This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Seeing the Trees for the Forest: Rethinking Weakly-Supervised Medical Visual Grounding

Ta Duc Huy[1] Duy Anh Huynh[1] Yutong Xie [2] Yuankai Qi[3] Qi Chen[1]
Phi Le Nguyen[4] Sen Kim Tran Son Lam Phung[5] Anton van den Hengel[1]
Zhibin Liao[1] Minh-Son To[6] Johan W. Verjans[1]
Vu Minh Hieu Phan [1*]

[1] Australian Institute for Machine Learning, University of Adelaide [2] Mohamed bin Zayed University of Artificial Intelligence
[3] Macquarie University [4] Hanoi University of Science and Technology [5] University of Wollongong [6] Flinders University

Abstract

Visual grounding (VG) is the capability to localize specific regions in an image associated with a particular text description. In medical imaging, VG enhances interpretability and cross-modal alignment of multimodal models by highlighting relevant pathological features corresponding to textual descriptions. Current vision-language models struggle to associate textual descriptions with disease regions due to inefficient attention mechanisms and a lack of fine-grained token representations. In this paper, we empirically demonstrate two key observations. First, current VLMs assign high norms to background tokens, diverting the model's attention from regions of disease. Second, the global tokens used for cross-modal learning are not representative of local disease tokens. This hampers identifying correlations between the text and disease tokens. To address this, we introduce simple, yet effective Disease-Aware Prompting (DAP) process, which uses the explainability map of a VLM to identify the appropriate image features. This simple strategy amplifies disease-relevant regions while suppressing background interference. Without any additional pixel-level annotations, DAP improves visual grounding accuracy by 20.74% compared to state-of-the-art methods across three major chest X-ray datasets.

1. Introduction

Visual grounding is a prominent research area at the vision-language intersection [15, 18, 33], enabling applications like visual question answering [24, 55], human-robot interaction [50, 54], and data annotation [26, 59]. Beside concept-based approaches [25], the ability of linking textual mentioned diseases to specific pathological image regions enhances interpretability [56], aiding radiologists in assessing model decisions. Additionally, VG integrates effectively with large language models (LLMs) to enhance multimodal understanding, e.g., refining 3D object localization through interleaved reasoning and grounding [68], summarizing long text with LLM to improve VG [69], and introducing iterative vision-language interaction with novel tokens for object detection (VG) and text generation [35]. In summary, VG enables transparency, assists localization, and enhances multi-modal interactions on the patch level of the emerging medical multi-modal models [12, 20, 41].

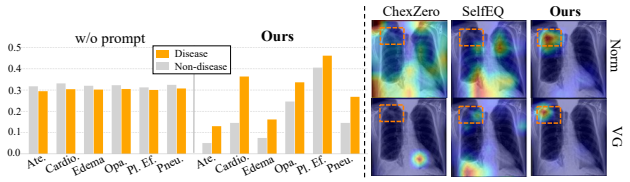


Figure 1. *Left*: Average cosine similarity between disease patch tokens and the corresponding disease texts of CheXzero [53] on MS-CXR [6] dataset. Without disease-aware prompting, the current VLM shows undifferentiated alignment scores between disease and non-disease patches with text. *Right*: Local feature norms, and visual grounding of CheXzero [53], Self-EQ [23], and our disease-aware prompted model.

Many methods [11, 22, 26, 41] perform *supervised* visual grounding on medical imaging. They rely on text-and-region correspondence annotations to train a cross-modal fusion module to ground text to specific disease regions. However, supervised approaches require dense annotations at the pixel level, which is expensive in the medical settings. Another work [5] proposes weakly-supervised VG for a pneumonia by using a pre-trained object detector.

Recently, weakly-supervised visual grounding methods [9, 30, 37–40, 46, 52] have leveraged strong vision-language models (VLMs) such as CLIP [45], as a backbone,

*Lead author.

and achieve a solid visual grounding performance without using text-to-region correspondence labels. Most methods obtain localization pseudo-labels using explainability methods [8, 36] applied to the vision-language models. Alternatively, mask-driven methods train a segmentation decoder to predict the foreground and background masks of CLIP [47, 48], or predict the masks used to mix two images [2]. Overall, current methods get the patch-level features from the pre-trained VLM backbone on the images then aim to improve local alignment. Unlike general images that typically contain multiple distinct objects , medical images typically contain more 'stuff'[7], and thus present a unique challenge: critical disease indicators often occupy only a small proportion of the image, while anatomical backgrounds dominate visual features. *The trees are lost in the forest*. As such, applied to medical imaging, current VLMs [4, 6, 44, 53] yield sub-optimal local features, which are less disease-focused.

In this work, we set out to analyze this limitation in current medical VLMs, identify whether it hinders visual grounding methods, and if so to propose improvements. Our analysis shows that: **1)** We observe that the L_2 norm of the token embeddings are high in background non-disease regions , as shown in Fig. 1. As the norm reflects token vector magnitude and representation strength, high background norms can divert attention from disease-relevant areas, impairing medical visual grounding. **2)** As disease regions are typically small, global image tokens are more strongly aligned with the major non-disease image patch tokens, and less representative of the fine-grained disease tokens, as shown in Fig. 8. As such, here is low alignment between disease image tokens and the global text and image tokens, which impairs visual grounding.

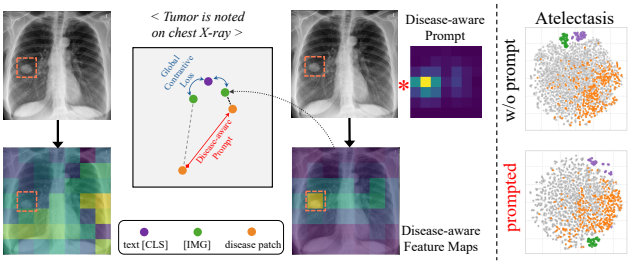


Figure 2. In current VLMs, text [CLS] (●) are pulled close to a global image token [IMG] (●) through contrastive loss, and thus they use the global token as a proxy to link with disease patch tokens (●). *Left*: w/o prompt, the global image token and the disease patch of the original input on the left show weak alignment, while our proposed *disease-aware prompting* enhances contrast for disease regions, aligning disease patches with both [IMG] and text [CLS]. *Right*: The original and prompted feature space for the images having disease 'atelectasis' on MS-CXR [6] dataset. The disease tokens are pulled towards the [IMG] upon prompted.

