Cyclic Vision-Language Manipulator: Towards Reliable and Fine-Grained Image Interpretation for Automated Report Generation

Yingying Fang[1*]**, Zihao Jin**[1*]**, Shaojie Guo**[2]**, Jinda Liu**[3]**, Zhiling Yue**[1]**, Yijian Gao**[1]**, Junzhi Ning**[1]**, Zhi Li**[2]**, Simon Walsh**[1]**, Guang Yang**[1†]

[1] Imperial College London, London, UK
[2] East China Normal University, Shanghai, China
[3] The Chinese University of Hong Kong, Hong Kong, China
{y.fang, g.yang}@imperial.ac.uk

Abstract

Despite significant advancements in automated report generation, the opaqueness of text interpretability continues to cast doubt on the reliability of the content produced. This paper introduces a novel approach to identify specific image features in X-ray images that influence the outputs of report generation models. Specifically, we propose Cyclic Vision-Language Manipulator (**CVLM**), a module to generate a manipulated X-ray from an original X-ray and its report from a designated report generator. The essence of CVLM is that cycling manipulated X-rays to the report generator produces altered reports aligned with the alterations pre-injected into the reports for X-ray generation, achieving the term "cyclic manipulation". This process allows direct comparison between original and manipulated X-rays, clarifying the critical image features driving changes in reports and enabling model users to assess the reliability of the generated texts. Empirical evaluations demonstrate that CVLM can identify more precise and reliable features compared to existing explanation methods, significantly enhancing the transparency and applicability of AI-generated reports.

1 Introduction

The automated and precise interpretation of chest X-rays has transformative potential, poised to enhance the consistency, quality, and efficiency of current interpretations conducted by human experts in healthcare. Over the past three years, substantial efforts have been invested in refining the language generation capabilities, aligning visual and linguistic features, and increasing the accuracy of clinical report findings. The advent of large language models has sparked another wave in report generation, prioritizing linguistic precision and sophistication [Lee *et al.*, 2023; He *et al.*, 2024; Liu *et al.*, 2024].

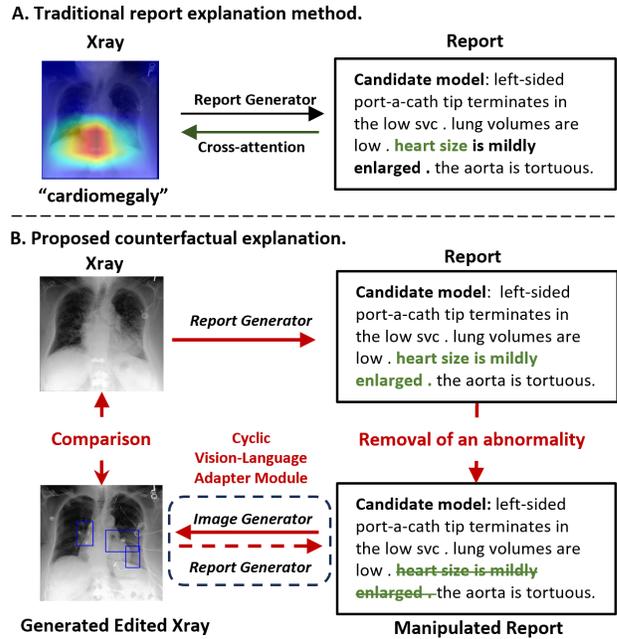[1]The appendix for this work is available at https://arxiv.org/abs/2411.05261.

Figure 1: The overview of using counterfactual explanation for decoding the report generated from a target report generator.

Despite these enhancements, the reports generated by these models often emerge as cryptic outputs from a "black box", leaving users with little understanding of the underlying processes. Furthermore, the proliferation of diverse models leads to inconsistent reports when analyzing identical X-rays, raising concerns about the reliability of these automated systems. This variability and lack of transparency have impeded their broader adoption in clinical settings [Hertz *et al.*, 2022; Müller *et al.*, 2024].

In response, several studies have adopted existing Explainable AI (XAI) techniques to uncover the visual features influencing generated content, thereby aiming to enhance the trustworthiness of the generated texts. The most widely used XAI methods in this field typically generate an attribution map to highlight the influential features related to the generated contents through cross-attention maps [Liu *et al.*, 2019; Cao *et al.*, 2023; Chen *et al.*, 2020] or the GradCAM-based

Figure 2: Overview of applying the proposed Cyclic Vision-Language Manipulator (CVLM) in explaining the report generator and the targets required for CVLM for counterfactual explanation. In this case, the decoded report generator identifies the abnormal heart contour as a signal for detecting "cardiomegaly". After reviewing the features associated with this finding, the radiologist agrees with the generated conclusion from the report generator.

methods [Alfarghaly *et al.*, 2021; Spinks and Moens, 2019; Wang *et al.*, 2024]. However, these attribution maps often fail to highlight fine-grained visual features, instead presenting large and coarse areas (see Fig. 1 (A)). More critically, there is currently no reliable method to verify whether these attribution maps accurately represent the information utilized by report generators in producing findings, as there is no ground truth to confirm what the model has actually learned [Chase Walker *et al.*, 2024]. These two limitations raise an important ambiguity: while the highlighted areas in the attribution maps may appear irrelevant to the generated findings, it remains unclear whether report generators rely on irrelevant features or whether the explanation methods misidentify visual features relative to the content being explained.

