DECOMPOSE AND DESCRIBE: NARROWING THE DIAGNOSIS-TO-REPORT GAP IN 3D CT REPORT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Automated 3D CT report generation could support radiology workflows, yet medical vision–language models (MedVLMs) often omit diagnostic findings from generated reports. We ask whether the diagnostic ability of these models carries over to free-form reporting. Across five MedVLMs, diagnostic VQA achieves higher macro-F1 than reports generated from the same CT studies and evaluated on the same disease categories. We call this task-dependent discrepancy the *diagnosis-to-report gap*. Conventional Report Adaptation improves reporting for most models but does not close the gap. To investigate and narrow it, we introduce ***DeCAD*** (*Decompose CT Analysis and Describe*), which progressively decomposes CT report generation at diagnostic, regional textual, and regional visual levels. The model identifies positive findings before writing a report, organizes diagnosis and description into regional tasks, and restricts each task to visual tokens from the corresponding region. Diagnostic and regional textual decomposition improve report-level diagnostic performance across the five MedVLMs, whereas regional visual decomposition provides limited additional gains. In a standard VLM without medical VQA pretraining, the first two levels remain effective; regional visual decomposition also provides further gains with a region-level contrastively pretrained encoder. Full DeCAD achieves state-of-the-art performance on CT-RATE and Merlin. Analyses show greater CT-token attention during diagnosis and higher recall of reference-positive findings in both the diagnostic output and final report than with Report Adaptation. These results show that explicitly organizing diagnosis improves diagnostic reporting and helps narrow the diagnosis-to-report gap. Code and models will be released upon acceptance.

1 INTRODUCTION

Automated 3D CT report generation could support radiology workflows by reducing reporting workload (Hamamci et al., 2024; Chen et al., 2025a), but generating diagnostically accurate reports from volumetric scans remains challenging (Sloan et al., 2024; Li et al., 2025; Chen et al., 2025a). A report-generation model must not only identify abnormalities throughout a volume, but also select and organize relevant findings into coherent clinical descriptions. Prior studies have shown that fluent reports do not necessarily ensure factual completeness or diagnostic accuracy (Miura et al., 2021; Chen et al., 2025b). Meanwhile, medical vision-language models (MedVLMs) have demonstrated increasingly strong image-interpretation capabilities (Li et al., 2023; Xu et al., 2025; Jiang et al., 2025; Hamamci et al., 2026b; Liu et al., 2026), including on medical visual question answering (VQA) benchmarks (Lau et al., 2018; Zhang et al., 2023; Bai et al., 2024). Advances in medical image-text pretraining (Li et al., 2024; Blankemeier et al., 2026), region-aware report generation (Tanida et al., 2023; Chen et al., 2025a), and volumetric representations (Hamamci et al., 2026a) have further improved multimodal understanding. These developments raise a fundamental question: *when a MedVLM can identify abnormalities under explicit diagnostic queries, does it reliably express the same information during free-form report generation?*

We investigate this question by comparing diagnostic VQA and free-form report generation on the same CT studies using aligned disease categories and evaluation metrics. Diagnostic VQA consistently achieves higher macro-F1 across all five MedVLMs on CT-RATE (Fig. 1b), with a similar

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

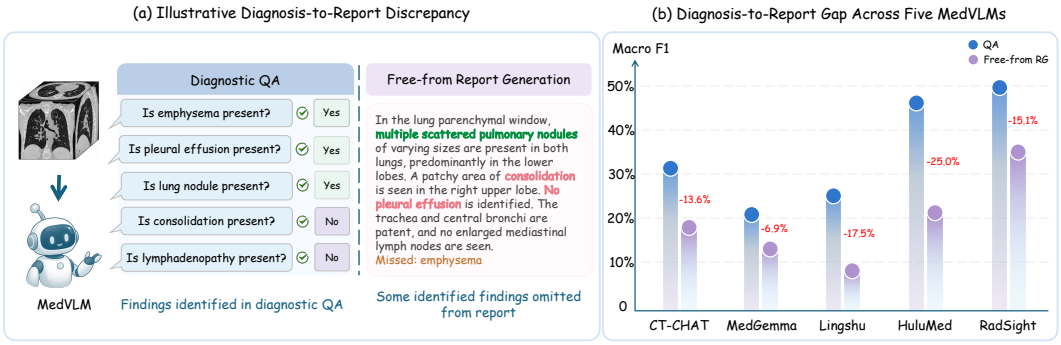

Figure 1: **Diagnostic ability and report expression in MedVLMs.** **(a)** Diagnostic QA responses and a free-form report generated from the same CT volume. **(b)** Diagnostic macro-F1 measured from QA answers and generated reports across five MedVLMs on CT-RATE.

discrepancy observed on Merlin (Table 2). We term this task-dependent discrepancy the *diagnosis-to-report gap*. The gap is particularly pronounced in recall, suggesting that free-form generation may fail to elicit and express findings that can be identified under explicit diagnostic prompts. Conventional fine-tuning on paired CT-report data, which we call *Report Adaptation*, improves diagnostic reporting but does not eliminate this discrepancy (*e.g.*, an 8.6-point macro-F1 gap for HuluMed-7B). These observations suggest that diagnostic capabilities elicited by explicit questions are not fully expressed in free-form reports, even after report-specific adaptation.

To investigate and narrow this gap, we introduce ***DeCAD*** (***De**compose **CT** **A**nalysis and **D**escribe*), a progressive adaptation strategy that structures CT report generation through diagnostic, textual, and visual decomposition. At the diagnostic level, the model first identifies positive findings from the CT volume and then generates a report conditioned on both the image and this diagnostic assessment. At the textual level, diagnostic assessment and report description are further organized into regional tasks, such as those for the lungs or heart, while retaining the full-volume CT as visual input. At the visual level, regional visual decomposition allows the language model to access only the visual tokens associated with the requested region. Across the evaluated MedVLMs, diagnostic and regional textual decomposition consistently improve reporting over direct image-to-report supervision, whereas regional visual decomposition provides limited additional benefit. These results suggest that direct reporting can underestimate the potential of existing MedVLMs: decomposition-guided adaptation can achieve state-of-the-art reporting performance without architectural modifications.

We next examine whether these benefits depend on the medical VQA pretraining of existing MedVLMs. When evaluated on a standard vision-language model without such pretraining, diagnostic and regional textual decomposition remain effective. This setting also allows us to revisit the gains from regional visual decomposition: with a region-level contrastively pretrained visual encoder, restricting visual conditioning to the requested region now provides further gains. Consequently, while diagnostic and textual decomposition remain effective without medical VQA pretraining, the efficacy of regional visual conditioning depends fundamentally on the alignment of the underlying visual representations.

To elucidate the mechanisms driving these reporting improvements, we probe the visual attention and finding expression within this standard VLM. Compared to conventional Report Adaptation, incorporating an explicit diagnostic assessment phase elicits stronger attention toward relevant CT visual tokens and significantly improves the recall of reference-positive findings—a benefit that cascades into the final generated report. These observations indicate that diagnostic decomposition systematically compels the model to ground its predictions in the visual input and articulate findings more comprehensively. Ultimately, this study demonstrates that optimizing the elicitation and structural organization of diagnostic information is as critical to automated multimodal reporting as advancing the core visual representations.

Our contributions are threefold:

- We identify a *diagnosis-to-report gap* across five MedVLMs through aligned evaluation of diagnostic VQA and free-form reporting on the same CT studies, and show that report-specific adaptation does not fully close it.
- We introduce *DeCAD*, a progressive framework for analyzing and narrowing this gap through diagnostic, regional textual, and regional visual decomposition. Diagnostic and regional textual decomposition substantially improve reporting over direct image-to-report supervision without changing model architectures.
- We show that these benefits extend to a standard vision-language model without medical VQA pretraining. With a regionally aligned visual encoder, regional visual decomposition provides further gains, enabling full DeCAD to achieve state-of-the-art reporting performance.

2 Do MedVLMs Report What They Can Diagnose?

We compare diagnostic VQA and free-form reporting to examine whether findings identifiable under explicit diagnostic queries are reflected in generated reports. We then test whether report-specific adaptation is sufficient to close the observed gap.

Finding 1: Existing MedVLMs exhibit a *diagnosis-to-report gap*: they identify more labeled findings when answering explicit diagnostic questions than they express in free-form reports from the same CT studies. *Report Adaptation* improves reporting for most models but does not close this gap.

2.1 Evaluation Protocol

Models and datasets. We evaluate five MedVLMs on CT-RATE (Hamamci et al., 2026b) across 18 thoracic diagnostic categories: CT-CHAT (Hamamci et al., 2026b), MedGemma-1.5-4B (Sellergren et al., 2026), Lingshu-7B (Xu et al., 2025), HuluMed-7B (Jiang et al., 2025), and RadSight-8B (Liu et al., 2026). We additionally evaluate Merlin (Blankemeier et al., 2026), HuluMed-7B, and RadSight-8B on the Merlin dataset across 30 abdominal diagnostic categories. We use official model checkpoints with their corresponding preprocessing and inference configurations.

Aligned evaluation. For each CT study, we compare diagnostic VQA and free-form reporting using the same disease categories and reference labels. Diagnostic VQA queries each category separately: *"Looking at these CT slices, is there <disease>? Answer yes or no."* For report generation, we use: *"Generate a comprehensive radiology report for this CT scan."* We extract disease predictions from generated reports using a publicly available RadBERT-based classifier for CT-RATE (Yan et al., 2022) and a ModernBERT-based classifier trained for Merlin (Warner et al., 2025). Both tasks are evaluated using macro-averaged precision, recall, and F1, with macro-F1 as the primary metric.