The two problems identified above illustrate the influence of non-disease background regions on the medical visual grounding challenge. To remedy their impact, we propose a simple-yet-effective *disease-aware feature prompting* approach, as illustrated in Fig. 2. We first extract the explainability map [8] of the pre-trained VLM for the given textual disease description. We then use this disease-aware map and prompt the vision models on the feature space, thus suppressing features of non-disease tokens and amplifying disease-related tokens. The model then extracts a global token, *i.e.* [IMG] token from this disease-focused feature map, bringing disease patches closer to the global image [IMG] token and improving its alignment with the text [CLS] token. This design is inspired by how radiologists modify the contrast to make the disease regions more salient and localize them. As shown in Fig. 1 (Right), our prompting strategy effectively assigns high norm values to disease regions, producing accurate visual grounding predictions. Without using any dense annotation, our method achieves superior performance compared to several state-of-the-art methods on multiple datasets.

Our main contributions are summarized as follows:

- We show that medical VLMs have high norm values in non-disease background regions, which misdirects the visual grounding of fine-grained disease regions.
- We further show that current VLMs suffer from *intra-modal misalignment*, where the global image token fails to capture the disease information from the local disease tokens. This weakens local patch-to-text alignment when training via global image-text contrastive losses.
- We introduce a simple-yet-effective *disease-aware feature prompting* for weakly-supervised medical visual grounding. Our technique suppresses the influence of non-disease local tokens to concentrate the model attention in relevant disease regions.
- Our proposed method achieves superior performance compared to several state-of-the-art methods for weakly-supervised medical visual grounding across three datasets. This paper is the first work, establishing a comprehensive medical weakly-supervised visual grounding benchmark on three datasets.

2. Related Works

Weakly-Supervised Visual Grounding aims to localize target objects mentioned in the text without segmentation labels. Detector-based methods [14, 21, 39, 40, 57, 58] adopts two-stage pipeline, using pre-trained object detectors to extract region of interests (RoIs), and then matching texts with the cropped RoI. A recent work [40] generates pseudo-captions for each RoI using BLIP, and matches between the text query and the pseudo-caption. RefCLIP [31], and QueryMatch [10] propose a one-stage pipeline, directly matching the text with the anchor point in the feature map or the DETR query. A state-of-the-art ground-

ing multimodal LLM [42] uses external segmentation models to train LLMs for grounding. Unfortunately, the reliability of medical-target detectors is not comparable with the general domain. In contrast, detector-free methods [2, 47, 48] leverage pseudo labels generated by the explainability map [8].G [48] and g++ [47] re-calibrate visual feature using the image-text similarity map. They also force the model to mask the background region to yield maximum similarity with the text. G++ [47] aggregates the self-attention map having high overlap with the grounding to create a more comprehensive pseudo-label. SelfEQ [23] encourages consistent visual grounding map with diverse text prompting. Existing WSVG methods rely on strong initial grounding map. Yet, current medical vision language models show low image-and-text alignment.

Vision-language models have garnered significant interest recently, largely since the success of CLIP [45]. Other notable VLMs include ALIGN [28], Florence [64], and Flamingo [1] have pushed forward research in this area. Building on these advances, several medical VLMs have emerged. These models learn the cross-modal information from the vision and language modalities such as chest radiographs and associated reports. CheXzero [53], MedCLIP [60], and BiomedCLIP [65] align images with raw medical reports with CLIP's contrastive learning objective, while BioVIL [6] and BioVIL-T [4] enhance the text encoding process. MedKLIP [61] and MAVL [44] leverage fine-grained report structures from RadGraph [27] to further improve performance. MedIM [62] introduces masked image modelling in medical VLMs. Given the fine-grained nature of medical images, where diseases often occupy small areas, current medical VLMs struggle to associate thefine-grained pathological regions with textual descriptions.

Prompt tuning is a strategy to adapt large-scale pre-trained models to new tasks. CoOp [67] and CoCoOp [66] prepend the text tokens with learnable parameters to adapt CLIP for many recognition tasks. The textual prompt tuning is widely used in open-vocabulary object detection [19] and semantic segmentation [63]. Several works [3, 29, 36] apply visual prompt tuning on the pixel space [3, 63] and feature space [29], to adapt a frozen pre-trained model. Previous works adopt learnable, and randomly initialized prompts for model adapting. In contrast, we adopt *informative* prompting, which enforces rich disease-aware prompts to guide medical visual grounding.

3. Proposed Method

Problem formulation. Given an input image $x_{\text{v}} \in \mathbb{R}^{C\times H\times W}$ and a text query x_{t} specifying an object in x_{v}, weakly supervised visual grounding aims to produce a segmentation map $y \in \mathbb{R}^{H\times W}$ for the referenced object, without dense segmentation labels. The visual grounding model leverages a VLM with an image encoder F_{v} and text encoder F_{t}, followed by a pixel decoder D. F_{v} provides a sequence of local visual token features $V = \{v_1, \ldots, v_n\}$ and the global image token `[IMG]` from x_{v} , while F_{t} yields text `[CLS]` token from x_{t}:

$$\{\texttt{[IMG]}, v_1, v_2, \ldots, v_n\} = F_{\text{v}}(x_{\text{v}}), \quad \texttt{[CLS]} = F_{\text{t}}(x_{\text{t}}). \tag{1}$$

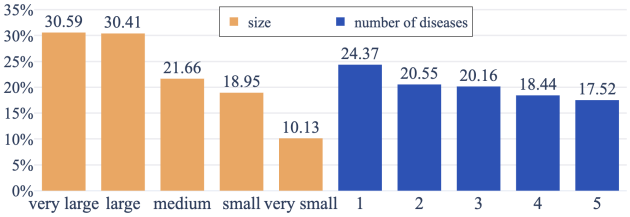


Figure 3. Performance analysis of BioVIL [6] on MS-CXR dataset as a function of (a) disease region size and (b) disease count per image. *Left:* Dice scores stratified by disease area, showing degraded performance for smaller regions. *Right:* Performance impact of concurrent diseases, demonstrating declining accuracy with increasing disease count. Disease areas are categorized into five groups by size percentiles.

3.1. Problem Analysis

Current VLMs [4, 44, 53] is efficient at identifying *what* diseases are present, *i.e*. image classification. However, it struggles in visual grounding, determining *where* to align text descriptions on the medical images. Here, we refer to the target disease regions as foreground (FG), and non-disease regions as background (BG). Fig. 1 (Left) shows the cosine similarity distributions of text tokens with the FG and BG tokens when using the state-of-the-art CheXzero [53]. It shows that BG tokens have a higher alignment with the text.

