Zero-Shot Medical Phrase Grounding with Off-the-shelf Diffusion Models

Konstantinos Vilouras ⓘ, Pedro Sanchez ⓘ, Alison Q. O'Neil ⓘ, and Sotirios A. Tsaftaris ⓘ, *Senior Member, IEEE*

Abstract— **Localizing the exact pathological regions in a given medical scan is an important imaging problem that traditionally requires a large amount of bounding box ground truth annotations to be accurately solved. However, there exist alternative, potentially weaker, forms of supervision, such as accompanying free-text reports, which are readily available. The task of performing localization with textual guidance is commonly referred to as phrase grounding. In this work, we use a publicly available Foundation Model, namely the Latent Diffusion Model, to perform this challenging task. This choice is supported by the fact that the Latent Diffusion Model, despite being generative in nature, contains cross-attention mechanisms that implicitly align visual and textual features, thus leading to intermediate representations that are suitable for the task at hand. In addition, we aim to perform this task in a zero-shot manner, i.e., without any training on the target task, meaning that the model's weights remain frozen. To this end, we devise strategies to select features and also refine them via post-processing without extra learnable parameters. We compare our proposed method with state-of-the-art approaches which explicitly enforce image-text alignment in a joint embedding space via contrastive learning. Results on a popular chest X-ray benchmark indicate that our method is competitive with SOTA on different types of pathology, and even outperforms them on average in terms of two metrics (mean IoU and AUC-ROC). Source code will be released upon acceptance at `https://github.com/vios-s`.**

Index Terms— **Deep Learning, Diffusion Models, Medical Imaging, Phrase Grounding, Zero-shot learning**

I. INTRODUCTION

THE rapid success of deep learning over the last few years has led to powerful data-driven models being deployed in real-world scenarios. Recently, by taking advantage of the scaling properties of popular deep learning methods
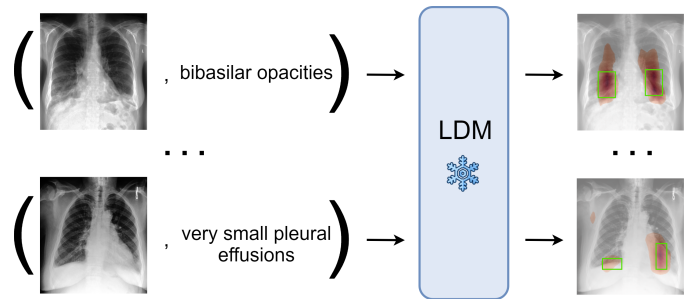
Fig. 1: High-level description of the zero-shot phrase grounding task. Given input pairs of an image (chest X-ray) and its accompanying text prompt, we leverage cross-modal feature alignment mechanisms within a frozen Latent Diffusion Model (LDM) to extract heatmaps, which indicate the regions where image and text are maximally aligned. Then, we evaluate the generated heatmaps based on ground truth bounding boxes (shown in green) for pathology detection. Our method, thus, is an illustration of using pre-trained LDMs for downstream applications in a zero-shot setting.

both in terms of learnable parameters and training data, we witness the era of Foundation Models (FMs) [1], i.e., large-scale neural networks that were trained on massive amounts of data. FMs have unprecedented capabilities: they can be readily applied to a wide variety of tasks as off-the-shelf solutions, or they can serve as a robust basis for training models for specific, potentially unseen, tasks and, plausibly, modalities (e.g., transferring knowledge from natural images to the medical domain). Among their many benefits, FMs provide machine learning practitioners and researchers with a universal tool that enables the widespread application of data-driven solutions to multiple scientific fields, as well as the development of a sound theoretical framework around well-known deep learning methodologies.

In this work, we investigate a specific type of FM: the Latent Diffusion Model (LDM) [2]. The LDM belongs to the class of probabilistic models and is considered one of the most groundbreaking methods for image synthesis. The versatile design of the LDM has significantly contributed towards its success, as external sources of information (e.g., text, segmentation masks, or any other type of a representation) can be easily incorporated into the model without any architectural changes. Here, we draw inspiration from research studies showing that diffusion models can solve downstream tasks

such as classification [3], [4] and segmentation [5] with little to no additional fine-tuning on target data. We attempt to validate this finding i.e., the re-usability of diffusion models for downstream tasks, in the medical imaging domain, using a model pretrained on medical images (i.e. chest X-rays) with readily available associated text (i.e. radiology reports) as conditioning.

In order to leverage a model trained with text as conditioning, we reframe the objection detection task as one of *phrase grounding*. In the context of diffusion models, instead of training a model that predicts bounding box coordinates [6], we propose to leverage cross-modal feature fusion mechanisms within an off-the-shelf LDM to directly perform phrase grounding, relying on the model having learned accurate fine-grained image-text alignment. Moreover, we consider the case of *zero-shot task transfer* [7] in our work, since the LDM's pre-training task (text-guided image generation) differs from the target task (phrase grounding). However, in contrast to [7] which requires further training using supervision across multiple tasks, we use an off-the-shelf model (LDM) with frozen weights to solve the task at hand.

Prior works have defined the phrase grounding task in multiple ways. Along with input image-text pairs, most methods further require either ground truth bounding boxes [8] or object detection models [9] during training, allowing test-time selection of the region proposal that most closely matches the input text. On the contrary, we recognise that manual ground truth bounding box annotations and external pathology detection models are typically difficult to acquire. Instead, we opt for an end-to-end system that extracts relevant information from natural language (e.g., location and severity modifiers of the underlying pathology) and, in turn, associates clinical findings with visual features corresponding to specific image regions. A high-level overview of our method for the phrase grounding task is shown in Fig. 1.

There are emergent properties of the LDM that are useful for the given task. First, the core of the LDM is a U-Net architecture [10] which, in turn, is equipped with inductive biases (e.g., multi-scale hierarchical feature learning) suitable for localisation tasks such as phrase grounding. This differentiates our method from other baselines that use image classification models such as ResNets [11], [12]. In addition, the LDM offers a sophisticated feature fusion mechanism: fusion layers (cross-attention) are incorporated at multiple levels of the architecture, and visual features evolve over time via the diffusion process. Therefore, this mechanism is expected to yield more refined representations compared to late fusion alone [11].

We perform extensive experiments on an established phrase grounding benchmark dataset, i.e., MS-CXR [11]. The results suggest that, despite being inherently a generative model, the LDM has learned high quality features for the task at hand. Our proposed approach, which tackles the extreme case where no additional fine-tuning is performed (we refer to it as a *zero-shot* scenario), yields, perhaps surprisingly, a highly competitive method that proves to be state-of-the-art in terms of two metrics (mean IoU and AUC-ROC) on average across 8 pathology labels.

