Mitigating Entity Hallucinations in 3D Radiology Report Generation via Dual-Stream Alignment

Lingyu Zhou[1*], **Yue Yu**[1*], **Zhang Yi**[1], **Xiuyuan Xu**[1†]

[1]Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu, China
scu_zly@stu.scu.edu.cn, xavier_yu@foxmail.com
zhangyi@scu.edu.cn, xuxiuyuan@scu.edu.cn

Abstract

Entity hallucination poses a major challenge in radiology report generation (RRG), particularly for 3D CT scans where complex spatial contexts amplify factual errors. To address this, medical entity phrases serve as key carriers for multi-modal prompting, integrating expert knowledge into the vision-language model. Current methods use unified cross-attention for volume-phrase alignment, failing to account for anatomical specificity during the alignment process. In this work, we introduce the **D**ual-stream **E**ntity **A**lignment **R**eporting network (DEAR) that separately models organ and lesion entities to resolve anatomical bias. Specifically, the dual-stream entity aligner is designed to partition medical entity phrases into organ and lesion streams, feeding them into separate cross-attention blocks in parallel to achieve fine-grained volume–phrase alignment. For structurally regular and spatially stable organ entities, an organ-guided cross-attention (OGCA) block is proposed to enforce structural consistency by retrieving the top-k voxel tokens via volume–phrase similarity and preserving spatial connectivity through morphological dilation. Meanwhile, a lesion-guided cross-attention (LGCA) block is introduced for structurally irregular and spatially variable lesion entities, enhancing anomaly sensitivity through phrase-weighted attention and refining discriminative boundaries via 3D residual Laplacian filtering. Experiments demonstrate that DEAR significantly reduces entity hallucinations and improves clinical factuality in 3D RRG benchmarks.

Code — https://github.com/SCU-zly/DEAR

Introduction

Hallucination commonly denotes the generation of content that, while maintaining logical coherence, deviates from factual correctness (Huang et al. 2025). The issue of hallucination becomes especially prominent when vision-language architectures, initially developed and refined for natural image processing, are adapted for radiology report generation (RRG).

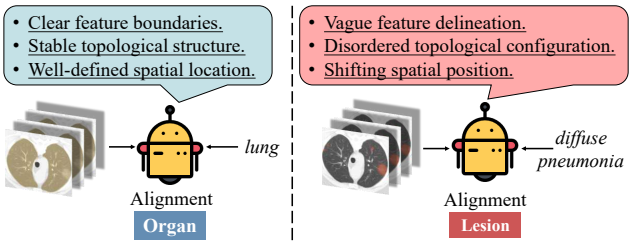

Figure 1: Visualization and interpretation of the anatomical discrepancies between organ and lesion entities. Different types of entity phrases tend to align with regions of interest that exhibit distinct anatomical characteristics.

While vision-language models have shown promise in various applications with inference efficiency comparable to expert radiologists, they frequently generate significant hallucinations, especially when producing critical medical entities. The challenge is further compounded in the context of three-dimensional medical imaging modalities, such as computed tomography (CT), where the vast amount of dense spatial contextual information introduces additional complexity. Consequently, entity hallucinations emerge as a major bottleneck in the deployment of reliable and clinically viable 3D RRG systems.

Leveraging the recent advances in large language models (LLMs), a growing body of research has explored vision-language prompting as an effective strategy to mitigate entity hallucination in radiology report generation. One straightforward approach involves introducing radiology expert report templates (Liu et al. 2024) for prompting. However, phrases within the templates that are not directly related to medical knowledge may interfere with the model's learning process. Meanwhile, (Li et al. 2023; Hou et al. 2024) et al. propose implicitly modeling medical domain knowledge through graph networks, aiming to reduce the impact of redundant phrase noise. Nevertheless, their approach fails to adequately consider visual features, potentially resulting in diminished performance when addressing entity hallucination issues. While some methods (Chen et al. 2024b; Xiao et al. 2025) incorporate region extractors to highlight key entities, the excessive cost of region annotations ultimately hinders the model's overall development.

[*]These authors contributed equally.
[†]Corresponding author.

Why not integrate medical entity phrases as a means of domain-specific knowledge? Entity phrases used in radiology reports are consistent and relatively easy to collect. Each phrase represents a critical entity in radiology reports, and simply highlighting the corresponding regions of interest within visual tokens could significantly reduce hallucinations (Li et al. 2024). Recent studies (Jin et al. 2024; Zhang et al. 2025) focus on modeling tokens aligned with medical entity phrases and visual features. Regrettably, such one-size-fits-all strategies are limited in their ability to effectively mitigate entity hallucination, particularly when it comes to preserving the anatomical fidelity required for high-stakes clinical reporting.

Medical entities fall into two main types: organs and lesions, each with distinct anatomical characteristics. As shown in Fig. 1, the organ *lung* typically occupies a consistent volume within the entire CT scan, displaying well-defined feature boundaries and a stable topological structure. In contrast, the lesion *diffuse pneumonia* typically occupies a variable volume, is characterized by indistinct boundaries, and exhibits a relatively disordered topological distribution.

Motivated by anatomical disparities, we propose a **D**ual-stream **E**ntity **A**lignment **R**eporting network (DEAR) that systematically models organ and lesion entities separately for radiology report generation. For any given 3D spatial features, DEAR aligns them in parallel using medical entity phrases. The novel dual-stream entity aligner we propose leverages the distinct anatomical properties of entities, utilizing a dual-stream approach for volume-phrase alignment, and designs different alignment mechanisms based on anatomical specificity. Specifically, for the organ phrase stream, DEAR introduces organ-guided cross-attention (OGCA) blocks to capture high-value volume tokens with respect to spatial structural consistency during the alignment process. OGCA first constructs a similarity matrix between volume and phrase to identify the top-k most relevant voxel tokens, and then applies morphological dilation to expand them into coherent spatial regions, enabling more comprehensive and anatomically consistent attention coverage. Concurrently, for the lesion phrase stream, DEAR incorporates a novel lesion-guided cross-attention (LGCA) block, designed to maintain sensitivity to anomalous and irregular spatial patterns. LGCA weights each volume token based on its similarity to lesion phrases, subtly shaping attention around abnormal regions, followed by a 3D residual Laplacian enhancement step to locally amplify boundary details of suspected lesions. Our contributions are as follows:

- We propose a dual-stream entity aligner that explicitly models organ and lesion entities according to their distinct anatomical characteristics, enabling separate volume-phrase alignments that reduce entity hallucination.
- We design an organ-guided cross-attention block that extracts coherent structural features for organs using spatial regularization and 3D dilated convolutions, suppressing redundant spatial noise while preserving the topological continuity of organs.
- We develop a lesion-guided cross-attention strategy to enhance lesion-specific attention through adaptive similarity weighting and 3D residual Laplacian refinement, highlighting the spatial context of lesions while enhancing discriminative boundary details.

