BEYOND THE EMBEDDING BOTTLENECK: ADAPTIVE RETRIEVAL-AUGMENTED 3D CT REPORT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Automated radiology report generation from 3D CT volumes often suffers from incomplete abnormality coverage. We provide empirical evidence that this limitation stems from a representational bottleneck: contrastive 3D CT embeddings encode discriminative abnormality signals, yet exhibit dimensional concentration, with CT-CLIP concentrating 90% of its variance in only 2 of 512 principal components. Corroborating this, scaling the language model yields no measurable improvement, suggesting that the bottleneck lies in the visual representation rather than the generator. This bottleneck limits both generation and retrieval. Image-based similarity often fails to isolate fine-grained abnormalities. We propose **AdaRAG-CT**, an adaptive augmentation framework that compensates for this visual bottleneck by introducing supplementary textual information through controlled retrieval and selectively integrating it during generation. On the CT-RATE benchmark, AdaRAG-CT reaches a Clinical F1 of 0.480. Ablations show that relevant retrieved context improves clinical efficacy, while adaptive triggering balances clinical and language quality. Code is available at `https://anonymous.4open.science/r/AdaRAG-CT-anonymous-5756`.

1 INTRODUCTION

Three-dimensional computed tomography (CT) is central to thoracic diagnosis, yet translating CT volumes into structured radiology reports remains labor-intensive. Recent work has explored vision-language models for this task. CT-CHAT (Hamamci et al., 2026) combines a contrastive 3D vision encoder with a large language model. CT-Agent (Mao et al., 2025) incorporates anatomy-aware retrieval and chain-of-thought reasoning. BTB3D (Hamamci et al., 2025) replaces contrastive encoders with reconstruction-based tokenization. The best reported Clinical F1 on CT-RATE remains below clinically acceptable levels, indicating that many abnormalities remain unreported.

To understand this limitation, we analyze the representation space of contrastive 3D CT embeddings. Linear probes (Alain & Bengio, 2016) achieve AUC scores between 0.59 and 0.97 across 18 abnormalities, confirming that discriminative signals are present. However, principal component analysis shows that most variance is concentrated in only a few dimensions (Jing et al., 2022; Aghajanyan et al., 2021), resulting in extremely low effective dimensionality. Organ-level encoders exhibit similar behavior. These embeddings can therefore distinguish coarse pathological states, but they lack the capacity to represent fine-grained report semantics. Scaling the language model from 8B to 70B parameters yields no measurable improvement (Tong et al., 2024; Hamamci et al., 2025), indicating that the bottleneck lies in the visual representation rather than the decoder.

When the visual channel cannot carry sufficient information, additional information must be introduced through another modality. Retrieval-augmented generation (Lewis et al., 2020; Gao et al., 2024; Fink et al., 2025) offers a mechanism for such supplementation: retrieving reference sentences from similar cases and injecting them as context supplies the generator with details not encoded in the visual embedding. However, the reliability of this textual channel depends on the same representation that constrains generation. If the embedding lacks capacity to separate fine-grained abnormalities, image-based retrieval cannot reliably isolate relevant cases. Image-based retrieval has limited abnormality-level precision, while overly frequent context injection can degrade generation.

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.

An empirical analysis across four 3D contrastive encoders confirms this link: abnormality-level retrieval precision remains low, particularly for organs with diverse abnormalities. As a result, image-based retrieval struggles to isolate relevant abnormalities when several co-occur within the same organ. Rather than replacing the encoder, as in concurrent reconstruction-based approaches (Hamamci et al., 2025), we retain existing contrastive models and design a controlled augmentation strategy that maximizes the utility of the textual channel while mitigating the misalignment inherited from the visual representation.

Our framework, AdaRAG-CT, introduces supplementary textual context through controlled retrieval and integrates it selectively during report generation. At the retrieval stage, we explore strategies that improve abnormality-level relevance beyond naive image-based retrieval. At the decoding stage, inspired by Self-RAG (Asai et al., 2024), a learned `[RAG]` token allows the model to decide when to retrieve and condition on the returned context, trained via an oracle-mixed regime that exposes the model to both ground-truth and imperfect retrievals.

This work makes three contributions. First, we diagnose a representational bottleneck in 3D medical contrastive embeddings: a PCA analysis across four encoders reveals effective dimensionalities of only a few dimensions, far below natural-image CLIP, providing empirical evidence that the visual channel alone is insufficient in our experiments. Second, we propose AdaRAG-CT, an adaptive augmentation framework that compensates for this bottleneck by introducing and controlling supplementary textual information: retrieval strategies improve abnormality-level relevance, while a learned adaptive mechanism enables the generator to selectively utilize retrieved context. Third, with adaptive retrieval augmentation, our model improves clinical efficacy on CT-RATE over the base model. Ablations show that retrieved context provides additional clinical evidence, while adaptive triggering controls when this context is used.