2.2 The Diagnosis-to-Report Gap in MedVLMs

We compare explicit diagnostic VQA with direct report generation. Under the aligned protocol, VQA achieves higher macro-F1 for all five models on CT-RATE, with gaps ranging from 6.9 to 25.0 points (Fig. 1b; Table 1). For HuluMed-7B, macro-F1 decreases from 46.3% under VQA to 21.3% under direct reporting, while recall decreases from 87.2% to 15.4%. Fig. 1a illustrates this discrepancy for an example CT volume.

The same pattern appears on abdominal CT. On Merlin, HuluMed-7B achieves a macro-F1 of 17.3% under VQA versus 3.8% under direct reporting, with recall of 78.1% versus 4.0% (Table 2). Together, these results characterize the *diagnosis-to-report gap*: although diagnostic information is accessible in visual representations and can be elicited by explicit queries, models express substantially less of this information in free-form reports.

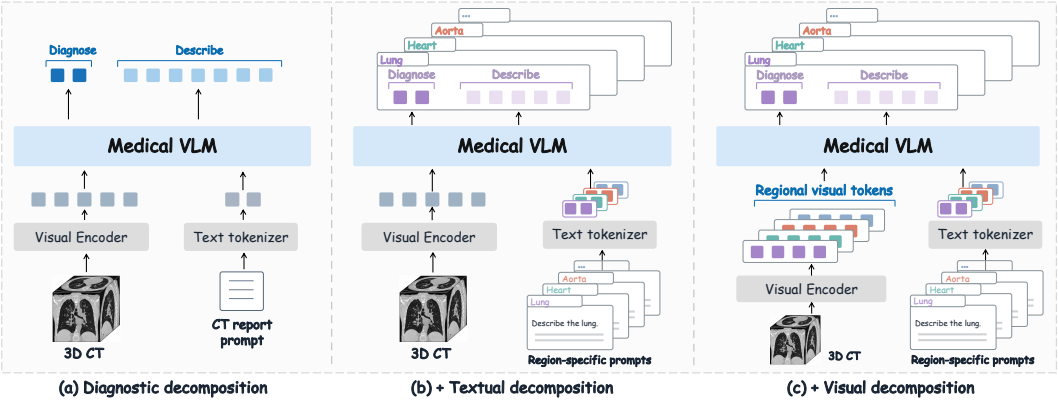

Figure 2: **DeCAD: progressive task decomposition for 3D CT report generation.** **(a)** Diagnostic decomposition generates an explicit diagnostic assessment before composing the report. **(b)** Regional textual decomposition organizes diagnostic assessment and report description by region, preserving the assessment-before-description sequence while retaining visual tokens from all target regions. **(c)** Regional visual decomposition further restricts each regional task to visual tokens associated with the requested region.

2.3 Is Report Adaptation Alone Enough?

The diagnosis-to-report gap may partly reflect insufficient adaptation to report generation. We therefore fine-tune each model on paired CT-report data using direct image-to-report supervision, termed *Report Adaptation*, while retaining its original architecture and visual input.

Results. Report Adaptation improves diagnostic macro-F1 for most models (Tables 1 and 2). On CT-RATE, Lingshu-7B improves from 7.1% to 21.5% and HuluMed-7B from 21.3% to 37.7%, but all five models remain below their diagnostic VQA scores. On Merlin, HuluMed-7B improves from 3.8% to 10.8% and RadSight-8B from 19.8% to 32.5%, still trailing their VQA scores by 6.5% and 7.9%, respectively. Thus, direct report supervision narrows but does not close the diagnosis-to-report gap, motivating explicit diagnostic assessment before report composition.

3 Can Task Decomposition Narrow the Diagnosis-to-Report Gap?

Report Adaptation leaves a residual diagnosis-to-report gap. We therefore examine whether decomposing report generation into intermediate diagnostic and regional tasks can further narrow this gap. *DeCAD* progressively decomposes report generation at the diagnostic, textual, and visual levels. We first make diagnostic assessment explicit before report composition, then organize assessment and description by region, and finally examine whether restricting visual conditioning to the corresponding region provides additional benefits.

Finding 2: Across five MedVLMs, (a) explicit *diagnostic decomposition* improves reporting over direct report supervision; (b) *regional textual decomposition* provides further gains; and (c) *regional visual decomposition* provides little incremental benefit with existing visual representations.

3.1 Diagnostic Decomposition

Under direct report supervision, a model generates a complete report r from a CT volume x:

$$x \rightarrow r. \tag{1}$$

Diagnostic assessment remains implicit within report generation. *Diagnostic decomposition* makes it explicit by introducing an intermediate assessment a. During training, the assessment target is constructed from positive disease findings extracted from the reference report, and the model learns

Table 1: Aligned diagnostic evaluation on CT-RATE. F1/P/R are macro-averaged diagnostic metrics evaluated against the same disease labels and reported as percentages. Within each model and metric, the best report-generation score is shaded purple and bold; the second-best distinct score is shaded blue and underlined. Ties share a rank. Diagnostic VQA is excluded from ranking.

	CT-CHAT			MedGemma-1.5-4B			Lingshu-7B			HuluMed-7B			RadSight-8B		
Setting	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R
Diagnostic VQA	31.4	22.7	85.0	20.6	33.7	27.9	24.6	27.5	59.3	46.3	33.4	87.2	49.7	45.1	66.8
Direct	17.8	37.4	18.1	13.7	39.4	12.4	7.1	28.6	5.4	21.3	52.5	15.4	34.6	43.9	30.9
Report Adapt.	17.7	39.7	15.1	19.8	46.0	19.5	21.5	56.2	17.8	37.7	53.8	33.2	37.2	43.8	33.7
Diagnostic Decomp.	30.7	**43.9**	30.1	32.3	**50.7**	28.9	30.8	47.2	27.8	45.0	57.4	42.9	42.2	45.3	40.6
Diag.+Text Decomp.	**36.6**	40.2	**42.0**	34.2	49.7	30.1	33.4	56.6	29.5	**51.4**	**61.2**	**48.6**	45.8	46.1	49.2
Diag.+Text+Visual	35.7	41.7	**42.0**	**34.3**	49.4	**32.3**	**33.9**	**58.2**	**30.3**	49.9	60.4	48.1	**46.4**	**46.6**	**50.6**

Table 2: Aligned diagnostic evaluation on Merlin. F1/P/R are macro-averaged diagnostic metrics evaluated against the same disease labels and reported as percentages. Within each model and metric, the best report-generation score is shaded purple and bold; the second-best distinct score is shaded blue and underlined. Ties share a rank. Diagnostic VQA is excluded from ranking.

	Merlin			HuluMed-7B			RadSight-8B		
Setting	F1	P	R	F1	P	R	F1	P	R
Diagnostic VQA	–	–	–	17.3	11.4	78.1	40.4	31.3	68.1
Direct	16.8	34.5	16.7	3.8	11.3	4.0	19.8	35.3	16.2
Report Adapt.	25.5	24.5	29.1	10.8	21.1	9.5	32.5	**38.1**	30.8
Diagnostic Decomp.	27.2	32.2	26.4	16.5	19.5	22.9	36.9	35.7	39.5
Diag.+Text Decomp.	**30.7**	**36.6**	**31.2**	**23.9**	29.0	26.2	**38.7**	35.2	**45.3**
Diag.+Text+Visual	29.0	34.5	29.5	22.2	**31.3**	**26.7**	33.4	37.3	35.1

to generate the assessment before the report:

$$x \to a \to r. \tag{2}$$

The report remains conditioned on both the CT input and the preceding assessment. At inference, the model autoregressively generates both without externally supplied diagnostic labels. Relative to *Report Adaptation*, this formulation retains the same architecture, visual input, CT-report training pairs, and report targets, while adding explicit diagnostic supervision before report composition.

Results. Diagnostic decomposition improves report generation over *Report Adaptation* across all evaluated models. On CT-RATE, macro-F1 increases by 13.0, 12.5, 9.3, 7.3, and 5.0 points for CT-CHAT, MedGemma-1.5-4B, Lingshu-7B, HuluMed-7B, and RadSight-8B, respectively, with recall also increasing for all five models (Table 1). On Merlin, macro-F1 increases by 1.7, 5.7, and 4.4 points for Merlin, HuluMed-7B, and RadSight-8B, respectively. Recall increases by 13.4 and 8.7 points for the latter two (Table 2). These results show that making diagnostic assessment explicit before report composition improves reporting across both chest and abdominal CT.

3.2 Regional Textual Decomposition

We next examine whether organizing diagnostic assessment and report description by anatomical region provides further benefits. *Regional textual decomposition* narrows each generation task to one region while preserving the order of diagnostic assessment followed by report description.

Let $\mathcal{R} = \{z_1, \ldots, z_K\}$ denote a predefined set of anatomical regions and q_z the instruction for region z. Given the full CT volume x and q_z, the model generates a regional diagnostic assessment a_z followed by the corresponding report description r_z:

$$(x, q_z) \to a_z \to r_z, \qquad z \in \mathcal{R}. \tag{3}$$

The same model is trained on paired regional diagnostic assessments and report descriptions and invoked separately for each region at inference. Regional descriptions are then assembled into a

Table 3: **Decomposition ablation without medical VQA pretraining.** Diag./Text/Visual mark the three decomposition levels; the first row is Report Adaptation. F1/P/R are macro-averaged percentages; bold marks the best in each column.

