SEEING IS NOT SAYING: CLOSING THE VERBALIZATION GAP IN 3D CT REPORT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work on 3D Computed Tomography (CT) radiology report generation has largely focused on improving visual representations and the evidence supplied to language models. We identify a substantial gap between what is recoverable from visual representations and what is ultimately expressed in generated reports. On the example of thorax abnormalities, lightweight probes across five frozen encoders recover substantially stronger abnormality signals than report generators express, including for encoders not trained with CT reports. We term this discrepancy the *verbalization gap* and address it by improving both the evidence available for generation and how that evidence is expressed. Our framework consolidates complementary CT representations by slot-wise convex weights learned over independently resampled tokens and introduces graded verbalization, which guides report generation with diagnostic predictions whose linguistic certainty reflects their class-level reliability. With a 3B language backbone, our system achieves 0.597 macro F1 and 0.526 CRG on CT-RATE, a relative improvement of 44% over AdaRAG-CT and 180% over BTB3D. Under zero-shot cross-dataset evaluation on RAD-ChestCT, it achieves 0.498 macro F1, 123% above BTB3D. These results indicate that 3D CT report generation is limited not only by what can be recovered from visual representations, but also by how effectively recoverable clinical findings are translated into language.

1 INTRODUCTION

Generating a radiology report from a 3D CT volume requires condensing hundreds of axial slices into a concise description of clinically relevant findings. Reliable automation could reduce the workload of radiologists and improve diagnostic efficiency. The growth of large-scale medical imaging resources (Littlejohns et al., 2020) and paired 3D CT report datasets (Hamamci et al., 2026b) has made model training and evaluation for this task feasible at scale. Recent advances have consequently centered on how to represent volumetric clinical evidence and convey it effectively to the report generator. One direction strengthens the visual representation itself: Hamamci et al. (2026a) replace slice-wise tokenization with compact frequency-aware volumetric tokens and show that precise 3D tokenization is critical for performance. Liang et al. (2026) identify severe dimensional concentration in contrastive CT embeddings, pair a global embedding with four organ-level embeddings, and add a textual channel retrieved on demand to address this representational bottleneck.

These directions ask how to provide the generator with better evidence. A separate question remains less explored: *when clinically relevant information is already recoverable from a visual representation, how much of it is actually expressed in the generated report?*

Table 1 reports the covariance spectrum of five frozen encoders trained under different objectives. Variance concentrates in all five, including the two that have never seen a CT report: $\dim_{90}$ occupies between 1.7% and 10.2% of the ambient dimension. The encoders further agree on which findings are hard, with per-abnormality AUROC rankings correlated at $\rho \geq 0.86$, *e.g.*, pleural effusion ranks easiest and lung nodule hardest for every one of them. This agreement does not extend to encoder-specific strengths. With per-class thresholds calibrated for each encoder, FlexiCT (Li et al., 2026) leads on twelve of the eighteen abnormalities by F1 and COLIPRI (Wald et al., 2025) achieves the best precision on eight, while SigVLP (Wang et al., 2026) reaches the highest recall on six. Their

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.

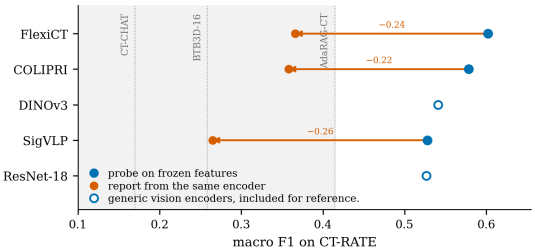


Figure 1: **The verbalization gap.** Abnormality information is more decodable from frozen CT features than is expressed in generated reports.

Table 1: **Effective dimensionality of 3D CT features on CT-RATE.** dim_{90}/dim_{95}: number of principal components for 90%/95% of the variance. ERank: effective rank. PR: participation ratio.

Representation	D	dim_{90}	dim_{95}	ERank	PR
COLIPRI	768	13	16	13.0	8.3
FlexiCT	864	34	59	29.7	14.7
SigVLP	1024	31	57	30.6	15.5
DINOv3	768	77	149	53.8	22.3
ResNet-18	512	52	97	36.6	14.9

representations appear to retain complementary task-relevant information, which motivates testing whether combining them can increase the evidence available to report generation.

How much of that clinically relevant evidence appears in the report is a separate question. We compare abnormality information recoverable from frozen features with that expressed in generated reports, using 18 CT-RATE abnormality labels and the same clinical-efficacy metric, with report-side labels extracted by the RadBERT-based evaluator commonly used for clinical-efficacy evaluation in CT report generation (Yan et al., 2022; Hamamci et al., 2026b). As shown in Figure 1, two-layer MLP probes on mean-pooled frozen CT features recover substantially stronger abnormality signals than the corresponding generated reports express, and abnormality labels remain strongly recoverable from DINOv3 (Siméoni et al., 2025) and an ImageNet-pretrained ResNet-18 (He et al., 2016), neither of which was trained on CT reports. We call this difference the *verbalization gap*.

Prior work has shown that diagnostic predictions can guide report generation. Jin et al. (2024) attach one of four states of 14 chest X-ray findings as diagnosis-driven prompt tokens, including an uncertain state derived from report annotations. That state records the hedging a radiologist chose for a particular image rather than the reliability of the diagnostic prediction itself. A report separates a pleural effusion that is present from bronchiectasis that cannot be excluded, and the hedge tells the reader how far to act on the sentence. A radiologist may hedge the description of a finding, while a diagnostic model may also vary in how reliably it predicts different findings. We therefore introduce *graded verbalization*, which uses diagnostic predictions to guide report generation while matching the linguistic certainty of each finding to its measured class-level reliability.

To address both the availability and expression of clinical evidence, we use a three-stage framework. The first widens the evidence through the consolidation of three encoders, which together produce up to 3.5×10^4 patch tokens, in a shared 512-dimensional space and resample each into 128 query tokens, producing a compact visual prefix. The second changes how that evidence is stated and through graded verbalization, we condition report generation on diagnostic predictions with reliability-aware linguistic certainty. The third adapts the language model with LoRA (Hu et al., 2022) to better incorporate this guidance into the generated report.

Our key contributions are as follows:

- **Characterization of the representation-to-report verbalization gap in 3D CT report generation.** We show that clinically relevant abnormality information can remain strongly decodable from frozen CT representations yet fail to appear in generated reports.
- **A framework that improves both visual evidence and its verbalization.** We consolidate complementary information from three CT encoders using a compact slot-wise convex combination, outperforming the best single encoder by 12% relative macro F1 and the strongest tested fusion alternative by 9%. We further introduce graded verbalization, which guides report generation with diagnostic predictions whose linguistic certainty reflects their class-level reliability.
- **State-of-the-Art Clinical Accuracy.** Our 3B model reaches 0.597 macro F1 and 0.526 CRG on CT-RATE, improving macro F1 by 44% over AdaRAG-CT and 180% over BTB3D, and achieves 0.498 macro F1 under zero-shot evaluation on RAD-ChestCT.

2 Related Work

Report generation from 3D CT volumes. Hamamci et al. (2024) introduce the task with an encoder-decoder trained from scratch, pairing an auto-regressive causal vision feature extractor with a decoder conditioned on relational memory. Subsequent systems redesign how volumetric evidence is represented and supplied to language decoders: Hamamci et al. (2026b) feed embeddings from a contrastive volume-report encoder; Hamamci et al. (2026a) replace slice-wise tokenization with frequency-aware volumetric tokens; Liang et al. (2026) add organ-level features together with a textual channel retrieved on demand; Bumgardner et al. (2026) graft a language-free self-supervised encoder into a decoder through a staged curriculum. Tian et al. (2026) instead condition generation on high-level scan-to-reference semantic differences encoded as visual prompts. Generalist medical models (Wu et al., 2025) cover CT alongside other modalities. These systems differ in how visual evidence is encoded, compressed or supplemented, and are scored on the findings a report states. How firmly each finding is stated is not part of their design.

Visual representations and multi-encoder fusion. Several encoders are pretrained specifically for CT: Wald et al. (2025) combine contrastive volume-report alignment with report-generation and vision-only self-supervised objectives, Li et al. (2026) pretrain agglomeratively from axial slices to anatomy to reports, and Wang et al. (2026) treat the axial dimension as unconstrained through rotary position embeddings. General-purpose encoders without medical-text training (Siméoni et al., 2025) are also used as visual backbones. Although representations learned under different objectives can exhibit substantial shared structure (Huh et al., 2024), differences in encoder design and training can still leave complementary strengths to exploit. Accordingly, combining several encoders is established practice in multimodal language models, where channel concatenation is reported to match more elaborate schemes (Shi et al., 2025a), and where alternatives consolidate encoders through a shared resampler (Kar et al., 2024), aggregate them spatially (Tong et al., 2024), interleave tokens after independent adaptation (Azadani et al., 2026), or distil them into a single model (Ranzinger et al., 2024). Recent work has studied visual representation and compression for 3D CT report generation through token-reducing projectors and anatomy-guided cropping (Suprijadi et al., 2026), 2D–3D fusion (Hosseini et al., 2025), and dual-3D encoders with spatial compression (Shi et al., 2025b); Baharoon et al. (2025) systematically study the broader 3D-MLLM design space.

Label- and structure-guided generation. Jin et al. (2024) condition report generation on four-state diagnostic prompts over chest X-ray findings, including an uncertain state derived from report annotations rather than measured prediction reliability. Cheng & Subramanian (2026) decompose a report into topic-specific findings, and Yan et al. (2026) construct abnormality proposals for each localized region and verify them with self-critiquing mechanisms. These methods structure which findings or topics enter generation, but do not explicitly tie the linguistic certainty of each finding to the measured reliability of its prediction. Confidence has instead been modeled as a property of the generator, with calibrated numeric scores produced alongside the report (Bani-Harouni et al., 2026). We take it as a property of the evidence, derived from an external classifier's per-class precision.