To address these limitations and provide a reliable attribution map that can facilitate the understanding of generated texts, we propose a model-agnostic counterfactual explanation method to generate attribution maps. This method relies on the creation of counterfactual images, which modify specific features in the images to elicit changes in the model's decision. Comparing the counterfactual image to the original enables the identification of crucial features responsible for the altered decision [Wachter *et al.*, 2017]. Specifically, we introduce a module called the Cyclic Vision-Language Manipulator (CVLM) to generate effective counterfactual images using a text-conditioned diffusion model. The core of CVLM is to achieve cyclic manipulation: the manipulation begins with the generated text, and the counterfactual images derived from this manipulated text must produce a regenerated report that accurately reflects the same alterations introduced in the initial report (see Fig. 2 (B)). The comparison between the cyclically generated counterfactual image and the original X-ray produces a difference map to identify fine-grained image features. Furthermore, the counterfactual images, coupled with the altered regenerated report, verify

that the identified features are being utilized by existing report generators in their text generation process. ***The major contribution of this paper*** is the development of a method capable of identifying feature attributions that can be verified as being used by the report generator for its generated content. The identified features may be consistent or inconsistent with human knowledge, as the absence of "ground truth" for the features learned by the model. The verified features identified by our method can assist model developers in locating relevant visual cues used by the generator, and help human experts evaluate the reliability of the generated findings by comparing these features with their own clinical knowledge. ***The key contributions of this work*** are summarized as follows:

- We propose a CVLM module for cyclic manipulation of a query X-ray and its generated reports from a designated report generator. Additionally, we introduce a **cyclic success rate** to ensure the generation of effective counterfactual images for decoding the report generation process and to quantitatively evaluate the manipulation effectiveness of CVLM.

- We propose an unsupervised feature localization framework based on the difference map between counterfactual and initial X-ray images, effectively filtering noise and identifying key features responsible for result changes, facilitating the understanding of explanation results without requiring additional human labeling.

- Experiments show that counterfactual images, generated from manipulated reports, achieve fine-grained feature identification and verification for the generated contents from the report generator, which is unattainable with existing methods.

- Experiments demonstrate that the model-agnostic nature of CVLM allows users to evaluate whether the vi-

sual features associated with the generated findings are meaningful for the corresponding content. This capability also enables users to compare the reliability of results across different report generators and make informed decisions on adopting suggestions from various models.

2 Related Work

2.1 Counterfactual Explanation

The most widely applied methods for visual explanation are post-hoc explanation methods, as their model-agnostic nature enables them to generalize across different models. Popular approaches include activation-based methods, backpropagation-based methods, and perturbation-based methods. With the recent evolution of generative AI models, counterfactual explanation methods have emerged as an advanced perturbation-based approach. These methods generate counterfactual images that elicit different findings from the model, enabling users to identify differences between similar classes—a common challenge in medical image classification tasks. Applications has emerged in explaining the classifiers with different modalities in medical image analysis, including X-rays [Atad *et al.*, 2022; Mertes *et al.*, 2022; Singla *et al.*, 2023; Schutte *et al.*, 2021; Sankaranarayanan *et al.*, 2022; Maksudov *et al.*, 2025], Magnetic Resonance Imaging [Tanyel *et al.*, 2023; Fontanella *et al.*, 2023], ultrasound [Reynaud *et al.*, 2022], histopathology images [Karras *et al.*, 2020; Schutte *et al.*, 2021], and Computed Tomography [Fang *et al.*, 2024a; Fang *et al.*, 2024b]. Over time, counterfactual generation methods have evolved from using variational autoencoders [Rodriguez *et al.*, 2021] and generative adversarial networks [Lang *et al.*, 2021; Atad *et al.*, 2022] to leveraging diffusion models [Rombach *et al.*, 2022; Fang *et al.*, 2024a; Fang *et al.*, 2024b]. While these methods typically generate counterfactual images to elicit changes under the guidance of a given classifier, in this work, we aim to develop a counterfactual generation method driven by text manipulation. The counterfactual X-ray is designed to achieve the desired content changes in the regenerated reports.

2.2 Explainability in Report Generation Models

Current report generator models primarily utilize cross-attention mechanisms to generate text from input images. Consequently, most report generation methods explain their outputs by identifying the most relevant image patches for keywords in the generated report. This is achieved by calculating distances within the cross-attention architecture, which serves as the basis for the explanations of the generated keywords [Wang *et al.*, 2023; Cao *et al.*, 2023; Chen *et al.*, 2023]. Another category of methods [Alfarghaly *et al.*, 2021; Spinks and Moens, 2019; Wang *et al.*, 2024] employs Grad-CAM, a widely used technique, to identify the most activated features within the image encoder layers and produce attribution maps. However, due to their dependence on intermediate layers of deep networks, these attribution maps often face challenges such as coarse localization and "unverifiable" accuracy in explaining keyword-specific regions.

In contrast to these approaches, Tanida *et al.* (2023) proposed a report generation method called RGRG, which significantly enhances the interpretability and transparency of generated reports by tailoring content generation to specific anatomical regions. However, this approach heavily relies on the preparation of a large paired dataset comprising anatomical regions and corresponding fine-grained reports for both the anatomical detection model and the report generation training. This requirement substantially increases manual labeling costs and limits the model's ability to incorporate larger X-ray-report datasets that lack fine-grained annotations. Additionally, the self-explanatory workflow of RGRG is not transferable to other report generators with different generation processes.

In this paper, we aim to propose an explanation method that achieves fine-grained visual feature localization for generated key findings as RGRG, while eliminating the need for extensive labeling. Being model-agnostic, our approach can enhance the interpretability and transparency of generated reports across a wide range of report generation models.

2.3 Text-controlled Image Editing

In recent years, text-guided image editing has garnered significant interest due to its ability to simplify image editing through natural language input. A substantial body of work leverages the alignment between text and image embeddings within the pretrained large vision-language model CLIP [Radford *et al.*, 2021]. These methods use changes in text embeddings before and after editing to infer corresponding changes in image embeddings, enabling the generation of edited images. *DiffusionCLIP* [Kim *et al.*, 2022] fine-tunes generative models using CLIP loss to align changes across the two modalities. Alternatively, other approaches map changes in text embeddings directly into the latent space of the image encoder to regenerate the image without altering network parameters [Patashnik *et al.*, 2021; Abdal *et al.*, 2022; Lyu *et al.*, 2023].

