ANATOBRIDGE: STRENGTHENING AND TRANSFERRING ANATOMICAL REPRESENTATIONS FOR 3D CHEST CT VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models remain challenged in understanding 3D chest computed tomography (CT), where clinically relevant evidence is volumetric, spatially localized, and often subtle. Existing models inadequately exploit the anatomical structure needed to ground such evidence: i) anatomical guidance is weakly integrated into visual encoding, ii) lesion supervision remains underused in representation learning, and iii) explicit anatomical outputs are not directly exposed to the LLM. In this paper, we present AnatoBridge, an anatomy-aware 3D chest CT VLM that incorporates anatomical information throughout visual encoding, representation learning, and language alignment to address these limitations. AnatoBridge strengthens the visual representations by integrating region-constrained anatomy queries throughout the transformer and using lesion-guided pretraining to enrich organ features and construct directly supervised disease-specific representations. Moreover, a dual-channel interface carries this information into language modeling to provide the LLM with complementary fine-grained visual information and interpretable anatomical evidence, where the feature channel projects the learned representations into continuous anatomy tokens and the prompt channel serializes explicit anatomical outputs as disease findings, organ geometry, and lesion grounding. Experimental results show that AnatoBridge outperforms the strongest baseline by 1.1 in AUC (85.5 vs. 84.4) in zero-shot classification on CT-RATE and yields a 0.7 AUC gain (76.1 vs. 75.4) on the external Rad-ChestCT cohort. Remarkably, it also exceeds RadSight-8B by an average 3.33 points across the three 3D-RAD tasks, by 1.80 points on CT-Spatial-VQA, and by 6.23 points in report-generation clinical efficacy F1.

1 INTRODUCTION

Multimodal large language models (MLLMs) have achieved strong results in general vision and 2D medical imaging (Li et al., 2023; Bai et al., 2025; Singh et al., 2025; Sellergren et al., 2025), yet comparatively few models directly process 3D chest computed tomography (CT). Extending these models to CT requires more than treating slices as additional image frames. A chest CT is a volumetric examination whose diagnostic content includes millimetre-scale structures—such as small nodules, airway walls, and interlobular septa—that must be interpreted jointly across neighboring slices and in relation to surrounding anatomy, while respecting spatial continuity and anisotropic resolution. Sparse slice sampling and 2D video-style adaptations (Xu et al., 2026; Lin et al., 2026) may therefore obscure or discard the fine-grained evidence needed for chest CT reasoning.

In parallel, dedicated 3D CT encoders have improved volumetric representations through CT–report pretraining and auxiliary objectives such as organ segmentation and disease classification (Shui et al., 2025; Zhang et al., 2026; Shi et al., 2026). Building on this progress, recent systems connect these encoders to LLMs for report generation and visual question answering (VQA) (Liu et al., 2026; Yuan et al., 2026). Yet important gaps remain in integrating anatomical knowledge across visual learning and language reasoning. Specifically, anatomical guidance is weakly integrated into visual encoding, while fine-grained lesion supervision remains underused in representation learning. At the language interface, existing systems typically pass only projected volumetric features to the LLM, leaving explicit organ-, disease-, and lesion-level predictions unavailable as structured evidence. We

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.

address these limitations by strengthening anatomical representations throughout visual learning and transferring continuous representations and explicit anatomical outputs to the LLM.

To this end, we present AnatoBridge, an anatomy-aware 3D chest CT VLM that incorporates anatomical information throughout visual encoding, representation learning, and language alignment, as illustrated in Figure 1. Organ and global queries are integrated into the visual transformer under anatomy-constrained attention, allowing region-specific evidence to accumulate across layers. Voxel-level lesion predictions are subsequently pooled into lesion-aware tokens that enrich organ representations and construct disease-specific features under a dedicated contrastive objective. To preserve this enhanced anatomical information during language alignment, a dual-channel interface conditions the LLM on continuous anatomy tokens projected from the learned representations together with structured descriptions of disease findings, organ geometry, and lesion grounding. Their combination provides fine-grained visual context and interpretable anatomical evidence for clinical question answering.

We evaluate both the pretrained encoder and the resulting VLM. On CT-RATE (Hamamci et al., 2024) and the external Rad-ChestCT (Draelos et al., 2021) cohort, AnatoBridge improves zero-shot disease classification, including a 4.8-point gain in lung-nodule AUC over CT-DiagVLM (Shi et al., 2026). On downstream 3D understanding, AnatoBridge-4B outperforms the larger RadSight-8B (Liu et al., 2026) across 3D-RAD (Gai et al., 2026), CT-Spatial-VQA (Monon et al., 2026), and CT-RATE report generation. On DeepChestVQA (Wang et al., 2026), it also improves recognition and medical reasoning over the strongest reported agent-based baseline in a single forward pass. Our contributions are:

- We propose an anatomy-constrained visual encoder that strengthens organ-specific representations across the full transformer depth;
- We introduce lesion-guided hierarchical representation learning that enriches organ features and constructs directly supervised disease-specific representations;
- We design a dual-channel anatomy-to-language interface that transfers enhanced continuous representations alongside configurable textual descriptions of explicit anatomical outputs.

2 Related Work

Vision–language pretraining for 3D CT with anatomical supervision. CT–report corpora have enabled volumetric vision–language models for zero-shot classification, retrieval, and report generation. CT-CLIP (Hamamci et al., 2024) establishes global contrastive pretraining on chest CT, Merlin (Blankemeier et al., 2026) adds structured clinical supervision, HLIP (Zhao et al., 2026) introduces hierarchical attention over the slice–scan–study structure for scalable pretraining on unfiltered data, and COLIPRI (Wald et al., 2026) blends paired and unpaired volumetric data under a summary-drafting objective. Region–sentence alignment, first studied in 2D chest X-rays (Huang et al., 2021; Müller et al., 2022), extends to 3D via fVLM's per-organ alignment (Shui et al., 2025) and ViSD-Boost's anatomy normality modeling (Cao et al., 2025). CT-DiagVLM (Shi et al., 2026) adds disease-level contrastive learning with query-based semantic extraction, and RADAR (Zhang et al., 2026) scales this to abdominal CT across 146 findings. Anatomical supervision from TotalSegmentator (Wasserthal et al., 2023) and ReXGroundingCT (Baharoon et al., 2026) has become central to fine-grained CT representation learning, yet anatomical signals remain encoded inside the visual stream rather than transferred as explicit evidence to language reasoning.

Multimodal language models for 3D medical imaging. General-purpose MLLMs (Bai et al., 2025; Wang et al., 2025; Singh et al., 2025) and 2D medical adaptations (Li et al., 2023; Moor et al., 2023; Chen et al., 2024; Sellergren et al., 2025; 2026) treat CT as 2D slices. Recent medical MLLMs target 3D data: sparse sampling (Hulu-Med (Jiang et al., 2025), Lingshu (Xu et al., 2026)) discards inter-slice findings; OmniCT treats slices as video frames (Lin et al., 2026); native volumetric models (M3D (Bai et al., 2024), RadFM (Wu et al., 2025), Photon (Fang et al., 2026)) train 3D encoders end-to-end; ClinFusion (Yuan et al., 2026) combines 3D and cascaded 2D encoders under spatial-aware fusion, following compositional multi-encoder designs (Tong et al., 2024); RadSight (Liu et al., 2026) uses dual 2D/3D architecture with perception-oriented tuning. Despite these advances, the encoder–LLM interface remains dominated by projected visual tokens, with anatomical predictions unavailable as evidence for downstream reasoning.