3 Method

Let $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{y}_i^{\text{cls}})\}_{i=1}^{N}$ denote a dataset of N CT studies, with volume $\boldsymbol{x}_i \in \mathbb{R}^{H\times W\times D}$, paired report $\boldsymbol{y}_i$, and $\boldsymbol{y}_i^{\text{cls}} \in \{0,1\}^{|\mathbb{C}|}$ the corresponding abnormality labels over a label set $\mathbb{C}$. Given E frozen visual encoders, we model $p_\Theta(\boldsymbol{y} \mid \boldsymbol{x})$ with a three-stage framework.

3.1 Slot-wise Consolidation of CT Encoders

Frozen encoders produce heterogeneous token sequences $\boldsymbol{Z}_e \in \mathbb{R}^{N_e\times d_e}$ that differ in both length and feature dimension. FlexiCT and COLIPRI produce 8,000 and 13,824 tokens, respectively, while SigVLP varies with the number of CT slices. Directly concatenating these representations would largely increase the visual context supplied to the language model. We instead first resample each encoder independently to a common token budget and then combine the resulting representations.

For each encoder e, a linear projection between two layer-normalization operations maps its features to a common width $d = 512$, $\tilde{\boldsymbol{Z}}_e \in \mathbb{R}^{N_e\times d}$. We then apply an encoder-specific Perceiver-style resampler with $T = 128$ learned queries, producing $\boldsymbol{Q}_e = \mathcal{R}_e(\tilde{\boldsymbol{Z}}_e) \in \mathbb{R}^{T\times d}$. Each resampler

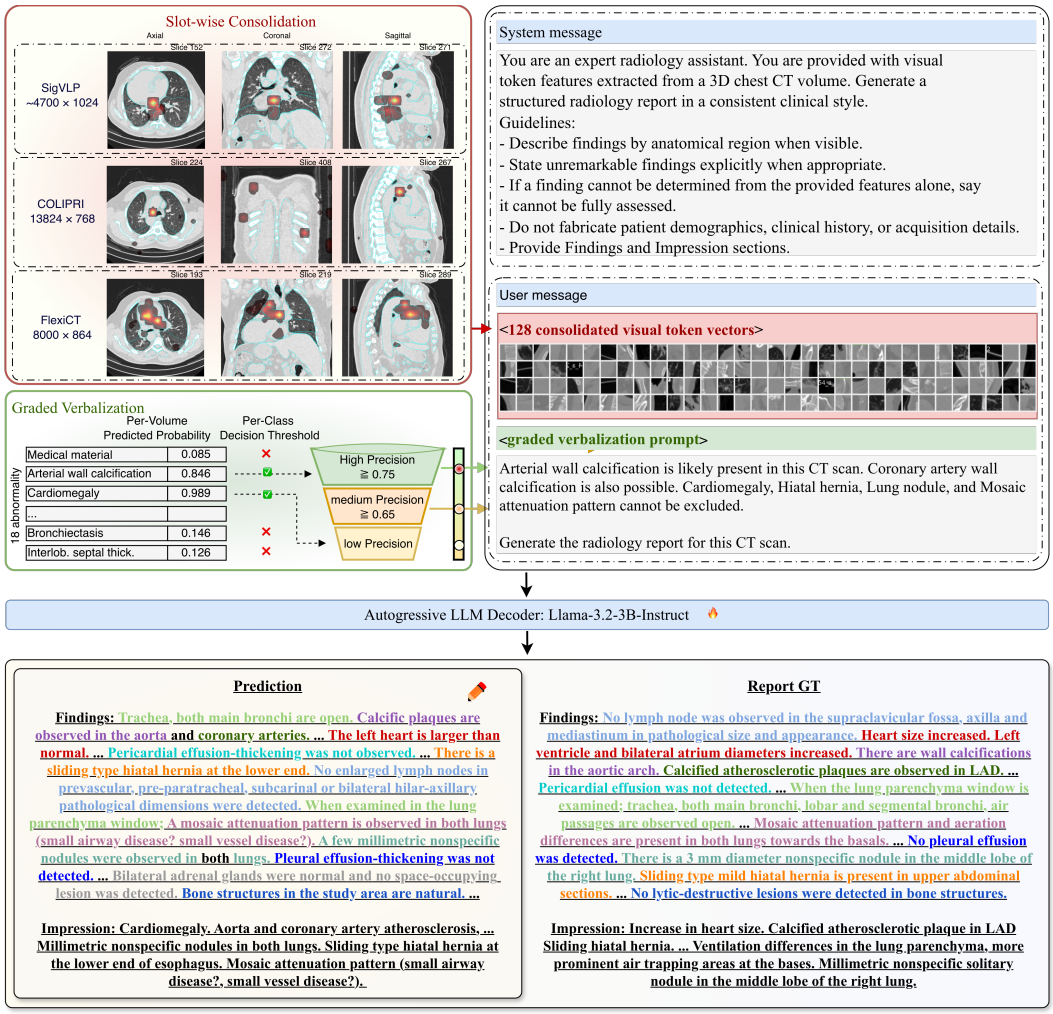


Figure 2: **Overview of the proposed report-generation framework.** Representations from multiple frozen CT encoders are independently resampled and consolidated into a 128-token visual prefix. The CT panels visualize generation-time spatial attention from each encoder after consolidation; for each anatomical view, the slice with the largest summed attention is displayed. Graded verbalization conditions generation on threshold-selected abnormalities with class-reliability-dependent linguistic certainty. The visual prefix and graded prompt jointly condition the language model. Matching colors in the example indicate corresponding clinical findings.

contains two cross-attention and feed-forward blocks with residual connections and eight attention heads. Using a separate resampler for each encoder allows each token distribution to be compressed independently. The resampled sequences are then consolidated slot by slot, where index t denotes a shared latent position across encoders in the visual prefix, and $\boldsymbol{Q}_{e,t}$ denotes the token produced by encoder e for slot t. We learn a separate set of encoder weights for each slot. Let $\boldsymbol{A} = \mathrm{softmax}_E(\boldsymbol{W}), \boldsymbol{W} \in \mathbb{R}^{T\times E}$, where the softmax is taken over the encoder dimension, so that $\boldsymbol{A}_{t,e} \geq 0$ and $\sum_{e=1}^{E} \boldsymbol{A}_{t,e} = 1$ for every slot t. The fused representation is

$$\boldsymbol{F}_t = \sum_{e=1}^{E} \boldsymbol{A}_{t,e}\boldsymbol{Q}_{e,t}, \qquad t = 1, \ldots, T, \tag{1}$$

with $\boldsymbol{F} \in \mathbb{R}^{T\times d}$. Each latent slot therefore learns its own weighting over the E encoders, while the fused sequence retains the token length and feature width of a single resampled encoder.

The weights $\boldsymbol{A}$ are shared across volumes rather than predicted separately for each input, so different latent slots learn different encoder preferences without introducing input-dependent routing. The encoder-specific resamplers and slot-wise mixing weights are optimized jointly under the report-generation objective, allowing the information assigned to the shared slots to be learned rather than prescribed. We initialize $\boldsymbol{W} = \mathbf{0}$, making the initial fusion exactly equivalent to uniform averaging across encoders. Deviations from uniform fusion are therefore learned during training rather than introduced by initialization.

An alignment adapter maps the fused representation $\boldsymbol{F}$ to the embedding width of the language model using a linear projection followed by a residual feed-forward block. This produces the visual prefix $\boldsymbol{P} \in \mathbb{R}^{T \times d_{\text{llm}}}$, which is prepended to the tokenized prompt and conditions report generation.

3.2 Graded Verbalization

The second stage converts classifier evidence into structured linguistic guidance for report generation. A multi-label classifier is trained separately from the report generator on frozen FlexiCT features. It pools the patch tokens using gated attention-based multiple-instance pooling and predicts a probability $p_c(\boldsymbol{x}) \in [0, 1]$ for each abnormality $c \in \mathbb{C}$. Classifier-predicted abnormalities that pass the calibrated class-specific thresholds are included in the prompt and expressed with assertion strengths determined by their class-level reliability.

For each class, we select a decision threshold θ_c that maximizes F_1 and record the corresponding precision π_c. At inference, the predicted finding set is $\hat{\mathbb{C}}(\boldsymbol{x}) = \{c \in \mathbb{C} : p_c(\boldsymbol{x}) \geq \theta_c\}$. Thus, θ_c controls *whether* a finding is verbalized, whereas π_c controls *how strongly* it is stated.

We discretise π_c into three assertion levels through a mapping τ: HIGH for $\pi_c \geq 0.75$, MED for $0.65 \leq \pi_c < 0.75$, and LOW otherwise. Selected findings are realized respectively as *likely present*, *possible*, and *cannot be excluded*. The reliability tier is class-specific and fixed across volumes; findings that do not pass the decision threshold are left unspecified rather than explicitly negated.

During training, prompt inclusion is determined by the positive CT-RATE labels, $\mathbb{C}^+(\boldsymbol{y}^{\text{cls}}) = \{c \in \mathbb{C} : y_c^{\text{cls}} = 1\}$, whereas inference uses the predicted set $\hat{\mathbb{C}}(\boldsymbol{x})$. The assertion mapping $\tau(\pi_c)$ is identical in both cases; only the source used to determine which findings enter the prompt changes.

3.3 Training

Classifier. The classifier of Section 3.2 is trained separately from the report generator, with the FlexiCT encoder frozen and only the gated pooling module and classification head optimized. For a training example $(\boldsymbol{x}, \boldsymbol{y}^{\text{cls}})$, it minimizes the class-weighted binary cross-entropy

$$\mathbb{L}_{\text{cls}} \;=\; -\sum_{c \in \mathbb{C}} \Big[\, w_c \, y_c^{\text{cls}} \log p_c(\boldsymbol{x}) \;+\; \big(1 - y_c^{\text{cls}}\big) \log\big(1 - p_c(\boldsymbol{x})\big) \Big], \qquad w_c \;=\; \frac{n_c^-}{n_c^+}, \tag{2}$$

where n_c^+ and n_c^- denote the numbers of positive and negative training volumes for class c. After training, the decision threshold θ_c for each class is calibrated on the early-stopping split by grid search to maximize F_1, and the corresponding precision π_c is recorded. The resulting θ_c and π_c remain unchanged at inference.

