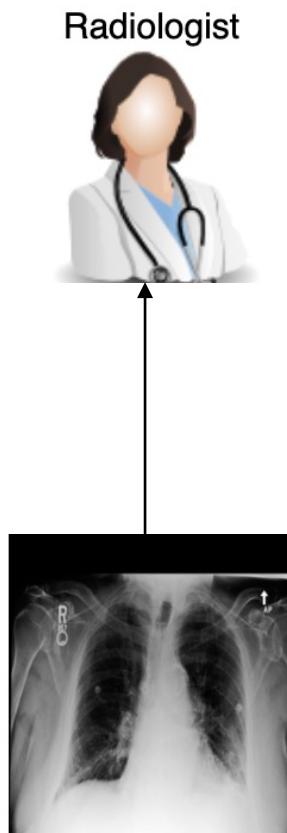




Can Image-Text Integration Improve Fact-Checking in Clinical Report Generation?

Jinge Wu
21 Nov 2023

Research Background



Chest radiograph

Radiology report

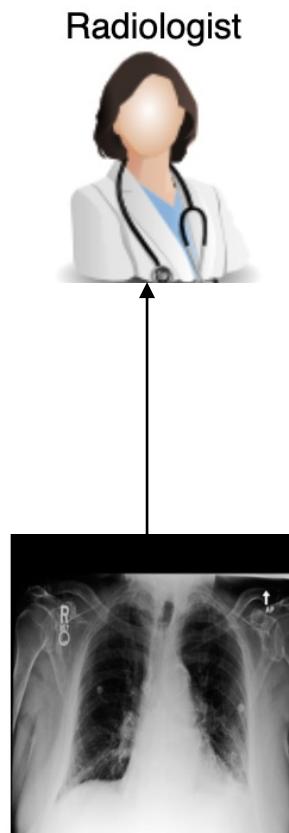
Findings: Lung volumes are very low, exaggerating cardiomedastinal silhouette and pulmonary vascular markings. Linear density in the right lower lung field very likely represents atelectasis. The patient's chin obscures the medial portions of the lung apices. No pleural effusion or pneumothorax is detected. Lower thoracic vertebra plana is chronic. Right humeral head hardware is noted.

Impression: Markedly low lung volumes, with right lung atelectasis.

detailed observations made by the radiologist when examining an imaging study

summary or interpretation of the findings

Research Background



Chest radiograph

Radiology report

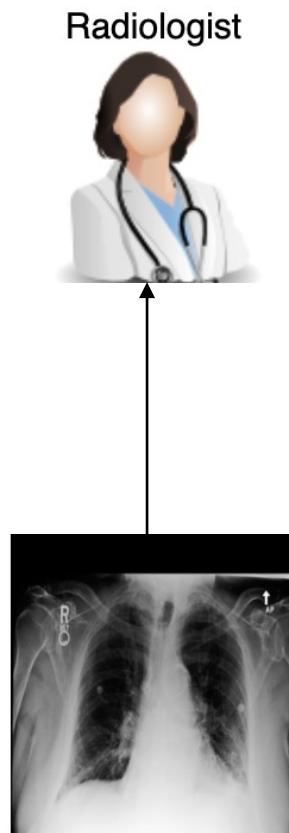
Findings: Lung volumes are very low, exaggerating cardiomediastinal silhouette and pulmonary vascular markings. Linear density in the right lower lung field very likely represents atelectasis. The patient's chin obscures the medial portions of the lung apices. No pleural effusion or pneumothorax is detected. Lower thoracic vertebra plana is chronic. Right humeral head hardware is noted.

Impression: Markedly low lung volumes, with right lung atelectasis.

Summarizer

Generated Impression: Markedly low lung volumes, with right lung atelectasis.

Research Background



Radiologist

Findings: Dual lead left chest wall pacing device is again seen. There are patchy bibasilar opacities identified, **left greater than right**. Superiorly the lungs are clear. Cardiomediastinal silhouette is stable. Median mediastinal clips again noted.

Impression: Bibasilar opacities more extensive on the **left than on the right**. Findings could be in part due to atelectasis given the **low lung volumes** although superimposed infection is possible.

Summarizer

Generated Impression: Bibasilar opacities potentially atelectasis noting that infection is not entirely excluded.

Challenge 1: Missing information which is
not given from the findings!

Chest radiograph

Research Background



Findings: Lung volumes are very low, exaggerating cardiomediastinal silhouette and pulmonary vascular markings. Linear density in the **right lower lung** field very likely represents **atelectasis**. The patient's chin obscures the medial portions of the lung apices. No pleural effusion or pneumothorax is detected. Lower thoracic vertebra plana is chronic. Right humeral head hardware is noted.

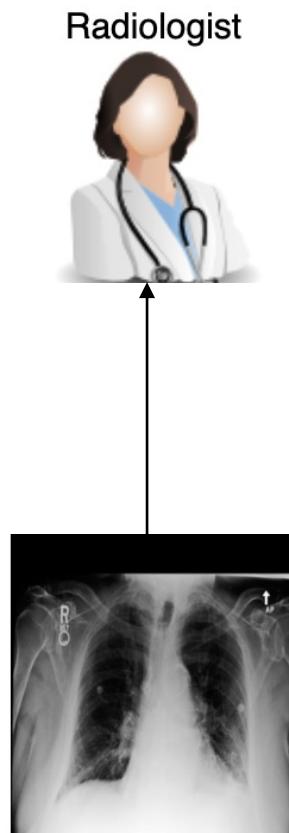
Impression: Markedly low lung volumes, with **right lung atelectasis**.

Summarizer

Generated Impression: Markedly low lung volumes, with **left lung atelectasis**.

Challenge2: factual inconsistency from generative AI model

Research Background



Chest radiograph

Findings: Lung volumes are very low, exaggerating cardiomediastinal silhouette and pulmonary vascular markings. Linear density in the right lower lung field very likely represents **edema**. The patient's chin obscures the medial portions of the lung apices. No pleural effusion or pneumothorax is detected. Lower thoracic vertebra plana is chronic. Right humeral head hardware is noted.

Impression: Markedly low lung volumes, with right lung **atelectasis**.

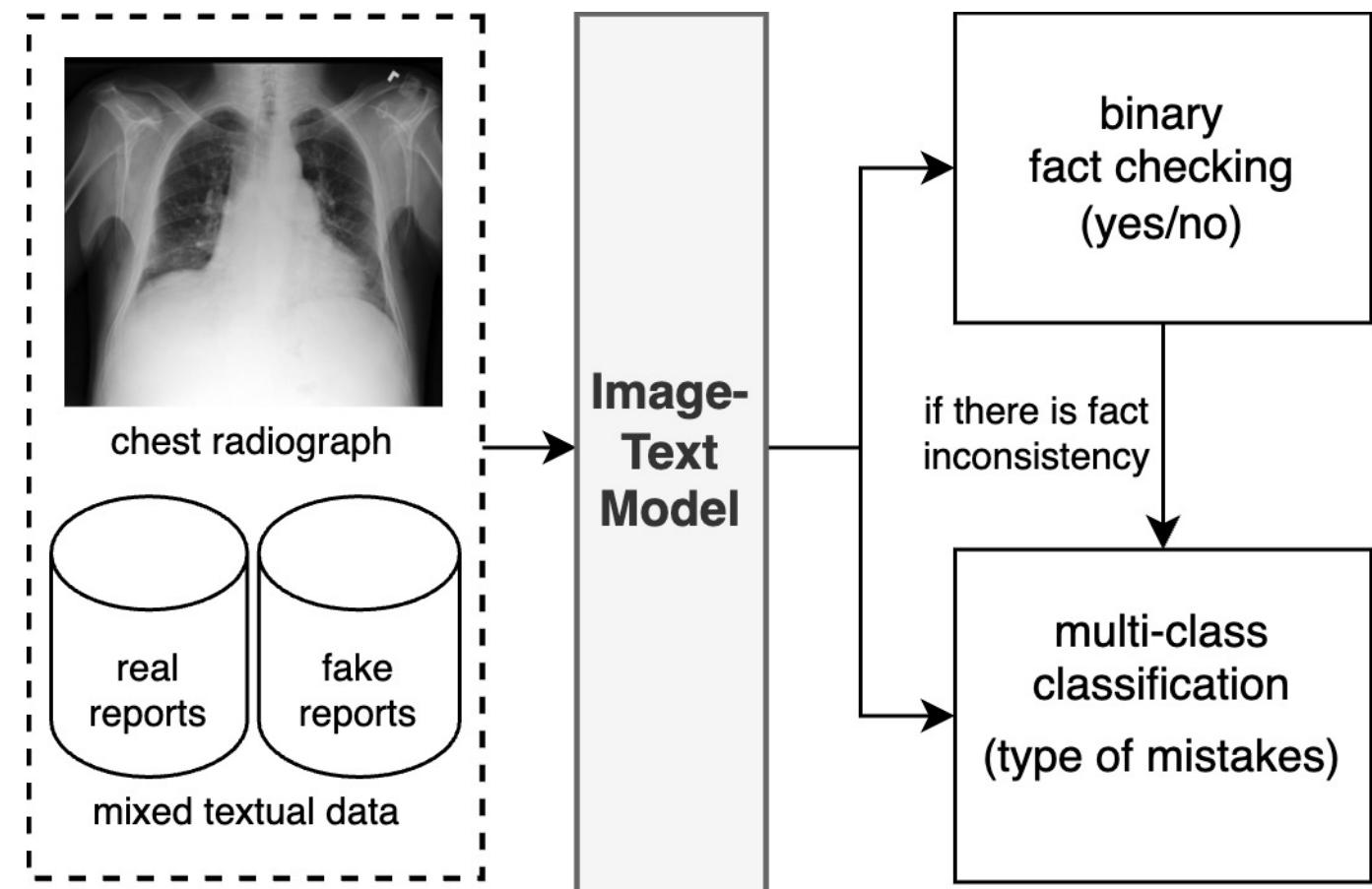
Summarizer

Generated Impression: Markedly low lung volumes, with right lung **edema**.

