# Graphical Abstract

**MedVista3D: Vision-Language Modeling for Reducing Diagnostic Errors in 3D CT Disease Detection, Understanding and Reporting**

Yuheng Li, Yenho Chen, Yuxiang Lai, Jike Zhong, Vanessa Wildman, Xiaofeng Yang

# Highlights

**MedVista3D: Vision-Language Modeling for Reducing Diagnostic Errors in 3D CT Disease Detection, Understanding and Reporting**

Yuheng Li, Yenho Chen, Yuxiang Lai, Jike Zhong, Vanessa Wildman, Xiaofeng Yang

- Research highlight 1
- Research highlight 2

# MedVista3D: Vision-Language Modeling for Reducing Diagnostic Errors in 3D CT Disease Detection, Understanding and Reporting

Yuheng Li[a], Yenho Chen[b], Yuxiang Lai[c], Jike Zhong[d], Vanessa Wildman[e], Xiaofeng Yang[a,e,f,*]

[a]*Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA, USA*
[b]*Department of Machine Learning, Georgia Institute of Technology, Atlanta, GA, USA*
[c]*Department of Radiation Oncology, Emory University School of Medicine, Atlanta, GA, USA*
[d]*Department of Computer Science, University of Southern California, Los Angeles, CA, USA*
[e]*Department of Radiation Oncology, Emory University School of Medicine, Atlanta, GA, USA*
[f]*Emory University School of Medicine, Atlanta, GA, USA*

**Abstract**

Radiologic diagnostic errors, such as under-reading errors, inattentional blindness, and communication failures, remain prevalent in clinical practice. These issues often stem from missed localized abnormalities, limited global context, and variability in report language. These challenges are amplified in 3D imaging, where clinicians must examine hundreds of slices per scan. Addressing them requires systems with precise localized detection, global volume-level reasoning, and semantically consistent natural language reporting. However, existing 3D vision-language models are unable to meet all three needs jointly—lacking local-global understanding for spatial reasoning and struggling with the variability and noise of uncurated radiology reports. We present MedVista3D, a unified semantic-enriched vision-language pretraining framework for 3D CT analysis. To enable joint disease detection and holistic interpretation, MedVista3D performs local and global image-text alignment for fine-grained representation learning within full-volume context. To address

*Corresponding author: xiaofeng.yang@emory.edu

report variability, we apply language model rewrites and introduce a Radiology Semantic Matching Bank for semantics-aware alignment. MedVista3D achieves state-of-the-art performance on zero-shot disease classification, report retrieval, and medical visual question answering, while transferring well to organ segmentation and prognosis prediction. Code and datasets will be released.

*Keywords:* Computed tomography, Vision language model, Foundation model, Report generation

---

## 1. Introduction

Despite decades of clinical experience, radiologic diagnostic errors remain common and pose a persistent source of patient harm Bruno et al. (2015). In a large-scale study Kim and Mansfield (2014), three categories of errors were found to be prevalent. **Under-reading errors** occur when abnormalities are simply missed even within the field of view, often due to insufficient attention to localized findings. **Inattentional blindness** arise due to tunnel vision or limited global context, missing lesions outside the area of focus or in underexamined slices. **Communication failures** occur when correctly identified findings are ineffectively conveyed, often due to ambiguous phrasing or inconsistent terminology in the radiology report Waite et al. (2018). Addressing these errors requires systems capable of precise local detection, comprehensive image understanding, and clear, consistent communication of findings.

The development of such systems is particularly crucial for 3D medical images, where physicians must examine hundreds of cross-sectional slices which remains both time-consuming and expertise-driven Shen et al. (2017). Fundamentally, radiologic image interpretation spans three related tasks: (1) **localized detection** of anomalies like tumors or opacities; (2) **global understanding** of disease patterns across whole volume, which informs tasks like disease classification and report retrieval; and (3) **reporting**, which involves accurately describing findings and answering clinical questions in natural language. Recent advances in medical vision-language models (VLMs) have shown promise in automating these components—enabling localized disease identification, global image-report retrieval, zero-shot classification, and report generation or visual question answering Shui et al. (2025); Li et al. (2024b); Thawkar et al. (2023); Zhang et al. (2023).

However, current medical VLMs cannot concurrently address these three diagnostic challenges, due to limitations in their training objectives and supervision data. **First**, existing models lack the capability to jointly perform local detection and global understanding, demonstrating under-reading and inattentional blindness for disease diagnosis. We analyze two state-of-the-art 3D CT VLMs (CT-CLIP Hamamci et al. (2024) and fVLM Shui et al. (2025)) under local and global settings for disease query. CT-CLIP is trained with a global-only objective, aligning entire volumes with full reports. As a result, it struggles to identify small and localized abnormalities due to insufficient local alignment. As shown in Figure 1 (top row, second column), its gradient activations focus on irrelevant regions when queried about gallbladder carcinoma, paralleling the under-reading error. Conversely, fVLM aligns organ-level features with their corresponding text descriptions, but lacks a mechanism for global understanding. As shown in Figure 1 (bottom row, third column), its activations neglect relevant organs under a global query, analogous to inattentional blindness where context beyond the region of focus is neglected. **Second**, the variability in real-world radiology reports could hinder learning consistent disease representations for effective reporting. Figure 2 (right) shows text examples from a large-scale public dataset (e.g. CT-RATE). Empirically, we find that unstructured reports often contain inconsistent interpretations, repetitive phrasing, and vague expressions that fail to clearly convey clinically significant findings, such as lymphadenopathy. These issues degrade the quality of learned representations and introduce ambiguity in downstream tasks such as report generation and visual question answering (VQA). Addressing them requires medical VLMs to combine unified visual grounding with semantically enriched alignment signals.

We propose **MedVista3D**, a 3D VLM to enhance the detection, understanding, and reporting of 3D CT image analysis. Our approach learns local-global representations while enhancing the disease-semantics understanding of the model. **First**, we derive a unified loss that simultaneously aligns CT volumes and organ-level features with their corresponding text descriptions. This maximizes the mutual information shared between CT images and corresponding text descriptions, enabling both local detection and global understanding of the model. We theoretically demonstrate that this unified loss captures more mutual information between global and local images and texts than single-scale alignment loss. We propose a novel dual-pathway vision encoder to jointly process global 3D CT volumes and local segmented organs. **Second**, we improve semantic supervision through unified semantic

