Does Grounding Improve Radiology Report Generation? An Empirical Study on PadChest-GR

Mohamed Aas-Alas[1] iD   MAASALA@PRHLT.UPV.ES
Alberto Albiol[1] iD   ALALBIOL@PRHLT.UPV.ES
Roberto Paredes[1] iD   RPAREDES@PRHLT.UPV.ES
[1] *Campus de Vera, Universitat Politècnica València, Camí de Vera s/n, 46022 Valencia, Spain*

Abstract

Radiology Report Generation (RRG) aims to automatically produce clinically accurate descriptions of medical images, yet current models often struggle with incomplete findings, generic phrasing, and hallucinations due to the absence of explicit grounding signals. To address these limitations, we propose a grounding-based RRG framework that integrates spatially localized visual evidence into the generation process. Our approach combines a vision encoder ViT with a language decoder LLM GPT-2 through a lightweight transformer-based bridging module inspired by Bridge-Enhanced Vision Encoder–Decoder (VED) architectures. Grounding is introduced using bounding boxes of anatomical regions and pathologies, enabling the model to attend to both global and localized features. We further define a region-to-text task, where the model generates findings directly from specific regions of interest. Experiments on the PadChest-GR dataset demonstrate that grounding substantially improves linguistic quality and clinical accuracy, with the full image plus grounding mask configuration achieving the strongest gains across BLEU, ROUGE-L, CIDEr, BERTScore, CheXbert F1, and RadGraph F1. Incremental masking analyses show that even partial grounding yields consistent benefits.

Keywords: Radiology Report Generation, Spatial Grounding, Region-to-Text Generation, Chest X-ray Interpretation.

1. Introduction

Radiology Report Generation (RRG) is a challenging task that aims to automatically produce detailed and clinically accurate textual descriptions of medical images, such as chest X-rays. This task requires not only recognizing visual abnormalities but also expressing them in a coherent and contextually appropriate medical narrative, similar to what a radiologist would write. Unlike traditional image captioning, RRG demands fine-grained understanding, reasoning, and domain-specific language generation.

However, RRG models face several problems because the decoder often implemented as a large language model (LLM) does not always know what to say or how to structure the report based solely on image features. Unlike Visual Question Answering (VQA), where a textual question provides a clear linguistic grounding and narrows the output space, RRG lacks an explicit guiding signal that constrains or directs the generation process. As a result, generated reports may be incomplete, overly generic, or contain hallucinated findings.

To address these challenges, we present a study on an RRG model that incorporates grounding information to guide the generation process. Specifically, we investigate whether

grounding the model with bounding boxes corresponding to detected anatomical regions and pathologies can enhance the factual accuracy, clinical relevance, and descriptive quality of the generated reports.

The key contributions of this work are twofold. First, we introduce a lightweight transformer-based bridging module that connects the visual encoder and language decoder, effectively integrating grounding information into the report generation process without significantly increasing computational cost. This component enables the model to attend to localized visual cues such as anatomical structures and pathological regions while maintaining global contextual awareness. Secondly, we propose a novel grounding method that leverages bounding boxes corresponding to detected anatomical regions and pathologies to explicitly guide the decoder toward relevant visual evidence during text generation. This approach also introduces a new task, Region-to-Text, in which the model generates a grounded finding directly from a specific region of interest.

For reproducibility and to facilitate further research in grounded Radiology Report Generation, a complete implementation of our approach is provided in a public repository, with all source code on GitHub[1]. We aim to promote transparency and foster future advancements by making our grounding-based RRG framework easily accessible to the research community.

2. Related work

Recent progress in multimodal models has centered on model architectures for vision–language fusion and strategies for spatial grounding. Advances in these areas have enabled finer alignment between visual features and textual reasoning. In radiology, models capable of generating bounding boxes or spatial cues such as RadVLM, MAIRA-2, Knowledge to Sight, and NOVA represent a crucial step toward grounded visual understanding. The task we address, region-to-text generation, differs in that it uses the extracted region features (e.g., bounding boxes) to generate grounded findings rather than predicting regions themselves. Nevertheless, these works remain highly relevant as they demonstrate effective methods for integrating spatial grounding with clinical reasoning. RadVLM introduces multitask setups that couple instruction following with localization to steer radiology-specific behaviors (Deperrois et al., 2025). MAIRA-2 emphasizes grounded report generation by conditioning text on image evidence and explicitly linking descriptions to visual findings (Bannur et al., 2024). Similarly, Knowledge to Sight explores how decomposing domain knowledge into visual attributes enhances abnormality grounding (Li et al., 2025), while NOVA extends these ideas to brain MRI, addressing rare anomaly localization and evaluation under distributional shift (Bercea et al., 2025).