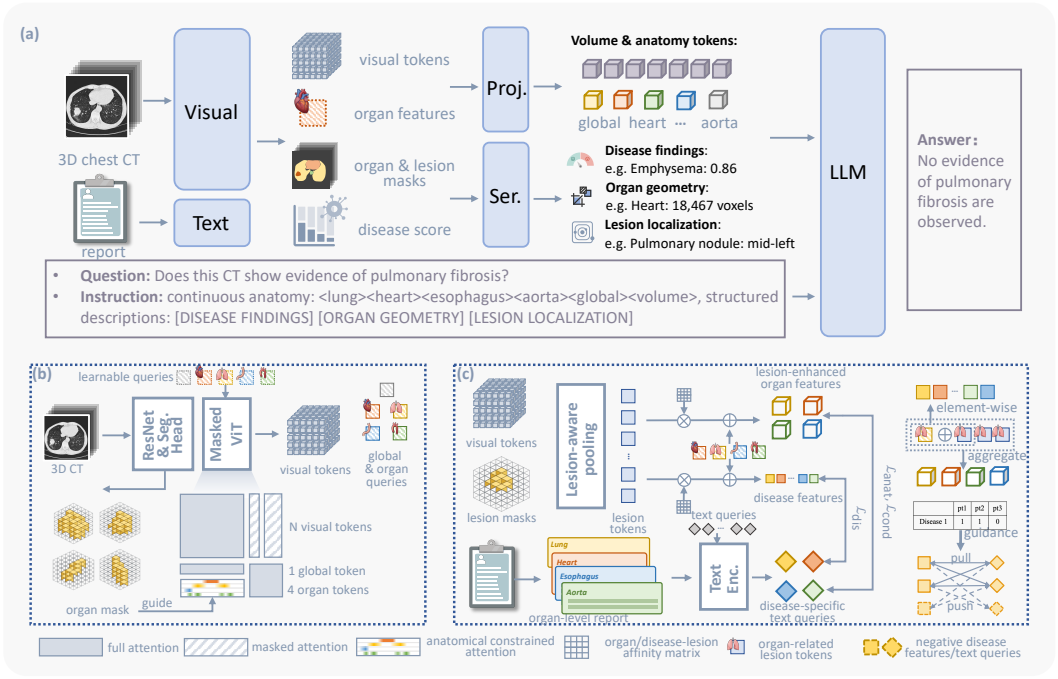


Figure 1: Overview of AnatoBridge. (a) Overall framework and dual-channel language alignment, which transfer continuous visual representations and serialized anatomical predictions to the LLM. Proj. and Ser. denote projection and serialization, respectively. (b) Anatomy-aware visual encoding with anatomy-constrained attention, where organ queries attend to their corresponding predicted regions. (c) Lesion-guided representation learning, which constructs lesion-enhanced organ features (f_o^{organ}) and disease-specific features (f_d^{disease}) from lesion-aware tokens. The right insets illustrate their hierarchical construction and the disease-specific contrastive objective ($\mathcal{L}_{\text{dis}}$).

3 Methodology

3.1 Overall Framework

Given a 3D chest CT volume $\boldsymbol{V} \in \mathbb{R}^{D\times H\times W\times C}$ and a clinical instruction $\boldsymbol{q}$, AnatoBridge follows a three-stage path from volumetric encoding to response generation; see Figure 1. First, anatomy-aware visual encoding maps $\boldsymbol{V}$ to dense visual tokens and organ- and global-query features, with predicted organ regions constraining how the organ queries aggregate evidence (Sec. 3.2). Second, lesion-guided representation learning pools the predicted lesion maps over the visual tokens and uses the resulting lesion-aware tokens to refine the organ representations and construct disease-specific features (Sec. 3.3). We denote these two pretrained visual stages jointly by Φ_θ:

$$\Phi_\theta(\boldsymbol{V}) = \left(\underbrace{\boldsymbol{Z}, \boldsymbol{F}^{\text{organ}}, \boldsymbol{f}^{\text{global}}}_{\text{continuous representations}}, \underbrace{\boldsymbol{s}^{\text{disease}}, \boldsymbol{P}^{\text{organ}}, \boldsymbol{Y}^{\text{lesion}}}_{\text{explicit anatomical outputs}} \right), \tag{1}$$

where $\boldsymbol{Z}$ denotes the dense visual tokens; $\boldsymbol{F}^{\text{organ}}$ and $\boldsymbol{f}^{\text{global}}$ are the lesion-enhanced organ features and the global feature; and $\boldsymbol{s}^{\text{disease}}$, $\boldsymbol{P}^{\text{organ}}$, and $\boldsymbol{Y}^{\text{lesion}}$ denote the disease scores, organ predictions, and lesion logits. The first group retains fine-grained visual information in continuous representations, whereas the second provides explicit anatomical and spatial evidence.

Third, dual-channel language alignment transforms these complementary outputs into context for the LLM (Sec. 3.4):

$$\boldsymbol{T}^{\text{anat}} = \Pi\big(\boldsymbol{F}^{\text{organ}}, \boldsymbol{f}^{\text{global}}\big), \qquad (P_D, P_O, P_L) = \mathcal{S}\big(\boldsymbol{s}^{\text{disease}}, \boldsymbol{P}^{\text{organ}}, \boldsymbol{Y}^{\text{lesion}}\big), \tag{2}$$

where Π projects the continuous features into anatomy tokens, while $\mathcal{S}$ serializes disease findings, organ geometry, and lesion grounding into task-configurable prompt blocks. The anatomy tokens

and structured prompts are combined with the volume tokens derived from $\boldsymbol{Z}$ and the instruction $\boldsymbol{q}$ to form the LLM input $\boldsymbol{X}$ in Eq. 15, from which the response $\boldsymbol{a} = (a_1, \ldots, a_T)$ is generated. Training follows the same flow: multi-task encoder pretraining learns the representations and predictions in Eq. 1, after which instruction tuning connects them to the LLM and optimizes the autoregressive objective in Eq. 16.

3.2 Anatomy-aware visual encoding

Our visual encoder builds on the CNN–ViT hybrid backbone of CT-DiagVLM (Shi et al., 2026), which combines a U-Net decoder for four-organ segmentation with post-hoc organ queries. Such queries extract regional features only after encoding completes, leaving the visual encoding anatomically uninformed. We instead integrate organ structure throughout the transformer, letting anatomical priors guide encoding from the earliest layers. Specifically, organ queries are updated across all transformer layers under region-constrained attention.

Anatomy queries inside the transformer. We prepend K learnable anatomy queries $\boldsymbol{Q} \in \mathbb{R}^{K\times C}$ to the visual-token sequence before the transformer. The backbone features are projected and flattened into N visual tokens $\boldsymbol{Z} \in \mathbb{R}^{B\times N\times C}$ on a $D_h \times D_w \times D_d$ grid, giving a total sequence length of $N + K$. The queries correspond to the four organs $\mathcal{O} = \{\text{lung}, \text{heart}, \text{esophagus}, \text{aorta}\}$ and one global query, giving $K = 5$. After the transformer, we obtain encoded organ query features $\boldsymbol{Q}^{\text{enc}} \in \mathbb{R}^{B\times 4\times C}$, with row vectors $\boldsymbol{q}_o^{\text{enc}}$, and an encoded global query $\boldsymbol{q}_{\text{g}}^{\text{enc}} \in \mathbb{R}^{B\times C}$.

Anatomy-constrained attention mask. Unconstrained organ queries may aggregate evidence from unrelated anatomical regions and alter the shared visual-token stream. We therefore apply an additive attention mask $\boldsymbol{M}_b \in \{0, -\infty\}^{(N+K)\times(N+K)}$ in every transformer block. Visual tokens retain unrestricted mutual attention but cannot attend to the anatomy queries. Conversely, each organ query attends only to visual tokens inside its down-sampled organ mask, while the global query attends to the full volume. Formally, let $\boldsymbol{S} \in \{0, 1\}^{B\times 4\times N}$ be the organ indicators down-sampled to the token grid. For batch element b, the mask is defined by

$$M_{b,v_i,x} = \begin{cases} 0, & x = v_j, \\ -\infty, & x = q_k, \end{cases} \tag{3}$$

$$M_{b,q_o,v_i} = \begin{cases} 0, & S_{b,o,i} = 1, \\ -\infty, & \text{otherwise}, \end{cases} \tag{4}$$

and

$$M_{b,q_{\text{g}},v_i} = 0, \qquad M_{b,q_k,q_l} = 0. \tag{5}$$

Here v_i, v_j index visual tokens, q_o denotes an organ query, and q_k, q_l denote any anatomy queries, including the global query q_{g}. Eq. 3–5 keep the visual stream independent of the added queries, constrain each organ query to its predicted region, and leave the global and query-to-query connections unmasked. The latter provides a valid attention path for empty organ masks and retains cross-region context.

The encoder outputs transformed dense visual tokens $\boldsymbol{Z}$, encoded organ query features $\boldsymbol{Q}^{\text{enc}}$, an encoded global query $\boldsymbol{q}_{\text{g}}^{\text{enc}}$, and voxel-wise organ predictions. The next stage further enriches the organ representations with lesion-specific evidence.

3.3 Lesion-guided representation learning

The region-constrained queries in Sec. 3.2 strengthen organ specificity, but an organ-level representation may still mix multiple coexisting abnormalities. We therefore use voxel-level lesion supervision to introduce a second, finer level of anatomical structure.

Voxel-level anatomical supervision. Alongside the inherited organ head, which predicts $\boldsymbol{P}^{\text{organ}} \in [0, 1]^{B\times 5\times D\times H\times W}$ for four organs and background, we add a lesion head that predicts a 14-class logit map $\boldsymbol{Y}^{\text{lesion}} \in \mathbb{R}^{B\times 14\times D\times H\times W}$. The organ loss is applied to all pretraining scans, whereas the lesion loss is evaluated only on the subset with ReXGroundingCT (Baharoon et al., 2026) annotations:

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{Dice}}\big(\boldsymbol{P}^{\text{organ}}, \boldsymbol{M}^{\text{organ}}\big), \quad \mathcal{L}_{\text{lesion}} = \mathcal{L}_{\text{Dice}}\big(\boldsymbol{Y}^{\text{lesion}}, \boldsymbol{M}^{\text{lesion}}\big). \tag{6}$$

The predicted lesion maps provide spatial weights for constructing lesion-aware tokens, which subsequently refine the organ- and disease-level representations. The same maps later provide explicit spatial evidence for the lesion-grounding prompt in Sec. 3.4.