With the emergence of text-conditioned image generation models such as Stable Diffusion [Rombach *et al.*, 2022], recent works have focused on more efficient text-guided image editing by directly modifying the text prompts used for image generation. However, a key challenge in these methods is that minor changes in prompts do not always lead to correspondingly minor changes in the generated images. To address this, *Prompt-to-Prompt* [Hertz *et al.*, 2022] achieved localized manipulation of generated images by merging intermediate feature maps during inference. Leveraging the paired images generated by *Prompt-to-Prompt*, *Pix2Pix* [Brooks *et al.*, 2023] further introduced a more user-friendly instructive editing network. *ControlNet* [Zhang *et al.*, 2023] trains an additional control network that uses auxiliary conditions, such as sketches or segmentation maps of key objects, to keep the objects unchanged while editing other image contents via text manipulation.

While the aforementioned methods edit images based on the semantic meaning of the modified text, CVLM differs in its manipulation goal. The manipulation of counterfactual images by CVLM, driven by the manipulated text, does not need to align with real-world semantics. Instead, it is specifically
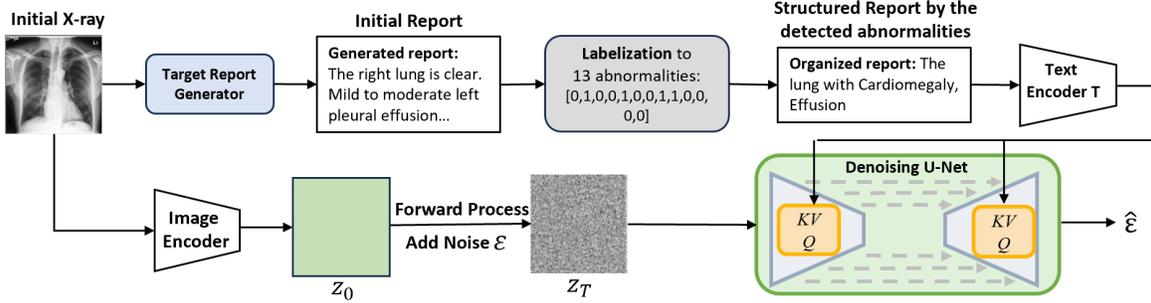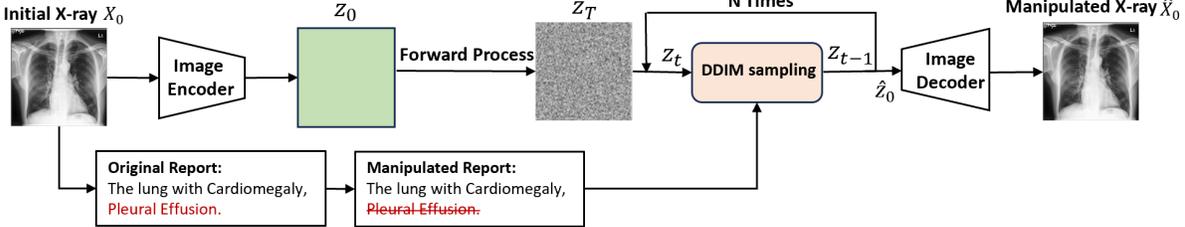
A. Training the Image Generator given a target <u>Report Generator</u>

B. Removal of a finding on a real query X-ray

Figure 3: Overview of applying the proposed Cyclic Vision-Language Manipulator in explaining the report generator and the challenges existing in developing the CVLM.

designed to elicit consistent changes in the regenerated reports produced by the target report generator. By achieving cyclic manipulation, the altered images are verified to influence on the report generator, thereby uncovering whether the generator relies on clinically recognized features or other biased features.

3 Method

The overall framework for CVLM and applying it for decoding a target report generator is illustrated in Fig. 3. In the following, we detail the components and application of CVLM to obtain tailored explanations for a target report generator.

3.1 Development of CVLM

The proposed CVLM module consists of an off-the-shelf report generator that produces reports from a query X-ray and a conditional diffusion model tailored to manipulate images based on the findings detected by the report generator. To achieve the explanation capability of the generated images, the image generators in CVLM are designed to meet three specific objectives, as illustrated in Fig. 2: (a) Reconstruction ability: Ensuring that query images can be accurately reconstructed from the unmanipulated generated reports. (b) Minimal image change: Ensuring that altering key findings in the generated report results in only minimal changes to the image. (c) Cyclic manipulation: Ensuring that the manipulated image results in consistent alterations in the regenerated report, completing the cyclic process. To effectively achieve these objectives, we implemented the following adaptations based on the advanced capabilities of the text-to-image Stable Diffusion model [Rombach *et al.*, 2022].

Data Preparation

To enable image manipulation based on the image features learned for the keywords generated by the report generator, rather than relying on real image features corresponding to medical terminology, we inferred the generated reports from the target report generator on the dataset on which it was trained. We paired the initial X-rays with the inferred results and trained the model to reconstruct the initial X-rays using the generated reports instead of the human-labeled ground-truth (GT) reports. This model is referred to as the tailored generation model, which can generate counterfactual X-rays by removing keywords from the generated reports, enabling the detection of specific differences responsible for changes in the findings produced by the target report generator.

Conditioning the reconstruction of the query image on the generated report is also critical for further feature identification, as explained in Fig. 2. For example, as shown in Fig. 4, training image generators with GT reports of X-ray images also achieve image manipulation via word manipulation. However, this approach fails to accurately reconstruct the original query X-ray. In this case, the reconstructed image from the GT model exaggerates features of "cardiomegaly" when it is present in the generated report but absent in the GT report. When "cardiomegaly" is removed from the text to observe its influence on the image, the GT reports also removes "cardiomegaly" but introduces more unintended changes to the initial query image. In contrast, the manipulated result from tailored model achieves more precise identification of the features in the initial image that caused the report generator to generate the finding of "cardiomegaly".