VLMs [4, 6, 53] are mostly trained using only the global image token `[IMG]`, *i.e*., extracting features from the entire image and aligning with the text `[CLS]`. However, medical images often feature *multiple* diseases with *small* regions, making it difficult to correctly assign text to the corresponding regions. This brings challenges in fine-grained visual grounding. Fig. 3 shows the visual grounding performance of the VL model on cases with multiple and small diseases. The Dice drops as the disease areas reduce, and the number of present diseases increases.

High norm on BG tokens, hampering the visual grounding capability. The L_2 norm of patch tokens $\|v\|_2$ indicates the activations of the local token v. Fig. 4 shows the norm of patch tokens of CheXzero. Tokens in BG regions are frequently activated with high values. Unlike natural images, medical images are highly canonical, with significant redundancy in BG areas. Hence, high-norm tokens often occur in these healthy BG regions rather than in the disease-associated FG areas. We then show that the high ac-

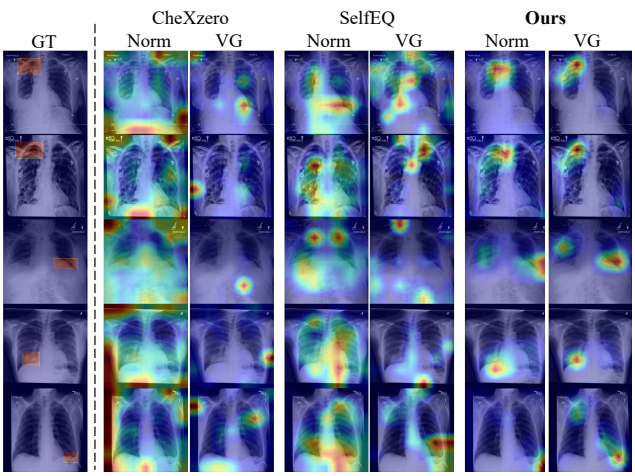


Figure 4. Comparative analysis of token norm distribution (Norm) and visual grounding (VG) performance across methods with representative cases show common chest pathologies. Left to right: Ground truth (GT), CheXzero [53] (foundational VLM), SelfEQ [23] (weakly-supervised baseline), and our disease-aware prompting approach. Our method produces more focused norm activations that better correspond to disease regions, leading to more precise grounding results. Colors indicate activation strength from low (blue) to high (red). Text descriptions are abbreviated.

tivation of BG tokens hampers the visual grounding. Fig. 4 further compares the norm map and the visual grounding of CheXzero [53]. The model tends to ground on background regions with high norm.

To quantify how spurious norm activation in the background (BG) regions affects visual grounding (VG), we calculate the L_2 norm of the tokens to form a norm map. We then evaluate the correlation between these norm maps and VG predictions. Fig. 5 shows two findings. *First*, VLMs show high norm activations in the background regions, evident by the high Dice score between the response map and the BG regions. *Second*, the Dice score between VG outputs on the norm maps is significantly higher than the Dice with the actual ground truth (GT). This shows that the models are inclined to ground on the highly activated regions of the norm map. These findings suggest that high norm activations in BG regions can divert the model away from disease areas, weakening VG accuracy.

Global image `[IMG]` token is non-representative of the local foreground disease tokens. Current VLMs mostly operate on the global image `[IMG]` token when aligning with the text token. Is `[IMG]` representative of local patch tokens v_i? This study shows a negative finding. Fig.6 investigates the intra-modal alignment between the global `[IMG]` token and foreground (FG) and background (BG) tokens using cosine similarity. In the top row, the `[IMG]` token aligns more closely with BG tokens than with FG tokens. This is further substantiated by the average cosine similarity between the global `[IMG]` token and individual

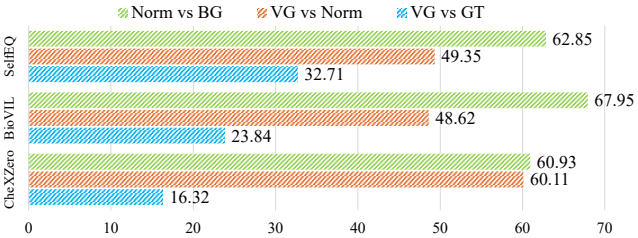


Figure 5. Dice score (%) between (i) Norm map and the background regions (BG); (ii) the visual grounding output (VG) and the norm map; and VG output and ground-truth (GT). Analysis of recent vision-language pre-training models [6, 53] and a recent weakly-supervised visual grounding method [23].

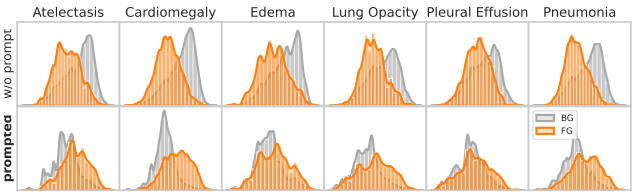


Figure 6. **Intra-modal alignment**. Distribution of cosine similarity scores between the global image `[IMG]` token and the image patch tokens, for original input and prompted input. Training with the disease-aware prompts boosts the intra-modal alignment between the global token `[IMG]` and disease patch tokens, indicated by the rightward skew in the distribution of cosine similarity between `[IMG]` and FG patches.

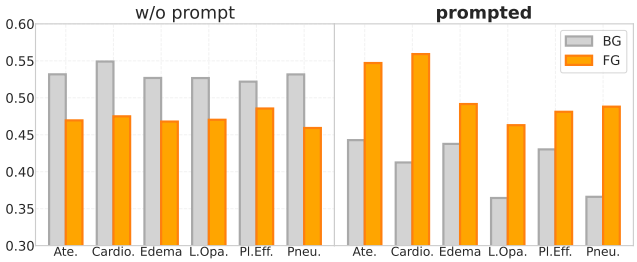


Figure 7. Average cosine similarity between the global image `[IMG]` token and the image patch tokens. Left: Original CheXzero model shows higher similarity to background without prompt than disease regions. Right: After prompting, the relationship inverts with stronger disease region alignment.

patches (see Fig. 7).

Current VLMs use the global image `[IMG]` token as an intermediary between the text `[CLS]` token and the local image patches. Hence, if the global image `[IMG]` token is *non-representative* of fine-grained local disease information, the alignment between the text `[CLS]` and FG tokens is suboptimal. To validate, we show the feature distribution of global image `[IMG]`, text `[CLS]`, BG, and FG tokens in Fig. 8. Due to the global cross-modal contrastive objective, the image and text `[CLS]` tokens are positioned closely to each other, yet both remain distant from FG disease tokens.