Lesion tokens. Because multiple abnormalities may coexist within the same organ, a single organ feature may not preserve lesion-specific evidence. We therefore use the predicted lesion probabilities to pool the visual-token grid into lesion-aware tokens:

$$\boldsymbol{t}_\ell = \frac{\sum_{i=1}^{N} s_{\ell,i} \boldsymbol{z}_i}{\sum_{i=1}^{N} s_{\ell,i} + \epsilon}, \tag{7}$$

where $\boldsymbol{s}_\ell \in [0,1]^N$ is the down-sampled probability of lesion class ℓ over the N visual tokens and ϵ is a small constant.

Hierarchical organ and disease representations. Let $\mathcal{L} = \{1, \ldots, 14\}$ denote the ReXGroundingCT lesion categories and $\mathcal{D}$ the 16 organ-associated CT-RATE diseases (Hamamci et al., 2024). Two fixed affinity matrices organize these tokens: $\boldsymbol{A}^{\text{organ}} \in \mathbb{R}^{4\times 14}$ aggregates lesion evidence by organ, and the sparse binary matrix $\boldsymbol{A}^{\text{disease}} \in \{0,1\}^{|\mathcal{D}|\times 14}$ selects the evidence associated with each disease. Appendix C details the complete cross-taxonomy mapping and its construction.

$$\boldsymbol{f}_o^{\text{organ}} = \boldsymbol{q}_o^{\text{enc}} + \boldsymbol{W}_o^{\text{lesion}} \Big(\sum_{\ell=1}^{14} \boldsymbol{A}_{o,\ell}^{\text{organ}} \boldsymbol{t}_\ell \Big), \tag{8}$$

$$\boldsymbol{f}_d^{\text{disease}} = \boldsymbol{q}_{o(d)}^{\text{enc}} + \boldsymbol{W}_d^{\text{dis}} \Big(\sum_{\ell=1}^{14} \boldsymbol{A}_{d,\ell}^{\text{disease}} \boldsymbol{t}_\ell \Big), \tag{9}$$

where $\boldsymbol{W}_o^{\text{lesion}}$ and $\boldsymbol{W}_d^{\text{dis}}$ are zero-initialized projections. The organ branch yields $\boldsymbol{F}^{\text{organ}} \in \mathbb{R}^{B\times 4\times C}$ by adding organ-associated lesion evidence to the corresponding encoded organ query through a residual connection. The resulting representation preserves organ-specific context while incorporating finer lesion-level evidence. The disease branch constructs a separate representation for each disease, enabling disease-specific supervision when multiple findings are associated with the same organ. If a disease has no lesion counterpart, its affinity row is zero and the representation reduces to the corresponding encoded organ query. The global feature bypasses lesion fusion, so $\boldsymbol{f}^{\text{global}} = \boldsymbol{q}_{\text{g}}^{\text{enc}}$.

Inherited organ-level and new disease-level objectives. We retain CT-DiagVLM's organ-level contrastive losses for organ-text and organ-disease alignment, $\mathcal{L}_{\text{anat}}$ and $\mathcal{L}_{\text{cond}}$ (Shi et al., 2026), but apply them to the lesion-enhanced features $\boldsymbol{F}^{\text{organ}}$ instead of the original organ-query representations. Thus, lesion supervision strengthens the regional features optimized by the inherited objectives, rather than acting only as an auxiliary segmentation signal. We additionally introduce a disease-specific contrastive loss that directly supervises $\boldsymbol{f}_d^{\text{disease}}$ and pulls same-disease patients together:

$$\mathcal{L}_{\text{I2T}}^{(d)} = -\frac{1}{|\mathcal{P}_d|^2} \sum_{p\in\mathcal{P}_d} \sum_{p'\in\mathcal{P}_d} \log \frac{\exp\big((\boldsymbol{f}_{p,d}^{\text{disease}})^\top \boldsymbol{q}_{p'}^{(d)}/\tau\big)}{\sum_r \exp\big((\boldsymbol{f}_{p,d}^{\text{disease}})^\top \boldsymbol{q}_r^{(d)}/\tau\big)}, \tag{10}$$

with $\mathcal{P}_d$ the patients positive for disease d, $\boldsymbol{f}_{p,d}^{\text{disease}}$ the disease-d visual feature of patient p, and $\boldsymbol{q}_p^{(d)}$ the corresponding positive text query; the text-to-image term is defined symmetrically with label weighting.

The full disease-specific loss averages both directions over the 16 organ-associated diseases:

$$\mathcal{L}_{\text{dis}} = \frac{1}{|\mathcal{D}|} \sum_{d\in\mathcal{D}} \frac{1}{2} \big(\mathcal{L}_{\text{I2T}}^{(d)} + \mathcal{L}_{\text{T2I}}^{(d)} \big). \tag{11}$$

The alignment loss combines the three terms:

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{anat}} + \mathcal{L}_{\text{cond}} + \mathcal{L}_{\text{dis}}. \tag{12}$$

Full pretraining objective. The encoder pretraining objective is

$$\mathcal{L}_{\text{pretrain}} = \lambda_{\text{seg}} \mathcal{L}_{\text{seg}} + \lambda_{\text{lesion}} \mathcal{L}_{\text{lesion}} + \lambda_{\text{align}} \mathcal{L}_{\text{align}}. \tag{13}$$

3.4 DUAL-CHANNEL LANGUAGE ALIGNMENT

The feature and prompt channels transfer complementary encoder outputs to the LLM in parallel: continuous features retain information beyond a fixed label vocabulary, while structured prompts expose explicit anatomical evidence.

Feature channel. We stack the four organ features and the global feature in a fixed anatomical order and project each feature into the LLM embedding space with a two-layer GELU MLP:

$$\boldsymbol{T}^{\text{anat}} = \text{MLP}_{\text{proj}}\big(\text{Concat}[\boldsymbol{F}^{\text{organ}}, \boldsymbol{f}^{\text{global}}]\big) = \big[\boldsymbol{t}^{\text{anat}}_{\text{lung}}; \boldsymbol{t}^{\text{anat}}_{\text{heart}}; \boldsymbol{t}^{\text{anat}}_{\text{esophagus}}; \boldsymbol{t}^{\text{anat}}_{\text{aorta}}; \boldsymbol{t}^{\text{anat}}_{\text{global}}\big] \in \mathbb{R}^{B\times 5\times H_{\text{LLM}}}. \tag{14}$$

The five rows of $\boldsymbol{T}^{\text{anat}}$ therefore correspond, in order, to the lung, heart, esophagus, aorta, and global anatomy tokens in the LLM input.

Prompt channel. The explicit anatomical outputs are serialized into three independently configurable text blocks: disease findings P_D, organ geometry P_O, and lesion grounding P_L. These blocks summarize predicted disease scores, organ size and location, and the spatial distribution of predicted lesions, respectively. Appendix D and Appendix E provide detailed procedures for prompt construction and spatial description serialization, respectively.

LLM input sequence and training. Let $\boldsymbol{T}^{\text{vol}}$ denote the volume-token sequence from the 3D visual backbone, and let $\text{Emb}(\cdot)$ denote tokenization followed by the LLM embedding lookup. The complete input sequence is

$$\boldsymbol{X} = \big[\boldsymbol{T}^{\text{anat}}\ ;\ \boldsymbol{T}^{\text{vol}}\ ;\ \text{Emb}(P_D)\ ;\ \text{Emb}(P_O)\ ;\ \text{Emb}(P_L)\ ;\ \text{Emb}(\boldsymbol{q})\big], \tag{15}$$

where semicolons denote concatenation along the token dimension. Conditioned on $\boldsymbol{X}$, the model is optimized with the standard autoregressive language-modeling objective:

$$\mathcal{L}_{\text{LM}} = -\sum_{t=1}^{T} \log P\big(a_t \mid a_{<t}, \boldsymbol{X}\big). \tag{16}$$

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and baselines. We evaluate encoder representations through zero-shot classification on CT-RATE and Rad-ChestCT, and evaluate language-aligned models on 3D-RAD, CT-Spatial-VQA, DeepChestVQA, and CT-RATE report generation. CT-Spatial-VQA and DeepChestVQA are used only for evaluation: no question–answer examples from either benchmark are included in instruction tuning. For direct comparison with RadSight, our instruction-tuning corpus follows the same data composition: 3D-RAD plus the auxiliary chest CT instruction data constructed by RadSight. Appendix B details all data sources and evaluation splits. Encoder baselines follow the CT-DiagVLM protocol (Shi et al., 2026); generative settings and reported baselines follow RadSight (Liu et al., 2026) and 3DMedAgent (Wang et al., 2026).