Challenge3: mistakes made by radiologist

What we do?

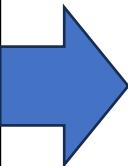
- Develop a fact checker with integration of images and textual reports
- Creation of benchmark dataset with different types of mistakes



Benchmark Data Creation

Findings: PA and lateral views of the chest provided. The lungs are adequately aerated. There is a **focal consolidation** at the **left lung base** adjacent to the lateral hemidiaphragm. There is mild vascular **engorgement**. There is bilateral apical pleural thickening. The cardiomedastinal silhouette is remarkable for aortic arch calcifications. The heart is top normal in size.

remove



Original data

Findings: PA and lateral views of the chest provided. The lungs are adequately aerated. There is a **focal consolidation** at the **left lung base** adjacent to the lateral hemidiaphragm. ~~There is mild vascular engorgement.~~ There is bilateral apical pleural thickening. The cardiomedastinal silhouette is remarkable for aortic arch calcifications. The heart is top normal in size.

add

Findings: PA and lateral views of the chest provided. The lungs are adequately aerated. There is a **focal consolidation** at the **left lung base** adjacent to the lateral hemidiaphragm. There is mild vascular **engorgement**. **The patient also shows signs of bronchitis.** There is bilateral apical pleural thickening. The cardiomedastinal silhouette is remarkable for aortic arch calcifications. The heart is top normal in size.

Data resources

Training data: MIMIC Chest X-ray (MIMIC-CXR)¹:

- 377,710 chest radiographs
- 227,835 radiograph studies

 Database  Credentialed Access

MIMIC-CXR Database

Alistair Johnson , Tom Pollard , Roger Mark , Seth Berkowitz , Steven Horng 

Published: Sept. 19, 2019. Version: 2.0.0

 Database  Credentialed Access

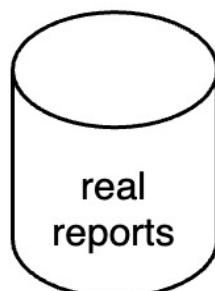
Radiographs + radiology reports

MIMIC-CXR-JPG - chest radiographs with structured labels

Alistair Johnson , Matt Lungren , Yifan Peng , Zhiyong Lu , Roger Mark , Seth Berkowitz , Steven Horng 

Published: Nov. 14, 2019. Version: 2.0.0

Radiographs + labels extracted from reports



Benchmark data: 49 paired radiology reports including 29 original reports + 20 modified dataset (10 with "remove" mistake+ 10 with "add" mistake)

¹<https://physionet.org/content/mimic-cxr/2.0.0/>, <https://physionet.org/content/mimic-cxr-jpg/2.0.0/>

Models

Image Models (training and inference):

CheXpert-DenseNet121-ensemble²:

gradient-weighted class activation map (Grad-CAMs) leveraged for visualization.

JFHealthcare-DenseNet121³ : a novel probabilistic class activation map (PCAM) is also leveraged during training.

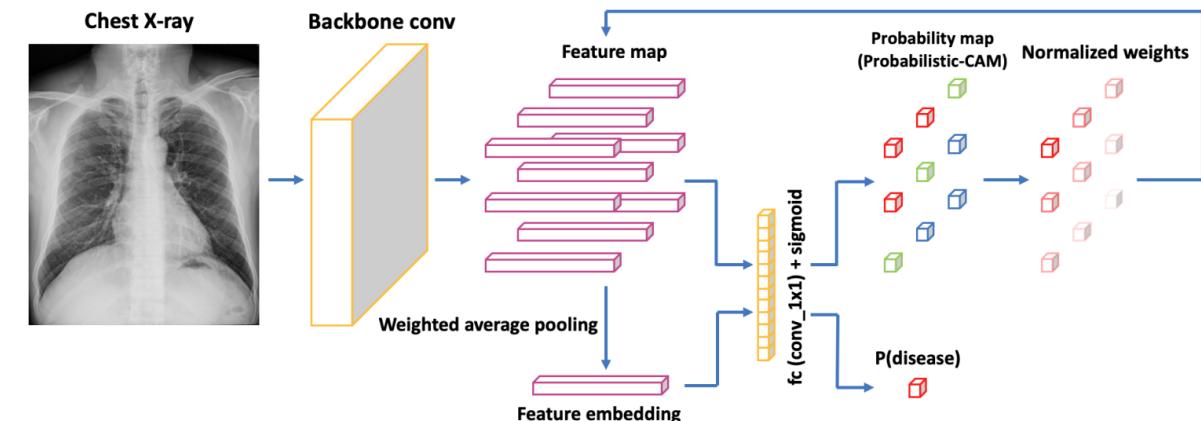


Figure 1: The framework of Probabilistic-CAM (PCAM) pooling.

Both models are fine-tuned for multi-label classification with five labels:
 “Consolidation”, “Atelectasis”, “Effusion”, “Cardiomegaly”, “Edema”.

²Irvin, Jeremy, et al. "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. No. 01. 2019.

³Ye, Wenwu, et al. "Weakly supervised lesion localization with probabilistic-cam pooling." *arXiv preprint arXiv:2005.14480* (2020).

Models

Language Models (inference only) :

Llama-7b⁴ : open source large language model with 7B parameters
Llama-13b⁴ : the same model with 13B parameters

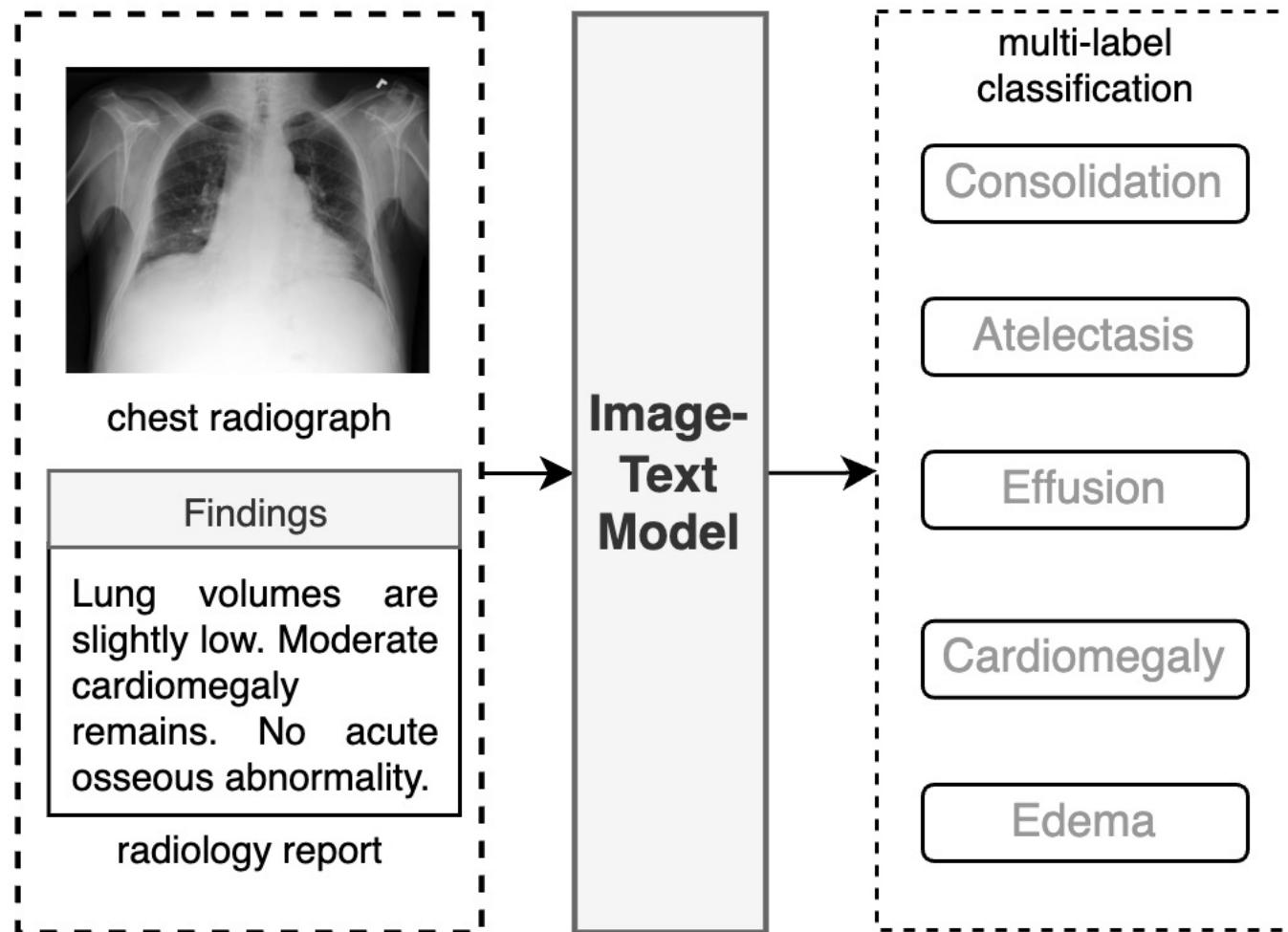
The models are trained with general language corpus.

