A. The Entity Description Base and Position Set

We leverage the rad-graph NER extraction results provided by [9] and further add extra descriptions. Tab. 1 shows the descriptions we used to translate different entities. We have kept 75 entities in entity query set Q, following [9], which covers 90% entities in reports. "Tail_abnorm_obs" entity represents some tailed entities and "exluded_obs" represents some entities useless for diagnosis. The last "covid-19" row is only referred for inference since it never appear in pre-train reports.

Table 1: The Entity description used for translate single entity name. The description can be easily found from website.

Entity	Description
normal	It means the absence of diseases and infirmity, indicating the structure is normal.
clear	The lungs are clear and normal. No evidence for other diseases on lung.
sharp	This means that an anatomical structure s boundary or edge is clear and normal, meaning it is free of diseases.
sharply	"Sharply seen means that an anatomical structure is clearly visible.
unremarkable	This represents some anatomical structures are normal, usually modifying cardiac and mediastinal silhouettes.
intact	The bonny structure is complete and normal, meaning no fractures.
stable	The modified anatomical structures are normal and stable. No evidence for diseases.
free	It usually refers to free air and is associate with pneumothorax, atelectasis, pneumoperitoneum and emphysema.
effusion	A pleural effusion is accumulation of excessive fluid in the pleural space, the potential space that surrounds each lung. A pleural effusion infiltrates the space between the visceral pleura and the parietal pleura.
opacity	It is defined as an area of hazy opacification due to air displacement by fluid, airway collapse, fibrosis, or a neoplastic process. It is causes include infections, interstitial lung disease, and pulmonary edema.
pneumothorax	A pneumothorax is an abnormal collection of air in the pleural space between the lung and the chest wall. It may be caused by pneumonia or fibrosis and other diseases.
edema	Pulmonary edema, also known as pulmonary congestion, is excessive liquid accumulation in the tissue and air spaces of the lungs. It will show fluid in the alveolar walls.
atelectasis	It is the collapse or closure of a lung resulting in reduced or absent gas exchange. Findings can include lung opacification and loss of lung volume.
tube	It is a surgical drain that is inserted through the chest wall and into the pleural space or the mediastinum to remove undesired substances such as air.
consolidation	It is a region of normally compressible lung tissue that has filled with liquid instead of air. Consolidation must be present to diagnose pneumonia: the signs of lobar pneumonia are characteristic and clinically referred to as consolidation.
process	Acute process means there is abnormality in the anotomy structure.
abnormality	It means the exist of diseases and infirmity, indicating the structure is abnormal.
enlarge	It usually modifies cardiac silhouette and heart. Cardiomegaly is a medical condition in which the heart is enlarged.
tip	It refers to the top head of the tube.
low	The presence of low lung volumes may be a sign of a restrictive lung condition such as pulmonary fibrosis or sarcoidosis.
pneumonia	Pneumonia is an inflammatory condition of the lung primarily small air sacs known as alveoli. Pneumonia may present with opacities. Complications such as pleural effusion may also be found increasing the diagnostic accuracy of lung consolidation and pleural effusion
line	It refers to venous access line ot PICC lines.
congestion	Pulmonary congestion is defined as accumulation of fluid in the lungs, resulting in impaired gas exchange and arterial hypoxemia.
catheter	catheter is a tube placed in the body to drain and collect urine from the bladder
cardiomegaly	Cardiomegaly (sometimes megacardia or megalocardia) is a medical condition in which the heart is enlarged.
fracture	Fracture is a break in a rib bone.
air	It refers to the free air or gas in pleural space, indicating pneumothorax. Air displacement by fluid may lead to opacity.
tortuous	The Aorta is slightly tortuous. Sometimes it may refer to varicose veins
lead	It refers to the leading head of the tube.
disease	It means the exist of diseases and abnormalty, indicating the structure is abnormal.
calcification	Pulmonary calcification is a common asymptomatic finding. Pulmonary calcifications are caused mainly by two mechanisms: the dystrophic form and the metastatic form
prominence	It means the exist of some observation.
device	It refer to some equipments like picc tub, valve catheter, pacemaker hardware, arthroplastmarker icd defib, device support equipment and mediport
engorgement	Pulmonary vascular engorgement means obstruction of the normal flux of blood within the blood vessel network of the lung resulting in engorgement of pulmonary vessels
picc	A peripherally inserted central catheter (PICC), also called a PICC line, is a long, thin tube that s inserted through a vein in your arm and passed through to the larger veins near your heart.
clip	Surgical clips or vascular clips usually represent the one kind of medical equipments.
elevation	If tissues or anatomical structures are elevated, they are raised up higher than the normal location.
expand	It means the lungs are normally expanded and clear, indicating the absence of pneumothorax.
nodule	A lung nodule or pulmonary nodule is a relatively small focal density in the lung. it may be confused with the projection of a structure of the chest wall or skin, such as a nipple, a healing rib fracture or lung cancer.
wire	Sternotomis wires means the center line of the chest.
fluid	It refers to the water of liquid in the lung and it may indicate edema and other diseases.
degenerative	Degenerative disease is the result of a continuous process based on degenerative cell changes
pacemaker	pacemaker device usually represents the one kind of medical equipments.