Related Work

Vision-Language Alignment

To mitigate hallucinations, early alignment strategies enhanced visual-semantic coupling through several methodologies. DCL injected pathological correlations via dynamic knowledge graphs (Li et al. 2023), RGRG achieved fine-grained locality by generating descriptions for predefined anatomical regions (Tanida et al. 2023), and AdaMatch associated variably-sized image patches with specific lexical items (Chen et al. 2024b). Addressing the limitations in semantic coherence of prior methods, the integration of LLMs as decoders advanced the generative and reasoning capabilities of RRG models. RadAlign combines a discriminative visual feature extractor with an LLM to achieve more robust alignment (Gu et al. 2025). While fVLMs attempt to mitigate hallucinations through region-of-interest annotation, this strategy incurs prohibitive annotation costs that hinder scalability (Shui et al. 2025).

Automated Radiology Report Generation

In 2D RRG, initial work introduced modules such as relational memory (Chen et al. 2020) and medical concept generation networks (Wang et al. 2022) to manage long-form reports. Subsequent efforts focused on more effective visual guidance, AM-MRG mines disease-aware tokens to enhance sensitivity (Wang et al. 2025a) and HC-LLM integrates longitudinal data to capture temporal changes (Liu et al. 2025b).

The transition to 3D RRG, however, introduces denser spatial contexts that significantly aggravate the hallucination problem. Existing studies (Hamamci, Er, and Menze 2024; Chen et al. 2024a) employ unified visual modeling for 3D radiological images, they predominantly disregard objective anatomical biases among key medical entities. Although methods like PromptMRG (Jin et al. 2024) and MEPNet (Zhang et al. 2025) enhance model performance by explicitly classifying features aligned with medical entities and incorporating expert knowledge bases. These methods fall short in thoroughly leveraging the anatomical context underlying entity-specific alignments. Instead, they indiscriminately accumulate parameters and inject redundant domain knowledge, imposing significant computational overhead and semantic overgeneralization, thereby undermining hallucination mitigation in complex medical scenarios.

Method

Overall Framework

As shown in Fig. 2, DEAR aims to generate expert-level radiology reports. Given a set of M 3D radiological volume images $\mathbb{V} = \{\mathcal{V}_1, \mathcal{V}_2, ..., \mathcal{V}_M\}$ and a collection of N frequently occurring medical entity phrases $\mathbb{E} = \{\mathcal{E}_1, \mathcal{E}_2, ..., \mathcal{E}_N\}$, the model aims to produce accurate textual descriptions. Mathematically, for each volume $\mathcal{V}_i \in \mathbb{V}$,

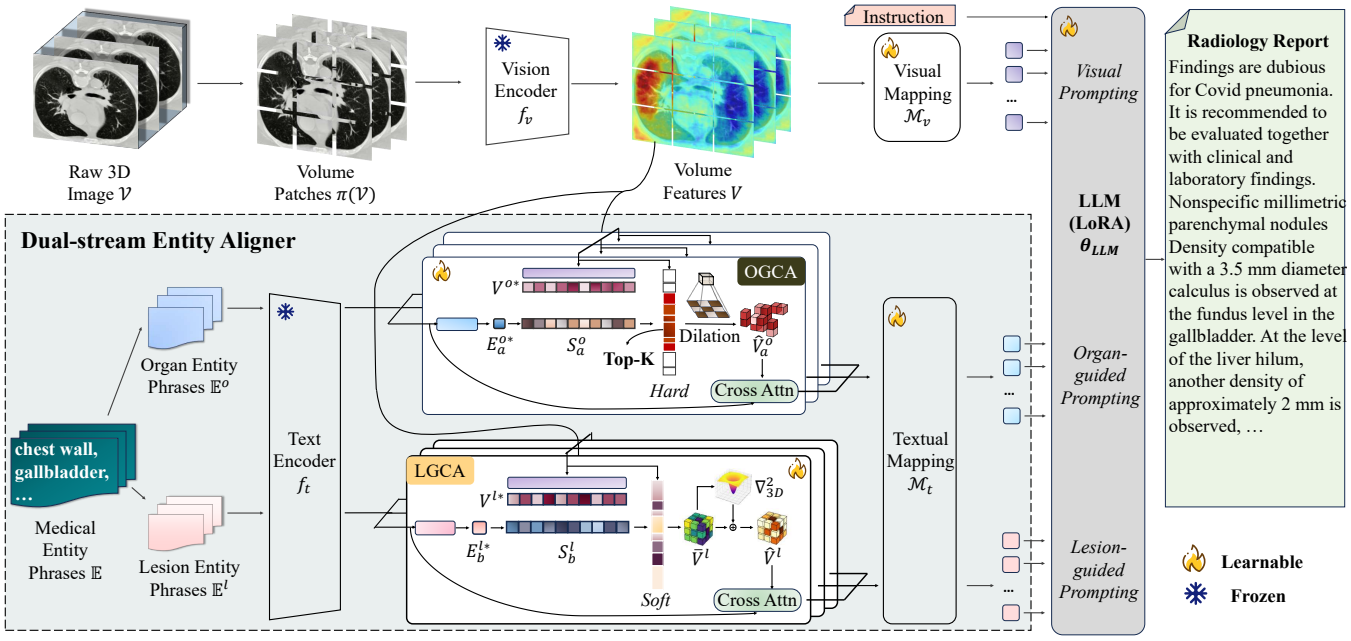

Figure 2: Overview of our proposed DEAR framework.

we have $\mathcal{V}_i \in \mathbb{R}^{H\times W\times S}$, indicating that each input is a 3D tensor of spatial resolution $H \times W$ with S slices. DEAR aims to approximate the ground-truth report labels $\mathbb{Y} = \{\mathcal{Y}_1, \mathcal{Y}_2, ..., \mathcal{Y}_M\}$, where each $\mathcal{Y}_j \in \mathbb{Y}$ belongs to a discrete textual space $\mathcal{T}$. Specifically, for each raw 3D image, DEAR uses the entire set $\mathbb{E}$ as prompting input to the text encoder to perform alignment. Finally, DEAR outputs the complete radiology report via the LLM θ_{LLM}.