Diag.	Text	Visual	CT-RATE			Merlin		
			F1	P	R	F1	P	R
–	–	–	15.9	37.6	15.1	16.6	32.6	14.1
✓	–	–	39.0	45.7	39.0	23.2	26.2	27.7
–	✓	✓	33.3	**58.0**	29.1	27.9	38.3	28.8
✓	✓	–	41.6	47.2	42.8	34.1	37.1	38.3
✓	✓	✓	**52.1**	55.6	**54.2**	**39.4**	**43.1**	**42.3**

Table 4: Report-generation comparison on CT-RATE. All scores are reported as percentages. F1/P/R are macro-averaged diagnostic metrics. MET, RaTE, RG, and GR denote METEOR, RaTEScore, RadGraph-F1, and GREEN. For each metric, the best distinct score is shaded purple and bold; the second-best is shaded blue and underlined. Ties share a rank.

Method	Diagnostic			Text		Clinical		
	F1	P	R	MET	BLEU	RaTE	RG	GR
Lingshu-32B (Xu et al., 2025)	9.3	19.7	8.4	10.4	1.0	50.4	5.7	10.8
RadFM (Wu et al., 2025)	11.6	19.5	11.2	3.7	0.1	34.8	0.8	1.3
MedGemma-1.5 (Sellergren et al., 2026)	13.7	39.4	12.4	7.1	0.7	49.6	10.3	11.1
Reg2RG (Chen et al., 2025a)	17.1	39.6	12.5	24.6	23.5	**66.9**	<u>25.2</u>	37.9
BTB3D (Hamamci et al., 2026a)	17.3	21.0	15.9	20.4	15.8	59.0	15.1	22.2
CT-CHAT (Hamamci et al., 2026b)	17.8	37.4	18.1	23.6	23.6	**66.9**	22.2	**42.2**
RadSight-8B (Liu et al., 2026)	34.6	43.8	30.9	23.0	21.2	65.7	21.6	38.6
HuluMed-32B (Jiang et al., 2025)	38.0	<u>48.0</u>	37.0	22.5	17.9	64.1	17.0	39.8
ViSD-Boost (Cao et al., 2025)	39.4	42.2	39.8	<u>24.8</u>	21.4	65.8	22.3	34.9
AdaRAG-CT-8B (Liang et al., 2026)	<u>41.8</u>	45.1	<u>46.7</u>	24.6	<u>24.2</u>	65.7	**25.8**	26.3
Ours (*full DeCAD*)	**52.1**	**55.6**	**54.2**	**26.3**	**27.1**	<u>66.5</u>	23.6	<u>42.0</u>

complete report. All tasks retain the full-volume visual input, allowing us to isolate textual decomposition while holding the visual input fixed.

Results. Compared with diagnostic decomposition, adding regional textual decomposition raises macro-F1 by 5.9, 1.9, 2.6, 6.4, and 3.6 points for CT-CHAT, MedGemma-1.5-4B, Lingshu-7B, HuluMed-7B, and RadSight-8B on CT-RATE, respectively. Recall increases by 1.2–11.9 points across all five models (Table 1). On Merlin, macro-F1 increases by 3.5, 7.4, and 1.8 points for Merlin, HuluMed-7B, and RadSight-8B, respectively, with corresponding recall gains of 4.8, 3.3, and 5.8 points (Table 2). Thus, organizing diagnostic assessment and report description by region further improves reporting across both datasets without changing the visual input.

3.3 Regional Visual Decomposition

Regional textual decomposition conditions each regional diagnostic assessment and report description on the full visual input. *Regional visual decomposition* instead allows the language model to access only the visual tokens associated with the requested region. The language model and regional diagnostic assessment and report description targets remain unchanged.

Results. Relative to regional textual decomposition, adding regional visual decomposition yields small macro-F1 changes on CT-RATE: -0.9, $+0.1$, $+0.5$, -1.5, and $+0.6$ points for CT-CHAT, MedGemma-1.5-4B, Lingshu-7B, HuluMed-7B, and RadSight-8B, respectively (Table 1). On Merlin, the corresponding changes are -1.7, -1.7, and -5.3 points for Merlin, HuluMed-7B, and RadSight-8B, respectively (Table 2). Thus, regional visual decomposition provides little incremental benefit with the visual representations of these existing MedVLMs, motivating the representation-alignment analysis.

4 Does Task Decomposition Require Medical VQA Pretraining for 3D CT Report Generation?

Diagnostic and regional textual decomposition improve reporting in existing MedVLMs, but these gains may depend on prior medical VQA training. We therefore examine *DeCAD* in a standard vision-language model without medical VQA pretraining. Its visual encoder is first trained with region-level CT-report contrastive learning to align regional visual representations with corresponding report descriptions. This setting allows us to test whether diagnostic and textual decomposition remain effective without medical VQA supervision and whether regional visual decomposition becomes beneficial with regionally aligned visual representations.

Finding 3: Diagnostic and regional textual decomposition remain effective without medical VQA pretraining. With a regionally aligned visual encoder, regional visual decomposition provides further gains. *Full DeCAD* achieves state-of-the-art reporting performance.

4.1 A Standard Vision-Language Model and Training Setup

Our standard VLM comprises a 3D ResNet-based visual encoder (Hara et al., 2018), a projection layer, and a general-purpose Qwen3-4B language model (Yang et al., 2025). The visual encoder undergoes region-level CT-report contrastive pretraining to align regional visual features with corresponding report descriptions. It produces 32 visual tokens per target region, which the projection layer maps into the language model's embedding space.

We then freeze the visual encoder and jointly train the projection layer and language model using either *Report Adaptation* or the decomposition formulations in Section 3. The architecture and pretrained visual encoder remain fixed across settings, allowing these formulations to be compared under the same visual representation. Further architecture and training details are provided in Appendix C.

4.2 Diagnostic and Textual Decomposition without Medical VQA Pretraining

We compare *Report Adaptation*, diagnostic decomposition, and diagnostic plus regional textual decomposition in the model without medical VQA pretraining (Section 4.1). All settings share the same region-level contrastively pretrained visual encoder, language-model initialization, and CT-report training pairs. Each receives visual tokens from all target regions, allowing us to evaluate diagnostic and textual decomposition while holding the visual input fixed.

Results. Diagnostic decomposition raises macro-F1 from 15.9% to 39.0% on CT-RATE and from 16.6% to 23.2% on Merlin (Table 3). Adding regional textual decomposition further raises these scores to 41.6% and 34.1%, gains of 2.6 and 10.9 points, respectively. Compared with Report Adaptation, the combined setting increases macro recall from 15.1% to 42.8% on CT-RATE and from 14.1% to 38.3% on Merlin. Macro precision also rises, from 37.6% to 47.2% and from 32.6% to 37.1%, respectively. Thus, diagnostic and regional textual decomposition improve the coverage of reference-positive findings without sacrificing precision, even without medical VQA pretraining.

4.3 Regional Visual Decomposition with Region-Level Contrastive Pretraining

Section 3.3 showed that regional visual decomposition provides little incremental benefit with the visual representations of existing MedVLMs. We therefore revisit it in our standard VLM, whose visual encoder undergoes region-level image-text contrastive pretraining. Keeping the regional diagnostic assessment and report description tasks fixed, we restrict the language model to the 32 visual tokens associated with the requested region.

Results. Adding regional visual decomposition to diagnostic and regional textual decomposition increases macro-F1 from 41.6% to 52.1% on CT-RATE and from 34.1% to 39.4% on Merlin, gains of 10.5 and 5.3 points, respectively (Table 3). The complete formulation also outperforms Text+Visual without diagnostic decomposition, which achieves macro-F1 scores of 33.3% and 27.9%, respec-

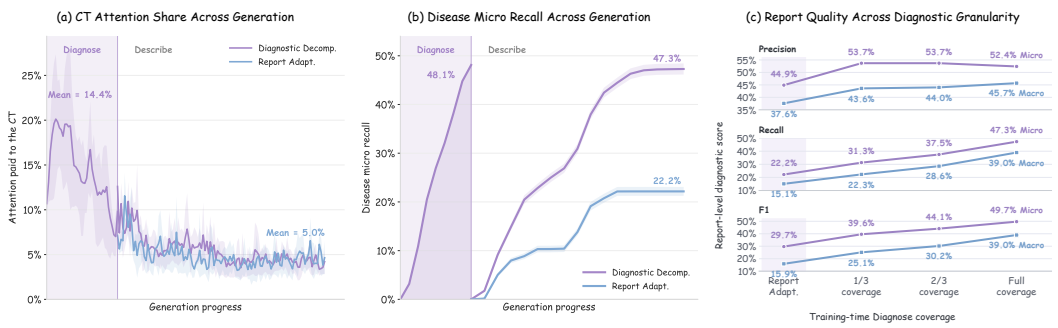

Figure 3: **Effects of Diagnostic Decomposition on report generation. (a)** Attention share assigned to CT visual tokens during *Diagnose* and *Describe*. **(b)** Micro recall of reference-positive findings across generated prefixes; the *Describe* curve counts findings in the report alone. **(c)** Report-level precision, recall, and F1 under different levels of training-time *Diagnose* coverage.

tively. Combining all three decomposition steps therefore yields the highest scores among these ablations.