Furthermore, to facilitate automated manipulation of the generated reports and align the image features to the generated contents in the generated reports, we removed the
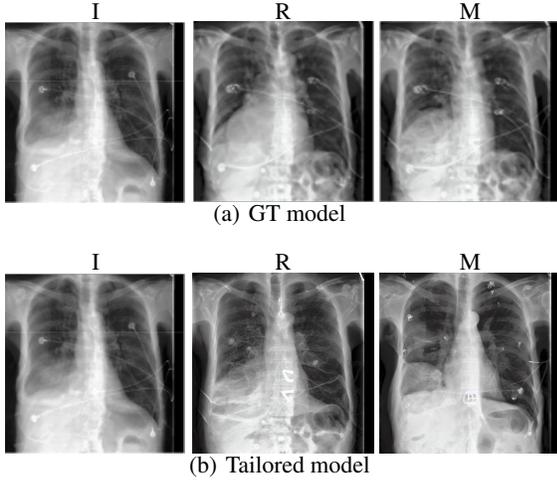
(a) GT model

(b) Tailored model

Figure 4: Reconstruction (R) and manipulation (M) of the initial image (I) by stable diffusion models trained with ground truth reports (GT model) and generated reports (tailored model), respectively. For both models, the reconstruction is conducted using the generated report, and the manipulation involves removing the presence of cardiomegaly from the prompt.

redundancy and simplified the reports into a predefined list of 13 abnormalities—Enlarged Cardiomediastinum, Cardiomegaly, Lung Opacity, Lung Lesion, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Effusion, Pleural Other, Fracture, and Support Devices—using the pretrained CheXbert classifier [Smit *et al.*, 2020], which is widely used for evaluating report correctness regarding major findings. We then organized the generated reports as a concatenated list of abnormalities, formatted as: "The lung with abnormalities $O_1, ..., O_N$" where O represents N abnormalities produced in the generated reports.

Training Objective

Our training objective follows the Stable Diffusion training procedure, which is given as below:

$$L_{LDM} := \mathbb{E}\left[\|\epsilon - \epsilon_\theta\left(z_T, T, \tau_\theta(y)\right)\|_2^2\right], \quad (1)$$

where z_T is the encoded feature of the initial real query X-ray image from the encoder of a variational autoencoder, τ_θ is the text encoder that transforms the prompt to the text embedding. During our training, we leverage the pretrained model weight for the text embedding and image autoencoder modules by a stable diffusion model pretrained on MIMIC [Liang *et al.*, 2023]. During training, we initialize the weight of UNET architecture by the stable diffusion pretrained weight "CompVis/stable-diffusion-v1-4" and freeze the parameter in the image autoencoder.

Real Image Manipulation

To enable the CVLM to explain the generated report of a real X-ray query, we employ Denoising Diffusion Implicit Models (DDIM), a non-stochastic variant of Denoising Diffusion Probabilistic Models (DDPMs), which performs the sampling process for T steps:

$$x_{t-1} = \sqrt{\overline{\alpha}_{t-1}}\left(\frac{x_t - \sqrt{1 - \overline{\alpha}_t} \cdot \epsilon_\theta(x_t, t)}{\sqrt{\overline{\alpha}_t}}\right)$$
$$+ \sqrt{1 - \overline{\alpha}_{t-1} - \sigma_t^2} \cdot \epsilon_\theta(x_t, t) + \sigma_t \epsilon_t \quad (2)$$

where $\epsilon_t \sim \mathcal{N}(0, \mathbf{I})$ represents a standard normal distribution, and σ_t controls the stochasticity of the forward process. Sharing the same inference formula as DDPM, DDIM sets σ_t in Eqn. (2) to zero, allowing for a deterministic reconstruction without randomness. To reconstruct the initial image, we approximate the initial noise x_T using DDIM Inversion, which introduces noise to the image through forward diffusion process.

3.2 Counterfactual Explanation

While the edited image reflects the manipulation in the report generator and achieves the desired changes in the regenerated report, as shown in Fig. 2, we refer to these manipulated X-rays as "cyclic" counterfactual images. These images are then utilized to decode the report generation process by identifying the visual feature changes that correspond to the modifications in the generated report for each query X-ray.

Removal of Visual Abnormality

To detect the underlying visual features associated with the context generated by the report generator, we modify the reorganized prompt by removing the findings produced in the generated report and send it to the image generation model for counterfactual generation, as shown in Fig. 3 (B). A successful cyclic counterfactual image is defined as one that successfully removes the target findings in the regenerated report. We then leverage these counterfactual images to detect the visual changes that lead to the reversal of the report findings.

Unsupervised Feature Identification Frame Generation

To facilitate the detection of crucial features that alter the findings in the regenerated report, we propose an unsupervised method for generating feature identification frames. This method calculates a frame based on the absolute difference map between the initial X-ray and its counterfactual, allowing the observation of visual alterations responsible for changes in the report while minimizing noise and excluding isolated points. Specifically, the absolute difference between the two images is first computed, followed by the application of a Gaussian blur with a kernel size of $H \times W$ and a threshold L to suppress noise in the difference map. To identify abnormalities that are semantically represented in the image, we extract the contours of isolated pixels, group them into connected components, and retain only the most significant ones by selecting contours with the largest areas. Finally, the difference frame is constructed by retaining the top K major components. An example of the entire workflow is illustrated in Fig. 5.

