Vision-language models for zero-shot multi-label fine-grained classification of chest X-ray images

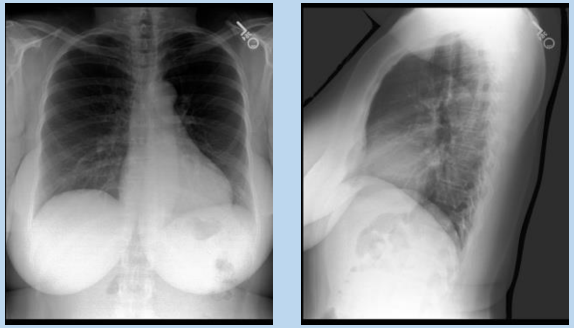
MICCAI 2024 Challenge: CXR-LT Task3

Yuyan Ge
University of Pennsylvania
yyge@seas.upenn.edu
Oct 10, 2024

Problem

- **Multi-label zero-shot classification**
  - Given chest X-ray images labeled with 40 diseases, learn classifier for 5 unseen diseases
- **Multi-label**: predict presence or absence of multiple diseases in each X-ray
- **Zero-shot**: predict 5 new diseases not seen during training

Example of one study



Multi-view X-ray images

Seen classes:
1. Adenopathy
2. Atelectasis
3. Azygos Lobe
4. Calcification of the Aorta

…

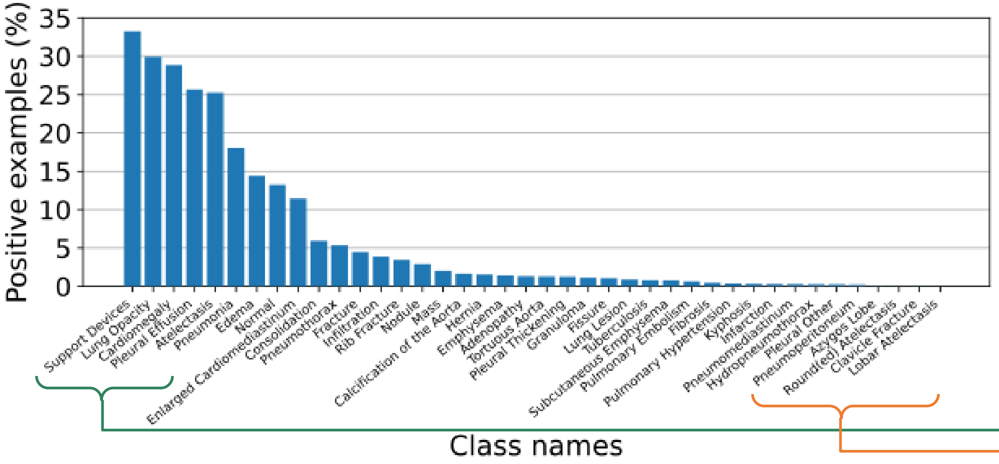
39. Tortuous Aorta
40. Tuberculosis

Target classes:
1. Bulla
2. Cardiomyopathy
3. Hilum
4. Osteopenia
5. Scoliosis

[1] CXR-LT Challenge website: https://bionlplab.github.io/2024_MICCAI_CXRLT/

Challenges

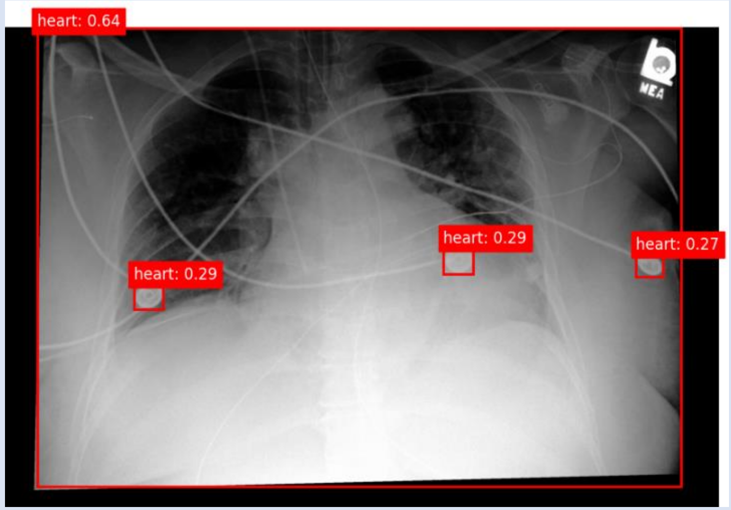
- **Zero-shot classification:** new classes are unseen and may be difficult to identify
- **Fine-grained classification:** class differences are nuanced, hence generalist models not trained on domain-specific data may fail
- **Long-tailed distribution:** percent of positive examples varies across classes (0.05% - 33%)