Report Generation. All three report-generation stages optimize the same autoregressive objective but differ in their conditioning prompts and trainable parameters. Let $\boldsymbol{s}^{(k)}$ denote the prompt used in stage k. For a report $\boldsymbol{y} = (y_1, \ldots, y_L)$, the objective is

$$\mathbb{L}(\Theta) \;=\; -\sum_{l=1}^{L} \log p_\Theta\big(y_l \mid \boldsymbol{y}_{<l},\, \boldsymbol{P}(\boldsymbol{x}),\, \boldsymbol{s}^{(k)}\big). \tag{3}$$

The loss is evaluated only over report tokens; the visual prefix and prompt serve as conditioning context. In Stage 1, $\boldsymbol{s}^{(1)}$ is a fixed report-generation instruction. The visual projections, encoder-specific resamplers, slot-wise weights $\boldsymbol{W}$, and alignment adapter are optimized, while the visual encoders and language model remain frozen. Stage 2 initializes from Stage 1 and optimizes the

same visual modules with the language model still frozen. The prompt $\boldsymbol{s}^{(2)}$ follows the graded-verbalization scheme of Section 3.2, with prompt inclusion determined by the positive reference abnormality labels during training. Stage 3 initializes from Stage 2 and retains the same graded conditioning. All visual modules are frozen, and only LoRA adapters inserted into the attention and feed-forward projections of the language model are optimized.

4 EXPERIMENTS

Datasets. We use two publicly available chest CT datasets, CT-RATE (Hamamci et al., 2026b) and RAD-ChestCT (Draelos et al., 2021), for our experiments. CT-RATE is one of the largest 3D medical imaging datasets with paired radiology reports. It contains 25,692 non-contrast chest CT scans from 21,304 patients and 50,188 reconstructed volumes, annotated with 18 report-derived abnormality labels. Following the official patient-level split, 1,304 patients (3,039 volumes) are used for internal validation. We further split the official training partition into 18,742 patients for model training and 1,258 patients (3,039 volumes) for early stopping. For external evaluation, we use RAD-ChestCT, an out-of-distribution set of 3,630 chest CT volumes with multi-label abnormality annotations. As free-text reports are unavailable in the public RAD-ChestCT release, this dataset is used only for evaluating clinical efficacy. Following Hamamci et al. (2026b), we map the RAD-ChestCT label space to the corresponding CT-RATE abnormality categories (full mapping provided in Appendix A) to enable cross-institutional generalization evaluation, yielding 16 scored classes.

Evaluation Metrics. We evaluate report generation performance in terms of clinical efficacy (CE) and natural language generation (NLG) quality. For CE, generated reports are converted into 18 binary abnormality labels using a RadBERT-based classifier and compared with the ground-truth labels. We report macro precision, recall, F_1, and accuracy to quantify agreement in clinically relevant findings. We additionally report the CRG score (Hamamci et al., 2025), which excludes true negatives and applies prevalence-aware weighting to penalize missed and spurious findings under class imbalance. For NLG quality, we report BLEU-1, BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Lavie & Denkowski, 2009) to measure lexical and sequence-level similarity between generated and reference reports.

Implementation Details. We use Llama-3.2-3B-Instruct (Grattafiori et al., 2024) as the language backbone and operate it in bfloat16 with gradient checkpointing. Training is performed on four NVIDIA A100 GPUs. All generator stages use AdamW with an initial learning rate of 10^{-4}, a 50-step linear warm-up followed by cosine decay to 10^{-6}, gradient clipping at 1.0, and a per-GPU batch size of one (effective batch size of four). The LoRA adapters use $r = 16$, $\alpha = 32$, and dropout 0.05, and are applied to all attention and MLP projections. The auxiliary classifier is trained separately with AdamW at 10^{-3}. Reports are generated using greedy decoding with at most 512 new tokens. For all baselines, we follow the official implementations and recommended inference settings, additional preprocessing details and baseline reproduction details are provided in Appendix B and C, respectively.

4.1 RESULTS

Comparison with state-of-the-art methods. Table 2 reports the quantitative results on the CT-RATE validation split. Our method achieves the best overall clinical performance, reaching 0.526 CRG and 0.597 macro-F_1, substantially improving over the strongest prior baseline, AdaRAG-CT (0.447 CRG and 0.414 F_1). Importantly, both precision and recall improve over all baselines, showing that the gain is not obtained by simply shifting the operating point toward either over-reporting or under-reporting findings. Table 11 reports per-class clinical efficacy for all 18 CT-RATE abnormalities. In terms of linguistic quality, our method also achieves the best BLEU-1 (0.487), BLEU-4 (0.242), and ROUGE-L (0.343) scores, while remaining competitive in METEOR (0.249 vs. 0.261 for AdaRAG-CT).

External validation on RAD-ChestCT. We further evaluate cross-dataset generalization on the external RAD-ChestCT dataset. As shown in Table 3, our method achieves the best external performance with 0.498 macro-F_1, 0.487 precision, and 0.622 recall, outperforming BTB3D-16 and CT-CHAT by 123.3% in F_1 and by 167.7% relative macro-F_1, respectively. Notably, recall remains nearly identical to that on CT-RATE (0.622 vs. 0.616), whereas precision decreases from 0.594 to

Table 2: **Report generation performance on the CT-RATE validation split.** All baseline results are reproduced or re-evaluated under our unified protocol; values may differ from originals. Results are reported with 95% bootstrap confidence intervals estimated using the percentile method over 1,000 resamples; METEOR is reported without intervals. Best results are highlighted in **bold**.

Method	Clinical Efficacy					NLG			
	CRG	F_1	Precision	Recall	Accuracy	BLEU-1	BLEU-4	ROUGE-L	METEOR
RadFM (Wu et al., 2025)	0.338 [0.337, 0.340]	0.036 [0.033, 0.038]	0.168 [0.050, 0.181]	0.043 [0.040, 0.047]	0.784 [0.779, 0.789]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.035 [0.034, 0.036]	0.018
M3D (Bai et al., 2024)	0.340 [0.339, 0.342]	0.112 [0.104, 0.119]	0.223 [0.197, 0.253]	0.084 [0.078, 0.090]	0.772 [0.768, 0.778]	0.003 [0.002, 0.003]	0.000 [0.000, 0.000]	0.057 [0.056, 0.058]	0.028
CT-CHAT (Hamamci et al., 2026b)	0.367 [0.364, 0.369]	0.170 [0.161, 0.177]	0.366 [0.320, 0.411]	0.158 [0.151, 0.165]	0.786 [0.781, 0.791]	0.389 [0.381, 0.396]	0.204 [0.197, 0.212]	0.333 [0.326, 0.341]	0.219
BTB3D-16 (Hamamci et al., 2026a)	0.354 [0.351, 0.357]	0.213 [0.204, 0.221]	0.210 [0.200, 0.219]	0.234 [0.224, 0.245]	0.703 [0.697, 0.708]	0.359 [0.354, 0.364]	0.130 [0.127, 0.134]	0.233 [0.230, 0.236]	0.193
Ker-VLJEPA-3B (Bumgardner et al., 2026)	0.393 [0.390, 0.396]	0.332 [0.324, 0.341]	0.246 [0.237, 0.253]	0.594 [0.583, 0.604]	0.586 [0.582, 0.591]	0.318 [0.313, 0.322]	0.068 [0.066, 0.069]	0.174 [0.173, 0.175]	0.169
AdaRAG-CT-base (Liang et al., 2026)	0.429 [0.426, 0.433]	0.391 [0.381, 0.400]	0.422 [0.409, 0.434]	0.392 [0.382, 0.402]	0.795 [0.790, 0.799]	0.436 [0.430, 0.443]	0.195 [0.189, 0.201]	0.297 [0.291, 0.302]	0.221
AdaRAG-CT (Liang et al., 2026)	0.447 [0.443, 0.451]	0.414 [0.405, 0.424]	0.451 [0.435, 0.468]	0.459 [0.449, 0.470]	0.801 [0.796, 0.806]	0.474 [0.470, 0.478]	0.235 [0.230, 0.240]	0.333 [0.328, 0.339]	**0.261**
Ours	**0.526** [0.521, 0.531]	**0.597** [0.587, 0.607]	**0.594** [0.582, 0.605]	**0.616** [0.604, 0.627]	**0.849** [0.845, 0.853]	**0.487** [0.482, 0.492]	**0.242** [0.236, 0.248]	**0.343** [0.337, 0.349]	0.249

Table 3: **External validation on RAD-ChestCT.** All baseline results are reproduced and evaluated under our unified protocol; values may differ from originals. Results are reported with 95% bootstrap confidence intervals estimated using the percentile method over 1,000 resamples. Best results in **bold**.

Method	F_1	Precision	Recall
RadFM	0.059 [0.057, 0.062]	0.268 [0.160, 0.307]	0.101 [0.098, 0.103]
M3D	0.152 [0.145, 0.159]	0.305 [0.289, 0.322]	0.113 [0.107, 0.119]
CT-CHAT	0.186 [0.182, 0.190]	0.203 [0.174, 0.233]	0.303 [0.300, 0.306]
BTB3D-16	0.223 [0.216, 0.230]	0.272 [0.263, 0.280]	0.222 [0.214, 0.231]
Ours	**0.498** [0.490, 0.505]	**0.487** [0.478, 0.494]	**0.622** [0.612, 0.631]

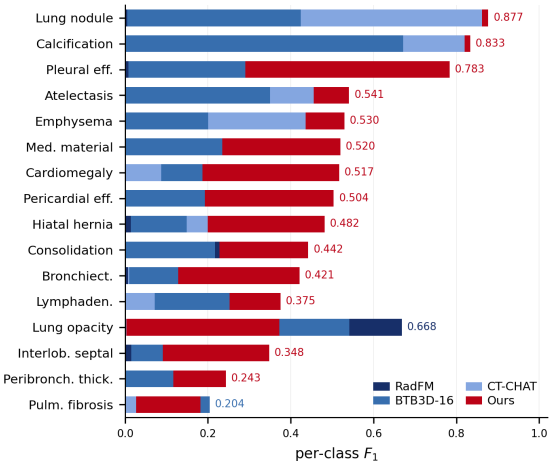


Figure 3: **Per-class F_1 on RAD-ChestCT.** Classes by the prevalence ratio of each abnormality.