Entity	Description
thicken	Pleural thickening is an increase in the bulkiness of one or both of the pulmonary pleurae. It may cause by pulmonary Infection, empyema, tuberculosis or lung cancer.
marking	It represents interstitial markings or bronchovascular markings
scar	A scar (or scar tissue) is an area of fibrous tissue that replaces normal tissues after an injury.
hyperinflate	Hyperinflated lungs are larger-than-normal lungs as a result of trapped air.
blunt	Blunting of the costophrenic angles is usually caused by a pleural effusion, as already discussed. Other causes of costophrenic angle blunting include lung disease in the region of the costophrenic angle, and lung hyperexpansion.
loss	The etiology of lung volume loss can be listed as follow: airway obstruction or compression, obesity, scoliosis, restrictive diseases such as pulmonary fibrosis and interstitial lung disease, tuberculosis.
widen	The mediastinum is not widened or enlarged.
collapse	Collapse lung refers to pneumothorax or atelectasis.
density	The density (more precisely, the volumetric mass density; also known as specific mass), of a substance is its mass per unit volume.
emphysema	Emphysema, or pulmonary emphysema, is a lower respiratory tract disease, characterized by air-filled spaces (pneumatosis) in the lungs, that can vary in size and may be very large.
aerate	Aeration (also called aerification or aeriation) is the process by which air is circulated through, mixed with or dissolved in a liquid or other substances that act as a fluid (such as soil).
mass	A lung mass is an abnormal growth or area in the lungs and it can also view as lung cancer.
crowd	Crowding of the bronchovascular structures is an important direct sign of volume loss. The atelectatic lung enhances densely after contrast administration because of closeness of the pulmonary arteries and arterioles within the collapsed lobe.
infiltrate	A pulmonary infiltrate is a substance denser than air, such as pus, blood, or protein, which lingers within the parenchyma of the lungs. Pulmonary infiltrates are associated with pneumonia, tuberculosis and sarcoidosis.
obscure	Some anatomy structures are not clear and is difficult to understand or see.
deformity	It means some body parts are abnormal or unjuried.
hernia	Lung hernia (Sibson hernia) is a protrusion of lung outside of thoracic wall. the hernia is noted after chest trauma, thoracic surgery or certain pulmonary diseases.
drainage	Tube drainage represents the one kind of medical equipment.
distention	Distension generally refers to an enlargement, dilation, or ballooning effect. It may refer to: Abdominal distension.
shift	The mediastinal shift is the deviation of the mediastinal structures towards one side of the chest cavity, usually seen on chest radiograph. It indicates a severe asymmetry of intrathoracic pressures.
stent	Tracheal stent represents the one kind of medical equipments
pressure	Pulmonary venous pressure is intermediate between mean PAP and LAP over all physiologic pressures
lesion	Lung nodules, pulmonary nodules, white spots, lesions—these terms all describe the same phenomenon: an abnormality in the lungs.
finding	Some observation on body parts, usually indicating abnormalty.
borderline	Borderline size of the cardiac silhouette means the cardiac silhouette is not enlarged and normal.
hardware	It represents the one kind of medical equipments.
dilation	The state of being larger or more open than normal
chf	Heart failure — sometimes known as congestive heart failure — occurs when the heart muscle doesn't pump blood as well as it should. When this happens, blood often backs up and fluid can build up in the lungs, causing shortness of breath.
redistribution	If the pulmonary edema is due to heart failure or fluid overload, you may also see cardiomegaly and distension of the pulmonary veins, particularly in the upper lung fields.
aspiration	Aspiration pneumonia occurs when food or liquid is breathed into the airways or lungs, instead of being swallowed.
tail_abnorm_obs	Some very rare diseases.
excluded_obs	Some meaningless observations.
covid-19	It is a contagious disease caused by a virus. Ground-glass opacities, consolidation, thickening, pleural effusions commonly appear in infection.