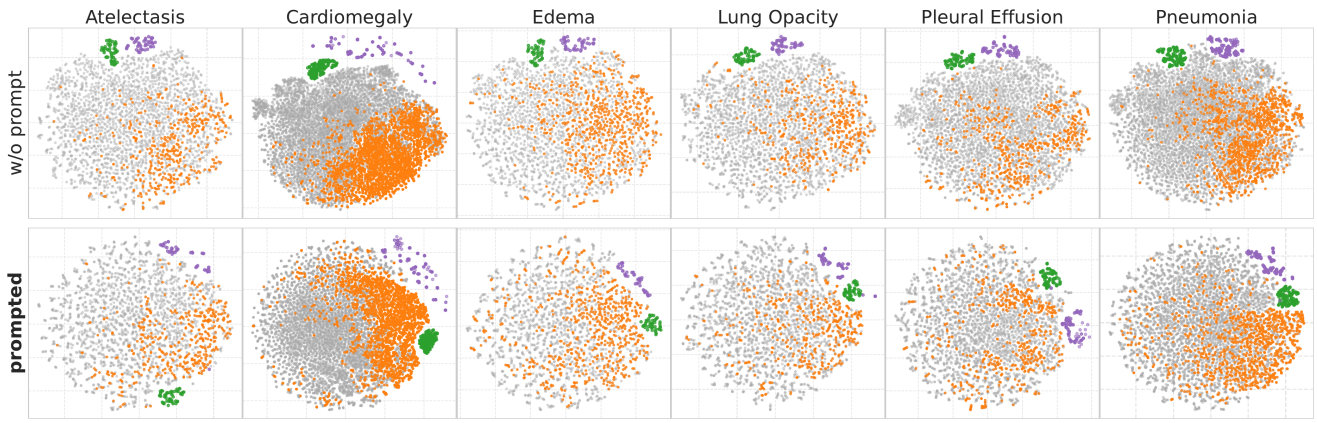


Figure 8. Feature space visualization for six common chest conditions without (Top) and with our disease-focused prompting(Bottom). The global image token (●) is close to the text token (●). However, the disease foreground patch (●) tokens are far from the global image token, thus being far from the text token. In contrast, our feature space shows higher alignment between local and global tokens, and local tokens and the text token.

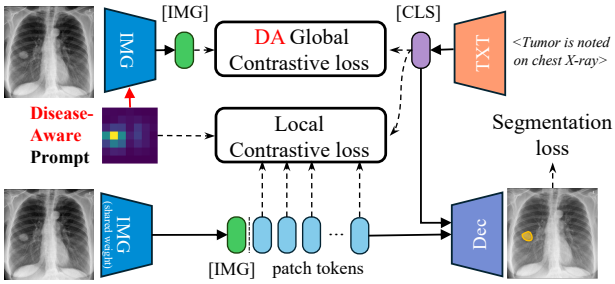


Figure 9. **Model architecture**. We use the disease-aware prompt to align the global image token [IMG] and [CLS] for the Disease-aware global contrastive loss, and to select FG/BG patch tokens for the Local Contrastive Loss. During inference, only the original image patch token and text [CLS] are used.

3.2. Disease-aware Prompting for Weakly-Supervised Visual Grounding

The study identifies two key issues: (1) weak alignment between the local tokens V and global [CLS] token, which fails to represent disease-relevant regions, and (2) high-norm background (BG) tokens that undesirably influence the visual grounding. As such, we propose a simple-yet-effective **disease-aware feature prompting** approach to enhance the influence of disease regions and reduce the impacts of background tokens. From the input image-text pair, we first extract the explainability map of the pre-trained VLM. Our framework uses this explainability map to prompt the image features, enhancing disease token contributions to the global [IMG], suppressing irrelevant BG's influence. Built upon this disease-aware prompting map, we adopt a hierarchical training objective with global and local contrastive losses as illustrated in Fig. 9.

Disease-aware feature prompting. To prioritize disease regions in the model's attention, we apply prompting to reweight the visual token patches so that disease FG regions have stronger impact on the global image [IMG] tokens within the Vision Transformer (ViT) via self-attention. Employing the VLM interpretability approach [8], we generate a 2D *Disease-aware Prompt* Φ as follows:

$$\Phi = \mathcal{R}\Big(V^l,\, \mathcal{M}\big(F_{\mathrm{v}}(x_{\mathrm{v}}),\, F_{\mathrm{t}}(x_{\mathrm{t}})\big)\Big), \tag{2}$$

where $V^l = \{v_1^l, \ldots, v_n^l\}$ denotes the intermediate visual tokens features at the layer l to interpret, $\mathcal{M}$ represents the matching function between the multi-modal outputs of the image and text encoders, and $\mathcal{R}$ applies the interpretability approach [8]. Although imperfect, Φ can effectively highlight disease-relevant FG regions indicated by the text x_{t}. We treat Φ as the disease-specific importance map for x_{t}, using it to weight the visual tokens features, thereby amplify the FG disease signals while suppressing BG noise:

$$\hat{V}^l = \Phi \cdot V^l. \tag{3}$$

Accordingly, the disease-aware prompted visual output at the final layer is:

$$\texttt{[IMG]}_\Phi^L, V_\Phi^L = F_{\mathrm{v}}(x_{\mathrm{v}}, \Phi), \tag{4}$$

The visual grounding output is produced by the pixel decoder, where we employed cross-attention to fuse the text and vision features:

$$\hat{y} = D\big(\mathrm{CrossAttn}(V^L, \texttt{[CLS]}^L)\big). \tag{5}$$

Loss objectives. We introduce hierarchical training objective with global and local losses to maximize granularity. First, we introduce the *disease-aware global contrastive loss*. After concentrating local disease information into the

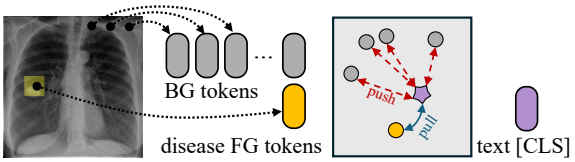


Figure 10. Local contrastive loss. The FG and BG tokens are selected using the Disease-aware prompt Φ.

global image `[IMG]` token via disease-aware prompting, we position disease-related local tokens close to the global image `[IMG]`. As such, using a global contrastive loss $\mathcal{L}_{\text{glb}}$, we bring the text `[CLS]` closer to the image `[IMG]`, which serves as a *proxy* to align disease-related local image tokens with the text `[CLS]`. Given the batch of image-text pairs $\{(x_{\text{i}}, x_{\text{t}})\}_{i=1}^{n}$, the cross-modal global contrastive loss is:

$$\mathcal{L}_{\text{glb}} = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\exp\left[\cos\left(\texttt{[IMG]}_{\Phi_i}^{L}, \texttt{[CLS]}_i\right)\right]}{\sum_{j\in\{x_t^-\}}\exp\left[\cos\left(\texttt{[IMG]}_{\Phi_i}^{L}, \texttt{[CLS]}_j\right)\right]}, \tag{6}$$

where $\{x_t^-\}$ indicates the set of negative text descriptions of x_i in the batch.