0.487. Figure 3 provides a per-class analysis on RAD-ChestCT. Our method achieves the highest F_1 in 14 of the 16 scored classes, including several lower-prevalence findings such as pericardial effusion and hiatal hernia. Lung opacity is a notable exception to the otherwise broad improvement. Our model achieves relatively high precision (0.728) but substantially lower recall (0.251), yielding an F_1 of 0.373.

Ablation study. Table 4 evaluates the contributions of graded verbalization and LoRA adaptation while keeping the consolidated visual prefix fixed. LoRA alone leaves clinical efficacy within the confidence intervals of the baseline (F_1: 0.410→0.421), although it improves all four NLG metrics. Introducing the graded-verbalization stage produces the largest clinical improvement, increasing macro-F_1 from 0.410 to 0.519, recall from 0.343 to 0.472, and CRG from 0.433 to 0.474, with precision remaining stable (0.584→0.600). When LoRA is applied together with graded verbalization, F1 increases further to 0.597 and recall to 0.616, while precision remains essentially unchanged. The full model's intervals do not overlap with either single-component variant, indicating that the two stages are complementary, *i.e.*, graded verbalization provides the generator with threshold-selected diagnostic findings expressed at reliability-dependent assertion levels, whereas LoRA adaptation further improves how effectively this structured guidance is incorporated into the generated report.

Closing the verbalization gap across abnormalities. Figure 4 quantifies how much of the observed representation-to-report gap is recovered across individual abnormalities. The baseline generator reaches 0.410 macro F1, compared with 0.636 for the auxiliary classifier, leaving a gap of 0.226. Graded verbalization raises report F1 to 0.519 and narrows this gap substantially. The full model

Table 4: **Ablation study on the CT-RATE validation split.** All variants share the same fused visual prefix and language model. Results are reported with 95% bootstrap confidence intervals estimated using the percentile method over 1,000 resamples.

Component		Clinical Efficacy					NLG			
Graded Verb.	LoRA	CRG	F_1	Precision	Recall	Accuracy	BLEU-1	BLEU-4	ROUGE-L	METEOR
		0.433 [0.429, 0.436]	0.410 [0.400, 0.420]	0.584 [0.566, 0.602]	0.343 [0.334, 0.352]	0.836 [0.831, 0.841]	0.411 [0.405, 0.418]	0.206 [0.200, 0.212]	0.328 [0.321, 0.334]	0.216
	✓	0.433 [0.430, 0.437]	0.421 [0.409, 0.431]	0.588 [0.570, 0.605]	0.351 [0.342, 0.362]	0.835 [0.831, 0.840]	0.426 [0.420, 0.433]	0.221 [0.215, 0.228]	**0.343** [0.336, 0.349]	0.231
✓		0.474 [0.469, 0.479]	0.519 [0.507, 0.529]	**0.600** [0.587, 0.613]	0.472 [0.460, 0.483]	0.843 [0.839, 0.847]	0.438 [0.433, 0.444]	0.217 [0.211, 0.223]	0.333 [0.327, 0.339]	0.227
✓	✓	**0.526** [0.521, 0.531]	**0.597** [0.587, 0.607]	0.594 [0.582, 0.605]	**0.616** [0.604, 0.627]	**0.849** [0.845, 0.853]	**0.487** [0.482, 0.492]	**0.242** [0.236, 0.248]	**0.343** [0.337, 0.349]	**0.249**

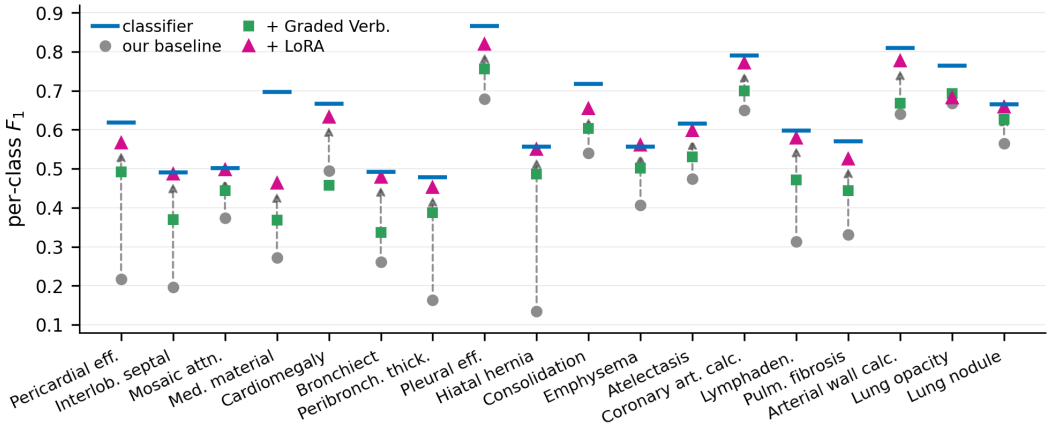


Figure 4: **Closing the observed representation-to-report gap across CT-RATE abnormalities.** Per-class F1 for the auxiliary classifier and report-generation variants across 18 abnormalities, ordered by validation-set positive support. Blue ticks give the F_1 of the classifier; dashed arrows indicate the baseline-to-full-model change. The full model improves all 18 abnormalities and reduces the macro-F1 gap to the classifier from 0.226 to 0.039.

reaches 0.597, leaving a gap of 0.039, approximately 83% smaller than the baseline-to-classifier gap, and improves all 18 abnormalities relative to the baseline. At this stage, macro precision is similar to that of the classifier (0.594 vs. 0.592), while most of the remaining difference is in recall (0.616 vs. 0.697). The improvement is not confined to high-support abnormalities. The nine classes with the lowest validation-set support improve by 0.239 F1 on average, compared with 0.135 for the nine higher-support classes, with particularly large gains for hiatal hernia, pericardial effusion, and peribronchial thickening. Missed findings fall from 6,256 to 3,642 while false positives rise from 2,697 to 4,643, reducing the missed-to-spurious finding ratio from 2.3 to 0.78.

Comparison of consolidation strategies. Table 5 compares several multi-encoder fusion strategies before report generation. Shared resampling follows BRAVE (Kar et al., 2024) in concatenating the projected encoder tokens before a single joint resampler, channel concatenation follows Eagle (Shi et al., 2025a) in fusing along the feature dimension after resampling each encoder independently, and sequence concatenation instead appends the independently resampled tokens along the sequence dimension. Among the single encoders, FlexiCT performs best clinically, with an F_1 of 0.366 at a precision of 0.629. Simply combining encoders does not guarantee improvement and averaging the three reduces F_1 to 0.325, while shared resampling and the two concatenation variants reach 0.370 to 0.375. In contrast, our method reaches an F_1 of 0.410 and a CRG of 0.433, improving F_1 by 12.0% over the best single encoder and by 9.3% over the strongest alternative fusion strategy. On NLG metrics sequence concatenation obtains the best scores, and no fusion strategy surpasses the single encoders on ROUGE-L.

Table 5: **Comparison of consolidation strategies on the CT-RATE validation split.** All variants use the same three encoders, the same language model and the same training schedule. Prefix gives the shape of the sequence passed to the language model.

		Clinical Efficacy					NLG			
Method	Prefix	CRG	F_1	Precision	Recall	Accuracy	BLEU-1	BLEU-4	ROUGE-L	METEOR
Single encoder, native resolution										
COLIPRI	$13{,}824 \times 768$	0.407	0.358	0.574	0.272	0.835	0.385	0.204	0.339	0.213
FlexiCT	$8{,}000 \times 864$	0.409	0.366	**0.629**	0.275	0.836	0.390	0.207	**0.342**	0.218
SigVLP	$N \times 1024$	0.384	0.265	0.462	0.204	0.815	0.397	0.207	0.332	0.214
Three encoders										
Average	128×512	0.413	0.325	0.564	0.265	0.828	0.411	0.207	0.323	0.216
Shared resampler	128×512	0.418	0.370	0.589	0.298	0.832	0.412	0.200	0.315	0.211
Sequence concat.	384×512	0.420	0.372	0.610	0.295	**0.840**	**0.426**	**0.221**	0.336	**0.222**
Channel concat.	128×512	0.422	0.375	0.552	0.315	0.830	0.424	0.213	0.329	0.220
Ours	128×512	**0.433**	**0.410**	0.584	**0.343**	0.836	0.411	0.206	0.328	0.216

Table 6: **Sensitivity to the language backbone and the token budget on the CT-RATE validation split.** All rows use slot-wise mixing without graded verbalization or LoRA, and differ only in the language model (top) or in T (bottom).

Setting	Clinical Efficacy					NLG			
	CRG	F_1	Precision	Recall	Accuracy	BLEU-1	BLEU-4	ROUGE-L	METEOR
Llama-3.2-3B	0.433	0.410	0.584	0.343	0.836	0.411	0.206	0.328	0.216
Llama-3.1-8B	0.414	0.371	0.551	0.305	0.822	0.408	0.190	0.300	0.209
Qwen2.5-3B	0.434	0.407	0.560	0.361	0.830	0.442	0.209	0.309	0.219
Token budget, Llama-3.2-3B									
$T = 64$	0.417	0.355	0.607	0.284	0.836	0.421	0.226	0.350	0.222
$T = 128$	0.433	0.410	0.584	0.343	0.836	0.411	0.206	0.328	0.216
$T = 256$	0.418	0.366	0.596	0.290	0.834	0.404	0.209	0.332	0.216
$T = 512$	0.423	0.381	0.571	0.317	0.830	0.419	0.205	0.322	0.217
$T = 1024$	0.423	0.389	0.582	0.319	0.835	0.409	0.205	0.325	0.218

Effect of language backbone and token budget. Table 6 evaluates whether the baseline performance depends on the language backbone or the size of the consolidated visual prefix. Replacing Llama-3.2-3B with the larger Llama-3.1-8B does not improve clinical efficacy (F_1: 0.410 vs. 0.371), while Qwen2.5-3B (Yang et al., 2024) yields comparable clinical performance (F_1: 0.407, CRG: 0.434). Simply scaling the language model does not close the clinical gap. Varying the token budget reveals a similar non-monotonic trend: $T = 128$ achieves the highest clinical F_1 (0.410) and CRG (0.433), whereas increasing the prefix to as many as 1,024 tokens provides no further benefit. Interestingly, $T = 64$ yields stronger NLG scores but substantially lower recall (0.284), suggesting that lexical similarity does not necessarily translate into more complete clinical reporting. Together, these controls suggest that the observed clinical gap is not resolved by simply increasing language-model capacity or visual-token budget within the tested range, motivating an explicit mechanism for carrying recoverable findings into the report.