Tasks:

1. Inference on multi-label classification
2. Inference on binary fact-checking

⁴Touvron, Hugo, et al. "Llama 2: Open foundation and fine-tuned chat models." arXiv preprint arXiv:2307.09288 (2023).

Models



Evaluation

1. Multi-label classification from both models:

“Consolidation”, “Atelectasis”, “Effusion”, “Cardiomegaly”, “Edema”.

2. Fact checking:

- Binary classification (Yes/No) from Llama only
- Multi-class classification (*Correct*, “add” mistake, “remove” mistake) for both models¹

¹With multi-label predictions from both models, we define the multi-class predictions as follows:

1. if *pred_image* is a subset of *pred_language* -> “add” mistake
2. If *pred_language* is a subset of *pred_image* -> “remove” mistake
3. If *pred_language* = *pred_image* -> “correct” no mistake

Results – training results for multi-label classification

Model	Precision	Recall	F1	Acc
Llama-7b	0.8265	0.8264	0.8265	0.7103
Llama-13b	0.8435	0.8016	0.8220	0.7103
CheXpert-DenseNet121-ensemble	1.0000	0.3306	0.4969	0.4413
JFHealthcare-DenseNet121	0.9106	0.9256	0.9180	0.8621

Table 1. Results for multi-label classification

Results – fact checking (binary/multi)

Model	Yes	No
Llama-7b	0.4545	0.4706
Llama-13b	0.6885	0.4000

Table 2. F1-score for binary classification of mistakes.

Model	Overall	Correct	Remove	Add
Llama-7b + CheXpert	0.1719	0.0000	0.2857	0.1579
Llama-13b + CheXpert	0.1821	0.0000	0.0714	0.3000
Llama-7b + JFhealthcare	0.2280	0.4091	0.3030	0.1000
Llama-13b + JFhealthcare	0.3694	0.3636	0.0690	0.2400

Table 3. F1-score for multi-class classification of mistakes. Overall means the average score of three different types “Correct”, “Remove” and “Add”.

Results-evaluation on the original report

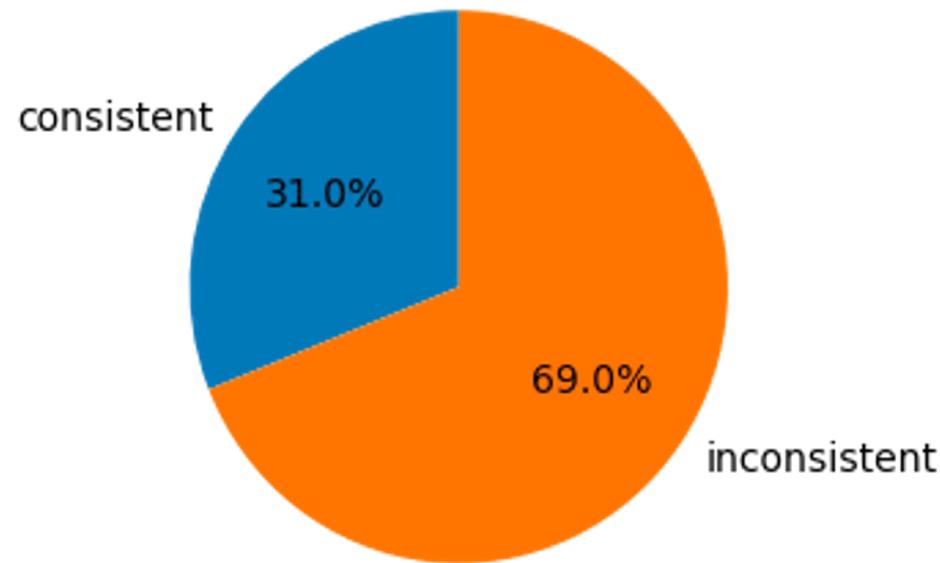


Figure 1. Fact-checking on the original dataset (all real reports).

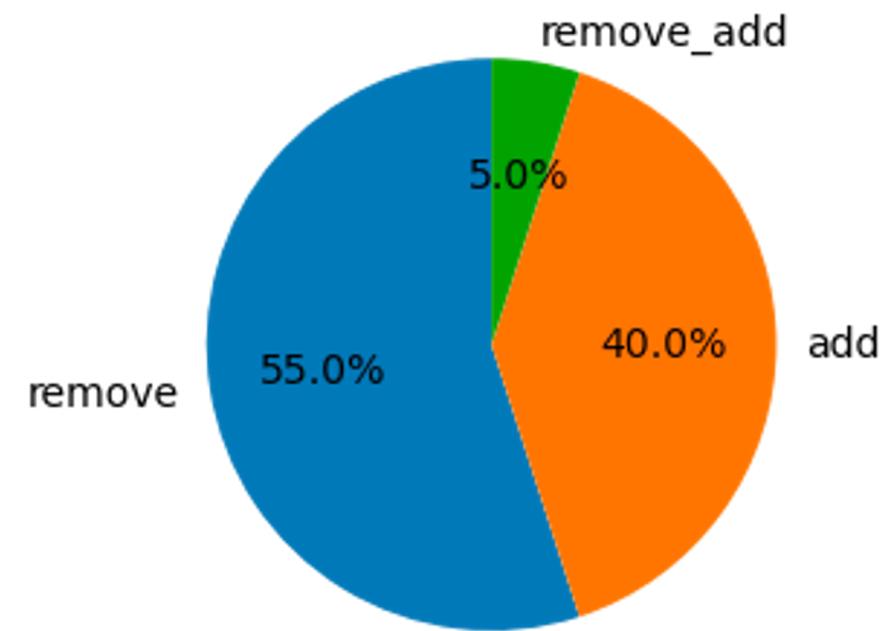


Figure 2. Different types of mistakes found from the original dataset (all real reports). *remove_add* indicates both “remove” and “add” mistakes detected.

Summary

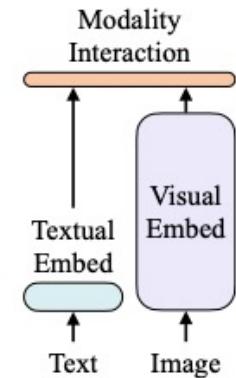
- Proposed a clinical fact checking benchmark with integration of images and texts

Future Plan

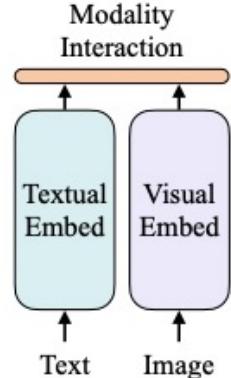
- Larger benchmark dataset with more different types of mistakes
- Human evaluations
- Vision-language models

A brief introduction to vision-language models

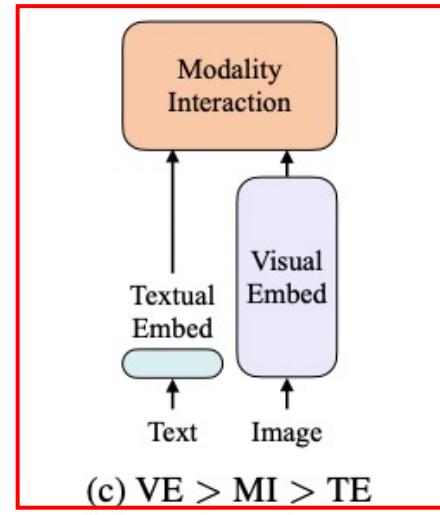
Vision-Language Models (VLMs)



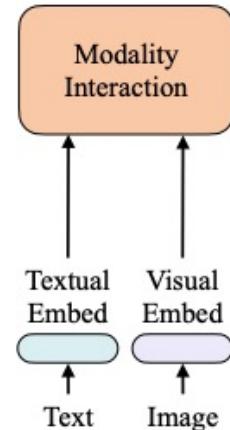
(a) $VE > TE > MI$
Eg. VSE



(b) $VE = TE > MI$
Eg. CLIP



(c) $VE > MI > TE$
Eg. OSCAR, ViLBERT, UNITER



(d) $MI > VE = TE$
Eg. ViLT

Figure 2. Four categories of vision-and-language models. The height of each rectangle denotes its relative computational size. VE, TE, and MI are short for visual embedder, textual embedder, and modality interaction, respectively.