Table 4 compares *full DeCAD*, combining region-level contrastive pretraining with all three decomposition steps, against prior methods on CT-RATE. Full DeCAD achieves the highest macro-F1, precision, and recall at 52.1%, 55.6%, and 54.2%, respectively. Its macro-F1 exceeds that of the strongest compared baseline, AdaRAG-CT-8B (Liang et al., 2026), by 10.3 points. It also achieves the highest METEOR (Banerjee & Lavie, 2005) and BLEU-4 (Papineni et al., 2002) scores, surpassing the next-best results by 1.5 and 2.9 points, respectively. Clinical scores remain competitive, trailing the best RaTEScore (Zhao et al., 2024), GREEN (Ostmeier et al., 2024), and RadGraph-F1 (Jain et al., 2021) by 0.4, 0.2, and 2.2 points, respectively. Overall, Full DeCAD achieves the strongest diagnostic performance while maintaining high text-generation quality and competitive clinical scores. Together with the controlled ablations in Table 3, these results show that the three decomposition steps improve reporting in this model without medical VQA pretraining.

4.4 How Does Diagnostic Decomposition Change Report Generation?

Visual attention and finding expression. To examine how diagnostic decomposition may affect report generation, we analyze visual attention across generation stages. We teacher-force each model on its own generated output and measure the attention share assigned to CT visual tokens:

$$v_t = \frac{1}{H} \sum_{h=1}^{H} \sum_{j \in \mathcal{V}} \alpha_{t,j,h}, \tag{4}$$

where H is the number of attention heads, $\mathcal{V}$ denotes visual-token positions, and $\alpha_{t,j,h}$ is the attention weight from position t to position j in head h of the penultimate language-decoder layer. After resampling trajectories within each phase and averaging across cases, the mean visual attention share reaches 14.4% during *Diagnose*, compared with 5.0% for Report Adaptation. During *Describe*, it remains comparable to Report Adaptation (Fig. 3a). This pattern is consistent with greater reliance on CT visual tokens during explicit diagnostic assessment.

We next examine finding expression during the stage with increased visual attention. We track micro recall of reference-positive findings in generated prefixes, reading the *Diagnose* list directly and using RadBERT for report text. The report curve resets at the onset of *Describe* and counts only findings expressed in the report. Recall reaches 48.1% during *Diagnose*, compared with 22.2% for Report Adaptation, and reaches 47.3% in the subsequent report (Fig. 3b). Thus, the diagnostic phase exhibits both greater visual attention and more complete identification of reference-positive findings, while the resulting report retains substantially higher finding recall than *Report Adaptation*.

Coverage of diagnostic supervision. We vary the proportion of training cases receiving explicit diagnostic assessment supervision, randomly selecting approximately one-third, two-thirds, or all cases. At inference, diagnostic assessments remain model-generated. Supervising one-third of

the cases raises report-level micro precision from 44.9% to 53.7% relative to Report Adaptation (Fig. 3c). As supervision coverage increases, micro recall rises from 31.3% to 37.5% and 47.3%, while precision remains near 53%. These results are consistent with a role for explicit diagnostic supervision in improving finding precision, while broader supervision coverage is associated with higher recall of positive findings in reports.

5 Related Work

LLM-based MedVLMs. MedVLMs acquire medical image-understanding capabilities through domain-specific pretraining and instruction tuning. LLaVA-Med and Med-Flamingo explore biomedical instruction following and multimodal few-shot learning (Li et al., 2023; Moor et al., 2023), while RadFM, M3D, and CT-CHAT extend medical image understanding to volumetric inputs (Wu et al., 2025; Bai et al., 2024; Hamamci et al., 2026b). Lingshu, HuluMed, and MedGemma 1.5 further expand task and modality coverage (Xu et al., 2025; Jiang et al., 2025; Sellergren et al., 2026). Other efforts emphasize specific capabilities, such as segmentation and localization in VividMed (Luo et al., 2025), perception-oriented training in RadSight (Liu et al., 2026), and diagnostic guidance in GSCo (He et al., 2026). These advances have strengthened medical image understanding, yet it remains unclear how much of MedVLMs' diagnostic ability is reflected in free-form CT reports. We characterize this *diagnosis-to-report gap* through aligned evaluation of diagnostic VQA and reporting on the same CT studies, and examine whether DeCAD improves the diagnostic content of reports generated by MedVLMs.

CT report generation. CT report generation has explored both visual representation learning and diagnostic guidance. CT2Rep generates reports from volumetric inputs using a visual encoder and autoregressive decoder (Hamamci et al., 2024), while Argus investigates visual encoder pretraining, token compression, and scaling strategies (Liu et al., 2025). Reg2RG combines global context with anatomical region features and aligns regional representations with report descriptions (Chen et al., 2025a). AdaRAG-CT addresses incomplete pathology coverage by adaptively retrieving supplementary text and selectively incorporating it during generation (Liang et al., 2026). For diagnostic guidance, Dia-LLaMA introduces disease prototype memory and disease-aware attention (Chen et al., 2025b), while CT-AGRG predicts abnormalities before generating reports (Di Piazza et al., 2025). These studies improve CT reporting through visual representations, regional information, retrieval, and diagnostic guidance, but do not systematically examine the diagnostic gains from decomposing the generation process. DeCAD progressively decomposes diagnostic identification, regional textual organization, and visual conditioning to assess each level's contribution to the model's diagnostic performance and its final reports.

6 Conclusion and Limitations

We identified a *diagnosis-to-report gap* in MedVLMs: diagnostic VQA consistently outperformed free-form reporting on aligned disease categories across five models on CT-RATE, with a similar discrepancy on Merlin. *Report Adaptation* improved reporting for most models but did not close the gap. Through *DeCAD*, we examined diagnostic, regional textual, and regional visual decomposition in sequence. Making diagnosis explicit before report composition improved reporting across the evaluated MedVLMs on CT-RATE, and organizing diagnosis and description by region provided further gains while retaining full-volume visual input. Regional visual decomposition added limited benefit in these models but provided further gains in a VLM with a visual encoder pretrained through region-level image–text contrastive learning. Without medical VQA pretraining, full DeCAD achieved the highest diagnostic macro-F1 among the compared report-generation methods on CT-RATE and Merlin, reaching 52.1% and 39.4%, respectively. These findings show that explicitly organizing diagnosis and description improves diagnostic reporting and helps narrow the diagnosis-to-report gap.

Limitations. Our diagnostic evaluation is restricted to the labeled disease categories in CT-RATE and Merlin and relies on RadBERT and ModernBERT for report labeling, respectively. Future work should extend evaluation to broader clinical settings and incorporate expert assessment.

AI USE STATEMENT

We used Qwen-Max to decompose reference CT reports into regional descriptions and to extract and normalize positive disease findings for constructing training targets, as described in Appendix B.1. We used ChatGPT to assist with manuscript organization and language editing. The authors reviewed the AI-assisted manuscript content and take responsibility for the data processing, experiments, results, and conclusions.

REFERENCES

Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.

Satanjeev Banerjee and Alon Lavie. 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*, pp. 65–72, 2005.

Louis Blankemeier, Ashwin Kumar, Joseph Paul Cohen, Jiaming Liu, Longchao Liu, Dave Van Veen, Syed Jamal Safdar Gardezi, Hongkun Yu, Magdalini Paschali, Zhihong Chen, et al. Merlin: a computed tomography vision–language foundation model and dataset. *Nature*, 652 (8112):1318–1328, 2026.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pp. 1–21. Springer, 2022.

Weiwei Cao, Jianpeng Zhang, Zhongyi Shui, Sinuo Wang, Zeli Chen, Xi Li, Le Lu, Xianghua Ye, Qi Zhang, Tingbo Liang, et al. Boosting vision semantic density with anatomy normality modeling for medical vision-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23041–23050, 2025.

Zhixuan Chen, Yequan Bie, Haibo Jin, and Hao Chen. Large language model with region-guided referring and grounding for ct report generation. *IEEE transactions on Medical Imaging*, 44(8): 3139–3150, 2025a.

Zhixuan Chen, Luyang Luo, Yequan Bie, and Hao Chen. Dia-llama: Towards large language model-driven ct report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 141–151. Springer, 2025b.

Theo Di Piazza, Carole Lazarus, Olivier Nempont, and Loic Boussel. Ct-agrg: Automated abnormality-guided report generation from 3d chest ct volumes. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pp. 01–05. IEEE, 2025.

Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. Ct2rep: Automated radiology report generation for 3d medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 476–486. Springer, 2024.

Ibrahim Ethem Hamamci, Sezgin Er, Suprosanna Shit, Hadrien Reynaud, Dong Yang, Pengfei Guo, Marc Edgar, Daguang Xu, Bernhard Kainz, and Bjoern Menze. Better tokens for better 3d: Advancing vision-language modeling in 3d medical imaging. *Advances in Neural Information Processing Systems*, 38:135074–135102, 2026a.

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, et al. Generalist foundation models from a multimodal dataset for 3d computed tomography. *Nature Biomedical Engineering*, pp. 1–19, 2026b.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 6546–6555. IEEE, 2018.

Sunan He, Yuxiang Nie, Hongmei Wang, Shu Yang, Yihui Wang, Zhiyuan Cai, Zhixuan Chen, Yingxue Xu, Linshan Wu, Ngai Shing Cheng, et al. Towards generalizable ai in medicine via generalist–specialist collaboration. *Nature Biomedical Engineering*, pp. 1–15, 2026.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*, 2021.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):180251, 2018.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in neural information processing systems*, 36:28541–28564, 2023.

Qingqiu Li, Xiaohan Yan, Jilan Xu, Runtian Yuan, Yuejie Zhang, Rui Feng, Quanli Shen, Xiaobo Zhang, and Shujun Wang. Anatomical structure-guided medical vision-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 80–90. Springer, 2024.