2 Related Work

3D Medical Report Generation. CT2Rep (Hamamci et al., 2024) introduced the first dedicated 3D CT report-generation framework using an autoregressive transformer with relational memory. Subsequent work paired 3D CT encoders with LLMs (CT-CHAT (Hamamci et al., 2026), Med3DVLM (Xin et al., 2026)), added anatomy-specific LoRA and exemplar retrieval (CT-Agent (Mao et al., 2025)), or explored multi-modal fusion (Luo et al., 2025; Lee et al., 2024) and memory-driven evidence aggregation (Chen et al., 2020; Wang et al., 2023). Several works directly recognise the visual representation as a bottleneck and propose alternative tokenization or aggregation schemes (BTB3D (Hamamci et al., 2025), SAMF (Hosseini et al., 2025)), or pursue broader 3D understanding via multi-task pre-training (COLIPRI (Wald et al., 2026), M3D (Bai et al., 2024), RadFM (Wu et al., 2025)). Our work complements this line by providing a quantitative diagnosis of *why* existing contrastive embeddings limit retrieval quality, namely that their effective dimensionality is strongly concentrated and poorly aligned with fine-grained abnormality semantics, and proposing an adaptive RAG framework that compensates through a textual channel rather than modifying the encoder.

Retrieval-Augmented Generation in Medical Imaging. RAG (Gao et al., 2024) grounds generative models in retrieved evidence, and adaptive variants decide when to retrieve: FLARE (Jiang et al., 2023) generates a tentative continuation, detects low-confidence tokens, backtracks, and regenerates with retrieved context. Self-RAG (Asai et al., 2024) trains reflection tokens to critique and select among retrieval-augmented candidates. In medical imaging, retrieval-based approaches have been explored for 2D chest X-ray report generation (Fink et al., 2025; Song et al., 2024), typically retrieving full reports or report sections as templates or auxiliary inputs, and generally assume that retrieval quality is sufficient. Our analysis shows that this assumption breaks down for 3D CT. Unlike FLARE, which triggers retrieval from low-confidence tokens, and Self-RAG, which critiques retrieval-augmented candidates, our model learns to emit a dedicated `[RAG]` token that decides when to retrieve.

Contrastive Representation Quality. Contrastive image–text learning (Radford et al., 2021) is known to exhibit failure modes in its representation spaces: dimensional collapse, where embeddings occupy a low-dimensional subspace (Jing et al., 2022); modality gap, where image and text embeddings cluster separately (Liang et al., 2022); and feature suppression, where task-

irrelevant variance dominates learned features (Zhang et al., 2025). In the medical domain, CT-CLIP (Hamamci et al., 2026) and ViSD-Boost (Cao et al., 2025) have trained 3D CT encoders with contrastive objectives on CT-RATE, but the retrieval-level consequences of their representation geometry have received little attention. Our analysis complements these works by showing that the dominant variance directions in 3D medical contrastive embeddings fail to encode abnormality semantics, an effect we term *variance-semantic misalignment*, and that this property limits downstream RAG pipelines.

Table 1: Effective dimensionality and train/validation stability of 3D CT embeddings. Gray brackets denote 95% intervals. $\dim_{90}$/$\dim_{95}$ count components explaining 90%/95% of variance; PR is the participation ratio.

Model	Embedding	Split	Total dim	$\dim_{90}$	$\dim_{95}$	PR
CLIP ViT-B/32 (ImageNet)	image emb.	reference	512	243.0	322.0	64.40
CT-CLIP (Zero-shot)	avg pool	train	512	2.0 [2.0,2.0]	4.0 [4.0,4.0]	1.43 [1.41,1.45]
	avg pool	valid	512	2.0 [2.0,2.0]	4.0 [4.0,4.0]	1.43 [1.41,1.45]
	max pool	train	512	109.8 [108.0,112.0]	181.3 [180.0,183.0]	6.70 [6.47,6.93]
	max pool	valid	512	103.6 [102.0,105.0]	171.1 [169.0,173.0]	6.65 [6.38,6.86]
CT-CLIP (VocabFine)	avg pool	train	512	1.0 [1.0,1.0]	2.0 [2.0,2.0]	1.16 [1.15,1.17]
	avg pool	valid	512	1.0 [1.0,1.0]	2.0 [2.0,2.0]	1.17 [1.16,1.18]
	max pool	train	512	2.5 [2.0,3.0]	40.7 [36.0,46.0]	1.26 [1.25,1.27]
	max pool	valid	512	2.8 [2.0,3.0]	40.4 [35.0,45.0]	1.26 [1.25,1.28]
ViSD-Boost	lung	train	256	9.3 [9.0,10.0]	11.6 [11.0,12.0]	6.30 [6.12,6.45]
	lung	valid	256	9.1 [9.0,10.0]	11.6 [11.0,12.0]	6.26 [6.09,6.45]
	heart	train	256	5.0 [5.0,5.0]	6.0 [6.0,6.0]	3.63 [3.53,3.73]
	heart	valid	256	5.0 [5.0,5.0]	6.0 [6.0,6.0]	3.62 [3.52,3.72]
FVLM	lung	train	256	6.0 [6.0,6.0]	8.0 [8.0,8.0]	4.94 [4.84,5.03]
	lung	valid	256	6.0 [6.0,6.0]	7.6 [7.0,8.0]	5.04 [4.93,5.14]
	heart	train	256	4.0 [4.0,4.0]	6.0 [6.0,6.0]	2.59 [2.53,2.65]
	heart	valid	256	4.0 [4.0,4.0]	6.0 [6.0,6.0]	2.56 [2.50,2.63]

3 Representational Bottleneck in 3D CT Embeddings

We first diagnose the representational bottleneck of 3D CT contrastive embeddings and its consequences for retrieval. The analysis uses the 46,407 training and 2,987 validation CT-RATE volumes (Hamamci et al., 2026), with paired reports and 18 binary abnormality labels, and three families of pre-trained 3D contrastive encoders: CT-CLIP (Hamamci et al., 2026) and its fine-tuned variant CT-CLIP (VocabFine), FVLM (Shui et al., 2025), and ViSD-Boost (Cao et al., 2025).