4 Experiments

In this section, we first outline the experimental settings, followed by the presentation of results. These results include the effectiveness of CVLM in explaining report contents, comparisons of explanation approaches, and ablation studies.
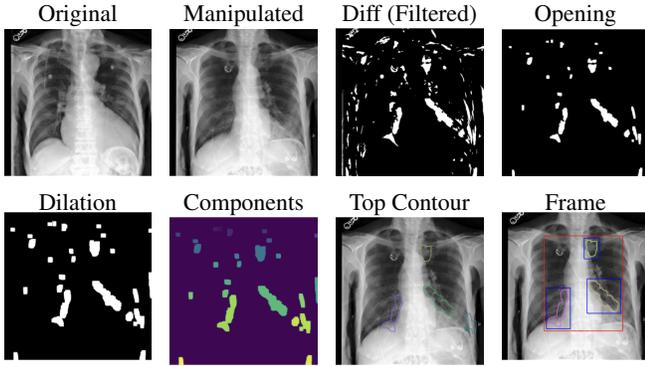
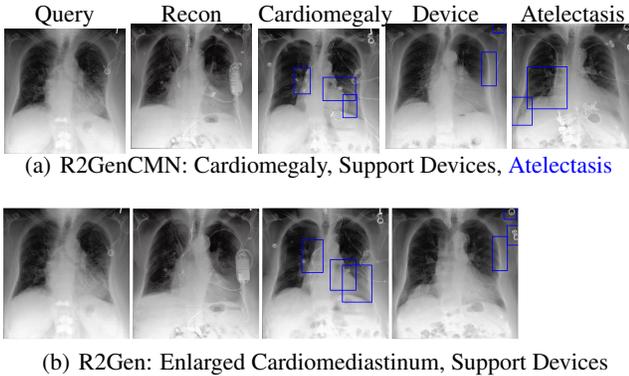Figure 5: Image processing pipeline in chest X-ray analysis.

(a) R2GenCMN: Cardiomegaly, Support Devices, Atelectasis

(b) R2Gen: Enlarged Cardiomediastinum, Support Devices

Figure 6: Explanation results for the same query X-rays with two different report generators. "(0)" denotes that the finding has been removed in the regenerated report; blue text indicates false positives against the ground-truth report.

4.1 Experimental Setting

Dataset and Report generators
In this paper, we developed and evaluated CVLM to decode two report generators, R2Gen [Chen *et al.*, 2020] and R2GenCMN [Chen *et al.*, 2022], in order to detect the visual features utilized by each within X-ray images. For each CVLM, we prepared the training dataset using MIMIC-CXR [Johnson *et al.*, 2019], as it was also used to train both report generators. The dataset comprises 473,057 chest X-ray images and 206,563 paired reports from 63,478 patients. Following the methodology in the two prior works, we utilized a subset of the dataset, consisting of 270,790 X-rays, to train CVLM. A validation set of 2,130 X-rays was used to select the optimal image generation model, while a test set of 3,858 images and reports was used to generate the corresponding counterfactual images.

Implementation Details
For developing CVLM, we froze the text encoders and trained the diffusion model using a batch size of 8 and a learning rate of 5×10^{-5} on a single A6000 GPU with 40 GB of memory. The model was trained for 100k steps over approximately one week. The final model for cyclic counterfactual generation was selected based on the highest PSNR of the reconstructed images on the validation set. For counterfactual generation,

the DDIM step was set to 50. For frame mask generation, a Gaussian blur of size 5×5 was applied, and the threshold was set to 95 ± 10, with the best value selected for each manipulated finding. The number of preserved tokens, K, was fixed at 5. More details are provided in the online appendix[1].

Evaluation Methods
We conduct a thorough evaluation of the proposed method in the following four steps: (1) We first propose a metric called ***Cyclic Rate*** (CR) to quantitatively assess the effectiveness of CVLM in generating cyclic counterfactual images for explanation. CR is defined as the success rate of achieving the intended manipulation in the regenerated report. Specifically, CR is calculated as the ratio of counterfactual images generated from reports where a finding is removed, and whose regenerated report successfully reflects the removal of that finding, to the total number of manipulations performed; (2) Next, we apply the feature identification framework to decode the visual features utilized by two individual report generators and compare the features they utilize; (3) To further illustrate the superiority of CVLM in identifying fine-grained and verifiable features for generated contents, we compare our proposed difference detection frame (introduced in Section 3.2) with a widely used model-agnosic explanation method Cross-Attention Map and the explanation results from a state-of-the-art self-explainable report generator RGRG [Tanida *et al.*, 2023]. The results are shown in Fig. 6. (4) Finally, we conduct ablation studies on the Cyclic Rate of CVLM, analyzing its performance with respect to inference times (T), training durations, and the use of raw reports versus structured reports.

4.2 Results

Success Rate of Cyclic Counterfactual Generation
In Table 1, we first investigate the impact of different training durations in the cyclic rate (CR) schedule of CVLM on generating cyclic counterfactual explanations. We then compare the CR settings for R2Gen and R2GenCMN. Specifically, we evaluate models that achieve the highest reconstruction quality (measured by PSNR) and those that achieve the lowest FID score. Our results show that the model with the best reconstruction ability, when paired with the generated text, leads to the highest cyclic manipulation effectiveness in explaining the reports for both R2Gen and R2GenCMN. We therefore select this checkpoint for report explanation. Both models achieve a success rate around 0.7, with CVLM for R2Gen showing a higher manipulation success rate overall.

Comparison of Different Report Generators
We present the visual explanation results from the cyclic counterfactual X-rays in Fig. 6 for R2Gen and R2GenCMN respectively. Specifically, we remove the abnormalities from their generated reports and generate the counterfactuals respectively, and resend the counterfactual images to their respective report generators to see if the abnormalities have been removed in the generated report. For the query X-ray in Fig. 6, R2GenCMN detected three abnormalities, while R2Gen detected two. Both models successfully removed the