Yilin Li, Chao Kong, Guosheng Zhao, and Zijian Zhao. Automatic radiology report generation with deep learning: a comprehensive review of methods and advances. *Artificial Intelligence Review*, 58(11):344, 2025.

Renjie Liang, Yiling Ma, Yang Xing, Zhengkang Fan, Jinqian Pan, Chengkun Sun, Li Li, Kuang Gong, and Jie Xu. Beyond the embedding bottleneck: Adaptive retrieval-augmented 3d ct report generation. *arXiv preprint arXiv:2603.15822*, 2026.

Che Liu, Zhongwei Wan, Yuqi Wang, Hui Shen, Haozhe Wang, Kangyu Zheng, Mi Zhang, and Rossella Arcucci. Argus: benchmarking and enhancing vision-language models for 3d radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16448–16460, 2025.

Jianqin Liu, Weiwei Cao, Wanxing Chang, Ruifeng Yuan, Bowen Shi, Zhilin Zheng, Xianjie Zhang, Ling Zhang, Peng Wang, and Jianpeng Zhang. Radsight: Towards perceptually reliable multimodal radiology image understanding. *arXiv preprint arXiv:2607.22293*, 2026.

Lingxiao Luo, Bingda Tang, Xuanzhong Chen, Rong Han, and Ting Chen. Vividmed: Vision language model with versatile visual grounding for medicine. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1800–1821, 2025.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5288–5304, 2021.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine learning for health (ML4H)*, pp. 353–367. PMLR, 2023.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson Md, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, et al. Green: Generative radiology report evaluation and error notation. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 374–390, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, et al. Medgemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*, 2026.

Zhongyi Shui, Jianpeng Zhang, Weiwei Cao, Sinuo Wang, Ruizhe Guo, Le Lu, Lin Yang, Xianghua Ye, Tingbo Liang, Qi Zhang, et al. Large-scale and fine-grained vision-language pre-training for enhanced ct image understanding. In *International Conference on Learning Representations*, volume 2025, pp. 24094–24107, 2025.

Phillip Sloan, Philip Clatworthy, Edwin Simpson, and Majid Mirmehdi. Automated radiology report generation: A review of recent advances. *IEEE Reviews in Biomedical Engineering*, 18:368–387, 2024.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7433–7442. IEEE, 2023.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 2526–2547, 2025.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866, 2025.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025.

An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y Chang, Amilcare Gentili, and Chun-Nan Hsu. Radbert: adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 4(4):e210258, 2022.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Ratescore: A metric for radiology report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15004–15019, 2024.

7 APPENDIX

A EXPERIMENTAL SETUP

Datasets. We evaluate on two paired 3D CT-report datasets. **CT-RATE** contains 25,692 non-contrast chest CT studies from 21,304 patients, paired with radiology reports and annotated with 18 thoracic abnormality categories. **Merlin** contains 25,494 abdominal CT scans paired with radiology reports. We use its 30 common imaging findings as the diagnostic ontology for evaluation. For both datasets, we follow the official data splits.

Existing MedVLMs. On CT-RATE, we evaluate HuluMed-7B, Lingshu-7B, CT-CHAT-8B, MedGemma-1.5-4B, and RadSight-8B. On the Merlin dataset, we evaluate the Merlin model, HuluMed-7B, and RadSight-8B. We follow each model's original CT preprocessing pipeline. HuluMed-7B uses its native 3D processor with 180 sagittal slices from a volume normalized to $[0, 1]$. Lingshu-7B uses 40 axial slices after percentile-based intensity stretching, resizing to 448×448, and grayscale-to-RGB conversion. CT-CHAT-8B encodes CT volumes in Hounsfield units (HU) into visual latents offline. MedGemma-1.5-4B follows its HU-based input pipeline, while RadSight-8B follows its official preprocessing configuration using only the 3D visual pathway. For each model, the same input pipeline is used across diagnostic VQA and report-generation settings.

All fine-tuned variants of the existing MedVLMs are trained for two epochs with a learning rate of 2×10^{-5}, using a cosine learning-rate schedule with a warmup ratio of 0.03, a per-GPU batch size of 1, and gradient accumulation over 8 steps. We use LoRA ($r = 128$, $\alpha = 256$) for Merlin, HuluMed-7B, Lingshu-7B, and RadSight-8B; LLaVA-style adaptation for CT-CHAT-8B; and full fine-tuning for MedGemma-1.5-4B.

Our standard VLM. Our standard VLM combines a 3D ResNet visual encoder with Qwen3-4B. We retain the original orientation of the preprocessed CT volumes, padding or cropping CT-RATE volumes to $112 \times 256 \times 352$ and Merlin volumes to $96 \times 256 \times 384$. Voxel intensities are clipped to $[-1000, 1000]$ HU and linearly mapped to $[0, 1]$. Training proceeds in two stages. We first pretrain the visual encoder for 10 epochs using region-level image-text contrastive learning with a peak learning rate of 1×10^{-4}. We then freeze the visual encoder and jointly train the visual projector and the full language model for five epochs with a learning rate of 2×10^{-5}. Stage-II training uses a cosine learning-rate schedule with a warmup ratio of 0.03, a per-GPU batch size of 1, and gradient accumulation over 8 steps.

All training runs use four nodes with 32 NVIDIA H20 96GB GPUs in total.

Evaluation metrics. We evaluate generated reports along three dimensions. For diagnostic accuracy, we extract disease labels from the generated `<report>` body using a RadBERT-based report classifier on CT-RATE and a ModernBERT-based report classifier on Merlin, and compare them against the ground-truth labels. We report macro-averaged F1, precision, and recall, with macro-F1 as the primary metric. For text similarity, we report METEOR (MET) and BLEU-4 (BLEU). For clinical report quality, we report RaTEScore (RaTE), RadGraph-F1 (RG), and GREEN (GR), using the official GREEN-RadLlama2-7B evaluator (Ostmeier et al., 2024).

B DATA CONSTRUCTION AND DISEASE VOCABULARIES

B.1 REGION-LEVEL REPORT DECOMPOSITION AND DISEASE EXTRACTION

To construct region-specific supervision, we adopt a two-stage processing pipeline for both CT-RATE and Merlin using Qwen-Max. The expanded vocabularies described below provide detailed diagnostic supervision for model training, while diagnostic evaluation uses the 18 CT-RATE and 30 Merlin disease categories introduced in Sec. A. We first decompose each radiology report into anatomical-region sections while retaining the clinical information relevant to each region. We then extract explicitly positive abnormalities from each regional section and normalize synonymous expressions into standardized disease labels. For clarity, Figs. 4 and 5 show the prompts instantiated

Prompt Design for Region-Level Report Extraction

You are a specialized chest CT radiology report parser. Split the provided Findings and Impression into region-level sections with zero information loss.

Region Mapping

lung: lung parenchyma, airways, bronchi, trachea, pleura
heart: heart, pericardium, cardiac valves, coronary arteries, pulmonary artery/trunk
aorta: thoracic aorta, mediastinal vessels, mediastinum, hilar/mediastinal lymph nodes
esophagus: esophagus, gastroesophageal junction, hiatal hernia
other: spine, chest wall, soft tissue, abdomen, and all other non-target findings

Rules

1. Preserve all clinical information exactly. Do not paraphrase, summarize, rewrite, or omit content. Keep measurements, comparisons, negations, and temporal descriptions intact.
2. Every piece of information must appear in at least one section.
3. Multi-organ statements must be duplicated to all relevant organs.
   "Calcified atherosclerotic changes in the aorta and coronary arteries." → aorta + heart
   "Cardiomegaly with bilateral pleural effusions." → heart + lung
   "Pacemaker on the left chest wall with leads extending to the ventricle." → heart + other
4. Keep related information together. Do not separate a finding from its measurement, comparison, cause, consequence, or associated finding.
   "Nodule measuring 38 mm, previously 32 mm." → keep intact
   "Compressive atelectasis associated with pleural effusion." → keep intact
5. Process Findings and Impression using the same rules. Impression statements should be assigned to the corresponding organ(s).
6. Preserve normal and negative findings.
   "Heart size is normal." → heart
   "No pleural effusion." → lung
   "Esophagus is unremarkable." → esophagus
7. Assign medical devices/materials by anatomical relevance.
   coronary stent → heart
   chest tube → lung
   aortic stent graft → aorta
   multi-organ devices → duplicate to all relevant sections
8. Preserve the original order of statements within each section.
9. Do not infer new findings or organ associations not explicitly stated in the report.

Output

Return only a valid JSON object:
{ "lung": "", "heart": "", "aorta": "", "esophagus": "", "other": "" }
Use the exact original wording. If a section has no relevant content, return an empty string.

Figure 4: **Prompt for region-level report decomposition.** The prompt divides the original chest CT Findings and Impression into region-specific report sections for the lung, heart, aorta/mediastinum, and esophagus. It instructs the model to preserve the source clinical information, assign each statement to its relevant anatomical region, and repeat findings involving multiple regions when necessary. The resulting sections provide region-level textual supervision.

for CT-RATE. Merlin follows the same procedure, with the region definitions and disease vocabulary adapted to its anatomical coverage.

B.2 CT-RATE Disease Vocabulary

For CT-RATE, we consider four anatomical regions and obtain **66 region-disease label assignments (64 unique disease labels)**: 42 for the lung, 12 for the heart, 9 for the aorta/mediastinum, and 3 for the esophagus.

