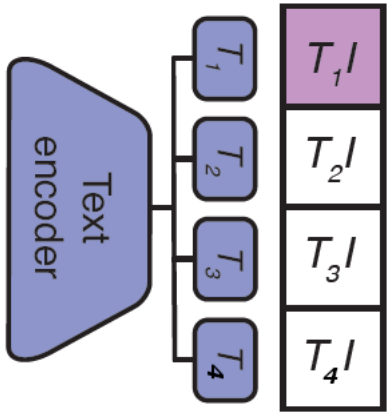


BiomedCLIP - Visual question answering





- T1: The myocardium is enlarged
- T2: This left ventricle is enlarged
- T3: The left atrium is enlarged
- T4: No structure is enlarged

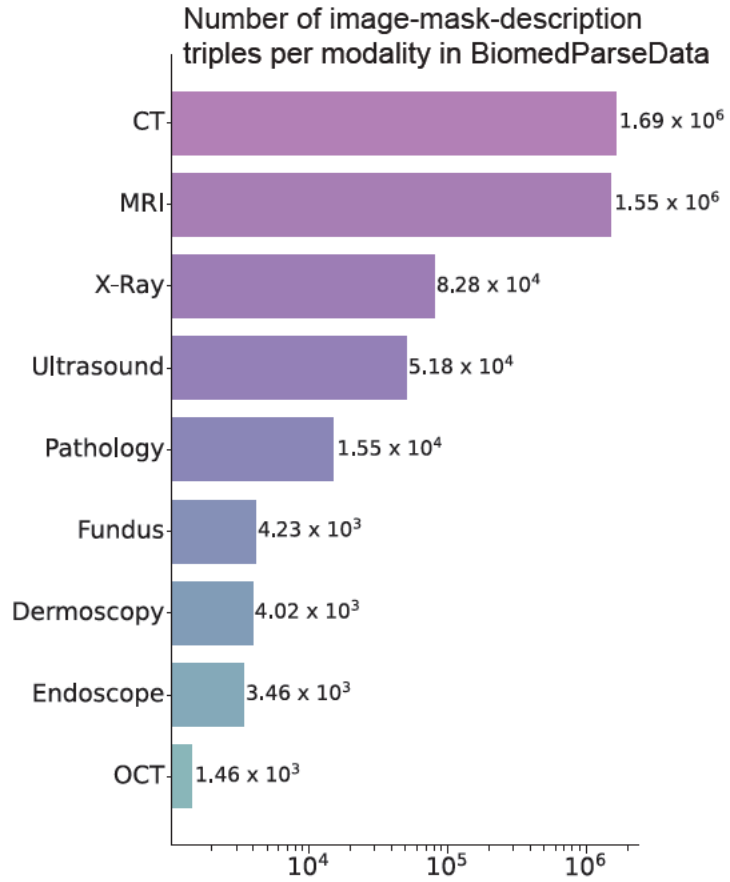


- ☑ The myocardium is enlarged (90 %)
- ☒ The left ventricle is enlarged (2 %)
- ☒ The left atrium is enlarged (3 %)
- ☒ No structure is enlarged (5 %)

Multimodal foundation models - BiomedParse

- 600 M of parameters
- Targeted downstream tasks:
  - segmentation / detection / recognition

- 1.1 M images across 9 imaging modalities
- All with manual /semi-manually segmented
- 3.4 M image-masks-label triplets
- 6.8 M image-masks-description triplets
- Across 9 imaging modalities



Multimodal foundation models - BiomedParse

Image encoder: EVA02-CLIP-L-14

- ✓ ViT-L/14 (24 layers, D=1024)
- ✓ 307 M parameters

Text encoder: CLIP text encoder

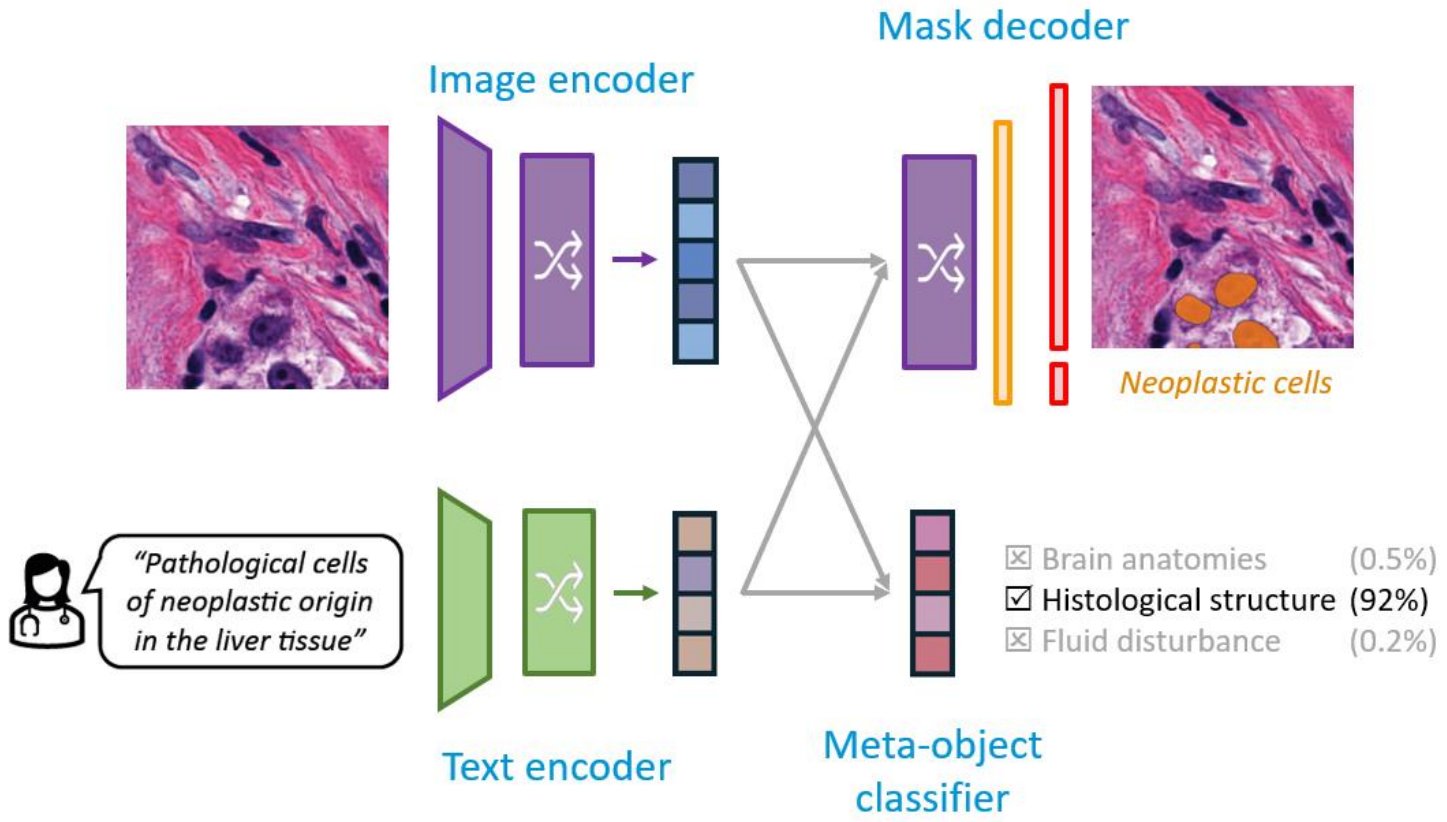
- ✓ Initialized with PubMedBert
- ✓ 124 M parameters

