

Region-Aware Multimodal Large Language Model via SlowFast Tokenization and Pseudo-Mask Guidance for 3D CT Report Generation

Sunggu Kyung¹, Jinyoung Seo², Hyunseok Lim¹, Dongyeong Kim³, Hyungbin Park², Jimin Sung², Wooyoung Jo², Yoojin Nam³, and Namkug Kim^{1*}

¹ Department of Convergence Medicine, University of Ulsan College of Medicine,
Asan Medical Center, Seoul, Republic of Korea

² University of Ulsan College of Medicine, Seoul, Republic of Korea

³ Department of Radiology and Research Institute of Radiology, University of Ulsan
College of Medicine, Asan Medical Center, Seoul, Republic of Korea
`babbu3682@gmail.com`, `namkugkim@gmail.com`

Abstract. Current CT report generation frameworks predominantly rely on global feature representations, often failing to capture region-specific details and potentially missing certain abnormalities. To overcome this limitation, we propose MedRegion-CT, a region-focused multimodal large language model framework featuring three key innovations. First, we revisit the SlowFast strategy to jointly model global and fine-grained information and adapt it to the medical domain via a Region-based SlowFast Tokenizer that extracts tokens guided by clinically meaningful regions. Second, generated pseudo-masks guide the model to attend to diagnostically important anatomical regions, facilitating a systematic understanding of the overall scan context. Third, quantitative lesion information, including size, diameter, and spatial location, is encoded as structured textual prompts, enabling context-aware and clinically informed report generation. To enable rigorous evaluation, we validate our framework on multi-institutional structured report generation benchmarks. Experimental results demonstrate that MedRegion-CT achieves state-of-the-art performance, outperforming existing approaches in both linguistic quality and clinical accuracy. All code is publicly available at: <https://github.com/babbu3682/MedRegion-CT>.

Keywords: Medical Report Generation · Multi-modal Large Language Models · Region-aware Representation Learning

1 Introduction

Computed tomography (CT) is a fundamental imaging modality widely used in clinical diagnosis, enabling detailed assessment of anatomical structures and pathological conditions [21]. However, interpreting large volumes of CT slices and composing radiology reports remains a time-intensive process that imposes

* Corresponding author

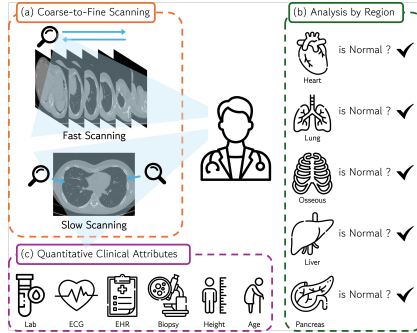


Fig. 1: Comprehensive radiological workflow for clinical diagnosis. (a) Coarse-to-Fine Scanning: fast and slow scanning techniques for preliminary image review. (b) Analysis by Region: systematic assessment of major organ systems. (c) Quantitative Clinical Attributes: consideration of various patient-specific attributes for comprehensive clinical assessment.

a substantial workload on radiologists [9, 13]. Consequently, automated systems that improve diagnostic efficiency while reducing clinical burden are highly desirable.

Recent advances in multimodal large language models (MLLMs) have achieved remarkable progress in complex multimodal reasoning tasks through strong language understanding capabilities and instruction tuning [6]. In particular, recent open-source 3D medical MLLMs [1, 16, 38, 44, 45] have demonstrated promising capabilities in jointly interpreting medical images and clinical information, supporting applications such as report generation and visual question answering.

Despite these advances, existing MLLMs primarily rely on global volume-level representations, which are insufficient for capturing fine-grained regional characteristics critical for accurate report generation. Advancing CT report generation therefore requires models that can effectively extract clinically meaningful regional information and integrate such localized evidence into the report generation process.

Radiologists initially review CT scans in a slice-wise manner to obtain a rapid holistic assessment, followed by focused examination of slices where lesions and organs are clearly visible for precise diagnosis (see Fig. 1-(a)). Inspired by this workflow, we revisit the SlowFast strategy [11] and introduce a Region-based SlowFast Tokenizer to steer the model’s focus toward clinically meaningful regions. Guided by a holistic foreground mask of major organs, global fast tokens use coarse-grained aggregation to capture volumetric context with a low token count per slice, while regional slow tokens employ fine-grained aggregation to extract detailed pathological information with a higher token count. By targeting diagnostically salient regions, this dual-pathway approach yields clinically grounded representations optimized for 3D chest CT.

Subsequently, radiologists perform systematic organ-focused analysis to obtain detailed diagnostic evidence (see Fig. 1-(b)). To emulate this process, we propose a Mask-Driven Visual Extractor. By leveraging pseudo-masks from a

universal segmentation model, it extracts region-aware visual tokens of key organs. These tokens are subsequently integrated into the large language model (LLM) to foster a systematic understanding of anatomical structures.

In addition, diagnostic decisions are often supported by quantitative clinical attributes (see Fig. 1-(c)). Accordingly, lesion pseudo-masks are processed using a deterministic algorithm to extract lesion attributes, including volume, diameter, and spatial location. These attributes are incorporated as structured textual prompts, enabling more precise and clinically informative report generation.

To facilitate rigorous evaluation, we construct multi-institutional structured report generation benchmarks, where each report is standardized into six clinically relevant, organ-specific sections.

In summary, we propose MedRegion-CT, an integrative framework for clinically precise CT report generation. Extensive experiments demonstrate that MedRegion-CT achieves state-of-the-art performance in structured report generation, producing reports with improved diagnostic accuracy and linguistic quality. Our contributions are as follows:

- We introduce a **Region-based SlowFast Tokenizer** that constrains the model’s focus to salient anatomical structures by utilizing a holistic foreground mask of major organs, enabling clinically-grounded learning of comprehensive and fine-grained 3D visual representations for CT report generation.
- We propose a **Mask-Driven Visual Extractor** that leverages pseudo masks of major organs to extract region-aware visual tokens, steering multimodal large language models to attend to major organs.
- We introduce a **Lesion Attribute Extractor** that derives quantitative lesion attributes, including volume, diameter, and spatial location, from pseudo-masks and encodes them as structured textual prompts to enhance clinically informative report generation.
- We construct a **Structured Chest CT Report Generation Benchmark** based on multi-institutional datasets and demonstrate state-of-the-art performance through extensive quantitative and qualitative evaluations across both linguistic and clinical assessment metrics.

2 Related Work

3D CT Visual Encoding for Medical MLLMs. In the 3D medical MLLM domain, the extraction of representative features remains a significant challenge. In the visual encoding stage, compared to the 2D MLLM domain, the 3D MLLM domain faces constraints in transfer learning due to the absence of robust and well-optimized encoders pretrained on large-scale datasets. Moreover, 3D medical images inherently produce numerous vision tokens or features, resulting in substantial computational costs and contextual limitations for LLMs.

To address these challenges, existing studies generally adopt two primary strategies for 3D visual encoding. The first approach combines 3D vision transformer (ViT) encoders with token compression mechanisms. Recent works [1, 16,

44] employ pretrained 3D ViT with input-level downsampling and compression modules such as attention pooling, which utilizes a small set of learnable latent queries, or spatial pooling in latent feature space to reduce token redundancy and computational cost. While this paradigm is intuitive and computationally efficient, aggressive token compression may inevitably discard fine-grained regional information that is critical for precise medical report generation.