Alternative prompting strategies have investigated how to elicit medically meaningful attention without requiring traditional box-level supervision. Visual Prompt Engineering for Vision–Language Models in Radiology demonstrates that structured prompts such as arrows or point-like cues can guide model attention and improve interpretability (Denner et al., 2025). While effective for classification, these strategies differ from region-to-text formulations where models must generate grounded findings from spatially localized visual evidence rather than classifying entire images.

---

1. https://github.com/LightVED-prhlt/RegionFormer

Architecturally, a common design pattern is to bridge frozen or semi-frozen vision encoders with powerful language decoders through learnable interfaces. BLIP-2 exemplifies this approach by introducing a lightweight Q-Former to adapt image features for large language models, reducing training cost while preserving generalization (Li et al., 2023). ClipCap similarly maps CLIP image embeddings into a prefix for autoregressive decoders, forming a minimal yet effective conduit between modalities (Mokady et al., 2021). Scaling this idea, the Matryoshka Query Transformer introduces query-based designs that enable flexible, nested subsets of capabilities, facilitating efficient inference while retaining strong alignment (Hu et al., 2024). These bridges operationalize a broad trend: transformer-based backbones with compact, trainable adapters that translate vision features into token-like inputs consumable by language models.

With respect to grounding, region-aware VLMs vary in how they represent and inject spatial cues. Some encode regions as textual coordinates within the prompt, letting the language stream parse boxes directly; notable examples include Kosmos-2, Shikra, and PINK, which format box coordinates or referential tags inline with text to steer attention (Peng et al., 2023; Chen et al., 2023b; Xuan et al., 2024). Others insert region features as dedicated tokens derived from RoI-aligned crops, masked patches, points, or masks; GPT4RoI, Ferret, and MG-LLaVA exemplify this feature-level injection, fusing localized embeddings alongside global image tokens to enable referential comprehension and region-conditioned responses (Zhang et al., 2025; You et al., 2023; Zhao et al., 2024). A complementary line modifies the vision encoder to bias attention spatially, (Chen et al., 2023a) incorporates positional encodings or masking schemes that privilege specific areas without requiring explicit box tokens. GRILL pursues fine-grained alignment during pre-training by jointly learning to match textual spans with corresponding image regions, enforcing region–phrase correspondence that later benefits downstream grounding (Jin et al., 2023). Finally, interactive systems such as RegionVLM encode user-provided points or scribbles as spatial trajectories, enabling iterative, fine-grained regional understanding without relying solely on bounding boxes and thus broadening the interface for clinical annotation and guidance (Lee et al., 2024).

Our focus aligns with region-to-text modeling in radiology, where the objective is to generate clinically grounded narratives directly from localized image features. Instead of predicting regions or classifying images, the model interprets predefined regions (e.g., bounding boxes) and produces text that explains the underlying visual evidence. This formulation bridges grounding and generation, allowing clinical reasoning to emerge from spatially constrained representations and moving toward interpretable, fine-grained medical report generation.

3. Our Approach To Grounding Radiology Report Generation

Our proposed model for the Grounded Radiology Report Generation task utilizes a Bridge-Enhanced Vision Encoder Decoder (VED) architecture as shown in Figure 1. It integrates the VIT (Dosovitskiy et al., 2021) as a vision model and GPT-2 (Radford et al., 2019) as the language model, both of which remain frozen during training while a lightweight mapping network is learned.

3.1. Vision Component

The Vision Transformer (ViT), more specifically the CLIP-based (Radford et al., 2021) ViT-B/32 model, is used as the vision encoder. Although architecturally similar to the standard ViT-B/32, the CLIP variant is trained from scratch on hundreds of millions of image–text pairs, learning a visual embedding space that aligns naturally with linguistic semantics. This makes CLIP-ViT particularly suitable for grounded reporting, where the model must connect radiological findings to localized visual evidence. Unlike ImageNet (Ridnik et al., 2021) pretrained ViTs, CLIP's visual encoder produces features that are better structured for downstream language conditioning and cross-modal reasoning.