Second, we employ the local contrastive loss $\mathcal{L}_{lcl}$ to align textual descriptions with disease regions in the image, as shown in Fig. 10. We threshold the disease-aware prompting mask Φ to select FG tokens $\text{FG}(V) = \{v^L\}_{\text{FG}}$ and draw them closer to the text `[CLS]`, while pushing BG tokens $\text{BG}(V) = \{v^L\}_{\text{BG}}$ away. The local contrastive loss for each image-text pair is defined as:

$$\mathcal{L}_{\text{lcl}} = -\frac{1}{|\text{FG}_i|}\sum_{v_{\text{fg}}\in\text{FG}_i}\log\frac{\exp\left[\cos\left(v_{\text{fg}}, \texttt{[CLS]}_i\right)\right]}{\sum_{v_{\text{bg}}\in\text{BG}_i}\exp\left[\cos\left(v_{\text{bg}}, \texttt{[CLS]}_i\right)\right]}.$$

Finally, we learn the pixel decoder with Φ as pseudo-label, using Dice loss $\mathcal{L}_{\text{seg}}$. The total loss objective is:

$$\mathcal{L} = \mathcal{L}_{\text{glb}} + \mathcal{L}_{\text{lcl}} + \mathcal{L}_{\text{seg}} \tag{7}$$

Note that, the extracted disease-aware prompt Φ is only needed during training to resolve the local misalignment of VLM. Our disease-aware contrastive loss $\mathcal{L}_{\text{glb}}$ pulls the text `[CLS]` close to the more disease-representative global token `[IMG]`, thus improving the local alignment between text and local disease tokens V. Hence, during inference, our model can ground disease patches on the original image without relying on the pre-extracted explainability map. Using this map to narrow down the *pathological search space*, our DAP enables the model to identify subtle and fine-grained disease features, thus further improving the disease visual grounding.

Discussion. While a recent supervised visual grounding method, ZSeg [63], also proposes to mask the input image, our disease-aware prompting differs. ZSeg masks in the pixel space; in contrast, we propose to prompt in the *feature* space to address the followings. First, pixel hard-masking can erase relevant context, while we multiply Φ with token features, preserves both local and global information. Second, pixel masking is insufficient to suppress high background norms as it only operates at the image level. Our method effectively reduces high background norm compared to pixel masking as shown in Fig. 12.

4. Experimental settings

Datasets. MIMIC-CXR v2 chest radiograph data [32] comprises 227,835 imaging studies from 65,379 patients with image-level labels. We include only frontal-view scans, utilizing in a train/validation set of 146.7k/22.2k samples. Evaluation was done using MS-CXR [6], RSNA Pneumonia [49] and COVID Rural [16] datasets. MS-CXR [6] is a subset of MIMIC, provides bounding boxes for 1162 images on 8 diseases. RSNA Pneumonia [49] includes 26k frontal chest X-rays with pneumonia masks, provided by the Radiological Society of North America. COVID Rural [16] contains over 200 annotated chest X-rays for COVID-19 segmentation. We use GPT4-o [43] to generate 50 unique text descriptions per disease label, which are reviewed by board-certified radiologists to create a refined text description database. During training, models are prompted with a random sample from its respective disease descriptions set.

Comparison approaches. We benchmark two groups. The first group is the medical VL pre-training models that are trained on a large-scale image-report pairs, including BioViL [6], BioViL-T [4], and CheXzero [53]. The second group includes state-of-the-art weakly visual grounding models, which are trained via the pseudo-labels generated from the pre-trained VLMs. These include Self-EQ [23], g [48], g++ [47], Transformer-MM [8], and VPT [36]. Since medical images do not have a strong and universal disease object detectors, we only benchmark against detector-free VG methods, and exclude detector-based methods [30, 39].

Metrics. Following recent medical visual grounding methods [4, 6], we adopt the contrast-to-noise ratio (CNR) as the main metric, which measures the difference between FG and BG activations. Besides, we adopt standard VG metrics, including pointing game (PG), Dice and IoU scores.

Implementation Details. Our experiments are conducted on 4 NVIDIA RTX A100 GPUs. Input images are resized to 224×224 and augmented with RandAugment [13]. We set up an Adam optimizer [17] with a learning rate of 0.008 and a batch size of 512 across all experiments. We cap the maximum length of each text to 77 as suggested by [53]. Our hyperparameter values and schedules were determined empirically on a small validation subset.