LLaVA^[1]: The first propose visual instruction prompting

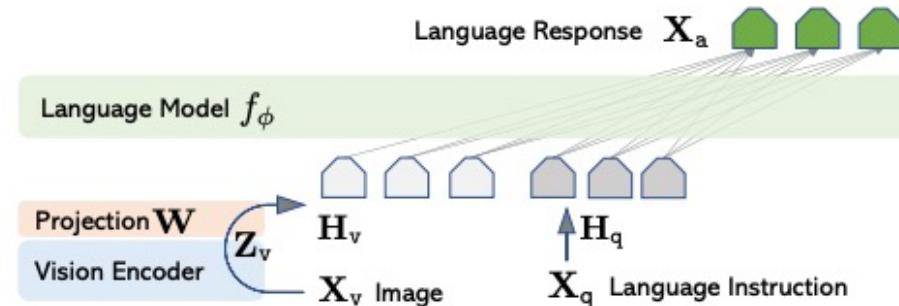


Figure 1: LLaVA network architecture.

Stage 1: Pre-training for Feature Alignment

Only the projection matrix is updated, based on a subset of CC3M

Stage 2: Fine-tuning End-to-End

Both the projection matrix and LLM are updated for two different use scenarios:

- **Visual Chat:** fine-tuned on our generated multimodal instruction-following data for daily user-oriented applications.
- **Science QA:** fine-tuned on multimodal reasoning dataset for the science domain

Context type 1: Captions

A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.



Context type 2: Boxes

person: [0.681, 0.242, 0.774, 0.694], person: [0.63, 0.222, 0.686, 0.516], person: [0.444, 0.233, 0.487, 0.34], backpack: [0.384, 0.696, 0.485, 0.914], backpack: [0.755, 0.413, 0.846, 0.692], suitcase: [0.758, 0.413, 0.845, 0.69], suitcase: [0.1, 0.497, 0.173, 0.579], bicycle: [0.282, 0.363, 0.327, 0.442], car: [0.786, 0.25, 0.848, 0.322], car: [0.783, 0.27, 0.827, 0.335], car: [0.86, 0.254, 0.891, 0.3], car: [0.261, 0.101, 0.787, 0.626]

Response type 1: conversation

Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV).
Question: Where is the vehicle parked?
Answer: The vehicle is parked in an underground parking area, likely in a public garage.

Question: What are the people in the image doing?
Answer: The people are trying to fit all of their luggage into the SUV, likely preparing for a trip.

Response type 2: detailed description

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip.

In the parking area, various luggage items are scattered around the vehicle. There are two backpacks, one located near the left rear wheel and the other closer to the right side of the vehicle. Additionally, there are two suitcases, one on the right side of the car and another further away near the center of the parking area. A bicycle can also be seen on the left side of the vehicle. Other cars are parked around the main SUV, with one car positioned behind it and slightly to the left, another behind and slightly to the right, and the third car further behind on the right side.

Response type 3: complex reasoning

Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings to accommodate. They might have to strategize and arrange the luggage efficiently to ensure that everything fits properly into the vehicle. Additionally, they need to consider the comfort of the passengers and visibility while driving, so the placement of the luggage must not obstruct the driver's view or make the passengers uncomfortable during the trip.

Table 1: One example to illustrate the instruction-following data. The top block shows the contexts such as captions and boxes used to prompt GPT, and the bottom block shows the three types of responses. Note that the visual image is not used to prompt GPT, we only show it here as a reference.

LLaVA 1.5^[1]: Achieves the best performance with the simplest architecture

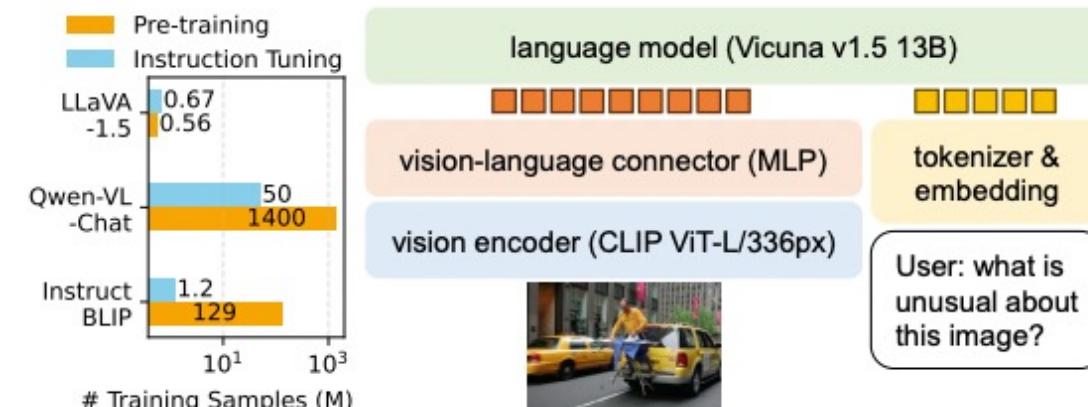
- MLP vision-language connector
- Academic task oriented data
- Additional scaling

Advantage:

- Zero-shot format instruction generalization.
- Zero-shot multilingual capability.

Disadvantage:

- Not yet capable of processing multiple images
- Not yet support for performing in-context learning (ICL)



LLaVA 1.5^[1]: Achieves the best performance with the simplest architecture

Visual input example, Tricky Question:



User If there are factual errors in the questions, point it out; if not, proceed answering the question.
What's happening in the desert?

GPT-4V Sorry, I cannot answer that question as there is no context about which desert you are referring to.

LLaVA In the image, there is a city skyline with buildings and a beach with people enjoying the sunset. The scene is set in a desert, which is unusual for a beach setting.

LLaVA-1.5 There are no deserts in the image. The image features a beach with palm trees, a city skyline, and a large body of water.

Visual input example, Constrained JSON Output:



User Please read the text in this image and return the information in the following JSON format (note xxx is placeholder, if the information is not available in the image, put "N/A" instead).

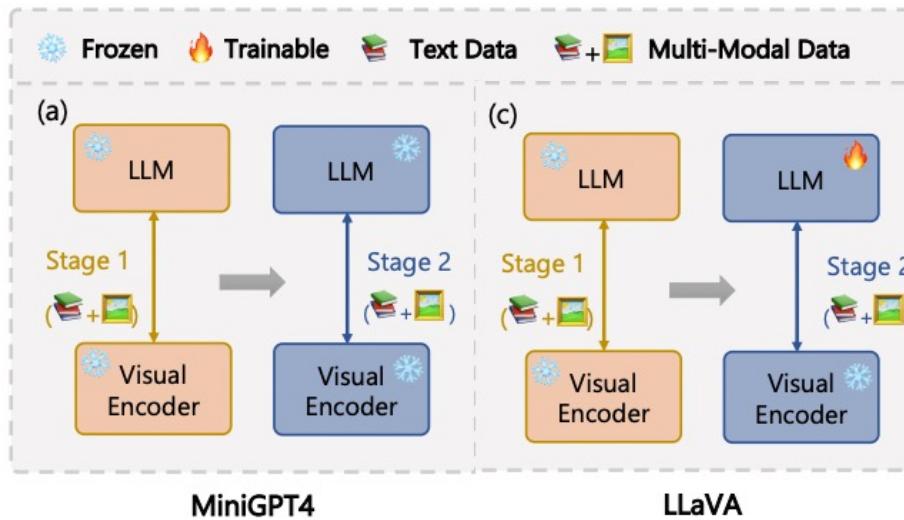
```
{"class": xxx, "DLN": xxx, "DOB": xxx, "Name": xxx, "Address": xxx, "EXP": xxx, "ISS": xxx, "SEX": xxx, "HGT": xxx, "WGT": xxx, "EYES": xxx, "HAIR": xxx, "DONOR": xxx}
```

GPT-4V {"class": "C", "DLN": "1234568", "DOB": "08/31/1977", "Name": "Ima Cardholder", "Address": "2570 24th Street, Anytown, CA 95818", "EXP": "08/31/2014", "ISS": "08/31/2011", "SEX": "F", "HGT": "5-05", "WGT": "125 lbs", "EYES": "BRN", "HAIR": "BRN", "DONOR": "VETERAN"}

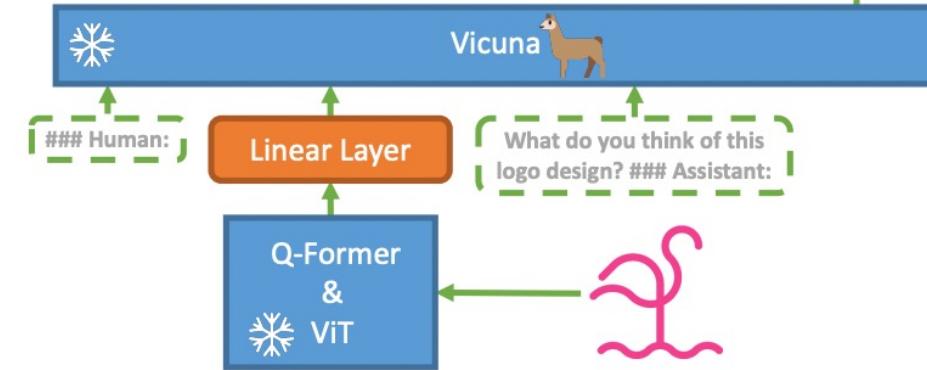
LLaVA {"class": "Driver License", "DLN": "CA123 456 789", "DOB": "01/01/1990", "Name": "Amy", "Address": "Anytown, CA", "EXP": "08/31/2023", "ISS": "Veteran", "SEX": "F", "HGT": "5'5", "WGT": "130", "EYES": "Brown", "HAIR": "Brown", "DONOR": "Unknown"}

LLaVA-1.5 {"class": "C", "DLN": "1132456789", "DOB": "08/23/1971", "Name": "Ima Cardholder", "Address": "Anytown, State, Zip", "EXP": "08/23/2014", "ISS": "California", "SEX": "F", "HGT": "5'5", "WGT": "123", "EYES": "Brown", "HAIR": "Brown", "DONOR": "N/A"}

Minigpt-4^[1]



The logo design is **simple** and **minimalistic**, with a **pink line drawing** of a **flamingo** **standing on one leg in the water**. The design is **clean** and **easy to recognize**, making it suitable for use in various contexts such as a logo for a beach resort or a flamingo-themed event. The use of a flamingo as a symbol adds a touch of **whimsy** and fun to the design, making it **memorable** and **eye-catching**. Overall, it's a **well-designed logo** that **effectively communicates the brand's message**.