To encode the discriminative spatial context of each volume image $\mathcal{V}$, DEAR first utilizes a 3D patch embedding module $\pi(\cdot)$ to divide the input into volume patches. A pretrained vision encoder f_v is then applied to extract volumetric features $V = \{v_1, v_2, ..., v_m\} \in \mathbb{R}^{b\times q\times d_v}$, where b denotes the batch size, q is the length of the token sequence, and d_v is the visual feature dimension. Following the same processing pipeline as in (Chen et al. 2024b; Wang et al. 2025b; Zhang et al. 2025), a learnable visual mapping layer $\mathcal{M}_v$ is introduced to transform the spatial visual representations into the LLM-compatible space with dimension d_{llm}.

Recently, formulating medical entity phrases as prompts has emerged as a novel paradigm for injecting expert medical knowledge into the report generation pipeline (Zhang et al. 2025; Mei et al. 2024; Altalla' et al. 2025). Specifically, we utilize a text encoder f_t to obtain textual feature queries. After the dual-stream entity aligner performs volume-phrase alignment, we project the token queries into the LLM feature space with dimension d_{llm} using an adaptive textual mapping module $\mathcal{M}_t$.

The final vision-language prompt Γ can be summarized into the following three components: (1) an $\{$*introduction*$\}$ that includes the basic task description and generation instructions; (2) $\{$*visual prompting*$\}$ enclosed by *[Img]* and *[/Img]* tags; and (3) entity promptings, which consist of two parts: the $\{$*organ-guided prompting*$\}$ representing the semantic features of organ-guided cross-attention, and the $\{$*lesion-guided prompting*$\}$ encoding the cross-modal alignment between visual and lesion phrase information. DEAR employs cross-entropy loss $\mathcal{L}$ to optimize the LLM on the word embeddings $\underbrace{\{y_1, y_2, \ldots, y_t\}}_{L}$ of each report label $\mathcal{Y}$:

$$\mathcal{L} = -\sum_{t=1}^{L} \log p(y_t | y_{1:t-1}, \Gamma, \theta_{LLM}). \quad (1)$$

Dual-stream Entity Aligner

The visual context discrepancy between organs and lesions often contributes to hallucination issues in RRG. This issue becomes especially pronounced in 3D RRG, where existing backbones demonstrate considerably higher rates of factual errors on medical entities. The problem stems not only from the vastly expanded spatial context in 3D compared to 2D, but also from the anatomical representational differences between organs and lesions. Organs tend to exhibit high spatial continuity and relatively stable topological structures in 3D space, while lesions are typically sparse, locally concentrated, and topologically irregular (Huang et al. 2024; Qin et al. 2022; Kerioui et al. 2022). Existing solutions (Zhao et al. 2024; Zhang et al. 2025; Jin et al. 2024) often focus on selecting or classifying aligned features, yet overlook the alignment of discriminative anatomical characteristics critical for reducing hallucinations.

To address this challenge, DEAR introduces a novel dual-stream entity aligner to perform volume-phrase alignment separately for organ and lesion phrases. We first divide the phrase features into two distinct query sets: $\mathbb{E}^o$ representing organ phrases, and $\mathbb{E}^l$ representing lesion phrases. For

any $\mathcal{E}_a^o \in \mathbb{E}^o$ and $\mathcal{E}_b^l \in \mathbb{E}^l$, the corresponding features E_a^o and E_b^l are obtained through the text encoder f_t. To handle the two types of phrase feature queries, we further design distinct cross-attention mechanisms tailored to attaching semantically specific medical entity representations.

Organ-Guided Cross-Attention

Naively employing volume-phrase alignment inevitably introduces redundant spatial context, and volume features that are irrelevant to the target entity phrases may further amplify hallucination. To address this, we propose organ-guided cross-attention (OGCA) to highlight discriminative volumes and enhance the structural coherence of organ-related representations.

For any organ phrase feature E_a^o, we project both volumetric visual features V^o and the phrase embedding into a shared vector space using learnable projectors. The projected organ phrase feature E_a^{o*} is then used to compute cosine similarity scores $s_{a\mu}^o$ with each projected volume token $v_\mu^{o*} \in V^{o*}$:

$$s_{a\mu}^o = \frac{(E_a^{o*})^\top \cdot v_\mu^{o*}}{||E_a^{o*}|| \cdot ||v_\mu^{o*}||}, \tag{2}$$

where $\mu \in \{1, 2, \ldots, \alpha\}$. We then select the top-k visual tokens that are most relevant to the phrase, following the procedure below:

$$\mathcal{G}_a = \text{TopK}_{\mu=1}^m(s_{a\mu}^o, k), \tag{3}$$

we obtain the top-k relevant visual token sequence $\mathcal{G}_a$. OGCA enhances both efficiency and robustness by selecting only the top-k volume tokens most relevant to the medical entity phrase E_a^o. This approach reduces computational overhead while minimizing interference from low-relevance spatial noise.

To maintain the structural connectivity of the 3D volumes, we apply 3D morphological dilation to the top-k selected tokens, expanding them into a coherent spatial region. For each selected token v_u^o with index $u \in \mathcal{G}_a$, we apply a 3D dilated convolution kernel $K \in \mathbb{R}^{d_k \times h \times w}$, where d_k, h, and w denote the kernel sizes along the depth, height, and width, respectively. Assuming a dilation factor δ that defines the spacing between neighboring elements in the kernel, the dilated token representation $\hat{v_u^o}$ can be computed as follows:

$$\hat{v_u^o} = \sum_{d_k,h,w} v(x+\delta\cdot d_k, y+\delta\cdot h, z+\delta\cdot w)\cdot K(d_k, h, w). \tag{4}$$

We then perform a cross-attention operation between the dilated visual token sequence and the organ phrase E_a^o, yielding the aligned output ϕ_a^o corresponding to E_a^o.

$$\phi_a^o = \sum_u \text{Softmax}(\frac{(W_Q^{(o,a)} E_a^o)^\top (W_K^{(o,a)} \hat{v}_u^o)}{\sqrt{d_\phi^{(o,a)}}}) \cdot (W_V^{(o,a)} \hat{v}_u^o), \tag{5}$$

$W_Q^{(o,a)}$, $W_K^{(o,a)}$, and $W_V^{(o,a)}$ denote the query, key, and value projection matrices for the cross-attention ϕ_a^o, respectively, and $d^{(o,a)}\phi$ represents the scaling factor for the feature dimension.