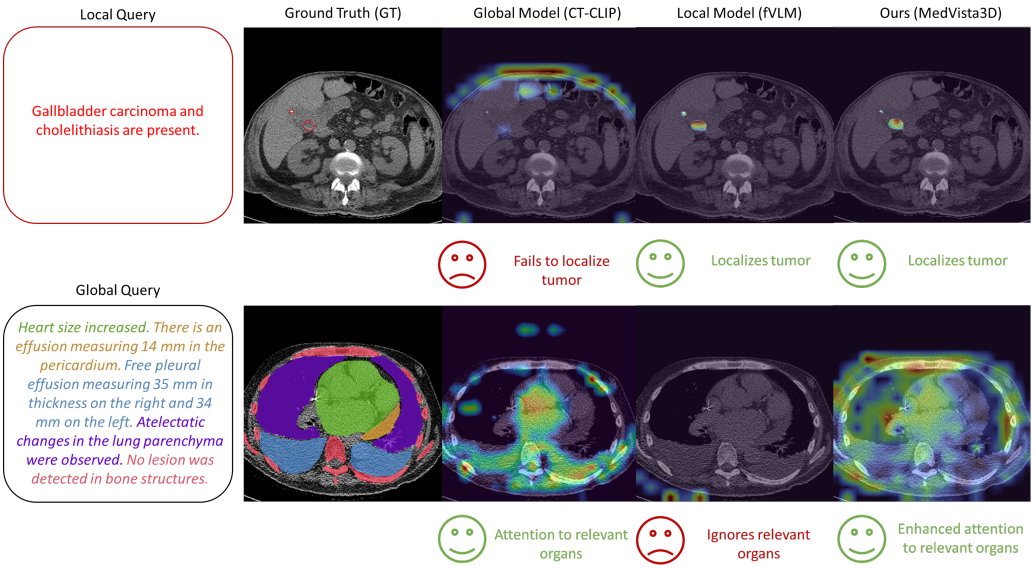


Figure 1: We visualize gradient activation maps for both global and local queries on global model (CT-CLIP) and local model (fVLM). Each row shows model attention for either a local (top) or global (bottom) query, with ground truth (GT) segmentations color-coded by sentence. The global model fails to detect tumor given a local query. The local model does not capture relevant anatomical regions given a global query. Our model effectively attends to relevant regions in both cases, demonstrating superior unified understanding.

alignment. We enhance radiology reports via Large Language Model (LLM) rewrites to emphasize the presence or absence of each disease to ensure consistency. We then propose the Radiology Semantic Matching Bank (RSMB) for additional semantic alignment at global and local scales. RSMB retrieves semantically matched disease descriptions via nearest-neighbor search, providing robust text supervision. As shown in Figure 2 (left), our MedVista3D considerably outperforms existing medical vision-language models on global disease zero-shot classification.

We summarize the following contributions:

- We identify the limitations of single-scale training objectives in existing 3D medical VLMs and derive a unified alignment loss. We theoretically demonstrate that our loss can capture more shared cross-modal information than single-scale losses. With a novel dual-pathway transformer, we jointly encode global-local representations using our unified objective.

- To address the variations in unstructured reports, we introduce unified

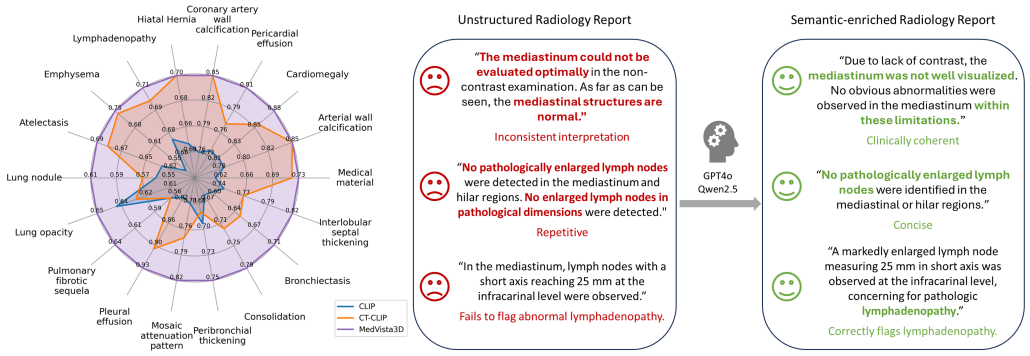


Figure 2: **Left:** *Global zero-shot performance of MedVista3D-ViT on CT-RATE.* AUC scores are reported per disease, reflecting the model's generalization across diverse pathologies. **Right:** *LLM-based refinement of radiology reports.* To address ambiguity and inconsistency in uncurated CT-RATE reports, we apply large language models (e.g., GPT-4o, Qwen2.5) to rewrite them with improved clarity and clinical coherence.

semantic supervision using LLM-rewritten reports and the Radiology Semantic Matching Bank, which retrieves semantically similar disease texts to enhance contrastive training across scales.

- We validate MedVista3D through comprehensive experiments across diverse medical tasks (e.g. *disease zero-shot detection*, *report retrieval*, *medical VQA*, *organ segmentation*, *disease classification*), achieving state-of-the-art performance through unified alignment.

## 2. Related work

### *2.1. Vision language models for medical imaging*

Previous medical VLMs predominantly employ global alignment, contrasting entire images and reports Chauhan et al. (2020); Zhang et al. (2024b); Lin et al. (2023); Stevens et al. (2024); Blankemeier et al. (2024); Hamamci et al. (2024). More recent works introduced local region-text alignment Lin et al. (2024); Shui et al. (2025) or local token-wise Wang et al. (2022); Huang et al. (2021) alignment to learn fine-grained visual features. However, our investigation reveals a large gap between representations learned from global model and local model (Figure 1). We motivate our approach by building a unified pretraining method to combine the strengths of each alignment.

### *2.2. Multi-scale alignment for VLM*

There remains limited research in multi-scale alignment for VLMs. Existing methods Huang et al. (2024); Du et al. (2024) focus on multiscale radiography-report alignment but lack the use of region masks or bounding boxes for fine-grained detection. Other approaches Chen et al. (2024); Zhang et al. (2022) combine image-text and region-text alignments using coarse bounding boxes, which are less effective for precise organ localization. In contrast, our work utilizes segmentation masks to extract fine-grained organ features, enabling more accurate local alignment.

### *2.3. Improving medical VLMs using synthetic data.*

Given the scarcity of annotated data and privacy concerns in medical imaging, synthetic data has been widely explored to augment images Koetzier et al. (2024); Özbey et al. (2023); Chlap et al. (2021). A few studies explored generating synthetic image or text data to support VLM pretraining Wu et al. (2023); Liu et al. (2024a); Bluethgen et al. (2024). MedKlip Wu et al. (2023) extracts named entities from reports and supervises using these disease-specific queries. However, this approach overlooks the context and completeness of a query sentence. The closest to our work are local VLMs Shui et al. (2025); Lin et al. (2024) that learn fine-grained representation using LLMs to decompose long reports into specific regions. However, these works do not learn unified representations for local-global disease understanding, nor do they address the text variations in radiology reports. We tackle this by enhancing disease semantics in reports via LLMs and performing semantic alignment using nearest-neighbor search in the text embedding space.

## 3. Method

**Overview** Compared to previous VLMs, our model performs alignment at four scales (Figure 3): (1) global volume with report, (2) local region with text, (3) global volume with a semantically enriched report, and (4) local region with a semantically enriched sentence. We also propose a simple and effective dual-pathway transformer to encode global-local information. The unified alignment strategy is detailed in Section 3.1, while semantic alignment using LLM-based rewrites and the Radiology Semantic Matching Bank is presented in Section 3.2.