3.1 Encoded but Narrow

Following Alain & Bengio (2016), we train linear probes (logistic regression) on frozen representations to test whether abnormality information is linearly accessible. Across the three encoders, probe AUC ranges from 0.59 to 0.97 for average-pooled embeddings across 18 abnormalities (Table 2), confirming that the contrastive objective successfully encodes diagnostic signals. However, high AUC does not imply rich information content. A principal component analysis (Jing et al., 2022; Aghajanyan et al., 2021) of the embedding spaces reveals that the effective dimensionality is remarkably low (Table 1). For each train/validation split, we report means and 95% intervals over 200 bootstrap resamples of 2,987 volumes. CT-CLIP with average pooling concentrates 90% of its variance in only 2 principal components ($\dim_{90} = 2$, PR = 1.4). Although max pooling inflates $\dim_{90}$ to about 110, the participation ratio remains only 6.7, indicating that most of this spread is long-tail noise rather than signal. Consistent with this, zero-shot CT-CLIP's macro AUC is lower for max pooling than average pooling (0.669 vs. 0.739; see Table 2). Organ-level embeddings from ViSD-Boost and FVLM fare similarly ($\dim_{90}$ is approximately 4–9 out of 256). A probe on the top 2 principal components of zero-shot CT-CLIP reaches a mean AUC of only 0.552, against 0.746 on the lower half. An embedding can achieve high binary AUC (e.g., "pleural effusion present vs. absent") using just one or two directions, yet may not reliably preserve the fine-grained semantic detail required for report generation, such as laterality, severity, and co-occurring abnormalities. This distinction between discriminability (high AUC) and information richness (low effective dimensionality) is central to our argument: the bottleneck is not that the LLM fails to extract available information, but that *the visual channel is too limited* to begin with. A natural-image baseline confirms

Table 2: Linear probe AUC on frozen contrastive embeddings.*

	CT-CLIP (Zero-shot)		CT-CLIP (VocabFine)			
Abnormality	avg	max	avg	max	FVLM	ViSD-Boost
Lung (11)						
Lung nodule	0.619	0.582	0.588	0.516	0.664	0.713
Mosaic attenuation	0.794	0.710	0.711	0.578	0.887	0.892
Peribronchial thick.	0.717	0.631	0.730	0.645	0.777	0.814
Consolidation	0.732	0.659	0.630	0.543	0.859	0.920
Bronchiectasis	0.679	0.605	0.624	0.564	0.752	0.790
Interlob. septal thick.	0.762	0.700	0.725	0.583	0.863	0.874
Emphysema	0.712	0.629	0.695	0.641	0.792	0.799
Atelectasis	0.680	0.627	0.667	0.561	0.738	0.797
Lung opacity	0.653	0.610	0.634	0.551	0.805	0.871
Pulm. fibrotic seq.	0.621	0.582	0.614	0.542	0.688	0.744
Pleural effusion	0.911	0.849	0.850	0.718	0.956	0.966
Heart (2)						
Cardiomegaly	0.891	0.794	0.836	0.666	0.943	0.952
Pericardial effusion	0.797	0.687	0.764	0.685	0.851	0.910
Aorta (2)						
Arterial wall calc.	0.818	0.734	0.808	0.676	0.898	0.913
Coronary artery calc.	0.812	0.737	0.797	0.664	0.879	0.903
Esophagus (1)						
Hiatal hernia	0.701	0.650	0.665	0.573	0.758	0.821
Other (2)						
Lymphadenopathy	0.688	0.621	0.648	0.586	—	—
Medical material	0.726	0.642	0.688	0.590	—	—
Macro avg	**0.739**	**0.669**	**0.704**	**0.604**	**0.823**	**0.853**

* "—": abnormality outside the organ encoder's scope.

the severity of this concentration: CLIP ViT-B/32 on ImageNet achieves $\dim_{90} = 243$ and PR = 64.4 on the same 512-dimensional space, about 45× higher in PR than CT-CLIP (Table 1). Two factors likely explain the gap. First, natural images span visually diverse categories, encouraging the encoder to use many dimensions. Chest CT volumes are structurally homogeneous, with pathological differences confined to subtle local variations, so a contrastive objective can separate most cases with very few directions. Second, existing 3D CT encoders are trained with a global contrastive loss that pairs an entire volume with an entire report, a granularity too coarse to incentivize fine-grained, abnormality-level representations. A scaling experiment corroborates this diagnosis: replacing the 8B LLM with a 70B variant yields Clinical F1 of 0.405 versus 0.455, showing no improvement despite an 8.75× increase in generator capacity. If the bottleneck were in the decoder, a larger model should improve performance. The absence of any gain confirms that the visual channel is the limiting factor. This observation is consistent with prior work: CT-CHAT reports near-identical metrics at 8B and 70B scales (Hamamci et al., 2026), and BTB3D similarly concludes that improved visual tokenization, rather than larger language backbones, is essential for 3D medical VLMs (Hamamci et al., 2025).

Table 3: Organ-level retrieval precision (Jaccard@10) with ViSD-Boost image embeddings and its text encoder on organ paragraphs. Upper Bound ranks by ground-truth label overlap.