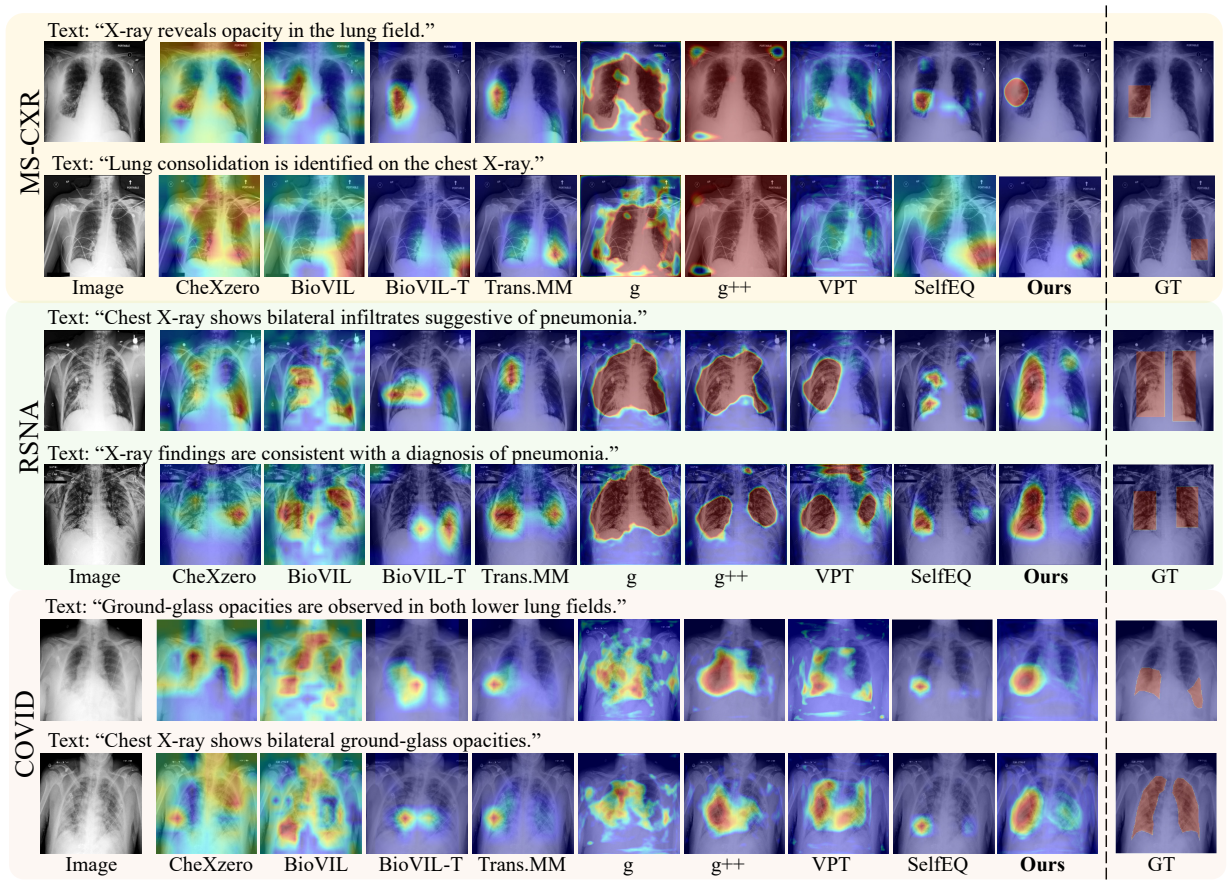


Figure 11. Qualitative results of recent VLMs, and weakly-supervised VG methods on MS-CXR[6] , RSNA [49] and Covid19 [16].

Table 1. Weakly-supervised visual grounding benchmark of different methods on MS-CXR [6] and RSNA [49] on CNR as our main metric, and other standard visual grounding metrics.

Dataset		**MS-CXR**				**RSNA Pneumonia**			
Method	Venue	CNR↑	PG ↑	Dice ↑	IoU ↑	CNR ↑	PG ↑	Dice ↑	IoU ↑
CheXzero [53]	Nat. BE 22	0.675	0.156	0.163	0.089	0.814	0.468	0.202	0.120
BioViL [6]	ECCV22	0.941	0.326	0.238	0.216	1.096	0.472	0.339	0.252
BioViL-T [4]	CVPR23	1.164	0.416	0.320	0.218	0.805	0.584	0.295	0.192
Trans-MM [8]	ICCV21	1.122	0.406	0.336	0.244	1.293	0.526	0.421	0.284
g [48]	NIPS22	0.672	0.304	0.290	0.186	1.461	0.284	0.450	0.315
g++ [47]	CVPR23	0.569	0.258	0.268	0.152	1.350	0.467	0.445	0.311
VPT [36]	ICASSP24	0.588	0.323	0.255	0.162	1.173	0.612	0.468	0.327
Self-EQ [23]	CVPR24	1.082	0.406	0.327	0.229	1.075	0.637	0.370	0.241
DAP	Proposed	**1.254**	**0.457**	**0.352**	**0.245**	**1.630**	**0.747**	**0.474**	**0.328**

5. Results and Discussions

Comparisons with state-of-the-art methods. Table 1 shows the results of recent methods on the two datasets. Our method demonstrates a notable improvement over the vision-language model BioViL, achieving a CNR boost of 7.73% and 48.72% respectively on MS-CXR and RSNA datasets. When compared with Self-EQ [23], a recent advancement in weakly-supervised visual grounding, our approach significantly outperforms it by 15.89% in CNR and 9.36% in the pointing game on MS-CXR. Furthermore, we surpass g++ [47] by 20.74% in CNR on the RSNA dataset.

Table 2 presents the comparisons on the Covid-19 segmentation dataset. We significantly outperforms the recent weakly-visual grounding methods, VPT, and Self-EQ by 17.7% and 33.26% in terms of PG. Fig. 11 shows that exist-

Table 2. Weakly-supervised visual grounding benchmark of different methods on Covid19 [16].

Method	Venue	CNR↑	PG ↑	Dice ↑	IoU ↑
CheXzero [53]	Nat. BE 22	0.923	0.235	0.165	0.097
BioViL [6]	ECCV22	0.623	0.438	0.177	0.201
BioViL-T [4]	CVPR23	0.739	0.469	0.297	0.220
Trans-MM [8]	ICCV21	0.955	0.406	0.325	0.223
g [48]	NIPS22	0.610	0.282	0.246	0.157
g++ [47]	CVPR23	0.702	0.469	0.311	0.205
VPT [36]	ICASSP24	0.739	0.531	0.332	0.220
Self-EQ [23]	CVPR24	0.659	0.469	0.304	0.190
DAP	Proposed	**1.018**	**0.625**	**0.391**	**0.267**

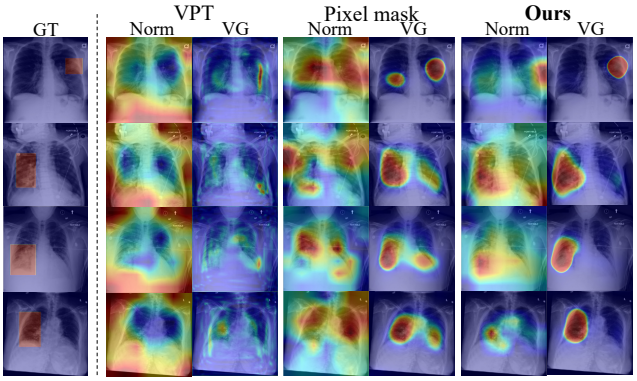


Figure 12. Token norm and VG analysis of prompt tuning methods: VPT [29], Pixel mask [63] versus our DAP.

ing visual grounding methods tends to over-segment the disease regions. These results underscore that existing weakly-supervised visual grounding methods fall short on medical imaging tasks, mainly due to the limited disease-focused local features within the VLM backbone. In contrast, ours show crisp localization on the fine-grained disease regions.

Ablation Study. *Norm and visual grounding of prompt tuning methods.* We evaluate the token norms of visual prompts in VPT [36] and pixel prompts in ZSeg [63].VPT only learns the prompt and does not overcome the problem of high norms in BG regions. Pixel prompting reduces BG noise at the pixel level, but its effect weakens in deeper layers. Our approach, which prompts directly in the feature space, provides consistent background suppression, with improved qualitative results demonstrated in Fig. 12.

