ChEX: Interactive Localization and Region Description in Chest X-rays

Philip Müller[1], Georgios Kaissis[1,2], and Daniel Rueckert[1,3]

[1] Technical University of Munich, Germany
[2] Helmholtz Munich, Germany
[3] Imperial College London, UK
{philip.j.mueller,g.kaissis,daniel.rueckert}@tum.de

Abstract. Report generation models offer fine-grained textual interpretations of medical images like chest X-rays, yet they often lack *interactivity* (*i.e.* the ability to steer the generation process through user queries) and *localized interpretability* (*i.e.* visually grounding their predictions), which we deem essential for future adoption in clinical practice. While there have been efforts to tackle these issues, they are either limited in their interactivity by not supporting textual queries or fail to also offer localized interpretability. Therefore, we propose a novel multitask architecture and training paradigm integrating textual prompts and bounding boxes for diverse aspects like anatomical regions and pathologies. We call this approach the *Chest X-Ray Explainer (ChEX)*. Evaluations across a heterogeneous set of 9 chest X-ray tasks, including localized image interpretation and report generation, showcase its competitiveness with SOTA models while additional analysis demonstrates ChEX's interactive capabilities. Code: https://github.com/philip-mueller/chex.

Keywords: Radiology Report Generation · Vision-Language Modeling

1 Introduction

The automatic interpretation of medical images, such as chest X-rays, holds immense promise for enhancing healthcare. Report generation models enable detailed textual interpretations beyond the capabilities of traditional image classification techniques alone. Despite significant progress in enhancing their accuracy [20], incorporating them in clinical practice remains a formidable challenge. A primary concern stems from the lack of transparency and interpretability surrounding the decision-making processes of these models, which poses obstacles for medical professionals seeking to validate their predictions, thereby hindering widespread adoption [11, 39]. Additionally, the non-interactive nature of most current models introduces risks in cases of inaccurate predictions, while interactive involvement in the generation process, *e.g.* through user prompts, may encourage the user to manually intervene in such cases.

Therefore, we advocate for two important aspects that can enhance the utility of these models in clinical practice: *(localized) interpretability and interactivity.*

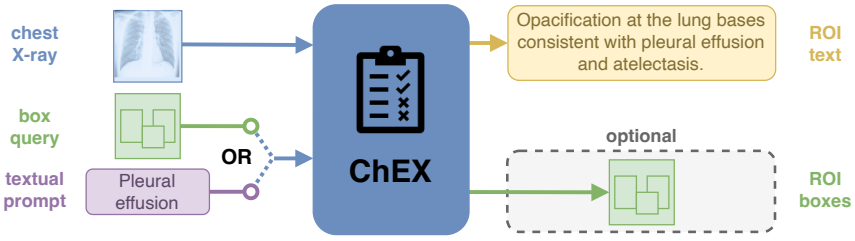


Fig. 1: Overview of ChEX. Given a chest X-ray and a user query, either as a textual prompt (*e.g.*, a pathology name, an anatomical region, or both) or as a bounding box, the model predicts a textual description of the queried region or aspect. For textual user prompts, it additionally predicts relevant bounding boxes. Thus, ChEX facilitates the interactive interpretation of chest X-rays while providing (localized) interpretability.

One pioneering effort in this direction is the work of Tanida *et al.* [59]. By incorporating bounding boxes of anatomical regions into the report generation process, their model offers enhanced interpretability. Furthermore, it supports bounding boxes as input queries, facilitating user interactivity. However, the model exhibits two major limitations: it lacks support for textual user prompts and focuses exclusively on anatomical regions during both training and inference. This narrow focus can lead to suboptimal predictions for other aspects, such as pathologies. In contrast, models like RaDialog [48], Med-PaLM M [61], and OmniFM-DR [72] support interactivity through textual prompts but do not provide bounding boxes for their textual answers or support them as queries.

In this work, we address these limitations by proposing a novel multitask architecture and training paradigm. Our approach integrates textual prompts and bounding boxes for various aspects, including anatomical regions, pathologies, and report sentences. We call this approach the *Chest X-Ray Explainer (ChEX)*. As illustrated in Figure 1, ChEX can be queried using textual prompts or user-defined bounding boxes and predicts individual descriptions for each specified region or aspect. For textual queries, the predictions are supplemented by bounding boxes to localize relevant regions for each answer. Therefore, ChEX offers a unique combination of interactivity – *reacting to user prompts* – and interpretability – *visually grounding the answers* – not provided by other report generation models. Furthermore, it additionally supports localized tasks beyond report generation (RG), including pathology (object) detection (OD), sentence grounding (SG), region classification (RC), and region explanation (RE).

Our contributions are as follows:

- We propose ChEX, an interactive and interpretable model for predicting visually grounded textual descriptions of chest X-rays based on user queries.
- We propose a multitask training paradigm, enabling ChEX to be jointly trained with diverse types of supervision from different datasets.
- We evaluate ChEX across 9 diverse chest X-ray tasks, spanning localized image interpretation and report generation functionalities. ChEX demonstrates competitive performance with specialized and general state-of-the-

art (SOTA) models despite being significantly smaller than some of these models, which have up to 80 times the size of ChEX. Notably, none of the baseline models is capable of performing all the tasks covered by ChEX.
– We conduct a thorough analysis of ChEX's interactive capabilities, demonstrating its proficiency in responding to specific user prompts.

2 Related Work

Medical Image-Text Models One prevalent category of medical image-text models adopts a CLIP-style [49] framework, leveraging contrastive image-text learning [3, 18, 30, 41, 55, 60, 62, 67, 80]. While most works focus on global image-text alignment, only some works consider more localized elements such as individual sentences or image patches [3, 18, 30, 41, 55, 62].

Report generation has recently received significant attention [13, 17, 20, 22, 28, 40, 46, 48, 58, 59, 61, 63, 66, 72]. Some of these approaches target specific aspects such as pathologies [17, 22, 58], anatomical regions [13, 59], or both [28]. However, support for interactive user queries remains rare among report generation models, with only the model RGRG [59] enabling queries via bounding boxes and a few others [48, 61, 72] supporting textual queries, albeit without the ability to jointly predict bounding boxes and textual descriptions. Notably, the model OmniFM-DR [72] supports the prediction of bounding boxes for textual prompts but without describing the content therein.

The ability of responding to textual prompts characterizes visual question answering (VQA) models, with some multitask models [26, 61, 72, 73] supporting zero-shot medical VQA, while others require fine-tuning [10, 16, 32, 57, 79]. However, unlike ChEX, these models lack localization capabilities for their responses. Moreover, it is important to note that while these methods rely on question-answer pairs for training, ChEX indirectly acquires its interactive capabilities through multitask training. Although localization tasks have been integrated into multitask training for natural images [64, 74, 78], such approaches remain scarce in the medical domain, with OmniFM-DR [72] being a notable exception.

Prompt-based Localization DETR [4] pioneered the use of token vectors for object detection, sparking a series of subsequent models following this approach [33, 38, 77, 83, 84]. However, these models typically employ tokens that are not input-dependent, often relying on learned or position-based tokens [33, 38, 77]. Visual grounding models [6, 7, 9, 29] predict bounding boxes for given textual phrases, relating them to image regions, and have been applied in medical imaging [5, 21]. GLIP [27, 78] unifies object detection and visual grounding using textual prompts, with applications emerging in the medical domain [15, 71]. Open-vocabulary object detection extends object detection models to unseen classes using textual prompts [14, 34, 37, 53, 70, 75, 76, 81, 82]. Similarly, prompt-based segmentation techniques have been explored [25] with applications in the medical domain [19, 36].