Implementation details. We pretrain the 3D CT encoder with the objective in Eq. 13 and initialize the task-specific 3D backbone from this checkpoint. The language backbone is Qwen3-VL-4B-Instruct by default, with its 8B counterpart also evaluated in Sec. 4.3, following the model choices of RadSight. Structured prompt statistics are generated offline by a frozen copy of the pretrained 3D encoder and cached per volume. Optimization settings are summarized in Appendix A

Evaluation metrics. Following previous works (Shui et al., 2025; Shi et al., 2026), zero-shot classification evaluates the 16 organ-associated labels and excludes the two global labels. We report macro-averaged AUC, accuracy (ACC), F1, and precision on the CT-RATE internal and Rad-ChestCT external test sets. For 3D-RAD, CT-Spatial-VQA and DeepChestVQA, we report accuracy. For report generation on CT-RATE, we report RaTEScore (Zhao et al., 2024) and clinical efficacy F1 (CE F1), which measures agreement in extracted clinical findings.

Table 1: Zero-shot classification on the CT-RATE internal and Rad-ChestCT external test sets under the 16-disease protocol. All baseline results are taken from Shi et al. (2026).

Method	Internal test (CT-RATE)				External test (Rad-ChestCT)			
	AUC	ACC	F1	Prec	AUC	ACC	F1	Prec
Random (Hamamci et al., 2024)	50.5	50.2	57.0	18.0	49.6	50.0	55.5	26.5
Supervised (Hamamci et al., 2024)	60.3	58.1	63.2	24.0	54.1	53.9	58.7	28.7
CT-CLIP (Hamamci et al., 2024)	73.3	66.9	70.8	32.6	63.2	59.9	64.7	34.1
CT-VocabFine (Hamamci et al., 2024)	76.0	70.4	73.8	35.6	65.7	62.1	66.8	35.6
CT-LiPro (Hamamci et al., 2024)	76.1	69.1	72.6	34.3	64.7	60.6	65.0	35.1
BIUD (Cao et al., 2024)	71.3	68.1	71.6	33.8	62.9	60.6	65.2	33.7
Merlin (Blankemeier et al., 2026)	72.8	67.2	70.9	33.7	64.4	61.9	66.3	34.8
HLIP (Zhao et al., 2026)	77.7	71.4	74.7	37.9	72.3	68.4	72.1	40.4
COLIPRI (Wald et al., 2026)	77.8	70.2	75.2	–	67.0	60.9	65.7	–
fVLM (Shui et al., 2025)	77.8	71.8	75.1	37.9	68.0	64.7	68.8	37.4
ViSD-Boost (Cao et al., 2025)	79.0	73.1	75.9	38.7	69.4	65.2	69.3	34.2
CT-DiagVLM (Shi et al., 2026)	84.4	78.1	80.3	45.2	75.4	69.8	73.4	39.2
AnatoBridge	**85.5**	**78.8**	**81.0**	**46.3**	**76.1**	**70.2**	**73.8**	**40.3**

Table 2: Main results on 3D medical VQA and report generation. All values are accuracy or score (higher is better), and the CT-RATE columns evaluate report generation. Evaluation settings and baseline values follow Liu et al. (2026). Exist. = Existence Detection, Static = Static Temporal Diagnosis, Long. = Longitudinal Temporal Diagnosis.

Method	3D-RAD			CT-Spatial-VQA	CT-RATE	
	Exist.	Static	Long.		RaTEScore	CE F1
Generalist MLLMs						
Qwen3-VL-8B (Bai et al., 2025)	54.14	14.20	37.70	38.19	63.45	9.39
InternVL-3.5-8B (Wang et al., 2025)	65.09	15.49	43.06	37.43	55.14	3.96
2D Medical MLLMs						
Lingshu-7B (Xu et al., 2026)	78.70	14.20	14.60	40.16	54.80	7.66
HuluMed-7B (Jiang et al., 2025)	81.60	58.00	79.40	42.96	64.09	23.92
MedGemma-1.5-4B (Sellergren et al., 2026)	66.68	17.47	19.32	38.24	55.35	11.62
OmniCT-7B (Lin et al., 2026)	67.92	47.09	69.23	36.39	54.28	10.19
3D Medical MLLMs						
RadFM-14B-FT (Wu et al., 2025)	29.20	44.11	42.99	31.49	–	–
M3D-Phi-3-4B-FT (Gai et al., 2026)	82.43	49.30	74.77	30.57	41.50	3.59
Photon-3B-FT (Fang et al., 2026)	83.07	52.86	77.01	33.25	–	–
Dedicated 3D CT MLLMs						
RadSight-4B (Liu et al., 2026)	85.14	63.80	81.69	44.33	60.73	36.45
RadSight-8B (Liu et al., 2026)	85.14	63.94	80.89	44.06	61.24	36.00
AnatoBridge-4B	87.38	**68.22**	**84.37**	45.86	65.36	42.23
AnatoBridge-8B	**87.43**	68.15	**84.37**	**46.97**	**66.71**	**43.11**

4.2 Zero-shot multi-disease classification

We first evaluate the encoder independently of the LLM. Under the 16-disease protocol, AnatoBridge achieves the best reported value for every metric on both the CT-RATE internal test set and the external Rad-ChestCT cohort (Table 1). Relative to CT-DiagVLM, AUC increases from 84.4 to 85.5 internally and from 75.4 to 76.1 externally, supporting improved cross-cohort transfer of the organ- and disease-level representations. Section 4.4 analyzes the effect on lung-nodule discrimination, and Appendix F provides complete per-disease results.

4.3 Main results on 3D medical understanding

Having established that anatomy-aware encoding and lesion-guided learning strengthen the encoder representations, we next examine whether these gains are retained after language alignment. Existing generalist and medical MLLMs remain particularly weak on temporal reasoning and report

Table 3: DeepChestVQA results: average accuracy per question type. All baseline results are taken from Wang et al. (2026).

Method	Recognition	Visual Reasoning	Medical Reasoning
Random	0.50	0.31	0.40
GPT-5 (Singh et al., 2025)	0.52	0.38	0.39
Qwen3-VL-30B (Bai et al., 2025)	0.53	0.32	0.45
MedGemma-27B (Sellergren et al., 2025)	0.50	0.31	0.41
HuatuoGPT-34B (Chen et al., 2024)	0.53	0.29	0.40
M3D (Bai et al., 2024)	0.52	0.32	0.38
RadFM (Wu et al., 2025)	0.47	0.30	0.35
3DMedAgent (Qwen3-VL-30B) (Wang et al., 2026)	0.68	**0.45**	0.52
AnatoBridge-4B	**0.72**	0.41	**0.54**
AnatoBridge-8B	0.70	0.44	**0.54**

Table 4: Encoder ablation on CT-RATE classification.

Configuration	AUC	Lung-Nodule AUC
Baseline (CT-DiagVLM (Shi et al., 2026))	84.4	66.8
+ Anatomy-aware visual encoder	84.8	66.9
w/o anatomy-constrained mask	83.9	64.4
+ Lesion-guided hierarchical features (w/o $\mathcal{L}_{\text{dis}}$)	84.6	71.3
+ Disease-specific contrastive loss (full)	**85.5**	**71.6**

generation: generalist models score below 16% on 3D-RAD static temporal diagnosis, while non-specialized 3D models reach at most 41.50 RaTEScore. AnatoBridge-4B outperforms all reported baselines across every evaluated metric, including the larger RadSight-8B. On 3D-RAD, it improves existence detection by $+2.24$, static temporal diagnosis by $+4.28$, and longitudinal temporal diagnosis by $+3.48$. On CT-Spatial-VQA, which tests spatial understanding of volumetric anatomy, it gains $+1.80$. For CT-RATE report generation, both RaTEScore ($+4.12$) and clinical efficacy F1 ($+6.23$) improve substantially. These consistent gains across temporal reasoning, spatial understanding, and structured generation suggest that region-constrained encoding and explicit anatomical evidence benefit diverse aspects of 3D medical understanding.

On DeepChestVQA (Table 3), AnatoBridge-4B exceeds the strongest reported agent-based baseline on recognition (0.72 vs. 0.68) and medical reasoning (0.54 vs. 0.52) in a single forward pass, but remains lower on visual reasoning (0.41 vs. 0.45). We attribute part of this gap to the difference in LLM capacity and examine this effect in the following scaling analysis.