---

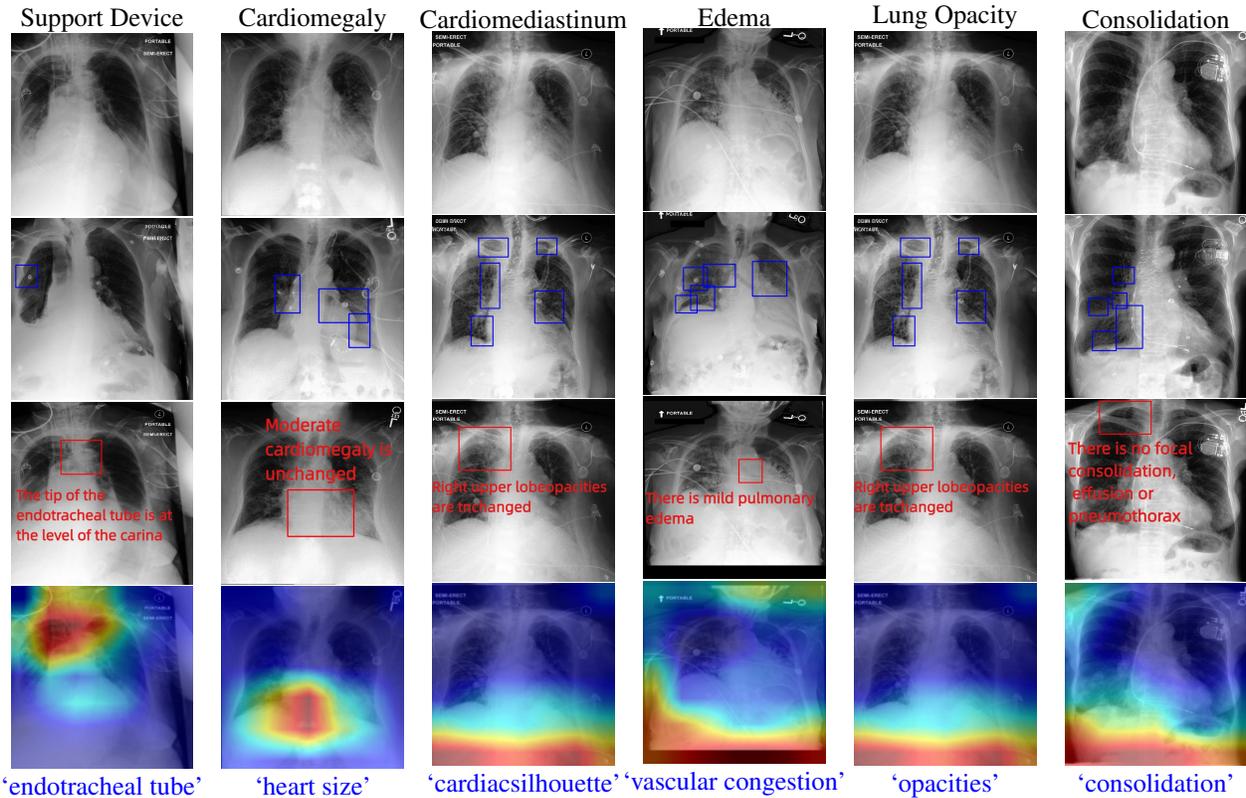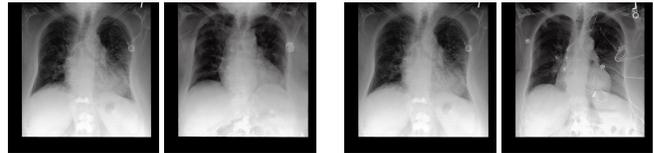[1]The appendix for this work is available at https://arxiv.org/abs/2411.05261.

Figure 7: Qualitative comparison with cross attention and RGRG methods on the MIMIC-CXR dataset. 1st row is the initial images, 2nd row is the counterfactuals generated by CVLM method, 3rd row is the images with the bounding box and the text generated by RGRG method, and 4th row is the heatmap with the attention entities (the blue text) generated by the cross attention method. Note: The counterfactuals in the figure all achieve the cyclic manipulation in the regenerated report. In this image, CVLM and attention are model-agnostic explanation methods applied to the findings in the generated reports from R2GenCMN, which are confirmed to exist in the GT report. In contrast, RGRG are explainable report generator which generates its report separately based on specific regions.

Remove_success	GT	Model_16k	Model_46k
R2GenCMN	0.655	0.690	0.595
Remove_success	GT	Model_14k	Model_42k
R2Gen	0.703	0.712	0.665

Table 1: Success rate of CVLM in explaining R2GenCMN and R2Gen across training checkpoints selected by PSNR and FID. Success rates are calculated over 569 findings from 400 images. 14k and 16k denote best-PSNR checkpoints; other iterations are selected based on lowest FID.

(a) Remove the "cardiomegaly" sentence from the raw report.

(b) Remove "cardiomegaly" from the reorganized report.

Figure 8: Manipulation from stable diffusion trained with raw reports and cleaned reports, respectively. For both, the manipulation is removing the contents about the existence of cardiomegaly from the prompt. The organized prompt which focuses on the findings brings significant change while the cardiomegaly is removed.

findings in their respective report generator outputs, and we can clearly observe the visual features contributing to the generated findings in their reports.

As "Atelectasis" is the point of disagreement between the two models, we included the opinions of two radiologists to provide their judgment and identify the features indicative of its presence. We found that their assessments and the identified features closely matched the features detected by R2GenCMN. This suggests that R2GenCMN utilizes clinically recognized features for generating this finding, even though the ground-truth report from the dataset did not in-

clude it. This also indicates that the explanations provided for the generated reports not only validate the findings of the models but also help identify potential missing findings in reports produced by humans. By presenting the generated report alongside its feature maps, our approach offers a more reliable and safer method for utilizing the generated reports in clinical practice. More cases are given in Appendix B[1].

Benchmark against State-of-the-art Explanations
Fig.7 shows the different explanations generated for various abnormalities. Compared to the cross-attention method, our approach produces more accurate localization of the major findings. The heatmaps produced by the cross-attention method appear unstable; for instance, the findings for cardiomediastinal silhouette, opacity, vascular congestion, and consolidation in Fig.7 are not localized within the lung areas.