Mask decoder: SEEM

- ✓ 50 M parameters

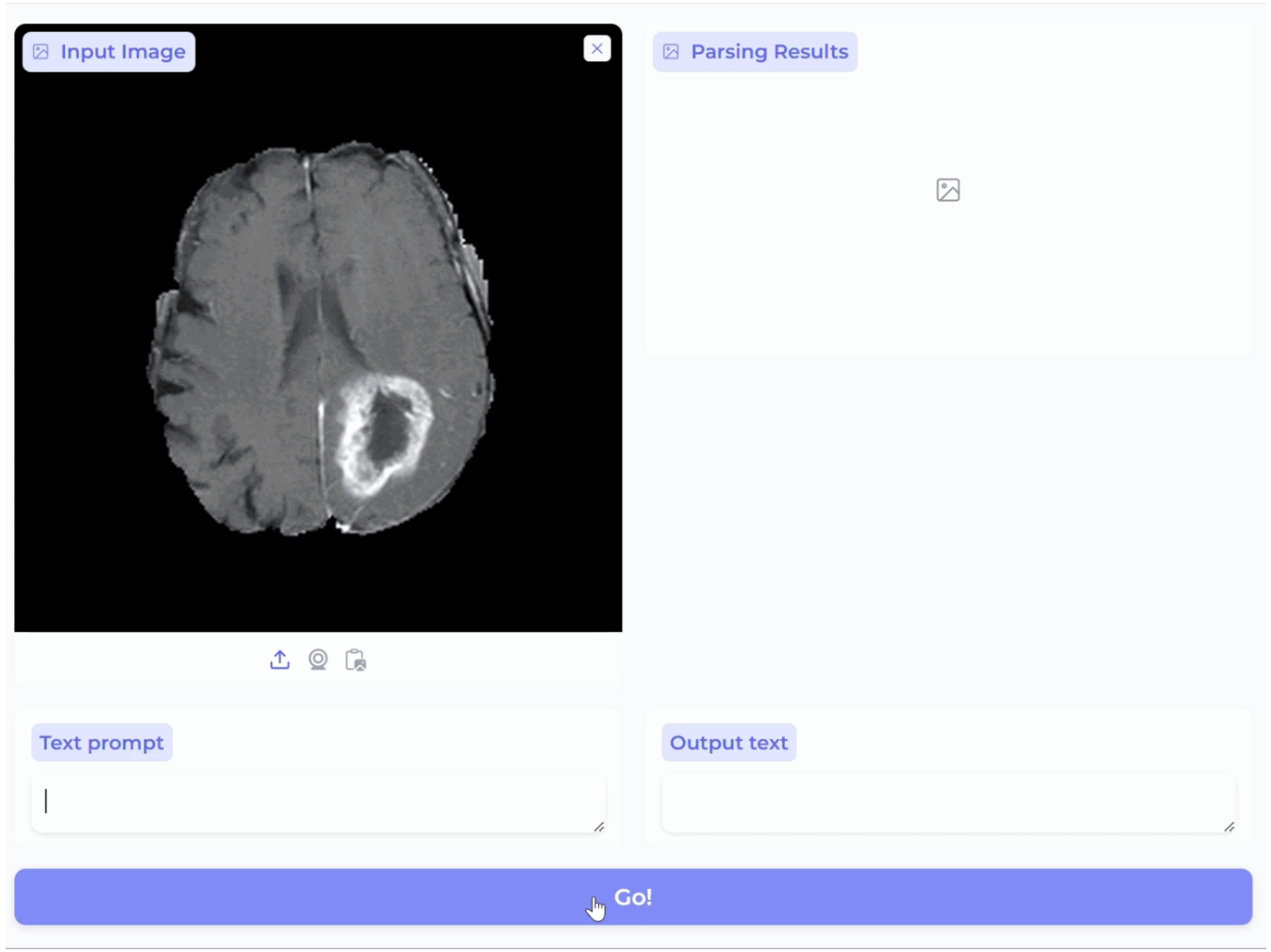
Meta-object classifier: simple MLP

- ✓ 15 meta-objects
- ✓ 300 K parameters



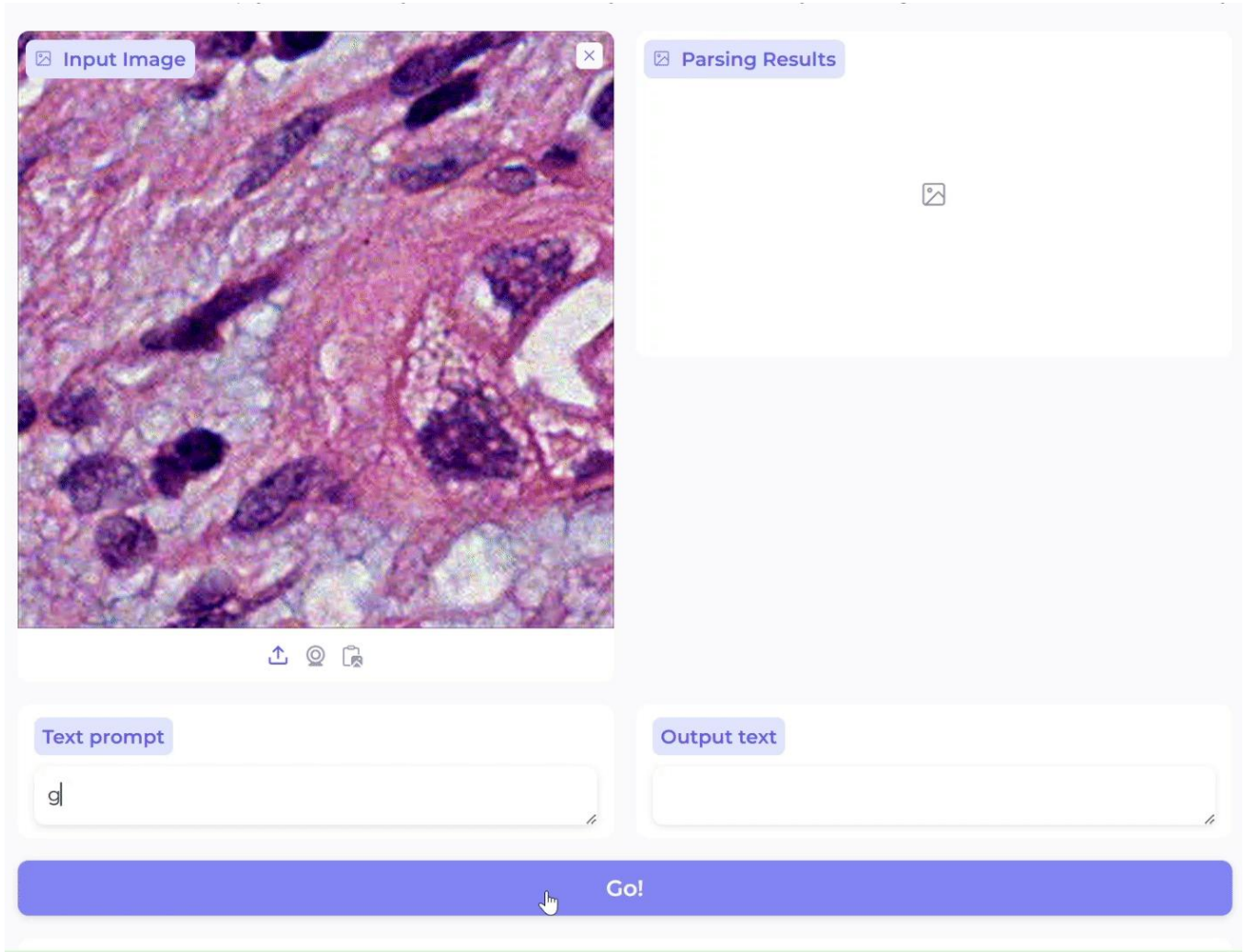
BiomedParse - Segmentation task

From https://microsoft.github.io/BiomedParse/



BiomedParse - Recognition task

From https://microsoft.github.io/BiomedParse/



Multimodal foundation models - EchoPrime

- 200 M of parameters
- Biomedical vision-language processing
- Four main steps:
  - Videos / reports alignment
  - View classification
  - Anatomy attention
  - Cross-modal retrieval
- Targeted task:
  - Automatic report generation

- 108 K patients
- 275 K echocardiographic exams
- 1 exam: >40 videos + 1 clinical report
- 12 M videos
- Structuration of the reports per anatomical section / interpretation task