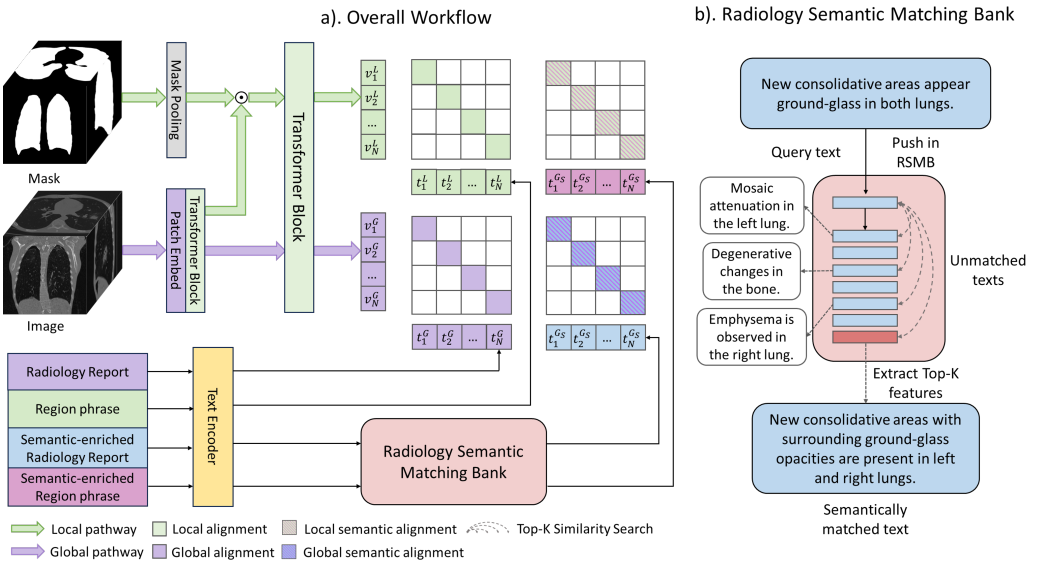


Figure 3: a). MedVista3D encodes 3D CT volumes at both global and local scales. For local alignment, visual organ embeddings are paired with organ and semantic-enriched phrases. For global alignment, global volume embedding is matched with the report embedding and its semantic-enriched versions augmented by LLMs. b). A radiology semantic matching bank maintains a queue of text embeddings from diverse radiology descriptions. For each query, a top-k similarity search retrieves semantically matching texts, filtering out less relevant ones.

### *3.1. Unifying Global and Local Alignment*

We propose unified alignment to leverage the complementary strengths of single-scale approaches. We first show that from a mutual information perspective, our proposed objective captures more shared cross-modal information than single-scale losses. Then, we propose a novel architecture dual-pathway transformer to jointly encode unified representations from 3D CT volumes.

**Connection to mutual information maximization (MI).** Contrastive VLM aims to learn positive-negative image-text embeddings by jointly training an image encoder $f_{\mathrm{I}}(\cdot)$ and a text encoder $f_{\mathrm{T}}(\cdot)$. MI quantifies the shared information between image and text by measuring how much knowing one variable reduces the uncertainty about the other. First, we define a common objective *global image-text alignment*, such as CLIP Radford et al. (2021). Given a dataset of $P$ pairs of CT image volumes and their corresponding radiology reports, $X = \{x_1, \ldots, x_P\}$ and $Y = \{y_1, \ldots, y_P\}$, the global em-

beddings for the $i$th volume-report pair can be obtained as $v_i^G = f_{\mathrm{I}}(x_i)$ and $t_i^G = f_{\mathrm{T}}(y_i)$, where $x_i \in \mathbb{R}^{1\times D\times H\times W}$ and $y_i \in \mathbb{R}^l$ represent the dimensions of the input CT volume and radiology report, respectively. To align image and text representations, a contrastive objective pushes the embeddings of matched volume-report pairs together while pushing those of unmatched pairs apart. Using InfoNCE loss Oord et al. (2018), the global alignment objective becomes,

$$\mathcal{L}_{\text{Global}} = \frac{1}{2}\left[\mathcal{L}_{I\to T}^G + \mathcal{L}_{T\to I}^G\right], \tag{1}$$

The first term consists of the global image-to-text loss, $\mathcal{L}_{I\to T}^G$, and is defined as,

$$\mathcal{L}_{I\to T}^G = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\text{sim}(v_i^G, t_i^G)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}(v_i^G, t_j^G)/\tau\right)}, \tag{2}$$

where $N$ is the batch size and $\text{sim}(\cdot,\cdot)$ is the similarity function and $\tau$ is a learnable logit. We omit $L_{T\to I}^G$ since it is symmetric.

However, global approach can overlook fine-grained similarities or differences among various organs. Alternatively, *local image-text alignment* identifies all possible regions in the CT image and extracts region-specific features Shui et al. (2025); Lin et al. (2024). Assuming the CT image can be divided into image regions $x_i^1, \ldots, x_i^r$, and radiology reports can also be decomposed into fine-grained captions $y_i^1, \ldots, y_i^r$ describing each organ, region-text pairs can be formed as $\{(x_i^1, y_i^1), \ldots, (x_i^r, y_i^r)\}$. For region $r$, $f_I$ extracts local image embedding $v_i^r$ and $f_T$ extracts local text embeddings $t_i^L$. The local alignment loss can be defined as:

$$\mathcal{L}_{\text{Local}} = \frac{1}{2}\left[\mathcal{L}_{I\to T}^L + \mathcal{L}_{T\to I}^L\right], \tag{3}$$

The local image-to-text loss can be written as:

$$\mathcal{L}_{I\to T}^L = -\frac{1}{RN}\sum_{r=1}^{R}\sum_{i=1}^{N}\log\frac{\exp\left(\text{sim}\big(v_i^r, t_i^r\big)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}\big(v_i^r, t_j^r\big)/\tau\right)}, \tag{4}$$

where $R$ is the total number of regions.

We make the connection between global/local objectives and MI using findings from Poole et al. (2019): contrastive InfoNCE loss estimates a lower

bound for MI. Extending this theorem to both global and local alignment yields:

$$I(X_G; Y_G) \geq -\mathcal{L}_{\text{Global}} + \log(N_G), \tag{5}$$

$$I(X_L; Y_L) \geq -\mathcal{L}_{\text{Local}} + \log(N_L), \tag{6}$$

where $N_G$ and $N_L$ are the number of negative samples in global and local alignment, respectively.

**A unified objective.** We observe that equations 5 and 6 can only capture partial structure as they focus exclusively on either local or global views separately. Our core insight is that capturing both the holistic contexts and fine-grained details requires maximizing a unified mutual information between the full set of global and local CT images $X = (X_L, X_G)$ and text reports $Y = (Y_L, Y_G)$, defined as,

$$I_{\text{Unified}}(X, Y) = I(X_G, X_L; Y_L, Y_G). \tag{7}$$

By the chain rule for mutual information, we have

$$I_{\text{Unified}}(X, Y) \geq \max\left\{I(X_L; Y_L),\ I(X_G; Y_G)\right\}, \tag{8}$$

indicating that the unified objective can capture more shared information between the modalities than considering either global or local inputs alone. This makes it better suited for learning representations that encode both global semantics and local alignment. However, directly optimizing $I_{\text{Unified}}(X, Y)$ is computationally intractable. Instead, we propose a unified contrastive loss that linearly integrates global and local alignment:

$$\mathcal{L}_{\text{unified}} = \frac{1}{2}\left[\mathcal{L}_{\text{Global}} + \mathcal{L}_{\text{Local}}\right]. \tag{9}$$

This objective is part of a valid lower bound,

$$I_{\text{Unified}}(X, Y) \geq -\mathcal{L}_{\text{unified}} + \frac{1}{2}\left[\log(N_L) + \log(N_G)\right], \tag{10}$$

and explicitly encourages the learned representation to jointly capture information from both global and local views of the input data.