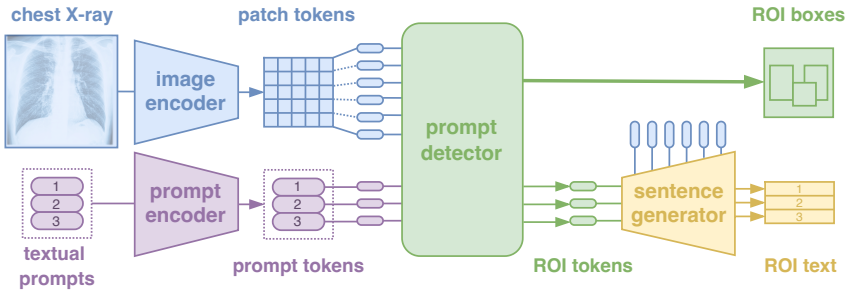


Fig. 2: Architecture of ChEX. The DETR-style prompt detector predicts bounding boxes and features for ROIs based on prompt tokens (textual prompts encoded by the prompt encoder) and patch features (from the image encoder). The sentence generator is then used to predict textual descriptions for each ROI independently.

3 ChEX: A Localized and Interactive Chest X-ray Description Model

Our model ChEX supports a wide range of tasks, spanning from localization to text generation, while considering different aspects like anatomical regions and pathologies. To facilitate efficient training and zero-shot inference across all tasks, ChEX employs a simple yet versatile architecture, outlined in Figure 2. First, an **image encoder** is used to extract patch tokens of the given chest X-ray, while each textual prompt is encoded using the **prompt encoder**, a frozen text encoder. Next, the **prompt detector**, a DETR [4]-style object detector, localizes the prompts within the image, predicting a set of bounding boxes for each prompt along with a single Region-of-interest (ROI) token vector per prompt. Finally, the **sentence generator** is conditioned on each ROI token independently as well as on all patches to predict a concise description of each ROI. Further details are given in the following Secs. 3.1 to 3.3. For reproducibility, we provide comprehensive implementation details in the supp. material.

3.1 Model Architecture

Image and Prompt Encoder We use CLIP [49] with ViT-B/32 [8] and the default text encoder, both pre-trained [60] on chest X-ray/report pairs from MIMIC-CXR [23]. Instead of using the `CLS` token, we use all patch tokens from the ViT and project them individually into the shared image-text space using the pre-trained projector from CLIP. Given Q textual prompts, each of them is embedded independently using the prompt encoder, *i.e.* the text encoder from CLIP, and projected to the shared image-text space. During training, we freeze the complete prompt encoder and the image encoder up to the 8th encoder layer.

Prompt Detector Using the Q prompt tokens from the output of the prompt encoder and the patch tokens from the image encoder, the prompt detector

independently localizes each prompt by predicting $M = 3$ bounding boxes per prompt, following the DETR [4] decoder with 6 layers. To support multiple boxes per prompt, we adopt [75] and additively combine each prompt token with M learned tokens, leading to a total of $Q \times M$ decoder query tokens. After decoding, we apply an MLP-based box predictor on each token feature to predict individual bounding boxes and box scores. We then use the bounding box parameters as a regional bias and compute box features by applying Gaussian ROI pooling [43] on the patch tokens. Additionally, we introduce a random skip connection to provide a direct path from the decoder layers. Finally, we compute the Q ROI tokens by aggregating the M box features per prompt using a weighted average based on box scores. When using bounding box queries instead of textual prompts, ROI features are directly computed with Gaussian ROI pooling on patch tokens using the given bounding box parameters and bypassing all decoder layers.

Sentence Generator For sentence generation, we use the GPT2-medium [50] model pre-trained on PubMed abstracts and condition it on each ROI token independently using P-tuning v2 [35] with MLP projection and without freezing parameters. To incorporate additional global context, we apply a *post decoder* comprising three transformer decoder layers, with the ROI tokens as queries and the patches as keys and values before feeding the features into the GPT2 model.

3.2 Multitask Training

We train ChEX in a multitask setting, where each sample provides one or more types of targets, including bounding boxes for pathologies or anatomical structures, pathology classification labels (per sample or per region), and report sentences (per sample or per region). To enable training with such a wide range of targets, we use three types of prompt tokens:

1. **Pathology tokens:** We define textual prompts for each pathology class (*e.g.*, "pleural effusion") and encode them using the prompt encoder.
2. **Anatomy tokens:** We define textual prompts for each anatomical region (*e.g.*, "right lung") and encode them using the prompt encoder.
3. **Sentence tokens:** We encode each individual sentence of the radiology report provided with the current sample.

For each sample, we only use the token types for which there are targets available. Using the encoded prompt tokens, the prompt detector predicts bounding boxes and ROI tokens for each of them, before the sentence generator (conditioned on the ROI tokens) predicts the target sentences.

ChEX does not only support textual queries, *i.e.* prompts, but also bounding box queries. Therefore, in some (randomly selected) batches, we train ChEX with query bounding boxes instead of prompt tokens. In such cases, we compute the ROI tokens based on the target bounding boxes directly using Gaussian ROI pooling, skipping the textual tokens and box prediction process.

Loss Functions We use the following loss functions to train our model:

1. **Bounding boxes:** We use a modified DETR loss [4], applying Hungarian matching independently per *pathology* or *anatomy* token until all predicted boxes are matched. We retain the L1 and gIoU losses, but omit the cross-entropy loss. Box scores are instead trained only for pathologies, using the focal loss [31] with positive targets for boxes matched in the first iteration.
2. **Pathology class labels:** We use an InfoNCE contrastive loss [47], pairing ROI tokens with textual pathology prompts according to the class labels. For example, if the pathology pleural effusion is present in the image (*pathology token*) or region (*anatomy token*), we build a positive pair with the prompt "pleural effusion", while negative pairs use the prompt "no pleural effusion" and prompts of non-present pathologies (*e.g.*, "pneumothorax", ...).
3. **Report sentences:** We apply autoregressive language modeling to predict region sentences (*anatomy token*) or to reconstruct the sentence (*sentence token*). For *sentence tokens*, we additionally use contrastive learning between ROI tokens and their corresponding textual sentence features. We also apply the CLIP loss [49] on average-pooled image patch and sentence features.

3.3 Zero-shot Inference

During inference, textual prompts like predefined pathology names, sentences, or user queries are encoded, and the prompt detector predicts bounding boxes and ROI tokens. These tokens are used by the sentence generator to predict descriptions of the detected regions. If regions require classification, such as for object detection or region classification, ROI tokens are classified based on cosine similarity with encoded prompts. When bounding boxes are given as queries, they're used for Gaussian ROI pooling, omitting other parts of the prompt detector and encoder if no textual prompts are provided. For full report generation, we use pre-defined sets of textual prompts and concatenate the predicted descriptions.