Additionally, we keep 51 positive positions, following [9], to form the position set P, as {*trachea, left_hilar, right_hilar, hilar_unspec, left_pleural, right_pleural, pleural_unspec, heart_size, heart_border, left_diaphragm, right_diaphragm, diaphragm_unspec, retrocardiac, lower_left_lobe, upper_left_lobe, lower_right_lobe middle_right_lobe, upper_right_lobe, left_lower_lung, left_mid_lung, left_upper_lung left_apical_lung, left_lung_unspec, right_lower_lung, right_mid_lung, right_upper_lung right_apical_lung, right_lung_unspec, lung_apices, lung_bases, left_costophrenic right_costophrenic, costophrenic_unspec, cardiophrenic_sulcus, mediastinal, spine clavicle, rib, stomach, right_atrium, right_ventricle, aorta, svc, interstitium, parenchymal, cavoatrial_junction, cardiopulmonary, pulmonary, lung_volumes, unspecified, other*}. "Other" is used to reprepresnt some tailed positions.

B. Implementation Details

Model architecture. As input to the model, images are resized into $224 \times 224 \times 3$. We use the first four layers of ResNet50 [4] as our visual backbone (Φ_{visual}), and adopt a MLP layer to transform the output feature dimension into $d = 256$. As a result, the output feature maps from visual encoder is $\mathcal{V} \in \mathbb{R}^{14 \times 14 \times 256}$. On the report side, we extract the triplets with a pre-trained NER module, as described in [6], and compute the entity and position embedding with a pre-trained ClinicalBERT [1], its

default embedding dim is $d' = 768$. We obtain $|Q| = 75$ entities and $|P| = 51$ positions that most frequently appear in the reports, following [9]. We sample $M = 7$ negative positions for each entity to calculate contrastive loss. In the fusion module, We adopt 4 Transformer Decoder layers with 4 heads.

Pre-training. At this stage, both the pre-process operation and language encoding use pre-trained networks, while the visual encoder and fusion module are trained end-to-end. We use AdamW [7] optimizer with $lr = 1\times10^{-4}$ and $lr_{\text{warm up}} = 1\times10^{-5}$. We train on a GeForce RTX 3090 GPU with batch size 32 for 60 epochs. First 5 epochs are set for warming up.

Fine-tuning. For the downstream tasks, with large amount of training data, we can fine-tune the model end-to-end, with our pre-trained visual backbone as initialization. Specifically, for image classification task, we adopt ResNet50 [4] and initialize its first four layers with our pre-trained visual encoder. For image segmentation task, we use ResUNet [3] as backbone and initialize its encoder with our pre-trained image encoder.

C. Ablation Study

Our final method mainly contains three key parts, transformer-based entity-query fusion module, position location contrastive loss (PosCL), and Entity Description (ED) encoder. We gradually remove the modules to analyze their effectiveness. "w/o (ED)" refers to removing descriptions and "w/o (PosCL + ED)" refers to only maintaining the fusion module with basic CE loss. We cannot further dismiss the transformer-based entity-query fusion module as it is the most basic module to support our pre-training. Tab. 2 and Tab. 3 shows the quantitative results.

Entity-query Fusion module. The lines about "w/o (PosCL + ED)" in tables demonstrate the performance of the basic model modified only by base entity existence CE loss. This model can exceed many former methods. This proves our assumption that the complex syntax will hurt the network to capture the useful entities significantly and our pre-process operation and entity-level supervision can greatly relieve the problem.

Position Contrastive Loss. The PosCL can significantly help the network to ground the abnormalities. As shown in the results by adding PosCL the classification results can be further improved, e.g., from 0.75 to 0.76 on AUC in ChestX-ray14 dataset. Besides classification, location contrastive loss can bring more gain in grounding. These results show position is another vital element in reports especially for grounding tasks. Our extracted triplets can conclude and clean the reports with little information loss and make the network learn the report information more straightforward.