During training, the vision encoder remains frozen to preserve its pretrained CLIP representations. All images (16-bit) are first loaded and normalized by converting them to float in the range [0,1], applying percentile normalization (99th percentile), clipping, and scaling to 8-bit. The resulting grayscale image is replicated into three channels to form an RGB image compatible with the vision encoder's input requirements. Standard image transformations are then applied to produce the input tensor for ViT-B/32. The encoder outputs a 512-dimensional embedding, which is L2-normalized before being used as input to the language component.

3.2. Language Component

For the language component, the pretrained GPT-2 model is used as the decoder. The visual embedding from the vision encoder is projected into the GPT-2 embedding space through a learnable mapping network, producing a sequence of prefix tokens that serve as conditioning context. The GPT-2 model receives these prefix embeddings concatenated with the token embeddings of the target finding text. During training the GPT-2 decoder is kept frozen; only the mapping network receives gradient updates. The model minimizes the next-token cross-entropy loss, ignoring padding tokens.
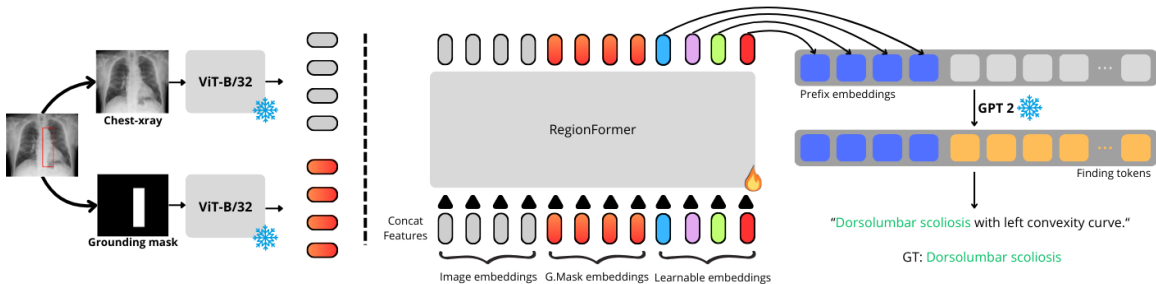
3.3. Region Former

Figure 1: Our proposed architecture for grounded radiology report generation.

The RegionFormer module integrates spatial grounding by using both the full image and its corresponding region representation (grounding mask) as visual inputs. Each sample contains a global image embedding and a grounding mask embedding obtained from

the same visual encoder. Depending on the configuration, the model can use only the image embedding, only the grounding mask embedding, or the concatenation of both. The concatenated vector is then mapped to a sequence of prefix tokens through a lightweight Transformer-based mapper. The resulting prefix embeddings condition GPT-2 to generate grounded radiology reports.

3.3.1. GROUNDING REPRESENTATION

Each Bounding box is represented using normalized coordinates $[x_1, y_1, x_2, y_2]$ within the range $[0, 1]$, defined relative to the image width and height. When rasterizing these bounding boxes, the normalized values are converted to pixel coordinates by scaling them to the image size, rounding to the nearest integer, and ensuring they fall within the image boundaries.

The Region image input $M \in \{0, 255\}^{H \times W}$ is a grounding mask, stencil, not a crop: all pixels inside any bounding box are set to 255 (white), and all others to 0 (black). The single channel is then replicated to RGB to match the vision encoder's expected 3-channel input.

The grounding mask M is passed through the same preprocessing pipeline as the original image I. This preserves spatial alignment under the input transform: the grounding mask undergoes the same transforms, ensuring that its embedding corresponds to the same geometric regions emphasized in the image representation.

3.4. PadChest-GR Dataset

PadChest-GR (de Castro et al., 2025) (Grounded-Reporting), derived from the original PadChest dataset (Bustos et al., 2020), is designed to support the training of Grounded Radiology Report Generation (GRRG) models for chest X-ray (CXR) interpretation. It consists of a publicly available bilingual dataset containing 4,555 CXR studies with grounded reports (3,099 abnormal and 1,456 normal). Each study includes complete sets of sentences describing individual present (positive) and absent (negative) findings in both English and Spanish. For this work, we rely exclusively on the English findings. In total, PadChest-GR provides 7,037 positive finding sentences and 3,422 negative ones. Each positive finding sentence is paired with up to two independent sets of bounding boxes annotated by different radiologists, along with categorical labels indicating the type of finding, anatomical region, and temporal progression. With its detailed localization information and rich annotations covering all clinically important findings, PadChest-GR constitutes a valuable dataset for developing and benchmarking GRRG models for CXR images.