Overall, our **contributions** can be summarised as follows:
- Building on the Latent Diffusion Model architecture, we gather semantically meaningful visual features both from multiple timesteps of the diffusion process and from various cross-attention layers. We target those layers since they inherently align information from the visual and textual stream, thus being suitable for the phrase grounding task.
- In contrast to conventional sampling methods for Latent Diffusion Models that utilise classifier-free guidance, our approach involves sampling from the unconditional model, while the conditional model (conditioned on a text input) is used merely to extract cross-attention maps.
- We perform extensive experiments on a medical dataset that provides ground truth bounding boxes for evaluation and test our method against several strong baselines. Results indicate that our method shows competitive performance against, and even exceeds in two metrics, state-of-the-art methods without any fine-tuning strategies.
- We perform an ablation study to justify the hyperparameter choices in our system.
- We further qualitatively analyze our proposed method, as well as the strongest available baselines, to provide useful insights for those approaches.

II. RELATED WORK

We now summarise the most relevant research. We start by mentioning prior influential works showing that diffusion models can be effectively applied to multiple downstream tasks in a zero- or few-shot setting. Next, due to their similarity to our approach, we also present methods tailored for image editing tasks that utilise Latent Diffusion Models. Then, we briefly discuss some of the most popular methods for phrase grounding in natural RGB images. Finally, we shift our focus to approaches related to the medical imaging domain, against which we compare our proposed method.

A. Downstream Application of Diffusion Models

Diffusion models have been successfully applied to a wide range of tasks. For instance, in the context of medical imaging, there exist works that train diffusion models from scratch to perform lesion localisation [13], [14], anomaly detection [15], and counterfactual generation [16], [17]. Recent studies have also revealed the ability of diffusion models to perform fairly well on downstream tasks with minimal supervision. For example, for classification tasks it has been shown that the posterior $p(\mathbf{c}|\mathbf{x})$ for all candidate classes $\mathbf{c}$ can be estimated from a diffusion model's residual errors at a given timestep without any requirement for further hyperparameters or training [3], [4]. Similarly, for segmentation, Baranchuk *et al.* [5] use intermediate visual representations extracted from a diffusion model. In this case, a few additional labelled images are required to train a shallow network that outputs pixel-wise predictions. In another work, Zhao *et al.* [18] experiment with the text feature extraction pipeline, as well as the choice of intermediate visual features, and train lightweight task-specific models for segmentation and depth estimation, respectively.

B. Manipulating Attention in Latent Diffusion Models

Closely related to our approach, yet developed for the task of image editing, recent studies use attention maps extracted from a Latent Diffusion Model to control the generated images. For example, Hertz *et al.* [19] showed that it is possible to apply global and local edits to an image by processing the cross-attention maps generated via source and target prompts, respectively. Patashnik *et al.* [20] aim to vary the shape of a target object in a generated image, and they use both cross- and self-attention maps to better preserve the shape and appearance for the rest of the image. Conversely, Tumanyan *et al.* [21] argue that only convolutional and self-attention features are useful for editing, since localised visual information is not described in text prompts.

C. Visual Phrase Grounding

Phrase grounding is a cross-modal reasoning task referring to the spatial localisation of objects present in an image given a relevant text description. Due to the widespread availability of image-text data sources, there already exist large-scale end-to-end models for the natural image domain.

In a discriminative learning scenario, ViLBERT [22] introduces the concept of co-attention, i.e., exchanging information between modalities within the transformer layers, and these co-attention mechanisms can be used to directly relate visual and textual tokens. Early fusion strategies are also adopted by MDETR [23] which enforces image-text alignment with appropriate learning objectives, and GLIP [24] in which image and language encoders are simultaneously trained to correctly assign a word to a specific image region.

In a generative learning scenario, Chen and Li [25] train a text-guided diffusion model to gradually recover ground truth bounding boxes from their noise-perturbed versions. On the other hand, Tang *et al.* [26] focus on a standard image generation model and demonstrate how various linguistic aspects of the input text prompt affect the generated image, by extracting and visualising the cross-attention heatmaps; this allows interpretation of stable diffusion's image generation process. They further evaluate use of the heatmaps for segmenting the objects, a localisation task which has parallels with phrase grounding, albeit the target image is a result of the text prompt rather than a parallel input. This approach is similar in principle to ours, but for the opposite process of denoising (corresponding to image generation).

D. Medical Visual Phrase Grounding

In the medical imaging domain, phrase grounding is considered a difficult task due to the inherent variation in textual information; radiologists commonly use domain-specific terms, describe the absence of pathological findings (e.g. "*No pneumothorax or pleural effusion*"), or use phrases that convey a level of uncertainty (e.g. "*Blunting of the right costophrenic angle is consistent with a small right pleural effusion*").

Earlier works have taken different approaches. Bhalodia *et al.* [27] extract pneumonia-related attributes from radiology reports, while a pre-trained bounding box detector is used to extract regions of interest (ROIs) and their associated features.

Then, using both streams of information, their system is trained to correctly classify attributes from visual features, as well as to maximise the similarity for a given image-text pair.

End-to-end discriminative methods to date are largely fully supervised [28], [29] or rely on self-supervised contrastive formulations [11], [12], [30]. Specifically, Chen *et al.* [28] train a vision-language Transformer model to directly predict bounding boxes, whereas Xu *et al.* [29] gather publicly available labelled chest X-ray datasets and train a single model via multi-task learning. LIMITR [30] is a self-supervised method that aligns local cross-modal representations that are further weighted via learnable significance scores. Lastly, BioViL [11] is an end-to-end model with a BERT text encoder finetuned on radiology reports that is optimised via both global and local cross-modal contrastive losses. Bannur *et al.* [12] extend the BioViL model to support longitudinal information across patients, and the resulting system (BioViL-T) achieves state-of-the-art performance on medical phrase grounding.

In contrast, we adopt a generative approach to this task. A concurrent work also applies diffusion models to phrase grounding [31]. Our approach differs in several ways: First, we use a pre-trained, publicly available diffusion model and propose mechanisms to perform phrase grounding whilst keeping the model frozen, whereas [31] focuses on training the model from scratch. Second, to define these mechanisms, we pay attention to how we select cross-attention layers and timesteps, whereas [31] simply average across both time and layers. Finally, we compare our method with more recent state-of-the-art and other, recently proposed, baselines. Nevertheless, the findings of [31] are beneficial for our work. They show that learning both text and visual encoders simultaneously severely degrades performance. In addition, regarding the textual information that is available for phrase grounding, they show that using sentences from the original radiology reports (as is the case, for example, for the MS-CXR [11] dataset) leads to the best overall results, outperforming both synthetically generated text (with ChatGPT) and the simplest case of using the class string as the input prompt.

III. METHODOLOGY

We adopt the following notation: x is a scalar, x denotes a vector, $\mathbf{x}$ is a tensor, $\mathcal{X}$ denotes a vector space, $\mathbf{f} : \mathcal{X} \rightarrow \mathcal{X}$ is the mapping between two vector spaces that is performed by a neural network $\mathbf{f}$, and $\mathbf{f}_\theta$ refers to the total number of learnable parameters (weights) in the network. We also consider a dataset $\mathcal{D} = \{\mathbf{x_i}, p_i\}_{i=1}^N$ with a total of N image-text pairs. For simplicity, we will next present our pipeline while considering a single input $(\mathbf{x}, p)$; the extension to a batch is trivial.