Lung (42): *lung nodule, ground-glass opacity, atelectasis, emphysema, pulmonary fibrotic sequela, consolidation, pleural effusion, peribronchial thickening, bronchiectasis, calcified nodule, interlobular septal thickening, mosaic attenuation pattern, pneumonia, centrilobular nodule, crazy-paving pattern, pleural thickening, air cyst, vascular enlargement, ground-glass nodule, bulla, lung mass, metastatic nodule, pleural calcification, dependent density increase, cavitary lesion, honeycombing, halo sign, lung opacity, mucus plugging, pneumothorax, bronchiolitis, pulmonary edema, parenchy-*

Prompt Design for Region-Level Disease Extraction

You are a medical terminology expert for chest CT reports. Extract all explicitly positive abnormalities from the provided Findings and Impression, and normalize them into standardized disease/condition names.

Extraction Rules

1. Positive findings only. Extract abnormalities explicitly stated as present.
   Exclude negated findings: no, without, negative for, normal, unremarkable.
   Exclude uncertain findings: possible, suspected, questionable, cannot exclude, may represent, differential diagnosis.
   If an uncertain diagnosis is based on a definite imaging finding, extract only the definite finding.
   Example: "Ground-glass opacity suggesting infection" → ground_glass_opacity, not infection.
2. Extract each distinct abnormality separately.
3. Normalize synonymous expressions and singular/plural variants to the same canonical disease key.
4. Deduplicate repeated findings. If the same abnormality appears in both Findings and Impression, output it only once.
5. Assign each finding to its anatomical organ.
   lung/airway/pleura → lung
   heart/pericardium/valves/coronary arteries → heart
   thoracic aorta/mediastinum/mediastinal or hilar lymph nodes → aorta
   esophagus/GE junction/hiatal hernia → esophagus
   medical devices/materials → medical_material, assigned to the organ where located

Standardization Rules

1. Output each finding as organ__disease_key.
2. Use lowercase English with underscores.
3. Do not merge findings that represent different radiological abnormalities.

Output

Return only a JSON array:
["lung__atelectasis", "heart__cardiomegaly"]
Return [] if no eligible positive findings are present. Do not output explanations, descriptions, or markdown.

Figure 5: **Prompt for region-level disease extraction.** The prompt extracts explicitly positive abnormalities from chest CT Findings and Impression, assigns each finding to its relevant anatomical region, and normalizes synonymous expressions into standardized disease labels while excluding negated or uncertain findings and repeated mentions within each region.

mal calcification, fissure thickening, pleural nodule, intrapulmonary lymph node, tracheal diverticulum, tracheomegaly, lymphangitis carcinomatosa, pulmonary artery dilation, empyema, and *bronchial obstruction*.

Heart (12): *coronary artery wall calcification, cardiomegaly, pericardial effusion, pulmonary artery dilation, medical material, valve calcification, aortic valve calcification, atrial enlargement, ventricular dilation, mitral valve calcification, pericardial thickening*, and *pulmonary hypertension*.

Aorta/Mediastinum (9): *arterial wall calcification, aortic dilation, calcified lymph node, lymphadenopathy, aortic aneurysm, medical material, aberrant subclavian artery, mediastinal mass*, and *pneumomediastinum*.

Esophagus (3): *hiatal hernia, esophageal wall thickening*, and *esophageal dilation*.

B.3 Merlin Disease Vocabulary

For Merlin, the same procedure is extended to its broader anatomical coverage. The resulting vocabulary contains **120 unique disease labels**, including **110 region-specific disease labels** across 15 anatomical regions and **10 additional study-level labels**. Unlike CT-RATE, each region-specific disease label in Merlin is assigned to a single anatomical region without duplication across regions.

Adrenal Gland (5): *adrenal nodule, adrenal adenoma, adrenal thickening, adrenal mass*, and *surgically absent adrenal gland*.

Aorta (5): *abdominal aortic aneurysm, atherosclerosis, aortic calcification, aortic dissection*, and *aortic occlusion*.

Bladder (6): *bladder wall thickening, bladder distension, bladder mass, bladder calculus, cystitis*, and *surgically absent bladder*.

Esophagus (5): *hiatal hernia, esophageal dilation, esophageal wall thickening, esophageal varices*, and *surgically absent esophagus*.

Gallbladder (8): *gallstones, surgically absent gallbladder, cholecystitis, gallbladder wall thickening, gallbladder distension, pericholecystic fluid, adenomyomatosis*, and *biliary stent*.

Heart (5): *cardiomegaly, pericardial effusion, aortic valve calcification, mitral valve calcification*, and *cardiac chamber enlargement*.

Kidney (11): *renal cyst, renal hypodensities, hydronephrosis, hydroureter, renal calculus, renal atrophy, renal mass, renal infarct, perinephric stranding, surgically absent kidney*, and *delayed nephrogram*.

Large Bowel (11): *submucosal edema, diverticulosis, diverticulitis, colonic wall thickening, appendicitis, colonic mass, pneumatosis, rectal wall thickening, fecal loading, surgically absent appendix*, and *surgically absent large bowel*.

Liver (14): *hepatic steatosis, hepatomegaly, hepatic cyst, hepatic metastasis, hepatic hemangioma, hepatic calcification, cirrhosis, biliary ductal dilation, intrahepatic biliary prominence, hepatic lesion, hepatic abscess, hepatic laceration, portal hypertension*, and *hepatic atrophy*.

Lumbar Vertebrae (5): *fracture, compression fracture, degenerative changes, lytic lesion*, and *sclerotic lesion*.

Lung (8): *atelectasis, pleural effusion, pulmonary nodule, consolidation, ground-glass opacity, bronchiectasis, pulmonary embolism*, and *pneumothorax*.

Pancreas (8): *pancreatic atrophy, pancreatic calcification, pancreatic cyst, pancreatic mass, pancreatic duct dilation, pancreatitis, pancreatic fatty replacement*, and *surgically absent pancreas*.

Small Bowel (6): *bowel obstruction, small bowel wall thickening, small bowel dilation, ileus, mesenteric stranding*, and *surgically absent small bowel*.

Spleen (6): *splenomegaly, splenic lesion, splenic calcification, splenic infarct, surgically absent spleen*, and *accessory spleen*.

Stomach (7): *gastric wall thickening, gastric distension, gastric mass, gastric bypass changes, gastric varices, gastric diverticulum*, and *surgically absent stomach*.

Additional Study-level Labels (10): *ascites, anasarca, coronary calcification, free air, osteopenia, lymphadenopathy, metastatic disease, prostatomegaly, thrombosis*, and *medical material*.

Together, the decomposed reports and standardized disease labels provide structured supervision for subsequent model training on both datasets.

C Model Architecture, and Training

Our standard VLM consists of a 3D visual encoder (Hara et al., 2018), a projection layer, and a Qwen3 language model (Yang et al., 2025). Training is performed in two stages. We first pretrain the visual encoder with region-level image-text contrastive learning and auxiliary organ segmentation. We then freeze the visual encoder and jointly optimize the projection layer and language model for report generation.

C.1 Stage I: Visual Encoder Pretraining

The image branch employs a 3D ResNet-18 backbone to extract multi-scale volumetric features, together with an auxiliary segmentation branch for organ localization. The text branch uses BiomedVLP-CXR-BERT (Boecking et al., 2022). Image and text representations are projected into a shared $d = 256$ dimensional embedding space.

The segmentation branch provides organ masks that localize each target region in the feature map. For organ o, we collect the spatial features within its corresponding mask as $X_o \in \mathbb{R}^{N_o \times d}$, where N_o denotes the number of spatial features assigned to that organ. A two-layer MLP predicts $M = 32$

Table 5: **Zero-shot disease AUC on CT-RATE.** AUC of the pretrained visual encoder across 16 region-specific disease categories.

Disease	AUC
Emphysema	0.755
Atelectasis	0.653
Lung nodule	0.615
Lung opacity	0.715
Pulmonary fibrotic sequela	0.648
Pleural effusion	0.815
Mosaic attenuation pattern	0.853
Peribronchial thickening	0.755
Consolidation	0.874
Bronchiectasis	0.766
Interlobular septal thickening	0.721
Cardiomegaly	0.942
Pericardial effusion	0.910
Coronary artery wall calcification	0.889
Hiatal hernia	0.804
Arterial wall calcification	0.916
Mean AUC	**0.789**

soft assignment maps, which aggregate these features into M region-level visual tokens for organ o:

$$A_o = \mathrm{softmax}_{N_o}\big(f_o(X_o)\big), \qquad Z_o = A_o^\top X_o, \tag{5}$$

where the softmax is applied over the N_o spatial tokens independently for each of the M slots, yielding $Z_o \in \mathbb{R}^{M\times d}$. A learned organ-specific query $\mathbf{u}_o$ further pools Z_o into a single embedding for contrastive alignment with the corresponding organ description extracted from the report. Only organs fully contained within the input volume are included in the contrastive objective. The pretraining loss is

$$\mathcal{L}_{\mathrm{pre}} = \sum_o \mathcal{L}_o^{\mathrm{ITC}} + \mathcal{L}_{\mathrm{seg}}, \tag{6}$$

where $\mathcal{L}_o^{\mathrm{ITC}}$ denotes the region-level image-text contrastive loss for organ o (Shui et al., 2025), and $\mathcal{L}_{\mathrm{seg}}$ is the Dice segmentation loss.

We pretrain a separate visual encoder for each dataset for 10 epochs, using a cosine learning-rate schedule with a peak learning rate of 10^{-4}, weight decay of 0.05, and a learnable contrastive temperature initialized to 0.07. The resulting encoders achieve mean AUCs of 0.789 on CT-RATE and 0.824 on Merlin; per-disease results are reported in Tables 5 and 6. For downstream report generation, we retain the unpooled region-level visual tokens Z_o for each organ, rather than the pooled contrastive embeddings.