Scaling to larger LLMs. Replacing the 4B LLM with its 8B counterpart (last rows of Tables 2 and 3) mainly benefits tasks with a substantial language-reasoning or generation component: CT-Spatial-VQA increases by 1.11 points, CT-RATE report generation improves by 1.35 RaTEScore and 0.88 CE F1, and DeepChestVQA visual reasoning rises from 0.41 to 0.44. The latter nearly matches the 0.45 achieved by 3DMedAgent with a 30B LLM, supporting the view that language-model capacity accounts for part of the 4B model's remaining gap. In contrast, performance on the three 3D-RAD tasks changes little, suggesting that LLM scaling is not the primary source of improvement on these tasks.

4.4 Ablation Study

Components of the anatomy-aware encoder. Table 4 separates the contributions of in-transformer anatomy queries, constrained attention, lesion-enhanced features, and disease-specific supervision. Removing the attention constraint while retaining the queries reduces AUC below the baseline (83.9 vs. 84.4), indicating that query placement alone is insufficient. Lesion-guided features mainly benefit fine-grained discrimination, increasing lung-nodule AUC from 66.8 to 71.3; the disease-specific contrastive loss then raises overall AUC to 85.5 while retaining the lung-nodule improvement (71.6).

Table 5: Effect of lesion fusion with and without structured prompts on 3D-RAD.

	3D-RAD		
Configuration	Exist.	Static	Long.
w/o lesion fusion, visual only	86.49	67.11	83.01
w/o lesion fusion, visual + prompt	87.12	67.13	83.54
w/ lesion fusion, visual only	86.69	67.56	83.50
w/ lesion fusion, visual + prompt (full)	**87.38**	**68.22**	**84.37**

Table 6: Ablation of anatomy tokens and prompt components on 3D-RAD. • indicates that a component is included.

					3D-RAD		
Volume	Anatomy tok.	Disease	Organ geom.	Lesion	Exist.	Static	Long.
•	–	–	–	–	86.42	66.79	82.60
•	•	–	–	–	86.69	67.56	83.50
•	•	•	–	–	87.24	67.94	83.64
•	•	•	•	–	87.22	68.19	83.85
•	•	•	•	•	**87.38**	**68.22**	**84.37**
–	–	•	•	•	86.69	66.86	83.22

Downstream effect of lesion fusion. Table 5 evaluates lesion fusion both with and without structured prompts. In the visual-only setting, lesion fusion improves existence detection from 86.49 to 86.69, static temporal diagnosis from 67.11 to 67.56, and longitudinal temporal diagnosis from 83.01 to 83.50. The gains remain when prompts are enabled, with improvements of $+0.26$, $+1.09$, and $+0.83$ points, respectively. Thus, the lesion-guided representations learned during encoder pretraining continue to benefit downstream reasoning rather than serving only the auxiliary segmentation objective.

Contributions of the feature and prompt channels. Table 6 incrementally adds the anatomy tokens from the feature channel and the disease, organ-geometry, and lesion-grounding blocks from the prompt channel. Anatomy tokens provide the largest single gains on static temporal diagnosis ($+0.77$) and longitudinal temporal diagnosis ($+0.90$), showing the value of transferring continuous organ representations. The explicit text blocks then contribute complementary information: disease findings improve existence detection most strongly ($+0.55$), organ geometry contributes spatial priors that benefit both diagnosis tasks, and lesion grounding produces the largest gain on longitudinal temporal diagnosis ($+0.52$). Using the three text blocks without the volume and anatomy tokens is consistently weaker than the complete model, indicating that explicit prompts complement rather than replace continuous visual features.

5 Conclusion

We introduced AnatoBridge, which strengthens anatomical representations during visual encoding and representation learning and transfers them to the LLM through complementary feature and prompt channels. The encoder integrates learnable anatomy queries into the transformer and constrains each organ query's direct visual access with an anatomy-constrained mask. Lesion-guided hierarchical pretraining converts voxel-level supervision into organ- and disease-level representations, while dual-channel transfer supplies the LLM with both continuous anatomy tokens and structured textual evidence. Across multi-disease classification, 3D medical VQA, spatial reasoning, and report generation, AnatoBridge consistently improves over the evaluated dedicated 3D CT MLLMs.

Limitations. Due to the availability of open medical imaging data, we evaluate AnatoBridge only on chest CT. Extending it to abdominal or whole-body CT would require new anatomical definitions, annotations, mappings, and validation. In addition, the compact encodings of organ geometry and lesion masks may not fully preserve their spatial detail, and the accuracy of the predicted anatomical priors could be further improved. We plan to address these limitations in future work.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology and experiments, clean and reformat data, support qualitative analysis, interpret results, improve the readability of the manuscript, and assist with translation. We did not use generative AI to design the research methodology or experiments or to implement the proposed methods. The generation of synthetic datasets, development of theoretical or conceptual frameworks, formulation or proof of mathematical claims, and proposal or refinement of hypotheses were not applicable to this work. All AI-assisted outputs were reviewed by the authors. Methodological feedback and result interpretations were assessed against the experimental evidence; data-processing outputs were checked for consistency and correctness; and translations and language edits were manually verified to preserve the intended technical meaning. We take full responsibility for the final content of this work, including all text, claims, analyses, and artifacts produced with the assistance of generative AI.

REFERENCES

Mohammed Baharoon, Luyang Luo, Michael Moritz, Abhinav Kumar, Sung Eun Kim, Xiaoman Zhang, Miao Zhu, Mahmoud H Alabbad, Maha S Alhazmi, Neel P Mistry, et al. Rexgroundingct: A 3d chest ct dataset for segmentation of findings from free-text reports. *NEJM AI*, 3(7): AIdbp2501220, 2026.

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.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Louis Blankemeier, Joseph Paul Cohen, Ashwin Kumar, Dave Van Veen, Syed Jamal Safdar Gardezi, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Cesar Truyts, Christian Bluethgen, Yufu Wu, Long Lian, Malte Engmann Kjeldskov Jensen, Sophie Ostmeier, Maya Varma, Jeya Maria Jose Valanarasu, Zhongnan Fang, Zepeng Huo, Zaid Nabulsi, Diego Ardila, Wei-Hung Weng, Edson Amaro Junior, Neera Ahuja, Jason Fries, Nigam H. Shah, Greg Zaharchuk, Marc Willis, Adam Yala, Andrew Johnston, Robert D. Boutin, Andrew Wentland, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, and Akshay S. Chaudhari. Merlin: a computed tomography vision–language foundation model and dataset. *Nature*, 652:1318–1328, 2026.

Weiwei Cao, Jianpeng Zhang, Yingda Xia, Tony CW Mok, Zi Li, Xianghua Ye, Le Lu, Jian Zheng, Yuxing Tang, and Ling Zhang. Bootstrapping chest CT image understanding by distilling knowledge from X-ray expert models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 11238–11247, 2024.

Weiwei Cao, Jianpeng Zhang, Zhongyi Shui, Sinuo Wang, Zeli Chen, Xi Li, Le Lu, Xianghua Ye, Qi Zhang, Tingbo Liang, Qi Zhang, and Ling Zhang. Boosting vision semantic density with anatomy normality modeling for medical vision-language pre-training. In *Int. Conf. Comput. Vis.*, pp. 23041–23050, 2025.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, et al. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 7346–7370, 2024.

Rachel Lea Draelos, David Dov, Maciej A Mazurowski, Joseph Y Lo, Ricardo Henao, Geoffrey D Rubin, and Lawrence Carin. Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. *Med. Image Anal.*, 67:101857, 2021.

Chengyu Fang, Heng Guo, Zheng Jiang, Chunming He, Xiu Li, and Minfeng Xu. Photon: Speedup volume understanding with efficient multimodal large language models. In *Int. Conf. Learn. Represent.*, 2026.

Xiaotang Gai, Jiaxiang Liu, Yichen Li, Zijie Meng, Jian Wu, and Zuozhu Liu. 3d-rad: A comprehensive 3d radiology med-vqa dataset with multi-temporal analysis and diverse diagnostic tasks. *Adv. Neural Inform. Process. Syst.*, 38, 2026.

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. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. *arXiv preprint arXiv:2403.17834*, 2024.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Int. Conf. Comput. Vis.*, pp. 3942–3951, 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.

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. *Adv. Neural Inform. Process. Syst.*, 36:28541–28564, 2023.

Tianwei Lin, Zhongwei Qiu, Wenqiao Zhang, Jiang Liu, Yihan Xie, Mingjian Gao, Zhenxuan Fan, Zhaocheng Li, Sijing Li, Zhongle Xie, et al. Omnict: Towards a unified slice-volume lvlm for comprehensive CT analysis. In *Int. Conf. Learn. Represent.*, 2026.

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.

Mashrafi Monon, Umaima Rahman, Asif Hanif, Numan Saeed, and Mohammad Yaqub. Lost in volume: The ct-spatialvqa benchmark for evaluating semantic-spatial understanding of 3d medical vision-language models. *arXiv preprint arXiv:2605.08787*, 2026.