Lesion-Guided Cross-Attention

A lesion is an area of abnormal tissue with highly variable shapes, which can occur in various parts of the body (Comelli et al. 2018). Applying a uniform cross-attention mechanism to align lesion entities may overlook critical yet subtle details. To address the irregular topological structures, varying spatial locations, and fuzzy boundaries typically associated with lesion features, we introduce a novel lesion-guided cross-attention (LGCA) mechanism. The LGCA block is designed to guide cross-attention, focusing on fine-grained and non-conventional abnormal regions, thereby enhancing sensitivity to lesion-specific patterns.

Given the lesion phrase feature E_b^{l*} and visual tokens V^{l*} projected into a shared space, we compute their pairwise similarity over β visual tokens as follows:

$$S_b^l = \Big\{ s_{b\tau}^l = \frac{(E_b^{l*})^\top \cdot v_\tau^{l*}}{||E_b^{l*}|| \cdot ||v_\tau^{l*}||} \Big| \tau \in \{1, 2, \ldots, \beta\},\ v_\tau^{l*} \in V^{l*} \Big\}. \tag{6}$$

Based on the similarity scores S_b^l, we compute the phrase-weighted feature $\bar{V}^l$ over the original pre-projection visual tokens V^l to obtain:

$$\bar{V}^l = \{\bar{v}_\tau^l = s_{b\tau}^l \cdot v_\tau^l | s_{b\tau}^l \in S_b^l, v_\tau^l \in V^l\}. \tag{7}$$

To further guide the model in learning the abnormal boundary and critical details of lesions, we highlight structural discontinuities in the visual features using a residual Laplacian approach. Specifically, for any phrase-weighted token $\bar{v}_\tau^l$, LGCA locally enhances edge information by applying a 3D Laplacian kernel (Pang et al. 2024), which accentuates abnormal spatial context. Inspired by the residual learning principle (He et al. 2016), the Laplacian-based edge enhancement is incorporated as a residual branch added to the visual features, resulting in a local enhanced token that is more sensitive to lesion-related abnormalities.

$$\hat{v}_\tau^l = v_\tau^l + \eta \cdot \nabla_{3\text{D}}^2 \bar{v}_\tau^l, \tag{8}$$

$\nabla_{3\text{D}}^2$ denotes a fixed 3D Laplacian convolution kernel, and η is a learnable scaling factor used to balance the influence of edge enhancement against the original semantic features. The enhanced token $\hat{v}_\tau^l$ is then fed into the cross-attention block for cross-modal alignment.

$$\phi_b^l = \sum_\beta \text{Softmax}(\frac{(W_Q^{(l,b)} E_b^l)^\top (W_K^{(l,b)} \hat{v}_\tau^l)}{\sqrt{d_\phi^{(l,b)}}}) \cdot (W_V^{(l,b)} \hat{v}_\tau^l). \tag{9}$$

where $W_Q^{(l,b)}$, $W_K^{(l,b)}$, $W_V^{(l,b)}$, and $d_\phi^{(l,b)}$ denote the parameters of the cross-attention ϕ_b^l.

Guided by the anatomical characteristics of lesions, LGCA explicitly emphasizes local structural discontinuities and lesion boundaries during the voxel–phrase alignment process, enhancing the interpretability of the model.

Experiment

Settings

We use M3D-CLIP (Bai et al. 2024) as the vision encoder f_v, while LLaMA3-8B (Dubey et al. 2024) with 8-bit quan-

Datasets	Type	Methods	B-1	B-4	MTR	R-L	F1	Rec.	Prec.	GREEN
BIMCV-R	2D‡	RADAR (ACL'2025)	0.391	0.248	0.402	0.453	0.155	0.104	0.301	0.258
		MLRG (CVPR'2025)	0.398	0.254	0.418	0.466	0.160	0.108	0.308	0.264
	3D	RadFM	0.404	0.278	0.455	0.475	0.165	0.111	0.319	0.291
		CT2-Rep† (MICCAI'2024)	0.410	0.285	0.463	0.483	0.125	0.115	0.326	–
		fVLM (ICLR'2025)	0.422	0.298	0.475	0.491	0.185	0.127	0.339	0.318
		Reg2RG (TMI'2025)	0.439	0.304	0.486	0.500	0.197	0.137	0.351	0.332
		DEAR (Ours)	**0.447**	**0.312**	**0.490**	**0.519**	**0.256**	**0.188**	**0.402**	**0.366**
CT-RATE	2D‡	RADAR (ACL'2025)	0.432	0.229	0.387	0.301	0.185	0.124	0.368	0.298
		MLRG (CVPR'2025)	0.438	0.230	0.393	0.309	0.191	0.128	0.374	0.312
	3D	RadFM†	0.442	0.237	0.399	0.315	0.195	0.131	0.382	–
		CT2-Rep† (MICCAI'2024)	0.460	**0.369**	0.295	0.459	0.534	0.548	0.749	–
		fVLM (ICLR'2025)	0.481	0.260	0.452	0.392	0.270	0.194	0.448	0.401
		Reg2RG† (TMI'2025)	0.473	0.249	0.441	0.367	0.253	0.181	0.423	–
		DEAR (Ours)	**0.491**	0.366	**0.462**	**0.472**	**0.677**	**0.598**	**0.782**	**0.511**
CTRG(Chest)	2D‡	RADAR (ACL'2025)	0.409	0.300	0.430	0.483	0.186	0.143	0.269	0.281
		MLRG (CVPR'2025)	0.418	0.304	0.428	0.474	0.187	0.142	0.273	0.285
	3D	RadFM†	0.487	0.309	0.492	0.491	–	–	–	–
		CT2-Rep† (MICCAI'2024)	0.423	0.292	0.470	0.502	0.139	0.102	0.325	–
		fVLM (ICLR'2025)	0.487	0.318	0.489	0.483	0.249	0.176	0.425	0.418
		Reg2RG† (TMI'2025)	0.496	0.320	0.497	0.478	–	–	–	–
		DEAR (Ours)	**0.502**	**0.329**	**0.501**	**0.514**	**0.306**	**0.225**	**0.478**	**0.463**