Organ	Img2Img	Img2Txt	Txt2Txt	Upper Bound
Lung	0.351	0.313	0.563	0.992
Heart	0.825	0.638	0.925	1.000
Esophagus	0.796	0.617	0.935	1.000
Aorta	0.741	0.617	0.666	1.000

3.2 Retrieval Under the Bottleneck

The dimensional concentration identified above has important consequences for retrieval. When most variance concentrates in a few dimensions, cosine similarity is dominated by these directions, collapsing fine-grained abnormality distinctions. We measure organ-level Jaccard@10, the intersection over union of abnormality label sets between the query and each of the top-10 cosine-nearest neighbours averaged over queries, and compare image-to-image, image-to-text (cross-modal), and text-to-text retrieval against an upper bound constructed from ground-truth abnormality label overlap (Table 3). Lung retrieval is weakest (Jaccard@10 = 0.351): with 11 distinct abnormalities, the

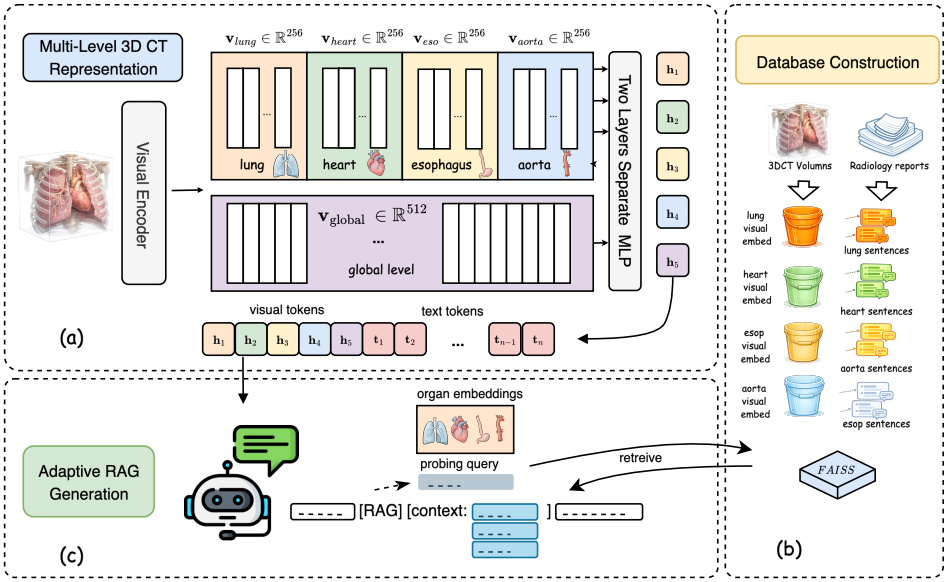


Figure 1: Overview of AdaRAG-CT. (a) A 3D CT volume is encoded into a global embedding (CT-CLIP) and four organ-specific embeddings (ViSD-Boost), which are projected into the LLM input space as visual tokens. (b) An organ-indexed sentence database is built from training-set reports and indexed with FAISS for efficient retrieval. (c) During report generation, the LLM autonomously emits a `[RAG]` token when it needs external evidence. Retrieved sentences are then injected into the context before the model continues generating.

Jaccard denominator grows quickly, requiring precise alignment across a combinatorially large label space. The upper bound of 0.992 for lung confirms that the database contains relevant matches. The bottleneck lies in the representation, not the coverage. By contrast, heart (2 abnormalities) and esophagus (1 abnormality) achieve much higher overlap simply because fewer labels need to agree. The aorta is an exception (Img2Img 0.741 $>$ Txt2Txt 0.666): its pathological vocabulary (arterial and coronary calcification) overlaps substantially with the heart section, reducing text-to-text discriminability, whereas the organ-specific image embeddings remain well-separated by construction.

A clear asymmetry emerges across modalities. Text-to-text cosine similarity outperforms image-to-image retrieval for lung (0.563 vs. 0.351), heart (0.925 vs. 0.825), and esophagus (0.935 vs. 0.796), indicating that the text encoder preserves abnormality semantics more effectively than the image encoder. Cross-modal retrieval (Img2Txt) is worse than image-to-image for all organs, indicating that direct cross-modal matching is unreliable in this setting. The gap between Txt2Txt and Img2Img suggests that the bottleneck originates in the image representation, not in cosine similarity per se. Conventional medical RAG systems query a text database with the image embedding (Img2Txt), which is directly constrained by this bottleneck. Text-to-text retrieval offers a way to reduce this dependence, suggesting that retrieval strategies using partial text may be more reliable than purely image-based approaches.

4 Method: AdaRAG-CT

AdaRAG-CT operates on top of a base vision-language model and augments it with adaptive retrieval-augmented generation.

4.1 Base Model

We adopt a LLaVA-style architecture (Liu et al., 2023) where a pre-trained LLM (Llama-3.1-8B-Instruct and Llama-3.3-70B-Instruct (Grattafiori et al., 2024)) receives visual input through learned projectors. Each 3D CT volume is represented by $K = 5$ embeddings: a global max-pooled embedding $\mathbf{v}_{\text{global}} \in \mathbb{R}^{512}$ from CT-CLIP and four organ-specific embeddings $\mathbf{v}_o \in \mathbb{R}^{256}$ ($o \in \{\text{lung}, \text{heart}, \text{esophagus}, \text{aorta}\}$) from ViSD-Boost. Each embedding is projected indepen-

dently as

$$\mathbf{h}_o = \text{MLP}_o(\mathbf{v}_o), \quad \mathbf{h}_o \in \mathbb{R}^{d_{\text{LLM}}}, \tag{1}$$

where MLP_o is a 2-layer projector specific to organ o. The resulting visual tokens $\{\mathbf{h}_o\}_{o=1}^{K}$ are incorporated into the input token sequence as visual context for the LLM. This base model, without any retrieval augmentation, serves as our primary baseline. This strong base is the result of design experiments on embedding combinations, organ-specific projectors, and two-stage LoRA training.