A. Overview of the Latent Diffusion Model

We follow the standard setup of the LDM architecture as proposed in [2]. The core components of the LDM are as follows. First, the visual encoder $\mathbf{E} : \mathcal{X} \rightarrow \mathcal{Z}$ yields a compressed latent representation z_0 of the input image x. Then, in the first part of the diffusion process (also called the forward process), Gaussian noise is added to z_0. This process
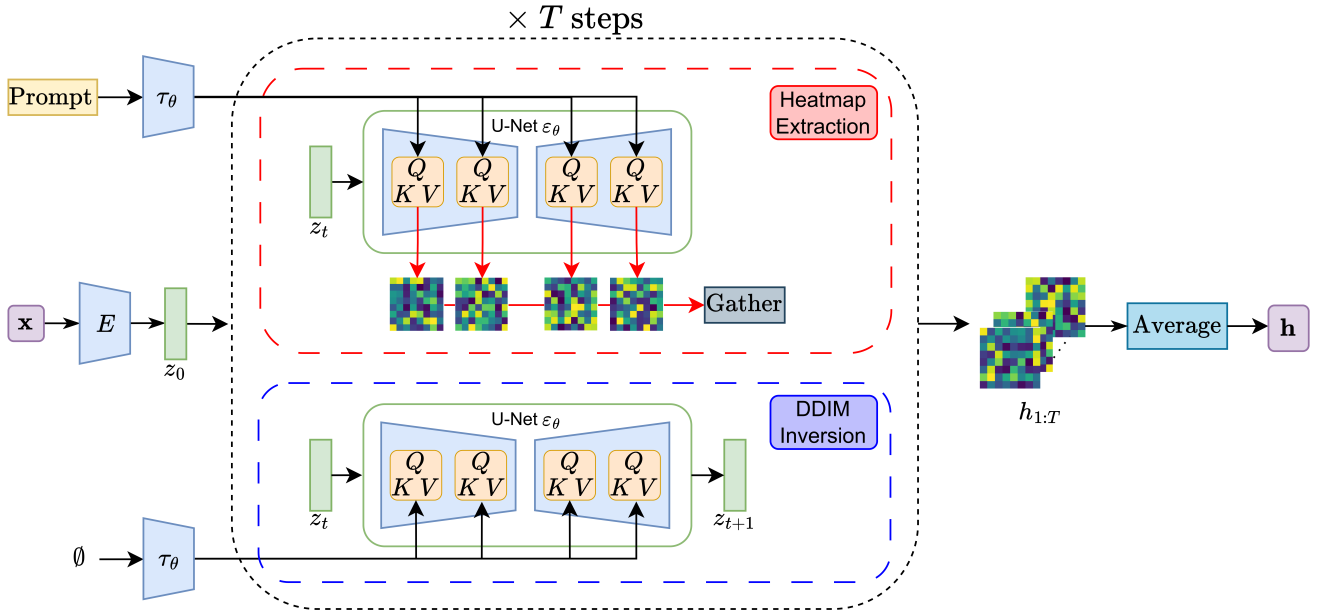
Fig. 2: Overview of our proposed phrase grounding pipeline based on the Latent Diffusion Model [2]. The input image-text pair is first processed via the encoders E and $\boldsymbol{\tau_\theta}$, respectively. Then, at each timestep of the diffusion process $t = 1, ..., T$, we gather cross-attention maps from the U-Net $\boldsymbol{\epsilon_\theta}$. The output heatmap $\mathbf{h}$ is generated by averaging the gathered attention maps.

is iterative, meaning that noise is repeatedly added over time to z_t ($t = 1, ..., T$), leading to a final latent representation z_T that is pure noise. Next, in the second part of the diffusion process (i.e., the reverse process), the goal is to repeatedly denoise the latent representations z_t ($t = T, ..., 1$) until z_0 is recovered. The denoising task is carried out by a U-Net model $\boldsymbol{\epsilon_\theta} : \mathcal{Z} \to \mathcal{Z}$ that predicts the noise introduced at timestep t. Finally, a visual decoder $\mathbf{D} : \mathcal{Z} \to \mathcal{X}$ generates the reconstructed image $\tilde{\mathbf{x}}$.

B. Image-Text Alignment via Cross-Attention

To aid reconstruction, information about the image $\mathbf{x}$ in text form is also injected in the U-Net model $\boldsymbol{\epsilon_\theta}$ via the cross-attention layers, which ultimately capture interactions between the visual and textual streams. Concretely, the prompt p that accompanies the image $\mathbf{x}$ is first split into S tokens and then passed through a text encoder τ_θ. The resulting text features serve as context in each cross-attention layer of the U-Net, i.e., they are used to generate keys $\mathbf{k}$ and values $\mathbf{v}$, whereas queries $\mathbf{q}$ are extracted from visual features. Intuitively, queries $\mathbf{q}$ provide a high-level description of the visual content per spatial location, keys $\mathbf{k}$ provide a high-level description of the textual content per token, and values $\mathbf{v}$ contain more detailed textual features per token. The cross-attention operation is defined in Eq. (1) below, where $\mathbf{P_K}$ and $\mathbf{P_V}$ denote linear projection layers for the keys and values, respectively, and d is a scaling factor.

$$\mathbf{k} = \mathbf{P_K}(\tau_\theta(p)) \qquad \mathbf{v} = \mathbf{P_V}(\tau_\theta(p))$$

$$\underbrace{A_p = \text{softmax}(\frac{\mathbf{q}\mathbf{k}^\top}{\sqrt{d}})}_{\text{cross-attention map}} \qquad \text{Attention}(\mathbf{q}, \mathbf{k}, \mathbf{v}) = A_p \cdot \mathbf{v}. \quad (1)$$

At each layer and timestep, cross-attention maps $A_p \in \mathbb{R}^{S \times B \times B}$ are computed according to Eq. (1), where S is the number of prompt tokens and B is the spatial size (either 16 or 32, depending on the layer's level in the U-Net architecture). Let N denote the total number of cross-attention layers in the U-Net and T the number of diffusion timesteps. That leaves us with $N \times T$ attention maps in total, where each one varies depending on the visual features (low-level features or higher-level semantics) and the noise level in each latent z_t. The extracted cross-attention maps are then used for phrase grounding.

C. Zero-shot Phrase Grounding with the LDM

We use a pre-trained LDM to perform zero-shot phrase grounding. More specifically, given an image and a prompt as input, we extract intermediate cross-attention maps from the U-Net and aggregate them into a heatmap $\mathbf{h}$. In the ideal case of perfect image-text alignment, $\mathbf{h}$ will be highly activated on the regions of interest specified by the prompt. An overview of our approach is presented in Fig. 2.

Specifically, our method is based on DDIM inversion [32] which maps input features z_0 into noise z_T. Note that we avoid classifier-free guidance during sampling as it would increase the accumulated error introduced by DDIM inversion [33]. To this end, we sample z_{t+1} from the LDM using only the unconditional inputs: the latent representation z_t and a null prompt (the empty string $\emptyset$). However, we gather the cross-attention maps using the previously sampled z_t and the conditioning input (prompt p). Our method is summarised in Alg. 1.