1. First pretraining stage

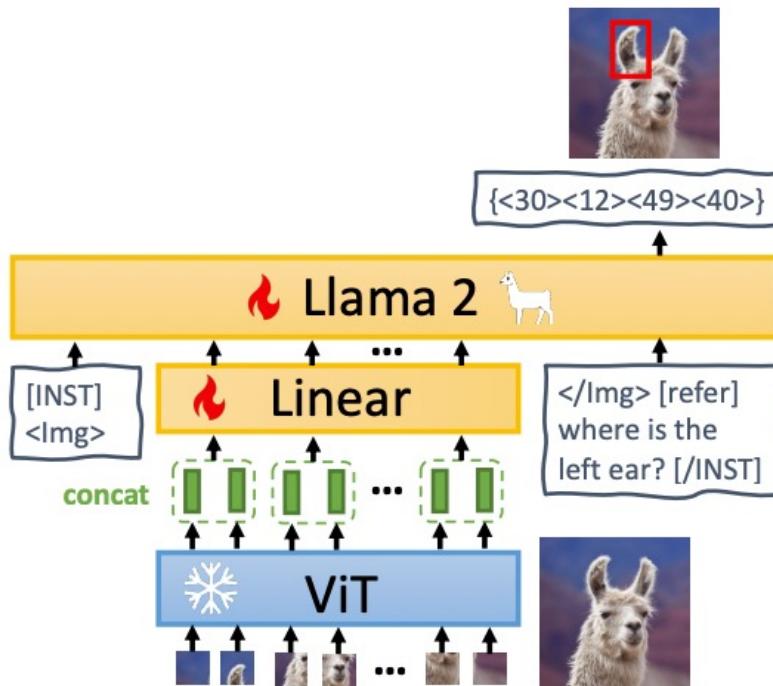
20,000 training steps with a batch size of 256, covering approximately 5 million image-text pairs

2. Fine-tuning with high-quality alignment dataset

Extend short answer, remove repetitive words, human annotations

[1] Zhu D, Chen J, Shen X, Li X, Elhoseiny M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592. 2023 Apr 20.

Minigpt-v2^[1]



- EVA visual backbone + LLaMA2
- Multi-task Instruction

[INST] <ImageFeature> [Task Identifier] Instruction [/INST]

- Special locations
- < Xleft >< Ytop >< Xright >< Ybottom >

Tasks	VQA	Caption	Grounded Caption	REC	REG	Object Parsing and Grounding
Identifiers	[vqa]	[caption]	[grounding]	[refer]	[identify]	[detection]

Table 1: Task identifier tokens for 6 different tasks, including visual question answering, image captioning, grounded image captioning, referring expression comprehension (REC), referring expression generation (REG), and object parsing and grounding (where the model extracts objects from the input text and determines their bounding box locations).

[1] Chen J, Zhu D, Shen X, Li X, Liu Z, Zhang P, Krishnamoorthi R, Chandra V, Xiong Y, Elhoseiny M. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint arXiv:2310.09478. 2023 Oct 14.

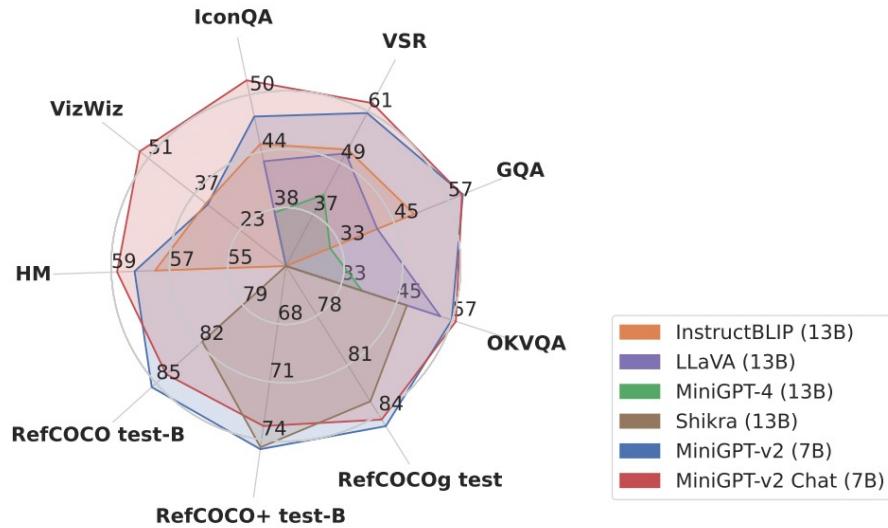
Minigpt-v2^[1]

Data types	Dataset	Stage 1	Stage 2	Stage 3
Weakly-labeled	GRIT-20M (REC and REG), LAION, CC3M, SBU	✓	✗	✗
Grounded caption	GRIT-20M	✓	✗	✗
Caption	COCO caption, Text Captions	✓	✓	✓
REC	RefCOCO, RefCOCO+, RefCOCOg, Visual Genome	✓	✓	✓
REG	RefCOCO, RefCOCO+, RefCOCOg	✓	✓	✓
VQA	GQA, VQAv2, OCR-VQA, OK-VQA, AOK-VQA	✓	✓	✓
Multimodal instruction	LLaVA dataset, Flickr30k, Multi-task conversation	✗	✗	✓
Langauge dataset	Unnatural Instructions	✗	✗	✓

Table 2: The training datasets used for our model three-stage training.

Method	Grounding	OKVQA	GQA	VSR (zero-shot)	IconVQA (zero-shot)	VizWiz (zero-shot)	HM (zero-shot)
Flamingo-9B	✗	44.7	-	31.8	-	28.8	57.0
BLIP-2 (13B)	✗	45.9	41.0	50.9	40.6	19.6	53.7
InstructBLIP (13B)	✗	-	49.5	52.1	44.8	33.4	57.5
MiniGPT-4 (13B)	✗	37.5	30.8	41.6	37.6	-	-
LLaVA (13B)	✗	54.4	41.3	51.2	43.0	-	-
Shikra (13B)	✓	47.2	-	-	-	-	-
Ours (7B)	✓	56.9	60.3	60.6	47.7	32.9	58.2
Ours (7B)-chat	✓	57.8	60.1	62.9	51.5	53.6	58.8

Table 3: **Results on multiple VQA tasks.** We report top-1 accuracy for each task. Grounding column indicates whether the model incorporates visual localization capability. The best performance for each benchmark is indicated in **bold**.