5 Conclusion

We identify a verbalization gap in 3D CT report generation, where clinically relevant abnormalities can remain recoverable from frozen visual representations yet fail to appear in generated reports. We address both evidence availability and evidence expression by consolidating complementary CT representations and conditioning generation on reliability-aware diagnostic guidance. On CT-RATE, the full system reaches 0.597 macro-F_1 and reduces the observed gap to the auxiliary classifier from 0.226 to 0.039, while retaining strong performance on mapped findings in RAD-ChestCT. These results suggest that improving 3D CT radiology reporting requires not only stronger visual evidence, but also mechanisms that reliably carry recoverable clinical information into language.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods, provide feedback on research methodology and experiments, assist with translation, support qualitative and thematic data analysis, and interpret results. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, propose or refine hypotheses, clean and reformat datasets; and generating synthetic data sets, formulating mathematical claims, providing critical ingredients for proving mathematical claims, and assisting in the writing of proofs are not applicable to this work. Additionally, we used generative AI tools to create and modify scientific figures, suggest experimental parameters, create and edit software code, draft parts of the research paper, summarize and analyze existing literature, discover research topics or identify gaps, brainstorm, source and search for information, identify relevant literature, and edit the research paper to improve readability. We have reviewed all AI-assisted work. LLM-generated code scaffolds were produced from detailed author specifications and were verified, corrected, and tested by the authors. AI-generated suggestions on experimental settings, result interpretation, related literature, and figure preparation were adopted only selectively and checked by the authors. AI-assisted drafting, editing, and translation were substantially revised by the authors before inclusion in the manuscript. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study uses only publicly released CT datasets and involves no new patient data collection or human-subject experiments. The proposed system is intended for research on automated radiology report generation and should not replace radiologist interpretation or independent clinical decision-making.

REPRODUCIBILITY STATEMENT

We provide complete descriptions of the model architecture, training objectives, and optimization procedure in Sections 3–3.3, and report dataset splits, evaluation metrics, and implementation settings in Section 4. Appendix B details encoder-specific preprocessing, Appendix A specifies the cross-dataset label mapping, and Appendix C documents baseline reproduction under the unified evaluation protocol. All datasets and pretrained models used in this work are publicly available and cited in the paper.

REFERENCES

Mozhgan Nasr Azadani, James Riddell, Sean Sedwards, and Krzysztof Czarnecki. Rethinking the Mixture of Vision Encoders Paradigm for Enhanced Visual Understanding in Multimodal LLMs. *Transactions on Machine Learning Research*, 2026. URL https://mlanthology.org/tmlr/2026/azadani2026tmlr-rethinking/.

Mohammed Baharoon, Jun Ma, Congyu Fang, Augustin Toma, and Bo Wang. Exploring the design space of 3d mllms for ct report generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15965 of *Lecture Notes in Computer Science*, pp. 237–246. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-032-04978-0_23.

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, 2024.

David Bani-Harouni, Chantal Pellegrini, Julian Lüers, Su Hwan Kim, Markus Baalmann, Benedikt Wiestler, Rickmer Braren, Nassir Navab, and Matthias Keicher. Calibrated confidence expression for radiology report generation. *arXiv preprint arXiv:2603.29492*, 2026.

V. K. Cody Bumgardner, Mitchell A. Klusty, Mahmut S. Gokmen, and Evan W. Damron. Curriculum-driven 3d ct report generation via language-free visual grafting and zone-constrained compression, 2026. URL https://arxiv.org/abs/2603.23308.

Sheng Cheng and Devika Subramanian. Rethinking radiology report generation: From narrative flow to topic-guided findings. In *The Fourteenth International Conference on Learning Representations*, 2026.

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. *Medical image analysis*, 67:101857, 2021.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. CT2Rep: Automated radiology report generation for 3D medical imaging. In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15012 of *Lecture Notes in Computer Science*, pp. 476–486. Springer Nature Switzerland, 2024.

Ibrahim Ethem Hamamci, Sezgin Er, Suprosanna Shit, Hadrien Reynaud, Bernhard Kainz, and Bjoern Menze. CRG score: A distribution-aware clinical metric for radiology report generation. In *Medical Imaging with Deep Learning-Short Papers*, 2025.

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*, 10:1610–1628, 2026b. doi: 10.1038/s41551-025-01599-y.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Abdullah Hosseini, Ahmed Ibrahim, and Ahmed Serag. From slices to volumes: Multi-scale fusion of 2d and 3d features for ct scan report generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15965 of *Lecture Notes in Computer Science*, pp. 268–277. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-032-04978-0_26.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022. URL https://mlanthology.org/iclr/2022/hu2022iclr-lora/.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642, 2024.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. PromptMRG: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 2607–2615, 2024.

Oğuzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. Brave: Broadening the visual encoding of vision-language models. In *European Conference on Computer Vision*, pp. 113–132. Springer, 2024.

Alon Lavie and Michael J. Denkowski. The meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23(2-3):105–115, 2009.

Yuheng Li, Yuan Gao, Haoyu Dong, Yuxiang Lai, Shansong Wang, Mojtaba Safari, James E Baciak, and Xiaofeng Yang. Universal ct representations from anatomy to disease phenotype through agglomerative pretraining. *arXiv preprint arXiv:2605.21906*, 2026.

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, 2026. URL https://arxiv.org/abs/2603.15822.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, 2004.

Thomas J. Littlejohns, Jo Holliday, Lorna M. Gibson, et al. The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nature Communications*, 11:2624, 2020. doi: 10.1038/s41467-020-15948-9.

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.

Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12490–12500, June 2024.

Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, Yilin Zhao, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, Bryan Catanzaro, Andrew Tao, Jan Kautz, Zhiding Yu, and Guilin Liu. Eagle: Exploring the Design Space for Multimodal LLMs with Mixture of Encoders. In *International Conference on Learning Representations*, 2025a. URL `https://mlanthology.org/iclr/2025/shi2025iclr-eagle/`.

Yanzhao Shi, Xiaodan Zhang, Junzhong Ji, Haoning Jiang, Chengxin Zheng, Yinong Wang, and Liangqiong Qu. Hsenet: Hybrid spatial encoding network for 3d medical vision-language understanding. *arXiv preprint arXiv:2506.09634*, 2025b.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Jonathan Suprijadi, Raphael Stock, Moritz Langenberg, David Zimmerer, Kim-Celine Kahl, Stefan Denner, Yannick Kirchhoff, Karol Gotkowski, Maximilian Rokuss, Jeremias Traub, Tassilo Wald, Constantin Ulrich, and Klaus Maier-Hein. Resolution meets reduction: Efficient visual context for 3d radiology report generation. *arXiv preprint arXiv:2608.08713*, 2026.

Yuhe Tian, Kun Zhang, Haoran Ma, Rui Yan, Yingtai Li, Rongsheng Wang, and Shaohua Kevin Zhou. Diffvp: Differential visual semantic prompting for llm-based ct report generation. In *European Conference on Computer Vision*, pp. 268–285. Springer, 2026.

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. *Advances in Neural Information Processing Systems*, 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. *arXiv preprint arXiv:2510.15042*, 2025.

Jiayi Wang, Hadrien Reynaud, Ibrahim Ethem Hamamci, Sezgin Er, Suprosanna Shit, Bjoern Menze, and Bernhard Kainz. Sigvlp: Sigmoid volume-language pre-training for self-supervised ct-volume adaptive representation learning. *arXiv preprint arXiv:2602.21735*, 2026.

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.

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.

Sixing Yan, Ziao Wang, Kejing Yin, William K. Cheung, Ka Chun Cheung, and Simon See. Learning self-critiquing mechanisms for region-guided chest x-ray report generation. In *The Fourteenth International Conference on Learning Representations*, 2026.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

A CT-RATE TO RAD-CHESTCT LABEL CORRESPONDENCE

RAD-ChestCT provides abnormality-location annotations over 84 abnormalities and 52 anatomical locations. We adopt the CT-RATE to RAD-ChestCT label mapping provided by the CT-CLIP authors for external validation, as summarized in Table 7.

Table 7: **CT-RATE to RAD-ChestCT label correspondence used for external evaluation.** Mosaic attenuation has no counterpart and is excluded from scoring. Arterial wall calcification and coronary artery wall calcification map to the same RAD-ChestCT labels and are scored once as arterial wall calcification, yielding 16 classes.

CT-RATE class	**RAD-ChestCT abnormalities**
Medical material	pacemaker_or_defib, catheter_or_port, hardware, stent, suture, staple, chest_tube, tracheal_tube, gi_tube, clip, breast_implant, heart_valve_replacement
Arterial wall calcification	calcification, scattered_calc
Cardiomegaly	cardiomegaly
Pericardial effusion	pericardial_effusion
Coronary artery wall calcification	calcification, scattered_calc
Hiatal hernia	hernia
Lymphadenopathy	lymphadenopathy
Emphysema	emphysema
Atelectasis	atelectasis
Lung nodule	nodule, nodulegr1cm, scattered_nod
Lung opacity	opacity
Pulmonary fibrotic sequela	fibrosis
Pleural effusion	pleural_effusion
Mosaic attenuation pattern	- (excluded)
Peribronchial thickening	bronchial_wall_thickening
Consolidation	consolidation
Bronchiectasis	bronchiectasis
Interlobular septal thickening	septal_thickening

B ADDITIONAL IMPLEMENTATION DETAILS

Each encoder is applied using its pretrained preprocessing pipeline, while all pretrained weights, classifier thresholds, and model hyperparameters are kept fixed across datasets. Table 8 summarizes the preprocessing used for CT-RATE. For RAD-ChestCT, only the data-loading and spacing-handling steps are adapted to account for differences in the released data format. Specifically, RAD-ChestCT volumes are provided as integer arrays with a known isotropic spacing of $0.8\,\text{mm}$ and are already clamped to $[-1000, 1000]$. SigVLP requires no spacing adaptation. For COLIPRI, we construct an affine using the known $0.8\,\text{mm}$ isotropic spacing before applying its standard orientation and resampling pipeline. For FlexiCT, volumes are resampled to $2\,\text{mm}$ isotropic spacing using trilinear interpolation, with all subsequent preprocessing unchanged.