High Positive Classes:
Support Devices
Lung Opacity
Cardiomegaly
Pleural Effusion
Atelectasis

Low Positive Classes:
Lobar Atelectasis
Clavicle Fracture
Round(ed) Atelectasis
Azygos Lobe
Pneumoperitoneum

Failed example

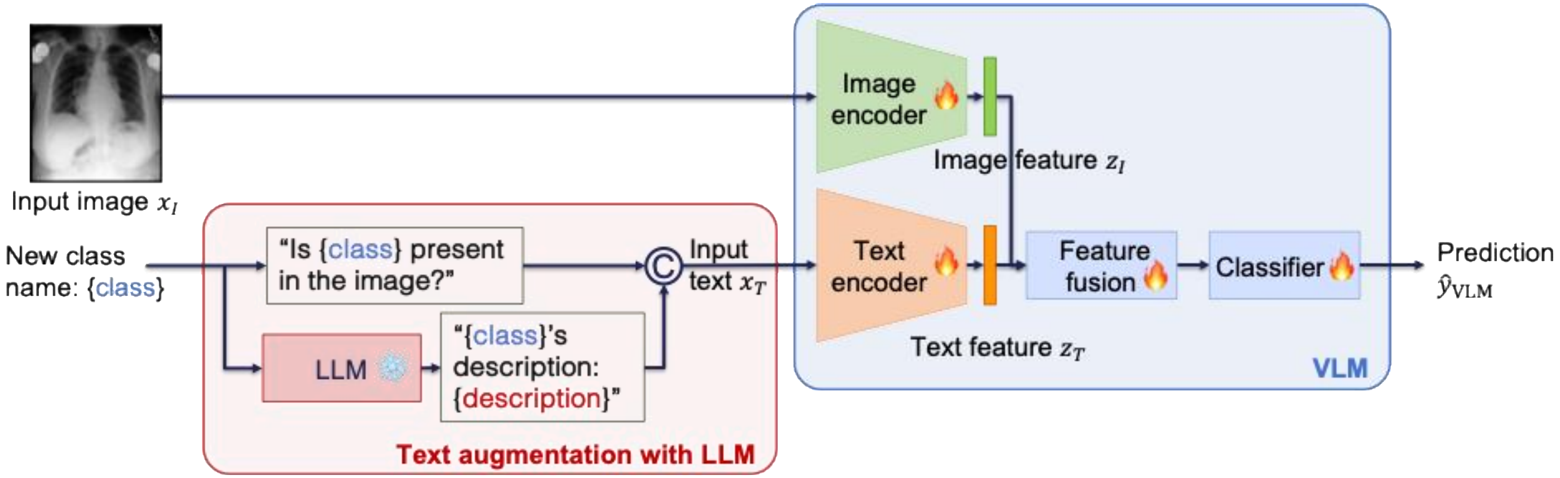


Model: GroundingDINO [1]
Input text: ‘heart. ’
Prediction: ['heart', 'heart', 'heart', 'heart']