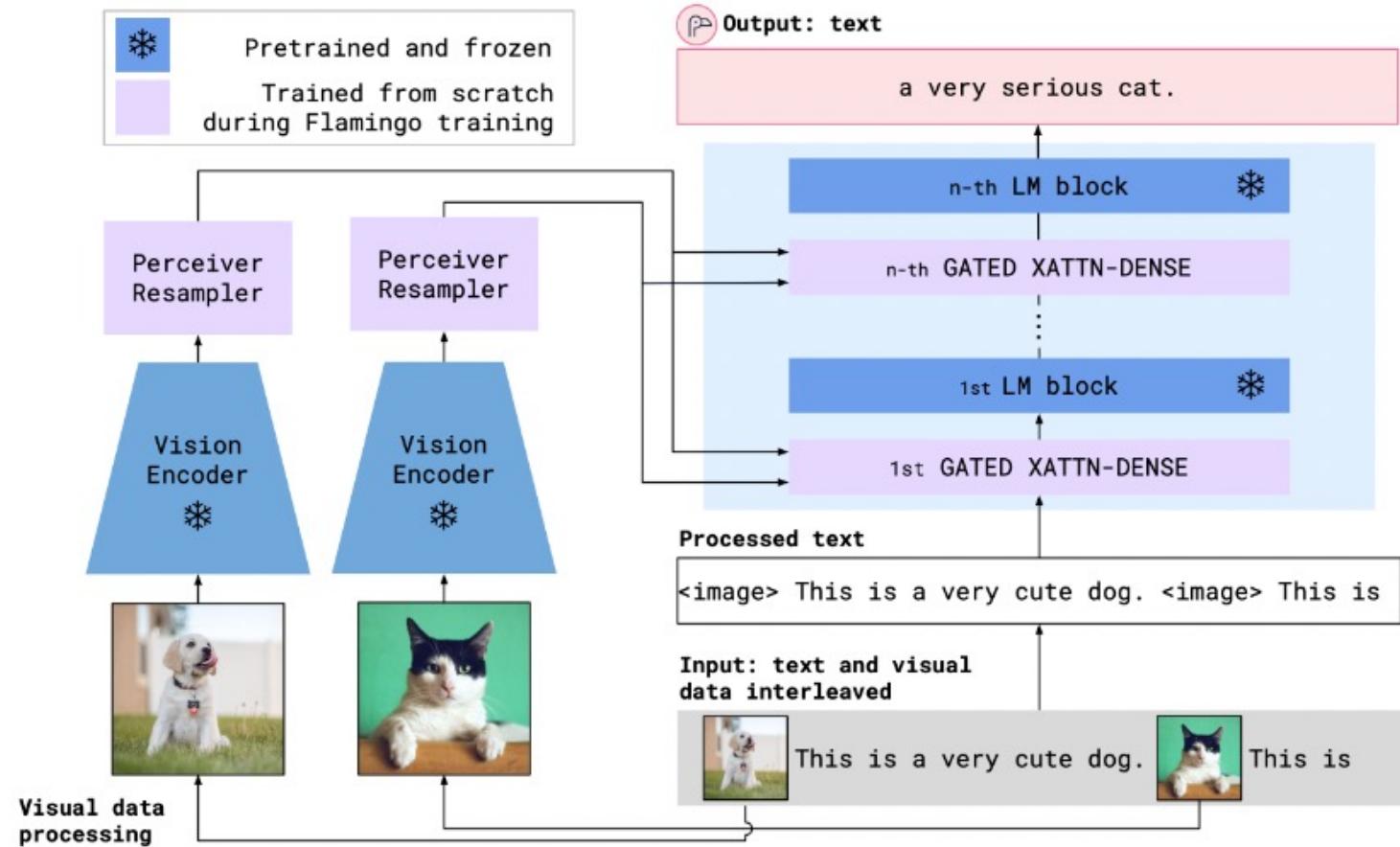
Method	CHAIR _I ↓	CHAIR _S ↓	Len
MiniGPT-4	9.2	31.5	116.2
mPLUG-Owl	30.2	76.8	98.5
LLaVA	18.8	62.7	90.7
MultiModal-GPT	18.2	36.2	45.7
MiniGPT-v2 (long)	8.7	25.3	56.5
MiniGPT-v2 (grounded)	7.6	12.5	18.9
MiniGPT-v2 (short)	4.4	7.1	10.3

Hallucination results

[1] Chen J, Zhu D, Shen X, Li X, Liu Z, Zhang P, Krishnamoorthi R, Chandra V, Xiong Y, Elhoseiny M. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint arXiv:2310.09478. 2023 Oct 14.

Openflamingo^[1]: open source version for DeepMind flamingo^[2]

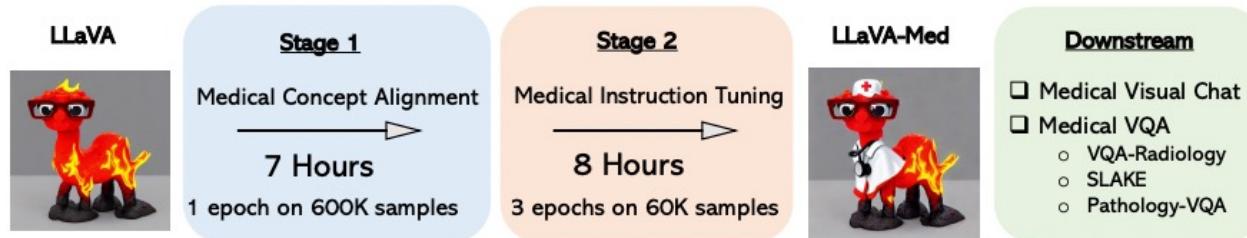
- Vision encoder: CLIP
- Language models: LLaMA
- **Perceiver resampler** : from varying-size large feature maps to few visual tokens
- **Gated cross-attention**



[1] Awadalla A, Gao I, Gardner J, Hessel J, Hanafy Y, Zhu W, Marathe K, Bitton Y, Gadre S, Sagawa S, Jitsev J. Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390. 2023 Aug 2.

[2] Alayrac JB, Donahue J, Luc P, Miech A, Barr I, Hasson Y, Lenc K, Mensch A, Millican K, Reynolds M, Ring R. Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems. 2022 Dec 6;35:23716-36.

LLaVA-Med^[1]: Biomedical visual instruction prompting



Stage 1: Biomedical Concept Feature Alignment.

only update the projection matrix

Filter 600K image-text pairs from PMC-15M converted to instruction-following data

Stage 2: End-to-End Instruction-Tuning.

update both projection layer and LM

Fine-tuning to Downstream Datasets.

fine-tune LLaVA-Med after the two-stage training on three biomedical VQA datasets

Dataset used: PMC-15M(stage 1); VQA-RAD; SLAKE; PathVQA

Biomedical Visual Instruction-Following Example

Caption:

Figure 4: Contrast-enhanced CT scan of the chest for patient #1. A large cavitating lesion was identified in the right upper lobe with pleural involvement and likely extension into the chest wall. In addition, right hilar lymphadenopathy was suspected. This was the primary malignant lesion in this patient.



Sentences with in-line figure mentions:

- Computerized tomography (CT) scans of the chest (Figure 4) were obtained.
- (Figure 4) and demonstrated a large cavitating lesion in the posterior aspect of the right upper lobe with probable pleural involvement likely representing the primary lesion.

Conversation:

User: Where is the lesion located?

Assistant: The lesion is located in the posterior aspect of the right upper lobe of the lung.

User: Is it extending into the chest wall?

Assistant: Yes, the CT scan suggests that the large cavitating lesion in the right upper lobe is likely extending into the chest wall.

User: Can you explain more about the cavitating lesion?

Assistant: Certainly. A cavitating lesion refers to an abnormal mass or growth within the lung tissue that has a cavity or hollow space inside it. This cavity can be formed due to the necrosis (death) of the tissue inside the lesion, leading to the breakdown and disintegration of the affected tissue. Cavitating lesions can be caused by various factors, such as infections (like tuberculosis or fungal infections), autoimmune diseases, or malignancies (cancer). In the context of the CT scan provided, the large cavitating lesion in the right upper lobe of the lung is suggestive of a malignant (cancerous) growth. The cavity within the lesion may be a result of the rapid and uncontrolled growth of cancer cells, which can lead to the destruction of surrounding healthy tissue. It's important to note that further diagnostic tests, such as a biopsy, would be required to confirm the malignancy and determine the specific type of cancer involved.

Med-flamingo^[1]

Build on openflamingo-9B

- LLaMA
- CLIP ViT

Dataset used for pretraining:

- MTB
- PMC-OA

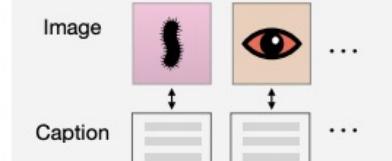
Dataset used for evaluation:

- VQA-RAD
- Path-VQA
- Visual USMLE

Focused on few-shot inference

1. Multimodal pre-training on medical literature

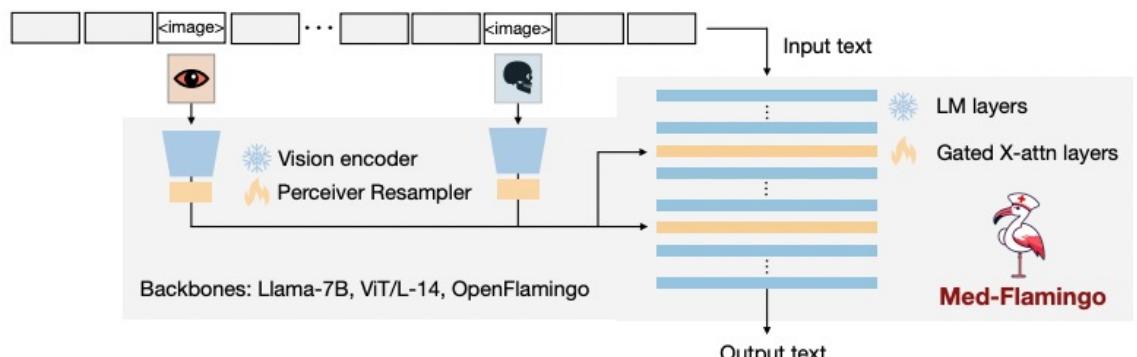
Paired data



Interleaved data

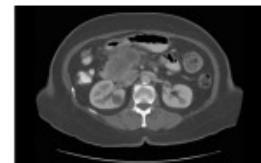


Tokenized data



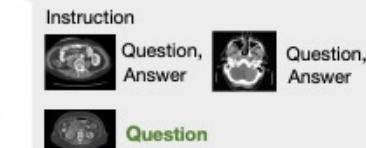
2. Few-shot generative VQA

Problem

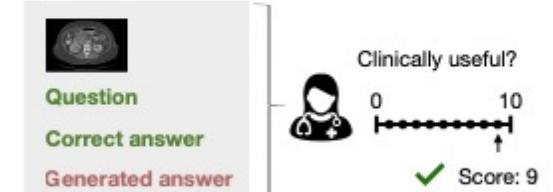


Question: What do the small white lesions in the aorta mean?