4.2 Organ-Indexed Sentence Database

We construct an organ-indexed sentence database from all training reports. Each report is first split into individual sentences. Every sentence is then annotated with its corresponding organ label (lung, heart, esophagus, aorta) and associated abnormalities, yielding approximately 572K sentences in total across both Findings and Impression sections. To represent each sentence, we employ a text encoder based on BiomedVLP-CXR-BERT (Boecking et al., 2022), followed by a learned 256-dimensional projection layer. The text encoder and projection layer are fine-tuned using a contrastive image–text objective, with frozen organ-level image embeddings serving as anchors. All sentence embeddings are stored in per-organ FAISS indices to enable efficient nearest-neighbour retrieval. Organ-level image embeddings from the training set are indexed for coarse candidate retrieval.

4.3 Adaptive Retrieval Training

Following Self-RAG (Asai et al., 2024), we introduce a special `[RAG]` token into the vocabulary. At inference, rather than injecting context at predetermined positions, the model generates `[RAG]` to request external context. During training, we run the base model on ground-truth reports and compute per-sentence perplexity. Sentences above the 70th percentile ($\text{ppl} \approx 1.935$) are selected as retrieval targets, with at most four targets per report. Intuitively, these high-perplexity sentences are the parts that the base generator finds difficult under visual conditioning alone, and therefore mark positions where additional evidence is likely to help. This threshold is used only to construct training supervision. For each target, we precompute oracle context using the ground-truth target sentence as a Text2Text query, and real context using image-based coarse filtering followed by text re-ranking. Final candidates are selected using Maximal Marginal Relevance (MMR) (Carbonell & Goldstein, 1998):

$$\text{MMR}(d_i) = \lambda \cdot \text{sim}(d_i) - (1-\lambda) \cdot \max_{d_j \in \mathcal{S}} \text{BLEU-2}(d_i, d_j), \tag{2}$$

where $\text{sim}(d_i)$ is the cosine similarity score and $\mathcal{S}$ is the set of already-selected sentences. Each training report receives oracle context with probability 0.7 and real context with probability 0.3, for both model sizes. This oracle-mixed regime exposes the model to high-quality signals while building robustness to imperfect retrieval. Oracle context is used only during training. Injected context tokens are masked from the generation loss.

4.4 Adaptive Retrieval Inference

During autoregressive generation, whenever the model emits `[RAG]`, it continues to generate the next sentence, which is then used as the retrieval query. The retrieved sentences are formatted and injected into the context, the initial generation is rolled back, and the model regenerates the sentence with the retrieved context. We explore two retrieval strategies for supplying candidate sentences. As shown in Fig. 2, both yield comparable generation quality, indicating that performance is robust to the choice of retrieval pipeline. Both strategies use MMR (Eq. (2)) with $K_{\text{fine}} = 3$.

Two-Stage retrieval. Given the organ currently being described, we first retrieve the $K_{\text{coarse}} = 20$ most similar training samples using image embedding cosine similarity, collecting all their sentences for the same organ as a candidate pool. This vision-guided stage narrows the search from the full database to patient-relevant candidates. Candidates are then re-ranked and selected via MMR (Eq. (2)) against the current query sentence.

Text2Text retrieval. We retrieve directly from the full per-organ sentence database using text-to-text cosine similarity, with the model's partial generation as query. Because this query already reflects the organ and tentative abnormality being described, Text2Text retrieval is less dependent on the low-dimensional image embedding. It is still conditioned on the generated report context, and therefore complements rather than replaces the visual input.

5 EXPERIMENTS

5.1 SETUP

Dataset. All experiments use the official CT-RATE (Hamamci et al., 2026) training and validation splits. After excluding CT volumes without pre-computed organ embeddings, we use 46,407 of 47,149 training volumes and 2,987 of 3,039 validation volumes. Each volume is paired with a radiology report, which we concatenate its Findings and Impression sections as the generation target.

Metrics. Following the CT-RATE evaluation protocol (Hamamci et al., 2026), we adopt Clinical Efficacy as the primary metric. A RadBERT-based classifier (Yan et al., 2022) extracts 18 binary abnormality labels (e.g., lung nodule, pleural effusion, cardiomegaly) from each generated report. These predictions are compared against the dataset-provided ground-truth annotations to compute per-abnormality Precision, Recall, and F1. We report the support-weighted average as Clinical Precision (Clin-P), Clinical Recall (Clin-R), and Clinical F1 (Clin-F1). We also report BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee & Lavie, 2005) for textual overlap, and LLaMA Score (Hamamci et al., 2026), a report quality rating by Llama-3.1-70B-Instruct.

Baselines. We compare against prior methods on CT-RATE report generation: CT2Rep (Hamamci et al., 2024), an encoder-decoder with relational memory; Merlin (Blankemeier et al., 2024), a 7B VLM for 3D CT understanding; CT-CHAT (Hamamci et al., 2026), which pairs a CT-ViT encoder with Llama-3.1 (8B/70B) via attention pooling; BTB3D (Hamamci et al., 2025), which uses reconstruction-based visual tokenization; SAMF (Hosseini et al., 2025), a 2D–3D feature fusion approach; and CT-Agent (Mao et al., 2025), an agent-based pipeline with chain-of-thought reasoning. All baseline numbers are taken from the original publications. We also report our own base model (no RAG) at 8B and 70B scales, trained in two stages: organ-specific projectors with frozen LLM weights, followed by LoRA fine-tuning.