The second approach adopts slice-wise 2D feature extraction followed by inter-slice aggregation. For instance, recent models aggregate slice-level representations using Z-formers [24], dual-encoding architectures [39], or attention mechanisms [38]. Benefiting from 2D vision encoders [10, 34] pretrained on large-scale datasets, such approaches effectively preserve slice-level details and inherit robust feature representations. However, such sequential 2D processing may overlook subtle spatial dependencies across the depth axis.

Recent advances in video recognition have demonstrated the effectiveness of the SlowFast dual-frame-rate strategy [11], which simultaneously models global temporal context and fine-grained motion information while maintaining computational efficiency [19, 46]. Inspired by this paradigm, we reinterpret the SlowFast strategy for 3D CT report generation and propose a Region-based SlowFast Tokenizer.

Region-Focused Medical MLLMs. Several studies [15, 29, 49] in general domains have explored the incorporation of local features into LLMs, demonstrating the potential of region-level understanding. Even in the medical domain, radiology reports describe multiple anatomical regions simultaneously; therefore, recent studies have moved beyond conventional approaches relying on global-level representations, incorporating region-specific interpretations.

To identify clinically significant anatomical regions, several studies [20, 42, 51] have constructed refer-and-ground instruction-following datasets that incorporate bounding box coordinates for model fine-tuning. Notably, MAIRA-2 [3] constructs a grounded radiology reporting dataset by meticulously annotating spatial locations corresponding to each CXR finding. This approach enables more precise and finding-specific interpretations.

Several studies [12, 40] have focused on leveraging region-of-interest (ROI) aligned features and integrating them into transformer-based language decoders to enhance radiology report generation. In particular, RGRG [40] employs a Faster R-CNN detector to extract anatomical region-level features from chest X-rays, enabling more fine-grained and clinically consistent descriptions.

Recent approaches integrate pixel-level features into language decoders using pseudo-masks. Specifically, MAIRA-SEG [37] extracts segmentation tokens from the pseudo-masks of major organs in chest X-ray images. Meanwhile, Reg2RG [7] utilizes a universal segmentation model to extract anatomical masks from chest CT images, integrating the resulting local features with global context for a cohesive contextual understanding.

However, comprehensive frameworks for 3D medical MLLMs that effectively leverage anatomical mask integration remain limited. To address this gap, we in-

roduce a novel framework that leverages universal segmentation models to generate pseudo-masks of key anatomical structures, effectively integrating region-specific features via these pseudo-masks.

Guided Textual Prompts for Reliable Medical MLLMs. Recent studies [25,41] have raised concerns regarding the reliability of visual representations in MLLMs, as these models may inaccurately encode visual information, failing to answer even elementary visual questions. To address these limitations, several recent studies have developed text prompting approaches. Text prompting is a natural language processing (NLP) technique developed to transform inputs into textual templates, thereby improving language models’ generalization capabilities and task adaptability, which has recently been systematized as guidelines for enhancing specific task performance [27].

Recent studies have improved model reliability by transforming clinical and visual attributes into textual prompts. For instance, diagnostic results [22], disease classifications from CT images [8], quantifiable morphological data [47], and voxel-level annotations [4] have been successfully encoded as text prompts to guide LLMs. Inspired by these studies, we extract additional clinical information from pseudo-masks and integrate it as textual prompts to the LLM to generate more reliable medical reports.

3 Method

The overall architecture of MedRegion-CT is illustrated in Fig. 2. Initially, pseudo-masks for six predefined major clinical regions—including the lungs, airway, mediastinum, heart, musculoskeletal system, and abdomen—are generated from CT scans using a universal segmentation model (see Fig. 2-(a)). Subsequently, slice-level visual features are extracted by processing 3D CT images slice-wise through a pre-trained 2D vision encoder (MedSigLIP [36]), which has been optimized through large-scale pre-training on millions of CT slices paired with their corresponding radiology reports. The extracted slice features are then transformed into 3D visual tokens, T_{vision} , via a Region-based Slow-Fast Tokenizer. This tokenizer integrates both global and regional tokens, each extracted based on the positions of an integrated multi-organ foreground mask (see Fig. 2-(b)). This approach provides an efficient compressed representation that preserves both inter-slice and intra-slice information while emphasizing key anatomical regions.

Multi-scale visual features and pseudo-masks are then fed into a Mask-Driven Visual Extractor to generate mask tokens, T_{mask} (see Fig. 2-(c)). This mechanism enables focused analysis of anatomical structures. Additionally, a Lesion Attribute Extractor, incorporating deterministic algorithms, is applied to the pseudo-masks to extract quantitative medical information, thereby generating lesion-specific attribute tokens, T_{attr} (see Fig. 2-(d)). This ensures the generation of reliable reports regarding clinically relevant organs and lesions.

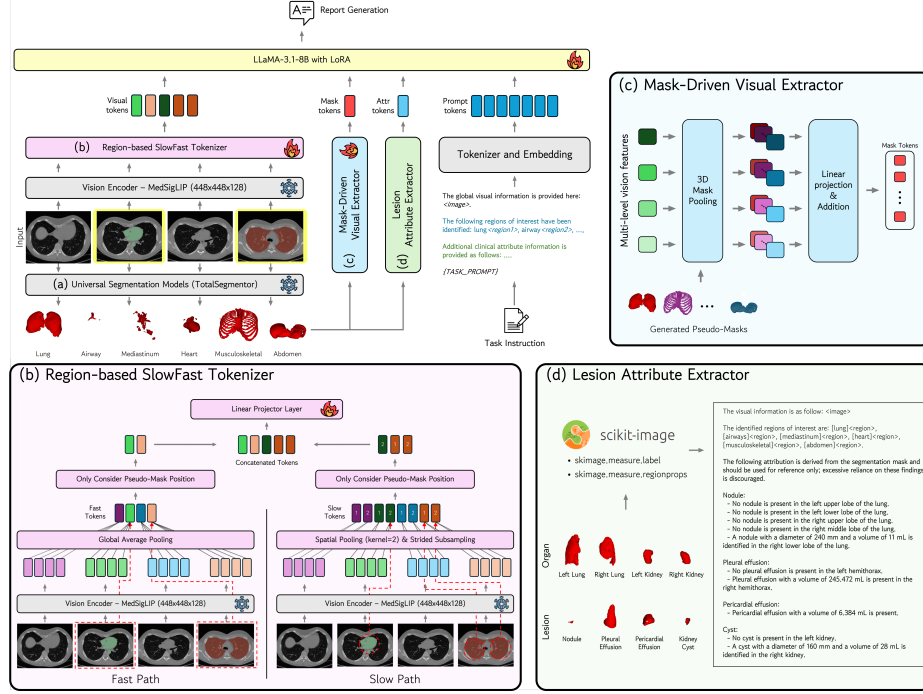


Fig. 2: Overview of MedRegion-CT. (a) Pseudo-masks for six major organs and lesions are generated using a universal segmentation model. (b) Slice visual features are extracted slice-wise and aggregated into 3D visual tokens via a Region-based SlowFast Tokenizer. (c) A Mask-Driven Visual Extractor fuses multi-scale features and pseudo-masks into segmentation tokens. (d) A Lesion Attribute Extractor derives clinical attributes from pseudo-masks, producing attribute tokens for specific lesions such as lung nodules, pleural effusion, pericardial effusion, and kidney cyst. All tokens are input to an LLM, guided by instruction prompts, for report generation.