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.

Philip Müller, Georgios Kaissis, Congyu Zou, and Daniel Rueckert. Joint learning of localized representations from medical images and reports. In *Eur. Conf. Comput. Vis.*, pp. 685–701. Springer, 2022.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

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.

Bowen Shi, Weiwei Cao, Ruifeng Yuan, Wanxing Chang, Wenrui Dai, Hongkai Xiong, Ling Zhang, and Jianpeng Zhang. Disease-centric vision-language pretraining with hybrid visual encoding for 3d computed tomography. In *Forty-third International Conference on Machine Learning*, 2026.

Zhongyi Shui, Jianpeng Zhang, Weiwei Cao, Sinuo Wang, Ruizhe Guo, Le Lu, Lin Yang, Xianghua Ye, Tingbo Liang, Qi Zhang, and Ling Zhang. Large-scale and fine-grained vision-language pre-training for enhanced CT image understanding. In *Int. Conf. Learn. Represent.*, 2025.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Adv. Neural Inform. Process. Syst.*, 37:87310–87356, 2024.

Tassilo Wald, Ibrahim Ethem Hamamci, Yuan Gao, Sam Bond-Taylor, Harshita Sharma, Maximilian Ilse, Cynthia Lo, Olesya Melnichenko, Anton Schwaighofer, Noel CF Codella, et al. Comprehensive language–image pre-training for 3d medical image understanding. In *Eur. Conf. Comput. Vis.*, pp. 388–406. Springer, 2026.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Ziyue Wang, Linghan Cai, Chang Han Low, Haofeng Liu, Junde Wu, Jinyu Wang, Rui Wang, Lei Song, Jiang Bian, Jingjing Fu, and Yueming Jin. 3DMedagent: Unified perception-to-understanding for 3d medical analysis. In *Forty-third International Conference on Machine Learning*, 2026.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023.

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: Generalist foundation model for unified multimodal medical understanding and reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Hangjie Yuan, Yichen Qian, Zhiwei Tang, Xianzhe Xu, Lirong Wu, Sicheng Yang, Jinwang Wang, Pengju Wang, Zhitao Zeng, Yizeng Han, et al. Clinfusion: A vision-centric multimodal llm system for holistic medical understanding. *arXiv preprint arXiv:2607.24743*, 2026.

Qi Zhang, Jianpeng Zhang, Weiwei Cao, Zilin Lu, Wanxing Chang, Haonan Ding, Cao Chen, Zhi Li, Xing Xue, Sinuo Wang, et al. An expert-level generalist ai for abdominal ct diagnosis. *Science*, 393(6817):eaec6129, 2026.

Chenhui Zhao, Yiwei Lyu, Asadur Zaman Chowdury, Edward S Harake, Akhil Kondepudi, Akshay T Rao, Xinhai Hou, Honglak Lee, and Todd C Hollon. Towards scalable language-image pre-training for 3d medical imaging. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856.

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.

A Reproducibility Details

Encoder pretraining. We adopt the preprocessing pipeline of CT-DiagVLM (Shi et al., 2026): volumes are resampled to $1\times1\times5$ mm resolution, Hounsfield-unit values are clipped to $[-1150, 350]$ and linearly scaled to $[0, 1]$, and a random crop of $112 \times 256 \times 352$ voxels is applied during training. This yields $N = 616$ visual tokens on a $7 \times 8 \times 11$ grid; the total transformer sequence length is $N + K = 621$. The text encoder is CXR-BERT, the five learnable anatomy queries are initialized from a truncated-normal distribution, and the lesion-fusion projections in Eqs. 8 and 9 are zero-initialized. Optimization uses Adam with a cosine-decay learning-rate schedule peaking at 1×10^{-4}. The loss weights in Eq. 13 are set to $\lambda_{\text{seg}} = \lambda_{\text{lesion}} = \lambda_{\text{align}} = 1$. Pretraining runs for 30 epochs with a total batch size of 48.

Instruction tuning. The 3D visual backbone carries over from the pretrained encoder, and a randomly initialized two-layer GELU projector maps 768-dimensional visual features to the LLM embedding dimension. All runs train for 2 epochs with a global batch size of 512 under BF16 precision. Learning rates are 1×10^{-5} for the LLM, 1×10^{-4} for the projectors, and 2×10^{-6} for the 3D encoder; the cosine-decay scheduler begins with a linear warmup over the first 3% of training steps. Unless otherwise stated, we use Qwen3-VL-4B-Instruct as the language backbone and apply full fine-tuning; LoRA ($r = 16$, $\alpha = 16$, dropout 0.05) is used only in the ablation study of Sec. G. Qwen3-VL-8B-Instruct appears only in the scaling experiments.

B Datasets, Splits, and Annotation Sources

B.1 Encoder Pretraining Data

We follow the data configuration of CT-DiagVLM (Shi et al., 2026) for encoder pretraining. CT-RATE (Hamamci et al., 2024) contains 50,188 chest CT volumes from 21,304 patients, each paired with a radiology report. Following its standard patient-level split, 20,000 patients are used for pretraining and the remaining 1,304 patients form the internal test set. The paired reports provide the textual supervision for organ- and disease-level alignment. Organ segmentation supervision is obtained from pseudo-labels generated by TotalSegmentator (Wasserthal et al., 2023).

Within the CT-RATE pretraining cohort, a subset of scans has additional lesion-grounding annotations from ReXGroundingCT (Baharoon et al., 2026). Specifically, ReXGroundingCT annotates 3,142 non-contrast CT-RATE scans with free-text findings and three-dimensional lesion instances, divided into 2,992 training, 50 validation, and 100 test scans. We apply the lesion-specific objective only to the 2,992 scans in the ReXGroundingCT training split; the remaining CT-RATE pretraining scans contribute to the other objectives without lesion supervision. Because the ReXGroundingCT annotations are instance-level, we convert them into 14-class semantic segmentation targets by taking the voxel-wise union of all instances assigned to the same lesion category within each volume. Its validation and test annotations are not used for model training.

B.2 Supervised Instruction-Tuning Data

Our supervised instruction-tuning corpus comprises 177,309 instruction–response pairs. Of these, 134,701 pairs (76.0%) come from the 3D-RAD training set. The remaining 42,608 pairs (24.0%) are drawn from the auxiliary chest CT instruction data constructed by RadSight (Liu et al., 2026). These data cover report generation (12,732 pairs), closed-ended abnormality detection (9,142), open-ended abnormality localization and characterization (8,054), multiple-choice abnormality recognition (4,971), 3D organ localization (3,925), and region-level content understanding (3,784). We include this additional data both to complement 3D-RAD with report-generation supervision and to enable a fair comparison with RadSight under a matched instruction-tuning data configuration.

B.3 Evaluation Data

Zero-shot disease classification. We evaluate the pretrained encoder on the CT-RATE internal test set and the external Rad-ChestCT (Draelos et al., 2021) cohort. The CT-RATE test set contains the 1,304 patients held out from encoder pretraining. Rad-ChestCT contains 3,630 CT volumes with paired reports and is used exclusively for external evaluation. Following CT-DiagVLM, both cohorts are evaluated under the same 16-disease protocol, which includes the labels associated with the four organs of interest and excludes the two global CT-RATE labels.

3D-RAD. The 3D-RAD test set (Gai et al., 2026) contains 33,712 question–answer pairs spanning both open- and closed-ended questions. Following the evaluation setting of RadSight (Liu et al., 2026), we report results on the closed-ended questions for the existence detection, static temporal diagnosis, and longitudinal temporal diagnosis subtasks.

CT-Spatial-VQA. CT-Spatial-VQA (Monon et al., 2026) evaluates semantic and spatial understanding of CT volumes through 8,953 open-ended test questions. We use no CT-Spatial-VQA training data and evaluate AnatoBridge directly in the zero-shot setting.

Table 7: Cross-taxonomy correspondence between the 14 ReXGroundingCT pulmonary finding categories and the organ-associated CT-RATE diseases. "None" denotes the absence of a semantically compatible disease label.