5.2 MAIN RESULTS

Table 4: Report generation performance on the CT-RATE validation set. Baseline numbers are from original publications.

		Clinical Efficacy			NLP				
Method	Params	F1	P	R	B-1	B-4	R-L	MET	LLaMA Score
CT2Rep (Hamamci et al., 2024)	—	0.160	0.435	0.128	0.372	0.213	—	0.197	—
Merlin (Blankemeier et al., 2024)	7B	0.160	0.295	0.112	0.231	0.099	—	0.148	—
CT-CHAT (Hamamci et al., 2026)	8B	—	—	—	0.494	—	0.584	0.311	7.440
CT-CHAT (Hamamci et al., 2026)	70B	0.184	0.450	0.158	0.498	—	0.581	0.311	7.429
BTB3D (Hamamci et al., 2025)	8B	0.258	0.260	0.260	0.439	0.213	—	0.223	—
SAMF (Hosseini et al., 2025)	3.8B	—	—	—	0.440	0.261	0.417	0.417	7.165
CT-Agent (Mao et al., 2025)	—	0.420	0.423	0.477	0.502	0.231	0.490	0.425	—
Ours (base)	8B	0.455	0.474	0.469	0.463	0.205	0.315	0.206	7.30
AdaRAG-CT	**8B**	**0.480**	**0.502**	**0.520**	**0.496**	**0.242**	**0.354**	**0.246**	**7.75**
Ours (base)	70B	0.405	0.468	0.373	0.449	0.213	0.334	0.208	7.10
AdaRAG-CT	**70B**	**0.426**	**0.483**	**0.413**	**0.497**	**0.250**	**0.361**	**0.232**	**7.53**

Note: Additional baseline results under our unified evaluation protocol are provided in the Supplementary Material.

Table 4 presents the main comparison. Our 8B base model obtains a Clinical F1 of 0.455 without retrieval. Scaling the LLM to 70B does not improve this figure. In fact, the 70B base model scores only 0.405, consistent with the PCA analysis showing severe dimensional concentration in the visual embeddings (Section 3.1). Both models read the same impoverished visual representations, so the larger decoder has no additional signal to exploit. The bottleneck is representational, not generative. Retrieval augmentation benefits both scales: the 8B model improves from 0.455 to 0.480 (+0.025)

and the 70B model from 0.405 to 0.426 (+0.021), confirming that the textual channel supplies information absent from the visual input. The 8B AdaRAG-CT (0.480) surpasses CT-Agent (0.420) and substantially outperforms CT-CHAT 70B (0.184) in Clinical F1, while also improving NLP metrics and LLaMA Score over our 8B base model.

5.3 Retrieval Pipeline Comparison

Figure 2 shows that both pipelines yield stable performance throughout training, with Clinical F1 virtually identical at every step and bootstrap confidence intervals consistently overlapping. A closer look at the precision–recall decomposition reveals a mild trade-off: Two-Stage retrieval, which first narrows candidates via image similarity before text re-ranking, tends to surface a broader set of organ-level abnormalities and thus favours recall. Text2Text retrieval, which matches directly against the model's partial generation, returns more topically precise context and favours precision. These complementary biases cancel out in F1, showing that AdaRAG-CT remains effective across both retrieval pipelines. This also explains why neither pipeline should be interpreted as strictly superior: they expose different useful contexts to the same generator, and the adaptive trigger determines when those contexts enter the report. This robustness is consistent with the bottleneck hypothesis: clinically relevant textual context can compensate for the limited visual signal under either retrieval design.

5.4 Context Utilization Ablation

Table 5: Context utilization ablation (8B model). All variants share the same retrieval pipeline. Δ is relative to the retrieval-disabled setting.

Utilization Strategy	Clin-F1	Δ	BLEU-4	ROUGE-L
Retrieval disabled	0.462	—	0.228	0.349
Fixed-interval (N=3)	0.453	−0.009	0.192	0.307
Fixed-interval (N=5)	**0.494**	**+0.032**	0.209	0.329
Fixed-interval (N=7)	0.482	+0.020	0.228	0.346
Adaptive RAG (ours)	0.480	+0.018	**0.242**	**0.354**

Table 5 ablates the context injection strategy while holding the retrieval pipeline fixed. The retrieval-disabled setting (0.462) reflects the RAG-trained model's own generation capability without retrieval context at inference, which differs slightly from the base model in Table 4 (0.455) that was trained without the RAG objective. Fixed-interval injection is sensitive to the choice of N: injecting too frequently (N=3) hurts performance below baseline, while N=5 maximises Clinical F1 (0.494) at the expense of NLP quality, and N=7 strikes a middle ground. This sensitivity to N highlights a practical limitation of fixed-interval strategies. The optimal frequency is dataset-dependent and requires