Using Alg. 1, we gather a total of $N \times T$ cross-attention maps $\mathbf{A}$. However, inspired by a prior work on semantic

---

Algorithm 1 Extracting cross-attention maps via DDIM inversion.

Require: image x, prompt p, encoder $\mathbf{E}$, U-Net $\epsilon_{\boldsymbol{\theta}}$, timesteps T, noise schedule parameters α_t

 1: $z_0 \leftarrow \mathbf{E}(x)$ ▷ initial visual features
 2: $\mathbf{A} \leftarrow []$ ▷ gathered cross-attention maps
 3: **for** $t \leftarrow 0$ **to** $T-1$ **do**
 4: $\quad \epsilon_{\emptyset}, A_{\emptyset,t} \leftarrow \boldsymbol{\epsilon_{\theta}}(z_t, t, \emptyset)$ ▷ unconditional path
 5: $\quad \epsilon_p, A_{p,t} \leftarrow \boldsymbol{\epsilon_{\theta}}(z_t, t, p)$ ▷ conditional path
 6: $\quad \mathbf{A} \leftarrow [\mathbf{A}, A_{p,t}]$ ▷ add current map to the stack
 7: $\quad z_{t+1} \leftarrow \sqrt{\alpha_{t+1}}\left(\frac{z_t - \sqrt{1-\alpha_t}\epsilon_{\emptyset}}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t+1}}\epsilon_{\emptyset}$
 8: $\quad z_t \leftarrow z_{t+1}$ ▷ update latents
 9: **end for**
10: **return** $\mathbf{A}$

---

segmentation with diffusion models [5], we aim to optimise phrase grounding performance by selecting attention maps only from *middle timesteps* and *middle cross-attention layers*. This choice is motivated by the following two observations. First, for $t \to T$, latents z_t are highly noisy and lose information about the image's global structure, thus attention maps do not convey any semantic meaning. On the contrary, for $t \to 0$, z_t contains fine-grained details about the input which, according to [5], hurt segmentation performance. Therefore, focusing on middle timesteps strikes a balance. Correspondingly, earlier (resp. final) layers are more informative for small (resp. large) objects in the image, meaning that those features would either activate over the entire image or only on large anatomical structures. Therefore, selecting the middle cross-attention layers balances between these two extreme cases.

The selected cross-attention maps are intensity-normalised to span the range [0,1] and resized to match the image's spatial dimensions with bilinear interpolation. The output heatmap $\mathbf{h}$ is formed by averaging maps across layers, timesteps and tokens. Denoting the selected cross-attention layers, timesteps, and all tokens excluding padding as L', T', S', respectively, Eq. (2) outlines this operation.

$$\mathbf{h} = \mathop{\mathbb{E}}_{l \in L'} \mathop{\mathbb{E}}_{t \in T'} \mathop{\mathbb{E}}_{s \in S'} \text{Resize}\left(\text{Norm}\left(\mathbf{A}_{s,\cdot,\cdot,l,t}\right)\right) \quad (2)$$

As a final step, results are refined by applying additional post-processing techniques on the generated heatmap $\mathbf{h}$ that do not involve any learnable parameters or any form of supervision, thus not violating the zero-shot scenario. Here, we use binary-Otsu thresholding [34] to separate the more strongly activating foreground from the more weakly activating background. The output binary mask is then applied to the heatmap $\mathbf{h}$ to suppress weak signals while leaving the foreground activations unaffected.

IV. EXPERIMENTAL RESULTS

A. LDM Pre-training Setup

First, we briefly discuss the LDM pre-training stage. Here, we use an open-source implementation of the LDM [35] along with a provided checkpoint. The model is pre-trained on MIMIC-CXR [36], a large-scale dataset of chest radiographs

accompanied by free-text reports. The training set consists of 368,960 chest X-ray images, whereas the text prompt for each image is a randomly sampled sentence from the corresponding radiology report, from either the *Impressions* or *Findings* section. Note that the pre-trained weights, as well as the configuration files with all training-related details, are publicly available in the MONAI Generative framework [37].

B. LDM Phrase Grounding Setup

We now instantiate our best performing setup for the phrase grounding task. We transform the input image into fixed spatial dimensions of 512×512. For the diffusion process, we set the total number of timesteps of the inverse DDIM scheduler to T = 300. In terms of text processing, we use the standard CLIP text encoder with frozen weights. Prior to encoding, the input prompt is tokenised and then either padded or truncated to match the maximum sequence length ($S = 77$ tokens). Since the CLIP text encoder is not trained on radiology reports, domain-specific medical terms are out-of-vocabulary (OOV); in practice this includes all of the pathology names except "pneumonia". The CLIP tokeniser handles any unknown OOV word by splitting it into multiple known sub-tokens.

The initial size of the gathered attention maps is $\mathbb{R}^{S \times B \times B \times L \times T}$, where $L = 11$ is the total number of cross-attention layers in the U-Net and $B \in \{16, 32\}$ denotes their spatial size (depending on the layer's level in the U-Net). After the selection stage described in Section III-C, we end up with attention maps collected from $L = 4$ layers (i.e., from the 3rd, 4th, 6th and 7th cross-attention layers), which automatically sets their spatial size to $B = 16$, and also from 60 different timesteps (i.e., from timestep 120 to 180) out of $T = 300$ steps in total. These choices are justified via the ablation study in Subsection IV-H. Note, however, that we did not perform an exhaustive search to find the optimal combination of selected layers and timesteps; in fact, it is possible that different settings per pathology might lead to better performance (we leave this for future work). Last, following [11], [12], the resulting heatmap $\mathbf{h}$ with resolution 512×512 is convolved with a Gaussian kernel ($\sigma = 2.5$) prior to Otsu thresholding.

C. Evaluation Dataset

We evaluate our proposed system on the MS-CXR benchmark [11] which consists of 1,158 image-sentence pairs with ground truth bounding boxes indicating the pathology. Note that this dataset is extracted from the official MIMIC-CXR test set. We further pre-process the dataset by merging entries corresponding to the same patient and the same sentence, i.e., an image might have more than one ground truth bounding box as reference for a given text prompt, where the pathology appears in multiple locations. We made the choice to compute metrics on a per-image basis (and not per bounding box) as this matched the performance reported by [11], [12].

D. Baselines

We compare our proposed method to 4 state-of-the-art baselines trained in a discriminative manner, using either

fully supervised or self-supervised learning. Fully supervised networks (*MedRPG* [28], *OmniFM-DR* [29]) are trained to correctly predict the ground truth bounding box, whereas the self-supervised networks (*BioViL* [11], *BioViL-T* [12]) are trained to maximise the cosine similarity between visual and textual features derived from the corresponding image and report. We evaluate our system against these 4 methods only since they were shown to outperform previous approaches, with *BioViL-T* [12] setting the current state-of-the-art on phrase grounding. For all methods, we use the publicly available pre-trained model checkpoints provided by the respective authors.