Code	ReXGroundingCT finding category	Associated CT-RATE disease(s)
1a	Bronchial wall thickening	Peribronchial thickening
1b	Bronchiectasis	Bronchiectasis
1c	Emphysema (including centrilobular, paraseptal, and bullous)	Emphysema
1d	Septal thickening (including interlobular thickening and reticulation)	Interlobular septal thickening
1e	Micronodules (including centrilobular, tree-in-bud, and perilymphatic)	None
1f	Other typically non-focal lung/airway/pleural abnormality	Mosaic attenuation pattern
2a	Linear opacity (including subsegmental atelectasis, scarring, and fibrosis)	Atelectasis; Pulmonary fibrotic sequela
2b	Atelectasis or consolidation	Consolidation
2c	Ground-glass opacity	Lung opacity
2d	Pulmonary nodules or masses	Lung nodule
2e	Pleural effusion or thickening	Pleural effusion
2f	Honeycombing	None
2g	Pneumothorax	None
2h	Other typically focal lung/airway/pleural opacity	None

DeepChestVQA. DeepChestVQA (Wang et al., 2026) contains 1,020 question–answer pairs derived from 892 chest CT scans and covers 17 capability dimensions. The benchmark is organized into three groups: Recognition comprises 3 subtypes and 180 questions, Visual Reasoning comprises 8 subtypes and 480 questions, and Medical Reasoning comprises 6 subtypes and 360 questions. We report average accuracy for each of the three question groups.

CT report generation. The CT-RATE test cohort comprises 1,304 patients excluded from the encoder-pretraining cohort. Following the evaluation protocol of RadSight (Liu et al., 2026), we retain one reconstructed volume for each CT examination, prioritizing the 3-mm reconstruction when available. Because some patients underwent multiple examinations, this procedure yields 1,564 CT volumes, each paired with its corresponding reference report, for report-generation evaluation.

C Disease–Lesion Affinity Matrix

ReXGroundingCT and CT-RATE organize pulmonary findings and diseases at different semantic granularities. We formulate their correspondence as a cross-taxonomy alignment represented by a sparse binary matrix $\boldsymbol{A}^{\text{disease}} \in \{0,1\}^{16\times 14}$. Qwen3.8-Max is used as a zero-shot semantic annotator to align the textual descriptions of the ReXGroundingCT categories with the CT-RATE disease terminology. The resulting correspondences are binarized and fixed throughout training, providing a semantic prior independent of patient-level label co-occurrence.

Specifically,

$$A^{\text{disease}}_{d,\ell} = \begin{cases} 1, & \text{if disease } d \text{ is associated with lesion category } \ell, \\ 0, & \text{otherwise.} \end{cases} \tag{17}$$

A lesion category may be linked to more than one disease, as in category 2a, whereas a disease without a direct lesion counterpart is assigned an all-zero row. Table 7 gives the complete sparse specification of the matrix: every association not listed in the last column is zero. The lesion category names follow the hierarchy defined by ReXGroundingCT (Baharoon et al., 2026).

The five remaining organ-associated CT-RATE diseases—cardiomegaly, pericardial effusion, coronary artery wall calcification, hiatal hernia, and arterial wall calcification—have no direct counterpart among these 14 pulmonary lesion categories. Their rows in $\boldsymbol{A}^{\text{disease}}$ are therefore zero, and

Eq. 9 reduces to the corresponding encoded organ-query feature. This fallback prevents unrelated pulmonary lesion tokens from being injected into non-pulmonary disease representations.

D Structured Prompt Construction

Section 3.4 introduces the disease, organ, and lesion prompt components, denoted by P_D, P_O, and P_L, respectively. This appendix specifies how the corresponding encoder outputs are converted into text and provides a representative example. All fields are derived from model predictions rather than reference annotations.

Disease findings. Let s_d denote the predicted score for disease d. The vocabulary contains all 18 CT-RATE labels: the 16 organ-associated diseases are scored from their corresponding disease features, and the two global diseases are scored from the global feature. We retain

$$\mathcal{D}^+ = \mathrm{TopK}_5\left\{(d, s_d) : s_d > 0.5\right\}, \tag{18}$$

and serialize the selected findings in descending order of confidence as disease–score pairs.

Organ geometry. For each organ $o \in \mathcal{O}$, we derive a hard mask $\widehat{M}_o$ from the predicted organ probabilities. Its geometry is summarized by the voxel count n_o, a normalized three-dimensional bounding box $\boldsymbol{b}_o \in [0,1]^6$, and a normalized centroid $\boldsymbol{c}_o \in [0,1]^3$. Together, these quantities encode organ size, spatial extent, and position without exposing the full segmentation mask to the LLM.

Lesion grounding. For each predicted lesion category, P_L reports the dominant anatomical zones obtained using the procedure in Appendix E. We retain at most five lesion categories, ranked by their predicted voxel volumes. This representation is particularly useful for multifocal or diffuse abnormalities for which a single centroid or bounding box is not informative.

Representative prompt. The following abridged example illustrates the textual information supplied to the LLM; line breaks are added for readability.

> **Disease findings:** Emphysema (0.86), Lung nodule (0.74), Atelectasis (0.67), Pleural effusion (0.58), Bronchiectasis (0.52).
> **Organ geometry:** Lung: 1,421,854 voxels, bbox $[0.02, 0.94, 0.12, 0.98, 0.01, 0.99]$, centroid $[0.48, 0.55, 0.50]$; Heart: 18,467 voxels, bbox $[0.31, 0.72, 0.38, 0.72, 0.34, 0.66]$, centroid $[0.51, 0.55, 0.50]$; Aorta: 21,336 voxels, centroid $[0.52, 0.44, 0.30]$.
> **Lesion grounding:** Emphysema (mid bilateral); Ground-glass opacity (lower-L 45%, lower-R 38%); Pulmonary nodule (mid-L); Pleural effusion (lower bilateral).
> **Question:** Does this patient have evidence of pulmonary fibrosis?

E Six-Zone Lesion Spatial Grounding

The lesion-grounding block introduced in Sec. 3.4 summarizes each predicted lesion mask over six thoracic zones. This representation is preferable to a single bounding box for multifocal or spatially diffuse abnormalities.

Let the preprocessed CT coordinates be normalized to $[0,1]^3$. We partition the craniocaudal axis into three equal intervals, corresponding to the upper, middle, and lower levels, and divide the transverse axis at the mid-sagittal plane into the patient's left and right sides. This gives the zone set

$$\mathcal{Z} = \{\mathrm{upper}, \mathrm{mid}, \mathrm{lower}\} \times \{\mathrm{L}, \mathrm{R}\}. \tag{19}$$

For lesion class ℓ, let $\widehat{M}_\ell(x)$ be its predicted binary mask and $\mathbf{1}_z(x)$ indicate whether voxel x belongs to zone z. We measure its relative occupancy in each zone as

$$r_{\ell,z} = \frac{\sum_x \widehat{M}_\ell(x)\mathbf{1}_z(x)}{\sum_x \widehat{M}_\ell(x) + \epsilon}, \qquad z \in \mathcal{Z}. \tag{20}$$

Only zones with $r_{\ell,z} \geq 0.15$ are retained. This removes minor scattered predictions while preserving the principal spatial distribution of each lesion category.

Table 8: Disease-specific zero-shot classification performance on the CT-RATE internal test set. Sens. and Spec. denote sensitivity and specificity, respectively. Diseases are ordered by AUC.

Organ	Disease	AUC	ACC	Sens.	Spec.
Lung	Pleural effusion	0.971	0.946	0.946	0.946
Heart	Cardiomegaly	0.947	0.895	0.850	0.901
Heart	Coronary artery wall calcification	0.940	0.881	0.905	0.872
Aorta	Arterial wall calcification	0.935	0.869	0.906	0.854
Lung	Consolidation	0.908	0.825	0.830	0.824
Heart	Pericardial effusion	0.894	0.818	0.848	0.816
Lung	Interlobular septal thickening	0.875	0.751	0.882	0.739
Lung	Mosaic attenuation pattern	0.860	0.790	0.751	0.794
Lung	Lung opacity	0.849	0.785	0.768	0.795
Esophagus	Hiatal hernia	0.831	0.782	0.740	0.789
Lung	Emphysema	0.814	0.732	0.741	0.730
Lung	Peribronchial thickening	0.806	0.720	0.758	0.716
Lung	Bronchiectasis	0.802	0.736	0.730	0.737
Lung	Atelectasis	0.786	0.718	0.729	0.715
Lung	Pulmonary fibrotic sequela	0.741	0.693	0.689	0.694
Lung	Lung nodule	0.716	0.661	0.609	0.703
Macro average		0.855	0.788	0.793	0.789

Table 9: Ablation of encoder tuning strategies on 3D-RAD.

	3D-RAD		
Strategy	Exist.	Static	Long.
Frozen (projector + LLM only)	87.23	67.35	83.54
LoRA	87.30	68.15	83.81
Full fine-tuning	**87.38**	**68.22**	**84.37**

To obtain concise bilateral descriptions, the left and right zones at level $h \in \{\text{upper}, \text{mid}, \text{lower}\}$ are merged when

$$r_{\ell,(h,\text{L})} \geq 0.4 \quad \text{and} \quad r_{\ell,(h,\text{R})} \geq 0.4. \tag{21}$$

The resulting finding is expressed as "h bilateral." Otherwise, retained zones are reported separately; when multiple zones are needed, the displayed percentage for zone z is $100 r_{\ell,z}\%$. Finally, lesion categories are ranked by total predicted voxel volume, and the five largest nonempty categories are included in the prompt.