Left ventricle (LV):
LV ejection fraction is 56%
The LV size is normal

Aorta valve:
A bioprosthetic valve is present

Multimodal foundation models - EchoPrime

Videos / report alignment

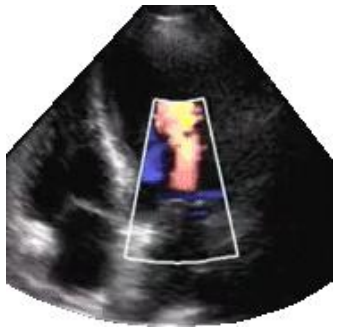
Videos encoder

- ✓ mViT-S (16 layers, D=768)
- ✓ 35 M parameters

Text encoder

- ✓ BioMedBERT
- ✓ 110 M parameters

Each videos is aligned with its report independently

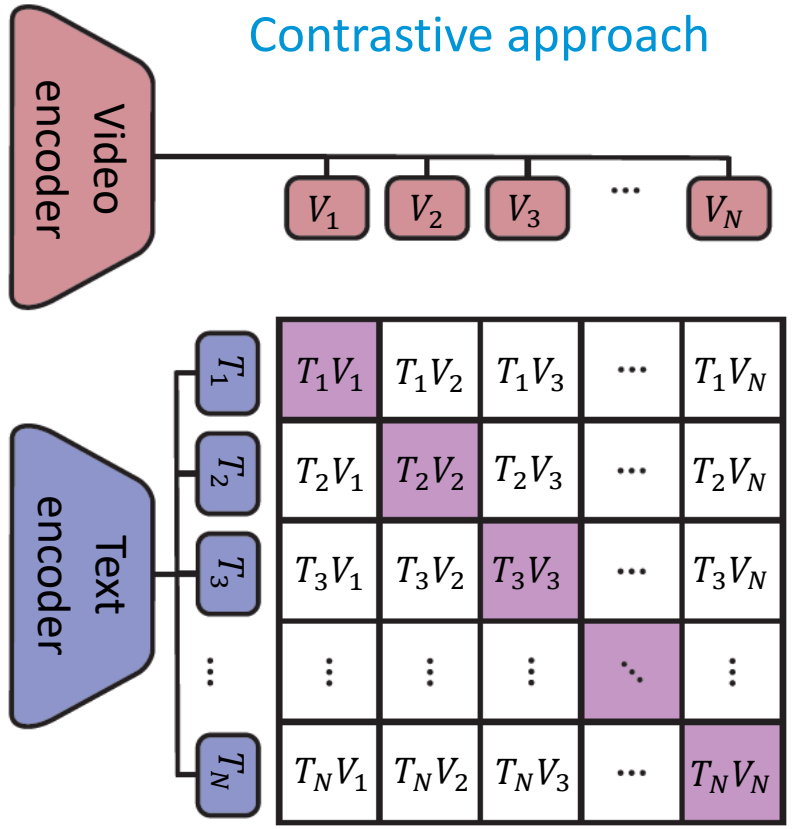


Left ventricle:
LV ejection fraction 60%
Mid diastolic dysfunction

Aortic valve:
A bioprosthetic valve is present

Pericardium:
Small pericardium effusion

CLIP-based method
Contrastive approach