For the training set, PadChest-GR comprises 3,185 studies, but since each study contains multiple findings, the data is divided into 7,342 individual entries, each corresponding to a specific image–finding pair. Similarly, the test set includes 915 studies, which, when divided per finding, result in 2,121 entries. This fine-grained structure enables models to learn more precise visual–textual alignments, improving their ability to associate image regions with specific radiological concepts.

3.5. Framework Overview

Our framework establishes a unified pipeline for Grounded Radiology Report Generation (GRRG), combining visual encoding, spatial grounding, and language generation within

a single end-to-end trainable model. Formally, given an input radiograph I and its set of bounding boxes $\mathcal{B} = \{b_k\}_{k=1}^K$, the model aims to generate a coherent finding $\hat{Y} = (y_1, \ldots, y_T)$ conditioned on both global and localized visual cues.

3.5.1. VISUAL ENCODING

Each image I and its corresponding grounding mask M are first projected into the visual embedding space through the vision encoder $\phi_{\text{VIT}}(\cdot)$.

$$v_{\text{img}} = \phi_{\text{VIT}}(I), \qquad v_{\text{g.mask}} = \phi_{\text{VIT}}(M)$$

Depending on the configuration, either the image embedding v_{img}, the grounding mask embedding $v_{\text{g.mask}}$, or their concatenation is used:

$$v = \begin{cases} v_{\text{img}}, & \text{image-only mode} \\ v_{\text{g.mask}}, & \text{grounding mask-only mode} \\ [v_{\text{img}}; v_{\text{g.mask}}], & \text{image+grounding mask mode} \end{cases}$$

All embeddings are L2-normalized such that $\|v\|_2 = 1$.

3.5.2. VISUAL–LANGUAGE MAPPING

The visual representation $v \in \mathbb{R}^d$ is mapped to the GPT-2 embedding space through a learnable Transformer-based mapper f_θ, which outputs a sequence of L_p prefix tokens $P = [p_1, \ldots, p_{L_p}] \in \mathbb{R}^{L_p \times h}$, where h is the GPT-2 hidden size:

$$P = f_\theta(v) = \text{Transformer}\big([Wv]_{L_c} \,\|\, E_{\text{prefix}}\big)_{L_p}.$$

Here, $W \in \mathbb{R}^{h \times d}$ is a linear projection that expands the single visual vector v into a sequence of L_c latent tokens compatible with GPT-2's embedding dimensionality. This projection is essential, as the CLIP-ViT produces a *single* 512-dimensional image descriptor, whereas GPT-2 operates over a *sequence* of h-dimensional token embeddings. The linear layer therefore serves as a learned interface that *unpacks* the compressed visual representation into multiple visual tokens that can be processed by the Transformer and effectively fused with the learnable prefix embeddings E_{prefix}. The Transformer then refines this joint representation through self-attention, and the final L_p tokens corresponding to the learned prefix positions form the output P, which serves as a conditioning context for the language model.

3.5.3. LANGUAGE GENERATION

Given token embeddings $E(y_1), \ldots, E(y_T)$ from GPT-2, the input sequence to the language model is constructed as

$$Z = [P; E(y_1), \ldots, E(y_T)]$$

GPT-2 then predicts the probability of each next token conditioned on both the prefix and the previous tokens:

$$P(y_t \mid y_{<t}, P) = \text{GPT2}(Z)_t$$

The model is trained by minimizing the standard next-token cross-entropy loss:

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, P)$$

During inference, reports are generated autoregressively using beam search.

3.5.4. GROUNDED GENERATION OBJECTIVE

Through the grounding mask embeddings, the model learns to associate textual tokens with localized image evidence. The overall generation function can be written as

$$\hat{Y} = \arg\max_{Y} P(Y \mid I, M) = \arg\max_{Y} P(Y \mid f_{\theta}([\phi_{\mathrm{VIT}}(I); \phi_{\mathrm{VIT}}(M)]))$$

thus ensuring that the generated finding $\hat{Y}$ is grounded both globally and regionally in the visual content.

Overall, our approach jointly models visual and textual modalities through learned prefix embeddings that bridge ViT visual representations and GPT-2 language generation, enabling spatially grounded and semantically coherent radiology report synthesis.