Multimodal few-shot prompt



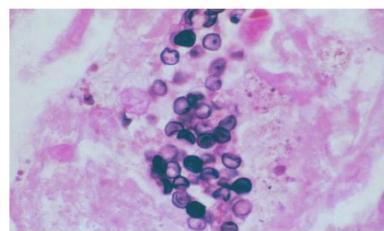
3. Human evaluation



[1] Moor M, Huang Q, Wu S, Yasunaga M, Zakka C, Dalmia Y, Reis EP, Rajpurkar P, Leskovec J. Med-flamingo: a multimodal medical few-shot learner. arXiv preprint arXiv:2307.15189. 2023 Jul 27.

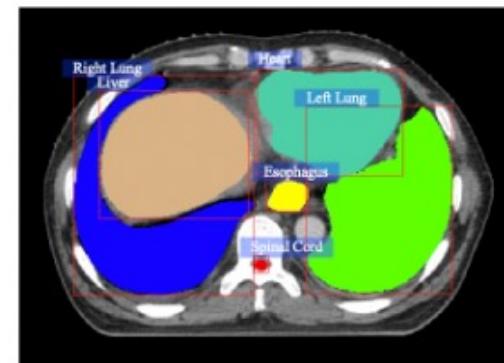
Some VQA datasets

Dataset	Modality	Source	Images	QA pairs
VQA-RAD [18]	Radiology	MedPix® database	0.3k	3.5k
PathVQA [12]	Pathology	PEIR Digital Library [14]	5k	32.8k
SLAKE [23]	Radiology	MSD [3], ChestX-ray8 [36], CHAOS [15]	0.7k	14k



PathVQA

Modality Q: Is this an MRI? A: no	Plane Q: Is this an axial image? A: yes	Organ System Q: What is the organ system? A: Gastrointestinal
Abnormality Q: Which organ is affected? A: pancreas	Color Q: Is the lesion more or less dense than the liver? A: less dense	
Object/Condition Presence Q: Is there gastric fullness? A: yes	Size Q: What is dilated? A: duodenum	
Positional Q: What is the location of the mass? A: head of the pancreas	Attribute (other) Q: Is the mass well circumscribed? A: No	
Counting Q: How many masses are there? A: yes	Other Q: How would you measure the length of the kidney A: unanswerable	



Knowledge-based:

- (En) What is the **function** of the **rightmost organ** in this picture?
(Zh) 图中是否有**器官**属于**呼吸系统**?
(Are there **organs** in this image belonging to the **respiratory system**?)

Vision-only:

- (En) Does the image contain **left lung**?
(Zh) 这张图片是关于**腹部**吗?
(Is this image about the **abdomen**?)

SLAKE

VQA-RAD

PMC-VQA^[1]

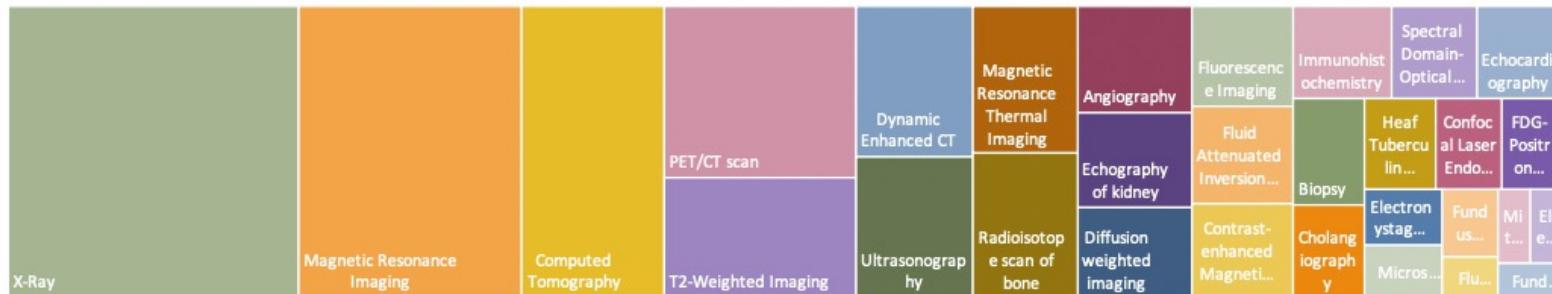


Figure 1: The top 20 figure types in PMC-VQA, cover a wide range of diagnostic procedures.

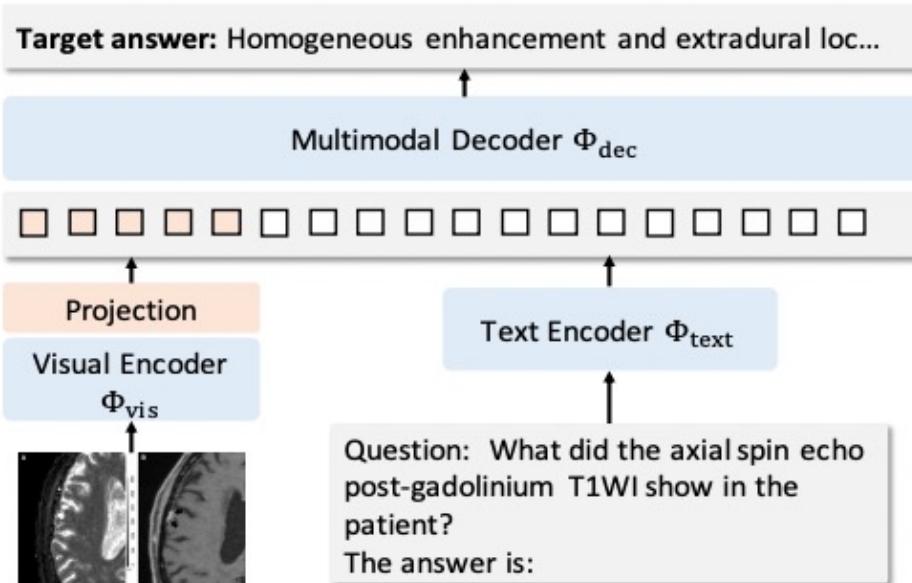
Dataset	Modality	Source	Images	QA pairs
VQA-RAD [18]	Radiology	MedPix® database	0.3k	3.5k
PathVQA [12]	Pathology	PEIR Digital Library [14]	5k	32.8k
SLAKE [23]	Radiology	MSD [3], ChestX-ray8 [36], CHAOS [15]	0.7k	14k
VQA-Med-2021 [5]	Radiology	MedPix® database	5k	5k
PMC-VQA	Mixture*	PubMed Central®	149k	227k

* Mixture: Radiology, Pathology, Microscopy, Signals, Generic biomedical illustrations, etc.

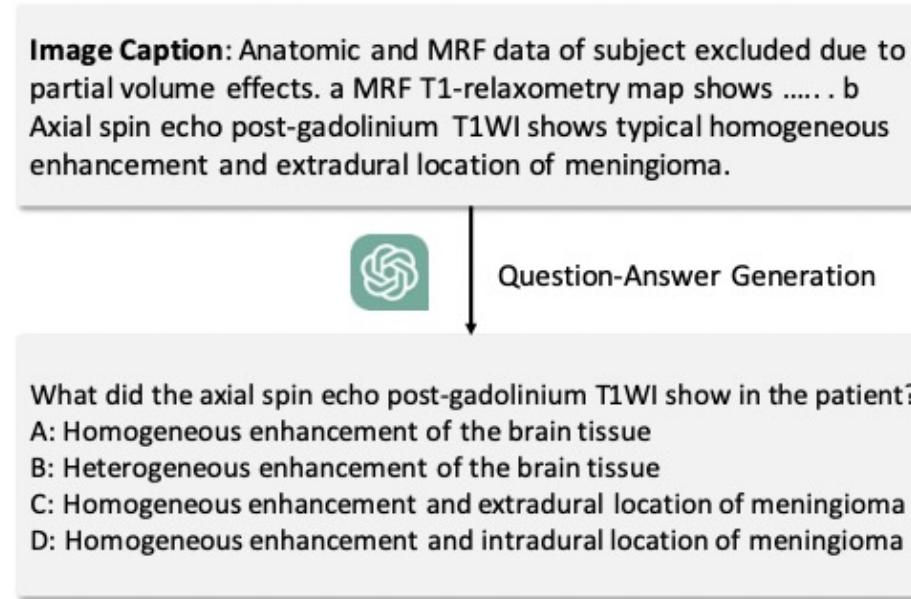
Created a new dataset PMC-VQA: contains 227k VQA pairs of 149k images

[1] Zhang X, Wu C, Zhao Z, Lin W, Zhang Y, Wang Y, Xie W. Pmc-vqa: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415. 2023 May 17.