**Dual-pathway transformer.** To jointly encode global and local information, we propose a novel dual-pathway transformer encoder for MedVista3D (Figure 3a). For global pathway, given CT volume $x_i$, the model first extracts patch embeddings $p_i \in \mathbb{R}^{c \times d \times h \times w}$ using a 3D convolutional layer. Transformer

blocks then generate a latent image embedding $v_i^L$, which the final transformer block refines into the global image embedding $v_i^G$. The radiology report is encoded by text encoder into a global text embedding $t_i^G$. For local pathway, given anatomical region $r$ and its segmentation map $M_i^r \in \{0,1\}^{D\times H\times W}$, mask pooling downsamples it to $\tilde{M}_i^r \in \{0,1\}^{d\times h\times w}$, matching the patch grid resolution. Active region tokens (threshold 0.5 in $\tilde{M}_i^r$) are selected from $p_i$ element-wise. These are processed by the last transformer block to produce the local image embedding $v_i^r$. The corresponding region-specific text phrase is encoded into a local text embedding $t_i^r$. This preserves the spatial relationships between global and local embeddings.

### *3.2. Radiology Semantic Enrichment and Alignment*

**Semantic enrichment of radiology reports.** While unified alignment yields richer representations, communication errors could still be caused by directly training on unstructured reports. Free-form radiology reports often suffer from length and inconsistent terminologies (e.g., *nodular opacities* vs. *lesions*). To address this, we prompt the LLMs to identify all possible abnormalities from the report and rewrite the findings as discrete, presence-or-absence statements. This process, applied to both global reports and local region phrases, yields standardized, succinct text descriptions where each sentence details at least one abnormality (prompts in Appendix D).

**Radiology semantics matching bank.** Building on these enriched texts, the RSMB provides robust supervision by retrieving semantically similar embeddings, addressing minor wording variations. We observe that enriched texts often describe the same findings with only minor variations in wording (e.g., "mild pleural thickening" vs. "slight pleural thickening"). RSMB is a 64k-sized first-in-first-out queue storing previously encoded enriched global and local text features. For a new enriched query text, its top-1 nearest neighbor (via cosine similarity) is retrieved from RSMB. The corresponding image embedding is then aligned with this retrieved text, ensuring robustness to text variations while maintaining consistent disease semantics.

**Unified semantic alignment.** Using RSMB and the enriched texts, we establish semantic alignment at two levels. For global-level, a semantically-enriched text embedding $\hat{t}_i^G$ queries RSMB to retrieve its nearest neighbor $\hat{t}_i^{G_{NN}}$. This embedding is aligned with global image embedding $v_i^G$ using contrastive loss:

$$\mathcal{L}_{\text{Global Semantic}} = \mathcal{L}_{I\to T}^{G_{NN}} + \mathcal{L}_{T\to I}^{G_{NN}}, \tag{11}$$

$$\mathcal{L}_{I\to T}^{G_{NN}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\text{sim}(v_i^G, \hat{t}_i^{G_{NN}})/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}(v_i^G, \hat{t}_j^{G_{NN}})/\tau\right)}. \tag{12}$$

Similarly for local-level, a semantically-enriched region embedding $\hat{t}_i^r$ queries RSMB for its neighbor $\hat{t}_i^{r_{NN}}$, which is aligned with the region image embedding $v_i^r$:

$$\mathcal{L}_{\text{Local Semantic}} = \mathcal{L}_{I\to T}^{L_{NN}} + \mathcal{L}_{T\to I}^{L_{NN}}, \tag{13}$$

$$\mathcal{L}_{I\to T}^{L_{NN}} = -\frac{1}{RN}\sum_{r=1}^{R}\sum_{i=1}^{N}\log\frac{\exp\left(\text{sim}\big(v_i^r, \hat{t}_i^{r_{NN}}\big)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}\big(v_i^r, \hat{t}_j^{r_{NN}}\big)/\tau\right)} \tag{14}$$

We propose the unified semantic objective as,

$$\mathcal{L}_{\text{unified Semantic}} = \mathcal{L}_{\text{Global Semantic}} + \mathcal{L}_{\text{Local Semantic}}. \tag{15}$$

Combining both unified and semantic alignment, our final pretraining objective is defined as,

$$\mathcal{L}_{\text{MedVista3D}} = \mathcal{L}_{\text{unified}} + \mathcal{L}_{\text{unified Semantic}}. \tag{16}$$

## 4. Experiments

### *4.1. Implementation details*

**Pretraining on CT-RATE:** We pretrain MedVista3D on the CT-RATE dataset Hamamci et al. (2024) using the training split (24,128 volumes) and perform testing on the internal test split (1,564 volumes). Local alignment uses Radgenome masks and region texts Zhang et al. (2024a). For dataset preprocessing, we use volumes resampled to 3.0 mm × 1.0 mm × 1.0 mm from Radgenome dataset Zhang et al. (2024a). We also use its segmentation masks and region sentences for regional image-text alignment. For intensity normalization, we follow CT-CLIP Hamamci et al. (2024) preprocessing. We uniformly resize all volumes to 96 × 320 × 320 using padding or center cropping. For text encoder, we use BiomedVLP-CXR-BERT-specialized Boecking et al. (2022). For vision encoder, we use 1). ViT-B, with embedding dimension as 768 and depth as 12; and 2). UniMISS-Small Xie et al. (2022).

We train MedVista3D-ViT using a batch size of 32 and MedVista3D-UniMISS using a batch size of 20 for a total of 16 epochs on our proposed loss. The optimizer is AdamW with the weight decay of 1e-5. We use a linear warmup with cosine decay scheduler for 200 steps and a learning rate of 5e-5. All experiments were conducted using NVIDIA A100 GPUs (80GB) on an internal cluster.

**Visual question answering using MedVista3D-LLaVA:** We use Llama-3.1-7B Dubey et al. (2024) as the language decoder and the pretrained MedVista3D-ViT as the vision encoder. For multi-modal projector, we use a two-layer MLP-GELU following LLaVA-1.5 Liu et al. (2024b). We follow the two-stage training strategy same as LLaVA: 1). First, we perform contrastive alignment using CT-RATE's volume-report pairs to tune the multi-modal projector; 2). Second, we perform supervised finetuning using LoRA Hu et al. (2021) with rank $r$ set to 128, scaling factor $\alpha$ set to 256, and a learning rate of 2e-5. We train a total of 10 epochs following CT-CHAT.

**Segmentation on TotalSegmentator:** We use pretrained MedVista3D-UniMISS as the segmentation encoder and attach STU-Net-B's decoder Huang et al. (2023). We use nn-UNet Isensee et al. (2021) to preprocess the TotalSegmentator dataset and train within their framework. We use a learning rate of 5e-5 and a batch size of 2. Input volumes are uniformly cropped to 128 $\times$ 128 $\times$ 128. We train for a total of 1000 epochs following the default setting.

**Classification on STOIC 2021:** MedVista3D-UniMISS is initialized with pretrained CT-RATE weights. We resample CT volumes to 3.0 mm $\times$ 1.0 mm $\times$ 1.0 mm and crop/pad to 96 $\times$ 320 $\times$ 320. From the full 2000 volumes, we randomly select 80% for training, 10% for validation and 10% for testing. We use a batch size of 96, a learning rate of 1e-4 and finetune for 10 epochs.