4 Experimental Setup and Evaluation

4.1 Training Dataset and Pre-processing

We train on the frontal chest X-rays from MIMIC-CXR [12, 23, 24] and VinDr-CXR [12, 44, 45]. MIMIC-CXR comes paired with radiology reports, of which we use the findings and impression sections and split them into individual sentences. We use additional supervision for the MIMIC-CXR images provided by the Chest ImaGenome (CIG) [12, 68, 69] dataset. It includes bounding boxes for 29 unique anatomical regions and additionally assigns report sentences as well as 53 unique findings and pathology labels to each of these regions. VinDr-CXR includes bounding boxes for 22 unique findings and pathologies. Overall, we train on 227,382 images from MIMIC-CXR and 15,000 images from VinDr-CXR, from their official train splits, but oversample VinDr-CXR samples to simulate equal size of both datasets. We randomly crop and resize all images to a resolution of 224×224 and then apply random horizontal flips, random affine transformations, contrast/brightness jittering, and random Gaussian blurring.

Table 1: Benchmark tasks with their datasets and evaluation metrics

Task Dataset	#Samples	#Classes	Evaluation Metrics
Sentence Grounding (SG): Predicting bounding boxes for given sentences			
MS-CXR [3]	169	none	mIoU, mAP
Pathology Detection (OD): Object detection of pathologies			
VinDrCXR [45]	1,500	top 15	mAP
NIH ChestXray (NIH8) [65]	448	8	mAP
MS-CXR [3]	169	8	mAP
Region Classification (RC): Classifying regions defined by given bounding boxes			
MS-CXR [3]	169	8	AUROC
Chest ImaGenome (CIG) [68]	3,402	53	weighted AUROC (wAUROC)
Region Explanation (RE): Predicting descriptions for regions defined by bounding boxes			
MS-CXR [3]	169	none	METEOR [1] Mic-F1-14†, Mac-F1-14†
Chest ImaGenome (CIG) [68]	3,402	none	METEOR [1] Mic-F1-14†, Mac-F1-14†
Full Report Generation (RG): Predicting full reports from chest X-rays			
MIMIC-CXR [23]	3,082	none	METEOR [1] Mic-F1-14†, Mac-F1-14†, Ex-F1-14† Mic-F1-5+†, Mac-F1-5+†

† Clinical efficacy (CE) metrics based on the CheXbert [56] classifier, micro-, macro-, and example-level-averaged over all 14 classes following Nicolson *et al.* [46] (Mic-F1-14, Mac-F1-14, Ex-F1-14), and averaged over 5 classes following Miura *et al.* [40] (Mic-F1-5+, Mac-F1-5+).

4.2 Benchmark Tasks

We evaluate the zero-shot performance of our model ChEX across the 9 chest X-ray tasks shown in Tab. 1. For details on the evaluation and dataset preparation, we refer to the supp. material. While no task-specific fine-tuning is applied, post-processing (*e.g.* box suppression and scaling of boxes) for ChEX and all baselines is adjusted to consider differences in annotation practices of datasets.

4.3 Benchmark Baselines

We consider a wide variety of specialist and generalist baselines in our benchmark. For SG, we consider the standard supervised visual grounding (SupVG) model TransVG [7] and the generative multitask model OmniFM-DR [72]. For OD and RC tasks, we compare against standard supervised object detection (SupOD) models including Faster R-CNN [54] and DETR-style models [4,38,83,84], and weakly-supervised object detection (WSupOD) including ADPD [42] and CheXNet [51], each of them with different (pre-trained) backbones. Additionally, we study the zero-shot capabilities of standard CLIP-style models for chest X-rays (BioVIL [3] and CheXzero [60]) on all SG, OD, and RC tasks. To the best of our knowledge, there is only one model supporting the RE tasks out-of-the-box, namely RGRG [59]. We also use the two CLIP-style baselines as image-text retrieval models and evaluate their performance on these tasks. For RG, we compare against several report generation models, including recent models like MAIRA-1 [20], Med-PaLM M [61], Prompt-MRG [22], and RaDialog [48].

5 Results

We compare ChEX with SOTA models (Sec. 5.1). Additionally, we study how ChEX reacts to (interactive) user queries (Secs. 5.2 and 5.3), show its interpretable and customizable report generation capabilities (Sec. 5.4), and study technical design choices (Sec. 5.5). ChEX is competitive with SOTA models while providing a high degree of interpretability and interactivity, therefore offering a promising path towards clinical application.

5.1 Comparison with SOTA

Fig. 3 and Tab. 2 provide an overview of the performance of our model ChEX compared to the best baselines, including specialized SOTA and common multitask models. On 8 of the 9 tasks, ChEX is competitive (within 1-std) or better than the best baseline on at least one metric. Only on pathology detection (OD) on VinDR-CXR is it outperformed by a specialized supervised object detection model. Note that none of the baselines is capable of performing all the tasks, as most of them are either focused solely on localization or generative tasks, but not both. Only the contrastive image-text (*i.e.* CLIP-style) models can perform a wide range of tasks but rely on retrieval for the generative tasks, thus showing poor performance on these tasks. Overall, our model ChEX shows excellent performance on a wide range of tasks, covering both localization as well as text generation, and is capable of replacing specialized models without major performance drops on most tasks.

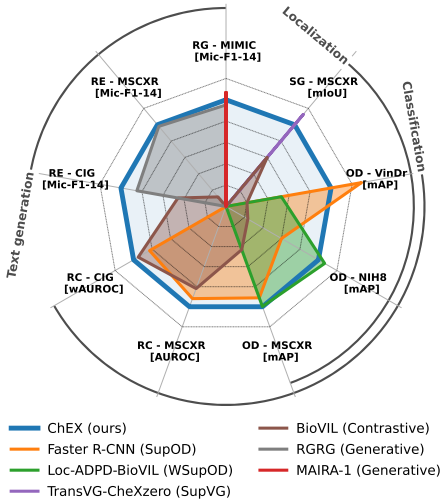


Fig. 3: Comparison of ChEX with specialized SOTA and common multitask models on 9 chest X-ray tasks, including sentence grounding (SG), pathology detection (OD), region classification (RC), region explanation (RE), and full report generation (RG). ChEX shows excellent performance on this wide range of tasks while none of the baselines is capable of even performing all of them. To improve readability, values are scaled relative to the results of ChEX.

Localization and Region Classification In *sentence grounding (SG)*, ChEX performs similarly to the generative model OmniFM-DR and the SupVG model TransVG (with TransVG showing an advantage in mIoU), despite TransVG being trained explicitly on this task. In *pathology detection (OD)*, ChEX is competitive on 2 out of 3 tasks. On VinDr-CXR, the SupOD model Faster R-CNN performs best, while on NIH8, ChEX almost doubles the performance of the

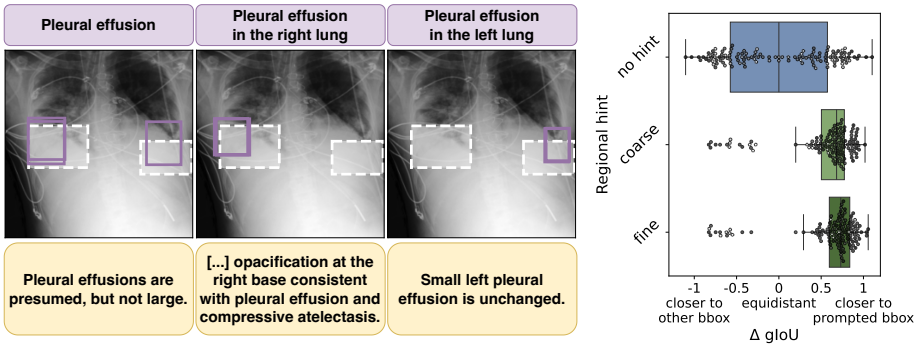
Table 2: Comparison of ChEX with the best-performing baselines for different types of models. On 8 of the 9 tasks, ChEX is competitive (within 1-std) or better than the best baseline on at least one metric, highlighting that ChEX is capable of replacing specialized models without major performance drops on most tasks. We indicate variability by std computed using bootstrapping and mark the best results as well as those within 1-std in bold. For detailed results, we refer to the supp. material.