Comparisons of different prompting. We compare VG performance of different prompting techniques, as shown in Table 3. Notably, we compares ours with pixel prompting adopted in ZSeg [63], which applies object binary masks on the image pixels. Ours significantly improves upon the pixel masking prompting by 32% in CNR metrics. Ours explicitly suppresses the non-disease regions, thus optimally alleviating the influence of background regions. Secondly, our disease-aware prompting outperforms learnable textual prompts (e.g., CoCoOp [67]), visual prompts [29], and multi-modal prompts [34]. This shows that *informative* disease-aware prompting is more suitable for fine-grained medical visual grounding.

Table 3. Comparisons of different prompt tuning techniques applied on weakly-supervised visual grounding on the MS-CXR dataset [6].

Prompting	CNR↑	PG↑	Dice↑	IoU↑
Red circle [51]	0.866	0.307	0.261	0.177
Pixel mask [63]	0.949	0.361	0.306	0.199
CoOp [67]	0.986	0.364	0.311	0.206
CoCoOp [66]	1.042	0.418	0.333	0.226
Visual prompt [29]	1.129	0.418	0.338	0.231
Multi-modal [34]	1.101	0.414	0.345	0.233
Disease-aware (Ours)	**1.254**	**0.457**	**0.352**	**0.245**

Table 4. Ablation study of the disease-aware prompting (DAP), the global contrastive loss, and the local contrastive loss on the MS-CXR dataset.

DP	$\mathcal{L}_{glb}$	$\mathcal{L}_{lcl}$	CNR↑	PG↑	Dice↑	IoU↑
	✓		0.878	0.376	0.279	0.153
✓	✓		1.061	<u>0.441</u>	<u>0.346</u>	0.237
		✓	<u>1.064</u>	0.434	0.343	<u>0.238</u>
✓	✓	✓	**1.254**	**0.457**	**0.352**	**0.245**

Ablation study of the proposed components. Table 4 shows the performance when ablating different components. First, adding disease-aware prompting into the global contrastive loss significantly improves the Dice score by 24.01%. This shows that using VLM with only global contrastive loss can lead to the low local alignment, hindering the visual grounding. Second, integrating DAP and global contrastive loss yields comparable performance with the local contrastive loss. Without fine-grained local objective, our disease-aware prompting helps pull the local patches closer to `[CLS]` only using the global token `[IMG]` as a proxy. Applying all components yield the highest performance, showing the complementary of disease-aware prompting and global-to-local objectives for fine-grained alignment.

6. Conclusion

This paper provides a diagnostic view of local alignment of medical VLM, hampering the weakly-supervised visual grounding. We show that (i) these models assign high norm values to non-disease regions, misdirecting attention, and (ii) intra-modal misalignment between global and local tokens weakens disease-specific grounding. As such, we propose DAP that strengthens disease-region representations while suppressing background interference. This paper constructs a comprehensive weakly-supervised medical VG on three datasets to catalyze the research, where DAP ours outperforms state-of-the-art methods by up to 20.74%.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 3

[2] Assaf Arbelle, Sivan Doveh, Amit Alfassy, Joseph Shtok, Guy Lev, Eli Schwartz, Hilde Kuehne, Hila Barak Levi, Prasanna Sattigeri, Rameswar Panda, et al. Detector-free weakly supervised grounding by separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1801–1812, 2021. 2, 3

[3] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022. 3

[4] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *CVPR*, pages 15016–15027, 2023. 2, 3, 6, 7, 8

[5] Riddhish Bhalodia, Ali Hatamizadeh, Leo Tam, Ziyue Xu, Xiaosong Wang, Evrim Turkbey, and Daguang Xu. Improving pneumonia localization via cross-attention on medical images and reports. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 571–581. Springer, 2021. 1

[6] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *ECCV*, pages 1–21. Springer, 2022. 1, 2, 3, 4, 6, 7, 8

[7] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 2

[8] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406, 2021. 2, 3, 5, 6, 7, 8

[9] Kan Chen, Jiyang Gao, and Ram Nevatia. Knowledge aided consistency for weakly supervised phrase grounding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4042–4050, 2018. 1

[10] Shengxin Chen, Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, and Rongrong Ji. Querymatch: A query-based contrastive learning framework for weakly supervised visual grounding. In *Proceedings of the ACM International Conference on Multimedia*, pages 4177–4186, 2024. 2

[11] Zhihao Chen, Yang Zhou, Anh Tran, Junting Zhao, Liang Wan, Gideon Su Kai Ooi, Lionel Tim-Ee Cheng, Choon Hua Thng, Xinxing Xu, Yong Liu, et al. Medical phrase grounding with region-phrase context contrastive alignment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 371–381. Springer, 2023. 1

[12] Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024. 1

[13] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 6

[14] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2601–2610, 2019. 2

[15] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779, 2021. 1

[16] Shivang Desai, Ahmad Baghal, Thidathip Wongsurawat, Piroon Jenjaroenpun, Thomas Powell, Shaymaa Al-Shukri, Kim Gates, Phillip Farmer, Michael Rutherford, Geri Blake, et al. Chest imaging representing a covid-19 positive rural us population. *Scientific Data*, 7(1):414, 2020. 6, 7, 8

[17] P Kingma Diederik. Adam: A method for stochastic optimization. *(No Title)*, 2014. 6

[18] Pelin Dogan, Leonid Sigal, and Markus Gross. Neural sequential phrase grounding (seqground). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4175–4184, 2019. 1

[19] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 3

[20] Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. Medrax: Medical reasoning agent for chest x-ray, 2025. 1

[21] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *European Conference on Computer Vision*, pages 752–768. Springer, 2020. 2

[22] Jinlong He, Pengfei Li, Gang Liu, and Shenjun Zhong. Parameter-efficient fine-tuning medical multimodal large language models for medical visual grounding. *arXiv preprint arXiv:2410.23822*, 2024. 1

[23] Ruozhen He, Paola Cascante-Bonilla, Ziyan Yang, Alexander C Berg, and Vicente Ordonez. Improved visual grounding through self-consistent explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13095–13105, 2024. 1, 3, 4, 6, 7, 8

[24] Pingping Huang, Jianhui Huang, Yuqing Guo, Min Qiao, and Yong Zhu. Multi-grained attention with object-level grounding for visual question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3595–3600, 2019. 1

[25] Ta Duc Huy, Sen Kim Tran, Phan Nguyen, Nguyen Hoang Tran, Tran Bao Sam, Anton van den Hengel, Zhibin Liao, Johan W. Verjans, Minh-Son To, and Vu Minh Hieu Phan. Interactive medical image analysis with concept-based similarity reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 30797–30806, 2025. 1