C.2 Stage II: Report Generation Training

After visual pretraining, the encoder is frozen and a learnable projection layer maps the visual tokens into the embedding space of the language model. Under direct report supervision, the model is trained to generate the reference report directly from the CT input. Diagnostic decomposition instead trains the model autoregressively on a target sequence consisting of a diagnostic assessment of positive findings followed by the report. For each target region, regional textual decomposition generates a diagnostic assessment before the corresponding report description. All settings are optimized using autoregressive negative log-likelihood over their respective target sequences. At inference, Report Adaptation generates the report directly. Settings with diagnostic decomposition first generate diagnostic text and then generate the report conditioned on both the CT input and that text.

Report Adaptation, diagnostic decomposition, and regional textual decomposition condition on visual tokens from all target organs. Regional visual decomposition instead restricts each regional task to the 32 region-level visual tokens associated with its requested organ. No external diagnostic

Table 6: **Zero-shot disease AUC on Merlin.** AUC of the pretrained visual encoder across 30 disease categories.

Disease	AUC	Disease	AUC
Pleural effusion	0.971	Renal cyst	0.832
Abdominal aortic aneurysm	0.968	Appendicitis	0.829
Bowel obstruction	0.955	Submucosal edema	0.800
Splenomegaly	0.948	Hepatomegaly	0.783
Surgically absent gallbladder	0.939	Anasarca	0.776
Cardiomegaly	0.938	Ascites	0.767
Hydronephrosis	0.938	Metastatic disease	0.756
Atherosclerosis	0.898	Free air	0.719
Aortic valve calcification	0.894	Fracture	0.715
Hiatal hernia	0.887	Atelectasis	0.708
Pancreatic atrophy	0.877	Renal hypodensities	0.701
Biliary ductal dilation	0.863	Osteopenia	0.695
Hepatic steatosis	0.861	Thrombosis	0.686
Gallstones	0.856	Lymphadenopathy	0.679
Coronary calcification	0.846	Prostatomegaly	0.643
		Mean AUC	**0.824**

Prompt for Region-level Report Generation

Below is the CT image of the organ: <volume>
Target organ: aorta

Identify which diseases are clearly present in this organ. List only the diseases that are present (do not list absent diseases). Then write a concise, coherent organ-level findings report consistent with your assessment. If no diseases are present, the assessment should be an empty array.

Respond in two sections:
<assessment>
["<disease>", ...] // only present disease names as a JSON array, [] if none
</assessment>
<report>
The organ-level findings report as natural radiology report prose.
</report>

Figure 6: Prompt for region-level report generation.

labels are provided at inference. Regional tasks are executed independently with the same model, and the resulting regional descriptions are concatenated to form the complete report. Intermediate assessment text is excluded from report evaluation.

C.3 Training and Inference Prompts

Figures 7 and 6 present the CT-level and region-level generation prompts, respectively. Each setting uses the same instruction during training and inference, requesting a complete report or a diagnostic assessment followed by a whole-volume or region-specific description. Regional visual decomposition uses the same prompts as regional textual decomposition; only the visual tokens supplied to each task differ. Assessments summarize positive findings; disease-specific yes/no questions are reserved for diagnostic VQA evaluation. Prompts for report decomposition and disease extraction are provided in Appendix B.1.

Prompt for CT-level Report Generation

Below is the CT scan: <volume>
Identify which diseases are clearly present in this CT scan. List only the diseases that are present (do not list absent diseases). Then write a concise, coherent radiology findings report consistent with your assessment.
If no diseases are present, the assessment should be an empty array.

Respond in two sections:
<assessment>
["<disease>", ...] // only present disease names as a JSON array, [] if none
</assessment>
<report>
The findings report as natural radiology report prose.
</report>

Figure 7: Prompt for CT-level report generation.

D Clinical-Efficacy Evaluation

D.1 Merlin Report Label Extractor

Model. We use ModernBERT-base (149M parameters) as the report-level multi-label classifier. Each report is encoded by ModernBERT, and the `[CLS]` representation is passed through a dropout layer followed by a linear classification head that produces 30 independent logits. The model is optimized with binary cross-entropy loss.

Training. Reports are tokenized and truncated to a maximum length of 1024 tokens. To address severe class imbalance, we apply a disease-specific positive-class weight, $w_i = \mathrm{clip}(N_i^{\mathrm{neg}}/N_i^{\mathrm{pos}}, 1, 50)$, in the binary cross-entropy loss, where N_i^{pos} and N_i^{neg} denote the numbers of positive and negative training examples for disease i, respectively. We also use **sentence-shuffle augmentation**, randomly permuting sentences within each report during training to reduce reliance on sentence order and encourage robustness to generated reports, which may present findings in a different order from reference reports. We train the model using AdamW with a learning rate of 2×10^{-4}, weight decay of 0.1, a dropout rate of 0.1, and a batch size of 32. The learning rate follows a cosine annealing schedule with warm restarts. We use validation macro-F1 for early stopping and select the checkpoint with the highest score.

Results. On the held-out Merlin test set, the resulting extractor achieves a macro AUC of 0.988 and a macro-F1 of 0.953 across the 30 disease categories (Table 7), supporting its use as the report-label extractor for downstream CE evaluation.

E Additional Results and Analysis

E.1 Micro-Averaged Clinical-Efficacy Results

The main paper reports macro-averaged metrics to give equal weight to each disease category despite class imbalance. As a complementary view, Table 8 reports micro-averaged F1, precision, and recall for all evaluated settings on CT-RATE.

Per-disease diagnostic performance. Table 9 breaks down the performance of full DeCAD across the 18 CT-RATE disease categories. The macro-F1 is 52.1, with substantial variation across diseases. Pleural effusion achieves the highest F1 (81.2), followed by arterial wall calcification (78.7) and coronary artery wall calcification (74.7). The precision-recall profiles also differ: hiatal hernia has relatively high recall (77.6) but low precision (31.4), whereas medical material has high precision (73.4) but low recall (15.2). These results show that the aggregate score does not fully capture category-specific errors and identify diseases for which missed findings or false positives remain prominent.

Table 7: **Report-label extraction performance on Merlin.** Per-disease AUC and F1 of the report-label extractor on the held-out test set across 30 disease categories.

Disease	AUC	F1	Disease	AUC	F1
Atelectasis	1.000	1.000	Metastatic disease	0.957	0.833
Coronary calcification	1.000	1.000	Osteopenia	0.994	0.883
Hepatic steatosis	1.000	1.000	Fracture	0.977	0.907
Pleural effusion	1.000	0.999	Abdominal aortic aneurysm	0.994	0.909
Surgically absent gallbladder	1.000	0.999	Prostatomegaly	0.979	0.896
Gallstones	1.000	0.997	Thrombosis	0.861	0.896
Renal cyst	1.000	0.996	Anasarca	0.985	0.892
Renal hypodensities	1.000	0.996	Lymphadenopathy	0.990	0.851
Splenomegaly	1.000	0.996	Bowel obstruction	0.998	0.918
Cardiomegaly	1.000	0.992	Free air	0.975	0.867
Hepatomegaly	1.000	0.991	Submucosal edema	0.993	0.968
Biliary ductal dilation	0.983	0.990	Hydronephrosis	0.996	0.959
Ascites	0.999	0.989	Hiatal hernia	0.999	0.975
Pancreatic atrophy	0.999	0.987	Appendicitis	0.996	0.941
Atherosclerosis	0.981	0.989	Aortic valve calcification	0.997	0.972
			Macro average	**0.988**	**0.953**

Table 8: Micro-averaged diagnostic performance on CT-RATE (%), computed using RadBERT. Within each model and metric, the best report-generation score is shaded purple and bold; the second-best distinct score is shaded blue and underlined. Diagnostic VQA is excluded from ranking.

	CT-CHAT			MedGemma-1.5-4B			Lingshu-7B		
Setting	F1	P	R	F1	P	R	F1	P	R
Diagnostic VQA	31.9	19.8	82.2	30.5	31.1	29.9	30.1	20.7	54.5
Direct	30.1	38.1	24.9	15.9	35.3	10.2	11.4	39.5	6.6
Report Adapt.	28.1	**44.3**	20.6	38.3	51.3	30.5	31.8	55.7	22.3
Diagnostic Decomp.	39.8	42.2	37.7	44.9	55.9	37.5	41.3	55.1	33.0
Diag.+Text Decomp.	**45.7**	43.2	48.6	44.4	**56.1**	36.8	44.3	61.5	34.6
Diag.+Text+Visual	44.5	40.3	**49.6**	**47.1**	55.5	**40.9**	**45.1**	**61.8**	**35.5**

	HuluMed-7B			RadSight-8B		
Setting	F1	P	R	F1	P	R
Diagnostic VQA	47.6	32.5	88.6	52.0	44.4	62.9
Direct	27.7	58.2	18.2	42.0	49.6	36.3
Report Adapt.	48.5	59.1	41.2	43.8	**50.6**	38.6
Diagnostic Decomp.	54.9	62.0	49.3	47.7	50.5	45.2
Diag.+Text Decomp.	**57.7**	**62.7**	**53.5**	51.0	48.7	53.6
Diag.+Text+Visual	56.6	60.6	53.0	**52.2**	49.1	**55.7**

E.2 Coverage of Diagnostic Supervision

How much diagnostic supervision is needed? The preceding results show that diagnostic decomposition improves report generation, but leave open how broadly disease-specific diagnostic supervision must be applied during training. We examine this question using the whole-image baseline and four assessment-then-report settings. The coarsest assessment states only whether the CT is normal or abnormal. For disease-specific diagnostic supervision, we randomly select approximately one-third, two-thirds, or all training cases to receive a full positive-finding assessment before report generation. At inference, the model generates its own assessment before generating the report.