Table 1: Comparisons on the three public datasets. Bold font indicates the best results for a given metric. † denotes results reported in the original literature. ‡ indicates that the model is adapted to the 3D datasets by replacing its original 2D encoder with the M3D-CLIP visual backbone.

tization is adopted as the language model. For efficient fine-tuning of the LLM θ_{LLM} for text generation, we employ LoRA (Hu et al. 2022) with a rank of 64. We collect 32 organ-related medical phrases and 48 lesion-related medical phrases from publicly available radiological text sources (Langlotz 2006; Li et al. 2024). The dilation factor δ in OGCA is set to 3. The batch size is set to 4, and we train the model for 50 epochs using the AdamW optimizer with a learning rate of 1e-4. All experiments are conducted using PyTorch 2.1.2 and Python 3.10, running on 2 NVIDIA GeForce RTX 4090 GPUs.

Datasets

We evaluated effectiveness and robustness using three publicly available 3D radiology report generation datasets, covering diverse clinical scenarios and data distributions.

CT-RATE The CT-RATE dataset (Hamamci et al. 2024) is a substantial collection of 25,692 non-contrast 3D chest CT scans from 21,304 unique patients, each paired with a detailed radiology report. For model training, we utilized the 'Findings' section of each report, which contains the detailed radiological observations.

BIMCV-R The BIMCV-R dataset (Tang et al. 2024) is an extension of the BIMCV-COVID-19+ collection, providing 8,069 3D CT volumes. It features wide pathological diversity, encompassing 96 distinct disease types such as oncological, infectious, cardiovascular, and respiratory conditions.

CTRG-Chest The CTRG-Chest-548K dataset (Chen et al. 2024c) consists of 1,804 3D CT volume-report pairs. Consistent with standard practice for report generation tasks, we chose the 'report' section from the reports for model training. All datasets were partitioned into training, validation, and testing sets using a 7:2:1 ratio.

Evaluation Metrics

To evaluate the performance of our model, we employ a comprehensive set of metrics covering both natural language generation (NLG) quality and clinical efficacy. For NLG assessment, we use BLEU-1 (B-1) and BLEU-4 (B-4) (Papineni et al. 2002), which measure the n-gram precision between generated and reference reports; METEOR (MTR) (Banerjee and Lavie 2005), which considers synonymy and stemming for a more semantically-aware comparison; and ROUGE-L (R-L) (Lin 2004), which evaluates the longest common subsequence to gauge recall. For clinical efficacy, we compute Recall (Rec.), Precision (Prec.), and the F1-score (F1) based on the accurate detection and description of key clinical keywords (medical entities) within the reports.

Furthermore, to more directly assess the clinical factual correctness of the generated reports, we adopt **GREEN** (Ostmeier et al. 2024), a specialized evaluation metric designed for radiology. GREEN uses an LLM to convert gener-

Methods	Generated Chest CT Reports	Organ Entity Phrase	Lesion Entity Phrase
MLRG	Trachea, both main bronchi are normal. Mediastinal main vascular structures, heart contour, size are normal. **No nodules or infiltrative lesions are detected in the lungs. Moderate pleural effusion is observed on the right.** Bone structures in the study area are natural.	*Mediastinal main vascular structures, heart, Thoracic aorta, Thoracic esophageal, lymph nodes, lung parenchyma, upper lobe, lower lobe, Upper abdominal organs, liver, adrenal glands, Bone structures*	*Pericardial effusion-thickening, tumoral wall thickening, enlarged lymph nodes, nonspecific nodules, mild recessions, infiltrative lesion, Pleural effusion-thickening, space-occupying lesion*
CT2Rep	**Trachea,** both **main bronchi are adequately visualized and appear patent. Mediastinal main vascular structures, heart contour is at the upper limits of normal.** Thoracic aorta diameter is normal. No enlarged lymph nodes were detected. When examined in the lung parenchyma window; **A few millimetric nonspecific nodules are observed in the upper lobe of the left lung.** Aeration of both lung parenchyma is normal and no infiltrative lesion is detected in the lung parenchyma. Pleural effusion-thickening was not detected.	*Mediastinal main vascular structures, heart, Thoracic aorta, Thoracic esophageal, lymph nodes, lung parenchyma, upper lobe, lower lobe, Upper abdominal organs, liver, adrenal glands, Bone structures*	*Pericardial effusion-thickening, tumoral wall thickening, enlarged lymph nodes, nonspecific nodules, mild recessions, infiltrative lesion, Pleural effusion-thickening, space-occupying lesion*
fVLM	Trachea, both main bronchi are open. Mediastinal main vascular structures, heart contour, size are normal. Thoracic aorta diameter is normal. No enlarged lymph nodes in prevascular, pre-paratracheal, subcarinal or bilateral hilar-axillary pathological dimensions were detected. When examined in the lung parenchyma window; **A few subcentimeter nonspecific nodules are observed in the upper lobe of the right lung.** Aeration of both lung parenchyma is normal and no infiltrative lesion is detected in the lung parenchyma. Pleural effusion-thickening was not detected. Upper abdominal organs are normal.	*Mediastinal main vascular structures, heart, Thoracic aorta, Thoracic esophageal, lymph nodes, lung parenchyma, upper lobe, lower lobe, Upper abdominal organs, liver, adrenal glands, Bone structures*	*Pericardial effusion-thickening, tumoral wall thickening, enlarged lymph nodes, nonspecific nodules, mild recessions, infiltrative lesion, Pleural effusion-thickening, space-occupying lesion*
DEAR	**Trachea, both main bronchi are open. Mediastinal main vascular structures, heart contour, size are normal.** Thoracic aorta diameter is normal. Pericardial effusion-thickening was not observed. No enlarged lymph nodes in prevascular, pre-paratracheal, subcarinal or bilateral hilar-axillary pathological dimensions were detected. When examined in the lung parenchyma window; **A few millimetric nonspecific nodule and mild recessions are observed in the upper lobe and lower lobe of the right lung**. Aeration of both lung parenchyma is normal. Pleural effusion-thickening was not detected. The included upper abdominal organs appear normal. No space-occupying lesion is seen in the liver within the scanned area. Both adrenal glands are normal with no detectable mass. Bone structures and vertebral body heights are maintained.	*Mediastinal main vascular structures, heart, Thoracic aorta, Thoracic esophageal, lymph nodes, lung parenchyma, upper lobe, lower lobe, Upper abdominal organs, liver, adrenal glands, Bone structures*	*Pericardial effusion-thickening, tumoral wall thickening, enlarged lymph nodes, nonspecific nodules, mild recessions, infiltrative lesion, Pleural effusion-thickening, space-occupying lesion*
Ground-Truth	Trachea, both main bronchi are open. Mediastinal main vascular structures, heart contour, size are normal. Thoracic aorta diameter is normal. Pericardial effusion-thickening was not observed. Thoracic esophageal calibration was normal and no significant tumoral wall thickening was detected. No enlarged lymph nodes in prevascular, pre-paratracheal, subcarinal or bilateral hilar-axillary pathological dimensions were detected. When examined in the lung parenchyma window; A few millimetric nonspecific nodules and mild recessions are observed in the upper lobe and lower lobe of the right lung. Aeration of both lung parenchyma is normal and no infiltrative lesion is detected in the lung parenchyma. Pleural effusion-thickening was not detected. Upper abdominal organs included in the sections are normal. No space-occupying lesion was detected in the liver that entered the cross-sectional area. Bilateral adrenal glands were normal and no space-occupying lesion was detected. Bone structures in the study area are natural.	*Mediastinal main vascular structures, heart, Thoracic aorta, Thoracic esophageal, lymph nodes, lung parenchyma, upper lobe, lower lobe, Upper abdominal organs, liver, Bilateral adrenal glands, Bone structures*	*Pericardial effusion-thickening, tumoral wall thickening, enlarged lymph nodes, nonspecific nodules, mild recessions, infiltrative lesion, Pleural effusion-thickening, space-occupying lesion*