The auxiliary classifier is trained independently on the official 18-label CT-RATE annotations using weighted binary cross-entropy, with the positive weight w_c of Eq. 2. We use AdamW with a learning

Table 8: **Encoder-specific preprocessing and tokenization for CT-RATE.** All three pipelines clamp CT intensities to $[-1000, 1000]$ HU before normalization. Token counts are per volume; D denotes the number of axial slices.

	SigVLP	**COLIPRI**	**FlexiCT**
Spacing	none	2 mm isotropic	2 mm isotropic
Orientation	none	SAR	none
Intensity	linear to $[-1, 1]$	linear to $[-1, 1]$	volume-wise z-score
Geometry	H, W to 256^2; centre-crop if $D > 800$	crop or pad to 192^3, minimum padding	symmetric pad to 160^3,then centre-crop
Input	$D \times 256 \times 256$	192^3	160^3
Patch	16^3	8^3	8^3
Tokens	$\lfloor D/16 \rfloor \times 16 \times 16$	13,824	8,000
Width	1024	768	864

rate of 10^{-3}, weight decay of 10^{-4}, a 600-step linear warm-up followed by cosine decay, and an effective batch size of 128.

C Baselines and reproduction

For comparisons under the unified evaluation protocol, the baselines are implemented according to their official codes and pre-trained weights. Generated reports are evaluated using the same clinical-efficacy and NLG pipelines under our unified evaluation protocol, including the RadBERT-based clinical label extraction and text preprocessing. We modify released implementations only when required to accommodate the dataset format or to resolve compatibility issues that otherwise prevent inference.

RadFM (Wu et al., 2025). We use the released checkpoint and official 3D report-generation prompt. The volumes already store rescaled HU values, no additional CSV intensity correction is applied; all other settings follow the released implementation. For RAD-ChestCT, the released volumes are passed through the same inference pipeline without additional intensity correction.

M3D (Bai et al., 2024). We follow the released inference configuration. CT-RATE volumes loaded from NIfTI are represented as (H, W, D), whereas M3D expects the slice dimension first; we therefore transpose them to (D, H, W) before applying the official preprocessing. No other model-specific changes are made. For RAD-ChestCT, the released volumes are already stored in (D, H, W) order, so this transpose is unnecessary; all other settings remain unchanged.

CT-CHAT (Hamamci et al., 2026b). We use the official checkpoint, inference pipeline, and the report-generation turns from the released `valid_vqa.json`. Two compatibility fixes are required by the public release: the stale vision-tower import is redirected to the released encoder module, and the token embeddings are resized after adding the four task tokens so that the released non-LoRA weights can be loaded. Neither change alters the intended model architecture or inference procedure. For RAD-ChestCT, we set the input spacing to the released 0.8 mm isotropic resolution and otherwise apply the same spatial and intensity preprocessing as for CT-RATE, without dataset-specific tuning.

Ker-VLJEPA-3B (Bumgardner et al., 2026). We use the released checkpoint together with the official preprocessing and decoding configuration without model-specific modification.

AdaRAG-CT (Liang et al., 2026). We use the report predictions released by the authors. The released predictions are matched to the CT-RATE validation cases and re-evaluated under the unified evaluation protocol. The values may differ from originals.

BTB3D-16 (Hamamci et al., 2026a). We reproduce BTB3D-16 using the official code and the released `16_16_8` checkpoint. As the released inference script does not specify the report-generation prompt, we use the per-volume natural-language prompts from CT-CHAT, following the setting described in the BTB3D paper. For RAD-ChestCT, we follow the official BTB3D preprocessing with two dataset-specific adaptations: we set `slope=1` and `intercept=0` because the released NPZ volumes are already in Hounsfield units, and use $(0.8, 0.8, 0.8)$ mm as the source spacing in place of CT-RATE's per-volume spacing metadata. We reproduce only the `16_16_8` variant; the `8_8_8` model expects 72-dim visual tokens, whereas the released encoding script produces 18-dim tokens and no corresponding preprocessing procedure is provided.

Table 9: **Robustness to tier boundaries on the CT-RATE validation split.** All rows use the same checkpoint initialisation, classifier, and training schedule, differing only in the precision boundaries $(\pi_{\text{hi}}, \pi_{\text{lo}})$ separating the HIGH, MED, and LOW tiers.

Tier Boundaries		Clinical Efficacy					NLG			
π_{hi}	π_{lo}	CRG	F_1	Precision	Recall	Accuracy	BLEU-1	BLEU-4	ROUGE-L	METEOR
0.65	0.50	0.469	0.507	0.603	0.453	**0.844**	0.434	0.209	0.324	0.223
0.70	0.55	0.466	0.497	0.594	0.445	0.842	0.433	0.210	0.326	0.223
0.70	0.60	0.469	0.514	0.600	0.461	0.842	0.435	0.213	0.328	0.224
0.75	0.55	0.471	0.509	0.600	0.461	0.844	**0.450**	**0.223**	**0.337**	**0.231**
0.75	0.60	0.474	0.517	0.604	0.467	0.844	0.432	0.210	0.327	0.224
0.75	0.65	**0.474**	**0.519**	0.600	**0.472**	0.843	0.438	0.217	0.333	0.227
0.80	0.55	0.463	0.499	**0.613**	0.441	0.843	0.440	0.222	0.337	0.227
0.80	0.60	0.473	0.517	0.606	0.467	0.843	0.441	0.215	0.330	0.226
0.80	0.65	0.468	0.508	0.605	0.453	0.843	0.438	0.213	0.327	0.225

D Qualitative Examples

Figure 5 provides a representative qualitative comparison between the reference report and reports generated by CT-CHAT, AdaRAG-CT, and our model. Among the compared methods, **Ours** provides more complete agreement with the reference, correctly reproducing both aortic measurements (41 mm ascending and 30 mm descending) and the atherosclerotic-calcification statement verbatim. It also captures the major emphysema, pulmonary nodule, pleural effusion, and sequelae findings more comprehensively than the competing models.

Some fine-grained discrepancies remain: both CT-CHAT and Ours omit the ground-truth qualifiers "paraseptal" and "calcific" for the emphysema and nodule findings, respectively; Ours describes the pleural effusion as unilateral on the right (7.5 mm) rather than bilateral and minimal, and reports sequelae changes without the ground-truth apical localization. AdaRAG-CT additionally localizes its sequelae finding to the right middle lobe, left lingular segment, and bilateral lower lobes rather than the apices.

Figure 6 complements this comparison by illustrating how the clinical content evolves across our pipeline. Compared with CT-CHAT, our baseline recovers the atelectatic changes missed by the external baseline and matches its coverage of the pulmonary nodules, but loses the emphysema finding captured by CT-CHAT. Adding Graded Verbalization recovers the emphysema and introduces the previously omitted sequelae finding. Subsequent LoRA adaptation further produces more detailed descriptions of the sequelae/fibrotic and atelectatic findings. The final report nevertheless retains residual errors, this progression illustrates that the main gain across stages comes from bringing previously omitted clinical findings into the report, while finer-grained localization and residual over-generation remain unresolved. These examples illustrate that our model can recover a larger fraction of the clinically relevant content and quantitative details in the reference report, while the remaining discrepancies are concentrated in finer-grained attributes.

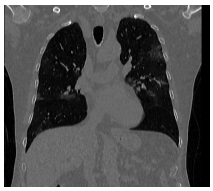
E Robustness to tier boundaries

Table 9 shows that graded verbalization is robust to the choice of tier boundaries. Across all tested settings, macro-F_1 remains between 0.497 and 0.519, consistently above the baseline without graded verbalization (0.410), with similar stability in CRG (0.463–0.474). The worst tested setting still exceeds AdaRAG-CT by 0.083 macro-F_1.

F LoRA configuration analysis

Table 10 varies one dimension of the LoRA configuration at a time. All tested configurations improve over the model without LoRA, with macro-F_1 ranging from 0.557 to 0.597. Performance varies little with adapter capacity: increasing the rank from 8 to 64 changes F_1 by only 0.007, while changing the target projections affects it by at most 0.018. In contrast, the learning rate has a larger

Ground Truth / Reference



Trachea and both main bronchi were in the midline and no obstructive pathology was detected in the lumen. In the non-contrast examination, the mediastinum was not evaluated optimally. As far as can be seen; **The anterior-posterior diameter of the ascending aorta was 41 mm, and the anterior-posterior diameter of the descending aorta was 30 mm, larger than normal**. Calibration of pulmonary arteries is natural. Heart contour, size is normal. Pericardial effusion-thickening was not observed. **Atherosclerotic wall calcifications were observed in the thoracic aorta and coronary arteries**. Thoracic esophagus calibration was normal and no significant tumoral wall thickening was detected. No enlarged lymph nodes were detected in prevascular, pre-paratracheal, subcarinal or bilateral hilar-axillary pathological dimensions. When examined in the lung parenchyma window; **Minimal pleural effusion was observed in both hemithorax**. **Multilobar-multisegmental vascular enlargement in both lungs, more diffuse central-peripheral vascular enlargement in the upper lobes and patchy ground glass consolidations with crazy paving pattern are observed, and the oulp appearance is consistent with Covid-19 pneumonia**. **Pleuroparenchymal fibroatelectasis sequelae changes were observed at the apical levels of both lungs**. **Some calcific millimetric nonspecific pulmonary nodules were observed in both lungs**. **Paraseptal emphysematous changes were observed in both apexes**. As far as can be seen on non-contrast sections, the upper abdominal organs are normal. No space-occupying lesion was detected in the liver that entered the cross-sectional area. Bilateral adrenal glands were normal and no space-occupying lesion was detected. Bone structures in the study area are natural. In the vertebra corpus end plateau, **degenerative osteophytic taperings were observed at the corners**. **Fusiform aneurysmatic dilatation in the thoracic aorta**, **atherosclerotic wall calcifications in the thoracic aorta and coronary arteries** **Bilateral smearing pleural effusion** and **findings consistent with Covid-19 pneumonia in the lung parenchyma** **Millimetric nonspecific pulmonary nodules in both lungs**, **reticuloseptic increase in apex, reticulonoidal fibromatous ammoniacal changes** **Mild spondylosis at the thoracic level**