[1] Liu et al., GroundingDINO. ECCV 2024. https://arxiv.org/abs/2303.05499

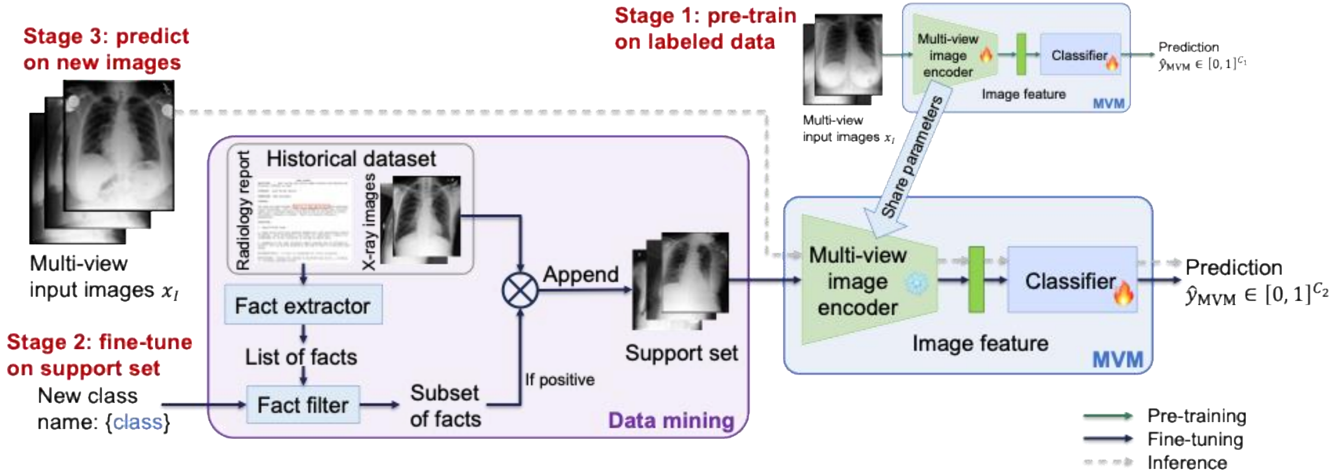
Method: Ensemble of two Models

- **Vision-Language Model (VLM)**



Pre-trained: domain specific data
Fine-tuned: 40-class dataset augmented with LLM-retrieved fine-grained class descriptions

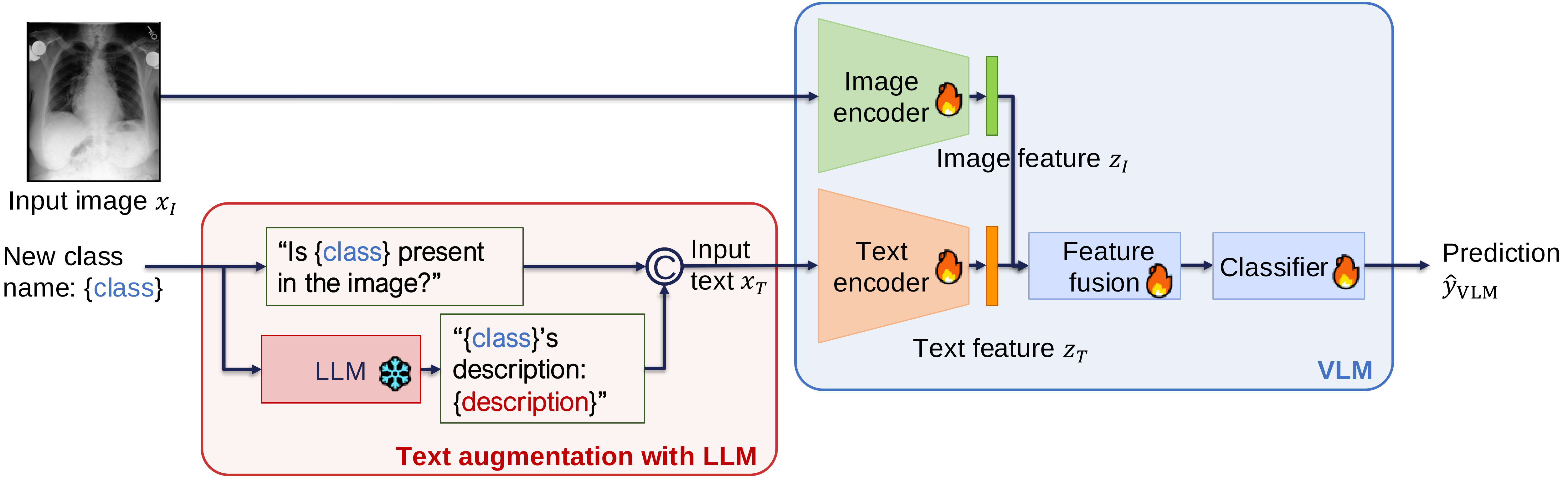
- **Multi-view Vision Model (MVM)**



Pre-trained: 40-class dataset
Fine-tuned: support set mined from historical data

Method 1: vision-language model

- **Key idea**: train a foundation vision-language model (VLM) to answer binary questions about the image content