Task	Metric	ChEX	SupOD	WSupOD	SupVG	Contrastive	Generative
Sentence Grounding (SG)							
MS-CXR	[mIoU]	47.52±1.45	–	–	**53.51**±1.53	28.57±1.31	46.2
	[mAP]	**44.47**±2.21	–	–	**44.05**±2.63	18.62±1.37	–
Pathology Detection (OD)							
VinDrCXR	[mAP]	14.12±0.95	**18.21**±1.20	7.44±0.88	–	2.82±0.25	–
NIH8	[mAP]	**11.14**±1.05	6.69±0.82	**11.89**±0.88	–	2.63±0.26	–
MS-CXR	[mAP]	**16.60**±1.38	**15.83**±1.42	**16.56**±1.06	–	7.15±0.52	–
Region Classification (RC)							
MS-CXR	[AUROC]	**82.33**±2.80	76.13±2.55	61.46±3.41	–	67.41±2.80	–
CIG	[wAUROC]	**70.46**±0.36	58.28±0.22	60.02±0.21	–	66.96±0.32	–
Region Explanation (RE)							
MS-CXR	[Mic-F1-14]	**49.97**±2.24	–	–	–	5.86±1.41	**48.97**±2.50
	[Mac-F1-14]	**20.50**±1.54	–	–	–	3.69±0.67	16.37±2.00
	[METEOR]	**8.79**±0.54	–	–	–	4.26±0.36	8.15±0.78
CIG	[Mic-F1-14]	**53.34**±0.43	–	–	–	24.40±0.38	45.26±0.44
	[Mac-F1-14]	**29.13**±0.35	–	–	–	8.93±0.15	20.88±0.19
	[METEOR]	**10.18**±0.13	–	–	–	3.82±0.03	**7.88**±0.10
Full Report Generation (RG)							
MIMIC-CXR†	[Mic-F1-14]	52.32±0.51	–	–	–	–	**55.7**
	[Mac-F1-14]	32.56±0.51	–	–	–	–	**39.83**
	[Ex-F1-14]	**58.76**±0.42	–	–	–	–	47.6
	[Mic-F1-5+]	**61.03**±0.56	–	–	–	–	58.8
	[Mac-F1-5+]	**55.85**±0.57	–	–	–	–	51.7
	[METEOR]	13.26±0.10	–	–	–	–	**33.3**

† Test splits and pre-processing can differ between models, leading to limitations in the exact comparison of results, as also acknowledged by [20, 59].

best SupOD model. On MS-CXR, ChEX is within 1-std of the best SupOD and WSupOD models. Zero-shot contrastive models perform poorly, relying on thresholding of noisy similarity maps. On *region classification (RC)* tasks, ChEX outperforms all baselines, with improvements of 8% on MS-CXR and 5% on CIG.

Text Generation On *Region explanation (RE)* tasks, ChEX performs similar or better than RGRG. On MS-CXR, ChEX improves by 25% on Mac-F1-14. On CIG, ChEX improves by 18% on Mic-F1-14, 40% on Mac-F1-14, and 29% on METEOR, although RGRG was explicitly trained on this task. Other report generation models cannot provide region-level descriptions while sentence-retrieval with contrastive baselines performs poorly on these tasks. For *full report generation (RG)* on MIMIC-CXR, ChEX sets a new state-of-the-art on the metrics Ex-F1-14 (+23%), Mic-F1-5+ (+4%), and Mac-F1-5+ (+8%). On the commonly used Mic-F1-14 metric, ChEX outperforms the 10 times larger model Med-PaLM M 12B and is only slightly outperformed by the 7 times larger SOTA model MAIRA-1. While ChEX shows limitations on Mac-F1-14, it still outperforms RGRG. However, ChEX performs relatively low on language-based metrics like METEOR due to its query-based report generation approach. Overall, ChEX demonstrates strong report generation performance, achieving results that are close or even better than SOTA models of up to 80 times the size of ChEX.



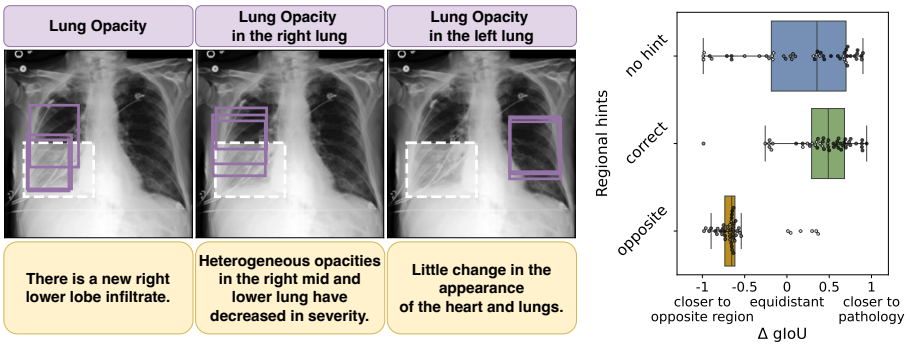
(a) Example chest X-ray with pleural effusion in both lungs. Ground-truth boxes are marked in white, queries are shown above, their predicted boxes in purple, and their predicted descriptions below.

(b) Relative distance (gIoU) between the queried and the other pathology instance. Darker points mark higher box scores.

Fig. 4: Effect of interactive prompting for multiregion disambiguation. In samples with the presence of the same pathology (*e.g.*, pleural effusion) in both lungs, using no regional hints in the textual query ("pleural effusion") detects both pathology instances equally well while adding a course regional hint ("pleural effusion in the right lung") or a fine regional hint ("pleural effusion in the right lower lung") steers the models towards selecting the queried pathology instance.

5.2 Interactive Prompting

ChEX aims to enable the interactive diagnosis of chest X-rays, going beyond the benchmarked tasks in Sec. 5.1. However, ChEX was trained only on simple prompts (*e.g.*, the name of a pathology). We thus validate that ChEX generalizes to more complex prompts by investigating the predicted bounding boxes in two scenarios: (i) multi-region disambiguation via regional hints and (ii) regional hints for negative regions. For scenario (i), shown in Fig. 4, we consider cases where a specific pathology is present in both lungs, *i.e.* where there are two boxes for the same pathology. We can observe that when using no regional hints in the textual query ("pleural effusion"), ChEX detects both pathology instances equally well while adding a course regional hint ("pleural effusion in the right lung") or a fine regional hint ("pleural effusion in the right lower lung"), steers the models towards selecting the queried pathology instance. For scenario (ii), shown in Fig. 5, we consider samples where the queried pathology is only present once (*i.e.* in one of the lungs). We again study the effect of using regional hints, either the correct hint ("lung opacity in the right lung", assuming there is a lung opacity only in the right lung) or the hint to the opposite lung ("lung opacity in the left lung"). We found that using no regional hint ("lung opacity") detects the pathology mostly correctly, while adding the correct regional hint improves the localization. If, on the other hand, we provide the regional hint for the opposite lung, the model is steered towards checking the queried anatomical region and thus away from the pathology. Overall, ChEX considers the user's intents of textual prompts very well and thus enables interactive usage patterns.