Color key: aortic dilation (ascending/descending) atherosclerotic calcifications (aorta/coronary) minimal pleural effusion Covid-19 pattern (GGO/crazy paving) apical fibroatelectasis sequelae calcific millimetric nodules paraseptal emphysema degenerative / spondylosis changes

CT-Chat

Findings: Trachea, both main bronchi are open. Mediastinal main vascular structures, heart contour, size are normal. Thoracic aorta diameter is normal. Pericardial effusion-thickening was not observed. Thoracic esophagus calibration was normal and no significant tumoral wall thickening was detected. No enlarged lymph nodes in prevascular, pre-paratracheal, subcarinal or bilateral hilar-axillary pathological dimensions were detected. When examined in the lung parenchyma window; There are **emphysematous changes in both lungs**. There are linear atelectasis in the lower lobes of both lungs. There are **minimal pleural effusion** and atelectasis in the lower lobe of the right lung. There are **millimetric nonspecific nodules in both lungs**. Upper abdominal organs included in the sections are normal. No space-occupying lesion was detected in the liver that entered the cross-sectional area. Bilateral adrenal glands were normal and no space-occupying lesion was detected. Bone structures in the study area are natural. Vertebral corpus heights are preserved. Impression: **Emphysematous changes in both lungs.** Atelectasis in both lungs. **Minimal pleural effusion** and atelectasis in the lower lobe of the right lung. **Millimetric nonspecific nodules in both lungs.**

AdaRAG-CT

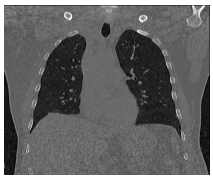
Trachea was in the midline of both main bronchi and no obstructive pathology was observed in its lumen. The mediastinum could not be evaluated optimally in the non-contrast examination. As far as can be seen; **The ascending aorta is wider than normal with an anterior-posterior diameter of 40 mm**. Calibration of other vascular structures of the mediastinum is natural. Heart contour, size is normal. Pericardial effusion-thickening was not observed. **Calcific atheroma plaques were observed on the walls of the thoracic aorta and coronary vascular structures**. Thoracic esophageal calibration was normal and no significant tumoral wall thickening was detected. No enlarged lymph nodes in prevascular, pre-paratracheal, subcarinal or bilateral hilar-axillary pathological dimensions were detected. When examined in the lung parenchyma window; **Multilobar-multisegmentary, central-peripheral localized crazy paving pattern accompanied by linear atelectasis and patchy ground glass consolidations showing signs of vascular enlargement were observed in both lungs, and the appearance is compatible with Covid-19 pneumonia during the resolution period**. It is recommended to evaluate clinically and laboratory together. Linear pleuroparenchymal fibroatelectatic sequelae changes were observed in the right lung middle lobe medial segment, left lung upper lobe inferior lingular segments, and lower lobes of both lungs. A mosaic attenuation pattern was observed in both lungs (small airway disease?, small vessel disease?). Liver contours show microlobulation as far as can be seen on non-contrast images. Left adrenal gland locus is hypodense lesion measuring approximately 20x15 mm in size. It has the appearance of an unenhanced fat. Bone structures in the study area are natural. Vertebral corpus heights are preserved. **Fusiform ectasia in the ascending aorta.** **Findings consistent with Covid-19 pneumonia in the resolution period in the lung parenchyma.** Mosaic attenuation pattern in both lungs (small airway disease?, small vessel disease?). **Sequelae changes in both lungs.** Microlobulation in liver contours, In the lateral segment of the left lobe of the liver, a stable hypodense area smaller than 5 mm in nonspecific appearance, which was also observed in the previous examination, is observed (cyst?). The left kidney cannot be evaluated within the limits of the examination. **Diffuse degenerative changes in bone structure.**

Ours

Trachea, both main bronchi are open. No occlusive pathology was observed in the lumen. The mediastinum could not be evaluated optimally in the non-contrast examination. As far as can be seen; **The anterior-posterior diameter of the ascending aorta was 41 mm, and the anterior-posterior diameter of the descending aorta was 30 mm**. Calibration of pulmonary arteries is natural. Heart contour, size is normal. Pericardial effusion-thickening was not observed. **Atherosclerotic wall calcifications were observed in the thoracic aorta and coronary arteries**. Thoracic esophagus calibration was normal and no significant pathological wall thickening was detected. Sliding type hiatal hernia was observed at the lower end of the esophagus. No enlarged lymph nodes in prevascular, pre-paratracheal, subcarinal or bilateral hilar-axillary pathological dimensions were detected. When examined in the lung parenchyma window; **Minimal emphysematous changes were observed in both lungs**. **Sequelae parenchymal changes were observed in both lungs**. **Multiple millimetric nonspecific nodules were observed in both lungs**. In the current examination, there are newly developed **ground-glass densities in both lungs, which are more clearly observed in the left upper lobe and bilateral lower lung lobes, and viral pneumonias are considered in the etiology of the findings. It is recommended to be evaluated together with clinical and laboratory findings in terms of Covid-19 pneumonia**. **A free pleural effusion measuring 7.5 mm was observed on the right at its deepest point**. No mass lesions were detected in both lungs. Upper abdominal organs included in the sections are normal. No space-occupying lesion was detected in the liver that entered the cross-sectional area. Bilateral adrenal glands were normal and no space-occupying lesion was detected. Calcified atheroma plaques were observed in the abdominal aorta and visceral branches. Minimal thickening was observed in the left adrenal gland. **Minimal osteodegenerative changes were observed in the thoracic vertebrae.** **Fusiform aneurysmatic dilatation in the thoracic aorta**, **atherosclerotic wall calcifications in the thoracic aorta and coronary arteries.** Hiatal hernia. **Emphysematous changes in both lungs**, **sequela parenchymal changes**, **millimetric nonspecific nodules in both lungs.** Newly developed **ground-glass densities in both**

Figure 5: **Qualitative comparison of CT report generation.** Colors indicate corresponding clinical findings across the reference and generated reports. The comparison highlights differences in clinical content coverage, quantitative measurements, and fine-grained descriptive details among the methods.

Ground Truth / Reference



Trachea and both main bronchi are open. No occlusive pathology was detected in the trachea and both main bronchi. There are appearances compatible with **pleuroparenchymal sequelae changes in both lung apexes, more prominent on the right**. There are appearances compatible with **sequelae changes in both lung lower lobes**. **Atelectasis, which was understood to be due to osteophyte compression, was observed in the medial superior segment of the lower lobe of the right lung. A similar appearance is also observed in the neighborhood of the anterior segment of the upper lobe of the right lung.** There are sometimes **linear atelectasis in both lungs**. There are **emphysematous changes in both lungs**. **Millimetric nodules were observed in both lungs. Mediastinal structures cannot be evaluated optimally because contrast material is not given.** As far as can be observed: **Heart contour and size are normal. The widths of the mediastinal main vascular structures are normal. No pleural or pericardial effusion was detected. No pathologically enlarged lymph nodes were detected in the mediastinum and hilar regions. Sliding type hiatal hernia was observed at the lower end of the esophagus. No upper abdominal free fluid-collection or pathologically enlarged lymph nodes were detected in the sections. No fractures or lytic-destructive lesions were detected in the bone structures within the sections. Emphysematous changes in both lungs. Stable millimetric nodules in both lungs. Sequelae changes** and **atelectasis in both lungs.**

Color key: emphysema sequelae/fibrotic changes atelectasis nodules hiatal hernia concordant normal/negative findings

CT-Chat

Findings: **Trachea, both main bronchi are open. Mediastinal main vascular structures, heart contour, size are normal. Thoracic aorta diameter is normal. Pericardial effusion-thickening was not observed.** Thoracic esophagus calibration was normal and no significant tumoral wall thickening was detected. **No enlarged lymph nodes in prevascular, pre-paratracheal, subcarinal or bilateral hilar-axillary pathological dimensions were detected.** When examined in the lung parenchyma window; There are **minimal emphysematous changes in both lungs**. There are minimal bronchiectasis in the central parts of both lungs. There are **millimetric nonspecific nodules in both lungs. Pleural effusion-thickening was not detected. Upper abdominal organs included in the sections are normal.** No space-occupying lesion was detected in the liver that entered the cross-sectional area. Bilateral adrenal glands were normal and no space-occupying lesion was detected. **Bone structures in the study area are natural.** Vertebral corpus heights are preserved. Impression: **Minimal emphysematous changes in both lungs.** Minimal bronchiectasis in the central parts of both lungs. **Millimetric nonspecific nodules in both lungs.**

Ours w/o Graded Verb.

Trachea, both main bronchi are open. Mediastinal main vascular structures, heart contour, size are normal. Thoracic aorta diameter is normal. Pericardial effusion-thickening was not observed. Thoracic esophagus calibration was normal and no significant tumoral wall thickening was detected. **No enlarged lymph nodes in prevascular, pre-paratracheal, subcarinal or bilateral hilar-axillary pathological dimensions were detected.** When examined in the lung parenchyma window; There are **atelectasis in the lower lobes of both lungs**. There are **linear atelectasis in the left lung lower lobe posterobasal and right lung middle lobe medial segment**. There are **millimetric nonspecific nodules in both lungs. Upper abdominal organs included in the sections are normal.** No space-occupying lesion was detected in the liver that entered the cross-sectional area. Bilateral adrenal glands were normal and no space-occupying lesion was detected. **Bone structures in the study area are natural.** Vertebral corpus heights are preserved. Not given.