Multimodal foundation models - EchoPrime

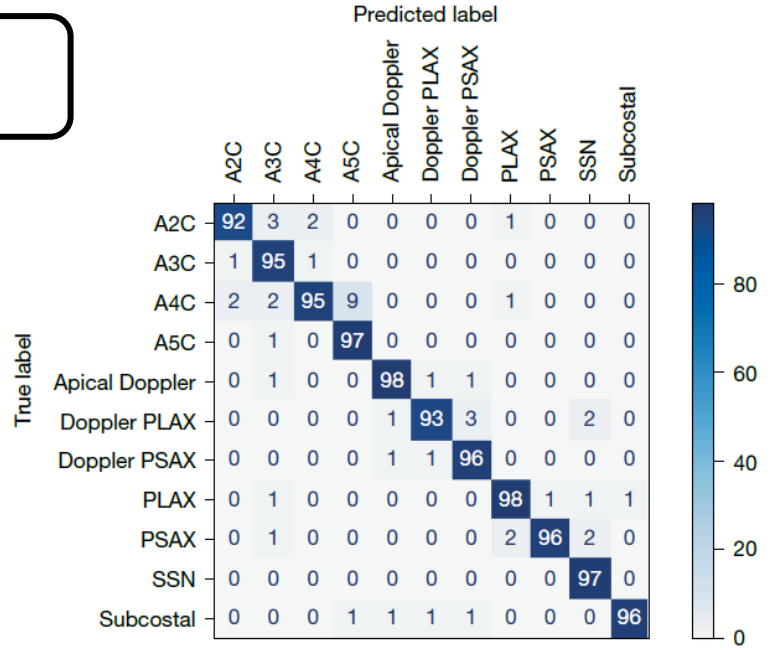
View classification

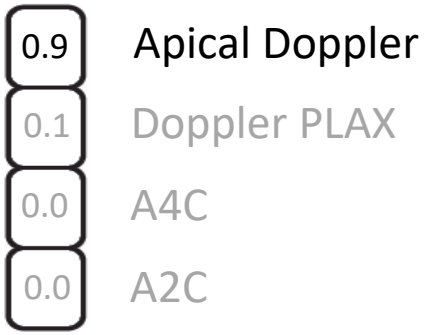
Model

- ✓ ConvNextBase
- ✓ 28 M parameters

Pipeline

- ✓ Video as input
- ✓ One-hot vector as output
- ✓ Classify 58 views

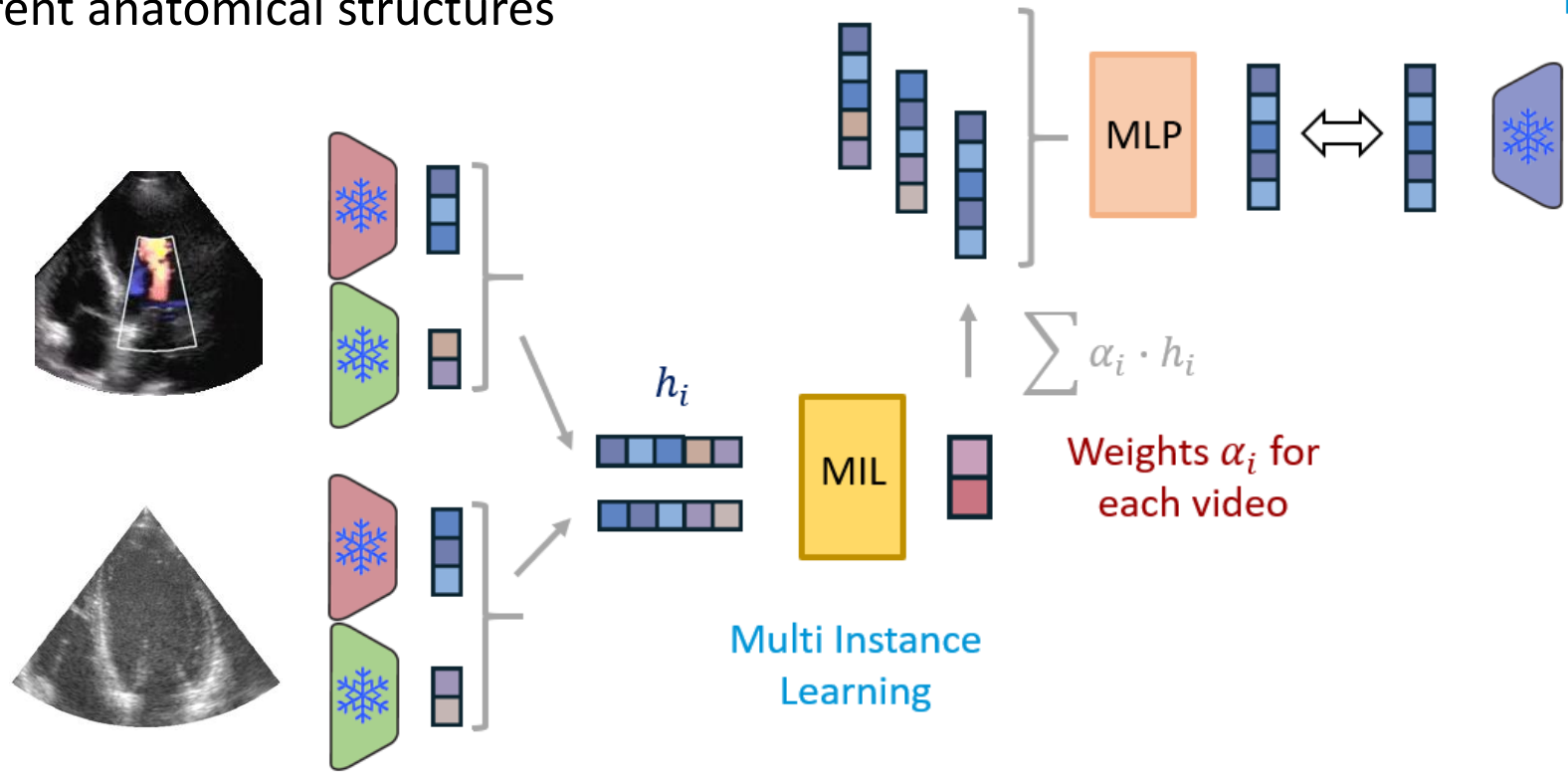




Multimodal foundation models - EchoPrime

Anatomical attention module

- Prioritization of echocardiographic views for different anatomical structures

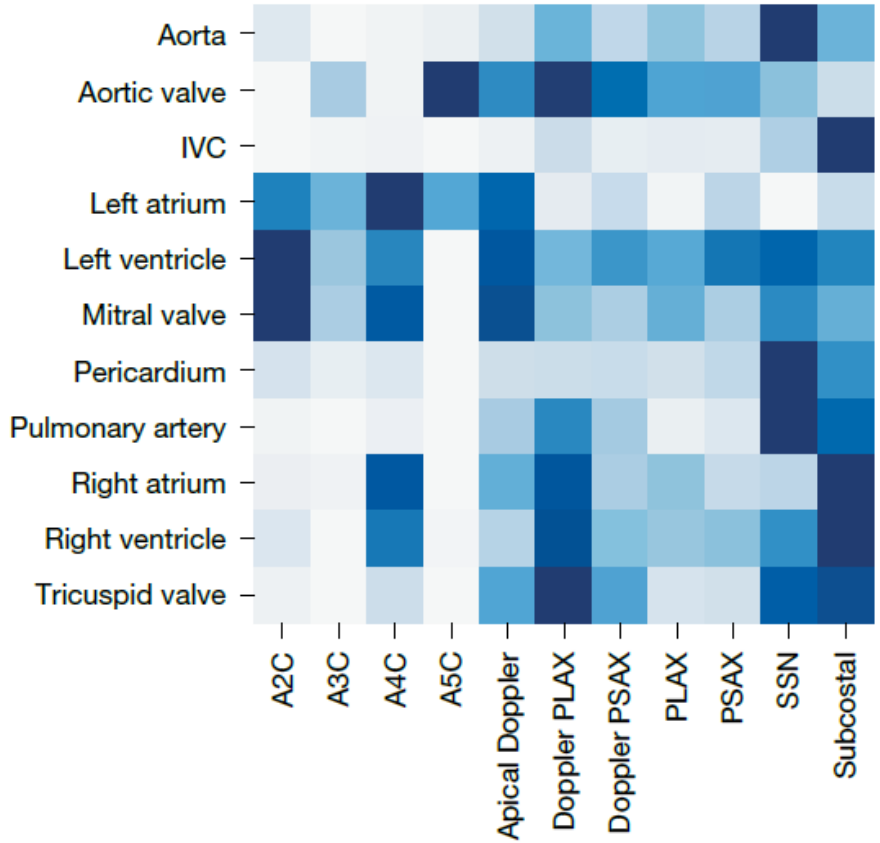


Multimodal foundation models - EchoPrime

Anatomical attention module

- Training procedure repeated for each interpretation task within each anatomical section