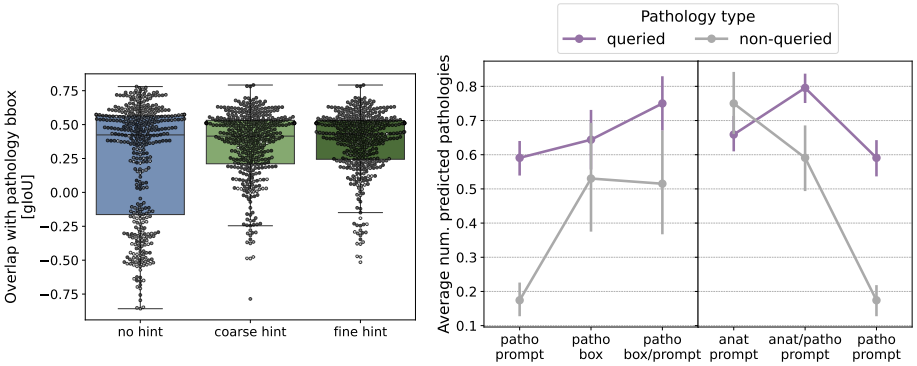
(a) Example chest X-ray with a lung opacity only in the right lung. Ground-truth boxes are marked in white, queries are shown above, their predicted boxes are in purple, and their predicted descriptions are shown below.

(b) Relative distance (gIoU) between the pathology and the opposite lung region. Darker dots mark higher box scores.

Fig. 5: Effect of interactive prompting with regional hints to negative regions. In samples with a pathology in only one of the lungs (*e.g.*, lung opacity in the right lung), using no regional hint ("lung opacity") detects the pathology mostly correctly. Adding the correct regional hint ("lung opacity in the right lung") improves the localization while the regional hint for the opposite lung ("lung opacity in the left lung") steers the model towards the queried anatomical region (away from the pathology), as expected.

5.3 Improvement through Precise Interactive Prompting

ChEX facilitates the interactive involvement of medical expert users. In Fig. 6, we investigate the impact of providing preciser prompts on the model's localization quality, its ability to accurately describe the queried pathology in the predicted sentence (prediction accuracy), and the inclusion of non-queried pathologies (query specificity). Providing coarse regional hints (*e.g.*, "pneumonia in the left lung") enhances localization and prediction accuracy compared to only providing the pathology name ("pneumonia"). Further refinement of regional hints (*e.g.*, "[...] in the left *upper* lung") offers minimal additional benefit on localization. In addition to textual prompts, the model can be queried using bounding boxes of relevant regions. Utilizing the bounding box of a pathology (instead of a textual prompt) improves the prediction accuracy but also includes additional aspects, rendering the prediction less query-specific. Combining textual and bounding box queries yields the best prediction accuracy, enhancing query specificity compared to using only bounding boxes. When employing a textual query of the associated anatomical region (*e.g.*, "left lung"), the present pathology is described adequately, but – as expected – descriptions are not specific to a single pathology. Using a more precise prompt (*e.g.*, "pneumonia in the left lung") enhances prediction accuracy and query specificity while relying solely on the pathology prompt increases query specificity further but marginally degrades the prediction accuracy. Despite ChEX demonstrating competitive performance even with simple prompts (Sec. 5.1), its responsiveness to more specific prompts enables even greater prediction accuracy when used interactively.



(a) Effect of prompting strategies on pathology localization quality (gIoU). Compared to using no regional hint (*e.g.* "pneumonia"), a course hint ("pneumonia in the left lung") or a fine hint ("pneumonia in the left upper lung") improves pathology localization. Darker dots mark higher box scores.

(b) Effect of prompting strategies on the presence of queried pathologies (*i.e.* defined in the prompt) and non-queried (*i.e.* any other positive or negative) pathologies in the predicted sentence. **Left:** Pathologies queried by its textual prompt (patho prompt), by its bounding box (patho box), or both. **Right:** Prompting the associated anatomical region of the pathology (anat prompt, *e.g.* "left lung"), the pathology with a regional hint ("penumonia in the left lung"), or only the pathology ("penumonia").

Fig. 6: Effect of precise prompting on localization (a) and sentence prediction (b). Preciser prompts improve localization and the description of queried pathologies, while more specific prompts reduce the description of additional, non-queried pathologies.

5.4 Interpretable and Customizable Report Generation

ChEX supports automatic full report generation using a pre-defined set of textual prompts based on which the model localizes and describes relevant regions. Unlike typical report generation models, the reliance on prompt sets for report generation offers high flexibility as the prompt sets can be customized without the need for re-training. We study (cf. supp. material) the effect of using only pathologies, anatomical regions, or both as prompt sets. The choice of the prompt set enables balancing precision and recall of the model. While all studied prompt sets lead to results competitive with the baselines, with Mic-F1-14 ranging from 50.08 to 52.37, using both prompt sets leads to optimal performance.

In Fig. 7, we show an example of a generated report. The prediction of bounding boxes for predicted sentences enables a high degree of transparency and interpretability – not provided by most report generation models – thus simplifying correctness checking and enabling an optimal clinical workflow.

5.5 Technical Insights

We additionally conducted ablation studies on several technical design decisions of our model and training paradigm. We summarize the key findings in the following. For detailed results we refer to the supp. material.

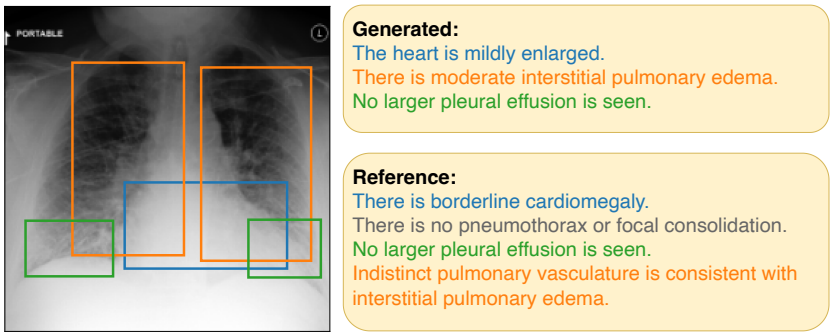


Fig. 7: Example of a generated report with predicted bounding boxes. Our model generates a concise and accurate report. By predicting bounding boxes for descriptions, ChEX provides a high level of interpretability and promotes easy checking of the generated report through radiologists, enabling an optimal workflow for clinical practice. For further qualitative examples, we refer to the supp. material.

The mixture of tokens for pathologies, anatomy, and report sentences enables multitask capabilities. We found that using all these three token types during training achieves the best multitask performance. Having both pathology and anatomy tokens is especially relevant for accurate localization (SG and OD tasks). Anatomy tokens are especially relevant for RC and RE tasks, likely due to the availability of pathology labels and sentences associated with bounding boxes. While sentence tokens can slightly harm localization quality, they are relevant to achieve the best possible text generation performance.