[1] Oquab et al., DINOv2: Learning Robust Visual Features without Supervision. 2024. http://arxiv.org/abs/2304.07193
[2] Benedikt et al., CXR-BERT-specialized. ECCV 2022. https://arxiv.org/pdf/2204.09817

Method 1: vision-language model

- Multi-stage approach to improve feature representations in CXR:

Stage 1: pretrain on multiple domain-specific, partially-labeled datasets

Stage 2: finetune on classification task with fine-grained description of labels

Stage 3: zero-shot inference for new classes (with descriptions)

Table. Stage 1 datasets

Name	Converted from	Number of text
MIMIC-CXR [1]	Radiology reports	583,202
CXR-Concepts	Question answering	4,145
Chest ImaGenome [2]	Scene graph	93
CXR-LT [3]	Classification	40
Total	--	587,480

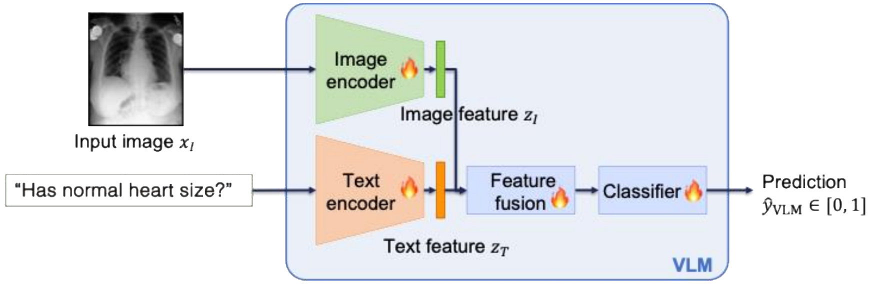


Figure. Stage 1 pretraining

[1] MIMIC-CXR. 2019. https://www.nature.com/articles/s41597-019-0322-0
[2] Chest ImaGenome. 2021. https://arxiv.org/pdf/2108.00316
[3] CXR-LT. 2024. https://codalab.lisn.upsaclay.fr/competitions/18604

Stage 2: finetune with class descriptions

- Use LLM knowledge retrieval to obtain fine-grained descriptions for each class

Prompt to ChatGPT-4o: "*You are an expert in radiology. I will give you a list of diseases related to chest x-ray. For each disease, please provide the visual description and visual facts of the disease. Please make sure the description is concise, and the visual facts are helpful to classify. Disease: {class name}.*"

Response from ChatGPT-4o:

Adenopathy:

- **Description:** Enlarged lymph nodes in the chest.
- **Key Features:** Visible as masses or enlarged areas near the mediastinum, often seen in the hilum region.

- Use class names + class descriptions to finetune VLM

Class name: Adenopathy

Input text to VLM:
"***Is** Adenopathy **present in the image?** Adenopathy**'s description:** Enlarged lymph nodes in the chest. Visible as masses or enlarged areas near the mediastinum, often seen in the hilum region.*"