### *4.2. Reducing under-reading and inattentional blindness via MedVista3D*

We evaluate how MedVista3D reduces under-reading and inattentional blindness by assessing both global understanding and local disease detection from CT volumes. We compare with global VLMs—trained on the entire CT volume and corresponding text (e.g., CLIP Radford et al. (2021), CT-CLIP Hamamci et al. (2024), Merlin Blankemeier et al. (2024))—and local VLMs aligning region features with region text (e.g., fVLM Shui et al. (2025)).

**Local Task**: Assess under-reading errors by evaluating localized disease detection within anatomical regions (lungs, heart, aorta, and esophagus).

Table 1: Performance comparison of VLMs at global tasks and local task on CT-RATE. **Blue** for global models and **green** for local model. **BOLD** means best result and underline second best. † : our implementation. ‡ : using official checkpoint.

| Method | Global | | | | | | Local | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Disease zero-shot | | | | Report retrieval | | Disease zero-shot | | | |
| | Precision | F1 | ACC | AUC | Recall 5 | Recall 10 | Precision | F1 | ACC | AUC |
| CLIP Radford et al. (2021) | 0.334 | 0.726 | 0.691 | 0.703 | 2.67% | 5.00% | 0.306 | 0.696 | 0.657 | 0.659 |
| Merlin‡ Blankemeier et al. (2024) | 0.229 | 0.612 | 0.558 | 0.578 | 1.11% | 2.02% | 0.199 | 0.479 | 0.433 | 0.538 |
| CT-CLIP Hamamci et al. (2024) | 0.306 | 0.691 | 0.651 | 0.704 | 2.34% | 3.95% | 0.297 | 0.678 | 0.636 | 0.645 |
| fVLM† Shui et al. (2025) | 0.293 | 0.684 | 0.641 | 0.644 | 1.82% | 3.06% | 0.372 | 0.752 | 0.722 | 0.759 |
| fVLM‡ Shui et al. (2025) | 0.248 | 0.684 | 0.600 | 0.591 | 0.32% | 1.09% | **0.379** | 0.751 | 0.718 | 0.778 |
| MedVista3D-ViT (ours) | 0.379 | 0.760 | 0.737 | 0.778 | **6.64%** | **10.68%** | 0.377 | **0.765** | **0.742** | **0.780** |
| MedVista3D-UniMISS (ours) | **0.385** | **0.770** | **0.745** | **0.782** | 5.01% | 8.65% | 0.372 | 0.754 | 0.726 | 0.753 |

Table 2: Generalization of VLMs at global tasks and local task on Rad-ChestCT. **Blue** for global models and **green** for local model. **BOLD** means best result and underline second best. † : our implementation. ‡ : using official checkpoint.

| Method | Global | | | | Local | | | |
|---|---|---|---|---|---|---|---|---|
| | Disease zero-shot | | | | Disease zero-shot | | | |
| | Precision | F1 | ACC | AUC | Precision | F1 | ACC | AUC |
| CLIP Radford et al. (2021) | 0.352 | 0.637 | 0.617 | 0.609 | 0.321 | 0.593 | 0.569 | 0.559 |
| Merlin‡ Blankemeier et al. (2024) | 0.339 | 0.605 | 0.581 | 0.596 | 0.210 | 0.562 | 0.513 | 0.552 |
| CT-CLIP Hamamci et al. (2024) | 0.339 | 0.648 | 0.599 | 0.632 | 0.334 | 0.608 | 0.584 | 0.689 |
| fVLM† Shui et al. (2025) | 0.314 | 0.587 | 0.562 | 0.518 | 0.315 | 0.596 | 0.571 | 0.524 |
| fVLM‡ Shui et al. (2025) | 0.332 | 0.561 | 0.535 | 0.544 | 0.374 | **0.688** | 0.647 | 0.680 |
| MedVista3D-ViT (ours) | **0.426** | **0.693** | **0.684** | 0.702 | **0.402** | 0.681 | **0.668** | **0.710** |
| MedVista3D-UniMISS (ours) | 0.393 | 0.664 | 0.646 | **0.713** | 0.378 | 0.650 | 0.628 | 0.697 |

1. *Disease zero-shot classification:* Given text prompts and segmentation masks, identify the presence of diseases. For global models, we crop the CT volume to the segmentation mask and perform padding. Metrics include AUC, ACC, precision, and weighted F1-score.

**Global Tasks**: Address inattentional blindness by evaluating the model's ability to detect findings outside expected regions and retrieve reports correctly describing relevant findings.

1. *Disease zero-shot classification:* Given text prompts, identify the presence of diseases in the CT volume without segmentation masks. We report the same metrics as in the local task.
2. *Report retrieval:* Given a CT volume, retrieve the corresponding radiology report from the entire dataset. We measure recall at top-5 and top-10.

Table 3: Comparison of various LLaVA architectures on medical VQA (long answer, short answer, report generation, and multiple choice) on CT-RATE. **BOLD** means best result and underline second best. **Blue** for 2D MLLMs and **green** for 3D MLLMs.

| Method | Long Answer | | | | Short Answer | | | |
|---|---|---|---|---|---|---|---|---|
| | BLEU_1 | METEOR | ROUGE_L | CIDER | BLEU_1 | METEOR | ROUGE_L | CIDER |
| CXR-LLaVA | 0.203 | 0.140 | 0.231 | 0.577 | 0.016 | 0.000 | 0.021 | 0.040 |
| LLaVA-Med | 0.137 | 0.156 | 0.202 | 0.315 | 0.014 | 0.051 | 0.025 | 0.007 |
| CT-CHAT | 0.480 | 0.294 | 0.512 | 3.100 | 0.280 | 0.160 | 0.598 | **1.821** |
| MedVista3D-LLaVA (ours) | **0.516** | **0.309** | **0.546** | **3.395** | **0.299** | **0.178** | **0.602** | 1.817 |

| Method | Report Generation | | | | Multiple Choice | | | |
|---|---|---|---|---|---|---|---|---|
| | BLEU_1 | METEOR | ROUGE_L | CIDER | BLEU_1 | METEOR | ROUGE_L | CIDER |
| CXR-LLaVA | 0.050 | 0.000 | 0.020 | 0.049 | 0.057 | 0.009 | 0.063 | 0.065 |
| LLaVA-Med | 0.002 | 0.024 | 0.056 | 0.000 | 0.085 | 0.175 | 0.135 | 0.151 |
| CT-CHAT | 0.381 | 0.217 | 0.334 | 0.221 | 0.838 | 0.578 | 0.895 | 7.850 |
| MedVista3D-LLaVA (ours) | **0.474** | **0.252** | **0.386** | **0.349** | **0.936** | **0.668** | **0.927** | **8.210** |

**Results for local task.** On localized zero-shot detection, both MedVista3D backbones match or surpass fVLM. Importantly, fVLM suffers from poor generalization to global tasks (AUC drops from 0.759 to 0.644), whereas MedVista3D maintains superior performance across both tasks. This demonstrates our model's ability to reduce under-reading errors.

**Results for global tasks.** Both MedVista3D-ViT and MedVista3D-UniMISS outperform all global models in disease zero-shot and report retrieval (Table 1). For global disease zero-shot, MedVista3D-UniMISS achieves the highest AUC (0.782) and F1 (0.770), outperforming CT-CLIP by 7.4 points in AUC and 6.9 points in F1. For report retrieval, MedVista3D-ViT surpasses CT-CLIP by 4.3% and 6.7% in top-5 and top-10 recall. These results validate our model's ability to jointly reduce inattentional blindness and under-reading errors.