Entity Description Encoder. By translate entities into description, we want to realize two goals. *First*, in addition to just learning from the image-report data, the network can actively learn the relationship between different entities based on the entity descriptions. As shown in tables, adding descriptions in most scenarios can help the network better understand the entity and bring gain to the final metric scores. *Second*, more importantly, the entity translation enables our model to **handle openset new diseases**. If excluding entity descriptions and prompting the entities only with their names as former works, the performance of our method will drop significant when facing unseen diseases which is discussed in zero-shot classification for Covid-19 at main body.

Dataset	RSNA Pneumonia			SIIM-ACR Pneumothorax			ChestX-ray14		
Methods	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑
w/o (PosCL + ED)	0.8532	0.6079	0.7669	0.8768	0.6672	0.8187	0.7502	0.2374	0.8541
w/o (ED)	0.8537	0.6241	**0.8146**	**0.9017**	**0.7008**	**0.8584**	0.7621	0.2452	0.8606
Ours	**0.8694**	**0.6342**	0.8002	0.8924	0.6833	0.8428	**0.7676**	**0.2525**	**0.8619**

Table 2: Ablation study on zero-shot classification task. AUC, F1 and ACC scores are reported. For ChestX-ray 14, the metrics all refer to the macro average on the 14 diseases.

Methods	Pointing Game↑	Recall↑	Precision↑	IoU↑	Dice↑
w/o (PosCL + ED)	0.7979	**0.8961**	0.4036	0.2783	0.4230
w/o (ED)	0.8424	0.8226	**0.6520**	0.3118	0.4610
Ours	**0.8721**	0.8661	0.6420	**0.3172**	**0.4649**

(a) Zero-shot grounding on Pneumonia

Methods	Pointing Game↑	Recall↑	Precision↑
w/o (PosCL + ED)	0.1786	0.3151	0.1336
w/o (ED)	**0.2080**	0.3178	0.1711
Ours	0.1975	**0.3562**	**0.1940**

(b) Zero-shot grounding on Pneumothorax

Table 3: Ablation study on zero-shot grounding tasks. (a) shows the results on RSNA Pneumonia dataset. (b) shows the results on SIIM-ACR Pneumothorax dataset.

D. Detailed results on ChestX-ray14

We further show the detailed performance of 14 different diseases on ChestX-ray14 dataset. Tab. 4 shows the results on the zero-shot setting. Our method can exceed the former methods for most diseases. The radar Fig. 1Y shows more visually how our model compares with other solutions under the zero-shot setting. Our method can exceed the former methods for most diseases. Under 100% fine-tuning settings, we achieved similarly excellent results as shown in Tab. 5.

Methods	Ate.	Car.	Eff.	Inf.	Mas.	Nod.	Pna.	Pnx.	Con.	Ede.	Emp.	Fib.	Thi.	Her.	AVG
ConVIRT [10]	0.4533	0.4601	0.7262	0.6238	0.6790	0.6322	0.6097	0.6698	0.6855	0.7699	0.4701	0.5293	0.6098	0.6220	0.6101
GLoRIA [5]	0.6680	0.7647	0.7975	0.6159	0.6722	0.5293	0.6755	0.4785	0.7306	0.8212	0.6033	0.5104	**0.6721**	0.7144	0.6610
BioViL [2]	0.5026	0.6328	0.7914	0.5791	0.7029	0.6126	0.6866	0.7516	0.7455	0.8533	0.7136	0.6751	0.6560	0.7692	0.6909
CheXzero [8]	0.7426	0.7956	0.8415	0.6223	0.7095	0.6666	0.7263	0.7679	0.7866	0.8862	0.6451	0.6402	0.6134	0.7704	0.7296
w/o (PosCL + ED)	0.7131	0.8100	0.8635	**0.6361**	0.7776	0.6740	0.6903	0.8124	0.7915	0.8869	0.7480	**0.6780**	0.6429	0.7784	0.7502
w/o (ED)	0.7420	0.8270	**0.8663**	0.6336	0.7867	**0.6974**	**0.7238**	0.8310	0.8037	0.8887	0.7865	0.6715	0.5414	0.8691	0.7621
ours	**0.7506**	**0.8299**	0.8636	0.6280	**0.7885**	0.6947	0.7236	**0.8361**	**0.8079**	**0.8888**	**0.7950**	0.6511	0.5783	**0.9097**	**0.7676**

Table 4: Comparison with other state-of-the-art methods on zero-shot ChestX-ray 14 diseases classification task. For each disease, AUC score is reported and the macro average AUC score is also reported. We use the first three letters to represent one disease but for "pneumonia" and "pneumothorax" we use the first two and the last letters.