The RGRG method provides reasonably interpretable results by associating findings with each anatomical region in its frame. However, upon examining the contents of its frame, it appears to have lower sensitivity to abnormalities, likely due to errors introduced by the pretrained detection model. While RGRG achieves internal interpretability, its framework cannot be adapted to other report generators that utilize different models or training datasets. In contrast, our proposed method offers precise localization explanations across various report generation models, as demonstrated in Fig. 6.

Additionally, CVLM not only demonstrates the ability to identify fine-grained features within X-rays but also effectively manipulates features occurring in multiple locations, such as opacity. It can also detect large-scale global abnormalities, such as enlarged cardiomediastinum, significantly surpassing the feature identification capabilities of existing explanation methods.

Ablation Study
We compared the manipulation method within CVLM to a direct report manipulation approach, where Stable Diffusion is trained directly with unprocessed reports without precleaning. The results in Fig. 8 indicate that using organized prompts focused on findings results in significantly better performance compared to removing entire sentences from unstructured reports. For quantitative comparison, out of a total of 400 manipulations, the structured report manipulation achieved a much higher success rate (0.712) than the direct report manipulation approach (0.449) for the R2Gen model.

Finally, we analyzed the influence of the denoising step T during diffusion model inference on achieving cyclic success. We evaluated T values ranging from 20 to 100 and found that $T > 50$ yields the highest success rates. For efficiency, $T = 50$ was selected. The computation cost and numerical results are provided in Appendix B[1].

4.3 Discussion

Broader Applications
While there is an abundance of X-ray report generators, existing evaluation metrics primarily focus on assessing the accuracy of the generated language and the detected abnormalities. Our proposed method introduces a novel metric by enabling the verification of identified features, facilitating the comparison of image features used for the generated findings in the report. This provides a unique approach to evaluating the trustworthiness of the generated content from different report generators by examining whether they rely on clinically consistent image features. The proposed method is modality-agnostic and can be easily extended to report generators for

---

[1]The appendix for this work is available at https://arxiv.org/abs/2411.05261.

---

other imaging modalities, such as MRI and CT scans, to improve the explainability of generated reports.

Limitation and Future Work
Currently, our method relies on CheXbert, which classifies only 13 diseases for reorganizing the report. This limitation restricts CVLM to explaining these 13 specific keywords. However, given the success of large language models (LLMs), such as GPT-4 [OpenAI, 2023], in accurately organizing reports, as demonstrated in a recent work [Zhang *et al.*, 2024], we can easily extend the proposed method to handle open vocabulary explanations by leveraging LLMs to extract a broader list of abnormalities from the generated reports, thereby enhancing the flexibility and applicability of CVLM. Besides, we will include more human experts and return them the findings from different report generators and their corresponding visual contexts, to further benchmark the reliability of the existing report generators.

5 Conclusion

In this paper, we propose a counterfactual generation method for query X-rays input to the report generator. These counterfactual X-rays modify specific visual features, resulting in the disappearance of certain major findings in the originally generated reports. This approach provides users with insights into the underlying feature patterns utilized by the report generators. Our method enhances the transparency of automated reports generated by existing report generators, serving as a valuable tool for experts to understand and evaluate the trustworthiness of AI-based reports. By improving interpretability and reliability, our approach strengthens confidence in the use of automated reports in real-world applications.

Contribution Statement

Yingying Fang and Zihao Jin (denoted by * on the first page) contributed equally to the conceptualization, methodology, investigation, and writing of this work. Shaojie Guo and Jinda Liu supported the benchmarking and visualization. Zhilin Yue contributed to the methodology and visualization. Yijian Gao, Junzhi Ning, and Zhili contributed to the review and editing of the manuscript. Simon Walsh provided clinical evaluation and review. Guang Yang (the corresponding author, denoted by † on the first page) contributed to the methodology, as well as the review and editing of the manuscript.

References

[Abdal *et al.*, 2022] Rameen Abdal, Peihao Zhu, John Femiani, Niloy Mitra, and Peter Wonka. Clip2stylegan: Unsupervised extraction of stylegan edit directions. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–9, 2022.

[Alfarghaly *et al.*, 2021] Omar Alfarghaly, Rana Khaled, Abeer Elkorany, Maha Helal, and Aly Fahmy. Automated radiology report generation using conditioned transformers. *Informatics in Medicine Unlocked*, 24:100557, 2021.

[Atad *et al.*, 2022] Matan Atad, Vitalii Dmytrenko, Yitong Li, Xinyue Zhang, Matthias Keicher, Jan Kirschke, Bene Wiestler, Ashkan Khakzar, and Nassir Navab. Chexplaining in style: Counterfactual explanations for chest x-rays using stylegan. *arXiv preprint arXiv:2207.07553*, 2022.

[Brooks *et al.*, 2023] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.

[Cao *et al.*, 2023] Yiming Cao, Lizhen Cui, Lei Zhang, Fuqiang Yu, Zhen Li, and Yonghui Xu. Mmtn: multimodal memory transformer network for image-report consistent medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 277–285, 2023.

[Chase Walker *et al.*, 2024] Dominic Simon Chase Walker, Kenny Chen, and Rickard Ewetz. Attribution quality metrics with magnitude alignment. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 530–538, 2024.

[Chen *et al.*, 2020] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020.

[Chen *et al.*, 2022] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. *arXiv preprint arXiv:2204.13258*, 2022.

[Chen *et al.*, 2023] Wenting Chen, Xiang Li, Linlin Shen, and Yixuan Yuan. Fine-grained image-text alignment in medical imaging enables cyclic image-report generation. *arXiv preprint arXiv:2312.08078*, 2023.

[Fang *et al.*, 2024a] Yingying Fang, Zihao Jin, Xiaodan Xing, Simon Walsh, and Guang Yang. Decoding decision reasoning: A counterfactual-powered model for knowledge discovery. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024.