To ensure standardized and clinically consistent outputs, we adopt a structured report format as the learning objective. This report is partitioned into six key anatomical sections, providing a comprehensive evaluation of the major clinical regions. By training the model on these structured labels, we explicitly guide the LLM to generate organized and domain-specific medical findings. Finally, T_{vision} , T_{mask} , and T_{attr} are concatenated and fed into the LLM along with an instruction prompt I . The model then generates a structured medical report R , which can be formulated as follows:

$$R = \text{LLM}(T_{vision}, T_{mask}, T_{attr}, I) \quad (1)$$

This strategy facilitates the generation of accurate, clinically grounded, and well-organized medical reports focused on the six predefined anatomical regions. The following subsections provide detailed explanations of each component.

3.1 Region-based SlowFast Tokenizer

While processing 3D CT images slice-wise through 2D encoders typically yields $D \times T$ tokens—where D denotes the number of axial slices and T represents the number of spatial tokens per slice—the direct input of such high-dimensional features into an LLM is computationally prohibitive. Existing SlowFast-based approaches [19, 46] attempt to mitigate this by combining spatially averaged tokens from all slices with tokens that are densely sampled from a subset of slices selected at uniform intervals. However, such uniform sampling strategies and the indiscriminate consideration of all slices are suboptimal for the 3D medical imaging domain.

To address this, we propose a Region-based SlowFast Tokenizer that leverages pseudo-masks to construct 3D visual tokens (T_{vision}) by isolating features that strictly correspond to predefined clinical regions. Let $f(i, j) \in \mathbb{R}^C$ denote the feature vector extracted from a 3D CT scan at the i -th depth slice ($i \in \{1, \dots, D\}$) and the j -th spatial token ($j \in \{1, \dots, T\}$), where C is the channel dimension. We pair each feature with a binary indicator $m(i, j) \in \{0, 1\}$ derived from the foreground mask of major organs, where $m(i, j) = 1$ if $f(i, j)$ belongs to a predefined anatomical region, and 0 otherwise. The tokenizer generates T_{vision} through two specialized pathways: the Fast Pathway and the Slow Pathway.

Fast Pathway (Global-Anatomical Context). The Fast tokens, T_{fast} , capture the essential global context of the relevant anatomy. Instead of applying spatial average pooling to every depth slice indiscriminately, we first use the foreground mask to retain only organ-bearing slices, i.e., slices where the foreground mask of major organs is active. For each selected slice, we aggregate its slice-level representation by average pooling over the T spatial tokens:

$$T_{fast} = \left\{ \text{AvgPool}_j(\{f(i, j)\}, \text{kernel} = T) \mid \sum_j m(i, j) > 0 \right\}. \quad (2)$$

This filters out non-informative background-only slices while preserving the global characteristics of each organ-bearing slice.

Slow Pathway (Local-Detailed Context). The Slow tokens, T_{slow} , preserve high-fidelity intra-slice details through region-guided spatial pooling. We harvest tokens strictly within the designated coordinates using masks with strided subsampling, and apply 2D spatial pooling with a kernel of 2:

$$T_{slow} = \left\{ \text{AvgPool}_j(\{f(i, j)\}, \text{kernel} = 2) \mid m(i, 2j) = 1 \right\}. \quad (3)$$

This focuses the model’s capacity on granular pathological nuances while discarding the redundant overhead of non-clinical areas.

Finally, the tokens from both pathways are concatenated to form the final 3D visual representation T_{vision} . This dual-pathway strategy prioritizes clinically relevant volumetric features. By leveraging key organ segmentation masks, we extract dynamic, patient-specific tokens, significantly reducing the total sequence length to a manageable size while ensuring the LLM receives concentrated diagnostic information.

3.2 Mask-Driven Visual Extractor

To enable focused analysis of specific anatomical structures, we incorporate a Mask-Driven Visual Extractor based on the Osprey framework [48]. While the Osprey employs explicit location tokens, extending them to 3D CT volumes incurs prohibitive computational costs. Therefore, we omit these tokens to maintain efficiency and extend its mask-driven approach with a Volumetric Mask Pooling mechanism to effectively capture dense anatomical information across the entire volume.

Building upon the visual encoder, we extract multi-level slice visual features from the 6th, 12th, 18th, and 24th layers of MedSigLIP and apply our Mask Pooling module independently to each layer to distill these dense features into region-specific representations using key anatomical region masks. Specifically, for a given key anatomical region o , we adapt its corresponding binary pseudo-mask $m(i, j, o) \in \{0, 1\}$ to the feature resolution through trilinear interpolation. The layer-specific mask token, $k_l(o)$, is computed by pooling the features f_l from layer l using the provided mask:

$$k_l(o) = \sum_{(i,j) \in \Omega} f_l(i, j) \cdot \frac{m(i, j, o)}{\sum_{(i',j') \in \Omega} m(i', j', o) + \epsilon} \quad (4)$$

where Ω defines the volumetric index set of size $D \times T$, $m(i, j, o)$ indicates the corresponding mask value for the target organ o , and ϵ is a small constant (10^{-8}).

Following the mask pooling for each layer, the resulting tokens $\{k_l(o)\}_{l=1}^4$ are passed through independent linear projectors, ϕ_l . The projected tokens are then aggregated via element-wise summation. Finally, this aggregated representation is passed through a final linear projector, ψ , to map the fused features into the embedding space of the LLM:

$$T_{mask} = \psi \left(\sum_{l=1}^4 \phi_l(k_l(o)) \right) \quad (5)$$

This hierarchical fusion strategy ensures that the final representation T_{mask} incorporates both fine-grained anatomical details and high-level semantic context. By representing each region with a single condensed mask token, we minimize prompt length and ensure scalability for multi-organ analysis within the fixed context length of the LLM.

The extracted mask tokens are injected into the LLM at fixed prompt positions designated by special `<region>` tokens. This structured integration allows the LLM to focus specifically on designated anatomical structures, facilitating consistent and region-enhanced clinical report generation.

3.3 Lesion Attribute Extractor

Transmitting sufficient visual features to LLMs remains challenging, as 3D CT volumes are often heavily downsampled or cropped due to computational constraints, potentially leading to the loss of fine-grained morphological details.

Relying solely on a vision encoder thus poses significant limitations for accurate diagnosis. To address this limitation, we integrate explicit descriptions of localized lesion characteristics into textual prompts, enabling the model to utilize these quantitative attributes that are crucial for determining disease states. Such characteristics—including volume, maximum diameter, and anatomical location—serve as vital diagnostic and prognostic biomarkers, particularly for assessing the malignancy risk of pulmonary nodules [30].