[26] Akimichi Ichinose, Taro Hatsutani, Keigo Nakamura, Yoshiro Kitamura, Satoshi Iizuka, Edgar Simo-Serra, Shoji Kido, and Noriyuki Tomiyama. Visual grounding of whole radiology reports for 3d ct images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 611–621. Springer, 2023. 1

[27] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. In *NeurIPS*, 2021. 3

[28] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 3

[29] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 3, 8

[30] Haojun Jiang, Yuanze Lin, Dongchen Han, Shiji Song, and Gao Huang. Pseudo-q: Generating pseudo language queries for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15513–15523, 2022. 1, 6

[31] Lei Jin, Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Annan Shu, and Rongrong Ji. Refclip: A universal teacher for weakly supervised referring expression comprehension. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2681–2690, 2023. 2

[32] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019. 6

[33] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 1

[34] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 8

[35] Junyan Li, Delin Chen, Yining Hong, Zhenfang Chen, Peihao Chen, Yikang Shen, and Chuang Gan. Covlm: Composing visual entities and relationships in large language models via communicative decoding. In *ICLR*, 2024. 1

[36] Pengyue Lin, Zhihan Yu, Mingcong Lu, Fangxiang Feng, Ruifan Li, and Xiaojie Wang. Visual prompt tuning for weakly supervised phrase grounding. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7895–7899. IEEE, 2024. 2, 3, 6, 7, 8

[37] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Dechao Meng, and Qingming Huang. Adaptive reconstruction network for weakly supervised referring expression grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2611–2620, 2019. 1

[38] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Li Su, and Qingming Huang. Knowledge-guided pairwise reconstruction network for weakly supervised referring expression grounding. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 539–547, 2019.

[39] Yongfei Liu, Bo Wan, Lin Ma, and Xuming He. Relation-aware instance refinement for weakly supervised visual grounding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5612–5621, 2021. 2, 6

[40] Yang Liu, Jiahua Zhang, Qingchao Chen, and Yuxin Peng. Confidence-aware pseudo-label learning for weakly supervised visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2828–2838, 2023. 1, 2

[41] Lingxiao Luo, Bingda Tang, Xuanzhong Chen, Rong Han, and Ting Chen. Vividmed: Vision language model with versatile visual grounding for medicine. *arXiv preprint arXiv:2410.12694*, 2024. 1

[42] Dung Nguyen, Minh Khoi Ho, Huy Ta, Thanh Tam Nguyen, Qi Chen, Kumar Rav, Quy Duong Dang, Satwik Ramchandre, Son Lam Phung, Zhibin Liao, et al. Localizing before answering: A hallucination evaluation benchmark for grounded medical multimodal llms. *arXiv e-prints*, pages arXiv–2505, 2025. 3

[43] OpenAI. Gpt-4 technical report, 2023. 6

[44] Vu Minh Hieu Phan, Yutong Xie, Yuankai Qi, Lingqiao Liu, Liyang Liu, Bowen Zhang, Zhibin Liao, Qi Wu, Minh-Son To, and Johan W Verjans. Decomposing disease descriptions for enhanced pathology detection: A multi-aspect vision-language pre-training framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11492–11501, 2024. 2, 3

[45] 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. 1, 3

[46] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *Computer Vision–ECCV 2016:*

14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, pages 817–834. Springer, 2016. 1

[47] Tal Shaharabany and Lior Wolf. Similarity maps for self-training weakly-supervised phrase grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6925–6934, 2023. 2, 3, 6, 7, 8

[48] Tal Shaharabany, Yoad Tewel, and Lior Wolf. What is where by looking: Weakly-supervised open-world phrase-grounding without text inputs. *Advances in Neural Information Processing Systems*, 35:28222–28237, 2022. 2, 3, 6, 7, 8

[49] George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1): e180041, 2019. 6, 7

[50] Mohit Shridhar and David Hsu. Interactive visual grounding of referring expressions for human-robot interaction. *arXiv preprint arXiv:1806.03831*, 2018. 1

[51] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11987–11997, 2023. 8

[52] Mingjie Sun, Jimin Xiao, Eng Gee Lim, Si Liu, and John Y Goulermas. Discriminative triad matching and reconstruction for weakly referring expression grounding. *IEEE transactions on pattern analysis and machine intelligence*, 43(11): 4189–4195, 2021. 1

[53] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12): 1399–1406, 2022. 1, 2, 3, 4, 6, 7, 8

[54] Giorgos Tziafas and Hamidreza Kasaei. Few-shot visual grounding for natural human-robot interaction. In *2021 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, pages 50–55. IEEE, 2021. 1

[55] Aisha Urooj, Hilde Kuehne, Kevin Duarte, Chuang Gan, Niels Lobo, and Mubarak Shah. Found a reason for me? weakly-supervised grounded visual question answering using capsules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8465–8474, 2021. 1

[56] Alan Q Wang, Batuhan K Karaman, Heejong Kim, Jacob Rosenthal, Rachit Saluja, Sean I Young, and Mert R Sabuncu. A framework for interpretability in machine learning for medical imaging. *IEEE Access*, 2024. 1

[57] Josiah Wang and Lucia Specia. Phrase localization without paired training examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4663–4672, 2019. 2

[58] Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. Improving weakly supervised visual grounding by contrastive knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14090–14100, 2021. 2

[59] Shijie Wang, Dahun Kim, Ali Taalimi, Chen Sun, and Weicheng Kuo. Learning visual grounding from generative vision and language model. *arXiv preprint arXiv:2407.14563*, 2024. 1

[60] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. 2022. 3

[61] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. MedKLIP: Medical knowledge enhanced language-image pre-training. In *ICCV*, 2023. 3

[62] Yutong Xie, Lin Gu, Tatsuya Harada, Jianpeng Zhang, Yong Xia, and Qi Wu. Rethinking masked image modelling for medical image representation. *Medical Image Analysis*, 98: 103304, 2024. 3

[63] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, Xiang Bai, et al. A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model. *arXiv preprint arXiv:2112.14757*, 3:2, 2021. 3, 6, 8

[64] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 3

[65] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023. 3

[66] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. 3, 8

[67] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3, 8

[68] Chenming Zhu, Tai Wang, Wenwei Zhang, Kai Chen, and Xihui Liu. Scanreason: Empowering 3d visual grounding with reasoning capabilities. In *European Conference on Computer Vision*, pages 151–168. Springer, 2024. 1

[69] HaiXiang Zhu, Lixian Su, ShuangMing Mao, and Jing Ye. Read before grounding: Scene knowledge visual grounding via multi-step parsing. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1136–1149, 2025. 1