E. Metrics

Given a predicted heatmap H (with values in $[0, 1]$ range) and ground truth binary segmentation mask M_{GT} which has ones within each bounding box area and zero otherwise, phrase grounding performance is measured via the following evaluation metrics:

1) Mean Intersection over Union (mIoU): mIoU is a standard metric to evaluate segmentation performance. The predicted binary mask at threshold thr is defined as $M_H = \{h \in H : h > thr\}$. Then, IoU at given threshold thr is calculated as

$$IoU_{@thr} = \frac{|M_H \cap M_{GT}|}{|M_H \cup M_{GT}|}. \tag{3}$$

Here, following [11], we calculate mIoU as the average over 5 different thresholds $thr \in [0.1, 0.2, 0.3, 0.4, 0.5]$.

2) Area Under ROC Curve (AUC-ROC): The area under the receiver operating characteristic curve is another method to estimate segmentation performance on a per-pixel basis given H and M_{GT}.

3) Contrast-to-Noise Ratio (CNR): CNR as used in [11] is a threshold-agnostic measure that reflects the distribution of raw heatmap activations over the entire input. Let A denote the area within each bounding box, whereas $\bar{A}$ is the rest of the heatmap, i.e., $H = A \cup \bar{A}$. Then, after calculating the mean μ and variance σ^2 of the raw heatmap scores for each area A and $\bar{A}$, respectively, CNR is defined as

$$CNR = \frac{\mu_A - \mu_{\bar{A}}}{\sqrt{\sigma_A^2 + \sigma_{\bar{A}}^2}}. \tag{4}$$

Note that this definition will penalise the case where $\mu_A < \mu_{\bar{A}}$. However, following [11], we also provide results for the absolute CNR evaluated as

$$|CNR| = \frac{|\mu_A - \mu_{\bar{A}}|}{\sqrt{\sigma_A^2 + \sigma_{\bar{A}}^2}}. \tag{5}$$

F. Evaluation Protocol

To ensure a fair comparison across all approaches, we adopt the following protocol:

- All metrics are computed on the original image dimensions. Therefore, for heatmap-based methods such as [11], [12], and our own method, we first perform nearest neighbor interpolation with appropriate zero padding to match the original image resolution. Note that each of those methods initially generates a fixed size heatmap

which matches the respective input size: for BioViL [11] the resolution is 480×480, for BioViL-T [12] it is equal to 448×448, while our LDM-based approach outputs a 512×512 heatmap per image. For methods [28] and [29] that predict bounding box coordinates, the predicted bounding box regions are resized to the width and height of the original image.

- Since heatmaps generated with *BioViL* [11] and *BioViL-T* [12] methods are in the $[-1, 1]$ range (due to cosine similarity), whereas our method yields heatmaps in the $[0, 1]$ range, we set all negative values for the aforementioned baselines [11], [12] to 0.
- Both *MedRPG* [28] and *OmniFM-DR* [29] are Transformer models that predict bounding box coordinates. Thus, we evaluate [28], [29] only in terms of mIoU.
- *MedRPG* [28] does not support the case where more than one bounding box exists per image. Therefore, for those images, we evaluate it on each bounding box separately and only report the maximum value of mIoU per image.

G. Results and Discussion

In this section we report results using the 4 aforementioned metrics (mIoU, AUC-ROC, |CNR| and CNR), and also discuss implications. Overall phrase grounding results on the MS-CXR database are reported in Table I. In Subsection IV-H we show the results of a separate ablation study.

We now provide an interpretation of the results shown in Table I. More specifically, we draw the following conclusions: First, our proposed method outperforms both supervised baselines MedRPG [28] and OmniFM-DR [29] by a large margin. This also holds for the baselines [11] and [12] trained on image-text pairs via self-supervision. This suggests that phrase grounding performance is largely affected by the size of the pre-training dataset, while bounding box annotations are typically scarce.

Second, our method is competitive with both BioViL variants [11], [12] for most pathologies on the MS-CXR dataset. In fact, our pipeline based on the LDM sets a new state-of-the-art in terms of both mIoU (0.9 % relative improvement to BioViL-T [12]) and AUC-ROC (0.2 % relative increase to BioViL-T [12]) metrics averaged across all classes. Note also that both BioViL models [11], [12] use radiology-specific text encoders which are expected to further improve performance, whereas our method relies on a frozen CLIP text encoder pre-trained on data collected from the Internet. In addition, unlike all other approaches that use discriminative models such as ResNets [11], [12] (or Transformers [28], [29]), the LDM is based on a U-Net for feature extraction, thus its representations are readily applicable to localisation tasks such as phrase grounding. We also observe that the results exhibit high variance across all methods. This is likely a data related issue and needs to be further investigated in the future.

Furthermore, given both definitions of the CNR metric presented in Eqs. (4) and (5), respectively, our method remains fairly robust between the two. Specifically, our approach yields the lowest difference between |CNR| and CNR for 7 out of 8 pathologies. This highlights that |CNR|, which is defined

TABLE I: Phrase grounding results on MS-CXR dataset. Mean results, and their respective standard deviation, are reported across the 8 pathologies of interest and also averaged (Avg). N denotes the sample size per pathology. Both |CNR| and CNR metrics are unbounded, i.e., in $(-\infty, \infty)$ range. mIoU and AUC-ROC are reported in (%). Higher is better for all metrics (↑). Note that, due to implementation details, only mIoU can be computed for methods [28] and [29]. Best metrics are highlighted with **bold**. Second best metrics are underlined.