Rather than relying on implicit features, we explicitly compute the geometric properties of lesions in the voxel space. Let $\mathcal{M}_{loc} \in \{0, 1\}^{D \times H \times W}$ and $\mathcal{M}_{les} \in \{0, 1\}^{D \times H \times W}$ denote the binary masks for an anatomical guidance region (e.g., right upper lobe) and a target lesion class (e.g., lung nodule), respectively. To isolate lesions strictly within a specific anatomical region, we compute their intersection and apply the largest connected component operator Φ to filter out segmentation noise:

$$\mathcal{M}_{target} = \Phi(\mathcal{M}_{loc} \odot \mathcal{M}_{les}) \quad (6)$$

For the isolated instance $\mathcal{M}_{target}$, the absolute volume V (in mL) and the maximum bounding-box diameter d_{max} (in mm) are derived as:

$$V = \left(\sum_{x,y,z} \mathcal{M}_{target}(x, y, z) \right) \times (s_x s_y s_z) / 10^3, \quad d_{max} = \max_{i \in \{x,y,z\}} (L_i \times s_i) \quad (7)$$

where s_x, s_y, s_z represent the normalized voxel spacing along each axis, and L_i denotes the voxel length of the bounding box of $\mathcal{M}_{target}$ along axis $i \in \{x, y, z\}$.

The specific types of detectable lesions are inherently determined by the universal segmentation model employed. Utilizing TotalSegmentator [43] for initial segmentation, alongside SimpleITK and scikit-image for morphological post-processing, we extract quantitative profiles for critical lesions including pulmonary nodules, pleural/pericardial effusions, and renal cysts. This metadata is restructured into a textual format (see Supplementary Algorithm 1) and encoded as attribute tokens T_{attr} for the LLM, as illustrated in Fig. 2-(d). This approach contributes to generating consistent and reliable region-focused reports while mitigating the variability inherent in human observations.

3.4 Structured Report Generation Benchmark

Publicly available radiology reports often lack systematic organization, hindering both supervised learning and consistent clinical evaluation. To address this, we establish a structured chest CT report generation benchmark by reorganizing unstructured data into a standardized format. We divide the anatomical field of view into six primary clinical regions (e.g., lung parenchyma, large airway, and mediastinum, among others). This categorization reflects the systematic approach employed by radiologists, ensuring thorough evaluation of relevant pathology within each region and enhancing overall report quality and consistency [5, 31]. Detailed specifications for these categories and their constituent sub-regions are provided in Supplementary Tab. S2.

To construct high-quality training labels, we utilized the RadGenome-Chest CT dataset [50], which provides multi-granularity grounded reports. While this dataset decomposes findings into numerous fine-grained structures, it lacks a unified representation for the six major clinical regions essential for comprehensive chest evaluation. Therefore, we developed a multi-stage NLP pipeline using the DSPy [23] framework and the Llama-3.3-70B-Instruct [14] model to selectively aggregate and refine these fragmented sentences. This pipeline systematically performs: (i) clinical cleansing to expand medical abbreviations; (ii) temporal refinement to focus on current findings by filtering out longitudinal comparison noise; and (iii) anatomical merging to integrate sub-region findings into the predefined clinical hierarchy.

To enable consistent evaluation, we applied this NLP pipeline to both the internal RadGenome-Chest CT dataset and the external Asan Medical Center (AMC) dataset. Reorganizing each ground-truth report into this unified format effectively constructs a standardized report generation benchmark for both in-domain and cross-institutional settings. These benchmarks facilitate systematic assessments of model performance across linguistic and clinical metrics, establishing a reliable foundation for automated structured reporting in 3D medical imaging.

4 Experiments

4.1 Datasets and Metrics

Datasets. We utilize two structured report generation benchmarks to comprehensively evaluate our model: the internal RadGenome-Chest CT benchmark for assessing in-domain performance, and the external AMC benchmark for evaluating zero-shot generalization under cross-institutional domain shift. Specifically, the RadGenome-Chest CT follows the official split, comprising 23,865 training/validation and 1,551 testing scans with corresponding multi-granularity reports. For the external AMC dataset, we incorporate non-contrast chest CT scans and paired reports collected from 4,014 patients between 2010 and 2021, from which a subset of 500 cases was randomly sampled. The retrospective use of the AMC dataset was approved by the institutional review board, and the requirement for informed consent was waived. All reports from both datasets were further processed into the structured format described in Sec. 3.4.

Metrics. Evaluating generated radiology reports requires assessing both linguistic similarity and clinical accuracy. Our methodology incorporates two metric categories: 1) Natural Language Generation (NLG) metrics include BLEU [33], ROUGE [26], and METEOR [2]. 2) Language Model (LM) based metrics consist of CRG [17], GREEN [32], and Clinical Accuracy (CA) [16] using the gpt-5-nano API. These metrics ensure comprehensive assessment across structural, linguistic, and clinical aspects essential for radiology reporting.

Table 1: Comparisons with state-of-the-art methods on the RadGenome-Chest CT and AMC structured report generation benchmarks. Brackets denote the 95% confidence interval (CI). **Bold** and underlined values indicate the best and second-best results, respectively.

Method	NLG Metrics			LM-based Metrics		
	BLEU	ROUGE	METEOR	GREEN	CRG	GPT-CA
<i>Structured RadGenome-Chest CT (Internal Dataset, n=1551)</i>						
RadFM [44]	0.3153 [0.3072, 0.3237]	0.3732 [0.3667, 0.3798]	0.4894 [0.4825, 0.4966]	0.3030 [0.2914, 0.3142]	0.3477 [0.3447, 0.3508]	0.2852 [0.2771, 0.2936]
M3D [1]	0.3177 [0.3083, 0.3277]	0.3973 [0.3889, 0.4058]	0.4938 [0.4860, 0.5018]	<u>0.4013</u> [0.3883, 0.4146]	0.3717 [0.3683, 0.3751]	0.3586 [0.3479, 0.3694]
MedM-VL [38]	0.3026 [0.2920, 0.3116]	0.3854 [0.3775, 0.3934]	0.4811 [0.4734, 0.4892]	0.3970 [0.3856, 0.4084]	0.3426 [0.3408, 0.3445]	0.3168 [0.3066, 0.3273]
Med3DVLM [45]	<u>0.3269</u> [0.3174, 0.3368]	<u>0.3991</u> [0.3905, 0.4078]	<u>0.5029</u> [0.4946, 0.5111]	0.3881 [0.3738, 0.4024]	<u>0.3866</u> [0.3828, 0.3907]	<u>0.3733</u> [0.3622, 0.3846]
CT-CHAT [16]	0.3081 [0.2980, 0.3175]	0.3900 [0.3815, 0.3987]	0.4831 [0.4749, 0.4917]	0.3966 [0.3843, 0.4092]	0.3425 [0.3408, 0.3443]	0.3265 [0.3155, 0.3374]
Ours	0.3435 [0.3325, 0.3534]	0.4225 [0.4132, 0.4318]	0.5205 [0.5118, 0.5291]	0.4555 [0.4411, 0.4697]	0.3959 [0.3919, 0.4003]	0.4003 [0.3884, 0.4124]
<i>Structured AMC (External Dataset, n=500, Zero-shot Inference)</i>						
RadFM [44]	0.1464 [0.1424, 0.1502]	0.1928 [0.1889, 0.1969]	0.4790 [0.4730, 0.4851]	0.0977 [0.0802, 0.1164]	0.3476 [0.3427, 0.3527]	0.2018 [0.1880, 0.2161]
M3D [1]	0.1185 [0.1137, 0.1218]	0.1565 [0.1537, 0.1593]	0.4457 [0.4399, 0.4515]	0.1102 [0.0901, 0.1317]	0.3725 [0.3666, 0.3787]	0.2243 [0.2104, 0.2389]
MedM-VL [38]	0.1235 [0.1216, 0.1254]	0.1543 [0.1517, 0.1569]	0.4430 [0.4371, 0.4489]	0.1115 [0.0909, 0.1336]	0.3472 [0.3436, 0.3508]	0.1963 [0.1792, 0.2143]
Med3DVLM [45]	0.1136 [0.1106, 0.1166]	0.1557 [0.1525, 0.1590]	0.4400 [0.4338, 0.4462]	<u>0.1138</u> [0.0933, 0.1361]	<u>0.3791</u> [0.3725, 0.3863]	<u>0.2366</u> [0.2215, 0.2527]
CT-CHAT [16]	0.1173 [0.1064, 0.1261]	0.1559 [0.1529, 0.1589]	0.4410 [0.4345, 0.4474]	<u>0.1138</u> [0.0926, 0.1363]	0.3487 [0.3449, 0.3525]	0.2053 [0.1870, 0.2247]
Ours	<u>0.1275</u> [0.1150, 0.1354]	<u>0.1572</u> [0.1544, 0.1601]	<u>0.4467</u> [0.4412, 0.4521]	0.1572 [0.1305, 0.1804]	0.3822 [0.3749, 0.3903]	0.2642 [0.2470, 0.2825]