**External validation.** To assess generalization, we perform external validation on the full Rad-ChestCT dataset Draelos et al. (2021) (3626 volumes), following CT-CLIP and fVLM. We evaluate on global and local zero-shot disease detection. Segmentation masks are obtained using TotalSegmentator model Wasserthal et al. (2023). As shown in Table 2, MedVista3D-ViT consistently outperforms existing global models (CLIP, Merlin, CT-CLIP) across all global metrics, achieving an AUC of 0.702. MedVista3D-UniMISS achieves the highest global AUC of 0.713. For local tasks, MedVista3D-ViT also surpasses the fVLM with an AUC of 0.710, demonstrating robust generalization capabilities.

**Qualitative results.** Figure 4 illustrates how region masking modulates the attention map of [CLS]-to-patch tokens in CLIP, fVLM, and our model. CLIP shows broad attention without masking, but it fails to localize the correct region with mask, reflecting under-reading and explaining its poor local detection performance. fVLM attends correctly with a mask but fixates on irrelevant, tiny background areas without it, indicating inattentional blindness and poor global understanding. In contrast, MedVista3D demonstrates both fine-grained attention with mask and global attention on the anatomy without mask, effectively mitigating both error types.

### *4.3. Mitigating communication errors with MedVista3D-LLaVA*

To evaluate how our model mitigates communication errors in CT reporting, we train MedVista3D-LLaVA, a multimodal large language model (MLLM), on the CT-RATE VQA dataset. The dataset includes long-answer questions, short-answer questions, multiple-choice questions, and report generation tasks. Following our pretraining setup, we train on the CT-RATE training split and validate on its internal validation split. Evaluation follows CT-CHAT Hamamci et al. (2024), using BLEU, METEOR, ROUGE_L, and CIDER scores. As shown in Table 3, our method consistently outperforms CT-CHAT as well as 2D multimodal assistants (LLaVA-Med Li et al. (2024a), CXR-LLaVA Lee et al. (2025)) by considerable margins. It achieves the best performance on multiple-choice questions (BLEU_1: 0.936, METEOR: 0.668, ROUGE_L: 0.927, CIDER: 8.21), and surpasses CT-CHAT by 3.6% and 1.9% BLEU_1 on long and short answer tasks, respectively. On the accuracy of multiple choice, our method achieves 91.5%. For report generation, our model shows a 9.3-point BLEU_1 improvement. These gains demonstrate the effectiveness of our unified alignment and semantic enrichment of radiology reports, which mitigate potential communication errors in diagnostic workflows.

Table 4: Ablation study on unified and semantic image-text alignment.

| Pretraining Strategy | Region phrase grounding | | Report retrieval | | Global disease zero-shot | | | |
|---|---|---|---|---|---|---|---|---|
| | Top 10 | Top 50 | Top 5 | Top 10 | Precision | F1 | ACC | AUC |
| Global Alignment | 0.04% | 0.36% | 4.53% | 7.88% | 0.293 | 0.689 | 0.633 | 0.675 |
| + Local Alignment | 0.19% | 0.76% | 4.98% | 8.32% | 0.281 | 0.676 | 0.634 | 0.664 |
| + Mask Pooling | 0.48% | 2.42% | 4.53% | 7.88% | 0.279 | 0.674 | 0.631 | 0.609 |
| + Global Semantic Alignment | 0.38% | 1.99% | 4.98% | 8.25% | 0.398 | 0.789 | 0.758 | 0.807 |
| + Local Semantic Alignment | 0.83% | 3.46% | 6.64% | 10.68% | 0.379 | 0.760 | 0.737 | 0.778 |

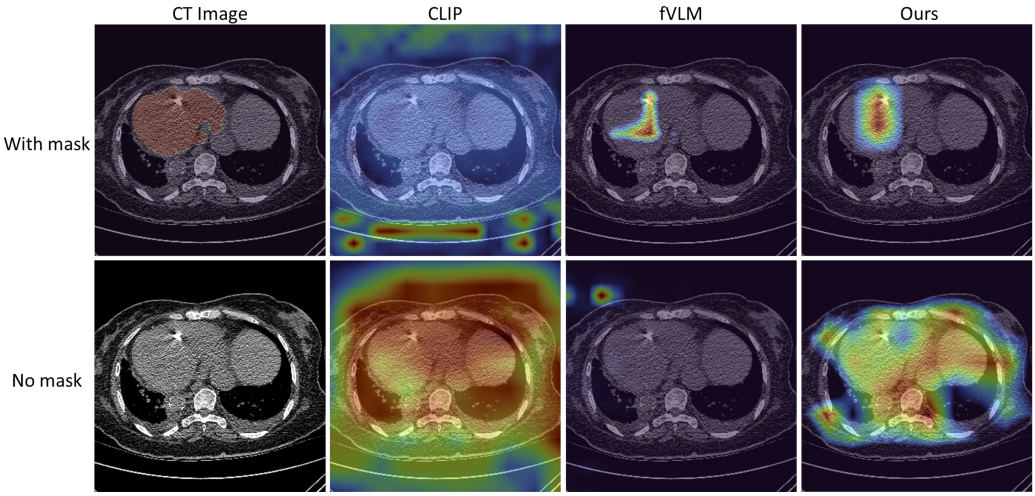


Figure 4: Impact of region masking on attention for CLIP, fVLM and MedVista3D (on CT-RATE). We visualize the attention maps of [CLS] token with other patch tokens given CT volume with (top) and without (bottom) region mask. MedVista3D remains focused on important organs regardless of masking. With mask CLIP shows diffuse attention; fVLM struggles without the mask.

### *4.4. Ablation Study*

**Ablation on unified and semantic alignment.** We conduct a detailed analysis on our proposed objective loss on both global and local tasks using CT-RATE. For local task, we evaluate region phrase grounding, where the model retrieves the correct region description given an image and a segmentation mask. As summarized in Table 4, we begin with a single global-level loss, with moderate global disease zero-shot and report retrieval performance, but lacking local capabilities. We further add local alignment which enables region grounding by explicitly learning organ-level embeddings. Next, we add mask pooling to allow more focused vision features on the segmentation mask, which further improve the region grounding but slightly compromising global zero-shot performance. Adding our proposed semantic alignment at global level considerably boosts global disease zero-shot and maintaining decent region retrieval. Finally, adding local semantic alignment yields the best overall balance between global and local tasks.

**Ablation on mask pooling.** We conduct a study on the mask pooling mechanism of our method on local disease zero-shot. Specifically, we choose various layers of vision transformer blocks to perform organ mask pooling. As shown in Table 5, we find that applying mask pooling before the last

transformer block yields the best performance. Applying mask pooling either before the first block or before the second block yields suboptimal performance.