Figure 3: Qualitative visualization from the CT-RATE dataset. Generated reports highlight factual errors (red) and their DEAR-corrected counterparts (green). The entity checklists provide a granular performance breakdown, with colors denoting the overall status of entity-related descriptions: green (correct), yellow (omission), and red (errors).

ated and reference reports into structured clinical statements, then identifies error types like hallucinations, omissions, and attribute errors. This provides a more precise, clinically relevant assessment of factual accuracy.

Comparison Studies

Quantitative Results We compare DEAR with advanced RRG models across three datasets. Specifically, we evaluate against RadFM (Wu et al. 2023), CT2Rep (Hamamci, Er, and Menze 2024), fVLM (Shui et al. 2025), and Reg2RG (Chen et al. 2025), which are tailored for 3D RRG. To ensure comprehensive evaluation, we also include leading 2D RRG models, RADAR (Hou et al. 2025) and MLRG (Liu et al. 2025a) that have demonstrated strong performance. As shown in Tab. 1, DEAR achieves outstanding and robust performance across all three datasets, validating the effectiveness of the dual-stream entity aligner in mitigating entity hallucination. In contrast, 2D models exhibit inherent limitations in capturing rich spatial contextual information. The retrieval-based architectures of RADAR and MLRG significantly underperform, with F1 score deficits of 0.492 and 0.487, respectively, on BIMCV-R compared to DEAR. Current state-of-the-art 3D RRG models, such as fVLM and Reg2RG, incorporate fine-grained region-level alignment. However, their unified multimodal feature alignment approaches limit their ability to capture the topological differences between regular organs and irregular lesions. This limitation results in significant F1-score decreases of 0.407 and 0.424 on CT-RATE when compared to DEAR. Moreover, in NLG metrics, DEAR surpasses the best-performing baselines on the three datasets by 0.0186, 0.0131, and 0.012, respectively, confirming the fluency of its generated text. The effectiveness of DEAR is demonstrated by several key performance metrics. In terms of the GREEN score, DEAR outperforms fVLM by substantial margins of 0.048, 0.11, and 0.045 across the three datasets, with F1 scores exhibiting a similar advantage. The aforementioned experimental results demonstrate that DEAR generates reports with both comprehensive diagnostic coverage and clinical reliability, showcasing the model's strong ability to reduce critical clinical errors.

Qualitative Results In Fig. 3, we provide qualitative visualizations of reports generated by DEAR and other SOTA methods to better illustrate the superiority of our approach. As depicted, the 2D model, MLRG, omits numerous crit-

Exp.	Dual-stream	OGCA	LGCA	B-1	B-4	MTR	R-L	F1	Rec.	Prec.	GREEN
1	✗	✗	✗	0.312	0.184	0.251	0.268	0.392	0.355	0.438	0.219
2	✓	✗	✗	0.354	0.213	0.287	0.301	0.428	0.392	0.472	0.247
3	✓	✓	✗	0.448	0.342	0.431	0.445	0.637	0.562	0.734	0.439
4	✓	✗	✓	0.472	0.351	0.445	0.457	0.565	0.498	0.653	0.381
5	✓	✓	✓	0.491	0.366	0.462	0.472	0.677	0.598	0.782	0.511

Table 2: Ablation study evaluating the contribution of each module to CT-RATE. *Exp*. denotes the experiment index. *Exp*. 1 uses vanilla cross-attention between unified medical entity phrases and visual tokens. *Exp*. 2 splits entity phrases into organ and lesion streams with separate visual alignments and no parameter sharing. *Exp*. 3 and *Exp*. 4 enhance *Exp*. 2 by implementing OGCA and LGCA modules respectively. *Exp*. 5 represents DEAR's final configuration.