Implementations. We implemented our framework using PyTorch and trained it on NVIDIA A100 GPUs. The input 3D CT volumes were standardized to a fixed resolution of $448 \times 448 \times 128$. We utilized the pre-trained MedSigLIP as the vision encoder and LLaMA-3.1-8B [14] as the language model, applying LoRA [18] for parameter-efficient fine-tuning. The model was trained in two stages: initially aligning the multimodal connector and mask projector, followed by fine-tuning on the structured report dataset. Detailed hyperparameters and preprocessing configurations are provided in Supplementary Sec. 1.

4.2 Comparison with State-of-the-Art 3D MLLMs

To evaluate our proposed framework, we compared it against five state-of-the-art 3D Medical MLLMs (RadFM [44], M3D [1], CT-CHAT [16], Med3DVLM [45], and MedM-VL [38]) on the internal RadGenome-Chest CT and external AMC benchmarks. All baselines were initialized with their official weights and

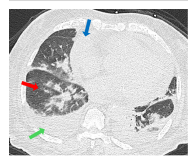
 CT scan → pericardial effusion → patchy consolidation → pleural effusion ■ correct sentence ■ wrong sentence	[Lung]: More prominent patchy consolidations are observed in the upper lobes of both lung parenchyma. Interlobular septa are thick. Compressive atelectasis and pleuroparenchymal density increases are observed in the lower lobes of both lungs. In addition, there is a slightly thick-walled air cyst of 4.3 cm in the laterobasal segment of the lower lobe of the right lung. There is mosaic attenuation consistent with small airway or small vessel disease in both lung parenchyma. There are bilateral pleural effusions measuring 2.5 cm in the thickest part on the right and 1.8 cm in the thickest part on the left, and passive atelectasis in the lung parenchyma adjacent to the effusion. [Heart]: Pericardial effusion measuring 2.7 cm in its thickest part is observed. The cardiothoracic index increased in favor of the heart.	[Lung]: When examined in the lung parenchyma window, There are subsegmental atelectasis in the lower lobes of both lungs. Minimal pleural effusion is observed on the left. Trachea, both main bronchi are open. [Heart]: Pericardial effusion-thickening was not observed. Mediastinal main vascular structures, heart contour, size are normal.	[Lung]: No mass or infiltrative lesion was detected in both lungs. There are millimetric nonspecific nodules in both lungs. The lower lobe of the left lung is almost completely atelectatic. There are atelectasis in the lower lobes of both lungs adjacent to the pleural effusion. There is minimal pleural effusion on the left. There is no pleural effusion on the right. [Heart]: As far as can be observed: The heart is larger than normal.	[Lung]: No mass or infiltrative lesion was detected in both lungs. There are minimal emphysematous changes in both lungs. There is a millimetric nonspecific nodule in the right lung. There are linear atelectasis in both lungs. No pleural or pericardial effusion was detected. [Heart]: As far as can be observed: Heart contour and size are normal.
	GT	RadFM	M3D	CT-CHAT
	Med3DVLM	MedM-VL	Ours	

Fig. 3: Qualitative comparison of generated radiology reports for a patient exhibiting significant pericardial and pleural effusions with associated pulmonary consolidations. We compare our proposed MedRegion-CT against state-of-the-art methods. Text is color-coded to indicate findings consistent with the ground truth (green) and findings that are inconsistent with the ground truth or represent hallucinations (yellow). For clarity, only the lung and heart sections of the report are visualized. MedRegion-CT demonstrates superior performance in both diagnostic accuracy and localization performance.

fine-tuned on our structured internal dataset to ensure optimal localized report generation. These models were then evaluated on the external AMC dataset without additional training to assess zero-shot generalization.

Quantitative Results. As summarized in Tab. 1, MedRegion-CT consistently outperforms the baselines across all evaluation categories on the internal benchmark. Our model achieves the highest NLG metrics (BLEU: 0.3435, ROUGE: 0.4225, METEOR: 0.5205), indicating strong lexical and semantic alignment with ground-truth reports. Furthermore, it achieves the best LM-based clinical scores (GREEN: 0.4555, CRG: 0.3959, GPT-CA: 0.4003). While M3D and Med3DVLM show competitive results in specific metrics, they fall short of our model’s overall performance. The substantial margin in clinical accuracy, particularly in GREEN, demonstrates that our approach more effectively captures essential diagnostic findings within the six predefined clinical regions (see Supplementary Tab. S1 for detailed region-wise results). A further size-stratified nodule analysis in Supplementary Tab. S3 confirms superior detection rates for mid-to-large nodules compared to existing baselines. Although performance decreases for small abnormalities ($\leq 5\text{mm}^3$), this is primarily bounded by the false negatives of the upstream off-the-shelf segmentation model.

Under zero-shot inference on the external AMC benchmark, MedRegion-CT demonstrates robust generalization against cross-institutional domain shifts. Although RadFM achieves the highest NLG scores, its lower LM-based clinical scores suggest that lexical-overlap metrics may favor institution-specific report-

ing styles over clinical faithfulness. In contrast, MedRegion-CT achieves the best LM-based performance (GREEN: 0.1572, CRG: 0.3822, GPT-CA: 0.2642) while maintaining the second-best NLG performance. This divergence highlights the limitation of surface-level lexical metrics in cross-institutional evaluations, where report phrasing and formatting vary substantially across hospitals.

Consistent with the results of the radiologist reader study (Supplementary Tab. S4), which highlight our model’s strong ability to capture key findings and minimize hallucinations, MedRegion-CT preserves clinically relevant information even when diverging from the exact wording of reference reports. Conversely, models relying primarily on global features (e.g., RadFM, CT-CHAT) exhibit lower clinical reliability in structured reporting, likely due to the difficulty of localized feature extraction in complex 3D volumes. Overall, our framework generates clinically grounded reports with strong zero-shot generalization.