Table 5: Ablation study on transformer block for mask pooling.

| | Local disease zero-shot | | | |
|---|---|---|---|---|
| Layer | Precision | F1 | ACC | AUC |
| 1st block | 0.336 | 0.709 | 0.656 | 0.703 |
| 2nd block | 0.342 | 0.716 | 0.666 | 0.732 |
| 12th block | 0.377 | 0.765 | 0.742 | 0.780 |

Table 6: Organ segmentation on TotalSegmentator and prognosis prediction on STOIC.

| Method | TotalSeg (DSC) | STOIC (AUC) |
|---|---|---|
| nnUNet | 0.852 | - |
| CT-CLIP | 0.805 | 0.631 |
| Merlin | <u>0.860</u> | <u>0.782</u> |
| M3D | 0.597 | 0.627 |
| RadFM | - | 0.649 |
| Ours | **0.872** | **0.807** |

### *4.5. Additional Applications*

**Organ Segmentation.** MedVista3D learns transferable representations for organ segmentation. We finetune our model using Totalsegmentator Wasserthal et al. (2023) which contains 1204 patients and 104 organs, covering a wide range of anatomical structures. We attach a U-Net decoder to our MedVista3D-UniMISS backbone for adaptation to segmentation task. We evaluate the segmentation performance using dice coefficient (DSC). We use nnUNet's default 5-fold cross-validation split for training and testing. MedVista3D-UniMISS achieves a DSC of 0.872 on TotalSegmentator, outperforming the state-of-the-art nnUNet by 2 points in DSC and Merlin by 1.2 point (Table 6).

**Prognosis prediction.** MedVista3D also enables accurate COVID prognosis prediction. We finetune using STOIC 2021 Revel et al. (2021) dataset for pneumonia severity prediction. We randomly select 80% for training, 10%

for validation and 10% for testing. A linear head is attached for classifying severe or non-severe (defined as death or need for intubation). AUC is used to evaluate the performance. MedVista3D-UniMISS achieves 0.807 AUC outperforming all comparable methods. These results show the adaptability of our method beyond multi-modal tasks.

## 5. Conclusion

We present **MedVista3D**, a 3D VLM using unified semantic alignment to address three major diagnostic errors in radiology: under-reading, inattentional blindness, and communication failures. To jointly support local detection and global understanding from 3D CT volumes, we propose a unified alignment loss based on mutual information maximization. To mitigate variability in report language, we leverage LLM-based rewrites and introduce a Radiology Semantic Matching Bank for robust semantic alignment. MedVista3D consistently outperforms existing 3D VLMs across multiple downstream tasks, including zero-shot disease classification, report retrieval, and VQA. It also demonstrates strong transferability to organ segmentation and prognosis prediction, highlighting its potential as a general-purpose foundation model for 3D medical imaging.

**Limitation and Future Work.** Our current pretraining is limited to chest CT, primarily due to the lack of large-scale, publicly available 3D image-report datasets for other anatomical regions. In future work, we aim to expand MedVista3D to include additional anatomical sites such as the brain, head-and-neck region, and pelvis. Moreover, we plan to extend our framework to other imaging modalities, such as MRI and PET, to further enhance its generalizability across clinical contexts.

## Appendix A. Prompting LLMs for improving disease semantics

We provide the prompts for LLM in Figure A.5.

## Appendix B. Training algorithm

We also provide training pseudo code in Algorithm 1.

**Algorithm 1** MedVista3D

**Require:** $(x_i^G, y_i^G)_{i=1}^B, (x_i^r, y_i^r)_{i=1}^B, (x_i^G, \hat{y}_i^G)_{i=1}^B, (x_i^r, \hat{y}_i^r)_{i=1}^B, f_I, f_T, \text{RSMB}$

**function** COMPUTE_MEDVISTA3D_LOSS($f_I, f_T$)
  $v_i^G, t_i^G \leftarrow f_I(x_i^G), f_T(y_i^G)$ ▷ Global features.
  $v_i^r, t_i^r \leftarrow f_I(x_i^r), f_T(y_i^r)$ ▷ Local features.
  $v_i^G, \hat{t}_i^G \leftarrow f_I(x_i^G), f_T(\hat{y}_i^G)$ ▷ Global semantic features.
  $v_i^r, \hat{t}_i^r \leftarrow f_I(x_i^r), f_T(\hat{y}_i^r)$ ▷ Local semantic features.

  $\hat{t}_i^{G_{NN}} \leftarrow$ Top-1 nearest neighbor of $\hat{t}_i^G$ from RSMB ▷ Global semantic query with RSMB.
  $\hat{t}_i^{r_{NN}} \leftarrow$ Top-1 nearest neighbor of $\hat{t}_i^r$ from RSMB ▷ Local semantic query with RSMB.

  Compute $L_{\text{Global}}$ from $(v_i^G, t_i^G)$ and $L_{\text{Local}}$ from $(v_i^r, t_i^r)$
  Compute $L_{\text{Global Semantic}}$ from $(v_i^G, \hat{t}_i^{G_{NN}})$ and $L_{\text{Local Semantic}}$ from $(v_i^r, \hat{t}_i^{r_{NN}})$
  Compute $L_{\text{MedVista3D}}$ from $L_{\text{Global}}, L_{\text{Local}}, L_{\text{Global Semantic}}$ and $L_{\text{Local Semantic}}$. ▷ Calculate the losses.
  Backward $L_{\text{MedVista3D}}$ and update $f_I, f_T$ ▷ Update the network.

  RSMB $\leftarrow$ Queue_Update(RSMB, $\hat{t}_i^G$)
  RSMB $\leftarrow$ Queue_Update(RSMB, $\hat{t}_i^r$) ▷ Update RSMB.
**end function**

**function** QUEUE_UPDATE(RSMB, $\hat{t}_i$)
  $B \leftarrow$ batch size of $\hat{t}_i$
  $ptr \leftarrow$ next free position in RSMB
  $S \leftarrow$ length of RSMB
  **if** $ptr + B \geq S$ **then** ▷ Queue size is exceeded.
    $RSMB[:, ptr : S] \leftarrow \hat{t}_i[:, 0 : (S - ptr)]$ ▷ Fill remaining slots.
    $ptr \leftarrow 0$ ▷ Reset pointer to the start.
  **else**
    $RSMB[:, ptr : ptr + B] \leftarrow \hat{t}_i$ ▷ Push embeddings into the queue.
    $ptr \leftarrow ptr + B$ ▷ Advance pointer by the batch size.
  **end if**
  **return** RSMB ▷ Return updated RSMB.
**end function**

You are a medical expert in radiology, specializing in chest CT. You will read concise image findings, and then rewrite with improved clarity. Always maintain faithfulness in your rewrite. Do not include new diagnostic information. Do not include bullet points or any other symbols, or explain your thinking. Only generate sentences.

Few-shot example 1: Subsegmental atelectatic changes were observed in the right lung middle lobe.

Few-shot example 2: Peripheral consolidations observed in the apicoposterior segment of the left upper lobe demonstrate a pattern suggestive of COVID-19 pneumonia.

Few-shot example 3: Atherosclerosis is present in the aorta and coronary arteries.

Figure A.5: Prompts for report-level rewrites to emphasize disease presences.

## References


Blankemeier, L., Cohen, J.P., Kumar, A., Van Veen, D., Gardezi, S.J.S., Paschali, M., Chen, Z., Delbrouck, J.B., Reis, E., Truyts, C., et al., 2024. Merlin: A vision language foundation model for 3d computed tomography. Research Square , rs–3.

Bluethgen, C., Chambon, P., Delbrouck, J.B., van der Sluijs, R., Połacin, M., Zambrano Chaves, J.M., Abraham, T.M., Purohit, S., Langlotz, C.P., Chaudhari, A.S., 2024. A vision–language foundation model for the generation of realistic chest x-ray images. Nature Biomedical Engineering , 1–13.

Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al., 2022. Making the most of text semantics to improve biomedical vision–language processing, in: European conference on computer vision, Springer. pp. 1–21.

Bruno, M.A., Walker, E.A., Abujudeh, H.H., 2015. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. Radiographics 35, 1668–1676.

Chauhan, G., Liao, R., Wells, W., Andreas, J., Wang, X., Berkowitz, S., Horng, S., Szolovits, P., Golland, P., 2020. Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23, Springer. pp. 529–539.

Chen, H.Y., Lai, Z., Zhang, H., Wang, X., Eichner, M., You, K., Cao, M., Zhang, B., Yang, Y., Gan, Z., 2024. Contrastive localized language-image pre-training. arXiv preprint arXiv:2410.02746 .

Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., Haworth, A., 2021. A review of medical image data augmentation techniques for deep learning applications. Journal of Medical Imaging and Radiation Oncology 65, 545–563.

Draelos, R.L., Dov, D., Mazurowski, M.A., Lo, J.Y., Henao, R., Rubin, G.D., Carin, L., 2021. Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. Medical image analysis 67, 101857.

Du, Y., Onofrey, J., Dvornek, N.C., 2024. Multi-view and multi-scale alignment for contrastive language-image pre-training in mammography. arXiv preprint arXiv:2409.18119 .

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 preprint arXiv:2407.21783 .

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 .

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2021. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 .

Huang, S.C., Shen, L., Lungren, M.P., Yeung, S., 2021. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3942–3951.

Huang, W., Li, C., Zhou, H.Y., Liu, J., Yang, H., Liang, Y., Shi, G., Zheng, H., Wang, S., 2024. Enhancing representation in medical vision-language foundation models via multi-scale information extraction techniques, in: 2024 IEEE International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 1–5.

Huang, Z., Wang, H., Deng, Z., Ye, J., Su, Y., Sun, H., He, J., Gu, Y., Gu, L., Zhang, S., et al., 2023. Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. arXiv preprint arXiv:2304.06716 .

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods 18, 203–211.

Kim, Y.W., Mansfield, L.T., 2014. Fool me twice: delayed diagnoses in radiology with emphasis on perpetuated errors. American journal of roentgenology 202, 465–470.

Koetzier, L.R., Wu, J., Mastrodicasa, D., Lutz, A., Chung, M., Koszek, W.A., Pratap, J., Chaudhari, A.S., Rajpurkar, P., Lungren, M.P., et al., 2024. Generating synthetic data for medical imaging. Radiology 312, e232471.

Lee, S., Youn, J., Kim, H., Kim, M., Yoon, S.H., 2025. Cxr-llava: a multimodal large language model for interpreting chest x-ray images. European Radiology , 1–13.

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J., 2024a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems 36.

Li, X., Zhao, L., Zhang, L., Wu, Z., Liu, Z., Jiang, H., Cao, C., Xu, S., Li, Y., Dai, H., et al., 2024b. Artificial general intelligence for medical imaging analysis. IEEE Reviews in Biomedical Engineering .

Lin, J., Xia, Y., Zhang, J., Yan, K., Lu, L., Luo, J., Zhang, L., 2024. Ct-glip: 3d grounded language-image pretraining with ct scans and radiology reports for full-body scenarios. arXiv preprint arXiv:2404.15272 .

Lin, W., Zhao, Z., Zhang, X., Wu, C., Zhang, Y., Wang, Y., Xie, W., 2023. Pmc-clip: Contrastive language-image pre-training using biomedical documents, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 525–536.

Liu, C., Wan, Z., Wang, H., Chen, Y., Qaiser, T., Jin, C., Yousefi, F., Burlutskiy, N., Arcucci, R., 2024a. Can medical vision-language pre-training succeed with purely synthetic data? arXiv preprint arXiv:2410.13523 .

Liu, H., Li, C., Li, Y., Lee, Y.J., 2024b. Improved baselines with visual instruction tuning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26296–26306.

Oord, A.v.d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 .

Özbey, M., Dalmaz, O., Dar, S.U., Bedel, H.A., Özturk, Ş., Güngör, A., Çukur, T., 2023. Unsupervised medical image translation with adversarial diffusion models. IEEE Transactions on Medical Imaging .

Poole, B., Ozair, S., Van Den Oord, A., Alemi, A., Tucker, G., 2019. On variational bounds of mutual information, in: International Conference on Machine Learning, PMLR. pp. 5171–5180.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR. pp. 8748–8763.

Revel, M.P., Boussouar, S., de Margerie-Mellon, C., Saab, I., Lapotre, T., Mompoint, D., Chassagnon, G., Milon, A., Lederlin, M., Bennani, S., et al., 2021. Study of thoracic ct in covid-19: the stoic project. Radiology 301, E361–E370.

Shen, D., Wu, G., Suk, H.I., 2017. Deep learning in medical image analysis. Annual review of biomedical engineering 19, 221–248.

Shui, Z., Zhang, J., Cao, W., Wang, S., Guo, R., Lu, L., Yang, L., Ye, X., Liang, T., Zhang, Q., Zhang, L., 2025. Large-scale and fine-grained vision-language pre-training for enhanced ct image understanding, in: The Thirteenth International Conference on Learning Representations.

Stevens, S., Wu, J., Thompson, M.J., Campolongo, E.G., Song, C.H., Carlyn, D.E., Dong, L., Dahdul, W.M., Stewart, C., Berger-Wolf, T., et al., 2024. Bioclip: A vision foundation model for the tree of life, in: Proceedings of

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19412–19424.

Thawkar, O., Shaker, A., Mullappilly, S.S., Cholakkal, H., Anwer, R.M., Khan, S., Laaksonen, J., Khan, F.S., 2023. Xraygpt: Chest radiographs summarization using medical vision-language models. arXiv preprint arXiv:2306.07971 .

Waite, S., Scott, J.M., Drexler, I., Martino, J., Legasto, A., Gale, B., Kolla, S., 2018. Communication errors in radiology–pitfalls and how to avoid them. Clinical imaging 51, 266–272.

Wang, F., Zhou, Y., Wang, S., Vardhanabhuti, V., Yu, L., 2022. Multi-granularity cross-modal alignment for generalized medical visual representation learning. Advances in Neural Information Processing Systems 35, 33536–33549.

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., 2023. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence 5.

Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W., 2023. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 21372–21383.

Xie, Y., Zhang, J., Xia, Y., Wu, Q., 2022. Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier, in: European Conference on Computer Vision, Springer. pp. 558–575.

Zhang, H., Zhang, P., Hu, X., Chen, Y.C., Li, L.H., Dai, X., Wang, L., Yuan, L., Hwang, J.N., Gao, J., 2022. Glipv2: unifying localization and vl understanding, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, pp. 36067–36080.

Zhang, K., Yu, J., Adhikarla, E., Zhou, R., Yan, Z., Liu, Y., Liu, Z., He, L., Davison, B., Li, X., et al., 2023. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. arXiv e-prints , arXiv–2305.

Zhang, X., Wu, C., Zhao, Z., Lei, J., Zhang, Y., Wang, Y., Xie, W., 2024a. Radgenome-chest ct: A grounded vision-language dataset for chest ct analysis. arXiv preprint arXiv:2404.16754 .

Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., Xie, W., 2024b. Development of a large-scale medical visual question-answering dataset. Communications Medicine 4, 277.