Method	Metric	Pneumonia (N=182)	Pneumothorax (N=243)	Consolidation (N=117)	Atelectasis (N=61)	Edema (N=44)	Cardiomeg. (N=333)	Lung Opac. (N=82)	Pleural Eff. (N=96)	Avg
MedRPG [28]	mIoU	11.3	5.8	6.1	12.2	5.1	16.3	6.5	11.7	9.4
OmniFM-DR [29]	mIoU	12.6	7.0	24.7	5.4	17.0	$\underline{30.9}$	15.4	5.1	14.8
BioViL [11]	|CNR|	$\underline{1.56}_{0.80}$	$\underline{0.78}_{0.56}$	$\mathbf{1.79}_{0.77}$	$\underline{1.37}_{0.65}$	$\mathbf{0.85}_{0.57}$	$0.81_{0.54}$	$\underline{1.24}_{0.81}$	$\underline{1.38}_{0.71}$	$\underline{1.22}_{0.35}$
	CNR	$\underline{1.49}_{0.89}$	$\underline{0.63}_{0.71}$	$\underline{1.73}_{0.83}$	$\underline{1.28}_{0.77}$	$\mathbf{0.77}_{0.60}$	$0.73_{0.64}$	$\underline{1.18}_{0.88}$	$\underline{1.33}_{0.78}$	$\underline{1.14}_{0.37}$
	mIoU	$\underline{27.3}_{16.0}$	$\underline{10.2}_{10.2}$	$\mathbf{31.8}_{14.3}$	$24.1_{14.2}$	$\underline{21.3}_{17.2}$	$22.0_{18.6}$	$15.0_{14.1}$	$\mathbf{20.4}_{11.6}$	$21.5_{6.31}$
	AUC-ROC	$76.5_{16.4}$	$\underline{66.3}_{17.5}$	$\underline{83.5}_{12.1}$	$\underline{76.4}_{15.1}$	$\underline{65.1}_{13.1}$	$64.5_{14.3}$	$68.9_{17.2}$	$\mathbf{76.6}_{15.8}$	$72.2_{6.50}$
BioViL-T [12]	|CNR|	$\mathbf{1.70}_{0.70}$	$\mathbf{1.01}_{0.63}$	$\mathbf{1.79}_{0.76}$	$\mathbf{1.47}_{0.66}$	$\underline{0.84}_{0.46}$	$\mathbf{1.06}_{0.50}$	$\mathbf{1.59}_{0.87}$	$\mathbf{1.55}_{0.67}$	$\mathbf{1.38}_{0.33}$
	CNR	$\mathbf{1.66}_{0.77}$	$\mathbf{0.91}_{0.75}$	$\mathbf{1.74}_{0.82}$	$\mathbf{1.45}_{0.69}$	$\mathbf{0.77}_{0.50}$	$\mathbf{1.05}_{0.52}$	$\mathbf{1.54}_{0.94}$	$\mathbf{1.53}_{0.70}$	$\mathbf{1.33}_{0.34}$
	mIoU	$\mathbf{29.0}_{13.9}$	$\mathbf{12.6}_{12.6}$	$\underline{30.2}_{12.9}$	$\underline{24.7}_{12.8}$	$19.0_{15.1}$	$23.5_{15.3}$	$\mathbf{17.5}_{12.6}$	$18.6_{10.2}$	$\underline{21.9}_{5.65}$
	AUC-ROC	$\mathbf{80.1}_{15.0}$	$\mathbf{70.1}_{18.0}$	$\mathbf{83.9}_{11.7}$	$\underline{76.4}_{14.4}$	$63.0_{11.9}$	$66.1_{12.6}$	$\underline{76.3}_{16.1}$	$73.8_{15.0}$	$\underline{73.7}_{6.58}$
Ours	|CNR|	$1.02_{0.46}$	$0.46_{0.31}$	$1.18_{0.55}$	$1.06_{0.50}$	$0.76_{0.36}$	$\underline{0.91}_{0.43}$	$1.12_{0.57}$	$0.88_{0.43}$	$0.92_{0.22}$
	CNR	$1.02_{0.47}$	$-0.08_{0.53}$	$1.16_{0.57}$	$1.06_{0.51}$	$0.71_{0.45}$	$\underline{0.90}_{0.44}$	$1.08_{0.63}$	$0.86_{0.46}$	$0.84_{0.37}$
	mIoU	$23.8_{11.9}$	$5.30_{5.30}$	$24.7_{14.4}$	$\mathbf{25.0}_{11.1}$	$\mathbf{30.5}_{19.6}$	$\mathbf{37.4}_{11.6}$	$\underline{16.7}_{13.0}$	$\underline{19.3}_{9.93}$	$\mathbf{22.8}_{8.94}$
	AUC-ROC	$\underline{78.9}_{9.78}$	$49.5_{14.8}$	$81.1_{9.98}$	$\mathbf{79.1}_{10.1}$	$\mathbf{72.4}_{11.4}$	$\mathbf{75.3}_{10.0}$	$\mathbf{79.6}_{13.0}$	$\underline{75.2}_{10.7}$	$\mathbf{73.9}_{9.62}$

in [11] and [12], overestimates performance, thus it is less reliable.

We also observe that every method performs poorly on the *Pneumothorax* class (our approach leads to a negative CNR value). We note that pneumothorax causes a dark air space (i.e., a region with low intensity pixels) in the position of the collapsed lung, as opposed to the other pathologies which manifest as "bright" regions. These dark regions may be more difficult to differentiate from normal lung.

H. Ablation Study

In Table II we provide an ablation study showing how different hyperparameters affect phrase grounding performance. To speed up experiments, we set the total number of timesteps to T=100, yet the empirical observations are expected to also hold for larger T. Starting from the initial setup of collecting all attention maps (L=11, T=100), we focus either on middle cross-attention layers (L=6) or middle timesteps (T=20). We see that both of the aforementioned choices have a positive impact on both metrics. Furthermore, we observe that the combination thereof (L=6, T=20) yields a substantial boost in performance compared to the initial setup (L=11, T=100). We also show the effect of using binary Otsu thresholding. In fact, although we notice a slight decrease in CNR, mIoU is increased by 3.5%. Note also how different pathologies might benefit from different setups.

Moreover, the ablation study presented in Table III shows the impact of prompt tokens on phrase grounding performance. To this end, while fixing all other components of our system (L=6, T=20 out of 100 timesteps in total), we discard cross-attention maps that are not related to pathology tokens i.e., the (sub-)tokens corresponding to the pathology label names. Also, for this experiment, we filtered out inputs that do not contain the pathology name in the prompt, thus reducing the dataset to 694 image-text pairs. The results in Table III

indicate that cross-attention maps corresponding to pathology tokens are not sufficient to perform phrase grounding (interestingly, when we only use pathology tokens, performance on *Pneumothorax* class slightly improves; however, this approach clearly underperforms on all other classes). This can be attributed to the fact that text prompts usually contain additional important information such as location (e.g., *right*, *left*, *bibasilar*) and severity (e.g., *mild*, *moderate*, *severe*) modifiers that can be used to localise the underlying pathology.

I. Qualitative Analysis

Fig. 3 depicts non cherry-picked examples from the MS-CXR dataset and the generated heatmaps for BioViL [11], BioViL-T [12], and our method. We observe that both BioViL models provide more densely localised results compared to our system, which activates on larger input areas. However, unlike BioViL(-T), our method does not miss an area of interest (cf. third row of Fig. 3 for pathology *Lung Opacity*). We note that our LDM-based method could also be focusing on less relevant anatomical regions in some cases (cf. first row of Fig. 3 for *Pneumothorax*) that can be recognised as easy false positives.

V. CONCLUSION

In this work we have presented a novel approach for performing phrase grounding with a pre-trained Latent Diffusion Model. In fact, we draw on the parts of the model that integrate visual and textual features, namely the cross-attention layers. These layers, as evidenced by our results, provide a rich source of information that can be readily used to solve the task at hand. Our proposed method does not alter the backbone generative model in any way, thus operating in zero-shot.

Our proposed system is limited by the computational cost of the LDM sampling process (one inference per timestep), which leads to slower inference speed compared to other baselines. We also identified certain pathologies (e.g., *Pneumothorax*)

TABLE II: Ablation study on the choice of layers L and timesteps T, as well as the effect of applying Otsu thresholding. Note that T=20 refers to the $[40, 60]$ range out of 100 steps in total, while L=6 includes layers with index in $\{3, ..., 8\}$ range out of 11 in total. CNR metric is unbounded, mIoU is in (%). Best metrics are highlighted with **bold**.