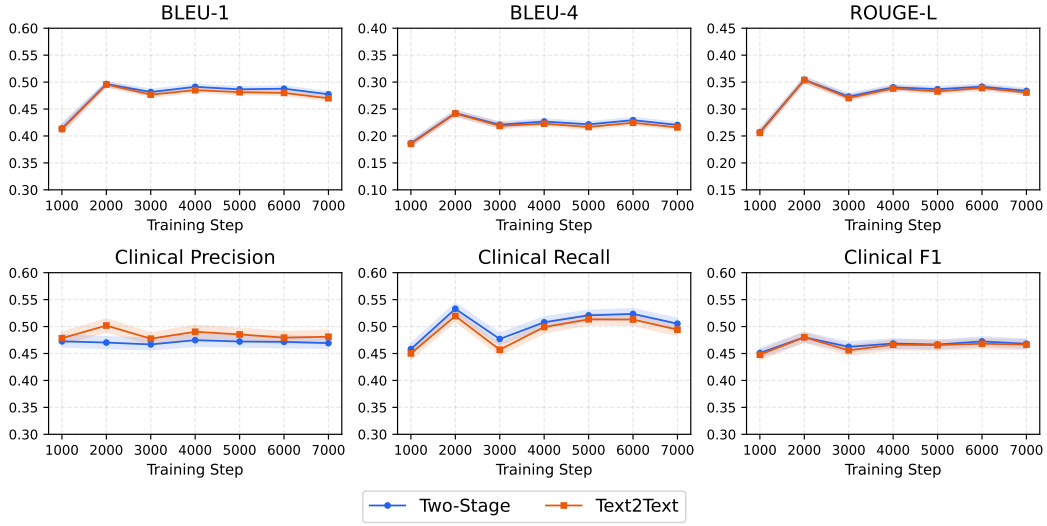


Figure 2: Six evaluation metrics across training steps for Two-Stage and Text2Text retrieval pipelines on the 8B model, with bootstrap 95% confidence intervals.

separate tuning. Adaptive RAG avoids fixed-interval tuning by triggering retrieval selectively: it achieves competitive Clinical F1 (0.480, +0.018) without any frequency hyperparameter, and consistently outperforms all fixed-interval variants on BLEU-4 and ROUGE-L. On average, the model emits 1.48 `[RAG]` triggers per report, confirming that retrieval is used sparingly and on demand rather than applied uniformly.

6 DISCUSSION

Representational bottleneck. The PCA and scaling results place the bottleneck in the visual representation. Retrieved text therefore acts as a second channel into the generator, one that can carry laterality, severity and co-occurring abnormalities that the embedding does not encode. This differs from the usual role of RAG, which is to supply external world knowledge. Reshaping the encoder directly did not help. Fine-tuning it with a PCA regulariser gave no stable gain in zero-shot classification or in report generation, which suggests that low effective dimensionality reflects missing lesion-level semantics, and a spectrum penalty alone does not supply them.

Adaptive triggering. The retrieval module opens the textual channel and the `[RAG]` token decides when it is used. Fixed-interval injection depends on its frequency: $N = 5$ reaches a higher Clinical F1 than adaptive triggering (0.494 against 0.480) but lower BLEU-4 and ROUGE-L, and N must be tuned per dataset. Oracle-mixed training is what makes the trigger usable. Seeing both oracle and real context teaches the generator to use retrieved evidence where it asks for it and to generate normally elsewhere. The mixing probability p_{oracle} is itself a hyperparameter whose best value may depend on the dataset. We also expect a 70B generator, with stronger language priors, to produce more confident hallucinations from the same visual input, which case-relevant evidence counters.

Retrieved content. The database is built from the training reports and adds no external medical knowledge. Because its sentences share the generator's training distribution, our results cannot fully separate the value of retrieval from the proximity of the retrieved cases to the training population. Many sentences are templates such as "No evidence of pleural effusion", which can fire the trigger without adding case-specific evidence and pull the output toward common phrasing. A median split on overlap with the database is consistent with this. High-overlap reports score higher on BLEU-4 and ROUGE-L (0.246 and 0.363 against 0.238 and 0.345) and lower on Clinical F1 (0.475 against 0.482), so template reuse mainly changes the wording.

Scope. All experiments use CT-RATE, a single-institution set of non-contrast chest CT, and other datasets and anatomical regions remain untested. Clinical F1 scores the presence of 18 labels. It is blind to size, location, laterality and severity, to temporal comparison and to abnormalities outside the 18 labels, so a report that flags a lung nodule without its size scores the same as one that gives it. The evaluation is retrospective, and a reader study is needed before clinical use.

7 CONCLUSION

Across 3D CT contrastive embeddings, PCA reveals severe dimensional concentration. CT-CLIP concentrates 90% of its variance in only 2 of 512 principal components. This representational poverty, not the generator's capacity, is the fundamental bottleneck: scaling the LLM from 8B to 70B parameters yields no improvement because both models read the same impoverished visual channel. AdaRAG-CT addresses this bottleneck by opening a supplementary textual channel through adaptive retrieval. A learned `[RAG]` token triggers case-relevant context injection during generation, achieving a Clinical F1 of 0.480 on CT-RATE. Ablations show that clinically relevant retrieved context provides the main clinical gain, while adaptive triggering offers a balanced way to control when this context is introduced. These results suggest that future 3D CT report-generation systems should evaluate not only decoder scale, but also the amount of clinically useful information available to the generator. Resolving the bottleneck at its source, through abnormality-aware visual pre-training or dense per-token visual representations, remains the longer-term solution. Until richer encoders become available, the textual retrieval channel provides an effective surrogate.

AI USE STATEMENT

We used AI tools in several parts of this work: to help implement and run the experiment code, for brainstorming, to assist in finding related work, to draft parts of the paper, and to edit and polish the writing. The authors designed the methodology and the experiments, interpreted the results, and reviewed and verified all AI-assisted code, text, and references. We take full responsibility for the final content.