4. Experimentation and Results

Two different studies were launched, each including multiple experiments, to thoroughly assess the improvement brought by adding grounding to the model. Each experiment was trained for 20 epochs using the AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate of 2e-5 and a linear warm-up of 500 steps. The batch size was set to 40. In this setup, the model generates 40 prefix embeddings encoding visual information before GPT-2 starts generating text.

For generation, we used beam search decoding with a beam size of 5, generation length of 15 tokens, and temperature set to 1.0. The best checkpoint was selected based on the lowest validation loss.

Our models are evaluated with two metric families. Among NLG metrics, BLEU(Papineni et al., 2002), ROUGE-L(Lin, 2004), and CIDEr(Vedantam et al., 2015) capture surface-level overlap and coverage of the reference text, while BERTScore(Zhang et al., 2020) provides a deeper semantic similarity measure. For clinical correctness, F1CheXbert(Smit et al., 2020) reflects agreement with clinical labels, and F1RadGraph(Jain et al., 2021) evaluates entity- and relation-level factual accuracy. Together, these metrics assess both linguistic quality and medical fidelity.

4.1. Study 1: Adding Grounding

The first study compared three configurations: using only the chest-xray images, using the grounding masks alongside the chest-xray images, and using only the grounding masks as input. This setup aimed to evaluate how explicit anatomical grounding affects finding generation and factual consistency.

As observed in Table 1 adding anatomical masks as a grounding signal significantly improved every metric compared to the baseline with only images. The model trained

Table 1: Comparison of model performance using only images, images and grounding masks, and only grounding mask. The best results are shown in **bold**.

Metric	Only Images	Images and Grounding Masks		Only Grounding Masks	
		Value	Δ	Value	Δ
BLEU-1	0.0911	**0.1860**	+0.0949	0.1516	+0.0605
BLEU-2	0.0481	**0.1296**	+0.0815	0.1004	+0.0523
BLEU-3	0.0300	**0.0912**	+0.0612	0.0687	+0.0387
BLEU-4	0.0209	**0.0640**	+0.0431	0.0474	+0.0265
ROUGE-L	0.1079	**0.2411**	+0.1332	0.2097	+0.1018
CIDEr	0.1636	**0.5875**	+0.4238	0.4324	+0.2688
BERTScore	0.2402	**0.3943**	+0.1541	0.3333	+0.0931
F1CheXbert	0.0768	**0.2546**	+0.1779	0.2143	+0.1375
F1RadGraph	0.0131	**0.0931**	+0.0799	0.0656	+0.0525

with images and grounding masks achieved the highest performance across all metrics, with especially strong gains in CIDEr (+0.42), BERTScore (+0.15), and CheXbert (+0.18), confirming that grounding improves both linguistic and factual accuracy. The only grounding masks configuration still outperformed the only images baseline, demonstrating that structural cues alone are informative, but performance slightly decreased without the raw image, suggesting that both modalities are complementary.

4.2. Study 2: A Percent-Scaled Analysis

Since the introduction of grounding masks provided consistent benefits, we performed a second study that incrementally increased the percentage of the train set used. Two configurations were evaluated: using only the grounding masks, and using both the grounding masks and the chest X-ray images. This experiment aimed to identify how much grounding is optimal before reaching performance saturation.

Table 2: Performance comparison across different percentage of the train set using only grounding masks, and images and grounding masks. The best results are shown in **bold**.

Experiment	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr	BERTScore	F1CheXbert	RadGraph
5% Only Grounding Masks	0.0958	0.0496	0.0196	0.0048	0.1102	0.0558	0.1889	0.0517	0.0063
5% Images and Grounding Masks	0.0729	0.0372	0.0080	0.0000	0.1000	0.0205	0.1069	0.0514	0.0044
10% Only Grounding Masks	0.1135	0.0627	0.0272	0.0000	0.1680	0.1039	0.2720	0.0624	0.0054
10% Images and Grounding Masks	0.1408	0.0821	0.0440	0.0000	0.2017	0.2042	0.2670	0.0619	0.0667
20% Only Grounding Masks	0.1610	0.0948	0.0543	0.0327	0.2063	0.3145	0.3433	0.1082	0.0443
20% Images and Grounding Masks	0.1675	0.1026	0.0599	0.0368	0.2182	0.3770	0.3300	0.1446	0.0424
50% Only Grounding Masks	0.1790	0.1151	0.0759	0.0498	0.2349	0.4739	0.3846	0.2083	0.0793
50% Images and Grounding Masks	0.1816	0.1241	0.0863	0.0605	0.2337	0.5265	0.3773	0.2264	0.0713
80% Only Grounding Masks	0.1344	0.0965	0.0677	0.0475	0.1865	0.4573	0.3558	0.2107	0.0774
80% Images and Grounding Masks	0.1792	0.1257	0.0899	**0.0650**	0.2389	0.5639	0.3923	0.2446	0.0862
100% Only Grounding Masks	0.1516	0.1004	0.0687	0.0474	0.2097	0.4324	0.3333	0.2143	0.0656
100% Images and Grounding Masks	**0.1860**	**0.1296**	**0.0912**	0.0640	**0.2411**	**0.5875**	**0.3943**	**0.2546**	**0.0931**