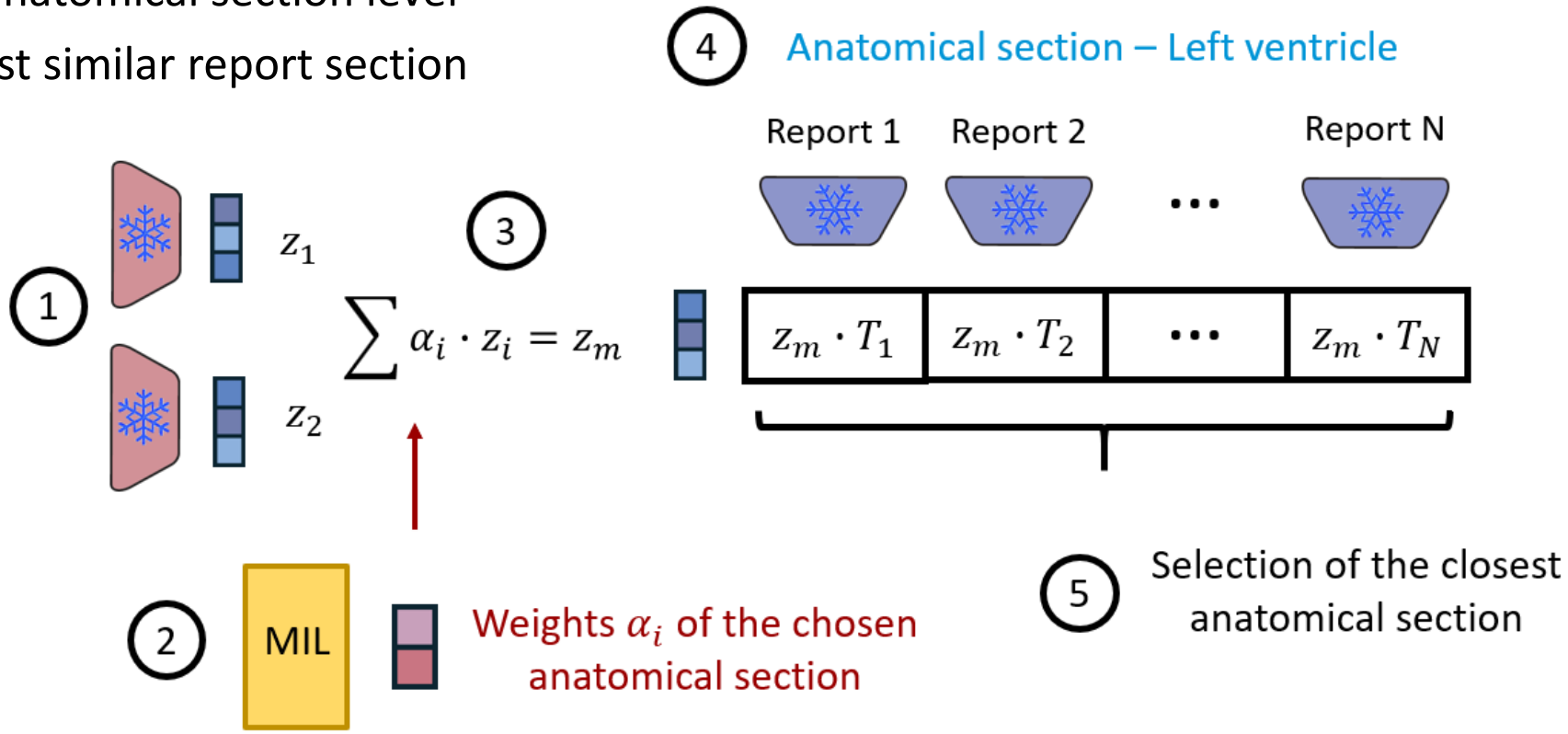
EchoPrime learned attention α_i



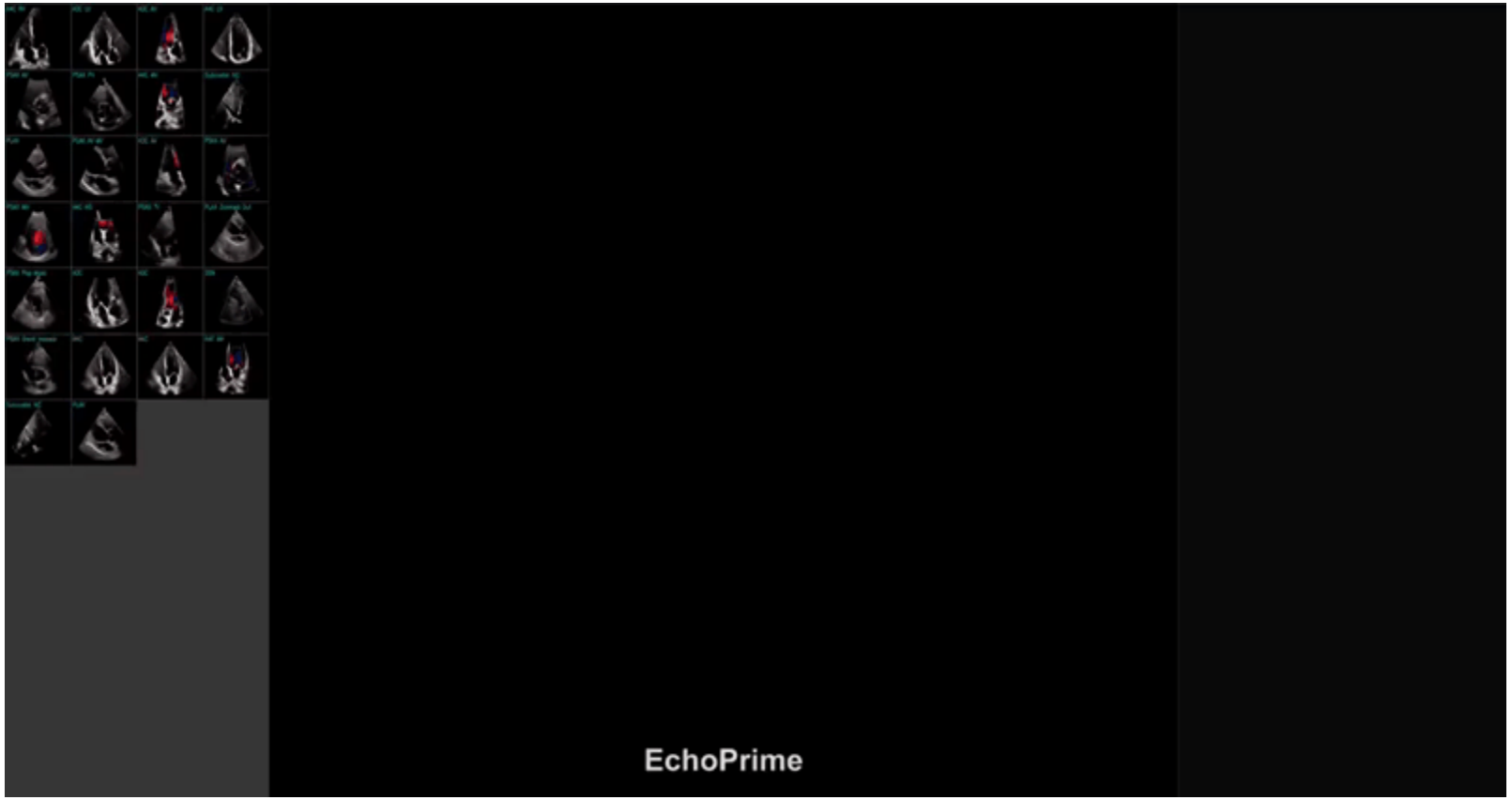
Multimodal foundation models - EchoPrime