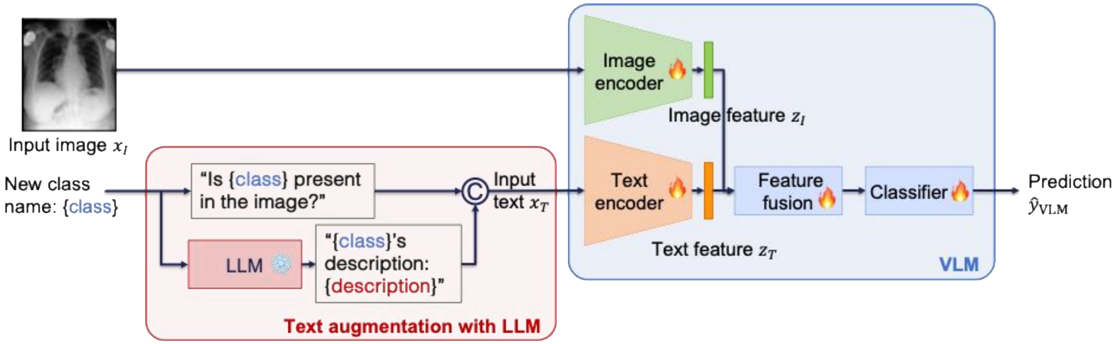
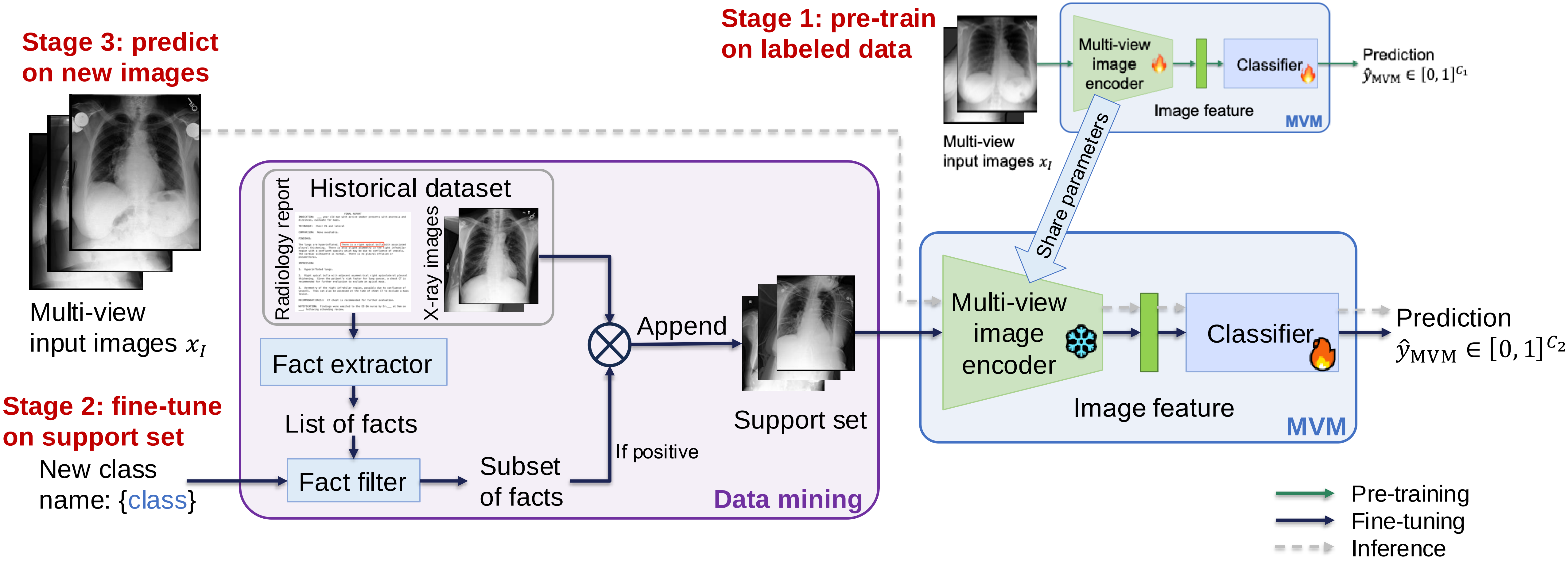