As observed in Table 2 Incrementally increasing the masking ratio produced consistent performance gains across most metrics. Taking F1CheXbert as the primary indicator of clinical factual accuracy, Figure 2 shows a clear trend: both the only grounding mask and the image and grounding mask configurations surpass the 100% only image baseline as early as 20% of the training set. This result has practical implications for dataset creation: to match or exceed the performance of a model trained exclusively on full-image data, it is not necessary to annotate the entire dataset. Instead, annotating approximately 20% of the chest X-ray scans with bounding boxes is sufficient to reach and even surpass the full-image baseline. This substantially reduces the annotation burden for radiologists while still achieving superior factual and semantic accuracy.

The highest overall performance was achieved using the complete images and grounding masks configuration, reinforcing that integrating spatial attention from segmented regions enhances both textual fluency and clinical accuracy.
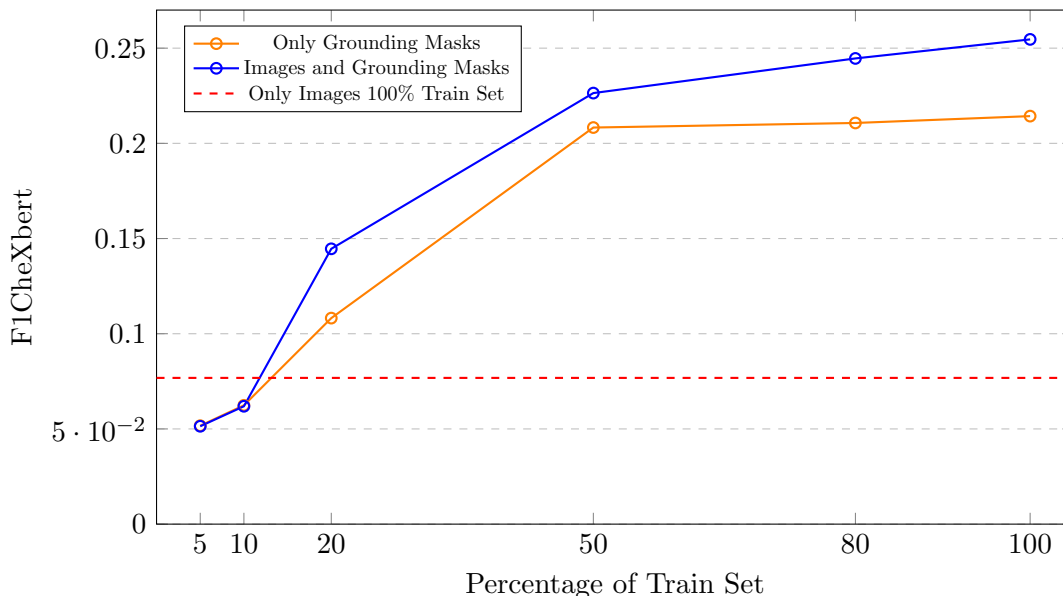

Figure 2: F1CheXbert comparison across training percentages.