Qualitative Results. Fig. 3 illustrates a qualitative comparison of generated reports on the internal structured benchmark. The ground-truth report for this case describes bilateral consolidations accompanied by significant pericardial and pleural effusions. MedRegion-CT demonstrates superior diagnostic fidelity by accurately identifying these multi-regional pathologies. Notably, it correctly captures the upper-lobe consolidations and reflects the presence of effusions with high precision, closely aligning with the ground-truth (GT).

In contrast, baseline models exhibit several clinical inaccuracies. RadFM and Med3DVLM miss the critical pericardial effusion, incorrectly reporting normal heart contours. M3D overlooks the primary parenchymal consolidations in favor of non-specific nodules, while MedM-VL provides an essentially normal assessment. These qualitative results further support the quantitative findings, confirming that MedRegion-CT effectively minimizes hallucinations and provides reliable clinical utility through precise regional localization.

4.3 Ablation Study

To evaluate the contribution of each module, we conducted an ablation study using a 30% training subset of the RadGenome-Chest CT dataset and validated zero-shot generalization on the AMC dataset. We compared our full configuration ($R+Attr+Mask$) against internal variants and external encoder baselines. Here, R , $Attr$, and $Mask$ denote the Region-based SlowFast Tokenizer, Lesion Attribute Extractor, and Mask-Driven Visual Extractor, respectively. $LITA_{enc}$. [19] and $M3D_{enc}$. [1] represent the conventional SlowFast strategy and the pre-trained 3D ViT encoder, respectively.

On the internal benchmark (Tab. 2, top), our full model outperforms all variants across both NLG (BLEU: 0.3273, ROUGE: 0.3988, METEOR: 0.5039) and clinical metrics (GREEN: 0.4025, CRG: 0.3861, GPT-CA: 0.3839), demonstrating its ability to generate fluent and clinically grounded reports. The incremental addition of $Attr$ and $Mask$ to the R baseline steadily improves performance. Notably, $Attr$ substantially enhances clinical consistency by integrating determinis-

Table 2: Ablation study of MedRegion-CT components on the RadGenome-Chest CT and AMC structured report generation benchmarks. Here, *R* denotes our Region-based SlowFast Tokenizer, *Attr* denotes our Lesion Attribute Extractor, and *Mask* denotes our Mask-Driven Visual Extractor. *LITA_{enc.}* and *M3D_{enc.}* indicate the conventional SlowFast tokenization strategy [19] and the pre-trained 3D ViT encoder from M3D [1], respectively.

Method	NLG Metrics			LM-based Metrics		
	BLEU	ROUGE	METEOR	GREEN	CRG	GPT-CA
<i>Structured RadGenome-Chest CT (Internal Dataset, n=1551)</i>						
<i>R+Attr+Mask</i>	0.3273 [0.3161, 0.3364]	0.3988 [0.3904, 0.4075]	0.5039 [0.4960, 0.5121]	0.4025 [0.3889, 0.4165]	0.3861 [0.3821, 0.3901]	0.3839 [0.3728, 0.3952]
<i>R+Attr</i>	0.3271 [0.3131, 0.3361]	0.3984 [0.3898, 0.4070]	0.5034 [0.4955, 0.5115]	0.4010 [0.3876, 0.4148]	0.3846 [0.3804, 0.3887]	0.3807 [0.3728, 0.3952]
<i>R+Mask</i>	0.3248 [0.3148, 0.3343]	0.3973 [0.3887, 0.4060]	0.4996 [0.4916, 0.5076]	0.3937 [0.3797, 0.4072]	0.3699 [0.3664, 0.3736]	0.3587 [0.3473, 0.3701]
<i>R</i>	0.3152 [0.3059, 0.3240]	0.3909 [0.3827, 0.3992]	0.4912 [0.4836, 0.4993]	0.3930 [0.3794, 0.4068]	0.3613 [0.3583, 0.3645]	0.3513 [0.3400, 0.3629]
<i>M3D_{enc.}</i>	0.3148 [0.3048, 0.3238]	0.3917 [0.3834, 0.4002]	0.4885 [0.4804, 0.4969]	0.4009 [0.3878, 0.4143]	0.3487 [0.3463, 0.3513]	0.3312 [0.3203, 0.3422]
<i>LITA_{enc.}</i>	0.3106 [0.3009, 0.3199]	0.3902 [0.3822, 0.3983]	0.4904 [0.4822, 0.4984]	0.3880 [0.3801, 0.4078]	0.3562 [0.3533, 0.3591]	0.3587 [0.3477, 0.3700]
<i>Structured AMC (External Dataset, n=500, Zero-shot Inference)</i>						
<i>R+Attr+Mask</i>	0.1137 [0.1112, 0.1160]	0.1577 [0.1548, 0.1607]	0.4514 [0.4461, 0.4568]	0.1425 [0.1188, 0.1678]	0.3772 [0.3709, 0.3843]	0.2564 [0.2404, 0.2731]
<i>R+Attr</i>	0.1076 [0.0963, 0.1151]	0.1570 [0.1540, 0.1601]	0.4493 [0.4438, 0.4549]	0.1365 [0.1130, 0.1611]	0.3762 [0.3698, 0.3838]	0.2530 [0.2368, 0.2703]
<i>R+Mask</i>	0.1128 [0.1098, 0.1160]	0.1628 [0.1595, 0.1662]	0.4207 [0.4149, 0.4267]	0.1313 [0.1084, 0.1552]	0.3616 [0.3569, 0.3665]	0.2323 [0.2160, 0.2494]
<i>R</i>	0.1081 [0.0913, 0.1229]	0.1588 [0.1557, 0.1620]	0.4465 [0.4406, 0.4522]	0.1280 [0.1057, 0.1514]	0.3645 [0.3584, 0.3707]	0.2244 [0.2063, 0.2390]
<i>M3D_{enc.}</i>	0.1126 [0.1000, 0.1216]	0.1566 [0.1537, 0.1595]	0.4450 [0.4393, 0.4509]	0.1221 [0.1010, 0.1443]	0.3576 [0.3522, 0.3638]	0.2126 [0.1966, 0.2296]
<i>LITA_{enc.}</i>	0.1079 [0.1053, 0.1105]	0.1578 [0.1547, 0.1607]	0.4473 [0.4420, 0.4528]	0.1264 [0.1042, 0.1498]	0.3597 [0.3547, 0.3657]	0.2241 [0.2078, 0.2414]

tic quantitative biomarkers. To further analyze our model’s dependence on segmentation quality, Supplementary Tab. S5 presents a mask perturbation stress test, showing that organ-mask distortions affect performance more strongly than lesion-mask distortions by weakening ROI guidance.

Zero-shot inference on the external AMC benchmark (Tab. 2, bottom) validates the architecture’s generalizability. The full model achieves the highest scores across most metrics, particularly in clinical evaluations (GREEN: 0.1425, CRG: 0.3772, GPT-CA: 0.2564). While *R+Mask* marginally outperforms the full model in ROUGE (0.1628 vs. 0.1577), integrating *Attr* is crucial for maintaining clinical accuracy on unseen distributions, evidenced by the distinct improvements in GREEN and GPT-CA. Finally, comparing our standalone regional tokenizer (*R*) with external encoders (*M3D_{enc.}*, *LITA_{enc.}*) demonstrates *R*’s superior transferability. This indicates that our region-prioritized architecture

provides a more effective foundation for 3D feature extraction by concentrating the model’s capacity on clinically salient anatomical areas, which directly translates to robust performance and generalization on external datasets.