Cross modal retrieval

- Operate at the anatomical section level
- Retrieve the most similar report section



Multimodal foundation models - EchoPrime

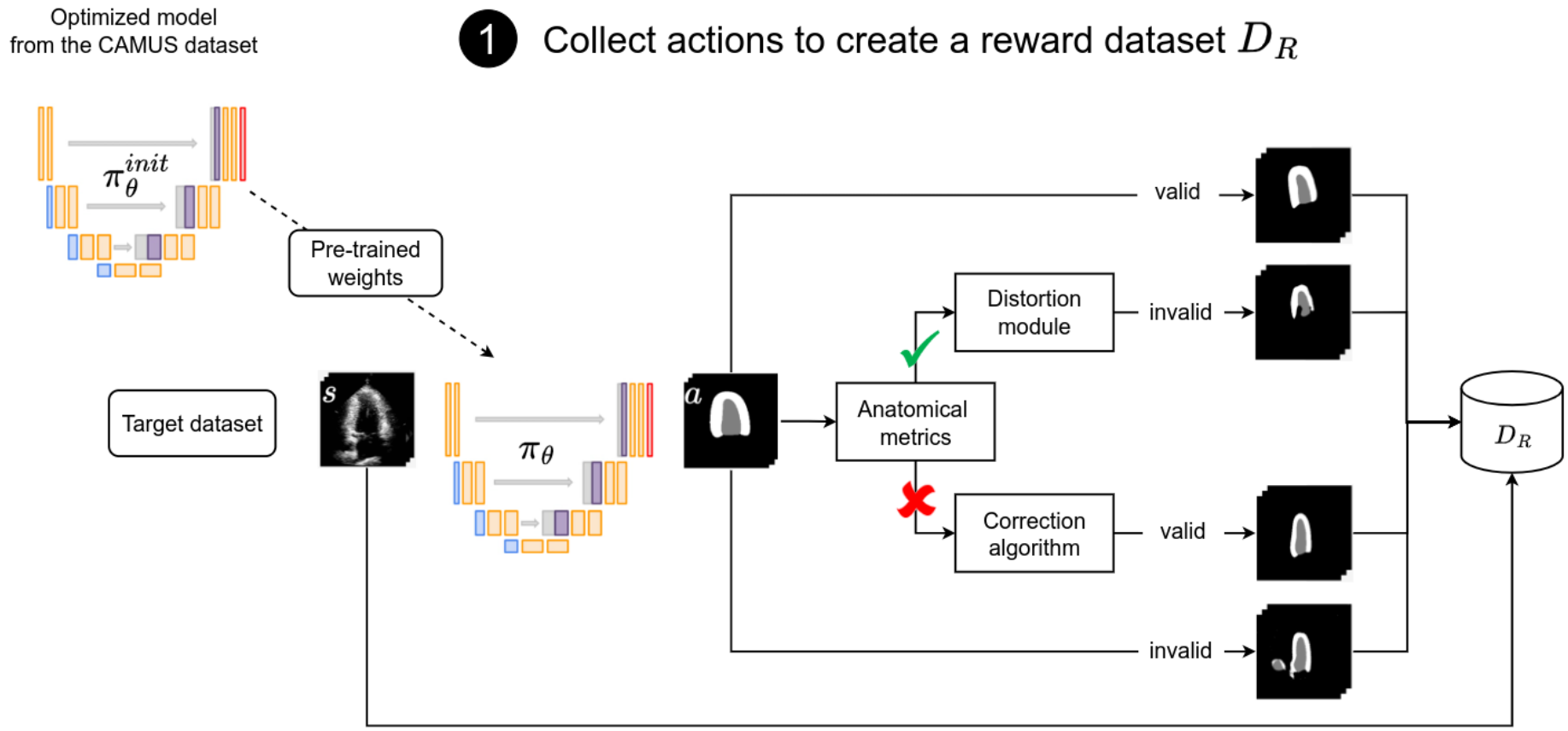


Conclusions

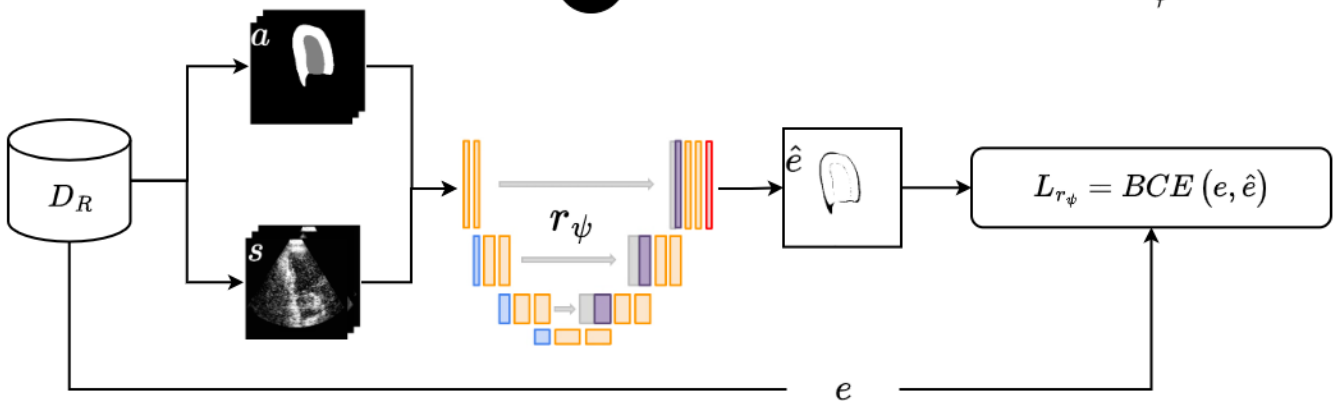
- ✓ Well-suited for semi-automatic annotation
- ✓ Strong generalization across tasks and modalities
- ✓ A powerful starting point for downstream tasks

- × Still less accurate than task-specific methods
- × Efficient multimodal integration remains an open challenge
- × Explainability and robustness remain key challenge
- × Clinical adaptation remains essential