Figure. Stage 2 finetuning

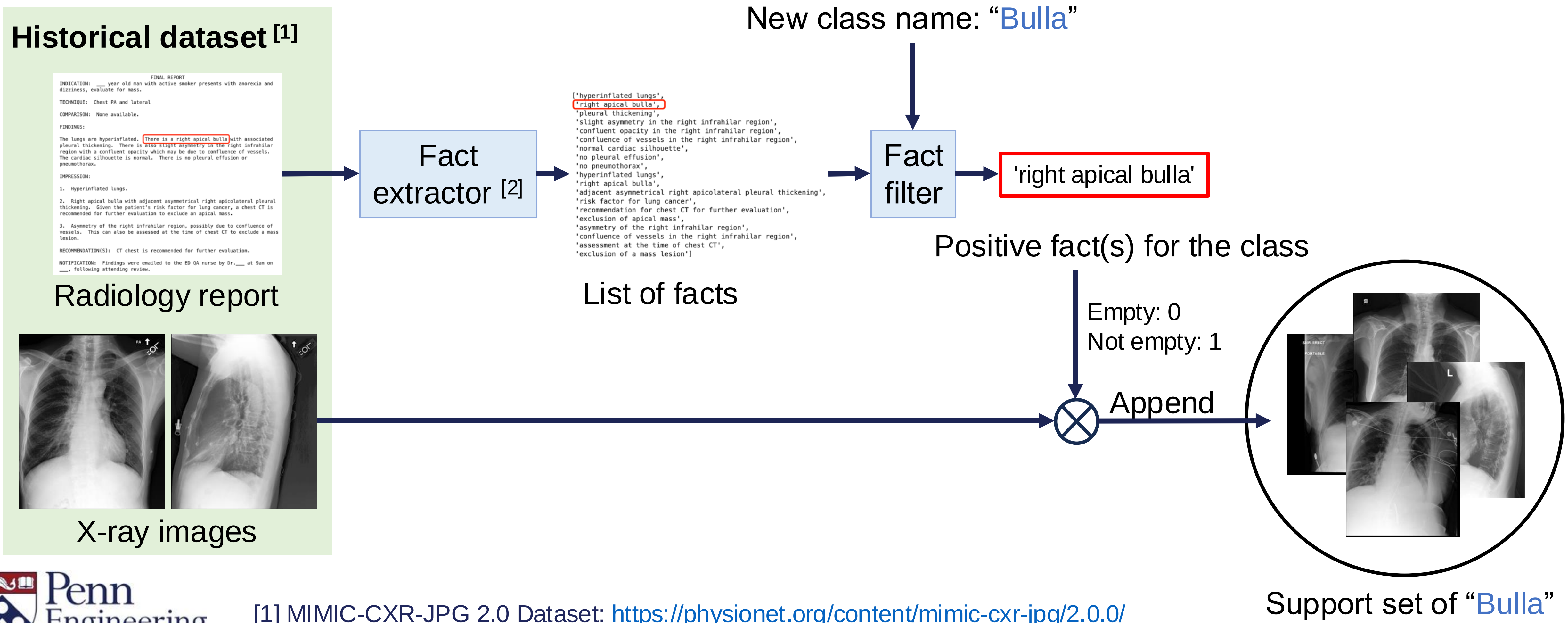
Method 2: multi-view vision model

- **Key idea**: convert zero-shot problem to few-shot problem by constructing support set for new classes