Top-*k*	3D dilation	B-4	Rec.	Prec.	GREEN
✗	$\delta = 1$	0.221	0.142	0.298	0.241
	$\delta = 2$	0.247	0.165	0.326	0.269
	$\delta = 3$	0.264	0.172	0.348	0.287
	$\delta = 4$	0.251	0.168	0.333	0.272
✓	$\delta = 1$	0.274	0.171	0.361	0.312
	$\delta = 2$	0.298	0.183	0.387	0.341
	$\delta = 3$	0.312	0.188	0.402	0.366
	$\delta = 4$	0.301	0.182	0.389	0.348

Table 3: Ablation study on BIMCV-R examining the effectiveness of each component in OGCA. We compare four different dilation factor configurations for the 3D dilation convolution.

Residual Laplacian	MTR	Rec.	Prec.	GREEN
✗	0.486	0.214	0.463	0.448
w/o Res.	0.492	0.221	0.452	0.453
✓	0.501	0.225	0.478	0.463

Table 4: Ablation study on CTRG(Chest) assessing the effectiveness of components in LGCA. "w/o Res." indicates the configuration using Laplacian filtering without residual connections.

ical findings. Although CT2Rep and fVLM generate more detailed reports, they suffer from gross factual errors, such as a missed diagnosis related to the "*lower lobe of the right lung*" and a misdiagnosis of "*subcentimeter nonspecific nodules*", indicating their deficiencies in describing fine-grained entity details and accurately localizing medical entities. In contrast, DEAR accurately and comprehensively describes all critical findings, correctly identifying the "*millimeter-scale nonspecific nodules*" and "*mild recessions*" located in the upper and lower lobes of the right lung. This demonstrates the efficacy of our dual-stream entity aligner. The DEAR achieves more accurate alignment between the generated text and visual evidence by independently modeling lesion and organ entity streams. This approach enables it to better handle irregular, subtle pathological patterns without neglecting the features of structurally stable organs.

Ablation Analysis

Influence of Each Module As demonstrated in Table 2, Exp. 1 without any module yields the lowest performance scores, confirming the absence of essential anatomical guidance. Adding the dual-stream mechanism in Exp. 2 produces modest gains (B-4 +0.029, GREEN +0.028 over Exp.1), demonstrating the presence of implicit anatomical priors. Incorporating OGCA in Exp. 3 significantly improves contextual metrics (B-4 +0.129, R-L +0.144) and GREEN (+0.192) compared to Exp. 2, while maintaining high organ precision at 0.734. Exp.4 with LGCA boosts lesion detection (Rec.+0.106, Prec.+0.181 over Exp.2) but reduces GREEN (–0.130 vs Exp.5), indicating stronger boundary focus but missing organ sensitivity. Exp. 5, which combines both OGCA and LGCA, achieves the highest overall scores, confirming the complementary benefits of organ-guided and lesion-guided alignment.

Evaluations in OGCA As shown in Table 3, removing Top-k selection consistently degrades all metrics. For instance, B-4 drops from 0.312 to 0.264 and GREEN from 0.366 to 0.287, confirming that filtering irrelevant tokens effectively mitigates hallucination. When $\delta = 1$, metrics decline sharply (GREEN 0.312 $\rightarrow$ 0.241) due to poor spatial continuity, while $\delta = 2$ and $\delta = 4$ yield moderate results with relatively stable B-4 but lower GREEN and Precision, indicating correct context but increased entity-level errors. Tests with different dilation factors show $\delta = 3$ yields optimal results.

Evaluations in LGCA As shown in Table 4, removing Laplacian filtering slightly reduces contextual consistency, indicating weaker boundary sensitivity. Using Laplacian without residual connections improves some metrics (MTR +0.006, GREEN +0.005), but causes an uneven effect with Precision dropping to 0.452, reflecting unstable edge enhancement. The full residual formulation achieves the best results, confirming that combining Laplacian cues with residual integration is critical for stable lesion-guided alignment.

Conclusion

DEAR improves 3D RRG by organ and lesion streams with OGCA's topology-aware selection and LGCA's boundary-enhancing filtering, yielding superior performance.

Acknowledgments

This work was supported in part by the Major Program of the National Natural Science Foundation of China under Grant 62495064, in part by the Key Research and Development Program of the Department of Science and Technology of the Tibet Autonomous Region under Grant XZ202402ZY0003, in part by the Innovation Research Team Program of the Science and Technology Department of Sichuan Province Grant 2024NSFTD0051, in part by in part by the Industrial Chain Collaborative Innovation Project of Science and Technology under Grant 2025-XT00-00018-GX, in part by the Sichuan Provincial Medical Imaging Clinical Medicine Research Center Open Project under Grant YXYX2412, and in part by the Clinical Medical Research Promotion Program of China Medical Foundation under Grant 2024CMFA10.

References

Altalla', B.; Ahmad, A.; Bitar, L.; Al-Bssol, M.; Al Omari, A.; and Sultan, I. 2025. Radiology Report Annotation Using Generative Large Language Models: Comparative Analysis. *International Journal of Biomedical Imaging*, 2025(1): 5019035.

Bai, F.; Du, Y.; Huang, T.; Meng, M. Q.-H.; and Zhao, B. 2024. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Chen, H.; Zhao, W.; Li, Y.; Zhong, T.; Wang, Y.; Shang, Y.; Guo, L.; Han, J.; Liu, T.; Liu, J.; et al. 2024a. 3d-ct-gpt: Generating 3d radiology reports through integration of large vision-language models. *arXiv preprint arXiv:2409.19330*.

Chen, W.; Shen, L.; Lin, J.; Luo, J.; Li, X.; and Yuan, Y. 2024b. Fine-Grained Image-Text Alignment in Medical Imaging Enables Explainable Cyclic Image-Report Generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9494–9509. Bangkok, Thailand: Association for Computational Linguistics.

Chen, Y.; Liu, C.; Liu, X.; Arcucci, R.; and Xiong, Z. 2024c. Bimcv-r: A landmark dataset for 3d ct text-image retrieval. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 124–134. Springer.

Chen, Z.; Bie, Y.; Jin, H.; and Chen, H. 2025. Large language model with region-guided referring and grounding for ct report generation. *IEEE Transactions on Medical Imaging*.

Chen, Z.; Song, Y.; Chang, T.-H.; and Wan, X. 2020. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*.