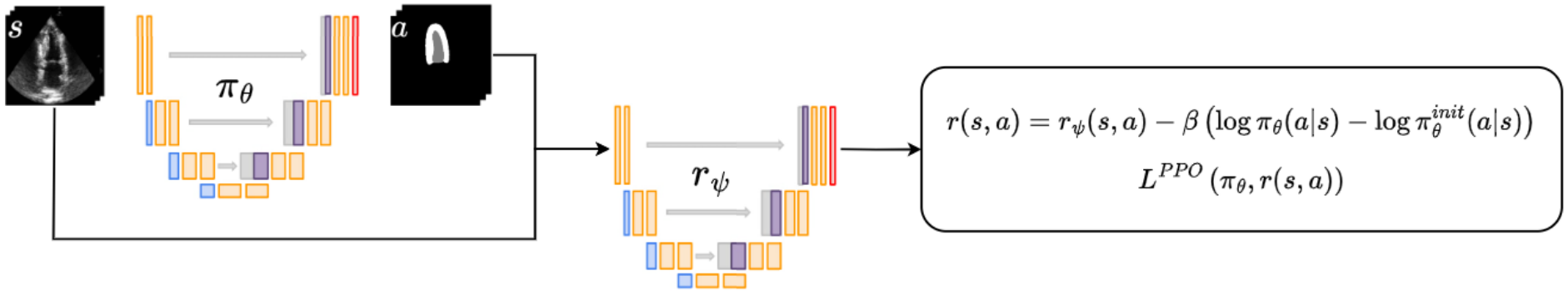
Perspectives - Domain adaptation through reinforcement learning



2 Train a reward network r_ψ



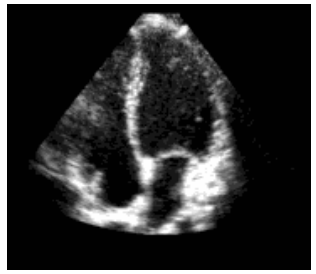
3 Fine-tune the policy network π_θ with PPO algorithm



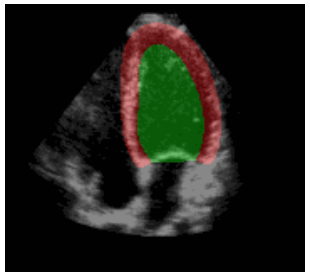
Perspectives - Domain adaptation through reinforcement learning

Successful case

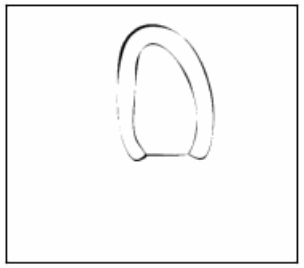
Input image



Segmentation

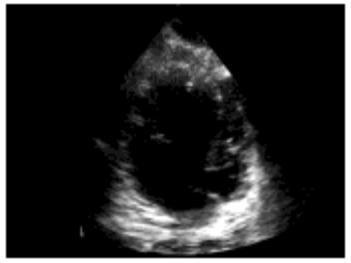


Reward

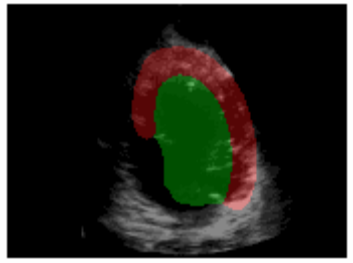


Detected failure case with correction

Input image



Segmentation



Reward

TTO correction