PMC-VQA^[1]



(a) Overall architecture of MedVInT



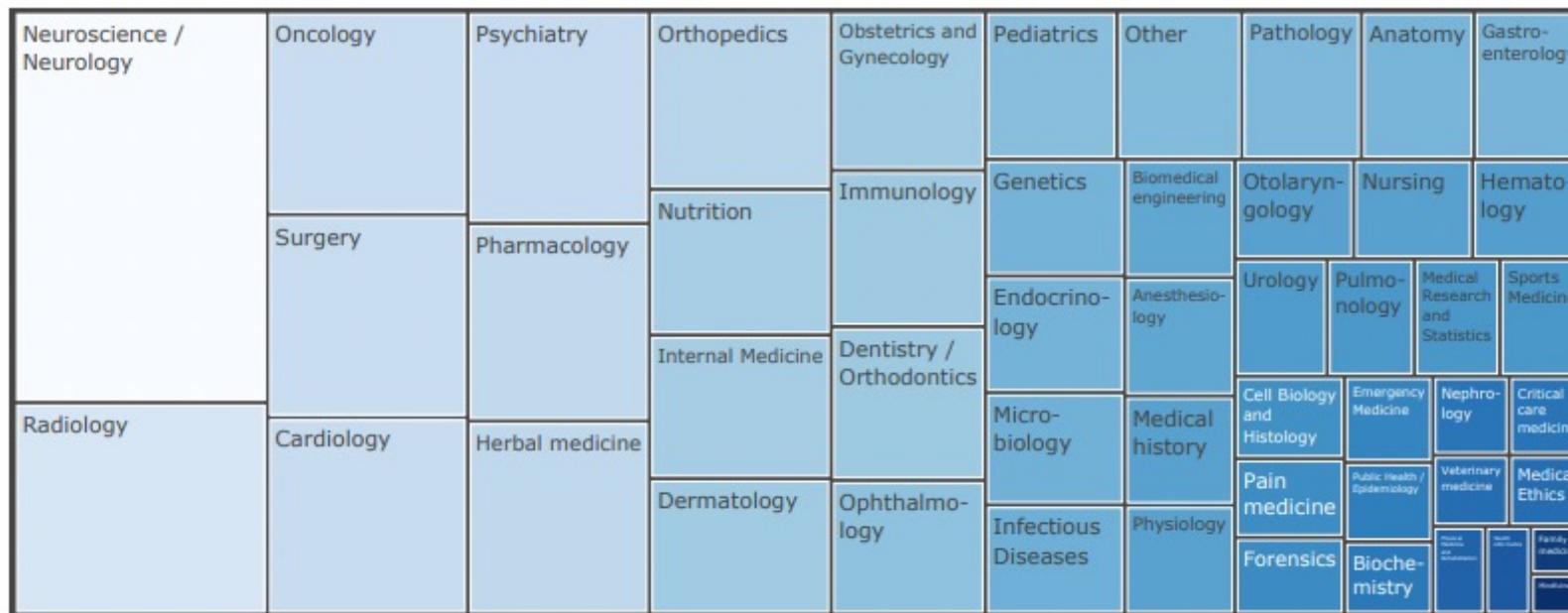
(b) Pipeline for PMC-VQA generation

Proposed medical visual instruction tuning model: MedVInT

[1] Zhang X, Wu C, Zhao Z, Lin W, Zhang Y, Wang Y, Xie W. Pmc-vqa: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415. 2023 May 17.

Med-flamingo^[1]

1. MTB data: 4721 textbooks from medical specialties



2. PMC-OA: 1.6M image-caption pairs from PubMedCentral

[1] Moor M, Huang Q, Wu S, Yasunaga M, Zakka C, Dalmia Y, Reis EP, Rajpurkar P, Leskovec J. Med-flamingo: a multimodal medical few-shot learner. arXiv preprint arXiv:2307.15189. 2023 Jul 27.

PMC-15M^[1]



Articles	Avg. figures per article	Avg. figure size (px)	Avg. caption length (token)	Figure-caption pairs	Training
3,298,780	4.6	582×702	110	15,282,336	13.9M

Table 1: PMC-15M statistics.

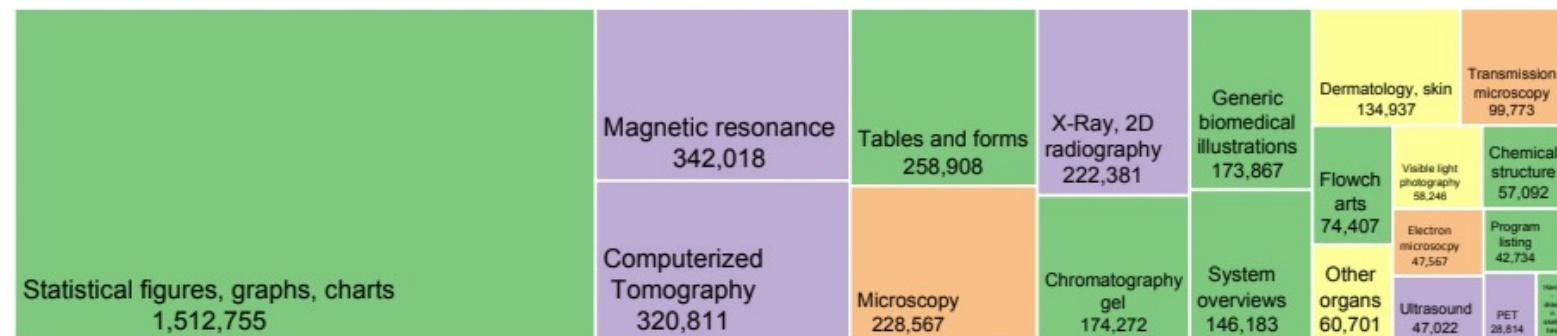
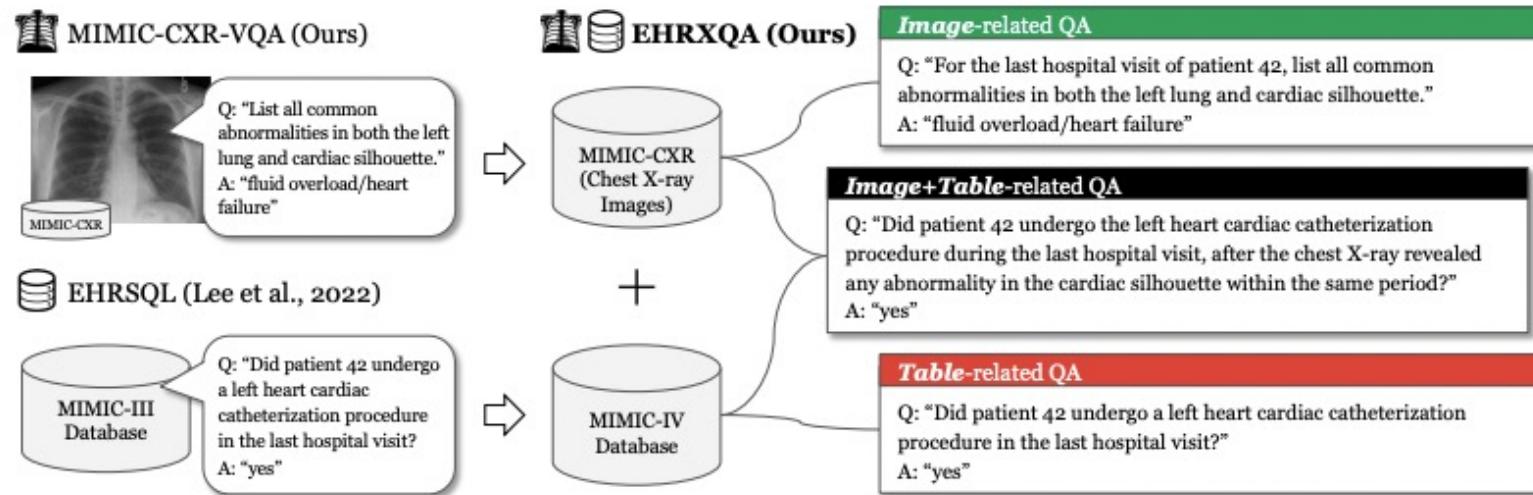


Figure 4: Estimate of PMC-15M figure type distribution based on the image taxonomy in García Seco de Herrera et al. (2015). Block size is proportional to keyword frequency for the corresponding image class.

[1] Zhang S, Xu Y, Usuyama N, Bagga J, Tinn R, Preston S, Rao R, Wei M, Valluri N, Wong C, Lungren MP. Large-scale domain-specific pretraining for biomedical vision-language processing. arXiv preprint arXiv:2303.00915. 2023 Mar 2.

MIMIC-CXR-VQA^[1]



- **complex (involving set and logical operations)**
- **diverse (with 48 templates)**
- **and large-scale (approximately 377K)**

[1] Bae S, Kyung D, Ryu J, Cho E, Lee G, Kweon S, Oh J, Ji L, Chang EI, Kim T, Choi E. EHRXQA: A Multi-Modal Question Answering Dataset for Electronic Health Records with Chest X-ray Images. arXiv preprint arXiv:2310.18652. 2023 Oct 28.