[Fang *et al.*, 2024b] Yingying Fang, Shuang Wu, Zihao Jin, Shiyi Wang, Caiwen Xu, Simon Walsh, and Guang Yang. Diffexplainer: Unveiling black box models via counterfactual generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 208–218. Springer, 2024.

[Fontanella *et al.*, 2023] Alessandro Fontanella, Antreas Antoniou, Wenwen Li, Joanna Wardlaw, Grant Mair, Emanuele Trucco, and Amos Storkey. Acat: Adversarial counterfactual attention for classification and detection in medical imaging. *arXiv preprint arXiv:2303.15421*, 2023.

[He *et al.*, 2024] Sunan He, Yuxiang Nie, Zhixuan Chen, Zhiyuan Cai, Hongmei Wang, Shu Yang, and Hao Chen. Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. *arXiv preprint arXiv:2404.15127*, 2024.

[Hertz *et al.*, 2022] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control.(2022). *URL https://arxiv. org/abs/2208.01626*, 2022.

[Johnson *et al.*, 2019] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

[Karras *et al.*, 2020] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Proc. NeurIPS*, 2020.

[Kim *et al.*, 2022] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2426–2435, 2022.

[Lang *et al.*, 2021] Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T Freeman, Phillip Isola, Amir Globerson, Michal Irani, et al. Explaining in style: Training a gan to explain a classifier in stylespace. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 693–702, 2021.

[Lee *et al.*, 2023] Suhyeon Lee, Won Jun Kim, Jinho Chang, and Jong Chul Ye. Llm-cxr: Instruction-finetuned llm for cxr image understanding and generation. *arXiv preprint arXiv:2305.11490*, 2023.

[Liang *et al.*, 2023] Kaizhao Liang, Xu Cao, Kuei-Da Liao, Tianren Gao, Wenqian Ye, Zhengyu Chen, Jianguo Cao, Tejas Nama, and Jimeng Sun. Pie: Simulating disease progression via progressive image editing. 2023.

[Liu *et al.*, 2019] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, pages 249–269. PMLR, 2019.

[Liu *et al.*, 2024] Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. Bootstrapping large language models for radiology report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18635–18643, 2024.

[Lyu *et al.*, 2023] Yueming Lyu, Tianwei Lin, Fu Li, Dongliang He, Jing Dong, and Tieniu Tan. Deltaedit: Exploring text-free training for text-driven image manipulation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6894–6903. IEEE, 2023.

[Maksudov *et al.*, 2025] Bulat Maksudov, Kathleen Curran, and Alessandra Mileo. Towards generating more interpretable counterfactuals via concept vectors: a preliminary study on chest x-rays. *arXiv preprint arXiv:2506.04058*, 2025.

[Mertes *et al.*, 2022] Silvan Mertes, Tobias Huber, Katharina Weitz, Alexander Heimerl, and Elisabeth André. Ganterfactual—counterfactual explanations for medical non-experts using generative adversarial learning. *Frontiers in artificial intelligence*, 5:825565, 2022.

[Müller *et al.*, 2024] Philip Müller, Georgios Kaissis, and Daniel Rueckert. Chex: Interactive localization and region description in chest x-rays. *arXiv preprint arXiv:2404.15770*, 2024.

[OpenAI, 2023] OpenAI. Gpt-4 technical report. https://arxiv.org/abs/2303.08774, 2023. arXiv:2303.08774.

[Patashnik *et al.*, 2021] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2085–2094, 2021.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Reynaud *et al.*, 2022] Hadrien Reynaud, Athanasios Vlontzos, Mischa Dombrowski, Ciarán Gilligan Lee, Arian Beqiri, Paul Leeson, and Bernhard Kainz. D'artagnan: Counterfactual video generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 599–609. Springer, 2022.

[Rodriguez *et al.*, 2021] Pau Rodriguez, Massimo Caccia, Alexandre Lacoste, Lee Zamparo, Issam Laradji, Laurent Charlin, and David Vazquez. Beyond trivial counterfactual explanations with diverse valuable explanations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1056–1065, 2021.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[Sankaranarayanan *et al.*, 2022] Swami Sankaranarayanan, Thomas Hartvigsen, Lauren Oakden-Rayner, Marzyeh Ghassemi, and Phillip Isola. Real world relevance of generative counterfactual explanations. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*, 2022.

[Schutte *et al.*, 2021] Kathryn Schutte, Olivier Moindrot, Paul Hérent, Jean-Baptiste Schiratti, and Simon Jégou. Using stylegan for visual interpretability of deep learning models on medical images. *arXiv preprint arXiv:2101.07563*, 2021.

[Singla *et al.*, 2023] Sumedha Singla, Motahhare Eslami, Brian Pollack, Stephen Wallace, and Kayhan Batmanghelich. Explaining the black-box smoothly—a counterfactual approach. *Medical Image Analysis*, 84:102721, 2023.

[Smit *et al.*, 2020] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*, 2020.

[Spinks and Moens, 2019] Graham Spinks and Marie-Francine Moens. Justifying diagnosis decisions by deep neural networks. *Journal of biomedical informatics*, 96:103248, 2019.

[Tanida *et al.*, 2023] Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7433–7442, 2023.

[Tanyel *et al.*, 2023] Toygar Tanyel, Serkan Ayvaz, and Bilgin Keserci. Beyond known reality: Exploiting counterfactual explanations for medical research. *arXiv preprint arXiv:2307.02131*, 2023.

[Wachter *et al.*, 2017] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[Wang *et al.*, 2023] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11558–11567, 2023.

[Wang *et al.*, 2024] Jun Wang, Abhir Bhalerao, Terry Yin, Simon See, and Yulan He. Camanet: class activation map guided attention network for radiology report generation. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[Zhang *et al.*, 2023] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[Zhang *et al.*, 2024] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Jiayu Lei, Ya Zhang, Yanfeng Wang, and Weidi Xie. Radgenome-chest ct: A grounded vision-language dataset for chest ct analysis. *arXiv preprint arXiv:2404.16754*, 2024.