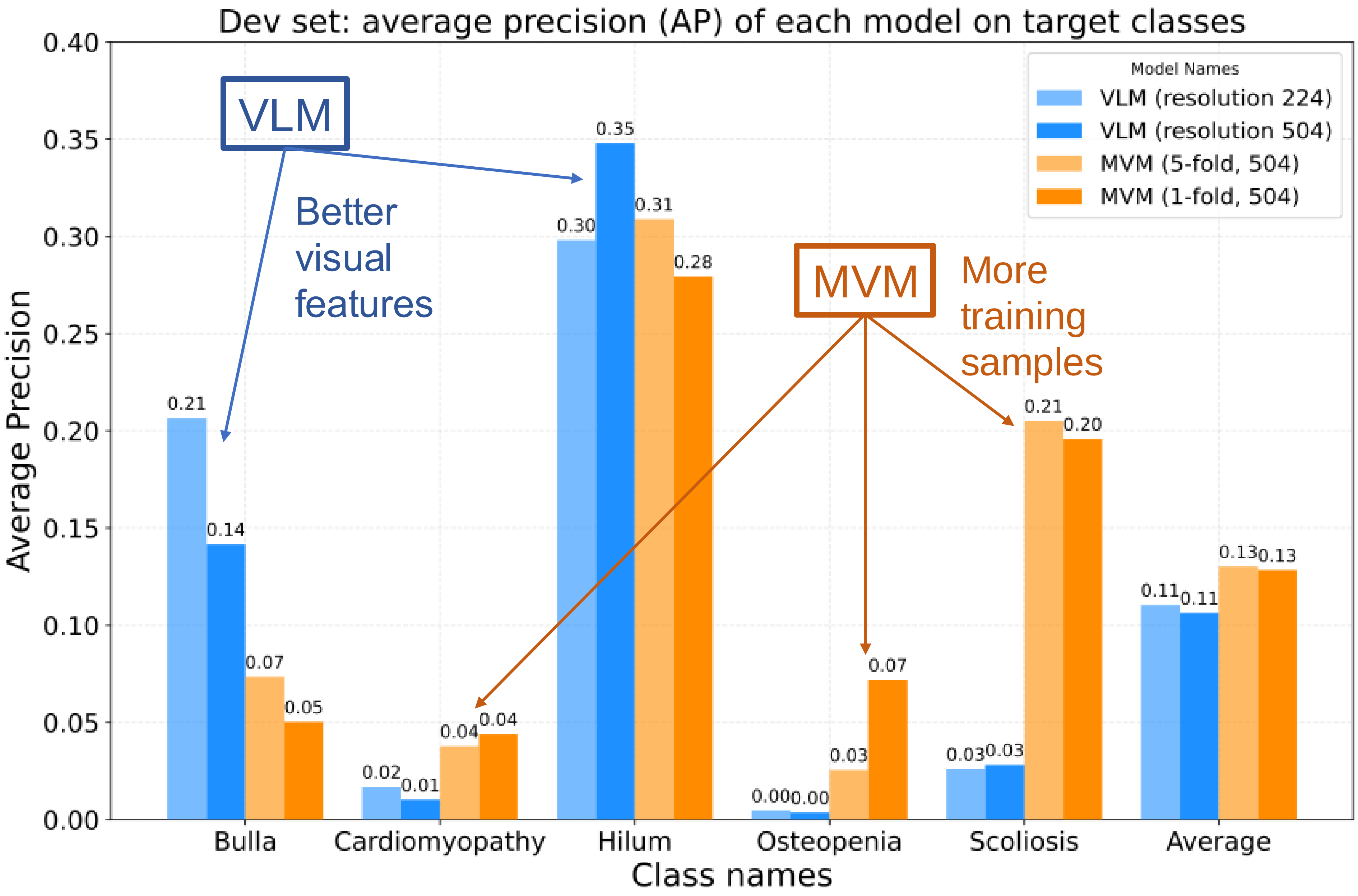
Stage 2: construct support set

- Mine positive samples from historical datasets



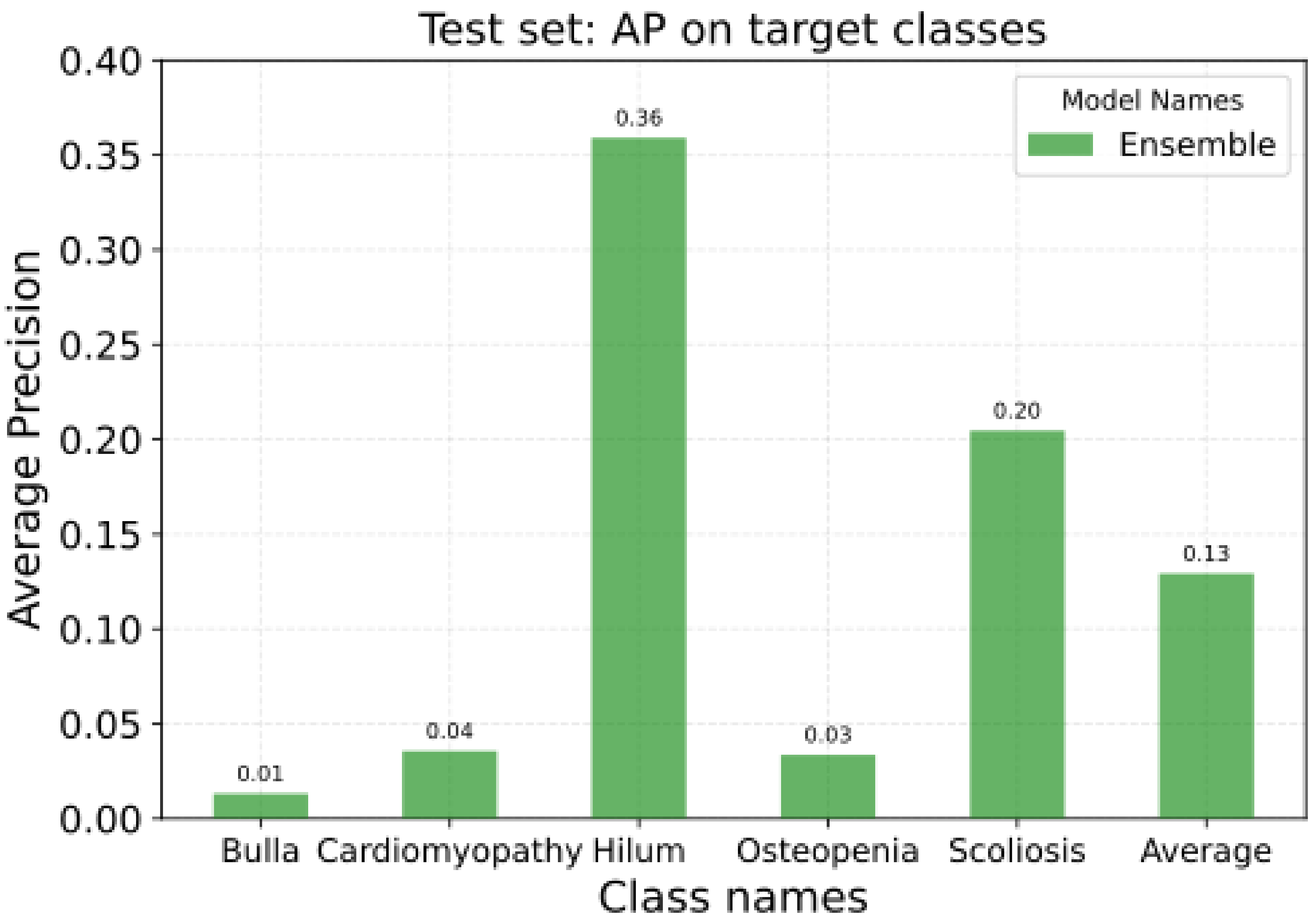
[1] MIMIC-CXR-JPG 2.0 Dataset: https://physionet.org/content/mimic-cxr-jpg/2.0.0/
[2] Messina et al., CXRFE. ACL 2024: https://arxiv.org/abs/2407.01948

Results



Dev set: average precision (AP) of each model on target classes
Average Precision
Class names
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Bulla
Cardiomyopathy
Hilum
Osteopenia
Scoliosis
Average
Model Names
VLM (resolution 224)
VLM (resolution 504)
MVM (5-fold, 504)
MVM (1-fold, 504)
0.21
0.14
0.07
0.05
0.02
0.01
0.04
0.04
0.30
0.35
0.31
0.28
0.00
0.00
0.03
0.07
0.03
0.03
0.21
0.20
0.11
0.11
0.13
0.13
VLM
Better visual features
MVM
More training samples

Results



Test set: AP on target classes
Model Names
Ensemble
Average Precision
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.01
0.04
0.36
0.03
0.20
0.13
Bulla
Cardiomyopathy
Hilum
Osteopenia
Scoliosis
Average
Class names

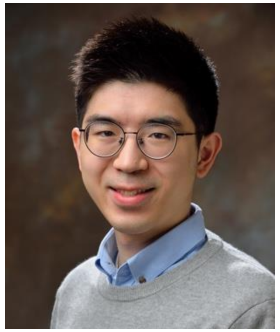
Conclusion

- VLM
  - Train on a unified domain-specific dataset
    - Incorporate generalization ability of LLMs
  - Use external knowledge
    - Incorporate domain knowledge from LLM
- MVM
  - Multi-view feature fusion
    - Aggregate multi-view images of one study
  - Support set
    - Mine positive samples for new classes from historical free-text datasets, converting zero-shot problem to few-shot problem

Acknowledgement

René Vidal
University of Pennsylvania

Pablo Messina
Pontificia Universidad Católica de Chile



Ryan Chan
University of Pennsylvania

Code

Code is available at
https://github.com/Glourier/MICCAI2024-ZeroShotCXR

For discussion, please feel free to reach out to
Yuyan Ge at *yyge@seas.upenn.edu*