4.3. Explainability

To analyze the effect of our grounding masks, we use an explainability approach that highlights where the model directs its attention. When grounding masks are included, the model places greater emphasis on the corresponding pathological regions, as illustrated in the attention maps in Figure 3. Moreover, the grounding masks not only highlight the relevant abnormalities but also guide the model on where to focus. This effect is evident in the final example, where the model generates new attention around the pacemaker region, even though it was not originally emphasized. This indicates that grounding masks not only boost finding generation accuracy but also improve interpretability by guiding the model to focus on clinically meaningful areas.
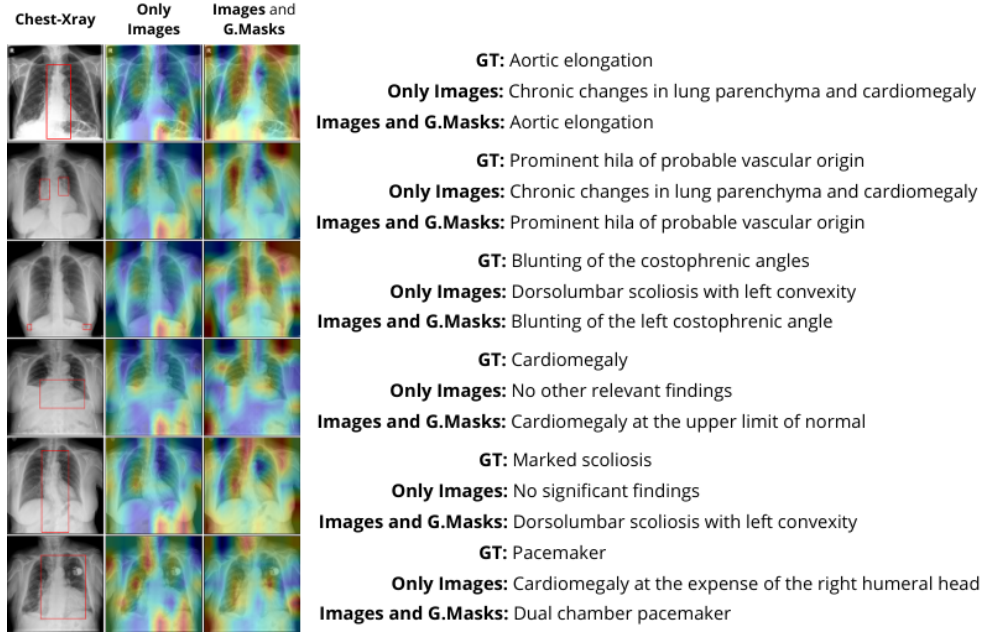
Figure 3: CXR images with attention maps using only image vs image and grounding mask.

5. Conclusion

This work shows that incorporating spatial grounding substantially enhances Radiology Report Generation by guiding models toward clinically relevant visual evidence. Although performance is limited by the small size of the PadChest-GR dataset (4,555 studies), grounding masks still yield clear gains in both linguistic quality and clinical accuracy.

By combining ViT embeddings with a transformer mapping module and leveraging grounding masks derived from annotated bounding boxes, our framework enables the decoder to generate more accurate, specific, and interpretable reports.

Experiments on PadChest-GR show consistent improvements across all metrics when using grounding. The configuration that only uses the grounding masks performs surprisingly well compared to the image-only configuration, which often struggles to align findings with text and produces unfocused descriptions. Grounding masks alone provide strong cues about the expected nature of the finding, even though this output is not derived from image features, as the model does not actually see the underlying image. Despite this, the grounding-based cues alone still perform substantially better than the image-only configuration. When image and mask are combined, the model benefits from both visual context and spatial guidance, achieving the best overall performance.

The proposed region-to-text formulation further demonstrates the model's ability to generate localized, evidence-based findings, advancing RRG toward more clinically meaningful and trustworthy generation.

Overall, our results highlight grounding as an essential component for advancing reliable and interpretable radiology report generation.

Acknowledgments

References

Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Noel C. F. Codella, Fabian Falck, Ozan Oktay, Matthew P. Lungren, Maria Teodora Wetscherek, Javier Alvarez-Valle, and Stephanie L. Hyland. Maira-2: Grounded radiology report generation, 2024. URL https://arxiv.org/abs/2406.04449.

Cosmin I. Bercea, Jun Li, Philipp Raffler, Evamaria O. Riedel, Lena Schmitzer, Angela Kurz, Felix Bitzer, Paula Roßmüller, Julian Canisius, Mirjam L. Beyrle, Che Liu, Wenjia Bai, Bernhard Kainz, Julia A. Schnabel, and Benedikt Wiestler. Nova: A benchmark for anomaly localization and clinical reasoning in brain mri, 2025. URL https://arxiv.org/abs/2505.14064.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797, December 2020. ISSN 1361-8415. doi: 10.1016/j.media.2020.101797. URL http://dx.doi.org/10.1016/j.media.2020.101797.