Supplementary Fig. S1 visualizes the diagnostic influence of each component. The full model ($R + Attr + Mask$) correctly identifies the hallmark ground-glass opacities (GGOs) in the peripheral subpleural area and establishes an accurate clinical diagnosis of viral pneumonia, closely matching the GT. When *Mask* is removed ($R + Attr$), the model still captures the viral pneumonia diagnosis but shows less precision in describing the subpleural distribution. Conversely, removing *Attr* ($R + Mask$) results in the identification of nodules and GGOs but introduces erroneous emphysematous changes not present in the GT. The baseline variants, including *R*, *M3D_{enc.}*, and *LITA_{enc.}*, fail significantly by either reporting normal aeration or identifying non-specific nodules while missing the primary GGO findings. This qualitative evidence confirms that our integrated approach is necessary to maintain both localization precision and diagnostic coherence across varying clinical scenarios.

5 Conclusion

In this work, we presented **MedRegion-CT**, a novel MLLM framework that shifts the paradigm from global volumetric encoding to region-aware 3D representation for CT report generation. By integrating a Region-based SlowFast Tokenizer, a Mask-Driven Visual Extractor, and a Lesion Attribute Extractor, our model effectively emulates the hierarchical diagnostic workflow of radiologists. These innovations enable precise capture of fine-grained pathological details and quantitative clinical metadata that are often overlooked by conventional global-feature approaches. Extensive evaluation on our Structured Chest CT Report Generation Benchmarks demonstrates that MedRegion-CT establishes a new state-of-the-art, significantly outperforming existing 3D MLLMs in both linguistic fluency and clinical accuracy. Our results confirm that regional guidance is essential for minimizing hallucinations and ensuring reliable diagnostic reporting. We believe this region-centric approach provides a scalable and robust foundation for future 3D medical image interpretation.

Acknowledgements

The authors would like to express their sincere gratitude to Jihyun Kim, Jun-seong Lee, and Jeong Min Song for their invaluable assistance with external data curation, as well as their comprehensive technical and experimental support throughout this work.

References

1. Bai, F., Du, Y., Huang, T., Meng, M.Q.H., Zhao, B.: M3d: Advancing 3d medical image analysis with multi-modal large language models. arXiv preprint arXiv:2404.00578 (2024)
2. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: IEEvaluation@ACL (2005), <https://api.semanticscholar.org/CorpusID:7164502>
3. Bannur, S., Bouzid, K., Castro, D.C., Schwaighofer, A., Bond-Taylor, S., Ilse, M., P'erez-Garc'ia, F., Salvatelli, V., Sharma, H., Meissen, F., Ranjit, M.P., Srivastav, S., Gong, J., Falck, F., Oktay, O., Thieme, A., Lungren, M.P., Wetscherek, M.T., Alvarez-Valle, J., Hyland, S.L.: Maira-2: Grounded radiology report generation. ArXiv **abs/2406.04449** (2024), <https://api.semanticscholar.org/CorpusID:270357817>
4. Bassi, P.R., Yavuz, M.C., Wang, K., Chen, X., Li, W., Decherchi, S., Cavalli, A., Yang, Y., Yuille, A.L., Zhou, Z.: Radgpt: Constructing 3d image-text tumor datasets. ArXiv **abs/2501.04678** (2025), <https://api.semanticscholar.org/CorpusID:275358308>
5. Cereser, L., Cortiula, F., Simiele, C., Peruzzi, V., Bortolot, M., Tullio, A., Como, G., Zuiani, C., Girometti, R.: Assessing the impact of structured reporting on learning how to report lung cancer staging ct: A triple cohort study on inexperienced readers. European journal of radiology **171**, 111291 (2024), <https://api.semanticscholar.org/CorpusID:266909282>
6. Chen, L., Wang, Z., Ren, S., Li, L., Zhao, H., Li, Y., Cai, Z., Guo, H., Zhang, L., Xiong, Y., et al.: Next token prediction towards multimodal intelligence: A comprehensive survey. arXiv preprint arXiv:2412.18619 (2024)
7. Chen, Z., Bie, Y., Jin, H., Chen, H.: Large language model with region-guided referring and grounding for ct report generation. ArXiv **abs/2411.15539** (2024), <https://api.semanticscholar.org/CorpusID:274234743>
8. Chen, Z., Luo, L., Bie, Y., Chen, H.: Dia-llama: Towards large language model-driven ct report generation. ArXiv **abs/2403.16386** (2024), <https://api.semanticscholar.org/CorpusID:268681763>
9. Claessens, Y.E., Debray, M.P., Tubach, F., Brun, A.L., Rammaert, B., Hausfater, P., Naccache, J.M., Ray, P., Choquet, C., Carette, M.F., Mayaud, C., Leport, C., Duval, X.: Early chest computed tomography scan to assist diagnosis and guide treatment decision for suspected community-acquired pneumonia. American journal of respiratory and critical care medicine **192** 8, 974–82 (2015), <https://api.semanticscholar.org/CorpusID:38597236>
10. Eslami, S., Meinel, C., De Melo, G.: Pubmedclip: How much does clip benefit visual question answering in the medical domain? In: Findings of the Association for Computational Linguistics: EACL 2023. pp. 1181–1193 (2023)
11. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019)
12. Gao, Y., Marshall, D., Xing, X., Ning, J., Papanastasiou, G., Yang, G., Komorowski, M.: Anatomy-guided radiology report generation with pathology-aware regional prompts. arXiv preprint arXiv:2411.10789 (2024)
13. Goergen, S.K., Pool, F.J., Turner, T.J., Grimm, J.E., Appleyard, M.N., Crock, C., Fahey, M.C., Fay, M.F., Ferris, N.J., Liew, S.M., et al.: Evidence-based guideline for the written radiology report: Methods, recommendations and implementation challenges. Journal of medical imaging and radiation oncology **57**(1), 1–7 (2013)

14. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
15. Guo, Q., Mello, S.D., Yin, H., Byeon, W., Cheung, K.C., Yu, Y., Luo, P., Liu, S.: Regiongpt: Towards region understanding vision language model. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 13796–13806 (2024), <https://api.semanticscholar.org/CorpusID:268247535>
16. Hamamci, I.E., Er, S., Almas, F., Simsek, A.G., Esirgun, S.N., Dogan, I., Dasdelen, M.F., Durugol, O.F., Wittmann, B., Amiranashvili, T., et al.: Developing generalist foundation models from a multimodal dataset for 3d computed tomography. arXiv preprint arXiv:2403.17834 (2024)
17. Hamamci, I.E., Er, S., Shit, S., Reynaud, H., Kainz, B., Menze, B.: Crg score: A distribution-aware clinical metric for radiology report generation. arXiv preprint arXiv:2505.17167 (2025)
18. Hu, J.E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Chen, W.: Lora: Low-rank adaptation of large language models. ArXiv **abs/2106.09685** (2021), <https://api.semanticscholar.org/CorpusID:235458009>
19. Huang, D.A., Liao, S., Radhakrishnan, S., Yin, H., Molchanov, P., Yu, Z., Kautz, J.: Lita: Language instructed temporal-localization assistant. ArXiv **abs/2403.19046** (2024), <https://api.semanticscholar.org/CorpusID:268733134>
20. Huang, X., Huang, H., Shen, L., Yang, Y., Shang, F., Liu, J., Liu, J.: A refer-and-ground multimodal large language model for biomedicine. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 399–409. Springer (2024)
21. Hussain, S., Mubeen, I., Ullah, N., Shah, S.S.U.D., Khan, B.A., Zahoor, M., Ullah, R., Khan, F.A., Sultan, M.A.: Modern diagnostic imaging technique applications and risk factors in the medical field: a review. BioMed research international **2022**(1), 5164970 (2022)
22. Jin, H., Che, H., Lin, Y.M., Chen, H.: Promptmrg: Diagnosis-driven prompts for medical report generation. ArXiv **abs/2308.12604** (2023), <https://api.semanticscholar.org/CorpusID:261100888>
23. Khattab, O., Singhvi, A., Maheshwari, P., Zhang, Z., Santhanam, K., Vardhamanan, S., Haq, S., Sharma, A., Joshi, T.T., Moazam, H., et al.: Dspy: Compiling declarative language model calls into self-improving pipelines. arXiv preprint arXiv:2310.03714 (2023)
24. Lee, C., Park, S., Shin, C.I., Choi, W.H., Park, H.J., Lee, J.E., Ye, J.C.: Read like a radiologist: Efficient vision-language model for 3d medical imaging interpretation. arXiv preprint arXiv:2412.13558 (2024)
25. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., rong Wen, J.: Evaluating object hallucination in large vision-language models. In: Conference on Empirical Methods in Natural Language Processing (2023), <https://api.semanticscholar.org/CorpusID:258740697>
26. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), <https://aclanthology.org/W04-1013/>
27. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM computing surveys **55**(9), 1–35 (2023)
28. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam. ArXiv **abs/1711.05101** (2017), <https://api.semanticscholar.org/CorpusID:3312944>

29. Ma, C., Jiang, Y., Wu, J., Yuan, Z., Qi, X.: Groma: Localized visual tokenization for grounding multimodal large language models. ArXiv **abs/2404.13013** (2024), <https://api.semanticscholar.org/CorpusID:269283071>
30. MacMahon, H., Naidich, D.P., Goo, J.M., Lee, K.S., Leung, A.N., Mayo, J.R., Mehta, A.C., Ohno, Y., Powell, C.A., Prokop, M., et al.: Guidelines for management of incidental pulmonary nodules detected on ct images: from the fleischner society 2017. *Radiology* **284**(1), 228–243 (2017)
31. Marcovici, P.A., Taylor, G.A.: Journal club: Structured radiology reports are more complete and more effective than unstructured reports. *AJR. American journal of roentgenology* **203** 6, 1265–71 (2014), <https://api.semanticscholar.org/CorpusID:207326910>
32. Ostmeier, S., Xu, J., Chen, Z., Varma, M., Blankemeier, L., Bluethgen, C., Michalson, A.E., Moseley, M.E., Langlotz, C.P., Chaudhari, A.S., Delbrouck, J.B.: Green: Generative radiology report evaluation and error notation. ArXiv **abs/2405.03595** (2024), <https://api.semanticscholar.org/CorpusID:269605082>
33. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Isabelle, P., Charniak, E., Lin, D. (eds.) *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (Jul 2002). <https://doi.org/10.3115/1073083.1073135>, <https://aclanthology.org/P02-1040/>
34. Pérez-García, F., Sharma, H., Bond-Taylor, S., Bouzid, K., Salvatelli, V., Ilse, M., Bannur, S., Castro, D.C., Schwaighofer, A., Lungren, M.P., et al.: Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence* pp. 1–12 (2025)
35. Rajbhandari, S., Rasley, J., Ruwase, O., He, Y.: Zero: Memory optimizations toward training trillion parameter models. *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis* pp. 1–16 (2019), <https://api.semanticscholar.org/CorpusID:269617042>
36. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
37. Sharma, H., Salvatelli, V., Srivastav, S., Bouzid, K., Bannur, S., Castro, D.C., Ilse, M., Bond-Taylor, S., Ranjit, M.P., Falck, F., et al.: Maira-seg: Enhancing radiology report generation with segmentation-aware multimodal large language models. arXiv preprint arXiv:2411.11362 (2024)
38. Shi, Y., Yang, S., Zhu, X., Wang, H., Fu, X., Li, M., Wu, J.: Medm-vl: What makes a good medical lvlm? In: *International Workshop on Agentic AI for Medicine*. pp. 290–299. Springer (2025)
39. Shi, Y., Zhu, X., Hu, Y., Guo, C., Li, M., Wu, J.: Med-2e3: A 2d-enhanced 3d medical multimodal large language model. arXiv preprint arXiv:2411.12783 (2024)
40. Tanida, T., Müller, P., Kaissis, G., Rueckert, D.: Interactive and explainable region-guided radiology report generation. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* pp. 7433–7442 (2023), <https://api.semanticscholar.org/CorpusID:258179419>
41. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* pp. 9568–9578 (2024), <https://api.semanticscholar.org/CorpusID:266976992>

42. Wang, L., Wang, H., Yang, H., Mao, J., Yang, Z., Shen, J., Li, X.: Interpretable bilingual multimodal large language model for diverse biomedical tasks (2025), <https://arxiv.org/abs/2410.18387>
43. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence* **5**(5), e230024 (2023)
44. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *arXiv preprint arXiv:2308.02463* (2023)
45. Xin, Y., Ates, G.C., Gong, K., Shao, W.: Med3dvlm: An efficient vision-language model for 3d medical image analysis. *IEEE Journal of Biomedical and Health Informatics* (2025)
46. Xu, M., Gao, M., Gan, Z., Chen, H.Y., Lai, Z., Gang, H., Kang, K., Dehghan, A.: Slowfast-llava: A strong training-free baseline for video large language models. *ArXiv abs/2407.15841* (2024), <https://api.semanticscholar.org/CorpusID:271329151>
47. Yeh, C.H., Wang, J., Graham, A.D., Liu, A.J., Tan, B., Chen, Y., Ma, Y., Lin, M.C.: Insight: A multi-modal diagnostic pipeline using llms for ocular surface disease diagnosis. *ArXiv abs/2410.00292* (2024), <https://api.semanticscholar.org/CorpusID:273022716>
48. Yuan, Y., Li, W., Liu, J., Tang, D., Luo, X., Qin, C., Zhang, L., Zhu, J.: Osprey: Pixel understanding with visual instruction tuning. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 28202–28211 (2023), <https://api.semanticscholar.org/CorpusID:266335219>
49. Zhang, S., Sun, P., Chen, S., Xiao, M., Shao, W., Zhang, W., Chen, K., Luo, P.: Gpt4roi: Instruction tuning large language model on region-of-interest. *ArXiv abs/2307.03601* (2023), <https://api.semanticscholar.org/CorpusID:259375716>
50. Zhang, X., Wu, C., Zhao, Z., Lei, J., Zhang, Y., Wang, Y., Xie, W.: Radgenome-chest ct: A grounded vision-language dataset for chest ct analysis. *arXiv preprint arXiv:2404.16754* (2024)
51. Zhou, H.Y., Acosta, J.N., Adithan, S., Datta, S., Topol, E.J., Rajpurkar, P.: Med-versa: A generalist foundation model for medical image interpretation (2025), <https://arxiv.org/abs/2405.07988>