Methods	Ate.	Car.	Eff.	Inf.	Mas.	Nod.	Pna.	Pnx.	Con.	Ede.	Emp.	Fib.	Thi.	Her.	AVG
Scratch	0.7835	0.8116	0.8563	0.6537	0.7788	0.6912	0.7004	0.8561	0.8090	0.8869	0.8564	0.7534	0.7454	0.9106	0.7924
ConVIRT [10]	0.8012	0.8360	0.8511	0.6613	0.8004	0.7490	0.6998	0.8666	0.8079	0.9023	0.9014	0.7933	0.7468	0.9627	0.8128
GLoRIA [5]	0.8263	0.8326	0.8596	**0.6641**	0.8179	0.7348	0.7104	0.8452	0.8129	0.8977	**0.9310**	0.7886	0.7608	0.9750	0.8184
BioViL [2]	0.8185	0.8543	0.8607	0.6660	0.8302	0.7633	0.7090	0.8595	**0.8287**	0.9031	0.9251	0.7912	0.7638	0.9696	0.8245
ours	**0.8291**	**0.8594**	**0.8719**	0.6565	**0.8382**	**0.7647**	**0.7378**	**0.8807**	0.8275	**0.9083**	0.9224	**0.7977**	**0.7784**	**0.9796**	**0.8323**

Table 5: Comparison with other state-of-the-art methods on fine-tuning ChestX-ray 14 diseases classification task. For each disease, AUC score is reported and the macro average AUC score is also reported. We use the first three letters to represent one disease but for "pneumonia" and "pneumothorax" we use the first two and the last letters.

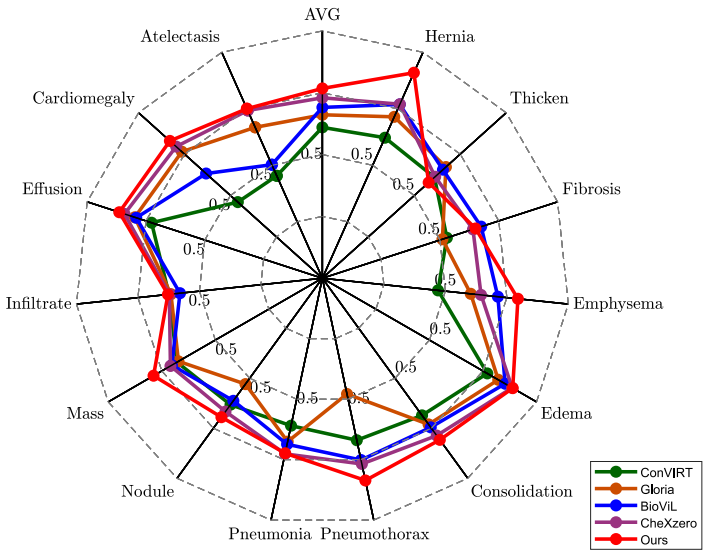

Figure 1: The radar figure of our method and other methods of ChestX-ray14 14 diseases. AUC scores are reported and, as shown, our method exceeds the previous state-of-the-art on most diseases.

E. Visualization Results

Fig. 2 shows visualization results of our model on zero-shot grounding task. As shown in figure, the ground truth of "Pneumonia" is given by bounding box and generally related to a large area region. Thus the metrics on this are higher than other two datasets. Our network captures its regions very well. For "Pneumothorax", its abnormality pattern is different from other diseases, which aim to capturing the collapsed part of the lung, rendering darker areas on the images rather than brighter opacity. Its ground-truth masks are generally thin and narrow while our network can still highlight its location. For "Covid-19", its image textual was similar to "Pneumonia", but since this is a totally new disease, grounding its regions is much more challenging. It requires the model to build relationships between them based on their complex definition and symptoms. The visualization results suggest that our model successfully achieve this, supporting that, for other unseen diseases, our model can also understand their complex descriptions.