These rules yield compact expressions such as "mid bilateral," "mid-L," or "lower-L 45%, lower-R 38%." A complete example in combination with the disease and organ blocks is provided in Appendix D.

F Complete Zero-Shot Classification Results

Table 8 reports the complete per-disease results on the CT-RATE internal test set under the 16-disease protocol used in Sec. 4.2. In addition to AUC and accuracy, we report sensitivity and specificity to characterize the class-wise operating points. Relative to the disease-specific results reported for CT-DiagVLM (Shi et al., 2026), AnatoBridge improves AUC on 11 of the 16 diseases and matches its performance on mosaic attenuation pattern. The largest gain is observed for lung nodule ($0.668 \rightarrow 0.716$), followed by emphysema ($0.786 \rightarrow 0.814$) and pulmonary fibrotic sequela ($0.719 \rightarrow 0.741$), supporting the benefit of lesion-guided representations for fine-grained pulmonary findings.

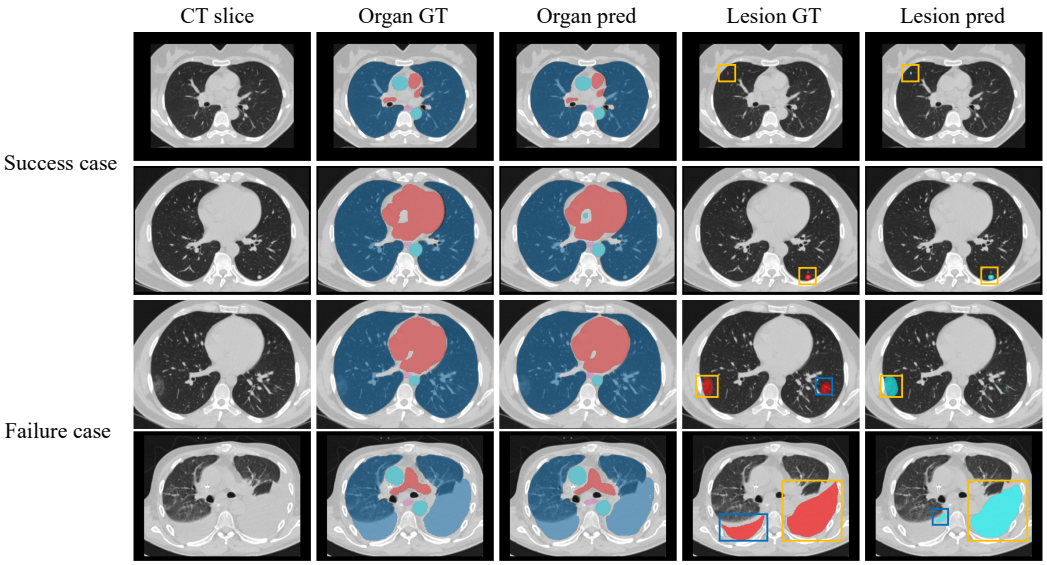


Figure 2: Qualitative organ and lesion segmentation on the ReXGroundingCT validation set. Rows 1–2: success cases with correct organ masks and small-lesion localization. Rows 3–4: failure cases showing a missed lesion and incomplete segmentation.

G Encoder Adaptation Strategies.

Table 9 compares freezing, LoRA, and full fine-tuning of the task encoder. Full fine-tuning achieves the best performance on all three 3D-RAD tasks but modifies the encoder weights, requiring an additional frozen copy for prior generation. LoRA recovers most of this gain with fewer parameters and no extra frozen copy, and even the fully frozen encoder remains competitive. We adopt full fine-tuning in the main experiments, while LoRA and freezing offer attractive alternatives for resource-constrained deployment.

H Qualitative Results and Failure Cases

Organ and lesion segmentation. Figure 2 visualizes organ and lesion segmentation predictions from the pretrained 3D encoder on the ReXGroundingCT validation set. The model produces accurate organ masks across all four cases, closely matching the ground truth for all four thoracic organs. The model also demonstrates meaningful lesion-localization capability. In the two success cases (rows 1–2), it correctly identifies small lesions highlighted by the yellow boxes in the ground truth. However, failure cases reveal two common error modes: missed detections, where a lesion present in the ground truth is not predicted (row 3, blue box), and incomplete segmentation, where the predicted mask covers only a portion of the lesion extent (row 4, blue box). Given the inherent difficulty of voxel-level lesion localization, we consider the current performance acceptable; the quantitative gains in Tables 4 and 5 confirm that the predicted lesion masks contribute positively to downstream tasks. A dedicated segmentation decoder could further improve lesion localization, but this is beyond the scope of the present work.

VQA qualitative analysis. Table 10 presents representative VQA examples where the full model succeeds. On these cases, the volume-only baseline fails across the board. For spatial reasoning (rows 1–2), the baseline reverses effusion laterality and detects only one lobe of diffuse nodules, while anatomy-aware encoding lets AnatoBridge reason about spatial layout. For existence detection (rows 3–4), the baseline misses both calcification and hiatal hernia; the disease representations in the prompt channel enable correct detection. For longitudinal diagnosis (rows 5–6), AnatoBridge combines sequence history with current CT to correctly classify resolved and new lesions, while the baseline confuses resolved with refractory and new with absent.

Table 10: Representative VQA examples from 3D-RAD. Rows 5–6: sequence [1] indicates the finding was previously present, [0] previously absent; the model must combine this history with the current CT to classify temporal status as A: Refractory (persistent/recurrent, now present), B: Resolved (previously present, now absent), C: New (absent previously, now present), or D: No abnormality (always absent).

Task	Question	GT	AnatoBridge	Vol.
Anomaly Detection	Where is the bilateral minimal pleural effusion more prominent?	Right side	Right side	Left side
Image Observation	Where are the metastatic nodules located?	Both lungs	Both lungs	Right lung upper lobe
Existence Detection	Is arterial wall calcification confirmed by the CT findings?	yes	yes	no
Existence Detection	Does the CT scan reveal hiatal hernia?	yes	yes	no
Longitudinal Diagnosis	How is lung nodule classified now? (sequence: [1])	B	B	A
Longitudinal Diagnosis	Current status of interlobular septal thickening? (sequence: [0])	C	C	D

Table 11: Representative failure cases in VQA. The disease score reflects the confidence of the disease representation in the prompt channel.

Task	Question	GT	AnatoBridge	Score
Existence Detection	Does this CT image present arterial wall calcification?	no	yes	0.90
Existence Detection	Is arterial wall calcification identifiable?	yes	no	0.76
Static Temporal	What is the current status of cardiomegaly?	C	A	–

Table 12: Stratified analysis of the prompt channel's contribution. Samples are grouped by whether the target disease appears in the top-5 retrieved findings and whether the ground truth is positive or negative. AnatoBridge denotes the full model with the prompt channel; Vol. denotes the volume-only baseline.

Stratum	n	AnatoBridge	Vol.	Δ
GT=yes, target in top-5	2,397	66.83%	56.49%	+10.35%
GT=yes, target above threshold, not in top-5	1,092	25.18%	25.73%	−0.55%
GT=yes, target below threshold	675	5.48%	5.19%	+0.30%
GT=no, false positive in top-5	2,636	80.12%	83.08%	−2.96%
GT=no, target above threshold, not in top-5	2,589	95.09%	93.55%	+1.55%
GT=no, target below threshold	13,830	99.78%	99.66%	+0.12%
Overall	23,219	87.38%	86.42%	+0.96%

Table 11 shows three representative failure modes. A high disease score (0.90) in the prompt channel misleads the model into a false positive (row 1), illustrating how retrieval errors propagate into wrong answers; conversely, a correct score (0.76) does not guarantee a correct response (row 2), indicating that the LLM may fail to properly integrate the prompt signal with its own reasoning. We quantify the broader pattern of error propagation in disease retrieval below. The third case (row 3) reflects a task setup limitation rather than a model failure: without prior scans, the model cannot determine temporal status.

Error propagation in disease retrieval. The prompt channel comprises disease findings, organ geometry, and lesion grounding; we stratify by disease findings as they are most relevant to existence detection. Table 12 shows that the prompt channel's contribution depends strongly on retrieval quality: correct retrieval within top-5 yields $+10.35$ points, while a false positive in top-5 for negative cases costs -2.96 points, with marginal gains (± 0.5) elsewhere. Notably, correct prompts yield only 66.83% accuracy rather than 100%, and misleading prompts still achieve 80.12% rather than

collapsing, indicating the model integrates disease findings with volume and anatomy tokens rather than following the prompt exclusively. The overall $+0.96$ point gain across most strata validates the reliability of incorporating disease representations into the prompt channel.