Ours w/o LoRA

Trachea and both main bronchi are open. No occlusive pathology was detected in the trachea and both main bronchi. There are **emphysematous changes in both lungs**. There are occasionally **linear atelectasis** and minimal ground glass densities in both lungs, especially in the lower lobes. **Millimetric nodules were observed in both lungs**. No mass or appearance compatible with pneumonic infiltration was detected in both lungs. **Mediastinal structures could not be evaluated optimally because no contrast agent was given.** As far as can be observed: **Heart contour and size are normal. The widths of the mediastinal main vascular structures are normal. There are no pathologically enlarged lymph nodes in the mediastinum and hilar regions.** No pathological increase in wall thickness was detected in the esophagus within the sections. **No upper abdominal free fluid-collection or pathologically enlarged lymph nodes were observed in the sections. No lytic-destructive lesions were detected in the bone structures within the sections. Emphysematous changes in both lungs. Millimetric nodules in both lungs. Sequelae changes in both lungs.**

Ours

Trachea, both main bronchi are open. Mediastinal main vascular structures, heart contour, size are normal. Thoracic aorta diameter is normal. Pericardial effusion-thickening was not observed. Thoracic esophagus calibration was normal and no significant pathological wall thickening was detected. **No enlarged lymph nodes in prevascular, pre-paratracheal, subcarinal or bilateral hilar-axillary pathological dimensions were detected.** When examined in the lung parenchyma window; **Emphysematous changes are observed in both lungs. Subpleural minimal reticular densities and subpleural minimal fibrotic densities are observed in the lower lobes of both lungs. Subsegmental atelectasis is observed in the right lung lower lobe posterobasal and left lung upper lobe inferior lingular segment.** There are **several nonspecific nodules in both lungs, the largest of which is 3 mm in the anterior segment of the upper lobe of the right lung**. In the upper abdominal organs included in the sections, there is a hypodense lesion of approximately 40 mm in diameter in the right kidney included in the examination. It could not be characterized in this examination. There are degenerative changes in the vertebrae. Widespread osteophytes extending anteriorly were observed in the vertebrae. **Emphysematous changes in both lungs. Sequelae fibrotic changes in both lungs, atelectasis. Several millimetric nonspecific nodules in both lungs.** Hypodense lesion (cyst?) in the right kidney. Degenerative changes in the vertebrae.

Figure 6: **Progressive qualitative comparison across our pipeline.** Colors indicate corresponding clinical findings between the reference and generated reports, with concordant normal or negative statements highlighted separately. CT-CHAT is shown as an external baseline, followed by the three stages of our method. The progression shows how the later stages progressively recover previously omitted clinical findings and refine their expression in the generated report.

Table 10: **Sensitivity to the LoRA configuration on the CT-RATE validation split.** All rows use slot-wise mixing with graded verbalization and Llama-3.2-3B, and vary one dimension at a time from a common sweep configuration ([targets, r, dropout, learning rate of the base]). The last row is the final model reported in the main text. Results are reported with 95% bootstrap confidence intervals estimated using the percentile method over 1,000 resamples.

	Clinical Efficacy				NLG			
	CRG	F_1	Precision	Recall	BLEU-1	BLEU-4	ROUGE-L	METEOR
w/o LoRA	0.474 [0.469, 0.479]	0.519 [0.507, 0.529]	0.600 [0.587, 0.613]	0.472 [0.460, 0.483]	0.438 [0.433, 0.444]	0.217 [0.211, 0.223]	0.333 [0.327, 0.339]	0.227
Target projections								
All	0.511 [0.506, 0.516]	0.578 [0.568, 0.588]	0.589 [0.577, 0.600]	0.580 [0.568, 0.592]	0.479 [0.474, 0.484]	0.232 [0.226, 0.237]	0.335 [0.329, 0.340]	0.243
$\{q, k, v, o\}$	0.501 [0.496, 0.507]	0.566 [0.555, 0.576]	0.599 [0.586, 0.610]	0.551 [0.539, 0.563]	0.465 [0.460, 0.471]	0.227 [0.222, 0.233]	0.335 [0.329, 0.340]	0.238
{gate,up,down}	0.499 [0.494, 0.504]	0.560 [0.549, 0.570]	0.589 [0.577, 0.602]	0.547 [0.535, 0.559]	0.468 [0.463, 0.473]	0.227 [0.221, 0.233]	0.332 [0.326, 0.338]	0.238
$\{q, v$,gate,up,down$\}$	0.505 [0.500, 0.510]	0.566 [0.556, 0.575]	0.590 [0.577, 0.602]	0.555 [0.543, 0.566]	0.465 [0.460, 0.471]	0.219 [0.213, 0.225]	0.323 [0.317, 0.328]	0.236
Rank								
$r = 8$	0.502 [0.497, 0.507]	0.567 [0.555, 0.576]	0.599 [0.586, 0.611]	0.549 [0.537, 0.561]	0.470 [0.464, 0.475]	0.227 [0.221, 0.233]	0.331 [0.325, 0.337]	0.239
$r = 32$	0.508 [0.503, 0.513]	0.571 [0.561, 0.581]	0.594 [0.581, 0.606]	0.564 [0.552, 0.576]	0.474 [0.469, 0.480]	0.232 [0.226, 0.238]	0.336 [0.330, 0.342]	0.241
$r = 64$	0.508 [0.503, 0.514]	0.574 [0.564, 0.584]	0.591 [0.579, 0.603]	0.572 [0.560, 0.583]	0.482 [0.477, 0.488]	0.235 [0.229, 0.241]	0.335 [0.329, 0.341]	0.244
Dropout								
0.00	0.500 [0.494, 0.505]	0.562 [0.552, 0.572]	0.595 [0.582, 0.607]	0.547 [0.535, 0.558]	0.473 [0.468, 0.478]	0.230 [0.224, 0.236]	0.334 [0.328, 0.340]	0.239
0.10	0.511 [0.505, 0.516]	0.580 [0.570, 0.590]	0.594 [0.582, 0.606]	0.579 [0.566, 0.590]	0.479 [0.474, 0.484]	0.233 [0.227, 0.239]	0.336 [0.331, 0.342]	0.244
0.15	0.511 [0.505, 0.516]	0.581 [0.570, 0.591]	0.594 [0.580, 0.605]	0.581 [0.569, 0.592]	0.478 [0.473, 0.483]	0.232 [0.226, 0.238]	0.336 [0.330, 0.342]	0.243
Learning rate								
$1 \cdot 10^{-5}$	0.497 [0.491, 0.502]	0.557 [0.545, 0.567]	0.594 [0.581, 0.607]	0.536 [0.524, 0.547]	0.471 [0.465, 0.476]	0.228 [0.222, 0.234]	0.333 [0.327, 0.339]	0.239
$5 \cdot 10^{-5}$	0.506 [0.501, 0.512]	0.573 [0.563, 0.583]	0.595 [0.583, 0.607]	0.569 [0.559, 0.582]	0.473 [0.468, 0.479]	0.233 [0.227, 0.239]	0.336 [0.330, 0.342]	0.241
$1 \cdot 10^{-4}$	0.526 [0.521, 0.531]	0.597 [0.587, 0.607]	0.594 [0.582, 0.605]	0.616 [0.604, 0.627]	0.487 [0.482, 0.492]	0.242 [0.236, 0.248]	0.343 [0.337, 0.349]	0.249

effect, increasing F_1 from 0.557 to 0.597 and recall from 0.536 to 0.616 between 10^{-5} and 10^{-4}. Precision remains tightly concentrated between 0.589 and 0.599 across all LoRA configurations. These results indicate that the third-stage gain is not primarily limited by adapter rank or placement; its variation is driven mainly by how strongly the language model is optimized to exploit the structured findings.

G PER-CLASS CLINICAL EFFICACY.

Tables 11 and 12 report the complete class-wise clinical results on CT-RATE and RAD-ChestCT, respectively.

Table 11: **Per-class clinical efficacy on CT-RATE.** Results are reported on the official validation split for all 18 abnormalities.

Finding	Precision	Recall	F_1	Accuracy
Medical material	0.6333	0.3642	0.4625	0.9128
Arterial wall calcification	0.7596	0.7947	0.7768	0.8697
Cardiomegaly	0.6127	0.6523	0.6319	0.9187
Pericardial effusion	0.5913	0.5442	0.5668	0.9381
Coronary artery wall calcification	0.7355	0.8105	0.7711	0.8789
Hiatal hernia	0.5102	0.5971	0.5503	0.8661
Lymphadenopathy	0.5604	0.5995	0.5793	0.7739
Emphysema	0.5007	0.6383	0.5612	0.8029
Atelectasis	0.5507	0.6550	0.5983	0.7937
Lung nodule	0.5489	0.8244	0.6590	0.6180
Lung opacity	0.7982	0.5946	0.6815	0.7835
Pulmonary fibrotic sequela	0.4796	0.5812	0.5256	0.7131
Pleural effusion	0.8347	0.8059	0.8200	0.9562
Mosaic attenuation pattern	0.5265	0.4704	0.4969	0.9207
Peribronchial thickening	0.4203	0.4901	0.4525	0.8615
Consolidation	0.6842	0.6265	0.6541	0.8733
Bronchiectasis	0.5336	0.4333	0.4783	0.8973
Interlobular septal thickening	0.4070	0.6064	0.4871	0.8954

Table 12: **Per-class clinical efficacy on RAD-ChestCT.** Results are reported for the 16 abnormalities retained after mapping the RAD-ChestCT label space to CT-RATE.

Finding	Precision	Recall	F_1	Accuracy
Medical material	0.6415	0.4369	0.5198	0.7358
Arterial wall calcification	0.8257	0.8411	0.8333	0.7625
Cardiomegaly	0.5293	0.5051	0.5169	0.8975
Pericardial effusion	0.6525	0.4103	0.5038	0.8747
Hiatal hernia	0.3926	0.6235	0.4818	0.8430
Lymphadenopathy	0.2435	0.8161	0.3751	0.5521
Emphysema	0.3649	0.9677	0.5299	0.5309
Atelectasis	0.4435	0.6932	0.5409	0.6493
Lung nodule	0.8372	0.9210	0.8771	0.7931
Lung opacity	0.7281	0.2506	0.3729	0.5460
Pulmonary fibrotic sequela	0.1102	0.5272	0.1823	0.3771
Pleural effusion	0.8124	0.7565	0.7835	0.9163
Peribronchial thickening	0.1673	0.4448	0.2431	0.7650
Consolidation	0.4915	0.4020	0.4423	0.8590
Bronchiectasis	0.3223	0.6087	0.4214	0.7353
Interlobular septal thickening	0.2269	0.7421	0.3476	0.8066