Chi Chen, Ruoyu Qin, Fuwen Luo, Xiaoyue Mi, Peng Li, Maosong Sun, and Yang Liu. Position-enhanced visual instruction tuning for multimodal large language models, 2023a. URL https://arxiv.org/abs/2308.13437.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic, 2023b. URL https://arxiv.org/abs/2306.15195.

Daniel Coelho de Castro, Aurelia Bustos, Shruthi Bannur, Stephanie L. Hyland, Kenza Bouzid, Maria Teodora Wetscherek, Maria Dolores Sánchez-Valverde, Lara Jaques-Pérez, Lourdes Pérez-Rodríguez, Kenji Takeda, José María Salinas-Serrano, Javier Alvarez-Valle, Joaquín Galant-Herrero, and Antonio Pertusa. Padchest-gr: A bilingual chest x-ray dataset for grounded radiology report generation. *NEJM AI*, 2(7), June 2025. ISSN 2836-9386. doi: 10.1056/aidbp2401120. URL http://dx.doi.org/10.1056/AIdbp2401120.

Stefan Denner, Markus Bujotzek, Dimitrios Bounias, David Zimmerer, Raphael Stock, and Klaus Maier-Hein. Visual prompt engineering for vision language models in radiology, 2025. URL https://arxiv.org/abs/2408.15802.

Nicolas Deperrois, Hidetoshi Matsuo, Samuel Ruipérez-Campillo, Moritz Vandenhirtz, Sonia Laguna, Alain Ryser, Koji Fujimoto, Mizuho Nishio, Thomas M. Sutter, Julia E. Vogt, Jonas Kluckert, Thomas Frauenfelder, Christian Blüthgen, Farhad Nooralahzadeh, and Michael Krauthammer. Radvlm: A multitask conversational vision-language model for radiology, 2025. URL https://arxiv.org/abs/2502.03333.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Wenbo Hu, Zi-Yi Dou, Liunian Harold Li, Amita Kamath, Nanyun Peng, and Kai-Wei Chang. Matryoshka query transformer for large vision-language models, 2024. URL https://arxiv.org/abs/2405.19315.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports, 2021. URL https://arxiv.org/abs/2106.14463.

Woojeong Jin, Subhabrata Mukherjee, Yu Cheng, Yelong Shen, Weizhu Chen, Ahmed Hassan Awadallah, Damien Jose, and Xiang Ren. Grill: Grounded vision-language pre-training via aligning text and image regions, 2023. URL https://arxiv.org/abs/2305.14676.

Jungbeom Lee, Sanghyuk Chun, and Sangdoo Yun. Toward interactive regional understanding in vision-large language models, 2024. URL https://arxiv.org/abs/2403.18260.

Jun Li, Che Liu, Wenjia Bai, Mingxuan Liu, Rossella Arcucci, Cosmin I. Bercea, and Julia A. Schnabel. Knowledge to sight: Reasoning over visual attributes via knowledge decomposition for abnormality grounding, 2025. URL https://arxiv.org/abs/2508.04572.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. URL https://arxiv.org/abs/2301.12597.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Ron Mokady, Amir Hertz, and Amit H. Bermano. Clipcap: Clip prefix for image captioning, 2021. URL https://arxiv.org/abs/2111.09734.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world, 2023. URL https://arxiv.org/abs/2306.14824.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pre-training for the masses, 2021. URL https://arxiv.org/abs/2104.10972.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert, 2020. URL https://arxiv.org/abs/2004.09167.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2015. URL https://arxiv.org/abs/1411.5726.

Shiyu Xuan, Qingpei Guo, Ming Yang, and Shiliang Zhang. Pink: Unveiling the power of referential comprehension for multi-modal llms, 2024. URL https://arxiv.org/abs/2310.00582.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity, 2023. URL https://arxiv.org/abs/2310.07704.

Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest, 2025. URL https://arxiv.org/abs/2307.03601.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. URL https://arxiv.org/abs/1904.09675.

Xiangyu Zhao, Xiangtai Li, Haodong Duan, Haian Huang, Yining Li, Kai Chen, and Hua Yang. Mg-llava: Towards multi-granularity visual instruction tuning, 2024. URL https://arxiv.org/abs/2406.17770.