REFERENCES

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7319–7328, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.568. URL https://aclanthology.org/2021.acl-long.568/.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2016. URL https://arxiv.org/abs/1610.01644.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

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. URL https://arxiv.org/abs/2404.00578.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss (eds.), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909/.

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, et al. Merlin: A vision language foundation model for 3D computed tomography. *Research square*, pp. rs–3, 2024.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. Making the most of text semantics to improve biomedical vision–language processing. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 1–21, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20059-5.

Weiwei Cao, Jianpeng Zhang, Zhongyi Shui, Sinuo Wang, Zeli Chen, Xi Li, Le Lu, Xianghua Ye, Qi Zhang, Tingbo Liang, et al. Boosting vision semantic density with anatomy normality modeling for medical vision-language pre-training, 2025. URL https://arxiv.org/abs/2508.03742.

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pp. 335–336, New York, NY, USA, 1998. Association for Computing Machinery. ISBN 1581130155. doi: 10.1145/290941.291025.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1439–1449. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.112.

Anna Fink, Alexander Rau, Marco Reisert, Fabian Bamberg, and Maximilian F Russe. Retrieval-augmented generation with large language models in radiology: from theory to practice. *Radiology: Artificial Intelligence*, 7(4):e240790, 2025.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024. URL https://arxiv.org/abs/2312.10997.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu

Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. CT2Rep: Automated radiology report generation for 3D medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 476–486. Springer, 2024.

Ibrahim Ethem Hamamci, Sezgin Er, Suprosanna Shit, Hadrien Reynaud, Dong Yang, Pengfei Guo, Marc Edgar, Daguang Xu, Bernhard Kainz, and Bjoern Menze. Better tokens for better 3D: Advancing vision-language modeling in 3D medical imaging. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 135074–135102. Curran Associates, Inc., 2025. doi: 10.52202/085713-4507. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/c4861cc81b768dc36a676fceed8f7672-Paper-Conference.pdf.

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, Weicheng Dai, Murong Xu, Hadrien Reynaud, Muhammed Furkan Dasdelen, Bastian Wittmann, Tamaz Amiranashvili, Enis Simsar, Mehmet Simsar, Emine Bensu Erdemir, Abdullah Alanbay, Anjany Sekuboyina, Berkan Lafci, Ahmet Kaplan, Zhiyong Lu, Malgorzata Polacin, Bernhard Kainz, Christian Bluethgen, Kayhan Batmanghelich, Mehmet Kemal Ozdemir, and Bjoern Menze. Generalist foundation models from a multimodal dataset for 3D computed tomography. *Nature Biomedical Engineering*, 10(8):1610–1628, February 2026. ISSN 2157-846X. doi: 10.1038/s41551-025-01599-y.

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.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.495. URL https://aclanthology.org/2023.emnlp-main.495/.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=YevsQ05DEN7.

Changsun Lee, Sangjoon Park, Cheong-Il Shin, Woo Hee Choi, Hyun Jeong Park, Jeong Eun Lee, and Jong Chul Ye. Read like a radiologist: Efficient vision-language model for 3D medical imaging interpretation, 2024. URL https://arxiv.org/abs/2412.13558.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17612–17625. Curran Associates, Inc., 2022. doi: 10.52202/068431-1280. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/702f4db7543a7432431df588d57bc7c9-Paper-Conference.pdf.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Lingxiao Luo, Bingda Tang, Xuanzhong Chen, Rong Han, and Ting Chen. VividMed: Vision language model with versatile visual grounding for medicine. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1800–1821, Albuquerque, New Mexico, April 2025. Association for Computational

Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.89. URL https://aclanthology.org/2025.naacl-long.89/.

Yuren Mao, Wenyi Xu, Yuyang Qin, and Yunjun Gao. Ct-agent: A multimodal-llm agent for 3d ct radiology question answering, 2025. URL https://arxiv.org/abs/2505.16229.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

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, 2025. URL https://arxiv.org/abs/2501.14548.

Steven Song, Anirudh Subramanyam, Irene Madejski, and Robert L. Grossman. LaB-RAG: Label boosted retrieval augmented generation for radiology report generation, 2024. URL https://arxiv.org/abs/2411.16523.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578, 2024. doi: 10.1109/CVPR52733.2024.00914.

Tassilo Wald, Ibrahim Ethem Hamamci, Yuan Gao, Sam Bond-Taylor, Harshita Sharma, Maximilian Ilse, Cynthia Lo, Olesya Melnichenko, Anton Schwaighofer, Noel C. F. Codella, Maria Teodora Wetscherek, Klaus H. Maier-Hein, Panagiotis Korfiatis, Valentina Salvatelli, Javier Alvarez-Valle, and Fernando Pérez-García. Comprehensive language-image pre-training for 3D medical image understanding, 2026. URL https://arxiv.org/abs/2510.15042.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. METransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11558–11567, 2023. doi: 10.1109/CVPR52729.2023.01112.

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), August 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-62385-7.

Yu Xin, Gorkem Can Ates, Kuang Gong, and Wei Shao. Med3DVLM: An efficient vision-language model for 3D medical image analysis. *IEEE Journal of Biomedical and Health Informatics*, 30 (3):2524–2536, 2026. doi: 10.1109/JBHI.2025.3604595.

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. doi: 10.1148/ryai.210258.

Jihai Zhang, Xiang Lan, Xiaoye Qu, Yu Cheng, Mengling Feng, and Bryan Hooi. Learning the unlearned: Mitigating feature suppression in contrastive learning. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 35–52, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73010-8.