Setup	Metric	Pneumonia	Pneumothorax	Consolidation	Atelectasis	Edema	Cardiomegaly	Lung Opacity	Pleural Eff.	Avg
L=11, T=100	CNR	0.87	-0.15	0.90	0.98	0.56	0.52	0.90	**0.89**	0.68
	mIoU	15.0	5.4	16.3	16.5	25.2	22.9	10.8	13.9	15.7
L=11, T=20	CNR	0.89	-0.15	0.94	0.97	0.66	0.60	0.91	0.88	0.71
	mIoU	16.1	5.3	17.3	17.2	25.7	24.3	11.8	14.0	16.5
L=6, T=100	CNR	0.89	-0.08	0.98	**1.03**	0.58	0.72	0.98	**0.89**	0.75
	mIoU	15.7	5.5	17.0	17.2	25.7	25.7	11.1	14.4	16.6
L=6, T=20	CNR	**0.92**	**-0.05**	1.02	**1.03**	**0.67**	0.79	**0.99**	**0.89**	**0.78**
	mIoU	15.8	**5.7**	17.4	17.0	26.8	27.5	11.6	13.9	17.0
L=6, T=20, Otsu	CNR	0.89	-0.14	1.00	1.00	0.62	**0.81**	0.96	0.85	0.75
	mIoU	**20.5**	5.1	**21.4**	**22.2**	**28.5**	**33.3**	**14.4**	**17.8**	**20.4**

TABLE III: Ablation study on the effect of the selected prompt tokens. In the first setup (*Pathology tokens*), cross-attention maps are extracted only from tokens related to the pathology, whereas the second setup (*All tokens*) considers the entire prompt. CNR metric is unbounded, mIoU is in (%). Best metrics are highlighted with **bold**.

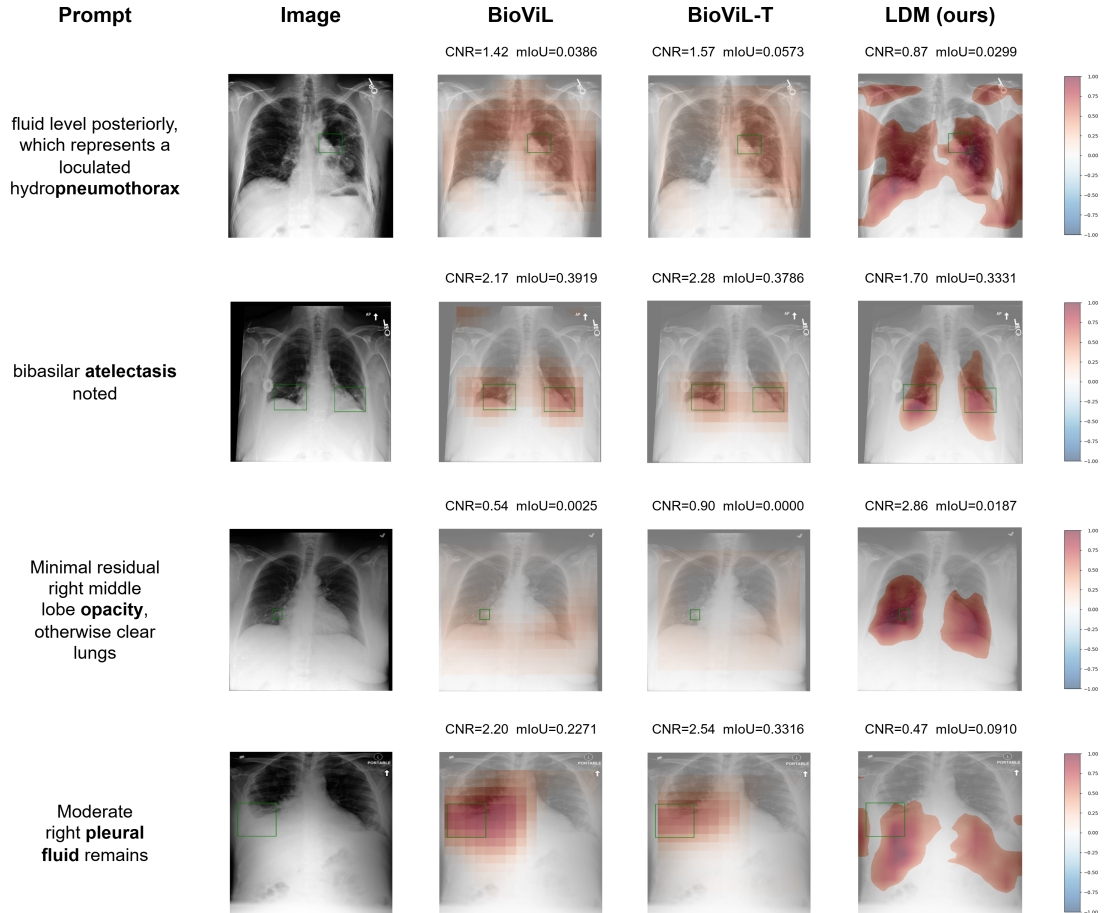| Setup	Metric	Pneumonia	Pneumothorax	Consolidation	Atelectasis	Edema	Cardiomegaly	Lung Opacity	Pleural Eff.	Avg
Pathology tokens	CNR	0.05	**0.25**	0.40	0.39	0.45	0.41	0.00	0.50	0.31
	mIoU	9.50	**6.76**	14.7	13.3	22.4	25.9	2.59	13.4	13.6
All tokens	CNR	**0.93**	-0.11	**1.12**	**1.04**	**0.64**	**0.96**	**0.59**	**0.89**	**0.76**
	mIoU	**21.8**	5.07	**21.9**	**22.9**	**29.5**	**41.4**	**6.55**	**17.8**	**20.9**

Fig. 3: Randomly selected results for the phrase grounding task. For each input image-prompt pair, we show the heatmaps generated from *BioViL* [11], *BioViL-T* [12] and our own method, respectively, overlaid on the original images. Ground truth classes are highlighted in **bold** within each prompt. Ground truth bounding boxes are depicted in green. For each method, we also provide the reported |CNR| and mIoU metrics (shown on top of each figure). Best viewed in colour.

where all models, including ours, underperform; this requires further analysis as it might indicate hidden data biases. We also expect that an LDM trained on chest X-rays will perform poorly in significantly different medical contexts (e.g., brain MRIs) without fine-tuning.

Regarding future work, it is worth experimenting with few-shot fine-tuning methods (e.g., low rank adaptation [38]) that would allow us to incorporate new knowledge to the pre-trained LDM, or even adapt it to unknown data distributions, with a small target dataset. In this direction, according to prior works in zero-shot domain adaptation, we would likely require access to either task-irrelevant target data [39] or target domain-specific prompts [40], [41] to make our zero-shot method robust to domain shifts. Moreover, devising faster sampling methods tailored for the phrase grounding task would render the LDM more efficient for real-world scenarios. We also believe that further improvements on the generative aspect of the LDM (e.g., mitigating various data biases [42]) will bring a positive effect on the model's downstream performance.