Localization targets are highly beneficial for all tasks, even for full report generation. We identified the importance of bounding box targets for all tasks. Pathology class labels are relevant mainly for OD tasks. Contrastive sentence supervision improves some OD tasks as well as MS-CXR-based RC and RE tasks, *i.e.* it helps with the understanding of pathology regions, while generative text supervision is not significantly beneficial for non-generative tasks.

Technical differences to baselines A key distinction between ChEX and most generative baselines lies in its aspect-level generation approach, where sentences are generated individually for specific findings and regions of the image. In contrast, models like MAIRA-1 [20], Med-PaLM M [61], or OmniFM-DR [72] generate the full report in a single shot. This approach is critical to ChEX's strong performance, as demonstrated by the good results of RGRG [59], which employs a similar region-level strategy. Compared to RGRG, ChEX introduces two key innovations: (i) pathology and sentence tokens, which extend RGRG's solely anatomy-based approach; and (ii) contrastive alignment of region features with their sentences and pathology prompts. Our ablation studies confirm that these differences are essential to ChEX's superior performance over RGRG.

6 Discussion

Importance of Semantically Meaningful Visual Features Many recent works on radiology report generation heavily focus on the utilization of large language models (LLMs) [20, 48, 61, 72]. We demonstrate that even smaller language models can achieve competitive performance when prioritizing improvements in image understanding aspects. Specifically, our approach places emphasis on semantic regions, leveraging bounding box supervision to enhance prediction quality. We argue that a synergistic combination of optimal image encoding with regional focus along with the deductive powers and knowledge of LLMs presents a promising path for improving report generation. Furthermore, our proposed training strategy enables the integration of multiple tasks and datasets, offering a viable approach to realizing this path.

Interpretability and Interactivity for Clinical Practice While the performance of report generation models is improving, even occasional inaccuracies in their predictions can limit their clinical applicability. We argue that a high degree of interpretability and the integration of medical experts in the generation process, *i.e.* interactivity, offers a more promising path to clinical application than solely improving prediction quality. Our model ChEX showcases a unique combination of interpretability and interactivity – not provided by other models – by providing bounding boxes for generated descriptions while enabling user guidance through textual prompts and bounding box queries.

Limitations Our training approach effectively utilizes the available datasets and eliminates the need for intricate data engineering. However, this comes at the cost of a more complex training process. Furthermore, the used datasets do not provide question-answer pairs, so textual queries are either based on pre-defined prompts or report sentences, and answers are always based on report sentences. This limits the types of supported textual queries, mainly regional or pathology hints, or both. At the same time, answers are plain descriptions of the queried regions or pathologies, while specific answers to more complicated questions (*e.g.* comparing regions) are not supported. Also, using report sentences as answers can lead to the hallucination of comparisons with previous images, although only a single image is used, a phenomenon common to report generation models [2, 20, 52]. Future work may use instruction tuning to tackle these issues.

Additionally, as we move towards clinical application, further evaluation from a radiologist's perspective is essential, and future work should include systematic studies on user experience to ensure seamless integration into clinical workflows.

Conclusion We proposed ChEX, a model for predicting visually grounded textual descriptions of chest X-rays based on user queries. Our analysis underscores ChEX's competitive performance against SOTA models across 9 tasks, and its responsiveness to user prompts, therefore laying a foundation for future advancements in interactive and localized text generation models.

Acknowledgements

The project was supported by ERC Grant Deep4MI (884622).

GK and DR received support from the German Federal Ministry of Education and Research and the Bavarian State Ministry for Science and the Arts under the Munich Centre for Machine Learning (MCML), from the German Ministry of Education and Research and the the Medical Informatics Initiative as part of the PrivateAIM Project, from the Bavarian Collaborative Research Project PRIPREKI of the Free State of Bavaria Funding Programme "Artificial Intelligence – Data Science", and from the German Academic Exchange Service (DAAD) under the Kondrad Zuse School of Excellence for Reliable AI (RelAI).

References

1. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
2. Bannur, S., Hyland, S., Liu, Q., Pérez-García, F., Ilse, M., Castro, D.C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., Schwaighofer, A., Wetscherek, M., Lungren, M.P., Nori, A., Alvarez-Valle, J., Oktay, O.: Learning to exploit temporal structure for biomedical vision-language processing. In: CVPR. pp. 15016–15027 (2023). https://doi.org/10.1109/CVPR52729.2023.01442
3. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., Poon, H., Oktay, O.: Making the most of text semantics to improve biomedical vision–language processing. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV. pp. 1–21. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-20059-5_1
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) ECCV. pp. 213–229. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13
5. Chen, Z., Zhou, Y., Tran, A., Zhao, J., Wan, L., Ooi, G.S.K., Cheng, L.T.E., Thng, C.H., Xu, X., Liu, Y., Fu, H.: Medical phrase grounding with region-phrase context contrastive alignment. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (eds.) MICCAI. pp. 371–381. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-43990-2_35
6. Deng, J., Yang, Z., Liu, D., Chen, T., Zhou, W., Zhang, Y., Li, H., Ouyang, W.: Transvg++: End-to-end visual grounding with language conditioned vision transformer. IEEE TPAMI **45**(11), 13636–13652 (nov 2023). https://doi.org/10.1109/TPAMI.2023.3296823
7. Deng, J., Yang, Z., Chen, T., Zhou, W., Li, H.: Transvg: End-to-end visual grounding with transformers. In: ICCV. pp. 1749–1759. IEEE (2021). https://doi.org/10.1109/ICCV48922.2021.00179, https://doi.org/10.1109/ICCV48922.2021.00179

8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
9. Du, Y., Fu, Z., Liu, Q., Wang, Y.: Visual grounding with transformers. In: ICME (2022)
10. Eslami, S., de Melo, G., Meinel, C.: Does CLIP benefit visual question answering in the medical domain as much as it does in the general domain? ArXiv preprint **abs/2112.13906** (2021), https://arxiv.org/abs/2112.13906
11. Geis, J.R., Brady, A.P., Wu, C.C., Spencer, J., Ranschaert, E., Jaremko, J.L., Langer, S.G., Borondy Kitts, A., Birch, J., Shields, W.F., et al.: Ethics of artificial intelligence in radiology: summary of the joint european and north american multisociety statement. Radiology **293**(2), 436–440 (2019)
12. Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., et al.: Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. Circulation [Online] **101**(23), 215—220 (2000)
13. Gu, T., Liu, D., Li, Z., Cai, W.: Complex organ mask guided radiology report generation. In: WACV. pp. 7995–8004 (January 2024)
14. Gu, X., Lin, T., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. In: ICLR (2022), https://openreview.net/forum?id=lL3lnMbR4WU
15. Guo, M., Yi, H., Qin, Z., Wang, H., Men, A., Lao, Q.: Multiple prompt fusion for zero-shot lesion detection using vision-language models. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (eds.) MICCAI. pp. 283–292. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-43904-9_28
16. He, J., Li, P., Liu, G., Zhao, Z., Zhong, S.: Pefomed: Parameter efficient fine-tuning on multimodal large language models for medical visual question answering (2024)
17. Hou, W., Xu, K., Cheng, Y., Li, W., Liu, J.: ORGAN: Observation-guided radiology report generation via tree reasoning. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 8108–8122. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-long.451, https://aclanthology.org/2023.acl-long.451
18. Huang, S., Shen, L., Lungren, M.P., Yeung, S.: Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: ICCV. pp. 3922–3931. IEEE (2021). https://doi.org/10.1109/ICCV48922.2021.00391, https://doi.org/10.1109/ICCV48922.2021.00391
19. Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., Liu, S., Chi, H., Hu, X., Yue, K., Li, L., Grau, V., Fan, D.P., Dong, F., Ni, D.: Segment anything model for medical images? Medical Image Analysis **92**, 103061 (2024). https://doi.org/https://doi.org/10.1016/j.media.2023.103061, https://www.sciencedirect.com/science/article/pii/S1361841523003213
20. Hyland, S.L., Bannur, S., Bouzid, K., Castro, D.C., Ranjit, M., Schwaighofer, A., Pérez-García, F., Salvatelli, V., Srivastav, S., Thieme, A., Codella, N., Lungren, M.P., Wetscherek, M.T., Oktay, O., Alvarez-Valle, J.: Maira-1: A specialised large multimodal model for radiology report generation (2023)
21. Ichinose, A., Hatsutani, T., Nakamura, K., Kitamura, Y., Iizuka, S., Simo-Serra, E., Kido, S., Tomiyama, N.: Visual grounding of whole radiology reports for 3d ct

images. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (eds.) MICCAI. pp. 611–621. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-43904-9_59
22. Jin, H., Che, H., Lin, Y., Chen, H.: Promptmrg: Diagnosis-driven prompts for medical report generation (2024)
23. Johnson, A., Pollard, T., Berkowitz, S., et al.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data **6**(317) (2019). https://doi.org/https://doi.org/10.1038/s41597-019-0322-0
24. Johnson, A., Pollard, T., Mark, R., Berkowitz, S., Horng, S.: Mimic-cxr database (version 2.0.0). PhysioNet (2019). https://doi.org/https://doi.org/10.13026/C2JT1Q
25. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything (2023)
26. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day (2023)
27. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., Chang, K.W., Gao, J.: Grounded language-image pre-training. In: CVPR. pp. 10955–10965 (2022). https://doi.org/10.1109/CVPR52688.2022.01069
28. Li, M., Lin, B., Chen, Z., Lin, H., Liang, X., Chang, X.: Dynamic graph enhanced contrastive learning for chest x-ray report generation. In: CVPR. pp. 3334–3343 (2023). https://doi.org/10.1109/CVPR52729.2023.00325
29. Li, M., Sigal, L.: Referring transformer: A one-step approach to multi-task visual grounding. In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (eds.) NeurIPS. pp. 19652–19664 (2021), https://proceedings.neurips.cc/paper/2021/hash/a376802c0811f1b9088828288eb0d3f0-Abstract.html
30. Liao, R., Moyer, D., Cha, M., Quigley, K., Berkowitz, S., Horng, S., Golland, P., Wells, W.M.: Multimodal representation learning via maximization of local mutual information. In: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (eds.) MICCAI. pp. 273–283. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_26
31. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. IEEE TPAMI **42**(2), 318–327 (2020). https://doi.org/10.1109/TPAMI.2018.2858826
32. Liu, J., Hu, T., Zhang, Y., Feng, Y., Hao, J., Lv, J., Liu, Z.: Parameter-efficient transfer learning for medical visual question answering. IEEE Transactions on Emerging Topics in Computational Intelligence pp. 1–11 (2023). https://doi.org/10.1109/TETCI.2023.3311333
33. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: DAB-DETR: dynamic anchor boxes are better queries for DETR. In: ICLR (2022), https://openreview.net/forum?id=oMI9PjOb9Jl
34. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection (2023)
35. Liu, X., Ji, K., Fu, Y., Du, Z., Yang, Z., Tang, J.: P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. CoRR **abs/2110.07602** (2021), https://arxiv.org/abs/2110.07602
36. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)

37. Maaz, M., Rasheed, H., Khan, S., Khan, F.S., Anwer, R.M., Yang, M.H.: Class-agnostic object detection with multi-modal transformer. In: ECCV. Springer (2022)
38. Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., Sun, L., Wang, J.: Conditional detr for fast training convergence. In: ICCV. pp. 3631–3640 (2021). https://doi.org/10.1109/ICCV48922.2021.00363
39. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence **267**, 1–38 (2019)
40. Miura, Y., Zhang, Y., Tsai, E., Langlotz, C., Jurafsky, D.: Improving factual completeness and consistency of image-to-text radiology report generation. In: NAACL. pp. 5288–5304 (2021)
41. Müller, P., Kaissis, G., Zou, C., Rueckert, D.: Joint learning of localized representations from medical images and reports. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV. pp. 685–701. Springer Nature Switzerland, Cham (2022)
42. Müller, P., Meissen, F., Brandt, J., Kaissis, G., Rueckert, D.: Anatomy-driven pathology detection on chest x-rays. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (eds.) MICCAI. pp. 57–66. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-43907-0_6
43. Müller, P., Meissen, F., Kaissis, G., Rueckert, D.: Weakly supervised object detection in chest x-rays with differentiable roi proposal networks and soft roi pooling (2024)
44. Nguyen, H.Q., Pham, H.H., Tuan Linh, L., Dao, M., Khanh, L.: Vindr-cxr: An open dataset of chest x-rays with radiologist annotations (version 1.0.0). PhysioNet (2021). https://doi.org/https://doi.org/10.13026/3akn-b287
45. Nguyen, H.Q., Lam, K., Le, L.T., Pham, H.H., Tran, D.Q., Nguyen, D.B., Le, D.D., Pham, C.M., Tong, H.T., Dinh, D.H., et al.: Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. Scientific Data **9**(1), 429 (2022). https://doi.org/https://doi.org/10.1038/s41597-022-01498-w
46. Nicolson, A., Dowling, J., Koopman, B.: Improving chest x-ray report generation by leveraging warm starting. Artificial Intelligence in Medicine **144**, 102633 (2023). https://doi.org/https://doi.org/10.1016/j.artmed.2023.102633, https://www.sciencedirect.com/science/article/pii/S0933365723001471
47. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv: 1807.03748 (2019)
48. Pellegrini, C., Özsoy, E., Busam, B., Navab, N., Keicher, M.: Radialog: A large vision-language model for radiology report generation and conversational assistance (2023)
49. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) ICML. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021), http://proceedings.mlr.press/v139/radford21a.html
50. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
51. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al.: Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225 (2017). https://doi.org/10.48550/arXiv.1711.05225