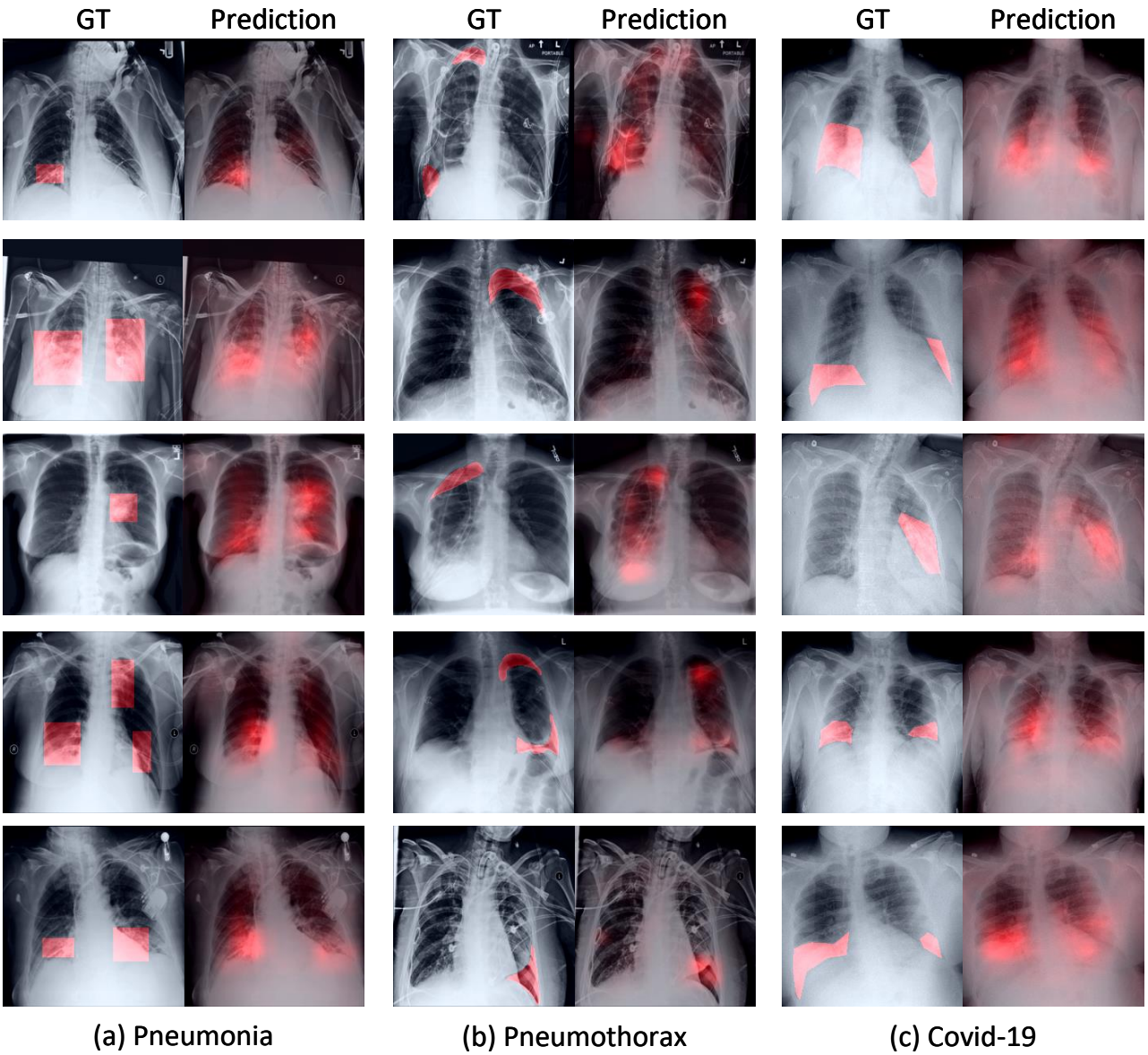

Figure 2: The visualization of zero-shot grounding results of our method. Each column represents the results on one disease and the left in it is the ground-truth and right is the heatmap predication of our model. The brighter the red on the figure, the more likely the model considering this region to be associated with abnormalities.

References

[1] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[2] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21, 2022. Official Implementation: `https://github.com/microsoft/hi-ml/tree/main/hi-ml-multimodal`.

[3] Foivos I Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951, 2021. Official Implementation: `https://github.com/marshuang80/gloria`.

[6] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, Curtis Langlotz, et al. Radgraph: Extracting clinical entities and relations from radiology reports. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[8] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, pages 1–8, 2022.

[9] Ke Yu, Shantanu Ghosh, Zhexiong Liu, Christopher Deible, and Kayhan Batmanghelich. Anatomy-guided weakly-supervised abnormality localization in chest x-rays. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 658–668. Springer, 2022.

[10] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare*, 2022. Highest Starred Implementation: `https://github.com/edreisMD/ConVIRT-pytorch`.