Table 10 shows that a binary normal/abnormal assessment does not improve on the whole-image baseline: micro-F1 is 29.6% versus 29.7%, and macro-F1 is 13.9% versus 15.9%. Compressing

Table 9: Per-disease diagnostic performance across 18 disease categories on CT-RATE, evaluated using the RadBERT-based report classifier. Diseases are ordered by F1. Scores are reported as percentages.

Disease	F1	P	R	Disease	F1	P	R
Pleural effusion	81.2	80.2	82.2	Emphysema	50.6	45.9	56.4
Arterial wall calcification	78.7	72.1	86.6	Hiatal hernia	44.7	31.4	77.6
Coronary artery calcification	74.7	67.3	83.9	Mosaic attenuation	41.2	45.8	37.4
Lung opacity	70.8	78.4	64.6	Pulmonary fibrotic sequela	38.5	51.4	30.8
Consolidation	66.3	65.1	67.5	Lymphadenopathy	37.3	53.4	28.7
Cardiomegaly	63.1	55.6	72.8	Peribronchial thickening	34.5	45.3	27.9
Lung nodule	62.8	56.3	71.1	Interlobular septal thickening	33.6	42.8	27.6
Pericardial effusion	53.6	49.0	59.0	Bronchiectasis	29.4	43.9	22.1
Atelectasis	51.5	42.8	64.6	Medical material	25.2	73.4	15.2
Micro average	**57.6**	**56.4**	**58.7**	**Macro average**	**52.1**	**55.6**	**54.2**

Table 10: Effect of diagnostic supervision coverage on whole-image report generation. P, R, and F1 are micro- or macro-averaged diagnostic metrics, reported as percentages. Disease-specific Diagnose supervision is applied to randomly selected subsets of the training cases.

	Micro			Macro		
Setting	P	R	F1	P	R	F1
Whole-image baseline	44.9	22.2	29.7	37.6	15.1	15.9
Normal/abnormal assessment	44.8	22.1	29.6	24.4	13.9	13.9
Diagnose on 1/3 of cases	53.7	31.3	39.6	43.6	22.3	25.1
Diagnose on 2/3 of cases	53.7	37.5	44.1	44.0	28.6	30.2
Diagnose on all cases	52.4	47.3	**49.7**	45.7	39.0	**39.0**

diverse findings into a single judgment therefore provides little disease-specific guidance for subsequent report generation. When disease-specific Diagnose supervision is introduced for one-third of the training cases, micro-precision increases from 44.9% to 53.7% and macro-precision from 37.6% to 43.6%, indicating that explicit diagnostic supervision improves the precision of the resulting reports even when applied to only a subset of the training data. As supervision coverage increases, recall rises steadily: micro-recall increases from 31.3% to 37.5% and 47.3%, while macro-recall increases from 22.3% to 28.6% and 39.0% for one-third, two-thirds, and all training cases, respectively. Precision remains broadly stable across these three settings. Together, these results suggest that disease-specific diagnostic supervision primarily improves report precision when first introduced, while broader supervision coverage progressively improves recall and reduces diagnostic omissions.

E.3 Qualitative Analysis

Figures 8 and 9 provide a qualitative comparison of Direct Report, Report Adaptation, and combined diagnostic and regional textual decomposition across five MedVLMs on the same CT study. The reference report contains six positive findings: arterial wall calcification, bronchiectasis, emphysema, hiatal hernia, lung nodule, and peribronchial thickening. Direct report generation frequently omits these abnormalities despite producing fluent radiological text. Conventional Report Adaptation recovers some findings for several models, but substantial omissions remain, and the improvement is not consistent across models. In contrast, diagnostic + textual decomposition increases the number of ground-truth findings expressed in the final report for all five MedVLMs. The recovered abnormalities span multiple anatomical regions and include pulmonary nodules, emphysematous and bronchiectatic changes, arterial calcification, hiatal hernia, and peribronchial thickening. These examples complement the quantitative results by illustrating the diagnosis-to-report gap at the level of individual findings: diagnostically relevant abnormalities can be omitted during free-form report composition, whereas explicitly decomposing diagnosis and description helps express more diagnostically relevant findings in the generated report.

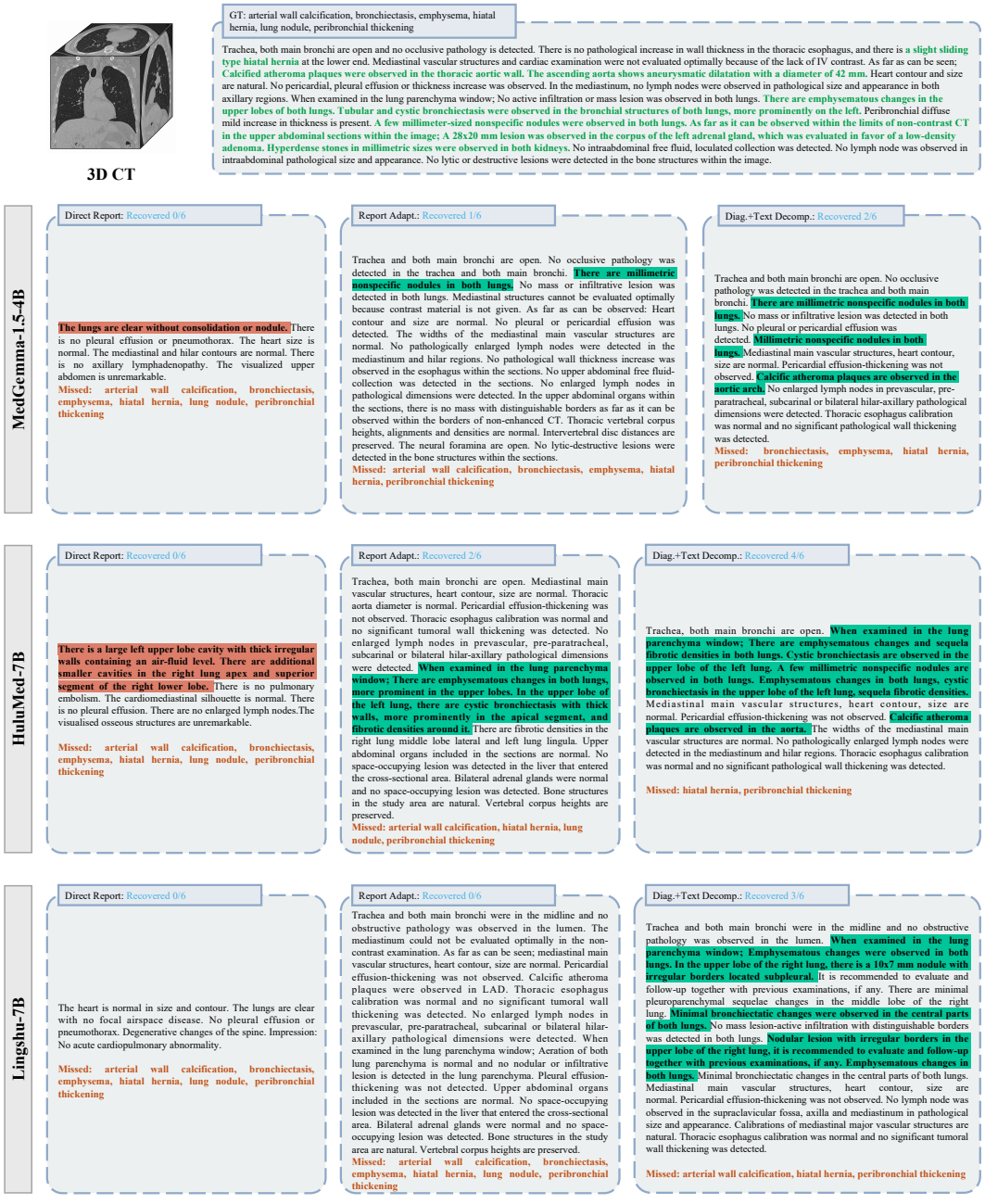

Figure 8: **Qualitative comparison on CT-RATE across MedGemma-1.5-4B, HuluMed-7B, and Lingshu-7B.** For the same CT study with six ground-truth positive findings, we compare Direct Report, Report Adaptation, and diagnostic + textual decomposition. Green highlights mark generated descriptions corresponding to ground-truth findings, orange text lists ground-truth findings omitted from each report, and red highlights indicate additional findings not described in the reference report. The header of each panel reports the number of ground-truth findings recovered. Across all three models, diagnostic + textual decomposition recovers more reference findings than the corresponding direct and adapted reports.



Figure 9: **Qualitative comparison on CT-RATE across RadSight-8B and CT-CHAT-8B.** We compare Direct Report, Report Adaptation, and diagnostic + textual decomposition on the same CT study with six ground-truth positive findings. Green highlights mark generated descriptions corresponding to ground-truth findings, orange text lists ground-truth findings omitted from each report, and red highlights indicate additional findings not described in the reference report. The header of each panel reports the number of ground-truth findings recovered. Diagnostic + textual decomposition increases the recovery of reference findings for both models, while conventional report adaptation does not consistently reduce diagnostic omissions.