52. Ramesh, V., Chi, N.A., Rajpurkar, P.: Improving radiology report generation systems by removing hallucinated references to non-existent priors. In: Machine Learning for Health. pp. 456–473. PMLR (2022)
53. Rasheed, H., Maaz, M., Khattak, M.U., Khan, S., Khan, F.S.: Bridging the gap between object and image-level representations for open-vocabulary detection. In: NIPS (2022)
54. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. NIPS **28** (2015)
55. Seibold, C., Reiß, S., Sarfraz, M.S., Stiefelhagen, R., Kleesiek, J.: Breaking with fixed set pathology recognition through report-guided contrastive training. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI. pp. 690–700. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_66
56. Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A.Y., Lungren, M.P.: Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert (2020)
57. van Sonsbeek, T., Derakhshani, M.M., Najdenkoska, I., Snoek, C.G.M., Worring, M.: Open-ended medical visual question answering through prefix tuning of language models. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (eds.) MICCAI. pp. 726–736. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-43904-9_70
58. Sun, J., Wei, D., Wang, L., Zheng, Y.: Lesion Guided Explainable Few Weak-Shot Medical Report Generation, p. 615–625. Springer Nature Switzerland (2022). https://doi.org/10.1007/978-3-031-16443-9_59, http://dx.doi.org/10.1007/978-3-031-16443-9_59
59. Tanida, T., Müller, P., Kaissis, G., Rueckert, D.: Interactive and explainable region-guided radiology report generation. In: CVPR. pp. 7433–7442 (2023). https://doi.org/10.1109/CVPR52729.2023.00718
60. Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P.: Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. Nature Biomedical Engineering **6**(12), 1399–1406 (2022)
61. Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.C., Carroll, A., Lau, C., Tanno, R., Ktena, I., Palepu, A., Mustafa, B., Chowdhery, A., Liu, Y., Kornblith, S., Fleet, D., Mansfield, P., Prakash, S., Wong, R., Virmani, S., Semturs, C., Mahdavi, S.S., Green, B., Dominowska, E., y Arcas, B.A., Barral, J., Webster, D., Corrado, G.S., Matias, Y., Singhal, K., Florence, P., Karthikesalingam, A., Natarajan, V.: Towards generalist biomedical ai. NEJM AI **1**(3), AIoa2300138 (2024). https://doi.org/10.1056/AIoa2300138
62. Wang, F., Zhou, Y., Wang, S., Vardhanabhuti, V., Yu, L.: Multi-granularity cross-modal alignment for generalized medical visual representation learning. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) NeurIPS (2022)
63. Wang, L., Ning, M., Lu, D., Wei, D., Zheng, Y., Chen, J.: An inclusive task-aware framework for radiology report generation. In: MICCAI. pp. 568–577 (2022)
64. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., Wei, F.: Image as a foreign language: Beit pretraining for vision and vision-language tasks. In: CVPR. pp. 19175–19186 (2023). https://doi.org/10.1109/CVPR52729.2023.01838
65. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised clas-

sification and localization of common thorax diseases. In: CVPR. pp. 2097–2106 (2017). https://doi.org/10.1109/CVPR.2017.369
66. Wang, Z., Liu, L., Wang, L., Zhou, L.: Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In: CVPR. pp. 11558–11567 (2023). https://doi.org/10.1109/CVPR52729.2023.01112
67. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: MedCLIP: Contrastive learning from unpaired medical images and text. In: Conference on Empirical Methods in Natural Language Processing. pp. 3876–3887. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (2022), https://aclanthology.org/2022.emnlp-main.256
68. Wu, J., Agu, N., Lourentzou, I., Sharma, A., Paguio, J.A., Yao, J.S., Dee, E.C., Kashyap, S., Giovannini, A., Celi, L.A., et al.: Chest imagenome dataset for clinical reasoning. In: NIPS (2021)
69. Wu, J.T., Agu, N.N., Lourentzou, I., Sharma, A., Paguio, J.A., Yao, J.S., Dee, E.C., Mitchell, W., Kashyap, S., Giovannini, A., et al.: Chest imagenome dataset (version 1.0.0). PhysioNet (2021). https://doi.org/https://doi.org/10.13026/wv01-y230
70. Wu, X., Zhu, F., Zhao, R., Li, H.: Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In: CVPR. pp. 7031–7040 (2023). https://doi.org/10.1109/CVPR52729.2023.00679
71. Wu, Y., Zhou, Y., Saiyin, J., Wei, B., Lai, M., Shou, J., Fan, Y., Xu, Y.: Zero-shot nuclei detection via visual-language pre-trained models. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (eds.) MICCAI. pp. 693–703. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-43987-2_67
72. Xu, L., Ni, Z., Liu, X., Wang, X., Li, H., Zhang, S.: Learning a multi-task transformer via unified and customized instruction tuning for chest radiograph interpretation (2023)
73. Xu, S., Yang, L., Kelly, C., Sieniek, M., Kohlberger, T., Ma, M., Weng, W.H., Kiraly, A., Kazemzadeh, S., Melamed, Z., Park, J., Strachan, P., Liu, Y., Lau, C., Singh, P., Chen, C., Etemadi, M., Kalidindi, S.R., Matias, Y., Chou, K., Corrado, G.S., Shetty, S., Tse, D., Prabhakara, S., Golden, D., Pilgrim, R., Eswaran, K., Sellergren, A.: Elixr: Towards a general purpose x-ray artificial intelligence system through alignment of large language models and radiology vision encoders (2023)
74. Yang, Z., Gan, Z., Wang, J., Hu, X., Ahmed, F., Liu, Z., Lu, Y., Wang, L.: Unitab: Unifying text and box outputs for grounded vision-language modeling. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV. pp. 521–539. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-20059-5_30
75. Zang, Y., Li, W., Zhou, K., Huang, C., Loy, C.C.: Open-vocabulary detr with conditional matching. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV. pp. 106–122. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-20077-9_7
76. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions. In: CVPR. pp. 14388–14397 (2021). https://doi.org/10.1109/CVPR46437.2021.01416
77. Zhang, G., Luo, Z., Yu, Y., Cui, K., Lu, S.: Accelerating detr convergence via semantic-aligned matching. In: CVPR. pp. 939–948 (2022). https://doi.org/10.1109/CVPR52688.2022.00102

78. Zhang, H., Zhang, P., Hu, X., Chen, Y.C., Li, L., Dai, X., Wang, L., Yuan, L., Hwang, J.N., Gao, J.: Glipv2: Unifying localization and vision-language understanding. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) NeurIPS. vol. 35, pp. 36067–36080. Curran Associates, Inc. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/ea370419760b421ce12e3082eb2ae1a8-Paper-Conference.pdf
79. Zhang, K., Yu, J., Adhikarla, E., Zhou, R., Yan, Z., Liu, Y., Liu, Z., He, L., Davison, B., Li, X., Ren, H., Fu, S., Zou, J., Liu, W., Huang, J., Chen, C., Zhou, Y., Liu, T., Chen, X., Chen, Y., Li, Q., Liu, H., Sun, L.: Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks (2024)
80. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Lipton, Z., Ranganath, R., Sendak, M., Sjoding, M., Yeung, S. (eds.) Machine Learning for Healthcare Conference. Proceedings of Machine Learning Research, vol. 182, pp. 2–25. PMLR (05–06 Aug 2022)
81. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., Gao, J.: Regionclip: Region-based language-image pretraining. In: CVPR. pp. 16772–16782 (2022). https://doi.org/10.1109/CVPR52688.2022.01629
82. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 350–368. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-20077-9_21
83. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: deformable transformers for end-to-end object detection. In: ICLR (2021), https://openreview.net/forum?id=gZ9hCDWe6ke
84. Zong, Z., Song, G., Liu, Y.: Detrs with collaborative hybrid assignments training. In: ICCV. pp. 6725–6735 (2023). https://doi.org/10.1109/ICCV51070.2023.00621