Comelli, A.; Stefano, A.; Benfante, V.; and Russo, G. 2018. Normal and abnormal tissue classification in positron emission tomography oncological studies. *Pattern Recognition and Image Analysis*, 28(1): 106–113.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

Gu, D.; Gao, Y.; Zhou, Y.; Zhou, M.; and Metaxas, D. 2025. Radalign: Advancing radiology report generation with vision-language concept alignment. *arXiv preprint arXiv:2501.07525*.

Hamamci, I. E.; Er, S.; and Menze, B. 2024. Ct2rep: Automated radiology report generation for 3d medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 476–486. Springer.

Hamamci, I. E.; Er, S.; Wang, C.; Almas, F.; Simsek, A. G.; Esirgun, S. N.; Doga, I.; Durugol, O. F.; Dai, W.; Xu, M.; et al. 2024. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. *arXiv preprint arXiv:2403.17834*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hou, W.; Cheng, Y.; Xu, K.; Li, H.; Hu, Y.; Li, W.; and Liu, J. 2025. RADAR: Enhancing Radiology Report Generation with Supplementary Knowledge Injection. *arXiv preprint arXiv:2505.14318*.

Hou, Z.; Yan, R.; Yan, Z.; Lang, N.; and Zhou, X. 2024. Energy-based controllable radiology report generation with medical knowledge. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 240–250. Springer.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2): 1–55.

Huang, Z.; Jiang, Y.; Zhang, R.; Zhang, S.; and Zhang, X. 2024. CAT: Coordinating Anatomical-Textual Prompts for Multi-Organ and Tumor Segmentation. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 3588–3610. Curran Associates, Inc.

Jin, H.; Che, H.; Lin, Y.; and Chen, H. 2024. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2607–2615.

Kerioui, M.; Desmée, S.; Mercier, F.; Lin, A.; Wu, B.; Jin, J.; Shen, X.; Le Tourneau, C.; Bruno, R.; and Guedj, J.

2022. Assessing the impact of organ-specific lesion dynamics on survival in patients with recurrent urothelial carcinoma treated with atezolizumab or chemotherapy. *ESMO open*, 7(1): 100346.

Langlotz, C. P. 2006. RadLex: a new method for indexing online educational materials.

Li, M.; Lin, B.; Chen, Z.; Lin, H.; Liang, X.; and Chang, X. 2023. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3334–3343.

Li, Y.; Wang, Z.; Liu, Y.; Wang, L.; Liu, L.; and Zhou, L. 2024. KARGEN: Knowledge-Enhanced Automated Radiology report generation using large language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 382–392. Springer.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, C.; Tian, Y.; Chen, W.; Song, Y.; and Zhang, Y. 2024. Bootstrapping large language models for radiology report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18635–18643.

Liu, K.; Ma, Z.; Kang, X.; Li, Y.; Xie, K.; Jiao, Z.; and Miao, Q. 2025a. Enhanced contrastive learning with multi-view longitudinal data for chest x-ray report generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 10348–10359.

Liu, T.; Wang, J.; Hu, Y.; Li, M.; Yi, J.; Chang, X.; Gao, J.; and Yin, B. 2025b. HC-LLM: Historical-constrained large language models for radiology report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5595–5603.

Mei, X.; Yang, L.; Gao, D.; Cai, X.; Han, J.; and Liu, T. 2024. Phraseaug: An augmented medical report generation model with phrasebook. *IEEE Transactions on Medical Imaging*, 43(12): 4211–4223.

Ostmeier, S.; Xu, J.; Chen, Z.; Varma, M.; Blankemeier, L.; Bluethgen, C.; Michalson, A. E.; Moseley, M.; Langlotz, C.; Chaudhari, A. S.; et al. 2024. Green: Generative radiology report evaluation and error notation. *arXiv preprint arXiv:2405.03595*.

Pang, B.; Zheng, Z.; Li, Y.; Wang, G.; and Wang, P.-S. 2024. Neural Laplacian Operator for 3D Point Clouds. *ACM Transactions on Graphics (TOG)*, 43(6): 1–14.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Qin, A.; Chen, S.; Liang, J.; Snyder, M.; and Yan, D. 2022. Evaluation of DIR schemes on tumor/organ with progressive shrinkage: Accuracy of tumor/organ internal tissue tracking during the radiation treatment. *Radiotherapy and Oncology*, 173: 170–178.

Shui, Z.; Zhang, J.; Cao, W.; Wang, S.; Guo, R.; Lu, L.; Yang, L.; Ye, X.; Liang, T.; Zhang, Q.; et al. 2025. Large-scale and fine-grained vision-language pre-training for enhanced ct image understanding. *arXiv preprint arXiv:2501.14548*.

Tang, Y.; Yang, H.; Zhang, L.; and Yuan, Y. 2024. Work like a doctor: Unifying scan localizer and dynamic generator for automated computed tomography report generation. *Expert Systems with Applications*, 237: 121442.

Tanida, T.; Müller, P.; Kaissis, G.; and Rueckert, D. 2023. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7433–7442.

Wang, X.; Wang, F.; Wang, H.; Jiang, B.; Li, C.; Wang, Y.; Tian, Y.; and Tang, J. 2025a. Activating associative disease-aware vision token memory for llm-based x-ray report generation. *arXiv preprint arXiv:2501.03458*.

Wang, Z.; Sun, Y.; Li, Z.; Yang, X.; Chen, F.; and Liao, H. 2025b. Llm-rg4: Flexible and factual radiology report generation across diverse input contexts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 8, 8250–8258.

Wang, Z.; Tang, M.; Wang, L.; Li, X.; and Zhou, L. 2022. A medical semantic-assisted transformer for radiographic report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 655–664. Springer.

Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Towards Generalist Foundation Model for Radiology by Leveraging Web-scale 2D&3D Medical Data. arXiv:2308.02463.

Xiao, T.; Shi, L.; Liu, P.; Wang, Z.; and Bai, C. 2025. Radiology report generation via multi-objective preference optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 8664–8672.

Zhang, X.; Shi, Y.; Ji, J.; Zheng, C.; and Qu, L. 2025. MEP-Net: Medical Entity-Balanced Prompting Network for Brain CT Report Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 25940–25948.

Zhao, J.; Zhou, Y.; Chen, Z.; Fu, H.; and Wan, L. 2024. Topicwise separable sentence retrieval for medical report generation. *IEEE Transactions on Medical Imaging*.