In terms of broader impact, our proposed framework might be used to automatically link reports to the relevant image locations, allowing fast inclusion of key images and easy navigation when reviewing a previous scan. We might also extend to the task of diagnosis by creating text prompts such as `"Where is {pathology_label}?"`, to achieve an off-the-shelf detector (beyond the scope of this paper).

REFERENCES

[1] R. Bommasani, D. A. Hudson, E. Adeli, *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[3] A. C. Li, M. Prabhudesai, S. Duggal, E. Brown, and D. Pathak, "Your diffusion model is secretly a zero-shot classifier," *arXiv preprint arXiv:2303.16203*, 2023.

[4] K. Clark and P. Jaini, "Text-to-image diffusion models are zero shot classifiers," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[5] D. Baranchuk, I. Rubachev, A. Voynov, V. Khrulkov, and A. Babenko, "Label-efficient semantic segmentation with diffusion models," *arXiv preprint arXiv:2112.03126*, 2021.

[6] S. Chen, P. Sun, Y. Song, and P. Luo, "Diffusiondet: Diffusion model for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 19 830–19 843.

[7] A. Pal and V. N. Balasubramanian, "Zero-shot task transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2189–2198.

[8] H. Zhao, J. T. Zhou, and Y.-S. Ong, "Word2pix: Word to pixel cross-attention transformer in visual grounding," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 2, pp. 1523–1533, 2022.

[9] X. Liu, S. Huang, Y. Kang, H. Chen, and D. Wang, "Vgdiffzero: Text-to-image diffusion models can be zero-shot visual grounders," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 2765–2769.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, Springer, 2015, pp. 234–241.

[11] B. Boecking, N. Usuyama, S. Bannur, *et al.*, "Making the most of text semantics to improve biomedical vision–language processing," in *European conference on computer vision*, Springer, 2022, pp. 1–21.

[12] S. Bannur, S. Hyland, Q. Liu, *et al.*, "Learning to exploit temporal structure for biomedical vision-language processing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 016–15 027.

[13] J. Wolleb, R. Sandkühler, F. Bieder, P. Valmaggia, and P. C. Cattin, "Diffusion models for implicit image segmentation ensembles," in *International Conference on Medical Imaging with Deep Learning*, PMLR, 2022, pp. 1336–1348.

[14] P. Sanchez, A. Kascenas, X. Liu, A. Q. O'Neil, and S. A. Tsaftaris, "What is healthy? generative counterfactual diffusion for lesion localization," in *MICCAI Workshop on Deep Generative Models*, Springer, 2022, pp. 34–44.

[15] W. H. Pinaya, M. S. Graham, R. Gray, *et al.*, "Fast unsupervised brain anomaly detection and segmentation with diffusion models," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2022, pp. 705–714.

[16] H. A. Bedel and T. Çukur, "Dreamr: Diffusion-driven counterfactual explanation for functional mri," *arXiv preprint arXiv:2307.09547*, 2023.

[17] P. Sanchez and S. A. Tsaftaris, "Diffusion causal models for counterfactual estimation," *arXiv preprint arXiv:2202.10166*, 2022.

[18] W. Zhao, Y. Rao, Z. Liu, B. Liu, J. Zhou, and J. Lu, "Unleashing text-to-image diffusion models for visual perception," *arXiv preprint arXiv:2303.02153*, 2023.

[19] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," *arXiv preprint arXiv:2208.01626*, 2022.

[20] O. Patashnik, D. Garibi, I. Azuri, H. Averbuch-Elor, and D. Cohen-Or, "Localizing object-level shape variations with text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 23 051–23 061.

[21] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel, "Plug-and-play diffusion features for text-driven image-to-image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1921–1930.

[22] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pre-training task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in neural information processing systems*, vol. 32, 2019.

[23] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, "Mdetr-modulated detection for end-to-end multi-modal understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1780–1790.

[24] L. H. Li, P. Zhang, H. Zhang, *et al.*, "Grounded language-image pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 965–10 975.

[25] S. Chen and B. Li, "Language-guided diffusion model for visual grounding," *arXiv preprint arXiv:2308.09599*, 2023.

[26] R. Tang, L. Liu, A. Pandey, *et al.*, "What the daam: Interpreting stable diffusion using cross attention," *arXiv preprint arXiv:2210.04885*, 2022.

[27] R. Bhalodia, A. Hatamizadeh, L. Tam, *et al.*, "Improving pneumonia localization via cross-attention on medical images and reports," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, Springer, 2021, pp. 571–581.

[28] Z. Chen, Y. Zhou, A. Tran, *et al.*, "Medical phrase grounding with region-phrase context contrastive alignment," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2023, pp. 371–381.

[29] L. Xu, Z. Ni, X. Liu, X. Wang, H. Li, and S. Zhang, "Learning a multi-task transformer via unified and customized instruction tuning for chest radiograph interpretation," *arXiv preprint arXiv:2311.01092*, 2023.

[30] G. Dawidowicz, E. Hirsch, and A. Tal, "Limitr: Leveraging local information for medical image-text representation," *arXiv preprint arXiv:2303.11755*, 2023.

[31] M. Dombrowski, H. Reynaud, J. P. Müller, M. Baugh, and B. Kainz, *Trade-offs in fine-tuned diffusion models between accuracy and interpretability*, 2023. arXiv: 2303.17908 [cs.CV].

[32] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[33] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, "Null-text inversion for editing real images using guided diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6038–6047.

[34] P.-S. Liao, T.-S. Chen, P.-C. Chung, *et al.*, "A fast algorithm for multilevel thresholding," *J. Inf. Sci. Eng.*, vol. 17, no. 5, pp. 713–727, 2001.

[35] W. H. Pinaya, P.-D. Tudosiu, J. Dafflon, *et al.*, "Brain imaging generation with latent diffusion models," in *MICCAI Workshop on Deep Generative Models*, Springer, 2022, pp. 117–126.

[36] A. E. Johnson, T. J. Pollard, N. R. Greenbaum, *et al.*, "Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs," *arXiv preprint arXiv:1901.07042*, 2019.

[37] W. H. Pinaya, M. S. Graham, E. Kerfoot, *et al.*, "Generative ai for medical imaging: Extending the monai framework," *arXiv preprint arXiv:2307.15208*, 2023.

[38] E. J. Hu, Y. Shen, P. Wallis, *et al.*, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[39] K.-C. Peng, Z. Wu, and J. Ernst, "Zero-shot deep domain adaptation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 764–781.

[40] M. Fahes, T.-H. Vu, A. Bursuc, P. Pérez, and R. De Charette, "Poda: Prompt-driven zero-shot domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 623–18 633.

[41] S. Yang, Z. Tian, L. Jiang, and J. Jia, "Unified language-driven zero-shot domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23 407–23 415.

[42] F. Pérez-García, S. Bond-Taylor, P. P. Sanchez, *et al.*, "Radedit: Stress-testing biomedical vision models via diffusion image editing," *arXiv preprint arXiv:2312.12865*, 2023.